# PREDICCIÓN DE PATOLOGÍAS ASOCIADAS A LA TIROIDES

**POR:**

Sara Camila Guarín Castillo, Carlos Hernan Molina Bustos

**MATERIA:**

Introducción a la inteligencia artificial

**PROFESOR:**

Raul Ramos Pollan


UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2022

**Preprocesamiento del dataset y avances del proyecto**

Se realiza la carga de los datos del dataset desde Kaggle a un notebook de colab, a partir del cual se inicia el análisis exploratorio del dataset.

En el dataframe se identifican 9172 filas y 31 columnas con información relevante de pacientes que tienen o han presentado enfermedades asociadas a la tiroides y se realiza una observación de toda la información contenida en este. Se hace una revisión del número de datos faltantes en el dataframe y se confirma que existen 14.629 valores *NaN* correspondientes al 5,136% del total de valores del dataframe (275.160). Se identifican las columnas que poseen valores nulos, encontrándose valores nulos en 7 de las columnas; luego se hace un promedio la cantidad de valores faltantes en cada uno de ellos y se observa que la columna TBG tiene la mayoría de sus valores nulos (96, 823%), por lo que se procede a eliminar dicha columna. Según lo anterior, el dataset cumple con las condiciones requeridas en cuanto a número de registros y columnas para su estudio.

De acuerdo con la información contenida en el dataframe, se identifican dos columnas de interés correspondientes a *query_hypothyroid* y *query_hyperthyroid*, las cuales son los registros booleanos de pacientes que presentan o no una enfermedad asociada a hipo o hipertiroidismo, donde true indica la presencia de la enfermedad y false su ausencia. Conocidas estas columnas, se indaga sobre la cantidad de pacientes que presentan o no una de las enfermedades de la tiroides mencionadas y se tiene que 630 pacientes presentan hipotiroidismo y 651 presentan hipertioidismo.

Debido a que *query_hypothyroid* y *query_hyperthyroid* son datos categóricos, para facilitar su estudio, se desea convertirlos en datos numéricos; es por ello que, mediante un mapeo, se convierten los valores *false* en 0 y *true* en 1.
Cabe aclarar que constantemente se realiza conteo y suma de los datos que se procesan para verificar que estos no se pierden.

Se realizan conteos de otras variables que son importantes para identificar la presencia de enfermedades de tiroides como lo son el embarazo, de la cual se tienen 9065 pacientes en estado de embarazo.

Debido a que existe interés en conocer cuántas mujeres y hombres tienen hipo o hipertiroidismo, se realizan varias agrupaciones de la siguiente manera:
Para dar una mejor interpretación de las variables que se estudian, se realizan gráficos que ayudan visualmente a su comprensión.

En la *Figura 1* se observa que la mayoría de los registros corresponden a mujeres pues, en el dataframe se tiene un conteo de 6073 mujeres y 2792 hombres.

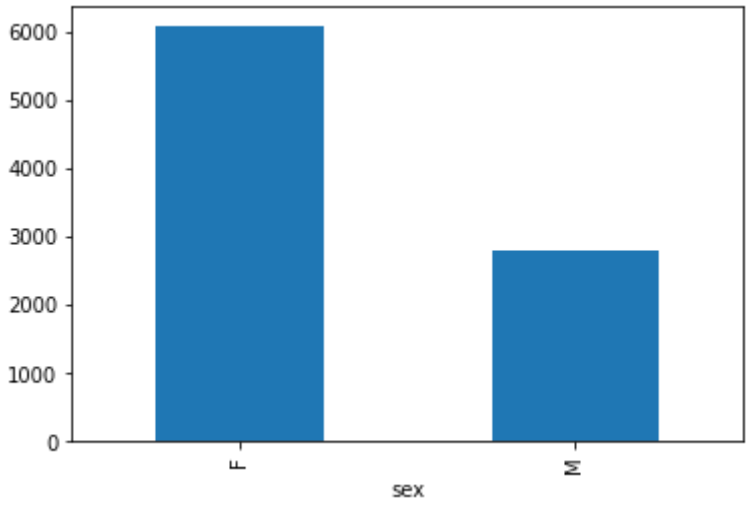


*Figura 1. Gráfico de barras indicando el sexo de los pacientes registrados en el dataframe.*

Además, se realiza un promedio de la edad según el sexo y como se observa en la *Figura 2* se tiene que el promedio de edad de las mujeres es aproximadamente 63 años y de los hombres es 99 años, indicando que los hombres elegidos para el estudio tienen una edad avanzada.

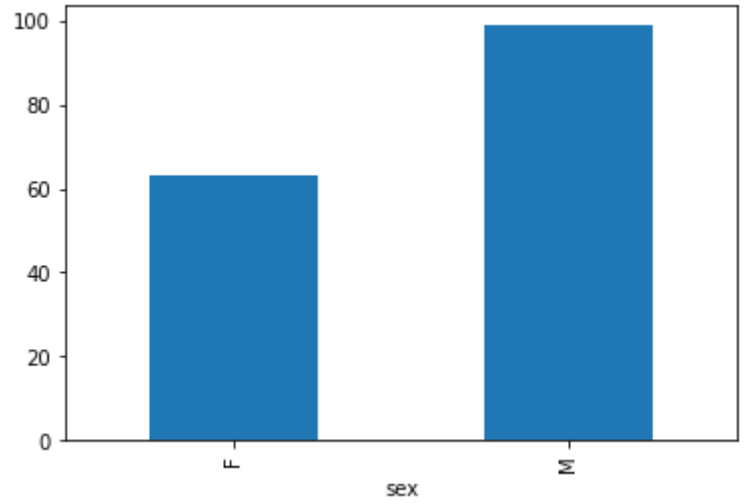


*Figura 2. Gráfico de barras indicando el promedio de edad de los pacientes registrados en el dataframe.*

Los registros de las columnas categóricas son reemplazados por valores binarios con el fin de hacer el tratamiento de estos datos en el modelado.
Ahora bien, las columnas con registros de tipo *object* son convertidos a datos numéricos.

Con los datos de las columnas, se visualizan histogramas para visualizar patrones o tendencias en los datos, como se observa en la *Figura 3*. Donde todas las columnas presentan datos recopilados en 0 y 1, excepto aquellos que tienen valores

numéricos diferentes como *T3, TT4, T4U* y *FTI l*os cuales muestran una distribución normal. Sin embargo, la edad no se comportó igual, pues los valores estuvieron concentrados en una sola barra, variable para la cual se espera una distribución normal, ver la *Figura 4*.

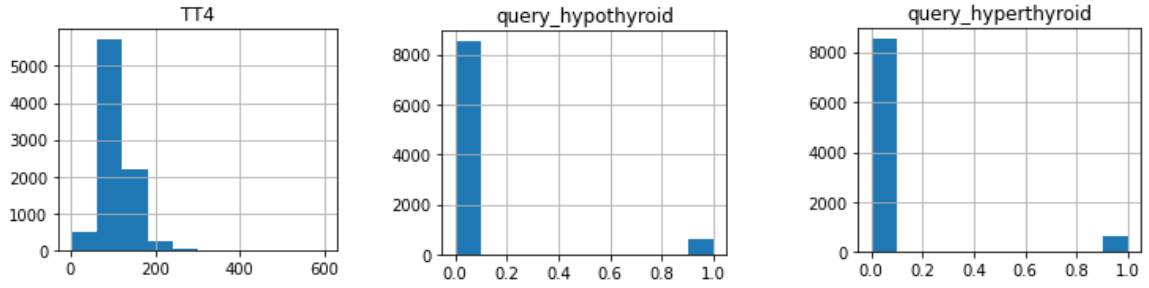


*Figura 3. Histogramas para sex, query_hypothyroid, query_hyperthyroid.*

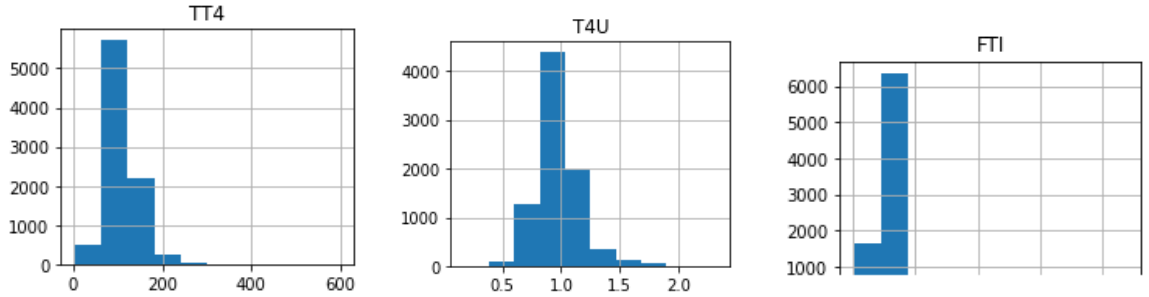


*Figura 3. Histogramas para TT4, T4U y FTI.*

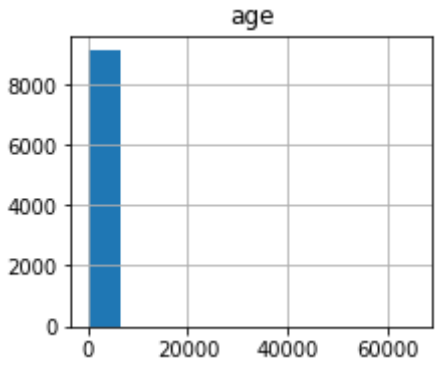


*Figura 4. Histograma de la edad.*

**Métricas de evaluación**

Como métricas de estudio para la entrega final se plantea emplear accuracy y *f1_score*; donde *accuracy* mide la exactitud del modelo (% de casos que el modelo ha acertado) y *f1_score* combina la precisión y la exhaustividad(*recall*) en un solo valor y se calcula haciendo la media armónica entre la precisión y la exhaustividad. Para el modelo se tendrán en cuenta el *query_hypothyroid* y el *query_hyperthyroid* que contienen información de interés sobre las personas que presentan o presentaron alguna vez enfermedades de la tiroides, hipotiroidismo o hipertiroidismo, respectivamente.

**Bibliografía**

**[1]** Werr, E. F. (2022). Thyroid Disease Data. Kaggle.com.

**[2]** Jose Martinez Heras. (2019, November 17). *Precision, Recall, F1, Accuracy en clasificación - IArtificial.net*. IArtificial.net. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/#:~:text=El%20valor%20F1%20se%20utiliza,la%20exhaustividad%20entre%20varias%20soluciones.&text=El%20valor%20F1%20asume%20que,la%20precisi%C3%B3n%20y%20la%20exhaustividad.

**[3]** merishnasuwal. (2020, November 10). *Diagnosing Hypothyroid disease using Deep Learning*. Kaggle.com; Kaggle. https://www.kaggle.com/code/merishnasuwal/diagnosing-hypothyroid-disease-using-deep-learning/notebook