



DIVISIÓN DE ESTUDIOS POSGRADO E INVESTIGACIÓN

Opción 1.- TESIS

TRABAJO PROFESIONAL

**“SISTEMA INTELIGENTE BASADO EN INDUSTRIA 4.0 PARA
MEJORAR LA GESTIÓN DE CONOCIMIENTO DENTRO DE LA
LOGÍSTICA DE SALUD PÚBLICA”**

**QUE PARA OBTENER EL GRADO DE:
DOCTOR EN CIENCIAS
DE LA INGENIERÍA**

PRESENTA:

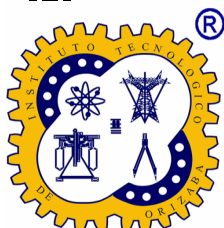
M.I.I. MARCO JULIO DEL MORAL ARGUMEDO

DIRECTOR DE TESIS:

Dr. ALBERTO ALFONSO AGUILAR LASSERRE

CODIRECTOR DE TESIS:

Dr. CARLOS FRANCISCO VÁZQUEZ RODRÍGUEZ



Orizaba, Veracruz, 20/11/2024
Dependencia: **División de Estudios de
Posgrado e Investigación**
Asunto: **Autorización de Impresión**
OPCION: I

C. MARCO JULIO DEL MORAL ARGUMEDO

Candidato a Grado de Doctor en:

CIENCIAS DE LA INGENIERÍA

P R E S E N T E.-

De acuerdo con el Reglamento de Titulación vigente de los Centros e Institutos Tecnológicos Federales del Tecnológico Nacional de México, de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

**"SISTEMA INTELIGENTE BASADO EN INDUSTRIA 4.0 PARA MEJORAR LA GESTIÓN DE CONOCIMIENTO
DENTRO DE LA LOGÍSTICA DE SALUD PÚBLICA"**

Comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

ATENTAMENTE

Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®



DRA. OFELIA LANDETA ESCAMILLA
JEFA DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN



Orizaba, Veracruz, 05/noviembre/2024
Asunto: Revisión de trabajo escrito

C. OFELIA LANDETA ESCAMILLA
JEFA DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
PRESENTE.-

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C.

MARCO JULIO DEL MORAL ARGUMEDO

La cual lleva el título de:

"SISTEMA INTELIGENTE BASADO EN INDUSTRIA 4.0 PARA MEJORAR LA GESTIÓN DE CONOCIMIENTO DENTRO DE LA LOGÍSTICA DE SALUD PÚBLICA"

Y concluyen que se acepta.

ATENTAMENTE
Excelencia en Educación Tecnológica®
CIENCIA-TÉCNICA-CULTURA®

PRESIDENTE: DR. ALBERTO ALFONSO AGUILAR LASSERRE



FIRMA

SECRETARIO: DR. RUBÉN POSADA GÓMEZ



FIRMA

VOCAL: DR. ULISES JUAREZ MARTINEZ



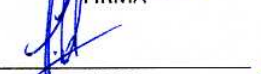
FIRMA

VOCAL: DR. OSCAR OSVALDO SANDOVAL GONZÁLEZ



FIRMA

VOCAL: DR. ALEJANDRO ALVARADO LASSMAN



FIRMA

VOCAL SUPLENTE: DR. ALBERTO OCHOA ZEZZATTI



FIRMA

TA-09 -18



“El camino del exceso conduce al palacio
de la sabiduría.”

- William Blake

Agradecimientos y dedicatoria

A mi madre. Por darme el coraje de enfrentarme a la vida con la potencia de un millón de soles, lleno de furor estelar.

A mi hermana. Mi alma gemela. La única persona que me entiende sin necesidad de palabras.

A mi abuela. Por cultivar y nutrir la preternatural semilla de mi intelecto.

A mi tía Sonia. Por ser mi segunda madre. Lo más cercano a la representación del amor incondicional.

A mi madrina, mi tercera madre. Nunca me pudiste domar, pero yo si te pude contagiar lo salvaje.

A mis hermanos: Dante, Alicia y en especial a Lalo, solo Dios sabe cuanto te extraño. A todos los Argumedo: Tío Pepe, mi tío Benja desde el paraíso. Mi tío Polo, nuestras almas comparten el mismo origen. A mi abuelo, el patriarca.

A mi papá y mi Tía Adriana. Por la oportunidad de continuar el legado familiar de científicos mexicanos.

A mis amigos. Por estar siempre a mi lado.

Al Doctor Alberto. Que le la vida me haya puesto en el camino de un mentor tan sublime como usted, es un regalo que no puedo agradecer lo suficiente.

CONTENIDO

RESUMEN	16
ABSTRACT	17
INTRODUCCIÓN	18
PLANTEAMIENTO DEL PROBLEMA	20
JUSTIFICACIÓN	28
HIPÓTESIS	31
OBJETIVOS	32
OBJETIVO GENERAL	32
OBJETIVOS ESPECÍFICOS	32
CAPÍTULO 1: ANTECEDENTES	33
1.1 Introducción	33
1.2 Instituto Mexicano del Seguro Social	33
1.3 Sistemas de Información en Salud	34
1.4 Datos	37
1.5 Conclusiones	38
CAPÍTULO 2: ESTADO DEL ARTE	39
2.1 Introducción	39
2.2 Análisis del Estado del Arte	39
2.2.1 Preguntas de investigación	42
2.3 Conclusión	44

CAPÍTULO 3: METODOLOGÍA	46
3.1 Introducción	46
3.2 Metodologías de trabajo	46
3.3 Conclusiones	51
CAPÍTULO 4: RESULTADOS	52
4.1 Introducción	52
4.2 Caso de estudio: Aplicación inteligente para determinar la existencia de medicinas asociadas con COVID19	54
4.2.1 Descripción del problema	55
4.2.2 Conceptualización	56
4.2.3 Discusión	65
4.3 Caso de estudio: Segmentación de Células Cervicales	67
4.3.1 Descripción del problema	68
4.3.2 Conceptualización	69
4.3.3 Discusión	83
4.4 Caso de estudio: Fiebre del Dengue	84
4.4.1 Descripción del problema	86
4.4.2 Conceptualización	91
4.4.3 Discusión	148
4.5 Arquitectura y diseño del sistema	150
4.5.1 Base de datos	153
4.5.2 Frontend	154

4.5.3	<i>Backend</i>	154
4.5.4	<i>Prototipo</i>	155
4.5.5	<i>Discusión</i>	165
CONCLUSIONES		168
RECOMENDACIONES.....		171
CONTRIBUCIÓN AL CONOCIMIENTO Y DESARROLLO TECNOLÓGICO		172
REFERENCIAS BIBLIOGRÁFICAS		175
PRODUCTOS ACADÉMICOS		187
ANEXO A: <i>Recopilación de datos</i>		192
A.1	<i>Datos epidemiológicos: SINAVE</i>	193
A.2	<i>Datos epidemiológicos: DGE</i>	196
A.3	<i>Datos geográficos: INEGI</i>	198
A.4	<i>Datos socio-demográficos: INEGI</i>	202
A.5	<i>Datos climatológicos: CONAGUA</i>	204
ANEXO B: <i>Desglose de revisión del estado del arte</i>		208
B.1	<i>Selección de artículos</i>	208
B.2	<i>Descripción de la metodología para revisión sistemática de literatura</i>	210
B.2.1	<i>Determinar preguntas de investigación</i>	212
B.2.2	<i>Extraer palabras clave y ejes de la investigación</i>	212
B.2.3	<i>Elección de buscadores pertinentes</i>	213
B.2.4	<i>Búsqueda secuencial con keywords</i>	213
B.2.5	<i>Aplicación de filtros</i>	214

B.2.6	Implementación de criterios de elegibilidad.....	215
ANEXO C:	Series de tiempo	217
C.1	Proceso de creación de algoritmo para predecir casos de Dengue y creación de prototipo	217
C.1.1	Datos utilizados para los experimentos.....	218
C.1.2	Métodos de Machine Learning para predicción de series de tiempo para dengue	219
C.1.3	Métodos de Deep Learning para predicción de series de tiempo para dengue.....	229
C.2	Integración de algoritmo de pronósticos	266
C.3	Conclusiones de los experimentos	268

LISTA DE FIGURAS

Figura 1.	Logo del IMSS	33
Figura 2.	Hospital de Herlev en Dinamarca (a) y el logo de la Universidad del Mar Egeo	38
Figura 3.	Diagrama de Venn	41
Figura 4.	Artículos por categoría.....	42
Figura 5.	Áreas de la sanidad	43
Figura 6.	Diagrama incremental de la metodología.....	47
Figura 7.	Diagrama de modelado	48
Figura 8.	Tipos de análisis	49
Figura 9.	Jerarquía <i>DIKW</i>	50
Figura 10.	Diagrama de resultados	53
Figura 11.	Descubrimiento de conocimiento	58
Figura 12.	Minería de datos	58
Figura 13.	Diagrama de flujo del Algoritmo K-means	60
Figura 14.	Diagrama de flujo del Algoritmo Cultural	62
Figura 15.	Resultados de la optimización.....	64
Figura 16.	Algoritmos culturales.....	65
Figura 17.	Muestreo de imágenes y máscaras	70
Figura 18.	Metodología para la segmentación de células cervicales	72
Figura 19.	Módulo MBConv	73
Figura 20.	Feature Pyramid Network, diagrama conceptual.....	74
Figura 21.	Creación de máscaras.....	75
Figura 22.	Imágenes aumentadas	76
Figura 23.	Comparación entre métodos de validación cruzada	77
Figura 24.	Pérdida en entrenamiento y validación	79
Figura 25.	Muestreo de células segmentadas	80
Figura 26.	Casos de dengue en el mundo.....	86
Figura 27.	Transmisión de la Fiebre del Dengue	87
Figura 28.	Ciclo de vida del <i>Aedes Aegypti</i>	88
Figura 29.	Casos de Fiebre del Dengue a nivel mundial.....	89
Figura 30.	Análisis demográfico	92
Figura 31.	Casos por grupo de edad.....	93

Figura 32.	Casos por grupo profesión	94
Figura 33.	Casos en municipios	96
Figura 34.	Densidad de población por municipio	97
Figura 35.	Regresión lineal	99
Figura 36.	Representación de la distancia esférica	100
Figura 37.	DBSCAN aplicado a Coatzacoalcos.....	102
Figura 38.	Casos de dengue por entidad	105
Figura 39.	Comparativa de series de tiempo	106
Figura 40.	Correlación de series de tiempo	108
Figura 41.	Descomposición de series de tiempo para los totales de casos de dengue.....	109
Figura 42.	Serie de tiempo de casos en Veracruz	110
Figura 43.	Comparativa de series de tiempo para Veracruz.....	111
Figura 44.	Análisis de correlación para Veracruz.....	112
Figura 45.	Comparativa de casos por año y mes para Veracruz.....	113
Figura 46.	Gráfico de caja para determinar la estacionalidad por mes	114
Figura 47.	Comparativa.....	115
Figura 48.	Descomposición de la serie de tiempo de casos en Veracruz	116
Figura 49.	Resultados de modelos predictivos	118
Figura 50.	Resultados del diagnóstico para modelo <i>SARIMAX</i>	119
Figura 51.	Estaciones climatológicas	122
Figura 52.	Comparativa de métodos de interpolación geoespacial.....	124
Figura 53.	Comparación entre resultados CONAGUA (a) y Kriging (b).....	125
Figura 54.	Proporción y número de casos por clase	132
Figura 55.	Proceso de validación cruzada anidada	134
Figura 56.	Resultados de la búsqueda en cuadrícula.....	135
Figura 57.	Representación de un bosque aleatorio	137
Figura 58.	Proceso de optimización.....	137
Figura 59.	Importancia de los hiperparámetros	138
Figura 60.	Rendimiento por hiperparámetro e iteración	139
Figura 61.	Rendimiento entre hiperparámetros	139
Figura 62.	Mapa de contorno para la relación de hiperparámetros.....	140
Figura 63.	Gráfico de coordenadas paralelas	140
Figura 64.	Curva de entrenamiento.....	143

Figura 65.	Gráficos comunes para evaluación de un modelo de clasificación.....	144
Figura 66.	Equidad para la combinación de todas las características.....	146
Figura 67.	Importancia de las características en las predicciones	147
Figura 68.	Evaluación probabilística.....	148
Figura 69.	Diagrama conceptual del sistema	152
Figura 70.	Arquitectura de la plataforma web para el Sistema Inteligente.....	155
Figura 71.	Arquitectura MTV de Django	156
Figura 72.	Análisis replicados en el prototipo	158
Figura 74.	Otros análisis posibles.....	160
Figura 75.	Mapeo de casos de dengue	161
Figura 76.	Arquitectura MTV de Django	162
Figura 77.	Visualización de series de tiempo	163
Figura 78.	Descomposición de series de tiempo	164
Figura 79.	Gráficos adicionales relevantes	165
Figura 80.	Captura de pantalla de la plataforma SINAVE.....	194
Figura 81.	Ficha informativa: SINAVE.....	194
Figura 82.	Ficha informativa: Anuario DGE.....	197
Figura 83.	Excalibur.....	198
Figura 84.	Ficha informativa: Marco Geoestadístico Nacional	199
Figura 85.	Área por municipio.....	202
Figura 86.	Ficha informativa: Censo 2020	203
Figura 87.	Captura del sitio.....	205
Figura 88.	Código fuente del sitio	206
Figura 89.	Ficha informativa: CONAGUA	207
Figura 90.	Diagrama del modelo prisma.....	211
Figura 91.	Método de conversión a aprendizaje supervisado.....	220
Figura 92.	Diagrama del árbol de decisión	222
Figura 93.	Importancia de las variables	224
Figura 94.	Comparativa entre serie y su predicción.....	225
Figura 95.	Importancia de variables: Ingeniería de características.....	226
Figura 96.	Comparativa de resultados: Ingeniería de características	227
Figura 97.	Importancia de las variables: variables climatológicas	228
Figura 98.	Comparativa de resultados: variables climatológicas	228

Figura 99.	Comparativa de modelos	231
Figura 100.	Método de escalamiento	232
Figura 101.	Comparativa de optimizadores	233
Figura 102.	Arquitecturas del perceptrón.....	234
Figura 103.	Comparativa de pérdida: perceptrón de 1 capa.....	235
Figura 104.	Evaluación secuencial: perceptrón de 1 capa.....	236
Figura 105.	Comparativa de rendimiento: perceptrón de 1 capa.....	236
Figura 106.	Comparativa de pérdida: perceptrón de 2 capas	237
Figura 107.	Evaluación secuencial: perceptrón de 2 capas.....	238
Figura 108.	Comparativa de rendimiento: perceptrón de 2 capas	238
Figura 109.	Comparativa de pérdida: perceptrón de 3 capas	239
Figura 110.	Evaluación secuencial: perceptrón de 3 capas.....	240
Figura 111.	Comparativa de rendimiento: perceptrón de 3 capas	240
Figura 112.	Convolución para series de tiempo.....	241
Figura 113.	Comparativa de pérdida: CNN de 1 capa.....	242
Figura 114.	Evaluación secuencial: CNN de 1 capa	242
Figura 115.	Comparativa de rendimiento: CNN de 1 capa.....	243
Figura 116.	Redes recurrentes y Long Term Short Memory Units.....	244
Figura 117.	Comparativa de pérdida: LSTM de 1 capa.....	245
Figura 118.	Evaluación secuencial: LSTM de 1 capa	245
Figura 119.	Comparativa de rendimiento: LSTM de 1 capa.....	246
Figura 120.	Comparativa de pérdida: LSTM de 2 capas.....	247
Figura 121.	Evaluación secuencial: LSTM de 2 capas	247
Figura 122.	Comparativa de rendimiento: LSTM de 2 capas.....	248
Figura 123.	Comparativa de pérdida: LSTM bidireccional	249
Figura 124.	Evaluación secuencial: LSTM bidireccional	249
Figura 125.	Comparativa de rendimiento: LSTM bidireccional	250
Figura 126.	Comparativa de rendimiento para modelos univariados	251
Figura 127.	Comparativa de pérdida: perceptrón de 1 capa multipaso.....	253
Figura 128.	Comparativa de pérdida: perceptrón de 1 capa multipaso.....	254
Figura 129.	Comparativa de pérdida: perceptrón de 1 capa multipaso.....	255
Figura 130.	Comparativa de pérdida: perceptrón de 1 capa multipaso.....	256
Figura 131.	Comparativa de pérdida: lstm de 1 capa multipaso	257

Figura 132.	Comparativa de pérdida: percentrón de 1 capa multipaso.....	258
Figura 133.	Comparativa de pérdida: percentrón de 1 capa multipaso.....	259
Figura 134.	Comparativa para modelos multipaso	260
Figura 135.	Comparativa de pérdida: percentrón de 2 capas multivariado.....	262
Figura 136.	Comparativa de pérdida: LSTM de 2 capas multivariado	263
Figura 137.	Comparativa de pérdida: percentrón de 2 capas multipaso y multivariable .	264
Figura 138.	Comparativa de pérdida: LSTM de 2 capas multipaso y multivariable	265
Figura 139.	Modelo de Facebook Prophet.....	267
Figura 140.	Tendencia y estacionalidad	267

LISTA DE TABLAS

Tabla 1.	Medicamentos para tratar enfermedades de las vías respiratorias	56
Tabla 2.	Características de los datos Herlev	70
Tabla 3.	Hiperparámetros del experimento	77
Tabla 4.	Comparación de los experimentos	78
Tabla 5.	Resultados finales.....	81
Tabla 6.	Comparación de rendimiento con otros trabajos	81
Tabla 7.	Flujo de datos para el caso de estudio	90
Tabla 8.	Horizontes y resolución	91
Tabla 9.	Cantidad de casos por municipios veracruzanos	95
Tabla 10.	Primeros diez elementos de conteo de puntos por polígonos.....	97
Tabla 11.	Densidad y número de casos	98
Tabla 12.	Correlación entre densidad y número de casos.....	98
Tabla 13.	Cantidad de casos por cluster	101
Tabla 14.	Casos en el mismo domicilio	101
Tabla 15.	Conjunto de datos de series de tiempo	103
Tabla 16.	Casos anuales en Veracruz.....	110
Tabla 17.	Resultados del ajuste modelo <i>SARIMAX</i>	120
Tabla 18.	Coeficientes del ajuste	120
Tabla 19.	Resultados del ajuste modelo <i>SARIMAX</i>	121
Tabla 20.	Mejores modelos y sus hiperparámetros	136
Tabla 21.	Mejores hiperparámetros para Bosque Aleatorio	141
Tabla 22.	Estadísticas de los árboles en el bosque	141
Tabla 23.	Resultados generales de evaluación	142
Tabla 24.	Resultados particulares de evaluación	143
Tabla 25.	Métricas de equidad	145
Tabla 26.	Muestra de datos de plataforma SINAVE.....	196
Tabla 27.	Tabla de artículos revisados	209
Tabla 28.	Palabras clave utilizadas.....	213
Tabla 29.	Tabla de resultados de búsqueda secuencial	214
Tabla 30.	Tabla de búsqueda con filtros	214
Tabla 31.	Tabla de artículos agrupados por eje	215

Tabla 32.	Tabla de análisis	216
Tabla 33.	Datos usados para el experimento	218
Tabla 34.	Comparativa del error	223
Tabla 35.	Ingeniería de características	226
Tabla 36.	Rendimiento por arquitectura: perceptrón de 1 capa	235
Tabla 37.	Rendimiento por arquitectura: perceptrón de 2 capas	237
Tabla 38.	Rendimiento por arquitectura: perceptrón de 3 capas	239
Tabla 39.	Rendimiento por arquitectura: CNN de 1 capa.....	241
Tabla 40.	Rendimiento por arquitectura: LSTM de 1 capa.....	244
Tabla 41.	Rendimiento por arquitectura: LSTM de 2 capas.....	246
Tabla 42.	Rendimiento por arquitectura: LSTM Bidireccional	248
Tabla 43.	Rendimiento por arquitectura: Perceptrón de 1 capa multipaso	252
Tabla 44.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	253
Tabla 45.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	254
Tabla 46.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	255
Tabla 47.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	256
Tabla 48.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	257
Tabla 49.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	258
Tabla 50.	Datos para series multivariable	261
Tabla 51.	Rendimiento por arquitectura: Perceptrón de 2 capas multipaso	262

RESUMEN

La gestión del conocimiento dentro de las cadenas logísticas de salud es crucial, ya que permite a los tomadores de decisiones convertir los datos generados en abstracciones, métodos y soluciones capaces de salvar vidas. Los Sistemas de Información en Salud captan estos datos y los almacenan de forma estructurada para su posterior consulta y operación. Sin embargo, estos sistemas son incapaces por sí solos de extraer conocimiento. Por ello, al combinarlos con los componentes de la *Industria 4.0*, se pueden convertir en auténticos Sistemas Inteligentes que, de manera dinámica, descubren patrones de comportamiento y proponen soluciones.

El presente trabajo es relevante porque define un Sistema Inteligente específico capaz de procesar datos de las cadenas logísticas de salud, integrando distintos algoritmos y técnicas computacionales. Se logró operar con datos médicos, climatológicos, demográficos y geoespaciales, integrándolos con algoritmos novedosos. Esto permite que el Sistema encapsule todo el proceso y refine datos en información, conocimiento y finalmente sabiduría.

Se analizaron tres casos de estudio distintos para modelar el flujo de datos dentro de la cadena logística y derivar el desarrollo del sistema. El primer caso conceptualizó una herramienta que coordinó distintos tipos de algoritmos para la administración del inventario médico y utilizó esta información para determinar el comportamiento del COVID-19 en una ciudad. El segundo caso de estudio usó algoritmos de *Aprendizaje Profundo*, uno de los más importantes dentro del paradigma de *Industria 4.0*, para extraer conocimiento de imágenes citológicas y segmentar células individuales en sus componentes. El tercer caso de estudio modeló el comportamiento de la fiebre del dengue a distintas resoluciones espaciales y temporales. Este modelado permitió encontrar fuentes de datos externas para apoyar la toma de decisiones epidemiológicas. Por último, los resultados de estos análisis fueron integrados en un prototipo del **Sistema Inteligente de Información en Salud** capaz de conjugar distintas fuentes de datos y operarlas dentro de la cadena logística de salud.

Palabras Clave: Industria 4.0, Sistemas de Información en Salud, Gestión de conocimiento, Logística de Sanidad Pública

ABSTRACT

The management of knowledge within health logistics chains is crucial, as it allows decision-makers to convert the data generated into abstractions, methods, and solutions capable of saving lives. Health Information Systems capture this data and store it in a structured manner for subsequent consultation and operation. However, these systems alone are incapable of extracting knowledge. Therefore, by combining them with the components of *Industry 4.0*, they can become true Intelligent Systems that dynamically discover behavioral patterns and propose solutions.

This work is relevant because it defines a specific Intelligent System capable of processing data from health logistics chains, integrating various algorithms and computational techniques. It successfully operated with medical, climatological, demographic, and geospatial data, integrating them with novel algorithms. This allows the System to encapsulate the entire process and refine data into information, knowledge, and finally wisdom.

Three different case studies were analyzed to model the data flow within the logistics chain and derive the system's development. The first case conceptualized a tool that coordinated different types of algorithms for medical inventory management and used this information to determine the behavior of COVID-19 in a city. The second case study used *Deep Learning* algorithms, one of the most important within the *Industry 4.0* paradigm, to extract knowledge from cytological images and segment individual cells into their components. The third case study modeled the behavior of dengue fever at different spatial and temporal resolutions. This modeling allowed the identification of external data sources to support epidemiological decision-making.

Finally, the results of these analyses were integrated into a prototype of the **Intelligent Health Information System** capable of combining different data sources and operating them within the health logistics chain.

Keywords: Industria 4.0, Sistemas de Información en Salud, Gestión de conocimiento, Logística de Sanidad Pública

INTRODUCCIÓN

La *Industria 4.0* es la implementación alemana de la **Cuarta Revolución Industrial** (Schawb, 2016, p. 7). Esta revolución promete traer cambios nunca antes vistos en la historia de la humanidad. Su poder transformador es enorme gracias a los avances en todas las áreas de la Ciencia y Tecnología. Sus pilares tecnológicos, como el Cómputo en la Nube, el Aprendizaje Automático, el Internet de las Cosas, la Manufactura Aditiva, la Robótica y la Ciencia de Datos, se integran bajo este paradigma industrial que los conjuga de una manera holística para crear un potencial de cambio donde el todo es mayor que la suma de sus partes.

En nuestro país, la *Industria 4.0* presenta un rezago significativo en su adopción. Esto impacta directamente en la competitividad de las industrias privadas y públicas, debido a la falta de políticas de incentivo tecnológico industrial. Todos los indicadores, incluso antes de la pandemia, mostraban una tendencia a la baja durante 2018 y principios de 2019. Esto evidencia la urgencia de implementar y desplegar este paradigma en nuestro país, tanto a nivel privado como gubernamental (Riquelme, 2019).

Uno de los mayores impulsores de la *Industria 4.0* es la gran cantidad de datos que se generan actualmente. Los sistemas de información, potenciados por los avances tecnológicos que les permiten ingerir datos en volúmenes y variedades enormes, se han transformado de meros repositorios a la parte central de muchas organizaciones, tanto públicas como privadas, a nivel internacional. Sin embargo, se requiere una actitud proactiva en su implementación para evolucionar estos sistemas y sacarlos de la pasividad y linealidad que los caracteriza. El paradigma de *Industria 4.0* tiene la capacidad de catalizar esta evolución. Esto es de suma importancia en el área de la logística, que tradicionalmente ha lidiado con problemas sumamente complejos, como la asignación de recursos, por ejemplo, las camas dentro de un hospital, lo cual es un problema *NP-completo* (Oddoye et al., 2007).

La reciente pandemia ha evidenciado la necesidad de construir sistemas de salud pública robustos que permitan a los tomadores de decisiones crear las mejores políticas de salud pública, las cuales se traducen directamente en salvar vidas. Tradicionalmente, los Sistemas de Información de Salud Pública se han comportado como repositorios pasivos donde solamente

se almacena la información, muchas veces coexistiendo con estándares distintos (Greenes, 2015). Los Sistemas de Información en Salud Pública guardan datos valiosos que, transformados en información y posteriormente en conocimiento, pueden ofrecer no solo análisis descriptivo, sino también otros tipos de análisis capaces de ofrecer soluciones y perspectivas hacia el futuro. Así, se puede extraer sabiduría y mejorar la gestión del conocimiento dentro de las cadenas logísticas de sanidad pública.

PLANTEAMIENTO DEL PROBLEMA

La pasividad de los Sistemas de Información en Salud y la carencia de una metodología se ha evidenciado por la pandemia de COVID-19. La falta de información en salud es una de las principales causas de la incapacidad para detener o contener la pandemia. Se aboga con urgencia por una forma estándar para el intercambio de información de salud y la inclusión de fuentes de datos externas (Jason, 2020). Estas fuentes de datos pueden ser públicas (como el Marco Geoestadístico Nacional del INEGI¹) o privadas (como las estadísticas de motores de búsqueda). Como proponen Perl et al. (Perl & Price, 2020), es de suma importancia tener acceso a la información de las personas para mejorar la respuesta ante epidemias.

Los Sistemas de Información en Salud desplegados en los países afectados por la pandemia de COVID-19 fueron cruciales para los tomadores de decisiones al desarrollar políticas y protocolos adecuados para responder ante el evento epidemiológico. Shiffman et al. (Shiffman & Shawar, 2020) mencionan que los datos son esenciales para entender las enfermedades. Además, enfatizan la disparidad que existe entre los países, lamentando que los ministros de salud de países como el nuestro no consideren necesario invertir en sus sistemas, limitando su capacidad de respuesta y creando así una cantidad enorme de muertes innecesarias (Salud, 2020b).

Con respecto a los datos, la pandemia fungió como una prueba de estrés de las políticas de captura de datos epidemiológicos. Actualmente, hay un auge en el análisis de datos referentes a este evento; gracias a ello, se ha obtenido un conocimiento sin precedentes sobre una pandemia en toda la historia de la humanidad. Sin embargo, existen diversos problemas debido a la falta de organización central dentro de las cadenas logísticas de sanidad pública. Algunos problemas incluyen la forma de tomar los datos (que puede crear patrones apócrifos), la falta de análisis de casos atípicos, la inclusión o no de un marco geográfico y la falta de verificación de resultados. Un Sistema Inteligente de Información en Salud puede mejorar una o más cadenas logísticas de salud (Geitgey, 2020), como el diagnóstico médico, el análisis epidemiológico o la respuesta ante emergencias médicas.

¹Instituto Nacional de Estadística y Geografía

Resolver el problema complejo de la logística de sanidad desde una perspectiva de datos requiere lidiar con problemas inherentes tanto a sus características primordiales como a su captura, almacenamiento, transmisión y procesamiento. A continuación, se enumeran algunos problemas presentes en los datos generados por organizaciones de salud:

- **Gran cantidad de datos:** Los datos se generan en gran cantidad en un corto periodo de tiempo, lo cual hace de su captura y procesamiento un problema no trivial. El *Big Data* ofrece una forma de captar estos datos y evitar su fuga.
- **Ruido:** El ruido en los datos no puede ser prevenido, ya que surge de cualquier proceso natural, sumado al error humano en la captura de datos. El análisis estadístico ofrece técnicas capaces de detectar valores atípicos en un conjunto de datos, lo que reduce la incertidumbre.
- **Falta de estructura:** Distintas fuentes de datos pueden tener diferentes estructuras. Lo que es importante para un especialista puede no serlo para otro. Es necesario implementar bases de datos capaces de manejar datos tanto estructurados como no estructurados.
- **Distintos tipos de datos:** Actualmente, la información médica se puede presentar en archivos tabulares o reportes escritos, pero también existen datos variados como señales o imágenes volumétricas. Algoritmos de *Aprendizaje Profundo* pueden ingerir imágenes casi sin procesamiento previo y realizar predicciones casi a nivel humano.
- **Multidimensionalidad:** Aun cuando solo se cuente con datos numéricos, es posible que estos tengan una gran cantidad de dimensiones, lo cual dificulta su procesamiento. La disponibilidad de algoritmos de reducción de dimensionalidad o dispositivos de hardware especializados es otro aporte ligado a la *Industria 4.0*.
- **Distintos tipos de variable:** Se pueden encontrar dentro de un mismo documento variables numéricas, discretas, continuas, categóricas, booleanas o inclusive difusas. Aplicar técnicas de Ciencia de Datos permite convertir datos tan dispares en información valiosa.

Los problemas dentro de la cadena logística de sanidad pública son sumamente complejos. La

forma de interactuar dentro de los elementos de la misma ofrece una dificultad especial, sobre todo si se toma en cuenta que muchas de las decisiones involucran vidas humanas. Esta es la razón por la cual se propone un **Sistema Inteligente** capaz de conjugar los componentes de *Industria 4.0* para ofrecer a los tomadores de decisiones herramientas analíticas y predictivas, con el apoyo de modernas técnicas algorítmicas que surgen de la *Inteligencia Artificial*.

De manera precisa, este trabajo doctoral conjuga los Sistemas de Apoyo a la Decisión y los Sistemas de Información con componentes de *Industria 4.0*, en un **Sistema Inteligente de Salud Pública** que tiene como objetivo mejorar la gestión de conocimiento dentro de uno o varios eslabones de la cadena logística de salud pública. Este **Sistema Inteligente** es capaz de extraer conocimiento de datos, generar soluciones y ofrecerlas a los tomadores de decisiones.

Cada uno de los casos de estudio se desarrolló de tal manera que apoyara la creación del **Sistema Inteligente**, ya sea mediante la generación de experiencia para implementar ciertos algoritmos, algún tipo de análisis particular o la manipulación de componentes complejos y la relación entre los mismos.

El primer caso de estudio es una propuesta para un Sistema Inteligente Híbrido, capaz de determinar el comportamiento de dispensarios médicos, ofreciendo como solución la identificación de puntos de interés y la gestión administrativa de inventario. Descubriendo información dentro de datos con herramientas computacionales adecuadas, como los *Data Warehouses* y *Metaheurísticas*. Este trabajo ofreció un acercamiento a la Logística Humanitaria y el impacto que pueden lograr los sistemas inteligentes en este tipo de aplicaciones. A continuación, se presentan las justificaciones para este caso de estudio, respaldadas por estadísticas aplicadas a México y la resolución de problemas generales y puntuales en las cadenas logísticas de salud:

- **Estadísticas de Salud en México:** Según datos del Instituto Nacional de Estadística y Geografía (INEGI), en 2020, el 17.2 % de la población en México no tenía acceso a servicios de salud. Además, la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2018 reportó que el 20 % de los pacientes experimentaron desabasto de medicamentos en las instituciones públicas de salud.
- **Optimización de la Cadena Logística:** La implementación de un Sistema Inteligente

Híbrido puede mejorar significativamente la eficiencia de la cadena logística de salud, reduciendo tiempos de espera y costos operativos. Esto es crucial en un país donde la logística de salud enfrenta desafíos como la dispersión geográfica y la variabilidad en la demanda de servicios médicos.

- **Identificación de Puntos de Interés:** El sistema propuesto puede identificar puntos de interés críticos, como áreas con alta demanda de servicios médicos o regiones con desabasto recurrente de medicamentos. Esto permite una mejor planificación y distribución de recursos.
- **Gestión Administrativa de Inventario:** La gestión eficiente del inventario es esencial para evitar el desabasto de medicamentos. En México, el 30 % de los medicamentos en instituciones públicas no están disponibles en el momento que los pacientes los necesitan, según la Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS).
- **Resolución de Problemas Generales:** Los sistemas inteligentes pueden abordar problemas generales en las cadenas logísticas de salud, como la falta de coordinación entre diferentes niveles de atención médica y la ineficiencia en la distribución de recursos. Esto es especialmente relevante en situaciones de emergencia, donde la rapidez y precisión en la distribución de medicamentos y suministros médicos son cruciales.
- **Problemas Puntuales en la Distribución de Medicinas:** En México, uno de los problemas puntuales es la distribución desigual de medicamentos entre zonas urbanas y rurales. Un Sistema Inteligente Híbrido puede utilizar algoritmos de optimización y análisis de datos para asegurar una distribución equitativa y oportuna de medicamentos, reduciendo así las disparidades en el acceso a la salud.
- **Impacto en la Logística Humanitaria:** La propuesta también tiene implicaciones importantes para la logística humanitaria, mejorando la capacidad de respuesta ante desastres naturales y emergencias sanitarias. La identificación rápida de necesidades y la gestión eficiente de inventarios pueden salvar vidas y reducir el sufrimiento humano.

Estas justificaciones destacan la importancia y el potencial impacto de implementar un Sistema Inteligente Híbrido en la gestión de dispensarios médicos en México, abordando tanto proble-

mas generales de la cadena logística de salud como problemas específicos en la distribución de medicinas.

El segundo caso de estudio se relaciona con la cadena logística de diagnóstico, donde se utilizaron arquitecturas novedosas de redes neuronales para extraer conocimiento de imágenes citológicas con el objetivo de segmentarlas en sus componentes morfológicos. Los algoritmos de *Aprendizaje Profundo* son una de las mejores herramientas computacionales para la extracción de conocimiento en datos que antes no podían ser procesados directamente. A continuación, se presentan las justificaciones para este caso de estudio, aplicadas en el contexto mexicano:

- **Alta Incidencia de Cáncer Cervicouterino:** En México, el cáncer cervicouterino es una de las principales causas de muerte por cáncer en mujeres. Según datos del Instituto Nacional de Estadística y Geografía (INEGI), en 2019 se registraron más de 4,000 muertes por esta enfermedad. La detección temprana a través de imágenes citológicas es crucial para reducir la mortalidad.
- **Limitaciones en el Diagnóstico Tradicional:** Los métodos tradicionales de diagnóstico citológico son laboriosos y dependen en gran medida de la experiencia del patólogo. La implementación de redes neuronales puede automatizar y mejorar la precisión del diagnóstico, reduciendo el margen de error humano.
- **Acceso a Tecnología Avanzada:** En muchas regiones de México, especialmente en áreas rurales, el acceso a tecnología avanzada y personal capacitado es limitado. Los sistemas basados en Aprendizaje Profundo pueden ser implementados en centros de salud con recursos limitados, mejorando el acceso a diagnósticos precisos.
- **Reducción de Costos:** La automatización del proceso de diagnóstico mediante redes neuronales puede reducir significativamente los costos asociados con el análisis citológico. Esto es especialmente relevante en un sistema de salud pública con recursos financieros limitados.
- **Eficiencia en la Cadena Logística:** La implementación de sistemas inteligentes en la cadena logística de diagnóstico puede mejorar la eficiencia en la gestión de muestras y

resultados, reduciendo los tiempos de espera para los pacientes y mejorando la calidad del servicio.

- **Capacidad de Manejo de Grandes Volúmenes de Datos:** Los algoritmos de Aprendizaje Profundo son capaces de procesar y analizar grandes volúmenes de datos de manera eficiente. Esto es crucial en un país con una alta demanda de servicios de salud, donde la capacidad de manejo de datos puede ser un factor limitante.
- **Mejora en la Toma de Decisiones Clínicas:** La extracción de conocimiento a partir de imágenes citológicas mediante redes neuronales puede proporcionar información valiosa para la toma de decisiones clínicas, permitiendo a los médicos implementar tratamientos más efectivos y personalizados.
- **Investigación y Desarrollo:** La aplicación de arquitecturas novedosas de redes neuronales en el diagnóstico citológico puede impulsar la investigación y desarrollo en el campo de la inteligencia artificial y la salud en México, posicionando al país como un líder en innovación tecnológica en salud.

Estas justificaciones subrayan la importancia y el impacto potencial de utilizar arquitecturas de redes neuronales en la cadena logística de diagnóstico en México, abordando tanto problemas generales del sistema de salud como problemas específicos en la detección y tratamiento del cáncer cervicouterino.

El último y principal caso de estudio trabajado fue sobre la Fiebre del Dengue. Este caso de estudio se relaciona directamente con la cadena logística epidemiológica y se refiere a la gestión tanto de casos individuales, su localización geográfica, relación con factores externos como el clima, así como su comportamiento a través del tiempo. Culminando con algoritmos inteligentes que apoyen a los tomadores de decisiones a predecir brotes de dengue a distintas resoluciones tanto espaciales como temporales o detectar condiciones graves de la enfermedad en pacientes vulnerables, entre otras. A continuación, se presentan las justificaciones para este caso de estudio, aplicadas en el contexto mexicano:

- **Alta Incidencia de Dengue en México:** Según la Secretaría de Salud, en 2019 se repor-

taron más de 268,000 casos de dengue en México, con una incidencia particularmente alta en estados como Veracruz, Jalisco y Chiapas. La gestión efectiva de estos casos es crucial para reducir la morbilidad y mortalidad asociadas con la enfermedad.

- **Impacto del Clima en la Propagación del Dengue:** El dengue es una enfermedad transmitida por mosquitos, cuya proliferación está fuertemente influenciada por factores climáticos como la temperatura y la humedad. La implementación de algoritmos inteligentes que integren datos climáticos puede mejorar la predicción de brotes y la planificación de intervenciones preventivas.
- **Localización Geográfica de Casos:** La identificación y seguimiento de la localización geográfica de los casos de dengue es esencial para entender los patrones de propagación de la enfermedad. Los sistemas inteligentes pueden analizar datos geoespaciales para identificar áreas de alto riesgo y focalizar los esfuerzos de control.
- **Predicción de Brotes:** Los algoritmos de aprendizaje automático pueden predecir brotes de dengue a distintas resoluciones espaciales y temporales, proporcionando a los tomadores de decisiones información valiosa para implementar medidas preventivas y de control de manera oportuna.
- **Detección de Condiciones Graves:** La capacidad de detectar condiciones graves de la enfermedad en pacientes vulnerables es crucial para reducir la mortalidad. Los sistemas inteligentes pueden analizar datos clínicos y epidemiológicos para identificar pacientes en riesgo y priorizar su atención.
- **Optimización de Recursos:** La gestión eficiente de recursos, como insecticidas y personal de salud, es fundamental para el control del dengue. Los sistemas inteligentes pueden optimizar la distribución y uso de estos recursos, asegurando una respuesta más efectiva y económica.
- **Mejora en la Toma de Decisiones:** La integración de datos epidemiológicos, geográficos y climáticos en un sistema inteligente proporciona a los tomadores de decisiones una visión completa y precisa de la situación, mejorando la efectividad de las intervenciones y políticas de salud pública.

- **Reducción de Costos:** La implementación de sistemas inteligentes para la gestión del dengue puede reducir significativamente los costos asociados con el control de la enfermedad, al permitir una planificación y ejecución más eficiente de las intervenciones.

Estas justificaciones destacan la relevancia de utilizar algoritmos inteligentes en la gestión de la Fiebre del Dengue en México, abordando tanto problemas generales de la cadena logística epidemiológica como problemas específicos en la predicción y control de brotes.

JUSTIFICACIÓN

La importancia del correcto manejo de las cadenas logísticas con métodos basados en la gestión de datos para su refinación en conocimiento, se puede enriquecer enormemente del enfoque de *Industria 4.0*. Conceptos tecnológicos propios como *Big Data*, *Internet de las Cosas* o *Inteligencia Artificial*, sumados a conceptos más abstractos como la seguridad, privacidad, transparencia, apertura y cooperación son capaces de potenciar la logística de salud en cualquier organización tanto privada como pública; es en estas últimas donde se puede obtener un mayor beneficio para la población de un país. Primero se comienza con un Sistema de Información para almacenar todos los datos generados orgánicamente por las organizaciones, para posteriormente dotarlo de la capacidad de presentar inteligencia y apoyar a la toma de decisiones.

Podemos encontrar múltiples iniciativas para implementar estos sistemas en países en vías de desarrollo como México. Así mismo, también existen ejemplos claros de los beneficios de su implementación. En la Ciudad de Nueva York, se obtuvo una mejora en los tiempos de respuesta ante emergencias implementando un sistema de información (Gotham et al., 2007). Desde el año 2012 se probó que la implementación general de Sistemas de Información incide positivamente en toda la cadena logística de salud (Massoudi et al., 2012).

Vest et. al. (Vest et al., 2014) realizaron un estudio donde se analizó que el uso de Sistemas de Información no siempre garantiza un mejor rendimiento, el problema no es tecnológico. Se requiere una mejora organizacional general en los procedimientos y manuales. Este es el enfoque holístico de que la *Industria 4.0* ofrece dentro de la cadena logística. Maximizar la usabilidad, inter-operabilidad y transparencia dentro de estos sistemas, es crucial para su correcto funcionamiento.

En nuestro país, los Sistemas de Información en Salud se encuentran desacoplados de las capas de datos públicas que también son generadas por otras entidades gubernamentales. Un ejemplo claro es la plataforma epidemiológica de SINAVE² (Salud, 2020a), que centraliza los datos epidemiológicos de distintas enfermedades, pero no sigue los paradigmas de *Industria*

²Sistema Nacional de Vigilancia Epidemiológica

4.0; concretamente, el sitio está creado con tecnologías obsoletas y es incapaz de integrar otras capas de datos como las demográficas o geoespaciales. Como ya se expuso, esto inhibe la capacidad de respuesta ante cualquier eventualidad epidemiológica, caso que se puede extrapolar a cualquier sub-cadena logística dentro de la salud y más en la pública.

Se han realizado estudios para determinar el impacto de una estrategia nacional basada en Sistemas de Información de Salud Pública para responder a emergencias, gracias a que permite tomar decisiones inteligentes y basadas en datos para la creación de política pública. Lamentablemente, los métodos de toma de decisiones y análisis se llevan a cabo fuera del sistema y no tienen el dinamismo que podemos observar en sistemas integrados bajo el paradigma de *Industria 4.0* (Gotham et al., 2015).

La mayoría de los esfuerzos de la *Inteligencia Artificial* dentro del área de salud se centran en algoritmos de *Aprendizaje Automático* o *Aprendizaje Profundo* para la creación de Sistemas de Diagnóstico Asistido por Computadora, dejando de lado otras cadenas logísticas como la de emergencias, epidemiología o administración de hospitales. Por sí solas, estas técnicas de *Inteligencia Artificial* solo permiten dar solución a sistemas puntuales, por lo que se sugiere implementarlas bajo un marco de trabajo general, en donde la interacción de distintos algoritmos y modelos permita a los tomadores de decisiones manejar un panorama general a distintas resoluciones espacio-temporales (Paton & Kobayashi, 2019).

Así mismo, la mayoría de los esfuerzos en implementar la *Industria 4.0* en el área de la salud, se han aplicado a la propagación de tecnologías que mejoren la interacción paciente–médico, la gestión de expedientes o la administración de unidades hospitalarias privadas. Sin embargo, en países en vías de desarrollo (Peters et al., 2008), el impacto de este paradigma puede ser mucho mayor si se aplica a la logística de salud pública, sobre todo para incrementar la atención médica en poblaciones vulnerables por sus condiciones socioeconómicas (Kale et al., 2013). Los Sistemas de Información en Salud, son cruciales para mejorar la calidad de vida de la población en países en vías de desarrollo con una proporción significativa de población en condiciones de pobreza (Clifford et al., 2008). En nuestro país, donde el 92 % de la población total tiene acceso a un teléfono inteligente (Transportes, 2018), pero solo el 52 % tiene acceso a internet. (INEGI, 2019), es más importante dirigir los esfuerzos en apoyar a los tomadores

de decisiones, en cualquier eslabón de la cadena logística de sanidad, como el diagnóstico, tratamiento, administración, etcétera

Un **Sistema Inteligente de Salud Pública**, que integre múltiples algoritmos (Ling, 2014), nos permitirá pensar en soluciones generales y no segregar el problema en objetivos inconexos. Darle la perspectiva de sanidad pública, nos permite domar los componentes de *Industria 4.0* aplicados a la salud y codificarlos en soluciones pertinentes con el fin de mejorar la gestión de conocimiento dentro de las cadenas logísticas. Esto se traduce directamente en una mejora en el servicio y calidad de vida de todos los mexicanos.

HIPÓTESIS

El desarrollo de un **Sistema Inteligente de Salud Pública** basado en *Industria 4.0* permitirá la mejora en la gestión de conocimiento dentro de las Cadenas Logísticas de Salud Pública, habilitando el refinar datos en información, conocimiento y sabiduría; procesando datos multidimensionales, con incertidumbre, masivos y de distintas fuentes para identificar patrones de comportamiento a través del tiempo, apoyar en las decisiones, realizar pronósticos precisos, predecir enfermedades y proponer soluciones.

OBJETIVOS

OBJETIVO GENERAL

Desarrollar un Sistema Inteligente basado en *Industria 4.0* para mejorar la gestión del conocimiento dentro de la Logística de Sanidad Pública. Este sistema apoyará distintos nodos de la cadena de valor y convertirá los datos generados en un apoyo para los tomadores de decisiones, ofreciendo una solución integral con capacidad de realizar análisis descriptivos, diagnósticos, predictivos y prescriptivos.

OBJETIVOS ESPECÍFICOS

- Analizar el estado del arte de *Industria 4.0* aplicada a la salud, con enfoque en Sistemas de Información, para evaluar su participación en las cadenas logísticas, especialmente en lo relacionado con la salud pública.
- Abstraer entornos generales de acción dentro de la cadena logística de salud pública para determinar el flujo de datos, cómo estos se convierten en conocimiento y cómo determinan el comportamiento de la cadena.
- Explorar el universo de datos públicos o privados accesibles que puedan ser integrados en el sistema, incluyendo datos demográficos, geográficos y de infraestructura.
- Diseñar el **Sistema Inteligente de Información de Salud Pública**, incluyendo la arquitectura computacional y tecnológica, bajo el paradigma de *Industria 4.0*, con la capacidad de ingerir datos masivos, variados y con alta incertidumbre.
- Desarrollar las arquitecturas y algoritmos de Inteligencia Artificial, incluyendo tanto algoritmos de aprendizaje automático como metaheurísticas, para predecir, optimizar y automatizar procesos dentro del sistema.
- Programar la plataforma web para el despliegue del **Sistema Inteligente de Salud Pública**. La plataforma debe ser capaz de integrar los algoritmos y técnicas desarrolladas, permitiendo la operación directa de datos tanto médicos como no médicos, así como el análisis de datos, visualización y toma de decisiones.

CAPÍTULO 1: ANTECEDENTES

1.1 Introducción

En este capítulo se presentan los antecedentes relevantes para el desarrollo del presente trabajo. Se aborda la historia y contribuciones del IMSS ¹ en el ámbito de la salud pública y la investigación en México, así como su colaboración con otras instituciones. Además, se discuten los sistemas de información en salud a nivel nacional e internacional, destacando sus fortalezas y debilidades. Finalmente, se describen las fuentes de datos utilizadas en este estudio y su importancia para la investigación.

1.2 Instituto Mexicano del Seguro Social

El IMSS (Figura 1) es una de las instituciones de salud pública más grandes de América Latina. Fue creado en la Ciudad de México el 19 de enero de 1943 por el presidente Manuel Ávila Camacho. Actualmente, ofrece sus servicios a más de 80 millones de derechohabientes. Desde sus inicios, el instituto ha reconocido la importancia de la investigación para el mejoramiento de la salud nacional. En 1961, con la creación del Centro Médico Nacional, se unificaron los esfuerzos en investigación. Uno de los frutos de esta consolidación fue la creación de la píldora anticonceptiva, un desarrollo que ha impactado la salud reproductiva no solo del país sino del mundo. Es la tercera fuerza de investigación en México y novena a nivel regional. Otra razón de su éxito es la vinculación con otras instituciones de investigación como el Instituto Politécnico Nacional y con la iniciativa privada (Adan, 2018).



Figura 1: Logo del IMSS

¹Instituto Mexicano del Seguro Social

Con el TECNM ², en particular el ITO ³, el IMSS ha celebrado múltiples convenios específicos y generales de colaboración, siendo estos convenios semilleros de investigación de vanguardia y de interés nacional. En estos convenios se han aplicado las más modernas técnicas de la ingeniería, como la *Inteligencia Artificial*, para solucionar problemas de salud con especial énfasis en la población más vulnerable.

Algunos trabajos previos de colaboración entre ambas instituciones fueron:

- Diagnóstico de Cáncer de Mama.
- Gestión y Detección de Enfermedades Degenerativas.
- Optimización de Recursos Hospitalarios.
- Trabajos relacionados con la Epidemiología de Enfermedades Transmitidas por Vector y su relación con ciertas poblaciones vulnerables.

Dos de los casos de estudio aquí presentados se derivan directamente de trabajos realizados anteriormente con el IMSS. El primero, relacionado con las Enfermedades Transmitidas por Vector, es el más complejo, ya que requiere la interacción entre múltiples cadenas logísticas y el análisis e interpretación de múltiples formas de datos, lo que convierte al flujo de información en una entidad de alta complejidad. El segundo se refiere a un sistema capaz de tomar una célula cancerígena y dividirla correctamente en sus partes, esta segmentación es crucial para el problema del Diagnóstico de Cáncer Cervicouterino.

1.3 Sistemas de Información en Salud

Los Sistemas de Información en Salud a nivel nacional están representados por los esfuerzos de la Secretaría de Salud para vigilar la salud de la población nacional a través del estudio epidemiológico de los datos recabados por el Sistema Nacional de Salud. El Comité Nacional de Vigilancia Epidemiológica gobierna la operación del Sistema Nacional de Vigilancia Epidemiológica. Este sistema tiene dentro de sus funciones proveer información de salud

²Tecnológico Nacional de México

³Tecnológico de Orizaba

a distintos niveles, concentrando la información de distintas jurisdicciones para crear un conjunto de datos nacional. Su objetivo es apoyar la toma de decisiones en políticas de salud pública (Salud, 2023).

Esta plataforma es un mero repositorio de datos, los cuales se almacenan de manera estructurada y luego son canalizados a entidades especializadas para su procesamiento. Los repositorios de datos de salud pública nacionales, en su implementación tradicional, son incapaces de gestionar información aparte de la epidemiológica, como por ejemplo, seguimiento de casos individuales, administración hospitalaria (como las urgencias) o de realizar análisis por sí mismos.

En las instituciones de salud privadas a nivel nacional, sucede todo lo contrario. Sus esfuerzos se enfocan principalmente en la gestión de expedientes hospitalarios y de pacientes, administración hospitalaria en general, costos e inventario, así como satisfacción al cliente. Los sistemas de información en salud que se implementan en las instituciones de salud privadas carecen de cualquier capacidad de realizar análisis médico u ofrecer conocimiento que pueda aportar a la salud, aunque son robustos para aportar información a la cadena logística de salud interna. Esto es problemático, ya que se ha encontrado que esto inhibe la capacidad de interoperabilidad de los datos (Jason, 2020).

A nivel internacional, los Sistemas de Información en Salud Pública han tenido éxito en países como Francia, donde su sistema es capaz de agregar datos nacionales para darle poder a uno de los mejores sistemas de salud pública a nivel nacional. Es capaz de realizar una trazabilidad de cualquier paciente desde su nacimiento hasta causas de mortalidad y se integra directamente con otras entidades gubernamentales relacionadas como la Seguridad Social. Este sistema se enfoca exclusivamente en salud y no permite la interacción con otras fuentes de datos como la geográfica o la demográfica (Goldberg et al., 2012).

Por otra parte, está demostrado que metodologías ágiles (no solo aplicadas a Software, sino también a otras áreas de la industria) son cruciales para el desarrollo, integración y adopción de Sistemas de Información en Salud Pública. Estas metodologías se enfocan en un trabajo colaborativo, autogestionado con énfasis en el usuario final. Un paso importante es la mejora

continua, la cual se agiliza al reducir la carga de gestión del proyecto (Roberts, 2018).

Un sistema híbrido inteligente es un tipo de sistema que combina diferentes enfoques o técnicas para aprovechar las fortalezas individuales de cada componente y mejorar el rendimiento global del sistema. Estos sistemas suelen integrar componentes basados en *Inteligencia Artificial*, *Aprendizaje Automático*, *Metaheurísticas* u otros métodos para abordar problemas complejos.

En el contexto de la *Inteligencia Artificial*, un sistema híbrido puede aprovechar la combinación de enfoques simbólicos y conexionistas, o integrar diferentes técnicas de aprendizaje, como aprendizaje supervisado y no supervisado. También puede incluir la combinación de sistemas basados en reglas y enfoques basados en modelos matemáticos.

Algunos ejemplos de Sistemas Híbridos Inteligentes incluyen:

- **Sistemas de Recomendación Híbridos:** Utilizan diferentes técnicas para recomendar elementos a los usuarios, como filtrado colaborativo y filtrado basado en contenido.
- **Sistemas de Diagnóstico Médico:** Pueden combinar reglas expertas con técnicas de aprendizaje automático para mejorar la precisión en el diagnóstico médico.
- **Sistemas de Control Inteligente:** En sistemas de control, se pueden combinar técnicas clásicas de control con métodos de aprendizaje automático para adaptarse a cambios en el entorno de manera más eficiente.
- **Sistemas de Optimización Híbridos:** En problemas de optimización, se pueden combinar algoritmos heurísticos con enfoques basados en modelos matemáticos para encontrar soluciones más rápidas y precisas.
- **Sistemas de Gestión de Conocimiento:** Pueden combinar diferentes técnicas para extraer conocimiento de datos, como minería de datos, aprendizaje automático y razonamiento basado en casos.

La idea detrás de los Sistemas Híbridos es capitalizar las fortalezas de diferentes enfoques para abordar las limitaciones individuales de cada uno. Esto puede conducir a un rendimiento

más robusto y eficiente, especialmente en situaciones donde un solo enfoque podría no ser suficiente para resolver un problema complejo.

1.4 Datos

Los datos analizados para la realización de este trabajo provienen de distintas fuentes y son de diversa forma, formato, cantidad, resolución y magnitud. Los datos nacionales corresponden a cuatro ramas distintas: *Salud*, *Geográficos*, *Demográficos* y *Climatológicos*. Estos datos, recabados por distintas dependencias de gobierno tanto a nivel federal como local, permiten realizar investigaciones de interés nacional ([Anexo A](#)).

El INEGI es el encargado a nivel nacional de recabar datos sobre las condiciones geográficas y demográficas del país para explorar el estado del mismo. Los datos geoespaciales que recaban sirven para la descripción total de la superficie nacional desde un punto de vista físico y no político o administrativo, y fungen como base para otros procesos como los Censos de Población y Vivienda que se realizan cada 10 años (u. INEGI, [2020](#)) (INEGI, [s.f.](#)).

El SMN ⁴, en su dependencia de la CONAGUA ⁵, concentra información climatológica del país. Están en constante monitoreo de las condiciones del país para objetivos como la prevención de desastres o la creación de pronósticos climatológicos. Contienen información sobre la temperatura del país, así como precipitación, entre otros (CONAGUA, [2020](#)).

Con respecto a datos de salud, se tienen como fuentes la DGE ⁶ en su sitio de datos públicos, y también una pequeña cantidad de datos no accesibles al público que provienen de la plataforma SINAVE. Estos datos se refieren a los casos de Enfermedades Transmitidas por Vector, en este caso, la Fiebre del Dengue (Salud, [2020a](#)).

Se agradece al Hospital Universitario de Herlev y a la Universidad del Mar Egeo ([Figura 2](#)) por la compilación y liberación al público del conjunto de datos de células cervicales que sirve como *Benchmark* para modelos de visión por computadora, procesamiento de imágenes y algoritmos de clasificación (Dounias, [2017](#)) (Jantzen et al., [2005](#)).

⁴Sistema Meteorológico Nacional

⁵Comisión Nacional del Agua

⁶Dirección General de Epidemiología



(a)



(b)

Figura 2: Hospital de Herlev en Dinamarca (a) y el logo de la Universidad del Mar Egeo

1.5 Conclusiones

En este capítulo se han presentado diversas fuentes de datos relevantes para el desarrollo del presente trabajo. Se ha destacado la importancia del SMN y su papel en la recopilación de datos climatológicos, así como la contribución de la DGE y la plataforma SINAVE en la provisión de datos de salud pública. Además, se ha reconocido la valiosa aportación del Hospital Universitario de Herlev y la Universidad del Mar Egeo en la liberación de conjuntos de datos de células cervicales, fundamentales para el avance en modelos de visión por computadora y algoritmos de clasificación. Estas fuentes de datos son esenciales para abordar problemas complejos de salud pública y desarrollar soluciones innovadoras que mejoren la gestión de la salud de la población.

CAPÍTULO 2: ESTADO DEL ARTE

2.1 Introducción

La *Industria 4.0* es un paradigma que actualmente está impulsando cambios enormes en todos los ámbitos de la sociedad. A diferencia de otros paradigmas industriales, la *Industria 4.0* ofrece enfoques conectivos que permiten construir sistemas capaces de solucionar problemas complejos dentro de las cadenas logísticas. Las cadenas logísticas de salud pública, soportadas por Sistemas de Información en salud, apoyan a los expertos en salud y a los tomadores de decisiones a hacer su trabajo de manera más eficaz y precisa. Mejorar estos sistemas con el paradigma de *Industria 4.0* puede incrementar su competitividad y derivar en mejores políticas de salud pública. El análisis aquí comprendido, entre los periodos marzo 2019 y marzo 2020, presenta la interacción entre la *Industria 4.0*, la *Salud*, los *Sistemas de Información* y la *Inteligencia Artificial* para resolver problemas de **Salud Pública**. Se encontró que existe un rezago en la adopción del paradigma de *Industria 4.0*, lo que demuestra un área de oportunidad para la investigación que puede mejorar la capacidad de respuesta de estos sistemas ante eventos como la pandemia actual.

A continuación, se presentan los resultados del análisis del Estado del Arte, mientras que el procedimiento desglosado se encuentra en el [Anexo B](#).

2.2 Análisis del Estado del Arte

Para contestar las preguntas de investigación y posicionar la propuesta dentro del estado del arte, se utilizaron los ejes de investigación: *Ind* (*Industria 4.0*), *IA* (*Inteligencia Artificial*), *HIS* (*Sistemas de Información en Salud* ¹), *DM* (*Toma de Decisiones* ²), y *Pub* (*Salud pública*).

Los resultados seleccionados tienen como objetivo determinar la relación de los artículos con las preguntas de investigación. Para ello, nos basamos en los cuatro ejes antes mencionados para realizar una comparativa cualitativa y encontrar cómo se posiciona cada uno de los artículos en relación con cada uno de estos conjuntos. Esta evaluación conceptual basada

¹Siglas en inglés: Health Information Systems

²Siglas en inglés: Decision Making

en conjuntos nos permite posicionar cada artículo dentro de uno o más conjuntos, así como visualizar fácilmente las intersecciones que manifiestan la interacción entre los conceptos comparados.

El primer conjunto *Ind* contiene los artículos que hacen mención de la *Industria 4.0* o alguno de sus componentes como *Internet de las Cosas* o *Sistemas Ciber-físicos*. El conjunto *IA* contiene aquellos que muestran *Inteligencia Artificial* (*Aprendizaje Automático*, *Aprendizaje Profundo*, etc.) así como *Metaheurísticas*, *Simulación* y *Ciencia de Datos*. *HIS* contiene aquellos que mencionan *Sistemas de Información en Salud*. *Pub* aquellos que se relacionan con Salud Pública. *DM* los que mencionan alguna técnica computacional específica para la *Toma de Decisiones* en conjunto con otros algoritmos.

Los artículos fueron analizados manualmente y asignados a uno o más ejes en función de su contenido. La **Tabla 27** muestra los artículos que pertenecen a cada conjunto así como la cardinalidad de los conjuntos; la mayoría de los artículos pertenecen al conjunto *Ind* mientras que el conjunto *Pub* solamente cuenta con una cardinalidad de dos.

El diagrama de Venn (**Figura 3**) evidencia fácilmente la interacción entre los artículos y los ejes. Observamos que la mayoría de los artículos se encuentran en el conjunto *Ind*, el cual tiene intersecciones importantes con los otros conjuntos. Los conjuntos *HIS*, *Pub* y *DM* se encuentran siempre en intersección con otro. El conjunto *IA* es el segundo con más miembros que no se intersecan con otros, pero tiene más intersecciones que *Ind*. Los conjuntos *DM* y *Pub* son los que menos representación tienen, siendo *Pub* el menos representado, existiendo solamente en intersección con *HIS* e *IA*. Por último, encontramos que la Presente Investigación (PI) se encuentra en la intersección de todos los conjuntos. Las letras A, B y C representan en qué motor de búsqueda fueron encontrados los artículos; siendo A para *ScienceDirect*, B para *GoogleScholar* y C para *MicrosoftAcademic*.

Es importante analizar la perspectiva no solo de la interacción entre los conjuntos, sino también la falta de la misma. La falta de interacción más clara es que ninguno de los artículos se encuentra en la intersección de todos los conjuntos, lo cual representa un área de oportunidad enorme para la investigación. Al realizar un análisis cuantitativo de la cantidad de artículos

presentes en cada conjunto, esta brecha se hace evidente, dada por la falta de trabajos que se enfoquen en el área pública tanto de la aplicación de *Industria 4.0* en salud, como en la aplicación de *Sistemas de Información en Salud e Inteligencia Artificial*.

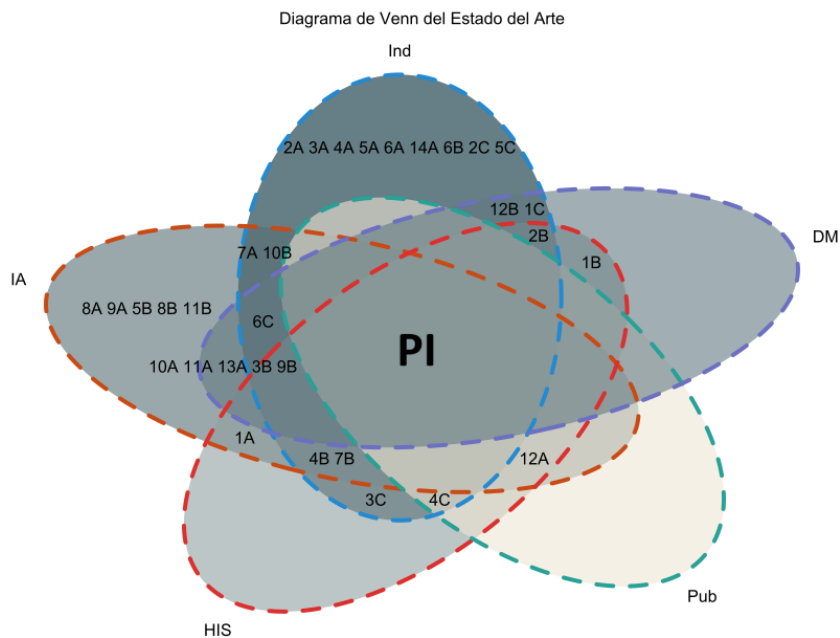


Figura 3: Diagrama de Venn

Como podemos observar en la [Figura 4](#), al contar los artículos que pertenecen a cada conjunto, encontramos una brecha importante. Esto evidencia la falta de trabajos que se enfoquen en el área pública tanto de la aplicación de la *Industria 4.0* en salud como en la aplicación de *Sistemas de Información e Inteligencia Artificial*.

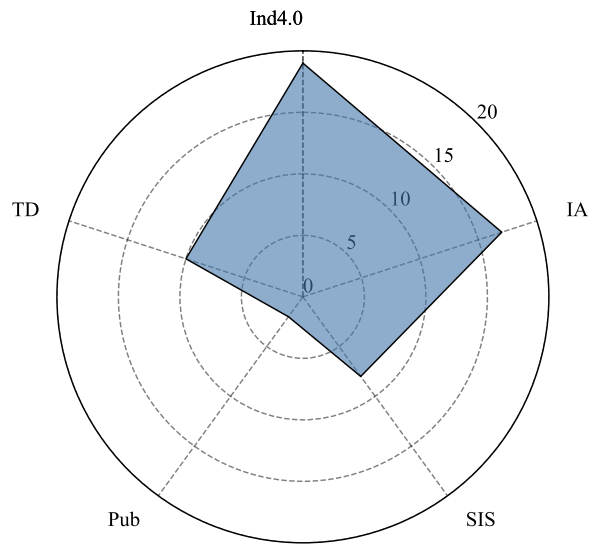


Figura 4: Artículos por categoría

2.2.1 Preguntas de investigación

Las preguntas de investigación son las siguientes:

¿Cuáles componentes de la *Industria 4.0* se han aplicado en el área de salud? Los componentes de *Industria 4.0* incluyen los conceptos y tecnologías asociados a este paradigma. El *Big Data*, el *Internet de las Cosas* y los *Sistemas Ciber-físicos* concentran la mayoría de los componentes de *Industria 4.0* aplicados a la salud. Llama la atención el componente de Economía Circular, que puede ser de especial atención en la sanidad pública para reducir costos e invertir el ahorro en personal o infraestructura, sobre todo en países en vías de desarrollo como México. Los chatbots y los wearables se enfocan en el paciente, estableciendo el primero un canal de comunicación y el segundo una forma de monitoreo remoto. Lamentablemente, por sus características, es difícil su adopción masiva en nuestro país.

¿Cuáles áreas de la logística de sanidad han sido tocadas por la *Industria 4.0*? Las áreas se refieren a las cadenas logísticas, tanto específicas como generales, presentes en la salud pública. La mayoría de los trabajos analizados solo atacan un área de la logística de salud, concentrándose en la cadena administrativa de procesos particulares más que en un enfoque

general y conectivo. Cuando se realizó la selección de artículos en los motores de búsqueda, en particular con la palabra clave *Inteligencia Artificial*, la mayor parte de los trabajos encontrados se enfocaban en el área de diagnóstico, en particular mediante Imagenología Médica, lo cual no se refleja en este análisis. De las áreas de Cirugía y Ortopedia, solo encontramos una mención; Ortopedia es la única área encontrada que menciona *Manufactura aditiva* o el concepto de *Smart Factory* (ver Figura 5).

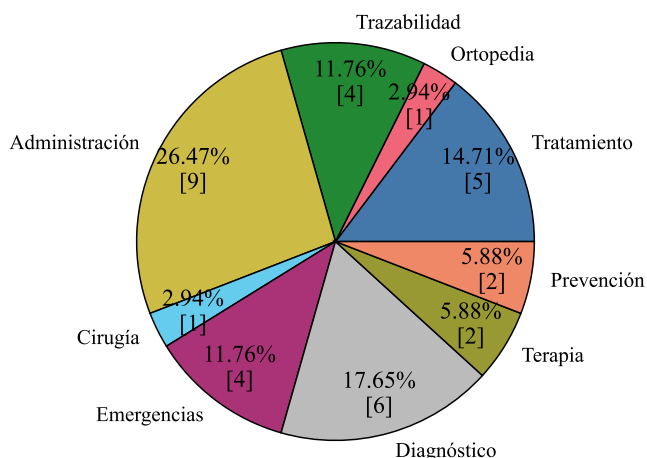


Figura 5: Áreas de la sanidad

¿Cuáles implementan o mencionan Sistemas de Información en Salud? Solo se encontraron 8 menciones de *Sistemas de Información en Salud*, siendo en su mayoría como apoyo a alguna técnica de *IA* o *DM*. El artículo 1A presenta una evaluación en la percepción del personal médico de estos sistemas, encontrándola favorable aunque ofrece algunas recomendaciones y retos.

¿Cuáles técnicas de IA, Toma de Decisiones, Machine Learning o Metaheurísticas se han implementado? En relación con *IA* y *DM*, se identificaron y analizaron los algoritmos específicos mencionados y utilizados en cada artículo. Se encuentran en su mayoría técnicas de *Aprendizaje Profundo*, siendo las Redes Neuronales Artificiales las más usadas. De estas redes neuronales, las Redes Neuronales Convolucionales son usadas principalmente cuando se procesan imágenes. El artículo 9A hace un análisis comparativo del rendimiento de varios

algoritmos. Se encontraron *Metaheurísticas* novedosas como Artificial Bee Colony. Las *Metaheurísticas* fueron usadas como método de IA, pero también para tomar decisiones. El artículo 3B utilizó Swarm Intelligence para optimizar los hiperparámetros de una Máquina de Soporte Vectorial, otro algoritmo bastante usado dentro de los artículos analizados.

¿Alguno aplica, menciona o se relaciona con salud pública? Únicamente se encontraron dos menciones a la salud pública dentro de los artículos. La mayoría no menciona si se enfocan a un área pública o privada, mientras que otros hablan específicamente del Servicio de Salud Privado. Los artículos hablan sobre la manera de interconectar distintas entidades dentro de la cadena logística nacional (hospitales, clínicas, laboratorios, dependencias) a escala regional y continental, y realizan un análisis de los retos que esto implica. Se hace énfasis en el manejo de los datos que se van a intercambiar dentro de estas cadenas, así como la mejora de la competitividad y eficiencia.

2.3 Conclusión

El propósito de este capítulo fue tomar una instantánea en el periodo de búsqueda sobre los artículos que relacionen conceptos plasmados en las palabras clave derivadas de las preguntas de investigación. Se realizó una búsqueda sistemática en tres buscadores de prestigio para incrementar la variedad de artículos para el análisis. En total, 31 artículos publicados entre marzo de 2019 y marzo de 2020 fueron seleccionados, procesados, categorizados y analizados para contestar las preguntas de investigación aquí propuestas.

Con los resultados se realizó una comparativa basada en ejes que asignaron cada artículo a uno o varios conjuntos para entender cómo se relacionan entre ellos en el espacio de temas que se analizaron. Posteriormente, se procedió a realizar una enumeración de componentes de *Industria 4.0*, áreas de la cadena logística, los algoritmos escogidos para labores de inferencia y predicción, así como una clasificación aparte de aquellos algoritmos utilizados directamente para asistir en la toma de decisiones.

Con respecto a los componentes de *Industria 4.0*, destaca la aparición de trabajos que incluyan conceptos abstractos como la sustentabilidad (Economía Circular), centrándose la mayoría

en el uso de dispositivos del Internet de las Cosas. Estos componentes de *Industria 4.0*, en su mayoría, enfocan sus esfuerzos en la relación Paciente-Personal médico, lo cual también muestra una brecha en la investigación para permitir que estos componentes sean implementados directamente en los nodos donde se toman las decisiones en la cadena logística. El *Aprendizaje Automático* tanto clásico como el *Aprendizaje Profundo* dominan las aplicaciones algorítmicas de *Inteligencia Artificial*, siendo las Redes Neuronales Artificiales el algoritmo más discutido entre los trabajos. Cabe mencionar que los autores tomaron especial empeño en demostrar cómo tales técnicas novedosas se distinguen de los algoritmos estadísticos tradicionales.

Es grato encontrar una saludable interacción entre el uso pasivo de algoritmos de *Inteligencia Artificial* para resolver un problema particular y concreto, con algoritmos capaces de asistir a la toma de decisiones en situaciones donde la complejidad del problema es tal que se requiere un análisis abstracto superior para poder encontrar soluciones pertinentes.

Con respecto a la Gestión de Conocimiento, se encontró que la mayoría de los trabajos no mencionan explícitamente la gestión de conocimiento, pero se puede inferir que la mayoría de los trabajos se enfocan en la generación de conocimiento a partir de los datos, lo cual es un paso importante para la gestión de conocimiento. La gestión de conocimiento es un área de oportunidad para futuras investigaciones, ya que se puede mejorar la toma de decisiones en la cadena logística de salud pública. Es significativo mencionar las metodologías de gestión de conocimiento utilizadas en los trabajos, puesto que no solo pueden ser de ayuda para futuras investigaciones basadas en los mismos, sino también para compaginar la producción de productos tecnológicos derivados por la investigación.

Principalmente, se encontró dentro de la literatura que existe un área de oportunidad para la realización de trabajos que incluyan tanto la toma de decisiones, los sistemas de información y su aplicación a la cadena logística de salud pública. Trabajos en esta área son de suma importancia, sobre todo porque proveen de los medios necesarios para que los países puedan crear políticas correctas de salud pública para aplicarlas en sus cadenas logísticas. Esto se traduce en una mejor respuesta ante los eventos cotidianos de la cadena logística, así como en habilitar a la cadena logística para responder correctamente ante eventos atípicos como la actual pandemia. Todo ello, con el fin último de mejorar y salvar vidas.

CAPÍTULO 3: METODOLOGÍA

3.1 Introducción

En este capítulo se describe la metodología propuesta para el desarrollo del presente trabajo de investigación. La metodología adoptada es incremental y puede contener fases iterativas, lo que permite una adaptación continua y mejora del proceso. Este enfoque es esencial para abordar la complejidad y dinamismo inherentes a los *Sistemas de Información en Salud* y su integración con los componentes de *Industria 4.0*.

La metodología se estructura en varias etapas clave, cada una de las cuales se centra en un aspecto específico del desarrollo del proyecto. Estas etapas incluyen la revisión del estado del arte, la modelación de casos de estudio, la determinación del flujo de datos, el diseño del prototipo y la construcción, validación e implementación del sistema. Cada una de estas etapas se detalla a continuación, proporcionando una visión clara y comprensiva del proceso metodológico.

El objetivo principal de esta metodología es garantizar que el sistema desarrollado sea robusto, seguro y capaz de integrarse de manera efectiva con las tecnologías emergentes de *Industria 4.0*. Además, se busca que el sistema facilite la gestión del conocimiento, permitiendo una toma de decisiones informada y eficiente en el ámbito de la salud pública.

3.2 Metodologías de trabajo

A continuación se muestra la metodología propuesta para este trabajo de investigación. Esta metodología es incremental y puede contener fases iterativas (**Figura 6**):

1. **Revisión del estado del arte:** Realizar un análisis iterativo de la literatura para determinar cómo evoluciona la generación de conocimiento relacionada con la solución del problema en estudio.
2. **Modelación de casos de estudio:** Para cumplir con los objetivos, se requiere integrar distintos casos de estudio dentro de un mismo sistema para permitirle realizar análisis, inferencia y predicciones generales.

3. **Determinación del flujo de datos:** Encontrar los caminos por donde fluyen los datos dentro de la cadena logística para encontrar formas de integración y posibles soluciones.
4. **Diseño del prototipo:** Diseñar los algoritmos, arquitecturas tecnológicas y computacionales, así como implementar metodologías de gestión de conocimiento.
5. **Construcción, validación e implementación:** Desplegar el sistema a los usuarios finales manteniendo la integridad del mismo mediante políticas continuas de validación de resultados.

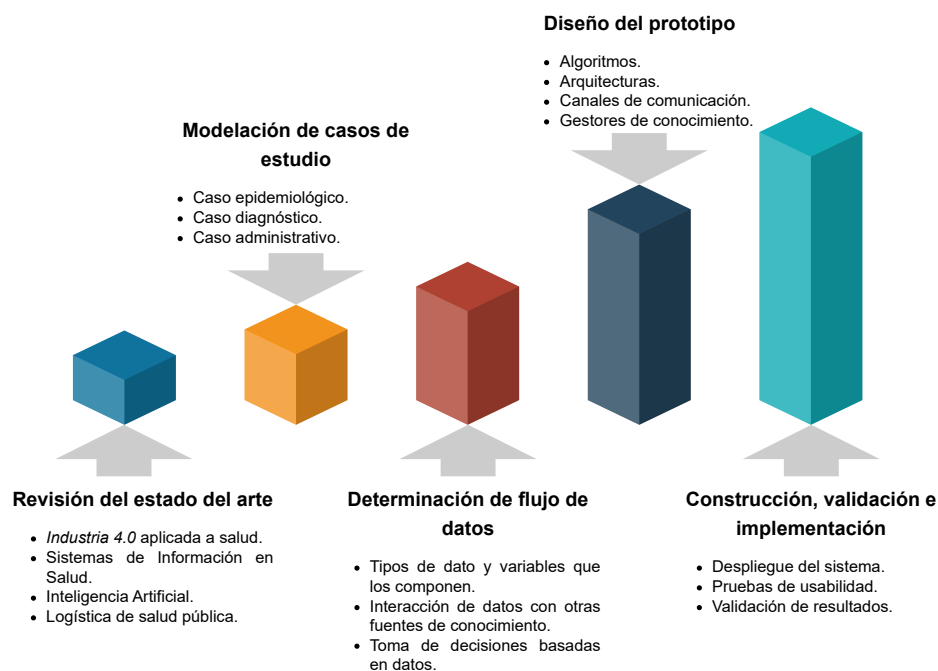


Figura 6: Diagrama incremental de la metodología

En la **Figura 7** se describe el proceso. Podemos observar cómo en cada paso se convierten datos crudos en conocimiento y posteriormente en sabiduría, y cómo este proceso tiene su paralelo en la programación del sistema.

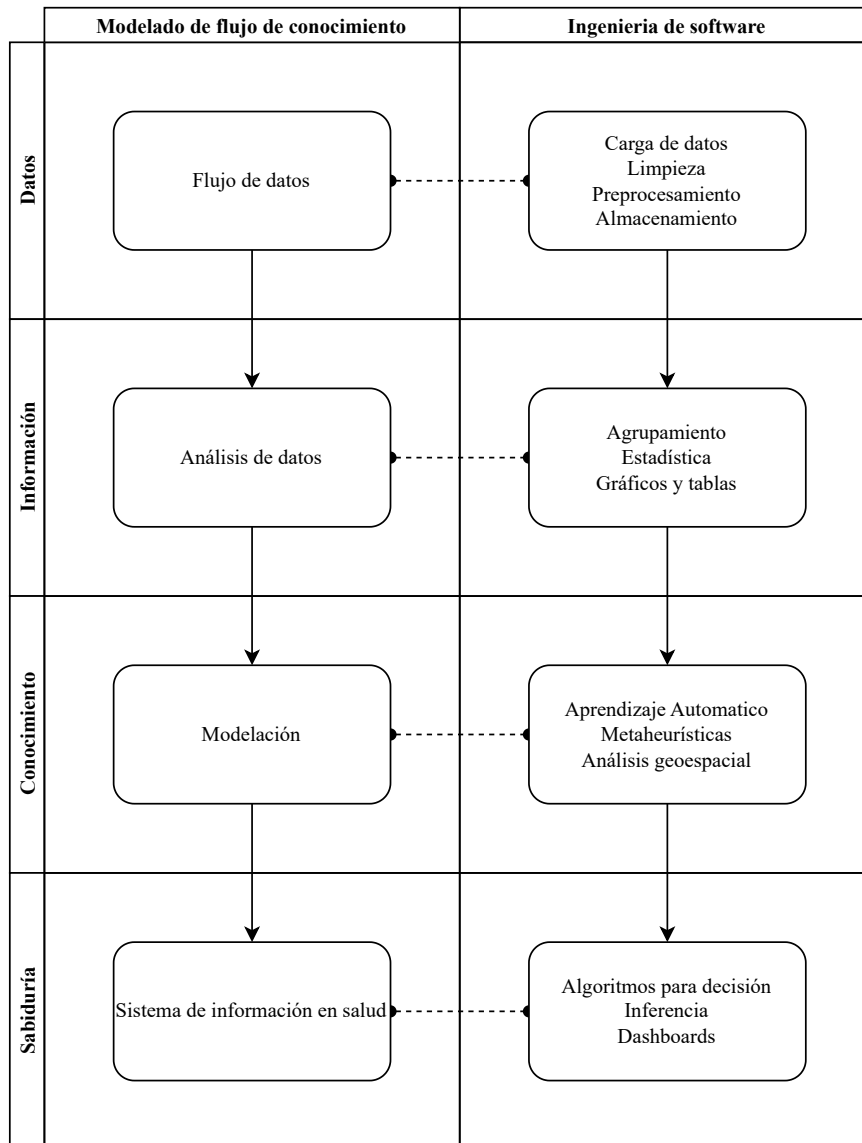


Figura 7: Diagrama de modelado

Los análisis basados en datos pueden ser de cuatro tipos, todos ellos pueden ser realizados dentro del sistema y están íntimamente relacionados con el sistema propuesto. El **Análisis Descriptivo** permite conocer qué sucede (con la *Industria 4.0*, esto puede ser en tiempo real) dentro de la cadena logística. Este análisis está basado en el procesamiento preciso y completo de los datos generados por eventos dentro de la cadena, por ejemplo, detectar un súbito incremento en la incidencia de cierta enfermedad. El **Análisis Diagnóstico** se enfoca en las causas iniciales de los eventos, analizando los datos de manera general pero aislando eventos

para proceder acorde. El **Análisis Predictivo** abre una ventana que ofrece vistazos ponderados al futuro, requiriendo el uso de datos históricos para detectar patrones de comportamiento. Finalmente, el **Análisis Prescriptivo** provee a los tomadores de decisiones dentro de la cadena logística de sugerencias sobre posibles soluciones en caso de que los análisis previos hayan encontrado problemas dentro de la cadena logística de salud (Figura 8). Podemos observar que a mayor complejidad, mayor valor.

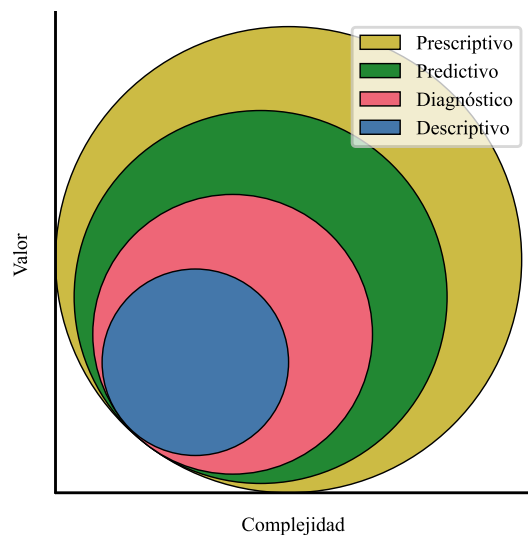


Figura 8: Tipos de análisis

Para el modelado del flujo de datos y su posterior conversión a conocimiento, nos apoyamos en el modelo *DIKW*¹ que significa: *Datos, Información, Conocimiento y Sabiduría*. Pretende encapsular la relación jerárquica entre estos componentes y define la extracción de sabiduría a partir de datos.

Los *Datos* son meras señales o símbolos, pueden ser útiles o puro ruido, carecen de organización y procesamiento; se podrían definir como inertes e inútiles. La *Información* surge cuando los datos se comienzan a hacer útiles; cuando se organizan, clasifican o agrupan, permitiendo el reconocimiento de patrones. Pasar al *Conocimiento* requiere aprendizaje, es decir, es un panorama mental que surge de la interacción entre la información y agentes que permiten su análisis. La *Sabiduría* es la punta de la jerarquía y sus tareas son realizar predicciones a futuro, ofrecer soluciones y razones de los eventos.

¹*Data, Information, Knowledge, Wisdom*

En el diagrama que muestra la **Figura 9**, podemos observar que el comportamiento de la decisión cambia de reacción o pasividad a anticipación entre más arriba nos encontremos en la jerarquía. Esto quiere expresar que entre más arriba se esté, mayor dinamismo existe. Cada uno de los escalones de la pirámide pretende contestar distintas preguntas y tratar distintos puntos temporales. Al subir en la pirámide también cambia el riesgo en la decisión, reduciéndose al tomar decisiones mejor informadas; el entendimiento también se incrementa más arriba. De la punta, surgen los modelos capaces de asistir a los tomadores de decisiones. (Ackoff, 1989)

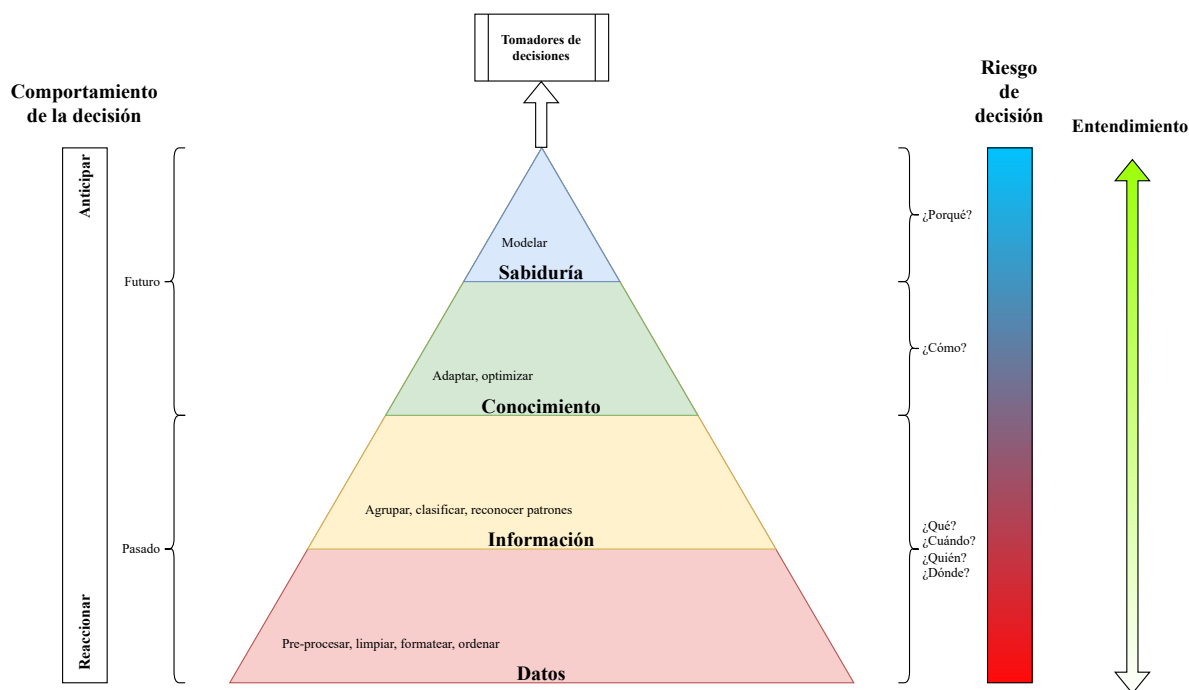


Figura 9: Jerarquía DIKW

Computacionalmente, los avances recientes en hardware y software han permitido bajar los costos de almacenamiento de datos, haciendo posible almacenar una gran cantidad. Esto da pauta a ejecutar mejores análisis basados en la jerarquía. Esta jerarquía tiene análogos en la *Ciencia de Datos* que van más allá del simple almacenamiento. Por ejemplo, los *Almacenes de Datos* ofrecen un repositorio de información, mientras que análisis estadísticos pueden generar conocimiento; culminando con la aplicación de algoritmos de *Inteligencia Artificial*, que pueden generar modelos capaces de encapsular *Sabiduría*. (IBM, 2020)

Este modelo ha sido aplicado con éxito en Salud y Ciencia de Datos, ofreciendo mejores

cimientos para modernizar la medicina basada en evidencia, proveer contexto en decisiones médicas y apoyar la toma de decisiones, muchas veces de vida o muerte. (Dammann, 2018)

3.3 Conclusiones

En este capítulo se ha presentado la metodología propuesta para el desarrollo del presente trabajo de investigación. La metodología es incremental y puede contener fases iterativas, lo que permite una adaptación continua y mejora del proceso.

En resumen, la metodología propuesta proporciona una estructura sólida y adaptable para abordar los desafíos del proyecto. La integración de componentes de *Industria 4.0* y el enfoque en la gestión de conocimiento son elementos clave para el éxito del trabajo. La implementación de esta metodología permitirá desarrollar soluciones innovadoras y efectivas para mejorar la gestión de la salud pública.

CAPÍTULO 4: RESULTADOS

4.1 Introducción

Para crear el **Sistema Inteligente**, fue necesario modelar correctamente el flujo de datos dentro de la cadena logística de salud. Esto incluyó detectar dónde surgen los datos, cómo manipularlos, determinar qué preguntas y respuestas se pueden extraer de ellos, identificar variables y datos auxiliares, y establecer la relación y jerarquía entre ellos. Asimismo, para la creación del prototipo, se requirió conocer los formatos de los datos, los requisitos para ingresarlos a una base de datos, el diseño de la misma, la arquitectura del sistema, y los módulos y bibliotecas de código. Por separado, estos procesos pueden consumir bastante tiempo. Es aquí donde la metodología de modelado (**Figura 7**) aplica para agilizar el proceso.

Cada uno de los casos de estudio se desarrolló de tal manera que apoyara la creación del Sistema Inteligente propuesto. Esto se logró mediante la generación de experiencia para implementar ciertos algoritmos, la realización de análisis particulares o la manipulación de componentes complejos y la relación entre ellos.

La pericia y el conocimiento generados en la resolución de cada caso de estudio permitieron diseñar un **Sistema Inteligente** más robusto que, dentro del prototipo, implementa todas las soluciones anteriormente presentadas.

En la **Figura 10**, se puede observar el diagrama conceptual de este proceso. Se desglosa la relación entre los Sistemas de Gestión y de Información, los problemas que presenta cada uno y cómo, al complementarlos con los componentes de *Industria 4.0*, se puede generar la implementación de un **Sistema Inteligente para la Logística de Salud Pública**. Para cada uno de los casos de estudio, se muestran los tipos de datos utilizados, los análisis realizados y la solución final propuesta.

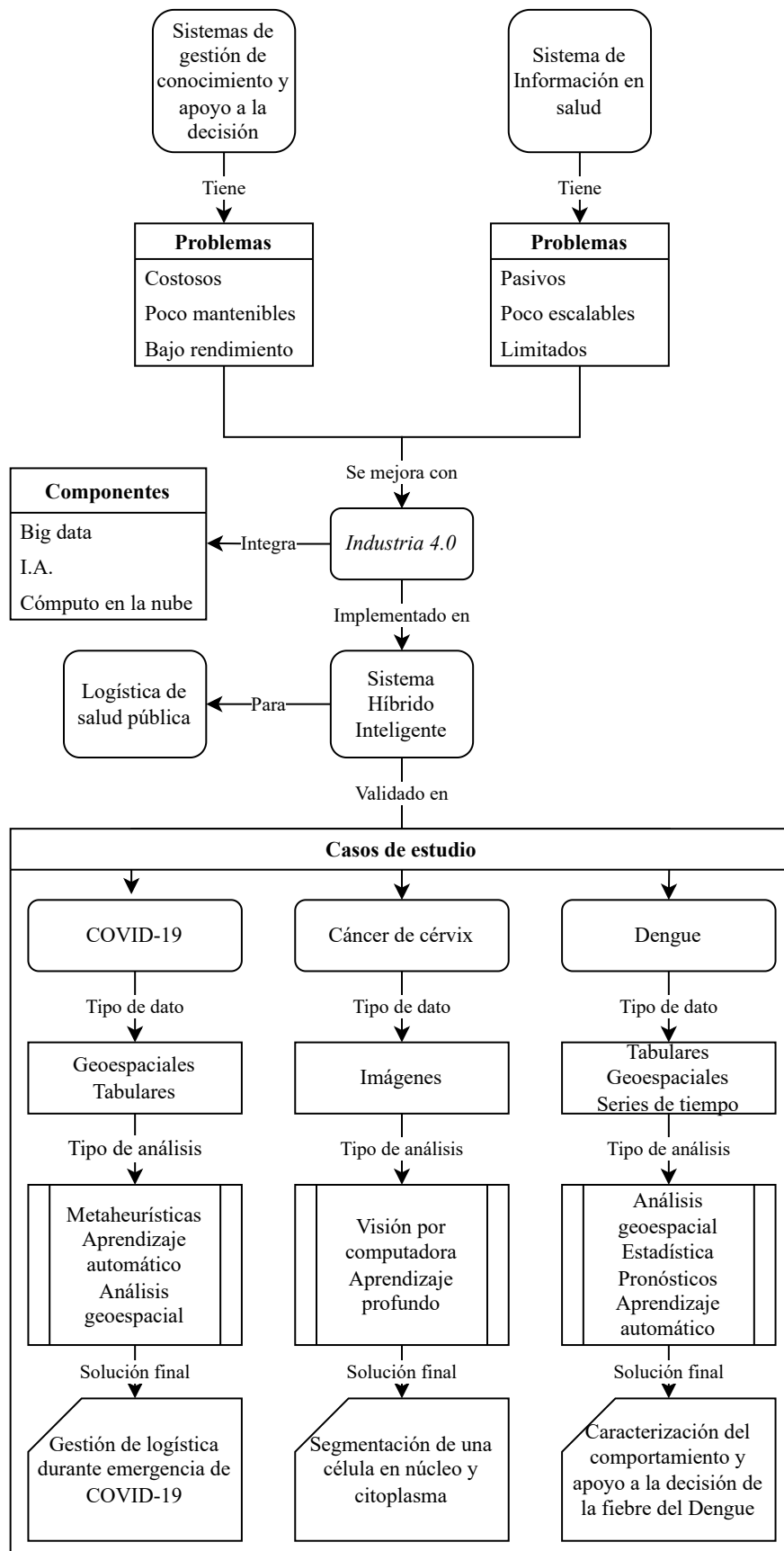


Figura 10: Diagrama de resultados

4.2 Caso de estudio: Aplicación inteligente para determinar la existencia de medicinas asociadas con COVID19

Los *Sistemas Híbridos Inteligentes* permiten la interacción de distintos algoritmos para realizar una o varias actividades en común. Al integrar diversas disciplinas, algoritmos y tecnologías, requieren de una orquestación correcta para que puedan funcionar de manera adecuada. El más crucial de estos procesos, es gestionar los datos de manera tal que puedan ser manipulados y transformados para iniciar el proceso de extracción de conocimiento. A continuación se expone la forma en la que, dada una base de datos, se pueden tomar estos y ser transformados de manera tal que de una misma fuente de datos se puedan aplicar distintos algoritmos inteligentes para proveer soluciones holísticas sobre algún problema en específico.

Los datos geoespaciales, son fundamentales para determinar el comportamiento de cualquier enfermedad. Estos se representan por coordenadas geográficas que pueden tomar forma de puntos, líneas, polígonos o mallas. Estos datos se pueden utilizar para determinar la ubicación de un evento, la relación entre eventos, la distribución de eventos, la frecuencia de eventos y, en particular, en la optimización de rutas de distribución.

Operativamente, los datos se concentran en repositorios de datos, los cuales pueden ser de distintos tipos. Los más comunes son los Data Warehouses, los Data Lakes y las Bases de Datos. Sin embargo, en su forma natural, no siempre es posible usarlos de manera directa, se requiere un paso previo para poder descubrir dentro de ellos, información relevante.

El problema de encontrar una ruta óptima, es un problema de tipo *NP-Completo*, es decir, pertenece a una clase de problemas que son sumamente difíciles de resolver debido a la gran complejidad computacional inherente a los mismos. Es por ello que se requieren algoritmos de optimización capaces de encontrar soluciones subóptimas en tiempos razonables. Estos algoritmos se conocen como metaheurísticas.

Este sistema es capaz de integrar distintos tipos de algoritmos para ofrecer una solución en particular. Sirviendo al **Sistema Inteligente** final como una prueba de concepto para la orquestación correcta de múltiples algoritmos y tecnologías para que actúen al unísono.

4.2.1 Descripción del problema

El crecimiento poblacional excesivo en las áreas metropolitanas ha superado las capacidades institucionales de los gobiernos, incrementando la demanda de servicios públicos, especialmente en el ámbito de la salud pública. Es urgente coordinar las “Zonas Metropolitanas Complejas”, como es el caso de León, donde la dinámica de la Zona Metropolitana excede las capacidades institucionales del municipio debido a la insuficiencia de recursos disponibles para atender a la población (Gräßer et al., 2017).

A nivel nacional, no existe un protocolo de inventario y distribución adecuado que permita obligar a las instituciones a organizarse para implementar planes y programas en beneficio de la ciudadanía. Así, la coordinación intergubernamental en el área metropolitana de León se produce por acuerdo de voluntades, más que en cumplimiento de la normativa existente.

Se conceptualizó el prototipo de una herramienta inteligente capaz de encontrar patrones ocultos en el inventario de los dispensarios, con la que se generarán las rondas a evaluar por una sociedad basada en *Algoritmos Culturales* para determinar la mejor ruta.

Este caso de estudio se basa en la necesidad de crear un concepto de *Sistema Híbrido Inteligente*, que reúna técnicas informáticas avanzadas como la *Minería de Datos* y los *Algoritmos Culturales* para determinar los sitios más susceptibles de la ciudad de León y además proporcionar a las autoridades un entorno óptimo para dar un mejor servicio a la sociedad mejorando su logística inteligente.

En particular, los problemas a resolver son los siguientes:

- Determinar los clústeres con mayor probabilidad de que se produzca un acto de infección en un momento dado.
- Optimizar la ruta logística para la correcta distribución de medicamentos.
- Generar un mapa que muestre la ruta que debe tomar cada vehículo asociado a la logística humanitaria.
- Implementar el sistema de una prueba de concepto.

4.2.2 Conceptualización

Se comenzó con la creación de una red que contenía la ubicación de las colonias en la ciudad de León y los lugares donde cada una de ellas estaba conectada, así como la distancia entre esos nodos dentro del mapa. Una vez terminada la red, se continuó con la construcción de los *Almacenes de Datos*, los cuales estaban integrados por los siguientes campos:

1. Año [Numérico].
2. Mes [Numérico].
3. Día [Numérico].
4. Hora [Numérico].
5. Riesgo [Numérico].
6. Inventario total de medicamentos de los dispensarios [Numérico].

Riesgo se refiere a la probabilidad de contagio basada en la cantidad de casos de dengue en la colonia. Mientras que el inventario total de medicamentos es una basada en la cantidad de medicamentos que comúnmente se utilizan para tratar los síntomas de las vías respiratorias (Tabla 1).

Tabla 1: Medicamentos para tratar enfermedades de las vías respiratorias

Categoría	Medicamentos
Antiinflamatorios	Ibuprofeno, Naproxeno, Paracetamol
Broncodilatadores	Salbutamol, Ipratropio, Teofilina
Antibióticos	Amoxicilina, Azitromicina, Cefuroxima
Antitusígenos	Dextrometorfano
Expectorantes	Guaifenesina, Ambroxol
Antihistamínicos	Loratadina, Cetirizina
Descongestionantes	Fenilefrina, Oximetazolina, Nafaazolina
Corticosteroides inhalados	Beclometasona, Budesonida, Fluticasona

El prototipo utilizó una base de datos generada aleatoriamente de 1000 registros para realizar funciones del sistema como la clasificación de datos, la agrupación, la generación de recorridos y la elaboración de mapas.

La *Minería de Datos* está dentro del conjunto de técnicas *KDD*¹. Fallad et. al. definen la *Minería de Datos* como “la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos” (Fayyad, 1996).

El *Almacén de Datos*² es de suma importancia en este tipo de problemas, este puede ser percibido como un concepto previo a la minería y se basa en ella para convertir datos en información. Un *Almacén de Datos* se compone de datos operativos y está diseñado para facilitar la toma de decisiones. La información almacenada en él es constantemente actualizada y solo se habilita para consultas; integra y hace consistentes los datos extraídos de las bases de datos operativas; por tanto, el DWH es un almacén de datos estructurado para la ejecución y generación de informes relacionados con la actividad de una organización. (Gill & Rao, 1996)

La *Minería de Datos* y el *KDD* integran desarrollos y conceptos procedentes de la estadística, el *Aprendizaje Automático*, la visualización de datos y la teoría de las bases de datos (Figura 11). Esta fusión de disciplinas muy diversas ha sido motivada (entre otras cosas) por el importante aumento del volumen de datos en todos los ámbitos de la actividad humana. Este un proceso iterativo de búsqueda de información no trivial en grandes volúmenes de datos que busca generar información similar a la que podría generar un experto humano: patrones, asociaciones, cambios, anomalías y estructuras significativas (Figura 12) (Abonyi et al., 2005).

¹Knowledge Discovery in Databases o Descubrimiento de conocimiento en bases de datos

²Data Warehouse en inglés

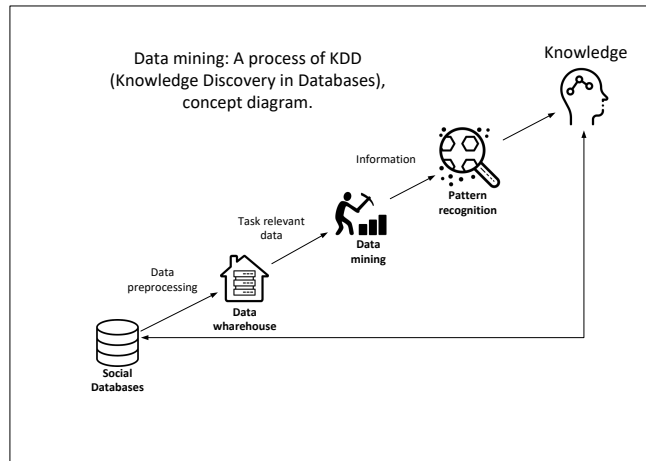


Figura 11: Descubrimiento de conocimiento usando minería de datos

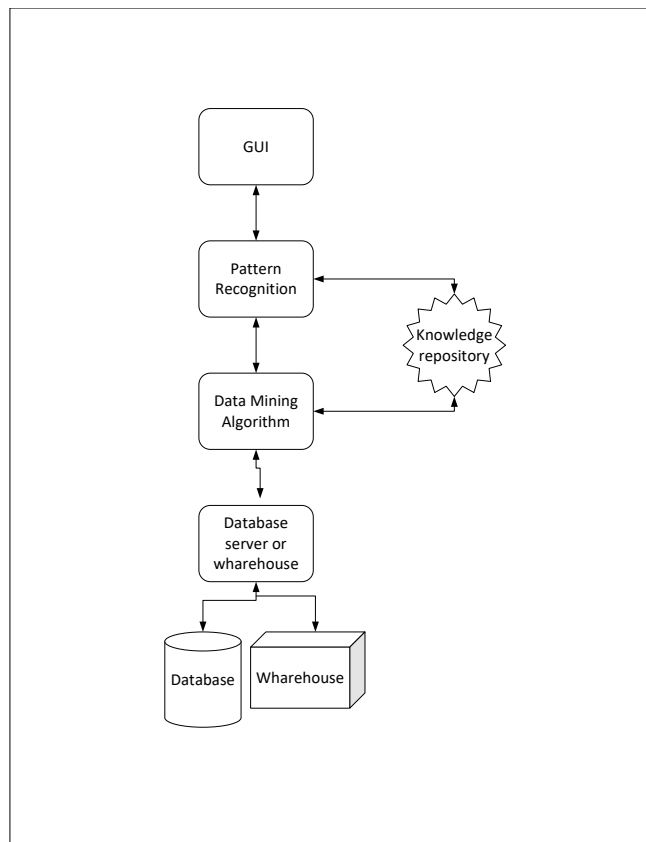


Figura 12: Proceso de minería de datos

El algoritmo *K-means* es un método de agrupación no supervisado que divide un conjunto de datos en K grupos (clúster) basándose en la similitud de los valores de los atributos de los

diferentes datos. El objetivo principal del algoritmo es minimizar la varianza total intragrupo, es decir, la suma de las distancias cuadradas entre los puntos de datos y el centroide de su grupo correspondiente. El proceso comienza con la selección de K centroides iniciales, que pueden ser elegidos de manera aleatoria o utilizando algún método heurístico. Estos centroides actúan como los puntos de referencia para los clústeres.

La ecuación del algoritmo es la siguiente:

$$\arg \min_S \sum_{i=1}^K \sum_{x \in S_i} \|x - \mu_i\|^2$$

donde:

- $S = \{S_1, S_2, \dots, S_K\}$ es la partición del conjunto de datos en K clusters.
- x es un punto de datos.
- μ_i es el centroide del cluster S_i .
- $\|x - \mu_i\|^2$ es la distancia euclidiana al cuadrado entre el punto de datos x y el centroide μ_i .

Una vez que los centroides iniciales han sido seleccionados, el algoritmo procede a asignar cada punto de datos al clúster cuyo centroide esté más cercano, utilizando una medida de distancia, como la distancia euclidiana. Esta asignación se realiza para todos los puntos de datos en el conjunto, formando así K clústeres iniciales. Después de la asignación, se recalculan los centroides de cada clúster como el promedio de los puntos de datos asignados a ese clúster. Este proceso de asignación y actualización de centroides se repite iterativamente.

El algoritmo continúa iterando entre la asignación de puntos a clúster y la actualización de centroides hasta que los centroides no cambien significativamente entre iteraciones consecutivas o se alcance un número máximo de iteraciones. El resultado final es una partición del conjunto de datos en K clústeres, donde cada clúster está representado por su centroide y los puntos de datos dentro de cada clúster son más similares entre sí que a los puntos de otros

clústeres. Este proceso permite identificar patrones y estructuras en los datos, facilitando el análisis y la interpretación de grandes volúmenes de información (Figura 13).

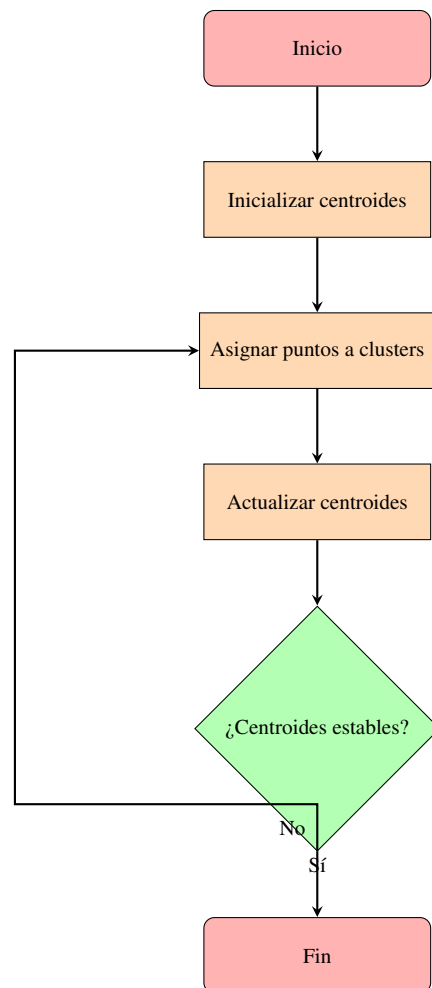


Figura 13: Diagrama de flujo del Algoritmo K-means

Con la aplicación de estas herramientas de minería de datos, el prototipo puede determinar (dada una fecha concreta) los centroides con mayor riesgo en un momento dado. Las colonias se agregan en los clústeres y este representa el punto de interés para la optimización de la ruta. Para determinar la colonia en específico, se toma el centroide del polígono de la colonia y se toma el clúster más cercano en distancia euclidiana.

Una vez determinadas estas colonias, se genera una sociedad basada en *Algoritmos Culturales*, que será el encargado de determinar, a lo largo del tiempo, la ruta más óptima. En la época cero, cada agente es inicializado con la información de cada colonia, a la que cada uno de

ellos propondrá una ruta, dando lugar a la negociación entre los agentes para seleccionar la mejor ruta propuesta en esa época, el espacio de creencias se actualizará solo cuando la ruta propuesta sea mejor que la previamente almacenada en el espacio de creencias, iniciando un ciclo de mejoras que se interrumpirá cuando se produzca M cantidad de épocas sin mejoras en las rutas (Hartigan & Wong, 1979).

Los *Algoritmos Culturales* son una clase de modelos computacionales derivados de la observación del proceso de evolución cultural en la naturaleza. Los *Algoritmos Culturales* tienen tres componentes principales: un espacio de población, un espacio de creencias y un protocolo que describe cómo se intercambian los conocimientos entre los dos primeros componentes. Los Algoritmos Culturales son un sistema patrimonial dual que caracteriza la evolución en la cultura humana a nivel macro-evolutivo, que ocurre dentro del espacio de creencias, y a nivel micro-evolutivo, que ocurre en el espacio poblacional. El conocimiento producido en la población que el espacio a nivel micro-evolutivo es aceptado o pasado al espacio de creencias y utilizado selectivamente para ajustar las estructuras de conocimiento (Maheri et al., 2021).

- **Espacio de Población:** Este componente representa a los individuos de la población, cada uno con sus propias soluciones candidatas al problema. La evolución a nivel micro-evolutivo ocurre dentro de este espacio, donde los individuos se someten a procesos de selección, reproducción y mutación.
- **Espacio de Creencias:** Este componente almacena el conocimiento acumulado de la población. La evolución a nivel macro-evolutivo ocurre dentro de este espacio, donde las creencias se actualizan y se utilizan para guiar la evolución de la población.
- **Protocolo de Intercambio de Conocimientos:** Este protocolo define cómo se transfieren los conocimientos entre el espacio de población y el espacio de creencias. El conocimiento producido en la población a nivel micro-evolutivo es aceptado o pasado al espacio de creencias y utilizado selectivamente para ajustar las estructuras de conocimiento.

El *Algoritmo Cultural* se describe en el diagrama de la **Figura 14**:

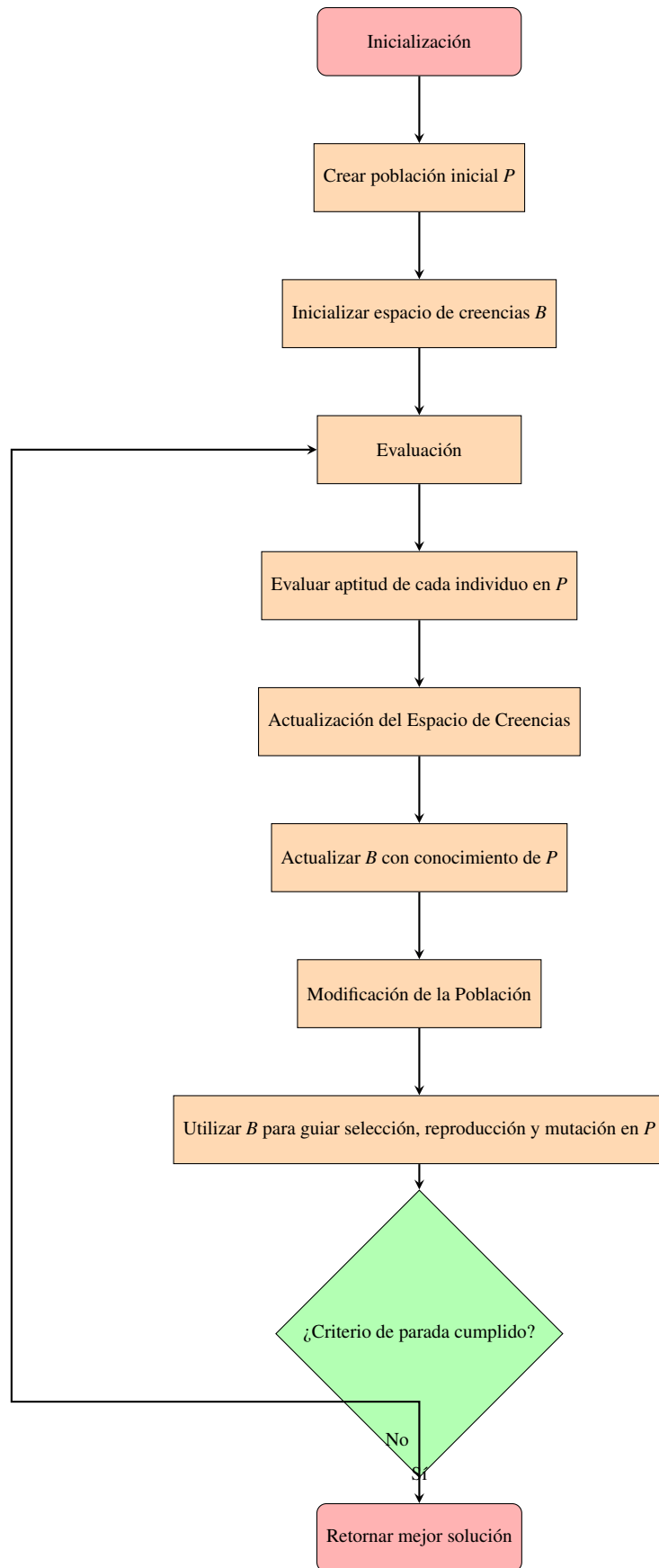
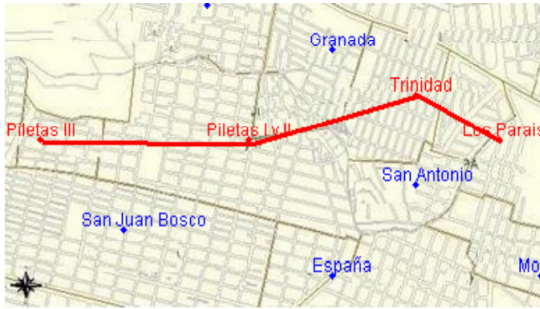


Figura 14: Diagrama de flujo del Algoritmo Cultural

Este conocimiento puede utilizarse entonces para influir en los cambios realizados por la población en la siguiente generación. Los *Algoritmos Culturales* utilizan cinco tipos de conocimientos básicos para resolver el problema, en lugar de limitarse al valor transmitido. Las fuentes de conocimiento incluyen el conocimiento normativo (rangos de comportamiento aceptable), el conocimiento circunstancial (ejemplos o recuerdos de soluciones exitosas y fallidas entre otros), el conocimiento del dominio (conocimiento de los objetos del dominio, sus relaciones y sus interacciones), el conocimiento histórico (patrones temporales de comportamiento) y el conocimiento topográfico (patrones espaciales de comportamiento).

Como ejemplo de los resultados, podemos citar el caso en el que el usuario solicita la generación de un recorrido proporcionando un rango de fechas y el número de vehículos asociados a la Logística Humanitaria requerida en la pandemia disponible en un rango de 1 a 20.

El sistema determinará las colonias necesarias para los actos de emergencia médica registrados en el rango de fechas especificado. Basándose en el número de colonias y el número de vehículos asociados a la Logística Humanitaria requerida en la pandemia. Utilizando el algoritmo *K-Means*, se generarán clústeres que serán entregados inmediatamente al Algoritmo Cultural, que determinará la ruta más óptima para cada clúster, generando un mapa que muestra la ruta que debe tomar cada vehículo asociado a la logística humanitaria requerida en la pandemia. Como se puede observar en la [Figura 15](#), al paso de seis periodos, el *Algoritmo Culturale* fue suficiente para determinar una ruta óptima entre las colonias dentro de los diferentes clústeres ([Figura 15a](#)), destacando como la distancia de la ruta disminuye a medida que avanzan los periodos ([Figura 15b](#),) y al no haber una ruta mejor propuesta en los últimos tres periodos, los agentes optan por finalizar y mostrar el recorrido ([Figura 15c](#)).



(a) Ruta entre colonias



(b) Avance en los periodos



(c) Ruta finalizada

Figura 15: Resultados de la optimización

Actualmente, el sistema está desplegado en una herramienta basada en línea de comandos. La cual recibe los siguientes parámetros: archivo de datos, directorio de salida, número de clústeres, número de épocas.

Para la ejecución del proceso se utilizaron 16 núcleos de procesamiento y 64GB de memoria RAM. Una ventaja de usar Java, es que es fácilmente paralelizable y puede gestionar correctamente los hilos de procesamiento, por lo que no se tuvo necesidad de implementar un código muy elaborado. El tiempo de ejecución fue de 2 minutos y 43 segundos, lo cual es aceptable para el tamaño de la base de datos utilizada (Figura 16).

```
WSL at ~ 45ms
00:51:38 bash compute.sh --input simulated_data.csv --output outputs-2/ --K 2 --epochs 10
Algoritmo Híbrido

Implementando K-Means/Algoritmos culturales.
Clustering ...
Optimizing ...
Writing file:
Image 1: Mon Nov 18 00:54:31 CST 2024
Image 2: Mon Nov 18 00:54:32 CST 2024
Image 3: Mon Nov 18 00:54:33 CST 2024
Output folder: outputs-2/
WSL at ~ 2m 43.382s
00:54:34
```

Figura 16: Ejemplo de uso de la línea de comandos

4.2.3 Discusión

El análisis de este caso de estudio demostró que no solo es factible la aplicación de la *Minería de Datos* a la información de los dispensarios de la ciudad de León, sino también su alto valor añadido para el análisis y generación de nuevo conocimiento que, junto con los *Algoritmos Culturales* encargados de la creación de rutas de distribución, fundamentan el nuevo conocimiento adquirido, proporcionando un soporte para ser realizado con base en hechos históricos, no al azar como ocurre en casi todas las organizaciones sanitarias, con ello, pretendemos reabastecer áreas importantes proporcionando un nuevo paradigma para una mejora sustancial en la logística de la sanidad pública. En nuestra investigación futura es considerable considerar un cuadro de mando inteligente o *dashboard* asociado a la Logística Humanitaria en un dispositivo móvil, similar al propuesto en (Xia & Lee, 2020).

Se requiere una exploración adicional de algoritmos para aumentar la capacidad de respuesta y las capacidades de la solución. Nuevas técnicas de *Inteligencia Artificial* como el *Aprendizaje Profundo*, han sido implementadas con éxito para detectar y diagnosticar COVID-19 de pacientes. Se proporciona un buen ejemplo, utilizando uno de los algoritmos más populares de Deep Learning para clasificar radiografías; una solución como esta puede ser implementada dentro de una aplicación de inteligencia ambiental (Dansana et al., 2020).

Los algoritmos de predicción también podrían servir de información previa al *Algoritmo Cultural* para mejorar su capacidad de una aplicación espacial a una temporal. Estos algoritmos podrían incluir *Aprendizaje Profundo* o algoritmos más maduros como *Lógica Difusa*. Aplicar la previsión a las series temporales permitiría a los responsables de la toma de decisiones

no solo planificar los casos de COVID-19, sino que también pueden determinar la tasa de agotamiento de medicamentos dentro de los dispensarios.

Por las condiciones de colaboración del proyecto, se tuvo que desarrollar en la herramienta Weka y ser programado en lenguaje Java. Este lenguaje, famoso por su robustez y seguridad, no es el más adecuado para el desarrollo de aplicaciones web, y carece de ciertas características que no lo hacen partícipe como principal dentro de la *Industria 4.0*. Esta importancia actualmente está reservada en lenguajes como Python, que se ha convertido en el lenguaje de programación más usado del mundo como ya se ha expuesto. Esto limita las capacidades de integración de la propuesta generada en este caso de estudio y evidencia la necesidad de elegir el conjunto de tecnologías adecuados para el desarrollo del **Sistema Inteligente**.

Con respecto al estado del arte, distintos trabajos utilizan *Metaheurísticas* en sus procesos de capitalización y extracción de conocimiento. (Al-Zoubi et al., 2019) utiliza un novedoso modelo de optimización basado en poblaciones de ciertos organismos marinos. Gergin y su equipo (Gergin et al., 2019) fueron capaces de utilizar el algoritmo inspirado por abejas para determinar la mejor locación para instalación de desechos médicos. *K-Means* fue utilizado (Knaus & Marks, 2019a) para la caracterización de fenotipos bacterianos. Ninguna de estas aplicaciones existe bajo un marco general de aplicación bajo el paradigma de *Industria 4.0*, mientras que esta conceptualización pretende existir dentro de un sistema basado en ese paradigma, en concreto, percibir las cadenas logísticas de salud dentro de una perspectiva de Sistemas Híbridos Inteligentes, lo cual hace de esta propuesta algo sumamente pertinente para el prototipo que se está desarrollando.

Finalmente, este caso de estudio aporta al **Sistema Inteligente** lo siguiente:

- Implementación de análisis geoespacial aplicado al área médica.
- Análisis geoespacial utilizando técnicas de aprendizaje automático.
- Optimización de problemas NP-completos usando *Metaheurísticas*.
- Pericia en la gestión de *Sistemas Inteligentes Híbridos*.
- Aplicación de los conceptos de *Minería de Datos* y *KDD*.

- Uso de algoritmos de *clustering* en el área médica.

4.3 Caso de estudio: Segmentación de Células Cervicales

Los datos tabulares son los más comunes y son la base de estudios estadísticos y de series de tiempo. Sin embargo, en aplicaciones reales, los datos generados dentro de cualquier cadena logística pueden tomar cualquier forma. En las cadenas logísticas médicas, los datos pueden tomar distintos formatos, los más extremos, son los datos en forma de señal como un electroencefalograma y los generados por procesos de imagenología médica. Estos últimos son de suma importancia para la cadena logística de diagnóstico, donde una correcta interpretación de la imagen es crucial para la salud del paciente.

El diagnóstico asistido por imágenes médicas, es un proceso generalizado y que existe en distintas áreas de la medicina. Las imágenes médicas pueden ser generadas por dispositivos como máquinas de rayos X, tomografías, resonancias magnéticas o distintos tipos de microscopio. Pueden capturar distintas estructuras dentro y fuera del cuerpo humano como huesos, tejidos o células individuales.

Poder convertir imágenes, representadas computacionalmente, en conocimiento médico, es tarea del procesamiento de imágenes médicas asistido por computadora, el cual es un subconjunto de la *Visión por Computadora*.

Este caso de estudio es una continuación de esfuerzos previos enfocados en asistir computacionalmente la toma de decisiones médicas, en particular, apoyo al diagnóstico de cáncer cervicouterino. Cuyo diagnóstico depende de la correcta caracterización de muestras celulares tomadas del paciente. Anteriormente, estos esfuerzos eran tediosos y poco prácticos, ya que requerían métodos compuestos de múltiples y complejos algoritmos operando al unísono para extraer información de las imágenes, procesarla y poder extraer conocimiento. En la actualidad, modernos algoritmos de *Inteligencia Artificial* (un componente importante de la *Industria 4.0*), son capaces de extraer conocimiento de las imágenes sin requerir de complejos pasos de procesamiento, permitiendo una integración eficiente dentro de las cadenas logísticas. Esto es relevante porque un proceso sencillo de operación, es más fácil de ser implementado y adoptado por los tomadores de decisiones.

Estos algoritmos, conocidos como *Aprendizaje Profundo* o *Deep Learning*, han detonado una innovación sin igual y básicamente gobiernan las aplicaciones de *Inteligencia Artificial* en el mercado. Las Redes Neuronales Convolucionales, han ofrecido los mejores rendimientos en todas las tareas de *Visión por Computadora*; este rendimiento es comparado y medido mediante la aplicación de *Benchmarks* como ImageNet.

Para la tarea de procesamiento computacional de imágenes citológicas, el *Benchmark* es el conjunto de datos compilados por el Hospital Universitario de Herlev. El trabajo realizado en este caso de estudio, fue tomar este conjunto de datos y convertirlo en información, segmentar cada imagen en sus componentes, clasificando pixel por pixel. Este caso de estudio, contribuye significativamente al área de estudio de dos maneras; la primera, hasta ahora es el primer trabajo donde se segmenta una célula cervical en núcleo y citoplasma utilizando *Aprendizaje Profundo*; y segundo, el modelo aplicado es novedoso para el problema, debido a que conjuga arquitecturas nunca antes usadas para la tarea.

En un marco más general, esto permite al **Sistema Inteligente** la gestión de conocimiento en una de las cadenas logísticas de salud más importante, la del diagnóstico. La ventaja del uso de algoritmos de *Aprendizaje Profundo*, es que su implementación, fuera de requerimientos específicos de Hardware, estos modelos se integran fácilmente dentro de cualquier sistema o plataforma web, lo cual los esenciales para el desarrollo de un **Sistema Inteligente**.

4.3.1 Descripción del problema

El cáncer es un término que se refiere a muchas enfermedades relacionadas, derivadas de un crecimiento anormal en una parte específica del cuerpo. Si el cáncer no se trata, se puede esparcir rápidamente a otras parte del cuerpo y matar al paciente. Una de las características que definen la letalidad de esta enfermedad, es que es difícil de diagnosticar y caro de tratar. A nivel mundial el cáncer cobra casi 10 millones de vidas, siendo los países en desarrollo aquellos que sufren más de esta enfermedad.

El cáncer cérvicouterino, es el crecimiento anormal de las células de la parte baja del útero. Este tipo de cáncer es la cuarta causa de muerte por cáncer en mujeres y su incidencia es mayor en países en vías de desarrollo. Evoluciona progresivamente desde pequeños cambios

en la morfología celular, estos cambios se derivan de múltiples causas, incluyendo infecciones virales. En países desarrollados, la vacunación contra el Virus del Papiloma Humano ha derivado en la caída en la incidencia y las muertes, mientras que en países en vías de desarrollo las campañas de detección son la principal causa del declive (Sung et al., 2021).

Las técnicas de detección manual como la prueba del Papanicolau son baratas, accesibles y pueden detectar hasta el 95 % de los casos de cáncer. El proceso involucra la toma de una muestra mediante raspado del cérvix, la preparación de la muestra para análisis citológico y la examinación por el experto; cuyo trabajo es la categorización de las células presentes en la muestra en normales y anormales o determinar el grado de lesión.

Al incrementar la capacidad de aplicación de la prueba de Papanicolau, se comienzan a hacer evidentes sus problemas inherentes. Los factores que determinan la calidad de una muestra son: presencia de células sobrepuestas, moco y otros tipos de residuos y tejidos. Aparte de ellos, el factor humano también es responsable en la variabilidad en la calidad de las muestras, ya que el entrenamiento y mantenimiento de expertos consume bastante tiempo y requiere una inversión significativa de recursos. Los errores humanos también pueden derivar de la fatiga o el sobre trabajo; es aquí donde el diagnóstico asistido por computadora puede apoyar al experto en reducir errores y la carga de trabajo, así como reducir el tiempo de análisis al identificar las características morfológicas en las células y analizar sus partes constituyentes: el núcleo y el citoplasma.

Las características físicas del núcleo y del citoplasma son cruciales para clasificar cada célula en normal y anormal. La correcta segmentación de estas partes ayuda a los procesos automáticos y a los expertos humanos a identificar células normales, anormales y cancerígenas. Es por ello que la segmentación de células cervicouterinas es una importante área de interés.

4.3.2 Conceptualización

El conjunto de datos de Herlev es un punto de referencia adecuado para probar una gran variedad de algoritmos (Jantzen et al., 2005). Fue elaborado por el Dr. Jan Jantzen en el Hospital Universitario de Herlev, en Dinamarca, en colaboración con G. Dounias y su grupo de investigación en la Universidad del Egeo (Dounias, 2017). Se diseñó principalmente para

evaluar los algoritmos de clasificación, incluyendo las máscaras segmentación para el fondo, núcleo y el citoplasma de las células individuales, adecuadas para las tareas de segmentación. Las características de este conjunto de datos se muestran en [Tabla 2](#).

Tabla 2: Características de los datos Herlev

Categoría	Cantidad	Clase	Total
Carcinoma in situ (CS)	150	Anormal	675
Light dysplastic (LD)	182	Anormal	
Moderate dysplastic (MD)	146	Anormal	
Severe dysplastic (SD)	197	Anormal	
Normal columnar (NC)	98	Normal	242
Normal intermediate (NI)	70	Normal	
Normal superficial (NS)	74	Normal	
Total	917		

Las muestras de las celdas y las máscaras pueden verse en la [Figura 17](#). También muestra algunos de los retos que presenta este conjunto de datos para la *segmentación semántica*. Por ejemplo, hay una cantidad significativa de residuos en las imágenes. Además, hay múltiples células en la imagen, pero las máscaras no están etiquetadas correctamente; también se encuentran células superpuestas y algunas imágenes tienen mala calidad. Por último, las máscaras pueden tener defectos, como píxeles no etiquetados o una región gris que representa una región desconocida.

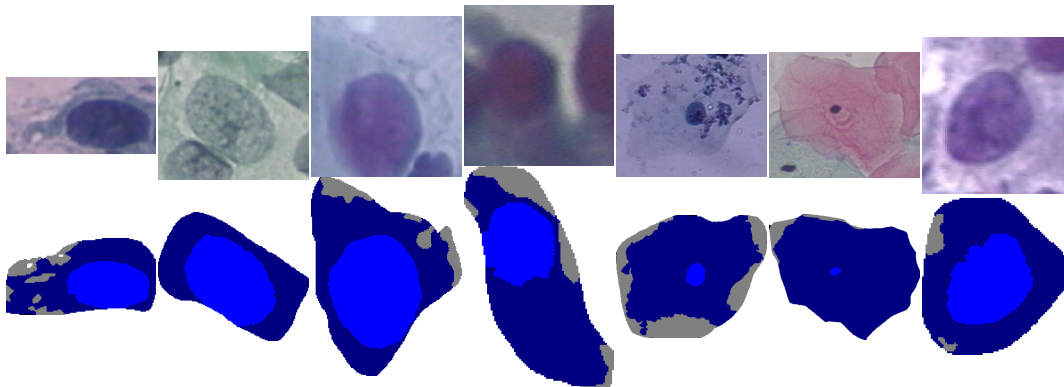


Figura 17: Muestreo de imágenes y máscaras

La *segmentación semántica* consiste en asignar a cada píxel de una imagen una etiqueta que represente la región de interés. Las *FCN*³ son arquitecturas particulares que constan de dos pasos integrados en una estructura decodificador/codificador: el *downsampling* y el *upsampling*, conectados por capas conocidas como cuello de botella. En primer lugar, el paso de *downsampling* reduce la resolución espacial tanto de la imagen como de la máscara, creando complejos mapeos de características. En segundo lugar, el *upsampling* remodela los mapeos a la resolución original, transformándolos en las máscaras de segmentación y recuperando la información perdida en el paso de *downsampling*. Esto se realiza gradualmente, apilando varias capas para capturar características cada vez más finas, como se necesita en una tarea de segmentación citológica. Ejemplos de estas arquitecturas son U-Net (Ronneberger et al., 2015), Feature Pyramid Networks (T.-Y. Lin et al., 2016), LinkNet (Chaurasia & Culurciello, 2017) y Pyramid Scene Parsing Networks (H. Zhao et al., 2016). Arquitecturas más complejas, como Mask-RCNN (He et al., 2020), combinan la *Segmentación Semántica* con la detección de objetos para producir la segmentación de instancias.

La **Figura 18** muestra un diagrama simplificado del proceso. En primer lugar, los datos de las imágenes del conjunto de datos de Herlev se preprocesan cambiando su tamaño y formato; las máscaras se adaptan para el proceso de aprendizaje creando matrices multidimensionales construidas concatenando matrices $2D$. A continuación, se introducen pesos de *Aprendizaje por Transferencia*. Por último, una etapa de evaluación en la que se aplican métricas para determinar el rendimiento del modelo.

³Redes Totalmente Convolucionales o *Fully Convolutional Networks*

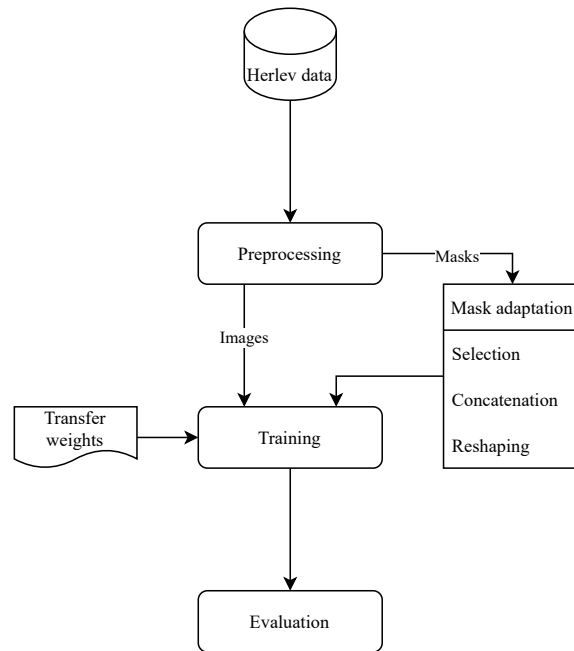


Figura 18: Metodología para la segmentación de células cervicales

Este proceso sirve como base incipiente para la creación de la metodología que se usó para el caso de estudio principal.

Como la tarea de segmentar los núcleos y el citoplasma es difícil, se necesita una arquitectura con gran capacidad. La selección cuidadosa de los componentes de la arquitectura es crucial para su rendimiento. Proponemos la construcción de un modelo de segmentación cervical basado en dos arquitecturas con rendimiento conocido dentro del estado del arte y una cuidadosa selección de hiperparámetros.

EfficientNet (Tan.2019) es una familia de arquitecturas de Redes Neuronales Convolucionales construidas para la clasificación de imágenes con un rendimiento superior en el Reto Imagenet (Russakovsky et al., 2014) y otros conjuntos de datos, superando a otras arquitecturas del estado del arte como Resnet (He.2015) y VGG19 (Simonyan & Zisserman, 2014). Alcanzó un 84.3 % de precisión en el top-1, y es 8.4 veces más pequeño y 6.1 veces más rápido que la segunda mejor arquitectura. Además, ajustando la escala del modelo, Mingxing Tan y su equipo produjeron varias arquitecturas de alto rendimiento variando características como la anchura, la profundidad, la resolución, etcétera. Su enfoque se basó en encontrar

un método sistemático para aumentar la capacidad de las redes neuronales convolucionales mediante el escalado y lograr una mayor precisión y eficiencia. *Compound Model Scaling* es un método capaz de maximizar la precisión mediante la optimización de la arquitectura de la red neuronal utilizando la memoria y los *FLOPs* como restricciones. Además, las arquitecturas de redes eficientes se componen de bloques básicos llamados *Inverted Bottleneck MBConv* (Figura 19) (Sandler et al., 2018).

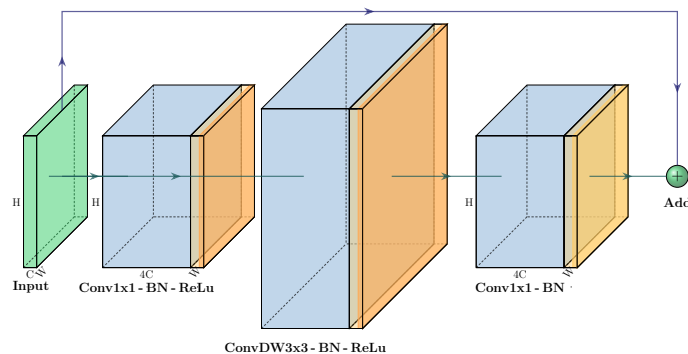


Figura 19: Módulo MBConv

El *Aprendizaje por Transferencia* permite a las redes reutilizar el conocimiento aprendido en otras tareas para mejorar su rendimiento en otra; esto se consigue transfiriendo los pesos del modelo y reentrenándolo en otro conjunto de datos. Cuanto mejor sea el rendimiento en una tarea, mejor será la transferencia y mejor será el rendimiento en la tarea posterior. (Kornblith et al., 2018). Qizhe Xie et. desarrolló un enfoque de aprendizaje semi-supervisado para mejorar el rendimiento en ImageNet utilizando arquitecturas existentes: *Noisy student* (Xie et al., 2019). Utilizando un conjunto de datos más amplio de imágenes sin etiquetar, su equipo mejoró el rendimiento del estado del arte hasta el 88.4 % de precisión en el top 1. Entrenaron a un agente profesor utilizando imágenes etiquetadas y lo utilizaron para generar pseudo-etiquetas asignadas a imágenes no etiquetadas, estas etiquetas pueden tener diferentes propiedades, luego entrenaron al agente estudiante y evaluaron el proceso utilizando tanto datos etiquetados como no etiquetados de manera iterativa.

Las *FPN* se desarrollaron inicialmente para la detección de objetos (T.-Y. Lin et al., 2016) y tienen un rendimiento de vanguardia en el desafío COCO (T. Y. Lin et al., 2014), un

Benchmark reconocido para la *detección de objetos* y la *Segmentación Semántica*. Esta arquitectura consigue extraer mapas de características proporcionales en múltiples niveles; al realizar este proceso independientemente del *downsampling*, puede generar fácilmente pirámides de características para la detección de objetos, la *Segmentación Semántica*. El *downsampling* construye una jerarquía de características; cada escala se correlaciona con un nivel de la pirámide, el *upsampler* recrea características de alta resolución, las conexiones laterales mejoran estas características y las fusionan con el *downsample*. Estas propiedades hacen del *FPN* un extractor de características genérico capaz de segmentar con precisión imágenes del mundo real (Figura 20).

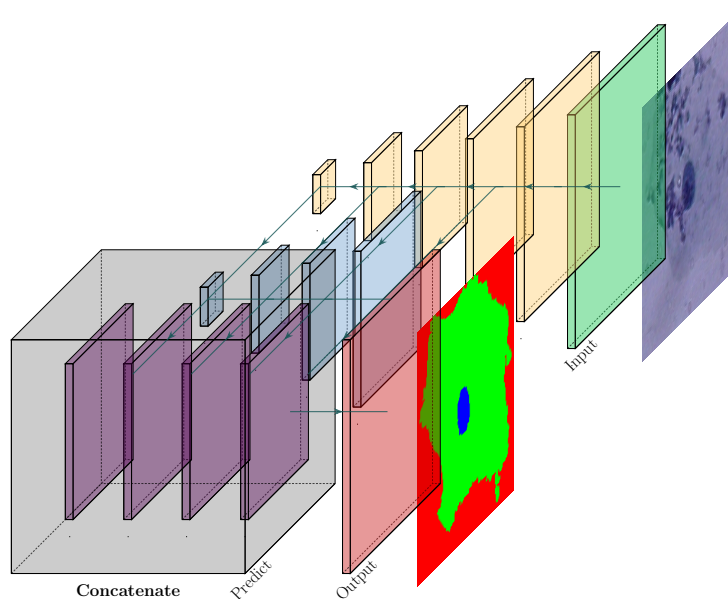


Figura 20: Feature Pyramid Network, diagrama conceptual

Las máscaras requieren un procesamiento adicional, ya que las imágenes no están adaptadas para el entrenamiento. El formato correcto es una matriz multidimensional de forma $H \times W \times N$, donde H significa el ancho de la imagen en píxeles, W significa la altura y N representa el número de clases si se trabaja con segmentación multiclase. Las regiones dentro de la máscara se extraen por colores en los que los tonos azules representan el núcleo, los verdes el citoplasma, el rojo es el fondo, el gris es una región desconocida y el negro es una región no etiquetada. La matriz final se construye con tres matrices, cada una de las cuales

consiste en una matriz unidimensional con 1 en los píxeles que pertenecen a la clase y 0 en el resto. Estas matrices se concatenan y se reconducen a una forma $128 \times 128 \times 3$. La primera (roja) corresponde al Fondo, la segunda (verde) al Citoplasma y la tercera (azul) al núcleo (Figura 21).

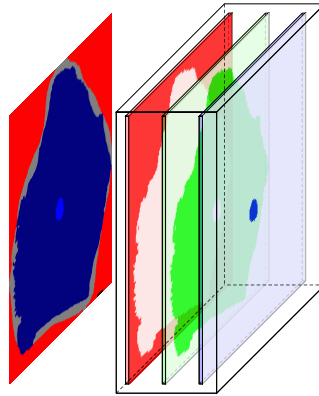


Figura 21: Creación de máscaras

El aumento de datos es capaz de mejorar tanto el rendimiento como el poder de generalización en los modelos de *Aprendizaje Profundo* (Taylor & Nitschke, 2017), también consigue reducir el sobreajuste. Se aplicaron técnicas tanto de transformaciones geométricas (rotación, escalado, traslación, reflexión y estiramiento) como de manipulación de la imagen (ruido y *CLAHE*). Cada tipo se seleccionó y aplicó aleatoriamente utilizando una distribución de probabilidad a cada imagen de entrenamiento. Este proceso se puede observar en la Figura 22.

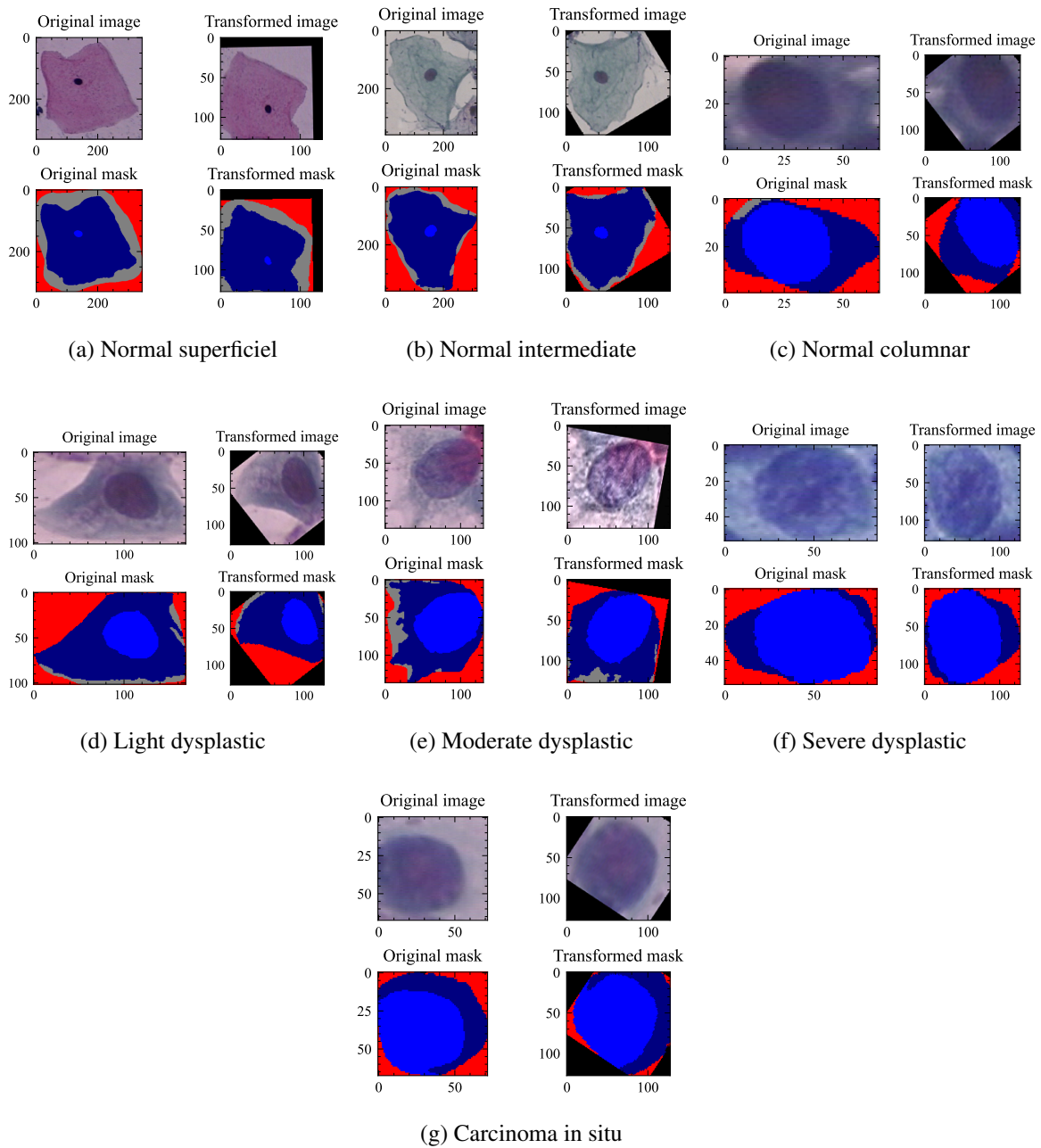


Figura 22: Imágenes aumentadas

El algoritmo se implementó en Python utilizando el módulo Tensorflow 2.0. Los experimentos se realizaron utilizando una GPU Nvidia Titan V con 12 GB de RAM en una estación de trabajo con 8 núcleos físicos y 8 virtuales con 64 GB de memoria del sistema. El modelo se entrenó siguiendo la técnica de validación cruzada estratificada (Figura 23).

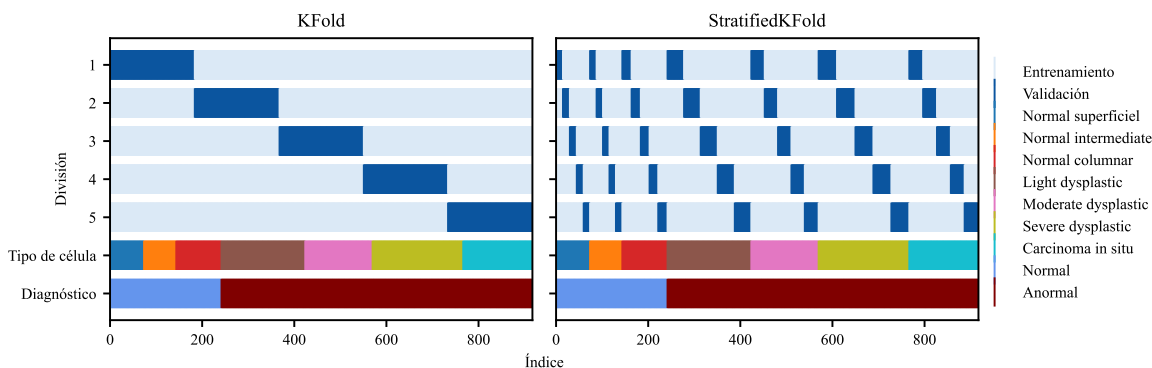


Figura 23: Comparación entre métodos de validación cruzada

Los hiperparámetros de entrenamiento se muestran en [Tabla 3](#). El tamaño del lote se mantiene bajo debido a las limitaciones de memoria. Se añade un término regularizador para reducir el sobreajuste en un conjunto de datos tan pequeño. El entrenamiento se lleva a cabo durante 150 épocas sin detenerse antes de tiempo. Se eligió la pérdida de datos por su rendimiento en un conjunto de datos desequilibrado (Ma, 2020), ya que el número de píxeles en las clases no es uniforme. El *Descenso de Gradiente Estocástico* es el optimizador estándar y con una alta tasa de aprendizaje e impulso evitamos caer en mínimos locales. Por último, los filtros se aumentan a 2048 para aumentar la capacidad del modelo.

Tabla 3: Hiperparámetros del experimento

Nombre	Valor
Batch size	8
Weight decay	0.001
Epochs	150
Loss	Dice
Learning Rate	0.01
Momentum	0.99
Optimizer	SGD
Pyramid Block Filters	2048

Se realizaron cinco experimentos para evaluar y comparar el rendimiento de nuestra propuesta frente a modelos con diferentes combinaciones de codificador/decodificador. La [Tabla 4](#) mues-

tra que la combinación de *FPN* con arquitecturas como VGG19 conduce a un rendimiento pobre; también, Resnet152 lucha por competir con EfficientNetB0, ambas redes fueron entrenadas usando los pesos de ImageNet para el aprendizaje de transferencia. Mientras tanto, U-Net y FPN logran un rendimiento sobresaliente en la tarea de segmentación cervical utilizando *EfficientNet* como codificador con pesos *Noisy Student* para el aprendizaje de transferencia. El aumento de la capacidad del modelo más allá de los valores seleccionados dio lugar a un sobreajuste, incluso con el decaimiento del aprendizaje y alta regularización.

La evaluación se realizó por subconjunto y luego por clase celular para cada métrica; luego, las métricas se agregaron promediando los resultados de las clases. El resultado final es la media de todas las métricas.

Tabla 4: Comparación de los experimentos

Experimento	Precision	Recall	F1-score	Specificity	IoU
VGG19 – FPN	0.764 ± 0.065	0.844 ± 0.053	0.79 ± 0.05	0.889 ± 0.027	0.672 ± 0.063
Restnet152 – U-Net	0.908 ± 0.024	0.907 ± 0.021	0.907 ± 0.014	0.952 ± 0.012	0.836 ± 0.023
EfficientNetB0 – FPN	0.907 ± 0.03	0.915 ± 0.023	0.91 ± 0.016	0.953 ± 0.009	0.841 ± 0.026
EfficientNetB7 – U-Net	0.911 ± 0.021	0.916 ± 0.023	0.913 ± 0.015	0.955 ± 0.009	0.846 ± 0.025
EfficientNetB7 – FPN*	0.913 ± 0.026	0.92 ± 0.017	0.916 ± 0.016	0.958 ± 0.008	0.851 ± 0.025

El rendimiento de las pérdidas por época se muestra en [Figura 24](#), el modelo está bien ajustado, ya que las pérdidas de entrenamiento y validación están cerca. El uso de otras pérdidas, como la de Jaccard, dio lugar a un rendimiento inferior.

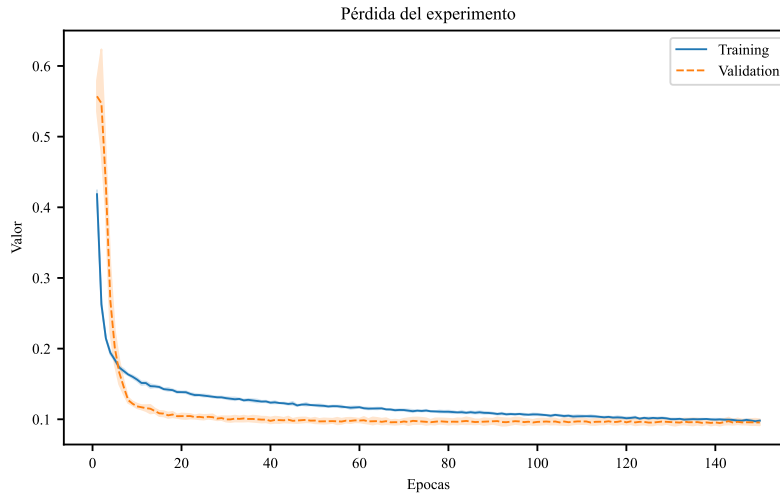


Figura 24: Pérdida en entrenamiento y validación

La **Figura 25** concentra las muestras tomadas para cada una de las clases. La primera columna muestra imágenes Herlev redimensionadas. La segunda columna muestra las máscaras, modificadas para representar el fondo como rojo o el primer canal de una imagen RGB estándar, el citoplasma como verde o el segundo canal, y los núcleos como azul (tercer canal). La predicción, en la tercera columna, es la salida del modelo remodelado a partir de un tensor. Finalmente, la última columna muestra los contornos de las máscaras superpuestas en la Imagen Herlev.

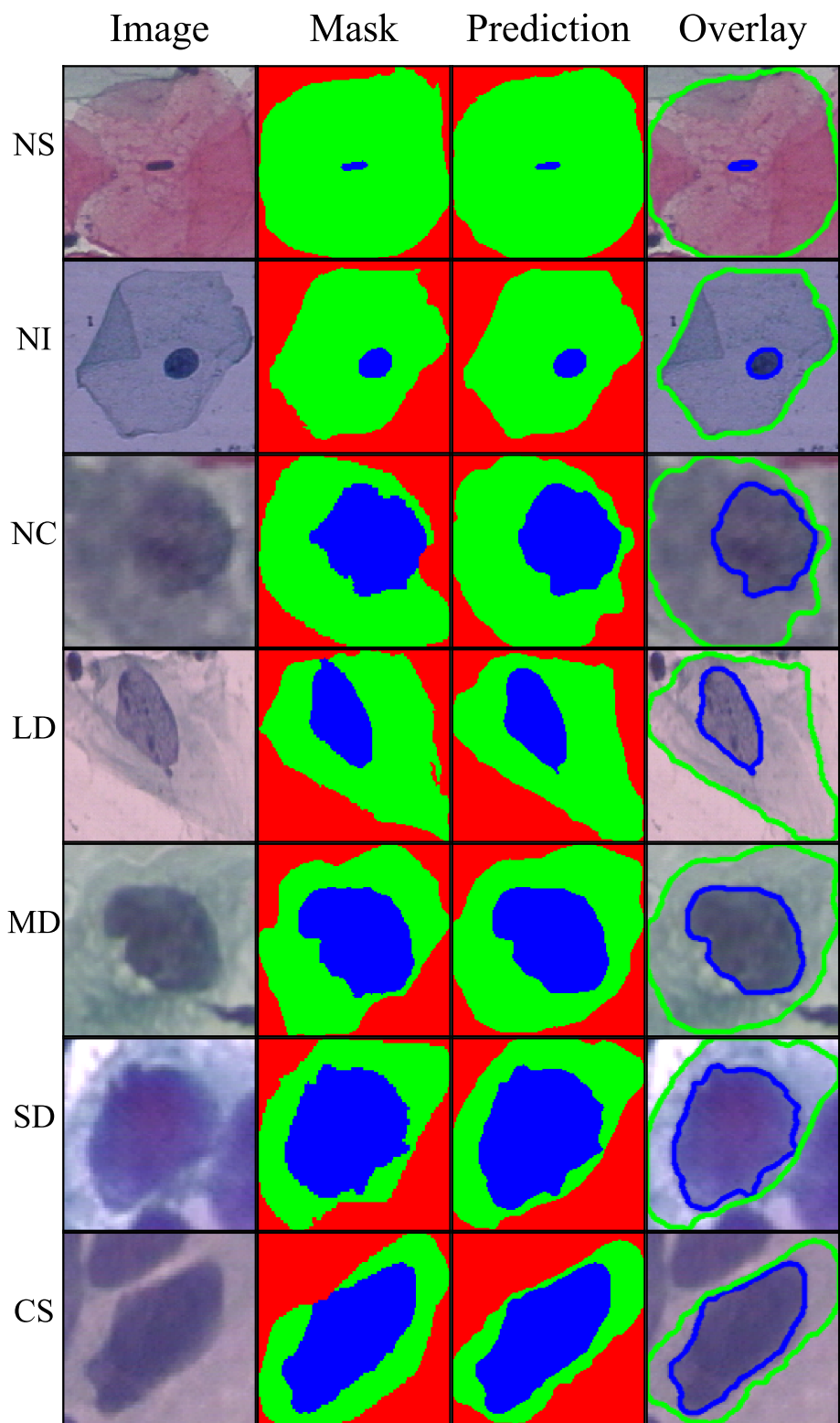


Figura 25: Muestreo de células segmentadas

Los resultados finales detallados se muestran en la **Tabla 5**.

Tabla 5: Resultados finales

Clase	Métrica	NS	NI	NC	LD	MD	SD	CS	Media
Citoplasma	F1-score	0.97 ± 0.01	0.98 ± 0.01	0.83 ± 0.02	0.9 ± 0.0	0.87 ± 0.02	0.83 ± 0.01	0.76 ± 0.03	0.88 ± 0.01
	IoU	0.94 ± 0.01	0.95 ± 0.01	0.71 ± 0.03	0.82 ± 0.01	0.77 ± 0.03	0.71 ± 0.01	0.62 ± 0.04	0.79 ± 0.02
	Precision	0.97 ± 0.01	0.98 ± 0.01	0.83 ± 0.04	0.9 ± 0.02	0.86 ± 0.03	0.82 ± 0.02	0.78 ± 0.04	0.88 ± 0.02
	Recall	0.97 ± 0.01	0.97 ± 0.01	0.84 ± 0.02	0.9 ± 0.02	0.88 ± 0.02	0.83 ± 0.02	0.75 ± 0.03	0.88 ± 0.02
	Specificity	0.94 ± 0.02	0.97 ± 0.01	0.91 ± 0.02	0.91 ± 0.01	0.91 ± 0.01	0.9 ± 0.01	0.92 ± 0.01	0.92 ± 0.01
Núcleo	F1-score	0.92 ± 0.05	0.95 ± 0.02	0.93 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.0	0.96 ± 0.01	0.95 ± 0.02
	IoU	0.86 ± 0.09	0.91 ± 0.04	0.86 ± 0.04	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.01	0.92 ± 0.02	0.9 ± 0.03
	Precision	0.91 ± 0.09	0.95 ± 0.05	0.92 ± 0.03	0.95 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.95 ± 0.02	0.94 ± 0.03
	Recall	0.94 ± 0.02	0.96 ± 0.03	0.94 ± 0.04	0.97 ± 0.01	0.96 ± 0.02	0.95 ± 0.01	0.96 ± 0.01	0.95 ± 0.02
	Specificity	1.0 ± 0.0	1.0 ± 0.0	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.0	0.98 ± 0.0	0.96 ± 0.01	0.99 ± 0.01
Media	F1-score	0.95 ± 0.03	0.96 ± 0.01	0.88 ± 0.02	0.93 ± 0.01	0.91 ± 0.01	0.89 ± 0.01	0.86 ± 0.02	0.91 ± 0.02
	IoU	0.9 ± 0.05	0.93 ± 0.03	0.79 ± 0.04	0.87 ± 0.01	0.85 ± 0.02	0.81 ± 0.01	0.77 ± 0.03	0.85 ± 0.03
	Precision	0.94 ± 0.05	0.97 ± 0.03	0.88 ± 0.04	0.93 ± 0.02	0.91 ± 0.02	0.89 ± 0.01	0.86 ± 0.03	0.91 ± 0.03
	Recall	0.96 ± 0.01	0.96 ± 0.02	0.89 ± 0.03	0.93 ± 0.01	0.92 ± 0.02	0.89 ± 0.01	0.86 ± 0.02	0.92 ± 0.02
	Specificity	0.97 ± 0.01	0.99 ± 0.0	0.94 ± 0.01	0.95 ± 0.01	0.95 ± 0.01	0.94 ± 0.01	0.94 ± 0.01	0.96 ± 0.01

Centramos nuestro análisis y comparación en trabajos que utilizan el conjunto de datos Herlev para la segmentación semántica cervical. Solo tres de los trabajos seleccionados para la comparación trataban de la segmentación multiclase (Wasswa et al., 2019)(Bandyopadhyay & Nasipuri, 2020)(L. Zhao et al., 2016), los demás trataban de la segmentación de núcleos de una sola clase y solo uno para la segmentación de células enteras Kurnianingsih et al., 2019. En **Tabla 6**, mostramos una comparación de una variedad de métodos aplicados a la *segmentación semántica* de células cervicales.

Tabla 6: Comparación de rendimiento con otros trabajos

Autor	Método	Núcleo					Citoplasma				
		Precision	Recall	F-1 score	Specificity	IoU	Precision	Recall	F-1 score	Specificity	IoU
Gautam et. al.(Gautam et al., 2018)	ConvNet	0.89	0.91	0.9	–	–	–	–	–	–	–
Zhang et. al.(Zhang et al., 2020)	DACN	0.95	0.95	0.95	–	–	–	–	–	–	–
Zhao et. al.(J. Zhao et al., 2019)	ConvNet ensemble	0.95	0.98	0.93	–	–	–	–	–	–	–
Zhao et. al.(J. Zhao et al., 2020)	PGU-net	0.9	0.96	0.92	–	–	–	–	–	–	–
Zhang et. al.(Zhang,2017a)	FCN	–	–	0.92	–	–	–	–	–	–	–
Sharma et. al.(Sharma & Mangat, 2016)	FCM	0.86	0.9	0.85	–	–	–	–	–	–	–
Chankong et. al.(Chankong et al., 2014)	FCM	0.85	0.83	0.8	–	–	–	–	–	–	–
Genctav et. al.(Genctav,2012)	Watershed	0.88	0.93	0.89	–	–	–	–	–	–	–
Bandyopadhyay et. al.(Bandyopadhyay & Nasipuri, 2020)	K-Means	–	–	–	–	0.72	–	–	–	–	–
Zhao et. al.(L. Zhao et al., 2016)	MRF	–	–	0.93	–	–	–	–	0.82	–	–
Wasswa et. al.(Wasswa et al., 2019)	PLC	1	0.98	0.99	0.98	–	0.98	0.99	0.99	–	–
Propuesta*	EfficientNet - FPN	0.94	0.95	0.95	0.99	0.9	0.88	0.88	0.88	0.92	0.79

De la **Tabla 5** podemos hacer varias observaciones: la desviación estándar es muy baja, la alta especificidad en todas las clases y regiones significa una detección negativa excepcional, *IoU* es la métrica más baja y tenemos la hipótesis de que es porque una región está dentro de otra, por lo que conduce a una superposición entre las máscaras, las características morfológicas de las clases de células explican la variación en el rendimiento de todas las métricas, el citoplasma es más difícil de segmentar y creemos que se debe a que está incrustado entre los núcleos y el fondo, las células normales son más accesibles de segmentar y el Carcinoma In Situ es la clase más difícil de segmentar porque los cambios morfológicos de la célula mutada alteran tanto los núcleos como el citoplasma, los núcleos múltiples y las regiones superpuestas no son infrecuentes.

Al revisar **Figura 25** queda claro que el modelo es robusto frente a los desechos y otras imperfecciones de la imagen, como la superposición; además, las características espaciales bien aprendidas permiten al modelo discriminar correctamente los núcleos y el citoplasma adicionales dentro de la muestra.

Después de la segmentación de células superpuestas, la segmentación multiclase es la tarea más ardua dentro de la segmentación de células cervicales. Se puede lograr una segmentación de clase precisa aumentando la capacidad del modelo y operando cuidadosamente el diseño de la arquitectura y el proceso de entrenamiento; esto es crucial para la detección del cáncer cervical y puede mejorar las tasas de detección. Las ventajas de nuestra propuesta *EfficientNetB7 – FPN* son las siguientes

1. Fase simple de entrenamiento.
2. Complejidad del modelo acotada.
3. Implementación en código abierto.
4. No se usó preprocesamiento complejo.
5. Gran poder de generalización.
6. Tiempo de entrenamiento moderado.
7. Proceso de segmentación totalmente automatizado.

8. Rendimiento por arriba del umbral.

Los resultados muestran que la eficacia de nuestra propuesta es sólida, y que puede proporcionar una *Segmentación Semántica* precisa de las células cervicales.

Se demostró el rendimiento de EfficientNetB7 y FPN investigando su rendimiento para la segmentación precisa de núcleos y citoplasma de células cervicales. Utilizando el conjunto de datos Herlev, un punto de referencia bien conocido y probado, como datos de entrenamiento, aprovechamos el aprendizaje de transferencia *Noisy student* y *Compound Model Scaling* para codificar con éxito los datos espaciales para alimentar las potentes Pyramid Features del decodificador. Nuestra metodología de evaluación conjuga métricas relevantes para evaluar correctamente el rendimiento del modelo. *EfficientNetB7 - FPN* tuvo un rendimiento medio, superando a modelos más complejos que requieren menos recursos computacionales y tiempo de entrenamiento.

En resumen, 0.91 ± 0.02 F1-score, 0.85 ± 0.03 IoU, 0.91 ± 0.03 Precisión, 0.92 ± 0.02 Recall, 0.96 ± 0.01 Especificidad con una desviación estándar muy baja. Cada métrica para cada región y clase de célula está por encima del umbral de eficacia de Zijdenbos sin necesidad de pasos complejos, tanto para el preprocesamiento como para la creación del modelo. Las limitaciones son la necesidad de un hardware especializado, como una GPU dedicada con gran cantidad de RAM. El trabajo futuro se centrará en mejorar la segmentación de la región del citoplasma.

4.3.3 Discusión

El análisis de este caso de estudio no solo derivó en un modelo con mejor rendimiento que otros trabajos especializados en la misma área. La comparación se realizó bajo los mismos términos para asegurar una evaluación libre de sesgos. También es el único trabajo encontrado que trata con la segmentación multiclase en el mismo conjunto de datos y la combinación de arquitecturas jamás se había usado para procesamiento de imágenes médicas.

Dentro del estado del arte, Lin et. al. (H. Lin et al., 2019a) se apoyan de Redes Neuronales Convolucionales para el proceso de diagnóstico de cataratas mediante imágenes, sin embargo,

limitan su modelo a la clasificación de una imagen y no extraen más información aparte del diagnóstico. Zeina Rayan et. al. (Rayan et al., 2019), demuestran que el uso de este tipo de Redes Neuronales son de gran utilidad para apoyar la toma de decisiones en el diagnóstico médico. Una mejora con respecto a estas aplicaciones, es que el análisis de este caso de estudio trata con un problema más difícil dentro de la *Visión por Computadora*, la *Segmentación Semántica*, así mismo, estas propuestas se encuentran desacopladas a cualquier y existen solamente como entidades aisladas monopropósito. Esta propuesta ha sido diseñada para que su operación sea dentro del Sistema. Trabajos previos con el conjunto de datos Herlev, demostraron la capacidad de desplegar modelos de apoyo al diagnóstico médico en dispositivos, derivando en la oportunidad de convertir este desarrollo en una *Smart tool*.

En concreto, las aportaciones de este caso de estudio son:

- Aplicación de algoritmos de visión por computadora al diagnóstico médico.
- Implementación y entrenamiento de modelos de Deep Learning.
- Generar pericia en entrenamiento de modelos grandes para reducir el sesgo en el rendimiento.
- Familiarización con métricas de evaluación para clasificación.

4.4 Caso de estudio: Fiebre del Dengue

Anteriormente, se trabajó en conjunto con el IMSS la problemática de la *Fiebre del Dengue*, su vector y su relación con la mortalidad de mujeres embarazadas. Para este caso en particular, se generalizó esta premisa para abarcar el problema de la localización de brotes de dengue de manera oportuna y generar conocimiento a partir de datos epidemiológicos para saber cómo se comporta la enfermedad y encontrar variables que incidan positiva o negativamente en su incidencia.

Al principio se tuvieron una cantidad limitada de datos epidemiológicos y no se tuvo la oportunidad de conseguir más, esto detonó un proceso de búsqueda exhaustiva para encontrar soluciones a estos contratiempos. En un trabajo previo de dengue en embarazadas ya se tenía

noción de que el análisis geoespacial era crucial para generalizar el problema a todos los sectores de la población; esto, sumado a la exploración en la literatura se pudo construir un puente para solventar los contratiempos.

La elección de este problema como principal caso de estudio fue sumamente acertada. Ya que es una enfermedad cuyo comportamiento se puede modelar utilizando distintas variables como la demográfica y la climatológica. Esto llevó a una incorporación temprana de fuentes de datos externas, lo importante es que se tiene entendimiento de estos datos, qué variables los componen, como se puede acceder a ellos y cuál es su estructura interna. Conocimiento que permite no solamente su uso para análisis estadístico, matemático y algorítmico, sino también es la piedra angular para construir una aplicación tecnológica como lo es un *Sistema de Información en Salud*.

No solo se solventaron los contratiempos, sino que se construyó un entendimiento superior y concreto para el desarrollo de la propuesta doctoral. Se probaron conceptos de muy alto nivel como *AutoML* y los *Sistemas Híbridos Inteligentes* y se demostró que, se puede mejorar una base de datos epidemiológica utilizando datos públicos disponibles y no solamente con datos generados por empresas de tecnología, casi siempre inaccesibles para el público en general.

El caso de estudio de *Fiebre del Dengue* se convirtió en el principal por varias razones:

- Trabajos anteriores con el IMSS.
- Complejidad de la cadena logística epidemiológica.
- Cantidad de datos que se pueden extraer de ella.
- Cantidad de datos disponibles.
- Cantidad de problemas que se pueden resolver con los datos disponibles.
- Facilidad de integración con datos climatológicos, geográficos, demográficos, etc.
- Impacto social de la enfermedad.
- Relación con factores externos como el cambio climático.

4.4.1 Descripción del problema

La *Fiebre del Dengue* es una Enfermedad Tropical Desatendida que puede infectar a casi 400 millones de personas al año. Los casos ocurren principalmente en zonas vulnerables como Latinoamérica, África, Asia y el Pacífico, concentrándose en áreas urbanas o semi urbanizadas (Figura 26). La enfermedad es causada por el virus del Dengue, que es capaz de infectar a una persona 4 veces, una por cada serotipo existente. Puede causar complicaciones serias en los infectados y es especialmente cruel con poblaciones vulnerables como las embarazadas (Control, European Centre for Disease Prevention and, s.f.).

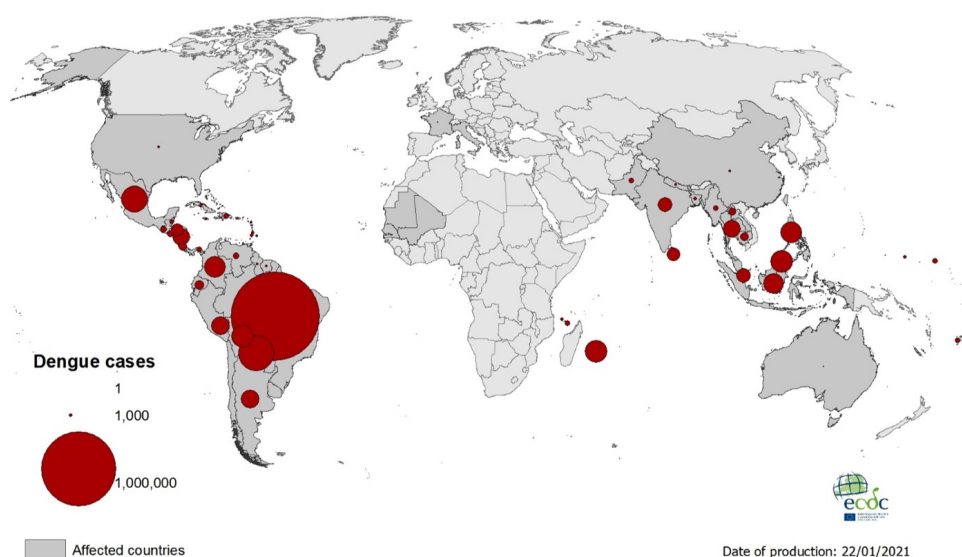


Figura 26: Casos de dengue en el mundo

4.4.1.1 Epidemiología

Es una enfermedad transmitida por vector, es decir, el virus reside dentro de un huésped que es el que propaga la enfermedad entre las personas. El vector de esta enfermedad es el mosquito *Aedes Aegypti*. Este mosquito también es capaz de transmitir otras enfermedades como el Zika y la Chikungunya. Al picar, el virus que reside dentro de las glándulas salivales, entra en el cuerpo humano iniciando la infección. Dentro del mosquito, el virus se replica en el estómago, alcanzando un alto grado de infección en entre 8 y 12 días, dependiendo de ciertas variables climáticas y genéticas. El virus no llega al mosquito por sí solo, sino que lo adquiere al picar

personas infectadas, creando así un círculo infeccioso entre las dos especies. En la actualidad, las hembras son capaces de transmitir el virus a sus crías (Figura 27). (CDC, 2020)

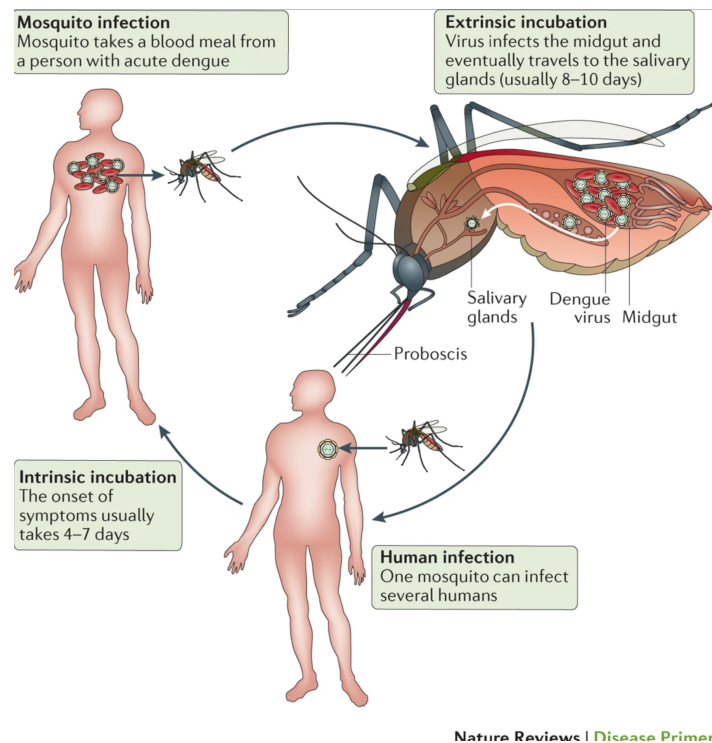


Figura 27: Transmisión de la Fiebre del Dengue

El mosquito tiene una relación **eusinantrópica**⁴ con el ser humano, absolutamente todo su ciclo de vida se desarrolla en agua estancada en contenedores. Anteriormente, era una especie selvática, pero en la actualidad su ciclo de vida es inseparable del humano. Con respecto al virus, estos se encuentran totalmente adaptados al humano. El ciclo de vida de la hembra transcurre en cuatro fases, como se muestra en la Figura 28 (Salas & Leal, 1995).

1. **Huevos:** Los huevos son depositados por una hembra fertilizada en una superficie de agua estancada. Dura 2 días.
2. **Larvas:** Las larvas viven en el agua durante 2 días.
3. **Pupas:** En siete días, las larvas se convierten en pupas.

⁴Se refiere a especies extremadamente adaptadas a convivir con el ser humano.

4. **Mosquito:** Cinco días después, surge un mosquito adulto que se alimentará de sangre durante 3 días, en un radio no mayor a 200 metros del nido, al segundo día comenzará a depositar huevos.

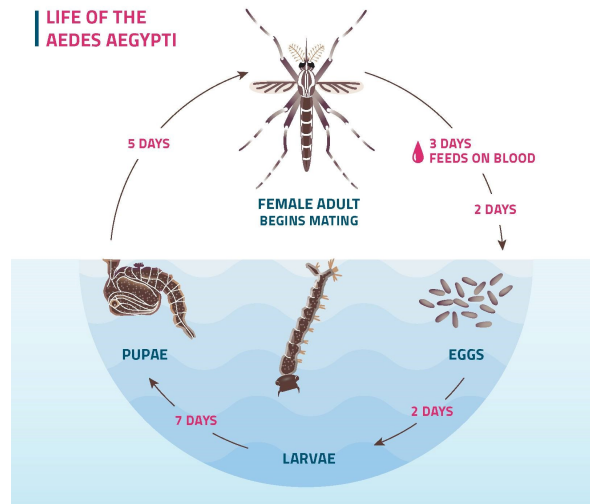


Figura 28: Ciclo de vida del *Aedes Aegypti*

Existe un marcado incremento en la cantidad a nivel mundial en todas las regiones (Figura 29), con incrementos significativos a partir de 1990. Esto es gracias al crecimiento económico que propició el aumento de la urbanización, la globalización y, lo más importante, el *cambio climático* (Gubler, 2011).

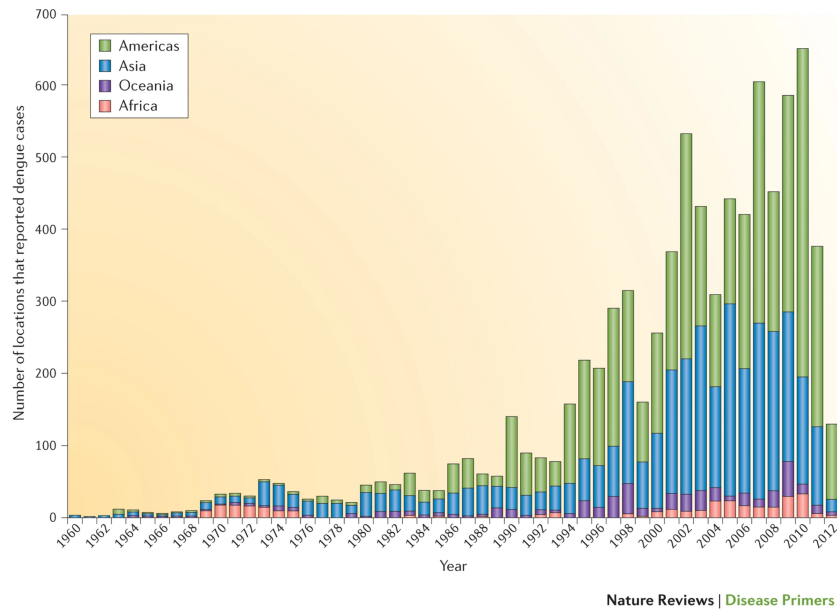


Figura 29: Casos de Fiebre del Dengue a nivel mundial

4.4.1.2 Síntomas

Por los síntomas, existen tres tipos de *Fiebre del Dengue*:

1. **Dengue normal:** Fiebre y alguno de los siguientes, náusea, vómito, sarpullido, calambres y dolores.
2. **Dengue con signos de alarma:** Dolor abdominal, vómito persistente, acumulación de fluidos, sangrado de mucosas, estado letárgico, hepatitis, plaquetas bajas.
3. **Dengue grave:** Extravasación capilar plasmática, arresto respiratorio, sangrado severo, daño en los órganos.

Es una enfermedad potencialmente letal, en la cual infecciones con otros serotipos de virus incrementa las posibilidades de desarrollar una forma grave de la enfermedad. No existe un tratamiento específico, sino que se trata con reposo y cuidados paliativos. Complicaciones especiales pueden surgir en pacientes vulnerables como las mujeres embarazadas, donde se puede incurrir en un mal diagnóstico con otras enfermedades del embarazo (Bhatt et al., 2013).

El primer síntoma siempre es una fiebre de intensidad variable. Generalmente, la enfermedad es corta, de menos de dos semanas de duración, en la mayoría de los casos aunque su dinamismo le

permite empeorar súbitamente en días posteriores. La muerte ocurre por choque hipovolémico y durante su convalecencia puede tener complicaciones serias en el cerebro o el corazón. El diagnóstico se hace no por los síntomas, sino por métodos virológicos o serológicos (Guzman et al., 2016).

4.4.1.3 Prevención

La forma más eficaz de combatir la Fiebre del Dengue, es atacando el mosquito que la transmite mediante la remoción de basura y cachivaches en los patios y azoteas, cubriendo los depósitos de agua limpia, mantenimiento correcto de construcciones urbanas y, sobre todo, campañas de fumigación especialmente diseñadas para atacar focos de infección.

4.4.1.4 Flujo de datos

Para este caso de estudio, el flujo de datos se modela tomando los casos individuales, capturados por la Secretaría de Salud y agregados al SINAVE(Salud, s.f.). Al agrupar estos datos por división geográfica o características demográficas del caso individual, se extrae información sobre el comportamiento de la enfermedad. Al analizar esta información a través del tiempo, tomando en cuenta su relación geoespacial y contrastándola con otros datos como el clima, podemos encontrar tendencias y patrones emergentes que sirven para apoyar a la toma de decisiones dentro de la cadena logística de salud pública. Por último, creando modelos predictivos, convertimos el conocimiento en sabiduría, permitiendo al sistema sugerir soluciones a los tomadores de decisiones (Tabla 7).

Tabla 7: Flujo de datos para el caso de estudio

Flujo	Representación
Datos	Casos individuales
Información	Agrupación geográfica y demográfica
Conocimiento	Comportamiento espaciotemporal y relaciones climatológicas
Sabiduría	Predicciones

Este análisis se pretende hacer a distintas resoluciones. Se nombra resolución a los niveles

espacio-temporales que representan los horizontes predictivos definidos para el análisis del flujo de datos. Las resoluciones propuestas son las siguientes, cabe mencionar que se pretende utilizar estos horizontes en conjunto, dando cabida a los niveles espacio-temporales antes mencionados, por ejemplo, predicciones por entidad por año o mes; y así consecutivamente (Tabla 8). Para esta etapa, los expertos en salud tomaron una pequeña muestra de datos de 52471 filas y 490 columnas.

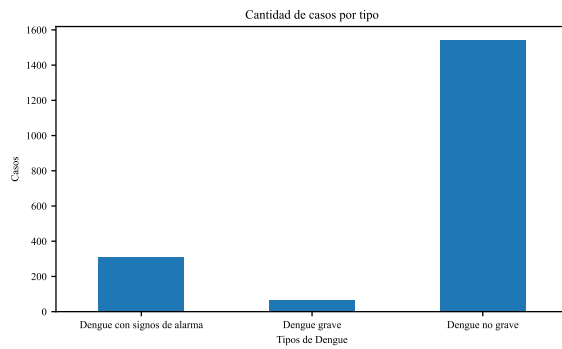
Tabla 8: Horizontes y resolución

Horizontes espaciales	Horizontes temporales
Entidad	Anual
Municipio	Mensual
AGEB	Diario
Manzana	Por hora

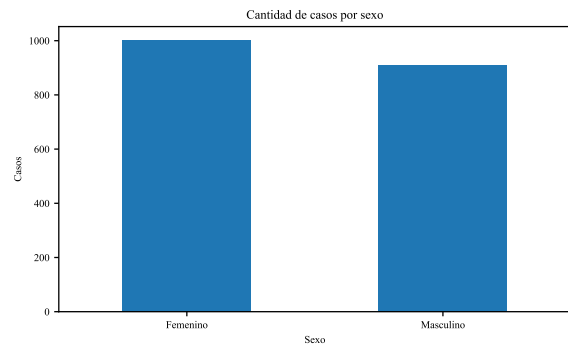
4.4.2 Conceptualización

4.4.2.1 Análisis demográfico

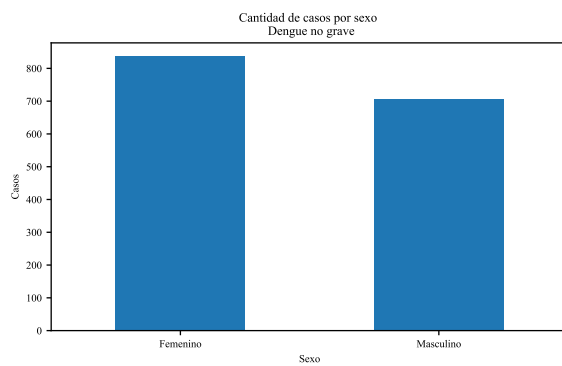
Al contar los casos de dengue por tipo de diagnóstico, observamos que la mayoría se concentra en casos no graves, mientras que la cantidad de casos graves es relativamente pequeña (Figura 30a). La cantidad de casos totales de dengue, agrupados por sexo, encontramos que la cantidad no es tan dispar entre ambos, con más casos de sexo femenino. Se procedió a analizar en función del sexo y del tipo de fiebre del dengue que aflige al caso (Figura 72c). En el caso de dengue no grave, la forma más común de la fiebre del dengue, encontramos que esta se concentra más en el sexo femenino. Es decir, de todos los casos, si el caso es de sexo femenino, es más probable que desarrolle un caso no grave que uno de cuidado (Figura 30c). En el caso de diagnóstico de dengue con signos de alarma. Observamos mayores casos en sexo masculino que en femenino (Figura 30d). Con respecto al dengue grave, podemos observar que este se presenta más en individuos del sexo masculino. Esto abre la oportunidad de buscar las causas de este fenómeno, para encontrar la relación entre el sexo masculino y la probabilidad de desarrollar una forma grave de la fiebre del dengue (Figura 30e).



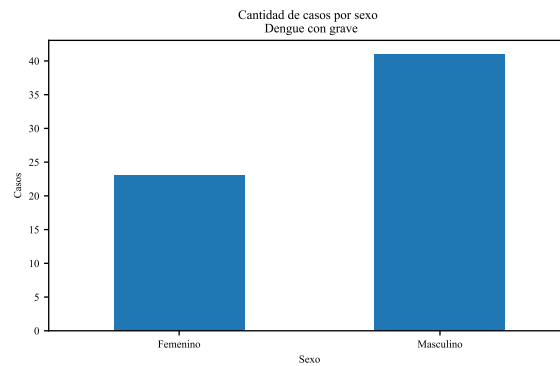
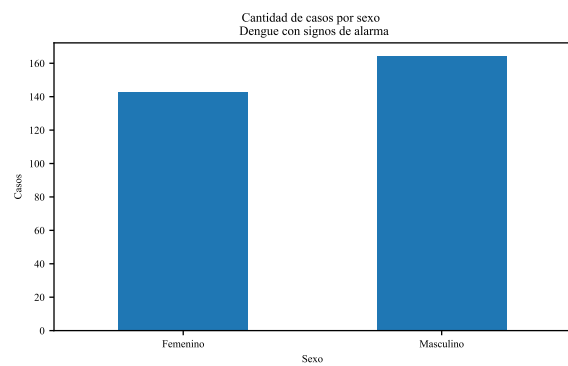
(a) Casos por diagnóstico



(b) Casos por sexo



(c) Casos por sexo y diagnóstico para dengue no grave (d) Casos por sexo y diagnóstico para dengue con signos de alarma



(e) Casos por sexo y diagnóstico para dengue grave

Figura 30: Análisis demográfico por diagnóstico y sexo

Se realizó un análisis χ^2 para determinar la correlación entre dos variables categóricas, en este caso, los diagnósticos finales y el sexo de los casos. Las hipótesis son las siguientes:

- **H0:** No existe relación entre el sexo y tener un dengue.
- **H1:** Existe una relación entre el sexo y tener dengue.

La prueba determinó, con valor de $p = 0.001$ a un nivel de significancia de 0.010, que se rechaza la hipótesis nula y existe una relación entre ambas variables.

Con respecto al análisis de casos contados por edad, en la siguiente gráfica (Figura 31) podemos observar que la mayoría de los casos se concentran en las edades de 6 a 19 años. Esto inspiró a realizar un análisis por profesión, ya que esta es la edad de la educación primaria y secundaria. El número de casos presenta un declive marcado conforme se incrementa la edad del caso.

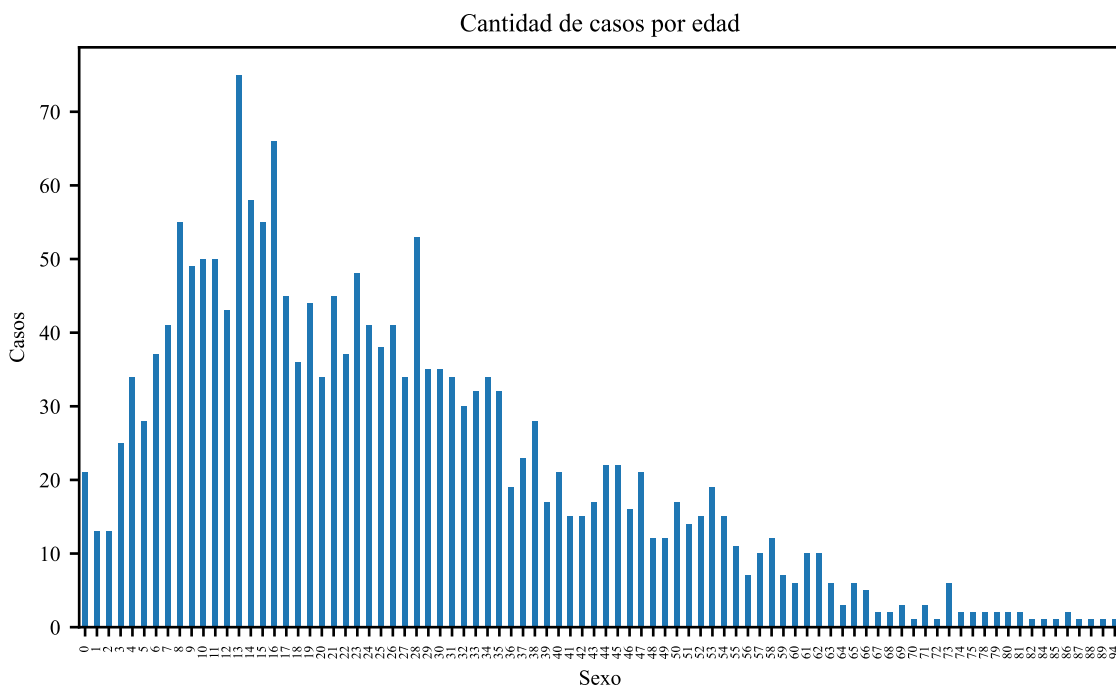


Figura 31: Casos por grupo de edad

Esto lo podemos corroborar agrupando los casos por profesión (Figura 32). Encontramos claramente que la mayoría de los casos son de estudiantes. Por lo cual se puede generar la hipótesis de que se contagiaron en su centro educativo. Las otras dos ocupaciones con alta incidencia son los empleados y trabajo en el hogar. Valdría la pena realizar un análisis geoespacial de la localización de las escuelas donde asisten los casos y cruzarla con el marco

geoestadístico nacional.

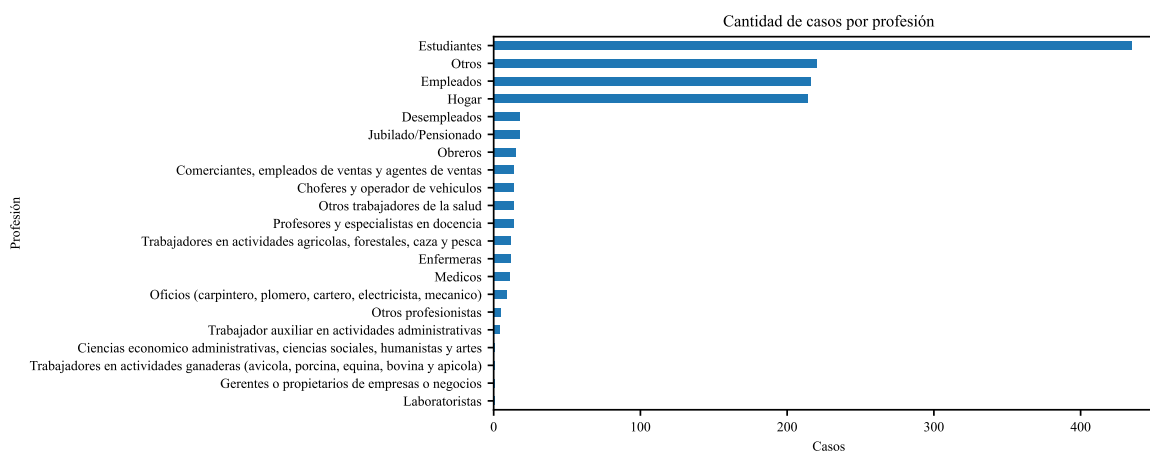


Figura 32: Casos por grupo profesión

Se encontró una relación entre las variables categóricas sexo y diagnóstico final mediante la prueba χ^2 cuadrada. Se deben hacer más pruebas estadísticas con mayor cantidad de datos para determinar la validez de esta relación. También, cruzar estos datos con datos demográficos de INEGI.

Con respecto a la edad, observamos que los datos se concentran en edades escolares, por lo cual es importante determinar con datos geoespaciales si la escuela a la que asisten es el foco epidemiológico. Este hallazgo es significativo, nos permite determinar posiciones puntuales de vigilancia epidemiológica dentro de las ciudades.

4.4.2.2 Análisis demográfico y geoespacial

Se realizó una exploración inicial de datos, primero, se cuenta por entidad federativa cuantos casos existen. Observamos que la mayoría de los casos se concentran en el estado de Veracruz, por lo que se trabajará solamente con esta entidad federativa.

Después del filtrado, quedan 3321 elementos, los cuales se agrupan por municipio de la siguiente manera, tomando los 10 municipios con mayor cantidad de casos. Esto lo podemos ver en la [Tabla 9](#).

Tabla 9: Cantidad de casos por municipios veracruzanos

Municipio	Cantidad
Veracruz	372
Papantla	204
Coatzacoalcos	153
Córdoba	141
Poza Rica de Hidalgo	138
Tuxpan	129
Orizaba	104

Para poder posicionar los datos en un mapa y realizar un análisis geoespacial, es necesario obtener las coordenadas de cada caso. Para ello, se unificaron las columnas dirección de los datos en una sola columna y luego, se procesó esta columna para convertir direcciones en coordenadas geográficas, utilizando un servicio externo de decodificación. *Geocodificar* es la acción de convertir una dirección de calle en un par de coordenadas geográficas, existen cantidad de servicios capaces de realizar esta tarea como por ejemplo *Google Geocode* o *Bing Maps*, lamentablemente estos servicios son de paga, por lo que buscó uno gratuito, afortunadamente se pudo encontrar el servicio de *HERE* para tal tarea. Se procedió a limpiar los datos eliminando aquellos que no cuenten con una dirección, obteniendo un total de 2715.

La respuesta del servicio de geocodificación es una estructura que contiene una o varias direcciones posibles, ordenadas por precisión de manera descendente. De esta estructura, se extraen las coordenadas del primer elemento, es decir, aquel con mejor precisión. Se crearon tres columnas, una que contiene ambas coordenadas y otras dos con cada coordenada individual. Observar que las coordenadas se tienen que invertir para ser mapeadas correctamente por la herramienta geoespacial, así mismo, se eliminaron los resultados con menos de 75 % de precisión.

La **Figura 33** muestra los polígonos de todos los municipios del estado de Veracruz, en una primera capa, mientras que en la segunda se colocan los puntos que representan cada caso de Dengue. Se nota a simple vista que existen agrupaciones importantes centradas en municipios

específicos.

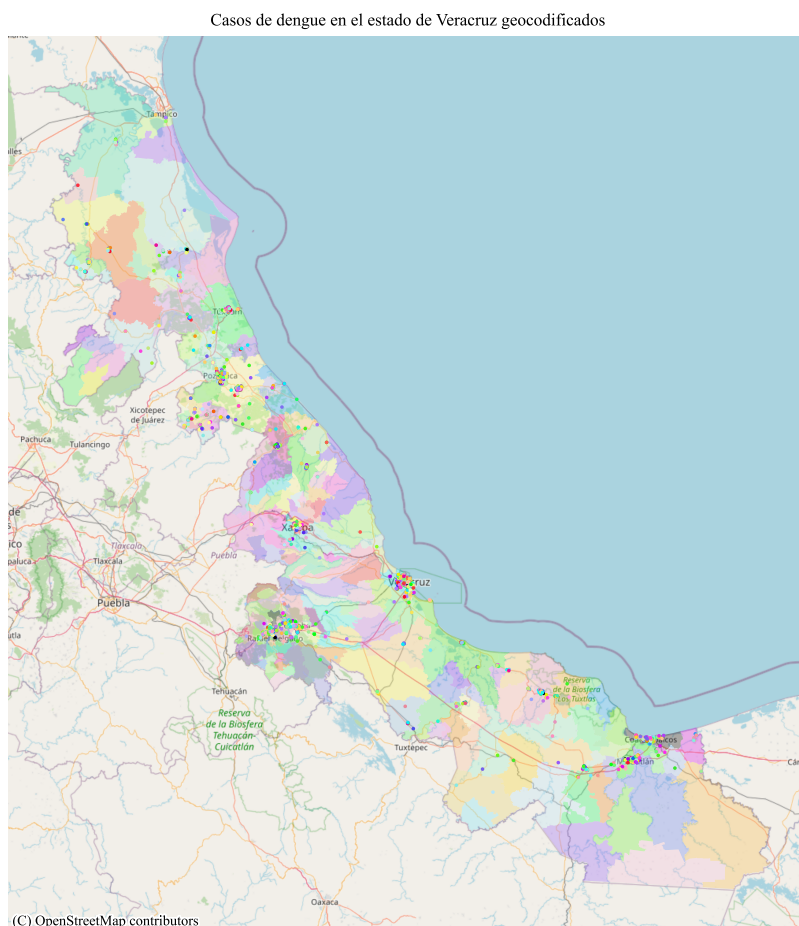


Figura 33: Casos en municipios

El censo contiene información a resolución por manzana, pero debido a que la geocodificación tiene cierto margen de error, es probable que algunos puntos no estén totalmente contenidos por los polígonos de manzana, es por ello que el análisis se realizó a una resolución de municipio. Luego, se calculó la densidad de población dividiendo la población total de cada municipio por su área.

Para poder visualizar, se creó un mapa de calor con los valores de densidad por municipio. Se puede observar claramente que la mayoría de la superficie del estado no fue censada en 2020. Esto deja de lado todas las poblaciones rurales (Figura 34). Sin embargo, esto no presenta ningún problema, ya que el dengue es una enfermedad urbana.

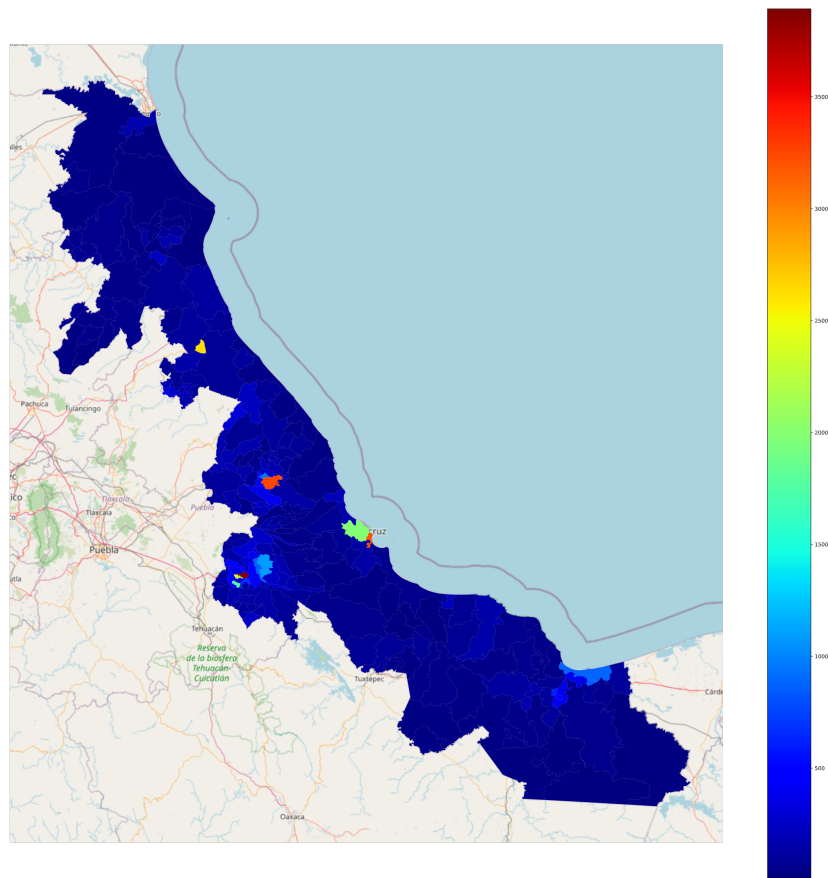


Figura 34: Densidad de población por municipio

Utilizando módulos de análisis geoespacial, se contaron los puntos por polígonos y se ordenan de mayor cantidad a menor cantidad. Podemos observar en la [Tabla 10](#), los datos de población total, área y densidad así como el número de casos por polígono, para los cinco municipios con mayor cantidad de casos.

Tabla 10: Primeros diez elementos de conteo de puntos por polígonos

Clave	Clave Municipio	Entidad	Municipio	Población	Área	Densidad	Puntos
30193	193	Veracruz	Veracruz	552156	278.5594	1982.184	324
30039	39	Veracruz	Coatzacoalcos	305260	347.262	879.0482	143
30131	131	Veracruz	Poza Rica de Hidalgo	193311	73.47472	2630.987	113
30028	28	Veracruz	Boca del Río	138058	42.87901	3219.711	104
30124	124	Veracruz	Papantla	158599	1666.596	95.16345	98

En la [Tabla 11](#) se puede observar una muestra de los datos ordenados de manera ascendente en referencia al número de puntos. A simple vista no se puede notar una relación evidente entre la densidad y el número de casos por municipio.

Tabla 11: Densidad y número de casos

Población	Área	Densidad	Puntos
552156	278.5594	1982.184124	324
305260	347.262	879.048173	143
193311	73.47472	2630.986683	113
138058	42.87901	3219.710868	104
158599	1666.596	95.163453	98

Al hacer un análisis de correlación ingenuo, es decir, sin comprobar supuestos de linealidad, homoscedasticidad, etc., encontramos que ambas columnas están ligeramente relacionadas, aunque esta relación puede ser netamente espuria ([Tabla 12](#)).

Tabla 12: Correlación entre densidad y número de casos

	Densidad	Puntos
Densidad	1	0.375524
Puntos	0.375524	1

Por completud, se hizo un análisis de *Regresión Lineal*, encontramos que si existe una ligera pendiente positiva en los datos, sin embargo, hay que tomar estos resultados como una exploración subjetiva. Podemos observar en la [Figura 35](#), el resultado de este análisis de regresión, dando como $R^2 = 0.286$.

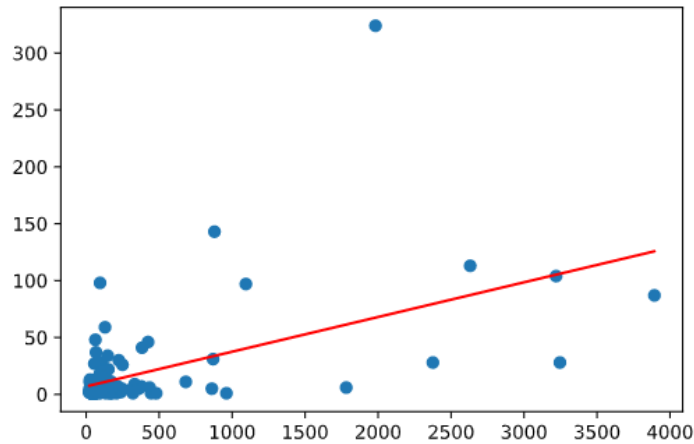


Figura 35: Regresión lineal

4.4.2.3 *Clustering*

El análisis de los datos se hizo bajo el contexto de municipio, en este caso, analizaremos uno de los municipios con mayor cantidad de casos: Coatzacoalcos. El análisis geoespacial dotó de sentido a los puntos geográficos transformando datos en información en función a su relación de distancia, debido a que la fiebre del Dengue es transmitida por un vector. La hembra de este mosquito es la única capaz de transmitir el virus y siempre lo hace en la vecindad de su nido. Expertos estiman el radio de alimentación en 300 metros.

El algoritmo escogido para el análisis es *DBSCAN*⁵, un algoritmo no supervisado de agrupamiento que crea grupos basados en densidad, es decir, la cantidad de puntos en un área determinada. Este algoritmo utiliza una métrica para relacionar los puntos, en nuestro caso, utilizaremos la fórmula del semiverseno para determinar la distancia entre los puntos debido a que la superficie terrestre es una esfera y no un plano (Figura 36).

⁵Density-based spatial clustering of applications with noise

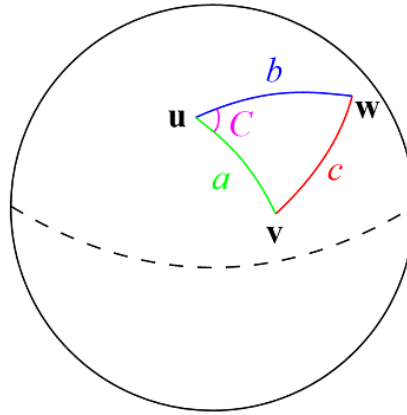


Figura 36: Representación de la distancia esférica

La implementación de *DBSCAN* utilizada (se probaron varias como *OPTICS*), recibe como hiperparámetros la distancia de relación (*eps*), el número mínimo de muestras relacionadas (*min_samples*), el algoritmo de procesamiento (*algorithm*) y la distancia a computar (*metrics*). El hiperparámetro *eps* se refiere a la distancia en la cual dos puntos se encuentran relacionados, en nuestro caso, al utilizar la fórmula de verseno para hacer los cálculos, los datos los debemos representar en radianes, es por ello que se procesa la distancia en kilómetros y las coordenadas. El número mínimo de muestras para considerarse un grupo, lo fijamos en 2. El algoritmo agrupa los puntos cumpliendo la relación que describen los hiperparámetros, los puntos que no entran en ningún grupo los agrega a un grupo especial que está denominado por el valor -1, es decir, aquellos puntos que no pertenecen a ningún grupo o clúster. (Ester et al., 1996)

Una vez entrenado el algoritmo, extraemos la cantidad de grupos o clústeres generados, en este caso 6. En la **Tabla 13** podemos observar los clústeres encontrados así como la cantidad de elementos, 31 de los cuales no fueron agrupados.

Tabla 13: Cantidad de casos por cluster

Cluster	Cantidad
-1	31
0	2
1	2
2	2
3	2
4	3
5	2

Se analizó la fecha de los casos así como la dirección donde ocurrieron, en función al grupo al que pertenecen. Como podemos observar, los casos del grupo 0 son casos reportados en un mismo domicilio, con un periodo de tiempo que va de un día a 8. Esto nos da un indicio de que cerca de esa posición geográfica, existe un criadero de mosquitos con capacidad infecciosa, un foco rojo epidemiológico (Tabla 14).

Tabla 14: Casos en el mismo domicilio

Fecha	Domicilio
10/09/2018	CALLE SAN JUAN COATZACOALCOS VERACRUZ
18/09/2018	CALLE SAN JUAN COATZACOALCOS VERACRUZ

En la Figura 37 podemos observar los clústeres encontrados en la ciudad de Coatzacoalcas. Se puede apreciar claramente la corta distancia de los casos agrupados por cluster, mientras que aquellos no agrupados se muestran bastante dispersos en la superficie.

Casos agrupados por DBSCAN en Coatzacoalcos

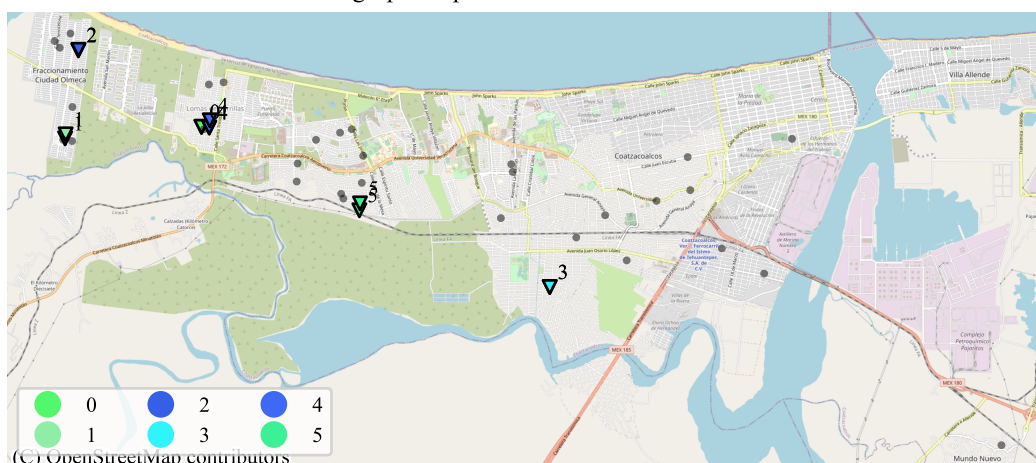


Figura 37: DBSCAN aplicado a Coatzacoalcos

El análisis de agrupamiento utilizando el algoritmo DBSCAN es adecuado para el análisis de coordenadas geoespaciales, conjuga a métrica correcta de distancia esférica y tiene ventajas con respecto a otros algoritmos como *K-Means*, una de ellas es que no requiere inicialización previa de los grupos y sus centros; el algoritmo es capaz de encontrarlos por sí mismo.

4.4.2.4 Análisis de series de tiempo

El análisis de series de tiempo se aplica a datos que se toman secuencialmente en el tiempo, es decir, tienen como índice principal un componente temporal. Este análisis permite extraer información de los datos temporales y se auxilia de distintas técnicas estadísticas, algorítmicas y de visualización de datos. Información tal como la estacionalidad, tendencia y estacionariedad son importantes para determinar el comportamiento de casos del dengue así como de la temperatura y precipitación. El análisis de series de tiempo nos permitirá comparar estos datos y relacionarlos para encontrar conocimiento. El objetivo final de este tipo de análisis, es poder hacer predicciones a futuro de las variables, es decir, puede ofrecer un estimado de casos en un periodo futuro de tiempo, permitiendo la toma de decisiones dentro de la cadena logística.

Utilizando los datos climatológicos de CONAGUA y epidemiológicos de la *DGE*, con una periodicidad de un mes del año 1985 al 2029, divididos por entidad federativa. Estos datos

son: temperatura y precipitación media, mínima y máxima; casos de dengue. Estos datos corresponden a:

1. Precipitación (mm).
2. Temperatura mínima (°C).
3. Temperatura máxima (°C).
4. Temperatura media (°C).
5. Casos de dengue.

A continuación se procedieron a empatar todas los archivos generados con base a su índice temporal para generar un conjunto de datos unificado para su análisis. Posteriormente, se formatea la columna fecha con el formato día/mes/año, siendo el mes en formato natural (Tabla 15). Luego se crearon columnas extras que contienen a la fecha descompuesta en sus tres componentes, en formato de cadena. Esto se realizó con dos objetivos:

- Poder agrupar fácilmente los datos por año y mes para poder realizar análisis de serie de tiempo y estadístico.
- Para contar con variables y características adicionales para el entrenamiento de algoritmos de machine learning.

Tabla 15: Conjunto de datos de series de tiempo

Fecha	Entidad	Casos	Precip	Tmax	Tmed	Tmin	Año	Mes
1985-01-01	Aguascalientes	8	4.91	20.71	11.97	3.22	1985	Enero
1985-01-01	Baja California	3	12.15	17.56	11.05	4.54	1985	Enero
1985-01-01	Baja California Sur	0	30.30	22.66	15.70	8.75	1985	Enero
1985-01-01	Campeche	0	20.94	29.95	22.93	15.91	1985	Enero
1985-01-01	Coahuila	0	40.71	15.06	8.44	1.82	1985	Enero

Este conjunto de datos representa una operación dentro de los datos sin procesar almacenados dentro del sistema. Suponiendo que se tienen datos de cada caso de dengue y de las mediciones de temperatura diarias, se calculan los valores anuales y a nivel nacional dentro del conjunto de datos bajo demanda. El siguiente análisis se realiza bajo este supuesto, pudiéndose cambiar la resolución de la serie de tiempo por año, mes, día y por entidad, municipio, localidad, AGEB o manzana, el proceso es el mismo.

Para iniciar el proceso de transformar datos de información, se hacen procedimientos de agregado estadístico y aritmético, para poder analizar el comportamiento de la fiebre del dengue a lo largo de los años, en cada entidad de la república.

Se observa que la entidad con mayor cantidad de casos fue Veracruz, siendo el segundo lugar, no por una gran diferencia, Jalisco. Es importante observar que la diferencia entre los dos primeros y el tercero es bastante significativa, lo cual queda abierto a investigación del porqué estas dos entidades cuentan con la mayor proporción de casos de dengue.

El histograma en la **Figura 38** nos permite visualizar la cantidad de casos de dengue por entidad. La diferencia de casos entre Veracruz y Jalisco con respecto al resto es bastante notoria.

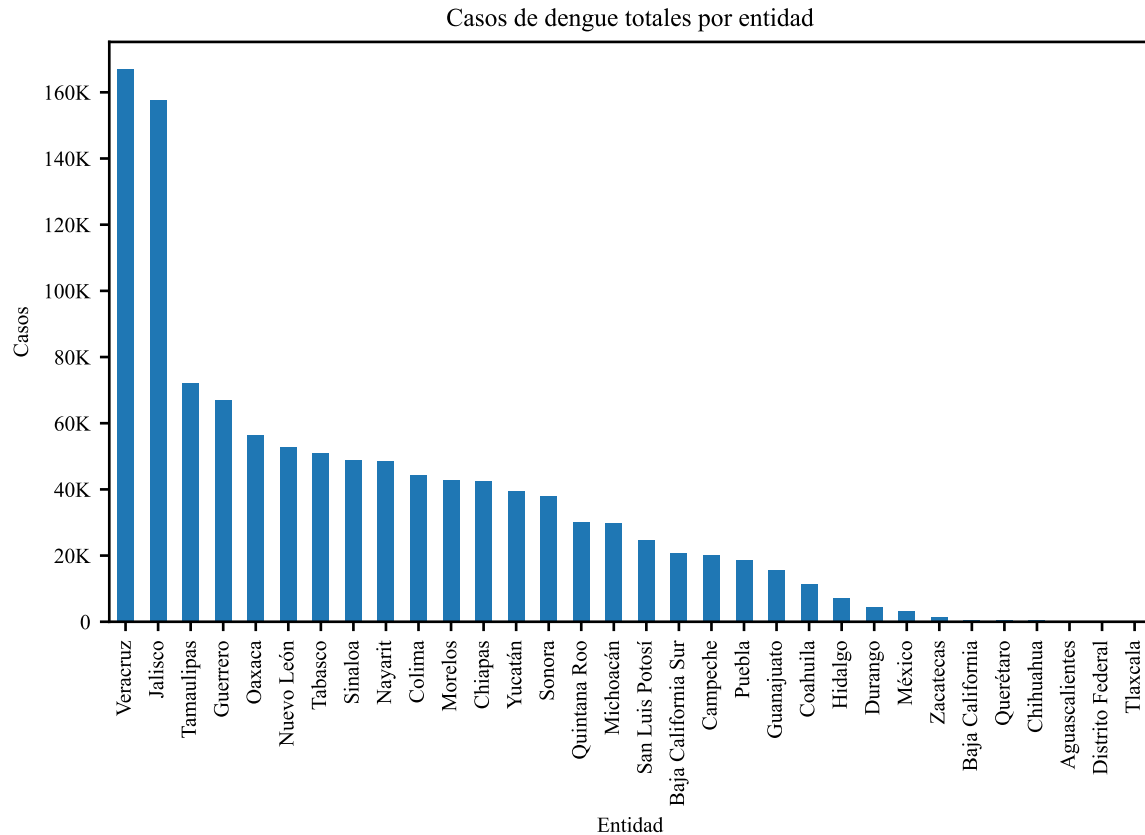


Figura 38: Casos de dengue por entidad

Se procedió a hacer un gráfico y análisis comparativo de las series de tiempo. En la [Figura 39](#) podemos observar las 5 series graficadas en conjunto, mostrando los máximos y mínimos de cada una. Varias observaciones se pueden realizar en esta gráfica:

1. Todos los máximos se encuentran a partir del año 2010.
2. Las tres series de temperatura muestra una clara tendencia hacia arriba.
3. La tendencia para Casos y Precipitación existe, pero no es tan clara, se tiene que descomponer la serie para encontrarla.
4. No se observan estacionalidades obvias.

A simple vista uno podría concluir directamente que las series no son estacionarias y que cuentan con tendencia y estacionalidad. Pero para confirmar estos supuestos se procederá a

realizar las pruebas estadísticas y análisis necesarios.

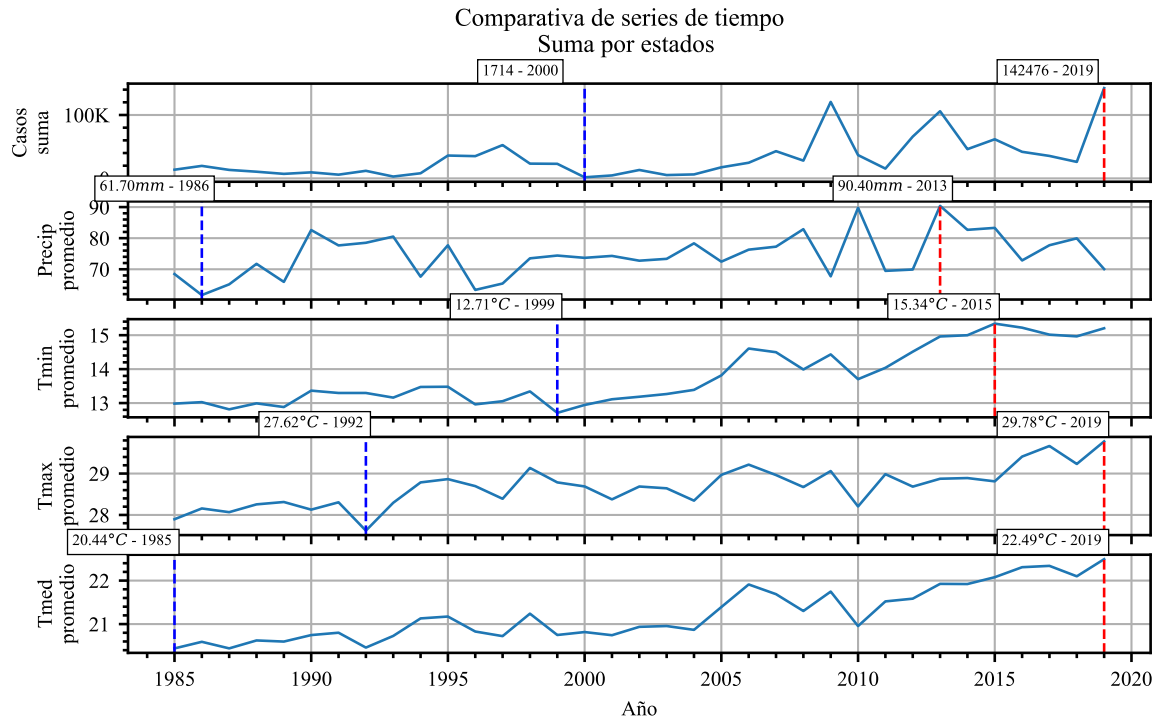


Figura 39: Comparativa de series de tiempo

Un proceso estacionario es aquel cuyas propiedades estadísticas no cambian con el tiempo. Esto no quiere decir que la serie se mantenga estática en el tiempo, sino que la forma en la que esta cambia se mantiene constante a lo largo del tiempo. Es por ello que un fenómeno estacionario es más fácil de analizar. Las series de tiempo estacionarias NO tienen tendencias o estacionalidades. Para poder determinar si existen tales patrones, se tiene que proceder a realizar un análisis de estacionariedad. Las pruebas a aplicar serán las siguientes:

- Prueba aumentada de *Dicker Fuller*.
- Prueba *KPSS* (*Kwiatkowski-Phillips-Schmidt-Shin*).

Siempre es bueno aplicar ambas pruebas, ya que si bien están relacionadas estadísticamente, cada una analiza cosas distintas dentro de la serie de tiempo. Los posibles resultados de este análisis pueden ser:

1. **Caso 1:** Ambas pruebas concluyen que la serie no es estacionaria = la serie no es estacionaria.
2. **Caso 2:** Ambas pruebas concluyen que la serie no estacionaria = la serie es estacionaria.
3. **Caso 3:** *KPSS* indica estacionariedad / *ADF* indica no estacionariedad = La serie es estacionaria bajo una tendencia. La tendencia debe ser removida para hacer la serie estrictamente estacionaria, luego, se prueba esta nueva serie para estacionariedad.
4. **Caso 4:** *KPSS* indica no estacionariedad / *ADF* indica estacionariedad = La serie es estacionaria en su diferencia, luego se diferencia la serie para hacerla estacionaria, se prueba esta nueva serie para estacionariedad.

Aparte de los casos acá presentados, existen dos implementaciones de la prueba *KPSS*, la primera revisa si la serie es estacionaria sobre una constante y la segunda sobre una tendencia. Se procede a realizar las tres pruebas:

- *ADF*: Serie no estacionaria.
- *KPSS* (c): Serie no estacionaria.
- *KPSS* (ct): Serie estacionaria con respecto a una tendencia determinista.

Por lo que se deduce que la serie puede hacerse estacionaria restando la tendencia.

Se realizan análisis de correlación y uno de correlación parcial sobre la serie de tiempo, si una serie está significativamente correlacionada, significa que los valores previos pueden ser útiles para predecir el valor actual. La correlación parcial hace lo mismo pero sin tomar los pasos intermedios. Podemos observar que la mayoría de los puntos se encuentran fuera del intervalo de confianza del 95 %, por lo que no tienen correlación significativa, por lo que podemos deducir que carece de estacionalidad (Figura 40).

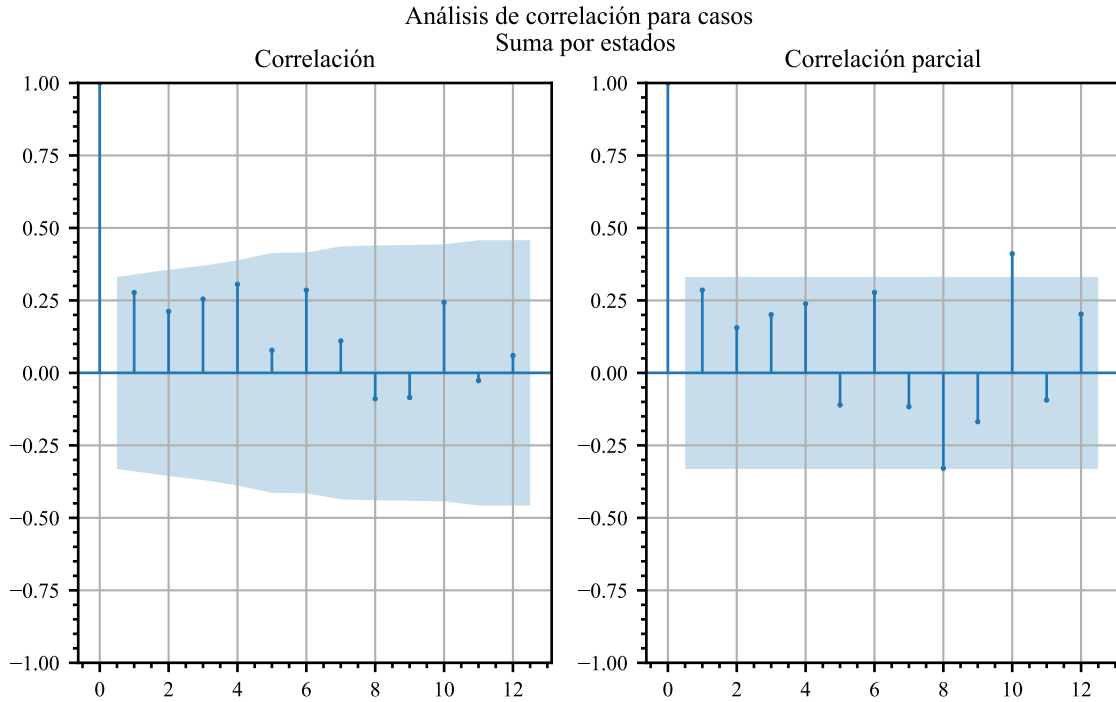


Figura 40: Correlación de series de tiempo

Toda serie de tiempo puede ser descompuesta en sus componentes:

1. **Caso aditivo:** El caso aditivo toma como supuestos que los componentes son independientes y sumados entre sí, la fórmula es la siguiente $X_t = T_t + S_t + R_t$.
2. **Caso multiplicativo:** Los componentes son dependientes y multiplicados entre sí, la fórmula es $X_t = T_t * S_t * R_t$.

Donde: $T = tendencia$, $S = estacionalidad$, $R = Residuales$.

La tendencia se presenta cuando existe una pendiente positiva o negativa dentro de la serie de tiempo, mientras que la estacionalidad se observa cuando se encuentran patrones distintos y repetidos en intervalos regulares, se asume que estos patrones suceden por factores estacionales, como el mes, año, fines de semana u hora del día.

Existen otros componentes que podrían incidir en las series de tiempo, uno de ellos es el ciclo, que se manifiesta cuando existe un patrón dentro de las series de tiempo que no depende

de factores estacionales. Debido a que existen puntos dentro de la serie de tiempo donde los casos son 0, no se utilizará el modelo multiplicativo. Se descompuso la serie de tiempo utilizando el modelo *STL*⁶, siendo Loess un método de estimación para relaciones no lineales. Tiene mejor rendimiento que otros métodos clásicos como los promedios móviles u otros más elaborados como *SEATS* o *X11*. Se realizaron descomposiciones a un periodo de 10 años. En la **Figura 41** podemos observar una tendencia en línea recta ascendente, lo que nos dice que la cantidad de casos está aumentando linealmente, no se detectaron estacionalidades marcadas aunque se observa un patrón a aumentar, por último, encontramos que los residuales difieren bastante de 0, por lo que hay comportamiento que no puede ser explicado por la descomposición (Cleveland et al., 1990).

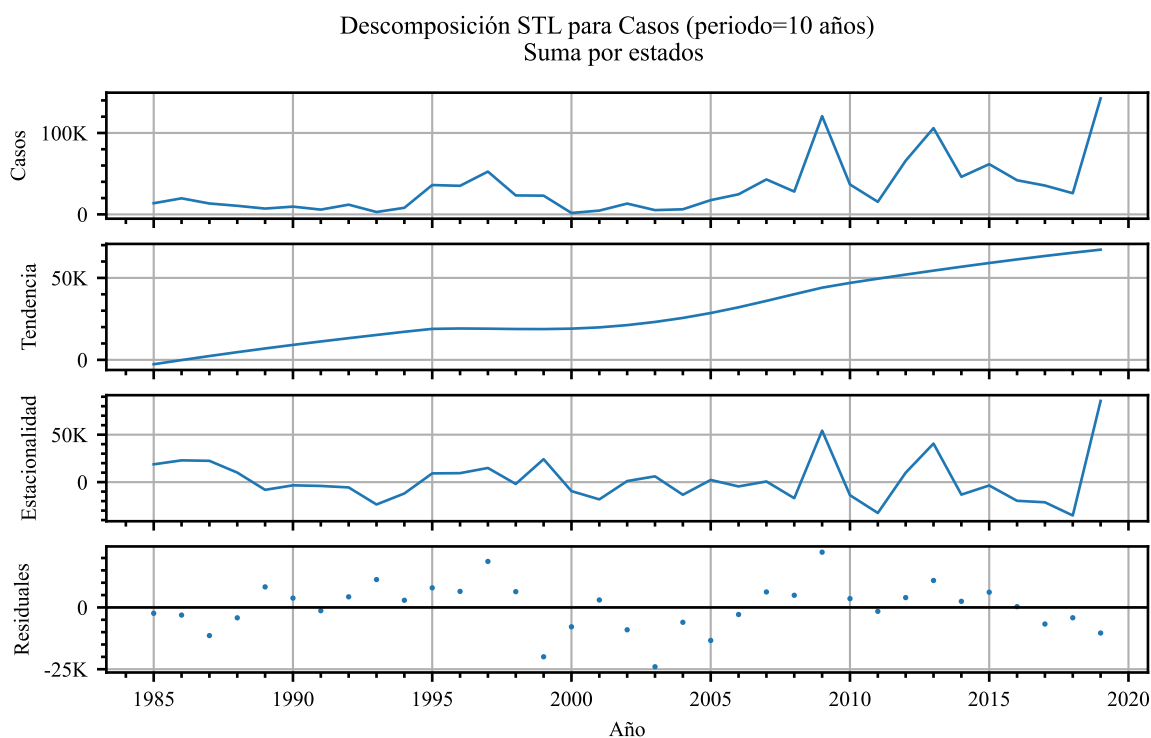


Figura 41: Descomposición de series de tiempo para los totales de casos de dengue

El propósito es mejorar la resolución tanto espacial como temporal, así que se creó otro conjunto de datos, esta vez, datos mensuales exclusivamente del estado de Veracruz. En la **Tabla 16** se observa una muestra.

⁶Seasonal and Trend decomposition using Loess

Tabla 16: Casos anuales en Veracruz

Fecha	Casos	Precip	Tmax	Tmed	Tmin	Año	Mes
1985-01-01	6	22.788344	23.231456	17.687719	12.143982	1985	Enero
1985-02-01	6	39.761199	24.168778	18.919350	13.669923	1985	Febrero
1985-03-01	6	31.119475	27.954705	22.238870	16.523035	1985	Marzo
1985-04-01	8	79.281711	29.761278	23.806926	17.852574	1985	Abril
1985-05-01	5	55.414756	31.719499	25.531910	19.344320	1985	Mayo

Al graficar esta serie de tiempo (**Figura 42**), son evidentes los picos de casos a través de los meses, con el máximo número de casos siendo agosto de 2019.

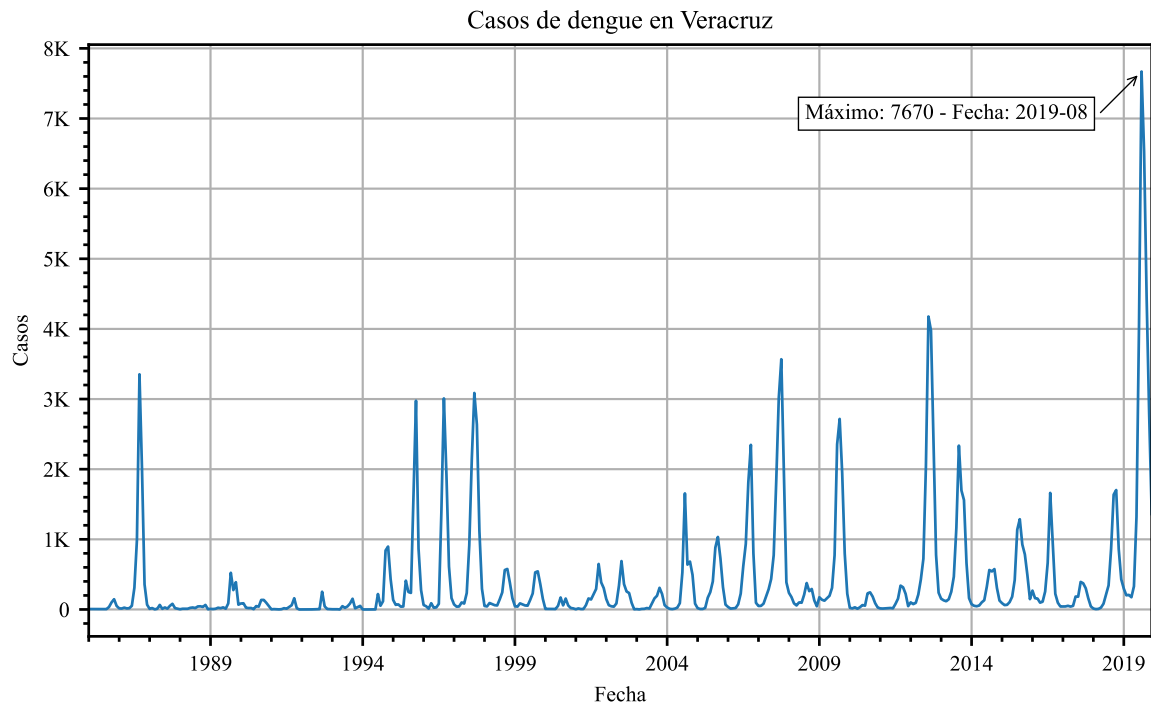


Figura 42: Serie de tiempo de casos en Veracruz

Al comparar las series de tiempo del estado, nos damos cuenta de la estacionalidad marcada, sin embargo, la tendencia no es fácilmente observable (**Figura 43**).

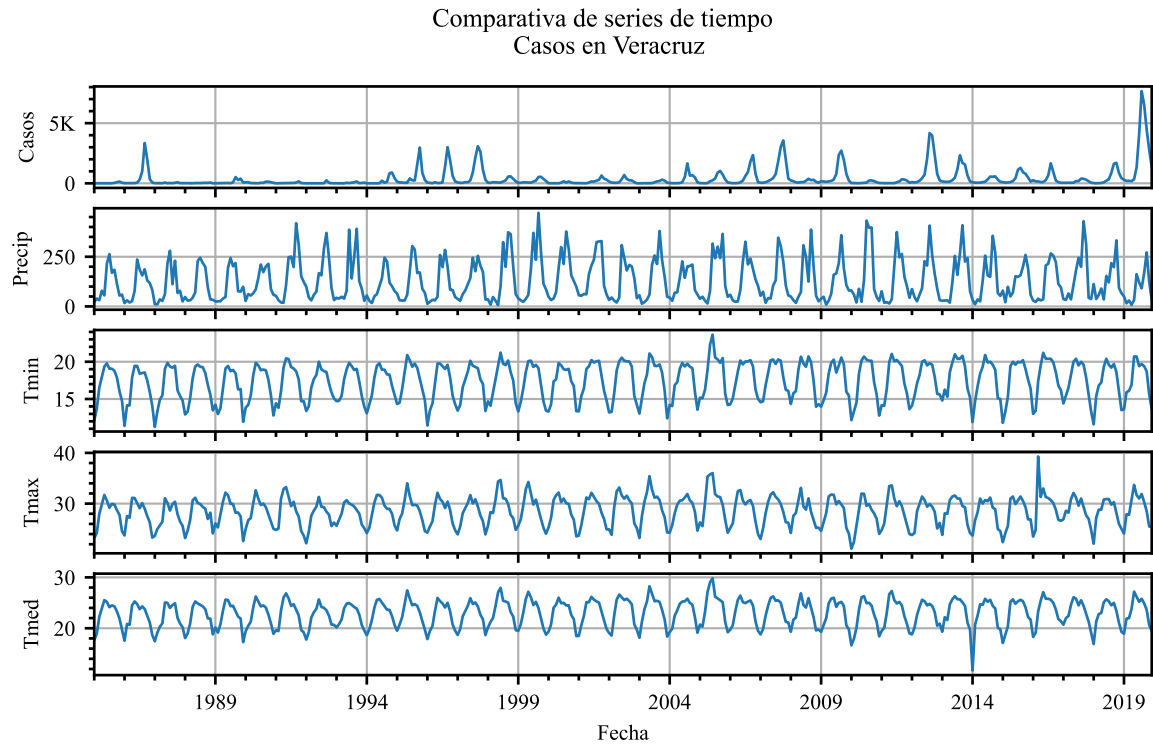


Figura 43: Comparativa de series de tiempo para Veracruz

Las pruebas indican que la serie es estacionaria en función a una tendencia determinista.

Podemos encontrar un claro componente estacional repetido cada 6 unidades (Figura 44).

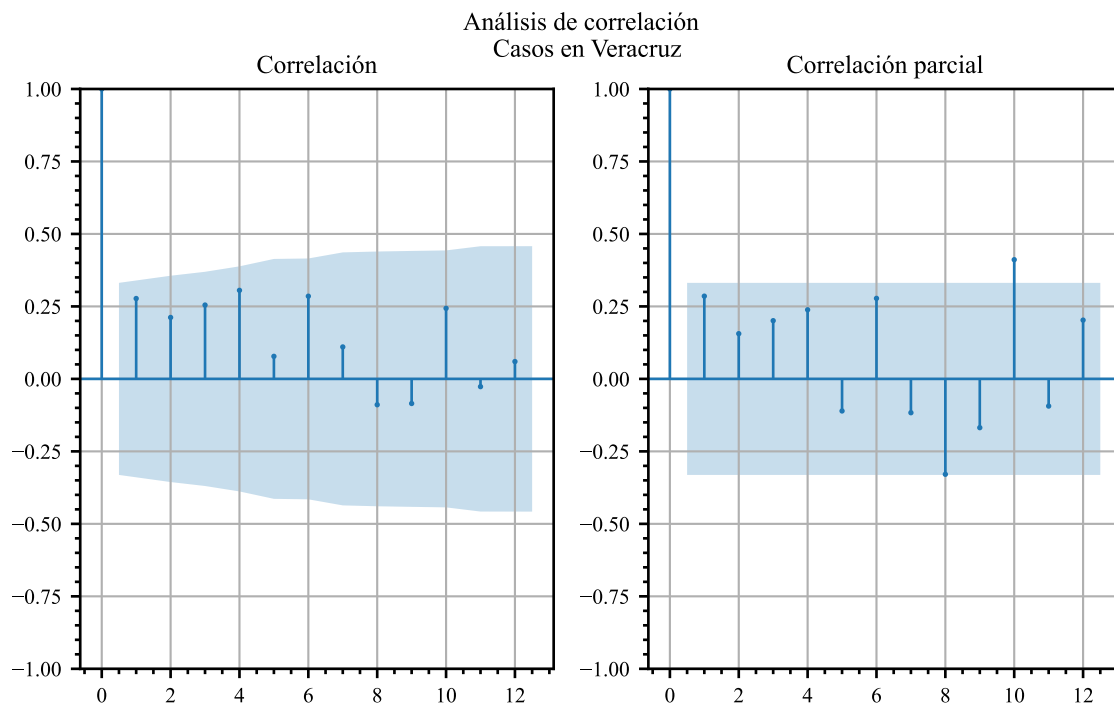


Figura 44: Análisis de correlación para Veracruz

Para encontrar los meses estacionales, hacemos una gráfica por año de todos los casos, donde podemos observar la mayor cantidad de casos en 2019, siendo los meses agosto, septiembre y octubre donde se concentran la mayoría de los casos (Figura 45).

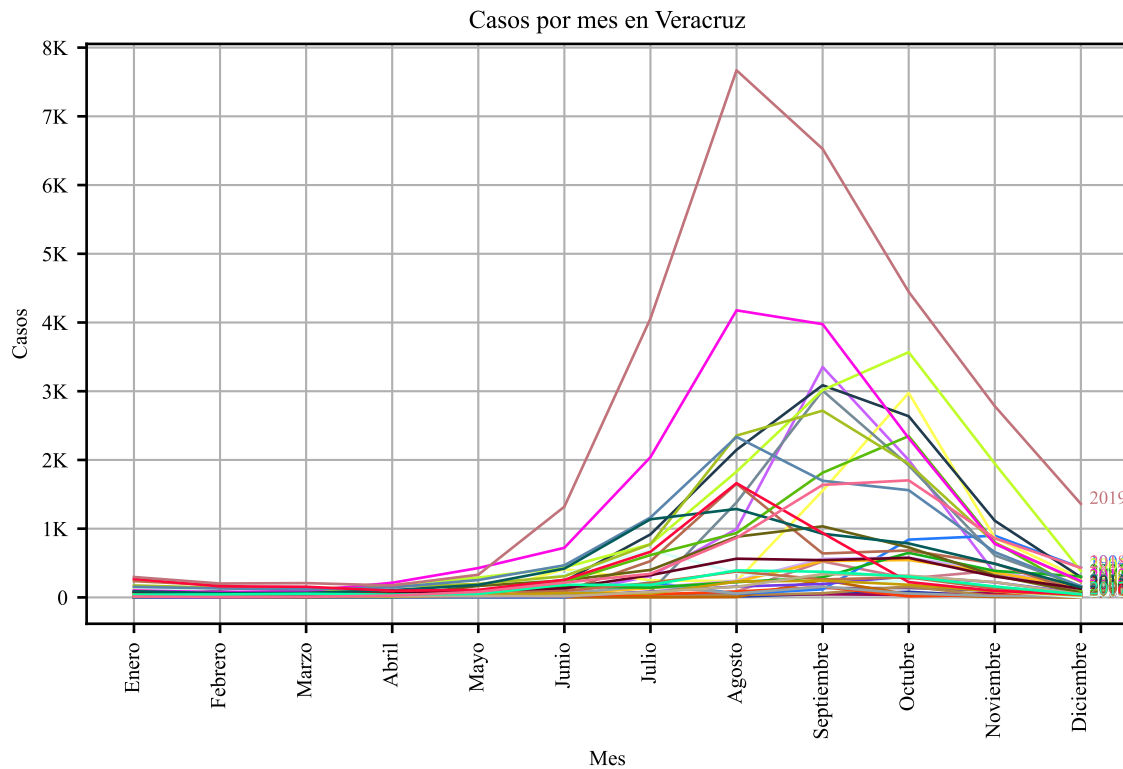


Figura 45: Comparativa de casos por año y mes para Veracruz

Al hacer un gráfico de caja agrupando datos por mes, se puede apreciar un patrón que relaciona la cantidad de casos con la precipitación (Figura 46).

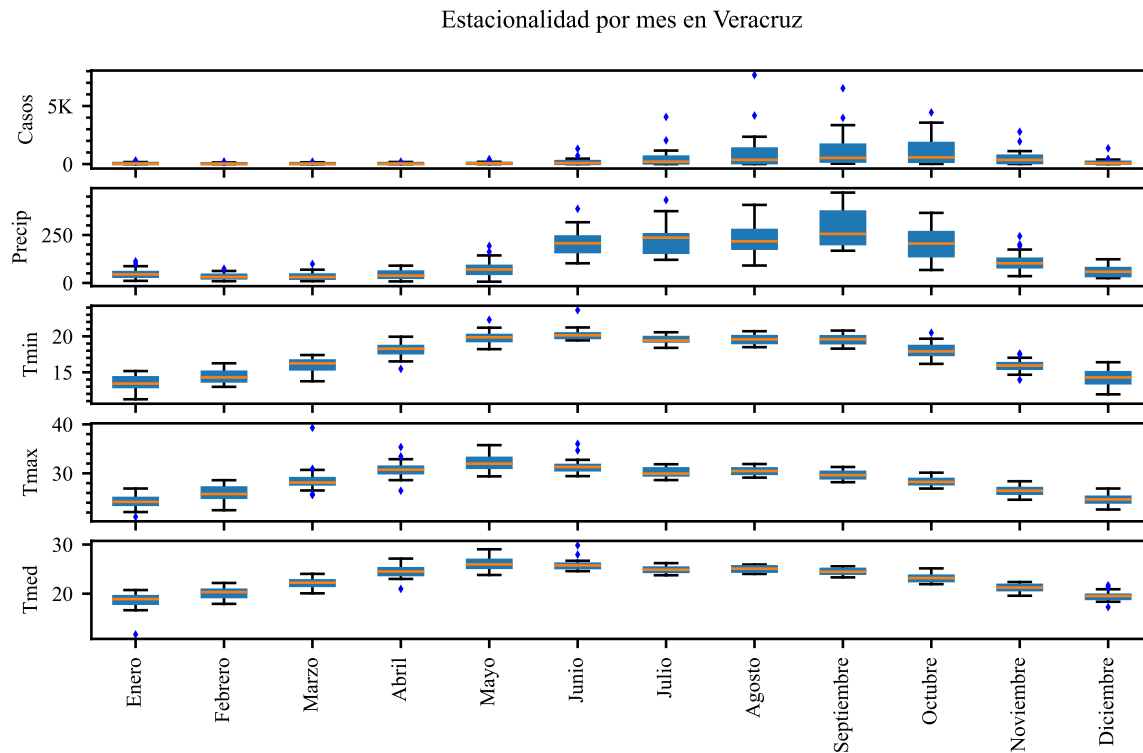


Figura 46: Gráfico de caja para determinar la estacionalidad por mes

También surgen patrones cuando se realiza este análisis agrupado por año [Figura 47](#). Se sugiere realizar un análisis más específico para determinar si existe una relación con fenómenos climatológicos globales o locales, como por ejemplo, el fenómeno de El Niño.

Estacionalidad por año en Veracruz

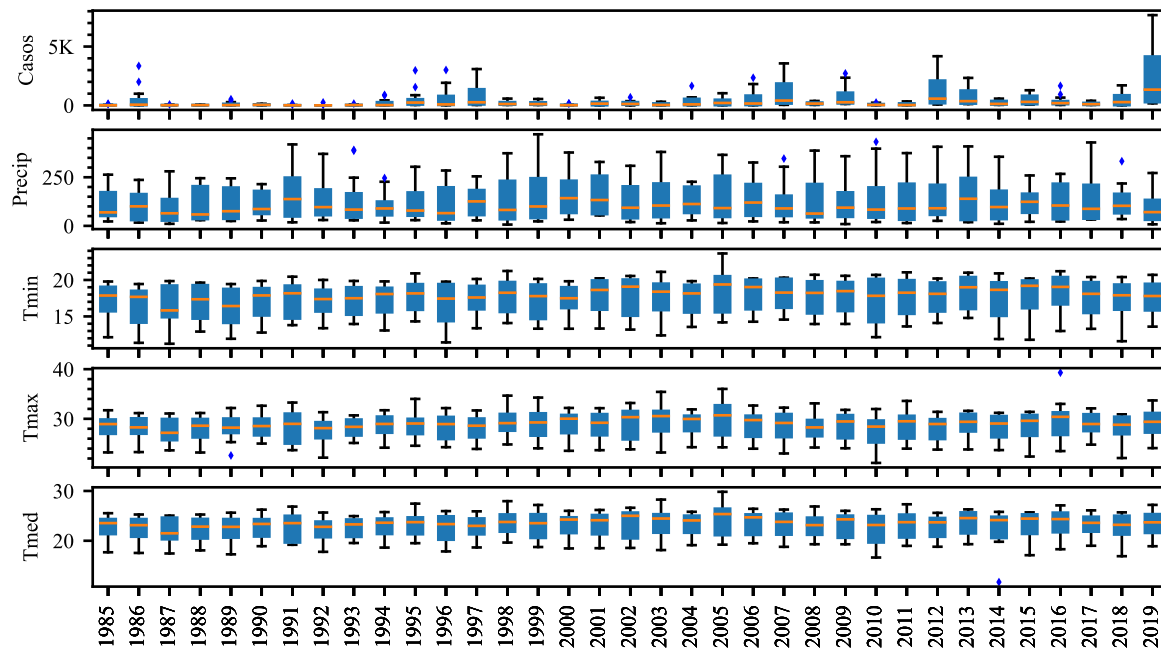


Figura 47: Comparativa

Finalmente, al descomponer los casos en el estado de Veracruz, en un periodo de 10 años, podemos observar una clara tendencia lineal positiva, lo que quiere decir que cada vez hay más casos de dengue. Mientras que curiosamente los residuales se presentan entre los años 1995 al 2009. De lo cual se deduce que los casos de dengue en el estado de Veracruz seguirán aumentando en los años venideros (Figura 48).

Descomposición STL para Veracruz (periodo=10 años)

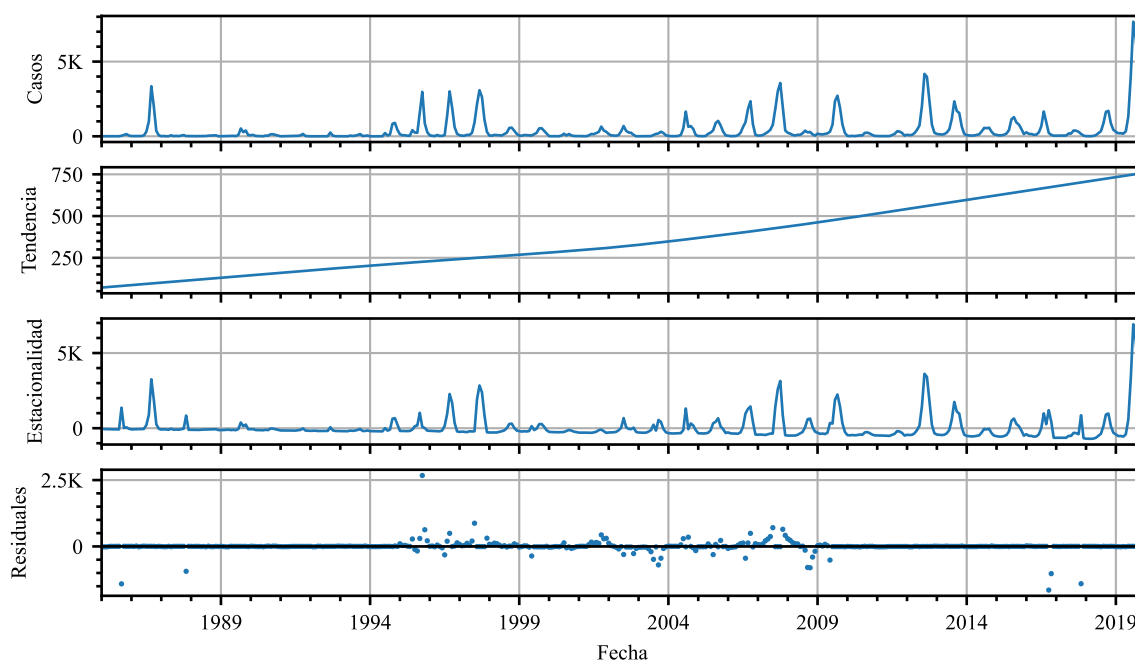


Figura 48: Descomposición de la serie de tiempo de casos en Veracruz

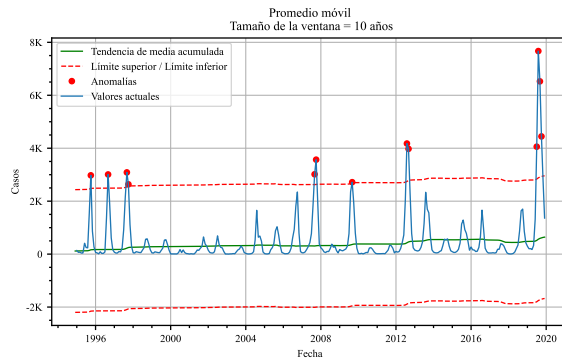
Para la predicción de casos de dengue a futuro en la serie de tiempo, existen numerosas técnicas capaces de ofrecer desde estimaciones sencillas hasta predicciones bajo certeza estadística. Primero se procedió a realizar un análisis con promedios móviles a la serie de tiempo de casos en Veracruz. Como se puede observar en la [Figura 49a](#), existen picos significativos que se pueden considerar como anomalías, los intervalos de confianza están formados por un análisis de la ventana móvil usando el error medio absoluto. La tendencia es lineal ligeramente ascendente.

Otro método relativamente sencillo para pronósticos en series de tiempo, es el suavizado exponencial. La [Figura 49b](#) muestra el proceso para dos valores de α . Este valor es el peso que se le asigna a cada observación, siendo las más recientes aquellas que cuentan más para el cálculo de la ventana. Si bien el suavizado captura relativamente bien las fluctuaciones en la serie de tiempo, no puede capturar los picos que fueron identificados como anomalías en el proceso anterior.

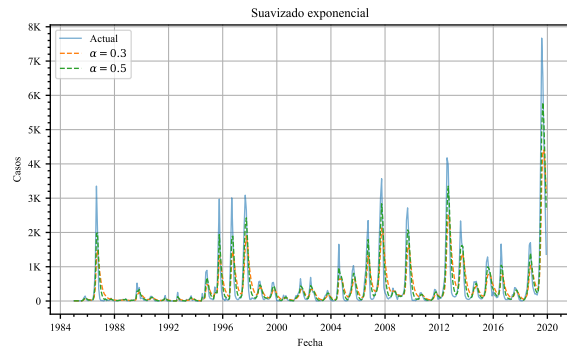
El primer modelo que se va a probar para realizar predicciones es el modelo de *Holt-Winters*, también conocido como Triple Suavizado Exponencial, es cual es capaz de capturar estacionalidad y tendencia. Este modelo fue optimizado usando métodos clásicos (newtonianos), aplicando una Validación Cruzada de 5 iteraciones. Como se puede observar en la [Figura 49c](#), el modelo no es capaz de generar un ajuste útil, las anomalías causan que el intervalo de confianza sea demasiado amplio. Se puede observar que la predicción a futuro, más allá de 2019, no es nada precisa (Winters, 1960).

Se probaron otros modelos para la predicción de casos de dengue, el proceso se desglosa en el [Anexo C](#). Este proceso incluyó distintas arquitecturas neuronales y modelos regresivos, sin embargo, su rendimiento no fue satisfactorio.

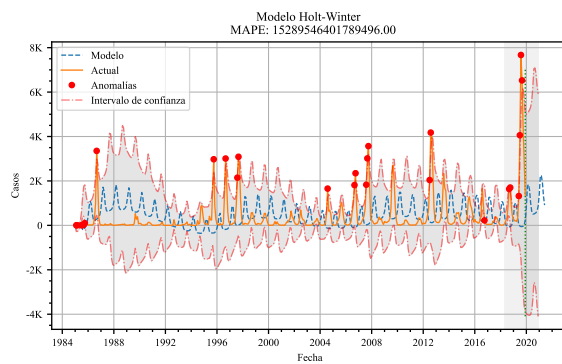
SARIMAX es un modelo estadístico de predicciones capaz de capturar información de la estacionalidad y la tendencia, es una extensión del clásico modelo *ARIMA*. Se representa la estacionalidad, MA representa los promedios móviles y X define su capacidad de procesar variables exógenas. Este modelo requiere del ajuste de varios hiperparámetros, los cuales se optimizaron utilizando algoritmos genéticos en un proceso similar al usado para el caso de estudio del COVID-19. Como se puede ver en la [Figura 49d](#), el modelo se ajusta de manera bastante aceptable a los datos de entrenamiento, mientras que en la predicción, captura los componentes estacionales de la serie de tiempo, sin embargo, poco puede hacer ante las anomalías que se manifestaron en los años 2018 y 2019, donde crecieron bastante los casos de dengue en el estado.



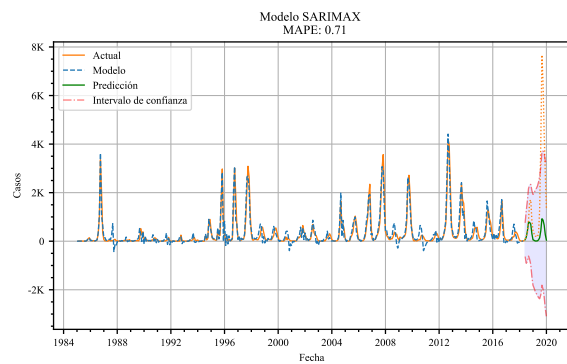
(a) Promedio móvil



(b) Suavizado exponencial



(c) Modelo Holt-Winter



(d) Modelo SARIMAX

Figura 49: Resultados de modelos predictivos

La Figura 50 nos ofrece un vistazo al rendimiento del modelo. Para los residuales estandarizados, observamos que fluctúan alrededor de la cero, sin embargo, tienen una varianza que no es uniforme. El histograma muestra un sesgo evidente hacia la izquierda, los datos no parecen ser normales. Los cuantiles muestran que la distribución está bastante sesgada.

Diagnóstico de modelo SARIMAX

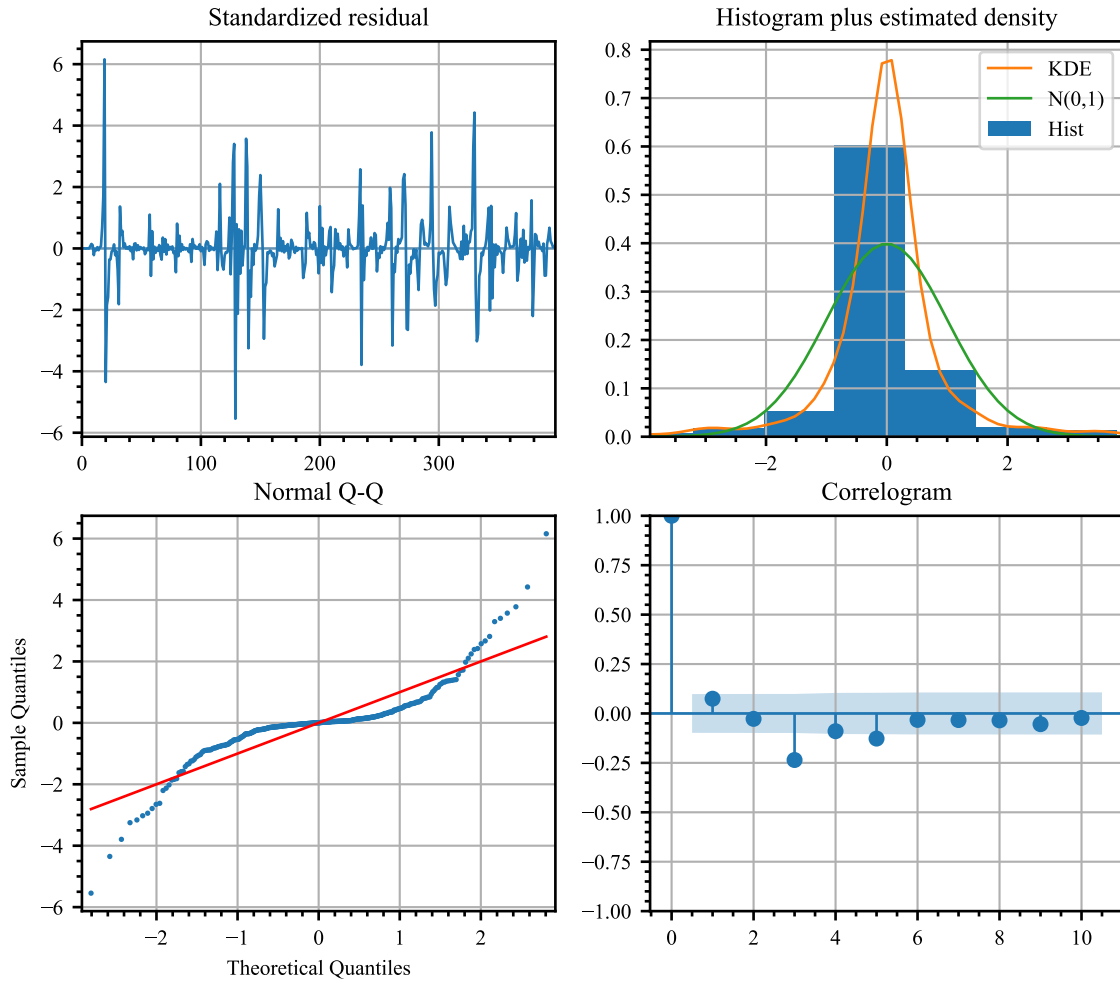


Figura 50: Resultados del diagnóstico para modelo *SARIMAX*

En la [Tabla 17](#) se concentran los resultados del ajuste del modelo, mostrándose los hiperparámetros utilizados para su ajuste, el tiempo de ajuste y el número de observaciones. Utilizando los criterios, se puede terminar que aunque el modelo es capaz de representar ciertos patrones dentro de la serie de tiempo, no es lo suficientemente potente como para capturar el comportamiento de las anomalías.

Tabla 17: Resultados del ajuste modelo *SARIMAX*

Dep. Variable:	y	No. Observations:	400
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-2911.878
Date:	Fri, 26 Nov 2021	AIC	5837.755
Time:	04:58:25	BIC	5865.678
Sample:	0 – 400	HQIC	5848.814
Covariance Type:	opg		

Los coeficientes se desglosan en la [Tabla 18](#) donde se desglosan los errores e intervalos de confianza de los coeficientes autorregresivos, promedios móviles, etcétera. El coeficiente ar.S.L.12 tiene un valor de P arriba de 0.05, lo que lo hace poco confiable.

Tabla 18: Coeficientes del ajuste

	coef	std err	z	P > z 	[0.025	0.975]
ma.L1	0.1882	0.023	8.150	0.000	0.143	0.233
ma.L2	-0.2881	0.031	-9.206	0.000	-0.349	-0.227
ar.S.L12	0.4303	0.163	2.644	0.008	0.111	0.749
ar.S.L24	0.5341	0.154	3.458	0.001	0.231	0.837
ma.S.L12	-0.2029	0.156	-1.297	0.195	-0.510	0.104
ma.S.L24	-0.5969	0.124	-4.812	0.000	-0.840	-0.354
sigma2	1.251e+05	5115.748	24.445	0.000	1.15e+05	1.35e+05

La [Tabla 19](#) ahonda sobre la composición de los residuales, si estos no son independientes, se puede tratar más al modelo y a la serie de tiempo para crear uno mejor. En la prueba de *Ljung-Box*, en la cual rechazamos la hipótesis nula y determinamos que los residuales no dependen del ruido. Por último, demostramos que los residuales no están distribuidos normalmente.

Tabla 19: Resultados del ajuste modelo SARIMAX

Ljung-Box (L1) (Q):	2.23	Jarque-Bera (JB):	1500.92
Prob(Q):	0.14	Prob(JB):	0.00
Heteroskedasticity (H):	0.99	Skew:	0.22
Prob(H) (two-sided):	0.95	Kurtosis:	12.49

4.4.2.5 Interpolación geoespacial de datos climatológicos

Derivado del trabajo previo, en función a la literatura encontrada para la predicción de eventos de dengue, se decidió realizar el intento para incrementar la resolución de las predicciones. La literatura revisada utiliza datos enfocados exclusivamente a entidades/estados de la República Mexicana, utilizando fuentes de datos públicos como Secretaría de Salud, Sistema Meteorológico Nacional, Comisión Nacional del Agua y el Instituto Nacional de Estadística y Geografía. La literatura se concentra exclusivamente en la resolución espacio-temporal de Entidad/Mes, debido a su falta de acceso a datos inmediatos epidemiológicos o climatológicos. Se trabajó en mejorar la resolución de los algoritmos predictivos, en un principio, a polígonos municipales; el problema surge, debido a que las distintas fuentes de información pública tienen diferentes resoluciones o carecen de características como históricos. Por ejemplo, CONAGUA expone un servicio web que devuelve datos climatológicos de las estaciones nacionales cada hora, pero sin ofrecer ningún histórico.

Es crucial contar con información de esta plataforma, porque nos permitiría detectar la cantidad de casos por ubicación específica donde ocurrieron, ofreciendo una resolución espacial de manzana, siendo el único factor de incertidumbre la precisión de los servicios externos de geocodificación utilizados para convertir las coordenadas geográficas en direcciones (la captura de datos en la plataforma SINAVE contempla la captura de coordenadas, el problema es en el 100 % de casos analizados no se cuenta con esta información, la captura de la dirección también carece de rigor lo cual impacta en la precisión).

Es por ello que los esfuerzos se volcaron en mejorar el entendimiento y los modelos matemáticos utilizados para la ingesta e inclusión de fuentes de datos externas. Como se comentó en anteriormente, las métricas ofrecidas por *Google Trends* se correlacionaban fuertemente con

la incidencia de la fiebre del dengue, lamentablemente este programa de vigilancia terminó, dejando un hueco bastante amplio en las fuentes de datos capaces de apoyar a los pronósticos epidemiológicos. Se realizó una indagatoria para encontrar formas de suplir esta fuente de datos, sin embargo, esta indagatoria no ha resultado fructífera.

La resolución espacial en referencia a datos climatológicos utilizados en los artículos previamente analizados, se concentró a nivel estatal. No se encontraron fuentes de información sobre clima a nivel municipal o resoluciones más bajas. Como Sistema Meteorológico Nacional, da información no solo por estado, sino también por cada una de las estaciones climatológicas nacionales, se buscó la forma de utilizar esta información para incrementar la resolución de los modelos a desarrollar. Las estaciones climatológicas nacionales no están distribuidas de una manera uniforme y algunas veces se encuentran concentradas en áreas geográficas relativamente pequeñas, como en el Valle de México (**Figura 51**).

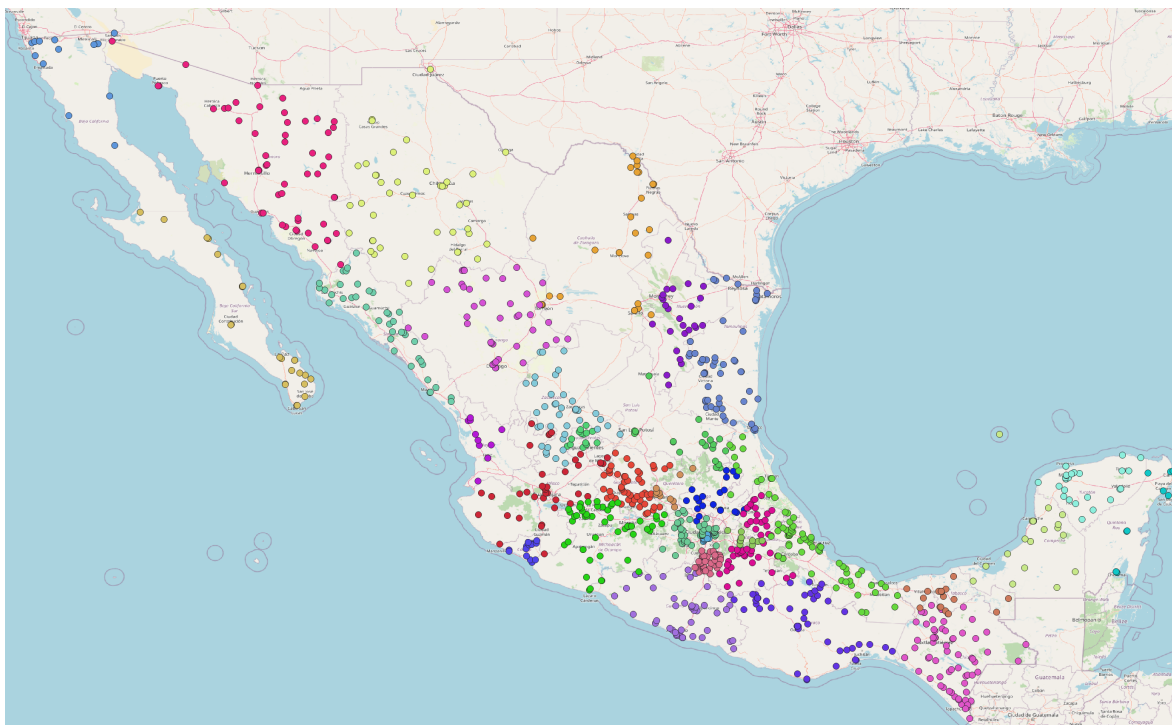


Figura 51: Estaciones climatológicas

Tomar una cantidad de valores y poder encontrar valores intermedios se conoce como interpolación. Existen distintos métodos de interpolación utilizados para climatología y análisis

geoespacial. Algunos como *Inverse Distance Weighting*, son deterministas y utilizan ponderaciones y promedios, comúnmente utilizados en aplicaciones climatológicas para generar mapas de calor; el problema de este método, es que no ofrece ninguna manera de evaluar si las predicciones son correctas o no (Esri, 2016).

Para interpolar los datos climatológicos de Temperatura Mínima mensual por estación, se decidió utilizar un algoritmo avanzado de interpolación no solo capaz de interpolar información, no crea puntos, sino que genera superficies predictivas capaces de dar información sobre las probabilidades y por consiguiente, la incertidumbre de las predicciones o valores interpolados, es decir, ofrece una medida de la confianza de que nuestra predicción sea cierta. Es importante aclarar que este algoritmo también se aplicó a las otras mediciones de temperatura y precipitación.

Este algoritmo requiere una entrada de datos, en este caso coordenadas geográficas y una variable Z a interpolar. *Kriging*, como se conoce a este algoritmo, es bastante similar a otros algoritmos de interpolación probabilística como los modelos Gaussianos, por lo que se podría considerar como un algoritmo de *Aprendizaje Automático*.

Los modelos de interpolación geoespacial usando el método de *Kriging* tienen como objetivo incrementar la resolución espacial de los datos climatológicos extraídos de CONAGUA, ya que las estaciones de medición se encuentran repartidas de manera heterogénea en el territorio nacional. Mejorar las resoluciones tanto espaciales como temporales mejorará la capacidad predictiva de los algoritmos permitiendo tomar mejores decisiones (Boulos et al., 2019).

La técnica de *Kriging*, se refiere a un algoritmo geoestadístico capaz de generar una superficie bidimensional a través de pares de coordenadas con un valor Z asignado, capturando el comportamiento espacial del fenómeno subyacente. El algoritmo realiza suposiciones sobre la interacción entre los puntos y el valor Z. Es un proceso complejo que consiste en varios pasos como el análisis estadístico exploratorio, generación de modelos de variograma, creación de la superficie y evaluaciones diagnósticas.

En la Figura 52 podemos observar una comparativa entre *Kriging* y otro algoritmo de interpolación, se ven claramente las ponderaciones que hace el algoritmo sobre todos los puntos para

determinar su relación espacial.

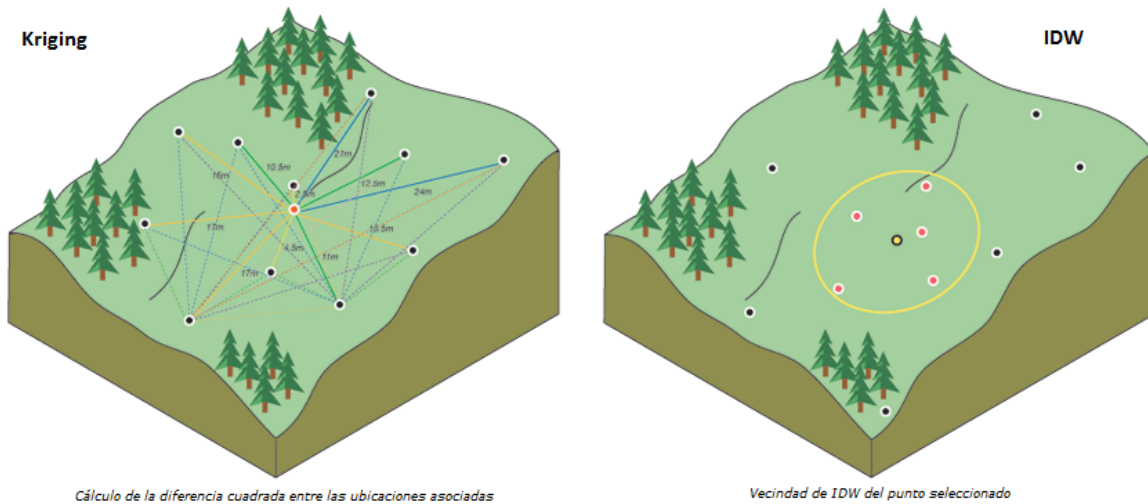


Figura 52: Comparativa de métodos de interpolación geoespacial

Uno de los problemas encontrados al implementar este algoritmo, es la cantidad de memoria RAM requerida para correrlo. El equipo de trabajo cuenta con el procesador más rápido desde que salió al mercado en su momento y viene acompañado con 64 GB de memoria RAM, el máximo soportado por la arquitectura. Estos recursos llegaron a su límite constantemente en lo que se entrenaba el algoritmo, requiriendo utilizar espacio de disco duro como apoyo a la memoria RAM, para ello se utilizó un disco de estado sólido de baja latencia como apoyo. Es importante tener en cuenta estas restricciones computacionales para la implementación posterior del **Sistema de Información**.

El algoritmo de *Kriging* cuenta con distintos hiperparámetros que hay que ajustar para encontrar un buen modelo. Estos hiperparámetros son: método, modelo de variograma y nlags. El espacio de búsqueda de hiperparámetros para cualquier algoritmo de entrenamiento computacional es bastante amplio, en este caso, se tiene la ventaja de que se está trabajando en el conjunto de los naturales y en variables que son categóricas. Para apoyarnos a encontrar los hiperparámetros que darán los mejores resultados, se aplicó un algoritmo de optimización para reducir el error en la estimación del modelo de *Kriging*. El algoritmo elegido fue el clásico *Algoritmo Genético* aplicado a optimización de hiperparámetros. El algoritmo genético trabajó con un tamaño de población de 50, una probabilidad de mutación de 0.10 y 10 generacio-

nes. Para validar este modelo se realizó una validación cruzada de 5 iteraciones. Primero se dividen los datos en $K \in \mathbb{N}$ y se realiza la optimización una vez por cada subconjunto, la salida será el rendimiento del experimento así como los hiperparámetros encontrados por el proceso (Kramer, 2017) (Vapnik, 1998).

Este proceso tiene como resultado un diccionario con los mejores hiperparámetros con los cuales se entrenará el modelo utilizando todos los datos. El modelo después se convertirá en una cadena binaria que funge como una captura del estado de memoria que almacenaba tal modelo, para posteriormente ser insertado en una tabla con un atributo para bytes, reduciendo la cantidad de espacio en disco necesaria para los modelos generados por todo este proceso.

Los hiperparámetros que reducen el error en el entrenamiento del modelo, obtenidos por el algoritmo genético fueron: *ordinary*, *spherical*, 46 y *euclidean*. Terminado el entrenamiento podemos hacer una comparativa entre los mapas de contorno generados por CONAGUA y el generado por nuestro modelo (Figura 53). Encontramos una similitud bastante grande. La ventaja de la metodología aplicada es que contamos un modelo general al que, al ingresar un par de coordenadas, nos devuelve la temperatura estimada en ese punto así como el nivel de confianza de esa estimación.

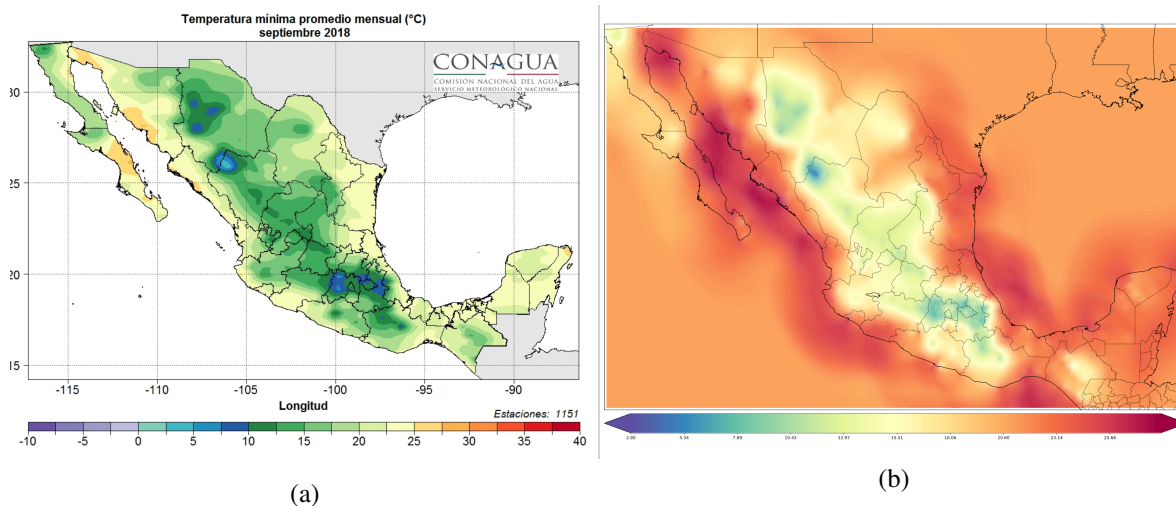


Figura 53: Comparación entre resultados CONAGUA (a) y Kriging (b)

Con el modelo de *Kriging* entrenado, contamos con una representación climatológica total de la superficie nacional y podremos encontrar cualquier variable climatológica. Esto nos ayudará

a incrementar la resolución espacio-temporal para posteriores aplicaciones algorítmicas.

Cabe destacar el uso de metaheurísticas para la optimización de algoritmos de entrenamiento, esto se conoce como AutoML, que es el área de la inteligencia artificial que trata de habilitar el entrenamiento acertado de algoritmos sin intervención humana, ideal para aplicaciones como los Sistemas de Información y Sistemas de Decisión.

4.4.2.6 Metodología final propuesta

Los casos de estudio previos y el análisis de datos de fuentes auxiliares, externas y las muestras obtenidas previamente, permitieron desarrollar técnicas, algoritmos y métodos para el procesamiento de datos epidemiológicos que surgen dentro de la cadena logística de salud pública, en particular, aquella que trata con la gestión de la Fiebre del Dengue.

Este proceso previo permitió determinar el flujo de datos y entender como transformarlos en información, conocimiento y, finalmente, sabiduría. Como vimos anteriormente, este proceso de extracción de conocimiento, aplica distintas metodologías y enfoques para gobernar la minería y el descubrimiento de conocimiento dentro de bastos conjuntos de datos. Si bien existen metodologías previas que tratan de encapsular este proceso como *CRISP-DM* (Shearer, 2000), pretenden ser demasiado generales por lo cual, en aplicaciones especializadas como el área médica, carecen del poder expresivo necesario para poder domar la complejidad de los datos.

Para esta última fase del caso de estudio, se trabajó directamente con la totalidad de datos que existen dentro de la plataforma SINAVE. Lo cual permitió probar y validar el trabajo previamente realizado, así como servir como base para la creación de una metodología particular para la gestión de datos epidemiológicos que surgen de la cadena logística de salud pública.

Lo importante de este proceso es que conjugó toda la pericia y experiencia previamente generadas de una manera directa, es decir, el mismo código utilizado para la muestra, fue utilizado para el conjunto total de datos. Así mismo, por su naturaleza, para poder trabajar y manipular esta gran cantidad de datos, se tuvieron que utilizar de manera frontal y directa

conceptos de *Industria 4.0* como el *Big Data* y el *Aprendizaje Automático*.

Una vez que se tuvo acceso a la totalidad de los datos, se desarrolló una metodología para poder aprovecharlos y refinarlos en sabiduría. La metodología propuesta para este caso en particular, cuenta con las siguientes etapas:

- **Obtención de datos:** Se compone de todos los pasos necesarios para la obtención de datos, incluyendo los repositorios donde se almacenarán, las rutinas necesarias para su descarga o cambios de localización de los mismos.
- **Manejo y gestión de datos:** Se refiere a convertir los datos de su forma básica en una forma apta para ser utilizados; por ejemplo, cambiar el formato de los datos, resolver problemas de codificación y versionado de los datos.
- **Descripción de datos:** Establece las rutinas básicas de análisis exploratorio para conocer las características de los datos; se extraen las características básicas como tipo de dato, cantidad, número de columnas/características y columnas objetivo.
- **Definición del problema:** Consiste en delimitar correctamente los alcances de uno o más análisis y permite determinar la familia de modelo o modelos a utilizar, las métricas objetivo, alcances del proceso, etcétera.
- **Limpieza y preparación de datos:** Encapsula los pasos para convertir una base de datos en un conjunto apto para la extracción de información, conocimiento y sabiduría; incluye etapas técnicas como la imputación de datos faltantes y etapas legales como la disociación, estipulada en la Ley Federal de Protección de Datos Personales(UNIÓN, s.f.).
- **Ingeniería de Características:** Determinar las características (columnas) necesarias para maximizar el rendimiento de uno o más modelos para la tarea definida en el primer paso; también puede transformar o generar nuevos datos para este propósito.
- **Ingeniería de modelos:** Se refiere a todas las disposiciones necesarias para entrenar un modelo asegurando una certeza estadística en su aprendizaje y permitiendo que su

poder de generalización sea el correcto; involucra técnicas como la validación cruzada y conceptos como sesgo y varianza.

- **Evaluación del modelo:** Determina si el rendimiento del modelo es adecuado en función al tipo de problema, las métricas definidas y la tarea a realizar; debido a la aplicación en el área médica se requiere un proceso de evaluación robusto.
- **Sesgos algorítmicos:** Debido a los entornos tecno–sociales de la aplicación del trabajo, se requiere determinar si los algoritmos entrenados tienen sesgos sociales o discriminan hacia uno o más sectores poblacionales; evitando así que los modelos reproduzcan algunas de las desigualdades estructurales que permean las sociedades modernas(Krishnapuram et al., 2016).
- **Explicabilidad:** Tiene como objetivo dilucidar como los modelos toman decisiones para transparentar el proceso y permitir que este sea integrado más fácilmente en los procesos de decisión realizados por los tomadores de decisiones.
- **Despliegue del modelo:** Una vez obtenido el modelo, se despliega dentro del Sistema Híbrido Inteligente para poder ser usado por los tomadores de decisiones; este proceso involucra la orquestación de distintas tecnologías y componentes de la *Industria 4.0*.
- **Monitoreo del modelo:** El ciclo de vida de un modelo no termina en su despliegue, es necesario continuar monitoreando el rendimiento para determinar si existe una degradación en su rendimiento por los cambios que surgen en los datos que se generan dentro de la cadena logística(K. Nelson et al., 2015).
- **Ajuste y calibración del modelo:** El paso anterior permite reajustar y calibrar el modelo para que su rendimiento se mantenga óptimo a lo largo de las aplicaciones del mismo.

Esta metodología propuesta, no solo expande los pasos básicos expuestos en *CRISP–DM*, sino que involucra conceptos “blandos” que anteriormente eran relegados a un segundo plano, como la explicabilidad y los sesgos algorítmicos. Estos conceptos son de suma importancia en aplicaciones médicas, ya que permiten que los tomadores de decisiones puedan confiar en los modelos y asegurar que toman la mejor decisión posible. Es aquí donde se hace evidente

que la *Industria 4.0* no solo es un paradigma industrial productivo, sino que también tiene el poder de incidir positivamente en temas sociales de crucial importancia. Sobre todo en países en vías de desarrollo como México, donde la magnitud de las desigualdades sociales aún no se describe en su totalidad.

Otra característica que distingue la metodología propuesta, es que la fase de Definición del Problema se realiza de manera posterior a la Descripción de los Datos. Lo que permite que el problema se defina en función a los datos y no al revés. Esto es de suma importancia, ya que en aplicaciones médicas, los datos son el resultado de un proceso de captura que no siempre es el más adecuado, por lo que es necesario conocerlos antes de definir el problema. También, permite ahorrar tiempo y recursos, puesto que no se definen problemas que no se pueden resolver con los datos disponibles, o que no tienen la suficiente información para ser resueltos o trabajar en problemas similares y así realizar solo una vez el proceso de extracción de conocimiento.

4.4.2.7 Modelos predictivos

Para validar esta metodología, se realizó todo el proceso de obtención y obtención de sabiduría usando los datos de la plataforma SINAVE.

A continuación se desglosan los resultados obtenidos en cada una de las etapas de la metodología.

4.4.2.8 Obtención de datos

Esta etapa involucró el acceso a la plataforma SINAVE para la descarga de datos. Solamente se tuvo un punto de acceso inicial; un equipo de cómputo de principios de la década de los 2000, situado en las instalaciones del Hospital de Especialidades del IMSS en la ciudad de Orizaba, Veracruz.

Para acceder a la plataforma, se requiere un conjunto de credenciales únicas para cada usuario, tales credenciales se generan de forma centralizada y el público en general no tiene acceso a las mismas. Gracias al convenio específico de investigación celebrado entre el TecNM y

el IMSS, se tiene permiso para la extracción de los datos y almacenamiento de los datos de manera privada, sin embargo, esto requería forzosamente la interacción supervisada con el equipo de cómputo en cuestión.

Por cada archivo, debido a las características el equipo y de la red, se tardaba un aproximado de 40 segundos en descargarse mediante la interfaz gráfica del explorador de internet. La plataforma no ofrece ninguna API para poder extraer los datos mediante el protocolo HTTP. Para evidenciar el problema que esto representa, los tiempos de descarga siguiendo este proceso serían: $70,804 \text{ archivos} \times 40 \text{ segundos} = 786 \text{ horas}$. Lo cual es prohibitivo y requiere una supervisión constante, cosa que es imposible debido a las responsabilidades del personal IMSS.

Para solventar este problema, se procedió a explorar el código fuente de SINAVE para poder explotar alguna vulnerabilidad utilizando técnicas de *Hacking ético* (Smith et al., 2002). Se encontraron diversas vulnerabilidades, sin embargo, la que se explotó para permitir la descarga de datos fue aquella que expone el URL particular de cada archivo existente dentro de la plataforma. Este URL carece de seguridad por lo que una petición desde cualquier parte del mundo puede acceder al recurso que identifica. Esto es una violación clara a los conceptos de ciberseguridad (Lezzi et al., 2018), mostrando de nueva cuenta lo imperioso que es la aplicación del paradigma de *Industria 4.0* en el sector público de salud.

Se creó un *Script* en Python multinúcleo, multihilo y asíncrono para la descarga de todos los datos. Tomando finalmente un total de tres horas en completar la tarea.

4.4.2.9 Manejo y gestión de datos

Los recursos presentes dentro de la plataforma almacenan varios estudios epidemiológicos, los concentran en archivos de texto plano y se comprimen en formato *.RAR*. Estos archivos fueron descomprimidos en directorios agrupados por entidad federativa. Los archivos de texto se encuentran codificados en *Latin-1* y separados por el delimitador *|*. Se procedió a convertirlos a formato *.CSV* y codificación *UTF-8* para su posterior procesamiento.

4.4.2.10 Descripción de datos

Los datos descomprimidos constituyen un total de 114 GB, 12 millones filas y 490 columnas. Contienen información de uno o varios estudios epidemiológicos e incluyen datos clínicos, organizacionales, personales, geográficos, de laboratorio, etcétera. Las columnas varían en cada archivo, por lo que se tuvo que unificar la información y fusionarlos en conjuntos de datos más grandes con las columnas corregidas.

Las características de estos datos los hacen candidatos ideales para el uso de técnicas de *Big Data*, ya que incluyen las tres Vs del paradigma: Volumen, Variedad y Velocidad (Wasem, s.f.). Su volumen hace que sea difícil trabajarlos en memoria RAM, por variedad se refiere a que los archivos no tienen una estructura uniforme y la velocidad se refiere a la velocidad de generación de los datos, que por lo que se ha podido ver, es de varios archivos con distintos estudios epidemiológicos por día.

Se identificaron varias columnas que pueden servir como pauta para establecer y definir problemas, por ejemplo, diagnosticar una persona con dengue, clasificar el tipo de dengue, determinar el fallecimiento. También se pueden transformar los datos en otras representaciones para crear otros análisis, por ejemplo, convertirlos en series de tiempo.

4.4.2.11 Definición del problema

Para este enfoque inicial se decidió trabajar con el problema de clasificar los tipos de dengue. En contraste a diagnosticar si una persona tiene dengue o no, este proceso requiere menos datos, ya que trabajamos con el subconjunto ya previamente diagnosticado.

Los datos contienen tres tipos de dengue: grave, no grave y con signos de alarma. Para este problema, se decidió convertir el problema en binario, agrupando los casos graves y con signos de alarma en una sola categoría. Esto se hizo con base a sugerencias del personal médico (Figura 54).

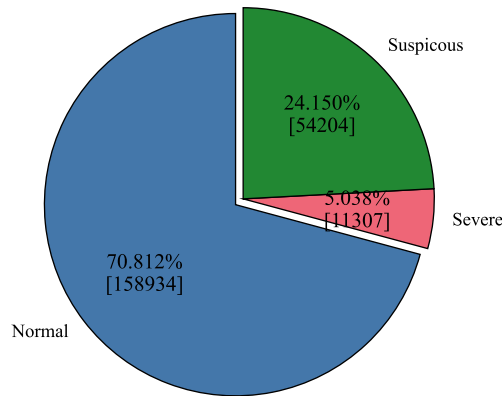


Figura 54: Proporción y número de casos por clase

4.4.2.12 Limpieza y preparación de datos

Primero se eliminaron las columnas relacionadas con la información del paciente, posteriormente se filtraron las columnas que contienen datos administrativos y organizacionales. Se eliminaron datos duplicados, casos diagnosticados con otras enfermedades.

Finalmente se obtuvieron aproximadamente 2M de datos viables. De estos datos se tomó una muestra del 10 % para trabajar con ellos, debido a que el proceso de entrenamiento de los modelos es computacionalmente costoso. La cantidad final de filas fue de 224,446.

Se decodificaron valores numéricos en las columnas por su descripción categórica, por ejemplo, Hombre en lugar de 1, etcétera. Se dejaron los valores faltantes como *NaN* para ser gestionados dentro del proceso de Ingeniería de Modelos. Con el objetivo de generar modelos robustos a la incertidumbre y falta de integridad en los datos.

Datos que se salen de rangos establecidos, como por ejemplo, la temperatura no puede ser mayor a cierto valor, fueron convertidos en faltantes para ser gestionados por los modelos.

Esta fase suele contener procesos como la Imputación y Escalamiento de los datos. Se delegó este procedimiento a la fase de Ingeniería de Modelos para probar distintas técnicas y ver como estas impactan en el rendimiento final.

4.4.2.13 Ingeniería de características

En esta fase se puso especial atención a los datos que contienen información de análisis clínicos. Se tienen datos de hasta 4 instancias de estudios hematológicos por paciente. Estas columnas fueron unificadas y promediadas con el objetivo de obtener concentrar la información en menos columnas y reducir la complejidad de la tarea. Otras columnas unificadas fueron las referentes a comorbilidades o enfermedades que no tienen una columna específica, sino que se encuentran en una columna de texto libre.

Se obtuvieron 80 columnas y una columna objetivo. Se apartaron el 20 % de los datos para la evaluación de los modelos. El 80 % restante se utilizó para la búsqueda de hiperparámetros y entrenamiento de los modelos. Este proceso se conoce como Validación Cruzada con Retención.

4.4.2.14 Ingeniería de modelos

El proceso de Ingeniería de Modelos se llevó a cabo siguiente dos enfoques: el primero, encontrar el mejor algoritmo para la tarea entre un conjunto de algoritmos preestablecidos y el segundo, encontrar los mejores hiperparámetros para tal algoritmo.

Para asegurar la certeza estadística, se aplicó el método más robusto posible de entrenamiento, la validación cruzada anidada. La cual consiste en separar los datos en $K \in \mathbb{N}$ subconjuntos de manera inicial y posteriormente, dividir esa partición en otros $K \in \mathbb{N}$. Cambiando en cada subconjunto los datos, modelos e hiperparámetros para reducir el sesgo y la varianza y tener una estimación precisa del rendimiento de cada modelo. Para el ciclo exterior se definió $K = 5$ y para el ciclo interior $K = 3$.

Debido a ciertos conceptos teóricos (Wolpert & Macready, 1997), es imposible saber *a priori* si un algoritmo se va a comportar mejor que otro para una tarea en particular. Es por ello que se probaron distintos algoritmos para dar con el mejor para el problema de clasificación binaria de tipos de fiebre del dengue. Los algoritmos probados fueron dos lineales: *Regresión Logística* y *Análisis Discriminante Lineal*; dos no lineales: *Máquinas de Soporte Vectorial* y *Perceptrón Multicapa*; y dos de ensamble: *Bosques Aleatorios* y *Potencialización del Gradiente mediante*

Histogramas.

Para encontrar el mejor modelo y sus hiperparámetros, se utilizó un método de fuerza bruta conocido como *Búsqueda en Cuadrícula*. Para optimizar los parámetros del mejor modelo, se utilizó un proceso de búsqueda bayesiana. La primera consiste en probar todas las combinaciones posibles de hiperparámetros dentro de un espacio de búsqueda predefinido, mientras que la segunda realiza una búsqueda aleatoria y posteriormente mejora en cada iteración utilizando métodos de *Inferencia Bayesiana* (Figura 55).

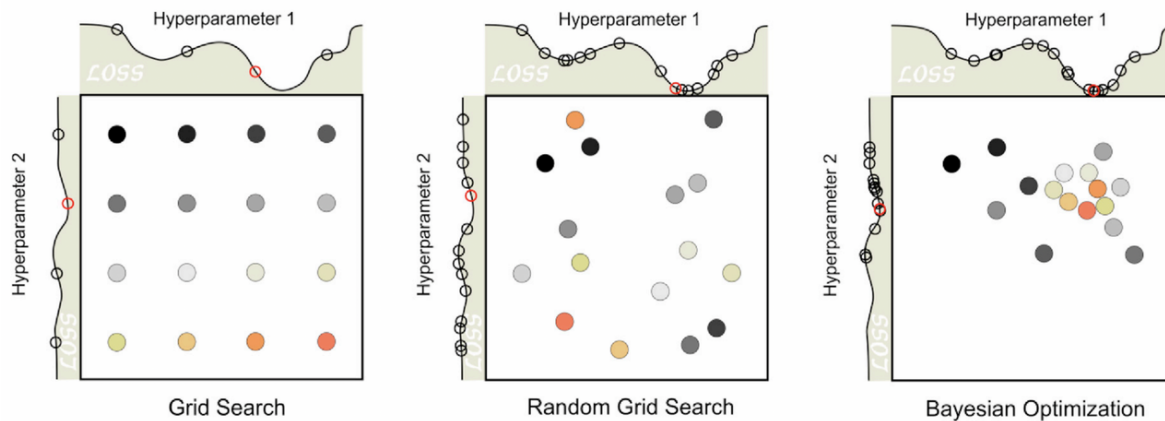


Figura 55: Proceso de validación cruzada anidada

Otros hiperparámetros que se optimizaron, pero no son específicos para cada algoritmo, fueron la técnica de imputación y el escalamiento. Para la imputación de datos numéricos, se probó con la media, mediana, moda y un valor constante -999 ; los datos categóricos se imputaron con la categoría “desconocido”. Esto hará al modelo robusto ante la incertidumbre y datos faltantes. Para el escalado de datos se probó con la normalización y estandarización.

Esta fase requirió un poder computacional enorme y aun así tardó 5 días en completarse. Los resultados de la validación cruzada se pueden observar en la Figura 56. Podemos ver que el mejor algoritmo fue el de *Bosques Aleatorios*, aunque tuvo una varianza bastante considerable, aunque no mayor que otros algoritmos como *KNN*.

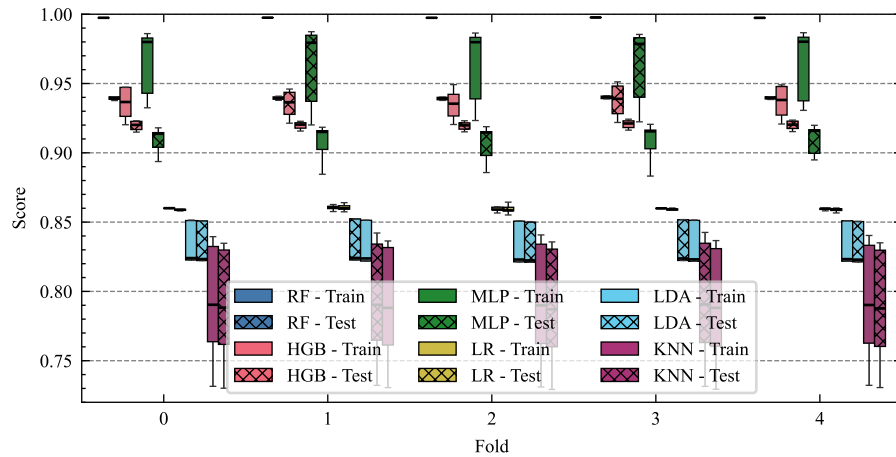


Figura 56: Resultados de la búsqueda en cuadrícula

Los resultados finales de esta optimización los encontramos en la [Tabla 20](#). El mejor modelo por mucho fue el *Bosque Aleatorio* ([Figura 57](#)), aunque también fue el que más tiempo requiere para su ajuste y evaluación. Uno pensaría que la implementación de *Potencialización del Gradiente mediante Histogramas*, fuese mejor, ya que involucra técnicas más avanzadas y aparece recurrentemente en la literatura como uno de los mejores algoritmos para datos tabulares.

Tabla 20: Mejores modelos y sus hiperparámetros

Estimator	Test score	Train score	Fit time	Score time	Best params	
RF	0.941 ± 0.002	0.998 ± 0.0	1169.288 ± 13.412	59.223 ± 3.25	Imputation	Constant
					Scaling	Standard
					Max feat	Sqrt
					Estimators	2000
HGB	0.924 ± 0.001	0.951 ± 0.004	43.815 ± 5.075	3.968 ± 0.308	Learning rate	0.1
					Max iter	1000
					Imputation	Median
					Scaling	Robust
MLP	0.921 ± 0.002	0.983 ± 0.002	631.375 ± 108.927	1.891 ± 0.072	Max iter	500
					Imputation	Median
					Scaling	Robust
					Layers	100,100,100
LR	0.871 ± 0.003	0.872 ± 0.002	429.85 ± 32.412	0.948 ± 0.165	Max iter	1000
					Imputation	Constant
					Scaling	Robust
					Solver	Newton-CG
LDA	0.852 ± 0.0	0.852 ± 0.001	5.208 ± 0.198	0.677 ± 0.087	Imputation	Constant
					Scaling	Robust
					Solver	SVD
KNN	0.837 ± 0.002	0.842 ± 0.0	1.282 ± 0.021	65.812 ± 0.132	Imputation	Constant
					Scaling	Standard
					Metric	Minkowski
					Neighbors	50

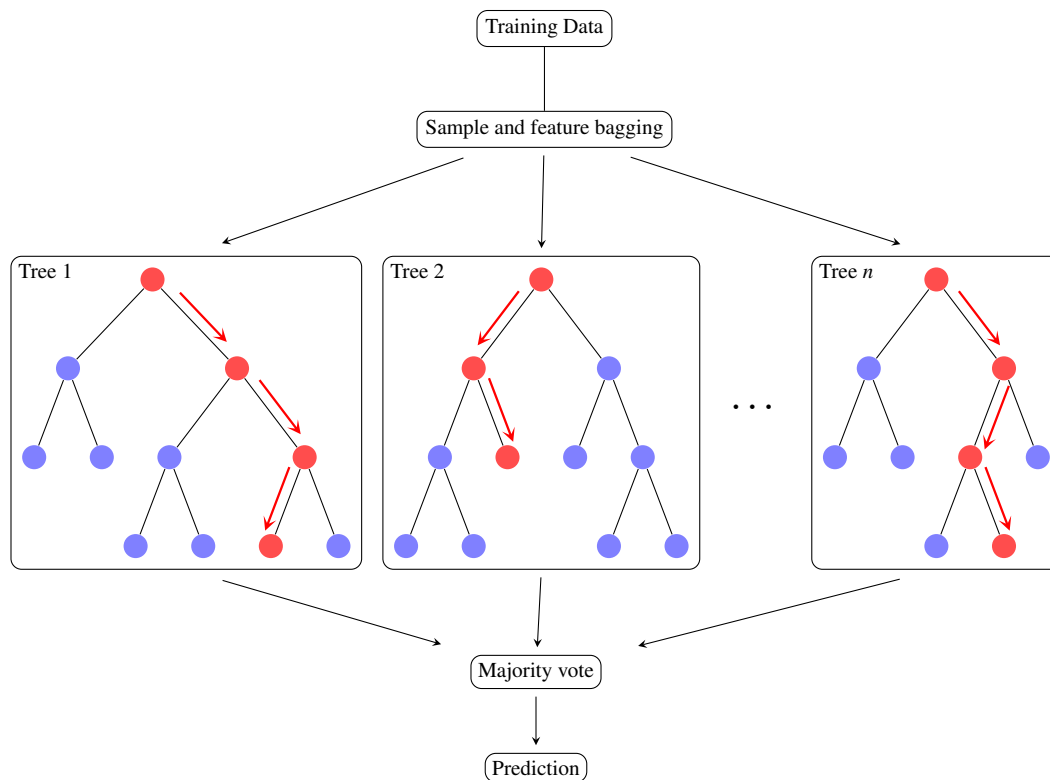


Figura 57: Representación de un bosque aleatorio

Para la búsqueda bayesiana de los mejores hiperparámetros para el *Bosque Aleatorio*, se aplicó el proceso durante 30 iteraciones. Como se puede observar en la [Figura 58](#), el proceso tuvo una convergencia consecutiva, lo cual es un buen indicador de que el proceso de optimización fue exitoso.

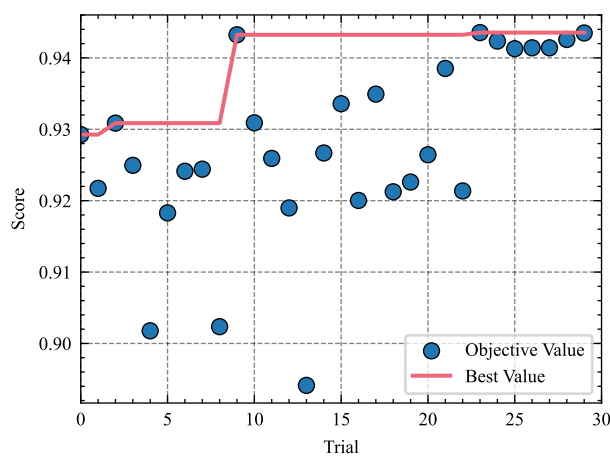


Figura 58: Proceso de optimización

Los hiperparámetros más importantes se pueden observar en la **Figura 59**. Siendo fundamental el peso de las clases; esto no es sorpresa debido a que estamos tratando con un problema binario desbalanceado. El siguiente más relevante se refiere al número mínimo de muestra para crear una nueva hoja en el árbol; naturalmente, al tratarse de muchos datos categóricos, se requieren muchas hojas para capturar el comportamiento de los datos.

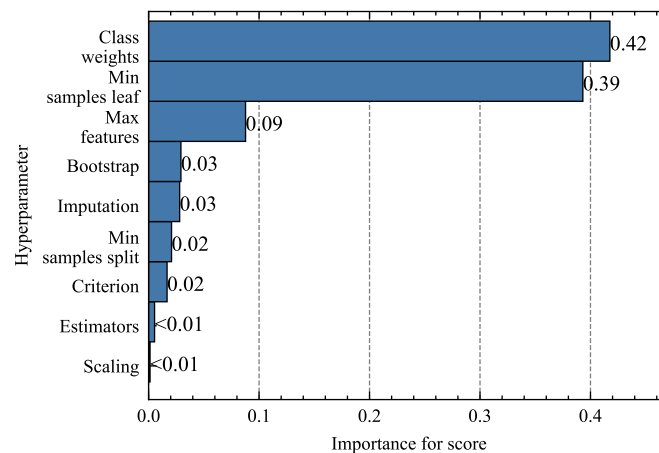


Figura 59: Importancia de los hiperparámetros

En la **Figura 60** podemos observar donde se concentran los hiperparámetros en relación con el rendimiento y la iteración. La búsqueda bayesiana permite reducir el espacio de búsqueda en cada iteración subsecuente, es por ello que se nota como los valores se van agrupando en ciertas regiones. La **Figura 61** muestra la interacción entre hiperparámetros numéricos y el rendimiento.

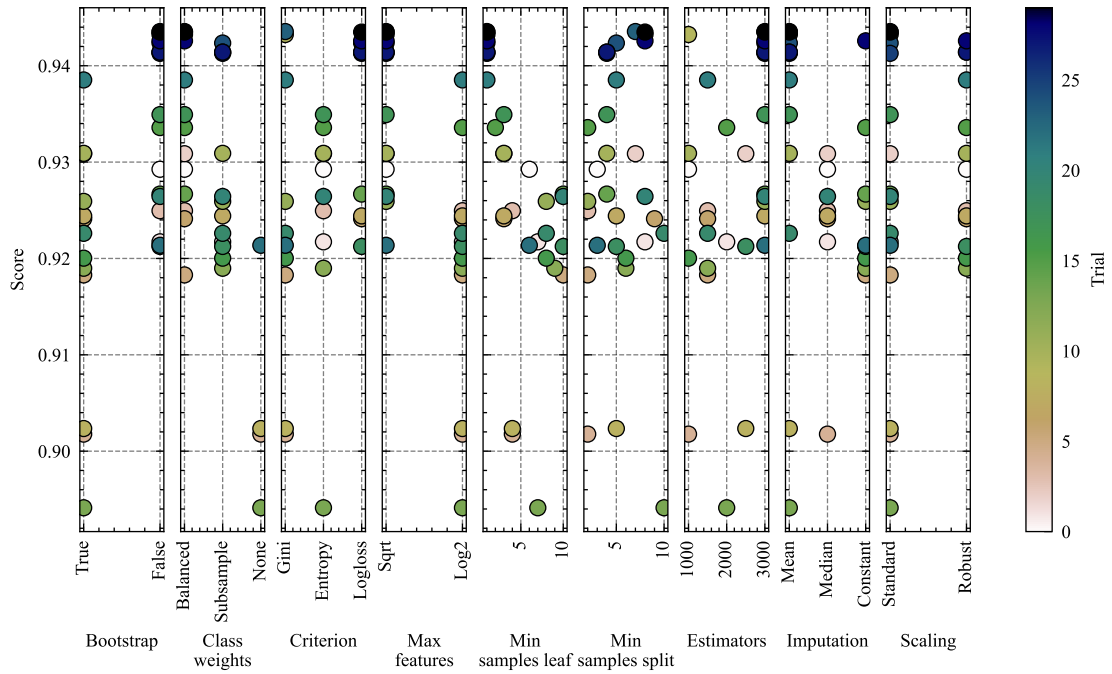


Figura 60: Rendimiento por hiperparámetro e iteración

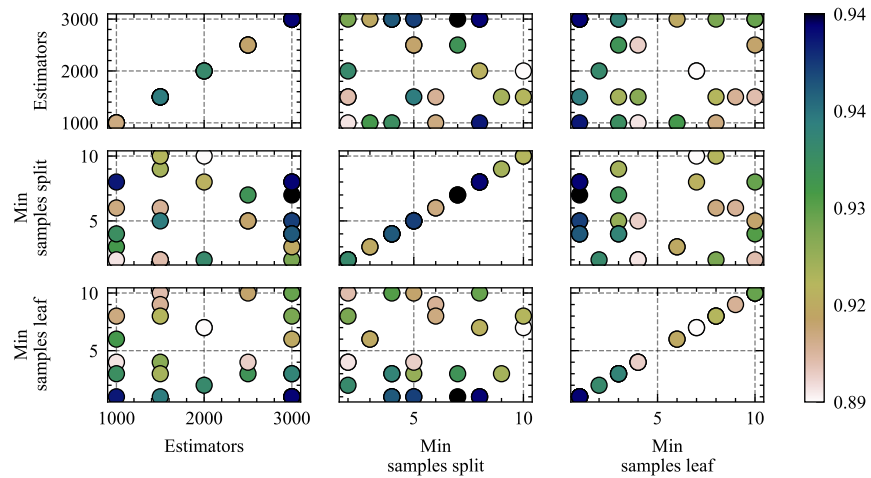


Figura 61: Rendimiento entre hiperparámetros

La [Figura 62](#) muestra la relación entre combinaciones de hiperparámetros con el rendimiento. Esto nos da un vistazo del espacio multidimensional de soluciones óptimas y pseudo óptimas en las que se desenvuelve el proceso de optimización para el *Bosque Aleatorio*.

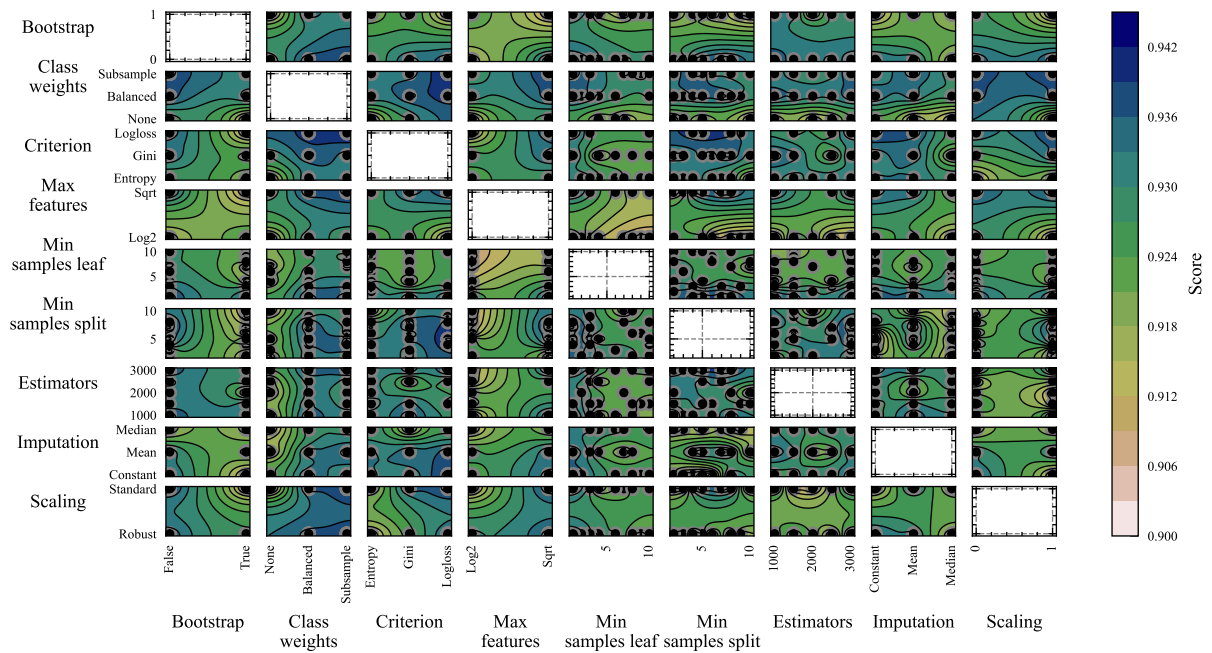


Figura 62: Mapa de contorno para la relación de hiperparámetros

El siguiente gráfico (Figura 63), muestra el flujo total de la combinación de todos los hiperparámetros, y los compara con otras combinaciones posibles. Con esto podemos trazar la influencia de cada uno de ellos en el rendimiento.

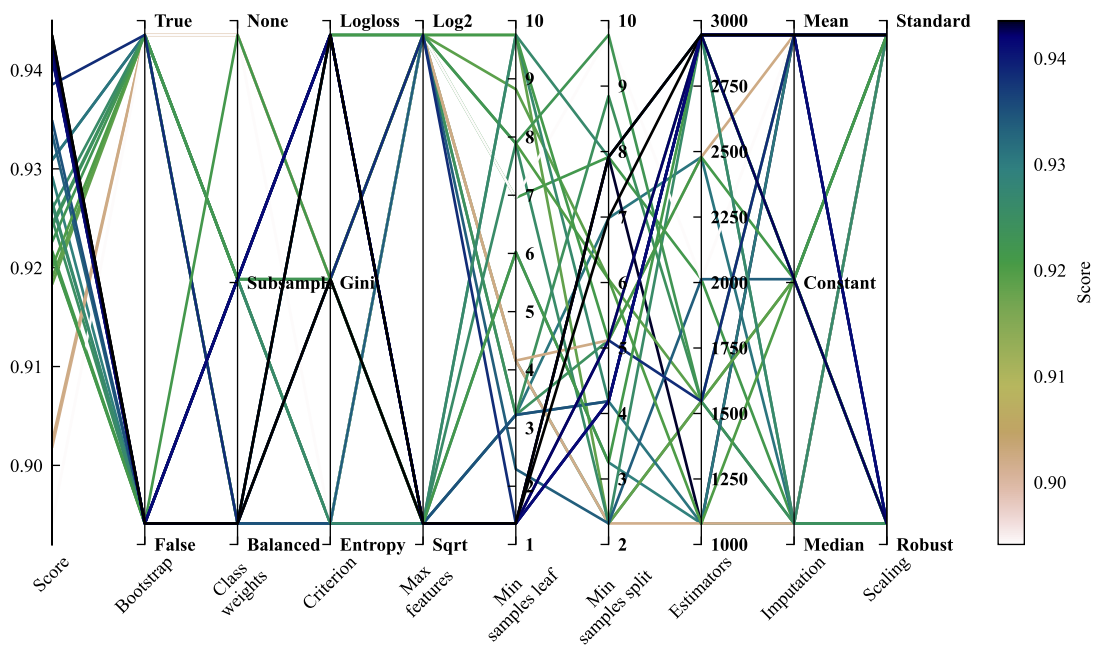


Figura 63: Gráfico de coordenadas paralelas

Finalmente, los mejores hiperparámetros para el algoritmo de Bosque Aleatorio los podemos ver en la [Tabla 21](#).

Tabla 21: Mejores hiperparámetros para Bosque Aleatorio

	Value
Estimators	3000
Criterion	Gini
Bootstrap	False
Scaling	Standard
Imputation	Mean
Max features	Sqrt
Min samples split	7
Min samples leaf	1
Class weights	Balanced
Score	0.940

Con respecto a las características generales de los árboles dentro del bosque, estas se pueden observar en la [Tabla 22](#). El modelo final es uno sumamente grande y complejo, capaz de capturar el comportamiento de los datos de manera precisa.

Tabla 22: Estadísticas de los árboles en el bosque

	Nodes	Depth
Mean	30395.617	70.382
Std	518.326	3.195
Min	28899	61
25 %	30023	68
50 %	30400	70
75 %	30757	72
Max	32229	85

4.4.2.15 Evaluación del modelo

La evaluación final del modelo se realizó después de entrenarlo con los parámetros optimizados con todos los datos de entrenamiento y posteriormente, aplicar métricas y gráficas para medir el rendimiento en el conjunto apartado en una de las etapas anteriores.

Las métricas de evaluación generales, se pueden observar en la [Tabla 23](#). El modelo tiene un rendimiento general bastante bueno. Las métricas que tienen que ser 1 presentan valores por arriba de 0.90, mientras que la última, que tiene que ser 0, presenta valores sumamente bajos. Esto nos da una idea del poder de generalización que tiene el modelo.

Con respecto a las métricas particulares ([Tabla 24](#)), desglosadas por clase y tipo de agrupación. Se observa que el modelo tiene mejor rendimiento para los casos de dengue Normal con respecto a los de dengue severo, esto puede ser debido al desbalance de datos entre ambas clases. Con valores consistentemente arriba del 0.95, el modelo presenta un rendimiento excepcional para el problema de clasificación binaria.

La curva de entrenamiento ([Figura 64](#)), permite entender como el modelo aprende a medida que se van presentando más datos para determinar si el modelo está sobreajustado o infraajustado a los datos de entrenamiento. En este caso, observamos que el modelo sigue aprendiendo conforme más datos vaya recibiendo, lo cual indica que el modelo no está sobreajustado y que puede generalizar bien a datos nuevos. Este fenómeno dio como pauta en la definición de las últimas dos etapas de la metodología anteriormente propuesta.

Tabla 23: Resultados generales de evaluación

BACC	ACC	Kappa	MCC	AUROC	AUPR	Brier
0.960	0.962	0.908	0.909	0.990	0.980	0.032

Tabla 24: Resultados particulares de evaluación

	F_1 score	G-mean	TPR	TNR	PPV	FPR	FNR
Normal	0.973	0.973	0.964	0.956	0.982	0.044	0.036
Severe	0.935	0.936	0.956	0.964	0.916	0.036	0.044
Macro	0.954	0.960	0.960	0.960	0.949	0.040	0.040
Micro	0.962	0.962	0.962	0.962	0.962	0.038	0.038

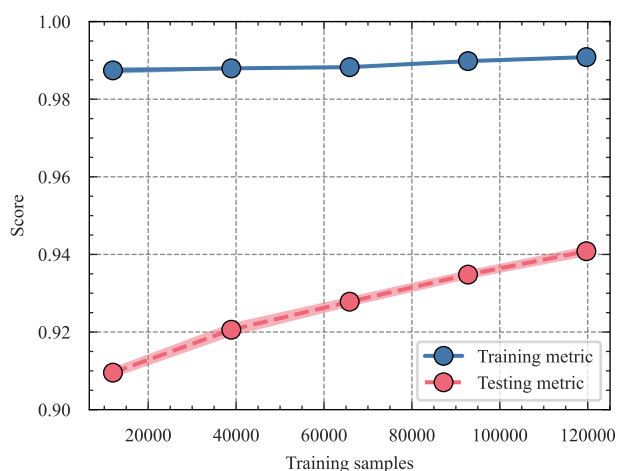
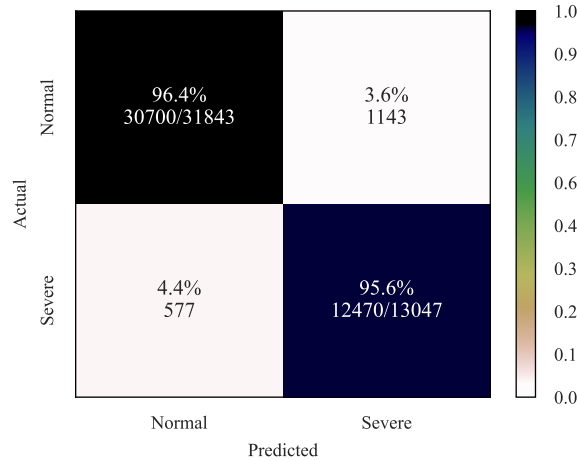
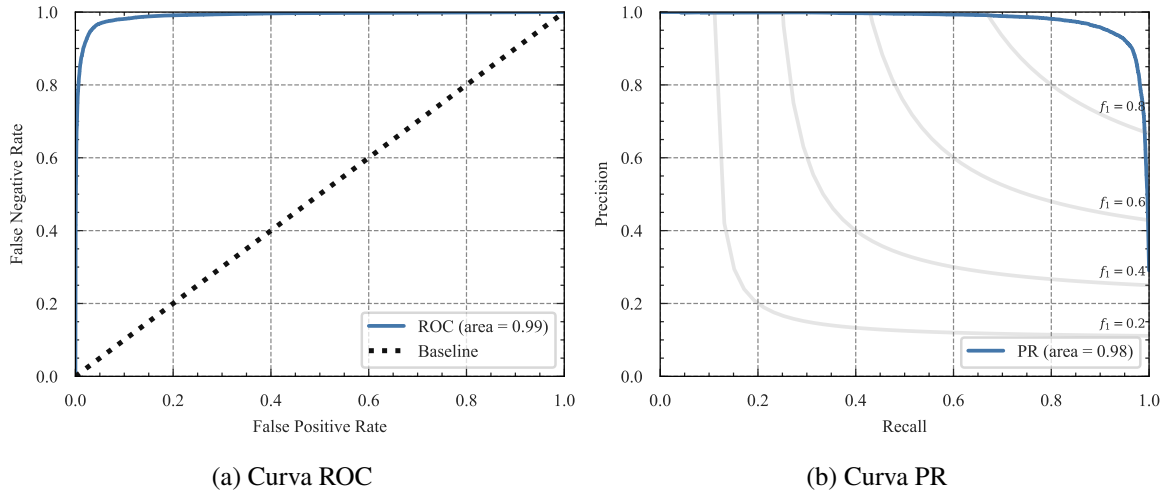


Figura 64: Curva de entrenamiento

Para la evaluación de modelos de clasificación, los gráficos comunes son la curva ROC, la curva PR y la matriz de confusión. Estos gráficos se pueden observar en la [Figura 65](#). La curva ROC presenta el rendimiento del modelo a distintos umbrales de clasificación, el área bajo esta curva debe ser lo más cercana a 1, lo cual indica que es un modelo robusto. La curva PR se comporta de manera similar, aunque es más robusta en casos donde existe un desbalance de clases. La matriz de confusión nos permite observar de manera clara los casos donde el modelo se equivoca y en que proporción. En el caso de diagnóstico médico, nos interesa que la tasa de Falsos Negativos sea bastante baja, ya que es más dañino confundir un caso de dengue grave por uno normal que viceversa. Afortunadamente, el modelo confunde solamente los casos severos en normales en 577 ocasiones, mientras que en el caso contrario, lo hace 1143 veces; este es el escenario que se busca en un modelo de clasificación médico.



(c) Matriz de confusión

Figura 65: Gráficos comunes para evaluación de un modelo de clasificación

El rendimiento del modelo es excepcional, lo cual aportará la capacidad de predicción necesaria para la toma de decisiones en la cadena logística de epidemiología.

4.4.2.16 Sesgos algorítmicos

La evaluación de sesgos algorítmicos se realizó utilizando dos características usadas para entrenar los modelos y una que se descartó en el proceso de Ingeniería de Características. Estas características son: sexo, edad e identidad indígena. Se evaluaron los sesgos en cada una de las características y en combinaciones de dos y tres características. Los resultados se pueden

observar en la [Tabla 25](#). El modelo no tiene ningún sesgo con la identidad indígena, pero sí se nota bastante un sesgo por edad. Al combinarse la edad con las otras dos características, el sesgo aumenta considerablemente; siendo las mujeres indígenas de entre 0 y 24 años las más afectadas ([Figura 66](#)).

Tabla 25: Métricas de equidad

	Demographic Parity		Equalized Odds	
	Difference	Ratio	Difference	Ratio
Indigenous	0.001	0.996	0.020	0.980
Sex	0.025	0.921	0.011	0.905
Ages	0.276	0.463	0.044	0.391
Indigenous,Sex	0.083	0.763	0.044	0.902
Indigenous,Ages	0.297	0.425	0.068	0.306
Sex,Ages	0.292	0.445	0.056	0.315
Indigenous,Sex,Ages	0.334	0.369	0.150	0.000

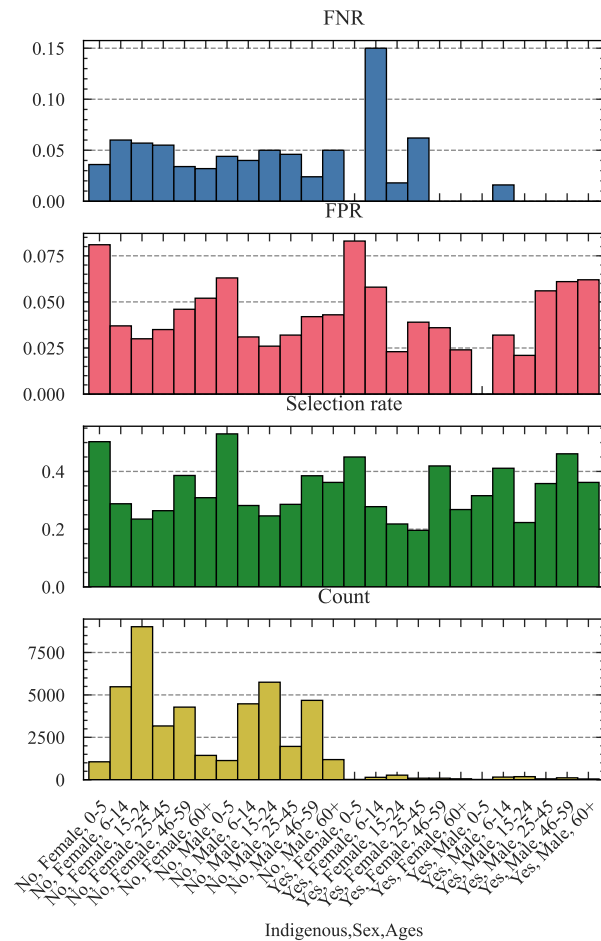


Figura 66: Equidad para la combinación de todas las características

Este fenómeno se puede deber a distintas causas, pero evidencia el poder de la metodología al ofrecer distintos niveles de análisis, desde el descriptivo al prescriptivo. Es obvio que existe un sesgo negativo en el modelo en el subconjunto antes mencionado. Por las características urbanas de la enfermedad y la poca cantidad de mujeres indígenas en esas condiciones, se demuestra que el sesgo es sumamente desproporcionado. Esto puede deberse a un sub muestreo de este estrato poblacional, a que tienen menor acceso a la atención médica ya sea por causas sociales o culturales, incluyendo manifestaciones discriminatorias o violencia de género. Evidencia de esto último se puede observar en la casi nula Tasa de Falsos Negativos para los varones indígenas a cualquier edad.

4.4.2.17 Explicabilidad

Entender como un modelo toma las decisiones y clasifica los datos de entrada es crucial en aplicaciones médicas. Afortunadamente, los Bosques de Decisión ofrecen una manera transparente para evaluar la importancia de las características en las predicciones finales. En la [Figura 67](#) se muestra la importancia de las mejores 10 características, sobre todos los árboles que componen el bosque de decisión. La cantidad de plaquetas, es la característica más importante para decidir si un caso de dengue es grave o no; seguida del dolor abdominal y el resto de las características que surgen de análisis clínicos. La importancia de esto es enorme, ya que estas columnas fueron imputadas y promediadas, por lo cual deducimos que el modelo es robusto a la incertidumbre y el ruido en los datos de entrada.

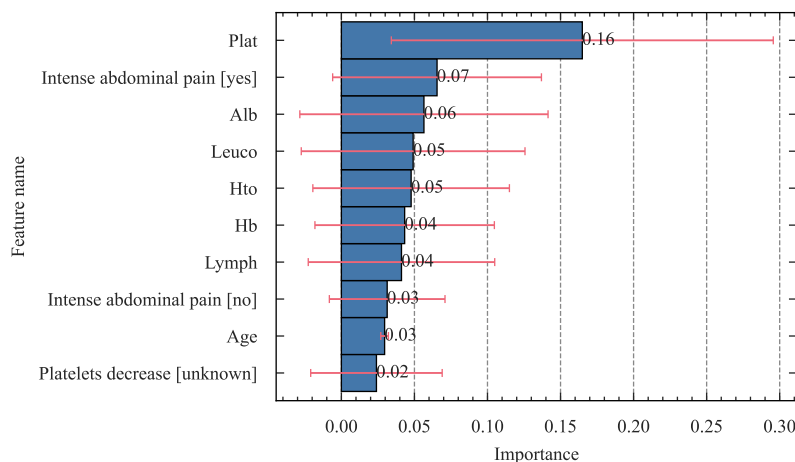


Figura 67: Importancia de las características en las predicciones

Una clasificación se puede definir como una función $f : \mathbb{R}^n \rightarrow \{0, 1\}$, donde n es el número de características. Para el caso binario, una probabilidad de 0.5 a 1 indicaría un hallazgo positivo, en este caso, la presencia de dengue grave. Analizar como se distribuyen estas probabilidades, nos da una idea de como toma el algoritmo sus decisiones. Al evaluar probabilísticamente los positivos con los valores predichos, generamos una curva de calibración ([Figura 68a](#)). Con respecto al histograma de las probabilidades ([Figura 68b](#)), es grato encontrar que estas se encuentran distribuidas en los extremos del intervalo, y no existe un pico en el centro que indique clasificaciones con poca confianza o certeza.

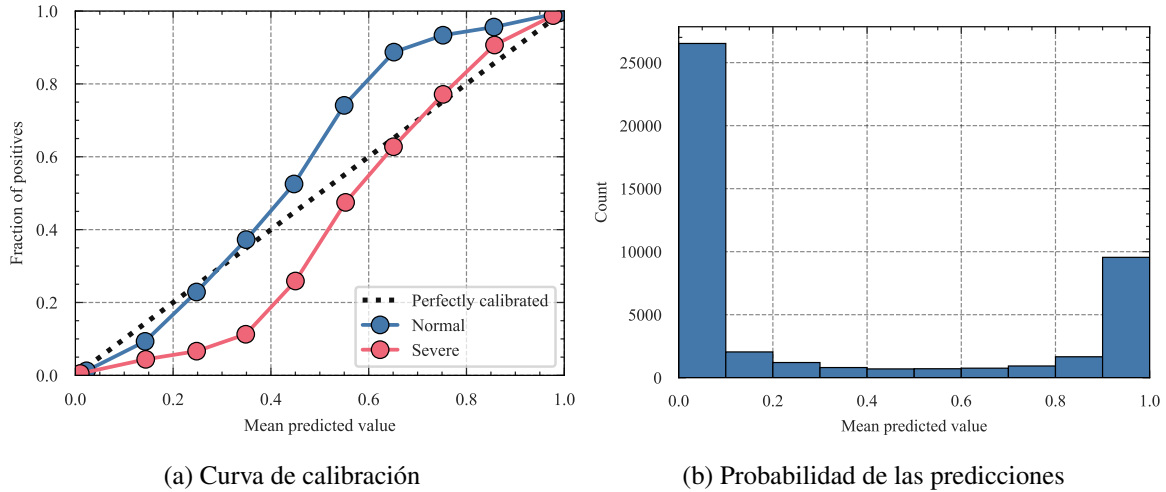


Figura 68: Evaluación probabilística

4.4.3 Discusión

Dentro del estado del arte, no se encontraron aplicaciones de análisis geoespacial relacionados con enfermedades (solo el trabajo de (Gergin et al., 2019), pero trata sobre instalaciones), aunque si estuvieron presentes algunos algoritmos de aprendizaje no supervisado (Du, 2018a) (Knaus & Marks, 2019a). Los pronósticos, como se ha analizado en semestres anteriores, tienen una presencia sumamente fuerte dentro del área médica (Rayan et al., 2019). Lo importante de la propuesta de este trabajo, es automatizar el proceso de análisis para que pueda ser realizado directamente por el experto médico, sin necesidad de tener conocimientos de programación o de análisis de datos.

En el estado del arte, solo el trabajo de (Pace et al., 2018) y colegas trata sobre análisis de series de tiempo. Es interesante su trabajo, ya que trabaja sobre un marco bien cimentado en la *Industria 4.0*. Ningún trabajo analizado se presenta como una solución integral que trate de capitalizar conocimiento dentro de la cadena logística de epidemiología. Otra diferencia radica en la aplicación de diferentes resoluciones a un mismo problema, es decir, analizar los mismos datos bajo distintas ventanas temporales y espaciales.

Con respecto al *Aprendizaje Automático*, tenemos que los trabajos (Rayan et al., 2019), (Mujumdar & Vaidehi, 2019a) y (Tanwar et al., 2020a); utilizan uno o más algoritmos para diagnosticar o clasificar enfermedades. Siendo (Mujumdar & Vaidehi, 2019a) aquel que tam-

bién aplica un conjunto de algoritmos y compara sus rendimientos para poder determinar cuál es el mejor. Las diferencias fundamentales radican no solo en los algoritmos escogidos, sino que trabajan con una cantidad de datos muchísimo menor (800), con muchas menos características (8), con datos balanceados, heterogéneos y son ningún faltante. Otra diferencia importante de los datos es que los usados en este trabajo fueron curados y con el objetivo principal de ser utilizados para la creación de modelos de clasificación; mientras que los datos utilizados en este trabajo son datos que surgen orgánicamente dentro de la cadena logística de salud pública, sin intermediarios, y se asume que representan objetivamente las distribuciones de probabilidad de las características de la población mexicana en general. Obvio los datos no están libres de sesgos, ya que es difícil asegurar la integridad de los mismos, por el hecho de que provienen de distintas fuentes, sin embargo, el personal médico da fe de que estos se comportan de manera real.

Otros trabajos que utilizan algún modelo de *Aprendizaje Automático* son el trabajo de (Santos et al., 2019), (H. Lin et al., 2019a), (Khaldi et al., 2019b). Sin embargo, sus objetivos no se centran en la etapa diagnóstica de la cadena logística de salud pública, (Santos et al., 2019) involucra la aplicación de algoritmos de *Inteligencia Artificial* a la salud, se enfoca principalmente en telemedicina, concepto que, a principio de este trabajo doctoral, se descartó por no estar en sintonía con el acceso tecnológico presente en la población mexicana; enfocarse en estos conceptos de industria 4.0, no solo es privativo, sino también discrimina a las poblaciones más vulnerables.

Khaldi et al. (Khaldi et al., 2019b) utilizan modelos para predecir las visitas de pacientes a emergencias. Este es un análisis de series de tiempo similar al que se pretende hacer con los datos recabados de SINAVE, aunque con un enfoque distinto. Sobre todo en la característica de la base de datos como se expuso anteriormente, pero también en los algoritmos que se utilizarán.

Esto evidencia la congruencia del trabajo realizado con respecto a lo encontrado en el estado del arte. Así mismo, también muestra las diferencias del mismo, por ejemplo, si bien (H. Lin et al., 2019a), (Thesmar et al., 2019a), (C.-H. Tseng et al., 2019) utilizan *Big Data* como componente de *Industria 4.0*, no lo hacen transparente, ya que no mencionan la cantidad de

datos ni su peso. Mientras que los trabajos que utilizan sistemas de información (Hathaliya et al., 2020a) no incorporan este componente en su desarrollo.

4.5 Arquitectura y diseño del sistema

En las secciones anteriores se describieron procesos que incorporaron varias metodologías al unísono, una aplicación eficiente de tecnologías y formas de investigación basada en código, así como la gestión de distintas disciplinas y fuentes de datos. Todo ello con el objetivo de modelar correctamente el flujo de datos dentro de la cadena logística de salud pública.

Para derivar la arquitectura y diseñar el sistema, se utilizó como base el caso de estudio de la *Fiebre del Dengue*; al tener una complejidad superior, mayor cantidad de partes y la densidad de la interacción entre las mismas, es más factible que funja como el origen de la programación de un Sistema de Híbrido Inteligente.

El propósito de este proceso, fue encontrar la estructura general de los datos para generar sus esquemas y construir arquitecturas computacionales, como las *Bases de Datos* y *Almacenes de Datos*. En primera instancia, nos permitió, de la manera más básica, generar las tablas que se incorporarán dentro de la base de datos, así como la descripción de los atributos necesarios; también, nos permitió conocer, de manera más abstracta, como se comportan los datos para determinar el tipo de arquitectura, tecnologías y paradigmas a aplicar para su correcta explotación.

La creación de las bases de datos tiene como uno de sus objetivos, proveer de datos a la siguiente capa, el *Almacén de Datos*, que convertirá los datos en información importante para los tomadores de decisiones. La creación de un *Almacén de Datos* tiene el propósito de culminar en la instauración de la salud basada en datos, concepto que se ha extendido en organizaciones de salud privada, donde se han utilizado desde para descubrir nuevos medicamentos hasta el cálculo general de costos de salud para aseguradoras y otras entidades corporativas. El *Almacén de Datos* representa el componente de *Big Data* de la **Industria 4.0**.

Algunas de las preguntas respondidas por el proceso anterior, están relacionadas con propiedades de los datos que surgen de la cadena logística así como de los datos externos a incorporar.

Al tener gran variación en su estructura (o carencia de ella) en los datos, sugieren la necesidad de un enfoque complejo que involucre distintas tecnologías. Presencia de datos geoespaciales, como el Marco Geoestadístico Nacional, series de tiempo como los datos climatológicos o los elementos no estructurados de los reportes epidemiológicos, fuerza al uso de tecnologías especializadas o de paga. Los nuevos Motores Gestores de Base de Datos, de código abierto, son capaces de gestionar eficientemente datos tabulares y son mejorados con extensiones para poder operar con datos geoespaciales y series de tiempo, también cuentan con formas computacionalmente viables para el uso de datos no estructurados, en su mayoría representados en formato *JSON*.

El **Sistema Inteligente**, es una plataforma web que conjuga los Sistemas de Información en Salud y los Sistemas de Apoyo a la Decisión con componentes de *Industria 4.0*. Estos componentes pueden ser para técnicos como el uso de algoritmos avanzados o tecnológicos como el cómputo en la nube, el *Big Data* o tecnologías móviles o abstractos, como la ciberseguridad y la usabilidad.

Su desarrollo está regido desde el principio para maximizar su operatividad, en concordancia con la *Industria 4.0* y está basado en los más modernos paradigmas en Ingeniería de Software, como lo es *DevOps* y desarrollo Ágil. En la actualidad, programas de este tipo se dividen en dos partes: Frontend y Backend.

El **Sistema Inteligente** debe de cumplir con características como ciberseguridad y encriptación, para proteger los datos de los pacientes y personal médico, pero también debe de permitir acceder a tal información a los algoritmos pero también a otros médicos o investigadores. El sistema fue desarrollado en su totalidad con herramientas *Open Source* y de licencia permissiva, también su código será abierto y almacenado en un repositorio público⁷. Se propone el uso del lenguaje Python, el cual es el lenguaje más usado en los campos de *Inteligencia Artificial* y, al ser un lenguaje de propósito general, también sirve para la creación de aplicaciones web. Conceptos como *DevOps* ayudarán a facilitar la implementación y despliegue de la plataforma web en la nube.

⁷https://github.com/marcodeimoral/sis_prototipo

El diagrama del **Sistema Inteligente** sugerido se observa en la **Figura 69**. El sistema en su capa exterior es una aplicación web que reside en la nube y puede ser accedida desde cualquier parte, esto incluye la operación del sistema como tal, así como el acceso y la ingesta de información. El subsistema de Toma de Decisiones gobierna, mediante la interfaz, predicciones e inferencias, el flujo de información y conocimiento dentro del sistema, cabe destacar que el sistema no es modular, sino que integra holística y dinámicamente distintos componentes, que sirven para propósitos distintos.

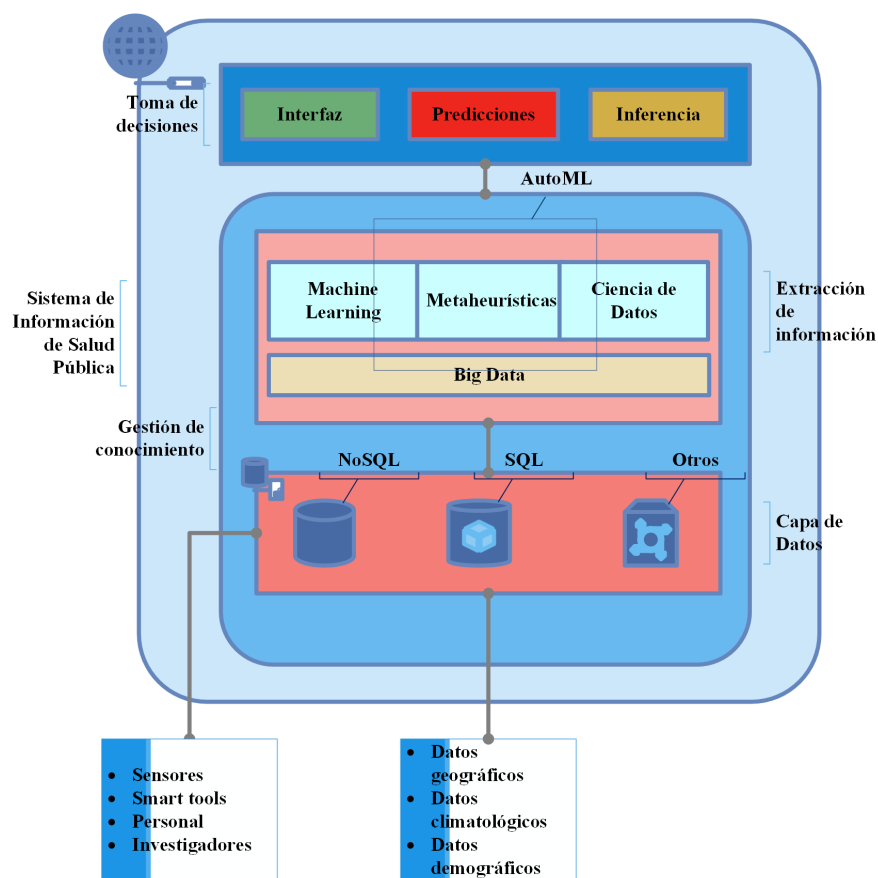


Figura 69: Diagrama conceptual del sistema

La gestión de conocimiento, que incluye el almacenamiento de datos y su transformación en información, cuenta con comunicación bidireccional con sensores, *Smart Tools* e individuos como el personal médico, pacientes o investigadores. Cabe recalcar la ingesta continua de datos geográficos, climatológicos y demográficos para enriquecer la Toma de Decisiones.

La Capa de Datos abstrae todo el proceso de *ETL*⁸ de los datos. Incluye una o varias bases de datos relacionales, no relacionales de esta forma como repositorios de datos como *Almacenes de Datos* o *Lagos de Datos*.

La Extracción de Información, se refiere a utilizar métodos matemáticos y algorítmicos para refinar datos en información. Convertir datos en información requiere aplicarles un preprocesamiento para ser transformados por algoritmos de IA. El objetivo es que estos componentes actúen en conjunto, un ejemplo de ello es *AutoML*, en el cual se entrenan algoritmos de *Aprendizaje Automático* de manera automática. *Big Data* gestiona la comunicación con la Capa de Datos para poder tratar gran cantidad de ellos.

Por último, la elección de tecnologías libres y modernas para el desarrollo del Sistema, nos ofrece una forma de trabajo sumamente poderosa, en la cual se operan ambos conceptos: el análisis de flujo de datos con la programación del sistema en paralelo, reutilizando código de un proceso en el otro y agilizar el trabajo en este ambicioso proyecto.

4.5.1 Base de datos

PostgreSQL es un Sistema de Gestión de Bases de Datos relacional orientado a objetos y de código abierto con gran capacidad de extensibilidad. Tiene capacidad de manejar datos geoespaciales con la extensión *PostGIS*, trae incluido el soporte para datos no estructurados en *JSON* y *Hstore*, optimizado para series de tiempo usando la extensión *Timescale*, es concurrente y puede ser distribuido.

El *backend* es la parte más complicada tanto en diseño como programación. Corresponde a todos los elementos que se tienen que orquestar correctamente para que un usuario pueda interactuar con el sistema propiamente. Esto incluye la aplicación web propiamente dicha, las bases de datos, scripts para ingesta de servicios externos y demás componentes de cómputo y administración.

En este diseño, se decidió unificar las bases de datos geoespaciales y tabulares en una sola gracias a las capacidades el motor gestor de base de datos elegido para la programación

⁸Extraer, Transformar y Adquirir; en inglés, Extract, Transform, and Load

del sistema. Este motor también es capaz de fungir como *Almacen de Datos* agregando la información que contiene en cubos de datos preprocesados listos para ser transmitidos a los procesos de análisis.

4.5.2 Frontend

Es la parte con la cual interactúa el usuario, esta se compone del sitio web propiamente dicho, pero también presenta funciones como la exposición de una API para usuarios externos, con el objetivo de, por ejemplo, la generación de reportes automatizados, así como canales de notificación externos que pueden ser mensajes de texto o WhatsApp para la propagación de eventos o noticias dentro del sistema. Distintos canales de interacción con el usuario dependen de distintas tecnologías, todas ellas relacionadas con el internet. De las tecnologías planeadas para el despliegue se tienen en mente el uso de la web tradicional así como frameworks capaces de unificar desarrollos multiplataforma, esto constituye la interfaz gráfica del usuario. Por otro lado, las interfaces no gráficas se representan en API y servicios web, que en este momento se encuentran implementados siguiendo los estándares web *RESTful* así como el esquema *OpenAPI* para la autodocumentación de tales servicios.

Son tres lenguajes los que dominan el desarrollo de *Frontend*: *HTML*, *CSS* y *JavaScript*.

Otro concepto importante para el desarrollo del Frontend, es la existencia de frameworks para la creación de aplicaciones web. Los más populares son *React*, *Angular* y *Vue*. Algunos de estos tienen capacidad de exportar lo programado para la web a otras plataformas, lo cual es bastante útil para desarrollar software de manera rápida y funcional, en concordancia con la *Industria 4.0*.

4.5.3 Backend

La aplicación web propiamente dicha fue programada usando el Framework web *Django*, programado en Python, permite una integración directa del código usado para el modelado de flujo de datos. Django es un Framework MTV: Model, Template, View. El Modelo es una Clase cuyos objetos representan filas individuales de la base de datos, se manifiesta gracias al *Mapeador Objeto Relacional* (ORM) y permite la consulta de los datos a un alto nivel. El

Template es la página web en HTML que interactúa con el usuario y la Vista controla el flujo de datos.

Una aplicación de Django, se divide en apps, que encapsulan funcionalidad, facilitan su gestión y modularizan el proyecto. En la **Figura 70** podemos observar el diagrama de la arquitectura de la aplicación web del Sistema de Información usando Django. A la derecha se tiene la base de datos que, mediante el ORM, se comunica con el control (Vistas), un paso de serialización convierte los datos en formatos legibles, el gestor de URL despacha las peticiones a las vistas y están regresan información mediante los Templates o la API. Los formularios permiten el ingreso de información al sistema. Arriba, están representadas las aplicaciones que se tienen diseñadas y en proceso de implementación, Core se refiere al sitio principal, gestión de usuarios y administración.

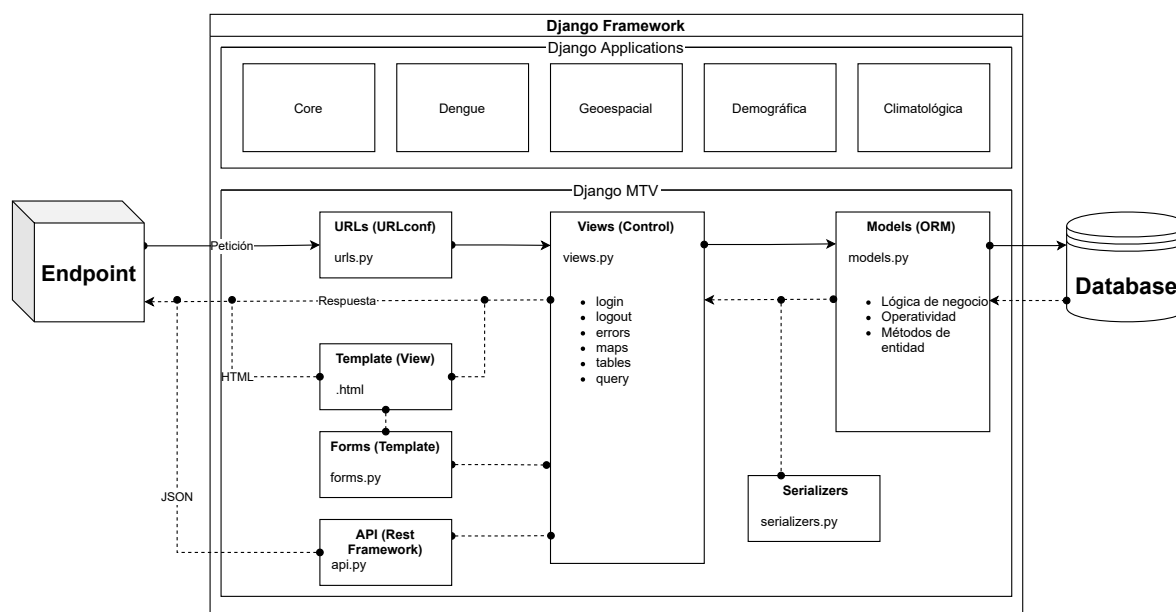


Figura 70: Arquitectura de la plataforma web para el Sistema Inteligente

4.5.4 Prototipo

Todas las etapas previas, metodologías, tipos de análisis, procesos de investigación y demás, se realizaron con el objetivo de crear un prototipo funcional del Sistema Inteligente propuesto. El prototipo es una aplicación web que permite realizar todas las premisas anteriormente expuestas.

Un componente sumamente importante en cualquier plataforma web, sobre todo en aquellas que van a ser usadas por muchos usuarios distintos, es la administración completa de la plataforma a través de un sitio especializado para tal tarea. Esto para facilitar el mantenimiento y operatividad de la misma. La **Figura 71** muestra la interfaz de administración de Django, la cual es bastante completa y permite la gestión de usuarios, permisos, grupos, aplicaciones, modelos, etcétera. La interfaz es bastante intuitiva y permite la gestión de la aplicación web de manera sencilla.

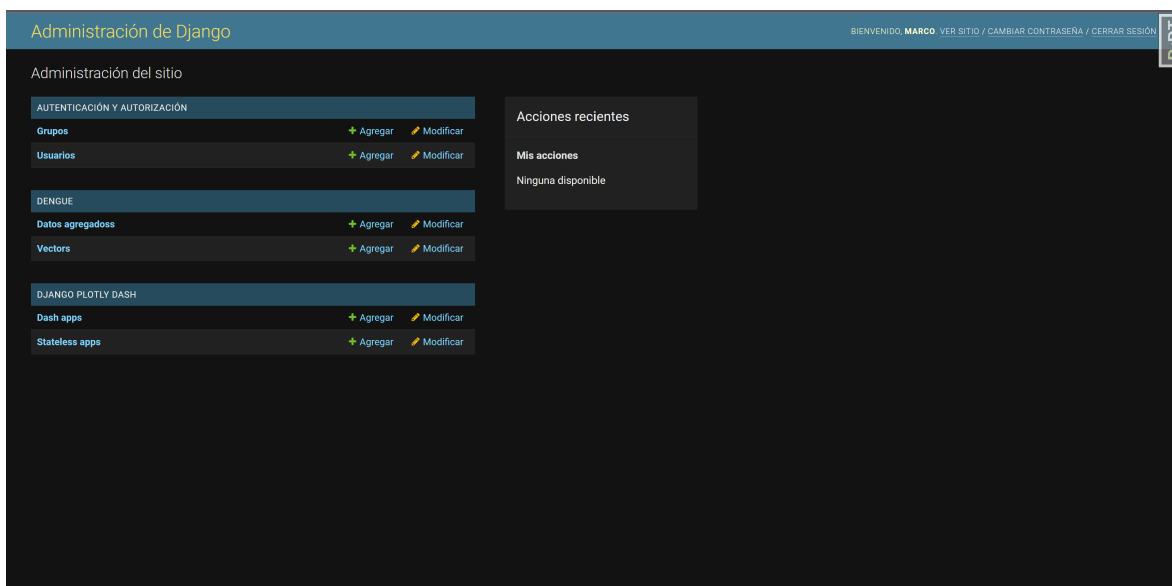


Figura 71: Arquitectura MTV de Django

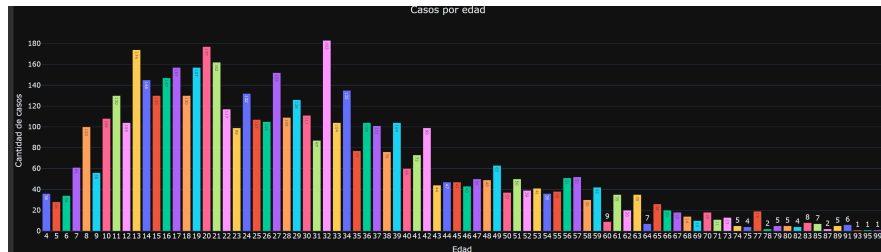
Las funcionalidades básicas del prototipo se basan en replicar los análisis realizados de manera estática y extrapolarlos a absolutamente todos los datos que existen dentro de la plataforma SINAVE, tanto pasados, presentes y futuros. Permitiendo a los tomadores de decisiones, desde ahora usuarios, realizar distintos análisis dinámicos y obtener sabiduría usando la plataforma web.

Los gráficos realizados por el sistema, así como las tablas dinámicas que provee la interfaz gráfica, constituyen un tablero de datos o *Dashboard* que maximiza la usabilidad del Sistema Inteligente. Incorporando componentes dinámicos como *tooltips* para mostrar toda la información que requiera el usuario.

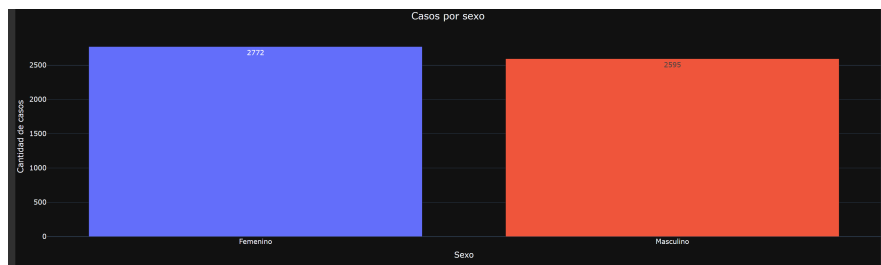
Algunos de los análisis replicados se pueden observar en la [Figura 72](#). Por ejemplo, se pueden realizar gráficos de barras para el diagnóstico final ([Figura 72a](#)), casos por edad ([Figura 72b](#)), por sexo ([Figura 72c](#)), por ocupación ([Figura 72d](#)) y por división geográfica ([Figura 72e](#)).



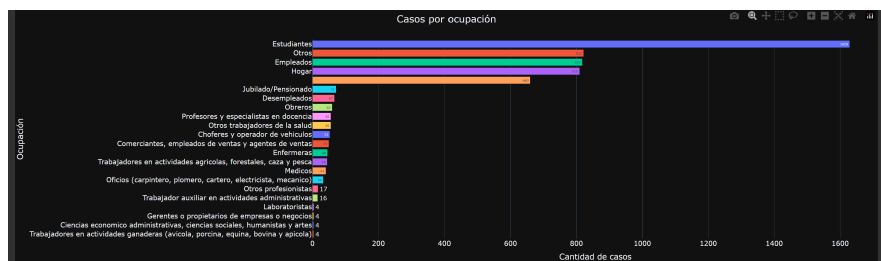
(a) Casos por diagnóstico



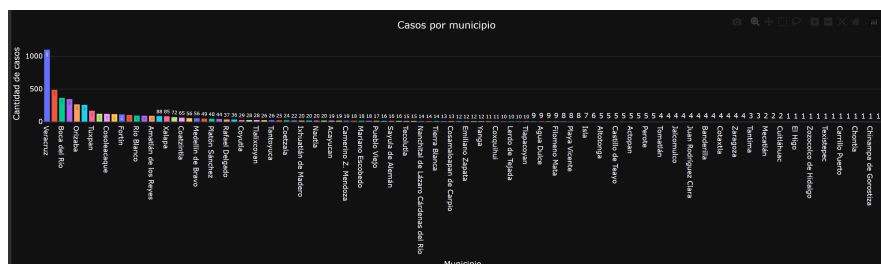
(b) Casos por edad



(c) Casos por sexo



(d) Casos por ocupación



(e) Casos por división geográfica

Figura 72: Análisis replicados en el prototipo

La plataforma ofrece gráficos dinámicos que permiten una visualización más profunda (Figura 73). Por ejemplo, apreciar a gran detalle la estratificación por características como sexo, diagnóstico y ocupación, se utiliza un gráfico *sunburst* interactivo.

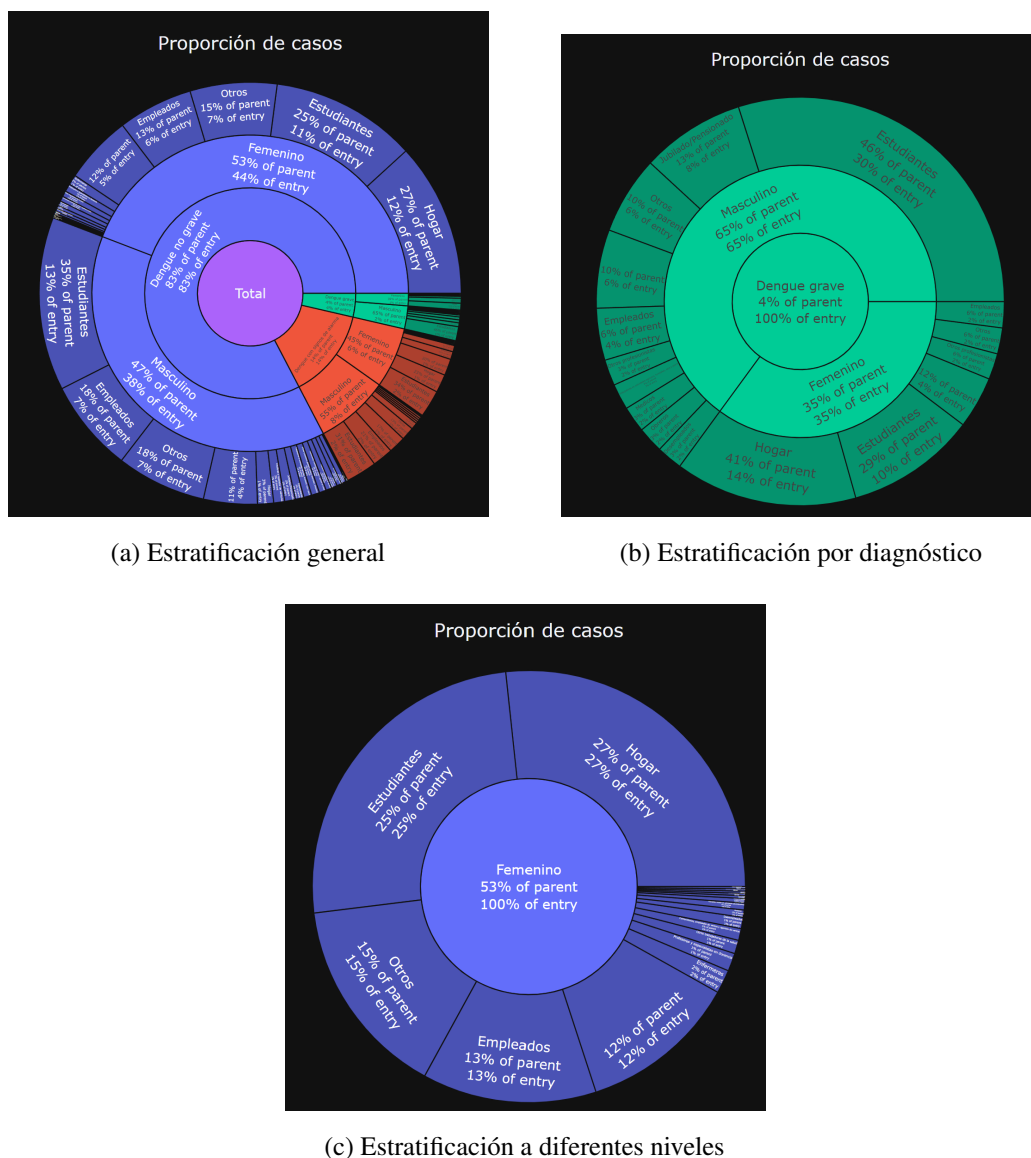
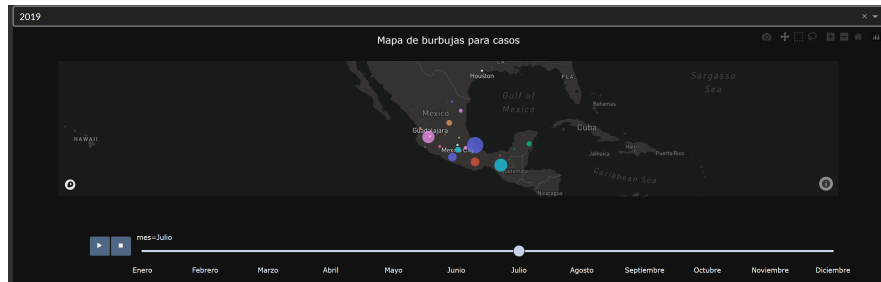
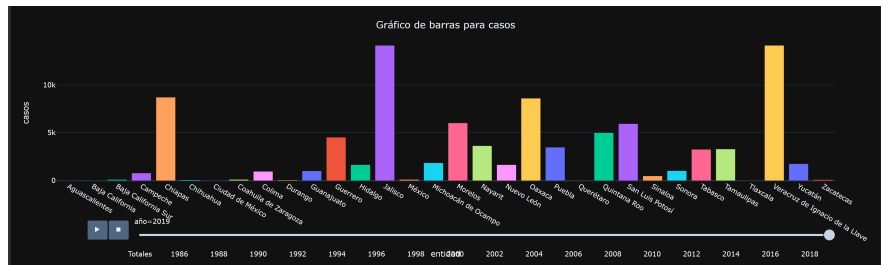


Figura 73: Ejemplos de gráficos dinámicos

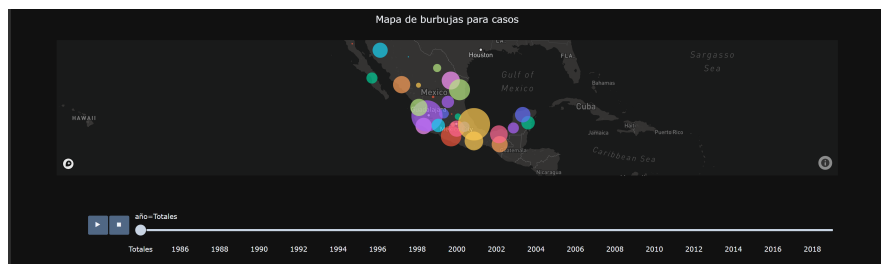
Otros tipos de análisis dinámicos pueden tener movimiento, para modelar y visualizar correctamente la evolución de los datos epidemiológicos. No solo con gráficos tradicionales como los gráficos de barras, sino también con visualización geoespacial y su transformación a través del tiempo (Figura 74).



(a) Mapa de burbujas mensual



(b) Casos por entidad a través del tiempo



(c) Mapa de burbujas anual

Figura 74: Otros análisis posibles

Unas de las funcionalidades más importantes de la plataforma, es poder realizar análisis geoespacial. Este proceso involucra geocodificar automáticamente cada caso que se almacena dentro de la plataforma, asegurando su integridad y utilidad, como por ejemplo, aplicando la proyección correcta a las coordenadas. En la [Figura 75](#), se observa la interfaz que permite mapear los puntos de casos de dengue, filtrarlos por fecha y ubicación, así como mostrar casos por diagnóstico.



Figura 75: Mapeo de casos de dengue

De las funciones implementadas más importantes, está la de detección de brotes mediante el algoritmo *DBSCAN*. Esta función aplica automáticamente el algoritmo para poder detectar brotes en función a la distancia de los casos, colorea los puntos en el mapa dependiendo si pertenecen a un grupo o no y muestra en el *tooltip* información como el nombre del paciente, la dirección, etcétera. Esta funcionalidad provee al Sistema Inteligente la capacidad de dar soluciones a la cadena logística; el usuario puede entonces tomar la decisión informada para poder aplicar una solución para evitar la transmisión de la enfermedad, por ejemplo, envío de personal de salud para fumigar. La **Figura 76** muestra la detección de brotes en el mapa. Esto es un Análisis Prescriptivo explícito.

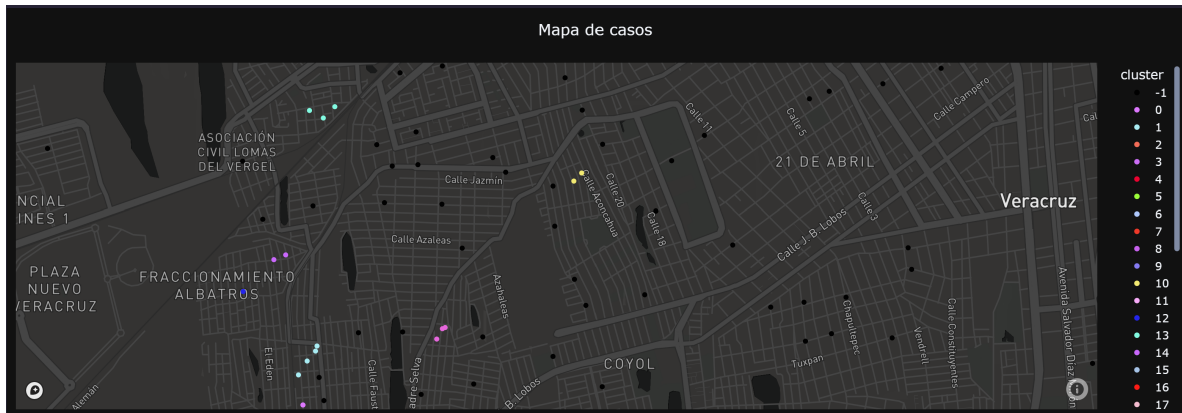
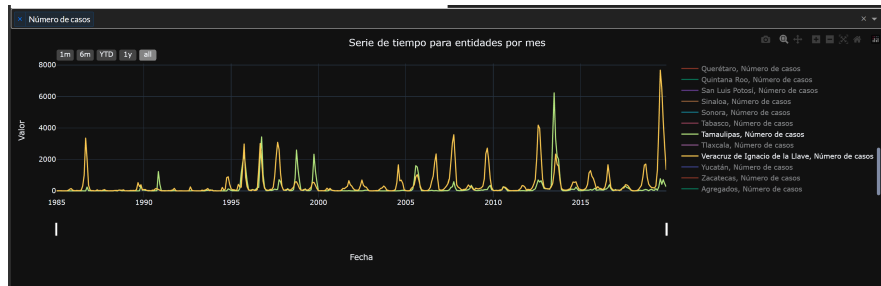
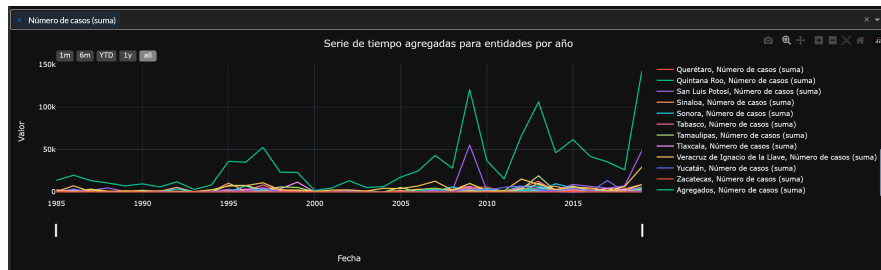


Figura 76: Arquitectura MTV de Django

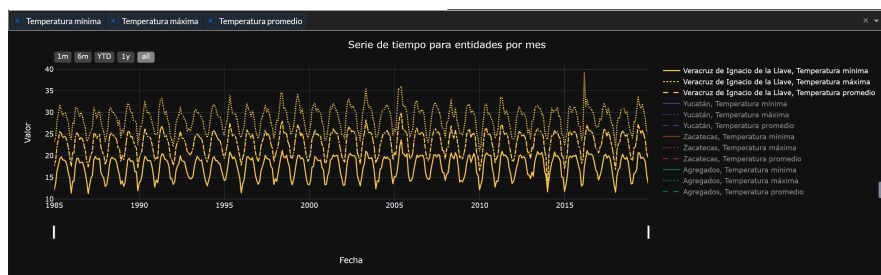
La interfaz para la visualización de series de tiempo [Figura 77](#) permite al usuario elegir entre distintas variables como el número de casos y variables climatológicas. También puede agrupar los datos de manera dinámica para observar su comportamiento en diferentes horizontes temporales. Podemos, en la misma gráfica, mostrar una o todas las variables, para encontrar patrones de comportamiento. La interfaz implementa controles de visualización dinámicos e intuitivos para maximizar la usabilidad, permitiendo a usuarios no expertos en replicar los análisis realizados pasivamente en la fase de análisis de flujo de datos.



(a) Casos por entidad



(b) Casos por entidad anual



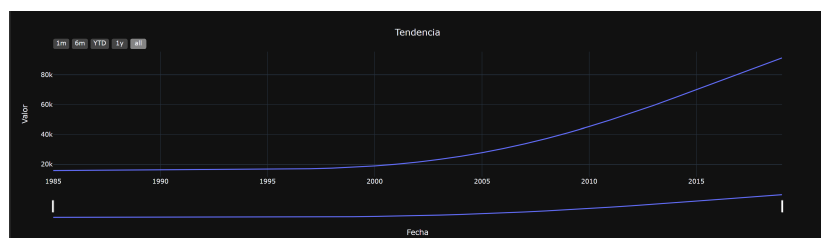
(c) Variables climatológicas

Figura 77: Visualización de series de tiempo

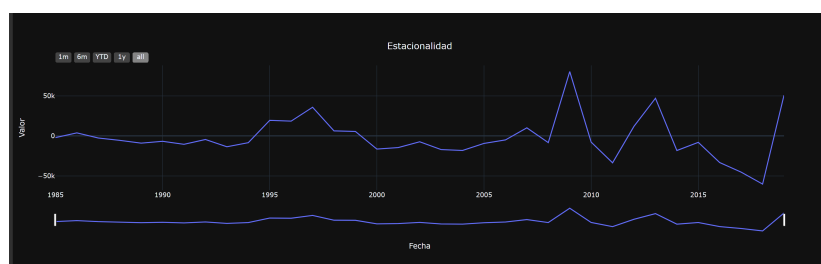
Siguiendo la línea de aplicar distintos tipos de análisis para incrementar el valor, la complejidad para generar conocimiento y sabiduría, se implementó la descomposición de series de tiempo en el prototipo para comprender mejor como se comporta la enfermedad a través del tiempo. Los datos utilizados para esta implementación provienen de datos pre agregados como se vio anteriormente, sin embargo, en la funcionalidad final, provendrán del DWH que automáticamente agrupará los datos en distintos horizontes temporales y espaciales (Figura 78).



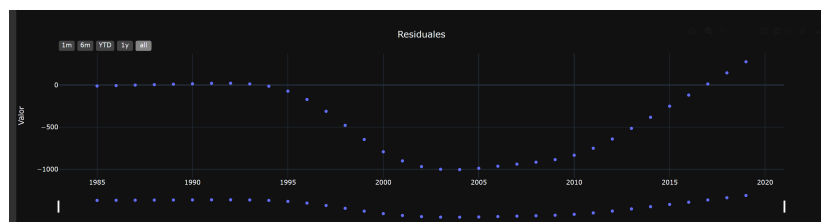
(a) Número de casos por año



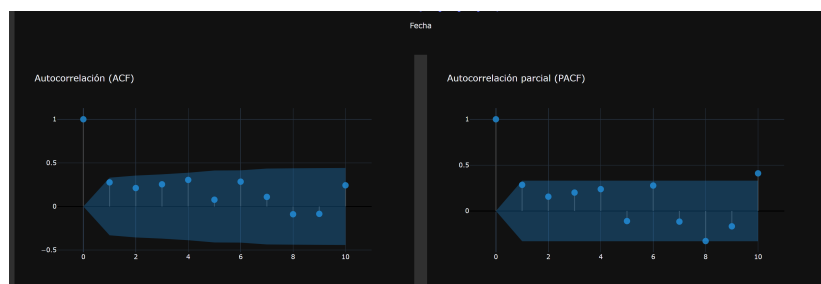
(b) Tendencia



(c) Estacionalidad



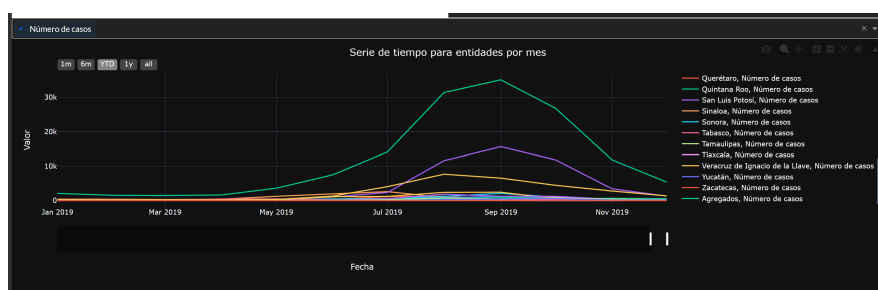
(d) Residuales



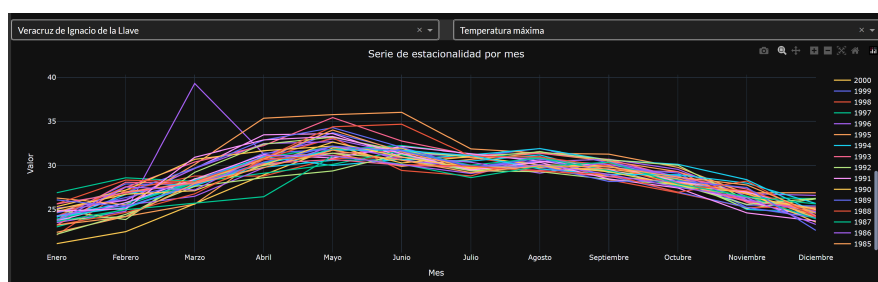
(e) Autocorrelaciones

Figura 78: Descomposición de series de tiempo

En la **Figura 79** se muestran algunos análisis adicionales de series de tiempo que puede realizar la plataforma. Descubriendo comportamiento haciendo comparaciones entre distintas variables. Las series de tiempo no solamente se pueden visualizar utilizando gráficos lineales, sino también estadísticamente al aplicar gráficos de caja para conocer la variabilidad de la enfermedad cuando los datos globales se agrupan por mes (**Figura 79c**).



(a) Comparativa de casos por entidad y mes



(b) Comparativa de casos por año y mes



(c) Estacionalidad global por mes

Figura 79: Gráficos adicionales relevantes

4.5.5 Discusión

Distintos trabajos en el estado del arte tratan sobre Sistemas de Información en Salud. Dos Santos y sus colegas (Santos et al., 2019), proponen un protocolo para medir el grado de

aceptación de la Inteligencia Artificial dentro del área de salud, en concreto, con la telemedicina. Una propuesta interesante es la que hace (Du, 2018a), para el uso de los sistemas de información para coordinar centros especializados en emergencias; este trabajo está definido exclusivamente a nivel regional y solo para la gestión de emergencias, nuestra propuesta pretende atacar distintas cadenas logísticas de salud. La discusión que realiza Poenarau (Cristina & Radu, 2019) en su artículo, fue de suma importancia para la conceptualización de este caso de estudio, por lo que su trabajo se extendió de una mera propuesta a un desarrollo tecnológico tangible. Thesmar (Thesmar et al., 2019a) reconoce varios retos en la gestión de datos generados en cadenas logísticas de salud, los más cruciales, relacionados con los derechos de los pacientes, pueden ser eliminados al incorporar regulaciones y protocolos que existen dentro de las instituciones públicas de salud, como esta propuesta. Esto es enfatizado por Chute et. al. (Chute & French, 2019a), donde se percibe la Industria 4.0 exclusivamente a entidades privadas de salud. El único trabajo que analiza los sistemas de información aplicados a la salud pública (Larrucea et al., 2020a) y como estos pueden coordinar distintos países bajo un mismo marco operativo, sin embargo, no propone modelos de extracción de información ni capitalización de conocimiento.

Gracias al trabajo de (Santos et al., 2019), sabemos que el personal médico tiene muchas ganas de adoptar tecnologías que les permitan mejorar su trabajo, pero que no necesariamente requieran de un gran esfuerzo de aprendizaje. Es por ello que se dedicó el tiempo suficiente para la programación y diseño de la interfaz gráfica. En específico, la tarea de visualización de datos es lo que permite que este trabajo brille en comparación con otros analizados dentro del estado del arte; una buena gráfica dice más que mil tablas.

Otros trabajos que tratan con sistemas de información en salud, se centran en la importantísima tarea de la gestión del expediente médico del paciente (Hathaliya et al., 2020a), sin embargo, ninguno trata directamente sobre traer la ciencia de los datos directamente al usuario final, en este caso, el personal médico.

La base de datos operativa que contiene los datos demográficos, geográficos y climatológicos está lista para ser utilizada para apoyar la toma de decisiones de cualquier otra enfermedad o tarea dentro de la cadena logística de salud pública; una aplicación directa e inmediata

sería para el apoyo en localización de clínicas o zonas vulnerables El prototipo implementa el **Sistema Inteligente** propuesto, pudiendo realizar la conversión de datos en sabiduría siguiendo las metodologías propuestas. Al tener la capacidad de realizar los cuatro tipos de análisis, el prototipo incide positivamente en la capacidad de toma de decisiones dentro de la cadena logística de salud pública.

El diseño del sistema sigue un patrón modular como sugiere Django, el cual centraliza las apps y permite al servidor primario gestionar su comportamiento. Sin embargo, la arquitectura general, presenta como esta centralización existe dentro de un ecosistema capaz de interactuar con agentes externos e internos de una forma dinámica. Utilizando la nube, el sistema cambia de un agente monolítico a una red de entidades multipropósito para la gestión de la salud pública. A futuro, la implementación final consistirá en el sistema principal y una nube de microservicios con capacidad de ejecución rápida y en paralelo.

Actualmente, los usuarios finales ya han sido expuestos al prototipo. Se han realizado pruebas de usabilidad y se han recibido comentarios positivos. El prototipo ha sido probado por usuarios no expertos en el área de la salud y han podido replicar los análisis realizados pasivamente en la fase de análisis de flujo de datos. Esto es un indicador de que el sistema es usable y que la interfaz gráfica es intuitiva. Ello gracias a que se buscó la retroalimentación directa con los expertos en salud para guiar el diseño total de la plataforma web.

CONCLUSIONES

Gracias a las metodologías de transformación de datos utilizadas, se pudo conjugar el proceso de análisis pasivo e integrarlo dentro de un prototipo del Sistema Híbrido Inteligente de Salud Pública. La metodología DIKW gobernó la forma en la que gestionamos y generamos el conocimiento a través de los datos que se generan orgánicamente dentro de la cadena logística de salud pública. No solamente podemos trabajar con datos históricos y describir el pasado, sino ahora también podemos anticipar eventos y responder en tiempo real a los eventos que surgen dentro de la cadena. Estos atisbos hacia el futuro que ofrece el sistema, permiten a las autoridades de salud pública tomar decisiones más informadas y con mayor certeza, lo cual se traduce en una mejor atención a la población. El sistema manifestado en el prototipo, es capaz de realizar los cuatro análisis descritos.

El diseño del sistema permitió generar una arquitectura para gestionar los datos de la cadena logística, de tal manera que de una fuente de datos nos permite realizar distintos tipos de análisis. Dentro del Sistema Inteligente, confluyen diversos algoritmos para tareas dispares. Operando con un espectro amplio de soluciones que solo el paradigma de industria 4.0 puede ofrecer. La integración entre distintas fuentes de datos, otorga un poder inigualable al sistema que permite observar el comportamiento de la cadena desde distintas perspectivas. En particular, el análisis estratificado de los datos en categorías poblacionales permite identificar sesgos y brechas de atención en la población.

Encontrar un área de oportunidad grande en los ejes de investigación, donde la brecha de trabajos relacionados es bastante evidente. La comparativa de los trabajos analizados en el estado del arte, nos permiten ver las fortalezas y el aporte de esta propuesta. La necesidad se evidencia apremiante en un momento histórico donde más que nunca se requieren sistemas robustos acoplados dentro de las cadenas logísticas de salud para poder solucionar problemas complejos, que en este caso, tienen la capacidad de salvar miles de vidas. Los avances en estos semestres se enfocaron en abstraer un caso particular para usarlo como origen del sistema, nuestro caso de estudio de Fiebre del Dengue. El objetivo final de este caso de estudio es generar un módulo de enfermedades transmitidas por vector para ser integrado dentro del

Sistema de Información en Salud. Durante este trayecto, se encontraron las fuentes de datos externas pertinentes que podrán ser ocupadas para todos los módulos a integrar dentro del sistema de información.

El componente más robusto e importante, aparte del predictor de brotes y el análisis de series de tiempo, es el clasificador de Fiebre del Dengue utilizando *Machine Learning*. No solo sirvió para validar la metodología y determinar si esta era capaz de encapsular todo el trabajo realizado, sino también permitió descubrir sesgos importantes en poblaciones vulnerables. Lo destacable es que esto se descubrió no realizando un análisis directo sobre los datos, sino que siguió un proceso que, de manera holística, culminó en confirmar la hipótesis inicial. El sistema es capaz de generar conocimiento de manera autónoma, sin necesidad de intervención humana, lo cual es una de las características más importantes de los sistemas inteligentes.

Los otros dos casos de estudio no solo formaron parte del pensamiento inductivo utilizado para el aporte al conocimiento, sino también por si mismos son aportaciones sumamente importantes. Sobre todo en la capacidad de dotar de inteligencia al sistema y generar pericia en la utilización de tecnologías y algoritmos sumamente complejos y difíciles de implementar.

En comparación con las plataformas pasadas, el Sistema de Información pasa de ser un repositorio pasivo de datos en un sistema activo y dinámico. Cuenta con una interfaz moderna e intuitiva que ha sido validada directamente por los usuarios finales. Al enfocarnos no solo en la parte técnica de la industria 4.0, sino también en los conceptos de seguridad y protección de datos, el sistema cumple con las normas y leyes vigentes nacionales para que los datos de los pacientes estén protegidos.

La meta siempre fue crear una plataforma que permita al experto en salud, realizar poderosos análisis y extraer conocimiento de una manera sencilla y amigable. No bastaba con encontrar patrones en los datos, sino permitir al experto decidir que tipos de análisis y a que resoluciones espacio-temporales para que el sistema se adapte a sus necesidades. El análisis pasivo inicial permitió encontrar patrones y generar hipótesis sobre el comportamiento de la enfermedad, por ejemplo, la concentración de casos en población estudiantil, o los sesgos poblacionales evidenciados contra la población indígena.

La metodología propuesta, no solamente se distingue por su enfoque y la integración de tecnologías de la industria 4.0; sino también integra conceptos de equidad social para enfocar los esfuerzos en mejorar la calidad de vida en poblaciones vulnerables. También, enfocarse en la explicabilidad de las predicciones de los algoritmos permite transparentar el proceso y no solamente mostrar resultados al experto, sino también retroalimentar el proceso mismo.

Lo importante de la metodología, es que se puede extrapolar a distintas enfermedades, por ejemplo, fiebre Chikungunya y Zika utilizando los datos de la plataforma SINAVE. Pero no solo se limita a una sola fuente de datos, es tan general que puede ser aplicada para gobernar gestión de conocimiento en diferentes eslabones de la cadena y está demostrado que funciona en distintas cadenas logísticas de salud: Tratamiento, Diagnóstico, Trazabilidad, Administración, entre otras.

La ventaja actual es que el prototipo es totalmente funcional y encapsula el Sistema de Información, la arquitectura y la metodología generada para poder mejorar la gestión del conocimiento dentro de las cadenas de logística de salud pública. Permitiendo a los expertos tomar mejores decisiones, basadas en sabiduría, para poder mejorar la calidad de vida de la toda la población. Todo ello cumpliendo los mejores estándares de seguridad y protección de datos; con una fuerte interacción con los conceptos de equidad social, sin distinción de origen étnico, género, o de clase social.

RECOMENDACIONES

Como recomendaciones finales, se proponen las siguientes:

- Crear un protocolo para migrar de la plataforma SINAVE al Sistema Inteligente propuesto, con el fin de eliminar la complejidad en los procesos de ETL y reducir la redundancia en los procesos de captura de datos que realiza la Dirección General de Epidemiología.
- Crear una política pública de salud para la estandarización de los datos de los sistemas de información en salud, con el fin de facilitar la interoperabilidad y reducir los errores de captura asociados a la heterogeneidad de los datos.
- Llevar los conceptos de equidad a un primer plano para la toma de decisiones en salud, con el fin de reducir las desigualdades en el acceso a los servicios de salud.
- Establecer un programa piloto para la implementación del sistema aquí expuesto para convertirlo de un prototipo a una plataforma final.
- Incrementar la vigilancia epidemiológica en escuelas, con especial atención a las de nivel primaria y secundaria, con el fin de eliminar brotes que se detectó que se originan en estos lugares.
- Analizar a conciencia la razón por la cual existe un sesgo sumamente marcado y alarmante en la población indígena femenina infante a nivel nacional.

CONTRIBUCIÓN AL CONOCIMIENTO Y DESARROLLO TECNOLÓGICO

La investigación aquí expuesta tiene como objetivo analizar las cadenas logísticas de sanidad pública para encontrar eslabones susceptibles a mejora, usando el paradigma de *Industria 4.0* y los **Sistemas Inteligentes de Información en Salud Pública** para mejorar la gestión del conocimiento, esto permitirá dar una solución integral a los problemas expuestos dentro de los casos de estudio.

Las aportaciones tanto tecnológicas como científica, son las siguientes:

Aportación tecnológica

Aportación tecnológica del trabajo: Un Sistema de Información en Salud Pública que integra los paradigmas de *Industria 4.0* para mejorar la manera en la que los datos se refinan en conocimiento y posteriormente sabiduría. Este sistema será especialmente robusto ante la incertidumbre para tomar mejores decisiones.

Aportación tecnológica de caso de estudio COVID-19:

- Sistema híbrido inteligente para la optimización de rutas de Logística Humanitaria, mediante el cual los usuarios (tomadores de decisiones) podrán obtener las rutas óptimas en situaciones de riesgo epidemiológico.
- Herramienta CLI, la cual permitirá a usuarios desarrolladores la integración del Sistema híbrido inteligente en sus propias aplicaciones.

Aportación tecnológica de caso de estudio Cáncer cervicouterino:

- Sistema de diagnóstico asistido por computadora para la segmentación semántica de células cervicales, el cual permitirá a los usuarios médicos, citotecnólogos, patólogos y oncólogos mejorar la precisión en el diagnóstico de cáncer cervicouterino.

Aportación tecnológica de caso de estudio Dengue:

- Prototipo de Sistema de Información en Salud basado en Industria 4.0 para la gestión total de la cadena logística epidemiológica de la Fiebre del Dengue, el cual permitirá a los usuarios (médicos, generadores de política pública, epidemiólogos, investigadores, etc.) mejorar la eficiencia en la prevención y control de la enfermedad.

Aportación al Conocimiento

Aportación científica del trabajo: Una arquitectura de logística de sanidad pública, que integre los paradigmas de *Industria 4.0* para mejorar la manera en la que los datos se refinan en conocimiento y posteriormente sabiduría. Esta arquitectura debe ser especialmente robusta ante la incertidumbre para tomar mejores decisiones.

Aportación científica de caso de estudio COVID-19:

- Metodología para la optimización de rutas de Logística Humanitaria, utilizando algoritmos de clústering (K-means) y metaheurística (Algoritmo Genético).

Aportación científica de caso de estudio Cáncer cervicouterino:

- Arquitectura neuronal para la segmentación semántica de células cervicales utilizando Feature Pyramid Networks y EffientNet.

Aportación científica de caso de estudio Dengue:

- Metodología para la creación de un modelo capaz de clasificar el tipo de fiebre del dengue. Optimizando los hiperparámetros de un Bosque de Decisión mediante métodos bayesianos.
- Metodología para la conjugación de datos epidemiológicos, demográficos y climatológicos para el análisis de la propagación de la fiebre del dengue.

- Metodología para la conjugación de datos geoespaciales, tabulares y series de tiempo para la caracterización de la fiebre del dengue.

REFERENCIAS BIBLIOGRÁFICAS

- Abonyi, J., Feil, B., & Abraham, A. (2005). Computational Intelligence in Data Mining. *Informatica (Slovenia)*, 29, 3-12. <http://www.informatica.si/index.php/informatica/article/view/13>
- Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16, 3-9. <http://www-public.imtbs-tsp.eu/%5Ctextbackslashstextasciitildegibson/Teaching/Teaching-ReadingMaterial/Ackoff89.pdf>
- Adan, V. S. (2018). *libro contribuciones del IMSS a la medicina mundial, pasado, presente y futuro*. <https://www.researchgate.net/publication/329841888>
- Al-Zoubi, A. M., Heidari, A. A., Habib, M., Faris, H., Aljarah, I., & Hassonah, M. A. (2019). Evolutionary Machine Learning Techniques, Algorithms and Applications. Springer, Singapore. https://doi.org/10.1007/978-981-32-9990-0_2
- Al-Zoubi, A. M., Heidari, A. A., Habib, M., Faris, H., Aljarah, I., & Hassonah, M. A. (2020). Salp Chain-Based Optimization of Support Vector Machines and Feature Weighting for Medical Diagnostic Information Systems. Springer, Singapore. https://doi.org/10.1007/978-981-32-9990-0_2
- Bandyopadhyay, H., & Nasipuri, M. (2020). Segmentation of Pap Smear Images for Cervical Cancer Detection. *2020 IEEE Calcutta Conference, CALCON 2020 - Proceedings*, 00, 30-33. <https://doi.org/10.1109/calcon49167.2020.9106484>
- Bates, M. (2019). Health Care Chatbots Are Here to Help. *IEEE Pulse*, 10(3), 12-14. <https://doi.org/10.1109/MPULS.2019.2911816>
- Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., & Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504-507. <https://doi.org/10.1038/nature12060>
- Boulos, M. N. K., Peng, G., & Vopham, T. (2019). An overview of GeoAI applications in health and healthcare. *International Journal of Health Geographics*, 18, 1-9. <https://doi.org/10.1186/s12942-019-0171-2>
- Cáceres, C., Rosário, J. M., & Amaya, D. (2019). Towards health 4.0: E-hospital proposal based industry 4.0 and artificial intelligence concepts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11526 LNAI, 84-89. https://doi.org/10.1007/978-3-030-21642-9_12
- CDC, u. (2020). Life Cycle of Aedes aegypti and Ae. albopictus Mosquitoes [https://www.cdc.gov/mosquitoes/pdfs/AedesLifeCycle_50https://www.cdc.gov/]. *U.S. Departement of Health and Human Services*, 1-2. <https://www.cdc.gov/mosquitoes/about/life-cycles/aedes.html>
- Chankong, T., Theera-Umpon, N., & Auephanwiriyakul, S. (2014). Automatic cervical cell segmentation and classification in Pap smears. *Computer Methods and Programs in Biomedicine*, 113(2), 539-556. <https://doi.org/10.1016/j.cmpb.2013.12.012>

- Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation [<http://dx.doi.org/10.1109/VCIP.2017.8305148>]. *2017 IEEE Visual Communications and Image Processing, VCIP 2017, 2018-Janua*, 1-4. <https://doi.org/10.1109/vcip.2017.8305148>
- Chute, C., & French, T. (2019a). Introducing Care 4.0: An Integrated Care Paradigm Built on Industry 4.0 Capabilities. *International Journal of Environmental Research and Public Health*, 16(12), 2247. <https://doi.org/10.3390/ijerph16122247>
- Chute, C., & French, T. (2019b). Introducing care 4.0: An integrated care paradigm built on industry 4.0 capabilities. *International Journal of Environmental Research and Public Health*, 16(12). <https://doi.org/10.3390/ijerph16122247>
- Clarke, M. (2000, febrero). The QUORUM statement. [https://doi.org/10.1016/S0140-6736\(05\)72172-3](https://doi.org/10.1016/S0140-6736(05)72172-3)
- Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess (with Discussion). *Journal of Official Statistics*, 6, 3-73.
- Clifford, G. D., Blaya, J. A., Hall-Clifford, R., & Fraser, H. S. (2008). Medical information systems: A foundation for healthcare technologies in developing countries. *BioMedical Engineering OnLine*, 7(1), 18. <https://doi.org/10.1186/1475-925x-7-18>
- CONAGUA, u. (2020). Resúmenes Mensuales de Temperaturas y Lluvia [<https://smn.conagua.gob.mx/es/climatologia/temperaturas-y-lluvias/resumenes-mensuales-de-temperaturas-y-lluvias%0Ahttp://smn.cna.gob.mx/es/clim>]. <https://smn.conagua.gob.mx/es/climatologia/temperaturas-y-lluvias/resumenes-mensuales-de-temperaturas-y-lluvias>
- Control, European Centre for Disease Prevention and. (s.f.). Geographical distribution of dengue cases reported worldwide, 2020. Consultado el 21 de noviembre de 2023, desde <https://www.ecdc.europa.eu/en/publications-data/geographical-distribution-dengue-cases-reported-worldwide-2020>
- CORE. (s.f.). CORE – Aggregating the world’s open access research papers. Consultado el 21 de mayo de 2020, desde <https://core.ac.uk/>
- Cristina, P., & Radu, D. (2019). Fractal techniques in Health Information Systems management: Can they be used? 00, 592-597. <https://doi.org/10.1109/cscs.2019.00108>
- Dammann, O. (2018). Data, Information, Evidence, and Knowledge:: A Proposal for Health Informatics and Data Science [<https://pmc/articles/PMC6435353/?report=abstract> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6435353/>]. *Online Journal of Public Health Informatics*, 10. <https://doi.org/10.5210/ojphi.v10i3.9631>
- Dansana, D., Kumar, R., Bhattacharjee, A., Hemanth, D. J., Gupta, D., Khanna, A., & Castillo, O. (2020). Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm. *Soft Computing*, 1-9. <https://doi.org/10.1007/s00500-020-05275-y>
- Daú, G., Scavarda, A., Scavarda, L. F., & Portugal, V. J. T. (2019). The healthcare sustainable supply chain 4.0: The circular economy transition conceptual framework with the corporate social responsibility mirror. *Sustainability (Switzerland)*, 11(12). <https://doi.org/10.3390/su11123259>

- Dautov, R., Distefano, S., & Buyya, R. (2019). Hierarchical data fusion for Smart Healthcare. *Journal of Big Data*, 6(1), 1-23. <https://doi.org/10.1186/s40537-019-0183-6>
- Del Moral, M., Ochoa, A., Lasserre, A., & Cedillo, G. (2022). Logistics of Hospitalization Patients with COVID and Ambulances Required. En *Studies in Systems, Decision and Control* (pp. 155-168, Vol. 347). Springer, Cham. https://doi.org/10.1007/978-3-030-68663-5_11
- Dos Santos, L. R. A., Bernardi, F. A., Silva Prado, G. C., Lima, V. C., Crepaldi, N. Y., Marçal, M. A., Lopes Rijo, R. P. C., Galliez, R. M., Ruffino-Netto, A., & Alves, D. (2019). The perception of health providers about an artificial intelligence applied to Tuberculosis video-based treatment in Brazil: A protocol proposal. *Procedia Computer Science*, 164, 595-601. <https://doi.org/10.1016/j.procs.2019.12.225>
- Dounias, G. (2017). MDE-Lab : The Management and Decision Engineering Laboratory. <http://mde-lab.aegean.gr/>
- Du, L. (2018a). Research on Key Technologies of Regional Synergy Emergency System Based on Medical Data Center. 154, 732-737. <https://doi.org/10.1016/j.procs.2019.06.114>
- Du, L. (2018b). Research on Key Technologies of Regional Synergy Emergency System Based on Medical Data Center. *Procedia Computer Science*, 154, 732-737. <https://doi.org/10.1016/j.procs.2019.06.114>
- Esri, u. (2016). How IDW works—Help | ArcGIS for Desktop. <https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-idw-works.htm>
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, 226-231. <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220>
- Fayyad, U. M. (1996). Advances in knowledge discovery and data mining, 611.
- Gautam, S., Bhavsar, A., Sao, A. K., & K.K., H. (2018). CNN based segmentation of nuclei in PAP-smear images with selective pre-processing. *Medical Imaging 2018: Digital Pathology*, 10581(May), 32. <https://doi.org/10.1117/12.2293526>
- Geitgey, A. (2020). Four Basic Data Science Lessons Illustrated by COVID-19 Data. <https://medium.com/@ageitgey/four-basic-data-science-lessons-illustrated-by-covid-19-data-7d94134a5b0e>
- Gergin, Z., Tunçbilek, N., & Esnaf, Ş. (2018). Clustering Approach Using Artificial Bee Colony Algorithm for Healthcare Waste Disposal Facility Location Problem. *International Journal of Operations Research and Information Systems*, 10(1), 56-75. <https://doi.org/10.4018/ijoris.2019010104>
- Gergin, Z., Tunçbilek, N., & Esnaf, Ş. (2019). Clustering Approach Using Artificial Bee Colony Algorithm for Healthcare Waste Disposal Facility Location Problem. *International Journal of Operations Research and Information Systems (IJORIS)*, 10(1), 56-75. <https://doi.org/10.4018/ijoris.2019010104>
- Gill, H. S., & Rao, P. C. (1996). *Data Warehousing, la integración de información para la mejor toma de decisiones*. Prentice Hall Hispanoamericana.
- Goldberg, M., Jougla, E., Fassa, M., Padieu, R., & Quantin, C. (2012). The French health information system. *Statistical Journal of the IAOS*, 28, 31-41. <https://doi.org/10.3233/sji-2012-0747>

- Google. (s.f.). Google Académico. Consultado el 21 de mayo de 2020, desde <https://scholar.google.com/>
- Gotham, I. J., Le, L. H., Sottolano, D. L., & Schmit, K. J. (2015). An informatics framework for public health information systems: a case study on how an informatics structure for integrated information systems provides benefit in supporting a statewide response to a public health emergency. *Information Systems and e-Business Management*, 13(4), 713-749. <https://doi.org/10.1007/s10257-014-0240-9>
- Gotham, I. J., Sottolano, D. L., Hennessy, M. E., Napoli, J. P., Dobkins, G., Le, L. H., Burhans, R. L., & Fage, B. I. (2007). An Integrated Information System for All‐Hazards Health Preparedness and Response. *Journal of Public Health Management and Practice*, 13(5), 486-496. <https://doi.org/10.1097/01.phh.0000285202.48588.89>
- Gräßer, F., Beckert, S., Küster, D., Schmitt, J., Abraham, S., Malberg, H., & Zaunseder, S. (2017). Therapy Decision Support Based on Recommender System Methods. *Journal of Healthcare Engineering*, 2017. <https://doi.org/10.1155/2017/8659460>
- Greenes, R. A. (2015, septiembre). Healthcare Information Management Systems, Cases, Strategies, and Solutions. En *Healthcare Information Management Systems: Cases, Strategies, and Solutions: Fourth Edition* (pp. 579-600). Springer International Publishing. https://doi.org/10.1007/978-3-319-20765-0_33
- Gubler, D. J. (2011). Dengue, Urbanization and globalization: The unholy trinity of the 21 st century. *Tropical Medicine and Health*, 39, 3-11. <https://doi.org/10.2149/tmh.2011-s05>
- Guzman, M. G., Gubler, D. J., Izquierdo, A., Martinez, E., & Halstead, S. B. (2016). Dengue infection. *Nature Reviews Disease Primers*, 2, 1-25. <https://doi.org/10.1038/nrdp.2016.55>
- Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 100-108. <https://doi.org/10.2307/2346830>
- Hathaliya, J. J., Tanwar, S., & Evans, R. (2020a). Securing electronic healthcare records: A mobile-based biometric authentication approach. *Journal of Information Security and Applications*, 53, 102528. <https://doi.org/10.1016/j.jisa.2020.102528>
- Hathaliya, J. J., Tanwar, S., & Evans, R. (2020b). Securing electronic healthcare records: A mobile-based biometric authentication approach. *Journal of Information Security and Applications*, 53, 102528. <https://doi.org/10.1016/j.jisa.2020.102528>
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 386-397. <https://doi.org/10.1109/tpami.2018.2844175>
- IBM, u. (2020). From data to knowledge – IBM Developer. <https://developer.ibm.com/articles/ba-data-becomes-knowledge-1/>
- INEGI. (s.f.). Diccionario de datos del Marco Geoestadístico. Consultado el 26 de noviembre de 2023, desde <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463906933>
- INEGI. (2019). Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) 2018. Comunicado de prensa. *Instituto Nacional de Estadística y Geografía (INEGI)*, 19.

https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2019/OtrTemEcon/ENDUTIH%5C_2018.pdf

- INEGI, u. (2020). Censo Población y Vivienda 2020. Población total [<https://www.inegi.org.mx/programas/ccpv/2020/>]. *Censo Población y Vivienda 2020*. <https://www.inegi.org.mx/programas/ccpv/2020/default.html%5C#Tabulados>
- Jantzen, J., Norup, J., Dounias, G., & Bjerregaard, B. (2005). Pap-smear Benchmark Data For Pattern Classification. *Proceedings of NiSIS 2005: Nature inspired Smart Information Systems (NiSIS)*, 1-9.
- Jason, C. (2020). COVID-19 Exposes Lack of Health Data Exchange, Interoperability. <https://ehrintelligence.com/news/covid-19-exposes-lack-of-health-data-exchange-interoperability>
- Javaid, M., & Haleem, A. (2020). Impact of industry 4.0 to create advancements in orthopaedics. *Journal of Clinical Orthopaedics and Trauma*. <https://doi.org/10.1016/j.jcot.2020.03.006>
- Kale, D., Hanlin, R., & Chataway, J. (2013). New drugs and health technologies for low-income populations: will the private sector meet the needs of low-income populations in developing countries? *Innovation and Development*, 3(1), 121-137. <https://doi.org/10.1080/2157930x.2013.764629>
- Khalidi, R., Afia, A. E., & Chiheb, R. (2019a). Forecasting of weekly patient visits to emergency department: Real case study. *Procedia Computer Science*, 148, 532-541. <https://doi.org/10.1016/j.procs.2019.01.026>
- Khalidi, R., Afia, A. E., & Chiheb, R. (2019b). Forecasting of weekly patient visits to emergency department: real case study. 148, 532-541. <https://doi.org/10.1016/j.procs.2019.01.026>
- Knaus, W. A., & Marks, R. D. (2019a). New Phenotypes for Sepsis. *JAMA*, 321(20), 1981-1982. <https://doi.org/10.1001/jama.2019.5794>
- Knaus, W. A., & Marks, R. D. (2019b, mayo). New Phenotypes for Sepsis: The Promise and Problem of Applying Machine Learning and Artificial Intelligence in Clinical Research. <https://doi.org/10.1001/jama.2019.5794>
- Kornblith, S., Shlens, J., & Le, Q. V. (2018). Do Better ImageNet Models Transfer Better? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019-June*, 2656-2666. <https://doi.org/10.1109/cvpr.2019.00277>
- Kramer, O. (2017). Genetic algorithms. En *Studies in Computational Intelligence* (pp. 11-19, Vol. 679). Springer, Boston, MA. https://doi.org/10.1007/978-3-319-52156-5_2
- Krishnapuram, B., Shah, M., Smola, A., Aggarwal, C., Shen, D., Rastogi, R., Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125-2126. <https://doi.org/10.1145/2939672.2945386>
- Kuo, Y. H., Chan, N. B., Leung, J. M., Meng, H., So, A. M. C., Tsoi, K. K., & Graham, C. A. (2020). An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department. *International Journal of Medical Informatics*, 139, 104143. <https://doi.org/10.1016/j.ijmedinf.2020.104143>

- Kurnianingsih, u., Allehaibi, K. H. S., Nugroho, L. E., Widyawan, u., Lazuardi, L., Prabuwo, A. S., & Mantoro, T. (2019). Segmentation and classification of cervical cells using deep learning. *IEEE Access*, 7, 116925-116941. <https://doi.org/10.1109/access.2019.2936017>
- Larrucea, X., Moffie, M., Asaf, S., & Santamaria, I. (2020a). Towards a GDPR compliant way to secure European cross border Healthcare Industry 4.0. *Computer Standards & Interfaces*, 69, 103408. <https://doi.org/10.1016/j.csi.2019.103408>
- Larrucea, X., Moffie, M., Asaf, S., & Santamaria, I. (2020b). Towards a GDPR compliant way to secure European cross border Healthcare Industry 4.0. *Computer Standards and Interfaces*, 69. <https://doi.org/10.1016/j.csi.2019.103408>
- Lezzi, M., Lazoi, M., & Corallo, A. (2018). Cybersecurity for Industry 4.0 in the current literature: A reference framework. *Computers in Industry*, 103, 97-110. <https://doi.org/10.1016/j.compind.2018.09.004>
- Lin, H., Li, R., Liu, Z., Chen, J., Yang, Y., Chen, H., Lin, Z., Lai, W., Long, E., Wu, X., Lin, D., Zhu, Y., Chen, C., Wu, D., Yu, T., Cao, Q., Li, X., Li, J., Li, W., . . . Liu, Y. (2019a). Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine*, 9, 52-59. <https://doi.org/10.1016/j.eclinm.2019.03.001>
- Lin, H., Li, R., Liu, Z., Chen, J., Yang, Y., Chen, H., Lin, Z., Lai, W., Long, E., Wu, X., Lin, D., Zhu, Y., Chen, C., Wu, D., Yu, T., Cao, Q., Li, X., Li, J., Li, W., . . . Liu, Y. (2019b). Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. *EClinicalMedicine*, 9, 52-59. <https://doi.org/10.1016/j.eclinm.2019.03.001>
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5), 740-755. https://doi.org/10.1007/978-3-319-10602-1_48
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2016). Feature Pyramid Networks for Object Detection. *arXiv*, 00, 1500-1504. <https://doi.org/10.1109/ispa-bdcloud-sustaincom-socialcom48970.2019.00217>
- Ling, S. (2014). Special issue on hybrid intelligent methods for health technologies. *Applied Soft Computing*, 20, 1-3. <https://doi.org/10.1016/j.asoc.2014.04.033>
- Luh, J. Y., Thompson, R. F., & Lin, S. (2019). Clinical Documentation and Patient Care Using Artificial Intelligence in Radiation Oncology. *Journal of the American College of Radiology*, 16(9), 1343-1346. <https://doi.org/10.1016/j.jacr.2019.05.044>
- Ma, J. (2020). Segmentation Loss Odyssey. *arXiv*. <http://arxiv.org/abs/2005.13449>

- Maheri, A., Jalili, S., Hosseinzadeh, Y., Khani, R., & Miryahiavi, M. (2021). A comprehensive survey on cultural algorithms. *Swarm and Evolutionary Computation*, 62, 100846. <https://doi.org/10.1016/j.swevo.2021.100846>
- Massoudi, B. L., Goodman, K. W., Gotham, I. J., Holmes, J. H., Lang, L., Miner, K., Potenziani, D. D., Richards, J., Turner, A. M., & Fu, P. C. (2012). An informatics agenda for public health: summarized recommendations from the 2011 AMIA PHI Conference. *Journal of the American Medical Informatics Association*, 19(5), 688-695. <https://doi.org/10.1136/amiainl-2011-000507>
- Microsoft. (s.f.). Home | Microsoft Academic. Consultado el 21 de mayo de 2020, desde <https://academic.microsoft.com/home>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Moreira, A., & Santos, M. F. (2020). Multichannel Interaction for Healthcare Intelligent Decision Support. *Procedia Computer Science*, 170, 1053-1058. <https://doi.org/10.1016/j.procs.2020.03.074>
- Mujumdar, A., & Vaidehi, V. (2019a). Diabetes Prediction using Machine Learning Algorithms. 165, 292-299. <https://doi.org/10.1016/j.procs.2020.01.047>
- Mujumdar, A., & Vaidehi, V. (2019b). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165, 292-299. <https://doi.org/10.1016/j.procs.2020.01.047>
- Nelson, A., Herron, D., Rees, G., & Nachev, P. (2019). Predicting scheduled hospital attendance with artificial intelligence. *npj Digital Medicine*, 2(1), 1-7. <https://doi.org/10.1038/s41746-019-0103-3>
- Nelson, K., Corbin, G., Anania, M., Kovacs, M., Tobias, J., & Blowers, M. (2015). Evaluating Model Drift in Machine Learning Algorithms. 2015 *IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1-8. <https://doi.org/10.1109/cisda.2015.7208643>
- Neumaier, M. (2019). Diagnostics 4.0: the medical laboratory in digital health. *Clinical Chemistry and Laboratory Medicine*.
- Oddoye, J. P., Yaghoobi, M. A., Tamiz, M., Jones, D. F., & Schmidt, P. (2007). A multi-objective model to determine efficient resource levels in a medical assessment unit. *Journal of the Operational Research Society*, 58(12), 1563-1573. <https://doi.org/10.1057/palgrave.jors.2602315>
- Pace, P., Aloï, G., Gravina, R., Caliciuri, G., Fortino, G., & Liotta, A. (2018). An Edge-Based Architecture to Support Efficient Applications for Healthcare Industry 4.0. *IEEE Transactions on Industrial Informatics*, 15(1), 481-489. <https://doi.org/10.1109/tii.2018.2843169>
- Pace, P., Aloï, G., Gravina, R., Caliciuri, G., Fortino, G., & Liotta, A. (2019). An Edge-Based Architecture to Support Efficient Applications for Healthcare Industry 4.0. *IEEE Transactions on Industrial Informatics*, 15(1), 481-489. <https://doi.org/10.1109/TII.2018.2843169>
- Paton, C., & Kobayashi, S. (2019). An Open Science Approach to Artificial Intelligence in Healthcare. *Yearbook of Medical Informatics*, 28(01), 047-051. <https://doi.org/10.1055/s-0039-1677898>

- Perl, T. M., & Price, C. S. (2020). Managing Emerging Infectious Diseases: Should Travel Be the Fifth Vital Sign? *Annals of Internal Medicine*, 172(8), M20-0643. <https://doi.org/10.7326/m20-0643>
- Peters, D. H., Garg, A., Bloom, G., Walker, D. G., Brieger, W. R., & Rahman, M. H. (2008). Poverty and Access to Health Care in Developing Countries. *Annals of the New York Academy of Sciences*, 1136(1), 161-171. <https://doi.org/10.1196/annals.1425.011>
- Poenaru, C. E., & Dobrescu, R. (2019). Fractal techniques in health information systems management: Can they be used? *Proceedings - 2019 22nd International Conference on Control Systems and Computer Science, CSCS 2019*, 592-597. <https://doi.org/10.1109/CSCS.2019.00108>
- Rayan, Z., Alfonse, M., & Salem, A. B. M. (2018). Machine Learning Approaches in Smart Health. *Procedia Computer Science*, 154, 361-368. <https://doi.org/10.1016/j.procs.2019.06.052>
- Rayan, Z., Alfonse, M., & Salem, A.-B. M. (2019). Machine Learning Approaches in Smart Health. 154, 361-368. <https://doi.org/10.1016/j.procs.2019.06.052>
- Riquelme, R. (2019). México llega con retraso a la Cuarta Revolución Industrial. *EL Economista*. <https://www.economista.com.mx/tecnologia/Mexico-llega-con-retraso-a-la-Cuarta-Revolucion-Industrial-20191009-0055.html>
- Roberts, M. W. (2018). Successful Public Health Information System Database Integration Projects: A Qualitative Study [pmc/articles/PMC6194103/?report=abstract <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6194103/>]. *Online Journal of Public Health Informatics*, 10. <https://doi.org/10.5210/ojphi.v10i2.9221>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. https://doi.org/10.1007/978-3-319-24574-4_28
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., ... Fei-Fei, L. (2014). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115, 211-252. <https://doi.org/10.1007/s11263-015-0816-y>
- Sait, M., Alattas, A., Omar, A., Almalki, S., Sharf, S., & Alsaggaf, E. (2019). Employing Virtual Reality Techniques in environment adaptation for autistic children. *Procedia Computer Science*, 163, 338-344. <https://doi.org/10.1016/j.procs.2019.12.116>
- Salas, I. F., & Leal, A. F. (1995). El papel del vector *Aedes aegypti* en la epidemiología del dengue en México. *Salud Pública de México*, 37, 45-52. <https://www.redalyc.org/articulo.oa?id=10609206>
- Salud, S. d. (s.f.). Sinave. Consultado el 21 de noviembre de 2023, desde <https://vectores.sinave.gob.mx/>
- Salud, S. d. (2020a). Datos Abiertos - Dirección General de Epidemiología | Secretaría de Salud | Gobierno | gob.mx. <https://www.gob.mx/salud/documentos/datos-abiertos-152127>
- Salud, S. d. (2020b, noviembre). Exceso de Mortalidad en México – Coronavirus. Consultado el 13 de noviembre de 2020, desde <https://coronavirus.gob.mx/exceso-de-mortalidad-en-mexico/>

- Salud, S. d. (2023). Sistema Nacional de Vigilancia Epidemiológica | Secretaría de Salud | Gobierno | gob.mx. Consultado el 26 de noviembre de 2023, desde <https://www.gob.mx/salud/acciones-y-programas/sistema-nacional-de-vigilancia-epidemiologica>
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4510-4520. <https://doi.org/10.1109/cvpr.2018.00474>
- Santos, L. R. A. d., Bernardi, F. A., Prado, G. C. S., Lima, V. C., Crepaldi, N. Y., Marçal, M. A., Rijo, R. P. C. L., Galliez, R. M., Ruffino-Netto, A., & Alves, D. (2019). The perception of health providers about an artificial intelligence applied to Tuberculosis video-based treatment in Brazil: a protocol proposal. *164*, 595-601. <https://doi.org/10.1016/j.procs.2019.12.225>
- Schawb, K. (2016). *The Fourth Industrial Revolution* (First). World Economic Forum.
- ScienceDirect. (2021). ScienceDirect.com | Science, health and medical journals, full text articles and books. Consultado el 1 de diciembre de 2021, desde <https://www.sciencedirect.com/>
- Sharma, B., & Mangat, K. K. (2016). An improved nucleus segmentation for cervical cell images using FCM clustering and BPNN. *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, 1924-1929. <https://doi.org/10.1109/icacci.2016.7732332>
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.
- Shiffman, J., & Shawar, Y. R. (2020). Strengthening accountability of the global health metrics enterprise [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7162633/]. *The Lancet*, 395(10234), 1452-1456. [https://doi.org/10.1016/s0140-6736\(20\)30416-5](https://doi.org/10.1016/s0140-6736(20)30416-5)
- Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1-14. <https://arxiv.org/abs/1409.1556v6>
- Smith, B., Yurcik, W., & Doss, D. (2002). Ethical Hacking: The Security Justification Redux. *IEEE 2002 International Symposium on Technology and Society (ISTAS'02). Social Implications of Information and Communication Technology. Proceedings (Cat. No.02CH37293)*, 374-379. <https://doi.org/10.1109/istas.2002.1013840>
- Sun, S., Zheng, X., Gong, B., Paredes, J. G., & Ordieres-Meré, J. (2020). Healthy operator 4.0: A human cyber-physical system architecture for smart workplaces. *Sensors (Switzerland)*, 20(7). <https://doi.org/10.3390/s20072011>
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249. <https://doi.org/10.3322/caac.21660>

- Tanwar, S., Parekh, K., & Evans, R. (2020a). Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *Journal of Information Security and Applications*, 50, 102407. <https://doi.org/10.1016/j.jisa.2019.102407>
- Tanwar, S., Parekh, K., & Evans, R. (2020b). Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *Journal of Information Security and Applications*, 50, 102407. <https://doi.org/10.1016/j.jisa.2019.102407>
- Taylor, L., & Nitschke, G. (2017). Improving Deep Learning using Generic Data Augmentation. *arXiv*. <http://arxiv.org/abs/1708.06020>
- Thesmar, D., Sraer, D., Pinheiro, L., Dadson, N., Veliche, R., & Greenberg, P. (2019a). Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges. *PharmacoEconomics*, 37(6), 745-752. <https://doi.org/10.1007/s40273-019-00777-6>
- Thesmar, D., Sraer, D., Pinheiro, L., Dadson, N., Veliche, R., & Greenberg, P. (2019b). Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges. *PharmacoEconomics*, 37(6), 745-752. <https://doi.org/10.1007/s40273-019-00777-6>
- Tiwari, K., Kumar, S., & Tiwari, R. (2020). FOG Assisted Healthcare Architecture for Pre-Operative Support to Reduce Latency. *Procedia Computer Science*, 167, 1312-1324. <https://doi.org/10.1016/j.procs.2020.03.447>
- Transportes, u. S. d. C. y. (2018). En México hay 74.3 millones de usuarios de internet y 18.3 millones de hogares con conexión a este servicio: ENDUTIH 2018 | Secretaría de Comunicaciones y Transportes | Gobierno | gob.mx. <https://www.gob.mx/sct/prensa/en-mexico-hay-74-3-millones-de-usuarios-de-internet-y-18-3-millones-de-hogares-con-conexion-a-este-servicio-endutih-2018-196013?idiom=es>
- Tseng, C. H., Lin, C., Chang, H. C., Liu, C. C., Serafico, B. M. F., Wu, L. C., Lin, C. T., Hsu, T., Huang, C. Y., & Lo, M. T. (2019). Cloud-Based Artificial Intelligence System for Large-Scale Arrhythmia Screening. *Computer*, 52(11), 40-51. <https://doi.org/10.1109/MC.2019.2933195>
- Tseng, C.-H., Lin, C., Chang, H.-C., Liu, C.-C., Serafico, B. M. F., Wu, L.-C., Lin, C.-T., Hsu, T., Huang, C.-Y., & Lo, M.-T. (2019). Cloud-Based Artificial Intelligence System for Large-Scale Arrhythmia Screening. *Computer*, 52(11), 40-51. <https://doi.org/10.1109/mc.2019.2933195>
- UNIÓN, C. D. D. D. H. C. D. L. (s.f.). Ley Federal de Protección de Datos Personales en Posesión de los Particulares. Consultado el 21 de noviembre de 2023, desde <https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf>
- Vapnik, V. (1998). *Statistical Learning Theory* (Vol. 39).
- Vest, J. R., Issel, L. M., & Lee, S. (2014). Experience of Using Information Systems in Public Health Practice: Findings from a Qualitative Study. *Online Journal of Public Health Informatics*, 5(3), 227-227. <https://doi.org/10.5210/ojphi.v5i3.4847>

- Wan, J., Tang, S., Li, D., Imran, M., Zhang, C., Liu, C., & Pang, Z. (2019). Reconfigurable smart factory for drug packing in healthcare industry 4.0. *IEEE Transactions on Industrial Informatics*, 15(1), 507-516. <https://doi.org/10.1109/TII.2018.2843811>
- Wasem, D. J. (s.f.). Mining of Massive Datasets. <https://api.semanticscholar.org/CorpusID:155331592>
- Wasswa, W., Obungoloch, J., Basaza-Ejiri, A. H., & Ware, A. (2019). Automated Segmentation of Nucleus, Cytoplasm and Background of Cervical Cells from Pap-smear Images using a Trainable Pixel Level Classifier. *Proceedings - Applied Imagery Pattern Recognition Workshop, 2019-Octob(1)*, 1-9. <https://doi.org/10.1109/aipr47015.2019.9174599>
- Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, 6, 324-342. <https://doi.org/10.1287/mnsc.6.3.324>
- Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67-82. <https://doi.org/10.1109/4235.585893>
- Xia, Y., & Lee, G. (2020). How to Return to Normalcy: Fast and Comprehensive Contact Tracing of COVID-19 through Proximity Sensing Using Mobile Devices [http://arxiv.org/abs/2004.12576]. <https://arxiv.org/abs/2004.12576v1>
- Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2019). Self-training with Noisy Student improves ImageNet classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10684-10695. <https://doi.org/10.1109/cvpr42600.2020.01070>
- Xu, S., & Chan, H. K. (2019). Forecasting medical device demand with online search queries: A big data and machine learning approach. *Procedia Manufacturing*, 39, 32-39. <https://doi.org/10.1016/j.promfg.2020.01.225>
- Zhang, M., Dong, B., & Li, Q. (2020). Deep Active Contour Network for Medical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12264 LNCS, 321-331. https://doi.org/10.1007/978-3-030-59719-1_32
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid Scene Parsing Network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua*, 6230-6239. <https://doi.org/10.1109/cvpr.2017.660>
- Zhao, J., Dai, L., Zhang, M., Yu, F., Li, M., Li, H., Wang, W., & Zhang, L. (2020). PGU-net+: progressive growing of U-net+ for automated cervical nuclei segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11977 LNCS, 51-58. https://doi.org/10.1007/978-3-030-37969-8_7
- Zhao, J., Li, Q., Li, X., Li, H., & Zhangs, L. (2019). Automated segmentation of cervical nuclei in pap smear images using deformable multi-path ensemble model. *Proceedings - International Symposium on Biomedical Imaging, 2019-April(Isbi)*, 1514-1518. <https://doi.org/10.1109/isbi.2019.8759262>

Zhao, L., Li, K., Wang, M., Yin, J., Zhu, E., Wu, C., Wang, S., & Zhu, C. (2016). Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search MRF. *Computers in Biology and Medicine*, 71, 46-56. <https://doi.org/10.1016/j.combiomed.2016.01.025>

PRODUCTOS ACADÉMICOS

Conferencia Arbitrada

- **Nombre del congreso:** Congreso Internacional de Logística y Cadena de Suministro.
- **Título:** *Intelligent Framework for Logistics 4.0: an approach for Healthcare.*
- **Fecha:** 7 al 9 de octubre de 2020.
- **Lugar:** San Luis Posotí.
- **Autores:** Marco Julio del Moral Argumedo, Alberto Alfonso Aguilar Lasserre, Carlos Alberto Ochoa Ortiz.



Congreso internacional

- **Nombre del congreso:** 20th International Conference on Hybrid Intelligent Systems (HIS 2020).
- **Título:** *Intelligent mobile application to determine the existence of medicines associated with COVID-19 in a social network of medical dispensaries in a Smart City.*
- **Fecha:** 14 al 15 de diciembre de 2020.
- **Lugar:** Virtual.
- **Autores:** Marco Julio del Moral Argumedo, Alberto Alfonso Aguilar Lasserre, Carlos Alberto Ochoa Zezzati (Del Moral et al., 2022).
- **Serie:** Studies in Systems, Decision and Control 347.
- **Procedimientos:** Technological and Industrial Applications Associated With Industry 4.0.
- **ISBN:** 978-3-030-68662-8.
- **Editorial:** Springer.



Revista científica arbitrada JCR

- **Journal:** Biomedical Signal Processing and Control Q1 (Ciencias Computacionales)
Factor de Impacto 5.367.
- **Título:** *A Deep Learning Approach for Automated Cytoplasm and Nuclei Cervical Segmentation.*
- **Autores:** Marco Julio del Moral Argumedo, Carlos Alberto Ochoa Zezzati, Rubén Posada Gómez, Ulises Juárez Martínez, Alberto Alfonso Aguilar Lasserre.
- **Estado:** Publicado.



- **Journal:** Biomedical Signal Processing and Control Factor de Impacto 5.076 Q1.
- **Título:** *A Machine Learning Approach for Severe Dengue Detection.*
- **Autores:** Marco Julio del Moral Argumedo, Carlos Vazquez, Alberto Alfonso Aguilar Lasserre.
- **Estado:** En revisión.

Registro de derechos de autor

- Sistema De Diagnostico Asistido Por Computadora Basado En Deep Learning Para La Clasificación Y Detección De Células De Cancer Cervicouterino En Un Examen De Papanicolau.
- Sistema De Información Geográfica Integrado A Una Plataforma Web Para Asistir En La Toma De Decisiones Epidemiológicas Con Respecto A Enfermedades Transmitidas Por El Vector Aedes Aegypti Y Su Relación Con El Riesgo Obstétrico.
- Sistema De Ayuda A Las Decisiones Clínicas (SADC) Para El Diagnóstico Y Control De Pacientes Con Diabetes Mellitus Tipo 2.

Estancia internacional de investigación

- **Lugar:** Toulouse, Francia.
- **Institución:** Escuela Nacional de Ingenieros en Artes Químicas y Tecnológicas.
- **Investigador responsable:** Catherine Azzaro-Pantel.
- **Proyecto:** Análisis de ciclo de vida del Paracetamol usando software libre.



ANEXO A: Recopilación de datos

En los artículos explorados, se encontraron los enlaces correctos de las dependencias y entidades gubernamentales que concentran datos pertinentes para la predicción de casos de dengue y que pueden fungir como base para el desarrollo de aplicaciones informáticas para la salud bajo el paradigma de industria 4.0 Vale la pena explicar los siguientes criterios de análisis: El estado de automatización se refiere a que si el url puede ser automatizado de manera sencilla, para cargar los datos que contiene en una base de datos para almacenar la información y proceder a realizar el análisis, de forma algorítmica sin intervención humana, gestionada por el Sistema de Información en Salud.

El proceso de recopilación de datos incluyó la identificación de datos importantes relacionados con el caso de estudio. Posteriormente, se buscaron estos recursos en internet y se identificó su disponibilidad para ser integrados dentro de este estudio.

Algunos procesos fueron sencillos, la lectura directa de los datos desde el archivo de origen; la mayoría requirió mejoras en la representación de caracteres especiales encontrados en el español. Otros requirieron la creación de metodologías complejas de extracción de información como el Reconocimiento Óptico de Caracteres (ROC) y en otras instancias, la descarga y obtención se realizó de manera semiautomática.

Cabe mencionar que los procesos de descarga, preprocesamiento y carga de información se describen *grosso modo* para no extender demasiado este documento. Pero cada proceso cuenta con sus respectivos algoritmos, funciones y código programado que automatiza estos tres procesos. Python es un lenguaje sumamente versátil y cuenta con las herramientas para realizar estos procesos de manera eficaz. Siendo las funciones programadas fácilmente transportables a otras aplicaciones, como la creación del Sistema web; actualmente, este código tiene más de 500 líneas. Lo más importante, es que los datos aquí expuestos se encuentran disponibles para ser operados en conjunto a cualquier otro caso de estudio, el cual podría ser otra enfermedad transmitida por el mismo vector u otros tipos de enfermedades como cáncer o diabetes.

A.1 Datos epidemiológicos: SINAVE

Los datos surgen de la plataforma de vigilancia epidemiológica de la Secretaría de Salud SINAVE (Sistema Nacional de Vigilancia Epidemiológica), concentrándose en el periodo 2017 a 2018. La búsqueda no se realizó personalmente, sino que nos apoyamos del IMSS-Orizaba, por lo que los parámetros utilizados para la búsqueda no están del todo claros.

En las **Figura 80** se observa una captura de pantalla del sitio web de la plataforma SINAVE. Se destaca el aviso frontal que menciona las restricciones de uso, por lo que se recurrió a colaborar con el IMSS para obtener acceso a una muestra de los datos contenidos en la plataforma. Así mismo, se puede apreciar en la interfaz del sitio la variedad de conjuntos de datos que contiene la plataforma, lamentablemente, solamente se tuvo acceso a los datos de Enfermedades Transmitidas por Vector y, en particular, casos de Fiebre del Dengue. Hay que tener en cuenta que, debido a las características de ciberseguridad de la plataforma, no se tuvo acceso directo a la plataforma y el filtrado de los conjuntos de datos y su descarga fueron hechos por otros colaboradores, es por ello que se tiene poco conocimiento de la operación de la plataforma y poco control en la cantidad y calidad de los datos a los que se tiene acceso (**Figura 81**).



Figura 80: Captura de pantalla de la plataforma SINAVE

SINAVE
▪ Institución de origen: Secretaría de salud, Dirección General de Epidemiología. ▪ URL: https://www.sinave.gob.mx/. ▪ Dato público: No. ▪ Resolución espacial: Punto – coordenada geográfica del caso. ▪ Resolución temporal: Por día. ▪ Variables: Múltiples (424 columnas). ▪ Automatización: Si*. ▪ Metadata: No. ▪ Extensión: .CSV. ▪ Formato: tabular. ▪ Descarga: Enlace directo.

Figura 81: Ficha informativa: SINAVE

Los datos de la plataforma SINAVE se descargaron directamente del sitio web. En total se recibieron cuatro archivos a los que se realizó un análisis para determinar sus características y

estructura.

Primero se leyó un archivo apartado para la exploración básica de la estructura de los datos con objetivo de obtener el significado de las columnas y valores importantes, como las claves de diagnóstico. La cantidad de columnas que contiene cada archivo es enorme, y enumerar cada una de ellas y describirla sale del enfoque de este documento; sin embargo, se tienen bien identificadas para ser utilizadas en otras iteraciones de solución. Los análisis se han centrado en un subconjunto específico de estas columnas, relacionadas con la posición geográfica y los diagnósticos de cada paciente.

Los archivos son un tipo de archivo de texto denominado *valores separados por delimitador*, que sirven para almacenar datos tabulares. Los tipos más comunes de delimitador son la coma (para archivos con extensión *.CSV*) o un tabulador (archivos *.TSV*). Los archivos descargados utilizan un delimitador especial, el símbolo: |. Así mismo, la codificación del archivo no es UTF-8 por lo que se tuvo que adecuar el proceso de lectura a estos criterios. Como se verá posteriormente, este tipo de criterios especiales y no estandarizados plagan las fuentes de datos gubernamentales.

De la lista de las columnas, nos centraremos en dos para determinar los diagnósticos finales emitidos para cada caso. La columna *DES_DIAG_FINAL* y *CVE_DIAG_FINAL* muestran la descripción del diagnóstico y la clave numérica que lo representa. Posteriormente, se determinó la variedad de diagnósticos y sus claves que se encuentran en el archivo. Los diagnósticos finales encontrados fueron: Dengue no grave, Dengue con signos de alarma, Dengue grave, Enfermedad por virus Zika, Fiebre Chikungunya y otros.

Se procedió a leer todos los archivos que contienen los datos y se unieron en un solo elemento de 40901 filas; posteriormente, al eliminar filas repetidas, nos quedamos con 17365 elementos. De los datos compilados, se procede a separar aquellos cuyo diagnóstico final haya sido dengue, en cualquier de sus variedades. Si bien el mosquito que trasmite el dengue *Aedes Aegypti* puede transmitir otras enfermedades como Zika o Chikungunya, se limitó el estudio al dengue, por lo que se quedan 3382 elementos.

La **Tabla 26** ofrece una muestra de los datos contenidos dentro de la plataforma SINAVE. Se

pueden observar las columnas con el ID o Folio, el diagnóstico final, el estado de residencia y la fecha de solicitud de atención.

Tabla 26: Muestra de datos de plataforma SINAVE

FOL_ID	DES_DIAG_FINAL	DES_EDO_RES	FEC_SOL_ATEN
30039SIOL600101	OTROS	VERACRUZ	2017-03-02
171PAFO840405	OTROS	VERACRUZ	2018-07-06
13028BARJ891103	DENGUE GRAVE	VERACRUZ	2017-08-07
223GASM970413	OTROS	VERACRUZ	2017-07-10
30044VEVL110203	DENGUE NO GRAVE	VERACRUZ	2018-10-10

A.2 Datos epidemiológicos: DGE

La Dirección General de Epidemiología (DGE) es fuente de tres conjuntos de datos integrados dentro del sistema.

El primero, gracias a un decreto publicado en el diario Oficial de la Federación, el día 20 de febrero del año de 2015 se pudieron poner a disposición de la ciudadanía, datos de casos asociados a las Enfermedades Transmitidas por Vector (Dengue) a partir del 03 de diciembre de 2020. Estos datos se encuentran publicados en función a la semana epidemiológica a la que corresponden y se encuentra en formato *.CSV*. La URL de los datos se encuentra bien formada por lo que su automatización fue sencilla. Cabe mencionar que los datos se encuentran anonimizados y no contienen información que permita relacionar el dato público a su caso en particular. Esto es interesante, ya que el respeto al paciente ha sido uno de los ejes principales de esta investigación y lo expuesto por las autoridades gubernamentales concuerda perfectamente con las intenciones finales de este trabajo. El análisis de estos datos se encuentra pendiente, puesto que el periodo de tiempo que abarcan es sumamente corto y se tienen mejores conjuntos de datos para realizar un estudio significativo.

El segundo, el boletín epidemiológico corresponde, a los reportes oficiales que hace la Secretaría de Salud cada semana para distintas enfermedades de interés. Su emisión ha sido desde 1981 a la fecha, lamentablemente, los datos presentados vienen en formato *.PDF*, siendo los anteriores al 2000 fotografías de documentos físicos, lo cual hizo bastante difícil su captura

y procesamiento, ya que muchos de estos documentos se encuentran en formatos dispares o simplemente están bastante deteriorados por lo que el algoritmo de reconocimiento de caracteres tuvo problemas para poner extraer la información de manera efectiva (Figura 82).

Dirección General de Epidemiología

- **Institución de origen:** Dirección General de Epidemiología, Secretaría de Salud.
- **URL:** <https://www.gob.mx/salud/documentos/datos-abiertos-152127>.
- **Dato público:** Si.
- **Resolución espacial:** Municipal, Estatal.
- **Resolución temporal:** Diaria, Mensual.
- **Variables:** Fecha, datos generales del caso, enfermedades adicionales.
- **Automatización:** Si.
- **Metadata:** Si.
- **Extensión:** *.PDF*.
- **Formato:** tabular.
- **Descarga:** Enlace directo.
- **Tamaño:** 50MB.

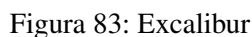
Figura 82: Ficha informativa: Anuario DGE

El tercero y último, contiene la cantidad de casos por Entidad por mes y año a partir de 1985. Estos datos sirvieron para el análisis de series de tiempo. Al combinarse con los datos de SINAVE, estos se agruparon por Entidad y por Año para tomar una forma igual a este conjunto y poder ser procesados usando el mismo código. Para poder ser usados, los datos epidemiológicos tuvieron que ser convertidos de formato *.PDF* a formato *.CSV*, capaz de ser ingresado y analizado bajo cualquier técnica de análisis matemático y estadístico.

Lamentablemente, los datos dentro del anuario epidemiológico de la secretaría de salud, se almacenan en un formato que no es el adecuado para guardar datos tabulares y que dificulta su análisis. También, el formato general de cada archivo va variando en cada año, por lo que se encuentran distintos formatos para presentar los mismos datos, lo que hace bastante difícil la implementación de un algoritmo totalmente automatizado para su extracción.

Para este paso, se utilizó un procedimiento semi-automatizado, haciendo uso de herramientas

Workspace - casos_1985.pdf



Uno de los principales análisis epidemiológicos que se pueden realizar es la cantidad de casos por división geográfica y por unidad de tiempo. Existen distintas maneras de realizar el análisis por división, el primero y más sencillo fue contabilizar los casos y agruparlos por las columnas que representen la división geográfica dentro de la tabla de datos de dengue. Sin embargo, este sistema se limita a las resoluciones espaciales de Entidad, Municipio y Localidad. Si bien esta

resolución es capaz de generar conocimiento pertinente, existen otras divisiones geográficas más precisas que pueden modelar el comportamiento de los casos dentro de la cadena logística de mejor manera, ofreciendo mejor información y conocimiento.

Otro problema con este enfoque, es la imposibilidad de determinar la relación geoespacial entre el punto del caso de dengue y el polígono que representa división geográfica, al menos de manera directa. Atributos del polígono como el área y el perímetro tuvieron que ser precalculados y agregados a la tabla de casos de dengue para poder determinar densidad de casos por Municipio, por poner un ejemplo (Figura 84).

El Marco Geoestadístico Nacional, según INEGI, es:

... un sistema único y de carácter nacional diseñado por el INEGI, el cual presenta la división del territorio nacional en diferentes niveles de desagregación para referir geográficamente la información estadística de los censos y encuestas institucionales y de las Unidades del Estado, que se integra al Sistema Nacional de Información Estadística y Geográfica (SNIEG).

Marco Geoestadístico Nacional

- **Institución de origen:** Instituto Nacional de Estadística y Geografía.
- **URL:** <https://www.inegi.org.mx/temas/mg/>.
- **Dato público:** Si.
- **Resolución espacial:** Polígono - División geográfica del país.
- **Resolución temporal:** Periodos mensuales.
- **Variables:** Múltiples (1 – 8).
- **Automatización:** Si.
- **Metadata:** Si.
- **Extensión:** .SHP.
- **Formato:** tabular.
- **Descarga:** Enlace directo.
- **Tamaño:** 11.9GB.

Figura 84: Ficha informativa: Marco Geoestadístico Nacional

El Marco Geoestadístico Nacional, es la compilación de las características geográficas nacionales realizada por el INEGI. Compila las divisiones geográficas del territorio nacional, entre otras características como vías pluviales o de comunicación. Con respecto a la división política del territorio nacional, el Marco Geoestadístico contiene los datos geospaciales, representados por polígonos, de todas las subdivisiones que tiene el país:

1. Entidades: Polígonos que representan las entidades federativas que componen al país, se identifican por un código numérico de dos dígitos.
2. Municipios: Polígonos que representan la división geográfica y administrativa de las entidades.
3. Localidades: Polígonos que representan asentamientos humanos dentro de los municipios y que reciben nombres particulares, pueden ser rurales y urbanas.
4. AGEB: Polígonos de las Áreas Geoestadísticas Básicas, pueden ser urbanas o rurales.
5. Manzana: Polígonos que representan las manzanas dentro de urbanizaciones.

Otros elementos contenidos dentro del Marco Geoestadístico son los territorios insulares, ejes de vialidad, servicios así como polígonos y frentes de manzana.

El Marco Geoestadístico Nacional, utiliza la siguiente proyección cartográfica:

- Cónica Conforme de Lambert (CCL)
- Dátum: ITRF2008.
- Elipsoide: GRS80.
- Meridiano origen: 102° 00' 00" W.
- Latitud origen: 12° 00' 00" N.
- Primer paralelo estándar: 17° 30' N.
- Segundo paralelo estándar: 29° 30' N.

- Falso Este: 2 500 000.
- Falso Norte: 0

Los datos se encuentran almacenados en formato *.SHP* el cual es propietario. Primero los archivos fueron convertidos a formato abierto *GeoJSON*. Se recodificaron los archivos para mejorar la interpretación de acentos y caracteres especiales.

Para poder visualizar correctamente, primero se proyectó al protocolo EPSG:4326 que es una representación geodésica de la superficie terrestre para posteriormente re-proyectar al protocolo EPSG:3857, que es una proyección de tipo Mercator. Esto se realizó debido a que EPSG:4326 maneja distancias en grados, mientras que EPSG:3857 los maneja en metros. La proyección inicial cónica de Lambert conserva correctamente las proporciones geográficas, pero deforma las áreas de los elementos geográficos.

Al estar los datos en una proyección tipo Mercator, podremos calcular el área de los polígonos de manera sencilla. Dividiendo por 10^6 para poner las unidades en Kilómetros cuadrados. Utilizando un mapa de calor, podemos visualizar el área de los municipios ([Figura 85](#)).

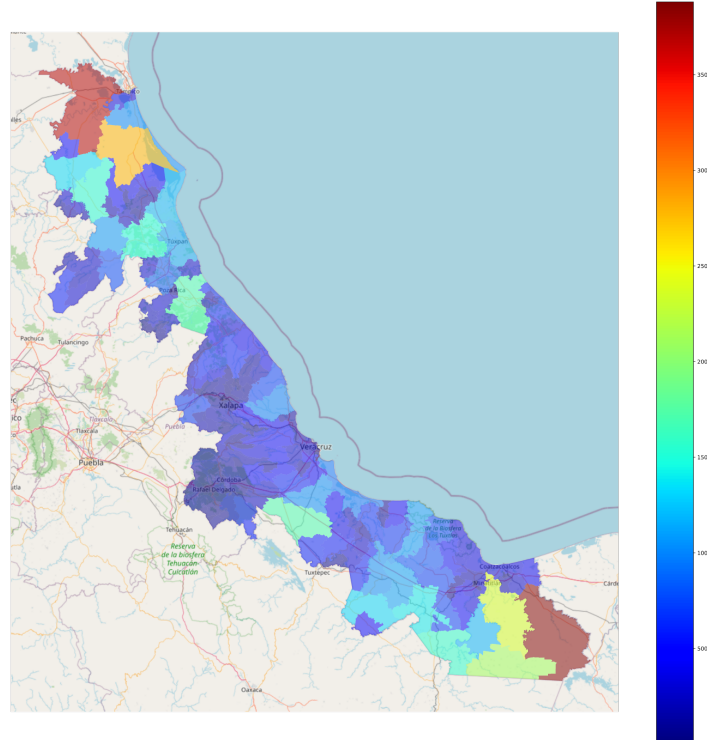


Figura 85: Área por municipio

Una parte importante del Margo Geoestadístico es la creación de un identificador único para cada dato. La clave geográfica se crea uniendo la sucesión numérica de la entidad, municipio, localidad, etcétera del objeto geográfico en cuestión.

A.4 Datos socio-demográficos: INEGI

Otra fuente de datos es el Censo Poblacional de 2020, el cual compila cantidad de habitantes por entidad, municipio, localidad, AGEB y manzana. El censo también segmenta la información poblacional en estratos como sexo o edad.

Lamentablemente, el censo solamente compila información sobre poblaciones urbanas, por lo que no contaremos con datos poblacionales para AGEB rurales. Otro indicio de que la falta de infraestructura en las áreas rurales hace difícil que puedan ser apoyadas por la ciencia de datos o la industria 4.0, al menos de manera directa. De las localidades rurales, se tiene información poblacional incompleta (Figura 86). Según INEGI el censo es:

El Censo de Población y Vivienda 2020 (Censo 2020) se realizó del 2 al 27 de marzo; en este participaron poco más de 147 mil entrevistadoras y entrevistadores que recorrieron los cerca de dos millones de kilómetros cuadrados del territorio nacional, visitando cada una de las viviendas para obtener información sobre estas, contar a la población que vive en México e indagar sobre sus principales características demográficas, socioeconómicas y culturales.

Marco Geoestadístico Nacional

- **Institución de origen:** Instituto Nacional de Estadística y Geografía.
- **URL:** <https://www.inegi.org.mx/programas/ccpv/2020/default.html>.
- **Dato público:** Si.
- **Resolución espacial:** Entidad, Municipio, Localidad, Ageb y Manzana.
- **Resolución temporal:** Decenio.
- **Variables:** Múltiples (222).
- **Automatización:** Si.
- **Metadata:** Si.
- **Extensión:** .CSV.
- **Formato:** tabular.
- **Descarga:** Enlace directo.
- **Tamaño:** 1.16GB.

Figura 86: Ficha informativa: Censo 2020

El Censo contiene más de 200 indicadores o columnas medidos a una resolución por manzana. Estos indicadores miden la cantidad de población dividida por distintos estratos como la edad, discapacidades, lenguas habladas. También contiene datos sobre las características de las viviendas que están en la manzana, como cantidad de dormitorios, materiales de construcción, cantidad de dispositivos electrónicos y acceso a internet. En el último censo, se agregaron indicadores tales como el uso de teléfono inteligente, computadora y consolas de videojuegos. Esto sirve para generar criterios socio-demográficos a partir de estos indicadores.

Cruzando los datos geográficos y demográficos, se construyó una superficie nacional virtual en la cual se tiene la descripción poblacional de cada manzana de la república. Ofreciendo una

herramienta poderosa para operarse con diversas aplicaciones de epidemiológicas y de salud.

A.5 Datos climatológicos: CONAGUA

El sitio web de la Comisión Nacional del Agua (CONAGUA), aparte de información relacionada con la dependencia gubernamental misma, presenta enlaces para acceder a distintas fuentes de datos climatológicos como históricos, datos presentes, pronósticos y sobre el cambio climático.

Los datos proporcionados por CONAGUA que se utilizaron para esta etapa del diseño de algoritmos predictivos, consisten en información climatológica sobre temperatura y precipitación para cada una de las estaciones pertenecientes al Sistema Meteorológico Nacional, dadas en una resolución temporal de un día y promedios por mes, que se conforman por máximos, mínimos y medias así como precipitación acumulada mensual. Estos datos posteriormente son utilizados, entre otras cosas, para construir modelos de interpolación geoespacial basados en *Machine Learning*.

Lamentablemente, el sitio web no contiene información de algún servicio web que exponga una API para extraer los datos de manera programática para facilitar la extracción sencilla de los mismos. Como se puede observar en la **Figura 87**, el sitio de CONAGUA muestra dos funcionalidades dentro de la misma página web. La primera arriba a la izquierda, es la descarga de datos por estación por año y por mes, utilizando un slider para generar mapas específicos para los datos consultados y descargados, estos mapas son generados dinámicamente mediante interpolación sencilla. La segunda función, arriba a la derecha, es la descarga de datos promedio por año por entidad.

Resúmenes Mensuales de Temperaturas y Lluvia

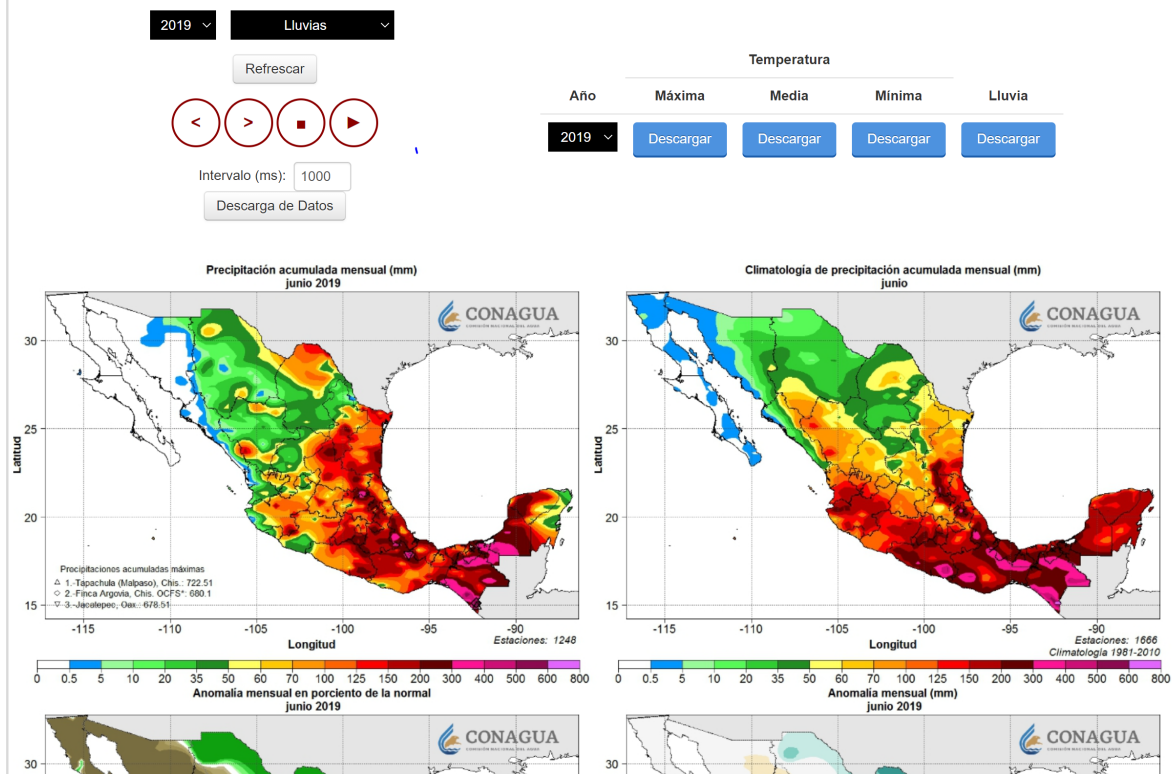


Figura 87: Captura del sitio

El primer problema que se encontró fue que, para los datos mensuales por estación, solamente permite descargar a partir del año 2018. Otro problema es la forma en la que está programada y diseñada la página web, ya que oculta la llamada al URL de descarga de los archivos .csv de los datos detrás de una llamada a función en lenguaje JavaScript. Haciendo una exploración del código fuente de la página y un análisis de su funcionamiento, se pudo encontrar como se construye esta URL de carga para poder automatizar el proceso. Para los datos promedio por mes y año, encontramos el URL expuesta en la barra de estado por lo cual se pudo deducir fácilmente como se construye la misma para automatizar la descarga; estos datos están en formato PDF por lo cual se requieren procesar para ser utilizados. La **Figura 88** muestra el punto exacto del código donde se pudo extraer el URL para la descarga, así como la forma de construir tal URL para realizar la consulta del dato. Cabe mencionar que el código fuente de la página tiene más de 7000 líneas de código. Este proceso fue automatizado utilizando Python, la actividad de obtener información mediante el análisis del código fuente de una página se le

Datos climatológicos

- **Institución de origen:** Comisión Nacional del Agua.
- **URL:** <https://smn.conagua.gob.mx/es/climatologia/temperaturas-y-lluvias/resumenes-mensuales-de-temperaturas-y-lluvias>.
- **Dato público:** Si.
- **Resolución espacial:** Puntual - Por estación, Estatal.
- **Resolución temporal:** Múltiple.
- **Variables:** Máximo, mínimo, promedio.
- **Automatización:** Si.
- **Metadata:** Si.
- **Extensión:** .CSV, .PDF.
- **Formato:** tabular.
- **Descarga:** Webscrapping.
- **Tamaño:** 800MB.

Figura 89: Ficha informativa: CONAGUA

ANEXO B: Desglose de revisión del estado del arte

B.1 Selección de artículos

La **Tabla 27** concentra los títulos de los artículos analizados, comenzando por el ID utilizado para la rápida identificación de los mismos, el título, la revista y su editorial. La mayoría de los artículos pertenecen a Editorial *Elsevier*, debido a que su buscador fue uno de los usados, sin embargo, la editorial, como era de esperarse, también apareció en los resultados de otros buscadores. El journal que más apareció en la búsqueda fue *Procedia Computer Science*. Con respecto a las editoriales, destaca la aparición de *The Lancet* y *Nature*, ambas son de las editoriales con mayor prestigio e impacto a nivel global.

Tabla 27: Tabla de artículos revisados

ID	Título	Journal	Editorial
1A	The perception of health providers about an artificial intelligence applied to Tuberculosis video-based treatment in Brazil: a protocol proposal (Dos Santos et al., 2019)	Procedia Computer Science	Elsevier
2A	Impact of industry 4.0 to create advancements in orthopaedics (Javaid & Haleem, 2020)	Journal of Clinical Orthopaedics and Trauma	Elsevier
3A	Blockchain-based electronic healthcare record system for healthcare 4.0 applications (Tawar et al., 2020b)	Journal of Information Security and Applications	Elsevier
4A	Securing electronic healthcare records: A mobile-based biometric authentication approach (Hathaliya et al., 2020b)	Journal of Information Security and Applications	Elsevier
5A	FOG Assisted Healthcare Architecture for Pre-Operative Support to Reduce Latency (Tiwari et al., 2020)	Procedia Computer Science	Elsevier
6A	Multichannel Interaction for Healthcare Intelligent Decision Support (Moreira & Santos, 2020)	Procedia Computer Science	Elsevier
7A	Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial (H. Lin et al., 2019b)	Ecclinical Medicine	The Lancet
8A	Machine Learning Approaches in Smart Health (Rayan et al., 2018)	Procedia Computer Science	Elsevier
9A	Diabetes Prediction using Machine Learning Algorithms (Mujumdar & Vaidehi, 2019b)	Procedia Computer Science	Elsevier
10A	An Integrated Approach of Machine Learning and Systems Thinking for Waiting Time Prediction in an Emergency Department (Kuo et al., 2020)	International Journal of Medical Informatics	Elsevier
11A	Forecasting Medical Device Demand with Online Search Queries: A Big Data and Machine Learning Approach (Xu & Chan, 2019)	Procedia Manufacturing	Elsevier
12A	Research on Key Technologies of Regional Synergy Emergency System Based on Medical Data Center (Du, 2018b)	Procedia Computer Science	Elsevier
13A	Forecasting of weekly patient visits to emergency department: real case study (Khalidi et al., 2019a)	Procedia Computer Science	Elsevier
14A	Employing Virtual Reality Techniques in environment adaptation for autistic children (Sait et al., 2019)	Procedia Computer Science	Elsevier
1B	Fractal Techniques in Health Information Systems Management: Can They Be Used? (Ponaru & Dobrescu, 2019)	International Conference on Control Systems and Computer Science	IEEE
2B	Hierarchical data fusion for Smart Healthcare (Dautov et al., 2019)	Journal of Big Data volume	Springer
3B	Salp Chain-Based Optimization of Support Vector Machines and Feature Weighting for Medical Diagnostic Information Systems (Al-Zoubi et al., 2020)	Evolutionary Machine Learning Techniques	Springer
4B	Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges (Thesmar et al., 2019b)	PharmacoEconomics	Springer
5B	Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis (Knaus & Marks, 2019b)	JAMA Sepsis Website	JAMA Network Open
6B	Diagnostics 4.0: the medical laboratory in digital health (Neumaier, 2019)	Clinical Chemistry and Laboratory Medicine	De Gruyter
7B	Cloud-Based Artificial Intelligence System for Large-Scale Arrhythmia Screening (C. H. Tseng et al., 2019)	DIGITAL HEALTH	IEEE
8B	Clinical Documentation and Patient Care Using Artificial Intelligence in Radiation Oncology (Luh et al., 2019)	Journal of the American College of Radiology	Elsevier
9B	Clustering Approach Using Artificial Bee Colony Algorithm for Healthcare Waste Disposal Facility Location Problem (Gergin et al., 2018)	International Journal of Operations Research and Information Systems (IJORIS)	IGI Global
10B	Health Care Chatbots Are Here to Help (A. Nelson et al., 2019)	IEEE Pulse	IEEE
11B	Predicting scheduled hospital attendance with artificial intelligence (Bates, 2019)	Digital Medicine	Nature
12B	An Edge-Based Architecture to Support Efficient Applications for Healthcare Industry 4.0 (Pace et al., 2019)	Transactions on Industrial Informatics	IEEE
1C	Reconfigurable Smart Factory for Drug Packing in Healthcare Industry 4.0 (Wan et al., 2019)	Transactions on Industrial Informatics	IEEE
2C	The Healthcare Sustainable Supply Chain 4.0: The Circular Economy Transition Conceptual Framework with the Corporate Social Responsibility Mirror (Daí et al., 2019)	Sustainability	MDPI
3C	Introducing Care 4.0: An Integrated Care Paradigm Built on Industry 4.0 Capabilities (Chute & French, 2019b)	International Journal of Environmental Research and Public Health	MDPI
4C	Towards a GDPR compliant way to secure European cross border Healthcare Industry 4.0 (Larrucea et al., 2020b)	Computer Standards & Interfaces	Elsevier
5C	Healthy Operator 4.0: A Human Cyber-Physical System Architecture for Smart Workplaces (Sun et al., 2020)	Sensors	MDPI
6C	Towards Health 4.0: e-Hospital Proposal Based Industry 4.0 and Artificial Intelligence Concepts (Cáceres et al., 2019)	Artificial Intelligence in Medicine in Europe	Springer

B.2 Descripción de la metodología para revisión sistemática de literatura

Para la elaboración del Estado del Arte, se realizó una búsqueda de artículos, basada en el concepto de *keywords* o *palabras claves*. Estas palabras pretenden encapsular tópicos que representan las áreas del conocimiento o de la técnica que se involucran en determinado trabajo, después, las palabras clave se transforman en los ejes de la investigación, para encontrar la forma en que el Estado de Arte se relaciona con la investigación aquí descrita. Se procedió a abordar la tarea mediante un método sistemático, inspirado en metodologías para *Reviews* y meta-análisis en medicina.

La metodología seguida para la realización del análisis sistemático, fue inspirada en la metodología PRISMA. Esta metodología fue creada para mejorar una metodología anterior llamada QUORUM, diseñada en 1996 para estandarizar y mejorar el reporte de meta análisis y reviews sistemáticos en ensayos clínicos. PRISMA propone cuatro fases: Identificación, Búsqueda, Elegibilidad e Inclusión; cada una con sus respectivos pasos, para este capítulo se escogió un subconjunto de los mismos. Cabe mencionar que PRISMA se enfoca exclusivamente en el acto de búsqueda de artículos, pero no sugiere la forma de realizar esta búsqueda, es por ello que a continuación se desglosan los pasos generales (Moher et al., 2009) (Clarke, 2000):

1. Determinar preguntas de investigación.
2. Extraer palabras clave y ejes de la investigación.
3. Elección de buscadores pertinentes.
4. Búsqueda secuencial con *keywords*.
5. Aplicación de filtros.
6. Implementación de criterios de elegibilidad.
7. Selección de artículos.

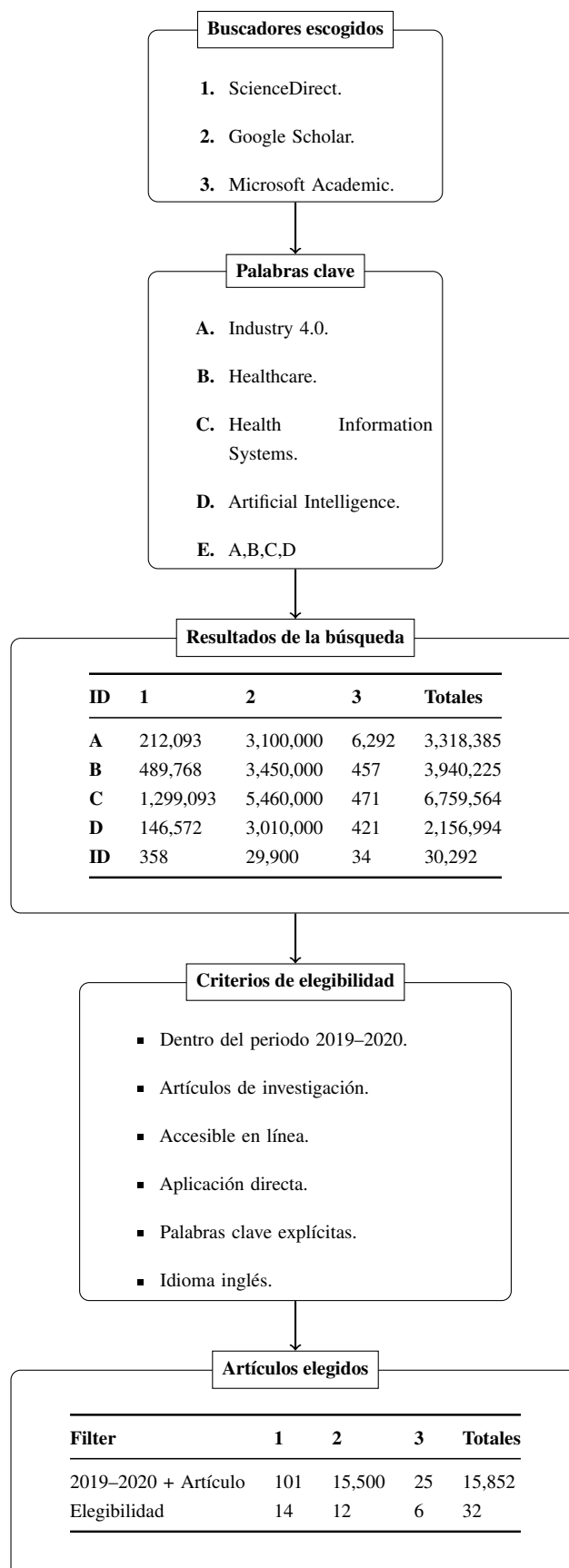


Figura 90: Diagrama del modelo prisma

B.2.1 Determinar preguntas de investigación

Las preguntas de investigación sirven como base para la búsqueda, ya que permiten guiar la búsqueda con el objetivo final de producir información útil sobre el estado de la literatura y su aplicación de la misma al tema analizado. Inicialmente, la búsqueda se había enfocado a conceptos de Logística 4.0 en lugar de Industria 4.0, lamentablemente, la búsqueda no arrojaba resultados adecuados o congruentes con los criterios de elegibilidad.

- ¿Cuáles componentes de la *Industria 4.0* se han aplicado en el área de salud?: Esta pregunta devuelve una respuesta tanto cualitativa como cuantitativa a la mención de cada componente de industria 4.0 (IoT, Big Data, etc.) en cada artículo analizado.
- ¿Cuáles áreas de la logística de sanidad han sido tocadas por la *Industria 4.0*?: Las áreas de la logística de sanidad son cualquier eslabón de la cadena logística, pueden ser conceptos específicos como Patología o conceptos generales como Administración o Trazabilidad.
- ¿Cuáles implementan o mencionan Sistemas de Información en Salud?: Esta pregunta evalúa la relación entre los componentes de Industria 4.0 y su relación con los Sistemas de Información en salud.
- ¿Cuáles técnicas de IA, Toma de Decisiones, Aprendizaje automático o Metaheurísticas se han implementado?: Esta pregunta también devuelve ambas respuestas (cualitativa y cuantitativa), enfocándose tanto en algoritmos modernos como redes neuronales artificiales o técnicas tradicionales de toma de decisiones como simulación o TOPSIS.
- ¿Alguno aplica, menciona o se relaciona con sanidad pública?: Este es un criterio binario, existe mención o no existe.

B.2.2 Extraer palabras clave y ejes de la investigación

Las palabras claves extraídas de las preguntas de investigación y sus criterios de presencia se muestran en la **Tabla 28**. Estas palabras fueron posteriormente derivadas en cuatro ejes de comparativa, que servirán como conjuntos para evaluar la posición de los artículos, estos

ejes son: Industria 4.0, Toma de Decisiones, Inteligencia Artificial, Sistemas de Información en Salud y Sanidad Pública. De las palabras clave, solamente Health Information System representa un concepto monolítico, es decir, que no puede ser separado en partes funcionales individuales, es por ello que su criterio se basa en la mera mención; las otras palabras clave se refieren a conceptos modulares que pueden ser separados y operar individualmente.

Tabla 28: Palabras clave utilizadas

Palabra	Criterio
Industry 4.0	Mención explícita o componente
Healthcare	Mención explícita o componente
Artificial Intelligence	Mención explícita o componente
Health Information System	Mención explícita

B.2.3 Elección de buscadores pertinentes

Para la búsqueda de los artículos, se escogieron tres buscadores de prestigio: *ScienceDirect* (A), *GoogleScholar* (B) y *MicrosoftAcademic* (C). Inicialmente, se incluyó el buscador *Core*, que concentra la mayor cantidad de artículos de acceso abierto, sin embargo, los resultados del proceso de filtrado arrojaron más de 4 millones de resultados, bastantes en otros idiomas diferentes al inglés o español, por lo que se optó solamente por los tres buscadores antes mencionados. (ScienceDirect, 2021) (Microsoft, s.f.) (Google, s.f.) (CORE, s.f.)

B.2.4 Búsqueda secuencial con *keywords*

El filtrado secuencial de palabras clave consiste en evaluar una por una cada palabra clave por cada buscador, capturando la cantidad de resultados que este devuelve y registrándose para tener una métrica somera del número de artículos relacionados por cada palabra clave. Esta búsqueda es realizada sin filtro alguno. Finalmente, se hace una búsqueda utilizando todas las palabras clave como entrada, los resultados de esta búsqueda representan la cantidad de artículos que tienen todas (*AND*) o una (*OR*) de las palabras clave. Los resultados se muestran en la **Tabla 29**:

Tabla 29: Tabla de resultados de búsqueda secuencial

ID	Keyword	ScienceDirect	GoogleScholar	MicrosoftAcademic	Totales
A	Industry 4.0	212093	3100000	6292	3318385
B	Healthcare	489768	3450000	457	3940225
C	Artificial Intelligence	1299093	5460000	471	6759564
D	Health Information Systems	146572	2010000	421	2156993
E	A,B,C,D	358	29900	34	30292

De esta tabla obtenemos las siguientes observaciones: De los buscadores, *Google Scholar* es el que más resultados arroja mientras que *Microsoft Academic* es el más limitado; ambos buscadores arrojan resultados de cualquier editorial que esté dada de alta en ese motor de búsqueda, mientras que *ScienceDirect* se concentra en resultados de la editorial *Elsevier*. De los tres, *Google Scholar* es el más flexible en sus criterios de búsqueda e inclusión. De las palabras clave, Inteligencia Artificial presenta la mayor cantidad de menciones, sumadas entre los buscadores; los sistemas de información en salud fueron los menos mencionados.

B.2.5 Aplicación de filtros

Estos artículos fueron filtrados por fecha, especificando el periodo 2019-2020 y, en donde se permitía, también se filtraron los resultados por tipo: *Research Article*. Los resultados los podemos ver en la [Tabla 30](#), se puede observar que GoogleScholar no cuenta con filtro específico para artículos de investigación. Posteriormente, se seleccionaron los artículos finales después de una lectura exploratoria.

Tabla 30: Tabla de búsqueda con filtros

Keyword	ScienceDirect	GoogleScholar	MicrosoftAcademic
Filtros	2019-2020, RA	2019-2020	2019-2020, RA
E	101	15500	250
Total	14	12	6

B.2.6 Implementación de criterios de elegibilidad

De los artículos filtrados, se hizo una selección final basada en los siguientes criterios de elegibilidad, estos criterios surgen para asegurar que los artículos filtrados están disponibles para análisis, cumplen un mínimo de características para ser relevantes y están dentro del espectro de la comunidad internacional. Un total de 32 artículos finales fueron considerados aptos para el análisis:

- Los artículos deben ser accesibles en línea.
- Deben ser aplicaciones directas.
- Mostrar explícitamente las palabras clave.
- Estar en idioma inglés.

Al aplicar los criterios de elegibilidad, nos encontramos artículos que no fueron filtrados por los pasos anteriores, artículos que si bien se encuentran indizados no están disponibles al público, otros cuentan con enlaces rotos o, aunque sus títulos están en idioma inglés, su contenido no. Otros se enfocan en evaluaciones teóricas y no muestran una explícita aplicación de las palabras clave.

Tabla 31: Tabla de artículos agrupados por eje

Conjunto	Artículos	Cantidad
Ind	2A, 3A, 4A, 5A, 6A, 7A, 14A, 2B, 4B, 6B, 7B, 10B, 12B, 1C, 2C, 3C, 4C, 5C, 6C	19
IA	1A, 7A, 8A, 9A, 10A, 11A, 12A, 13A, 3B, 4B, 5B, 7B, 8B, 9B, 10B, 11B, 6C	17
HIS	12A, 1B, 2B, 4B, 7B, 3C, 4C, 1A	8
PUB	12A, 4C	2
DM	10A, 13A, 1B, 2B, 3B, 9B, 12B, 1C, 6C, 11A	10

Después del análisis y lectura de los artículos, se procedió a contestar la preguntas de investigación. La **Tabla 32** concentra los resultados de los componentes de *Industria 4.0*, áreas de la logística de sanidad tocadas, algoritmos de *Inteligencia Artificial* así como los de Toma

de Decisiones. Podemos observar que, si bien la búsqueda se realizó tomando en cuenta *Industria 4.0*, algunos de los artículos no muestran claramente el uso de algún componente, la mayoría toma alguna técnica de IA, de Tecnologías de la Información sin adoptar plenamente el paradigma.

Tabla 32: Tabla de análisis

ID	Componente	Área	IA	DM
1A	NA	Tratamiento	CNN	NA
2A	Additive Manufacture	Ortopedia, Tratamiento, Terapia	NA	NA
3A	Blockchain	Trazabilidad, Administración	NA	NA
4A	Wearables	Administración	NA	NA
5A	IoT	Cirugía	NA	NA
6A	IoT	Administración	NA	NA
7A	Big Data	Tratamiento	CNN	NA
8A	NA	Diagnóstico, Emergencias	SVM, ANN, CNN	NA
9A	NA	Diagnóstico, Tratamiento, Trazabilidad	SVM, Misceláneo Forest, Decision Tree, Extra Tree, Ada Boost , ANN, Linear Discriminant Analysis , Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging, Gradient Boosting	NA
10A	NA	Emergencias	Linear Regression, ANN, SVM, Gradient Boosting	Systems Thinking
11A	NA	Administración	ANN	Ensemble Empirical Mode Decomposition
12A	NA	Emergencias	Fuzzy C-means	NA
13A	NA	Emergencias	ANN	ARIMA
14a	Virtual Reality	Terapia	NA	NA
1B	NA	Diagnóstico, Administración	NA	Fractal Multi Agent System
2B	IoT	Administración	NA	Complex Event Processing
3B	NA	Diagnóstico	SVM	Swarm Intelligence
4B	Big Data	Diagnóstico	No especifica	NA
5B	NA	Tratamiento	K-means	NA
6B	Digital Twins	Diagnóstico	NA	NA
7B	Big Data	Diagnóstico	DSP	NA
8B	NA	Diagnóstico, Trazabilidad	NLP	NA
9B	NA	Administración	Artificial Bee Colony	Facility Location Problem
10B	Chatbots	Administración	NLP	NA
11B	NA	Administración	Simulación	NA
12B	IoT	Prevención	NA	Time Series
1C	CPS	Administración	NA	Ontology
2C	Circular Economy	Administración	NA	NA
3C	Care 4.0	Administración	NA	NA
4C	Cybersecurity	Trazabilidad, Administración	NA	NA
5C	CPS	Prevención	NA	NA
6C	CPS	Emergencias	Simulación	Agentes

ANEXO C: Series de tiempo

C.1 Proceso de creación de algoritmo para predecir casos de Dengue y creación de prototipo

El Caso de estudio de Cáncer de Cérvix se descartó debido a que es un problema sumamente puntual, se podría expandir consiguiendo datos estadísticos sobre la distribución de la enfermedad pero lamentablemente no se han tenido acceso a este tipo de datos. Es por ello que el caso de estudio elegido fue el de Dengue, ya que cuenta con distintos tipos de datos y, debido a que la incidencia de esta enfermedad ha ido aumentando, se determinó que es excelente candidato para plasmar el prototipo basado en el Framework que se está desarrollando.

Se lograron implementar las bases de datos demográficas, geoespaciales y climatológicas, una interfaz gráfica web para visualizar y operar los datos, así como los cimientos de la arquitectura modular, implementando algoritmos de entrenamiento en microservicios externos para ser operados por la aplicación principal. Es por ello que se comenzó con el análisis y evaluación de distintos algoritmos, metodologías y arquitecturas para realizar la tarea de predicción de casos de Dengue.

Como se verá posteriormente, se probaron algoritmos tanto de Machine Learning como de Deep Learning para la tarea de predicción de series de tiempo. También se probaron distintas arquitecturas y se crearon los métodos de evaluación pertinentes para poder comparar los algoritmos de manera adecuada. Este proceso incluyó la revisión de literatura realizada en semestres anteriores así como una búsqueda más general sobre aplicaciones modernas de algoritmos aplicados a la tarea de predicción en series de tiempo (Rayan et al., 2019).

Muchas series de tiempo pueden ser muy volátiles para que se les aplique de manera arbitraria un algoritmo de pronósticos, así mismo, no existe una garantía para que un algoritmo, por más potente que sea, funcione para todos los problemas o inclusive instancias del mismo problema, esta es la razón por la cual se hicieron estos experimentos (Khaldi et al., 2019b).

C.1.1 Datos utilizados para los experimentos

Los datos utilizados fueron explorados en semestres anteriores y contienen la información sobre los casos de Fiebre del Dengue en cada uno de los Estados de la República, así como datos climatológicos; Estos datos fueron obtenidos de fuentes públicas.

Para el entrenamiento de los modelos, se utilizaron solamente datos del Estado de Veracruz, ya que es uno de los estados con mayor cantidad de casos. No se realizó un análisis a nivel país, ya que el comportamiento es sumamente dispar entre cada entidad federativa.

El rango de los datos va desde el año de 1985 hasta el 2019, de los cuales se agregaron los datos en periodicidad mensual. No se tomó el cuenta el año 2020 ya que la conversión de estos datos de formato pdf a formato tabular se realiza de manera semi automática, así que se trabajó con lo que ya se tenía.

Estos conjuntos de datos representan las agregaciones que se realizarían dentro del sistema al computar totales por aspectos geoespaciales y temporales, en este caso, por entidad federativa a intervalo mensual. Dentro del sistema este proceso se realizaría combinando los datos individuales de casos que se van almacenando en la plataforma, y el cómputo de los agregados se realizaría dentro de los Data Warehouses diseñados para la Minería de Datos y Conocimiento que realizará el sistema (Tabla 33).

	Casos	Precip	Tmax	Tmed	Tmin
Fecha					
1985-01-01	6	22.788344	23.231456	17.687719	12.143982
1985-02-01	6	39.761199	24.168778	18.919350	13.669923
1985-03-01	6	31.119475	27.954705	22.238870	16.523035
1985-04-01	8	79.281711	29.761278	23.806926	17.852574
1985-05-01	5	55.414756	31.719499	25.531910	19.344320

Tabla 33: Datos usados para el experimento

C.1.2 Métodos de Machine Learning para predicción de series de tiempo para dengue

Aplicar algoritmos de Machine Learning novedosos es el siguiente paso al uso de modelos como SARIMAX para series de tiempo. Los algoritmos de Machine Learning, son capaces de aprender los patrones dentro de cierto conjunto de datos con el objetivo de generalizar predicciones, clasificaciones o regresiones sobre conjuntos de datos que no hayan visto.

Este tipo de algoritmos ha sido usado con bastante éxito en muchas aplicaciones, incluyendo visión por computadora. La desventaja es que requieren complejos pasos de preprocesamiento y extracción de características ya que solo aceptan datos que se encuentran en forma tabular/matricial.

Existen numerosos algoritmos de Machine Learning que se han aplicado exitosamente a la predicción con series de tiempo, algunos ejemplos son:

1. K-Nearest Neighbor
2. Kernel Regression
3. Radial Basis Functions
4. Regression Trees
5. Support Vector Regression
6. Gaussian Processes

C.1.2.1 Convertir serie de tiempo en aprendizaje supervisado

Para la aplicación de algoritmos de Machine Learning a la predicción de series de tiempo, se requiere convertir el problema a una aplicación de aprendizaje supervisado.

Esto se hace creando una serie de ventanas deslizantes que se van moviendo para generar muestras con muestras asociadas a una variable objetivo, esto se realiza de manera secuencial como se observa en la **Figura 91**.

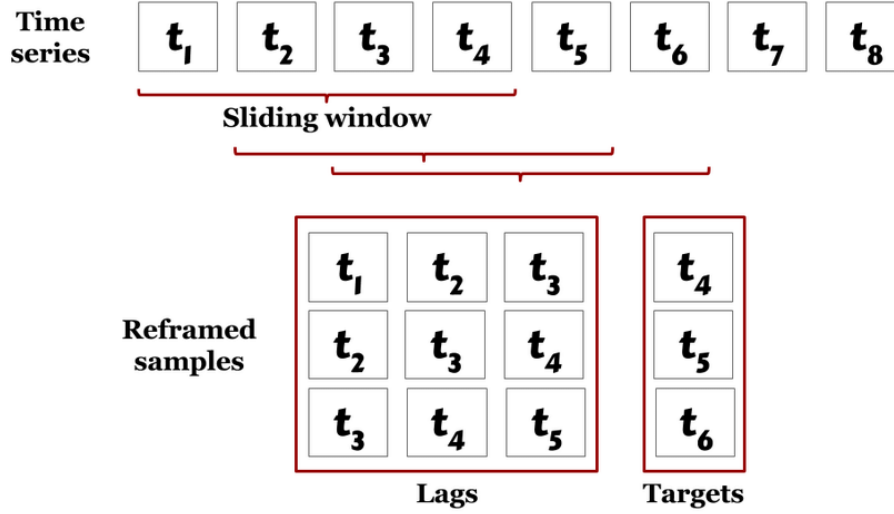


Figura 91: Método de conversión a aprendizaje supervisado

Con esto se pretende mostrar al algoritmo como los pasos previos afectan los posteriores de manera consecutiva sobre toda la serie.

C.1.2.2 Métricas de evaluación

Para el entrenamiento de este modelo y los subsecuentes, se utilizará la métrica Raíz del Error Cuadrático Medio, la cual es una métrica y función de pérdida ampliamente utilizada para la evaluación de modelos de regresión y procesamiento de series de tiempo. La fórmula se muestra a continuación.

$$RMSE = \sqrt{\sum_{n=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Esta fórmula será utilizada para todos los experimentos posteriores como la función de pérdida a minimizar y métrica de evaluación.

C.1.2.3 XGBoost: Árboles de Decisión y Potenciamiento del Gradiente.

XGBoost es uno de los algoritmos más populares actualmente para las tareas de regresión y clasificación realizadas sobre datos tabulares. Es una extensión de los modelos clásicos

de Árboles de Decisión y mejora los procesos de otras metodologías como los Bosques de Decisión. En aplicaciones prácticas, este algoritmo a sustituido muchos otros clásicos como SVM, es por ello que se eligió para representar los modelos de Machine Learning para esta tarea de predicción de casos de dengue.

Este algoritmo es un meta-algoritmo de ensamble, es decir, junta distintos modelos y los une de manera holística, es decir, el todo es mayor a la suma de las partes. Este proceso conjuga múltiples clasificadores débiles para generar un clasificador robusto. Esta implementación de Potenciamiento de Gradiente difiere de la original, ya que incluye optimizaciones de segundo orden que mejoran su rendimiento.

Una ventaja de XGBoost y otros algoritmos de Machine Learning, es que tienen una capacidad de explicabilidad superior a las redes neuronales, ya que un humano puede seguir el razonamiento usado por el algoritmo para generar las salidas.

Conjuntos de prueba y entrenamiento Para el correcto entrenamiento de los modelos, el conjunto de datos inicial se divide en dos subconjuntos: entrenamiento y prueba. Con el objetivo de determinar de manera precisa el poder de generalización de los modelos. En este caso, se utilizarán los últimos 12 registros para la evaluación del rendimiento de los modelos.

Para XGBoost, se utilizarán las variables que representan el número de casos, la precipitación y la temperatura mínima. Esta variable de temperatura fue la más significativa que se encontró en los estudios previos analizados.

Creación del modelo regresor El modelo regresor de XGBoost utilizado tiene como objetivo la minimización de cierta métrica de evaluación. Uno de los hiperparámetros de este algoritmo es el número de estimadores, que representa la cantidad de Árboles de Decisión que se utilizarán para armar el ensamble de modelos predictivos débiles y con los cuales se generará un predictor estadísticamente fuerte. Para este caso se utilizaron 3000 estimadores (Figura 92).

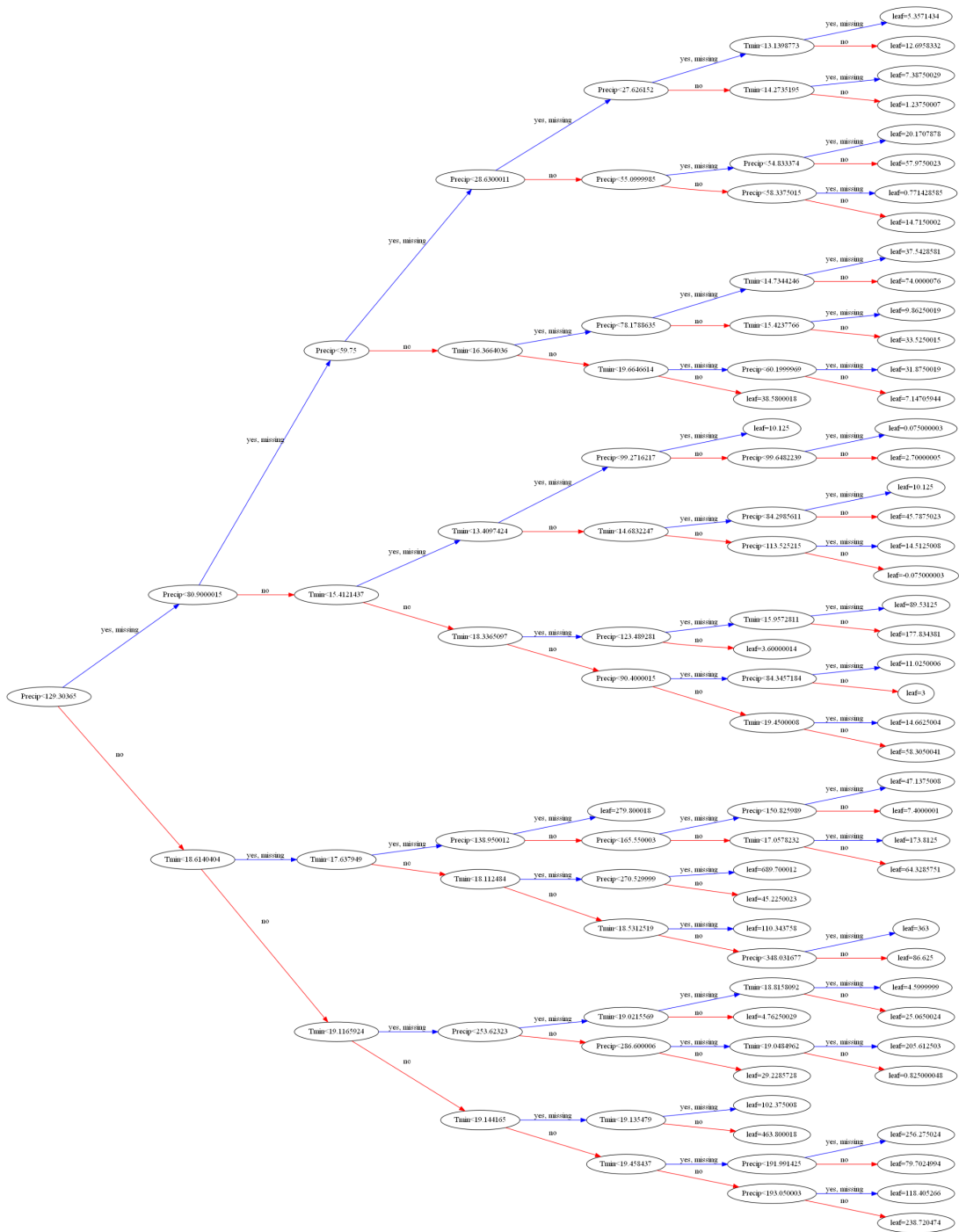


Figura 92: Diagrama del árbol de decisión

Como se puede observar, el modelo fue ajustado durante 100 épocas. El RMSE generado

por la validación dentro del conjunto de entrenamiento decrece constantemente, sin embargo, la validación en el conjunto de prueba se estanca a un nivel tres órdenes de magnitud arriba del entrenamiento, esto quiere decir que el modelo tiene un rendimiento bastante pobre para predecir los casos de dengue.

En la comparativa entre el RMSE de las predicciones sobre los datos de prueba, comparada con la media de los mismos, encontramos que el error es más grande que la media, por lo que el algoritmo no es capaz de determinar correctamente el comportamiento de la serie (Tabla 34).

RMSE	Media de las predicciones
3182.154408	2477.91666

Tabla 34: Comparativa del error

Importancia de las variables Una de las ventajas de la implementación de XGBoost, es que es capaz de determinar la importancia de las variables dentro de los árboles de decisión generados. Podemos observar claramente que la Precipitación es la variable más importante para la predicción de casos de dengue (Figura 93).

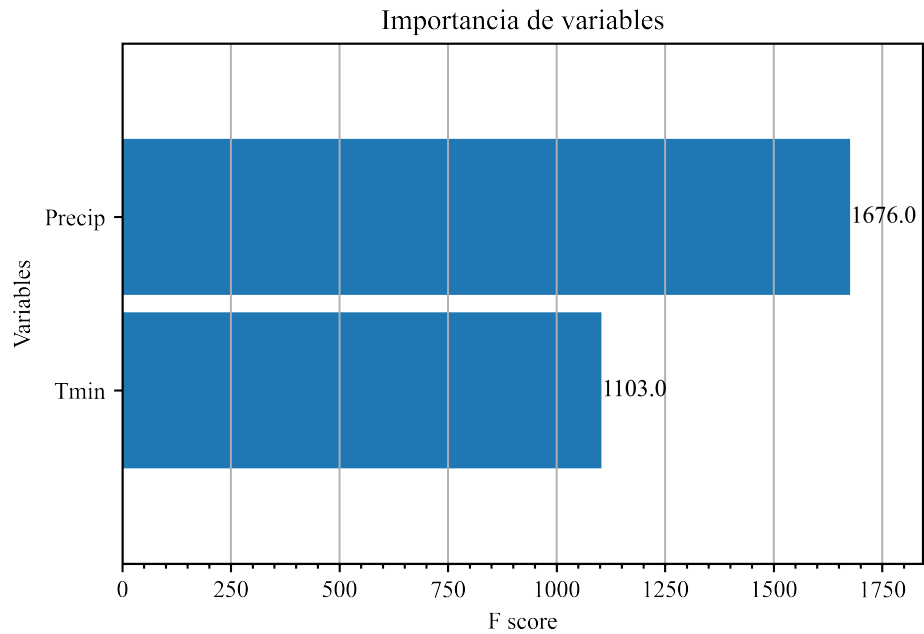


Figura 93: Importancia de las variables

Comparativa de predicciones En la siguiente gráfica se comparan las predicciones realizadas con el conjunto de prueba comparadas con los valores reales. Es evidente que el modelo es incapaz de determinar correctamente la cantidad de casos, aunque si fue capaz de encontrar cierto patrón en la serie de tiempo y representarlo aunque de manera somera.

Debido a esto se realizaron otros experimentos cambiando variables y generando variables y características nuevas para tratar de capturar el comportamiento intrínseco del fenómeno subyacente (Figura 94).

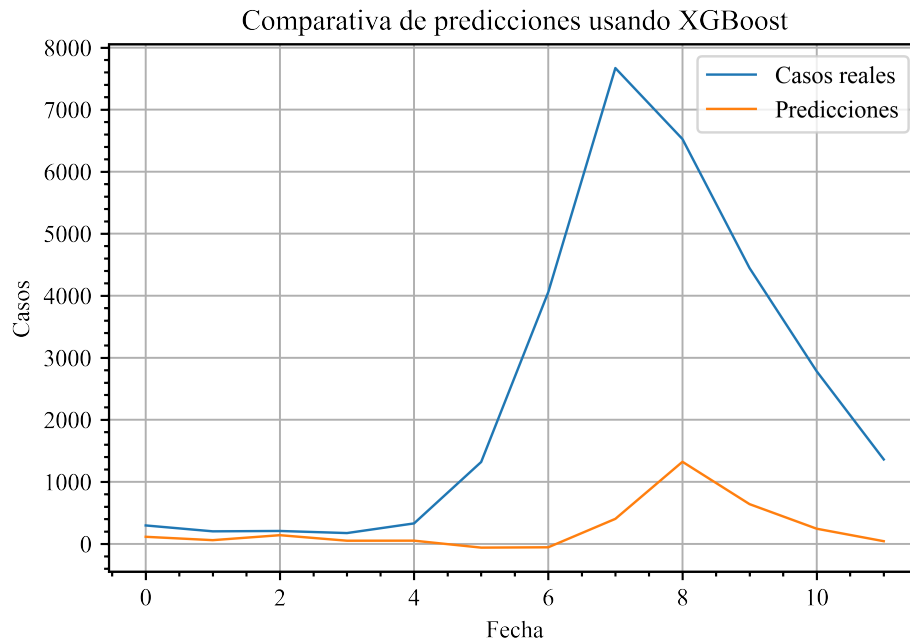


Figura 94: Comparativa entre serie y su predicción

Ingeniería de Características La Ingeniería de Características es un proceso bastante utilizado en Machine Learning. Este proceso consiste en generar nuevas variables o columnas a partir de los datos, para poder capturar mejor el comportamiento y los patrones que existen dentro de los mismos. Este proceso se ha definido como una mezcla entre técnica y arte y puede ser bastante complejo y tedioso.

Para las series de tiempo, las características y variables generadas pueden consistir en la hora, el minuto, el día de la semana, el cuarto del año, el número de mes, etcétera. Esto sirve para que el algoritmo entienda las relaciones que tienen los aspectos temporales y su relación con la variable de respuesta.

En este caso, al ser datos mensuales, solo se pueden generar las variables número de mes, año y cuarto del año ([Tabla 35](#)).

	Precip	Tmin	quarter	month	year
Fecha					
1985-01-01	22.788344	12.143982	1	1	1985
1985-02-01	39.761199	13.669923	1	2	1985
1985-03-01	31.119475	16.523035	1	3	1985
1985-04-01	79.281711	17.852574	2	4	1985
1985-05-01	55.414756	19.344320	2	5	1985

Tabla 35: Ingeniería de características

Se puede observar que el comportamiento es bastante similar al experimento anterior, aunque si se pudo reducir levemente el RMSE tanto en los conjuntos de entrenamiento y validación.

Las variables climatológicas son más importantes para la predicción de los casos que las características temporales generadas (Figura 95).

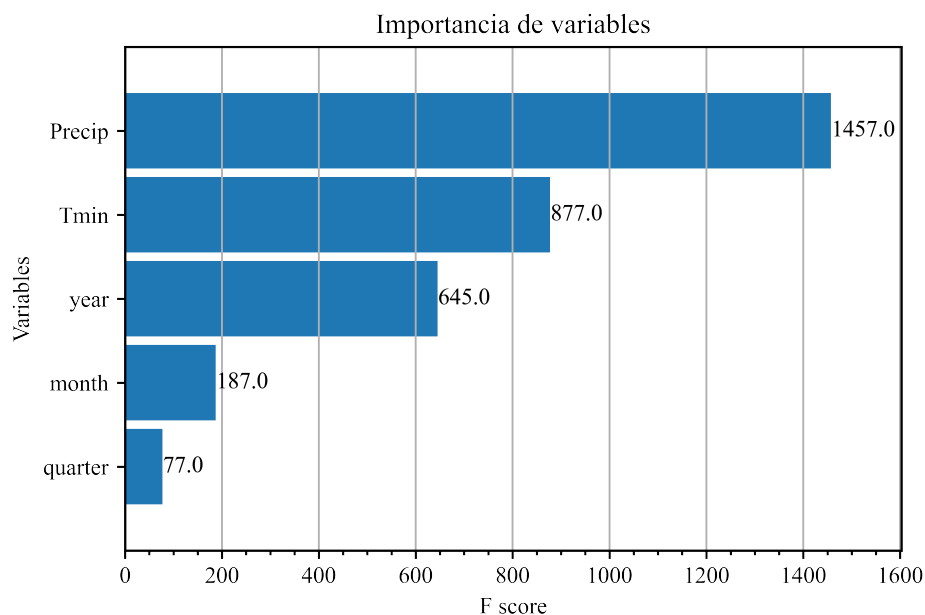


Figura 95: Importancia de variables: Ingeniería de características

Al evaluar el modelo con las características generadas, encontramos que tampoco fue capaz de

realizar predicciones útiles sobre el conjunto de pruebas (Figura 96).

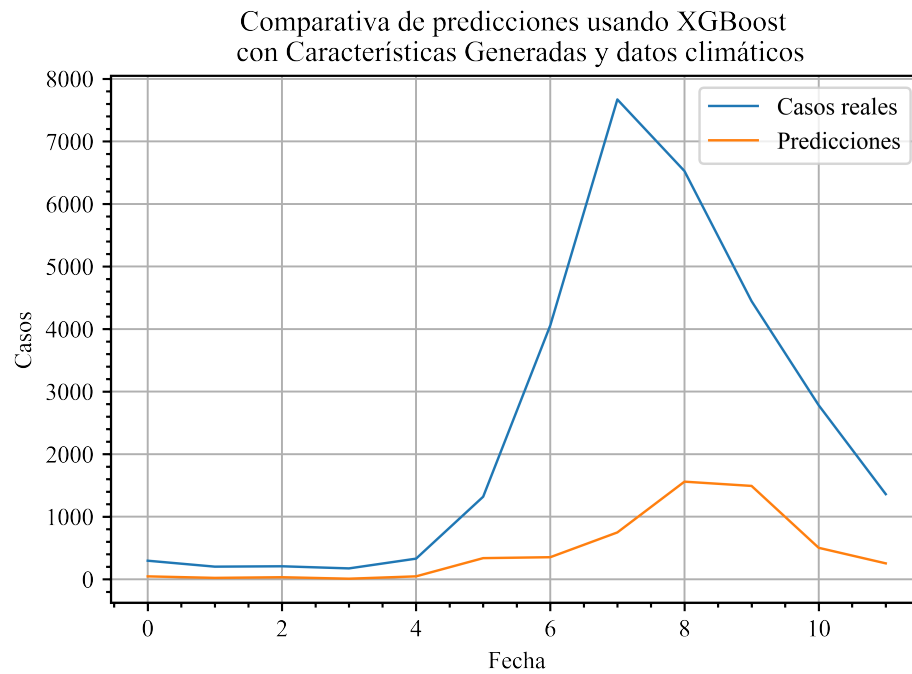


Figura 96: Comparativa de resultados: Ingeniería de características

Por completud, se realizó un experimento removiendo las variables climatológicas y usando solamente las características temporales generadas (Figura 97).

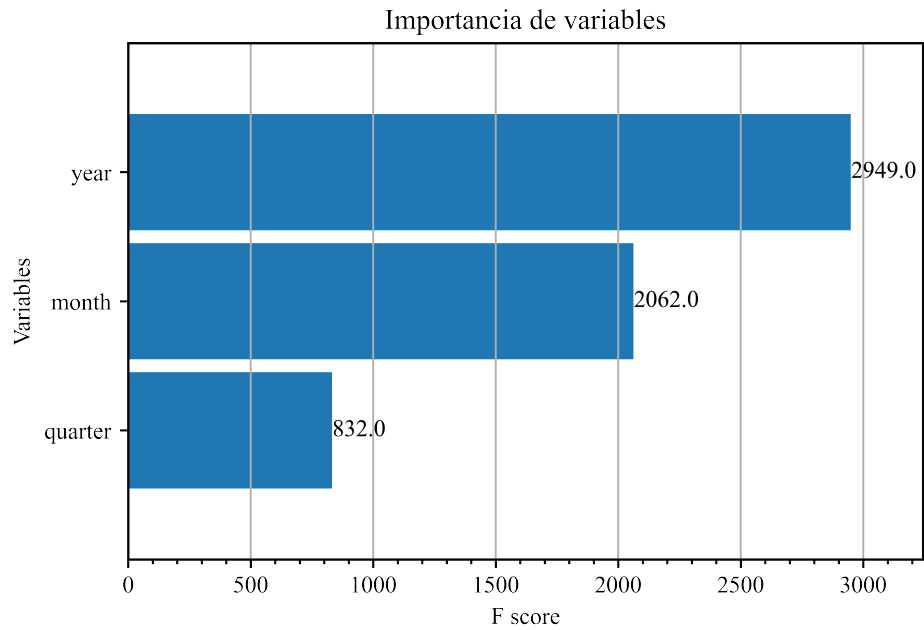


Figura 97: Importancia de las variables: variables climatológicas

La gráfica muestra que este modelo tampoco fue exitoso (Figura 98).

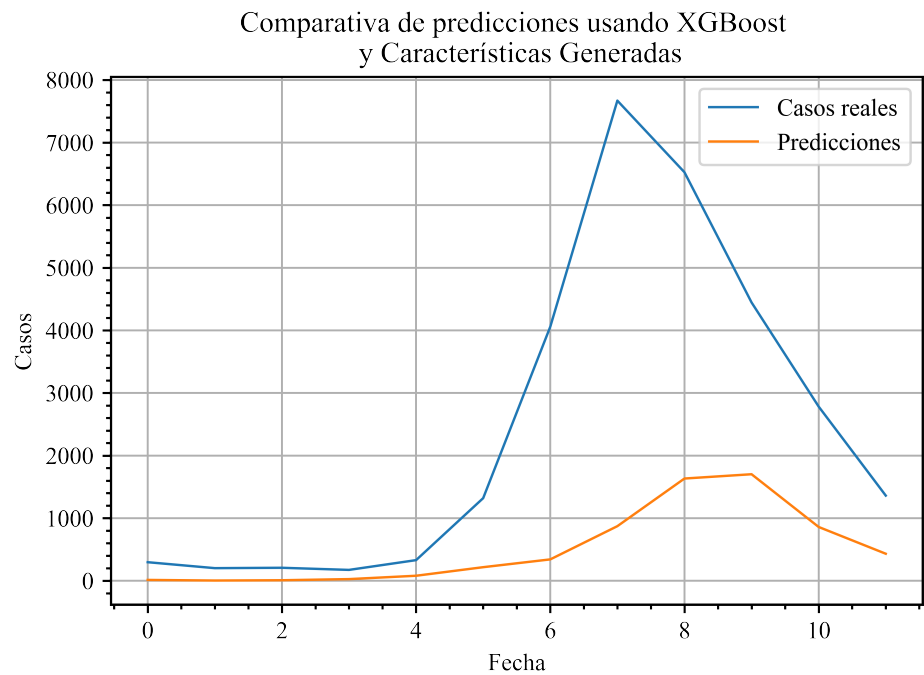


Figura 98: Comparativa de resultados: variables climatológicas

C.1.2.4 Resultados del modelado con Machine Learning

XGBoost es un algoritmo moderno y sumamente potente que ha reemplazado muchos otros modelos de Machine Learning para el aprendizaje supervisado. Es especialmente exitoso cuando se trata de regresión y clasificación de datos tabulares. Lamentablemente, fue incapaz de abstraer correctamente el comportamiento de las series de tiempo para la predicción de casos del dengue.

Es por ello que se procedió a realizar modelos de Deep Learning para la predicción de series de tiempo. La ventaja de estos modelos que es son capaces de encontrar los patrones importantes por si solos y no se requiere de un proceso complejo de ingeniería de características previo.

C.1.3 Métodos de Deep Learning para predicción de series de tiempo para dengue

El Deep Learning tiene ventajas sobre el Machine Learning, ya que en distintos tipos de tareas como visión por computadora o procesamiento de lenguaje natural, no requiere de un paso previo de Ingeniería de Características o de Extracción de patrones para poder ser aplicados.

Estos algoritmos, como se ha tratado en trabajos anteriores, son capaces de procesar datos más allá de información tabular y en la actualidad tienen el mejor rendimiento en este tipo de tareas.

C.1.3.1 Tareas de Deep Learning para series de tiempo

Las tareas de Deep Learning para la predicción de series de tiempo son:

- Modelos univariados.
- Modelos multivariados.
- Modelos univariados multipaso.
- Modelos multivariados multipaso.

El número de variables se refiere a la cantidad de series de tiempo que se ingresan al modelo

para realizar sus predicciones. Mientras que los pasos se refiere a cuantas unidades a futuro se realizarán las predicciones.

En nuestro caso, las variables de entrada son las series de tiempo de casos de dengue así como las series de datos climatológicos.

C.1.3.2 Modelos de Deep Learning explorados

Los modelos utilizados para esta exploración fueron los siguientes:

- Perceptrón.
- Redes Neuronales Convolucionales.
- Redes Neuronales Recurrentes

Estos modelos son representativos de las Redes Neuronales y Deep Learning en general, siendo las arquitecturas modificaciones estructurales de tales modelos. Para el Perceptrón se modifican las capas y el número de neuronas en cada capa, para las Redes Convolucionales se modifican las capas, tipo de capa y número de filtros, mientras que en las Redes Recurrentes se cambian el número de capas y de unidades por capa.

En la siguiente imagen encontramos la comparativa entre estos modelos. Las CNN no son exclusivas para imágenes, también pueden ser usadas para señales, pero las RNN no pueden ser usadas para imágenes. Todos estos modelos son difíciles de entrenar, una de las razones es que todos sufren de problemas con los gradientes (Figura 99).

	MLP	RNN	CNN
Data	Tabular data	Sequence data (Time Series, Text, Audio)	Image data
Recurrent connections	No	Yes	No
Parameter sharing	No	Yes	Yes
Spatial relationship	No	No	Yes
Vanishing & Exploding Gradient	Yes	Yes	Yes

Figura 99: Comparativa de modelos

De cada uno de estos modelos se probaron arquitecturas distintas, se evaluaron bajo las mismas condiciones con el conjunto de prueba junto con los hiperparámetros y se compararon sus rendimientos para determinar el mejor modelo y la mejor arquitectura.

Los datos utilizados para este proceso son los mismos que se usaron para Machine Learning.

C.1.3.3 Escalamiento de los datos

Como se puede observar, los datos de las columnas se encuentran repartidos en distintos rangos. Esto puede causar un problema al entrenar modelos, ya que cada variable contribuirá de manera distinta, en función a su escala al aprendizaje del modelo, lo cual puede crear sesgos e impactar negativamente en el rendimiento final del modelo. Esto es bastante en modelos basados en el algoritmo de retropropagación, en donde el proceso se hace más estable y potencialmente más rápido, también mitiga el problema de los gradientes.

Para ello es recomendable escalar y/o normalizar los datos. En este caso, se usó el escalador MinMaxScaler. La fórmula para el escalamiento de los datos se muestra a continuación.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Este proceso toma los mínimos y máximos encontrados en el conjunto de datos y los utiliza para reducir los datos a un intervalo $[0, 1]$. Esto se puede ver claramente en la siguiente figura (Figura 100).

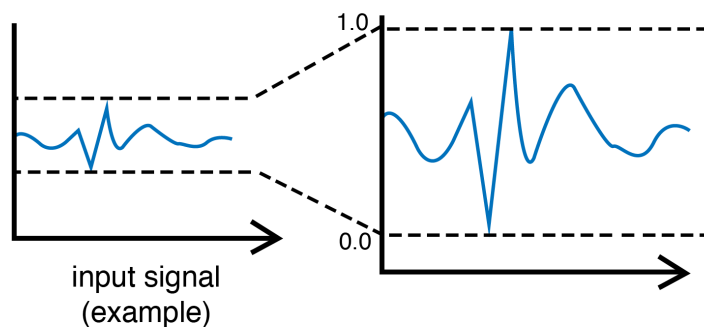


Figura 100: Método de escalamiento

Este proceso es sensible a los valores extremos dentro del conjunto de datos, sin embargo, no se eliminaron porque se pretende que el modelo pueda capturar la totalidad del comportamiento de la serie de tiempo de casos de dengue.

Otra forma de reducir el rango de los datos es normalizarlos, sin embargo, esto requiere ciertas condiciones estadísticas que no están aseguradas para el conjunto de datos de casos de dengue.

C.1.3.4 Predicción de series univariadas

Primero se probó la tarea de series univariadas, es decir, una sola variable para predecir una sola unidad de tiempo a futuro.

Hiperparámetros de entrenamiento Para el entrenamiento de las arquitecturas subsecuentes, se utilizaron los siguientes hiperparámetros:

- Datos de validación: 12 meses.

- Número de pasos de entrada: 6 meses.
- Tasa de aprendizaje: 0.001
- Optimizador: Nesterov-accelerated Adaptive Moment Estimation.
- Función de pérdida: Raíz del error cuadrático medio.
- Épocas: 200
- Tamaño de lote: 512

Los pasos de entrada se refiere a que se utilizarán entradas de 6 meses para predecir el próximo mes. Esto conforme al proceso de conversión de serie de tiempo a aprendizaje supervisado.

La tasa de aprendizaje se dejó fija para alcanzar una convergencia mejor y una comparativa más equitativa entre todos los modelos.

Vale la pena observar la elección del optimizador. Nadam es un optimizador que incorpora métodos para evitar mínimos locales. Tiene un mejor rendimiento que el optimizador clásico SGD y otros de los optimizadores más populares como Adam (Figura 101).

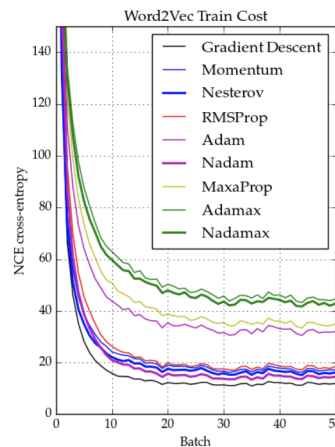


Figura 101: Comparativa de optimizadores

El resto de los hiperparámetros se fijó usando una heurística personal.

Perceptrón Multicapa Estos dos modelos difieren en la cantidad de capas ocultas y no en la cantidad de neuronas en cada capa. Se crearon múltiples arquitecturas desde 1 capa hasta 3 capas ocultas, probando variadas configuraciones de neuronas en las mismas. Todas utilizaron la función de activación ReLU en cada una de sus capas.

Como se puede ver en la siguiente imagen (Figura 102), un perceptrón multicapa se convierte en una Red Neuronal Profunda al agregar más capas ocultas.

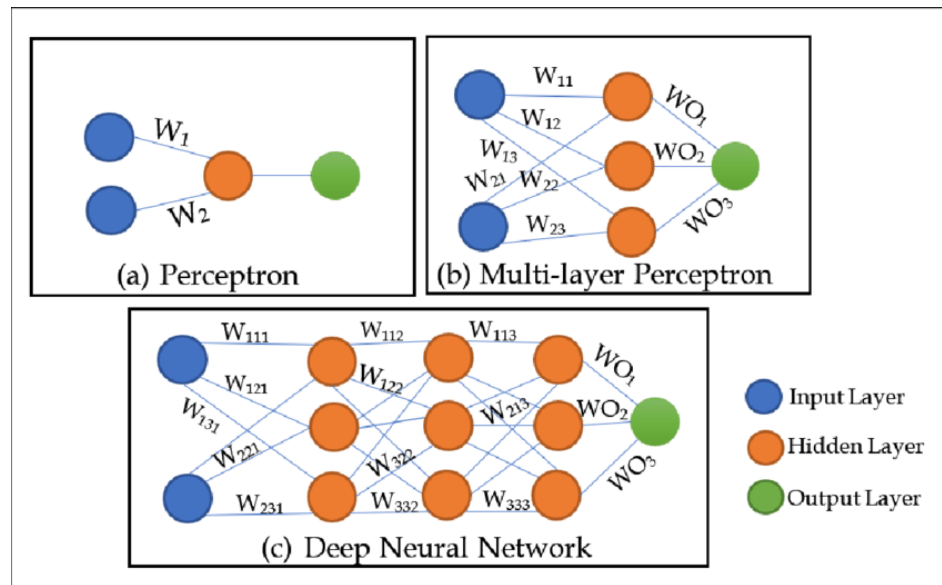


Figura 102: Arquitecturas del perceptrón

El número de neuronas se tuvo que mantener debajo de cierto rango, ya que las capas incrementan exponencialmente la cantidad de memoria que usa el modelo.

Perceptrón de 1 capa Para este modelo se probaron arquitecturas con 512, 1024, 2048 y 4085 neuronas [Tabla 36](#).

Neuronas	RMSE
512	0.27799729977459775
1024	0.2525902194760879
2048	0.24079683427627366
4086	0.27173173006586393

Tabla 36: Rendimiento por arquitectura: perceptrón de 1 capa

Podemos observar en la [Figura 103](#) del entrenamiento que la pérdida es muy cercana a cero en el conjunto de entrenamiento pero se estanca en el conjunto de prueba.

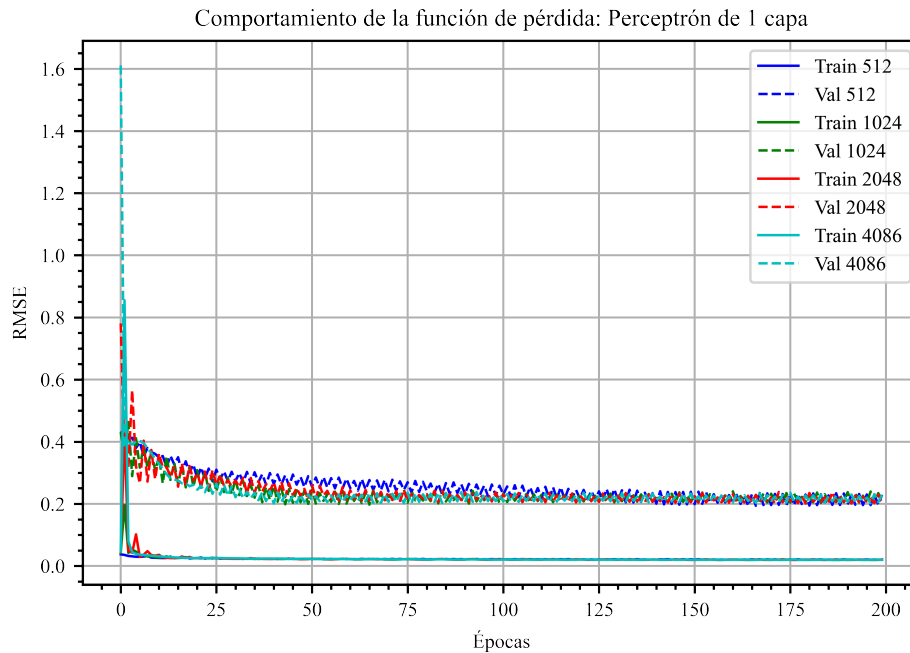


Figura 103: Comparativa de pérdida: perceptrón de 1 capa

El rendimiento de las arquitecturas fue mejor con un número de neuronas moderado ([Figura 104](#)).

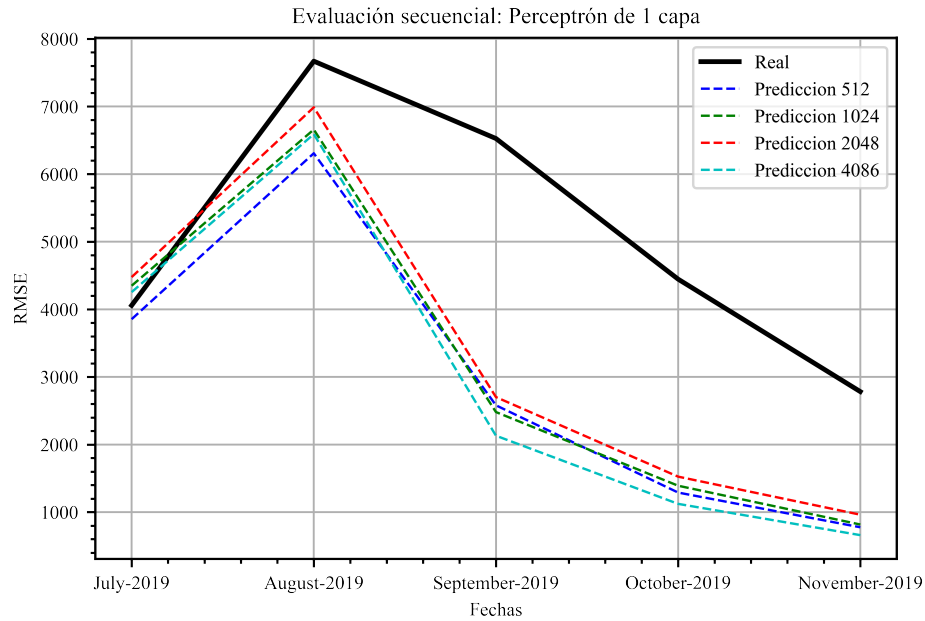


Figura 104: Evaluación secuencial: perceptrón de 1 capa

La comparativa de evaluación muestra que los modelos fueron capaces de encontrar una estructura similar en el conjunto de prueba, sin embargo, no logra capturar la totalidad del comportamiento del fenómeno (Figura 105).

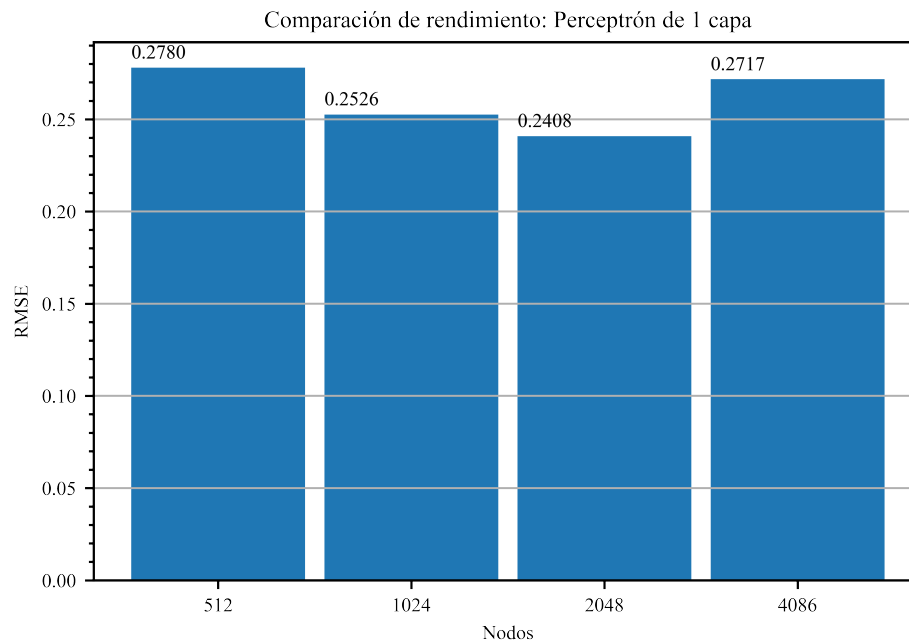


Figura 105: Comparativa de rendimiento: perceptrón de 1 capa

Perceptrón de dos capas Para este modelo se probaron (512, 256), (1024, 512), (2048, 1024) para cada una de las dos capas (Tabla 37).

Neuronas	RMSE
(512, 256)	0.2663535622773258
(1024, 512)	0.25390077227930835
(2048, 1024)	0.2526750762556925

Tabla 37: Rendimiento por arquitectura: perceptrón de 2 capas

Agregar una capa al perceptrón mejora el rendimiento, sin embargo, incrementa bastante el uso de memoria. Incrementar el tamaño de las capas ocultas impacta negativamente en el rendimiento.

El modelo es capaz de encontrar patrones pertinentes, sin embargo, en septiembre en lugar de predecir un incremento, predice una menor cantidad de casos (Figura 106, Figura 107, Figura 108).

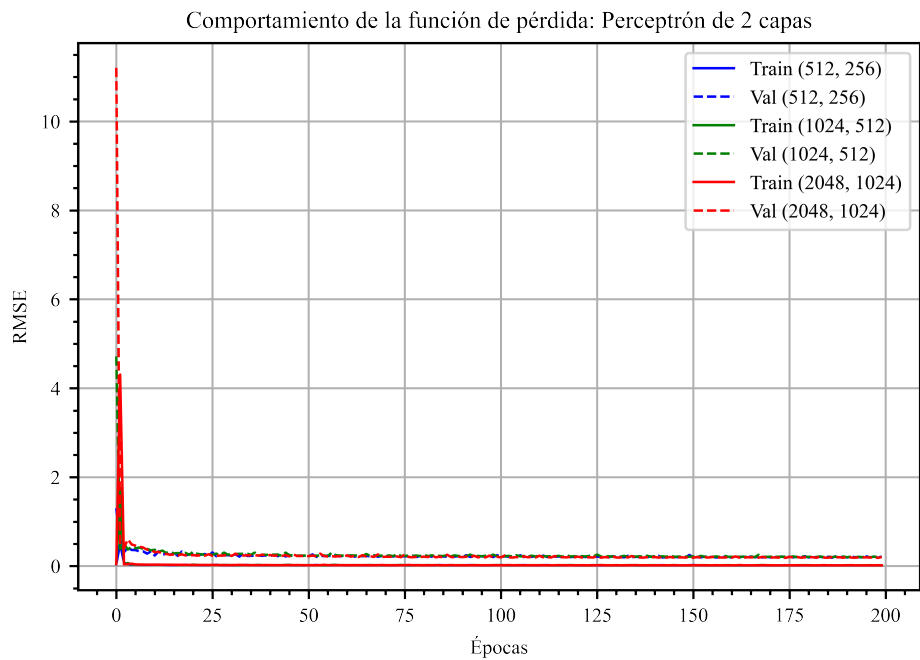


Figura 106: Comparativa de pérdida: perceptrón de 2 capas

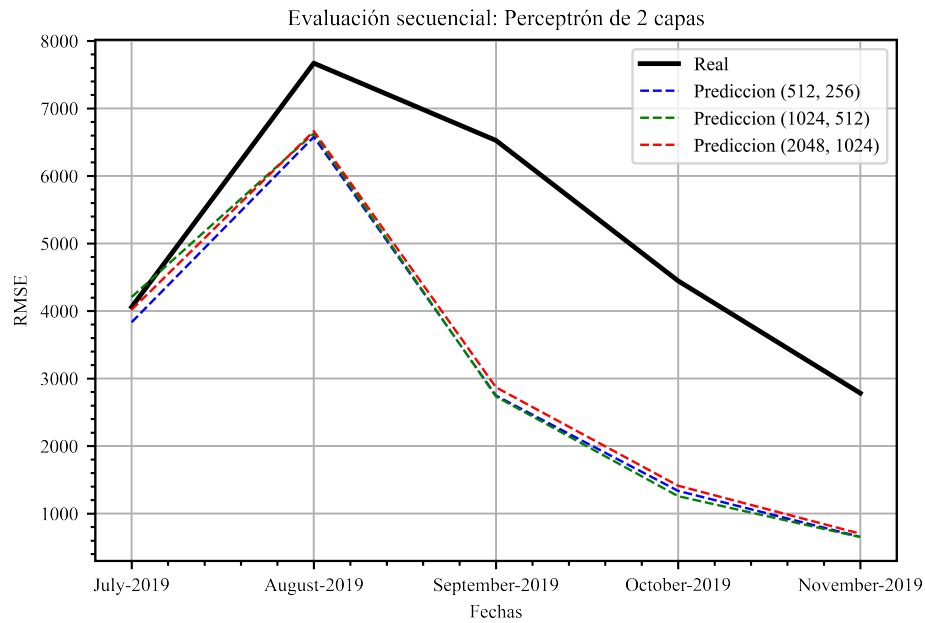


Figura 107: Evaluación secuencial: perceptrón de 2 capas

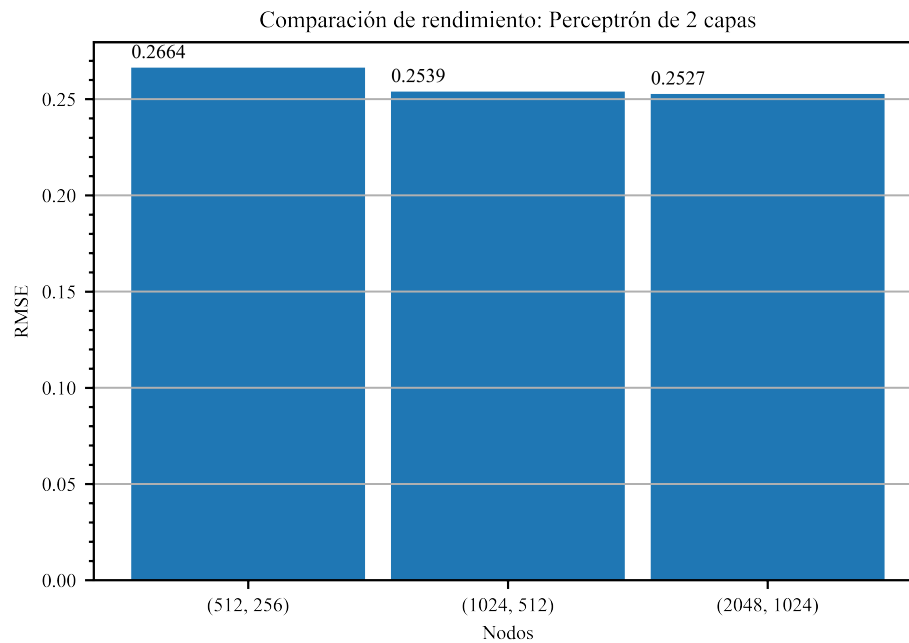


Figura 108: Comparativa de rendimiento: perceptrón de 2 capas

Perceptrón de 3 capas Finalmente, se usaron (512, 256, 128), (1024, 512, 256), (2048, 1024, 512) estas neuronas para las tres capas de este modelo.

Neuronas	RMSE
(512, 256, 128)	0.27361407189882936
(1024, 512, 256)	0.2764884314567098
(2048, 1024, 512)	0.28553395655019054

Tabla 38: Rendimiento por arquitectura: perceptrón de 3 capas

Agregar una capa extra mejoró el rendimiento del modelo, pero agregar más neuronas a las capas tiene un impacto negativo. El modelo también captura el comportamiento pero sigue sin predecir correctamente los picos de casos ([Figura 109](#), [Figura 110](#), [Figura 111](#)).

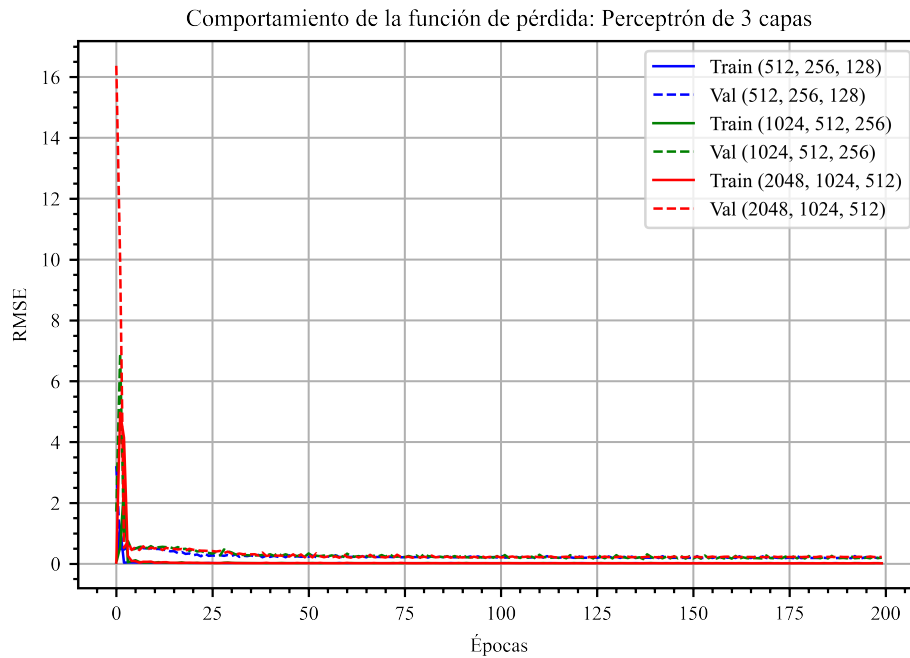


Figura 109: Comparativa de pérdida: perceptrón de 3 capas

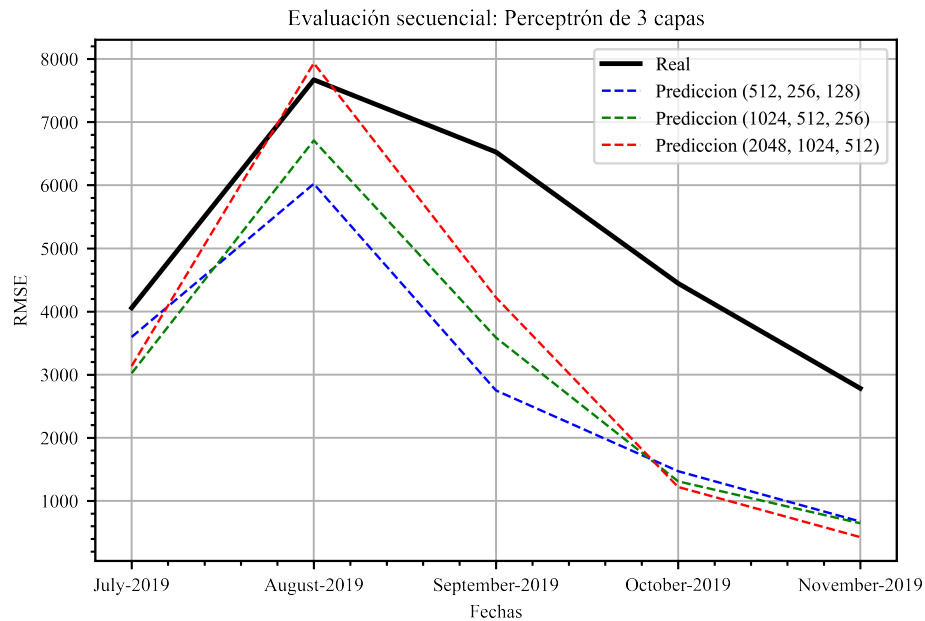


Figura 110: Evaluación secuencial: perceptrón de 3 capas

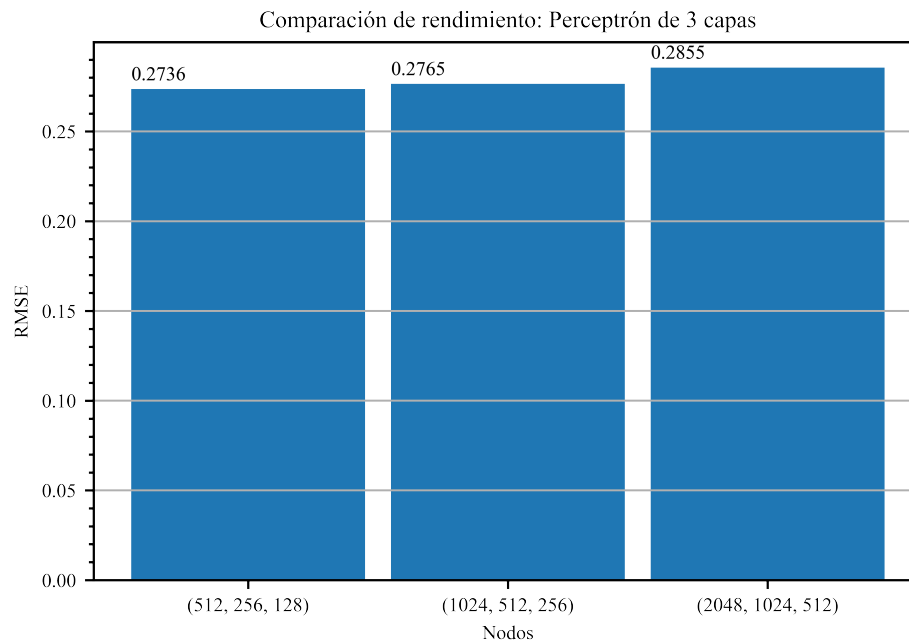


Figura 111: Comparativa de rendimiento: perceptrón de 3 capas

Red Neuronal Convolutacional Este tipo de redes ha sido usado exitosamente para el procesamiento de datos complejos como las imágenes. Son capaces de encontrar por si mismas los

patrones y características importantes en los datos de entrenamiento.

Para las series de tiempo, se utiliza una convolución 1D que opera secuencialmente sobre los datos convertidos a aprendizaje supervisado **Figura 112**.

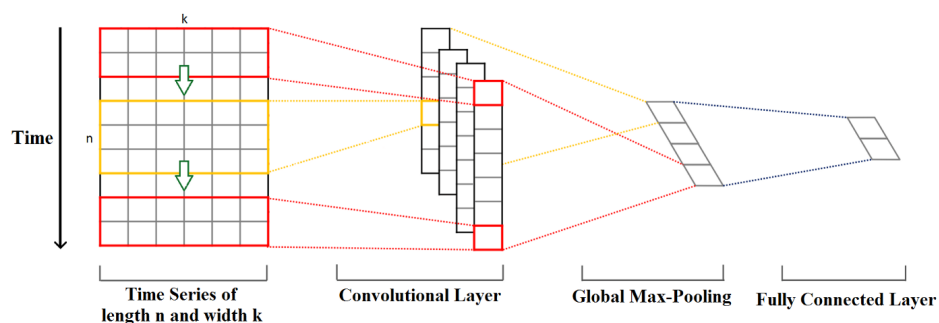


Figura 112: Convolución para series de tiempo

La arquitectura se compone de una dos capas convolucionales 1D, una capa de MaxPooling y una capa densa para realizar las predicciones. Los parámetros a variar en la arquitectura son el número de filtros en cada capa Conv1D: 128, 256, 512, 1024. El número de kernels se mantuvo constante con 3.

Filtros	RMSE
128	0.2816954587419522
256	0.3168345098889043
512	0.29728972469882464
1024	0.3370583974537395
2048	0.29529985856324725
4096	0.4788805257015482

Tabla 39: Rendimiento por arquitectura: CNN de 1 capa

Estos modelos tienen un rendimiento menor al perceptrón, sin embargo, a un número de filtros alto, son capaces de modelar de manera similar el comportamiento de la serie, el problema

radica en que las predicciones se quedan bastante cortas al real. Se puede observar que a partir de ciertos filtros, estos resultan contraproducentes (Figura 113, Figura 114, Figura 115).

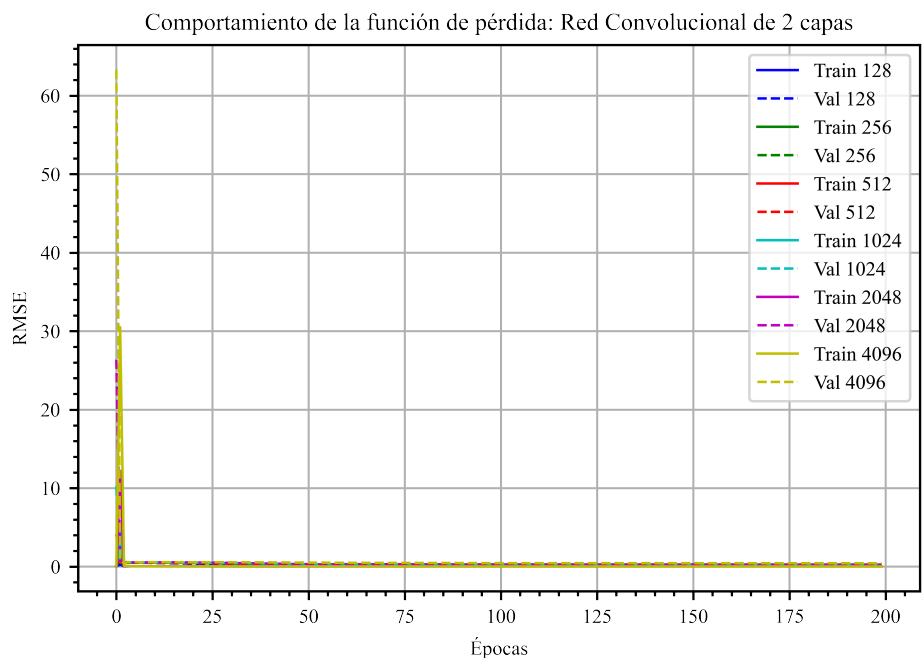


Figura 113: Comparativa de pérdida: CNN de 1 capa

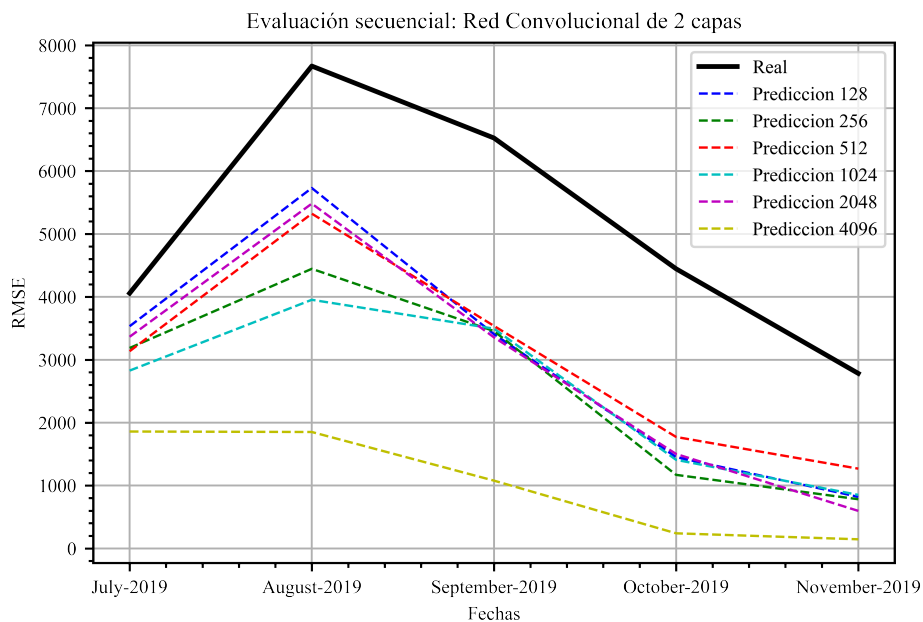


Figura 114: Evaluación secuencial: CNN de 1 capa

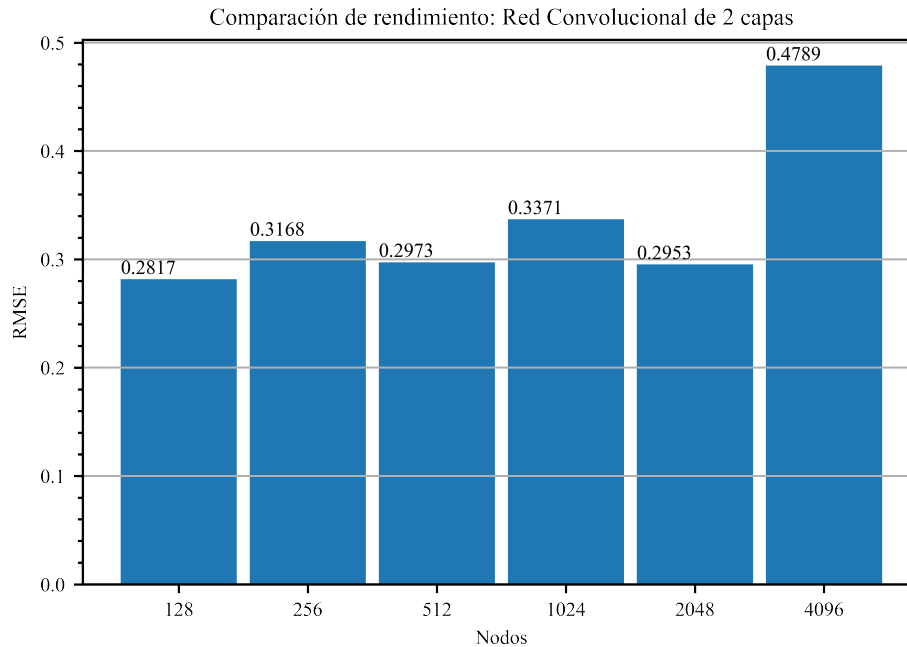


Figura 115: Comparativa de rendimiento: CNN de 1 capa

Redes Neuronales Recurrentes En este tipo de red, las conexiones de los nodos forman un grafo a través de una secuencia temporal, lo que les permite tener un comportamiento dinámico. Cuentan con un estado interno que les permite "recordar" valores previos para realizar predicciones a futuro.

Para el problema de series de tiempo para dengue, se usó el modelo Long short-term memory o LSTM de redes neuronales recurrentes. Este modelo ha tenido bastante éxito en procesamiento de datos similares a una serie de tiempo, como por ejemplo, reconocimiento del habla o clasificación de señales.

En la siguiente imagen podemos observar la complejidad cada una de las celdas o capas LSTM que componen estas arquitecturas (Figura 116).

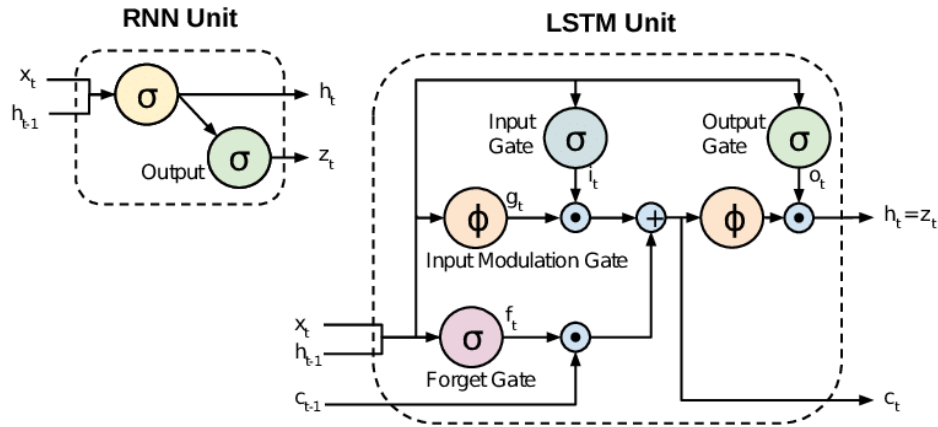


Figura 116: Redes recurrentes y Long Term Short Memory Units

Así como en otro tipo de redes, se pueden apilar capas para incrementar la capacidad predictiva del modelo.

LSTM de 1 capa Para estos modelos se probaron 50, 100, 150, 200, 250, 300 y 350 cantidad de unidades (Tabla 40).

Unidades	RMSE
50	0.2980782533259861
100	0.33794608818415806
150	0.35515983381380706
200	0.3376980674259327
250	0.33411806535280697
300	0.33611282051257346
350	0.3113676586199027

Tabla 40: Rendimiento por arquitectura: LSTM de 1 capa

Podemos observar que la pérdida decrece hasta aproximadamente 125 épocas, donde comienza a explotar. Esto se debe a que probablemente el modelo esté sobreajustando sobre el conjunto de entrenamiento. Se pueden aplicar técnicas de regularización para corregir este problema.

El modelo es incapaz de capturar correctamente el comportamiento, aunque ha dado predicciones sumamente precisas con respecto a los casos reales (Figura 117, Figura 118, Figura 119).

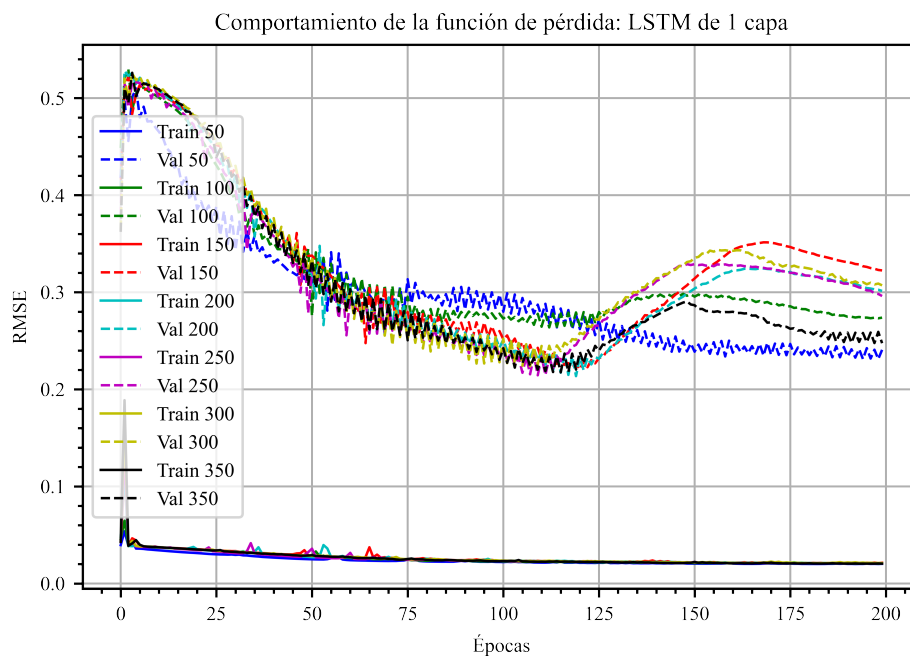


Figura 117: Comparativa de pérdida: LSTM de 1 capa

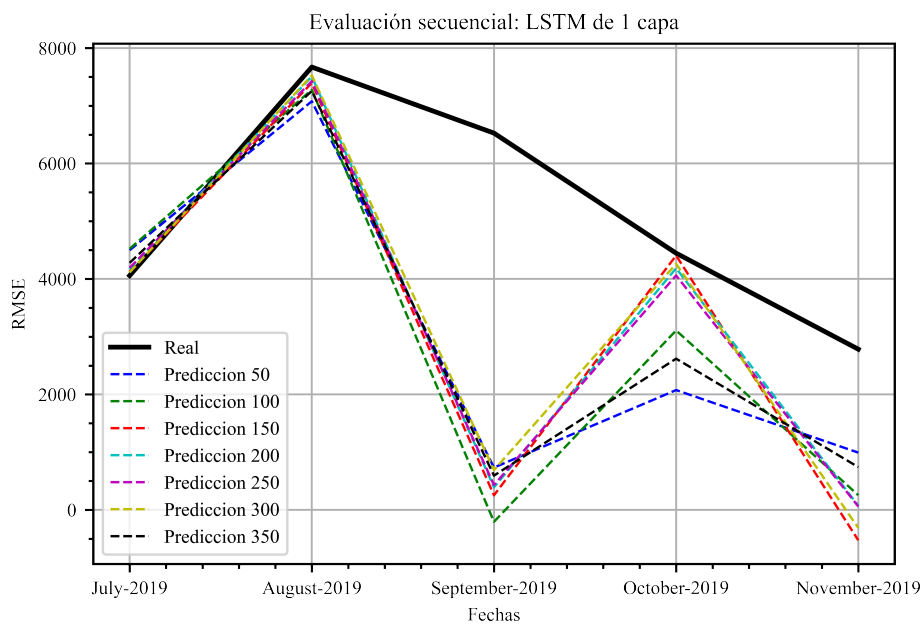


Figura 118: Evaluación secuencial: LSTM de 1 capa

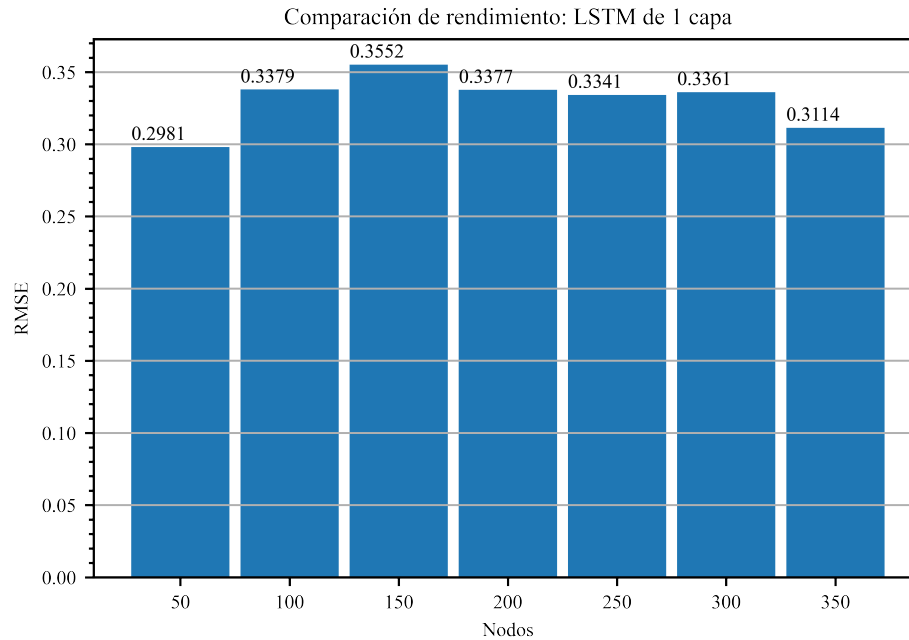


Figura 119: Comparativa de rendimiento: LSTM de 1 capa

LSTM de 2 capas Se agregó una capa extra al modelo, con las mismas unidades 50, 100, 150, 200, 250, 300 (Tabla 41).

Unidades	RMSE
50	0.3660073458967233
100	0.34317870980674964
150	0.2747619104346175
200	0.3602888091802829
250	0.4895612048990135
300	0.3100472777866654

Tabla 41: Rendimiento por arquitectura: LSTM de 2 capas

Agregar otra capa no mejora el rendimiento del modelo, el cual comienza a sobreajustar aproximadamente al mismo tiempo. Con una capa extra, el modelo da sobrestimaciones de casos (Figura 120, Figura 121, Figura 122).

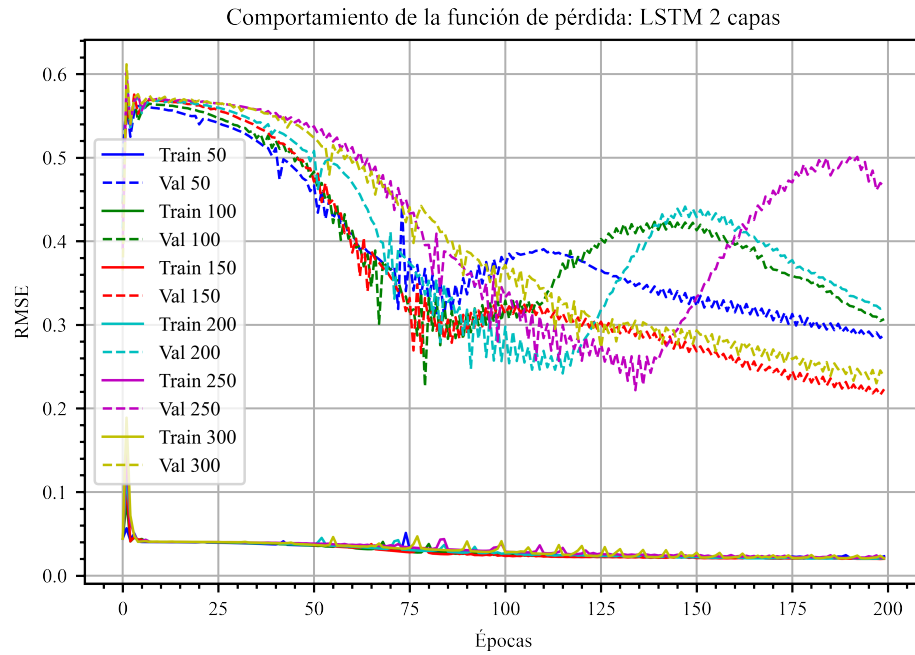


Figura 120: Comparativa de pérdida: LSTM de 2 capas

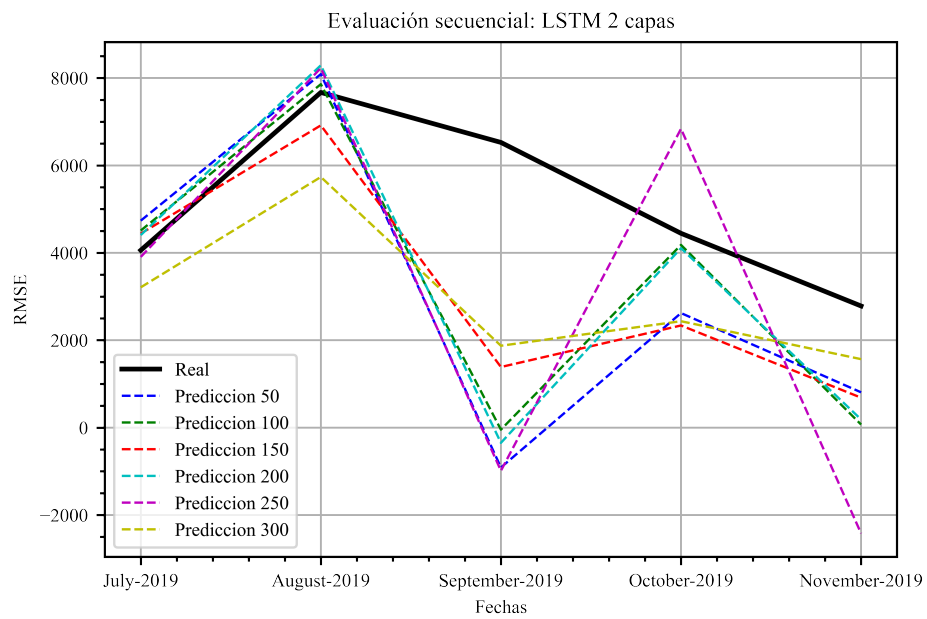


Figura 121: Evaluación secuencial: LSTM de 2 capas

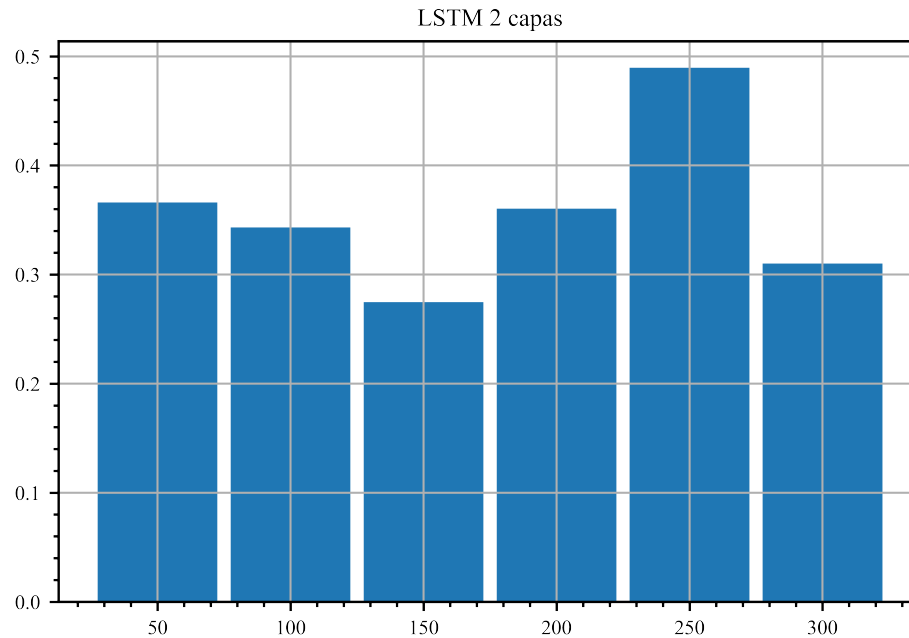


Figura 122: Comparativa de rendimiento: LSTM de 2 capas

LSTM Bidireccional Este tipo de modelos puede capturar información temporal tanto hacia el pasado como hacia el futuro (Tabla 42).

Unidades	RMSE
50	0.3201135864439049
100	0.3219668986009985
150	0.3162823909908708
200	0.3377198361549139
250	0.30567806829994854
300	0.32000950385692173

Tabla 42: Rendimiento por arquitectura: LSTM Bidireccional

El modelo sobreajusta antes en la arquitectura bidireccional. Podemos ver como las pérdidas explotan a partir de las 100 épocas. Este modelo tampoco tiene un rendimiento aceptable (Figura 123, Figura 124, Figura 125).

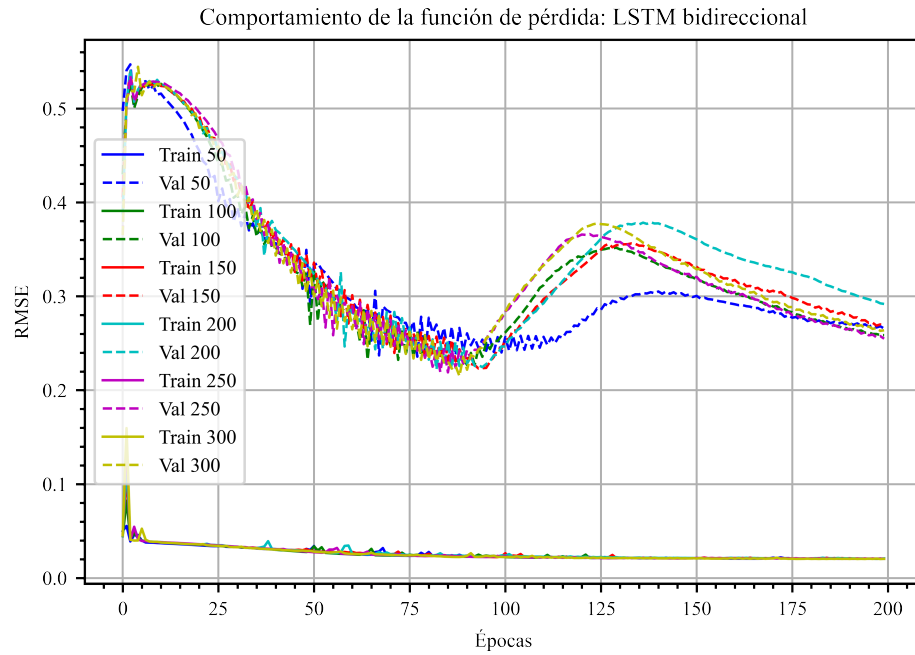


Figura 123: Comparativa de pérdida: LSTM bidireccional

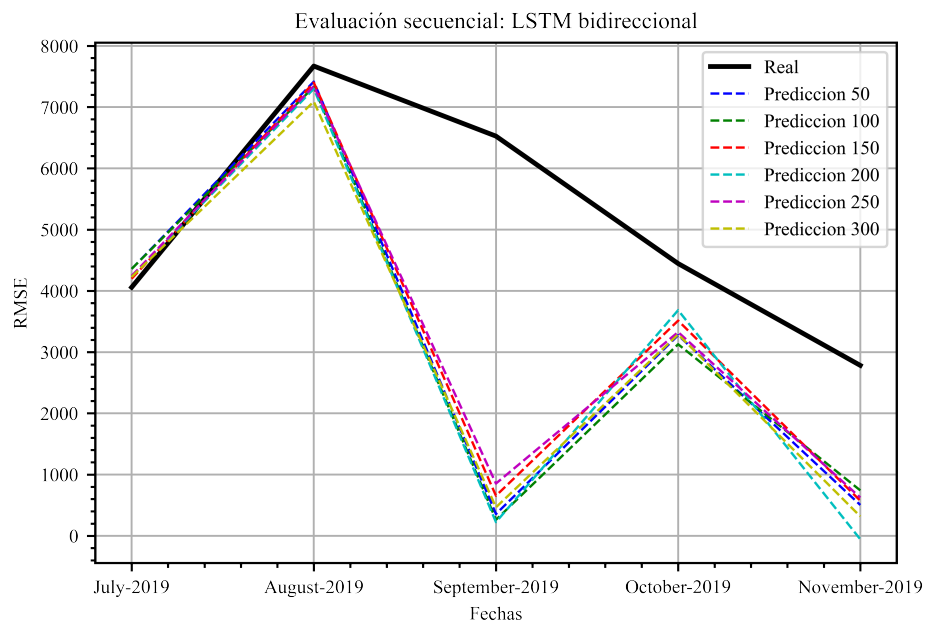


Figura 124: Evaluación secuencial: LSTM bidireccional

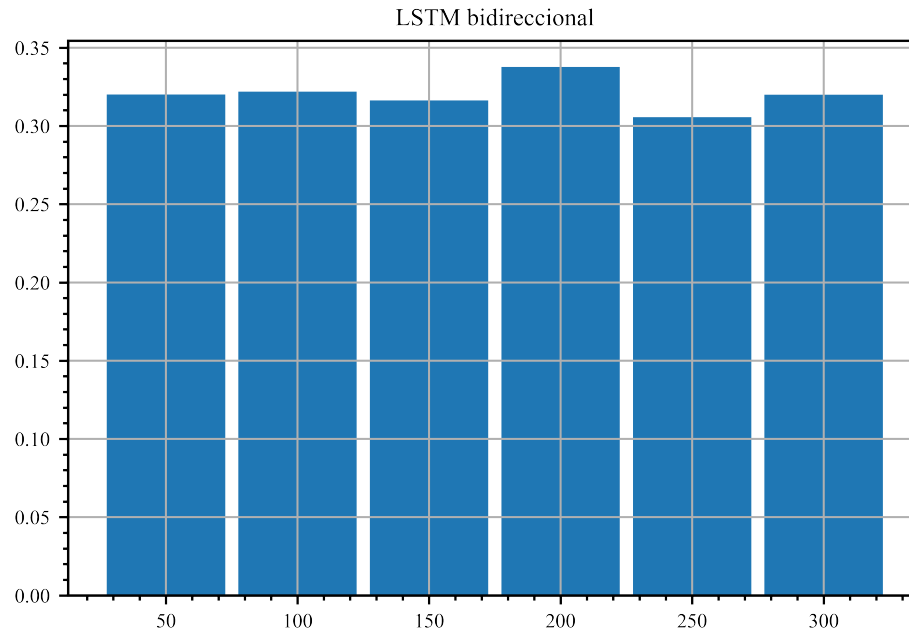


Figura 125: Comparativa de rendimiento: LSTM bidireccional

C.1.3.5 Conclusiones para modelos univariados

Los modelos multivariados con Deep Learning fueron capaces de encontrar patrones subyacentes del fenómeno dentro de la serie de tiempo de casos de dengue. Lamentablemente el poder de generalización de los modelos no fue el adecuado (Figura 126).

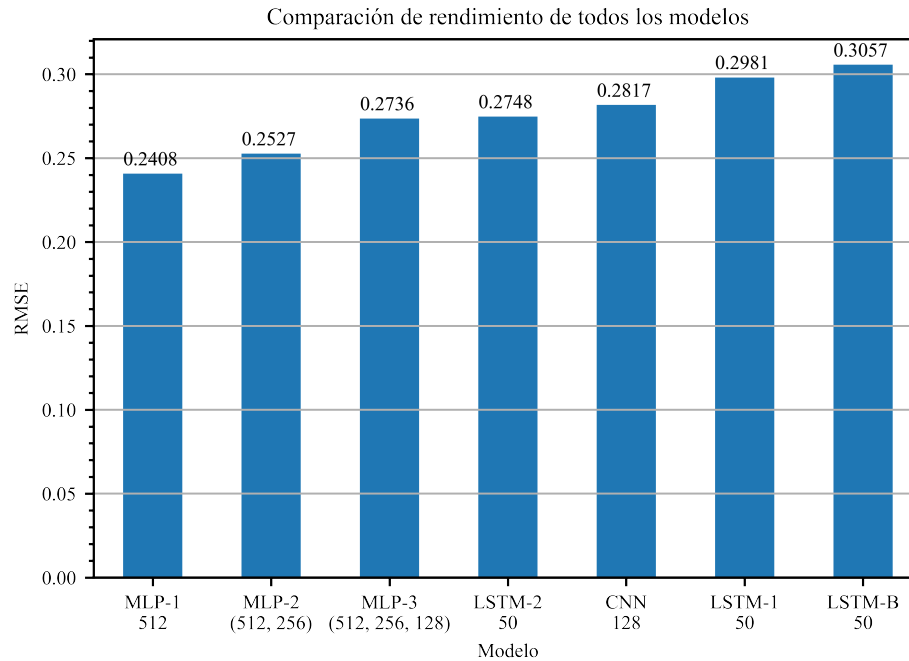


Figura 126: Comparativa de rendimiento para modelos univariados

El Perceptrón de 1 capa fue el algoritmo con mejor rendimiento, aunque no fue capaz de capturar correctamente el comportamiento de la serie. La única arquitectura que fue capaz fue CNN, aunque sus predicciones se quedaban cortas con respecto al conjunto de pruebas. Casi todos los modelos no pudieron capturar el comportamiento de casos que ocurrió en septiembre de 2019.

C.1.3.6 Predicción de series multipaso

Estos experimentos se refieren a la predicción de múltiples unidades de tiempo de una sola variable.

Hiperparámetros de entrenamiento Para el entrenamiento de las arquitecturas para modelos multipaso se utilizaron:

- Datos de validación: 12 meses.
- Número de pasos de entrada: 6 meses.

- Numero de pasos de salida: 3 meses.
- Tasa de aprendizaje: 0.001
- Optimizador: Nesterov-accelerated Adaptive Moment Estimation.
- Función de pérdida: Raíz del error cuadrático medio.
- Épocas: 200
- Tamaño de lote: 512

Esto quiere decir que dados 6 meses de entrada, las predicciones se realizarán para los próximos 3 meses.

Perceptrón de 1 capa Se probaron 512, 1024, 2048, 4086 neuronas en la capa oculta.

Neuronas	RMSE
512	0.45391453939520277
1024	0.46403363918194385
2048	0.4518420795763441
4086	0.4535882467275815

Tabla 43: Rendimiento por arquitectura: Perceptrón de 1 capa multipaso

Podemos la pérdida en el conjunto de validación se estanca en 0.4, por lo que el modelo no es capaz de representar correctamente la serie de tiempo (**Figura 127**).

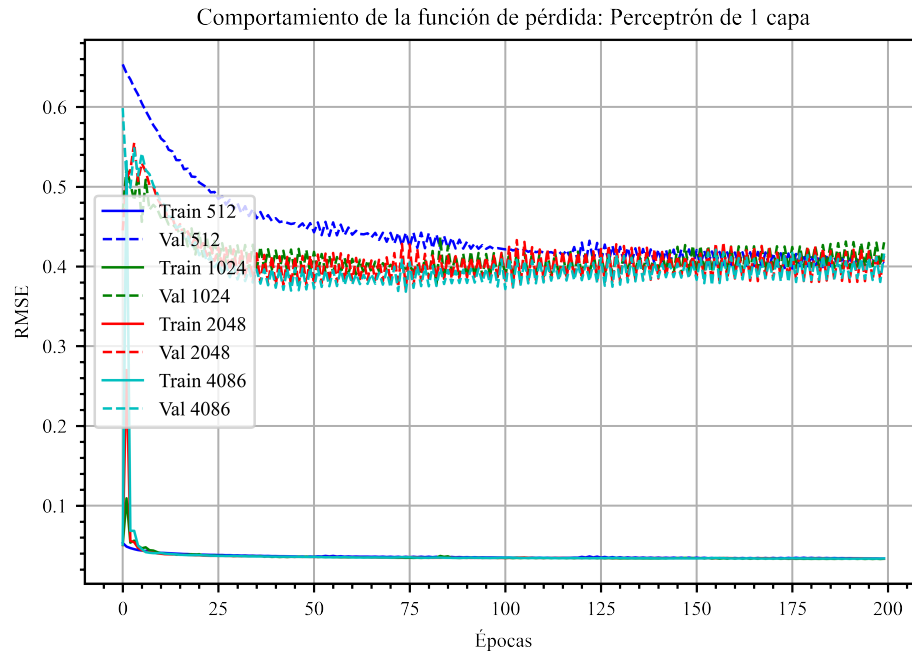


Figura 127: Comparativa de pérdida: perceptrón de 1 capa multipaso

Perceptrón de 2 capas Se usaron (512, 256), (1024, 512), (2048, 1024) neuronas para cada capa.

Neuronas	RMSE
512, 256	0.42559652933834735
1024, 512	0.425660597454121
2048, 1024	0.41025176743298136

Tabla 44: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Con un valor de 0.4, la arquitectura con (1024, 512) neuronas tuvo el mejor rendimiento (Figura 128).

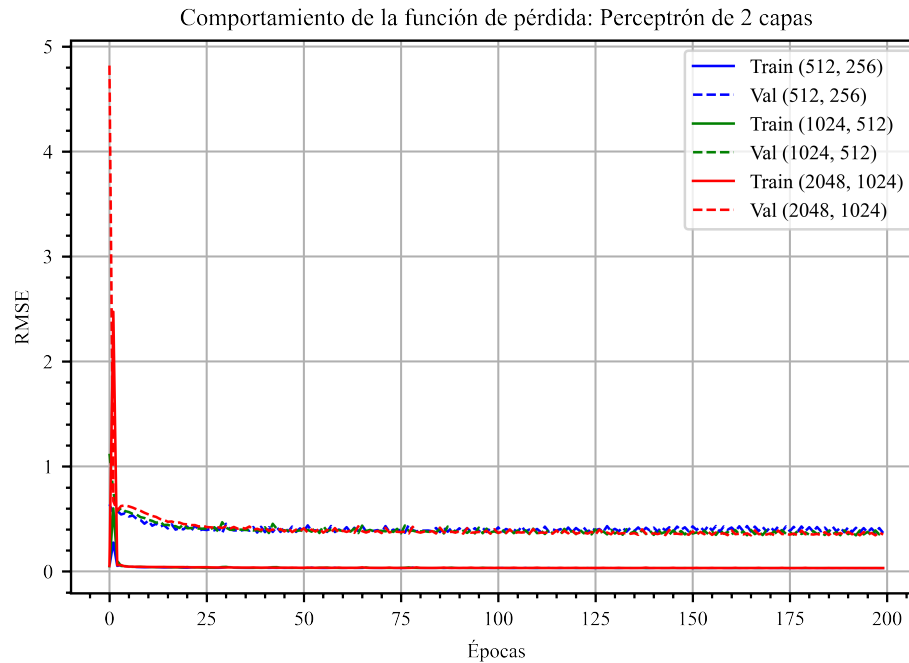


Figura 128: Comparativa de pérdida: perceptrón de 1 capa multipaso

Perceptron de 3 capas Se usaron los siguientes valores de neuronas (512, 256, 128), (1024, 512, 256), (2048, 1024, 512) para generar las arquitecturas.

Neuronas	RMSE
512, 256, 128	0.4676440760889091
1024, 512, 256	0.44931998548666313
2048, 1024, 512	0.4167934522083748

Tabla 45: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Estas arquitecturas no tuvieron un mejor rendimiento que las anteriores (Figura 129).

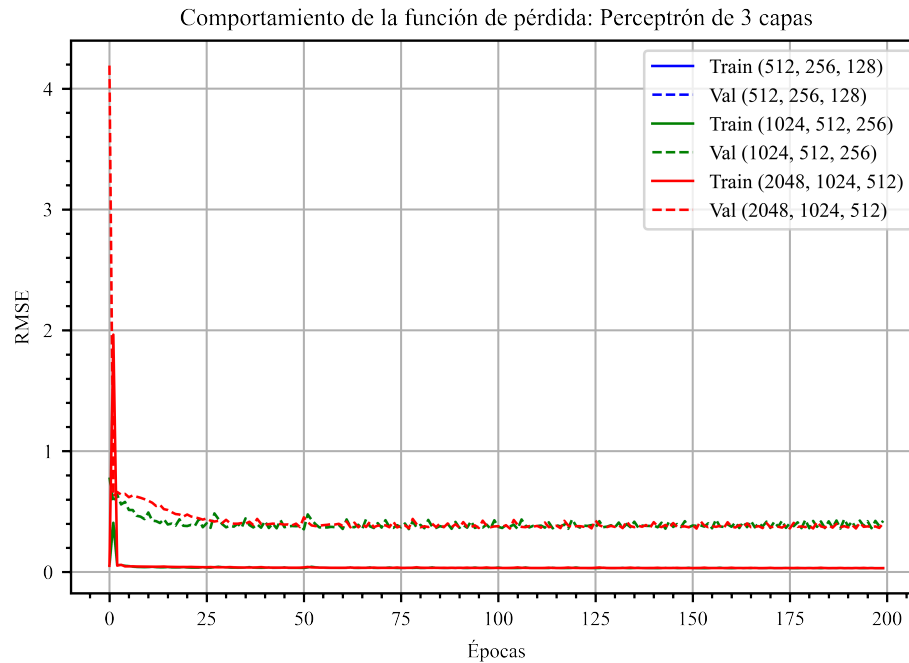


Figura 129: Comparativa de pérdida: perceptrón de 1 capa multipaso

Red neuronal convolucionada de 1 capa En la tarea anterior esta fue uno de los modelos neuronales con mejor rendimiento. Para estos experimentos se usaron 128, 256, 512, 1024, 2048, 4096 filtros en la capa 1D.

Filtros	RMSE
128	0.4798883801936636
256	0.4693336579534952
512	0.4694775167898302
1024	0.47691469919224166
2048	0.4749426934366457
4096	0.4606187496675405

Tabla 46: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Como en los anteriores experimentos la pérdida en validación deja de crecer.

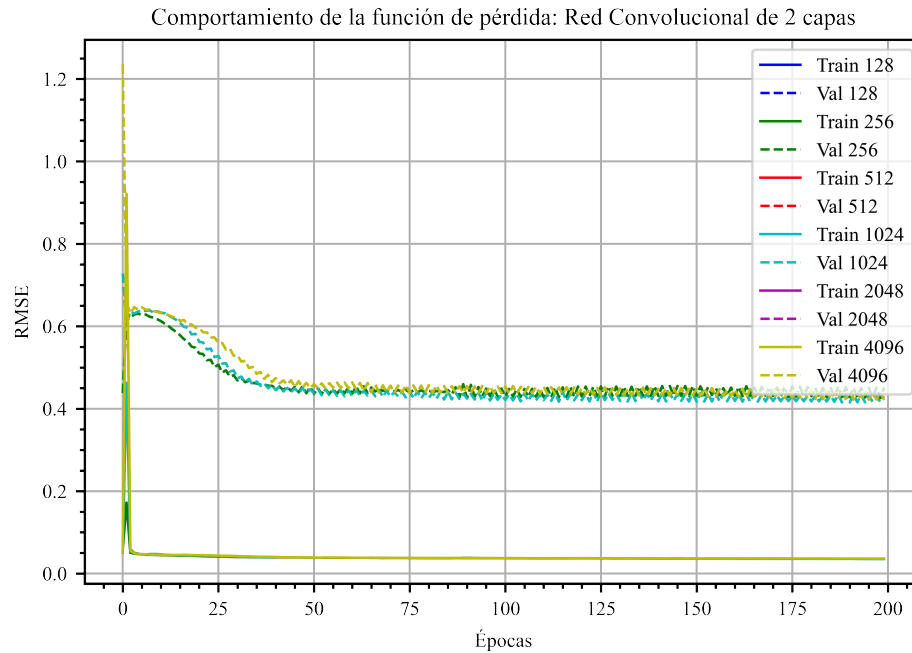


Figura 130: Comparativa de pérdida: perceptrón de 1 capa multipaso

LSTM 1 capa Para este entrenamiento se usaron 50, 100, 150, 200, 250, 300 unidades.

Unidades	RMSE
50	0.4773354495513072
100	0.4667563385889995
150	0.4655414259213768
200	0.463272620754545
250	0.4636408491798103
300	0.4602537022118665

Tabla 47: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Todas las arquitecturas tuvieron un comportamiento similar.

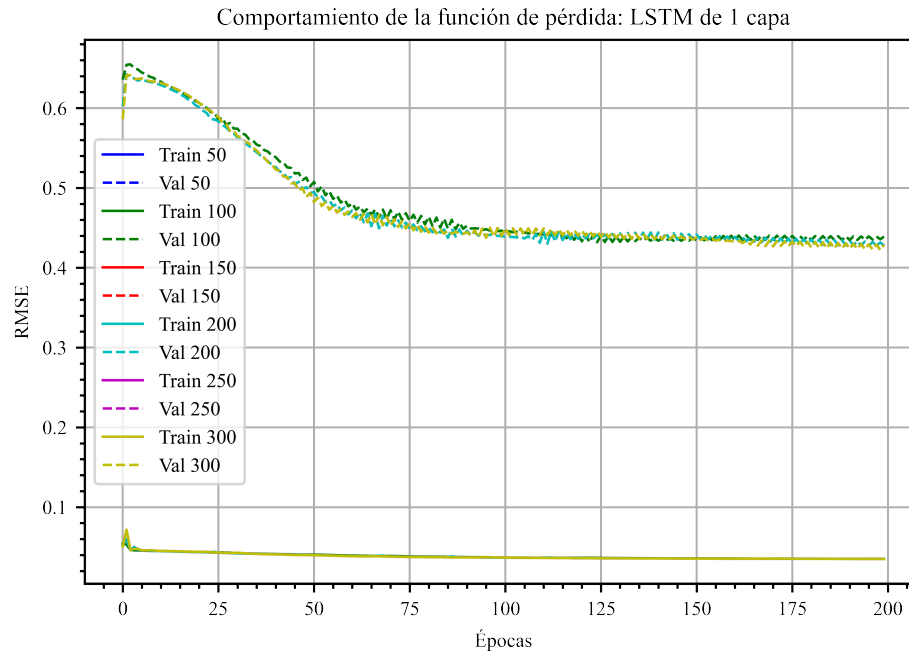


Figura 131: Comparativa de pérdida: lstm de 1 capa multipaso

LSTM 2 capas Se usaron las mismas unidades para la arquitectura bicapa.

Unidades	RMSE
50	0.466485239080832
100	0.4348202935645056
150	0.42851725876204544
200	0.47392337564551634
250	0.47085351452453006
300	0.43857006004886734

Tabla 48: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

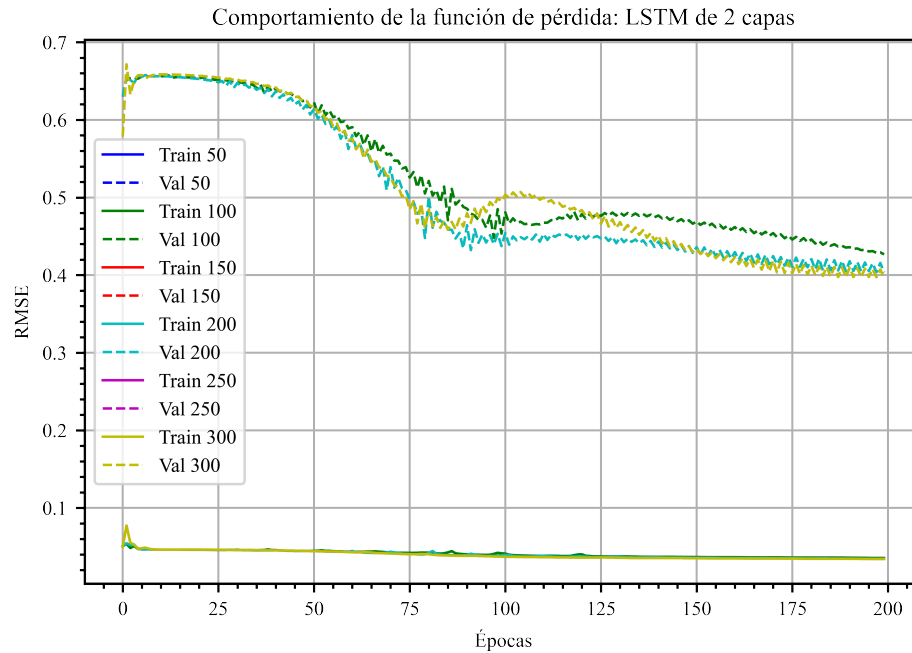


Figura 132: Comparativa de pérdida: perceptrón de 1 capa multipaso

LSTM 3 capas Se usaron las mismas unidades para la arquitectura tricapa.

Unidades	RMSE
50	0.43632648312117334
100	0.4742875653861627
150	0.46127929432665143
200	0.4632974490353326
250	0.4442014027345487
300	0.44361705970769577

Tabla 49: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Agregar más capas a los modelos de redes recurrentes no incrementa la capacidad de generalización de la red.

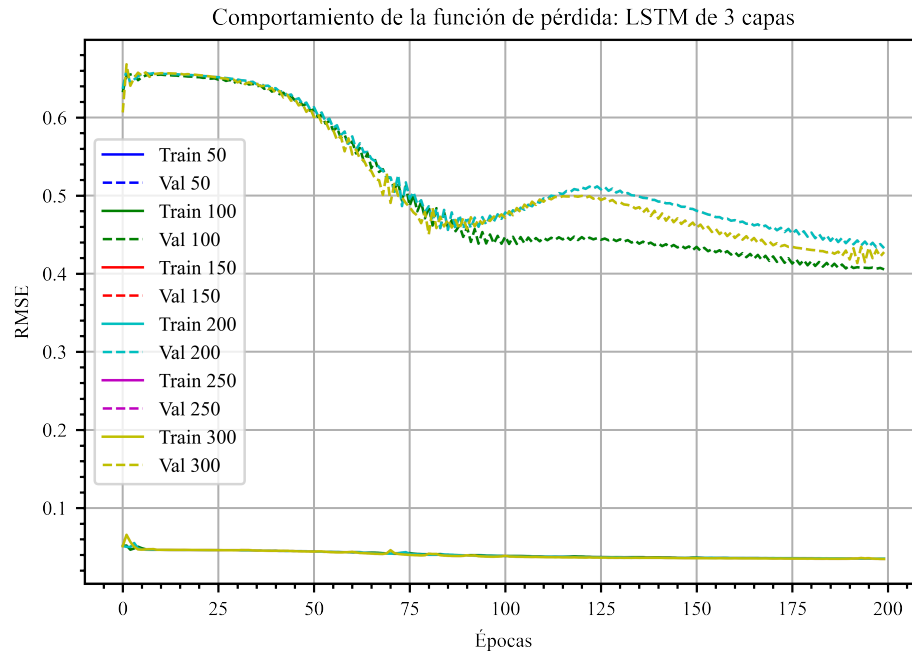


Figura 133: Comparativa de pérdida: percentrón de 1 capa multipaso

C.1.3.7 Conclusión para modelos multipaso

El rendimiento de estos modelos fue debajo de lo esperado, aunque se compensa con la dificultad de predecir varios valores a futuro en lugar de un valor discreto.

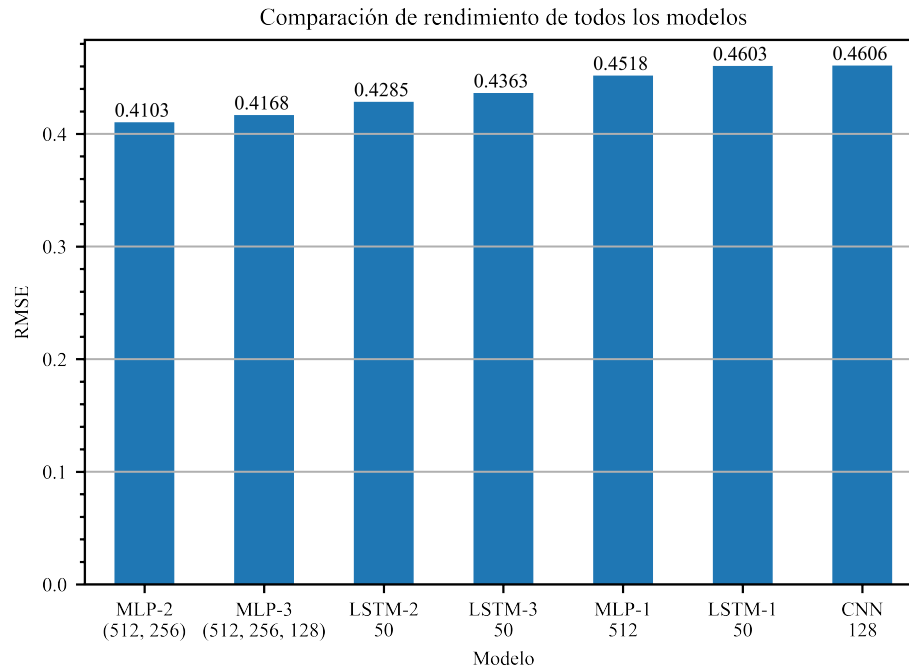


Figura 134: Comparativa para modelos multipaso

Otra vez, de todas las arquitectura, el perceptrón multicapa de dos capas, fue el modelo con mejor rendimiento. Se observa que el peor fueron las redes convolucionales. Aumentar más capas a los modelos LSTM no ayudó a mejorar el rendimiento para la predicción de casos de dengue.

C.1.3.8 Predicción de series multivariable

Estos experimentos se refieren a la predicción de múltiples variables a un solo horizonte temporal.

Fecha	Casos	Precip	Tmax	Tmed	Tmin	Año	Mes
1985-01-01	6	22.788344	23.231456	17.687719	12.143982	1985	January
1985-02-01	6	39.761199	24.168778	18.919350	13.669923	1985	February
1985-03-01	6	31.119475	27.954705	22.238870	16.523035	1985	March
1985-04-01	8	79.281711	29.761278	23.806926	17.852574	1985	April
1985-05-01	5	55.414756	31.719499	25.531910	19.344320	1985	May

Tabla 50: Datos para series multivariable

Hiperparámetros de entrenamiento Para el entrenamiento de las arquitecturas para modelos multipaso y multivariable se utilizaron:

- Datos de validación: 12 meses.
- Número de pasos de entrada: 6 meses.
- Tasa de aprendizaje: 0.001
- Optimizador: Nesterov-accelerated Adaptive Moment Estimation.
- Función de pérdida: Raíz del error cuadrático medio.
- Épocas: 200
- Tamaño de lote: 512

Esto quiere decir que dados 6 meses de entrada, las predicciones se realizarán para el siguiente mes de las tres series de tiempo.

Perceptrón de dos capas Como en el proceso anterior, se probaron distintas arquitecturas de perceptrones multicapa. Se probaron [(512, 256), (1024, 512), (2048, 1024)] neuronas para las capas ocultas.

Neuronas	RMSE
512, 256	0.04847761389618609
1024, 512	0.08946642313471759
2048, 1024	0.1057068512762241

Tabla 51: Rendimiento por arquitectura: Perceptrón de 2 capas multipaso

Se puede observar que tanto el conjunto de entrenamiento como el conjunto de prueba convergen de manera bastante cercana, esto indica un buen rendimiento del modelo para series de tiempo multivariadas.

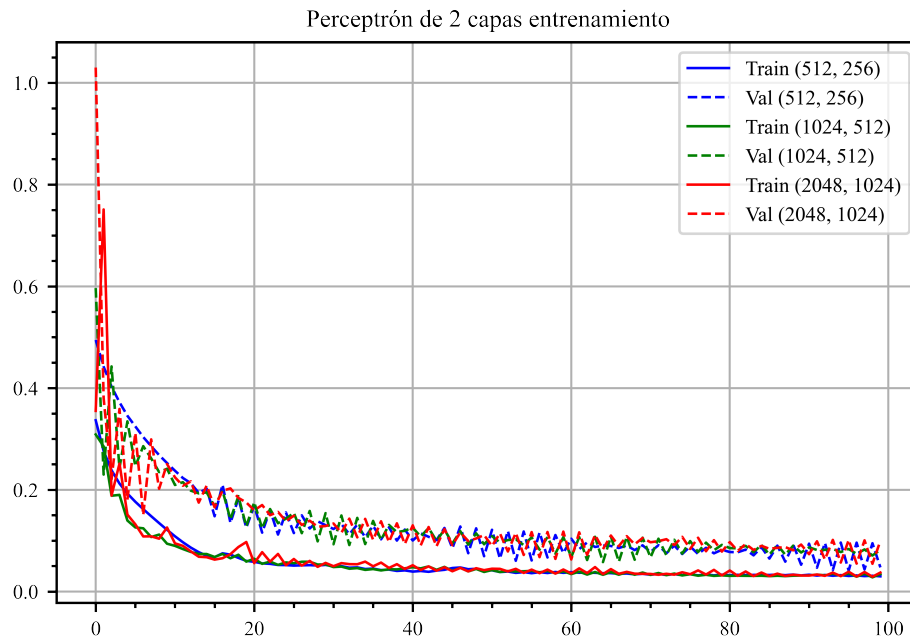


Figura 135: Comparativa de pérdida: perceptrón de 2 capas multivariado

LSTM de 2 capas Para este entrenamiento se usaron 100, 125, 150, 175, 200 unidades.

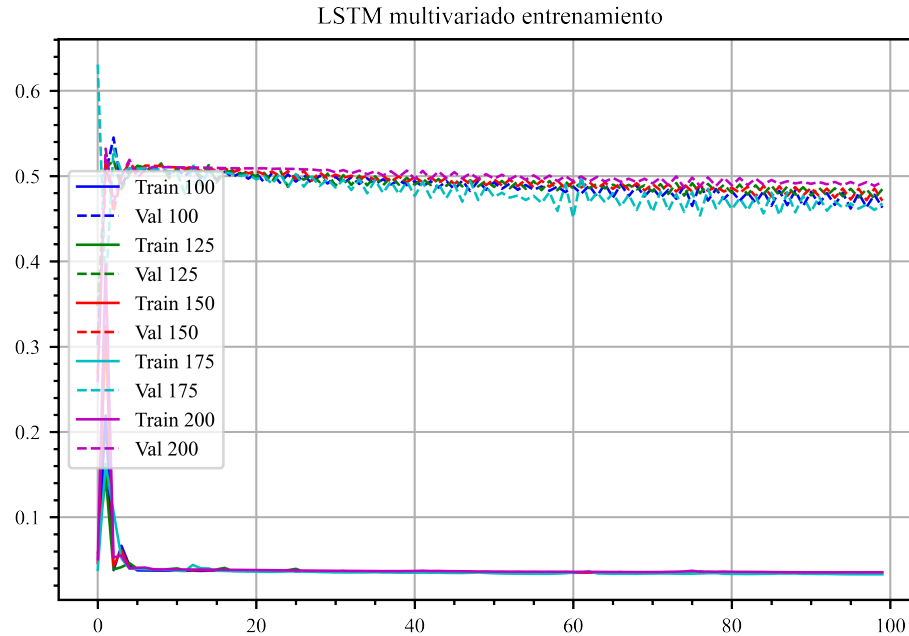


Figura 136: Comparativa de pérdida: LSTM de 2 capas multivariado

Conclusión para modelos multivariable El rendimiento con los perceptrones multicapa fue bastante aceptable, faltan realizar evaluaciones y gráficas pertinentes para determinar correctamente el rendimiento para este problema. Las arquitecturas LSTM sorprendentemente no se comportaron lo suficientemente bien como para ser implementadas para el problema del dengue.

C.1.3.9 Predicción de series multivariable y multipaso

Estos experimentos se refieren a la predicción de múltiples unidades de tiempo de múltiples variables.

Hiperparámetros de entrenamiento Para el entrenamiento de las arquitecturas para modelos multipaso y multivariable se utilizaron:

- Datos de validación: 12 meses.
- Número de pasos de entrada: 6 meses.

- Número de pasos de salida: 3 meses.
- Tasa de aprendizaje: 0.001
- Optimizador: Nesterov-accelerated Adaptive Moment Estimation.
- Función de pérdida: Raíz del error cuadrático medio.
- Épocas: 200
- Tamaño de lote: 512

Esto quiere decir que dados 6 meses de entrada, las predicciones se realizarán para los siguientes tres meses de las tres series de tiempo.

Perceptrón de dos capas Se probaron (512, 256), (1024, 512), (2048, 1024) neuronas para las capas ocultas como en el experimento multivariado.

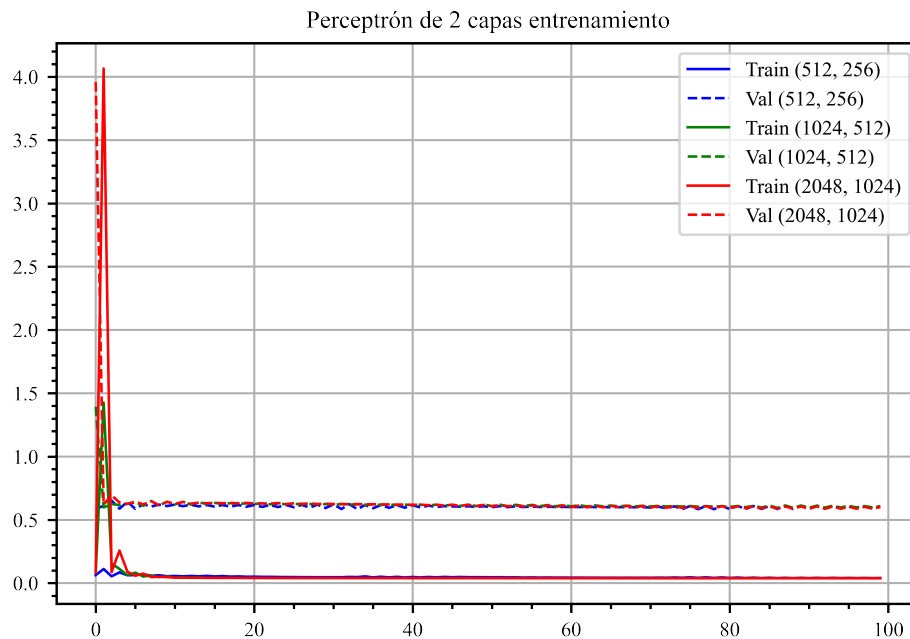


Figura 137: Comparativa de pérdida: perceptrón de 2 capas multipaso y multivariable

EL rendimiento del perceptrón en este problema fue consistente pero los conjuntos no convergen de la manera en que se esperaría. Se puede decir que los modelos de perceptrón trabajan

bien cuando se trata de múltiples variables pero su rendimiento se reduce al tratar de predecir horizontes temporales amplios.

Como podemos observar en esta comparativa de los vectores de salida de prueba y el conjunto base, el modelo está infraestimando los casos de dengue, la precipitación y la temperatura mínima, pero es capaz de capturar levemente el comportamiento incremental de las series de tiempo.

LSTM de dos capas Para este entrenamiento se usaron 100, 125, 150, 175, 200 unidades.

El rendimiento de LSTM fue bastante pobre en el conjunto de prueba. Esto se puede observar en la comparativa de vectores de salida, donde se infravalora bastante la precipitación así como los casos de dengue.

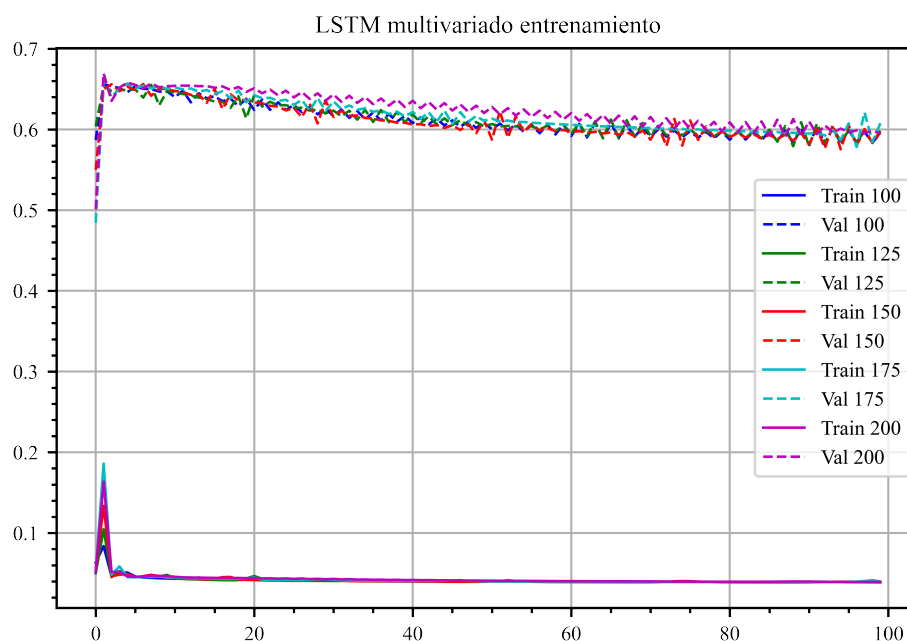


Figura 138: Comparativa de pérdida: LSTM de 2 capas multipaso y multivariable

Conclusión para modelos multivariable y multipaso Ninguna de los modelos probados y arquitecturas tuvo un rendimiento aceptable, también hace falta realizar la evaluación concreta y aplicar las métricas.

C.2 Integración de algoritmo de pronósticos

En el reporte anterior, realicé el entrenamiento de una gran variedad de algoritmos de Deep Learning y otros tipos aplicados a pronósticos y predicción de series de tiempo, tanto univariado como multivariado. Encontré que los modelos basados en Redes Neuronales no tenían un rendimiento adecuado debido a la presencia de valores atípicos en las series de tiempo, en particular, el crecimiento desmedido en casos gracias al cambio climático; estos valores atípicos inducen ruido en el modelo y evitan que capture adecuadamente el comportamiento del fenómeno, son capaces de capturar la tendencia de la serie, pero no tienen un buen rendimiento en predecir el número de casos (Moreira & Santos, 2020).

Realicé también otra exploración de algoritmos encontrando varios novedosos con resultados sumamente prometedores, algunos precisamente fueron desarrollados por grandes empresas como Facebook y Google, pude probar algunos, pero tengo que familiarizarme con ellos para poder acertar en su aplicación al problema del dengue. Algunos de estos modelos son: Facebook Prophet, BATS/TBATS, NBEATS, NHiTS, TCNModel, TFTModel, Transformer Model. Estos últimos, están basados en una nueva interpretación de la red neuronal artificial, los modelos transformadores, que están teniendo resultados sumamente prometedores en varias áreas incluidas los pronósticos y la visión por computadora. Hice pruebas con el algoritmo de Prophet, que es univariado, sin embargo, no fue capaz de determinar correctamente el número de casos (Figura 139), aunque sí pudo generar un entendimiento sobre la tendencia (Figura 140).

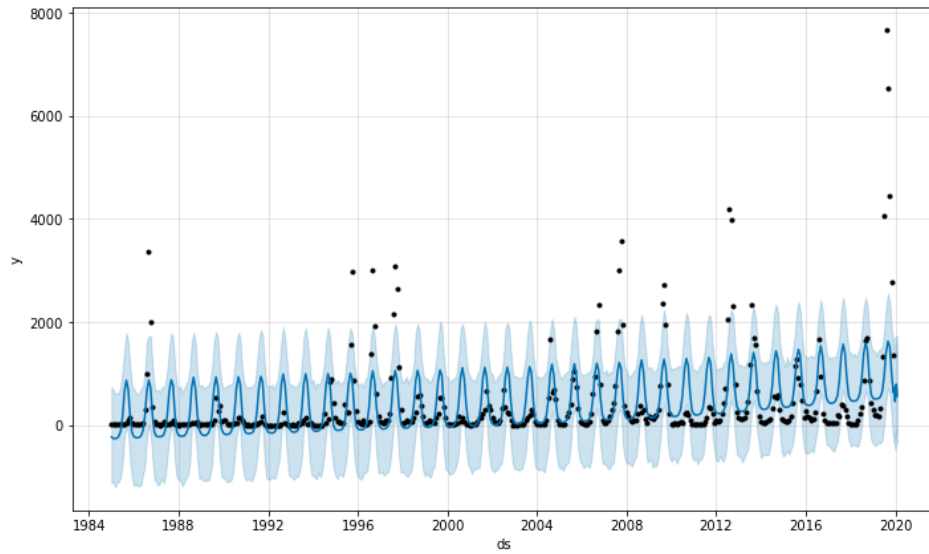


Figura 139: Modelo de Facebook Prophet

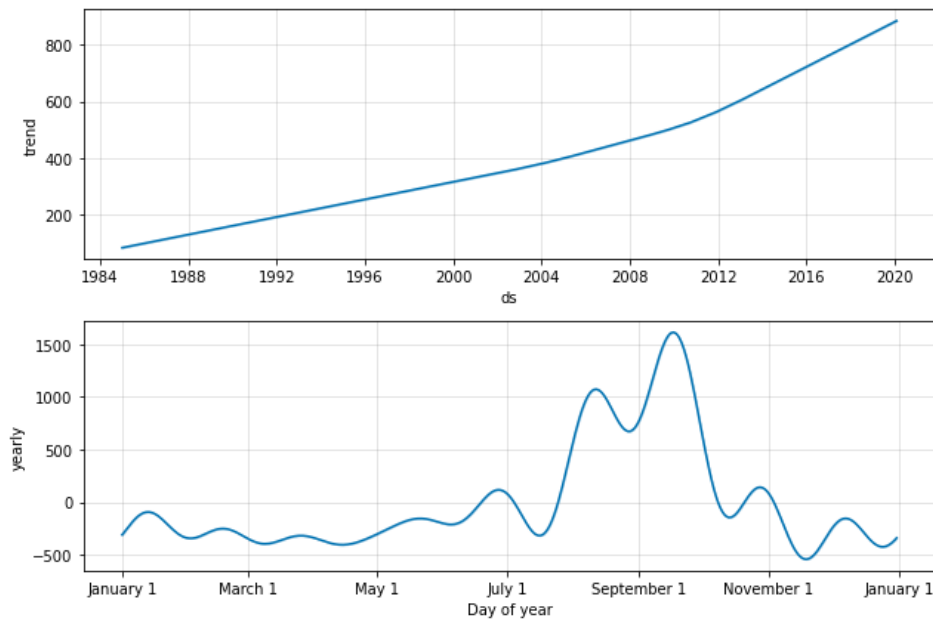


Figura 140: Tendencia y estacionalidad

Debido a la complejidad de integrar algoritmo de Deep Learning y su pobre rendimiento en el problema de pronóstico, decidí que el modelo a implementar sería una variación más poderosa del clásico modelo ARIMA para series de tiempo. La versión ocupada se llama SARIMAX; esta versión permite al modelo capturar comportamiento estacional con variables exógenas. El

rendimiento de este modelo fue adecuado para la visualización rápida del comportamiento de la serie de tiempo, también cuenta con otras ventajas. Como mencionó el Dr. Oscar en el predoctoral, se tiene que establecer un balance entre el rendimiento de un modelo, su facilidad de uso y entrenamiento, así como los recursos que se tienen que invertir para poder generarlo; todo ello en función a que su salida y rendimiento sea capaz de informar correctamente al tomador de decisiones (Thesmar et al., 2019a).

Una cuestión importante que hay que considerar para SARIMAX, es que requiere el ajuste de los siguientes parámetros: (p,d,q) , (P,D,Q,s) . Los primeros son los parámetros del modelo tradicional ARIMA u orden y los segundos son los parámetros de estacionales. Por lo tanto, se optimizarán 7 parámetros. La metodología de optimización de los mismos fue por grid search, en la cual se probaron las combinaciones posibles sin exhaustividad para encontrar aquellas que dieran el menor Error Medio Absoluto. Para la optimización, se utilizaron los criterios AIC y BIC.

La integración de SARIMAX se encuentra terminada, actualmente es capaz de entrenar el modelo y optimizar los parámetros usando la metodología expuesta dentro del prototipo. A futuro, se pretende que esta optimización se realice mediante algoritmos más poderosos de optimización como metaheurísticas o búsquedas bayesianas, que tienen la ventaja de ser más rápidas y ofrecer mejor rendimiento nominal.

C.3 Conclusiones de los experimentos

Como se pudo observar en los anteriores experimento, la tarea de predecir el número de casos de dengue es bastante compleja. Primero se trató de replicar los resultados de trabajos previos analizados en el segundo semestre y se encontraron bastantes discrepancias entre los trabajos y los experimentos. Una de las causas de esta discrepancia es que el año 2019 fue un valor extremo en las tres series de tiempo ya que tuvo una cantidad bastante más elevada de casos que en los años anteriores, lo cual introdujo cierto ruido a los modelos predictivos; cuando se realizaron los mismos experimentos eliminando el año 2019, se tuvo muchísimo mejor rendimiento.

Otra discrepancia es que, en los trabajos con mejor rendimiento, se utilizaron bases de datos externas de origen privado pero accesibles al público, por ejemplo, la base de datos de dengue de google generada por la búsqueda de ciertas palabras clave analizadas por región, lamentablemente, esta base de datos ya no se encuentra actualizada por lo que no sirve para crear un modelo robusto de uso continuo.

Uno de los hallazgos más importantes es que el rendimiento de los algoritmos basados en el Perceptrón tuvieron un rendimiento bastante bueno, esto es importante, ya que son algoritmos más probados, requieren menos memoria, recursos computacionales y son relativamente fáciles y rápidos de entrenar. Estas características los hacen candidatos para ser implementados dentro del sistema. Aquí, se tiene que realizar un análisis costo-beneficio entre las necesidades de los expertos y la facilidad de generar modelos. Se pueden generar algunas preguntas como: ¿Qué es mejor, entrenar múltiples algoritmos para distintas evaluaciones o entrenar un solo algoritmo sumamente grande que se actualize a si mismo?

También se descartaron los algoritmos de Machine Learning ya que requieren de fases exhaustivas de ingeniería de características para ser usados de manera efectiva, sin que esto asegure un correcto rendimiento.

Aún quedan por explorar múltiples arquitecturas para cada una de las tareas, por ejemplo, redes ConvLSTM o Encoder-Decoder, también se tiene que trabajar en los métodos de evaluación. También se pueden ajustar otros hiperparámetros dentro de cada capa, aumentar o disminuir su número, agregar otras capas complementarias para integrar regularización para reducir el sobreajuste, como Dropout o BatchNormalization.

Estos experimentos fueron bastante enriquecedores para dilucidar el comportamiento del problema a solucionar, se generó entendimiento sobre como las arquitecturas, capas e hiperparámetros inciden positiva o negativamente en el rendimiento de los modelos. A futuro, esto es un conocimiento invaluable ya que se pretende automatizar lo más posible el proceso para que los modelos se generen y entrenen de la manera más transparente posible y puedan ser utilizados por el usuario final.

Otro problema es con el conjunto de datos, ya que son pocos para que algoritmos con mayor

capacidad puedan ser entrenados sin sobreajustarse. Esto se podrá mejorar teniendo datos más granulares, por ejemplo, semanales o inclusive diarios. Así mismo, la resolución espacial es bastante amplia, por lo que sería mejor tener múltiples modelos pequeños, entrenados para cada horizonte temporal y espacial.