# **Universidad Nacional de Rosario**

# Facultad de Ciencias Exactas, Ingeniería y Agrimensura



Tecnicatura Universitaria en Inteligencia Artificial

## **Clasificación Supervisada y Métodos de Ensamble**

### 

### **TUIA - IA 4.3 Minería de Datos**

### **Trabajo Práctico Nro. 3**

***Fecha*:** 25/11/2023

***Docentes:*** ***Grupo:***

* *Flavio Spetale*   *- Revello Simon*
* *Dolores Biondo - Giampaoli Fabio*
* *Facundo Vasquez*
* *Sofia Errecarte*

***Objetivo:***

El objetivo de este trabajo práctico es integrar los conocimientos adquiridos en las unidades 5 (Clasificación supervisada) y 6 (Métodos de Ensamble) en un problema real asociado a la determinación del color de los granos de café mediante la medición atributos característicos

***Actividades:***

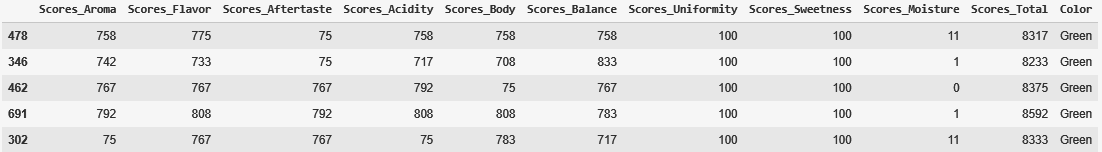
1. Descargar el conjunto de CoffeeRatings.csv, para realizar el trabajo práctico. Analizar los atributos del conjunto de datos (distribuciones, valores, outliers, tipos de datos, etc.).
2. Realizar la predicción del atributo Color utilizando máquinas de vectores con kernel lineal analizando el parámetro costo. Mostrar los resultados sobre los conjuntos de test (Precisión, Exhaustividad y Exactitud) utilizando validación cruzada con k =5.
3. Realizar la predicción del atributo Color utilizando máquinas de vectores con kernel gaussiano analizando los parámetros costo y gama. Mostrar los resultados sobre los conjuntos de test (Precisión, Exhaustividad y Exactitud) utilizando validación cruzada con k=5.
4. Realizar la predicción del atributo Color utilizando Random Forest analizando los parámetros cantidad de estimadores y la máxima profundidad de los árboles. Mostrar los resultados sobre los conjuntos de test (Precisión, Exhaustividad y Exactitud) utilizando validación cruzada con k =5.

***Resolución***

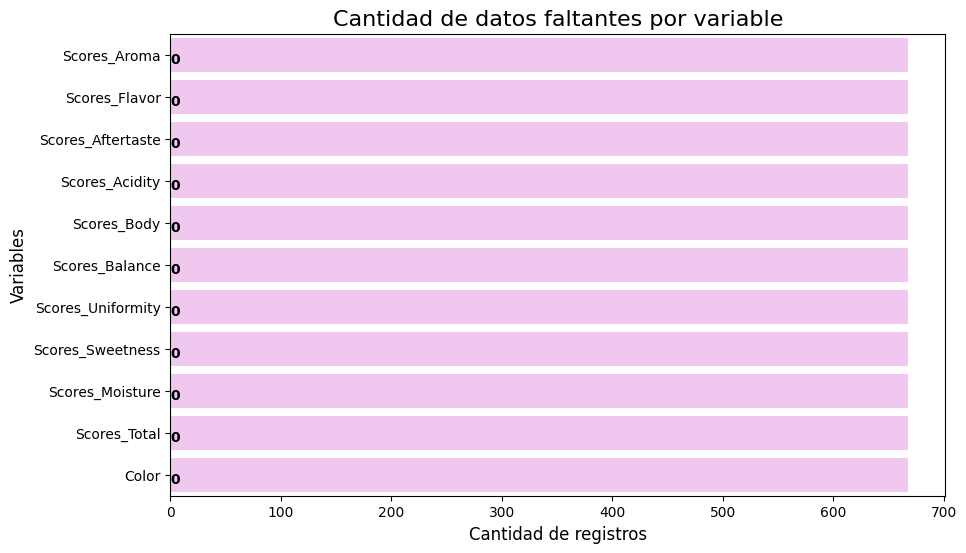
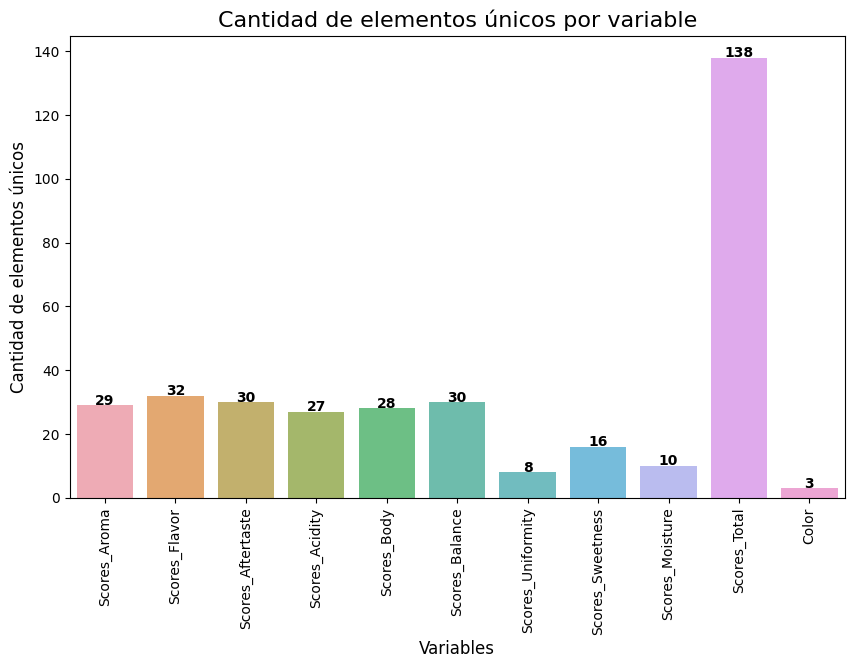
El dataset cuenta inicialmente con un total de 835 filas y 11 columnas. Cada columna representa una variable numérica discretaque representan puntajes de cada característica de un tipo de cadena particular. Pero también posee una variable categórica de tres clases, ‘Color’, que representa el color del grano del café.

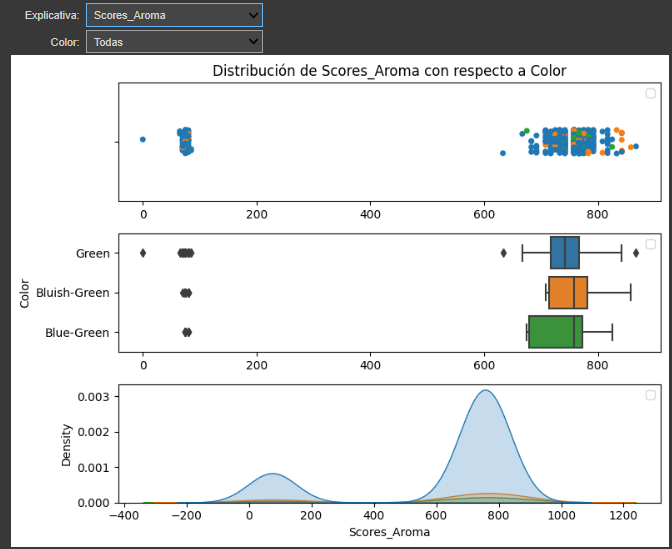
Planteamos la división inicial del conjunto de datos, con el fin de no perturbar el conjunto de validación que deberíamos tratar como si no lo conociéramos de antemano, para que el modelo de clasificación no se contamine con datos que todavía no conoce. Por ello, se dividen las variables explicativas y la variable objetivo (Color), y de ellos se toma el 80% de los registros para entrenamiento del modelo, y el restante para validación de las predicciones.

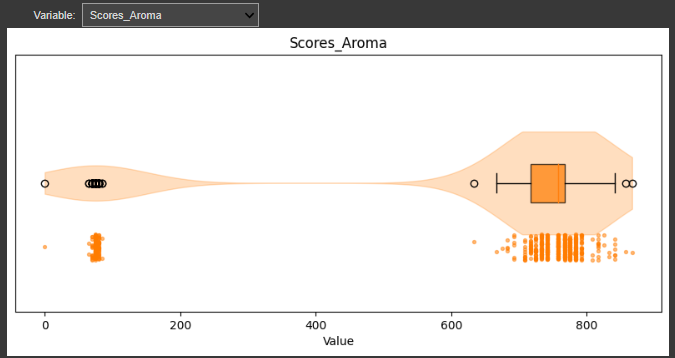
Esta es una muestra del conjunto de entrenamiento (concatenación de X e y):

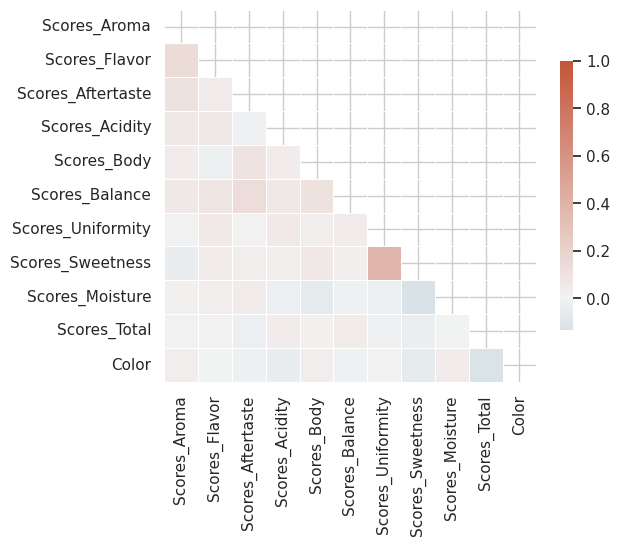


Notamos primeramente que los datos están pre procesados, ya que no hay valores nulos en el dataset (imagen izquierda). Además podemos notar las proporciones de valores únicos de cada variable (imagen derecha)::

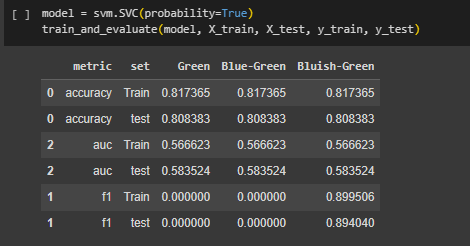


Como parte del análisis descriptivo, podemos comprender la distribución de las variables y de cada clase con un gráfico interactivo que permite la elegir la variable de interés a lo largo del eje x, y la o las clases que se desean visualizar:  


Notamos en su mayoría que las variables cumplen esta distribución de muchos valores concentrados en valores altos o bajos, y algunos grupos de registros en el lado extremo, pero en las tres clases a la vez. Esto lo notamos mejor en un distribuciones de las variables considerando todas las clases, nuevamente en un gráfico interactivo:  
  
Esto sugiere que tendremos que hacer un tratamiento o consideraciones para los valores atípicos, debido a que su presencia perturba las distribuciones de las variables, y posiblemente las estimaciones del modelo de clasificación que no son robustos.

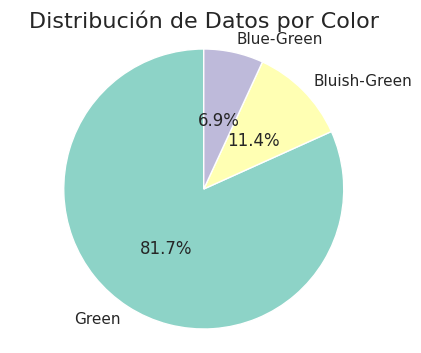


Una matriz de correlación indica la una baja correlación lineal entre todas las variables en general.



Para controlar que las transformaciones del dataset resulten eficientes y maximicen la explicabilidad de la variable objetivo, se crea un modelo base de clasificación con Soporoso Vector Machines con sus parámetros por default. De esta manera cada conjunto que se le pasa por parámetro obtiene métricas de validación para clasificación.

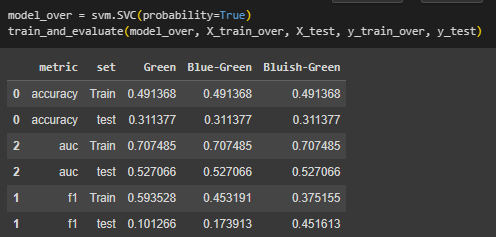
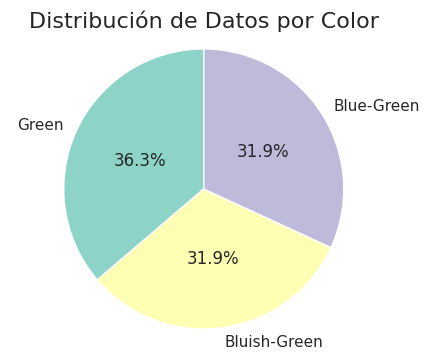
La imagen de la derecha muestra métricas base para el dataset sin procesamiento.



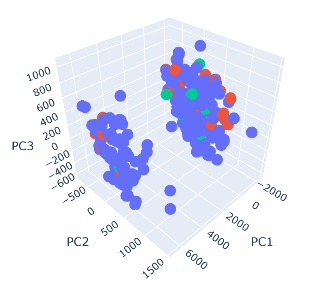
Notamos principalmente que la combinación de la precisión y exhaustividad (f1) es muy grande para solo una de las clases, y nula para las restantes. Esto sugiere que hay un gran desbalance de las clases en la variable objetivo. Esto lo podemos notar rápidamente en un gráfico de proporciones.

Tratar el desbalance de las clases es importante para que el modelo de clasificación no tome decisiones sesgadas para predecir debido a la falta de entrenamiento en las clases minoritarias. Se propone la utilización de SMOTE para balancear el dataset mediante oversailing por cercanía de datos, aumentando el número de datos para las clases minoritarias a un número dado.

Pero en esta instancia, el balance no es favorable debido a que disminuye las métricas del modelo base, y además causa sobreajuste en los datos. Algo similar ocurre con la técnica de undersampling para balancear.

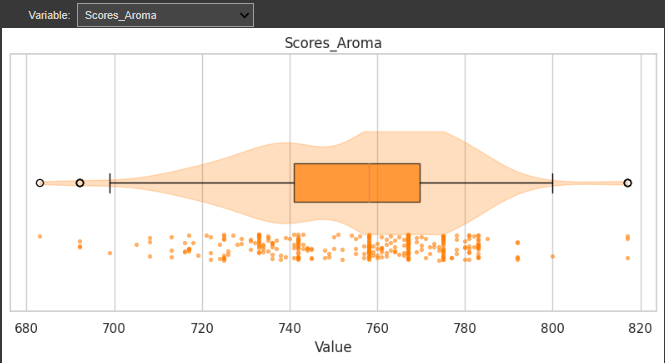


Por lo que se evita continuar este enfoque de tratar el desbalance.



Mediante tecnica de reduccion de dimensionalidad lineal como PCA, se pueden visualizar las distribucion de los datos, donde los diferentes colores representan las clases de Color.

Notamos nuevamente presencia de valores atípicos pertenecientes a las tres clases, y que parecen manejar un rango de valores particular que los distancia del resto.

Para el tratamiento manual de valores atípicos se usa un enfoque manual de tratarlos, mediante la simple eliminación de los mismos.Esto nos deja con distribuciones de variables más uniformes:  


Pero las métricas del modelo base disminuyen fuertemente, por lo que resolveremos este enfoque de tratamiento de valores atípicos mediante parametrización de los modelos, en lugar de tratarlos directamente los datos.

***Conclusiones***