TRABAJO PRÁCTICO N° 3

MINERIA DE DATOS

2023

INTEGRANTES:

- PONCE, DANIEL
- YAÑEZ, MIRIAN

TRABAJO PRÁCTICO N° 3

Minería de Datos

Investigación sobre el tema:

Indagando un poco acerca del tema encontramos este sitio web con la siguiente información:

https://primerocafe.com.mx/mundo-barista/10-virtudes-cafe-califican-cata/

Los atributos de un café dan voz a la historia que hay detrás de él: su especie, la tierra donde fue cultivado, sus condiciones de almacenamiento, su proceso de tueste. Esos elementos se conjugan para dar origen a un universo de sabor y aromas que se juzgan en una cata para determinar la calidad del grano.

De acuerdo con los protocolos de catación de la Specialty Coffee Asociation, la calidad de un café se determina tras evaluar **10 categorías que pueden alcanzar un puntaje del 10 al 100**.

A partir de 80 puntos se considera café de especialidad. Abajo de esa cantidad se cataloga como café común o comercial.

Categorías a evaluar en una cata

1. Fragancia y aroma

Se evalúa en tres momentos: el olor del café molido fresco antes de añadir el agua (fragancia), el olor del café cuando se mezcla con agua (aroma) y el olor al romper la costra formada en la superficie con la cuchara de cata.

2. Sabor

Al sorber el café se califican principalmente las notas que aparecen entre la primera impresión olfativa y las que se perciben en el retrogusto final. Un truco para detectar la riqueza aromática es sorber la infusión tapándose la nariz. Después, ya con la nariz libre, se vuelve a dar un sorbo para notar mucho más sabor.

3. Retrogusto

Se trata de cuántos minutos o segundos permanece en la boca el sabor del café después de ser ingerido. Si su duración es fugaz o deja sensaciones desagradables se califica con un puntaje bajo.

4. Acidez

Evalúa la nitidez de la bebida. Por lo general recibe el calificativo de “brillante” cuando la calificación es positiva y de “agria” cuando es negativa. Se juzga que la acidez sea balanceada y tenga estos cuatro componentes: intensidad; jugosidad, relacionada con la salivación producida; dulzor, referente al tipo de acidez (limón, mandarina, naranja) y brillo, el cual es la sensación agradable en el paladar.

5. Cuerpo

Es la densidad de la bebida. Las sensaciones de peso y volumen que el café deja en la boca. Se identifica como intenso o ligero.

6. Balance

Es la combinación del sabor, regusto, acidez y cuerpo. La manera en que estos elementos crean una bebida armónica y compleja. Si presenta una escasa cantidad de los atributos esperados según su especie o tipo de tueste; o bien, si un sólo atributo resalta de manera exagerada, la bebida recibe una baja puntuación.

7. Dulzor

Es la percepción de la boca que deja la presencia de algunos carbohidratos que contiene el grano. Se analiza que la bebida no sea agria, astringente o amarga para obtener una alta calificación.

8. Limpieza

Es la falta de impresiones negativas en el café infusionado, desde el primer sorbo hasta el retrogusto final. Se reprueba la infusión que presenta sabores ajenos.

9. Uniformidad

Es la consistencia de las propiedades que se detectan con los sentidos a través de las distintas tazas que se prueban de la misma muestra. Se califica que los cafés de una misma variedad sean uniformes. Si tienen cualidades o dejan sensaciones muy diferentes obtienen pocos puntos.

10. Impresión general

Es la apreciación de cada catador sobre la bebida. La calificación global que otorga de acuerdo a qué tan agradable o desagradable le resultó la muestra que degustó. Un café que refleja las cualidades de su origen y tueste recibe una alta valoración.

Análisis del conjunto de datos (distribuciones, valores, outliers, tipos de datos, etc.)

Visualizamos los datos y los nombres de las variables:

```
     Scores_Aroma  Scores_Flavor  Scores_Aftertaste  Scores_Acidity  ...  Scores_Sweetness  Scores_Moisture  Scores_Total         Color
0              85             85                 80              80  ...               100               12          8692         Green
1              85            817                 80             775  ...               100               12          8642         Green
2             833             80                 80              80  ...               100               11          8608    Blue-Green
3              80             80                 80             767  ...               100               11          8542    Blue-Green
4              80            792                775             775  ...               100               11          8492         Green
..            ...            ...                ...             ...  ...               ...              ...           ...           ...
830           758             70                675             692  ...               100               11          7917         Green
831           758            767                742             742  ...               867                1          7908         Green
832             0              0                  0               0  ...                 0               12             0         Green
833           767            775                783             767  ...               792                1           825  Bluish-Green
834            75            742                708             742  ...               100               11          8158    Blue-Green

[835 rows x 11 columns]
lista de columnas
Index(['Scores_Aroma', 'Scores_Flavor', 'Scores_Aftertaste', 'Scores_Acidity',
       'Scores_Body', 'Scores_Balance', 'Scores_Uniformity',
       'Scores_Sweetness', 'Scores_Moisture', 'Scores_Total', 'Color'],
      dtype='object')
```

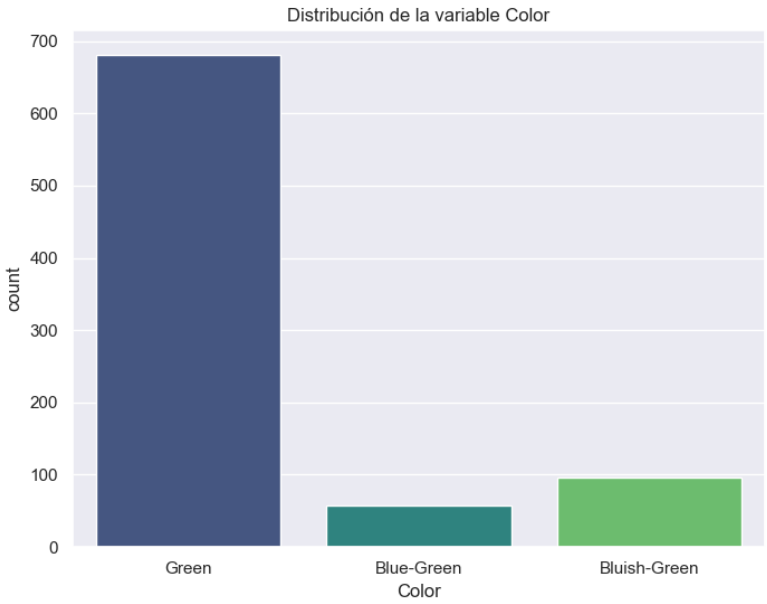
```
RangeIndex: 835 entries, 0 to 834
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Scores_Aroma       835 non-null    int64
 1   Scores_Flavor      835 non-null    int64
 2   Scores_Aftertaste  835 non-null    int64
 3   Scores_Acidity     835 non-null    int64
 4   Scores_Body        835 non-null    int64
 5   Scores_Balance     835 non-null    int64
 6   Scores_Uniformity  835 non-null    int64
 7   Scores_Sweetness   835 non-null    int64
 8   Scores_Moisture    835 non-null    int64
 9   Scores_Total       835 non-null    int64
 10  Color              835 non-null    object
dtypes: int64(10), object(1)
```

```
       Scores_Aroma  Scores_Flavor  Scores_Aftertaste  Scores_Acidity  ...  Scores_Uniformity  Scores_Sweetness  Scores_Moisture  Scores_Total
count    835.000000     835.000000         835.000000      835.000000  ...         835.000000        835.000000       835.000000    835.000000
mean     623.726946     615.576048         611.677844      617.116168  ...         191.758084        157.644311         8.231138   6675.440719
std      273.720152     274.245606         263.865747      274.480062  ...         257.583828        203.439529         5.130245   3007.519639
min        0.000000       0.000000           0.000000        0.000000  ...           0.000000          0.000000         0.000000      0.000000
25%      717.000000     708.000000         683.000000      717.000000  ...         100.000000        100.000000         1.000000   7792.000000
50%      758.000000     742.000000         733.000000      742.000000  ...         100.000000        100.000000        11.000000   8183.000000
75%      775.000000     767.000000         758.000000      767.000000  ...         100.000000        100.000000        12.000000   8325.000000
max      875.000000     883.000000         867.000000      875.000000  ...         933.000000        933.000000        17.000000   9058.000000
```

El conjunto de datos consta de 835 registros y 11 variables, con 10 de ellas siendo variables numéricas (de tipo int64) y 1 variable categórica (de tipo object). Un aspecto destacado es la ausencia de datos nulos en el conjunto.

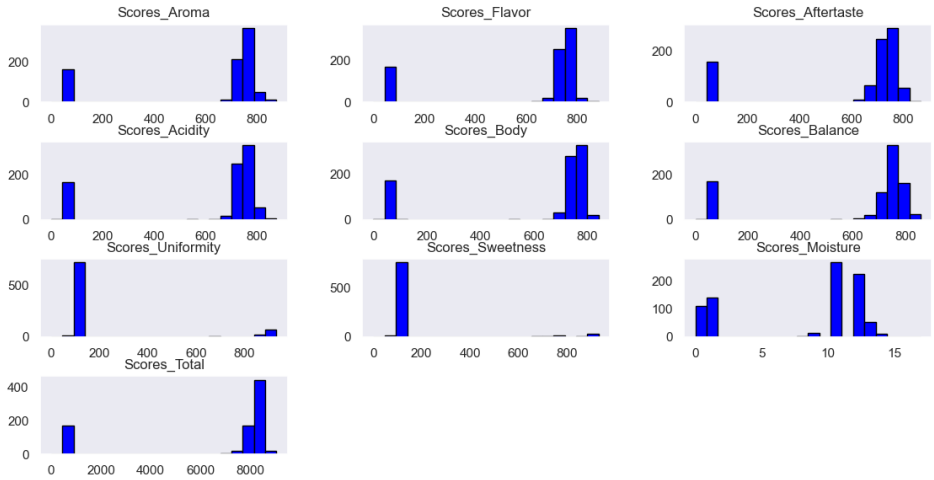
Al profundizar en el tema, identificamos que los valores de las variables deben encontrarse en el rango de 10 a 100, según la información disponible. En línea con esta observación, optamos por realizar una depuración de datos para filtrar aquellos que se encuentren fuera de este rango. Un ejemplo de esta acción es la eliminación de la fila con índice 832, donde todos los valores son igual a cero (0). Esta fila se percibe como atípica o como una entrada errónea, ya que la presencia exclusiva de valores numéricos nulos no parece tener coherencia en el contexto del conjunto de datos que evalúa la calidad del café. Eliminar estos datos atípicos contribuirá a mejorar la calidad y confiabilidad del conjunto de datos para análisis y modelado subsiguientes.

Distribución de la variable Color:



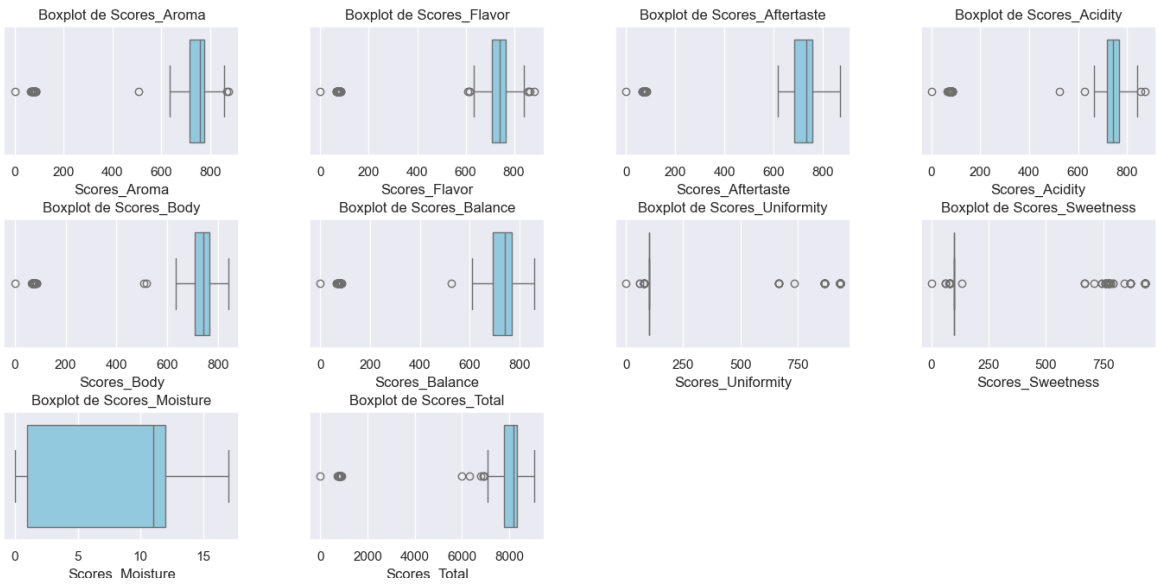
Se puede observar que el dataset está desbalanceado, hay una gran cantidad de datos de color green, alrededor de 680, mientras que de Blue-Green apenas llega a 50 y Bluish-Green a 100.

Distribución de las variables numéricas:



En la representación gráfica de los datos, se evidencia la presencia de dos picos prominentes en los extremos.

Boxplot para visualizar outliers:



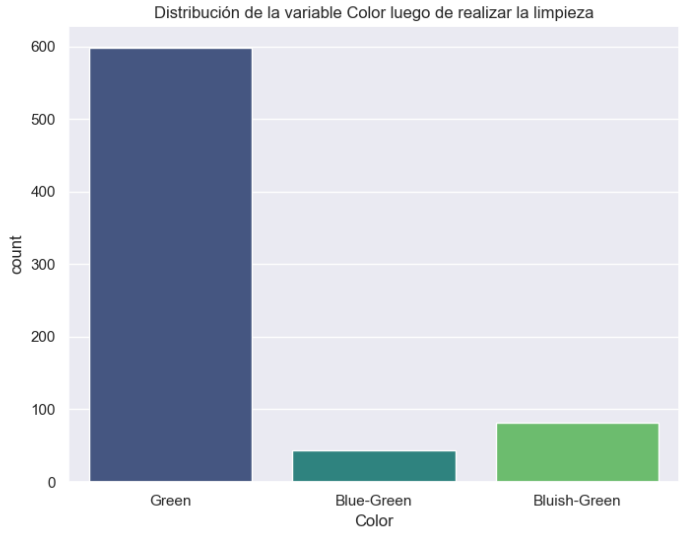
La visualización revela una marcada presencia de valores atípicos en la mayoría de las columnas, a excepción de la columna Scores_Moisture.

Realizamos una limpieza a partir de esto y vamos a visualizar nuevamente como quedan nuestros datos:

```
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Scores_Aroma        722 non-null    float64
 1   Scores_Flavor       722 non-null    float64
 2   Scores_Aftertaste   722 non-null    float64
 3   Scores_Acidity      722 non-null    float64
 4   Scores_Body         722 non-null    float64
 5   Scores_Balance      722 non-null    float64
 6   Scores_Uniformity   722 non-null    float64
 7   Scores_Sweetness    722 non-null    float64
 8   Scores_Moisture     722 non-null    int64
 9   Scores_Total        722 non-null    int64
 10  Color               722 non-null    object
 11  Color_Blue-Green    722 non-null    bool
 12  Color_Bluish-Green  722 non-null    bool
 13  Color_Green         722 non-null    bool
dtypes: bool(3), float64(8), int64(2), object(1)
```

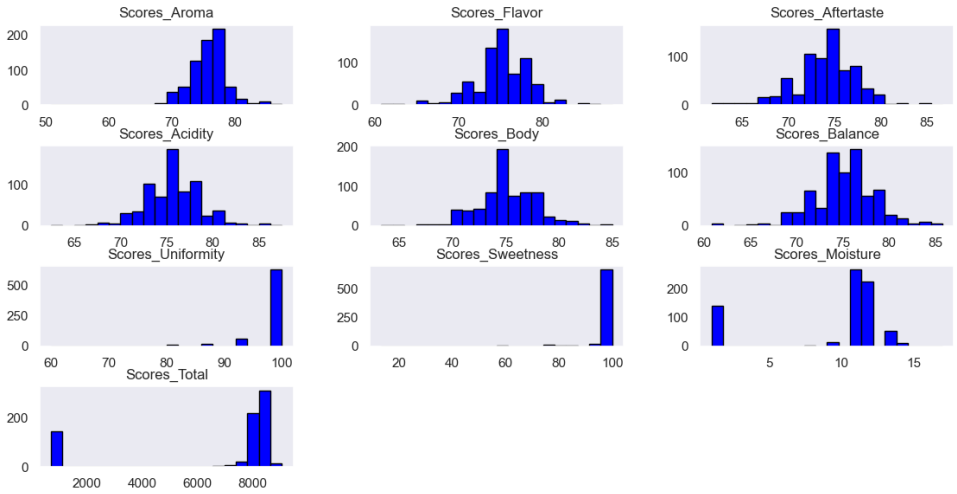
Después de la transformación de la variable categórica "Color" en variables dummy, y habiendo realizado la limpieza necesaria, observamos que el conjunto de datos resultante consta de 722 filas y 14 columnas.

Distribución de la variable Color luego de realizar la limpieza:

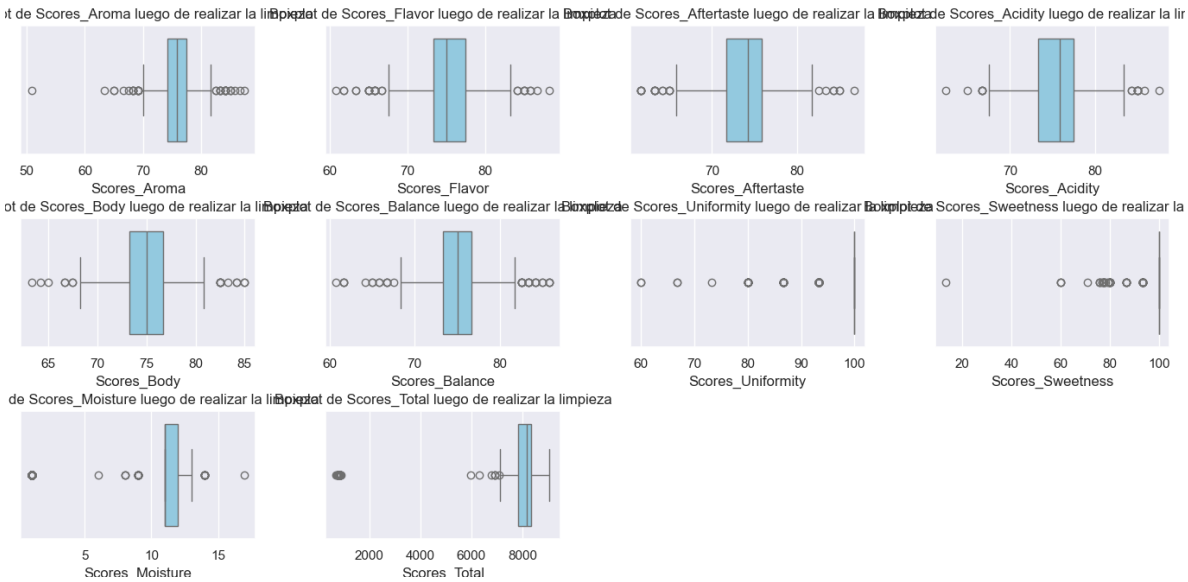


Luego de la limpieza se puede observar que el dataset continúa desbalanceado, la cantidad de datos de color green, bajo a 600, mientras que de Blue-Green continua alrededor de 50 y Bluish-Green por debejo de 100.

Distribución de las variables numéricas luego de realizar la limpieza:



Boxplot para visualizar outliers luego de realizer la limpieza:



Se puede observar cómo mejoró la distribución de los datos, asemejándose más a una distribución normal.

SVM LINEAL

```
SVM LINEAL
Resultados de la validación cruzada (k=5) para C=0.01 (SVM Lineal):
Precisión promedio: 0.8275862068965518
Precisión por partición: [0.82758621 0.82758621 0.82758621 0.82758621 0.82758621]
Exactitud: 0.9448275862068966
Exhaustividad: 0.9448275862068966
Precisión: 0.8931034482758621
```

Parámetro C utilizado: 0.01

Resultados de la validación cruzada (k=5):

- Precisión promedio: 82.76%
- Precisión por partición:82.76%
- Exactitud: 94.48%
- Exhaustividad: 94.48%
- Precisión: 89.31%

SVM GAUSSIANO

```
SVM GAUSSIANO
Resultados de la validación cruzada (k=5) para C=0.01, gamma=scale (SVM Gaussiano):
Precisión promedio: 0.8275862068965518
Precisión por partición: [0.82758621 0.82758621 0.82758621 0.82758621 0.82758621]
Exactitud: 0.8275862068965517
Exhaustividad: 0.8275862068965517
Precisión: 0.6848989298454221
```

Parámetros utilizados: C=0.01, gamma=scale

Resultados de la validación cruzada (k=5):

- Precisión promedio: 82.76%
- Precisión por partición: 82.76%
- Exactitud: 82.76%
- Exhaustividad: 82.76%
- Precisión: 68.49%

RANDOM FOREST

```
RANDOM FOREST
Resultados de la validación cruzada (k=5) para n_estimators=200, max_depth=1 (Random Forest):
Precisión promedio: 0.8413793103448276
Precisión por partición: [0.82758621 0.82758621 0.86206897 0.82758621 0.86206897]
Exactitud: 0.8275862068965517
Exhaustividad: 0.8275862068965517
Precisión: 0.6848989298454221
```

Parámetros utilizados: n_estimators=200, max_depth=1

Resultados de la validación cruzada (k=5):

- Precisión promedio: 84.14%
- Precisión por partición: 82.76%, 82.76%, 86.21%, 82.76%, 86.21%
- Exactitud: 82.76%
- Exhaustividad: 82.76%
- Precisión: 68.49%

Conclusión:

SVM Lineal: El modelo muestra buenos resultados, con una alta precisión y exactitud. La exhaustividad también es alta, indicando que el modelo puede identificar correctamente la mayoría de las instancias positivas. La precisión es un poco más baja, lo que podría sugerir la presencia de falsos positivos.

SVM Gaussiano: Aunque tiene una precisión promedio similar a la SVM lineal, la precisión, exactitud y exhaustividad son más bajas. Esto sugiere que este modelo podría tener un rendimiento ligeramente inferior en la clasificación de las clases.

Random Forest: Muestra resultados competitivos, con una precisión promedio ligeramente superior. La variabilidad en la precisión entre los pliegues de validación cruzada puede indicar cierta sensibilidad a la partición de los datos.

Para este caso, se observa que Random Forest tiene un rendimiento similar pero ligeramente superior en comparación con SVM Lineal y Gaussiano.