La gran pandemia

Modelado

**Repositorio =** [**GitHub**](https://github.com/Vosinepi/covid_data_coderhouse_2023)

Al querer predecir la variable fallecidos nos sucede algo muy recurrente cuando se quieren realizar modelos con datos médicos. La variable objetivo generalmente está muy desbalanceada.

Por esta razón aplicamos métodos de balanceo. En este caso el método es de submuestreo. Lo que hacemos es quedarnos con los fallecidos y equiparar la cantidad de sobrevivientes con estos.

Total datos dataset normalizado: **1048575**

Total datos luego del rebalanceo: **153884**

Variables utilizadas:

| features = ['USMER', 'SEX','INTUBED', 'PNEUMONIA','AGE', 'PREGNANT', 'DIABETES', 'COPD', 'ASTHMA', 'INMSUPR', 'HIPERTENSION', 'OTHER\_DISEASE', 'CARDIOVASCULAR', 'OBESITY', 'RENAL\_CHRONIC', 'TOBACCO', 'ICU', 'PATIENT\_TYPE']  target = 'fallecidos' |
| --- |

Modelamos utilizando Best Param

Mejores hiperparámetros: {'learning\_rate': 0.1, 'max\_depth': 3, 'n\_estimators': 300}

Probamos el modelo con nuestros datos de test

Accuracy: 0.9110049712447607

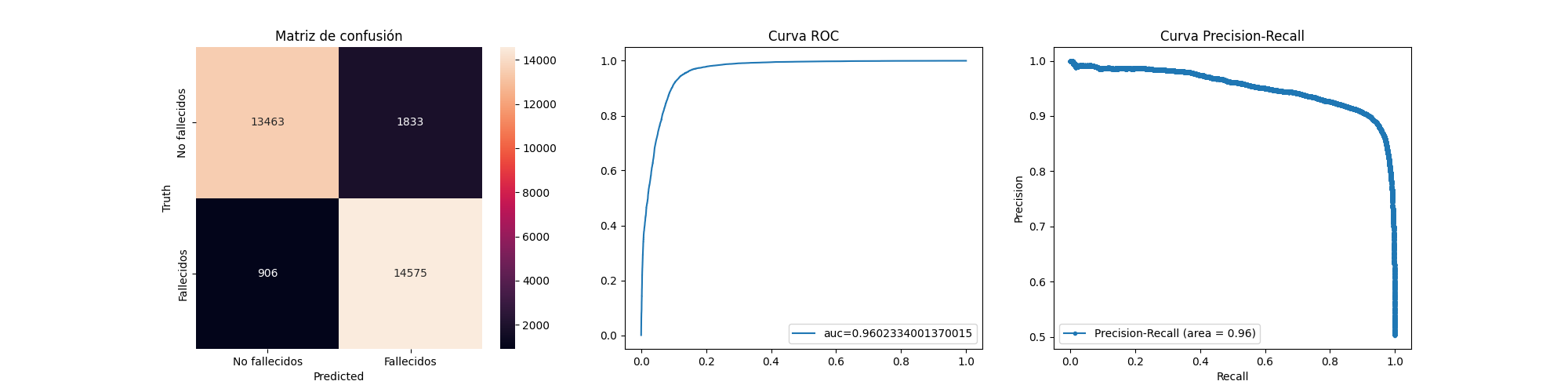
Precision: 0.8882862018527548

Recall: 0.9414766487952975

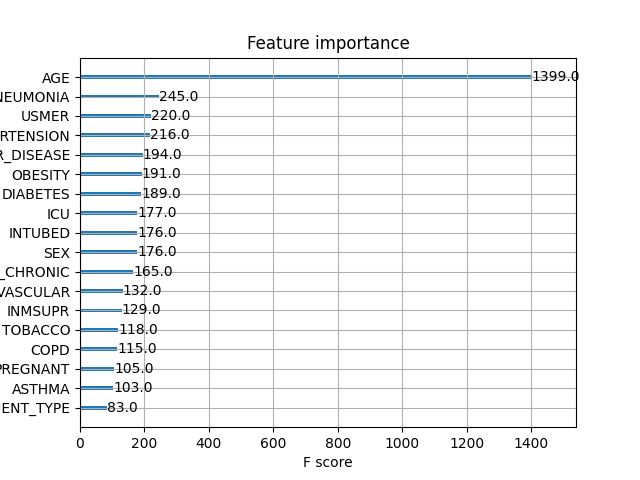
F1: 0.9141083132114523

Matriz de confusión: [[13463 1833] [ 906 14575]]

En este caso, el modelo clasificó correctamente 13463 instancias como negativas (verdaderos negativos) y 14575 instancias como positivas (verdaderos positivos). Sin embargo, cometió errores en 1833 instancias clasificándolas erróneamente como positivas (falsos positivos) y en 906 instancias clasificándolas erróneamente como negativas (falsos negativos).

Obtuvimos un Recall del 94% esto es aceptable para un modelo de estas características donde tratamos de evitar tener que dar falsos positivos como resultados ya que no hay que olvidar que estamos hablando de personas que contrajeron una enfermedad que muchas veces requiere de atención. 

Importancia de las variables



Descontando la edad que es la variable que más influye en los fallecidos de covid vemos que si se contrae Neumonía el riesgo aumenta significativamente, le siguen la hipertensión, obesidad y diabetes.

Análisis de Falsos Positivos y Negativos

Lo que podemos inferir observando los datos de los Falsos es que por más que tengamos en cuenta la mayor cantidad posible de variables en torno a una enfermedad aún quedan muchas por fuera. Analizando los gráficos observamos que por ejemplo los falsos positivos fueron perfectamente interpretados por el modelo, mostrando que la evolución y desenlace de una enfermedad aunque muchas veces es lineal siempre hay factores de aleatoriedad que van a terminar en un resultado diferente al predicho. Siempre vamos a tener pacientes perfectamente sanos que no superan la enfermedad y pacientes con pre existentes riesgosas que afronten la enfermedad y la superen.

