ProyectoAlcoffe

Integrantes:

- **Miguel López Vélez - Bioingeniería 1001014378**
- **Maria Fernanda Villarreal Teherán - Bioingeniería**

El proyecto consiste en predecir la calidad del café arábigo según algunas características como altura, origen, variedad, método de proceso, procedencias, entre otras. La calidad de este se mide por medio del aroma, textura, acidez, dulzura, etc., para ello se usarán dichas columnas con valores de 1-10. Además de las columnas de origen y altura para tener una mejor estimación, inicialmente se piensa trabajar con los 1312 datos, pero esto se podrá reducir dependiendo de la capacidad computacional para procesarlos. En este caso el modelo se evaluará con varias pruebas, las que incluyen MAE, MSE, y R2 score. Finalmente se enfoca en nuevos caficultores que desean saber qué impacto tendrá su café, si las ventas de café no aumentan a base de 10% con las predicciones para la adquisición de nuevos cafés el proyecto no vale la pena.

Para esto primero se leyeron los datos obtenidos a través de la base de datos del URL: https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi?select=arabica_data_cleaned.csv , obteniendo inicialmente 1311 filas y 43 columnas de las cuales no todas las columnas serán de nuestro interés ya que no arrojaban datos relevantes como por ejemplo el número del lote, nombre de la granja, entre otras, estas columnas se procedieron a eliminar, dejando de esta manera solamente 24 columnas.

La primera variable que se analizó fue “Total.Cup.Points” por medio de un histograma (siendo esta la variable de interés para predecir), este se utiliza para agrupar los datos en x y contar la cantidad en cada contenedor, al ser el “bins” un número entero, se define de igual ancho en el rango. Se procedió entonces a calcular la simetría, la cual al ser menor a cero y con lo observado en el histograma se concluye que presenta una asimetría hacia la derecha, luego de esto, se realiza una transformación de 1 sobre la variable para normalizar y finalmente se eliminaron todos los valores en cero para que no diera infinito, obteniendo los resultados mostrados en la figura 1, de igual manera se hace el procedimiento con las diferentes variables (sabor, acidez, cuerpo etc) por la relación entre estas y la variable de interés.

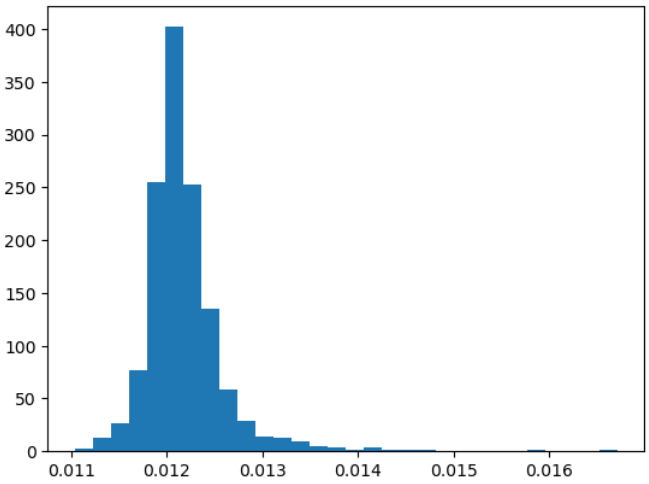

Figura 1. Histograma de la variable de interés normalizada.

La segunda variable a analizar fue el país de origen, obteniendo mayor cantidad de resultados en países latinoamericanos tales como México, Colombia y Guatemala (Figura 2), sin embargo, al realizar la comparación del origen con el promedio del "Total Cup Points" estos países no destacaron, mientras que, si lo hicieron Estados Unidos, Ethiopia y Papua New Guinea, esto se debe a que de estos países se encuentran menos registros en el data set (Figura 3).

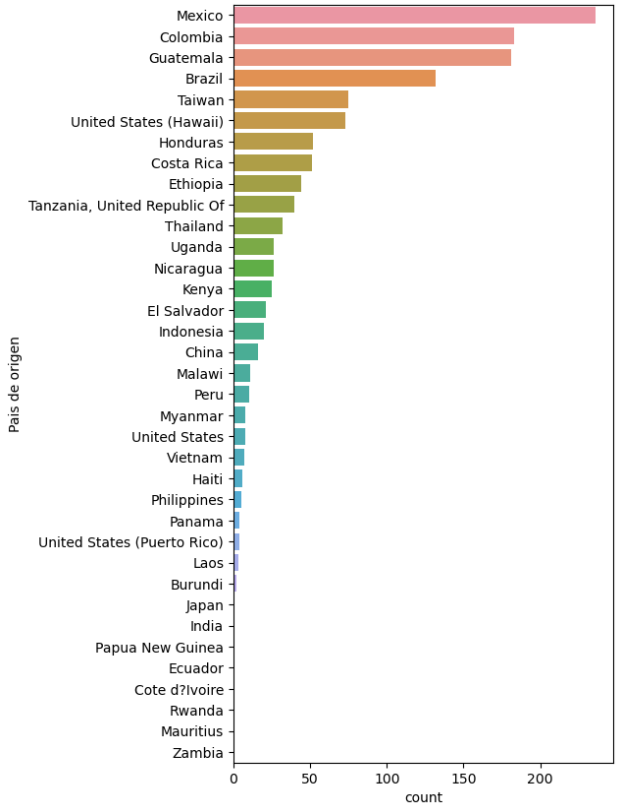

Figura 2. Cantidad de datos según el país de origen

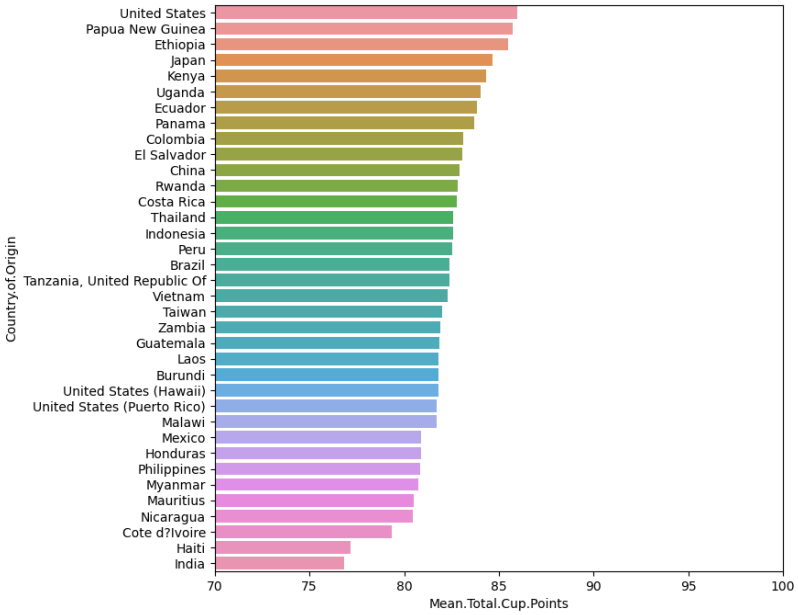

Figura 3. Promedio de la variable de interés respecto al país de origen del café

Para realizar un análisis más preciso se tuvieron en cuenta los valores faltantes en las variables, obteniendo que los porcentajes de valores perdidos más altos los tiene la altitud media, alta y baja, seguido por el color, la variedad y el método de procesamiento, en variables con valores perdidos también se incluyen, aunque en menor medida, la región, el año de cosecha y quakers o imperfecciones en el grano de café (Figura 4).

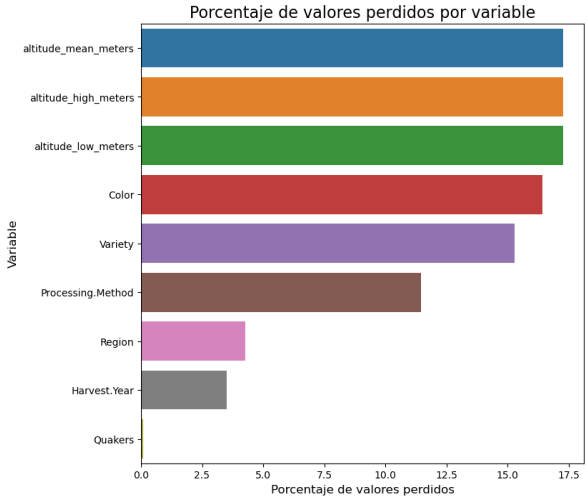

Figura 4. Porcentaje de valores perdidos por variable

En este caso para solucionar el problema se reemplazaron los datos faltantes por promedio según cada país de origen, de esta misma forma se reemplazaron los datos faltantes en variables como variedad, método de procesado y color, sin embargo, para los países en los que el promedio no es posible por falta de datos se reemplaza la por la moda del data set en su campo particular. Para el año de

cosecha se tomó un promedio de los años presentes y se aproximaron a su entero más cercano.

Para el preprocesamiento lo primero que se realizó fue organizar los datos de acuerdo a un formato estándar de año, esto debido a que algunos datos estaban en un formato diferente. Luego, se discretizaron los valores de países de origen y regiones obteniendo los resultados mostrados en la figura 5.

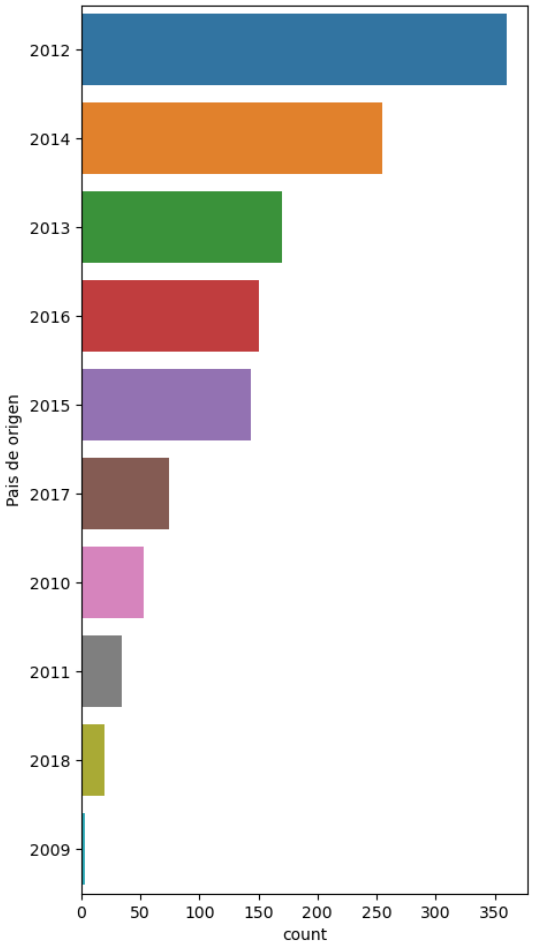

Figura 5.

Lo siguiente que se realizó fue una matriz de correlaciones, esto se hizo con el fin de observar qué tan relacionadas estaban cada par de variables, estos valores de correlación se ubican entre –1 y +1 y el valor de correlación es positivo si las dos variables tienden a aumentar o disminuir al mismo tiempo. También se analizó la correlación entre cada variable y nuestra variable de interés que es “Total Cup Points” obteniendo que la mayoría de las variables tenían una alta relación (Figura 6) destacándose el sabor, el regusto y el balance, mientras que, la altitud media, alta y baja y el método de procesamiento tienen poca correlación y quakers y el color al tener una correlación negativa indica que los dos valores tienen tendencias contrarias.

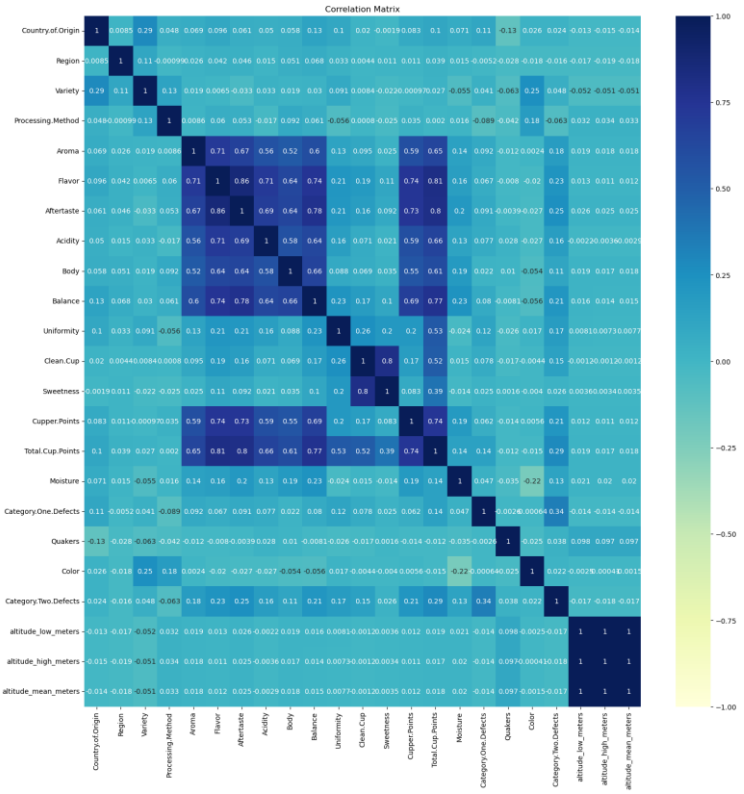

Figura 6. Matriz de correlación

Luego de esto las filas son desorganizadas y los datos son divididos con la finalidad de tener cierta cantidad para realizar el entrenamiento y el resto para realizar las pruebas.