Informe - Clasificación de Péptidos Antimicrobianos

# Descripción del problema

Encontrar nuevas moléculas antimicrobianas que puedan ser usadas contra bacterias resistentes a los antibióticos actuales se ha tornado en un problema farmacéutico mundial. Los péptidos antimicrobianos han tomado importancia en el desarrollo de nuevos antibióticos, dado que tienen grandes propiedades que actúan como medio de defensa en contra de enfermedades producidas por microorganismos.

Existe una gran cantidad de péptidos presentes en diferentes partes como plantas, mamíferos, insectos, entre otros [1]. Como hay tantos agentes presentes en la naturaleza y medio ambiente, su identificación se ha convertido en unos de los más grandes retos de la inteligencia artificial con miras a construir soluciones que apoyen la investigación y desarrollo de nuevas moléculas antimicrobianas.

# Metodología implementada

Para efectos de lograr identificar los péptidos antimicrobianos se implementaron tres diferentes técnicas de machine learning, Random Forest, Redes Neuronales y Regresión logística, teniendo en cuenta:

1. Uso de datos de entrenamiento con información de péptidos antimicrobianos y no antimicrobianos, adicionalmente datos de validación independientes a los de entrenamiento.
2. División de los datos de entrenamiento en dos, un subconjunto de datos utilizado netamente para entrenar los modelos de machine learning y otro para probar que tan efectivos son dichos modelos clasificando los péptidos.
3. Selección de características, teniendo en cuenta que se dispone de 1760 características relacionadas a los péptidos. Encontrar cuales de ellas describen mejor a péptidos con propiedades microbianas. La selección de características se realiza hallando el valor de la información (IV) usando el peso de la evidencia (WOE).
4. Comparar resultado de ejecutar los modelos con las 1760 características y solo con aquellas que describen mejor las propiedades microbianas.
5. Balanceo de los péptidos antimicrobianos y no antimicrobianos, con el fin de que sean comparables la dos clases y así obtener resultados más confiables.
6. Normalización de la información con el fin de que cada una de las características sean homogéneas entre sí.
7. Evaluar cada métrica otorgada por las técnicas de machine learning para concluir sobre cada una de ellas y su poder de clasificación.

# Análisis y discusión de los resultados

## Prueba 1. Bosques aleatorios

### Técnica de machine learning

Bosques aleatorios o random forest por su nombre en inglés fue usado como clasificador

### Hiperparámetros ajustados

### Análisis de métricas

### Tiempos de entrenamiento

## Prueba 2. Regresión logística

### Técnica de machine learning

La técnica de regresión logística es un método de clasificación binaria

### Hiperparámetros ajustados

Los hiperparámetros ajustados de una regresión logística son:

**Penalidad.**

Toma los valores de l1 regresión Lasso (reduce el coeficiente de las características menos importantes a cero) y l2 regresión Ridge

**Parámetro de regularización** (C = 1 / λ).

Se prueban con los valores [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]. Para valores pequeños de C, aumentamos la intensidad de la regularización, lo que creará modelos simples que no satisfacen los datos (no se ajustara). Para valores grandes de C, reducimos el poder de regularización, lo que implica que el modelo puede aumentar su complejidad y, por lo tanto, sobreajustar los datos o ajustar mejor los datos

**class\_weight.**

Énfasis sobre cada una de las clases en caso tal de que las clases no estén balanceadas. Para efectos de la selección se toman los posibles valores [{1:0.5, 0:0.5}, {1:0.4, 0:0.6}, {1:0.6, 0:0.4}, {1:0.7, 0:0.3}]. Este parámetro no es necesario puesto que en preprocesamiento se balancean las clases.

**Solver.**

Algoritmo usado en el problema de optimización. Trata de encontrar el peso del parámetro que minimiza la función de costo. Opciones **[**'liblinear', 'saga', 'sag', lbfgs, 'newton-cg']

Una vez ejecutado hypertuning para determinar los hiperparametros que dan mejor precisión, se tiene:

Mejor puntaje: 1.0

Mejores parámetros: {'solver': liblinear, 'penalty': 'l1', 'C': 0.001, dual: False, max\_iter: 100-110, }

### Análisis de métricas

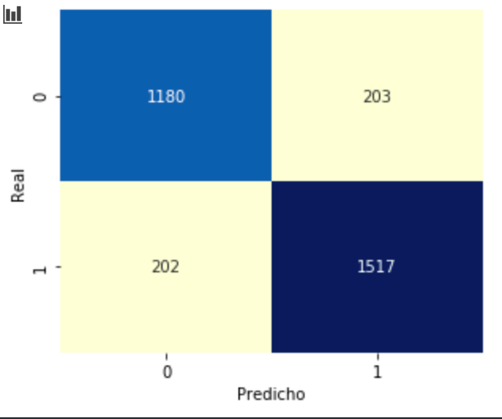
Al probar el modelo con los datos de validación dados en el ejercicio se evidencia un problema de performance mismatch, por lo cual se abordan los siguientes alternativas:

Sobreajuste

Datos de validación no representativos

Algoritmo estocástico

Modelo sin ajustar los hiperparametros



Modelo sin selección de características

Modelo con selección de características

### Tiempos de entrenamiento

# Conclusiones

# Referencias

[1]. Péptido antimicrobiano. <https://es.wikipedia.org/wiki/P%C3%A9ptido_antimicrobiano#Clasificaci%C3%B3n>