Informe Final: Clasificación de Proteínas con Machine Learning

Ciencia e Ingeniería Computacional

Profesora: Carol Moraga.

Nombres:
Luis Miranda
Iván Bozo

Rancagua, May 2, 2024

Contents

1 | Introducción

1.1 | Hipótesis de Trabajo

Se ha planteado que el uso de modelos avanzados de aprendizaje automático puede permitir una clasificación precisa y eficiente de las familias de secuencias de proteínas, basándose únicamente en su secuencia de aminoácidos. Esta hipótesis se sustenta en la premisa de que las secuencias de proteínas, a pesar de su diversidad, exhiben patrones y similitudes estructurales y funcionales que pueden ser identificados y explotados por algoritmos de aprendizaje automático. Se argumenta que estas secuencias, al ser analizadas a través de técnicas de aprendizaje profundo y análisis de patrones, revelarán características distintivas que no solo permitirán clasificarlas en familias existentes, sino que también podrían descubrir nuevas agrupaciones o subfamilias basadas en características subyacentes no evidentes a simple vista. Además, se considera que esta aproximación podría revelar insights sobre la relación entre la secuencia de aminoácidos y la función biológica de las proteínas, contribuyendo así a una comprensión más profunda de los procesos biológicos.

1.2 | Objetivos

1.2.1 | Objetivo General

El objetivo general de esta investigación es desarrollar y validar un modelo de machine learning que sea capaz de clasificar con alta precisión y eficiencia las secuencias de proteínas en sus respectivas familias, utilizando exclusivamente la información contenida en su secuencia de aminoácidos. Este modelo buscará establecer un nuevo estándar en la clasificación de proteínas, proporcionando una herramienta valiosa para la investigación biomédica y la biotecnología.

1.2.2 | Objetivos Específicos

- Seleccionar los modelos de aprendizaje supervisado junto con técnicas de procesamiento de datos más adecuados para la tarea de clasificación de secuencias de proteínas.
- Desarrollar un pipeline de procesamiento de datos que incluya la limpieza, normalización y codificación adecuada de las secuencias de aminoácidos para su uso en modelos de aprendizaje automático.
- Implementar y entrenar varios modelos de aprendizaje supervisado, luego evaluando su eficacia en la clasificación de secuencias de proteínas en familias.
- Evaluar el rendimiento de los modelos utilizando métricas estándar como precisión, recall, accuracy y f1-score.
- Analizar los patrones y características aprendidos por el modelo para obtener insights sobre las relaciones estructurales y funcionales entre diferentes secuencias de proteínas.
- Documentar y comunicar los hallazgos, destacando tanto las implicaciones prácticas para la clasificación de proteínas como las posibles contribuciones teóricas al campo de la bioinformática y la biología molecular.

2 | Marco Teórico

2.1 | Materiales

En el contexto de la programación en Python, se emplean diversas bibliotecas especializadas que potencian el análisis de datos, la creación de modelos de aprendizaje automático y el desarrollo de aplicaciones científicas. Una de las ventajas principales de Python radica en su amplia gama de librerías de código abierto que facilitan la implementación de algoritmos complejos con una sintaxis simple y legible.

2.1.1 | Librerías a Utilizar

- **Pandas:** Proporciona estructuras de datos flexibles y herramientas para la manipulación eficiente de conjuntos de datos estructurados. Permite realizar operaciones de limpieza, filtrado y agregación de datos de manera sencilla y eficaz.

- **NumPy:** Destaca por su capacidad para trabajar con arreglos y matrices multidimensionales. Ofrece una amplia variedad de funciones matemáticas y herramientas para la manipulación de datos numéricos, lo que resulta crucial en la implementación de algoritmos de aprendizaje automático.
- **Scikit-learn:** Ofrece una colección integral de algoritmos para realizar tareas de clasificación, regresión, clustering y preprocesamiento de datos. Integrando herramientas como CountVectorizer para la conversión de texto a representaciones numéricas y train_test_split para dividir conjuntos de datos en entrenamiento y prueba.
- **Matplotlib y Seaborn:** Facilitan la visualización de datos y la evaluación de modelos. Permiten crear visualizaciones claras y expresivas, así como explorar datos y presentar resultados de manera efectiva.
- **MultinomialNB:** Utilizados para la implementación de modelos de Naive Bayes multinomial.
- **TensorFlow y Keras:** Proporcionan un entorno flexible para el desarrollo de modelos de redes neuronales artificiales. Permiten crear modelos secuenciales mediante capas densas y capas de dropout de manera eficiente.
- **LabelEncoder:** Facilita la codificación de etiquetas categóricas en forma numérica para su uso en modelos de aprendizaje automático. Resulta crucial en la preparación de datos para su procesamiento por parte de los algoritmos de aprendizaje.

2.1.2 | Métricas de Desempeño

1. **Precisión:**
   - **Definición:** Mide la proporción de instancias verdaderamente positivas entre las instancias que el modelo predijo como positivas.
   - **Fórmula:** $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
   - **Uso:** Evalúa la exactitud de las predicciones positivas.
2. **Recall (Sensibilidad o Tasa de Verdaderos Positivos):**
   - **Definición:** Mide la proporción de instancias verdaderamente positivas que el modelo identifica correctamente.
   - **Fórmula:** $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
   - **Uso:** Evalúa la capacidad del modelo para encontrar todas las instancias positivas.
3. **F1-score:**
   - **Definición:** Es la media armónica entre precision y recall, proporcionando un equilibrio entre ambas métricas.
   - **Fórmula:** $\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$
   - **Uso:** Resume la precisión y el recall en una sola métrica, ideal para comparar modelos en clasificaciones desbalanceadas.
4. **Accuracy (Exactitud):**
   - **Definición:** Mide la proporción de predicciones correctas sobre el total de predicciones.
   - **Fórmula:** $\text{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Samples}}$
   - **Uso:** Evalúa la tasa global de predicciones correctas.
5. **Matriz de Confusión:**
   - **Definición:** Una tabla que muestra el desempeño del modelo en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
   - **Uso:** Proporciona una visión detallada de las predicciones del modelo.

Estas métricas son fundamentales para evaluar y comparar modelos de clasificación. La elección de la métrica adecuada depende del contexto del problema y la importancia relativa de los errores de clasificación. En un problema desbalanceado, donde una clase es más relevante que otra, se pueden priorizar precision, recall o f1-score para esa clase en particular.
Comparar modelos implica analizar no solo la precisión general, sino también el rendimiento específico en clases individuales. Es posible que un modelo tenga una alta precisión global pero bajas métricas en clases críticas. Por eso, la comparación detallada de métricas por clase es esencial para elegir el modelo más apropiado para un caso particular.

2.2 | Base de Datos

Los datos utilizados en este estudio constituyen un conjunto de proteínas recuperadas del Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). El archivo del PDB es un repositorio de coordenadas atómicas y otra información que describe proteínas y otras macromoléculas biológicas importantes. Métodos como la cristalografía de rayos X, la espectroscopía de resonancia magnética nuclear (NMR) y la criomicroscopía electrónica se emplean para determinar la ubicación de cada átomo en relación con los demás en la molécula. El PDB, en constante crecimiento, refleja la investigación que se lleva a cabo en laboratorios de todo el mundo. Esto puede resultar tanto emocionante como desafiante al utilizar la base de datos en la investigación y educación. Se dispone de estructuras para muchas de las proteínas y ácidos nucleicos involucrados en los procesos centrales de la vida, permitiendo acceder a estructuras de ribosomas, oncogenes, objetivos de fármacos e incluso virus completos. Sin embargo, puede ser un desafío encontrar la información necesaria, dado que el PDB almacena una amplia variedad de estructuras. Es común encontrar múltiples estructuras para una molécula dada, o estructuras parciales, o estructuras modificadas o inactivadas con respecto a su forma nativa.

- **pdb_data_no_dups.csv:** Contiene metadatos de proteínas que incluyen detalles sobre la clasificación de la proteína, métodos de extracción, entre otros.
- **data_seq.csv:** Contiene 400,000 secuencias de estructuras de proteínas.

2.3 | Estado del Arte

El campo del análisis de estructuras de proteínas ha experimentado una evolución significativa a lo largo de los años. Inicialmente, los modelos de predicción se centraban en categorizar aminoácidos en estructuras helicoidales, empleando enfoques simplificados basados principalmente en las propiedades fisicoquímicas de los aminoácidos. Sin embargo, estas técnicas iniciales tenían limitaciones, particularmente en su capacidad para abordar la complejidad y la variabilidad de las estructuras proteicas.
Investigaciones recientes han destacado la importancia de las interacciones de la cadena principal en la determinación de la estructura secundaria de las proteínas, más que en las cadenas laterales. Esto ha llevado al desarrollo de modelos más sofisticados que utilizan métodos de regresión para predecir contactos entre residuos. A pesar de estos avances, las técnicas estadísticas tradicionales enfrentan limitaciones significativas. Estas incluyen la dependencia de un conjunto reducido de proteínas conocidas y la suposición de linealidad en las predicciones, lo que puede no ser adecuado para capturar la complejidad de las interacciones interresiduales.
En este contexto, el avance en el análisis de datos biológicos ha impulsado la evolución de algoritmos en Machine Learning. Métodos como minería de datos, inteligencia artificial y heurísticas complejas han emergido para gestionar volúmenes extensos de datos y mejorar la eficiencia en el análisis de datos biológicos. El aprendizaje automático, en particular, ha demostrado ser una herramienta poderosa en el análisis de estructuras de proteínas, ofreciendo la capacidad de aprender patrones complejos y no lineales a partir de grandes conjuntos de datos.
Estos avances indican un cambio significativo en la forma en que se abordan las predicciones de estructuras de proteínas, pasando de modelos estadísticos basados en suposiciones simplificadas a enfoques de aprendizaje automático que pueden manejar la complejidad y la heterogeneidad de los datos biológicos de manera más efectiva.

2.3.1 | Machine Learning

El Machine Learning es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones a partir de datos y realizar tareas sin una

programación explícita. En términos matemáticos, se trata de desarrollar sistemas capaces de aprender a partir de datos de entrenamiento, mejorando su desempeño con la experiencia.

2.3.2 | Redes Neuronales

Las redes neuronales artificiales (RNA) son aplicadas ampliamente con el objetivo de lograr predictores con un alto nivel de generalización y robustez para la predicción de estructuras de proteínas. Se han desarrollado distintos modelos como las redes feed-forward (ANN), modelos recurrentes (RNN), y aquellos que emplean funciones de base radial (RBFNN), cada uno con sus especificaciones y aplicaciones particulares. Estos modelos son capaces de aprender patrones complejos en los datos y pueden ser entrenados mediante diferentes técnicas, como el algoritmo back-propagation o el uso de algoritmos genéticos para optimizar los parámetros iniciales.
En el contexto de la predicción de aminoácidos y estructuras de proteínas, las RNA ofrecen un enfoque potente debido a su capacidad para modelar y aprender las complejas relaciones entre las secuencias de aminoácidos y sus conformaciones estructurales resultantes. Las RNA pueden captar dependencias sutiles y no lineales que son críticas para determinar la estructura secundaria y terciaria de las proteínas, algo que los métodos estadísticos tradicionales no siempre pueden lograr efectivamente.
Las arquitecturas de RNA pueden variar en complejidad, desde modelos simples con una única neurona de salida para predicciones binarias, hasta redes más complejas con múltiples neuronas de salida para representar estructuras proteicas en 2D o 3D. Además, se han propuesto esquemas de combinación de redes neuronales por niveles, donde una RNA en un primer nivel se encarga de la predicción secuencia-estructura para determinar la estructura secundaria, y otra RNA en un nivel posterior realiza predicciones más detalladas.

2.3.3 | Árboles de Decisión

Los Árboles de Decisión son estructuras de árbol que representan reglas de decisión. Se utilizan para clasificación y regresión mediante la partición recursiva del conjunto de datos en subconjuntos más pequeños, basándose en las características más informativas. Cada nodo interno representa una característica, cada rama representa una regla de decisión y cada hoja representa una salida. La construcción de estos árboles busca maximizar la información ganada en cada partición y se puede realizar mediante algoritmos como ID3, C4.5 o CART.

2.3.4 | Naive Bayes

Naive Bayes es un método probabilístico basado en el teorema de Bayes y asume independencia condicional entre las características. Se utiliza para clasificación y predicción, calculando la probabilidad de que una instancia pertenezca a una clase particular dado un conjunto de características. Este método asigna una clase a una instancia basándose en la probabilidad a posteriori, que se calcula mediante la probabilidad a priori y la verosimilitud condicional de las características dadas las clases.

- Distribución Gaussiana Multinomial La Distribución Gaussiana Multinomial es una extensión de la distribución gaussiana a múltiples dimensiones, lo que la hace especialmente útil en el contexto del aprendizaje automático para modelar conjuntos de datos con múltiples características. En la predicción de estructuras de proteínas, este enfoque puede ser empleado para analizar la distribución de diferentes tipos de aminoácidos y sus propiedades, o para modelar la relación entre diferentes características de las proteínas y su conformación estructural.

2.4 | Retos en la Predicción de Estructura de Proteínas

2.4.1 | Complejidad de las Proteínas

Las proteínas son extremadamente diversas en su estructura y función. Esta variabilidad representa un desafío significativo para predecir su conformación tridimensional a partir de su secuencia de aminoácidos. Cada proteína puede adoptar diferentes estructuras según el entorno y las interacciones con otras moléculas, lo que añade un nivel adicional de complejidad a la predicción. Además, las proteínas no son estructuras estáticas; su función depende de su capacidad para cambiar de forma y interactuar dinámicamente con otras moléculas. A pesar de los avances en técnicas experimentales, aún existe una carencia de datos estructurales detallados para muchas proteínas, lo que limita el entrenamiento y la validación de modelos de aprendizaje automático.

2.4.2 | Importancia en la Investigación Biomédica

La predicción precisa de estructuras de proteínas es crucial para el diseño de nuevos medicamentos, ya que permite identificar potenciales sitios de unión para fármacos. Además, muchas enfermedades están asociadas con el mal funcionamiento de proteínas específicas. Comprender su estructura es fundamental para desarrollar terapias dirigidas.

2.5 | Ventajas y Limitaciones

2.5.1 | Ventajas

El aprendizaje automático ofrece varias ventajas significativas en la predicción de estructuras de proteínas:

- **Análisis de Grandes Volúmenes de Datos**: Los algoritmos de aprendizaje automático pueden procesar y analizar grandes conjuntos de datos estructurales y secuenciales, una tarea que sería prácticamente imposible de realizar manualmente. Esto es crucial en la bioinformática, donde los datos son abundantes y complejos.
- **Descubrimiento de Nuevas Relaciones**: El aprendizaje automático puede identificar patrones y relaciones no evidentes en los datos, lo que puede llevar al descubrimiento de nuevos insights sobre las estructuras de las proteínas y sus funciones.
- **Eficiencia de Costos y Tiempo**: Comparado con los métodos experimentales tradicionales, como la cristalografía de rayos X y la resonancia magnética nuclear, el aprendizaje automático es mucho más rápido y menos costoso para predecir estructuras de proteínas.
- **Mejora Continua**: Los modelos de aprendizaje automático pueden mejorar continuamente su desempeño a medida que se les alimenta con más datos, lo que significa que su precisión y eficacia pueden aumentar con el tiempo.

2.5.2 | Limitaciones

A pesar de sus ventajas, el aprendizaje automático también enfrenta varias limitaciones en la predicción de estructuras de proteínas:

- **Dependencia de Datos de Calidad**: La precisión de los modelos de aprendizaje automático depende en gran medida de la calidad y cantidad de los datos de entrenamiento. Datos insuficientes o de baja calidad pueden llevar a predicciones inexactas.
- **Interpretación de Modelos**: Algunos modelos complejos, especialmente las redes neuronales profundas, actúan como ”cajas negras” y pueden ser difíciles de interpretar en términos de cómo llegan a sus predicciones, lo que plantea desafíos en la explicación de los resultados.
- **Generalización**: Los modelos pueden tener dificultades para generalizar a nuevas proteínas que tienen características significativamente diferentes de aquellas en el conjunto de datos de entrenamiento.
- **Sobreajuste**: Existe el riesgo de sobreajuste, especialmente en situaciones donde la cantidad de datos es limitada en comparación con la complejidad del modelo, lo que puede llevar a modelos que funcionan bien en datos de entrenamiento pero mal en datos no vistos.

3 | Metodología

La metodología empleada para este estudio se dividió en varias etapas, cada una enfocada en aspectos específicos del análisis y modelado de los datos.

3.1 | Plan de Trabajo

3.1.1 | Filtrado y Pre Procesamiento de los Datos

Inicialmente, se cargaron los datos desde la fuente correspondiente utilizando pandas, importando las bibliotecas necesarias para el análisis, entre ellas: pandas, numpy, sklearn, matplotlib, seaborn y

tensorflow.keras. Posteriormente, se procedió a unir dos conjuntos de datos relevantes para el estudio utilizando la información de estructura relevante. Se identificó la necesidad de trabajar exclusivamente con proteínas, por lo que se realizó un filtrado de datos para seleccionar únicamente las observaciones correspondientes a este tipo de macromoléculas, descartando otras categorías. Además, se procedió con la limpieza de registros nulos para asegurar la calidad de los datos utilizados en el análisis.

3.1.2 | Análisis Exploratorio de los Datos

Para comprender mejor la distribución de clases dentro del conjunto de datos, se realizó un análisis exploratorio inicial. Se calcularon las frecuencias de las distintas clases de proteínas presentes en el conjunto de datos. Dada la gran cantidad de clases con datos insuficientes, se filtró el conjunto de datos, conservando únicamente aquellas clases con un número mínimo de registros (> 1000), con el objetivo de mejorar la capacidad predictiva de los modelos de clasificación.

3.1.3 | Obtención de Conjuntos de Entrenamiento y Prueba

Después del proceso de filtrado, se procedió a la división de los datos para la creación de conjuntos de entrenamiento y prueba. Este paso fue crucial para el entrenamiento y la validación de los modelos. Para la extracción de características relevantes de las secuencias de proteínas, se empleó CountVectorizer, con un ngram_range de (4,4), permitiendo así la generación de características basadas en secuencias de aminoácidos.

3.1.4 | Modelos de Machine Learning

Se implementaron diferentes modelos de aprendizaje automático para la clasificación de proteínas:

1. **Redes Neuronales:**
   - Configuración de una red neuronal con capas densas y capas de desactivación.
   - Uso de la función de activación 'relu'.
   - Compilación del modelo con el optimizador 'adam' y la función de pérdida.
   - Entrenamiento del modelo con el conjunto de datos de entrenamiento.
   - Evaluación del desempeño con los datos de prueba.
2. **Naive Bayes:**
   - Creación de un modelo de clasificación Naive Bayes con la distribución Multinomial.
   - Entrenamiento y evaluación del modelo con los conjuntos de datos de entrenamiento y prueba, respectivamente.

3.1.5 | Análisis de Resultados

Se generaron y visualizaron las matrices de confusión para cada uno de los modelos implementados, utilizando Seaborn para representar gráficamente la efectividad de la clasificación realizada por cada modelo.

3.2 | Carta Gantt

La carta gantt asociada al trabajo de investigación, está basada en el plan de trabajo propuesto en la sección anterior, por lo tanto, es un recurso visual que muestra como se desarrolló el proceso para llevar a cabo este proyecto de investigación. Se presenta a continuación:

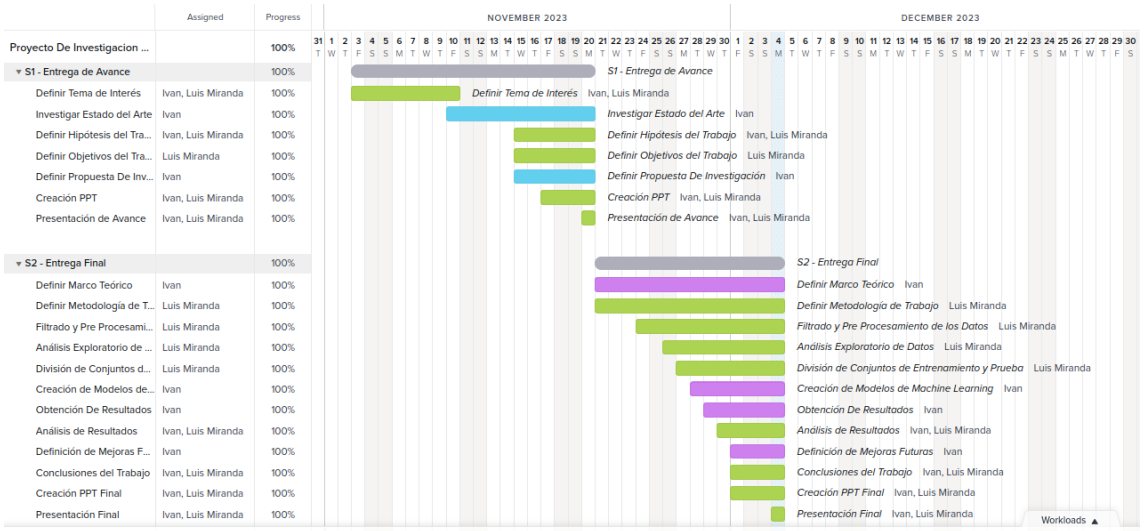

Figure 3.1: Carta Gantt del Plan De Trabajo

- S1 - Entrega de Avance (100%)
  - Definir Tema de Interés (100%)
  - Investigar Estado del Arte (100%)
  - Definir Hipótesis del Trabajo (100%)
  - Definir Objetivos del Trabajo (100%)
  - Definir Propuesta de Investigación (100%)
  - Creación PPT (100%)
  - Presentación de Avance (100%)
- S2 - Entrega Final (100%)
  - Definir Marco Teórico (100%)
  - Definir Metodología de Trabajo (100%)
  - Filtrado y Pre Procesamiento de los Datos (100%)
  - Análisis Exploratorio de Datos (100%)
  - División de Conjuntos de Entrenamiento y Prueba (100%)
  - Creación de Modelos de Machine Learning (100%)
  - Obtención de Resultados (100%)
  - Análisis de Resultados (100%)
  - Definición de Mejoras Futuras (100%)
  - Conclusiones del Trabajo (100%)
  - Creación PPT Final (100%)
  - Presentación Final (100%)

4 | Resultados de Los Modelos

4.1 | Modelo de Red Neuronal

Este modelo de clasificador obtuvo los siguientes resultados:

- **Reportes de Clasificación:**

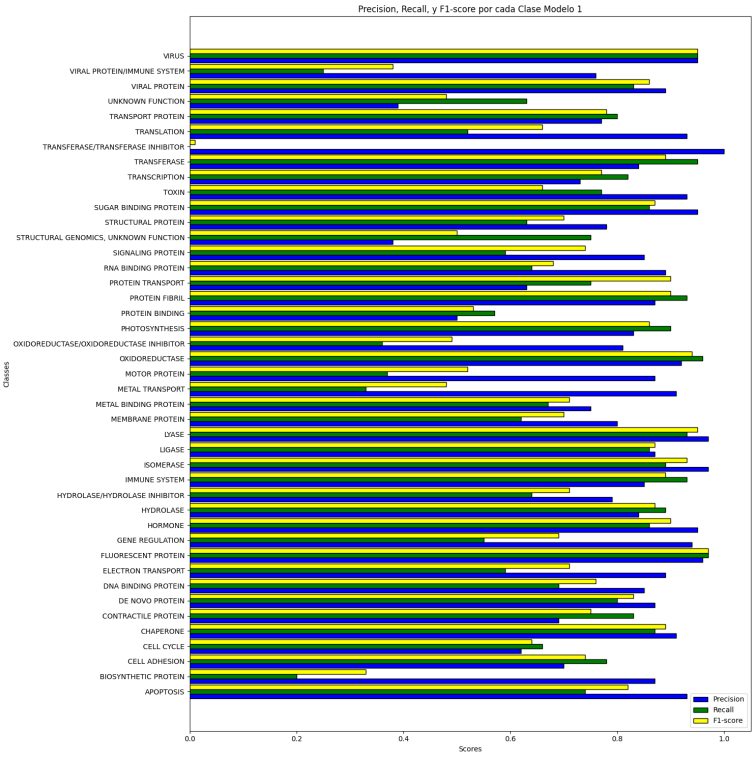

Figure 4.1: Reporte de Clasificación Modelo 1

- **Accuracy:** 83%.
- **Matriz de Confusión:**

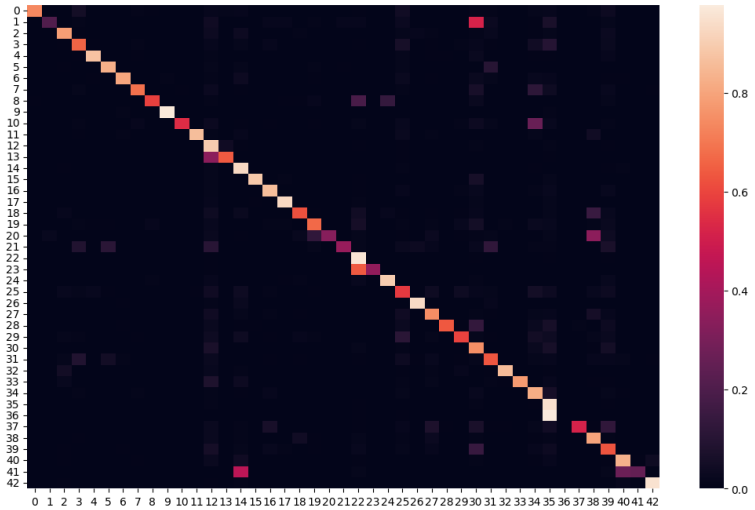

Figure 4.2: Matriz de Confusión Modelo 1

4.2 | Modelo de MultinomialNB

Este modelo clasificador obtuvo los siguientes resultados:

- **Reportes de Clasificación:**

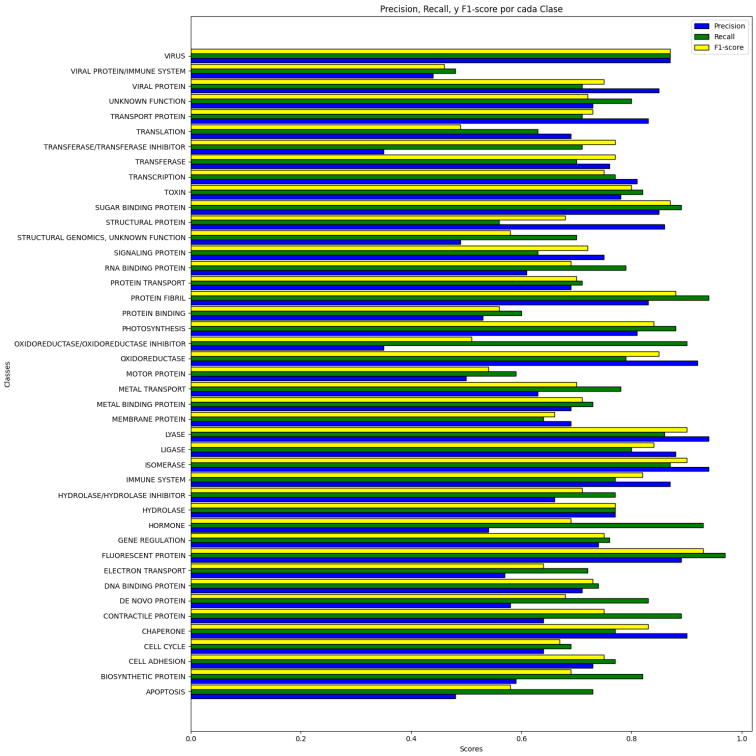

Figure 4.3: Reporte de Clasificación Modelo 2

- **Accuracy:** 76%.
- **Matriz de Confusión:**

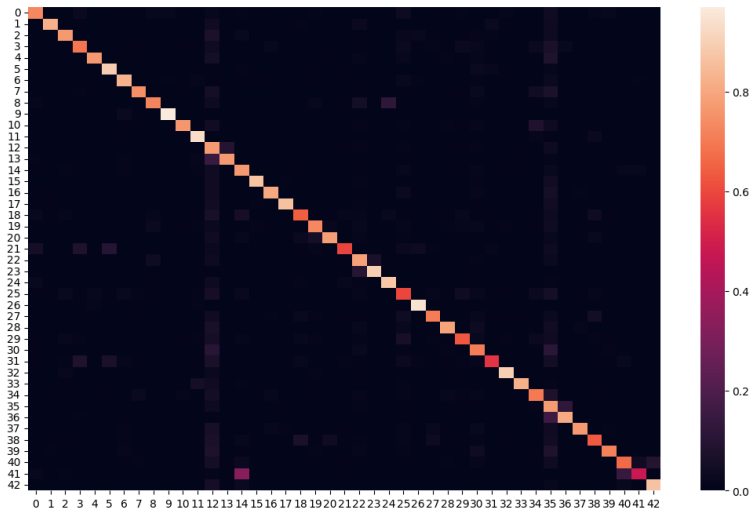

Figure 4.4: Matriz de Confusión Modelo 2

5 | Análisis y Trabajo Futuro

5.1 | Reportes de Clasificación y Accuracy

El modelo de Red Neuronal tiene una precisión total (accuracy) del 83%, mientras que el modelo de Multinomial NB alcanza una precisión total del 76%. Basándonos únicamente en la precisión total, el modelo de Red Neuronal parece tener un desempeño superior en comparación con el modelo de Multinomial NB.
Sin embargo, para una comparación más detallada y completa, es necesario evaluar otros aspectos además de la precisión, como la precisión, recall y f1-score para cada clase. Estos valores brindan una comprensión más profunda del rendimiento del modelo en cada categoría.
En términos generales, el modelo de Red Neuronal parece tener un mejor rendimiento basándose en como lucen los gráficos de cada reporte de clasificación, aunque es un poco más difícil entrar en el detalle de analizar clase por clase, por la cantidad de distintos tipos de proteínas obtenidas en esta investigación.

5.2 | Matrices de Confusión

Según esta métrica de desempeño, también se hace bastante difícil dilucidar cuál modelo es mejor que el otro, puesto que los dos modelos tienen pro y contra en cuanto a como lucen los mapas de calor de sus matrices de confusión.
Para el modelo 1 (Red Neuronal), se puede observar una clara línea diagonal, que representa las clases efectivamente clasificadas como Verdaderos Positivos, pero hay muchas clases que están siendo clasificadas de forma errónea, entre las más notorias serían la clase 30 clasificada como 1 y la clase 14 clasificada como 41.
Para el modelo 2 (MultinomialNB), se puede observar una línea diagonal bastante uniforme, lo que muestra una buena clasificación de clases; sin embargo, es mucho más notorio en esta matriz de confusión, que las clases 12 y 35 están siendo constantemente clasificadas de forma errónea, al observarse dos líneas rectas verticales en el gráfico.

5.3 | Posibles fuentes de Error

1. **Complejidad del Modelo:** La red neuronal (Modelo 1) puede ser más compleja y flexible en la captura de patrones complejos en los datos. Sin embargo, esta flexibilidad puede conducir a una mayor propensión a sobreajustar los datos y, por ende, a una clasificación incorrecta de algunas clases.

   El modelo MultinomialNB (Modelo 2), al ser un modelo más simple, tiende a tener un desempeño más uniforme en la clasificación, pero puede no capturar la sutileza de las relaciones complejas entre las características, lo que lleva a errores sistemáticos en ciertas clases.

2. **Desbalance de Datos:** La distribución desigual de ejemplos entre clases puede influir en el desempeño de los modelos. Clases con pocos ejemplos pueden ser más difíciles de clasificar correctamente, lo que lleva a errores específicos para esas clases.

3. **Selección de Características:** La selección o extracción de características puede afectar la capacidad de los modelos para generalizar. Si algunas características importantes no están siendo consideradas por los modelos, podrían surgir errores de clasificación en ciertas clases.

4. **Hiperparámetros del Modelo:** Los hiperparámetros, como la tasa de aprendizaje en la red neuronal o la suavización en el modelo MultinomialNB, podrían no estar optimizados, lo que afecta la capacidad de los modelos para aprender y generalizar adecuadamente.

5. **Sensibilidad a ciertos Patrones:** Los modelos pueden ser más o menos sensibles a ciertos tipos de patrones. Por ejemplo, uno puede ser mejor en la detección de características específicas en una clase, mientras que el otro modelo podría desempeñarse mejor en otra.

5.4 | Trabajo Futuro

El modelo actual enfrenta desafíos debido a la complejidad inherente y la similitud entre proteínas, especialmente en clases donde las proteínas comparten características análogas o dominios funcionales. Por ejemplo, las proteínas dentro de la misma categoría funcional, como enzimas o reguladores, a menudo

muestran similitudes estructurales y funcionales, lo que conduce a clasificaciones erróneas observadas en la matriz de confusión y el mapa de calor.

5.4.1 | Posibles Mejoras:

1. **Ampliación del Espacio de Características:** La incorporación de factores como el pH, el peso molecular y otros componentes relevantes más allá del conjunto actual de características de cuatro aminoácidos podría mejorar la capacidad del modelo para discernir diferencias intrincadas entre proteínas. Estas características adicionales podrían ofrecer información más profunda sobre agrupaciones familiares y categorías funcionales distintas, mejorando potencialmente la precisión de la clasificación.

2. **Aprovechamiento de la Información de Aminoácidos:** Explorar el uso de una gama más amplia de aminoácidos podría potenciar las capacidades predictivas del modelo. Extender el rango de ngramas más allá de cuatro caracteres facilitaría capturar interacciones más intrincadas entre aminoácidos, reflejando las complejidades inherentes en las estructuras y funcionalidades de las proteínas presentes en la realidad.

3. **Aumento de la Complejidad del Modelo:** Evaluar la posibilidad de utilizar modelos más sofisticados u optimizar los existentes podría abordar potencialmente las similitudes matizadas entre proteínas. Los modelos que pueden adaptarse a las relaciones intrincadas y características compartidas dentro de familias de proteínas podrían mejorar la precisión de la clasificación.

Al integrar características adicionales, ampliar el espacio de características y explorar técnicas de modelado más complejas, el objetivo es mejorar el poder discriminatorio del modelo y permitirle diferenciar mejor entre proteínas estrechamente relacionadas, elevando en última instancia su precisión y rendimiento en la clasificación de proteínas en diversas categorías funcionales.

6 | Conclusiones

El estudio buscaba validar la hipótesis de que mediante el uso de modelos avanzados de aprendizaje automático basados únicamente en secuencias de aminoácidos, sería posible clasificar con precisión las familias de secuencias de proteínas. A pesar de la diversidad de estas secuencias, se consideraba que patrones y similitudes estructurales y funcionales podían ser identificados por algoritmos de aprendizaje automático.
El modelo de Red Neuronal logró una precisión del 83%, superando al modelo MultinomialNB con un 76% en términos de accuracy. Aunque esta métrica podría sugerir una ventaja del modelo de Red Neuronal, un análisis más detallado reveló desafíos complejos al considerar precision, recall y f1-score para cada clase. Ambos modelos mostraron dificultades en la clasificación de clases específicas. El modelo de Red Neuronal presentó errores notorios en ciertas clases, mientras que el modelo MultinomialNB mostró patrones consistentes de clasificación errónea en otras clases.
Las posibles fuentes de error incluyen la complejidad de los modelos, el desbalance de datos y la limitación en la selección de características. Además, la restricción del espacio de características a solo cuatro aminoácidos fue una limitación evidente.
Para futuras investigaciones, se propone ampliar el espacio de características, incluyendo factores como pH y peso molecular, y explorar una gama más amplia de aminoácidos. Además, se considera evaluar modelos más sofisticados para mejorar la capacidad de clasificación.
Estos hallazgos sugieren que, aunque se logró una precisión considerable, el desafío de clasificar proteínas basándose únicamente en secuencias de aminoácidos requiere un enfoque más sofisticado e inclusivo para mejorar la precisión y su aplicabilidad en áreas como la investigación biomédica y la biotecnología.

Referencias

1. Wu, S., & Zhang, Y. (2009). Chapter 11. Protein Structure Prediction. In *Bioinformatics: Tools and Applications* (D. Edwards, Ed.). Springer Science+Business Media, LLC, pp. 225–242.
2. Walsh, I., Bau, D., Martin, A. J. M., Mooney, C., Vullo, A., & Pollastri, G. (2009). Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. *BMC Structural Biology, 9*(5), 1–38.
3. Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *J. Mol. Biol., 120*, 97–120.
4. Qian, N., & Sejnowski, T. J. (1988). Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol., 202*, 865–884.
5. Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Biophysics, 86*(January), 152–156.
6. Marrero-ponce, Y., Casañola-martín, G. M., Tareq, M., Khan, H., Rescigno, A., & Abad, C. (2010). Ligand-Based Computer-Aided Discovery of Tyrosinase Inhibitors. Applications of the TOMOCOMD-CARDD Method to the Elucidation of New Compounds. *Current Pharmaceutical Design, 16*, 2601–2624.
7. Armañanzas, R., Inza, I., Santana, R., Saeys, Y., Flores, J. L., Lozano, J. A., ... Larrañaga, P. (2008). A review of estimation of distribution algorithms in bioinformatics. *BioData Mining, 1*(6), 1–12.
8. Mitra, S. (2006). Bioinformatics With Soft Computing. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, 36*(5), 616–635.
9. Ouzounis, C. A., & Valencia, A. (2003). Early bioinformatics: the birth of a discipline — a personal view. *Bioinformatics, 19*(17), 2176–2190.
10. Zhang, Z. (2002). An Overview of Protein Structure Prediction: From Homology to Ab Initio. In *Bioc218*, pp. 1–10.