EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Aplicación de Técnicas de Reducción de la Dimensionalidad en Datos de Anticuerpos del Virus SARS-CoV-2

presentada por

ING. Samuel Isaí Narciso Galván

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Directora de tesis

Dra. María Yasmín Hernández Pérez

Codirector de tesis

Dr. Javier Ortiz Hernández

Cuernavaca, Morelos, México. Noviembre de 2024.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación
y Desarrollo Tecnológico
Dirección

Cuernavaca, Mor., 29/Octubre/2024

OFICIO No. DCC/239/2024

Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **SAMUEL ISAI NARCISO GALVÁN** con número de control, **M22CE051**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"APLICACIÓN DE TÉCNICAS DE REDUCCIÓN DE LA DIMENSIONALIDAD EN DATOS DE ANTICUERPOS DEL VIRUS SARS-COV-2"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

María Yasmín Hernández Pérez
Directora de tesis

Javier Ortiz Hernández
Codirector de tesis

Alicia Martínez Rebollar
Revisor 1

Eddie Helbert Clemente Torres
Revisor 2

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación
y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca, Mor., 31/octubre/2024
No. De Oficio: SAC/339/2024
Asunto: Autorización de impresión de tesis

SAMUEL ISAI NARCISO GALVÁN
CANDIDATO AL GRADO DE MAESTRO
CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"APLICACIÓN DE TÉCNICAS DE REDUCCIÓN DE LA DIMENSIONALIDAD EN DATOS DE ANTICUERPOS DEL VIRUS SARS-COV-2"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica·
"Conocimiento y Tecnología al Servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

Dedicatoria

A mi hermano Raúl Omar, por ser mi inspiración y motivación constante. Tu apoyo incondicional y tu ejemplo de perseverancia y superación me han impulsado a alcanzar esta meta, recordándome siempre el valor del esfuerzo y la dedicación.

A mis padres, Catalina y Raúl, quienes, con su amor, paciencia y guía, han sido el cimiento de cada paso en mi vida. Gracias por creer en mí y por enseñarme el verdadero significado de la dedicación y el compromiso. Esta tesis es un reflejo de todo lo que me han enseñado y del amor que siempre me han brindado.

Agradecimientos

Deseo expresar mi más sincero agradecimiento a todas las personas que, de diversas formas, contribuyeron al desarrollo de este trabajo y a la culminación de mis estudios de posgrado.

En primer lugar, agradezco profundamente a mi hermano Raúl Omar, cuyo apoyo y motivación fueron el impulso necesario para iniciar mis estudios de posgrado en el Centro Nacional de Investigación y Desarrollo Tecnológico. Su inquebrantable aliento fue esencial para la exitosa culminación de mi maestría, y sin su respaldo incondicional, este logro no habría sido posible.

Extiendo mi más sentido agradecimiento a mis padres, Catalina y Raúl, cuyo apoyo constante, guía y preocupación por mi bienestar han sido pilares fundamentales en mi vida académica y personal. Sin su apoyo incondicional, alcanzar el título de Maestro en Ciencias de la Computación habría sido una meta inalcanzable.

Agradezco de manera especial a la doctora María Yasmín Hernández Pérez, mi directora de tesis, quien me apoyó en todo momento durante mis estudios de posgrado. Su guía, paciencia y dedicación fueron invaluables en cada etapa de este proceso, y su constante apoyo hizo posible que superara cada obstáculo con determinación y confianza.

De igual manera, agradezco a mis amigos, colegas, docentes y mentores, quienes, durante el transcurso de mis estudios y aún después de ellos, me brindaron su guía, conocimientos y apoyo en aspectos tanto personales como profesionales. Su impacto en mi vida permanecerá siempre en mi memoria y en mi corazón.

Agradezco también al Tecnológico Nacional de México, TecNM, por brindarme la oportunidad de cursar mis estudios de posgrado y por facilitar las instalaciones de su Centro Nacional de Investigación y Desarrollo Tecnológico, CENIDET, las cuales fueron esenciales para el desarrollo de mi formación académica y profesional.

Finalmente, expreso mi más profundo agradecimiento al Consejo Nacional de Humanidades, Ciencias y Tecnologías, CONAHCyT, por la beca otorgada durante mis dos años de estudios de maestría, la cual fue fundamental para la realización de esta etapa de mi formación.

Resumen

Esta tesis examina el impacto de las técnicas de reducción de dimensionalidad en la clasificación de anticuerpos contra el SARS-CoV-2, utilizando datos de secuencias genéticas de anticuerpos disponibles en la base de datos *Observed Antibody Space*, OAS.

En particular, se enfoca en la transformación de secuencias de aminoácidos, específicamente de la región determinante de la complementariedad, CDR3, en representaciones vectoriales, *word embeddings,* para su posterior procesamiento en modelos de aprendizaje automático. Esta transformación permite el uso de datos no etiquetados y de alta dimensionalidad, pero plantea el desafío de la maldición de la dimensionalidad, la cual puede afectar la precisión y eficiencia de los modelos predictivos.

Para abordar este problema, se aplican y evalúan dos técnicas de reducción de dimensionalidad: *Principal component analysis* PCA y *Uniform Manifold Approximation and Projection*, UMAP. La investigación desarrolla 36 modelos de clasificación utilizando algoritmos de Máquinas de Soporte Vectorial, Bosques Aleatorios y K Vecinos Más Cercanos, probando cada uno en conjuntos de datos originales y en conjuntos reducidos por PCA y UMAP. Se buscó determinar si la reducción de dimensionalidad mejora el rendimiento de los modelos en términos de precisión, eficiencia y generalización en el contexto de la clasificación de anticuerpos.

Los resultados obtenidos se analizan para identificar el algoritmo y técnica de reducción de dimensionalidad más efectivos para el conjunto de datos de anticuerpos. La investigación concluye con recomendaciones sobre el uso de PCA y UMAP en tareas de clasificación de secuencias de anticuerpos, ofreciendo recomendaciones sobre cómo estas técnicas pueden facilitar el análisis predictivo en bioinformática y contribuir al desarrollo de modelos eficientes para la identificación de anticuerpos relevantes en inmunología.

Abstract

This thesis examines the impact of dimensionality reduction techniques on the classification of SARS-CoV-2 antibodies, utilizing genetic sequence data from the Observed Antibody Space, OAS, database.

Specifically, it focuses on transforming amino acid sequences from the Complementarity Determining Region, CDR, into word embeddings for subsequent processing in machine learning models. This transformation enables the use of unlabeled, high-dimensional data but presents the challenge of the curse of dimensionality, which can affect the models' accuracy and efficiency.

To address this problem, two dimensionality reduction techniques are applied and evaluated: Principal Component Analysis, PCA, and Uniform Manifold Approximation and Projection, UMAP. The research develops 36 classification models using Support Vector Machines, Random Forests, and K-Nearest Neighbors algorithms, testing each on original datasets and on reduced datasets with PCA and UMAP. The objective is to determine whether dimensionality reduction improves model performance in terms of accuracy, recall, and generalization within the context of antibody classification.

The results are analyzed to identify the most effective algorithm and dimensionality reduction technique for the antibody dataset. The study concludes with recommendations on the use of PCA and UMAP for antibody sequence classification, providing insights into how these techniques can facilitate predictive analysis in bioinformatics and contribute to the development of efficient models for identifying relevant antibodies in immunology.

Tabla de contenido

Lista de figuras

Lista de tablas

Presentación

En esta tesis se investigó la utilidad de aplicar técnicas de reducción de dimensionalidad, específicamente Análisis de Componentes Principales y Uniform Manifold Approximation and Projection, en el desarrollo de modelos predictivos a partir de conjuntos de datos de anticuerpos del SARS-CoV-2. El objetivo principal fue evaluar si estas técnicas mejoran el rendimiento de los modelos de clasificación en términos de precisión y eficiencia, dado que los conjuntos de datos originales presentan alta dimensionalidad debido a la complejidad de las secuencias de aminoácidos.

Capítulo 1 Se presenta el problema principal de la investigación, los objetivos, el alcance del trabajo y las limitaciones que enfrenta la tesis. Se establece el contexto y la relevancia del análisis de secuencias de anticuerpos del SARS-CoV-2.

Capítulo 2 Este capítulo ofrece una explicación detallada de los conceptos clave necesarios para comprender el trabajo, incluyendo la reducción de dimensionalidad, el procesamiento de lenguaje natural, y los fundamentos de los anticuerpos y sus secuencias genéticas.

Capítulo 3 Aquí se revisan los avances más recientes de otros investigadores en temas relacionados con la clasificación de anticuerpos y la aplicación de técnicas de reducción de dimensionalidad. Se describe el contexto teórico y práctico de las herramientas y metodologías utilizadas en esta tesis.

Capítulo 4 Se describen los experimentos realizados para construir 36 modelos de clasificación utilizando algoritmos de Máquinas de Soporte Vectorial, Bosques Aleatorios y K Vecinos Más Cercanos. Se explica en detalle cómo se construyeron los conjuntos de datos, aplicando técnicas de PLN y *word embeddings* para transformar secuencias de aminoácidos en datos numéricos procesables.

Capítulo 5 En este capítulo se analizan los resultados obtenidos de los modelos, comparando el desempeño de cada uno según el tipo de conjunto de datos utilizado, sin reducción de dimensionalidad, con PCA y con UMAP. Se identifica el algoritmo de clasificación más efectivo para cada tipo de conjunto.

Capítulo 6 Se da una conclusión de lo logrado en esta investigación y se detalla el mejor modelo de clasificación construido, destacando su rendimiento y la configuración de datos que resultó ser más eficiente.

Capítulo 1

Introducción

“'La biología permite, la cultura prohíbe”

Sapiens: a brief history of humankind (Harari, 2015).

1. Introducción

La minería de datos es un campo multidisciplinario que abarca diversas áreas de investigación y aplicación, como el análisis predictivo, el modelado estadístico, la extracción de patrones y la bioinformática. La minería de datos, junto con el aprendizaje automático, ha sido fundamental para resolver tareas complejas en ciencias biológicas, tales como la predicción de estructuras proteicas, la identificación de genes asociados a enfermedades y el análisis de expresión génica (Zaki et al., 2003).

Dentro de la bioinformática, la minería de datos ha ganado relevancia al permitir el análisis profundo de los datos genómicos, lo que facilita la comprensión de procesos biológicos a nivel molecular. Una de las áreas emergentes más desafiantes es la clasificación de anticuerpos en función de los antígenos a los que se dirigen. Los anticuerpos, componentes clave del sistema inmunitario, son proteínas que se unen a antígenos específicos para neutralizar patógenos, y la clasificación de estos anticuerpos es crucial para el desarrollo de terapias y vacunas.

Para abordar la clasificación de anticuerpos, es necesario contar con conjuntos de datos que proporcionen información tanto estructural como funcional de los anticuerpos. Una fuente clave de datos en este ámbito es *The Observed Antibody Space,* OAS, una base de datos que recopila secuencias genómicas de anticuerpos de diversas especies, incluidas aquellas de relevancia clínica, como humanos y murciélagos, en el contexto de enfermedades como el VIH, ébola y SARS-CoV-2. OAS incluye una enorme cantidad de información sobre la estructura de los anticuerpos, como las cadenas pesadas y ligeras, lo que la convierte en un recurso valioso para el análisis biológico y la investigación en inmunología.

Sin embargo, aunque los datos de OAS son ricos y detallados, su formato no es inmediatamente apto para técnicas de aprendizaje automático, ya que las secuencias genómicas están compuestas por aminoácidos y nucleótidos, en lugar de datos numéricos. Además, los conjuntos de datos no incluyen directamente etiquetas que clasifiquen los anticuerpos en función de los antígenos que combaten. Para hacer que estos datos sean útiles en los modelos de aprendizaje automático, es necesario filtrar la información irrelevante y centrarse en las partes más críticas del anticuerpo, en particular las regiones determinantes de la complementariedad, CDR, por sus siglas en inglés.

Las CDR son regiones clave en los anticuerpos, responsables de la unión al antígeno. Entre ellas, la CDR3 es la región más variable y la que proporciona mayor información relevante para la detección de antígenos (Xu & Davis, 2000). Dado su rol crucial, la región CDR3 es de especial interés para el análisis computacional y la clasificación de anticuerpos. Sin embargo, para que estas secuencias de aminoácidos puedan ser procesadas mediante algoritmos de aprendizaje automático, es necesario transformarlas en representaciones numéricas.

Originalmente desarrolladas para el análisis de texto, las técnicas de PLN han sido adaptadas para el análisis de secuencias biológicas. En este contexto, las secuencias de aminoácidos de la región CDR3 pueden transformarse en *word embeddings*, una representación vectorial densa que captura las relaciones semánticas entre secuencias, similar a cómo las palabras en PLN capturan significado en textos.

Sin embargo, la transformación de secuencias biológicas en *word embeddings* presenta un desafío adicional: la alta dimensionalidad de los embeddings generados, que suelen tener hasta 100 dimensiones. Este nivel de dimensionalidad puede aumentar la complejidad computacional y reducir la capacidad de los modelos de aprendizaje automático para generalizar. Para abordar este problema, es necesario aplicar técnicas de reducción de dimensionalidad, que permiten conservar únicamente las dimensiones más relevantes, reduciendo el ruido y optimizando el rendimiento del modelo sin sacrificar información crítica.

En este documento, se propone un enfoque para procesar los datos de anticuerpos derivados de OAS, específicamente aquellos relacionados con el SARS-CoV-2, transformando las secuencias de la región CDR3 en *word embeddings*. Además, se evaluará el impacto de la reducción de dimensionalidad sobre el desempeño de los modelos de clasificación, comparando el rendimiento de los modelos entrenados con datos originales frente a aquellos con datos reducidos. Esta evaluación permitirá determinar la efectividad de las técnicas de reducción de dimensionalidad para mejorar la precisión y eficiencia de los modelos predictivos en el análisis de anticuerpos.

1.1. Planteamiento del problema

En la investigación de anticuerpos y su clasificación, el enfoque tradicional para construir modelos de aprendizaje automático se basa en conjuntos de datos previamente etiquetados y preparados por expertos. Estos conjuntos de datos contienen información numérica que facilita la integración directa con algoritmos de clasificación, lo que permite desarrollar modelos precisos y eficientes para la identificación y predicción de respuestas inmunitarias. Sin embargo, este enfoque tiene limitaciones importantes, ya que depende de la disponibilidad de conjuntos de datos etiquetados y estructurados, lo que no siempre es posible en estudios emergentes o cuando se trabaja con nuevos patógenos, como es el caso de SARS-CoV-2.

Este trabajo se enfrenta a un problema fundamental, ¿cómo construir modelos de clasificación de anticuerpos a partir de conjuntos de datos no etiquetados?, los cuales están compuestos en su mayoría por secuencias génicas de aminoácidos, los cuales no están inmediatamente listos para ser utilizados en modelos de aprendizaje automático. Las secuencias biológicas, como las que se encuentran en la base de datos de OAS, contienen una gran cantidad de información sobre los anticuerpos, incluyendo las secuencias CDR. Sin embargo, estas secuencias no están etiquetadas ni en un formato numérico adecuado para modelos de clasificación directa.

Para resolver este problema, se propone transformar estas secuencias biológicas en *word embeddings* mediante técnicas de procesamiento de lenguaje natural (PLN). Esta transformación permite convertir las secuencias de aminoácidos en vectores numéricos, haciendo que los datos sean utilizables por los modelos de aprendizaje automático. No obstante, este enfoque introduce un nuevo desafío, los *word embeddings* generados suelen ser de alta dimensionalidad, específicamente, de hasta 100 dimensiones, lo que complica el procesamiento y análisis eficiente de los datos.

La alta dimensionalidad de los *word embeddings* puede afectar negativamente el rendimiento de los modelos de clasificación, ya que aumenta la complejidad computacional y puede reducir la capacidad de generalización del modelo. Este fenómeno, conocido como *curse of dimensionality* o maldición de la dimensionalidad, afecta la precisión y la capacidad del modelo para identificar patrones significativos en los datos. Para mitigar este problema, se propone el uso de técnicas de reducción de dimensionalidad, tales como el Análisis de Componentes Principales y Uniform Manifold Approximation and Projection. Estas técnicas permiten conservar las dimensiones más relevantes de los *word embeddings*, reduciendo la complejidad de los datos sin sacrificar la información crítica.

El problema central de esta investigación, por lo tanto, radica en cómo transformar secuencias genómicas no etiquetadas en un formato adecuado para el aprendizaje automático, utilizando *word embeddings* y, al mismo tiempo, abordar el desafío de la alta dimensionalidad mediante técnicas de reducción. El objetivo es construir modelos de clasificación adecuados que puedan identificar anticuerpos según el antígeno que combaten, mejorando rendimiento y precisión del modelo a través de la reducción de dimensionalidad.

1.2. Justificación

El análisis y clasificación de anticuerpos es una tarea fundamental en la investigación biomédica y la inmunología, especialmente para el desarrollo de terapias y vacunas contra patógenos emergentes como el SARS-CoV-2. En este contexto, el uso de técnicas avanzadas de minería de datos y aprendizaje automático ha sido esencial para resolver problemas complejos en bioinformática, desde la predicción de estructuras proteicas hasta la identificación de genes y mutaciones relacionadas con enfermedades (Zaki et al., 2003). Sin embargo, la construcción de modelos de clasificación en este campo ha dependido tradicionalmente de conjuntos de datos etiquetados y estructurados, lo cual limita su aplicabilidad en escenarios emergentes, como en el caso de nuevos patógenos, donde la obtención de etiquetas precisas puede ser un proceso costoso y prolongado.

Esta investigación justifica su relevancia al proponer la aplicación de técnicas de reducción de dimensionalidad en datos de anticuerpos del SARS-CoV-2 para mejorar la construcción de modelos de clasificación a partir de secuencias genómicas no etiquetadas. Utilizando los datos ofrecidos por el *Observed Antibody Space* OAS, una de las bases de datos más completas en secuencias de anticuerpos de diversas especies, esta tesis plantea una solución para transformar estas secuencias de aminoácidos en representaciones numéricas mediante técnicas de procesamiento de lenguaje natural, PLN, como los *word embeddings.*

Esta metodología permite aprovechar secuencias no etiquetadas en modelos de clasificación, expandiendo el alcance de los datos en bioinformática y facilitando la identificación de patrones relevantes en los anticuerpos, en particular en las regiones determinantes de la complementariedad (CDR), como la CDR3, crítica en la detección de antígenos (Xu & Davis, 2000).

Aunque la transformación de secuencias en *word embeddings* facilita su uso en modelos de aprendizaje automático, introduce un desafío importante: la alta dimensionalidad de estas representaciones, que puede dificultar el entrenamiento de los modelos y limitar su capacidad de generalización (Angermueller et al., 2016). Por ello, esta tesis evalúa la aplicación de técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales y *Uniform Manifold Approximation and Projection*, para reducir la complejidad de los datos y mejorar la eficiencia de los modelos de clasificación sin perder información biológicamente relevante (McInnes et al., 2018). La reducción de dimensionalidad no solo simplifica los datos, sino que también optimiza el rendimiento de los algoritmos de clasificación, una característica esencial en contextos clínicos y de investigación, donde la rapidez y precisión son determinantes.

La importancia de esta investigación reside en ofrecer un enfoque flexible y eficiente para la clasificación de anticuerpos, prescindiendo de conjuntos de datos etiquetados y ampliando la aplicabilidad de estas metodologías en otros contextos que implican secuencias genómicas no etiquetadas, como en virología e inmunología. Al combinar técnicas de reducción de dimensionalidad con métodos de PLN, esta tesis contribuye al avance del aprendizaje automático y la minería de datos en biología computacional, facilitando el análisis de grandes volúmenes de datos biológicos de alta complejidad y promoviendo nuevas oportunidades para el estudio y clasificación de anticuerpos del SARS-CoV-2 y otros patógenos.

1.3. Objetivos

1.3.1. Objetivo general

Aplicar las técnicas de reducción de la dimensionalidad PCA y UMAP para conformar conjuntos de datos de anticuerpos del SARS-CoV-2 con menor dimensionalidad, que permitan el desarrollo de modelos de clasificación.

1.3.2. Objetivos específicos

1. Identificar características en los conjuntos de datos de anticuerpos del SARS-CoV-2 mediante un análisis de la literatura, para identificar cuál aporta más información acerca de los anticuerpos y ser utilizada en los modelos de clasificación.
2. Evaluar las técnicas de reducción de dimensionalidad PCA y UMAP para seleccionar la adecuada que se puede usar en el conjunto de datos de anticuerpos del SARS-CoV-2.
3. Desarrollar modelos de aprendizaje utilizando conjuntos de datos con y sin dimensionalidad reducida, con el fin de determinar la eficacia de las técnicas aplicadas en el desempeño de los modelos.

1.4. Metodología de solución

En esta sección se introduce y describe la metodología utilizada para abordar y resolver el problema planteado en esta investigación. La metodología adoptada sigue un enfoque estructurado, diseñado para tratar la clasificación de secuencias génicas de anticuerpos del SARS-CoV-2 y otros virus, aprovechando técnicas de procesamiento de lenguaje natural y aprendizaje automático.

El flujo completo de la metodología se representa gráficamente en la Figura 1.1, proporcionando una vista general de cada etapa y la secuencia que se sigue para resolver el problema de clasificación de secuencias génicas de anticuerpos.

La metodología implementada en esta investigación consta de cinco etapas, en la cual entran conjuntos de datos de secuencias de anticuerpos del SARS-CoV-2 en forma de aminoácidos y se obtiene el mejor modelo predictivo.

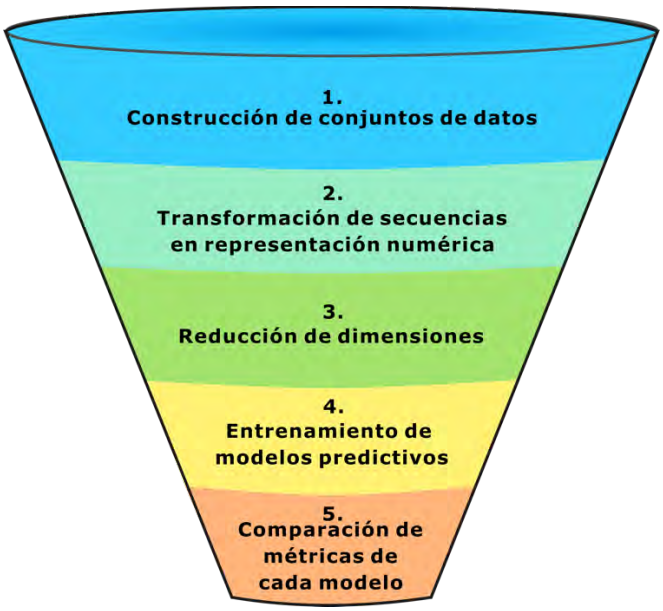

Figura 1.1 Metodología de solución

Capítulo 2

Marco teórico

2. Marco teórico

2.1. El virus del SARS-CoV-2

El virus de SARS-CoV-2 fue responsable de la pandemia de COVID-19, lo cual ha tenido un impacto profundo y global desde su aparición en diciembre de 2019. Este virus es un miembro de la familia Coronaviridae, específicamente del género Betacoronavirus, el SARS-CoV-2 comparte similitudes con otros coronavirus humanos como el SARS-CoV y el MERS-CoV, pero su alta capacidad de transmisión y su adaptabilidad ocasionaron una propagación más rápida y extensa. Este virus se caracteriza por un genoma de ARN de cadena simple y sentido positivo, que codifica cuatro proteínas estructurales principales: la espiga (Spike, S), la nucleocápside (N), la envoltura (E) y la membrana (M), además de varias proteínas accesorias y no estructurales que juegan un papel crucial en su ciclo de vida y en la evasión del sistema inmunitario del huésped (Wu et al., 2020).

La investigación sobre el SARS-CoV-2 ha sido fundamental para la comprensión de las bases moleculares de la infección viral, la patogénesis, y la respuesta inmunitaria. El estudio de este virus impulsa avances significativos en varias áreas de la ciencia, incluida la virología, la inmunología, la epidemiología, y la bioinformática. La rápida secuenciación del genoma del SARS-CoV-2, publicada poco después de la identificación del virus, permitió a los investigadores desarrollar rápidamente pruebas diagnósticas, estudiar la estructura del virus, y comenzar el diseño de vacunas (Hoffmann et al., 2020).

La investigación científica sobre el SARS-CoV-2 también ha revelado la importancia de las variantes del virus, que surgen de mutaciones en su genoma y pueden afectar la transmisibilidad y la eficacia de las respuestas inmunitarias y las vacunas. Las variantes que presentan mutaciones en la proteína Spike, como la variante delta o la variante ómicron, han mostrado diferentes grados de evasión inmunitaria, lo que ha planteado desafíos continuos en el control de la pandemia. La vigilancia genómica continua es esencial para identificar y caracterizar estas variantes, lo que subraya la importancia de herramientas avanzadas de bioinformática y aprendizaje automático para analizar los grandes volúmenes de datos genómicos generados (Harvey et al., 2021).

2.2. Aprendizaje automático

El aprendizaje automático o *machine learning* es un campo interdisciplinario que combina estadística, informática e inteligencia artificial con el objetivo de desarrollar algoritmos capaces de aprender y hacer predicciones a partir de datos. En su núcleo, el aprendizaje automático implica la creación de modelos que pueden identificar patrones complejos en los datos, realizar clasificaciones y predicciones, y mejorar continuamente su rendimiento a medida que se les proporciona más información. Este campo se divide en varias subáreas clave, cada una de las cuales aborda diferentes aspectos del aprendizaje y la toma de decisiones basados en datos.

En el aprendizaje supervisado, el algoritmo se entrena utilizando un conjunto de datos etiquetados, donde las entradas están asociadas con salidas conocidas. El objetivo es aprender una función que pueda mapear entradas nuevas a las salidas correctas. Este enfoque es ampliamente utilizado en tareas de clasificación y regresión. La clasificación se refiere a la asignación de etiquetas o categorías a los datos. Por su parte, la regresión implica predecir valores continuos, como la expresión génica, en función de ciertos marcadores biológicos.

A diferencia del aprendizaje supervisado, el aprendizaje no supervisado se utiliza cuando los datos no están etiquetados. Su objetivo es descubrir estructuras ocultas o patrones en los datos. La técnica más común es el agrupamiento. El agrupamiento o *Clustering*,

identifica grupos de datos similares, lo cual es útil para descubrir nuevas clases o subtipos en los datos genéticos.

El enfoque del aprendizaje semisupervisado combina aprendizaje supervisado y no supervisado, utilizando una pequeña cantidad de datos etiquetados junto con una gran cantidad de datos no etiquetados. Es útil en situaciones donde el etiquetado completo de los datos es costoso o difícil de obtener.

En el aprendizaje por refuerzo, un agente aprende a tomar decisiones mediante la interacción con un entorno, recibiendo recompensas o castigos según las acciones que realiza.

2.2.1. Técnicas de clasificación en aprendizaje automático

En el campo del aprendizaje automático, existen varias técnicas de clasificación que se aplican para categorizar datos en diferentes clases basadas en características aprendidas a partir de datos de entrenamiento. En la genómica, estas técnicas son esenciales para el análisis de secuencias genéticas, como las del SARS-CoV-2, ayudando a identificar patrones y realizar predicciones sobre la eficacia de respuestas inmunitarias. A pesar de que existe una amplia diversidad de técnicas de clasificación de aprendizaje automático supervisado y no supervisado, puede ser complejo saber por qué técnica decantarse, algunas características útiles a considerar en la selección de las técnicas adecuadas son, una clase definida en los datos, la linealidad de los datos, el tipo de datos usado, la interpretabilidad.

Siguiendo estas características, y a partir del análisis del estado del arte, se identificaron tres técnicas, las cuales han sido ampliamente documentadas en el uso de clasificación de anticuerpos. K-Vecinos Más Cercanos, Máquinas de soporte vectorial, y Bosque aleatorio.

K-Vecinos más cercanos, KNN por sus siglas en inglés, es un algoritmo de clasificación no supervisado que se basa en la proximidad en el espacio de características para asignar una clase a un nuevo punto de datos. El principio subyacente es que los puntos de datos que están cerca unos de otros en el espacio de características son más propensos a compartir la misma clase. Para clasificar un nuevo dato, se siguen tres pasos 1. Calcular la distancia entre el nuevo punto de datos y todos los puntos en el conjunto de entrenamiento. La distancia puede calcularse usando varias métricas, siendo las más comunes la distancia euclidiana, Manhattan, o Minkowski. 2 Identificar los k vecinos más cercanos al nuevo punto de datos, donde k es un parámetro predefinido por el usuario. 3 Asignar la clase más común entre los k vecinos al nuevo punto de datos. Este proceso se conoce como votación mayoritaria.

Máquinas de soporte vectorial, SVM por sus siglas en inglés, es un algoritmo supervisado diseñado para resolver problemas de clasificación y regresión. SVM busca un hiperplano en un espacio de alta dimensionalidad que separe los datos en diferentes clases con un margen máximo. Los pasos clave que sigue este algoritmo son 1. Definir un hiperplano que maximice el margen, es decir, la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase, conocidos como vectores de soporte. 2. Optimización de la posición del hiperplano para maximizar esta distancia del marge, lo que mejora la generalización del modelo. 3. Manejar problemas no lineales mediante kernels, SVM utiliza funciones kernel que transforman el espacio de características original en un espacio de mayor dimensionalidad donde las clases pueden ser linealmente separables.

Bosque aleatorio, RF por sus siglas en inglés (*Random Forest*), es un algoritmo de aprendizaje supervisado que opera construyendo múltiples árboles de decisión y combinando sus resultados para mejorar la precisión y reducir el riesgo de sobreajuste. El uso subyacente de este algoritmo consiste en: 1. Construcción de árboles de decisión, donde cada árbol es entrenado en un subconjunto aleatorio de los datos de entrenamiento y un subconjunto aleatorio de características. 2. Votación por mayoría para clasificar un nuevo punto de datos, cada árbol individual en el bosque realiza una predicción, y la clase final se asigna basada en la votación mayoritaria entre todos los árboles. Random Forest es altamente robusto frente al sobreajuste,

lo que lo convierte en una opción preferida en escenarios con muchos datos y características. RF es capaz de manejar datos faltantes y es menos sensible a los valores atípicos en comparación con otros métodos. Sin embargo, la interpretación de modelos RF puede ser compleja debido a la naturaleza del modelo ensamble.

2.2.2. Optimización de hiperparámetros

La optimización de hiperparámetros es el proceso de ajustar los parámetros de un modelo de aprendizaje automático que no se aprenden directamente de los datos, como el número de vecinos en KNN o el margen en SVM, para mejorar el rendimiento del modelo. Estos hiperparámetros son cruciales porque determinan la capacidad del modelo para generalizar a datos no vistos. Se utiliza en la fase de ajuste del modelo para encontrar la combinación óptima de parámetros que minimice el error de validación y evite el sobreajuste. Métodos como la búsqueda en cuadrícula, *grid search* o la búsqueda aleatoria, *random search* son comunes, aunque pueden ser computacionalmente costosos.

Tree-structured Parzen Estimator, TPE es un método bayesiano de optimización de hiperparámetros que modela la función de pérdida como una distribución probabilística, permitiendo una búsqueda más eficiente en espacios de hiperparámetros complejos en comparación con métodos tradicionales. Este método de optimización resulta especialmente útil en escenarios donde el espacio de hiperparámetros es vasto y complejo, como en modelos de aprendizaje profundo o en aplicaciones genómicas donde el ajuste fino de los hiperparámetros puede marcar una diferencia significativa en el rendimiento. TPE ha sido utilizado para optimizar modelos como KNN y RF, ajustando hiperparámetros clave para mejorar la precisión de la clasificación y la capacidad de generalización del modelo (Bergstra et al., 2011).

2.3. Conjuntos de datos genómicos de anticuerpos

Los conjuntos de datos genómicos de anticuerpos han emergido como un recurso clave en la investigación de diversos virus, dentro de los cuales se encuentra el SARS-CoV-2, proporcionando información vital para entender su estructura y desarrollar terapias y vacunas. Estos datos consisten en secuencias de genes que codifican anticuerpos, comúnmente representadas en forma de secuencias de aminoácidos.

El análisis de estos conjuntos de datos permite identificar patrones en las respuestas inmunitarias y predecir la eficacia de los anticuerpos en la neutralización del virus. Sin embargo, la naturaleza compleja y de alta dimensionalidad de estos datos presenta desafíos significativos en su manejo y análisis, especialmente cuando se usan técnicas de aprendizaje automático.

2.3.1. Características de los conjuntos de datos de anticuerpos

Los conjuntos de datos de anticuerpos contienen secuencias genéticas que han sido observadas en estudios experimentales de respuesta inmune. Estas secuencias pueden variar significativamente en longitud y estructura, lo que refleja la diversidad del repertorio de anticuerpos en los seres humanos.

La característica principal de estos datos es su alta variabilidad, ya que cada secuencia de aminoácidos puede tener cientos de posiciones, y cada posición puede estar ocupada por uno de 20 aminoácidos diferentes. Además, la naturaleza de las secuencias implica que son datos no estructurados, lo que hace que su procesamiento y análisis requieran técnicas avanzadas de bioinformática y aprendizaje automático. Tales características hacen que los conjuntos de datos de anticuerpos sean particularmente desafiantes de manejar, ya que las

relaciones entre diferentes secuencias no son triviales y requieren métodos especializados para su análisis (Raybould et al., 2019).

2.3.2. Fuentes y métodos de obtención de datos genómicos

Los conjuntos de datos genómicos de anticuerpos provienen de diversas fuentes, incluyendo estudios de secuenciación de próxima generación, NGS, experimentos de selección de anticuerpos a partir de bibliotecas combinatorias, y bases de datos especializadas.

La principal fuente de datos para esta investigación fue *The Observed Antibody Space.* OAS, la cual es una de las bases de datos más extensas y detalladas que contiene secuencias de anticuerpos. Recopila millones de secuencias de anticuerpos observadas en humanos y otros organismos, generadas principalmente a través de NGS. OAS proporciona una plataforma para la investigación de la diversidad del repertorio de anticuerpos, facilitando estudios que analizan cómo diferentes individuos o poblaciones responden a infecciones como la del SARS-CoV-2. La base de datos también permite el análisis comparativo entre diferentes estados de salud y enfermedad, proporcionando una base para entender cómo se desarrolla y mantiene la inmunidad.

The International ImMunoGeneTics Information System, IMGT, esta es otra base de datos exhaustiva que incluye información sobre secuencias inmunogenéticas, incluyendo genes de inmunoglobulinas, TCR, receptores de células T, y MHC, complejo mayor de histocompatibilidad. IMGT es ampliamente utilizada para estudios de filogenia, diversidad genética y estructura de anticuerpos (Lefranc et al., 2015).

VDJdb, es otra base de datos y aunque se centra en las secuencias de receptores de células T, TCR, es relevante para estudios inmunológicos amplios, incluidos aquellos que exploran la interacción entre TCR y epítopos virales. Dado que los TCR y los anticuerpos son componentes críticos de la respuesta inmunitaria adaptativa, VDJdb proporciona un recurso valioso para entender cómo las células T reconocen y responden a las infecciones virales, incluyendo el SARS-CoV-2. Esto es particularmente importante para el desarrollo de inmunoterapias y vacunas que puedan aprovechar la respuesta de las células T junto con la de los anticuerpos (Shugay et al., 2018).

A pesar de la riqueza y diversidad de los datos presentes en estas fuentes, un desafío común en su análisis es la necesidad de un preprocesamiento exhaustivo antes de aplicar técnicas de aprendizaje automático.

2.3.3. Manejo de conjuntos de datos de alta dimensionalidad

El manejo de conjuntos de datos de anticuerpos de alta dimensionalidad es uno de los principales desafíos en la bioinformática y en la aplicación de técnicas de aprendizaje automático. La alta dimensionalidad se refiere al gran número de características que cada secuencia de aminoácidos puede tener, lo que puede llevar a problemas como la maldición de la dimensionalidad, *curse of dimensionality*, donde el rendimiento de los modelos de aprendizaje automático se deteriora debido a la dificultad de encontrar patrones significativos en un espacio de características tan amplio.

Además, algo a tener en cuenta es que las secuencias genómicas en su forma de aminoácidos o nucleótidos no son directamente utilizables por los algoritmos de aprendizaje automático. Por esta razón, es necesario un preprocesamiento extenso, que incluye la transformación de estas secuencias en vectores numéricos mediante técnicas de procesamiento de lenguaje natural, como *word embeddings* de secuencias proteicas. Estas transformaciones permiten que los modelos de aprendizaje automático puedan procesar y analizar los datos, pero también introducen la necesidad de técnicas de reducción de dimensionalidad para manejar el tamaño y la complejidad de los vectores resultantes.

Métodos como el Análisis de Componentes Principales, t-SNE, o autoencoders son comúnmente empleados para reducir la dimensionalidad, facilitando así el descubrimiento de patrones relevantes sin perder información crítica (AlQuraishi, 2021).

2.4. Anticuerpos y antígenos

Los anticuerpos, también conocidos como inmunoglobulinas, son proteínas especializadas producidas por las células B del sistema inmunitario en respuesta a la presencia de antígenos, que en el caso del SARS-CoV-2, son principalmente las proteínas del virus, como la proteína Spike. La interacción entre anticuerpos y antígenos es fundamental para la neutralización del virus, ya que los anticuerpos se unen a epítopos específicos en la superficie del antígeno, bloqueando su capacidad de infectar las células huésped. Esta unión es altamente específica, permitiendo que los anticuerpos identifiquen y neutralicen efectivamente los patógenos, facilitando su eliminación del cuerpo.

La estructura de los anticuerpos es clave para su función. Están compuestos por dos cadenas pesadas y dos cadenas ligeras, formando una estructura en forma de Y, como se muestra en la Figura 2.1. Cada brazo de la Y contiene un sitio de unión al antígeno, compuesto por regiones variables que interactúan directamente con el epítopo del antígeno, es decir con la parte que será reconocida por un anticuerpo y a la cual se unirá. Estas regiones variables están formadas por dominios estructurales llamados región determinante de la complementariedad, Complementarity Determining Regions CDRs, que son los responsables de la especificidad del anticuerpo.

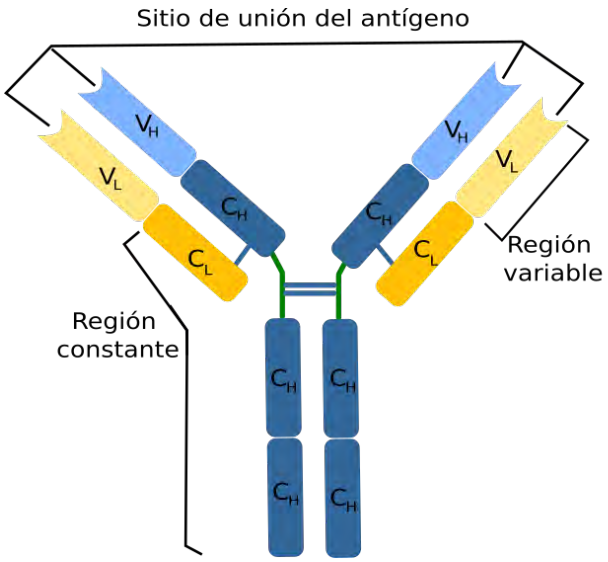

Figura 2.1. Estructura de un anticuerpo.

Como se observa en la Figura 2.2, los CDRs se dividen en tres regiones en cada cadena pesada y ligera: CDR1, CDR2 y CDR3. De estas, el CDR3 es particularmente relevante debido a su alta variabilidad y capacidad para formar estructuras que pueden interactuar con una amplia gama de antígenos. El CDR3 en las cadenas pesadas, CDR-H3 es la región más variable y juega un papel crítico en la determinación de la especificidad del anticuerpo, especialmente en la identificación de nuevos epítopos en antígenos virales como la proteína Spike del SARS-CoV-2. La longitud y la secuencia del CDR-H3 varían significativamente entre diferentes anticuerpos, lo que le permite adaptarse a la gran diversidad de antígenos que el sistema inmunitario puede encontrar.

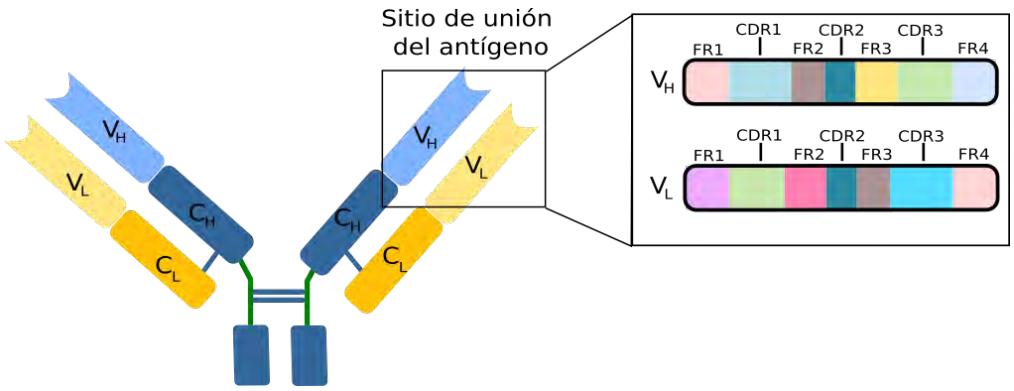

Figura 2.2. regiones de las cadenas del anticuerpo.

El CDR3 es esencial en la respuesta inmunitaria adaptativa porque su variabilidad permite la generación de un repertorio diverso de anticuerpos, capaces de reconocer y neutralizar una amplia gama de antígenos. En el contexto del SARS-CoV-2, la capacidad del CDR-H3, es decir del CDR3 perteneciente a la cadena pesada, para reconocer epítopos específicos de la proteína Spike es crucial para la eficacia de la respuesta inmunitaria. Las mutaciones en la proteína Spike pueden modificar estos epítopos, lo que potencialmente podría disminuir la capacidad de los anticuerpos de unirse eficazmente al virus y neutralizarlo. Esta capacidad de escape inmune es una de las razones por las cuales variantes como delta y ómicron han presentado desafíos significativos en la eficacia de las vacunas y las terapias basadas en anticuerpos (Greaney et al., 2021). Además de su importancia en la neutralización directa del virus, la especificidad de los anticuerpos, especialmente a través del CDR-H3, se ha aprovechado en técnicas avanzadas de identificación de antígenos.

2.5. Reducción de la dimensionalidad

La reducción de dimensionalidad es una técnica esencial en el aprendizaje automático, al igual que en el análisis de datos genómicos, ya que permite manejar la complejidad inherente a los conjuntos de datos de alta dimensionalidad, como aquellos derivados de secuencias de anticuerpos del SARS-CoV-2.

Estos datos, que pueden incluir cientos o incluso miles de variables, y presentan desafíos significativos para los modelos de aprendizaje automático, ya que la alta dimensionalidad puede conducir a problemas como la maldición de la dimensionalidad, *curse of dimensionality*, donde el rendimiento del modelo se degrada debido a la dificultad de generalizar en un espacio de características tan vasto.

La reducción de dimensionalidad aborda estos desafíos al simplificar los datos, manteniendo la mayor cantidad de información relevante posible, lo que a su vez mejora la eficacia de los modelos, facilita la visualización y permite la identificación de patrones y relaciones significativas.

Las técnicas de reducción de dimensionalidad se dividen generalmente en dos categorías principales, como se puede observar en la Figura 2.3: selección de características y extracción de características. Ambas tienen el objetivo de reducir el número de variables en el conjunto de datos, pero lo hacen de maneras distintas.

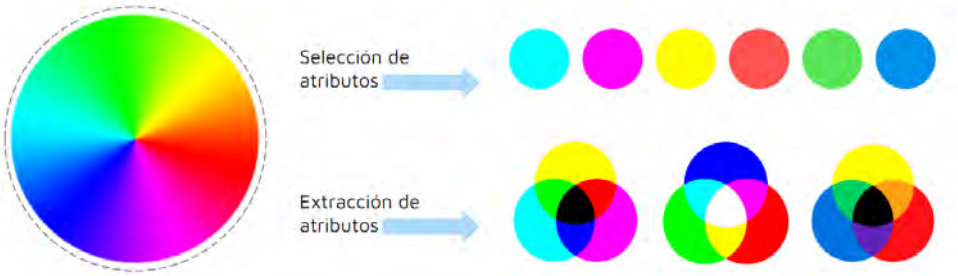

Figura 2.3. Selección y Extracción de atributos.

La selección de características implica identificar y retener un subconjunto de las características originales que son más relevantes para el problema en cuestión. Este enfoque no altera las características seleccionadas; simplemente descarta aquellas que se consideran menos importantes o redundantes. Técnicas como la selección basada en la varianza, los métodos de selección secuencial y los algoritmos basados en importancia de características.

Por otro lado, la extracción de características crea nuevas características a partir de combinaciones lineales o no lineales de las características originales, generando un conjunto de variables de menor dimensión que captura la mayor parte de la varianza en los datos. Ejemplos de técnicas de extracción incluyen el Análisis de Componentes Principales, t-distributed Stochastic Neighbor Embedding, y Uniform Manifold Approximation and Projection.

Las técnicas de extracción de características suelen ser preferibles especialmente cuando se trabaja con representaciones complejas como los *word embeddings* generados a partir de secuencias de aminoácidos. Los *word embeddings*, que son representaciones vectoriales de secuencias generadas a través de técnicas de procesamiento de lenguaje natural, pueden ser altamente no interpretables debido a la naturaleza densa y distribuida de los vectores generados. Esto significa que cada dimensión en un embedding no corresponde directamente a una característica biológica específica, sino que representa combinaciones complejas de varias características subyacentes.

Debido al problema de la interpretabilidad de los *word embeddings*, la extracción de características es una mejor opción en muchos casos. Las técnicas de extracción, como PCA, transforman el espacio de alta dimensionalidad en un espacio de menor dimensionalidad, donde las nuevas dimensiones, o componentes principales, son combinaciones lineales de las dimensiones originales. Estas nuevas dimensiones están ordenadas según la cantidad de varianza de los datos que explican, lo que significa que las primeras dimensiones capturan la mayor parte de la información relevante. Esto no solo reduce la complejidad del modelo, sino que también facilita la interpretación al enfocar el análisis en un conjunto reducido de dimensiones que capturan las variaciones más significativas en los datos.

2.5.1. Principal Component Analysis

El Análisis de Componentes Principales, PCA, por sus siglas en inglés, es una técnica estadística para la reducción de dimensionalidad, especialmente relevante en el análisis de datos complejos, como pueden ser aquellos datos genómicos derivados de secuencias genéticas y de anticuerpos. PCA transforma un conjunto de variables originales, que pueden estar correlacionadas entre sí, en un nuevo conjunto de variables no correlacionadas denominadas componentes principales. Estos componentes principales son combinaciones lineales de las variables originales y están ordenados de tal manera que el primer componente principal captura la mayor parte de la varianza presente en los datos, el segundo componente principal captura la mayor parte de la varianza restante, y así sucesivamente. El proceso de PCA se puede dividir en los siguientes pasos: 1. Dado que PCA se basa en la varianza de los datos, es

fundamental que las variables sean estandarizadas, especialmente si tienen diferentes unidades de medida. La estandarización implica restar la media de cada variable y dividirla por su desviación estándar, asegurando que todas las variables contribuyan de manera equitativa al análisis.

2. Una vez estandarizados los datos, se calcula la matriz de covarianza, que describe cómo las variables varían conjuntamente. La covarianza positiva entre dos variables indica que tienden a aumentar o disminuir juntas, mientras que una covarianza negativa indica una relación inversa.

3. A partir de la matriz de covarianza, se calculan los valores propios, *eigenvalues*, y los vectores propios, *eigenvectors*. Los valores propios indican la cantidad de varianza capturada por cada componente principal, mientras que los vectores propios representan la dirección de cada componente principal en el espacio de características.

4. Los componentes principales se ordenan en función de sus valores propios, de mayor a menor. Los primeros componentes principales son los que capturan la mayor parte de la varianza en los datos. La selección de los componentes principales se basa en aquellos que retienen una proporción significativa de la varianza total, lo que se determina típicamente a través de un gráfico del codo, que muestra el número de componentes frente a su valor propio.

5. Finalmente, los datos originales se proyectan en el espacio definido por los componentes principales seleccionados, reduciendo así la dimensionalidad mientras se conserva la mayor parte de la información relevante.

El análisis de la varianza es un aspecto crucial en PCA, ya que permite determinar cuánta de la variabilidad total en los datos originales es capturada por cada componente principal. Al priorizar aquellos componentes que retienen una proporción significativa de la varianza total, PCA facilita la selección de las características más informativas, lo que permite reducir la dimensionalidad del conjunto de datos sin perder información crítica. Sin embargo, uno de los desafíos inherentes a PCA es que, al reducir la dimensionalidad, se puede perder información biológicamente significativa si no se lleva a cabo un análisis cuidadoso de cuánta varianza es realmente importante para el fenómeno biológico que se está estudiando.

2.5.2. t-distributed Stochastic Neighbor Embedding

T-distributed Stochastic Neighbor Embedding es una técnica de reducción de dimensionalidad no lineal desarrollada para facilitar la visualización de datos de alta dimensionalidad en un espacio de menor dimensión, como dos o tres dimensiones. A diferencia de métodos lineales como PCA, t-SNE se enfoca en preservar las relaciones locales en los datos, lo que significa que los puntos de datos que están cercanos en el espacio original permanecerán cercanos en el espacio reducido, mientras que las distancias más grandes entre puntos no necesariamente se preservan.

t-SNE opera en dos fases principales. En el espacio de alta dimensionalidad, t-SNE calcula una probabilidad para cada par de puntos que refleja lo cercanos que están en términos de su similitud. Esto se hace utilizando una distribución de probabilidad gaussiana centrada en cada punto, lo que permite modelar las relaciones de proximidad.

En el espacio de baja dimensionalidad, t-SNE busca un mapeo de los puntos que minimice la divergencia de Kullback-Leibler entre las distribuciones de probabilidad de los pares de puntos en los dos espacios, alta y baja dimensionalidad. En lugar de utilizar una distribución gaussiana, t-SNE utiliza una distribución t de Student con un grado de libertad para calcular las probabilidades en el espacio reducido. Esta elección es crucial porque la distribución t tiene colas más largas que una gaussiana, lo que ayuda a mantener la separación de puntos dispares, evitando la sobrepoblación en el espacio reducido.

Aunque t-SNE es extremadamente útil para la visualización, presenta varias limitaciones. 1. No genera un conjunto de características o variables reducidas que puedan ser

directamente utilizadas en modelos de aprendizaje automático. Su propósito es puramente visual y exploratorio, lo que significa que no puede ser utilizado para reducir dimensionalidad de una manera que optimice el rendimiento de modelos predictivos. 2. Requiere la configuración de varios hiperparámetros, siendo el más crítico el perplexity, que controla el equilibrio entre la preservación de las relaciones locales y globales en los datos. La elección inadecuada de este o de otros hiperparámetros puede llevar a visualizaciones que no representen fielmente la estructura de los datos, lo que podría inducir a errores en la interpretación. 3 Es computacionalmente intensivo y no escala bien con conjuntos de datos extremadamente grandes, lo que puede ser un problema en estudios genómicos donde el volumen de datos es muy alto.

2.5.3. Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection es una técnica de reducción de dimensionalidad no lineal que ha ganado reconocimiento en la comunidad científica, particularmente en el análisis de datos genómicos y de anticuerpos. Al igual que t-SNE, UMAP se utiliza principalmente para la visualización de datos de alta dimensionalidad en espacios de menor dimensión, generalmente en dos o tres dimensiones. Sin embargo, UMAP ofrece varias ventajas distintivas que lo hacen especialmente útil para la exploración de datos complejos en el contexto de la genómica, incluyendo un mejor equilibrio entre la preservación de la estructura global y local del espacio de datos, mayor velocidad computacional, y una menor sensibilidad a la elección de hiperparámetros (McInnes et al., 2020). Basándose principalmente en teoría de grafos, álgebra lineal y teoría de homología, lo que le permite mapear los datos de alta dimensionalidad en un espacio de baja dimensionalidad, preservando tanto la estructura local como global de los datos.

UMAP comienza construyendo un grafo de vecindad ponderado en el espacio de alta dimensionalidad. Esto implica identificar los vecinos más cercanos para cada punto de datos, basándose en una métrica de distancia, como la euclidiana y establecer conexiones ponderadas entre ellos. Estas conexiones reflejan la proximidad de los puntos en el espacio de alta dimensionalidad.

Posteriormente, asume que los datos residen en una manifold de baja dimensionalidad que está incrustada en un espacio de alta dimensionalidad. El algoritmo estima la estructura local de esta manifold utilizando el grafo de vecindad construido en el paso anterior.

Una vez que la estructura local ha sido estimada, UMAP proyecta los datos en un espacio de baja dimensionalidad, preservando tanto las distancias locales entre puntos cercanos como las relaciones globales en la manifold. Para lograr esto, UMAP minimiza una función de costo que busca alinear las distancias en el espacio de baja dimensionalidad con las probabilidades de vecindad en el espacio de alta dimensionalidad.

UMAP y t-SNE son técnicas populares para la visualización de datos de alta dimensionalidad, pero difieren en varios aspectos clave que pueden hacer que UMAP sea más adecuado en ciertos contextos, como el análisis de datos de anticuerpos del SARS-CoV-2:

UMAP es generalmente más rápido que t-SNE, especialmente en conjuntos de datos grandes. Esto se debe a su enfoque en la construcción del grafo de vecindad y su método de optimización más eficiente. En estudios genómicos, donde los conjuntos de datos pueden ser masivos, la mayor velocidad de UMAP es una ventaja significativa.

Mientras que t-SNE está diseñado principalmente para preservar relaciones locales, es decir, puntos cercanos permanecen cercanos, UMAP equilibra mejor la preservación de la estructura local y global. Esto significa que UMAP es más capaz de mantener la topología general del espacio de alta dimensionalidad en la proyección de baja dimensionalidad, lo que puede ser crucial para identificar agrupaciones más significativas y estructuras jerárquicas en los datos.

2.6. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural, NLP, por sus siglas en inglés, es un subcampo de la inteligencia artificial y la lingüística computacional que se centra en la interacción entre las computadoras y el lenguaje humano. Combinando técnicas de lingüística, aprendizaje automático y ciencias de la computación para analizar, comprender y generar lenguaje natural, es decir, el lenguaje que los humanos usamos para comunicarnos. Desde su origen se ha utilizado principalmente para tareas relacionadas con el análisis de textos, tales como la traducción automática, la corrección gramatical, el análisis de sentimientos, la clasificación de textos, la generación de resúmenes automáticos y la comprensión del lenguaje natural.

Algunas tareas destacadas con NLP son: Análisis sintáctico y semántico, el cual se refiere a la construcción de árboles sintácticos que representan la estructura gramatical de una oración. El análisis semántico, por otro lado, busca interpretar el significado de las palabras y frases en un contexto determinado. Estas técnicas son fundamentales para comprender la estructura y el significado subyacente en el lenguaje.

Los modelos de lenguaje, estos algoritmos que predicen la probabilidad de una secuencia de palabras en un lenguaje natural. Los modelos como el n-gram, redes neuronales recurrentes, y más recientemente, los transformadores como GPT, *Generative Pretrained Transformer* y BERT, *Bidirectional Encoder Representations from Transformers*, han avanzado significativamente en la capacidad de los sistemas de NLP para comprender y generar texto coherente y contextualizado.

Con el tiempo, las técnicas desarrolladas en NLP para el análisis de textos se han adaptado para abordar problemas en otros dominios, incluyendo el análisis de secuencias biológicas como las secuencias de ADN, ARN y proteínas. Estas secuencias, aunque no son textos en el sentido tradicional, comparten ciertas propiedades con el lenguaje natural, como se puede observar en la Figura 2.4, están compuestas de unidades básicas, nucleótidos o aminoácidos que forman cadenas lineales con un orden específico, y la combinación de estas unidades puede llevar a la formación de *palabras* o *motifs* que tienen significados biológicos específicos.

Figura 2.4. Similitud del NLP y las secuencias de aminoácidos de los anticuerpos.

2.6.1. Word Embedding: representación vectorial de secuencias

Word Embedding o incrustaciones de palabras son una técnica fundamental en el procesamiento de lenguaje natural que permite transformar palabras o secuencias de caracteres en representaciones vectoriales de alta dimensión. En NLP tradicional, *word embedding* convierte palabras en vectores numéricos que capturan relaciones semánticas entre ellas, permitiendo que los modelos de aprendizaje automático comprendan y manipulen el lenguaje natural de manera más efectiva.

Las técnicas de *word embedding* han sido adaptadas para su uso en el análisis de secuencias biológicas, como las secuencias de aminoácidos en proteínas, facilitando su estudio

y análisis computacional. Esta adaptación es crucial porque permite que las secuencias proteicas, que son cadenas de aminoácidos, sean representadas en un espacio vectorial que puede ser manipulado por algoritmos de aprendizaje automático.

Al igual que en NLP, donde las palabras que aparecen en contextos similares tienden a tener representaciones vectoriales similares, en las secuencias proteicas, los aminoácidos o combinaciones de aminoácidos que tienden a aparecer juntos en secuencias funcionalmente relacionadas tendrán representaciones vectoriales similares.

2.6.2. ProtVec: Adaptación de word2Vec para secuencias proteicas

ProtVec es una técnica basada en la adaptación del modelo Word2Vec, que originalmente se utilizó en el procesamiento de lenguaje natural para representar palabras en un espacio vectorial de alta dimensión. *ProtVec* está específicamente diseñado para el análisis de secuencias proteicas, convirtiendo tripletes de aminoácidos, k-mers, en vectores densos que capturan la información contextual de las secuencias biológicas. Demostrando ser especialmente útil en la identificación de patrones funcionales en proteínas, lo que es fundamental para comprender la actividad biológica y la función de las proteínas, así como para el desarrollo de nuevas terapias y vacunas.

ProtVec inicia segmentando las secuencias de proteínas en tripletes de aminoácidos, conocidos como k-mers. Estos tripletes actúan como las "palabras" en un lenguaje proteico. Cada triplete representa una combinación específica de tres aminoácidos consecutivos en la secuencia proteica, lo que permite capturar la información contextual y estructural de la secuencia de proteínas.

Similar a Word2Vec, *ProtVec* utiliza modelos de aprendizaje no supervisado, como Skip-gram, para entrenar los vectores que representan los k-mers. Durante el entrenamiento, el modelo ajusta estos vectores de tal manera que los k-mers que tienden a aparecer en contextos similares, es decir, en secuencias funcionalmente relacionadas, tengan representaciones vectoriales cercanas en el espacio vectorial. Esto permite que los vectores capturen la semántica de los k-mers en términos de su función biológica o su papel en la estructura proteica.

Una vez entrenados, los vectores de los k-mers se combinan para representar secuencias proteicas completas. La combinación de estos vectores a lo largo de una secuencia permite obtener un embedding global que captura la información funcional y estructural de la proteína. Estos embeddings pueden luego ser utilizados en diversas aplicaciones de análisis proteico.

Finalmente, al igual a otras técnicas de embedding, uno de los principales desafíos de *ProtVec* es la interpretabilidad de los vectores generados, ya que estos representan combinaciones abstractas de características biológicas que pueden ser difíciles de interpretar de manera directa. Así mismo, el rendimiento de *ProtVec* depende de la calidad del conjunto de datos de entrenamiento. Secuencias mal alineadas o con ruido pueden afectar la calidad de los embeddings y, en consecuencia, las predicciones realizadas a partir de ellos.

2.7. Limitaciones y desafíos en la reducción de dimensionalidad

La reducción de dimensionalidad es una técnica ampliamente usada en el análisis de datos genómicos, donde los conjuntos de datos son de alta dimensionalidad y complejidad. Sin embargo, este proceso no está exento de limitaciones y desafíos que deben ser considerados cuidadosamente.

Una de las limitaciones más significativas de la reducción de dimensionalidad es la posible pérdida de información crítica. Durante el proceso de reducción, se eliminan componentes o características que se consideran menos relevantes para explicar la variabilidad en los datos originales. Sin embargo, existe el riesgo de que esta eliminación incluya información que, aunque aparentemente menos significativa, sea crucial para la comprensión

completa de los mecanismos biológicos subyacentes (Ringnér, 2008). Este problema es particularmente relevante cuando se utilizan métodos lineales como el Análisis de Componentes Principales, que puede no capturar adecuadamente las interacciones no lineales entre variables.

Otro factor relevante es la calidad del conjunto de datos, ya que impacta directamente en la efectividad de la reducción de dimensionalidad. Los datos de baja calidad, que pueden estar contaminados con ruido o tener valores faltantes, pueden conducir a resultados engañosos o erróneos cuando se aplican técnicas de reducción de dimensionalidad. Además, el sesgo en los datos, que puede surgir de una representación desproporcionada de ciertas variantes virales o respuestas inmunitarias, también puede afectar la validez de los resultados obtenidos tras la reducción de dimensionalidad (Ioffe & Szegedy, 2015).

Capítulo 3

Estado del arte

“En la transición del campo a la ciudad, nos volvimos más libres, pero también más solitarios”

El dilema de la pareja (Perel, 2017).

3. Estado del arte

En este tercer capítulo se hace una revisión del estado del arte sobre los trabajos más relevantes en el campo de la reducción de dimensiones y el tratamiento de secuencias biológicas, destacando los avances y las limitaciones de las técnicas actuales, así como su aplicación en dominios específicos como el análisis de proteínas, secuencias biológicas, y conjuntos de datos celulares.

3.1. Métodos y enfoques de reducción de dimensiones

Sharma y Saroha (2015) investigan las metodologías utilizadas para la reducción de dimensiones en conjuntos de datos, destacando el fenómeno conocido como la "maldición de la dimensionalidad". Este fenómeno ocurre debido al elevado número de atributos presentes en los conjuntos de datos, que generalmente son grandes porque se recopila mucha información sin considerar previamente el modelo que se desea generar. La maldición de la dimensionalidad complica la tarea de clasificación en estos conjuntos de datos. El estudio también discute diversas técnicas de reducción de dimensionalidad. Entre las técnicas lineales, se destacan el Análisis de Componentes Principales y el Análisis Discriminante Lineal. Por otro lado, se mencionan técnicas no lineales como el Clustering Espectral, el Mapeo Isométrico, los Eigenmaps Laplacianos, la Incrustación Lineal Local y el Kernel PCA. Los algoritmos lineales ofrecen ventajas como un bajo costo computacional y menor sensibilidad al ruido en los datos, aunque presentan la desventaja de asumir que las características con mayor varianza son las más importantes, y pueden tener dificultades para evaluar la matriz de covarianza. En contraste, los algoritmos no lineales son más adecuados para conjuntos de datos con estructuras no lineales, aunque implican un mayor costo computacional. Finalmente, los autores recomiendan considerar tanto el tipo de conjunto de datos como los requisitos específicos del algoritmo de aprendizaje automático antes de seleccionar una técnica de reducción de dimensiones.

Velliangiri et al. (2019) exploran dos enfoques principales para la reducción de dimensiones en el análisis de datos: la selección de atributos y la extracción de atributos. Estas técnicas son fundamentales para mejorar el rendimiento y la precisión de los modelos de clasificación construidos a partir de conjuntos de datos reducidos. El enfoque de selección de atributos incluye métodos como la Correlación de Pearson, el Análisis Discriminante Lineal, la Selección Progresiva de Características, la Eliminación Regresiva de Características y la Eliminación Recursiva de Características. Por otro lado, el enfoque de extracción de atributos abarca técnicas como el Análisis de Componentes Principales, el Análisis Discriminante Generalizado y el Análisis de Componentes Independientes. Los autores destacan que el uso de estas técnicas no solo simplifica los modelos, sino que también optimiza su rendimiento y precisión, haciendo que los algoritmos de clasificación sean más eficientes y efectivos. Sin embargo, no promueven una técnica en particular, sino que subrayan que la elección de la técnica depende del conjunto de datos y del objetivo del estudio. Cada técnica presenta ventajas y desventajas, y no existe una solución única que se adapte a todos los problemas.

Nguyen y Holmes (2019) ofrecen diez consejos esenciales para realizar una reducción de dimensiones adecuada, interpretar correctamente los resultados y comprender su significado. Los consejos destacados incluyen: 1. Elige un método apropiado: Las técnicas lineales, como el PCA, preservan la estructura global del conjunto de datos, mientras que las técnicas no lineales son más eficaces en la representación de interacciones locales. 2. Decide el número de dimensiones a retener de manera consciente: No se debe seleccionar el número de dimensiones de manera arbitraria, incluso cuando el propósito sea la visualización en 2D o 3D. Es crucial asegurarse de que este número sea adecuado para el problema en cuestión para evitar la pérdida de información valiosa. 3. Encuentra las señales ocultas: La reducción de dimensiones debe captar de manera eficaz las fuentes dominantes de variación, permitiendo

identificar patrones ocultos o pequeños clusters que no se observaban previamente. Se recomienda realizar un análisis de datos tanto antes como después de la reducción de dimensiones para descubrir relaciones que podrían haber pasado desapercibidas. Otros consejos incluyen la selección cuidadosa de métodos según el tipo de datos, el manejo adecuado de datos categóricos, la utilización de métodos embebidos para optimizar similitudes y disimilitudes, y la importancia de comprender el significado de las nuevas dimensiones generadas. Los autores también subrayan la necesidad de comprobar la robustez de los resultados y cuantificar las incertidumbres para garantizar la validez del análisis.

Wang (2019) presenta un método para seleccionar el número óptimo de dimensiones en incrustaciones de palabras, *word embeddings* utilizando el Análisis de Componentes Principales. Este método se basa en eliminar dimensiones una por una hasta alcanzar la mejor precisión en el modelo. El proceso comienza transformando la incrustación mediante PCA. A continuación, se eliminan las dimensiones de manera progresiva, comenzando por aquellas que contribuyen menos a la explicación de la varianza. Aunque cada dimensión aporta de manera diferente a la varianza explicada, todas contribuyen de igual forma al cálculo del producto interior. Las dimensiones con menor varianza, pero igual ponderación en el producto interior pueden disminuir el poder discriminativo del modelo. Por lo tanto, eliminarlas permite centrarse en las dimensiones más discriminativas. Este enfoque ofrece una alternativa al uso tradicional de la varianza acumulada. Aunque el método no ha sido probado específicamente con secuencias de aminoácidos, podría ser útil en este contexto.

Zebari et al. (2020) realizan una revisión de la literatura abarcando artículos publicados entre 2017 y 2020, centrada en las técnicas de reducción de dimensiones. La revisión distingue entre dos enfoques principales: técnicas de selección de características y técnicas de extracción de características. La diferencia clave entre estos dos grupos radica en que, en la selección de características, se puede perder información, ya que algunas características se excluyen durante el proceso de selección del subconjunto. Por otro lado, en la extracción de características, es posible reducir la dimensión sin perder una parte significativa del conjunto de datos original. En la mayoría de los artículos revisados sobre selección de características, se comparan nuevas técnicas o combinaciones de técnicas, que luego se evalúan utilizando algoritmos de clasificación. Por su parte, en los artículos revisados sobre extracción de características, el Análisis de Componentes Principales se destaca como la técnica más común. Al igual que con la selección de características, el desempeño de las técnicas de extracción se evalúa mediante algoritmos de clasificación. Además, la revisión revela que los algoritmos de clasificación más utilizados en los estudios revisados son las Máquinas de Soporte Vectorial y los K-Vecinos más Cercanos. Asimismo, se observa que la métrica más comúnmente empleada para evaluar la efectividad de las técnicas de reducción de dimensiones es la precisión.

Arowolo et al. (2021) realizan una revisión de estudios que aplican técnicas de reducción de dimensionalidad, con un enfoque en aquellas que optimizan el agrupamiento de datos de manera eficiente, reducen el tiempo de procesamiento computacional y mejoran la clasificación en la secuenciación de ARN, RNA-Seq. La mayoría de los estudios revisados desarrollan variantes de técnicas existentes, siendo las más comunes las variantes de PCA, que luego son comparadas con otras técnicas de reducción de dimensiones como el Análisis de Componentes Principales, t-Distributed Stochastic Neighbor Embedding, Locally Linear Embedding, Isomap, Diffusion Maps y Laplacian Eigenmaps. En cuanto a las técnicas de clasificación, los estudios revisados muestran una preferencia por el uso de Máquinas de Soporte Vectorial y redes neuronales de Perceptrón Multicapa. Estas técnicas se destacan por su capacidad para manejar la complejidad y la alta dimensionalidad de los datos RNA-Seq, contribuyendo a mejorar la precisión de la clasificación en los análisis de vectores de malaria.

3.1.1. Conclusiones de la sección

Los artículos de este grupo abarcan principalmente el uso de técnicas tradicionales de reducción de dimensiones, tanto lineales como no lineales. Las técnicas más discutidas incluyen Análisis de Componentes Principales y Análisis Discriminante Lineal como herramientas estándar para datos lineales, mientras que para estructuras no lineales se mencionan métodos como Kernel PCA, LLE y Clustering Espectral.

Un problema recurrente es la maldición de la dimensionalidad, la cual hace que sea difícil trabajar con datos de alta dimensionalidad sin perder información clave. Aunque las técnicas lineales como PCA son ampliamente utilizadas por su bajo costo computacional, los artículos subrayan la importancia de elegir adecuadamente la técnica según la naturaleza del conjunto de datos y del problema que se busque solucionar a partir de estos datos.

Tabla 3.1 Resumen de los artículos de métodos y enfoques de reducción de dimensiones

Autores	Año	Técnicas usadas	Enfoque	Aplicación	Conclusión
Sharma y Saroha	2015	PCA, LDA, Isomap, Laplacian Eigenmaps, LLE, Kernel PCA	Extracción	Conjuntos de datos de gran tamaño y conjuntos de datos lineales o no lineales	Las técnicas presentadas se pueden aplican en distintos conjuntos de datos, dependiendo de la linealidad del mismo.
Velliangiri et al.	2019	Pearson Correlation, LDA, Selección progresiva de características, Eliminación recursiva de características, LDA, PCA, ICA,	Selección y Extracción	Conjuntos de datos de gran tamaño, en donde identificar las características más relevantes es importante	Las técnicas listadas se adaptan a las necesidades del problema, ya sea solo reducir dimensiones o identificar las dimensiones más relevantes
Nguyen y holmes	2019	PCA, LDA	Extracción	Conjuntos de datos de alta dimensionalidad	La reducción de dimensiones no es una tarea trivial y no depende solo de la técnica, también influye el preprocesamiento de los datos
Wang	2019	PCA	Extracción	Conjuntos de datos lineales	La selección adecuada del número de componentes de PCA puede mejorar la precisión de los algoritmos de aprendizaje automático
Zebari et al.	2020	Selección de atributos y Extracción de atributos	Selección y Extracción	Diversos conjuntos de datos de alta dimensionalidad, en los cuales el enfoque es identificar atributos más relevantes	La técnica idónea dependerá del problema, se debe tener en cuenta que la reducción de dimensiones perderá una cantidad distinta de información relevante

Arowolo et al.	2021	PCA. t-SNE, LLE, Isomap, Laplacian Rigenmaps	Extracción	Conjuntos de datos de RNA-Seq	La reducción de dimensiones es recomendada en este tipo de datos de alta dimensionalidad, lo cual ayudará al desempeño de los métodos de aprendizaje automático

3.2. Representación y análisis de secuencias biológicas mediante técnicas de aprendizaje automático y NLP

Asgari y Mofrad (2015) presentan un enfoque innovador para representar proteínas y otras secuencias biológicas mediante incrustaciones de palabras, utilizando técnicas de aprendizaje profundo inspiradas en el procesamiento de lenguaje natural. En este enfoque, las secuencias de proteínas se tratan como "oraciones", donde tríos de aminoácidos actúan como "palabras". Los investigadores proponen *ProtVec*, el cual se basa en el modelo Word2Vec, comúnmente utilizado en NLP, para convertir secuencias de proteínas en vectores densos que capturan las propiedades semánticas de los aminoácidos. A través de un entrenamiento no supervisado en una base de datos de proteínas no etiquetadas, *ProtVec* aprende representaciones que reflejan tanto las propiedades físico-químicas de los aminoácidos como su relación con la estructura y función de las proteínas. Los experimentos realizados demuestran que *ProtVec* permite la clasificación de proteínas.

Diggins et al. (2015) proponen un flujo de trabajo diseñado para analizar conjuntos de datos de células de alta dimensionalidad, combinando herramientas de aprendizaje automático supervisado y no supervisado. Este enfoque tiene como objetivo mejorar la comprensión de las poblaciones celulares. El flujo de trabajo consta de cinco etapas principales: Recolección de datos: Esta etapa inicial requiere la colaboración con expertos para asegurar que se recolectan datos relevantes para la investigación antes del análisis. Procesamiento de datos: En esta fase, los datos se estandarizan, normalizan y escalan, además de realizarse tareas comunes del procesamiento de datos en minería de datos tradicional. Análisis inicial: Aquí se busca identificar poblaciones celulares. Los expertos en biología molecular deben tener un conocimiento profundo de los datos y de las poblaciones celulares de interés, asegurando que los datos recolectados en la primera etapa contengan la información necesaria. Visualización de grupos celulares: Esta etapa emplea técnicas de reducción de dimensiones y visualización de datos, como t-SNE, viSNE, ISOMAP, LLE y PCA. Estas técnicas permiten representar datos de alta dimensionalidad en pocas dimensiones. Tras la identificación de grupos celulares, se realiza un refinamiento utilizando herramientas de agrupamiento como SPADE, Misty Mountain y Citrus, que asignan automáticamente células a grupos específicos. Caracterización de grupos celulares: En la etapa final, los grupos celulares descubiertos se caracterizan mediante la comparación de características utilizando mapas térmicos, gráficos de violín y superposiciones de histogramas para la visualización, además de realizar modelado de datos y otros análisis estadísticos. Aunque este flujo de trabajo fue inicialmente diseñado para el análisis de datos de citometría de masas, su flexibilidad permite su adaptación a otros tipos de datos similares, dictando pasos generales que pueden modificarse según sea necesario en cada etapa.

Fischer et al. (2020) desarrollaron y entrenaron múltiples arquitecturas de aprendizaje profundo para modelar la interacción entre el receptor de células T, TCR y el complejo mayor de histocompatibilidad peptídico, pMHC, incorporando covariables específicas de la célula para reducir la variabilidad en los datos de células individuales. Utilizando conjuntos de datos

públicos que incluyen más de 100,000 células T y pares TCR-antígeno provenientes de bases de datos como IEDB y VDJdb, evaluaron diversas arquitecturas de redes neuronales, incluyendo modelos que consideran tanto las cadenas α como β del TCR. Los modelos que tratan los antígenos como variables categóricas mostraron un rendimiento superior en comparación con aquellos que modelan conjuntamente las secuencias del TCR y del antígeno. La inclusión de covariables específicas de la célula, como la identidad del donante y la abundancia de proteínas de superficie, mejoró significativamente la precisión de las predicciones, alcanzando un área bajo la curva de 0.87, en comparación con 0.33 en modelos sin covariables. Estos resultados demuestran la posibilidad de imputar la especificidad antigénica de células T a partir de sus secuencias de TCR, añadiendo así una capa de información fenotípica valiosa a los estudios de secuenciación de ARN de células individuales.

Ostrovsky-Berman et al. (2021) desarrollaron una técnica de incrustación de palabras denominada immune2vec, la cual es una técnica de incrustación de palabras que aplica principios de procesamiento de lenguaje natural específicamente a secuencias de receptores de células B y T. Este método convierte las secuencias de aminoácidos de estos receptores en representaciones numéricas densas, capturando patrones y relaciones contextuales entre las diferentes secuencias. La metodología propuesta para utilizar immune2vec sigue los siguientes pasos: 1. Recolección de datos de distintos conjuntos de secuencias. 2. Generación de vectores numéricos para cada secuencia utilizando el modelo entrenado de immune2vec. 3. Reducción de dimensiones de los vectores. 4. Obtención de la familia IGHV utilizando el paquete Alakazam de R. 5. División del conjunto de vectores en subconjuntos de entrenamiento y prueba. 6. Entrenamiento de algoritmos de clasificación para distinguir entre las familias IGHV. Esta metodología, junto con el modelo immune2vec, fue probada en cinco conjuntos de secuencias de células receptoras B o T. Las secuencias de cada conjunto se transformaron en vectores numéricos utilizando immune2vec. Dado que los vectores generados tenían 100 dimensiones, se utilizó Bosques Aleatorios para reducirlas, quedando las 18 dimensiones más relevantes. Los algoritmos de clasificación empleados incluyeron Árboles de Decisión, Bosques Aleatorios y K-Vecinos Más Cercanos. Los modelos generados mostraron una precisión superior al 70%, lo que indica un buen desempeño del modelo immune2vec en la transformación de secuencias de células B o T en vectores numéricos.

Ofer et al. (2021) exploran una nueva forma de interpretar secuencias de aminoácidos tratándose como si fueran palabras de un lenguaje natural común, utilizando técnicas de procesamiento de lenguaje natural. Este enfoque permite aplicar algoritmos computacionales a las secuencias de proteínas, facilitando su comprensión como un conjunto coherente. Entre las técnicas de PNL aplicables a secuencias de aminoácidos se destacan: 1. Tokenización: Esta técnica segmenta las secuencias en subgrupo, tokens que pueden interpretarse como palabras, facilitando su análisis. 2. Bolsa de Palabras, Bag of Words: Parte de la premisa de que tokens similares comparten características comunes, como la procedencia de la proteína o el antígeno que combaten. 3. Incrustación de Palabras *word embedding*: Permite representar palabras, tokens como vectores de tamaño fijo, donde aquellos con valores similares tienen características comunes y se ubican más cerca entre sí en el espacio vectorial, mientras que los tokens no relacionados se encuentran más distantes. 4. Modelos Profundos de Lenguaje: Estos modelos tienen una mayor capacidad para comprender el contexto de las palabras dentro de las secuencias, lo que permite inferir propiedades específicas de cada secuencia. Los autores señalan que, aunque estas no son las únicas técnicas disponibles, son las más comunes y cuentan con mayor documentación, además de haber sido probadas en diversas áreas de investigación.

Valkiers et al. (2021) desarrollaron un algoritmo de agrupamiento centrado en los receptores de células T, TCR, diseñado para agrupar secuencias de aminoácidos sin necesidad de conocimiento previo sobre su especificidad antigénica. Esta herramienta permite identificar

grupos de células relacionadas que no están previamente etiquetadas. El algoritmo de agrupamiento consta de dos etapas principales: Creación de superclusters: En esta etapa, las secuencias de aminoácidos se transforman en embeddings o representaciones numéricas que reflejan sus propiedades fisicoquímicas y capturan las relaciones entre las secuencias. Luego, se localizan los centroides de los superclusters utilizando el algoritmo de k-means. Refinamiento de grupos: En la segunda etapa, los superclusters se reagrupan en clusters más pequeños y específicos, refinando así los grupos iniciales. A partir de estos grupos refinados, se construyen grafos donde los nodos representan las secuencias y los bordes indican la similitud entre ellas. Estos grafos, organizados por similitud, permiten identificar posibles grupos específicos de epítopos, ya que las secuencias similares tienden a compartir bordes dentro del grafo, lo que revela relaciones basadas en las secuencias en el repertorio. Este enfoque facilita la identificación de patrones en secuencias CDR3 sin la necesidad de etiquetas específicas, proporcionando una herramienta poderosa para el análisis de receptores de células T.

Adjuik y Ananey-Obiri (2022) desarrollaron un modelo de clasificación eficiente y preciso para secuencias del virus COVID-19, utilizando vectores numéricos de proteínas generados mediante la técnica de incrustación de palabras o *word embedding*. Se emplearon dos conjuntos de datos: el primero contenía secuencias de proteínas del virus COVID-19 y el segundo fue obtenido de la plataforma NCBI. Las secuencias de aminoácidos fueron transformadas en vectores numéricos utilizando el modelo de Bolsa Continua de Palabras, CBOW. Dado que los vectores iniciales tenían 200 dimensiones, se aplicó la técnica de reducción de dimensiones PCA, reduciendo los vectores a solo 10 dimensiones. Estos nuevos vectores se utilizaron para entrenar varios modelos de clasificación, incluyendo Regresión Logística, Bosques Aleatorios, Máquinas de Soporte Vectorial, K-Vecinos Más Cercanos y Análisis Discriminante Lineal. Tras la construcción de los modelos, se evaluaron en función de su precisión, destacando el modelo basado en Bosques Aleatorios como el de mejor rendimiento.

Brandes et al. (2022) presentan ProteinBERT, un modelo de lenguaje profundo diseñado para capturar las características únicas de las proteínas. ProteinBERT se entrena previamente combinando el modelado de lenguaje con una tarea de predicción de anotaciones de la Ontología de Genes, lo que permite al modelo aprender tanto la estructura de las secuencias de proteínas como su función biológica. Este enfoque permite al modelo procesar tanto representaciones locales, detalles específicos de la secuencia como globales, contexto más amplio, ofreciendo un análisis integral de las secuencias de proteínas. Los experimentos realizados muestran que ProteinBERT es altamente efectivo en la predicción de propiedades de proteínas, incluso cuando se utilizan conjuntos de datos limitados, lo que sugiere su potencial como una herramienta valiosa en bioinformática para la predicción de funciones y características de proteínas. Sin embargo, debido a la naturaleza de aprendizaje profundo del modelo, es importante considerar su alto costo computacional.

Weber et al. (2024) Llevan a cabo un análisis exhaustivo, exploran los avances recientes en el campo de la predicción de la unión de receptores de células T mediante técnicas de aprendizaje automático. El rápido desarrollo de las técnicas de secuenciación inmune y métodos experimentales ha generado una vasta cantidad de datos sobre el repertorio de TCR, lo que ha impulsado la creación de modelos predictivos cada vez más sofisticados. La evolución de estas técnicas ha seguido un camino que va desde enfoques no supervisados de agrupamiento hasta modelos supervisados de clasificación, culminando en las aplicaciones más recientes de Modelos de Lenguaje de Proteínas, PLMs. El análisis destaca el impacto significativo de los modelos basados en transformadores en la bioinformática. Estos modelos, preentrenados en grandes colecciones de secuencias de proteínas no etiquetadas, permiten convertir secuencias de aminoácidos en embeddings vectorizados que capturan propiedades

biológicas relevantes. Los intentos recientes de aprovechar los PLMs han logrado resultados competitivos en tareas relacionadas con la predicción de la unión de TCR. Sin embargo, se enfatiza la necesidad urgente de mejorar la interpretabilidad de estos modelos, que a menudo operan como cajas negras, limitando la comprensión del proceso predictivo.

3.2.1. Conclusiones de la sección

Los artículos destacan el uso de técnicas de aprendizaje automático y profundo, así como de técnicas de procesamiento de lenguaje natural para analizar secuencias biológicas, como proteínas y receptores de células, en su representación de secuencias de aminoácidos.

Las metodologías en su mayoría incluyen el uso de incrustaciones de palabras o *word embeddings* para transformar secuencias en vectores numéricos, utilizando modelos como *ProtVec* y immune2vec, basados en algoritmos como Word2Vec. Estos enfoques permiten la representación semántica de secuencias de proteínas y la reducción de dimensiones mediante técnicas como PCA o algoritmos de clasificación como Bosques Aleatorios, Máquinas de soporte vectorial, entre otras.

Los problemas observados en este grupo incluyen la complejidad computacional asociada con el uso de modelos profundos como ProteinBERT, que aunque son efectivos, requieren gran capacidad de procesamiento. Además, otro desafío es la interpretabilidad de las nuevas dimensiones generadas por estas técnicas avanzadas, ya que, en algunos casos, las representaciones generadas pueden ser difíciles de explicar biológicamente.

Tabla 3.2.1 Resumen de la sección

Autores	Año	Técnica principal	Tipo de secuencia	Modelo de clasificación o agrupamiento	Resultado
Asgari y Mofrad	2015	ProtVec	Proteínas	Clasificación	*ProtVec* permite transformar proteínas en *word embeddings*, útiles para alimentar modelos de aprendizaje automático
Diggins et al.	2015	t-SNE, viSNE, Isomap, LLE, PCA	Células	Agrupamiento	El flujo de trabajo propuesto mejora en la comprensión de poblaciones celulares
Fischer et al.	2020	Redes Neuronales	Receptores de células T	Clasificación	El uso de redes neuronales permite obtener mayor información subyacente a los datos, pero a costa de un mayor costo computacional.
Ostrovsky-berman et al.	2021	Immune2vec	receptores de células B y T	Clasificación	Immune2vec permite la transformación de secuencias de células receptoras B y T en vectores numéricos
Ofer et al.	2021	Tokenización, Bolsa de palabras, Word Embeddings	Aminoácidos	N/A	El uso de PLN sobre secuencias de aminoácidos mejora su interpretabilidad y uso con algoritmos de aprendizaje automático

Autores	Año	Técnica principal	Tipo de secuencia	Modelo de clasificación o agrupamiento	Resultado
Valkiers et al.	2021	Algoritmo de clustering para secuencias TCR	Receptor de células T	Agrupamiento	El algoritmo de agrupamiento permite encontrar relaciones en secuencias no etiquetadas
Adjuik y ananey-obiri	2022	*Word Embeddings*	Proteínas	Clasificación	La construcción de *word embeddings* a partir de proteínas permite la clasificación de las mismas
Brandes et al.	2022	ProteinBERT	Proteínas	Clasificación	El uso de ProteinBERT permite una alta precisión en la clasificación y predicción de proteínas
Weber et al.	2024	modelos de lenguaje de proteínas	receptores de células T	clasificación y agrupamiento	La predicción de células T es un campo en constante crecimiento y el uso de modelos de lenguaje de proteínas es uno de los avances más recientes

3.3. Evaluación comparativa de técnicas de reducción de dimensiones y preprocesamiento de datos

Obaid et al. (2019) investigan el efecto de tres técnicas de preprocesamiento de datos, Min-Max, Z-Score, y Escalado Decimal, y una técnica de discretización, log2 en el rendimiento del algoritmo de clasificación J48. Además, evalúan el impacto de las técnicas de reducción de dimensiones Principal Component Analysis y Linear Discriminant Analysis. Los experimentos se realizaron utilizando el conjunto de datos NSL-KDD. En el primer experimento, que comparó el desempeño de los modelos construidos con datos normalizados y discretizados, se observó que el modelo basado en datos discretizados con el algoritmo log2 obtuvo el mejor rendimiento en términos de precisión y tiempo de construcción, superando a los modelos normalizados. El segundo experimento comparó las técnicas de reducción de dimensiones PCA y LDA, encontrando que PCA proporcionó el mejor desempeño. Con solo 5 dimensiones, el modelo basado en PCA logró una precisión del 99.28%, destacándose como la técnica más efectiva en este contexto.

Becht et al. (2019) comparan varias técnicas de reducción de dimensiones no lineales aplicadas a conjuntos de datos biológicos, con el objetivo de identificar cuál de estas técnicas ofrece mejores resultados en términos de tiempo de ejecución, calidad, y comprensión de los datos tras la reducción de dimensiones. Las técnicas evaluadas incluyen UMAP, t-Distributed Stochastic Neighbor Embedding, Fit-SNE y un autoencoder SCVIS. Los experimentos se llevaron a cabo utilizando el algoritmo Phenograph, basado en el clustering de Louvain y el etiquetado manual de clusters, clasificando los datos en seis grandes grupos celulares. Tanto UMAP como t-SNE lograron representar adecuadamente las relaciones identificadas por Phenograph; sin embargo, se observó que t-SNE generó conjuntos más impuros, es decir,

grupos con datos mal clasificados que pertenecían a otros grupos. En contraste, UMAP produjo grupos más puros y alineados con las clasificaciones manuales realizadas con Phenograph. Aunque ambas técnicas representaron los datos en solo dos dimensiones, UMAP generó visualizaciones más comprensibles, mientras que las representaciones de t-SNE resultaron más difíciles de interpretar en cuanto a la estructura y las relaciones de los conjuntos. Debido a su capacidad para crear representaciones más claras y precisas, los autores destacan UMAP como la técnica más efectiva y comprensible entre las comparadas.

Enríquez et al. (2021) comparan las técnicas de reducción de dimensiones PCA y el Algoritmo CUR aplicadas a datos de pruebas de COVID-19 obtenidos de un laboratorio clínico en Ibarra, Ecuador. El conjunto de datos utilizado constaba de pocas instancias y dimensiones. Aunque PCA está tradicionalmente orientado a la extracción de atributos, el objetivo del experimento era identificar los atributos más relevantes, es decir, realizar una selección de atributos. PCA redujo las 7 dimensiones originales a solo 3, manteniendo el 80% de la varianza de los datos. Sin embargo, los resultados obtenidos con PCA fueron más difíciles de interpretar, complicando la identificación de los atributos más relevantes. En contraste, el Algoritmo CUR mostró una mayor claridad en la relevancia de cada atributo, también reduciendo a 3 dimensiones, pero con una mejor capacidad para identificar el atributo más significativo en el conjunto de datos. Por esta razón, los autores destacan el Algoritmo CUR por su superior interpretabilidad en comparación con PCA.

Khalilian et al. (2022) presentan un modelo de reducción de dimensiones de aprendizaje profundo denominado VAEResDR, que combina un Autoencoder Variacional, VAE, con una Red Neuronal Residual, ResNet. Este modelo fue diseñado para analizar secuencias CDR3 de la cadena pesada con el objetivo de identificar patrones ocultos y mejorar la agrupación de secuencias CDR3 de anticuerpos y nanocuerpos. Los experimentos se realizaron utilizando varios conjuntos de datos con un alto número de instancias, y las secuencias de cadenas de texto fueron convertidas a vectores numéricos mediante la técnica de One-Hot Encoding. Se llevaron a cabo dos experimentos: en el primero, se comparó la capacidad de diferentes técnicas de reducción de dimensiones para mantener la relación de los datos dentro de un mismo cluster después de reducir las dimensiones a solo dos. En el segundo experimento, se agregó un grupo adicional y se repitió el análisis. Las técnicas comparadas incluyeron PCA, t-SNE, UMAP, VAE, scCCESS(AE), VAEDR y la técnica propuesta VAEResDR. En ambos experimentos, VAEResDR demostró ser superior, ya que los clusters mantenían mejor su estructura y la relación entre los datos en comparación con las otras técnicas de reducción de dimensiones.

Rustam et al. (2022) desarrollaron un algoritmo de reducción de dimensiones denominado Weiszfeld Modificado, MWA, el cual fue comparado con otros algoritmos de reducción de dimensiones, como PCA, Classical Multidimensional Scaling, CMDS, Laplacian Eigenmaps y Locally Linear Embedding, LLE. Los experimentos se realizaron utilizando un conjunto de datos de metabolitos provenientes de brotes de clavo indonesios Para evaluar la calidad de los datos tras la reducción de dimensiones, se analizó la calidad de los clusters generados utilizando el algoritmo fuzzy c-means. Los resultados mostraron que MWA y PCA fueron las técnicas más efectivas, ya que lograron reducir las dimensiones a un número menor mientras mantenían la correcta relación entre los datos. Además, los clusters formados por estas técnicas contenían datos más coherentes y relacionados entre sí, en comparación con los clusters generados por las otras técnicas evaluadas.

3.3.1. Conclusión de la sección

Los artículos evaluaron diferentes técnicas de preprocesamiento y reducción de dimensiones, destacando nuevamente el uso de PCA como una de las herramientas más eficaces para la reducción de dimensiones en diversos contextos. Se compararon técnicas como UMAP, t-SNE, y autoencoders, mostrando que UMAP ofrece mejores visualizaciones y pureza de

agrupamientos en algunos casos, aunque PCA sigue destacando por su robustez y facilidad de implementación.

En términos de preprocesamiento, técnicas como la normalización y la discretización fueron fundamentales para mejorar el rendimiento de los algoritmos de clasificación. Un problema común fue la dificultad para interpretar las nuevas dimensiones cuando se utilizan técnicas más complejas como los autoencoders o los algoritmos basados en transformadores, los cuales, aunque efectivos, tienden a comportarse como "cajas negras".

Tabla 3.3.1 Resumen de la sección

Autores	Año	Técnicas de reducción de dimensiones	Conjunto de datos utilizado	Métrica evaluada	Conclusiones
Obaid et al.	2019	PCA, LDA	NSL-KDD	Precisión y sensibilidad	PCA es superior que LDA
Becht et al.	2019	UMAP, t-SNE, SCVIS, Autoencoder	Datos biológicos	Calidad y pureza de los clusters	UMAP fue la técnica superior en comparación con las demás
Enríquez et al.	2021	PCA, CUR algorithm	Pruebas de covid-19	Interpretabilidad de las dimensiones	CUR algorithm permite mayor interpretabilidad de las dimensiones reducidas, ya que hace selección de atributos
Khalilian et al.	2022	PCA, t-SNE, UMAP, VAEResDR, VAER	Secuencias proteicas	Relación entre clusters	VAEResDR mantiene una mejor estructura en los clusters generados
Rustam et al.	2022	MWA, PCA, CMDS, LLE, Laplacian eigenmaps	Metabolitos de clavo	Calidad de clusters	PCA reduce mejor la dimensionalidad manteniendo las relaciones existentes en los datos

3.4. Conclusión del estado del arte

La reducción de dimensiones es fundamental cuando se trabaja con datos masivos y de alta dimensionalidad, como en el caso de secuencias biológicas, donde se recopilan grandes cantidades de información sin una clara preselección de atributos relevantes.

El fenómeno conocido como la maldición de la dimensionalidad es un problema constante en este tipo de datos, ya que un número elevado de atributos dificulta el procesamiento y análisis eficiente de los datos. Ante esta situación, el uso de técnicas de reducción de dimensiones no solo ayuda simplificando los datos con los que se entrenarán los modelos, sino que también mejora la precisión, interpretabilidad y velocidad de procesamiento de los mismos, haciendo que los algoritmos de clasificación y agrupamiento funcionen de manera más eficaz.

Entre las herramientas más utilizadas se encuentra, Análisis de Componentes Principales, PCA, la cual es la técnica dominante debido a su bajo costo computacional y su capacidad para capturar gran parte de la varianza en los datos con una pequeña cantidad de componentes. Aunque PCA es particularmente eficaz para datos lineales, presenta limitaciones al suponer que las dimensiones con mayor varianza son las más importantes, lo que no siempre es cierto para datos biológicos o secuencias genéticas.

Por otro lado, técnicas no lineales como t-SNE y UMAP están ganando popularidad en áreas donde los datos tienen estructuras no lineales complejas. Los resultados de los estudios comparativos indican que UMAP ofrece una ventaja sobre t-SNE en términos de la claridad de las visualizaciones y la pureza de los clusters formados. Sin embargo, t-SNE sigue siendo útil para la exploración inicial de datos, especialmente cuando se trata de datos con una distribución compleja en espacios de alta dimensionalidad.

Las técnicas basadas en autoencoders y modelos profundos, como los Autoencoders Variacionales, VAE, han demostrado ser eficaces para mantener la estructura de los datos tras la reducción de dimensiones, aunque presentan desafíos en cuanto a la interpretabilidad y los recursos computacionales que requieren.

Otro aspecto relevante tratado en el estado del arte es el preprocesamiento de datos, que incluye técnicas como la normalización, discretización y el escalado. Estas técnicas son esenciales para garantizar que las entradas a los modelos de reducción de dimensiones estén preparadas adecuadamente, evitando sesgos que puedan afectar el rendimiento del análisis.

Al igual que las técnicas de reducción de dimensiones propuestas han incrementado, otro campo en rápida expansión es la aplicación de técnicas de procesamiento de lenguaje natural, NLP, y aprendizaje profundo al análisis de secuencias biológicas. Herramientas como ProtVec y Immune2vec, basadas en la transformación de secuencias de aminoácidos en incrustaciones numéricas, *word embeddings*, permiten capturar la semántica y propiedades fisicoquímicas de las secuencias biológicas. Esto ha abierto nuevas oportunidades para mejorar la clasificación y agrupamiento de proteínas, receptores de células T, TCR, y otras secuencias biológicas, ya que estas técnicas permiten una representación densa y significativa de las secuencias, lo que facilita su análisis en un espacio de dimensiones reducidas.

Además, el desarrollo de modelos de lenguaje profundo como ProteinBERT representa un avance importante en la predicción de propiedades y funciones de proteínas. Estos modelos son capaces de aprender tanto la estructura local como global de las secuencias, proporcionando una visión integral de las proteínas. Sin embargo, el uso de estos modelos profundos plantea un desafío en cuanto a su alto costo computacional y su limitada interpretabilidad, lo que puede restringir su aplicación en escenarios donde la transparencia del modelo es crucial.

Finalmente, a pesar de los avances, persisten varios problemas. Uno de los desafíos principales es la interpretabilidad de los modelos y las nuevas dimensiones generadas. A medida que se emplean técnicas más complejas como los autoencoders o los modelos basados

en transformadores, el entendimiento de cómo las dimensiones resultantes capturan la variabilidad de los datos se vuelve más difícil. Esta falta de interpretabilidad limita la capacidad de los investigadores para extraer conclusiones biológicamente relevantes, lo cual es crucial en el ámbito de la bioinformática.

Otro desafío es el costo computacional, especialmente con los modelos profundos. Aunque estos modelos son capaces de capturar patrones complejos en los datos, su entrenamiento y optimización requieren recursos significativos, lo que puede ser un obstáculo en contextos donde el acceso a infraestructura computacional es limitado.

Por último, la selección de la técnica adecuada sigue siendo un tema abierto. No existe una técnica de reducción de dimensiones que sea universalmente aplicable a todos los problemas. Cada técnica tiene sus propias fortalezas y limitaciones, y la elección debe basarse en las características específicas del conjunto de datos, el objetivo del análisis y las restricciones computacionales.

Capítulo 4

Construcción de modelos predictivos

"Una reacción anormal ante una situación anormal es un comportamiento normal."

El hombre en busca de sentido (Frankl, 1946).

4. Construcción de modelos predictivos

En este capítulo se detallan las actividades llevadas a cabo durante el desarrollo de la tesis, siguiendo la metodología de investigación establecida en el proyecto. Estas actividades, que fueron esbozadas previamente en la Figura 1.1 del Capítulo 1, abarcan desde la obtención de datos relacionados con el SARS-CoV-2 hasta la creación de conjuntos de datos basados en la región más variable de los anticuerpos, conocida como CDR3. Además, se describe el tratamiento de estas secuencias de aminoácidos utilizando técnicas de procesamiento de lenguaje natural y la posterior generación de modelos de clasificación. Estos modelos se desarrollaron tanto con conjuntos de datos originales, sin reducción de dimensiones, como con datos cuya dimensionalidad fue reducida mediante técnicas como el Análisis de Componentes Principales y la Aproximación y Proyección Uniforme de la Variedad, UMAP, por sus siglas en inglés.

En la Figura 4.1. al igual que en la sección 1.4 se muestra la metodología implementada en esta investigación, la cual consta de cinco etapas, las cuales se describen a continuación.

1. Construcción de conjuntos de datos

   En esta primera etapa, se reúnen y preparan los conjuntos de datos necesarios para el análisis. Se seleccionan y organizan los datos que contienen las secuencias de interés, como los CDR3 de anticuerpos, agrupándolos según criterios específicos para asegurar su utilidad en las etapas posteriores.

2. Transformación de secuencias en representación numérica

   Las secuencias de los conjuntos de datos se convierten en representaciones numéricas, *word embeddings*, facilitando su procesamiento por modelos de aprendizaje automático. Este paso es esencial para que las secuencias puedan ser interpretadas y analizadas computacionalmente.

3. Reducción de dimensiones

   En esta etapa, se aplica reducción de dimensiones a los datos transformados, utilizando técnicas como PCA o UMAP. Esto ayuda a simplificar la representación de los datos, manteniendo solo la información más relevante y mejorando la eficiencia del modelo.

4. Entrenamiento de modelos predictivos

   Durante esta etapa se entrenan los modelos predictivos con los conjuntos de datos previamente transformados y reducidos en dimensiones. Los modelos predictivos que se usarán son máquinas de soporte vectorial, k-vecinos más cercanos y bosque aleatorio, cada uno de los modelos se ajusta mediante optimización de hiperparámetros para aprender correctamente los patrones en los datos, con el objetivo de hacer predicciones precisas en nuevas muestras.

5. Comparación de métricas de cada modelo

   En la etapa final, se comparan las métricas de desempeño, precisión, sensibilidad, y puntaje F1 de los modelos entrenados. Esto permite evaluar cuál modelo funciona mejor y seleccionar el más adecuado para el problema.

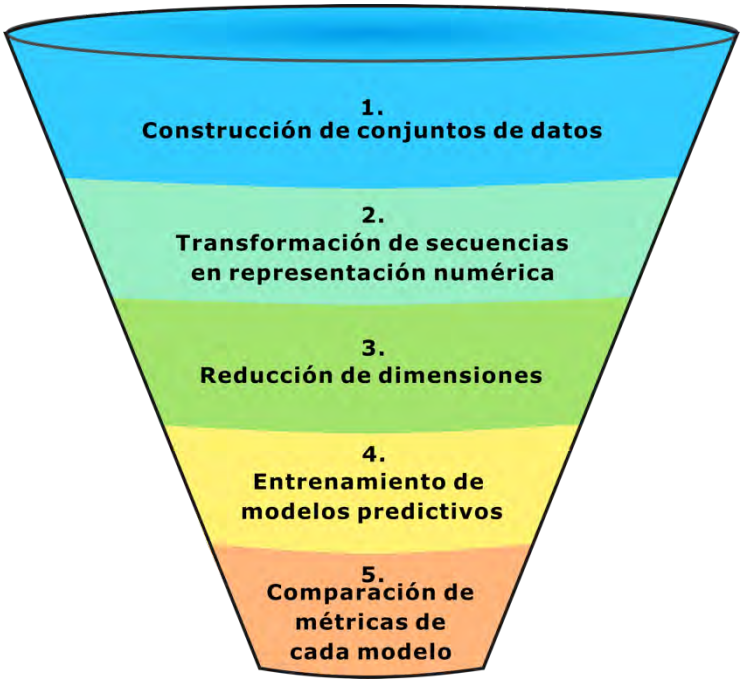

Figura 4.1. Metodología de solución

4.1. Construcción de conjuntos de datos

El proyecto *Observed Antibody Space*, OAS desempeña un papel fundamental en esta investigación, ya que proporciona una vasta recopilación de datos de anticuerpos generados por diversas especies, incluyendo humanos, ratones, murciélagos y conejos. Los anticuerpos recopilados abarcan diversas enfermedades, como el ébola, el VIH, el dengue y el SARS-CoV-2, entre otras. Dado que la investigación se centra exclusivamente en anticuerpos humanos específicos del virus SARS-CoV-2, fue necesario realizar un proceso de filtrado cuidadoso de los conjuntos de datos disponibles en OAS, los atributos buscados se encuentran en la Tabla 4.1. Este filtrado permitió seleccionar solo aquellos conjuntos de datos que se alineaban con los objetivos de la investigación, asegurando la relevancia y pertinencia de los datos utilizados.

La búsqueda y descarga de datos se realizó en la página web de OAS, accesible a través de la URL: https://opig.stats.ox.ac.uk/webapps/oas/oas_unpaired/. El filtro de conjuntos de datos se aplicó utilizando los siguientes parámetros de búsqueda:

Tabla 4.1 Parámetros de búsqueda dentro de OAS

Atributo	Sin Evidencia	Con evidencia
Especie	humano	humano
Enfermedad	Ninguna	SARS-CoV-2
Vacuna	Ninguna	Ninguna
Cadena	Pesada	Pesada
Isotipo	IGHG	IGHG
Forma de extracción de datos	PBMC	PBMC

Tras la búsqueda, se consiguieron los siguientes conjuntos de datos:

Personas sanas: Se encontraron 468 conjuntos de datos donde el parámetro de enfermedad estaba marcado como "Ninguna". Estos conjuntos corresponden a individuos que no mostraban evidencia de enfermedad en el momento de la recolección de los datos.

Personas enfermas: Cuando se seleccionó el parámetro de enfermedad como SARS-CoV-2, se obtuvieron 266 conjuntos de datos correspondientes a personas que habían dado positivo para el virus en el momento de la recolección de los datos.

Posterior a la descarga de estos conjuntos de datos, se llevó a cabo un análisis para contabilizar los registros en cada uno de los 734 conjuntos, 468 de personas sanas y 266 de personas enfermas. El objetivo principal de este análisis fue identificar y excluir aquellos conjuntos de datos que contuvieran menos de 100 registros. Este umbral se estableció para asegurar la suficiencia de datos en las etapas posteriores de la investigación, dado que cada conjunto representa un individuo y cada registro corresponde a un anticuerpo presente en dicho individuo.

El proceso de filtrado dio como resultado la reducción del número total de conjuntos de datos, pasando de 734 a 395, de los cuales 172 correspondían a personas sanas y 223 a personas enfermas. La Figura 4.1 ilustra la reducción de estos conjuntos, destacando el proceso de selección que permitió centrar la investigación en un subconjunto de datos más manejable y relevante.

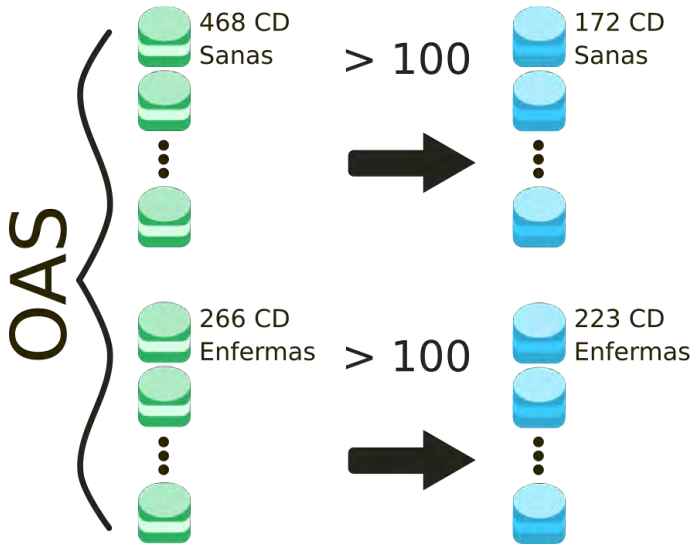

Figura 4.2. Filtros de conjuntos de datos descargados

4.1.1. Construcción de subconjuntos de anticuerpos del SARS-CoV-2

Para facilitar la comprensión de los atributos y su organización en los conjuntos de datos provistos por OAS, se han incluido dos anexos relevantes. El Anexo A presenta una tabla descriptiva que clasifica los 98 atributos de los conjuntos de datos en 21 categorías. Estos atributos son comunes a todos los conjuntos de datos proporcionados por OAS, lo que simplificó su procesamiento y permitió un enfoque uniforme durante el análisis. Además, el Anexo B muestra una tabla representativa de un conjunto de datos descargado de OAS, donde se ilustran los 98 atributos junto con tres registros seleccionados. Esta tabla sirve como referencia para entender el tipo de información contenida en cada conjunto de datos.

Los anticuerpos son componentes clave del sistema inmunológico y desempeñan tres funciones fundamentales: reconocimiento, unión de antígenos y función efectora. El reconocimiento y la unión a antígenos se llevan a cabo a través de los extremos aminoterminales, también conocidos como la región variable de las cadenas pesadas y ligeras

de los anticuerpos. Por otro lado, la función efectora es realizada por el extremo carboxiterminal, correspondiente a la región constante de las cadenas pesadas (García Merino, 2011). Dentro de la región variable, que es la parte más diversa entre diferentes anticuerpos, existen segmentos cruciales para la unión al antígeno. Esta región está compuesta por siete segmentos de aminoácidos, de los cuales tres son conocidos como regiones hipervariables o CDR, del inglés Complementarity *Determining Regions*, y los cuatro restantes conforman las regiones framework o de estructura (Gorshtein, 2022). Las CDR son especialmente importantes, ya que son las partes del anticuerpo que establecen contacto directo con el antígeno, desempeñando un papel activo en su reconocimiento y unión. Sin embargo, no todas las regiones CDR ofrecen la misma cantidad de información sobre el antígeno. Según estudios previos, la región CDR3 es la más diversa y, por lo tanto, la más informativa y distintiva en comparación con las demás (Xu & Davis, 2000).

Dado que el CDR3 pertenece a la región hipervariable y es la secuencia que proporciona más información respecto al antígeno que detecta, se decidió enfocar el análisis y experimentación exclusivamente en este atributo de los conjuntos de datos extraídos de OAS, descartando los demás atributos. Centrando todo el esfuerzo en estos registros, los cuales tienen la información más relevante para la investigación. En la Figura 4.2 se presenta un ejemplo representativo de una secuencia de anticuerpo, donde se muestra la cadena de texto correspondiente y se resalta en amarillo la región CDR3. Cabe destacar que la longitud y el contenido del CDR3 pueden variar considerablemente entre diferentes anticuerpos, así como su posición de inicio y fin. Sin embargo, estos detalles, al igual que la secuencia en nucleótidos y aminoácidos, son proporcionados por los investigadores de OAS, como se puede observar en el Anexo B, lo que facilita su identificación y análisis en los datos disponibles.

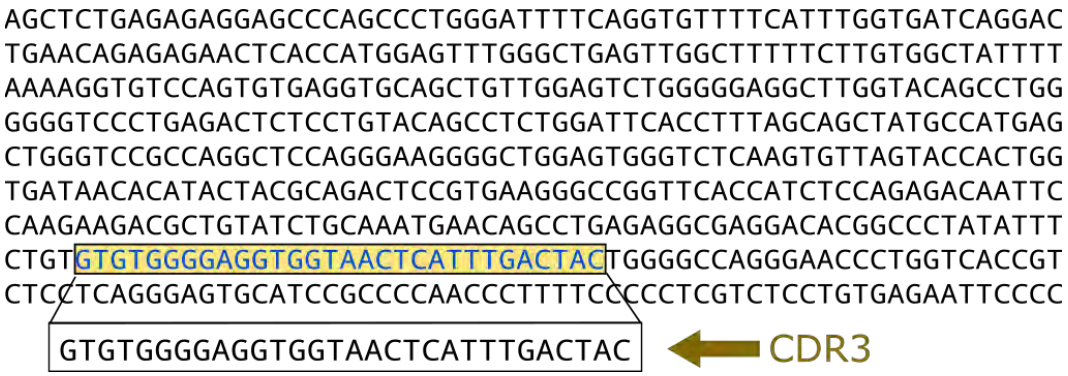

Figura 4.3. Anticuerpo y región CDR3

A partir de los 223 conjuntos de datos correspondientes a personas enfermas, se construyó subconjuntos de datos que se enfocan principalmente en anticuerpos específicos del SARS-CoV-2. Este enfoque se basa en la premisa de que, cuando un individuo es infectado por un virus, en este caso SARS-CoV-2, su sistema inmunitario responde aumentando significativamente la producción de anticuerpos específicos para atacar y neutralizar al virus (Stafford et al., 2016). Por lo tanto, aunque los conjuntos de datos originales contengan anticuerpos que respondan a otras enfermedades, se espera que los anticuerpos más repetidos en estos conjuntos sean aquellos específicos contra el SARS-CoV-2, ya que las muestras fueron recolectadas de personas que estaban infectadas con este virus en el momento de la toma de muestras. Este mismo enfoque se aplicó también a los 172 conjuntos de datos de personas sanas, aunque se anticipa que la distribución de anticuerpos será diferente, reflejando una menor presencia de anticuerpos específicos.

Para la construcción de los subconjuntos se agrupan y ordenan las secuencias CDR3 según su frecuencia de aparición en cada conjunto de datos. Luego, se crean los subconjuntos específicos descritos a continuación:

1. El anticuerpo más frecuente de cada individuo, CD1:
   Este conjunto, CD1, se construyó tomando la CDR3 con mayor frecuencia en cada uno de los 395 conjuntos de datos, 223 de personas enfermas y 172 de personas sanas, dando como resultado un conjunto final compuesto por 395 registros, representando el anticuerpo predominante de cada individuo.
2. Los diez anticuerpos más frecuentes de cada individuo, CD10:
   Para el conjunto, CD10, se seleccionaron las diez CDR3 más frecuentes de cada conjunto de datos, lo que generó un total de 3,950 registros, 395 individuos × 10 anticuerpos por individuo.
3. Los cien anticuerpos más frecuentes de cada individuo, CD100:
   Este último subconjunto, CD100, incluye los cien anticuerpos más frecuentes de cada conjunto de datos, dando como resultado un conjunto de datos con un total de 39,500 registros, 395 individuos × 100 anticuerpos por individuo.

La Figura 4.3 ilustra el proceso de selección y agrupación de los anticuerpos en función de su frecuencia de aparición en los diferentes conjuntos de datos. Este enfoque sistemático de agrupación asegura que los conjuntos de datos resultantes sean representativos de la respuesta inmune específica frente al SARS-CoV-2.

Figura 4.4. Proceso para generar los tres conjuntos de datos a partir de los 395 conjuntos de datos.

Para enriquecer el análisis de los tres nuevos conjuntos de datos, CD1, CD10 y CD100, se añadieron tres atributos a cada registro: Longitud_CDR, Frecuencia, y Clase. Estos atributos complementan la información contenida en los conjuntos de datos y permiten un análisis más detallado y preciso. *Longitud_CDR*: Este atributo se calculó determinando el número de aminoácidos presentes en la secuencia del CDR3. Dado que la longitud del CDR3 puede variar significativamente entre diferentes anticuerpos, este atributo es relevante para entender cómo las variaciones en la longitud pueden estar relacionadas con la especificidad y la eficacia de la unión del anticuerpo al antígeno. *Frecuencia*: Indica el número de veces que un CDR3 específico aparece en su conjunto de datos original. *Clase*: Este atributo clasifica cada registro según el origen del conjunto de datos. Se asignó la clase SARS a los registros provenientes de conjuntos de datos de personas enfermas, y la clase NoSARS a aquellos registros procedentes de personas sanas.

La Tabla 4.2 muestra un ejemplo representativo de cómo estos atributos se integran en los registros de los conjuntos de datos, proporcionando una visión clara de la estructura de los datos enriquecidos.

Tabla 4.2 Extracto del conjunto de datos CD1

	CDR3	Longitud_CDR	Frecuencia	Clase
1	ARVFPRWLQFDPYFDY	16	4949	SARS
2	AKGATKVDY	9	612	SARS
3	ARGEDSAKLGKGNY	14	470	No SARS
4	AREGYNYFDT	10	684	No SARS

El atributo más relevante en los tres conjuntos de datos es la CDR3, la cual es una secuencia de aminoácidos representada como una cadena de texto. Esta secuencia es fundamental porque contiene la información clave sobre la especificidad del anticuerpo hacia el antígeno, en este caso, el virus SARS-CoV-2. Sin embargo, los algoritmos de aprendizaje automático no pueden procesar directamente estas cadenas de texto en su forma de aminoácidos. Por lo tanto, es necesario convertir las secuencias de CDR3 en una representación numérica que pueda ser utilizada para la construcción de modelos de aprendizaje automático.

4.2. Transformación de secuencias en representación numérica

La transformación de secuencias de CDR3 a una representación numérica se realiza mediante técnicas avanzadas de procesamiento de lenguaje natural, que permiten convertir estas cadenas en vectores numéricos utilizables por modelos de aprendizaje automático (Patil et al., 2023). Una de las técnicas más destacadas en este ámbito es Word2Vec, ampliamente utilizada para representar palabras como vectores en espacios de alta dimensionalidad. Word2Vec se basa en la idea de que las palabras que aparecen en contextos similares tienden a compartir representaciones vectoriales similares, capturando así relaciones semánticas a partir de grandes conjuntos de datos textuales (Mikolov et al., 2013).

Sin embargo, Word2Vec fue diseñado para trabajar con texto convencional, donde las palabras tienen un significado semántico en el contexto de una oración y siguen una estructura gramatical bien definida. Este enfoque no se aplica directamente a las secuencias de CDR3, que no siguen las mismas reglas de construcción que las palabras en el lenguaje natural. Las secuencias CDR3 no tienen un significado semántico convencional y no forman parte de oraciones en un sentido lingüístico. Por lo tanto, la aplicación de Word2Vec a estas secuencias puede no capturar adecuadamente la complejidad y la variabilidad que caracteriza a los CDR3. Para superar esta limitación, los investigadores Asgari y Mofrad (2015) desarrollaron una variante específica para secuencias biológicas denominada *ProtVec*. *ProtVec* es una red neuronal que transforma secuencias de aminoácidos, como las secuencias CDR3, en vectores numéricos de alta dimensionalidad. *ProtVec* fue entrenado utilizando la base de datos Swiss-Prot, que contiene información sobre 7,027 familias distintas de proteínas, lo que permite que la red aprenda representaciones numéricas significativas para las secuencias de aminoácidos.

A diferencia de las palabras en el lenguaje natural, las secuencias de aminoácidos de los anticuerpos no tienen una estructura gramatical y no forman "oraciones" en el sentido lingüístico. Sin embargo, *ProtVec* aborda esta diferencia al tratar las secuencias de aminoácidos de manera similar a cómo Word2Vec trata las oraciones, pero adaptado al contexto biológico. En lugar de palabras, *ProtVec* utiliza 3-gramas, tripletes de aminoácidos para capturar la información contenida en las secuencias de proteínas. La Figura 4.4 ilustra esta analogía, donde

las secuencias de aminoácidos se comparan con oraciones y los 3-gramas con palabras. De esta manera, *ProtVec* puede descubrir patrones y relaciones dentro de las secuencias de CDR3, superando las limitaciones que presenta Word2Vec en este contexto.

El resultado de aplicar *ProtVec* a las secuencia CDR3 es una representación numérica en forma de *word embeddings*, que son vectores de 100 dimensiones en los que cada vector representa una secuencia CDR3. Estos embeddings permiten a los modelos de aprendizaje automático procesar y analizar los CDR3 como datos numéricos. Sin embargo, es importante destacar que el significado de cada una de las 100 dimensiones de los *word embeddings* generados por *ProtVec* no es interpretable en un sentido directo. Esto se debe a que los vectores se generan a partir de valores iniciales aleatorios que luego se ajustan durante el entrenamiento de la red neuronal. A pesar de esta falta de interpretabilidad directa, los *word embeddings* resultantes han demostrado ser altamente efectivos para capturar la complejidad y diversidad de las secuencias de aminoácidos, lo que los convierte en una herramienta valiosa para el análisis de datos inmunológicos como los CDR3.

Figura 4.5. Similitud del lenguaje natural y las secuencias de aminoácidos de los anticuerpos

4.2.1. Pruebas con proteínas *Cuticle*, *GalKase* y *Heliz*

Antes de aplicar *ProtVec* a los conjuntos de datos de SARS-CoV-2, se llevó a cabo una evaluación preliminar de la herramienta para verificar su capacidad de representar adecuadamente las proteínas en forma de vectores numéricos. Esta evaluación fue fundamental para asegurar que *ProtVec* pudiera capturar correctamente las características de las secuencias de aminoácidos antes de aplicarlo a las secuencias de CDR3. Para ello, se seleccionaron tres familias de proteínas de la base de datos Pfam, una base de datos reconocida que contiene múltiples secuencias de estructuras de proteínas, abarcando diversos dominios y familias proteicas comunes (Sonnhammer et al., 1997).

El motivo de utilizar Pfam para esta evaluación, en lugar de los conjuntos CD1, CD10 o CD100, es que Pfam proporciona secuencias de aminoácidos que están claramente etiquetadas según la familia proteica a la que pertenecen. Esto es esencial, ya que las secuencias de CDR3 en los conjuntos de datos de personas enfermas pueden incluir anticuerpos de diferentes tipos, lo que complica su etiquetado preciso. Trabajar con Pfam garantiza que cada secuencia de aminoácidos esté correctamente clasificada, permitiendo una evaluación más precisa de la eficacia de *ProtVec*.

Para esta evaluación, se seleccionaron tres familias de proteínas: *Cuticle*, *GalKase* y *Heliz*, las cuales tienen una longitud media similar en sus secuencias de aminoácidos, con un promedio de 49 caracteres. Específicamente, la familia Cuticle cuenta con 100 registros, GalKase con 192 registros, y Heliz con 101 registros. La similitud en la longitud de las secuencias entre las tres familias minimiza la influencia que las diferencias en la longitud podrían tener en los resultados, permitiendo una comparación más equitativa. Cada familia fue sometida al siguiente proceso de tratamiento de datos:

Transformación a su representación numérica con ProtVec: Las secuencias de aminoácidos de cada familia fueron transformadas en vectores numéricos utilizando *ProtVec*, lo que permitió capturar las relaciones subyacentes en estas secuencias.

Asignación de su respectiva clase según la familia a la que pertenece: Se etiquetó cada vector numérico con la clase correspondiente a su familia proteica, *Cuticle, GalKase* o *Heliz,*, facilitando la identificación posterior de las diferencias entre las familias.

Reducción de dimensiones con t-SNE de 100 a 2 dimensiones: Para el análisis visual de los resultados, se empleó la técnica de reducción de dimensiones t-SNE (van der Maaten & Hinton, 2008). t-SNE permite representar datos de alta dimensionalidad en solo dos dimensiones, lo que facilita la identificación visual de patrones y la separación entre las diferentes familias de proteínas sin la necesidad de entrenar un modelo de clasificación.

La Figura 4.5 muestra el resultado de la transformación numérica de las secuencias de aminoácidos y su posterior reducción a dos dimensiones utilizando t-SNE. Como se puede observar en la figura, existe una clara separación entre las tres familias de proteínas. Este resultado es un fuerte indicativo de que *ProtVec* está transformando correctamente las secuencias de aminoácidos en representaciones numéricas, capturando de manera efectiva las relaciones semánticas entre las secuencias de proteínas.

Este proceso de evaluación con proteínas *Cuticle, GalKase* y *Heliz* proporciona una validación inicial de la eficacia de *ProtVec* para el análisis de secuencias de aminoácidos, lo que respalda su aplicación en la transformación de secuencias CDR3 en los conjuntos de datos de SARS-CoV-2 en etapas posteriores del proyecto.

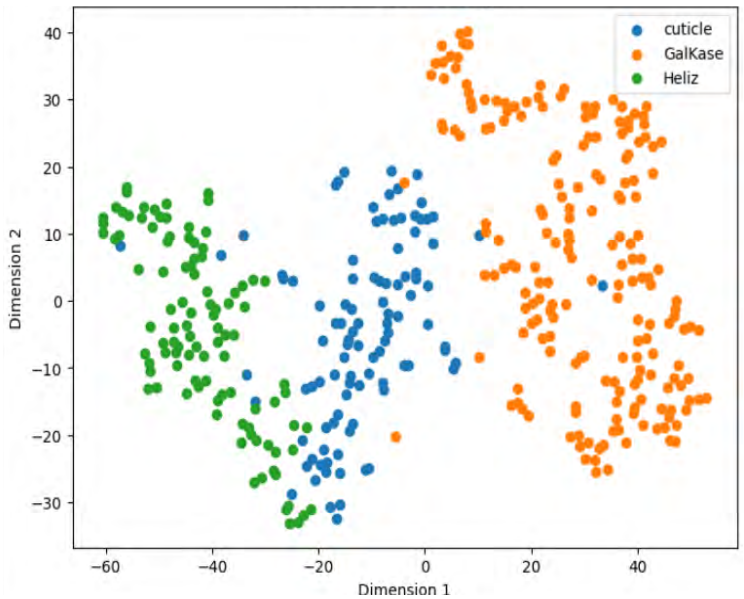

Figura 4.6. Proteínas *Cuticle*, *GalKase* y *Heliz* representadas en dos dimensiones

4.3. Entrenamiento de modelos predictivos

Para la configuración experimental de este estudio, se seleccionaron tres algoritmos de clasificación ampliamente reconocidos por su eficacia en la clasificación de representaciones numéricas de palabras o *word embedding*: Máquinas de Soporte Vectorial (Mammone et al., 2009) (Mammone et al., 2009), Bosques Aleatorios (Rigatti, 2017) y K Vecinos Más Cercanos (Kramer, 2013). Estos algoritmos han sido elegidos basándose en una revisión exhaustiva de la literatura, donde se ha demostrado su capacidad para manejar datos complejos y proporcionar resultados precisos en tareas de clasificación. (Arowolo et al., 2021; Asgari & Mofrad, 2015; Fischer et al., 2020; Ostrovsky-Berman et al., 2021; Weber et al., 2024).

Antes de aplicar estos algoritmos, los datos se sometieron a un proceso de normalización utilizando el Min-Max Scaler. Este método ajusta los valores de las características a un rango entre 0 y 1, lo que es crucial para algoritmos como SVM y KNN, que son sensibles a las magnitudes de las características. La normalización asegura que todas las características contribuyan de manera equitativa al proceso de clasificación, evitando que las características con rangos más amplios dominen el análisis.

Dado que la clase SARS tenía más registros que la clase NoSARS, se aplicó una técnica de *undersampling* utilizando *RandomUnderSampler*. Este método se utilizó para equilibrar el número de muestras entre las dos clases, reduciendo el número de registros de la clase mayoritaria, SARS para igualarlo con la clase minoritaria, NoSARS. El objetivo del *undersampling* es evitar que el modelo de clasificación se sesgue hacia la clase más representada, lo que podría comprometer la precisión y generalización del modelo.

Adicionalmente, se aplicó SeqLogo para identificar los n-gramas que más se repiten al inicio y al final de las secuencias CDR3. SeqLogo es una herramienta bioinformática comúnmente utilizada para visualizar la frecuencia de residuos, aminoácidos, nucleótidos, etc. en posiciones específicas de secuencias alineadas, lo que facilita la identificación de patrones conservados (Crooks et al., 2004). En este estudio, SeqLogo permitió identificar los n-gramas recurrentes en las secuencias de CDR3. Los n-gramas que se repiten de manera consistente en todas las secuencias fueron considerados redundantes, ya que no aportan suficiente información distintiva para el proceso de clasificación. Por lo tanto, estos 2-gramas del inicio de las secuencias y 3-gramas presentes al final de las secuencias fueron eliminados para evitar que introduzcan ruido en el modelo de clasificación y mejorar la capacidad del modelo para diferenciar entre las clases SARS y NoSARS.

Para mejorar la robustez y la validez de los resultados obtenidos, el conjunto de datos se dividió en un 80% para entrenamiento y un 20% para prueba. Esta división inicial asegura que el modelo se entrene en una muestra representativa de los datos mientras se reserva un conjunto independiente para evaluar su desempeño. Además de esta división, se realizó un cross-validation con 10 *folds* en cada uno de los algoritmos de clasificación seleccionados. Esta técnica implica dividir el conjunto de datos de entrenamiento en 10 partes o *folds*. En cada iteración del proceso, uno de los *folds* se utiliza como un conjunto de prueba interno, mientras que los otros nueve se emplean para entrenar el modelo. Este procedimiento se repite 10 veces, asegurando que cada *fold* sea utilizado una vez como conjunto de prueba. Al final, se promedian los resultados de las 10 iteraciones, proporcionando una estimación más confiable del rendimiento del modelo.

Estas configuraciones y preprocesamientos aseguran que los modelos de clasificación sean entrenados con datos optimizados y balanceados, lo que es fundamental para obtener resultados precisos y confiables en la tarea de clasificación de secuencias CDR3 en el contexto del SARS-CoV-2.

4.3.1. Optimización de hiperparámetros

La optimización de hiperparámetros es un paso crucial en la creación de modelos de clasificación efectivos, ya que permite ajustar los parámetros de los algoritmos para maximizar su rendimiento en tareas específicas. En este estudio, se realizó una optimización de hiperparámetros para cada uno de los algoritmos seleccionados con el objetivo de garantizar un desempeño óptimo en la clasificación de anticuerpos a partir de las secuencias de CDR3. Para llevar a cabo esta optimización, se empleó la optimización bayesiana, una técnica avanzada que utiliza un modelo probabilístico para seleccionar los hiperparámetros de manera más eficiente que los métodos tradicionales como la búsqueda en cuadrícula o la búsqueda aleatoria. En particular, se utilizó la variante del Estimador de Parzen Estructurado en Árbol, TPE, que ha demostrado ser eficaz para encontrar configuraciones óptimas de hiperparámetros en espacios de búsqueda complejos y de alta dimensionalidad (Yang & Shami, 2020).

La Tabla 4.3 muestra los hiperparámetros optimizados para cada algoritmo, junto con una breve descripción de cada uno. Estos hiperparámetros fueron ajustados para mejorar el rendimiento en las tareas de clasificación, considerando tanto los conjuntos de datos completos CD1, CD10, CD100 como los conjuntos con reducción de dimensionalidad CD1_PCA90, CD10_PCA90, CD100_PCA90, CD1_PCA95, CD10_PCA95, CD100_PCA95, CD1_UMAP, CD10_UMAP, CD100_UMAP. La reducción de dimensionalidad se llevó a cabo utilizando técnicas como PCA con retención del 90% y 95% de la varianza, y UMAP, para reducir el número de características y mejorar la manejabilidad de los datos sin perder información relevante.

Es importante destacar que, en el caso del algoritmo de Máquinas de Soporte Vectorial, no se realizó una optimización de hiperparámetros. Durante la experimentación inicial, se observó que la configuración por defecto del algoritmo ofrecía un rendimiento similar al obtenido mediante la optimización de hiperparámetros. Por lo tanto, se decidió mantener la configuración predeterminada, evitando un proceso de optimización que no aportaría mejoras significativas.

Tabla 4.3 Descripción y configuración de los hiperparámetros

Algoritmo	Hiperparámetro	Descripción
K vecinos más cercanos	*n_neighbors*	Número de vecinos que se contemplarán para determinar la clasificación de un punto de consulta específico.
Máquinas de soporte vectorial	*C*	El parámetro de regularización en C gestiona la ponderación entre maximizar el margen y minimizar el error de clasificación.
	Kernel	Determina el tipo de límite de decisión utilizado para separar las clases en el espacio de entrada.
Bosque aleatorio	*n_estimators*	Número de árboles de decisión considerados para entrenar el modelo.
	max_features	Número de características utilizadas para entrenar cada árbol.
	max_deph	Profundidad máxima de cada árbol de decisión.
	min_samples_leaf	Número mínimo de instancias necesarias en un nodo hoja.
	min_samples_split	Número mínimo de muestras necesarias para dividir un nodo interno en dos sub-nodos en cada árbol.

	Criterion	Función utilizada para evaluar la calidad de una división.

Las Tablas 4.4, 4.5 y 4.6 detallan los hiperparámetros utilizados para cada conjunto de datos, diferenciando entre los conjuntos completos y los conjuntos con reducción de dimensiones. Estos parámetros incluyen ajustes específicos para cada algoritmo, como el número de estimadores en Bosques Aleatorios, el valor de k en KNN, entre otros.

Tabla 4.4 Configuración de hiperparámetros del conjunto CD1, CD1_PCA90 y CD1_PCA95

Algoritmo	Hiperparámetro	CD1	CD1_PCA90	CD1_PCA95	UMAP
RF	*Criterion*	*entropy*	*entropy*	*gini*	*gini*
	max_deph	*9*	*26*	*23*	*29*
	max_features	*59*	*3*	*24*	*17*
	min_samples_leaf	*3*	*3*	*1*	*8*
	min_samples_split	*2*	*8*	*2*	*11*
	n_estimators	*47*	*52*	*16*	*45*
KNN	*n_neighbors*	*16*	*4*	*15*	*5*

Tabla 4.5 Configuración de hiperparámetros del conjunto CD10, CD10_PCA90 y CD10_PCA95

Algoritmo	Hiperparámetro	CD10	CD10_PCA90	CD10_PCA95	UMAP
RF	*Criterion*	*entropy*	*entropy*	*entropy*	*gini*
	max_deph	*40*	*9*	*15*	*37*
	max_features	*59*	*44*	*60*	*29*
	min_samples_leaf	*8*	*5*	*3*	*4*
	min_samples_split	*6*	*9*	*9*	*9*
	n_estimators	*98*	*11*	*96*	*62*
KNN	*n_neighbors*	*15*	*2*	*8*	*9*

Tabla 4.6 Configuración de hiperparámetros del conjunto CD100, CD100_PCA90 y CD100_PCA95

Algoritmo	Hiperparámetro	CD10	CD10_PCA90	CD10_PCA95	UMAP
RF	*Criterion*	*entropy*	*entropy*	*entropy*	*gini*
	max_deph	*40*	*7*	*41*	*37*
	max_features	*59*	*50*	*45*	*29*
	min_samples_leaf	*8*	*6*	*3*	*4*
	min_samples_split	*6*	*5*	*7*	*9*
	n_estimators	*98*	*80*	*72*	*62*
KNN	*n_neighbors*	*15*	*8*	*6*	*9*

4.3.2. Construcción de modelos sin reducción de dimensiones

En esta etapa del estudio, se crearon nueve modelos de clasificación con el objetivo de establecer un marco de referencia que permita comparar el rendimiento de los modelos que incorporan técnicas de reducción de dimensiones en fases posteriores. Para ello, se utilizaron tres algoritmos de clasificación ampliamente reconocidos: Máquinas de Soporte Vectorial, K-Vecinos Más Cercanos y Bosque Aleatorio. Cada uno de estos algoritmos fue entrenado utilizando tres conjuntos de datos diferentes: CD1, CD10 y CD100, los cuales tienen la misma cantidad de atributos y varía el número de registros.

Dado que la clase de mayor interés en este estudio es SARS, el análisis se centró en determinar cuál de los modelos ofrece el mejor rendimiento en la identificación de esta clase, sin embargo, también se tomará como referencia la clase NoSARS evaluando su desempeño entre modelos con el propósito de determinar si el desempeño de un modelo es equilibrado en ambas clases. Aunque se evalúan varias métricas de rendimiento, el puntaje F1, *F1-score*, se consideró la métrica más importante, ya que combina la precisión, *precision,* la proporción de verdaderos positivos entre los positivos identificados y la sensibilidad, *recall,* la proporción de verdaderos positivos entre todos los positivos reales. Esta se calcula como el promedio armónico de la precisión y la sensibilidad, ofreciendo un balance entre ambas métricas, lo cual es crucial en el contexto de la detección de SARS-CoV-2, donde tanto la identificación precisa como la cobertura de los casos positivos son fundamentales para evaluar la efectividad de los modelos de clasificación.

La Tabla 4.7 muestra el desempeño de los modelos entrenados utilizando el conjunto de datos CD1, que contiene 395 registros. Se observan dos modelos con buenos resultados. En primer lugar, el modelo RF destacó al obtener una precisión del 93% y una sensibilidad del 76% en la clase SARS, lo que indica un sólido rendimiento en la identificación de casos positivos. Además, este modelo logró un F1-Score de 0.83, lo que refuerza su balance entre precisión y sensibilidad. Este comportamiento es consistente en la clase NoSARS, donde RF también mostró una alta sensibilidad del 94% y una precisión del 81%, alcanzando un F1-Score de 0.87.

Por otro lado, KNN mostró un rendimiento equilibrado, con una precisión del 63% y una sensibilidad del 0.73 en la clase SARS, su F1-Score es de 0.68. En la clase NoSARS, KNN tuvo una precisión de 0.71 y una sensibilidad de 0.61, con un F1-Score del 0.66, lo que indica un desempeño razonable, aunque menos competitivo comparado con el modelo RF.

Finalmente, el modelo SVM presentó un desempeño inferior. Aunque alcanzó una precisión del 0.72 en la clase SARS, su sensibilidad fue significativamente menor, con un 0.55, resultando en un F1-Score del 0.62. Este resultado sugiere que SVM podría no ser tan eficaz en la identificación de casos SARS en comparación con RF y KNN.

El clasificador RF se destaca como el mejor modelo de clasificación para el conjunto CD1, especialmente en la clase SARS, gracias a su alta precisión y buen equilibrio entre las métricas. KNN y SVM, aunque presentan resultados aceptables, no alcanzan el mismo nivel de rendimiento que RF, especialmente en la identificación de casos SARS.

Tabla 4.7 Desempeño de los modelos sin reducción de dimensiones con el CD1

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.67	0.71	0.61	0.66
	SARS		0.63	0.73	0.68
RF	NoSARS	0.86	0.81	0.94	0.87
	SARS		0.93	0.76	0.83
SVM	NoSARS	0.68	0.66	0.81	0.73
	SARS		0.72	0.55	0.62

La Tabla 4.8 presenta el desempeño de los tres modelos de clasificación entrenados utilizando el conjunto de datos CD10, que contiene 3,950 registros. Al igual que con el conjunto CD1, el modelo de RF volvió a destacar, logrando una precisión del 0.92 y una sensibilidad del 0.71 en la clase SARS. Estos resultados indican que RF mantiene un sólido rendimiento en la identificación de casos positivos, demostrando su eficacia en escenarios con un mayor volumen de datos. Además, RF alcanzó un F1-Score de 0.80, lo que subraya su equilibrio entre precisión y sensibilidad, asegurando una buena performance en la clasificación de anticuerpos específicos de SARS-CoV-2.

El rendimiento de RF en la clase NoSARS también fue notable, con una sensibilidad del 0.95 y una precisión del 0.79, alcanzando un F1-Score de 0.86. Este comportamiento consistente sugiere que RF es altamente efectivo no solo en la detección de casos positivos, SARS, sino también en la identificación correcta de casos negativos, NoSARS.

Por otro lado, el modelo de KNN mostró un rendimiento inferior en comparación con RF, obteniendo una precisión del 67% y una sensibilidad del 71%, junto con un F1-Score de 0.68 en la clase SARS, lo cual sugiere un rendimiento aceptable, pero significativamente inferior al observado con RF. En la clase NoSARS, KNN tuvo un F1-Score del 0.72, lo que indica un desempeño razonable, aunque claramente inferior contra RF.

El modelo con el menor desempeño con este conjunto fue SVM, el cual mostró un desempeño inferior en comparación con RF y cercano al obtenido con KNN. En la clase SARS, SVM alcanzó una precisión del 0.80, pero su sensibilidad fue considerablemente menor, con un 0.52, dando un F1-Score del 0.63. Este resultado indica que SVM podría no ser tan eficaz en la identificación de casos SARS, lo que puede limitar su utilidad en aplicaciones donde la detección precisa de SARS es crítica. Sin embargo, SVM demostró un mejor rendimiento en la clase NoSARS, logrando un F1-Score de 0.77, lo que sugiere que SVM es más confiable

para detectar casos negativos en comparación con KNN, aunque con un menor rendimiento en la clase de mayor interés, SARS.

Tabla 4.8 Desempeño de los modelos sin reducción de dimensiones con el CD10

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.70	0.72	0.71	0.72
	SARS		0.67	0.71	0.68
RF	NoSARS	0.84	0.79	0.95	0.86
	SARS		0.92	0.71	0.80
SVM	NoSARS	0.72	0.68	0.89	0.77
	SARS		0.80	0.52	0.63

La Tabla 4.9 presenta el desempeño de los tres modelos de clasificación entrenados utilizando el conjunto de datos CD100, que contiene 39,500 registros. Al igual que en los casos anteriores, el modelo de RF volvió a destacarse, logrando la mayor precisión con un 0.93 y una sensibilidad del 0.67, dando como resultado un F1-Score del 0.78 en la clase SARS. En la clase NoSARS, RF mostró un rendimiento igualmente destacable, con un F1-Score de 0.83. Estos resultados indican que RF continúa siendo el modelo más efectivo, inclusive para este conjunto de datos con mayor número de registros, demostrando un rendimiento consistentemente equilibrado en ambas clases.

El modelo KNN nuevamente mostró el segundo mejor desempeño de los tres modelos. En la clase SARS, obtuvo una precisión del 0.87 y una sensibilidad del 0.63, con un F1-Score de 0.73. En la clase NoSARS, KNN alcanzó un F1-Score de 0.80.

Nuevamente, el modelo SVM se ve opacado por el desempeño de los otros modelos, ya que su precisión fue del 0.75, pero su sensibilidad solamente alcanzó el 0.52, ocasionando un F1-Score del 0.68. Se debe destacar que en la clase NoSARS su F1-Score fue de 0.81, siendo mejor que el obtenido por KNN, pero igualmente inferior que RF.

El modelo más destacado para este conjunto de datos vuelve a ser RF, lo cual indica que a pesar del incremento en el número de registros, este modelo puede aprender adecuadamente de ellos.

Tabla 4.9 Desempeño de los modelos sin reducción de dimensiones con el CD100

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.77	0.71	0.90	0.80
KNN	SARS		0.87	0.63	0.73
RF	NoSARS	0.81	0.75	0.95	0.83
RF	SARS		0.93	0.67	0.78
SVM	NoSARS	0.76	0.88	0.75	0.81
SVM	SARS		0.75	0.52	0.68

4.3.3. Construcción de modelos con reducción de dimensiones con PCA

En esta sección, se describe el impacto de la aplicación del Análisis de Componentes Principales en la reducción de la dimensionalidad de los conjuntos de datos y cómo esta técnica afecta el desempeño de los modelos de clasificación de anticuerpos. El objetivo principal de utilizar PCA es simplificar los conjuntos de datos al reducir el número de dimensiones, manteniendo al mismo tiempo la mayor cantidad posible de información relevante para la clasificación.

Se realizaron dos grupos de pruebas aplicando PCA, utilizando dos niveles diferentes de varianza acumulada: 90% y 95%. La intención es evaluar si un mayor porcentaje de varianza acumulada, 95%, mejora significativamente el rendimiento de los modelos de clasificación en comparación con un porcentaje de 90%, o si el nivel de varianza más bajo es suficiente para mantener un buen rendimiento con una menor complejidad computacional.

En la Figura 4.6, se presenta una representación gráfica de la varianza acumulada, mostrando cómo en los tres conjuntos de datos el 90% de la varianza se concentra en los primeros dos componentes principales, mientras que el resto se distribuye en los 98 componentes restantes.

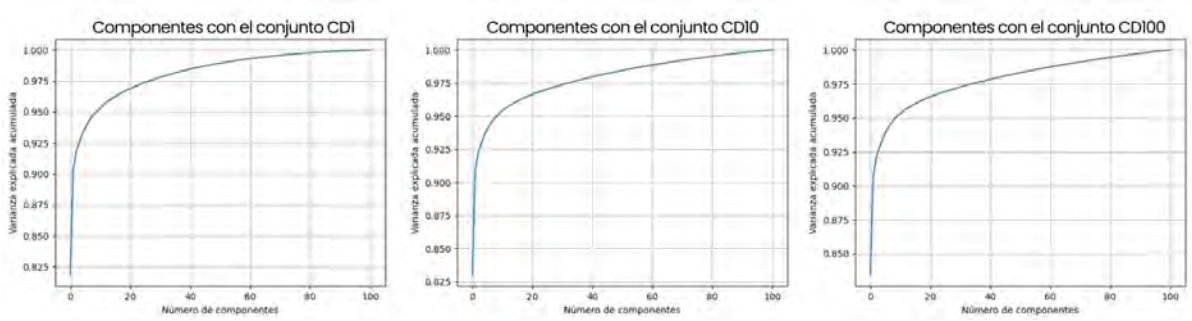

Figura 4.7. Número de componentes en los tres conjuntos de datos

La Tabla 4.10 complementa la figura 4.6 mostrando el número de componentes en cada nivel de varianza y la cantidad de varianza que se acumula en cada una de ellas, aunque solo se muestran las primeras 15 dimensiones son suficientes para observar el 90% y 95% de varianza acumulada. Esta tabla es importante para entender cómo se distribuye la información a través de las distintas dimensiones y cómo la reducción de dimensiones a dos componentes principales puede afectar el modelo.

Tabla 4.10 Varianza acumulada explicada en los tres conjuntos de datos.

CD1		CD10		CD100	
Componentes	*Varianza acumulada*	*Componentes*	*Varianza acumulada*	*Componentes*	*Varianza acumulada*
1	0.8186	1	0.8303	1	0.8345
2	0.9032	2	0.9092	2	0.9071
3	0.9182	3	0.9226	3	0.9207
4	0.9253	4	0.9295	4	0.9282
5	0.9320	5	0.9356	5	0.9340
6	0.9372	6	0.9402	6	0.9390
7	0.9420	7	0.9442	7	0.9432
8	0.9453	8	0.9472	8	0.9462
9	0.9483	9	0.9500	9	0.9490
10	0.9511	10	0.9523	10	0.9514
11	0.9536	11	0.9542	11	0.9533
12	0.9557	12	0.9558	12	0.9551
13	0.9576	13	0.9574	13	0.9566
14	0.9595	14	0.9589	14	0.9581
15	0.9611	15	0.9602	15	0.9594

4.3.3.1. Análisis de componentes principales con 90% de varianza

El análisis del desempeño de los modelos de clasificación cuando se entrenaron utilizando conjuntos de datos reducidos mediante PCA, manteniendo el 90% de la varianza acumulada, se realiza en esta sección. Este nivel de reducción de dimensiones tiene como objetivo capturar la mayor parte de la información significativa del conjunto de datos original mientras se descartan las características menos relevantes o redundantes.

Los conjuntos de datos CD1, CD10 y CD100 originales, que inicialmente contenían 100 dimensiones, fueron reducidos a tan solo 2 dimensiones mediante PCA. Esta drástica reducción en el número de dimensiones simplifica los datos, lo que puede facilitar el entrenamiento de los modelos y reducir el riesgo de sobreajuste. Los tres conjuntos de datos reducidos fueron nombrados como CD1_PCA90, CD10_PCA90 y CD100_PCA90 para reflejar que se aplicó PCA con el 90% de varianza acumulada. La Figura 4.7 ilustra este proceso de reducción de dimensiones.

Figura 4.8. Reducción de dimensiones con PCA 90% de varianza

En la Tabla 4.11, se presentan los resultados de los modelos entrenados con el conjunto CD1_PCA90, que fue reducido a 2 dimensiones y contiene 395 registros. Los modelos KNN y RF destacaron particularmente en la clase SARS, ambos alcanzando una precisión del 0.76, junto con una sensibilidad del 0.79, por lo cual muestran un F1-Score del 0.78. Estos resultados indican que tanto KNN como RF son altamente efectivos y similares en la identificación de casos positivos de SARS, manteniendo un equilibrio adecuado entre precisión y sensibilidad.

Por otro lado, el modelo SVM mostró un desempeño más modesto en la clase SARS. SVM logró una precisión del 0.75 y una sensibilidad del 0.71, lo que dio como resultado un F1-Score de 0.73. Aunque estos resultados son competitivos, SVM no logra igualar el rendimiento de KNN y RF en la identificación de casos positivos de SARS.

En la clase NoSARS, sin embargo, SVM se destacó, alcanzando un F1-Score del 0.84. Estos resultados sugieren que SVM es muy eficaz en la identificación de casos negativos, mostrando un rendimiento sobresaliente en esta clase. No obstante, dado que la clase SARS es la más relevante en el contexto de esta investigación, el menor rendimiento de SVM en la detección de casos positivos lo posiciona como el modelo con el desempeño menos favorable en comparación con los modelos más destacados KNN y RF.

Tabla 4.11 Desempeño de los modelos con el conjunto CD1_PCA90

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.78	0.80	0.78	0.79
	SARS		0.76	0.79	0.78
RF	NoSARS	0.78	0.80	0.78	0.79
	SARS		0.76	0.79	0.78
SVM	NoSARS	0.80	0.95	0.75	0.84
	SARS		0.75	0.71	0.73

La Tabla 4.12 presenta el desempeño de los tres modelos de clasificación entrenados utilizando el conjunto de datos CD10_PCA90, que contiene 3,950 registros. El modelo KNN muestra el mejor rendimiento en la clase SARS, logrando una precisión del 0.86 y una sensibilidad del 0.61, dando un F1-Score de 0.72.

En contraste, El modelo de RF destacó en la clase SARS, logró una alta precisión del 0.97, pero, su sensibilidad fue baja, logrando solo un 0.55, ocasionando un F1-Score del 0.71, lo cual es ligeramente inferior a KNN.

El modelo de SVM, por su parte, tuvo el desempeño más bajo en la clase SARS, con una precisión del 75%, una sensibilidad del 57%, y un F1-Score de 0.65. Estos resultados posicionan a SVM como el modelo menos efectivo en la detección de casos SARS. En la clase NoSARS, tanto RF como SVM presentaron los valores F1-Score más altos.

En este conjunto de datos, el modelo destacado es KNN, ya que tiene un buen equilibrio en sus métricas de la clase SARS y un destacado desempeño en la clase NoSARS.

Tabla 4.12 Desempeño de los modelos con el conjunto CD10_PCA90

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.77	0.73	0.92	0.81
	SARS		0.86	0.61	0.72
RF	NoSARS	0.78	0.72	0.98	0.83
	SARS		0.97	0.55	0.71
SVM	NoSARS	0.76	0.90	0.75	0.82
	SARS		0.75	0.57	0.65

La Tabla 4.13 muestra el rendimiento de los modelos de clasificación entrenados con el conjunto de datos CD100_PCA90, que contiene 39,500 registros. Nuevamente, el modelo KNN presenta el mejor rendimiento, destacándose por su equilibrio en la clase SARS, ya que logró una precisión del 0.86 y una sensibilidad del 0.62, con un F1-Score de 0.72. Aunque su precisión es menor en comparación con el modelo RF, KNN logra un mejor equilibrio entre precisión y sensibilidad.

El modelo RF, por otro lado, alcanzó la precisión más alta entre los tres modelos, con un 0.96 en la clase SARS. Sin embargo, su sensibilidad fue de solo 0.57, lo que refleja un desequilibrio significativo entre precisión y sensibilidad, tiene un F1-Score de 0.71. Estos resultados indican que, aunque RF es extremadamente preciso al identificar los casos positivos de SARS, tiene dificultades para capturar todos los casos.

El modelo SVM, aunque competitivo, se posiciona como el de menor desempeño en la clase SARS, ya que obtuvo una precisión del 0.75 y una sensibilidad del 0.66. Aunque su sensibilidad es superior a la de RF, SVM no logra alcanzar la misma precisión, lo que lo coloca en desventaja en escenarios donde ambas métricas, precisión y sensibilidad, son cruciales.

Para este conjunto de datos con un alto número de registros, el modelo más destacado es KNN, ya que ofrece un buen equilibrio entre precisión y sensibilidad en la clase SARS

Tabla 4.13 Desempeño de los modelos con el conjunto CD100_PCA90

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.76	0.71	0.90	0.79
	SARS		0.86	0.62	0.72
RF	NoSARS	0.77	0.70	0.97	0.81
	SARS		0.96	0.57	0.71
SVM	NoSARS	0.77	0.95	0.75	0.84
	SARS		0.75	0.66	0.70

4.3.3.2. Análisis de componentes principales con 95% de varianza

En este apartado se continúa con la técnica de reducción de dimensiones mediante Análisis de Componentes Principales, pero en esta ocasión se evalúa el desempeño de los algoritmos de clasificación cuando se entrenan con conjuntos de datos que retienen el 95% de la varianza acumulada. El objetivo es determinar si al capturar un porcentaje mayor de la varianza, y, por lo tanto, más información del conjunto de datos original, se mejora el rendimiento de los modelos de clasificación y si es que puede superar el desempeño de los modelos cuando fueron entrenados con los conjuntos de datos con todas sus dimensiones.

Con esta configuración, los conjuntos de datos CD1, CD10 y CD100 redujeron sus dimensiones de manera significativa, pero con un mayor número de componentes en comparación con el enfoque del 90% de varianza acumulada. Los conjuntos construidos son CD1_PCA95, CD10_PCA95 y CD100_PCA95, constando de 10, 9 y 10 dimensiones respectivamente.

Estos cambios en la dimensionalidad se ilustran en la Figura 4.7. Al comparar estos resultados con los obtenidos con un 90% de varianza, se podrá evaluar si el incremento en las dimensiones retenidas conlleva mejoras significativas en el rendimiento de los modelos, o si el aumento de la complejidad no justifica la ganancia en precisión y sensibilidad.

Figura 4.9. Reducción de dimensiones con PCA 95% de varianza

La Tabla 4.14 muestra el desempeño de los modelos de clasificación entrenados con el conjunto CD1_PCA95, el cual tiene 10 dimensiones de las 100 iniciales. Se destaca el modelo RF como el mejor clasificador.

En la clase SARS, RF logró una precisión del 0.87 y una notable sensibilidad del 0.82, alcanzando un F1-Score de 0.84. Este rendimiento sugiere que RF no solo es eficaz en la detección de casos positivos, sino que también mantiene un buen equilibrio, lo que lo convierte en la opción más robusta en este conjunto de datos.

El modelo KNN también mostró un desempeño sólido, en la clase SARS, ya que logró una alta precisión del 0.96 y sensibilidad del 0.70, con un F1-Score de 0.81. Su rendimiento en la clase SARS, aunque fuerte, es ligeramente inferior al de RF en términos de equilibrio general entre las métricas.

Finalmente, el modelo SVM mostró un rendimiento aceptable pero inferior en comparación con RF y KNN. En la clase SARS, SVM alcanzó una precisión del 0.72 y una sensibilidad del 0.74, resultando en un F1-Score de 0.73. En la clase NoSARS, SVM logró un F1-Score de 0.84, lo que indica un buen desempeño, aunque su capacidad para balancear precisión y sensibilidad en la clase SARS es menor en comparación con los otros modelos.

Tabla 4.14 Desempeño de los modelos con el conjunto CD1_PCA95

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.84	0.78	0.97	0.86
	SARS		0.96	0.70	0.81
RF	NoSARS	0.86	0.84	0.89	0.86
	SARS		0.87	0.82	0.84
SVM	NoSARS	0.80	0.92	0.77	0.84
	SARS		0.72	0.74	0.73

En la Tabla 4.15 se muestra el desempeño de los modelos sobre el conjunto CD10_PCA95, el modelo RF por segunda ocasión se posiciona como el mejor clasificador. En la clase SARS, RF alcanzó una precisión del 94% y una sensibilidad del 65%, llegando a un F1-Score de 0.77.

En segundo lugar, el modelo KNN también mostró un rendimiento competitivo, con una precisión del 0.86, una sensibilidad del 0.60, y un F1-Score de 0.71, sin embargo, en la clase NoSARS su F1-Score alcanzó el 0.81.

Finalmente, el modelo SVM mostró el rendimiento más bajo entre los tres. En la clase SARS, SVM logró una precisión del 0.77 y una sensibilidad del 0.58, con un F1-Score de 0.66, el más bajo de la tabla. En la clase NoSARS, aunque SVM alcanzó una precisión del 0.92, su F1-Score fue de 0.82, siendo superior a lo obtenido por KNN en esta clase, indicando un buen rendimiento en general, pero con menos balance en comparación con RF.

Tabla 4.15 Desempeño de los modelos con el conjunto CD10_PCA95

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.77	0.73	0.91	0.81
	SARS		0.86	0.60	0.71
RF	NoSARS	0.82	0.76	0.96	0.85
	SARS		0.94	0.65	0.77
SVM	NoSARS	0.77	0.92	0.74	0.82
	SARS		0.77	0.58	0.66

En la Tabla 4.16, se muestra el desempeño de los modelos en el conjunto CD100_PCA95, nuevamente el modelo RF se posiciona como el mejor clasificador, en la clase SARS, RF alcanzó una precisión del 0.94 y una sensibilidad del 0.68, con un F1-Score de 0.79, mientras que en la clase NoSARS logró un F1-Score del 0.84.

El segundo modelo con mejor rendimiento es KNN, que mostró un desempeño sólido en la clase SARS, con una precisión del 0.88 y una sensibilidad del 0.65, alcanzando un F1-Score de 0.74, mientras que en la clase NoSARS su F1-Score fue del 0.81.

Al igual que en los casos anteriores, el modelo SVM mostró un rendimiento aceptable, pero inferior en comparación con RF y KNN. En la clase SARS alcanzó una precisión del 0.71 y una sensibilidad del 0.73, logrando un F1-Score de 0.72. En la clase NoSARS, SVM obtuvo una precisión del 91% y un F1-Score de 0.83, lo cual es superior al obtenido por KNN, pero inferior a SVM.

Tabla 4.16 Desempeño de los modelos con el conjunto CD100_PCA95

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.78	0.72	0.91	0.81
	SARS		0.88	0.65	0.74
RF	NoSARS	0.82	0.75	0.96	0.84
	SARS		0.94	0.68	0.79
SVM	NoSARS	0.77	0.91	0.76	0.83
	SARS		0.71	0.73	0.72

4.3.4. Construcción de modelos con reducción de dimensiones con UMAP

En esta sección se describe la aplicación de la técnica de reducción de dimensiones UMAP, la cual difiere significativamente de PCA en su enfoque para reducir la dimensionalidad de los datos. A diferencia de PCA, UMAP no cuenta con un método directo y cuantitativo, como el valor de varianza acumulada, para determinar el número óptimo de dimensiones que capturan la mayor cantidad de información relevante del conjunto de datos.

Debido a esta particularidad de UMAP, fue necesario adoptar un enfoque empírico para identificar el número adecuado de dimensiones que permita a los modelos de clasificación alcanzar su mejor desempeño. Para ello, se generaron múltiples conjuntos de datos con distintas dimensionalidades a partir de los conjuntos iniciales CD1, CD10 y CD100. Específicamente, las 100 dimensiones originales se redujeron en un rango que va desde 1 hasta 90 dimensiones, como se ilustra en la Figura 4.9.

Una vez generados estos 90 conjuntos de datos derivados de cada conjunto inicial, se construyó los modelos de clasificación para cada uno de ellos. Se llevó a cabo una evaluación exhaustiva del desempeño de cada modelo, con el objetivo de identificar la configuración de dimensiones que ofreciera los mejores resultados en términos de precisión, sensibilidad y otras métricas relevantes.

Posteriormente, se seleccionaron los conjuntos de datos que presentaron el mejor desempeño para cada conjunto inicial CD1_UMAP el cual consta de 5 dimensiones, CD10_UMAP constando de 6 dimensiones y finalmente CD100_UMAP el cual mantiene 7 dimensiones. Los resultados detallados de esta selección se presentan en las Tablas 4.17, 4.18 y 4.19, donde se resume el rendimiento de los tres modelos de clasificación sobre los tres conjuntos de datos más optimizados. Este enfoque permitió maximizar la eficacia de UMAP en la reducción de dimensiones, logrando un equilibrio entre la simplicidad del modelo y la retención de información crítica, lo que se traduce en un mejor rendimiento general de los modelos de clasificación.

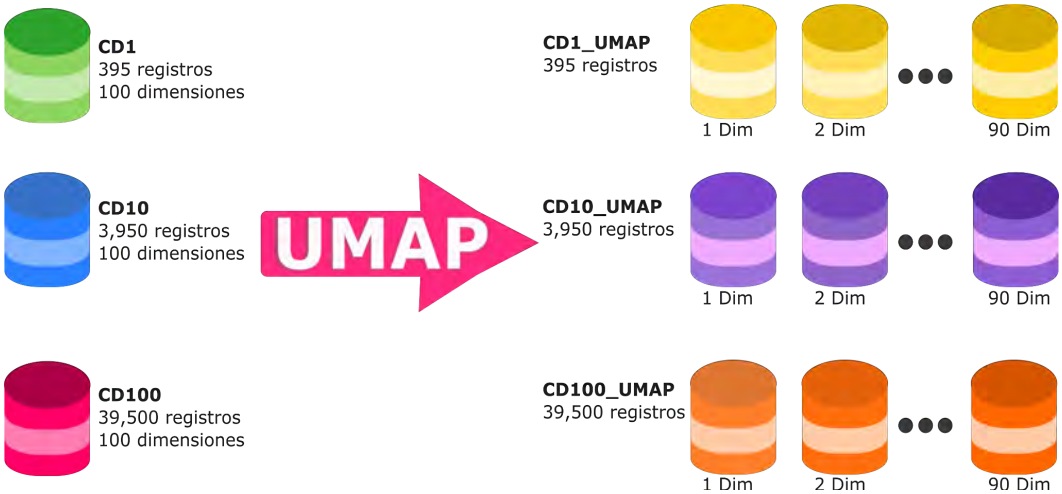

Figura 4.10. Reducción de dimensiones con UMAP desde 1 hasta 90 dimensiones

La Tabla 4.17 presenta el desempeño de los modelos de clasificación entrenados con el conjunto de datos CD1_UMAP.

El mejor modelo con este conjunto de datos es SVM, dado que su desempeño en la clase SARS alcanzó una precisión del 0.52, una sensibilidad del 0.85 y un F1-Score del 0.64. El F1-Score de la clase NoSARS alcanzó el 0.53.

El modelo RF logró el segundo mejor desempeño en la clase SARS, obteniendo una sensibilidad del 0.76 y un F1-Score de 0.63, a pesar de su baja precisión del 0.54. Estos sugiere que RF es eficaz para identificar la mayoría de los casos positivos de SARS, aunque tiene una tasa significativa de falsos positivos, lo cual se refleja en su moderada precisión.

Por otro lado, el modelo KNN obtuvo el rendimiento más bajo en la clase SARS, con una precisión del 52% y una sensibilidad del 65%, lo que dio como resultado un F1-Score de 0.58. En cuanto a la clase NoSARS, obtuvo una precisión del 68%, una sensibilidad de 0.56 y un F1-Score de 0.61.

Tabla 4.17 Desempeño de los modelos con el conjunto CD1_UMAP

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.59	0.68	0.56	0.61
	SARS		0.52	0.65	0.58
RF	NoSARS	0.62	0.74	0.51	0.61
	SARS		0.54	0.76	0.63
SVM	NoSARS	0.59	0.78	0.40	0.53
	SARS		0.52	0.85	0.64

La Tabla 4.18 presenta el rendimiento de los modelos de clasificación utilizando el conjunto de datos CD10_UMAP.

El modelo RF se destaca ligeramente sobre los otros en la clase SARS, logrando una precisión del 0.70 y una sensibilidad del 0.78, lo que se traduce en un F1-Score de 0.73. Esto sugieren que RF ofrece un equilibrio decente entre la capacidad de identificar correctamente los casos positivos de SARS y minimizar los falsos positivos.

El modelo KNN mejoró con respecto a su desempeño en el modelo anterior, con una precisión del 69% y una sensibilidad del 71%, obteniendo un F1-Score de 0.70. Aunque su desempeño con respecto al modelo entrenado con el conjunto CD1 mejoró, su capacidad para identificar correctamente los casos positivos es limitada, lo que lo coloca detrás del modelo RF en términos de rendimiento general para la clase SARS.

Por otro lado, el modelo SVM mostró un desempeño inferior. En la clase SARS, alcanzó una precisión del 0.57 y una sensibilidad del 0.83, con un F1-Score de 0.68. Si bien su sensibilidad fue la más alta de los tres modelos, su precisión también fue la más baja, lo que en promedio lo vuelve el modelo con menor desempeño.

Tabla 4.18 Desempeño de los modelos con el conjunto CD10_UMAP

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.66	0.63	0.61	0.62
	SARS		0.69	0.71	0.70
RF	NoSARS	0.69	0.68	0.58	0.63
	SARS		0.70	0.78	0.73
SVM	NoSARS	0.56	0.51	0.22	0.31
	SARS		0.57	0.83	0.68

La Tabla 4.19 muestra el rendimiento de los modelos de clasificación con el conjunto de datos CD100_UMAP. El modelo RF nuevamente demostró ser el más eficaz en la clase SARS, con una precisión del 0.72 y una sensibilidad del 0.75, lo que se refleja en un F1-Score de 0.73.

El modelo KNN presentó un rendimiento más moderado en la clase SARS, con una precisión del 0.71 y una sensibilidad del 0.73, dando un F1-Score de 0.72. Aunque su desempeño es aceptable, se queda por debajo del modelo RF, especialmente en términos de sensibilidad, lo que puede limitar su eficacia en la detección de todos los casos positivos.

Finalmente, el modelo SVM mostró el menor desempeño en la clase SARS, alcanzando una precisión del 0.59, pero con una sensibilidad notablemente mayor que la de los otros modelos, del 85%, lo que dio un F1-Score de 0.69. Sin embargo, aunque su sensibilidad fue muy buena en la clase SARS, en la clase NoSARS, alcanzó solo un 0.24, y un F1-Score del 0.34, este desempeño tan desequilibrado posicionan este modelo como el de peor calidad.

Tabla 4.19 Desempeño de los modelos con el conjunto CD100_UMAP

Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
KNN	NoSARS	0.68	0.64	0.61	0.63
	SARS		0.71	0.73	0.72
RF	NoSARS	0.69	0.66	0.62	0.64
	SARS		0.72	0.75	0.73
SVM	NoSARS	0.58	0.55	0.24	0.34
	SARS		0.59	0.85	0.69

4.5. Comparación de métricas de cada modelo

En este capítulo se ha presentado la construcción y evaluación de modelos de clasificación utilizando los conjuntos de datos iniciales CD1, CD10 y CD100, los cuales cuentan con 395, 3,950 y 39,500 registros respectivamente, y se entrenaron los algoritmos Máquinas de Soporte Vectorial, K-Vecinos Más Cercanos y Bosque Aleatorio. Estos modelos sirvieron como un marco de referencia para comparar su desempeño con modelos entrenados utilizando conjuntos de datos con menor dimensionalidad.

A lo largo del desarrollo experimental, se exploraron dos enfoques principales de reducción de dimensiones con el objetivo de mejorar el desempeño de los modelos de clasificación:

Reducción de Dimensiones mediante PCA: Este enfoque se implementó en dos variantes, utilizando 90% y 95% de varianza acumulada para determinar el número de dimensiones retenidas, con el propósito de identificar si mantener un mayor número de varianza acumulada mejora en el desempeño de los modelos.

En el caso del 90% de varianza acumulada, se construyeron los conjuntos CD1_PCA90, CD10_PCA90 y CD100_PCA90, que redujeron las 100 dimensiones originales a solo 2 dimensiones. Esta drástica reducción simplificó significativamente los conjuntos de datos, y los modelos entrenados con estos conjuntos mostraron un desempeño superior en la mayoría de los casos en comparación con los modelos entrenados con todas las dimensiones. Lo cual sugiere que la reducción de dimensiones para este tipo de datos mejora la capacidad de los modelos de aprendizaje automático para generalizar mejor.

Para el 95% de varianza acumulada, se crearon los conjuntos CD1_PCA95, CD10_PCA95 y CD100_PCA95, que retuvieron 10 dimensiones. Aunque esta configuración implicó un aumento en el número de dimensiones, los modelos mostraron una mejora significativa en el desempeño comparado con los modelos entrenados con todas las dimensiones y con aquellos que utilizaban solo 2 dimensiones. El incremento en la retención de varianza permitió conservar más información relevante, lo que se tradujo en un mejor rendimiento general en comparación con todos los modelos anteriores, aunque a costa de una menor reducción dimensional.

Reducción de Dimensiones mediante UMAP: Se aplicó la técnica UMAP para construir los conjuntos CD1_UMAP, CD10_UMAP y CD100_UMAP. A diferencia de PCA, UMAP no ofrece un método directo para seleccionar el número óptimo de dimensiones, por lo que se experimentó con una amplia gama de reducciones dimensionales, desde 1 hasta 90 dimensiones, para identificar el número de dimensiones que optimizara el rendimiento de los modelos.

Los modelos construidos con los conjuntos de datos reducidos por UMAP no lograron superar el rendimiento general de los modelos entrenados con PCA o los modelos iniciales con todas sus dimensiones. Sin embargo, en la métrica de sensibilidad en la clase SARS, UMAP se destacó por encima de todos los demás modelos. A pesar de este logro, el desempeño en la clase NoSARS fue inferior al de todos los otros modelos, lo que sugiere que UMAP no mantiene un buen equilibrio en las métricas de ambas clases.

Por lo tanto, aunque UMAP puede ser útil en situaciones específicas donde la detección precisa de SARS es prioritaria, no se aconseja su uso general en este contexto debido a su desequilibrio en el rendimiento entre las dos clases.

La Tabla 4.20 presenta el desempeño de todos los modelos exclusivamente en la clase SARS, donde se observa que en la mayoría de las métricas, los modelos reducidos por PCA con 90% y 95% de varianza acumulada demostraron tener el mejor rendimiento. Esto refuerza la conclusión de que PCA, y particularmente la configuración con 95% de varianza acumulada, es la técnica más eficaz para la reducción de dimensiones en el contexto de la clasificación de secuencias de anticuerpos del SARS-CoV-2.

Tabla 4.20 Desempeño de todos los modelos en la clase SARS

		Exactitud				Precisión				Sensibilidad				F1-Score			
Conjunto	***Modelo***	***SRD***	***PCA90***	***PCA95***	***UMAP***	***SRD***	***PCA90***	***PCA95***	***UMAP***	***SRD***	***PCA90***	***PCA95***	***UMAP***	***SRD***	***PCA90***	***PCA95***	***UMAP***
CD1	*KNN*	0.67	0.78	0.84	0.59	0.63	0.76	0.96	0.52	0.73	0.79	0.70	0.65	0.68	0.78	0.81	0.58
	RF	0.86	0.78	0.86	0.62	0.93	0.76	0.87	0.54	0.76	0.79	0.82	0.76	0.83	0.78	0.84	0.63
	SVM	0.68	0.80	0.80	0.59	0.72	0.75	0.72	0.52	0.55	0.71	0.74	0.85	0.62	0.73	0.73	0.64
CD10	*KNN*	0.70	0.77	0.77	0.66	0.67	0.86	0.86	0.69	0.71	0.61	0.60	0.71	0.68	0.72	0.71	0.70
	RF	0.84	0.78	0.82	0.69	0.92	0.97	0.94	0.70	0.71	0.55	0.65	0.78	0.80	0.71	0.77	0.73
	SVM	0.72	0.76	0.77	0.56	0.80	0.75	0.77	0.57	0.52	0.57	0.58	0.83	0.63	0.65	0.66	0.68
CD100	*KNN*	0.77	0.76	0.78	0.68	0.87	0.86	0.88	0.71	0.63	0.62	0.65	0.73	0.73	0.72	0.74	0.72
	RF	0.81	0.77	0.82	0.69	0.93	0.96	0.94	0.72	0.67	0.57	0.68	0.75	0.78	0.71	0.79	0.73
	SVM	0.76	0.77	0.77	0.58	0.75	0.75	0.71	0.59	0.52	0.66	0.73	0.85	0.68	0.70	0.72	0.69

Capítulo 5

Resultados

"La suerte solo favorece a la mente preparada"

Pasteur, 1854.

5. Resultados

En este capítulo se presenta la comparación del desempeño de los diversos modelos de clasificación que se construyeron utilizando diferentes enfoques de reducción de dimensionalidad. Esta evaluación se realiza sobre los tres grupos de modelos en contextos distintos:

Modelos entrenados con los conjuntos sin reducción de dimensionalidad: Estos modelos se basaron en los conjuntos iniciales CD1, CD10 y CD100, y sirven como una base de comparación de desempeño frente a los modelos construidos con conjuntos de datos reducidos.

Modelos entrenados con conjuntos reducidos mediante PCA: Aquí se aplicaron dos niveles de retención de varianza acumulada, 90% y 95%, para reducir las dimensiones de los conjuntos de datos. Estos modelos permiten evaluar cómo la cantidad de varianza retenida afecta el rendimiento de los clasificadores.

Modelos entrenados con conjuntos reducidos mediante UMAP: En este enfoque, se utilizó la técnica UMAP para reducir la dimensionalidad, generando diferentes versiones de los conjuntos de datos originales para identificar cómo UMAP impacta en el rendimiento comparado con PCA y los datos sin reducción.

En la Tabla 5.1 se listan los mejores modelos seleccionados con base en la métrica F1-Score. Esta métrica fue elegida por su capacidad para proporcionar una visión equilibrada entre precisión y sensibilidad, ofreciendo un panorama general del rendimiento de cada modelo. Para mejorar la comprensión de los resultados, cada modelo ha sido nombrado de la siguiente manera: comienza con las siglas del nombre del algoritmo utilizado, seguido del conjunto de datos empleado, y termina con la técnica de reducción de dimensiones aplicada sobre el conjunto. Para los modelos base que fueron creados sin reducción de dimensiones, se utiliza la etiqueta SRD, sin reducción de dimensiones.

Este esquema de nomenclatura facilita la identificación y comparación directa de los modelos bajo diferentes condiciones, permitiendo una evaluación clara y precisa de cómo las técnicas de reducción de dimensiones influyen en el rendimiento de los modelos de clasificación de anticuerpos del SARS-CoV-2.

Tabla 5.1 Nombre de los mejores modelos construidos con cada conjunto

Algoritmo de clasificación	Conjunto de datos	Técnica de reducción	Título de los modelos
RF	CD1	SRD	RF_CD1_SRD
RF	CD10	SRD	RF_CD10_SRD
RF	CD100	SRD	RF_CD100_SRD
KNN	CD1_PCA90	PCA90	KNN_CD1_PCA90
KNN	CD10_PCA90	PCA90	KNN_CD10_PCA90
KNN	CD100_PCA90	PCA90	KNN_CD100_PCA90

Algoritmo de clasificación	Conjunto de datos	Técnica de reducción	Título de los modelos
RF	CD1_PCA95	PCA95	RF_CD1_PCA95
RF	CD10_PCA95	PCA95	RF_CD10_PCA95
RF	CD100_PCA95	PCA95	RF_CD100_PCA95
SVM	CD1_UMAP	UMAP	SVM_CD1_UMAP
RF	CD10_UMAP	UMAP	RF_CD10_UMAP
RF	CD100_UMAP	UMAP	RF_CD100_UMAP

5.1. Comparación del desempeño de modelos entrenados con datos sin reducción de dimensiones

En la sección 4.5 del Capítulo 4 se construyeron nueve modelos de clasificación utilizando los algoritmos Máquinas de Soporte Vectorial, SVM, K-Vecinos Más Cercanos, KNN y Bosques Aleatorios, RF. Estos modelos fueron entrenados con los conjuntos de datos CD1, CD10 y CD100, que contienen las mismas 100 dimensiones, pero difieren en el número de registros contenidos en cada conjunto.

Los resultados obtenidos para cada uno de estos modelos se presentan en las Tablas 4.11, 4.12 y 4.13 del Capítulo 4. La tabla 5.2 de esta sección lista el mejor modelo de cada uno de los conjuntos. En todos los casos, el algoritmo que mostró el mejor desempeño fue RF, de manera que los modelos con el mejor rendimiento en esta etapa son: RF_CD1_SRD, RF_CD10_SRD y RF_CD100_SRD.

El modelo RF_CD1_SRD se destaca en color verde por ser el que obtuvo el mejor rendimiento entre los tres modelos sin reducción de dimensiones. Este modelo, RF_CD1_SRD, no solo sobresale en la clase SARS en comparación con los otros modelos, sino que también muestra un excelente rendimiento en la clase NoSARS, lo que indica un buen equilibrio en la clasificación de ambas clases. Esto es particularmente importante en el contexto de clasificación de secuencias de anticuerpos, donde el balance entre precisión y sensibilidad en ambas clases es crucial para asegurar un modelo robusto.

Una observación adicional sobre el F1-Score de los modelos entrenados es que este disminuye a medida que aumenta el número de registros en los conjuntos de datos, pasando de 0.83 en RF_CD1_SRD a 0.78 en RF_CD100_SRD. Aunque esta disminución no es drástica, es significativa y podría indicar que al aumentar el número de registros se podrían haber incluido registros mal etiquetados o con mayor ruido, lo que afecta el rendimiento del modelo.

Tabla 5.2 F1-Score de los mejores modelos construidos con conjuntos sin reducción de dimensiones.

Conjunto	Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
CD1	RF	NoSARS	0.86	0.81	0.94	0.87
		SARS		0.93	0.76	0.83
CD10	RF	NoSARS	0.84	0.79	0.95	0.86
		SARS		0.92	0.71	0.80
CD100	RF	NoSARS	0.81	0.75	0.95	0.83
		SARS		0.93	0.67	0.78

5.3. Comparación del desempeño de modelos entenados con datos reducidos con PCA

5.3.1. Modelos entrenados con conjuntos reducidos con PCA con 90% de varianza

En la sección 4.6 del Capítulo 4, se construyeron nueve modelos de clasificación utilizando los algoritmos SVM, KNN y RF. Estos modelos fueron entrenados con los conjuntos de datos CD1_PCA90, CD10_PCA90 y CD100_PCA90, que se redujeron de 100 dimensiones a solo 2 dimensiones con PCA, manteniendo el 90% de la varianza acumulada.

Los resultados obtenidos para cada uno de estos modelos se detallan en las Tablas 4.14, 4.15 y 4.16 del Capítulo 4. La Tabla 5.3 lista los modelos con mejor desempeño en cada uno de los conjuntos. A diferencia del caso anterior, donde los tres mejores modelos fueron los basados en RF, en este caso, el algoritmo KNN mostró un mejor desempeño. En particular, el modelo KNN_CD1_PCA90 se destaca en verde como el mejor de los tres, alcanzando un F1-Score de 0.78, siendo superior a los otros dos modelos.

Aunque en la Tabla 5.3 solo se muestra el mejor desempeño de cada conjunto, es importante mencionar que los modelos construidos con el algoritmo RF tuvieron un desempeño muy cercano al de los modelos KNN. Sin embargo, se observa nuevamente el fenómeno que ocurrió con los modelos sin reducción de dimensiones: el F1-Score disminuye a medida que el número de registros aumenta. Esto reafirma la idea de que incrementar el número de registros podría haber introducido datos mal etiquetados o con mayor ruido, lo que afectó el rendimiento.

Otra característica a destacar es que, aunque los mejores modelos construidos con PCA y el 90% de varianza tuvieron un desempeño ligeramente inferior en comparación con los modelos sin reducción de dimensiones, estos modelos mostraron un rendimiento muy sólido con solo dos dimensiones. Este resultado subraya la efectividad de la reducción de dimensiones en mantener un buen rendimiento mientras simplifica el modelo.

Por otro lado, al comparar el desempeño de los modelos que no figuran en la Tabla 5.3, es decir, los que no fueron los mejores, se observa un aumento considerable en su rendimiento. Esto significa que, aunque el rendimiento de los mejores modelos decreció ligeramente, el desempeño general de los otros modelos mejoró, lo que sugiere que la reducción de dimensiones mediante PCA no solo simplifica los modelos, sino que también puede ayudar a equilibrar el rendimiento entre los diferentes algoritmos y conjuntos de datos.

Tabla 5.3 F1-Score de los mejores modelos construidos con conjuntos reducidos con PCA reteniendo el 90% de varianza

Conjunto	Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
CD1	KNN	NoSARS	0.78	0.80	0.78	0.79
		SARS		0.76	0.79	0.78
CD10	KNN	NoSARS	0.77	0.73	0.92	0.81
		SARS		0.86	0.61	0.72
CD100	KNN	NoSARS	0.76	0.71	0.90	0.79
		SARS		0.86	0.62	0.72

5.3.2. Modelos entrenados con conjuntos reducidos con PCA con 95% de varianza

En la sección 4.7 del Capítulo 4, se construyeron nueve modelos de clasificación utilizando los algoritmos SVM, KNN y RF. Estos modelos fueron entrenados con los conjuntos de datos CD1_PCA95, CD10_PCA95 y CD100_PCA95, cuya dimensionalidad se redujo de 100 dimensiones a 10, 9 y 10 dimensiones respectivamente, manteniendo el 95% de la varianza acumulada. Este enfoque representó un aumento en la dimensionalidad comparado con los conjuntos donde se mantuvo solo el 90% de varianza y se redujo a 2 componentes en cada conjunto.

Los resultados obtenidos para cada uno de estos modelos se detallan en las Tablas 4.17, 4.18 y 4.19 del Capítulo 4. La Tabla 5.4 muestra los mejores resultados de cada uno de los conjuntos. Nuevamente, el mejor modelo fue aquel entrenado con un menor número de registros, destacándose el modelo RF_CD1_PCA95, resaltado en color verde, ya que mostró el mejor desempeño de los tres.

Una vez más, se percibe un decremento en el desempeño a medida que el número de registros aumenta. Al comparar el desempeño de los mejores modelos construidos con los conjuntos reducidos manteniendo el 95% de varianza contra aquellos que mantuvieron el 90% de varianza, se observa que el incremento del 5% de varianza acumulada mejora significativamente los resultados. De hecho, el modelo RF_CD100_PCA95, que tuvo el menor desempeño de los tres modelos en este grupo, fue aún superior al mejor modelo del grupo entrenado con el 90% de varianza, que fue KNN_CD1_PCA90. Esto demuestra que el uso del 95% de la varianza aporta una mejora significativa en los modelos de clasificación.

Por otro lado, en comparación con el desempeño de los modelos sin reducción de dimensiones, que utilizan las 100 dimensiones originales para entrenar sus modelos, los modelos con el 95% de varianza superan a los modelos sin reducción de dimensiones, estos son mejores no solo en el caso de los tres modelos más destacados, sino tambien en aquellos modelos que no figuran en la Tabla 5.4, incluso esos modelos incrementaron significativamente su desempeño en comparación con los modelos construidos en los casos anteriores. Esto significa que, utilizando solo una décima parte de las dimensiones originales, se puede obtener un mejor desempeño en los modelos de clasificación.

Estos resultados subrayan la importancia de retener una mayor cantidad de varianza al realizar la reducción dimensional, ya que no solo se logra compresión significativa de los datos, sino que también se mejora el rendimiento general de los modelos de clasificación en este tipo de datos.

Tabla 5.4 F1-Score de los mejores modelos construidos con conjuntos reducidos con PCA reteniendo el 95% de varianza

Conjunto	Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
CD1	RF	NoSARS	0.86	0.84	0.89	0.86
		SARS		0.87	0.82	0.84
CD10	RF	NoSARS	0.82	0.76	0.96	0.85
		SARS		0.94	0.65	0.77
CD100	RF	NoSARS	0.82	0.75	0.96	0.84
		SARS		0.94	0.68	0.79

5.4. Comparación del desempeño de modelos entenados con datos reducidos con UMAP

En la sección 4.7 del Capítulo 4, se construyeron los últimos nueve modelos de clasificación utilizando los algoritmos SVM, KNN y RF. Estos modelos fueron entrenados con los conjuntos de datos CD1_UMAP, CD10_UMAP y CD100_UMAP, que mantenían 5, 6 y 7 dimensiones respectivamente. En la Tabla 5.5 se muestran los tres mejores modelos de los nueve construidos.

A diferencia de los casos anteriores, donde el F1-Score tendía a disminuir a medida que aumentaba el número de registros, en este caso, el F1-Score no disminuyó, sino que incrementó. Además, el mejor desempeño entre los modelos se dio entre aquellos con un mayor número de registros, destacándose los modelos RF_CD10_UMAP y RF_CD100_UMAP, cuyos resultados fueron casi similares.

Sin embargo, a pesar de que los modelos reducidos con UMAP muestran un desempeño adecuado en la clase SARS, presentan un rendimiento significativamente inferior en la clase NoSARS. Esto indica que no existe un equilibrio en el desempeño de ambas clases, lo cual es un factor crítico en la evaluación de la eficacia de los modelos de clasificación. Un modelo que no mantiene un buen equilibrio entre las clases puede ser menos confiable en situaciones donde es esencial detectar tanto casos positivos como negativos con alta precisión.

Además, es importante destacar que, aunque los modelos construidos con los conjuntos de datos reducidos con UMAP mostraron un desempeño aceptable, no alcanzaron el desempeño de los modelos sin reducción de dimensiones ni de aquellos reducidos mediante PCA. Esto sugiere que, aunque UMAP puede ser útil en ciertos contextos, en este caso no proporcionó una ventaja competitiva significativa en términos de rendimiento global.

Una métrica en la que los modelos entrenados con UMAP sí destacaron fue la sensibilidad en la clase SARS, que fue la más alta en comparación con los demás modelos. Sin embargo, este incremento en la sensibilidad vino acompañado de una precisión mucho más baja, lo que resultó en una reducción considerable del F1-Score. Este desequilibrio entre sensibilidad y precisión contribuyó a que estos modelos fueran los de menor desempeño general entre todos los evaluados.

Tabla 5.5 F1-Score de los mejores modelos construidos con conjuntos reducidos con UMAP

Conjunto	Modelo	Clase	Exactitud	Precisión	Sensibilidad	F1-Score
CD1	SVM	NoSARS	0.59	0.78	0.40	0.53
		SARS		0.52	0.85	0.64
CD10	RF	NoSARS	0.69	0.68	0.58	0.63
		SARS		0.70	0.78	0.73
CD100	RF	NoSARS	0.69	0.66	0.62	0.64
		SARS		0.72	0.75	0.73

5.5. Conclusión de la evaluación de resultados

En este capítulo se ha evaluado el desempeño de los mejores modelos de clasificación construidos a partir de diferentes conjuntos de datos: sin reducción de dimensionalidad y con reducción mediante las técnicas de PCA y UMAP. Esta doble comparación tenía dos objetivos principales: identificar el modelo más destacado en cada grupo de reducción de dimensiones y determinar en qué situaciones la reducción de dimensionalidad resulta beneficiosa.

Los modelos con los F1-Score más altos dentro de cada técnica de reducción de dimensiones y sin reducción fueron: *RF_CD1_SRD, KNN_CD1_PCA90, RF_CD1_PCA95, y RF_CD10_UMAP*. Estos modelos destacaron por su alto rendimiento en sus respectivas configuraciones. En la mayoría de los casos, el algoritmo de clasificación con mejor desempeño fue el de Bosques Aleatorios, RF, seguido de K-Vecinos más Cercanos, KNN.

De las técnicas de reducción de dimensionalidad, PCA se destacó como la más efectiva para mantener una buena representación de los datos, al mismo tiempo que reducía considerablemente el número de dimensiones. Especialmente, cuando se conservó el 95% de la varianza acumulada, de esta manera los conjuntos de datos reducidos ocupaban solo una décima parte de las dimensiones originales, y su desempeño era superior al de los modelos entrenados con los datos sin reducción. Este resultado es significativo, ya que demuestra que PCA no solo reduce la complejidad de los datos, sino que también puede mejorar el rendimiento de los modelos.

El análisis también reveló un patrón claro que, el rendimiento de los modelos tendía a decrecer a medida que aumentaba el número de registros en el conjunto de datos. Los modelos construidos con conjuntos de datos más pequeños lograron mejores resultados, mientras que aquellos entrenados con los conjuntos de datos más grandes presentaron un peor desempeño. Esto sugiere que la presencia de registros mal etiquetados es más problemática en los conjuntos de datos grandes, lo que afecta negativamente la precisión del modelo.

En términos generales, la técnica de reducción de dimensiones recomendada es PCA, ya que permitió obtener buenos resultados tanto con el 90% y los mejores resultados con el 95% de varianza acumulada. Cuando se retiene el 90% de la varianza, se reduce el conjunto de datos a solo dos dimensiones, lo que es útil para análisis exploratorios. Por otro lado, reteniendo el 95% de la varianza, se mantiene una décima parte de las dimensiones originales, ofreciendo el mejor equilibrio entre rendimiento y simplicidad.

Finalmente, aunque la Tabla 5.1 destaca los mejores modelos, cabe mencionar que otros modelos también mostraron mejoras notables. Por ejemplo, los modelos de KNN incrementaron significativamente su rendimiento cuando se construyeron con un menor número de dimensiones, lo que sugiere que todos los algoritmos de clasificación se beneficiaron de la reducción de dimensionalidad en términos de desempeño.

Capítulo 6

Conclusiones y trabajo futuro

“Slow down you're doing fine”

Vienna (Billy Joel, 1977)

6.1. Conclusiones

El problema que motivó esta investigación fue la necesidad de desarrollar modelos de aprendizaje automático capaces de clasificar secuencias génicas de anticuerpos a partir de grandes conjuntos de datos genómicos. Estos conjuntos de secuencias de anticuerpos, diseñados para ser analizados por bioinformáticos o biólogos moleculares, requerían una solución para extraer información útil y clasificar las secuencias del SARS-CoV-2 sin la intervención constante de expertos humanos para clasificar cada secuencia manualmente. La principal complejidad de los datos radicaba en su representación inicial en aminoácidos y su alta dimensionalidad, lo que representaba un desafío significativo para los métodos tradicionales de análisis.

La metodología implementada en este trabajo proporcionó una solución clara a este desafío. Se combinó el procesamiento de lenguaje natural enfocado en datos génicos, lo que permitió transformar las secuencias de aminoácidos en representaciones numéricas, como *word embeddings*. Además, se emplearon modelos avanzados de aprendizaje automático configurados mediante la configuración de hiperparámetros, lo que permitió construir modelos capaces de clasificar secuencias de aminoácidos pertenecientes a anticuerpos del SARS-CoV-2 y diferenciarlas de las que no pertenecen a este virus.

A partir del análisis del estado del arte se identificaron diversos enfoques para abordar problemas similares en genómica, que van desde métodos tradicionales de selección de características a partir de datos distintos a secuencias de aminoácidos hasta técnicas recientes basadas en *embeddings* de secuencias y modelos de deep learning. Sin embargo, la combinación de PLN y aprendizaje automático presentada en este estudio demostró ser especialmente efectiva para el análisis de secuencias biológicas. El uso de *ProtVec* facilitó la captura de patrones funcionales en las secuencias de anticuerpos, permitiendo su conversión en vectores numéricos densos que mejoraron la posterior clasificación.

Los tres modelos de aprendizaje automático, Bosques Aleatorios, K-Vecinos Más Cercanos y Máquinas de Soporte Vectorial utilizados en este trabajo demostraron ser adecuados para la clasificación precisa de los datos de anticuerpos pertenecientes al SARS-CoV-2. Además, la reducción de dimensionalidad no solo mejoró la velocidad de procesamiento de los modelos, sino que también aumentó la precisión general. En particular, el uso de PCA con un 95% de varianza acumulada se destacó como la mejor técnica de reducción dimensional, ya que redujo significativamente la complejidad del conjunto de datos mientras mantenía un alto nivel de rendimiento de clasificación.

El mejor modelo en este estudio fue Random Forest aplicado a los datos del conjunto CD1 reducidos mediante PCA manteniendo el 95% de la varianza acumulada, RF_CD1_PCA95. Este modelo alcanzó una exactitud del 86%, una precisión del 87%, una sensibilidad del 82% y un F1-Score del 84%, superando a otros modelos. Esto confirma que, en este caso, Random Forest fue la mejor estrategia para abordar el problema inicial de clasificar eficientemente secuencias de anticuerpos en un contexto de alta dimensionalidad.

Además, los mejores resultados se obtuvieron cuando se utilizó PCA para reducir las dimensiones, lo que indica que, para este tipo de datos, representaciones numéricas de secuencias biológicas, PCA es la técnica que mejor reduce sus dimensiones mientras mantiene las relaciones semánticas subyacentes en los datos.

6.2. Trabajo futuro

Durante el desarrollo de esta tesis, se han identificado áreas de investigación que podrían ser exploradas por futuros estudiantes de maestría. Uno de los temas más relevantes es la identificación de herramientas para el remuestreo de datos en secuencias biológicas.

El trabajo futuro en esta línea tendría como objetivo identificar o desarrollar un método o herramienta para remuestrear secuencias biológicas que han sido transformadas en representaciones numéricas o *word embeddings*. Este desafío surge debido a que los *word embeddings* generados con herramientas como *ProtVec* o similares a Word2Vec producen arreglos numéricos cuyas dimensiones contienen información que es difícil de interpretar. Estas dimensiones no siempre tienen un significado claro, lo que dificulta la generación de valores aleatorios o la modificación directa de estas representaciones numéricas sin comprometer su coherencia semántica o su precisión.

Abordar este problema implicaría investigar y desarrollar nuevas técnicas de remuestreo capaces de manejar las características específicas de los datos biológicos transformados en secuencias numéricas. Actualmente, no existen soluciones adecuadas para el remuestreo de secuencias biológicas representadas como *word embeddings*, por lo que cualquier avance en este campo podría tener un impacto significativo en el análisis de datos biológicos y genómicos.

Este enfoque no solo contribuiría al campo de la bioinformática, sino que también tendría aplicaciones en otras áreas como el procesamiento de lenguaje natural, donde los *word embeddings* son utilizados extensamente. Al proporcionar herramientas más precisas y efectivas para el remuestreo de datos en estos dominios, se facilitaría un análisis más robusto y confiable de las secuencias biológicas transformadas en datos numéricos, permitiendo a los modelos de aprendizaje automático trabajar de manera más eficiente con datos balanceados y mejor representados.

El desarrollo de nuevas herramientas y técnicas de remuestreo mejoraría la capacidad de los investigadores para manejar grandes conjuntos de datos biológicos transformados, proporcionando modelos más precisos y menos propensos a sesgos, lo que, a su vez, podría acelerar descubrimientos y avances en áreas relacionadas con la salud y la genómica.

6.3. Actividades Académicas Adicionales

Tabla 6.2 Actividades académicas

Semestre	**Actividad**
Segundo	Autor del artículo: *Revisión de la literatura sobre técnicas de reducción de la dimensionalidad y su aplicación en datos de anticuerpos de SARS-CoV-2* Co-Autor del artículo: *Modelado conceptual del proceso de anotación genómica del segmento V*
Tercero	Autor del artículo: *Técnicas de reducción de la dimensionalidad en la predicción de anticuerpos del SARS-CoV-2* Co-Autor del artículo: *Revisión de la literatura de sistemas de anotación automática de genes V(D)J* Co-Autor del artículo: *Modelo de aprendizaje automático para predecir la deserción escolar* Instructor del curso: *Curso básico de lengua de señas mexicanas*
Cuarto	Co-Autor del artículo: *School Dropout Prediction with class balancing and hyperparameter configuration*

Referencias bibliográficas

Adjuik, T. A., & Ananey-Obiri, D. (2022). Word2vec neural model-based technique to generate protein vectors for combating COVID-19: A machine learning approach. *International Journal of Information Technology*, *14*(7), 3291-3299. https://doi.org/10.1007/s41870-022-00949-2

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1-8. https://doi.org/10.1016/j.cbpa.2021.04.005

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Arowolo, M. O., Adebiyi, M. O., Aremu, C., & Adebiyi, A. A. (2021). A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector. *Journal of Big Data*, *8*(1), 50. https://doi.org/10.1186/s40537-021-00441-x

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, *10*(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38-44. https://doi.org/10.1038/nbt.4314

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, *24*. https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A

universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102-2110. https://doi.org/10.1093/bioinformatics/btac020

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator: Figure 1. *Genome Research*, *14*(6), 1188-1190. https://doi.org/10.1101/gr.849004

Diggins, K. E., Ferrell, P. B., & Irish, J. M. (2015). Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. *Methods*, *82*, 55-63. https://doi.org/10.1016/j.ymeth.2015.05.008

Enríquez, M., Naranjo, S., Amaro, I., & Camacho, F. (2021). Dimensionality Reduction Using PCA and CUR Algorithm for Data on COVID-19 Tests. En M. Botto-Tobar, H. Cruz, & A. Díaz Cadena (Eds.), *Artificial Intelligence, Computer and Software Engineering Advances* (Vol. 1326, pp. 121-134). Springer International Publishing. https://doi.org/10.1007/978-3-030-68080-0_9

Fischer, D. S., Wu, Y., Schubert, B., & Theis, F. J. (2020). Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Molecular Systems Biology*, *16*(8), e9416. https://doi.org/10.15252/msb.20199416

Frankl, V. (1946). *El hombre en busca de sentido*. Herder Editorial.

García Merino, A. (2011). Anticuerpos monoclonales. Aspectos básicos. *Neurología*, *26*(5), 301-306. https://doi.org/10.1016/j.nrl.2010.10.005

Gorshtein, G. (2022, agosto 10). *Structure and Function of Antibodies*. Rapid Novor. https://www.rapidnovor.com/structure-and-function-of-antibodies/

Greaney, A. J., Loes, A. N., Crawford, K. H. D., Starr, T. N., Malone, K. D., Chu, H. Y., & Bloom, J. D. (2021). Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host & Microbe*, *29*(3), 463-476.e6.

https://doi.org/10.1016/j.chom.2021.02.003

Harari, Y. N. (2015). *Sapiens: A brief history of humankind* (First U.S. edition). Harper.

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, *19*(7), 409-424. https://doi.org/10.1038/s41579-021-00573-0

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*(2), 271-280.e8. https://doi.org/10.1016/j.cell.2020.02.052

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (No. arXiv:1502.03167). arXiv. https://doi.org/10.48550/arXiv.1502.03167

Khalilian, S., Nasr Isfahani, M., Moti, Z., Baloochestani, A., Chavosh, A., & Hemmatian, Z. (2022). *A Deep Dimensionality Reduction method based on Variational Autoencoder for Antibody Complementarity Determining Region Sequences Analysis*. 116-105. https://doi.org/10.29007/x25c

Kramer, O. (2013). K-Nearest Neighbors. En O. Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors* (Vol. 51, pp. 13-23). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38652-7_2

Lefranc, M.-P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., Hadi-Saljoqi, S., Sasorith, S., Lefranc, G., & Kossida, S. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Research*, *43*(D1), D413-D422.

https://doi.org/10.1093/nar/gku1056

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *WIREs Computational Statistics*, *1*(3), 283-289. https://doi.org/10.1002/wics.49

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (No. arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.48550/ARXIV.1301.3781

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, *15*(6), e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Obaid, H. S., Dheyab, S. A., & Sabry, S. S. (2019). The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning. *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, 279-283. https://doi.org/10.1109/IEMECONX.2019.8877011

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750-1758. https://doi.org/10.1016/j.csbj.2021.03.022

Ostrovsky-Berman, M., Frankel, B., Polak, P., & Yaari, G. (2021). Immune2vec: Embedding B/T Cell Receptor Sequences in ℝN Using Natural Language Processing. *Frontiers in Immunology*, *12*, 680687. https://doi.org/10.3389/fimmu.2021.680687

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*, *11*, 36120-36146. https://doi.org/10.1109/ACCESS.2023.3266377

Perel, E. (2017). *The state of affairs: Rethinking infedility* (First edition). Harper, an imprint of HarperCollinsPublishers.

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., & Deane, C. M. (2019). Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(10), 4025-4030. https://doi.org/10.1073/pnas.1810576116

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303-304. https://doi.org/10.1038/nbt0308-303

Rustam, Gunawan, A. Y., & Kresnowati, M. T. A. P. (2022). Data dimensionality reduction technique for clustering problem of metabolomics data. *Heliyon*, *8*(6), e09715. https://doi.org/10.1016/j.heliyon.2022.e09715

Sharma, N., & Saroha, K. (2015). Study of dimension reduction methodologies in data mining. *International Conference on Computing, Communication & Automation*, 133-137. https://doi.org/10.1109/CCAA.2015.7148359

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., … Chudakov, D. M. (2018). VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, *46*(D1), D419-D427. https://doi.org/10.1093/nar/gkx760

Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*(3), 405-420.

https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

Stafford, P., Wrapp, D., & Johnston, S. A. (2016). General Assessment of Humoral Activity in Healthy Humans. *Molecular & Cellular Proteomics*, *15*(5), 1610-1621. https://doi.org/10.1074/mcp.M115.054601

Valkiers, S., Van Houcke, M., Laukens, K., & Meysman, P. (2021). ClusTCR: A python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics*, *37*(24), 4865-4867. https://doi.org/10.1093/bioinformatics/btab446

van der Maaten, L., & Hinton, G. (2008). Viualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, *165*, 104-111. https://doi.org/10.1016/j.procs.2020.01.079

Wang, Y. (2019). Single Training Dimension Selection for Word Embedding with PCA. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3595-3600. https://doi.org/10.18653/v1/D19-1369

Weber, A., Pélissier, A., & Rodríguez Martínez, M. (2024). T-cell receptor binding prediction: A machine learning revolution. *ImmunoInformatics*, *15*, 100040. https://doi.org/10.1016/j.immuno.2024.100040

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265-269. https://doi.org/10.1038/s41586-020-2008-3

Xu, J. L., & Davis, M. M. (2000). Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity*, *13*(1), 37-45. https://doi.org/10.1016/S1074-7613(00)00006-6

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316. https://doi.org/10.1016/j.neucom.2020.07.061

Zaki, M. J., Wang, J. T. L., & Toivonen, H. T. T. (2003). Data mining in bioinformatics: Report on BIOKDD'03. *ACM SIGKDD Explorations Newsletter*, *5*(2), 198-199. https://doi.org/10.1145/980972.981006

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56-70. https://doi.org/10.38094/jastt1224

Anexos

Anexo A. Atributos de los conjuntos de datos de OAS

Los 98 atributos que describe cada uno de los registros en los conjuntos de datos de OAS se agrupan en 22 categorías, las cuales descritas en la tabla A.1, el agrupamiento se hizo con base en la similitud de contenido de los atributos, es decir, los atributos que contienen un dato similar, serán agrupadas en la misma categoría, por ejemplo, los indicadores de inicio y fin de las secuencias de alineamiento de los segmentos V, D y J se encuentran en la misma categoría, dado que estos 6 atributos contendrán siempre la posición del nucleótido que da inicio o con la que concluye la secuencia.

Tabla A.1 Categorías de los atributos en los conjuntos de datos de OAS

Categoría	**Atributos**	**Descripción**
Secuencia de anticuerpo	sequence	Secuencia de nucleótidos de consulta
Clasificación de locus	locus	Clasificación genérica del locus
Banderas de contenido en la secuencia	stop_codon, vj_in_frame, v_frameshift, productive, rev_comp, complete_vdj	Valores booleanos que indican diversas caracteristicas en la secuencia
V, D y J Call	v_call, d_call, j_call	Gen V, D y J con alelo
Alineamiento y línea germinal de secuencia	sequence_alignment, germline_alignment, sequence_alignment_aa, germline_alignment_aa	Parte alineada de la secuencia y la linea germinal, incluidas las correcciones indel o espaciadores de numeración, como los espacios IMGT
Inicio y fin del alineamiento de secuencia V, D y J	v_alignment_start, v_alignment_end, d_alignment_start, d_alignment_end, j_alignment_start, j_alignment_end	Posición inicial y final de la alineación del gen V, D y J
Secuencias de alineamiento y línea germinal V, D y J	v_sequence_alignment, v_sequence_alignment_aa, v_germline_alignment, v_germline_alignment_aa, d_sequence_alignment, d_sequence_alignment_aa, d_germline_alignment, d_germline_alignment_aa, j_sequence_alignment, j_sequence_alignment_aa, j_germline_alignment, j_germline_alignment_aa	Porción alineada de la secuencia de consulta asignada al gen V, D y J

Inicio y fin de las secuencias de alineamiento y línea germinal V, D y J	v_sequence_start, v_sequence_end, v_germline_start, v_germline_end, d_sequence_start, d_sequence_end, d_germline_start, d_germline_end, j_sequence_start, j_sequence_end, j_germline_start, j_germline_end	Posición inicial y final del gen V, D y J en la secuencia
framework en nucleótidos y aminoácidos	fwr1, fwr1_aa, fwr2, fwr2_aa, fwr3, fwr3_aa, fwr4, fwr4_aa	Secuencia de nucleótidos y aminoácidos de la región framework 1,2,3 y 4
Inicio y fin del framework	fwr1_start, fwr1_end, fwr2_start, fwr2_end, fwr3_start, fwr3_end, fwr4_start, fwr4_end	Posición inicial y final de framework 1,2,3 y 4 en la secuencia
CDRs en nucleótidos y aminoácidos	cdr1, cdr1_aa, cdr2, cdr2_aa, cdr3, cdr3_aa	Secuencia de nucleótidos y aminoácidos de la región CDR1, CDR2 y CDR3 alineada.
Inicio y fin de los CDRs	cdr1_start, cdr1_end, cdr2_start, cdr2_end, cdr3_start, cdr3_end	Posición inicial y final de CDR1, CDR2 y CDR3 en la secuencia
Junction	junction, junction_length, junction_aa, junction_aa_length	Secuencia de nucleótidos y aminoácidos de la región de unión
Puntuación V, D y J	v_score, d_score, j_score	Puntuación de alineación del gen V, D y J
Cadena CIGAR V, D y J	v_cigar, d_cigar, j_cigar	Cadena CIGAR para la alineación del gen V, D y J
Support V, D y J	v_support, d_support, j_support	Alineamiento del gen V, D y J Valor E, valor p, verosimilitud, probabilidad u otra medida similar de respaldo para la asignación del gen V, D y J
Identidad V, D y J	v_identity, d_identity, j_identity	Identidad fraccionaria para la alineación del gen V, D y J
NP y NP length	np1, np1_length, np2, np2_length	Secuencia de nucleótidos de la región N/P combinada entre el gen V y el primer alineamiento del gen D o entre los alineamientos del gen V y el gen J
Estado y numeración ANARCI	ANARCI_numbering, ANARCI_status	Residuos inusuales, falta de cisteínas conservadas, eliminaciones e inserciones fuera de las CDR, truncamiento de las estructuras 1 o 4 y si la CDR3 tiene más de 37 residuos

Redundancia	Redundancy	Número de veces que se encontró la secuencia dada en la ejecución
Región C	c_region	Nucleótidos que forman parte del dominio conservado

Anexo B. Ejemplo de dataset OAS

La tabla B.1 muestra de manera transpuesta tres registros extraídos de un conjunto de datos que se usó en este trabajo, el propósito de transponer la tabla original fue debido al espacio de la tabla original, dado que contiene 98 atributos y el contenido del mismo es extenso se decidió por esta representación. El atributo principal es *sequence* y del mismo se desprenden los demás, con el fin de describirlo. Se observa su secuencia de alineamiento, la línea germinal en nucleótidos y en aminoácidos, entre otros atributos que describen al inicial.

Tabla B.1 Ejemplos de registros de un conjunto de datos original de OAS

Atributo	**Registro 1**	**Registro 2**	**Registro 3**
sequence	CTTCGGAGACCCT GTCCCTCACCTGC GGTGTCTATGGTG GGTCCTTCAGTGG TTACTACTGGAGC TGGATCCGCCAGC CCCCAGGGAAGG GGCTGGAGTGGAT TGGGGAAATCAAT CAGAGTGGAAGT ACCAACTACAACC CGTCCCAGAAGAG TCGAGTCACCATA TCAGTAGACACGT CCAAGAACCAGTT CTCCCTGAAAGTG GACTCTGTGACCG CCGCGGACACGG GTGTTTATTACTG TGCGACTCCCCGC TATAGGGCGACCT ACTACCCCCTCGA CTTCTGGGGCCAG GGAACCCTGGTCA CCGTCTCCTCAGC CTCCACCAAGGGC CCATCGGTCTTCC CCCTGGCACCCTC CTCCAAGAGCACC TCTGGGGGGTCCC TGAGACTCTCCTT AAGTGAGCTAGCC TGAAGCGTGCTGG AC	CTGGGGGGTCCCT GAGACTCTCCTGT GCAGCCTCTGGGT TCAATTTCAGCGG CTCTGCTATCCAG TGGGTCCGCCAGC CTTCCGGGAAAGG GCTGGAGTGGATT GGCCGTATCAGAA GCAAGCCTAAAG GTTACGCGACAAC ATATGCTGCGTCC CTAAAAGGCAGAT TCGTTATCTCCAG AGATGATTCAAGG AACACGGCATATC TCCAGATACACAG CCTAAAAATCGAG GACATGGCCGTGT ATTATTGTACCGG AGACTATTTATAC TGGGACCAGGGA ACCCTGGTCTCCG TCTCCTCAGCCTC CACCAAGGGCCCA TCGGTCTTCCCCC TGGCACCCTCCTC CAAGAGCACCTCT GGGGGCACAGCG GCCCTGGGCTGCC TGGTCAAGGACTA CTTCACATTGAAT CGCAGCCTGAAGC GTGCTGGACA	CTTCGGAGACCCT GTCCCTCACCTGC ACTGTCTCCGGTG TCTCCATCAGTAT TTACTACTGGACC TGGATCCGGCAGC CCCCAGGGAAGG GACTGGAGTGGAT TGGTTATATCTAT TACAGTGGGAGCA CCACCTACAACCC TTCCCTCAAGAGT CGAGTCACCTTGT CAGCAGACACGTC CAAGAACCAGTTC TCCCTGAAGCTGA GTTCTGTGACTGC TGCGGACACGGCC GTCTATTATTGTG CGAGAGATGTTGG TATGGACGTCTGG GGCCAAGGGATC ACGGTCACCGTCT CCTCAGCCTCCAC CAAGGGCCCATCG GTCTTCCCCCTGG CACCCTCCTCCAA GAGCACCTCTGGG GGCACAGCGGCCC TGGGCTGCCTGGT CAAGGACTACTAA CAAACGGACAAA AGCCTGAAGCGTG CTGGACA
locus	H	H	H
stop_codon	F	F	F

vj_in_frame	T	T	T
v_frameshift	F	F	F
productive	T	T	T
rev_comp	F	T	T
complete_vdj	F	F	F
v_call	IGHV4-34*01	IGHV3-73*01	IGHV4-59*01
d_call	IGHD1-26*01	IGHD4-17*01	
j_call	IGHJ4*02	IGHJ4*02	IGHJ6*02
sequence_alignment	CTTCGGAGACCCTGTCCCTCACCTGCGGTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCAGAGTGGAAGTACCAACTACAACCCGTCCCAGAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAAGTGGACTCTGTGACCGCCGCGGACACGGGTGTTTATTACTGTGCGACTCCCCGCTATAGGGCGACCTACTACCCCCTCGACTTCTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	CTGGGGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGGTTCAATTTCAGCGGCTCTGCTATCCAGTGGGTCCGCCAGCCTTCCGGGAAAGGGCTGGAGTGGATTGGCCGTATCAGAAGCAAGCCTAAAGGTTACGCGACAACATATGCTGCGTCCCTAAAAGGCAGATTCGTTATCTCCAGAGATGATTCAAGGAACACGGCATATCTCCAGATACACAGCCTAAAAATCGAGGACATGGCCGTGTATTATTGTACCGGAGACTATTTATACTGGGACCAGGGAACCCTGGTCTCCGTCTCCTCAG	CTTCGGAGACCCTGTCCCTCACCTGCACTGTCTCCGGTGTCTCCATCAGTATTTACTACTGGACCTGGATCCGGCAGCCCCCAGGGAAGGGACTGGAGTGGATTGGTTATATCTATTACAGTGGGAGCACCACCTACAACCCTTCCCTCAAGAGTCGAGTCACCTTGTCAGCAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGTTCTGTGACTGCTGCGGACACGGCCGTCTATTATTGTGCGAGAGATGTTGGTATGGACGTCTGGGGCCAAGGGATCACGGTCACCGTCTCCTCA

germline_alignment	CTTCGGAGACCCTGTCCCTCACCTGCGCTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCCGCGGACACGGCTGTGTATTACTGTGCGANNNNNNNNTATAGTGGGAGCTACTACNNNNNNGACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	CTGGGGGGTCCCTGAAACTCTCCTGTGCAGCCTCTGGGTTCACCTTCAGTGGCTCTGCTATGCACTGGGTCCGCCAGGCTTCCGGGAAAGGGCTGGAGTGGGTTGGCCGTATTAGAAGCAAAGCTAACAGTTACGCGACAGCATATGCTGCGTCGGTGAAAGGCAGGTTCACCATCTCCAGAGATGATTCAAAGAACACGGCGTATCTGCAAATGAACAGCCTGAAAACCGAGGACACGGCCGTGTATTACTGTACCGGTGACTANNNNTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	CTTCGGAGACCCTGTCCCTCACCTGCACTGTCTCTGGTGGCTCCATCAGTAGTTACTACTGGAGCTGGATCCGGCAGCCCCCAGGGAAGGGACTGGAGTGGATTGGGTATATCTATTACAGTGGGAGCACCAACTACAACCCCTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCTGCGGACACGGCCGTGTATTACTGTGCGAGAGANNNNGGTATGGACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCA
sequence_alignment_aa	SETLSLTCGVYGGSFSGYYWSWIRQPPGKGLEWIGEINQSGSTNYNPSQKSRVTISVDTSKNQFSLKVDSVTAADTGVYYCATPRYRATYYPLDFWGQGTLVTVSS	GGSLRLSCAASGFNFSGSAIQWVRQPSGKGLEWIGRIRSKPKGYATTYAASLKGRFVISRDDSRNTAYLQIHSLKIEDMAVYYCTGDYLYWDQGTLVSVSS	SETLSLTCTVSGVSISIYYWTWIRQPPGKGLEWIGYIYYSGSTTYNPSLKSRVTLSADTSKNQFSLKLSSVTAADTAVYYCARDVGMDVWGQGITVTVSS
germline_alignment_aa	SETLSLTCAVYGGSFSGYYWSWIRQPPGKGLEWIGEINHSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAXXXYSGSYYXXDYWGQGTLVTVSS	GGSLKLSCAASGFTFSGSAMHWVRQASGKGLEWVGRIRSKANSYATAYAASVKGRFTISRDDSKNTAYLQMNSLKTEDTAVYYCTGDXXYWGQGTLVTVSS	SETLSLTCTVSGGSISSYYWSWIRQPPGKGLEWIGYIYYSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARXXGMDVWGQGTTVTVSS
v_alignment_start	1	1	1
v_alignment_end	249	256	253
d_alignment_start	258	257	
d_alignment_end	275	265	
j_alignment_start	282	270	258

j_alignment_end	321	306	302
v_sequence_alignment	CTTCGGAGACCCTGTCCCTCACCTGCGGTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCAGAGTGGAAGTACCAACTACAACCCGTCCCAGAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAAGTGGACTCTGTGACCGCCGCGGACACGGGTGTTTATTACTGTGC	CTGGGGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGGTTCAATTTCAGCGGCTCTGCTATCCAGTGGGTCCGCCAGCCTTCCGGGAAAGGGCTGGAGTGGATTGGCCGTATCAGAAGCAAGCCTAAAGGTTACGCGACAACATATGCTGCGTCCCTAAAAGGCAGATTCGTTATCTCCAGAGATGATTCAAGGAACACGGCATATCTCCAGATACACAGCCTAAAAATCGAGGACATGGCCGTGTA	CTTCGGAGACCCTGTCCCTCACCTGCACTGTCTCCGGTGTCTCCATCAGTATTTACTACTGGACCTGGATCCGGCAGCCCCCAGGGAAGGGACTGGAGTGGATTGGTTATATCTATTACAGTGGGAGCACCACCTACAACCCTTCCCTCAAGAGTCGAGTCACCTTGTCAGCAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGTTCTGTGACTGCTGCGGACACGGCCGTCTATTATTGTGCGA
v_sequence_alignment_aa	SETLSLTCGVYGGSFSGYYWSWIRQPPGKGLEWIGEINQSGSTNYNPSQKSRVTISVDTSKNQFSLKVDSVTAADTGVYYCA	GGSLRLSCAASGFNFSGSAIQWVRQPSGKGLEWIGRIRSKPKGYATTYAASLKGRFVISRDDSRNTAYLQIHSLKIEDMAVYYCT	SETLSLTCTVSGVSISIYYWTWIRQPPGKGLEWIGYIYYSGSTTYNPSLKSRVTLSADTSKNQFSLKLSSVTAADTAVYYCAR
v_germline_alignment	CTTCGGAGACCCTGTCCCTCACCTGCGCTGTCTATGGTGGGTCCTTCAGTGGTTACTACTGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAAATCAATCATAGTGGAAGCACCAACTACAACCCGTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCCGCGGACACGGCTGTGTATTACTGTGCGA	CTGGGGGGTCCCTGAAACTCTCCTGTGCAGCCTCTGGGTTCACCTTCAGTGGCTCTGCTATGCACTGGGTCCGCCAGGCTTCCGGGAAAGGGCTGGAGTGGGTTGGCCGTATTAGAAGCAAAGCTAACAGTTACGCGACAGCATATGCTGCGTCGGTGAAAGGCAGGTTCACCATCTCCAGAGATGATTCAAAGAACACGGCGTATCTGCAAATGAACAGCCTGAAAACCGAGGACACGGCCGTGTATTACTGTAC	CTTCGGAGACCCTGTCCCTCACCTGCACTGTCTCTGGTGGCTCCATCAGTAGTTACTACTGGAGCTGGATCCGGCAGCCCCCAGGGAAGGGACTGGAGTGGATTGGGTATATCTATTACAGTGGGAGCACCAACTACAACCCCTCCCTCAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGCTCTGTGACCGCTGCGGACACGGCCGTGTATTACTGTGCGAGAGA

v_germline_alignment_aa	SETLSLTCAVYGGSFSGYYWSWIRQPPGKGLEWIGEINHSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCA	GGSLKLSCAASGFTFSGSAMHWVRQASGKGLEWVGRIRSKANSYATAYAASVKGRFTISRDDSKNTAYLQMNSLKTEDTAVYYCT	SETLSLTCTVSGGSISSYYWSWIRQPPGKGLEWIGYIYYSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR
d_sequence_alignment	TATAGGGCGACCTACTAC	CGGAGACTA	
d_sequence_alignment_aa	YRATYY	GD	
d_germline_alignment	TATAGTGGGAGCTACTAC	CGGTGACTA	
d_germline_alignment_aa	YSGSYY	GD	
j_sequence_alignment	GACTTCTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	TACTGGGACCAGGGAACCCTGGTCTCCGTCTCCTCAG	GGTATGGACGTCTGGGGCCAAGGGATCACGGTCACCGTCTCCTCA
j_sequence_alignment_aa	DFWGQGTLVTVSS	YWDQGTLVSVSS	GMDVWGQGITVTVSS
j_germline_alignment	GACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	TACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG	GGTATGGACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCA
j_germline_alignment_aa	DYWGQGTLVTVSS	YWGQGTLVTVSS	GMDVWGQGTTVTVSS
fwr1	CTTCGGAGACCCTGTCCCTCACCTGCGGTGTCTAT	CTGGGGGGTCCCTGAGACTCTCCTGTGCAGCCTCT	CTTCGGAGACCCTGTCCCTCACCTGCACTGTCTCC
fwr1_aa	SETLSLTCGVY	GGSLRLSCAAS	SETLSLTCTVS
cdr1	GGTGGGTCCTTCAGTGGTTACTAC	GGGTTCAATTTCAGCGGCTCTGCT	GGTGTCTCCATCAGTATTTACTAC
cdr1_aa	GGSFSGYY	GFNFSGSA	GVSISIYY
fwr2	TGGAGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGGAA	ATCCAGTGGGTCCGCCAGCCTTCCGGGAAAGGGCTGGAGTGGATTGGCCGT	TGGACCTGGATCCGGCAGCCCCCAGGGAAGGGACTGGAGTGGATTGGTTAT
fwr2_aa	WSWIRQPPGKGLEWIGE	IQWVRQPSGKGLEWIGR	WTWIRQPPGKGLEWIGY
cdr2	ATCAATCAGAGTGGAAGTACC	ATCAGAAGCAAGCCTAAAGGTTACGCGACA	ATCTATTACAGTGGGAGCACC

cdr2_aa	INQSGST	IRSKPKGYAT	IYYSGST
fwr3	AACTACAACCCGTCCCAGAAGAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGAAAGTGGACTCTGTGACCGCCGCGGACACGGGTGTTTATTACTGT	ACATATGCTGCGTCCCTAAAAGGCAGATTCGTTATCTCCAGAGATGATTCAAGGAACACGGCATATCTCCAGATACACAGCCTAAAAATCGAGGACATGGCCGTGTATTATTGT	ACCTACAACCCTTCCCTCAAGAGTCGAGTCACCTTGTCAGCAGACACGTCCAAGAACCAGTTCTCCCTGAAGCTGAGTTCTGTGACTGCTGCGGACACGGCCGTCTATTATTGT
fwr3_aa	NYNPSQKSRVTISVDTSKNQFSLKVDSVTAADTGVYYC	TYAASLKGRFVISRDDSRNTAYLQIHSLKIEDMAVYYC	TYNPSLKSRVTLSADTSKNQFSLKLSSVTAADTAVYYC
fwr4	TGGGGCCAGGGAACCCTGGTCACCGTCTCCTCA	TGGGACCAGGGAACCCTGGTCTCCGTCTCCTCA	TGGGGCCAAGGGATCACGGTCACCGTCTCCTCA
fwr4_aa	WGQGTLVTVSS	WDQGTLVSVSS	WGQGITVTVSS
cdr3	GCGACTCCCCGCTATAGGGCGACCTACTACCCCCTCGACTTC	ACCGGAGACTATTTATAC	GCGAGAGATGTTGGTATGGACGTC
cdr3_aa	ATPRYRATYYPLDF	TGDYLY	ARDVGMDV
junction	TGTGCGACTCCCCGCTATAGGGCGACCTACTACCCCCTCGACTTCTGG	TGTACCGGAGACTATTTATACTGG	TGTGCGAGAGATGTTGGTATGGACGTCTGG
junction_length	48	24	30
junction_aa	CATPRYRATYYPLDFW	CTGDYLYW	CARDVGMDVW
junction_aa_length	16	8	10
v_score	355.528	304.11	352.411
d_score	17.992	12.223	
j_score	71.827	60.291	81.44
v_cigar	40N249M179S4N	40N256M180S6N	40N253M180S
d_cigar	257S2N18M153S	256S6N9M171S1N	
j_cigar	281S8N40M107S	269S11N37M130S	257S17N45M131S
v_support	4.64E-100	1.42E-84	4.08E-99
d_support	7.314	40.62	
j_support	9.89E-17	2.99E-13	1.28E-19

v_identity	95.582	87.891	94.466
d_identity	83.333	88.889	
j_identity	97.5	94.595	97.778
v_sequence_start	1	1	1
v_sequence_end	249	256	253
v_germline_start	41	41	41
v_germline_end	289	296	293
d_sequence_start	258	257	
d_sequence_end	275	265	
d_germline_start	3	7	
d_germline_end	20	15	
j_sequence_start	282	270	258
j_sequence_end	321	306	302
j_germline_start	9	12	18
j_germline_end	48	48	62
fwr1_start	1	1	1
fwr1_end	35	35	35
cdr1_start	36	36	36
cdr1_end	59	59	59
fwr2_start	60	60	60
fwr2_end	110	110	110
cdr2_start	111	111	111
cdr2_end	131	140	131
fwr3_start	132	141	132
fwr3_end	245	254	245
fwr4_start	288	273	270
fwr4_end	320	305	302
cdr3_start	246	255	246
cdr3_end	287	272	269
np1	CTCCCCGC		TGTT
np1_length	8	0	4

np2	CCCCTC	TTTA	
np2_length	6	4	
c_region	CCTCCACCAAGGG CCCATCGGTCTTC CCCCTGGCACCCT CCTCCAAGAGCAC CTCTGGGGGGTCC CTGAGACTCTCCT TAAGTGAGCTAGC CTGAAGCGTGCTG GAC	CCTCCACCAAGGG CCCATCGGTCTTC CCCCTGGCACCCT CCTCCAAGAGCAC CTCTGGGGGCACA GCGGCCCTGGGCT GCCTGGTCAAGGA CTACTTCACATTG AATCGCAGCCTGA AGCGTGCTGGACA	GCCTCCACCAAGG GCCCATCGGTCTT CCCCCTGGCACCC TCCTCCAAGAGCA CCTCTGGGGGCAC AGCGGCCCTGGGC TGCCTGGTCAAGG ACTACTAACAAAC GGACAAAAGCCT GAAGCGTGCTGGA
Redundancy	1	63	5

ANARCI_numbering	{'fwh1': {'16 ': 'S', '17 ': 'E', '18 ': 'T', '19 ': 'L', '20 ': 'S', '21 ': 'L', '22 ': 'T', '23 ': 'C', '24 ': 'G', '25 ': 'V', '26 ': 'Y'}, 'cdrh1': {'27 ': 'G', '28 ': 'G', '29 ': 'S', '30 ': 'F', '35 ': 'S', '36 ': 'G', '37 ': 'Y', '38 ': 'Y'}, 'fwh2': {'39 ': 'W', '40 ': 'S', '41 ': 'W', '42 ': 'I', '43 ': 'R', '44 ': 'Q', '45 ': 'P', '46 ': 'P', '47 ': 'G', '48 ': 'K', '49 ': 'G', '50 ': 'L', '51 ': 'E', '52 ': 'W', '53 ': 'I', '54 ': 'G', '55 ': 'E'}, 'cdrh2': {'56 ': 'I', '57 ': 'N', '58 ': 'Q', '59 ': 'S', '63 ': 'G', '64 ': 'S', '65 ': 'T'}, 'cdrh3': {'105 ': 'A', '106 ': 'T', '107 ': 'P', '108 ': 'R', '109 ': 'Y', '110 ': 'R', '111 ': 'A', '112A': 'T', '112 ': 'Y', '113 ': 'Y', '114 ': 'P', '115 ': 'L', '116 ': 'D', '117 ': 'F'}, 'fwh4': {'118 ': 'W', '119 ': 'G', '120 ': 'Q', '121 ': 'G', '122 ': 'T', '123 ': 'L', '124 ': 'V', '125 ': 'T', '126 ': 'V', '127 ': 'S', '128 ': 'S'}}	{'fwh1': {'16 ': 'G', '17 ': 'G', '18 ': 'S', '19 ': 'L', '20 ': 'R', '21 ': 'L', '22 ': 'S', '23 ': 'C', '24 ': 'A', '25 ': 'A', '26 ': 'S'}, 'cdrh1': {'27 ': 'G', '28 ': 'F', '29 ': 'N', '30 ': 'F', '35 ': 'S', '36 ': 'G', '37 ': 'S', '38 ': 'A'}, 'fwh2': {'39 ': 'I', '40 ': 'Q', '41 ': 'W', '42 ': 'V', '43 ': 'R', '44 ': 'Q', '45 ': 'P', '46 ': 'S', '47 ': 'G', '48 ': 'K', '49 ': 'G', '50 ': 'L', '51 ': 'E', '52 ': 'W', '53 ': 'I', '54 ': 'G', '55 ': 'R'}, 'cdrh2': {'56 ': 'I', '57 ': 'R', '58 ': 'S', '59 ': 'K', '60 ': 'P', '61 ': 'K', '62 ': 'G', '63 ': 'Y', '64 ': 'A', '65 ': 'T'}, 'cdrh3': {'105 ': 'T', '106 ': 'G', '107 ': 'D', '115 ': 'Y', '116 ': 'L', '117 ': 'Y'}, 'fwh4': {'118 ': 'W', '119 ': 'D', '120 ': 'Q', '121 ': 'G', '122 ': 'T', '123 ': 'L', '124 ': 'V', '125 ': 'S', '126 ': 'V', '127 ': 'S', '128 ': 'S'}}	{'fwh1': {'16 ': 'S', '17 ': 'E', '18 ': 'T', '19 ': 'L', '20 ': 'S', '21 ': 'L', '22 ': 'T', '23 ': 'C', '24 ': 'T', '25 ': 'V', '26 ': 'S'}, 'cdrh1': {'27 ': 'G', '28 ': 'V', '29 ': 'S', '30 ': 'I', '35 ': 'S', '36 ': 'I', '37 ': 'Y', '38 ': 'Y'}, 'fwh2': {'39 ': 'W', '40 ': 'T', '41 ': 'W', '42 ': 'I', '43 ': 'R', '44 ': 'Q', '45 ': 'P', '46 ': 'P', '47 ': 'G', '48 ': 'K', '49 ': 'G', '50 ': 'L', '51 ': 'E', '52 ': 'W', '53 ': 'I', '54 ': 'G', '55 ': 'Y'}, 'cdrh2': {'56 ': 'I', '57 ': 'Y', '58 ': 'Y', '59 ': 'S', '63 ': 'G', '64 ': 'S', '65 ': 'T'}, 'cdrh3': {'105 ': 'A', '106 ': 'R', '107 ': 'D', '108 ': 'V', '114 ': 'G', '115 ': 'M', '116 ': 'D', '117 ': 'V'}, 'fwh4': {'118 ': 'W', '119 ': 'G', '120 ': 'Q', '121 ': 'G', '122 ': 'I', '123 ': 'T', '124 ': 'V', '125 ': 'T', '126 ': 'V', '127 ': 'S', '128 ': 'S'}}
ANARCI_status	|Deletions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 73|||Shorter than IMGT defined: fw1|	|Deletions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 73|||Shorter than IMGT defined: fw1|	|Deletions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 73|||Shorter than IMGT defined: fw1|