**PROYECTO DE INTELIGENCIA ARTIFICIAL**

**Tópico:**  COVID 2020-2021

**Dataset:** <https://www.kaggle.com/datasets/mykeysid10/covid19-dataset-for-year-2020>

OBJETIVO

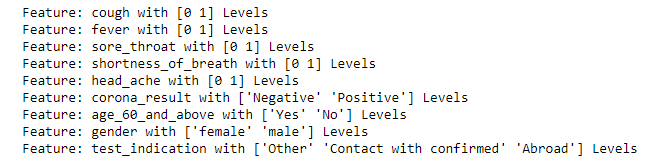
detectar si un paciente es Covid positivo o negativo en función de los síntomas, el sexo, la edad y las indicaciones de la prueba.

DESCRIPCION DE LOS CAMPOS

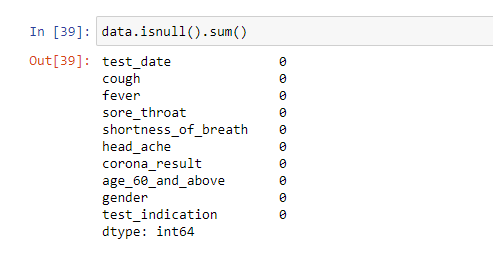
|  |  |
| --- | --- |
| test\_date  ,cough,  fever,  sore\_throat,  shortness\_of\_breath,  head\_ache,  corona\_result,age\_60\_and\_above,  gender,  test\_indication | fecha\_prueba,  tos,  fiebre,  dolor\_de\_garganta,  falta\_de\_respiración,  dolor\_cabeza,  resultado\_corona,edad\_60\_y\_mayores,  sexo,  indicación\_prueba |

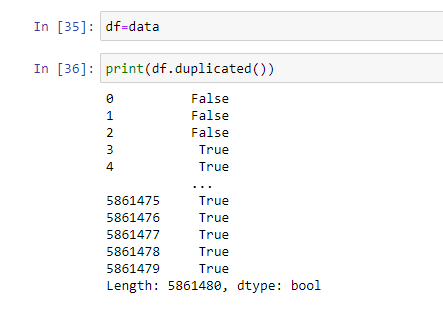
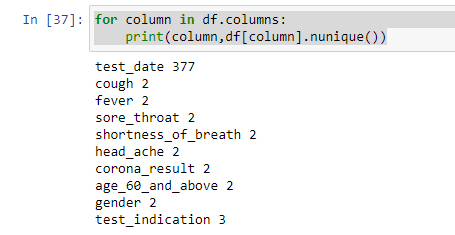
PREPROCESAMIENTO

CARACTERSTICA DE LOS DATOS

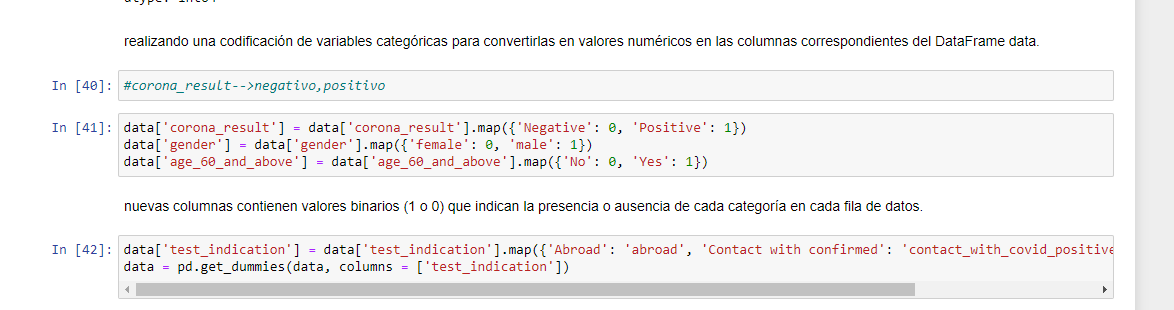


NO EXISTE VALORES NULOS

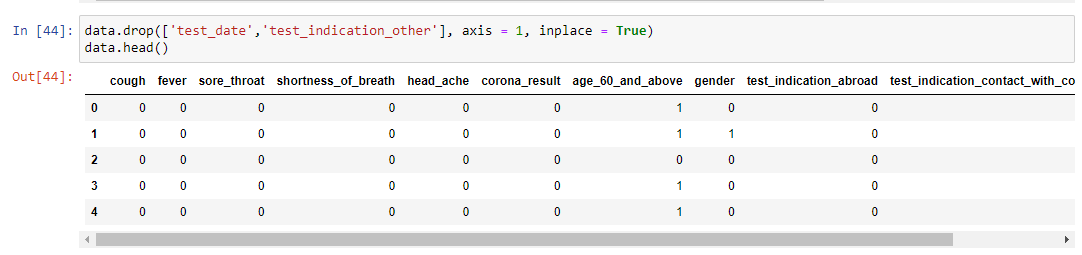


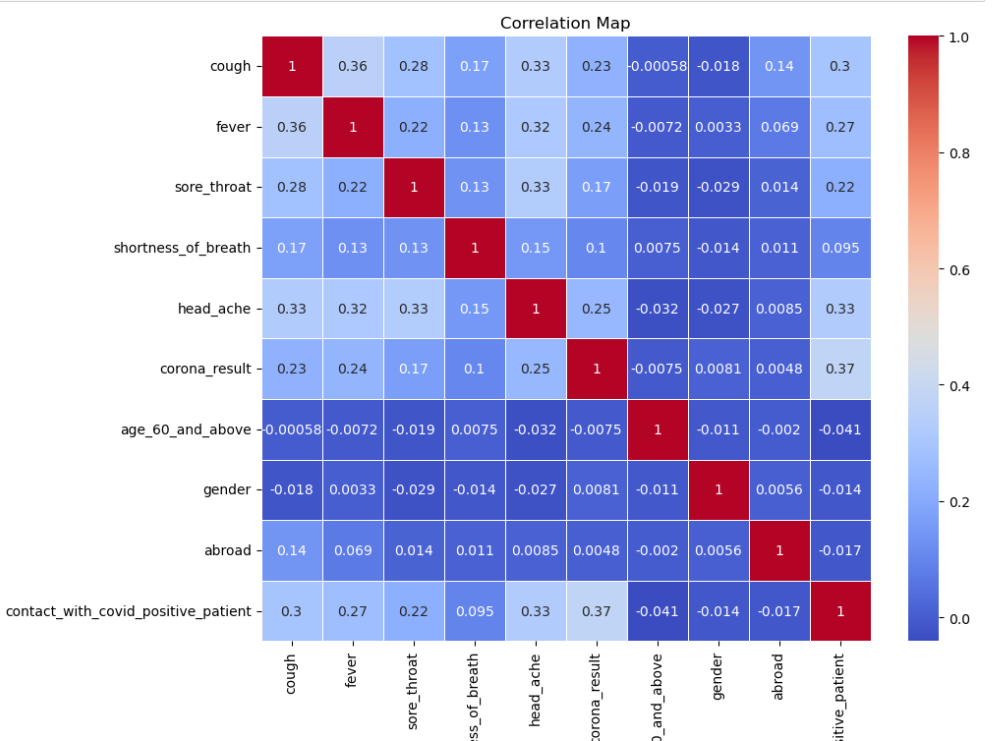
Conversión de variables categóricas para convertirlas en valores numéricos



Elimnacion de las colmnas test\_date, test indication\_other

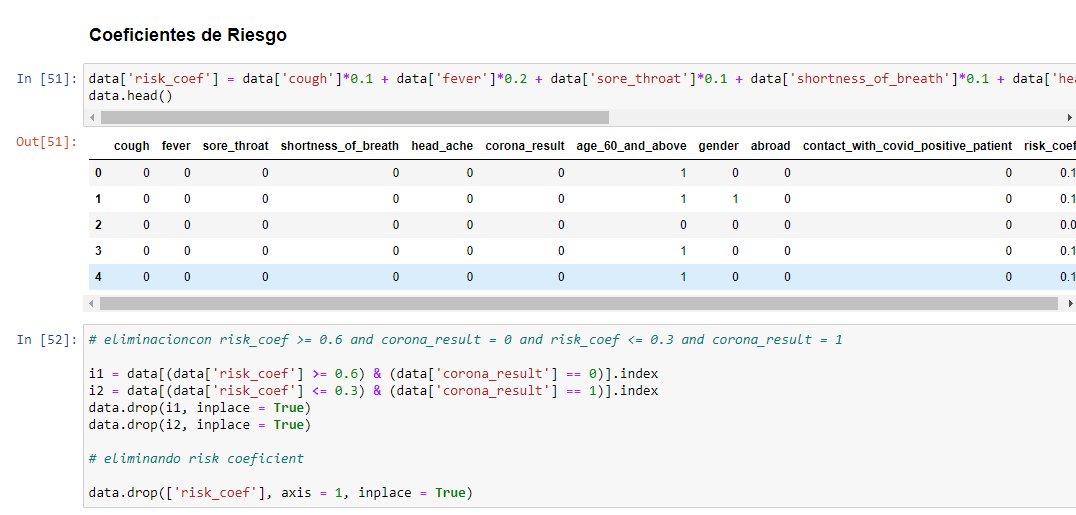


Matriz de correlación el cual indica que las variables no están fuertemente correlacionadas



Creación de coeficientes de riesgo

Va asignar aquellos registros que tengan acaracteristcas asociadas con un mayor riesgo Es decir si se tiene tos,fiebre,dificultad de respirar🡪mayor riesgo -🡪1



CLASIFICADOR

la aplicación de un clasificador de bosques aleatorios (Random Forest Classifier) con los mejores parámetros sugeridos por GridSearch.

En este caso, se crea una instancia del clasificador RandomForestClassifier con los siguientes parámetros:

n\_estimators = 200: Indica el número de árboles en el bosque.

max\_depth = 8: Especifica la profundidad máxima de cada árbol en el bosque.

criterion = 'gini': Indica el criterio utilizado para medir la calidad de la división de los nodos del árbol.

Además, se establece random\_state = 42 para asegurar la reproducibilidad de los resultados.

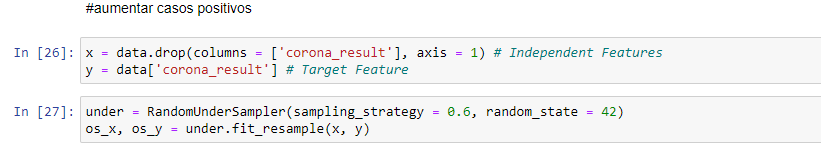
A continuación, se ajusta el clasificador utilizando el método fit con los conjuntos de entrenamiento X\_train y Y\_train.

Luego, se realizan predicciones tanto en el conjunto de entrenamiento (X\_train) como en el conjunto de prueba (X\_test) utilizando el método predict.

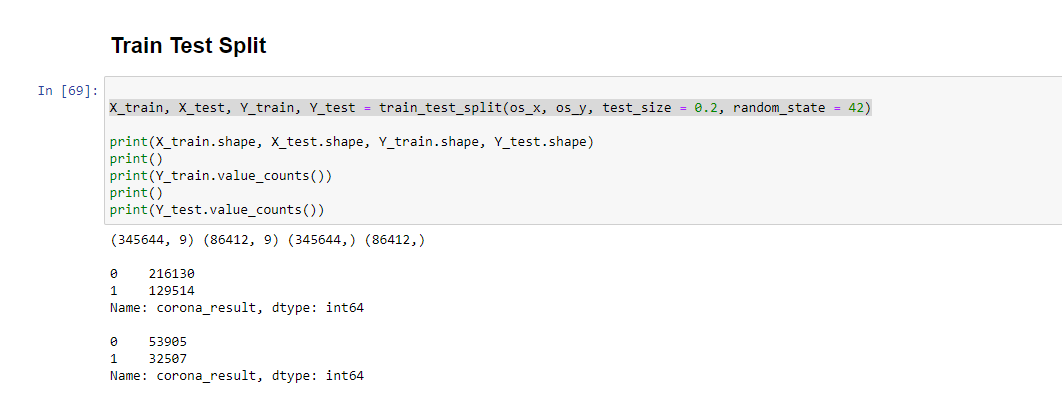
Finalmente, se evalúa el rendimiento del modelo utilizando la función evaluation\_parametrics, que probablemente incluye métricas como precisión, recall, puntuación F1 y exactitud para comparar las predicciones con los valores reales tanto en el conjunto de entrenamiento como en el conjunto de prueba.



Son datos diferentes a la vida real por lo tanto al ve que existen mas casos negativos, la idea es aumentar los casos positivos



ENTRENAMIENTO



"(345644, 9)" indica que el conjunto de datos original tiene 345644 filas y 9 columnas.

"(86412, 9)" indica que el conjunto de prueba (X\_test) tiene 86412 filas y 9 columnas.

"(345644,)" indica que las etiquetas del conjunto de datos original (Y\_train) tienen 345644 elementos.

"(86412,)" indica que las etiquetas del conjunto de prueba (Y\_test) tienen 86412 elementos.

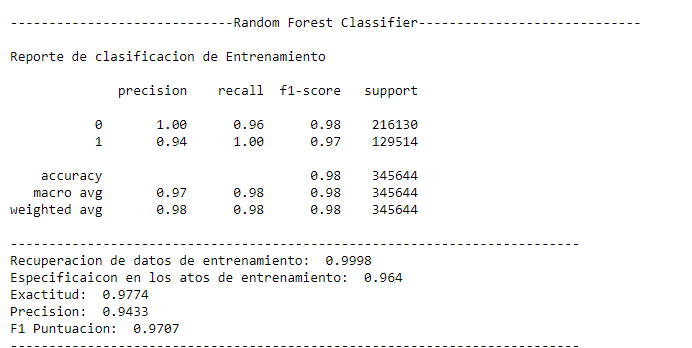
A continuación, se muestra la distribución de las etiquetas en el conjunto de datos original y en el conjunto de prueba:

En el conjunto de datos original, hay 216130 muestras etiquetadas como 0 (Negativo) y 129514 muestras etiquetadas como 1 (Positivo).

En el conjunto de prueba, hay 53905 muestras etiquetadas como 0 (Negativo) y 32507 muestras etiquetadas como 1 (Positivo).

Estos números indican la cantidad de muestras para cada clase en los conjuntos de datos, lo cual puede ser útil para comprender la distribución de los datos y el equilibrio de clases en el problema de clasificación.

Usando el clasificador



Precision: La precisión es la proporción de muestras correctamente clasificadas como positivas en relación con todas las muestras clasificadas como positivas. En este caso, la precisión para la clase 0 es del 100% y para la clase 1 es del 94%. Esto significa que el modelo tiene una alta precisión en predecir correctamente los casos negativos y una precisión ligeramente inferior en predecir correctamente los casos positivos.

Recall (Recuperación): El recall es la proporción de muestras positivas correctamente identificadas en relación con todas las muestras que realmente son positivas. En este caso, el recall para la clase 0 es del 96% y para la clase 1 es del 100%. Esto indica que el modelo tiene un alto recall en identificar correctamente los casos positivos en comparación con los casos negativos.

F1-Score: El F1-Score es una medida de la precisión y el recall combinados, calculada como la media armónica de ambos valores. En este caso, el F1-Score para la clase 0 es de 0.98 y para la clase 1 es de 0.97. Esto indica que el modelo tiene un buen equilibrio entre la precisión y el recall en ambas clases.

Accuracy (Exactitud): La exactitud es la proporción de muestras correctamente clasificadas en relación con todas las muestras. En este caso, la exactitud es del 98%, lo que indica que el modelo tiene un alto rendimiento general en la clasificación correcta de las muestras.

En resumen, el modelo Random Forest Classifier muestra un buen rendimiento en términos de precisión, recall, F1-Score y exactitud en el conjunto de entrenamiento. Sin embargo, es importante tener en cuenta que estos resultados se refieren al conjunto de entrenamiento y es necesario evaluar el rendimiento del modelo en el conjunto de prueba para obtener una evaluación completa de su desempeño.