# Clasificación de péptidos antimicrobianos con funcionalidades específicas

Integrantes:
Santiago Rivera Montoya (santiago.riveram@udea.edu.co)
Emanuel López Higuita (emanuel.lopezh@udea.edu.co)
Tutor:
Carlos Andrés Mera (carlos.mera@udea.edu.co)

Proyecto Integrador I - 2508700 – G17
Repositorio

**Resumen.** Este informe detalla el desarrollo de un proyecto centrado en la clasificación de Péptidos Antimicrobianos (AMPs) según sus funciones específicas utilizando técnicas de Aprendizaje Automático. Para ello, se utilizaron dos bases de datos (UNAL y Starpep), que fueron integradas y depuradas mediante un exhaustivo preprocesamiento que incluyó limpieza, eliminación de duplicados, normalización y selección de características fisicoquímicas, como composición de aminoácidos y propiedades moleculares. El análisis exploratorio reveló patrones significativos en la composición de aminoácidos y las propiedades fisicoquímicas, destacando la importancia de leucina y lisina en todas las funcionalidades. Para la clasificación, se diseñó un modelo jerárquico en tres fases, evaluando algoritmos como Random Forest, Logistic Regression, LightGBM, XGBoost y ExtraTrees. Los resultados evidenciaron un excelente rendimiento en la identificación general de capacidad antimicrobiana (F1-score: 0.93 con Random Forest), pero un desempeño limitado en las clases minoritarias, como anticancerígena y antitumoral (F1-score: 0.49 con LightGBM), en la tercera fase, enfocada en clasificar los AMPs según funciones específicas (F1-score: 0.78 con LightGBM), destacando en categorías como antifúngicos y antibacterianos gracias a una mayor representatividad en los datos, además que los puntajes se mantuvieron constantes incluso con datos que los modelos no habían visto. Sin embargo, se observó que las funcionalidades menos representadas continuaron afectando el rendimiento del modelo, lo que resalta la necesidad de bases de datos más equilibradas para mejorar el desempeño en estas categorías específicas. Esto último subraya los retos de manejar conjuntos de datos desbalanceados en problemas de clasificación multietiqueta. Aunque se implementaron estrategias para abordar el desbalance, como sobremuestreo y reducción de etiquetas, estas no lograron mejorar significativamente los resultados. No obstante, el modelo jerárquico demostró ser prometedor, ofreciendo una base sólida para futuras investigaciones en la clasificación de

AMPs y destacando el potencial del Aprendizaje Automático para impulsar el descubrimiento de nuevos péptidos funcionales.

**Palabras clave:** Péptidos antimicrobianos, clasificación multietiqueta, aprendizaje supervisado.

## INTRODUCCIÓN

En la literatura actual, se han desarrollado diversos clasificadores para determinar si un péptido posee capacidad antimicrobiana, e incluso se han propuesto métodos para su generación artificial mediante técnicas de inteligencia artificial. Sin embargo, son pocos los trabajos que se han enfocado en clasificar los Péptidos Antimicrobianos (AMPs) según sus funciones específicas, utilizando las bases de datos ya existentes creadas por expertos.

Este proyecto tiene como objetivo abordar esta necesidad, tomando dichas bases de datos y aplicando técnicas de Aprendizaje Automático para clasificar los AMPs según sus funciones específicas, tales como anticancerígenos, antifúngicos, antivirales, entre otros. En este informe, se presenta el análisis de datos realizado y la extracción de características, los cuales sientan las bases para la construcción de modelos de clasificación que permitirán identificar las funciones específicas de estos péptidos.

## DESCRIPCIÓN DE LOS DATOS

Se cuenta con dos bases de datos, una proveniente de la Universidad Nacional que cuenta con 31659 registros, de los cuales, 18209 son no AMP y 13448 AMP y 20 variables o columnas y la otra la base de datos de Starpep que cuenta con 45119 registros y diversas funcionalidades antimicrobianas, que en particular solo 23156 cuentan con las funcionalidades antimicrobianas necesarias para este modelo.

## LIMPIEZA Y PREPARACIÓN DE LOS DATOS

El proceso comienza con el preprocesamiento de los conjuntos de datos. En la base de datos de la UNAL, se eliminaron las características fisicoquímicas, conservando únicamente las variables objetivo (anticáncer, antifúngico, antiviral, entre otras) y los identificadores de cada péptido.

En el caso de la base de datos de Starpep, se extrajo la metadata y se agrupó utilizando variables dummy, lo que permitió conservar solo las variables objetivo y el identificador de cada péptido.

El segundo paso implica filtrar péptidos según su capacidad antimicrobiana, solo se dejan aquellos con las funcionalidades: antifúngico, antiviral, antibacteriano, anti-Gram positivo, anti-Gram negativo, anticancerígeno, anti-VIH, antiparasitario y antitumoral. Después, se verifica la existencia de secuencias idénticas en ambas bases de datos. En caso de encontrar péptidos duplicados, se eliminan para evitar duplicidades al combinar las bases de datos.

El tercer paso consiste en el renombramiento de las columnas en ambas bases de datos por separado, facilitando su posterior combinación.

En el cuarto paso, tras la refactorización de los nombres y el orden de las columnas se combinaron ambas bases de datos, resultando en una base de datos exclusivamente de péptidos AMP y otra que incluye tanto péptidos con capacidad

antimicrobiana (que pueden inhibir el crecimiento o eliminar microorganismos, bacterias, hongos, virus, etc.), como quienes no la poseen.

Luego, se realizó un ajuste en ambas bases de datos para asegurar la coherencia en la clasificación de los péptidos. Si un péptido no tenía un "1" en la columna "antimicrobiano", pero sí lo tenía en alguna de las columnas relacionadas con capacidad antimicrobiana (antifúngico, antiviral, antibacteriano, anti-Gram positivo, anti-Gram negativo, anti-VIH, antiparasitario), se asignaba un "1" en la columna "antimicrobiano".

Posteriormente, ambos conjuntos de datos fueron filtrados para eliminar péptidos con secuencias menores a 7 aminoácidos o mayores a 100. Como resultado, no se encontraron secuencias mayores a 100 aminoácidos; sin embargo, se eliminaron 498 péptidos con menos de 7 aminoácidos. También se descartaron péptidos que contenían aminoácidos no naturales y aquellos cuya secuencia tenía una variación menor a tres tipos de aminoácidos. Estas modificaciones se realizaron en preparación para la futura extracción de características.

Finalmente, se exportan los archivos en formato .fasta para llevar a cabo la extracción de características fisicoquímicas. Estas características se obtienen mediante las librerías propy3 en Python3 para extraer descriptores como la composición de aminoácidos (AAC), composición de dipéptidos (DPC), autocorrelación de Moreau-Broto normalizada (MBauto), autocorrelación de Moran (Moranauto), autocorrelación de Geary (Gearyauto), composición, transición y distribución (CTD), números de acoplamiento del orden de secuencia (SOCN), cuasi-orden de secuencia (QSO) y composición de pseudo- aminoácidos (PAAC)., y la librería Peptides en R para calcular la longitud, índice de Boman, carga, punto isoeléctrico e hidrofobicidad.

Al concluir este proceso, se obtienen dos dataframes finales:

Df_final_AMP: 22,192 filas x 1507 columnas

Df_final: 40,840 filas x 1507 columnas

Al hacer un conteo de los péptidos AMP según sus funcionalidades específicas se obtuvieron los siguientes resultados.

**Tabla 1. Cantidad de péptidos por actividad antimicrobiana**

| Funcionalidad | Cantidad |
|---|---|
| Antimicrobiano | 21709 |
| Antifúngico | 5763 |
| Antiviral | 4134 |
| Anti-Gram + | 8078 |
| Anti-Gram - | 8141 |
| Anticancerígeno | 1555 |
| Antibacteriano | 8151 |
| Anti-VIH | 953 |
| Antiparasitario | 425 |
| Antitumoral | 550 |

Este proceso de preprocesamiento y extracción de características fisicoquímicas ha permitido generar dos conjuntos de datos depurados y estructurados que son fundamentales para realizar el análisis exploratorio de los datos y para los futuros modelos de aprendizaje automático orientados a la clasificación de péptidos.

## ANÁLISIS EXPLORATORIO DE DATOS

Tras unificar los conjuntos de datos con el proceso realizado anteriormente, se procedió con el análisis mediante gráficos;

el de cajas y bigotes arrojó que parece haber una gran cantidad de datos "atípicos", pero se conoce que las propiedades fisicoquímicas de los péptidos tienen una gran variabilidad según la cantidad y variedad de aminoácidos que posean, así que con el gráfico no hay suficiente evidencia para eliminarlos. Además, se sacaron los datos que parecían ser atípicos, sin embargo, dado que, tras analizar cada columna, se concluye que estos datos atípicos no se deben a errores de medición, cálculo o datos incorrectos. Por el contrario, son fuentes únicas de información que enriquecerán los modelos, aumentando la robustez de estos.

Con el mapa de calor, se analizó la correlación entre columnas, lo que arrojó que las correlaciones entre aminoácidos (A, R, N, etc.) suelen ser bajas, lo que indica que la presencia de un aminoácido específico no afecta significativamente a otro en este conjunto de datos. Sin embargo, algunas combinaciones de aminoácidos tienen correlaciones negativas o positivas moderadas, lo que podría reflejar patrones específicos de interés en la secuencia de aminoácidos.

Por otro lado, las propiedades fisicoquímicas como Hidrofobicidad, Índice de Boman, y Carga muestran correlaciones diversas con los diferentes aminoácidos. Por ejemplo, la correlación negativa entre Hidrofobicidad y el aminoácido R sugiere que cuando R es alto, la Hidrofobicidad tiende a ser baja.

Al agrupar los péptidos según su funcionalidad específica y calcular la media de cada aminoácido en cada categoría, se observó que la L (Leucina) y la K (Lisina) destacan en todas las categorías funcionales, siendo los aminoácidos que aparecen en mayor cantidad promedio en cada funcionalidad. Esto sugiere que tanto la leucina como la lisina podrían desempeñar roles críticos en la eficacia de los péptidos en sus respectivas funciones biológicas.

Por último, al agrupar los péptidos según su función específica, y analizarlos con las propiedades de: carga, longitud, índice de boman, e hidrofobicidad, mediante el gráfico de cajas y bigotes se encontró que por índice de Boman, carga e hidrofobicidad, los tres gráficos arrojan que no hay mucha variación según sus características, no obstante, por longitud Los AMP antifúngicos y antibacterianos tienden a ser más largos y mostrar mayor variabilidad en comparación con otros y por punto isoeléctrico el anti VIH tiene un menor punto isoeléctrico en comparación con los demás.

**NORMALIZACIÓN Y SELECCIÓN DE CARACTERÍSTICAS**

Primero, se realizó la prueba de Shapiro para evaluar si las variables seguían una distribución normal, resultando que no la seguían. Esto determinó el tipo de normalización a aplicar. Se observó una gran variabilidad entre las escalas de las variables sin normalizar, lo que justificó la necesidad de normalización previa a la selección de características.

Se aplicó la normalización Min-Max por ser la más adecuada para los datos, descartando otras opciones como Z-score, robust scaling y log transformation, debido a su alto costo computacional o la falta de distribución normal en los datos.

Posteriormente, se analizó la varianza de las columnas normalizadas y se utilizó el método VarianceThreshold, con un umbral de 0.0025, logrando reducir de 1507 a 903 columnas, manteniendo un equilibrio entre la reducción de dimensionalidad y la retención de información relevante.

Como segundo método, se empleó la selección basada en árboles (Tree-based), utilizando la variable objetivo antimicrobiano en el primer nivel de nuestro modelo multinivel, y 50 árboles para optimizar el tiempo de ejecución. Este método eliminó 554 variables, quedando con 349.

Finalmente, se aplicó SelectKBest con el parámetro k = 245, basándonos en un análisis PCA que sugiere que con 245 componentes se puede explicar un alto porcentaje de la variabilidad (90%), sin embargo, falta evaluar los modelos con estas características y ver que tan eficientes son en el entrenamiento y la predicción.

## METRICAS DE EVALUACIÓN

Dado que el problema es de clasificación multiclase y multietiqueta con clases desbalanceadas, se optó por utilizar el F1-score como métrica de evaluación. Esta métrica, ideal para casos de desequilibrio entre clases, se calcula como la media armónica de la precisión y el recall lo que penaliza a los modelos que tienen un desequilibrio en estas dos métricas. La siguiente ecuación (1) muestra el F1-score en términos de los elementos de la matriz de confusión

$$F1 - score = 2 \times \frac{\mathrm{Pr}\,ecision \times Recall}{\mathrm{Pr}\,ecision + Recall} \quad (1)$$

## ENTRENAMIENTO DE MODELOS

Dado que el problema es de clasificación multietiqueta, donde un péptido puede tener múltiples actividades simultáneamente, se implementaron técnicas especializadas para manejar esta complejidad. Estas técnicas permiten manejar de forma eficiente la naturaleza compleja del problema, capturando las relaciones entre las distintas etiquetas y las características de los datos.

Se desarrolló un modelo jerárquico de tres fases para abordar esta tarea:

**Primera fase:** Clasificación general de capacidad antimicrobiana. En esta etapa, el objetivo fue identificar si un péptido tiene capacidad antimicrobiana o no. Para ello, se evaluaron como principales candidatos los algoritmos Random Forest Classifier y Logistic Regression. Estas técnicas son reconocidas por su capacidad de manejar problemas complejos; Random Forest se destaca por su robustez frente al sobreajuste gracias a la combinación de múltiples árboles de decisión, mientras que Logistic Regression es eficaz para problemas lineales y ofrece interpretabilidad en las predicciones.

**Segunda fase:** Clasificación de actividad antitumoral y anticancerígena. En esta fase, se buscó identificar si un péptido presentaba actividades antitumorales o anticancerígenas. Se seleccionaron como técnicas candidatas LightGBM, XGBoost (XGB) y ExtraTrees. Estas herramientas son ampliamente utilizadas para tareas de clasificación complejas:

LightGBM y XGB destacan por su eficiencia y capacidad de manejo de grandes volúmenes de datos, aprovechando técnicas de ensamble y gradient boosting. ExtraTrees, por su parte, genera árboles completamente aleatorizados que proporcionan una alta capacidad de generalización.

**Tercera fase:** Clasificación de péptidos por funcionalidades específicas. En esta última etapa, se abordó la clasificación específica de las funcionalidades en los péptidos previamente identificados como antimicrobianos. Al igual que en la segunda fase, se emplearon LightGBM, XGB y ExtraTrees como principales

técnicas candidatas debido a su capacidad de modelar interacciones complejas entre las características y las etiquetas.

Para la creación de los modelos, se realizó una partición inicial de los datos en un 80% para entrenamiento y prueba, y un 20% destinado a la validación final del rendimiento del modelo utilizando datos no vistos previamente. Esta estrategia permite evaluar la capacidad del modelo para generalizar nuevos datos.

Además, para mejorar la robustez y fiabilidad del modelo, se implementó una validación cruzada utilizando StratifiedKFold. Esta técnica garantiza que cada pliegue contenga una distribución representativa de las clases. En el caso de problemas multiclase, se optó por el ajuste MultilabelStratifiedKFold, que asegura que cada partición mantenga la proporción de etiquetas en tareas de clasificación multietiqueta. El proceso se realizó con 5 pliegues (folds), lo que facilita una evaluación más confiable y reduce el riesgo de sobreajuste.

## RESULTADOS Y ÁNALISIS

Al inicio del proceso, se realizó una búsqueda exhaustiva de los mejores hiperparámetros para cada modelo utilizando GridSearchCV, una técnica que evalúa sistemáticamente combinaciones específicas de valores de hiperparámetros para encontrar la configuración óptima. Garantizando así que se identifiquen los parámetros que maximizan el rendimiento de los modelos. A continuación, se presentan los hiperparámetros resultantes:

**Tabla 2. Fase 1**

| Modelo | Mejores hiperparámetros |
|---|---|
| Random Forest | n_estimators = 100 |
| | criterion = ‘gini’ |
| Logistic Regression | penalty = 'l2’ |

**Tabla 3. Fase 2 y 3**

| Modelo | Mejores hiperparámetros |
|---|---|
| LightGBM | n_estimators = 1000 |
| | learning_rate = 0.1 |
| | boosting_type = ‘gbdt’ |
| XGBoost | n_estimators = 500 |
| | learning_rate = 0.1 |
| | eval_metric = 'mlogloss' |
| ExtraTrees | n_estimators = 300 |
| | criterion = ‘entropy’ |
| | |

Después de determinar los mejores hiperparámetros para cada modelo, se evaluaron nuevamente utilizando validación cruzada y datos no vistos previamente. La siguiente tabla (Tabla 4) presenta el promedio de los resultados obtenidos en los pliegues de la validación cruzada y el puntaje alcanzado con el conjunto de validación.

**Tabla 4. Promedio de los modelos con validación cruzada y datos no vistos previamente**

| **Fase 1** | | |
|---|---|---|
| | F1-score | |
| | Validación cruzada | Validación final |
| Random Forest | 0.93 | 0.93 |
| Logistic Regression | 0.88 | 0.89 |
| **Fase 2** | | |
| LightGBM | 0.49 | 0.49 |
| XGBoost | 0.47 | 0.47 |
| ExtraTrees | 0.45 | 0.43 |
| **Fase 3** | | |
| LightGBM | 0.78 | 0.78 |
| XGBoost | 0.78 | 0.78 |
| ExtraTrees | 0.78 | 0.78 |

| | | |
|---|---|---|

Con base en los resultados presentados en la tabla anterior (Tabla 4), se observa que en la **primera fase** ambas técnicas implementadas lograron un buen rendimiento al clasificar si un péptido posee capacidad antimicrobiana. Sin embargo, el Random Forest destacó como la técnica con el mejor desempeño, evidenciando un mejor F1 score, indicando un equilibrio superior entre precisión y sensibilidad. Esto puede atribuirse a la capacidad del modelo para manejar relaciones no lineales entre las características y su robustez frente a datos desequilibrados, lo que lo convierte en una opción ideal para esta etapa.

Para la **segunda fase**, se observó una notable caída en la capacidad de clasificación para las clases "anticancerígeno" y "antitumoral". Incluso con el mejor modelo, LightGBM, se obtuvo un F1 Score de 0.49 tanto en la validación cruzada como en los datos de validación no utilizados previamente. Este bajo rendimiento se atribuye a la escasez de instancias de péptidos con actividad en estas dos clases, lo que dificultó que los modelos identificaran patrones significativos necesarios para un buen desempeño.

Se exploraron diversas estrategias para mitigar este problema. Por un lado, se implementó un RandomOversampler, que duplicaba aleatoriamente instancias de las clases minoritarias y las añadía al conjunto de entrenamiento. Por otro lado, se realizó un experimento en el que, si un péptido presentaba ambas actividades (anticancerígena y antitumoral), se eliminaba una de ellas, dejando únicamente la otra con la esperanza de mejorar la capacidad predictiva del modelo. Sin embargo, ambos enfoques resultaron infructuosos, ya que no mejoraron los resultados y, en algunos casos, incluso los empeoraron. Esto evidencia la dificultad de abordar problemas de desbalance extremo en clasificación multiclase y multi etiqueta.

En la fase final, los tres modelos lograron un F1 Score similar de 0.78, lo cual, a primera vista, podría considerarse un buen resultado. Sin embargo, como se analizará más adelante (Tabla 5), persisten deficiencias en la clasificación de ciertas clases. Estas falencias están directamente relacionadas con los problemas previamente mencionados, particularmente la escasez de péptidos en algunas clases, lo que dificulta que los modelos identifiquen patrones consistentes para una clasificación precisa.

Al comparar los F1 Scores obtenidos en la validación cruzada y en la validación final (datos desconocidos), se observa que los modelos de cada fase logran un desempeño consistente. Esto indica que los modelos generalizan bien a datos nuevos y no presentan un sesgo significativo al clasificar instancias desconocidas, obteniendo resultados muy similares en ambas evaluaciones.

A continuación (Tabla 5), observaremos un análisis detallado del desempeño específico para cada clase en el conjunto de validación final, enfocándose únicamente en los mejores modelos seleccionados en cada una de las fases.

**Tabla 5. Puntajes del conjunto de validación para cada característica**

| **Fase 1 Random Forest** | **precision** | **recall** | **f1-score** |
|---|---|---|---|
| Antimicrobiano | 0.93 | 0.92 | 0.93 |
| **Fase 2 LigthGBM** | | | |
| Anticancerígeno | 0.82 | 0.35 | 0.49 |

| | | | |
|---|---|---|---|
| Antitumoral | 0.77 | 0.35 | 0.48 |
| **Fase 3 LigthGBM** | | | |
| Antibacteríano | 0.84 | 0.82 | 0.83 |
| Anti Gram + | 0.82 | 0.83 | 0.83 |
| Anti Gram - | 0.83 | 0.83 | 0.83 |
| Antifúngico | 0.77 | 0.62 | 0.69 |
| Antiviral | 0.80 | 0.64 | 0.71 |
| AntiVIH | 0.81 | 0.48 | 0.61 |
| Antiparasitario | 0.56 | 0.16 | 0.25 |

En la clasificación de actividad antimicrobiana de un péptido refleja un excelente balance entre precisión y recall. Indicando un gran desempeño para identificar correctamente los péptidos antimicrobianos, con pocos falsos positivos y negativos.

Para la siguiente fase, ambas clases tienen un F1-score medio tendiendo a bajo, causado principalmente por un recall reducido (0.35). Aunque la precisión es relativamente alta, el modelo no está capturando correctamente muchos casos positivos, posiblemente debido a un desequilibrio en la cantidad de datos de estas clases.

Finalmente, podemos observar que el desempeño del modelo mejora notablemente en la clasificación de la funcionalidad específica de los péptidos antimicrobianos cuando cuenta con una mayor cantidad de instancias por clase. Esto es evidente en clases como Antibacteriano, Anti Gram + y Anti Gram -, donde el modelo logra un F1-score sólido y equilibrado. Sin embargo, se presentan grandes dificultades en la identificación de clases como AntiVIH y Antiparasitario, debido a la escasez de datos representativos, con solo 953 y 425 instancias respectivamente.

La falta de datos en las clases con bajo F1-score, como Anticancerígeno, Antitumoral, AntiVIH y Antiparasitario, afecta directamente el desempeño de los modelos al dificultar la detección de patrones relevantes. Esto genera un desequilibrio en las métricas de clasificación, reflejado principalmente en un bajo recall. En consecuencia, los modelos tienden a omitir casos positivos, mostrando una incapacidad para generalizar correctamente debido a la limitada cantidad de datos disponibles para estas categorías.

## CONCLUSIÓN

En este proyecto se implementaron diversas estrategias para abordar un desafío actual y relevante: la identificación de péptidos con funcionalidades específicas, lo cual es clave para el diseño de alternativas terapéuticas más efectivas. No obstante, la escasez de datos en algunas clases, incluso con el uso de técnicas avanzadas de modelado, representó un obstáculo significativo para obtener clasificaciones precisas y generalizables. Con el fin de mitigar este desafío, se planea, en un futuro proyecto, la implementación de técnicas de deep learning orientadas a la generación de péptidos sintéticos. Estas técnicas podrían aportar nuevas secuencias y patrones únicos, enriqueciendo las clases minoritarias y contribuyendo a resolver el problema del desbalanceo en los datos.

## BIBLIOGRAFÍA

1. Organización Mundial de la Salud. Plan de acción mundial sobre la resistencia a los antimicrobianos. Organización Mundial de la Salud. 2016. https://iris.who.int/handle/10665/255204.
2. Vélez A, Mera C, Orduz S , Branch JW. Generación de péptidos antimicrobianos

mediante redes neuronales recurrentes. Revista DYNA. 88(216), pp. 210-219. 2021.
3. Wu, Q., Ke, H., Li, D., Wang, Q., Fang, J., Zhou, J.: Recent progress in machine learning-based prediction of peptide activity for drug discovery. Current topics in medicinal chemistry 19(1), 4–16 (2019). https://doi.org/10.2174/1568026619666190122151634
4. Waghu, F.H., Barai, R.S., Gurung, P., Idicula-Thomas, S.: CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Research 44(D1), D1094–D1097 (2016). https://doi.org/10.1093/NAR/GKV1051
5. Szymczak, Paulina & Szczurek, Ewa. (2023). Artificial intelligence-driven antimicrobial peptide discovery. Current opinion in structural biology. 83. 102733. 10.1016/j.sbi.2023.102733.