Predicción Regiones Epítopos en Antígenos para la Producción de Anticuerpos Contra el Virus SARS-CoV-2

DOCUMENTO EJECUTIVO

CODERHOUSE – DATA SCIENCE
NOVIEMBRE 2021
BUENOS AIRES

Ing Omar Ovalles

CONTENIDO

INTRODUCCIÓN AL PROBLEMA

La ciencia de datos ha probado ser bastante útil en proveer una perspectiva nueva a algunos de los problemas más complicados de la actualidad, incluyendo aquellos relacionados con la elaboración de vacunas para diferentes enfermedades cuyos efectos han mostrado ser bastante preocupantes. Entre estas enfermedades, se ha destacado la reciente pandemia declarada por la OMS (por alcance y peligrosidad) la cual es la enfermedad Covid-19 ocasionada por el virus SARS-CoV-2.

Se han observado abundantes ejemplos del uso de la ciencia de datos para predecir brotes locales (epidemias) con el objetivo de realizar una mejor administración de los recursos para atender los nuevos casos, o incluso se ha utilizado para métodos de diagnóstico más eficientes, Sin embargo, un enfoque diferente que se ha utilizado dentro de empresas farmacéuticas es el de utilizar estos métodos para agilizar el proceso de fabricación de vacunas contra el virus.

En este proyecto nos dedicaremos a este último enfoque.

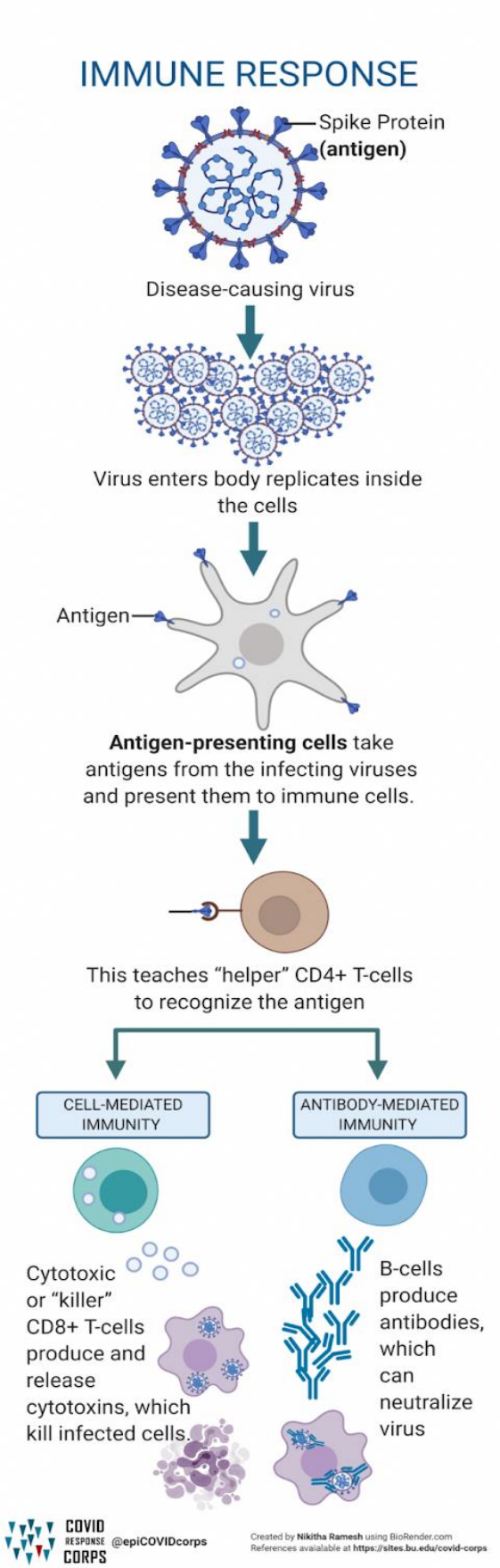
OBJETIVO DEL PROYECTO

- Analizar y predecir las regiones que podrían funcionar como Epítopos (puntos de adherencia) para anticuerpos generados por las “células-B” del sistema inmunológico adaptativo.
- Agilizar la producción de vacunas contra el virus SARS-CoV-2.

USUARIO FINAL

Organizaciones dedicadas a la salud, productores de fármacos, instituciones de investigación biológica y afines.

BASES TEÓRICAS

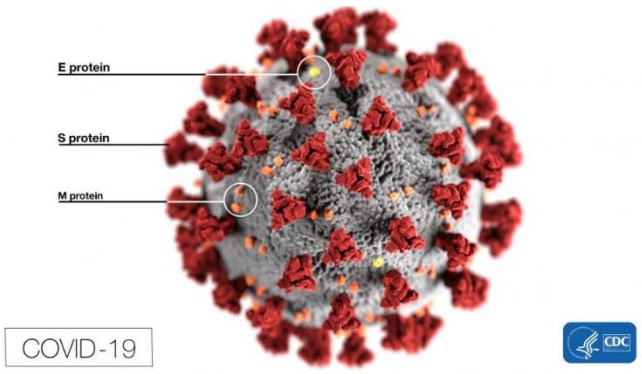


Para entender el análisis realizado al dataset, se debe conocer inicialmente el funcionamiento básico del sistema inmune adaptativo o lento:

1. El cuerpo es expuesto a un patógeno (agente externo que puede causar enfermedades), en este caso un virus, el cual inicia a replicarse tan pronto pueda adherirse a un tipo específico de célula en el cuerpo y utilizar sus recursos para esta tarea.

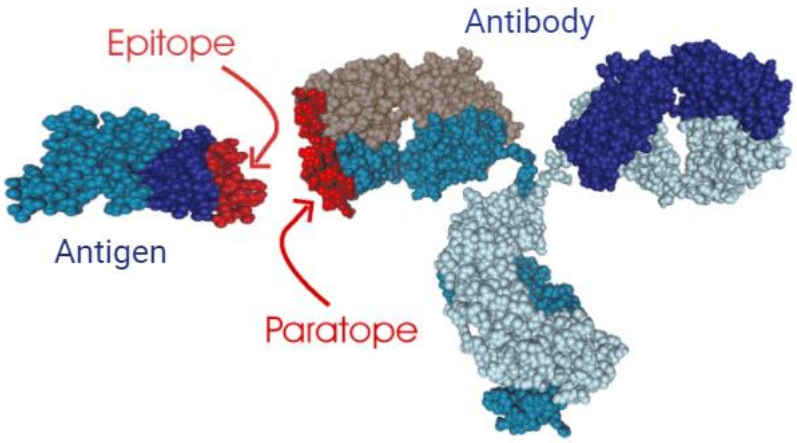
2. Existen células del sistema inmune generalmente conocidas como APC (Antigen Presenting Cells) las cuales toman una sección del patógeno (moléculas antígenas) para presentárselas a las células denominadas "CD4+ Helper T-cell".

3. Estas células "Helper T" luego pueden realizar diferentes tipos de activaciones, pero en este caso nos concentraremos en la estimulación de células "B" para la producción de anticuerpos específicamente diseñados para adherirse al patógeno dada la forma de sus componentes exteriores (usualmente picos en virus).



Las células “B” pueden neutralizar moléculas patógenas atacándolas con mucha especificidad utilizando receptores en su superficie (anticuerpos). Esto se logra mediante interacciones moleculares entre el parátopo (sección del anticuerpo involucrado en la unión) y la región de interacción (epítopo) de su molécula objetivo (antígeno).

Una vez que el epítopo y el receptor del anticuerpo se vinculan (como un rompecabezas), la producción de más anticuerpos se estimula con el objetivo específico de adherirse a tantos antígenos como pueda para neutralizar la acción del patógeno.



BASE DE DATOS

Descripción de la base de datos

Se utilizó una base de datos pública, disponible en la plataforma online "Kaggle" en el siguiente enlace

Proveedor: Future Corp.

Basado en el siguiente estudio

Datasets

La base está compuesta por las siguientes 3 tablas:

1. input_bcell: dataset de entrenamiento que posee la variable objetivo.
2. input_sars: dataset de entrenamiento que posee la variable objetivo.
3. input_covid: dataset de datos reales al cual aplicar el modelo predictivo.

Detalle de datasets

1. input_bcell

Contiene el listado de países con su código como clave primaria, el continente que integra y el código de este último.

Campo	Descripción	Tipo
parent_protein_id	ID único de la proteína a la cual pertenece el péptido sobre el cual se podría adherir el anticuerpo.	String
protein_seq	Secuencia de aminoácidos que compone el polipéptido completo (proteina).	String
start_position	Posición de inicio de la secuenca del péptido (en conteo de aminoácido).	Iint
end_position	Posición final de la secuenca del péptido (en conteo de aminoácido).	Int
peptide_seq	Secuencia de aminoácidos del péptido sobresaliente que forma el epítopo del cual se podría adherir el anticuerpo requerido.	String
chou_fasman	Característica del péptido (giro β), una	Float
emini	Característica del péptido, accesibilidad relativa de la superficie.	Float
kolaskar_tongaonkar	Característica del péptido, antigenicidad o capacidad para funcionar como antígeno.	Float
parker	Característica del péptido, hidrofobicidad o repelencia al agua (sólo en el péptido que podría formar el epítopo).	Float

isoelectric_point	Característica del péptido, pH al cual una molécula no posee carga eléctrica neta o es eléctricamente neutra (en media estadística). A este valor de pH, la solubilidad de la sustancia es casi nula.	Float
aromacity	Característica de la proteina, estabilidad inherente a estructuras resonantes cíclicas en compuestos orgánicos.	Float
hydrophobicity	Característica del péptido, hidrofobicidad o repelencia al agua (en la proteína completa para el caso de input_sars & covid).	Float
stability	Característica general a la estabilidad de la proteína, incluyendo resistencia a la temperatura, resistecia a la acción de enzimas destructoras de enlaces péptidos (peptidasa), resistencia a la agregación con proteínas mal formadas, etc.	Float

2. input_sars

Esta tabla posee la misma estructura que la anterior, sin embargo, se destaca el hecho de que los valores correspondientes a los últimos cuatro features (sin contar la variable objetivo), contienen el mismo valor, debido a que están basados en la proteína principal y este dataset contener solamente la proteína pico del virus SARS (Coronavirus PC4-205).

3. input_covid

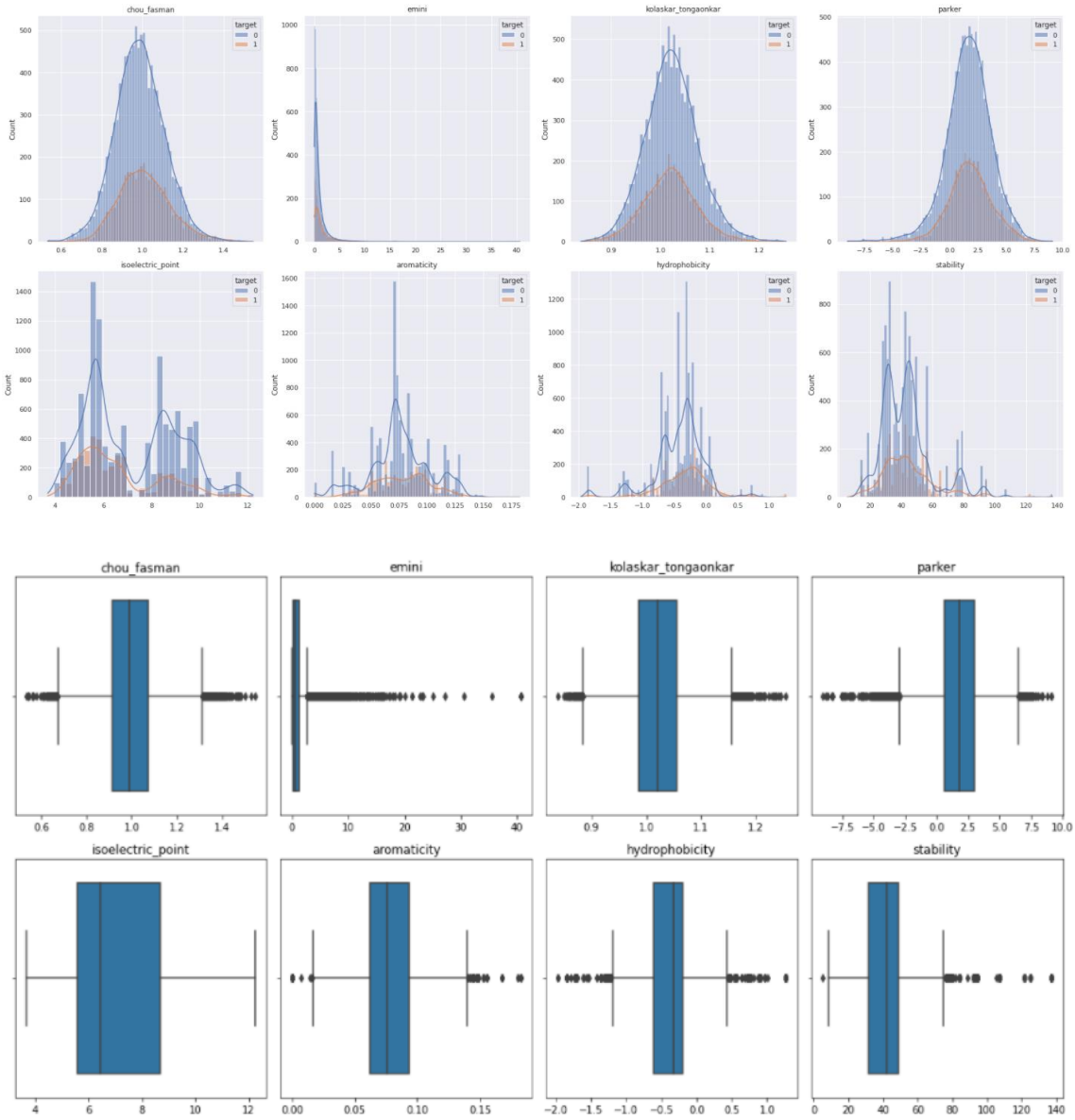
Finalmente, esta tabla es más parecida a la segunda (input_sars) sin embargo, su característica principal es que no posee la variable objetivo (target) por lo que se utilizará para la predicción final.

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Univariado

Se observó que las variables que proporcionan datos realmente valiosos poseen valores numéricos reales, los cuales se encuentran en las columnas *chou_fasman, emini kolaskar_tongaonkar, parker, isoelectric_point, aromaticity, hydrophobicity* y *stability*.

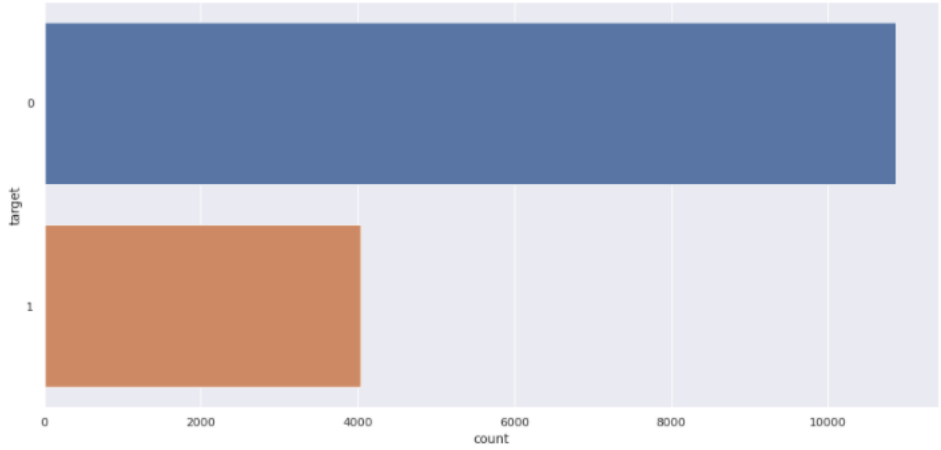
Las mismas poseen las siguientes distribuciones:



Podemos observar la distribución de cada una de las variables (primera imagen) y la ubicación de los valores observados con respecto a los rangos intercuartílicos (segunda imagen), de esto último se destacan que existen muchos valores “outliers” o atípicos (muy desviados de la media) principalmente en la variable “emini”, factor que hay que tomar en cuenta ya que estos

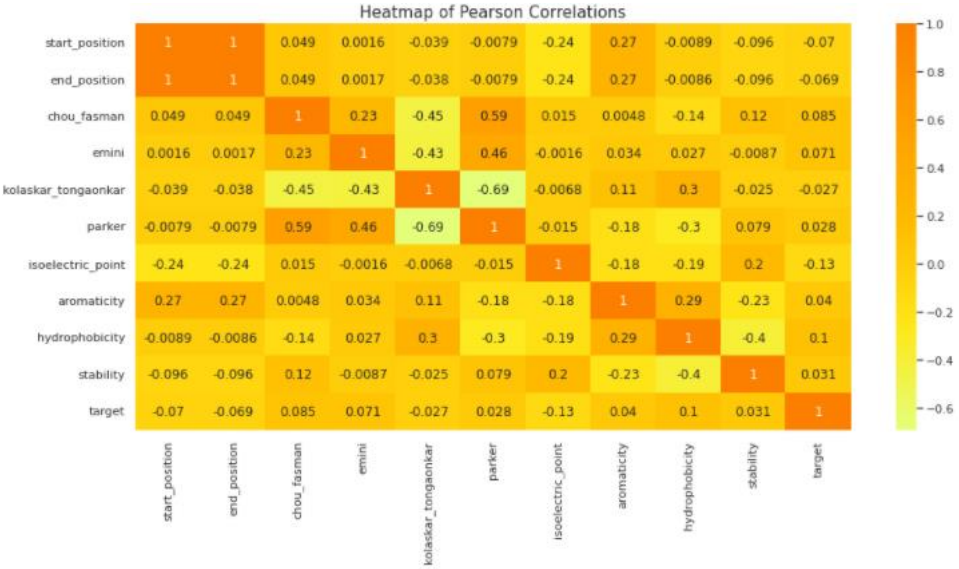
valores no corresponden a errores de ingreso en la base de datos porque todos son cálculos directos y por lo tanto poseen información relevante, sin embargo, las técnicas de escalado más comunes son susceptibles a outliers.

Un aspecto muy importante para realzar es el desbalanceo que presenta la variable objetivo, la cual posee casi el triple de valores negativos que positivos, por lo que hay que tomar este comportamiento en cuenta más adelante para segmentar los sets de entrenamiento y prueba cuidando de mantener la proporción.

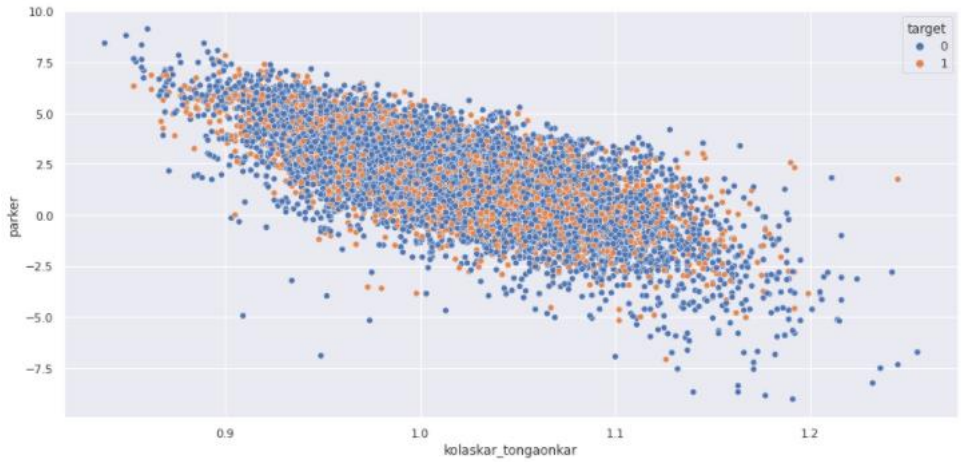


Bivariado/Multivariado

Se intentó determinar si existía alguna correlación de tipo lineal (Pearson) entre alguna de las variables, por lo que se calculó tal índice y se graficó en el siguiente mapa de calor:



Las relaciones más altas según este índice se dan entre las variables "kolaskar_tongaonkar" y "parker" en forma decreciente, sin embargo, no se puede apreciar visualmente una relación evidente entre estas dos variables, por lo que establecer conclusiones en este punto sería inadecuado.



DESARROLLO DEL MODELO

Ya que el problema principal presentado corresponde a un problema de clasificación, inicialmente se intentaron cada uno de los modelos vistos en clase individualmente para evaluar cuál de ellos presenta la mejor métrica

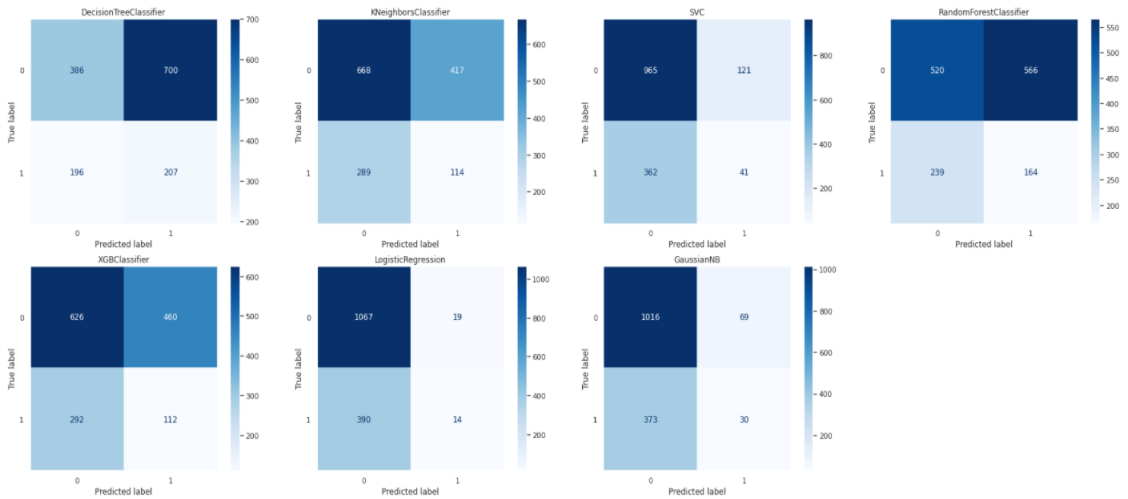
Métrica de evaluación

Al determinar cuál de los modelos probados es preferible, se debe definir la métrica de evaluación, la cual en este caso corresponde a la **cantidad de Falsos Positivos**, es decir, la cantidad de predicciones clasificadas como "1" que en realidad eran "0" tanto sobre el set de entrenamiento como sobre el set de prueba.

Dado esto, se probaron los siguientes algoritmos:

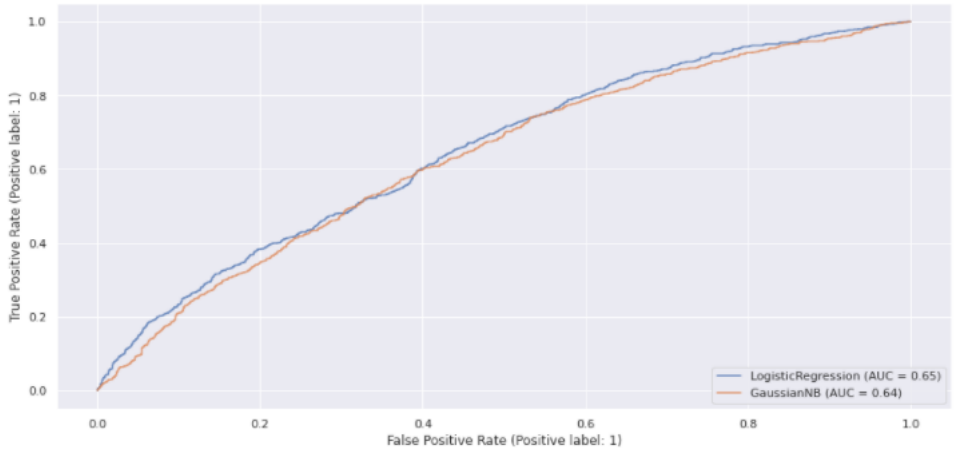
- Árbol de Decisión
- KNN (K Vecinos Cercanos)
- SVC (Support Vector Classifyer)
- Random Forest
- XGBoost
- Regresión Logística
- Naive Bayes

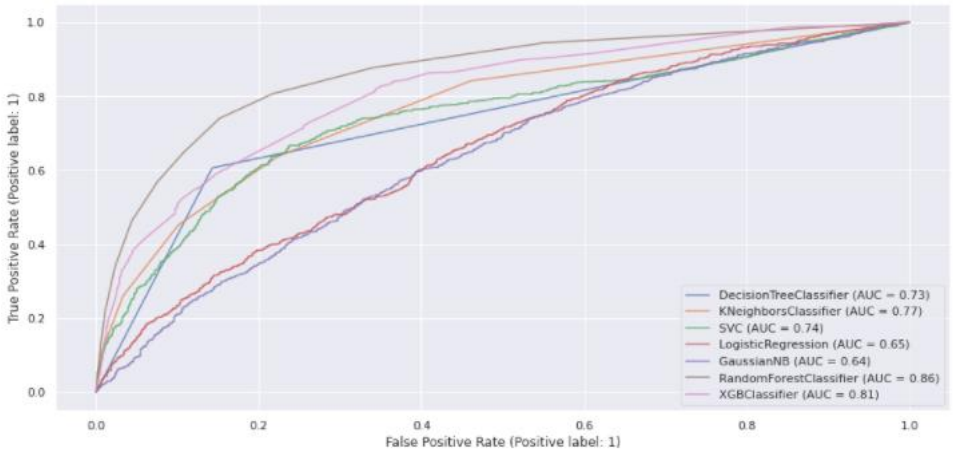
Para cada uno de los algoritmos se plotearon sus respectivas matrices de confusión calculadas con los valores promedio de Verdaderos Positivos (TP), Falsos Positivos (FP), Verdaderos Negativos (TN) y Falsos Negativos (FN).



Basándonos en la métrica definida anteriormente y los resultados obtenidos, se podría concluir que los dos modelos que proveen la mejor métrica son Regresión Logística y Naive Bayes a pesar de que otros modelos poseen mejores valores en métricas convencionales como el accuracy.

Sin embargo, basándonos en la visualización de la curva ROC para ambos modelos y su área bajo la curva, se puede identificar que poseen características muy similares y es poco probable que se pueda realizar una mejoría significativa en al optimizar los hiperparámetros de ambas, por lo que intentaremos comparar las mismas curvas para todos los modelos anteriores, lo cual debe hacerse utilizando únicamente un fold de separación train/test.





Podríamos inferir que debido a que la curva ROC correspondiente al modelo Random Forest posee un mayor área bajo la curva que la misma para el modelo de Regresión Logística, se podría hipotéticamente obtener una mejor relación entre la tasa de TP/FP incluso teniendo un menor número absoluto de FP.

Es importante resaltar el hecho de que el algoritmo Random Forest tiene la capacidad de sobre ajustarse a los datos de entrenamiento y por lo tanto tener mucho sesgo basado en la selección del dataset para tal fin.

OPTIMIZACIÓN DE HIPERPARÁMETROS

Se utilizó el método de Grid Search (búsqueda por grilla) para optimizar los hiperparámetros de los modelos seleccionados con las siguientes características:

- Regresión Logística: se debe utilizar una lista de diccionarios, ya que no todos los parámetros son intercambiables con el resto y corresponden a un orden predefinido

```
parameters = [{'solver': ['newton-cg'], 'penalty': ['l2']},
              {'solver': ['lbfgs'], 'penalty': ['l2']},
              {'solver': ['liblinear'], 'penalty': ['l1','l2']},
              {'solver': ['sag'], 'penalty': ['l2']},
              {'solver': ['saga'], 'penalty': ['elasticnet','l1','l2']
}]
```

- Random Forest:

```
parameters = {'n_estimators': (1,5,10,15,20),
              'criterion': ('gini', 'entropy'),
              'class_weight':('balanced', 'balanced_subsample'),}
```

Los resultados obtenidos de tal proceso fueron los siguientes:

- Regresión Logística:

```
FP on train set:  8.6
Best Parameters:  {'penalty': 'l1', 'solver': 'liblinear'}
```

- Random Forest:

```
FP on train set:  60.4
Best Parameters:  {'class_weight': 'balanced_subsample', 'criterion':
'gini', 'n_estimators': 20}
```

PREDICCIÓN FINAL

Según lo observado en las últimas dos secciones, se podría concluir que, para este caso y siguiendo esta metodología, el algoritmo que presenta las mejores métricas basándose únicamente en la minimización de la cantidad absoluta de "Falsos Positivos" es la Regresión Logística, realizando la optimización de hiperparámetros por el método de "Grid Search", se obtuvo el modelo a utilizar en nuestro dataset de datos reales (sin muestras de la variable objetivo).

Utilizando el mismo modelo en el dataset de datos reales a evaluar (input_bcell) se obtuvieron las siguientes predicciones, separadas por cantidad en cada categoría (“0” péptido no viable para ser región epítopo, “1” péptido viable para ser región epítopo)

Regresión Logística

target	
0	20262
1	50

Random Forest

target	
0	13646
1	6666

Ya que nuestro objetivo es obtener con la mayor certeza posible la cantidad de predicciones clasificadas con un uno (1), haremos una unión entre las predicciones de ambos modelos para obtener los índices correspondientes a tal unión, los cuales generan los siguientes resultados:

start_position	end_position	peptide_seq
163	180	YYHKNNKSWMESEFRVYS
163	181	YYHKNNKSWMESEFRVYSS
466	485	GNYNYLYRLFRKSNLKPFER
467	486	NYNYLYRLFRKSNLKPFERD
825	834	LPDPSKPSKR
826	839	PDPSKPSKRSFIED
828	835	PSKPSKRS
1158	1176	DPLQPELDSFKEELDKYFK
1158	1177	DPLQPELDSFKEELDKYFKN
1161	1177	QPELDSFKEELDKYFKN
1161	1178	QPELDSFKEELDKYFKNH
1161	1179	QPELDSFKEELDKYFKNHT
1161	1180	QPELDSFKEELDKYFKNHTS
1162	1179	PELDSFKEELDKYFKNHT
1162	1180	PELDSFKEELDKYFKNHTS
1162	1181	PELDSFKEELDKYFKNHTSP
1163	1181	ELDSFKEELDKYFKNHTSP
1163	1182	ELDSFKEELDKYFKNHTSPD

start_position	end_position	peptide_seq
1164	1182	LDSFKEELDKYFKNHTSPD
1165	1180	DSFKEELDKYFKNHTS
1165	1181	DSFKEELDKYFKNHTSP
1165	1182	DSFKEELDKYFKNHTSPD
1165	1183	DSFKEELDKYFKNHTSPDV
1165	1184	DSFKEELDKYFKNHTSPDVD
1166	1182	SFKEELDKYFKNHTSPD
1166	1184	SFKEELDKYFKNHTSPDVD
1167	1182	FKEELDKYFKNHTSPD
1168	1181	KEELDKYFKNHTSP
1168	1182	KEELDKYFKNHTSPD
1168	1184	KEELDKYFKNHTSPDVD
1168	1187	KEELDKYFKNHTSPDVDLGD
1169	1182	EELDKYFKNHTSPD

APÉNDICE

Fuentes de Información

- https://sites.bu.edu/covid-corps/projects/science-communication/types-of-vaccines-infographics/
- https://www.rcsb.org/structure/6VYB
- https://www.frontiersin.org/articles/10.3389/fimmu.2019.00298/full
- https://covalx.com/epitope-mapping-overview.php
- https://scikit-learn.org/stable/

Herramientas Tecnológicas Implementadas

- Pandas
- Scikit Learn
- Matplotlib, Seaborn
- Google Collab