Proyecto final

Técnicas de aprendizaje de máquina

Andres Useche

Juan Manuel

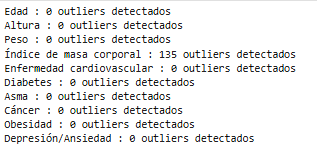
Rafael Torregroza

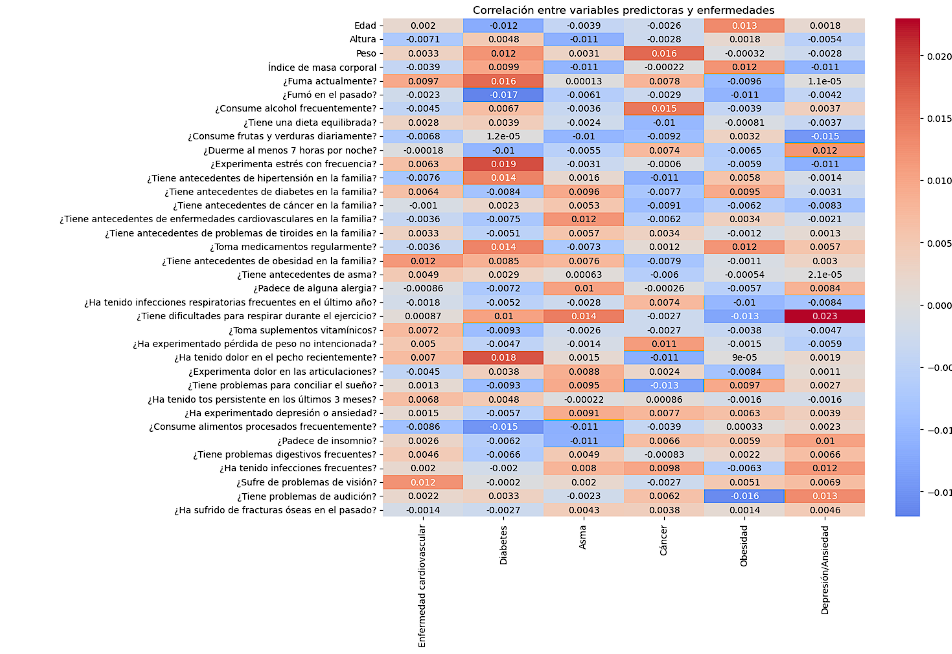
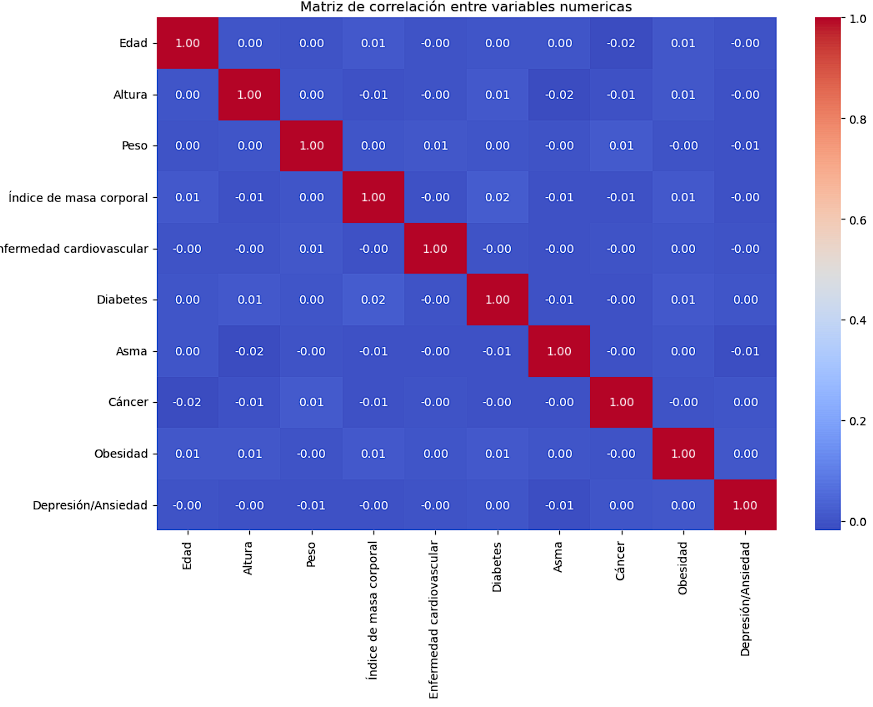
1. **Preprocesamiento y exploración:**

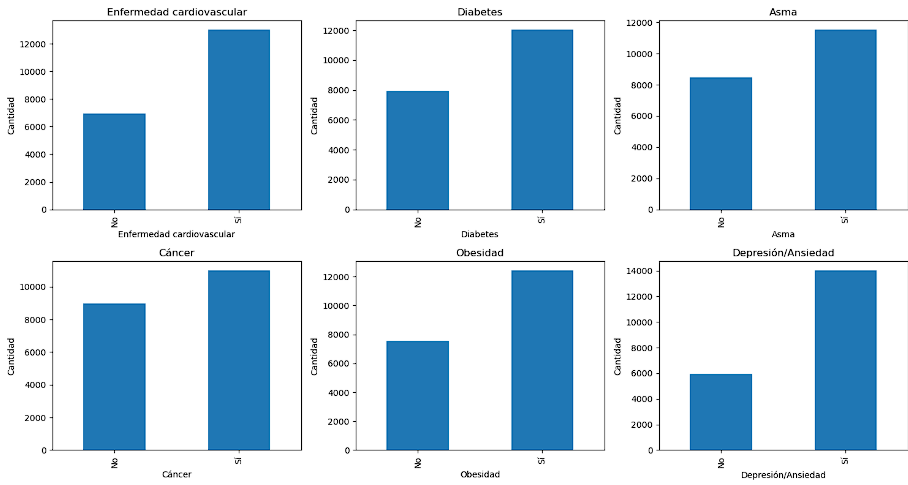
Para el preprocesamiento y la exploración de los datos se revisaron si el dataset contenía valores atípicos, nulos, si había correlación entre ellos, y si las clases estaban desbalanceadas entre ellas. De este análisis inicial se encontró que el dataset contiene principalmente variables categóricas, no contiene valores nulos, con una correlación entre ellas bastante baja casi nula, se encontró una muy pequeña cantidad de valores atípicos(135 pertenecientes solo a la variable Índice de masa corporal), por lo que decidimos no eliminarlos, aparte se encontró que había un desbalance entre las clases de las variables objetivo.

Primero verificamos si el dataset contenía:

* Valores atípicos → identificamos algunos pocos, pero no los eliminamos porque no representaban un problema grave ni eran errores evidentes.



* Valores nulos → encontramos que no había, lo que facilitó mucho el preprocesamiento.
* Correlación entre variables → detectamos que la mayoría de las variables no presentaban correlaciones significativas entre sí (es decir, eran bastante independientes).
* Desbalance entre clases → encontramos que algunas enfermedades, como el cáncer, tenían muchos menos casos positivos en comparación con la clase negativa, lo que indicaba un desbalance importante.



* Además, descubrimos que las variables eran principalmente categóricas (por ejemplo, preguntas de sí/no, niveles de actividad, antecedentes familiares), lo que implicó trabajar con codificación adecuada para los modelos posteriores.

Adicionalmente se realizaron análisis descriptivos y gráficas de dispersión para entender un poco más los datos. Estas gráficas se encuentran dentro del notebook.

1. **Selección de variables**

Aunque la matriz de correlación mostró relaciones débiles en general, identificamos algunas variables que tenían sentido clínico y mostraban asociaciones relevantes.  
 Por ejemplo:

* Obesidad ↔ índice de masa corporal, antecedentes familiares de obesidad
* Asma ↔ antecedentes familiares de asma
* Diabetes ↔ antecedentes familiares, dieta, peso

Estas variables se seleccionaron como claves, no solo por la estadística, sino porque son coherentes según el conocimiento médico del problema.

**Análisis de Componentes Principales (PCA)** Aplicamos PCA para reducir la dimensionalidad de los datos y extraer componentes principales que explicarían la mayor cantidad de varianza. Aunque cada componente explicaba solo cerca del 2.7% de la varianza, el conjunto total nos permitió generar representaciones latentes útiles, especialmente para modelos complejos como SVM o redes neuronales.

La decisión final fue:  
 - Para modelos interpretables (como regresión logística, árboles) → usar solo las variables seleccionadas por correlación.  
 - Para modelos complejos (SVM, redes neuronales) → usar el conjunto completo reducido por PCA, ya que estos modelos pueden trabajar bien con patrones combinados.

1. **Modelos de clasificación**

Los modelos utilizados para trabajar el problema fueron

* Bagging
* Redes Neuronales
* SVM
* Boosting

-GBM

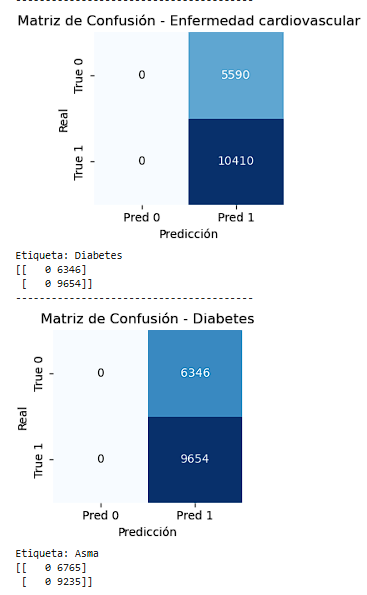
-XGboost

* Para el modelo de Bagging se utilizó random forest y bagging con árboles modelo el cual obtuvo los mejores resultados y se utilizó como modelo final
* Para Redes Neuronales se utilizó un modelo de red neuronal multicapa dado que el problema implica clasificar varias etiquetas, Para este modelo se utilizaron las variables de PCA
* También se utilizó un modelo de SVM para este modelo se utilizaron las variables de PCA
* Para Modelos de Boosting se utilizó el modelo de XG Boost y de GBM

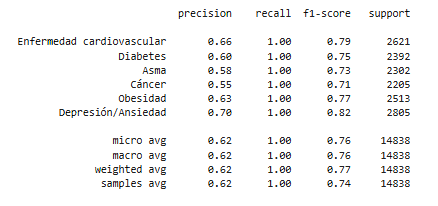
Para todos los modelos se definieron un rango de parámetros con los cuales utilizando grid search se buscó cuál era la mejor combinación de estos, para así obtener el mejor modelo posible.

Patrones de predicción:

En todos los modelos en el conjunto de entrenamiento clasificaba muy bien las personas que estaban enfermas, pero no las personas que no estaban enfermas(las predecía como enfermas). Como se ve acontinuación



En los conjuntos de entrenamiento, todos los modelos seguir en promedio estas métricas



Lo cual reafirma lo anteriormente mencionado

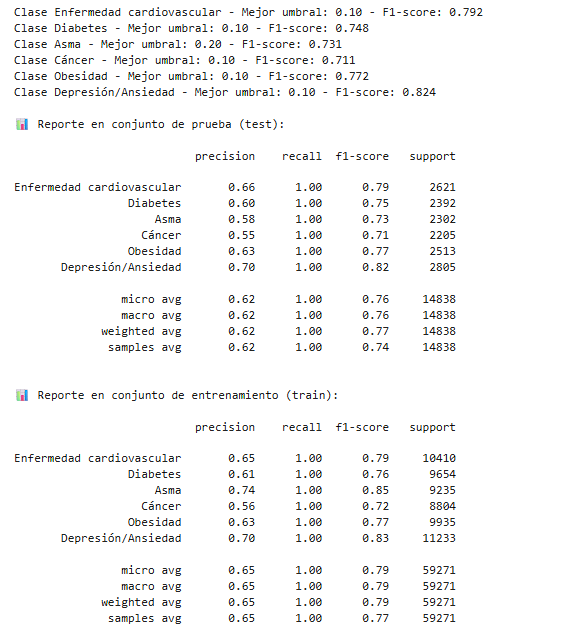
Para corroborar si el desbalance de clases estaba afectando el modelo se ajustaron los umbrales de probabilidad de predicción por cada enfermedad en los mejores modelos. En bagging y en GBM.

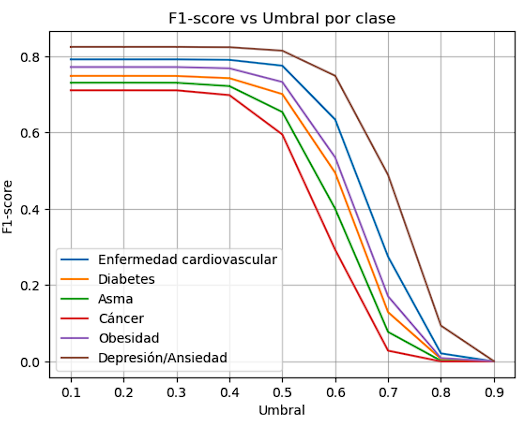
Estos fueron los parámetros y las métricas:

Bagging:

Tiempo de entrenamiento: 92.95 segundos

Mejores parámetros: {'estimator\_\_estimator\_\_min\_samples\_split': 10, 'estimator\_\_max\_features': 0.5, 'estimator\_\_max\_samples': 0.5, 'estimator\_\_n\_estimators': 50}



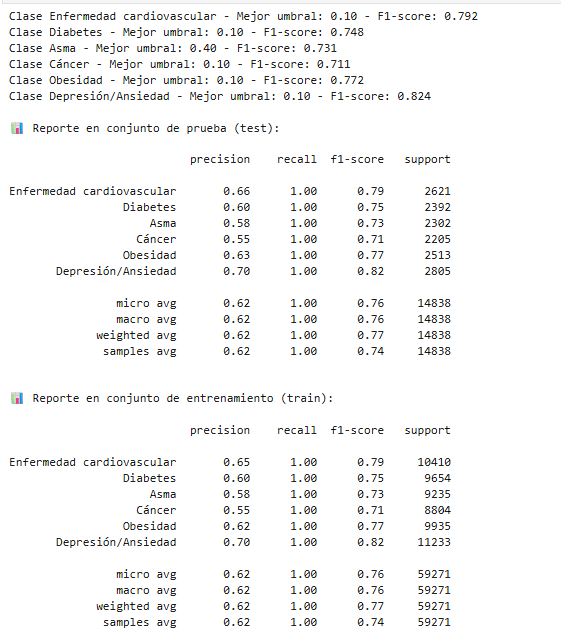


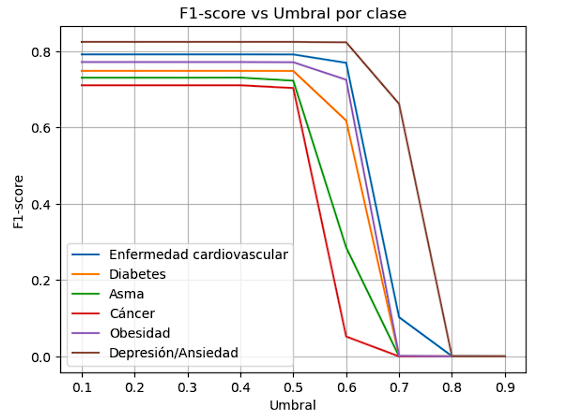
GBM

Tiempo de entrenamiento: 126.95 segundos

✅ Mejores parámetros: {'estimator\_\_learning\_rate': 0.01, 'estimator\_\_max\_depth': 3, 'estimator\_\_n\_estimators': 50}

📊 Reporte de clasificación:





El modelo utilizado finalmente fue el modelo de Bagging con árboles de decisión, y para intentar mejorar las predicciones de este modelo se probó y busco el mejor punto de corte en cuanto a la probabilidad de las clases predichas, iterando varias veces se encontró un punto de corte ideal para cada etiqueta tanto para el Precicion y para el F1 Score.

1. **Evaluación**

Para evaluar el rendimiento de los modelos aplicamos las métricas solicitadas en el enunciado del proyecto: precisión (accuracy), recall y F1-score.

Se compararon los resultados por modelo y por cada una de las enfermedades objetivo.  
 También tomamos en cuenta los tiempos de entrenamiento, dado que algunos modelos, como las redes neuronales y el SVM, requieren más tiempo y recursos computacionales que modelos como Random Forest.

Los resultados mostraron que el modelo de bagging (en particular, Random Forest) logró un equilibrio sólido entre precisión y capacidad de generalización, superando a los otros modelos en la mayoría de las enfermedades, especialmente en aquellas con clases desbalanceadas. El modelo de bagging con árboles de decisión que se utilizó al final también se evaluó según el mejor punto de corte para cada variable predicha con lo que mostró una gran capacidad para predecir la clase positiva para enfermedad, y un poco de deficiencia en predecir a los saludables.

El modelo de redes neuronales mostró un buen rendimiento en enfermedades donde los patrones eran menos lineales, pero fue más sensible al ajuste de hiper parámetros. El SVM tuvo un rendimiento aceptable, aunque con tiempos de entrenamiento más largos, y el boosting también mostró buenos resultados, destacándose en recall, pero requiriendo más ajuste fino.

1. **Conclusiones**

De los modelos utilizados se pudo encontrar que el modelo de bagging con árboles de decisión fue el mejor modelo para el problema, y para adaptar mejor el modelo al problema multietiqueta podemos utilizar puntos de corte diferentes para cada variable haciendo un modelo más robusto. En conclusión, fue posible implementar un modelo que clasificara con bastante predicción si una persona padece alguna o todas las enfermedades, descritas en el dataset y aunque tenga la tendencia a clasificar más positivos, esto tiene más valor que clasificar más personas sanas dado que equivocarse en clasificar una persona enferma tiene implicaciones más graves.

Enlace de la presentación

* <https://youtu.be/2M6sITk9-GA>