-

# íNDICE

[1. INTRODUCCIÓN](#_heading=h.3zc1r1s7qdym)

[1.1. Presentación del equipo de trabajo](#_heading=h.ffg961xxbs2g)

[1.2. Temática](#_heading=h.745o4poj1sgd)

[1.3. Problemática](#_heading=h.fkkbw2tvlvsd)

[1.4. Objetivo](#_heading=h.mcnmx3wiyxt4)

[1.5. Hipótesis](#_heading=h.bqx1b2bkrlcn)

[1.6. Alcance](#_heading=h.eb0t9d3lb5pi)

[1.7. Stack tecnológico](#_heading=h.xqsrp1ybtsod)

[1.8. Tabla de versionado](#_heading=h.a83z02hhj41n)

[2. BASE DE DATOS](#_heading=h.6zp8tm17yk6m)

[2.1. Dataset](#_heading=h.600xt940pqjb)

[3. SECCIÓN COLABORATIVA](#_heading=h.99ra7f2u2guv)

[3.1. Data wrangling](#_heading=h.c76v5ns30shf)

[3.2. Análisis Univariado](#_heading=h.ukrc1nvnatyf)

[3.3. Análisis Bivariado](#_heading=h.2lc86uskvhdw)

[3.4. Análisis Multivariado](#_heading=h.erpye2vqyhhx)

[3.5. Algoritmos de entrenamiento](#_heading=h.unqa2aj1xytf)

[3.5.1. Algoritmos de clasificación](#_heading=h.jude4sqmcgo)

[3.5.1.1. Decision Tree](#_heading=h.n86200w3puyo)

[3.5.1.2. Random Forest](#_heading=h.7sevp1qxlo5n)

[3.5.1.3. KNN](#_heading=h.mmjbimwjhyk)

[3.5.1.4. Conclusión de Algoritmos de Clasificación](#_heading=h.8wr7bls7n650)

[3.5.1.5. Algoritmo adecuado](#_heading=h.tvjlx9h7s1mt)

[4.CONCLUSIÓN FINAL](#_heading=h.u8zsd1wxru9i)

# 1. INTRODUCCIÓN

## 1.1. Presentación del equipo de trabajo

La consultora “*CoderData*” está conformada por:

* **Pedro Di Liscia:** Hidrogeólogo de la Universidad Nacional de La Pampa.
* **Oscar Nieto:** Ingeniero de datos en Datalytics.
* **Alfredo Parente:** Analista de Datos Calidad Interna en Claro.
* **Sofía Vila:** Analista financiero en Finca Flichman.

## 1.2. Temática

Una empresa internacional de distribución de vinos con origen italiano llamada “VinTernazionali” presenta una disminución lenta pero sostenida de sus ventas desde hace 10 años.

Luego de realizar una auditoría interna, notaron que la causa de esta disminución es que el sistema de clasificación de vinos que daban a sus clientes no era adecuado, ya que solamente se basaba en la relación calidad-precio. Ese sistema había sido eficiente en el pasado, pero actualmente con la aplicación de nuevas tecnologías, cambios metodológicos en la industria y aparición de nuevas bodegas, la mayoría de los vinos no puede ser clasificado solamente utilizando estas dos variables.

Es por esa razón que actualmente la empresa está intentando aplicar un nuevo sistema de clasificación de sus vinos para retomar el vínculo con sus clientes.

La empresa nos provee un dataset ([https:// www.kaggle.com/ zynicide/ wine- reviews ?select= winemag -data -130k-v2.csv](https://www.kaggle.com/zynicide/wine-reviews?select=winemag-data-130k-v2.csv)) el cual contiene todas las reviews realizadas a los vinos que ellos comercializan.

La consultora “*CoderData*” trabajará con el dataset provisto e intentará generar ese nuevo sistema de clasificación solicitado por la empresa.

## 1.3. Problemática

El cliente sostiene que la relación calidad precio ya no es un parámetro viable para la clasificación rápida de los vinos, y en consecuencia de eso, sus ventas disminuyen año a año.

## 1.4. Objetivo

Crear un nuevo sistema de clasificación rápida de vinos que pueda ser aplicado por el consumidor final al momento de la compra del producto. Es por ello que este sistema debe utilizar información visible en la etiqueta (bodega, año de cosecha, precio, etc) y no considerar parámetros físico-químicos (acidez, pH, etc) que para el consumidor final promedio, no tienen ningún significado. Adicionalmente, este sistema debe ser capaz de clasificar nuevos vinos que ingresen al catálogo de la empresa.

## 1.5. Hipótesis

Las hipótesis que se plantean para el trabajo son las siguientes:

* La relación calidad-precio es un buen sistema de evaluación rápida de los vinos.
* Los vinos de alta calidad son los más caros.
* El precio no es la única variable que define si un vino es bueno, regular o malo.

## 1.6. Alcance

El proyecto tiene un alcance internacional, destinado a la empresa distribuidora “VinTernazionali”.

## 1.7. Stack tecnológico

En el proyecto serán utilizadas las siguientes herramientas:

1. Google Colab
2. Jupyter Notebook
3. Paquete Office

## 1.8. Tabla de versionado

| **ENTREGA** | **FECHA** | **VERSIÓN** |
| --- | --- | --- |
| Primer Entrega del Proyecto Final | 21/03/2022 | 01.2022 |
| Segunda Entrega del Proyecto Final | 18/04/2022 | 02.2022 |
| Tercera Entrega del Proyecto Final | 04/05/2022 | 03.2022 |
| Cuarta Entrega del Proyecto Final | 18/05/2022 | 04.2022 |

# 2. BASE DE DATOS

## 2.1. Dataset

El Dataset empleado contiene:

* **Número de Columnas:** 19
* **Número de Filas:** 129971

Modelado del Dataset:

* Se procedió a concatenar 4 columnas accesorias que estaban en .csv que fueron divididas por presentar una "," + Tabulación:
  + Columna 1: 2956
  + Columna 2: 159
  + Columna 3: 12
  + Columna 4: 1
* Se procedió a modelar el dataset para poder obtener el dato de género de los Sommeliers para poder usar el algoritmo KNN para problemas de clasificación:
  + Se validó que en la columna taste\_name no tuviera datos NaN, vacíos o corruptos.
  + Se agregó una nueva columna en la que se diferencia el género del Sommelier.
  + Tras concatenar el dato se procedió a confeccionar la nueva columna "winery".
  + Se validó que no hubiera datos numéricos en las columnas adyacentes a “points” y “price”.
* Se procedió a mover un total de 1000 datos.
* Se usó Power Query, la opción de agregar una columna a través del ejemplo para poder traer el dato del año de producción del vino.
* Se validó que en la columna points no tuviera datos NaN, vacios o corruptos.
* Se generó una Tabla Dinámica para validar el recuento de cada puntaje.
* Se leyó reseñas en las que se obtuvo la conclusión que los puntajes que califican en su mayoría a los buenos vinos eran aquellos que tenían un puntaje de 93 o más.
* En base al último punto se generó un método de clasificación:
  + Los vinos que tuvieran un puntaje de 93 o más se les otorgó el valor 1
  + Los vinos que tuvieran un puntaje de 92 o menos se les otorgó un puntaje de 2
* Se generó una nueva columna con este puntaje denominada: "aprobó"

Datos NA:

* Country: 64
  + Se completa los registros con el top 12 países
* Price: 8996
  + =SI([@price]="";ALEATORIO.ENTRE(15;50);[@price])
* Taster\_name: 26284
  + Se completó con el top 15 tester.
* Year: hay 4415 que se generó una fórmula aleatoria.
  + Se procedió a limpiar el dato en las celdas corruptas.
  + Se aplica la fórmula =SI(@year=””;ALEATORIO.ENTRE(2006;2014);@year), el dato de los años se sacó a través de una Tabla dinámica para poder saber los años con mayor número de puntaje.

Descripción de los campos:

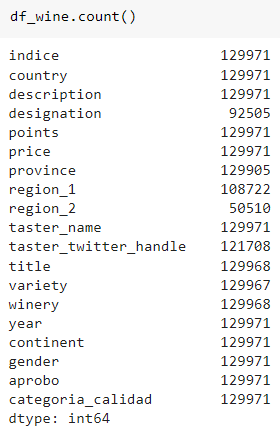
* **Country:** país de origen.
* **Description:** breve descripción del vino.
* **Designation:** viñedo de donde proviene el vino.
* **Points:** puntaje otorgado por el sommelier.
* **Price:** precio internacional del vino.
* **Province:** provincia de donde proviene el vino.
* **Region\_1:** provincia o estado de donde proviene el vino.
* **Region\_2:** región más específica de donde proviene el vino.
* **Taster\_name:** nombre del sommelier.
* **Taster\_twitter\_handle:** cuenta de twitter del sommelier.
* **Title:** etiqueta del vino.
* **Variety:** varietales.
* **Winery:** bodega.
* **Year:** año de producción.
* Continent: continente al que pertenece el país de origen del vino.
* Gender: género del sommelier que realizó el testing del vino.
* Aprobó: en base al puntaje que obtuvo el vino se divide en 1 (aprobó>92 puntos) y 2 (desaprobó<92 puntos).
* Categoría de calidad: En base al puntaje, se procedió a Clasificar el vino en 3 categorías: A (>=95 puntos), B (90-94,9 puntos), C (<89,9 puntos).

# 3. SECCIÓN COLABORATIVA

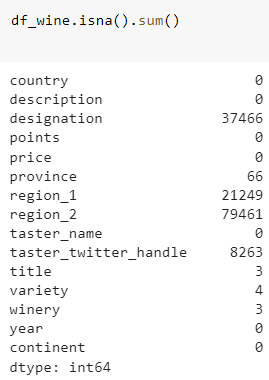
## 3.1. Data wrangling

Una vez importado el dataset, se creó el dataframe “df\_wine” y se obtuvo la siguiente información:

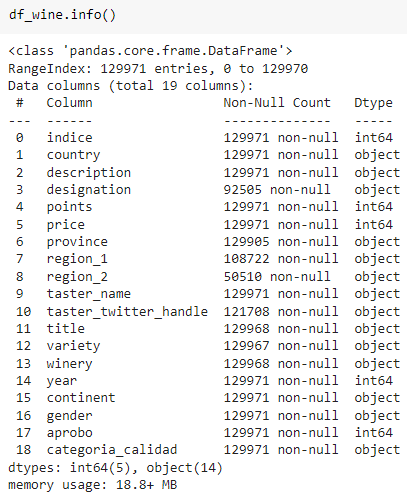
* El comando “.count()” cuenta la cantidad de datos que hay en una columna.



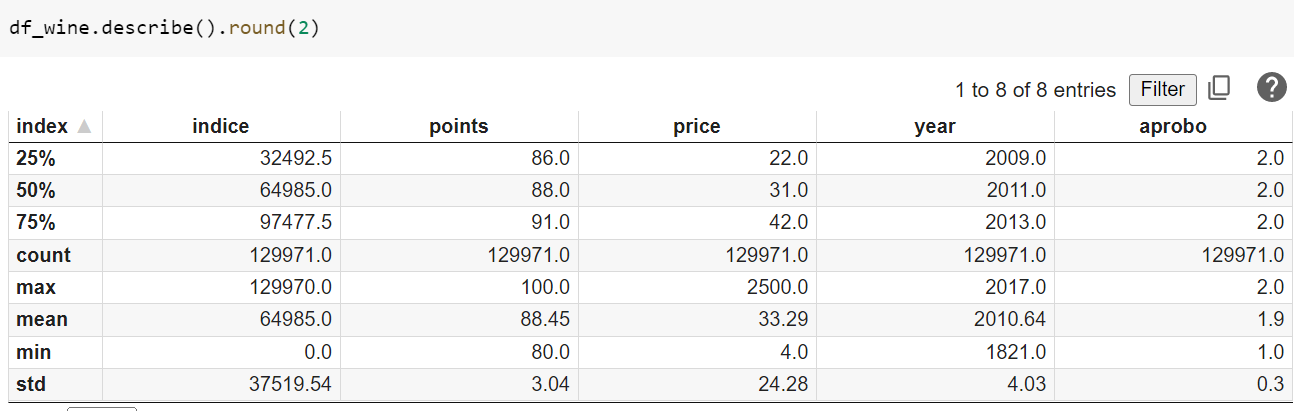
* Mientras que “.isna().sum()” nos dice la cantidad de datos nulos que hay en una columna.



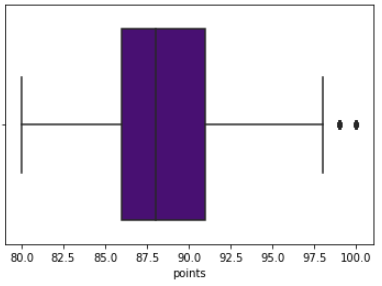
* Con “.info()” podemos observar un resumen de las características principales (cantidad de datos nulos y tipos de datos) del dataframe.



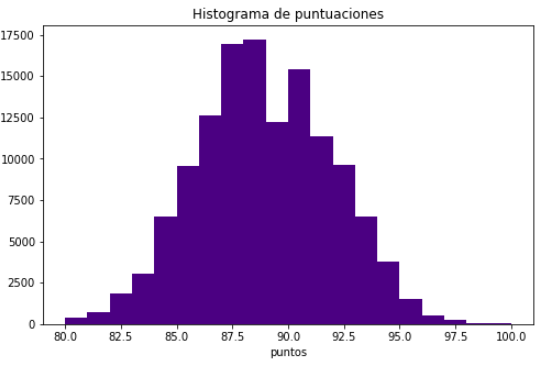
* Con “.describe” se puede ver el comportamiento estadístico de las variables numéricas del dataframe.



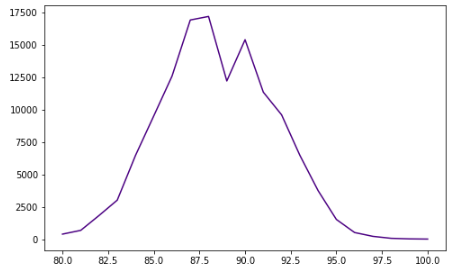
## 3.2. Análisis Univariado

Consiste en el análisis de cada una de las variables por separado, por lo que se basa exclusivamente en una única variable. 

En este boxplot observamos que el puntaje se concentra mayoritariamente en un rango entre 86 y 91 puntos.



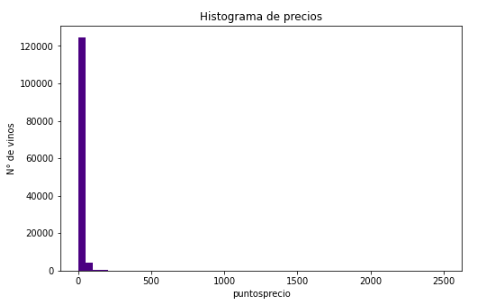
Este histograma nos confirma lo mostrado en el boxplot.



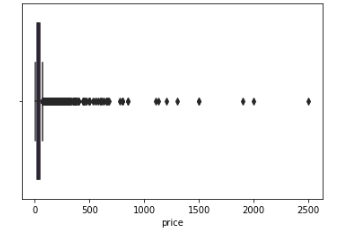
Otra forma de ver la distribución de los datos.



Este histograma nos muestra que la gran mayoría de los años de cosecha es posterior al año 2000

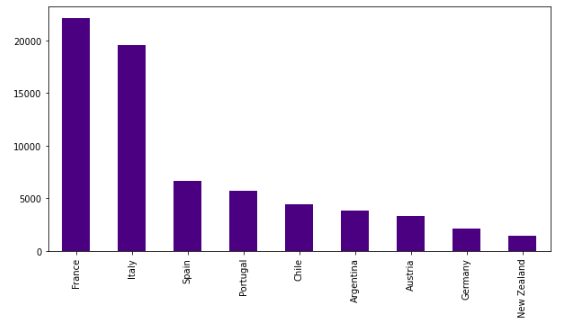


Observamos que la gran mayoría de los precios tienen un valor bajo.



En este boxplot observamos que el precio tiene una variabilidad muy grande.

El grueso de los precios son bajos, pero hay algunos que superan los 500 dólares.

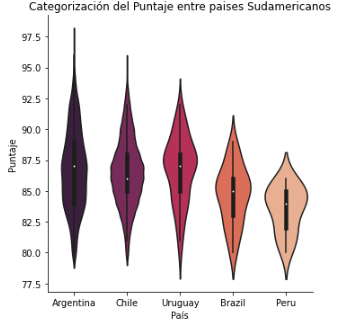


Aquí observamos el país de origen de los vinos que tienen más reviews.

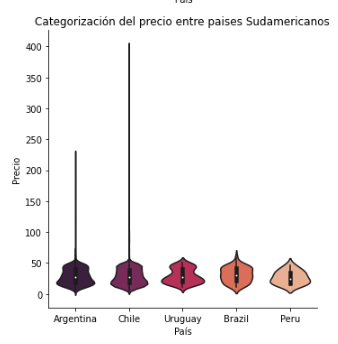
## 3.3. Análisis Bivariado

El análisis bivariado es el empleo de dos o más variables , buscando conocer causalidad, efectos o correlaciones.

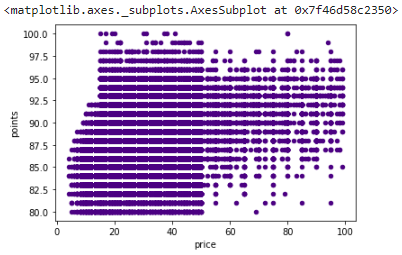
Comparamos los precios y puntajes de los vinos de los países de américa del sur en un gráfico de violín, se puede distinguir que la variabilidad del puntaje de los vinos argentinos es mayor y que los vinos de Perú tienen un puntaje máximo de 87.5

En lo que respecta a los precios, los vinos de Argentina, Chile y Uruguay tienen un comportamiento bimodal, los precios de Argentina y Chile cuentan con outliers, sin embargo, la media de los precios es similar en los cinco países.

Comparamos los puntajes con un gráfico de violín.

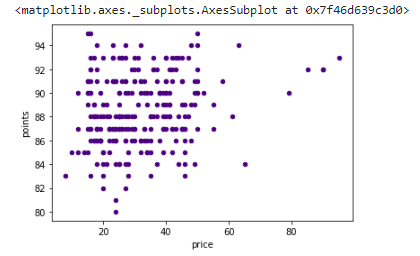


Es destacable que los precios en Argentina, Chile y Uruguay tienen un comportamiento bimodal



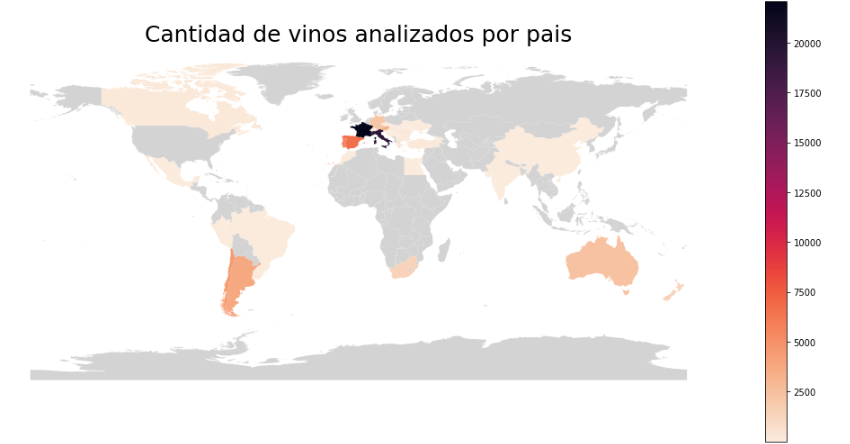
Hacemos scatter plot con las variables de puntos y precio (menor a 100 dólares).

La elevada cantidad de puntos no nos permite distinguir ningún rasgo



Haciendo el mismo gráfico pero con una muestra de 250 datos, se observa que en el extremo. Inferior izquierdo hay mayor densidad de datos, también se ve una leve correlación directa.

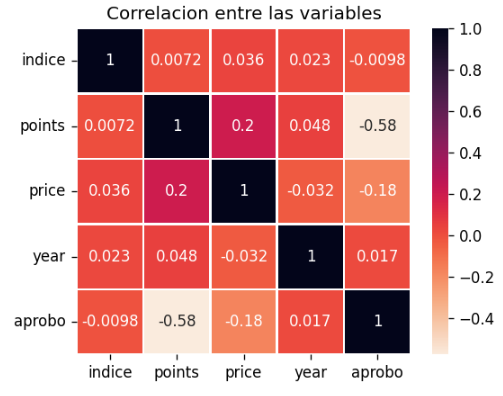
## 

Con un gráfico de dispersión hexagonal (hexbin) es posible notar que la gran mayoría de los vinos tiene puntaje cercanos a 87.5 y un precio que ronda los 20 a 40 dólares.

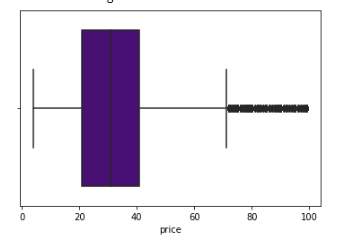
Generamos un mapa en donde vemos el origen de los vinos que tienen reviews.

## 3.4. Análisis Multivariado

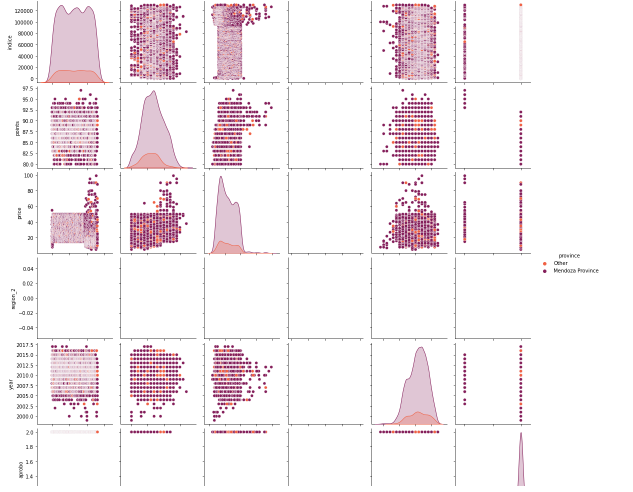
En este crosstab podemos observar los nombres de los catadores, el puntaje que otorgaron y que tan frecuentemente otorgan ese puntaje.



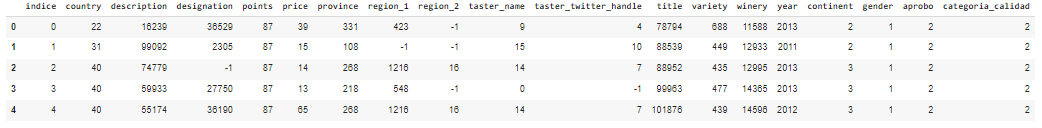
Generando una matriz de correlación se observa claramente que los valores tienen una baja de adecuación entre sí.



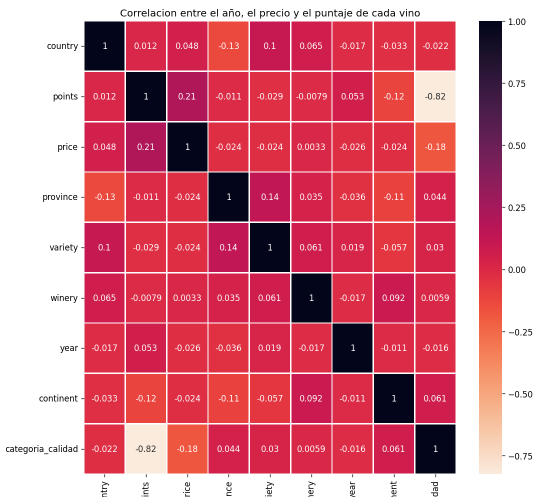
Decidimos fijar el precio máximo en 100 dólares creando el dataset “df\_wine\_100usd”, de esa manera eliminamos los precios más elevados (outliers) y analizamos de mejor forma los vinos que tienen un precio más accesible y llegan al público masivo.



Posteriormente, se crea el dataset “df\_wine\_100usd\_arg”, el cual contiene los vinos con precios más bajos a 100 dólares y provenientes de Argentina. Con ese dataset creamos un pairplot y diferenciamos por provincia de origen, en donde se observa que las variables no tienen una correlación entre sí.



Transformamos todo en variables categóricas para poder entrenar de mejor manera el modelo.



Se observa que la correlación entre las variables es muy baja

## 

## 3.5. Algoritmos de entrenamiento

Elegimos algoritmos de entrenamiento y preparamos los datos para el proceso de entrenamiento del modelo. También evaluamos los indicadores de desempeño predictivo del modelo y realizamos optimizaciones.

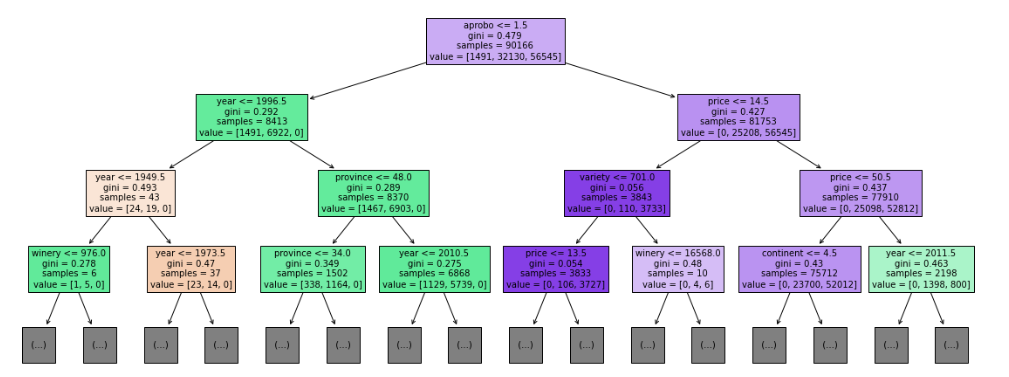
### 3.5.1. Algoritmos de clasificación

En base a los análisis de variables realizados previamente, quitamos del dataset las columnas irrelevantes que no aportan información al modelo. Las columnas que eliminamos son: indice, region\_1, region\_2, taster\_name, taster\_twitter\_handle, description, title, gender.

#### 3.5.1.1. Decision Tree

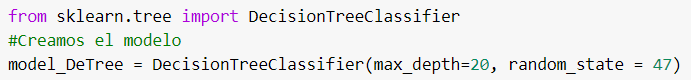
Los árboles de decisión son representaciones gráficas de posibles soluciones a una decisión basadas en ciertas condiciones, es uno de los algoritmos de aprendizaje supervisado más utilizados en machine learning y pueden realizar tareas de clasificación o regresión. Comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades. Esto le da una forma similar a la de un árbol.

Representación gráfica de nuestro árbol de decisión:



Resultados y optimización de Decision Tree:

Se corrió el algoritmo con las siguientes especificaciones



Obteniendo como resultado

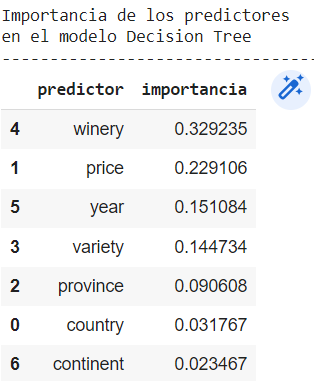


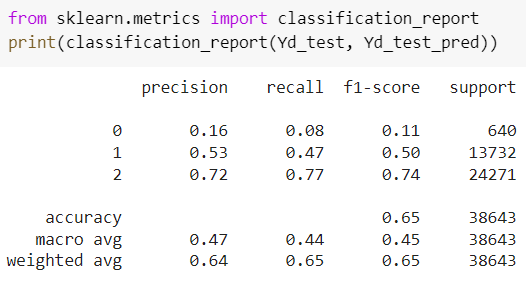
Posteriormente se realizó la optimización, y los mejores parámetros resultaron ser:



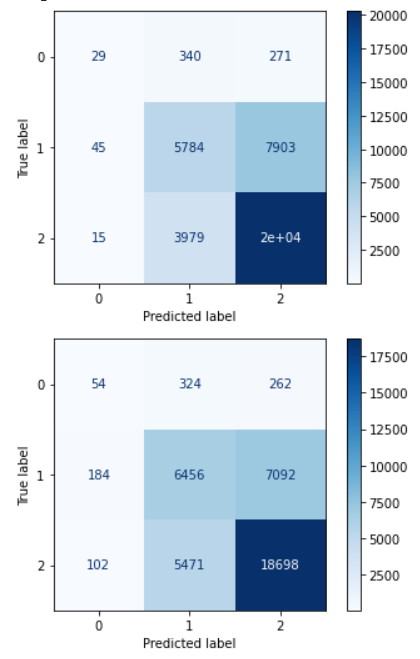
Se corrió el algoritmo con esos parámetros, y se obtuvo como resultado:







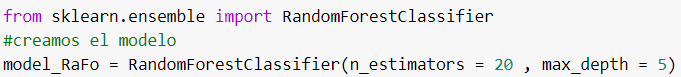
La matriz de confusión permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Realizamos las matrices de confusión del mismo algoritmo, optimizado y sin optimizar.



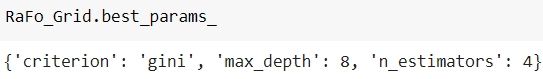
Conclusión: En la matriz de confusión se puede observar que el algoritmo clasifica de mejor manera los vinos de peor calidad.

#### 3.5.1.2. Random Forest

Para poder usar el algoritmo de Random Forest se modeló el dataset para obtener en base al puntaje que alcanzó el vino por parte del Sommelier una clasificación:

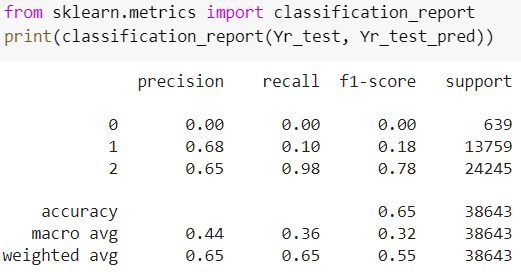


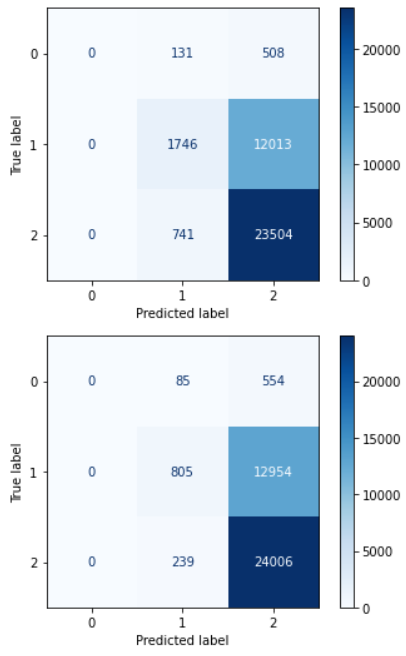








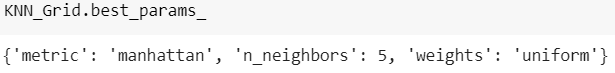




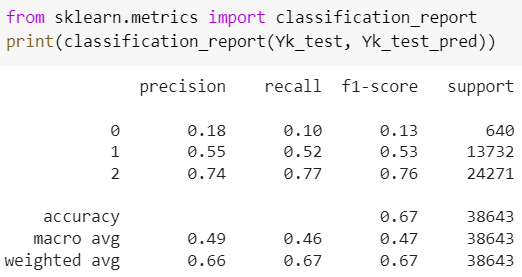
Conclusión: En la matriz de confusión se puede observar que el algoritmo clasifica de mejor manera los vinos de peor calidad..

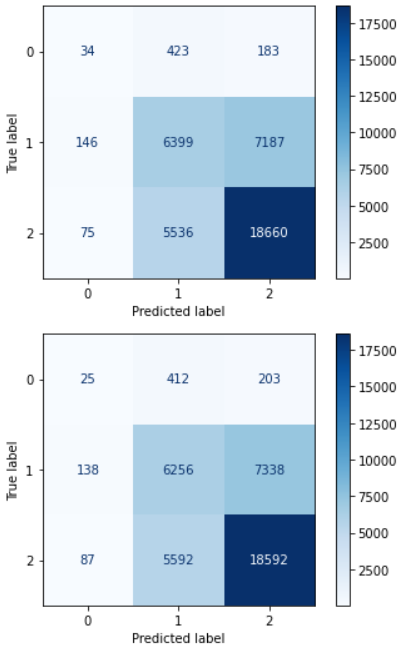
#### 3.5.1.3. KNN









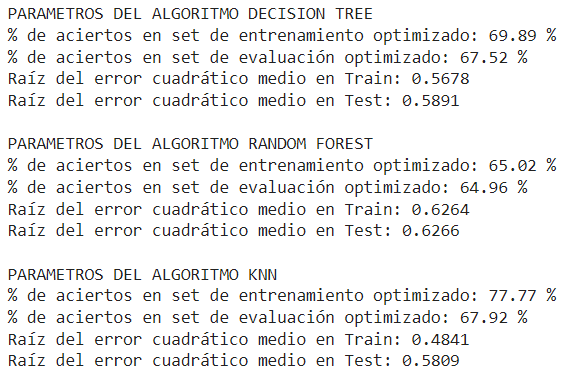


Conclusión: En la matriz de confusión se puede observar que el algoritmo clasifica de mejor manera los vinos de peor calidad.

#### 3.5.1.4. Conclusión de Algoritmos de Clasificación

Podemos ver cómo han mejorado los modelos con respecto al utilizado en el comienzo del preprocesado.

Si bien no hay tanta diferencia en el resultado entre los tres modelos utilizados (KNN, Decision Tree, Random Forest), podemos indicar que, el que arrojó mejor resultado fue el Decision Tree con Grid Search, que presente un RMSE bajo y, además no presenta Over o Under Fiting.



#### 3.5.1.5. Algoritmo adecuado

En un principio, nuestro enfoque fue realizar tres modelos diferentes esperando obtener resultados distintos entre los mismos. Una vez entrenados y optimizados, notamos que no hay tanta diferencia en los resultados de cada uno de ellos.

# 4.CONCLUSIÓN FINAL

Defendemos la idea de que no hay relación entre la calidad y el precio en un vino. Ya que, existen vinos muy buenos a muy buen precio y no tienen por qué costar más de 20 dólares para ser considerados excelentes en la relación calidad-precio. Como es el caso de:

