Métodos Cuantitativos y Simulación
Proyecto Final

Equipo Azul:

Juan Camilo Granados A01023421
Alberto Treviño González A00824493
David Sanchez Ortiz A01283105
Luis Yerik Arámbula Barrera A00825080

Etapa 1: Exploración

Descripción de los datos

Seleccionamos la base de datos de Vinos ya que es un tema que nos interesa y nos gustaría saber un poco más a detalle. Los datos que incluye son: Alcohol, Chlorides, Citric Acid, Density, Fixed Acidity, Free sulfur dioxide, PH, Quality, residual sugar, sulphates, total sulfur dioxide, volatile acidity. Estos datos conforman la estructura de un vino, por lo cual las analizaremos para ver cómo cada valor afecta en la calidad del vino.

Tabla 1.1 Variables de los datos y su descripción

Alcohol	El grado de alcohol del vino.
Chlorides	Los cloruros son el mayor contribuidor a la sal.
Citric Acid	La cantidad de ácido cítrico presente.
Density	Densidad en g/ml.
Fixed Acidity	Nivel de ácidos no volátiles, es decir, no se evaporan fácilmente
Free Sulfur Dioxide	La cantidad de SO2 “libre” para reaccionar químicamente. Esta sustancia cuenta con propiedades germicidas y antioxidantes.
PH	Medida de acidez o alcalinidad de una sustancia.
Quality	Calidad en una escala del 0 al 10.
Residual Sugar	Cantidad de azúcar restante, proveniente de las uvas después de fermentar.
Sulphates	El vino es fermentado usando levadura la cual desprende sulfato. El sulfato se usa para prevenir la oxidación del vino, puede afectar el color y el sabor, previene el crecimiento de microorganismos, mejora los compuestos que desprenden las uvas.
Total Sulfur Dioxide	La cantidad de SO2 total, contando el “libre”.
Volatile Acidity	La cantidad de ácido acético o ácidos gaseosos. En cantidades altas da un mal sabor.

Objetivo

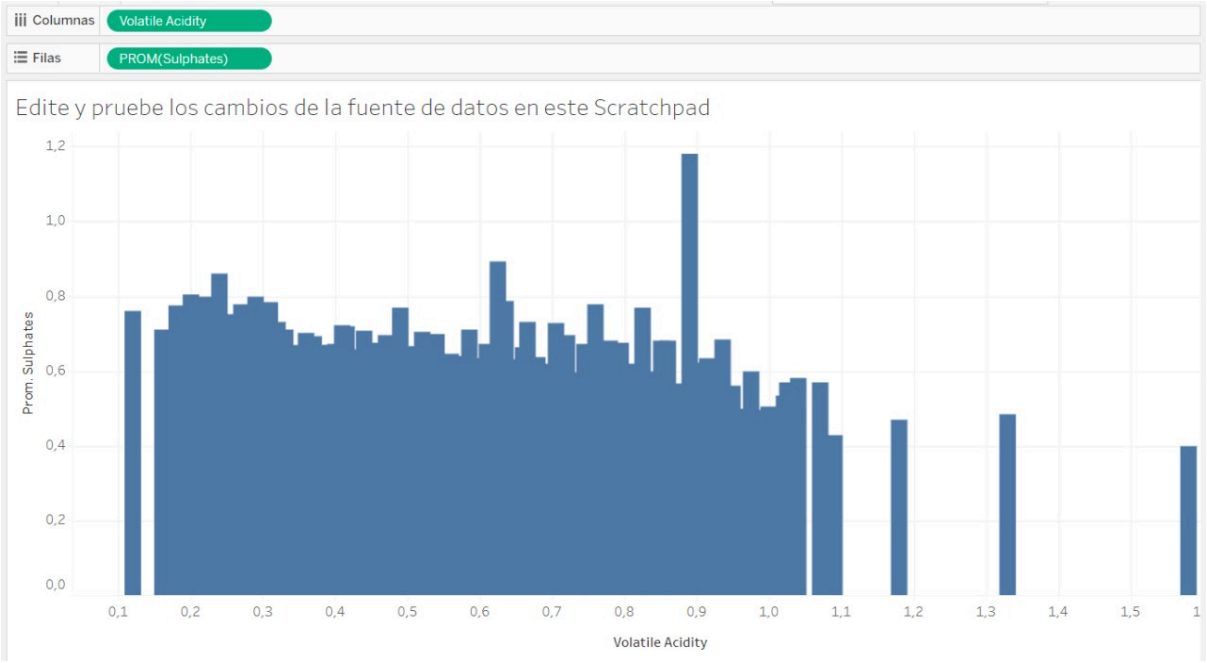
Nos interesa saber cómo ciertos factores afectan la calidad del vino, por ejemplo, nos gustaría saber la relación entre los cloruros con el azúcar residual, el alcohol con el ph, el volatile acidity con el sulphate. Nos intriga saber cómo reaccionan unos con los otros y qué posible relación pueden tener entre sí y en el vino.

Selección

Los datos que encontramos estaban limpios; solamente el ID de los vinos no era necesario para nuestro análisis. La mayoría de nuestros datos son relevantes ya que son parte de la elaboración del vino, sus características y dependen unos de los otros.

Exploración

Gráfica 1.2 Comparación (Acidity/Sulphates)

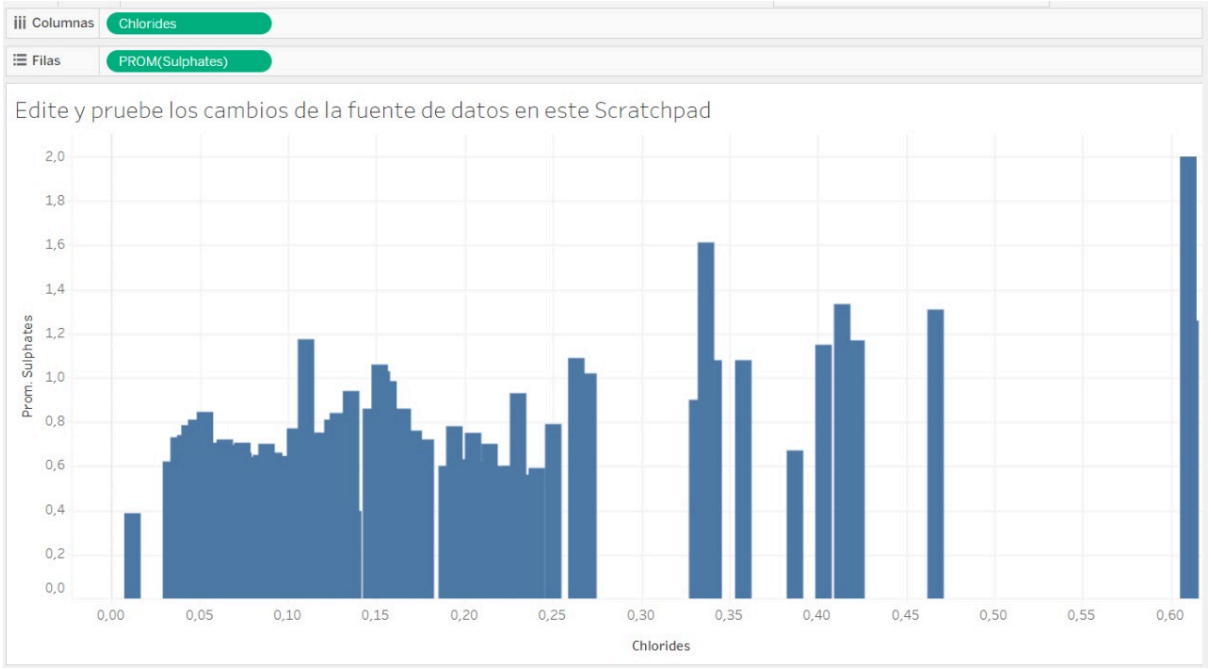


Comparación entre Acidez y Sulfatos

Queremos ver como la acidez afecta al sulfato, ya que es el encargado del color y sabor del vino, y se quiere comprobar si al momento de tener una acidez alta o baja,

puede llegar a afectar dichas características. Se puede observar que el sulfato va disminuyendo poco a poco conforme se eleva la acidez.

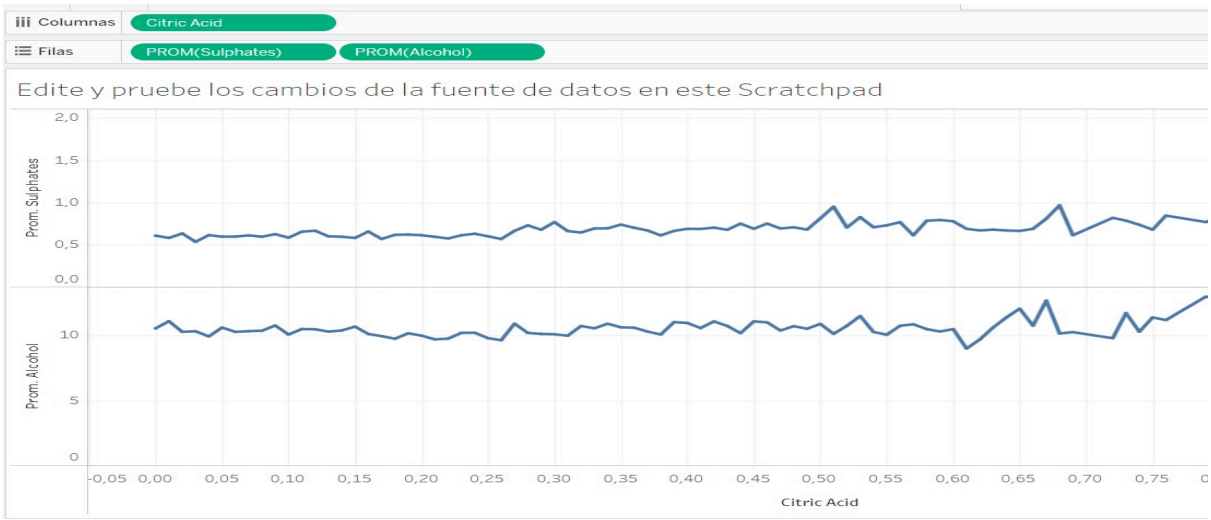
Gráfica 1.3 Comparación (Sulphates/Chlorides)



Comparación entre Sulfatos y Cloruros

Tenemos la suposición de que entre más concentración de sulfatos que desprende la levadura al ser fermentada se agregan más cloruros para nivelar el nivel de salinidad del Vino. En la gráfica podemos observar que en los niveles de promedio del sulfato se tiene una medida de cloruros, pero mientras se vaya aumentando el ulfato al mismo tiempo se aumentan los Cloruros del vino.

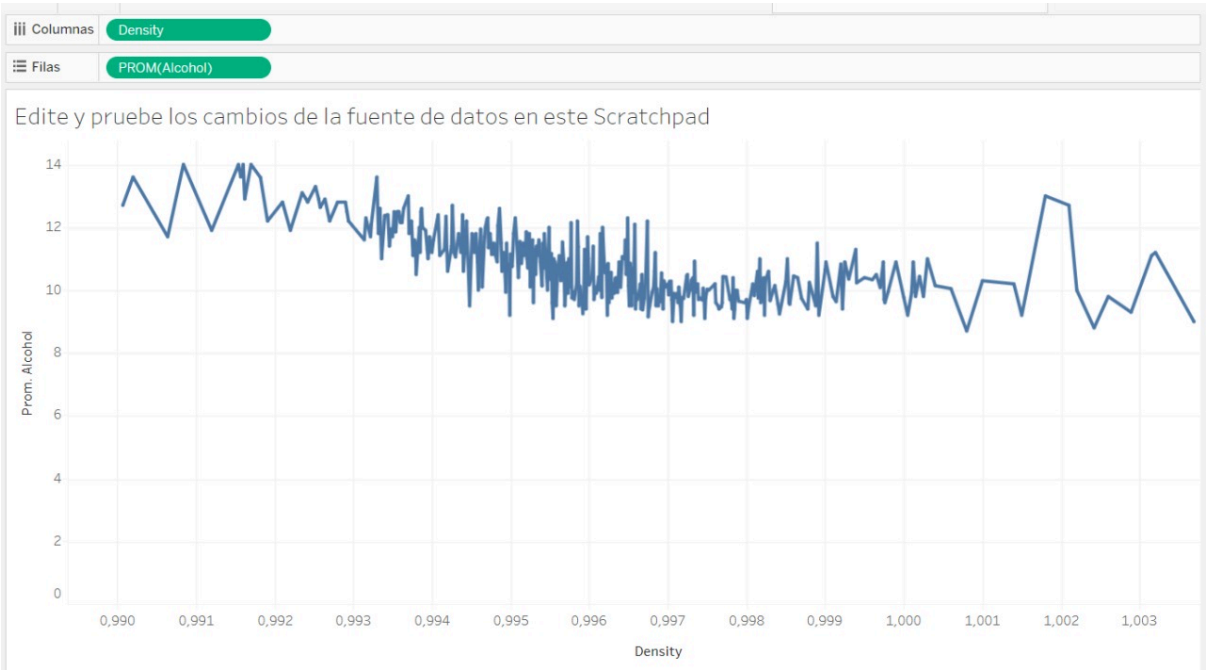
Gráfica 1.4 Comparación (Citric Acid/ Sulphates/Alcohol)



Comparación entre Ácido cítrico, sulfatos y alcohol

Existe una relación entre el Ácido cítrico, los sulfatos y el alcohol, y es que a medida que aumenta el ácido cítrico, también van aumentando tanto los sulfatos como el alcohol dentro del producto.

Gráfica 1.5 Comparación (Alcohol/Density)



Comparación entre Alcohol y Densidad

Teníamos la idea de que mientras más alcohol tuviera el vino, se podría representar con una mayor densidad pero en la gráfica podemos observar lo contrario, mientras más denso es el vino, menos alcohol tiene. Por lo cual era una dato que no nos esperábamos.

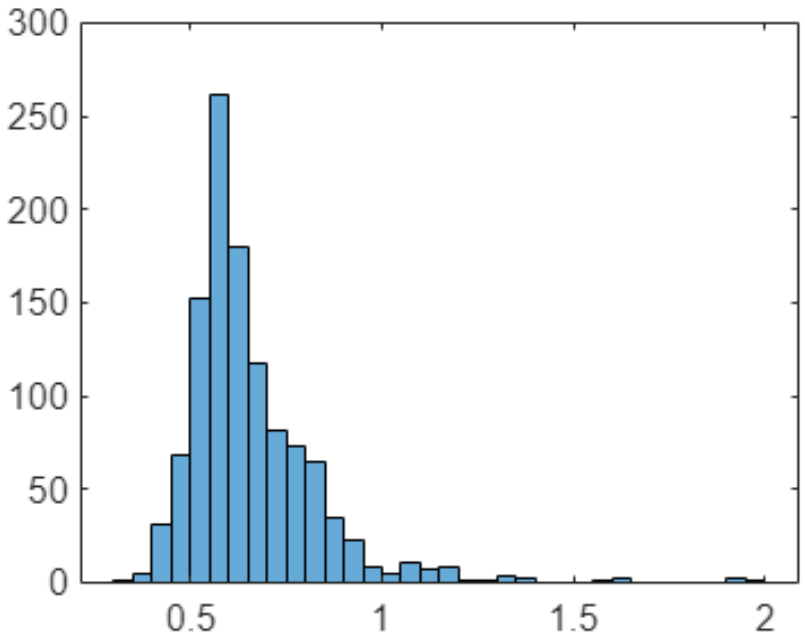
Etapa 2: Análisis Estadístico

Análisis estadístico descriptivo

Para esta sección elegimos cuatro variables las cuales consideramos dado el análisis previo como unas de las más importantes. Analizaremos a detalle cada una de estas variables a continuación.

Sulfatos

Gráfica 2.1 Histograma de Sulfatos



Media: 0.6577

Varianza: 0.0290

Desviación Estándar: 0.1704

Simetría: 2.4940

Curtosis: 14.9596

Hipótesis: Con los datos observados podemos ver que la gráfica está sesgada a la izquierda, por lo cual asumimos que se encuentra mayor cantidad de datos en esta zona. Además, la curtosis al ser muy alta, y viendo la gráfica, podemos observar que la curva se desplaza bastante y hay datos extremos muy lejanos a la línea media.

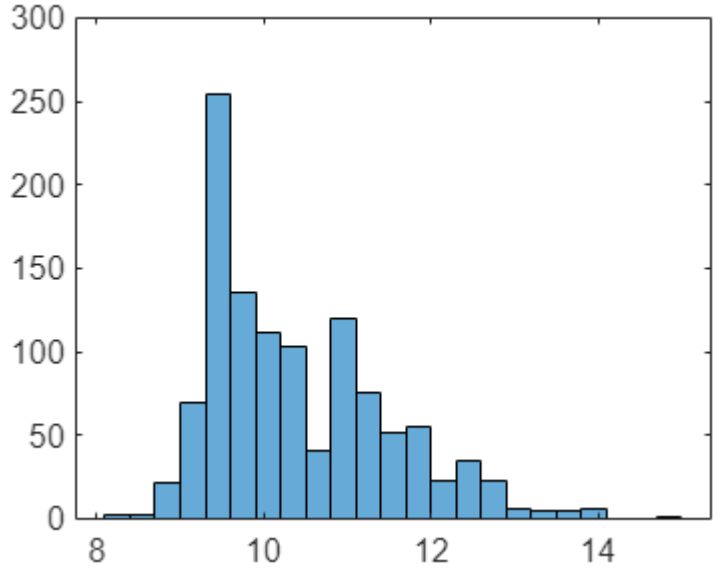
Interpretaciones:

En la media podemos observar que el valor esperado de nuestros datos se encuentra en 0.6577, con esto y con la varianza y desviación estándar la cual es bastante baja, podemos interpretar que la gran mayoría de datos se encuentran bastante pegados a la media, además, con la simetría podemos observar que no es perfectamente simétrica nuestra distribución de datos. Debido a que la curtosis es

alta, tenemos algunos datos atípicos alejados de nuestra gráfica estándar, estos probablemente no se vayan a tomar en cuenta en un futuro.

Alcohol

Gráfica 2.2 Histograma de Alcohol



Media: 10.4421

Varianza: 1.1711

Desviación Estándar:1.0822

Simetría: 0.8622

Curtosis: 3.2150

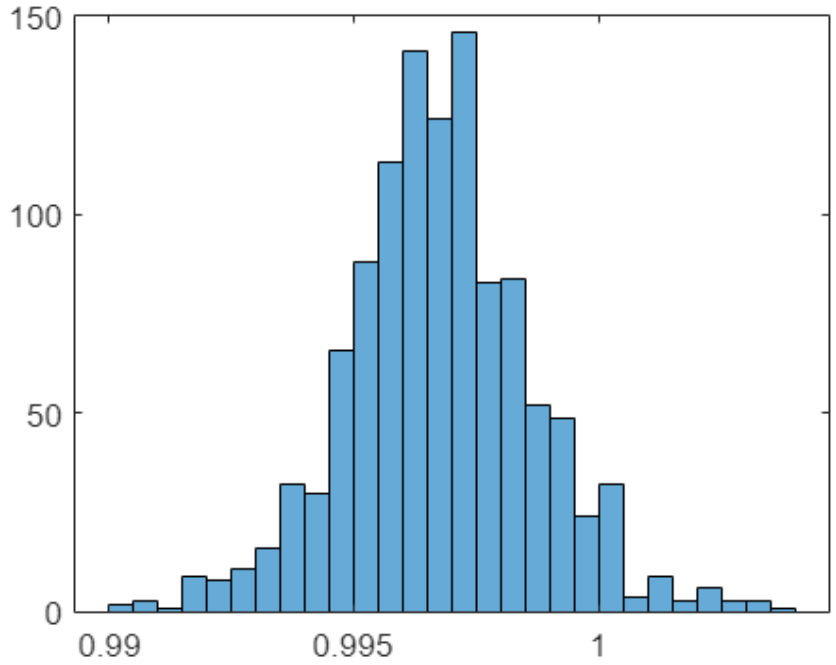
Hipótesis: Al observar nuestros datos y analizarlo de la mano con nuestro histograma, podemos observar que la varianza entre cada dato es bastante alta, lo cual nos indica que los datos varían mucho en esta cuestión, y no siguen una distribución tan exacta como las demás, esto en gran parte debe ser porque el nivel de alcohol en un vino varía mucho para satisfacer los gustos de los clientes.

Interpretaciones:

La con la media podemos observar que el valor esperado de nuestros datos se encuentra en 10.4421, con esto y con la varianza y desviación estándar la cual es un poco baja, podemos interpretar que la mayoría de datos se encuentran bastante pegados a la media, además, con la simetría podemos observar que es algo simétrica nuestra distribución de datos. Debido a que la curtosis es baja, tenemos nuestros datos mayormente concentrados cerca de la distribución principal.

Density

Gráfica 2.3 Histograma de Densidad



Media: 0.9967

Varianza: 0.0003860

Desviación Estándar: 0.0019

Simetría: 0.1023

Curtosis: 3.8790

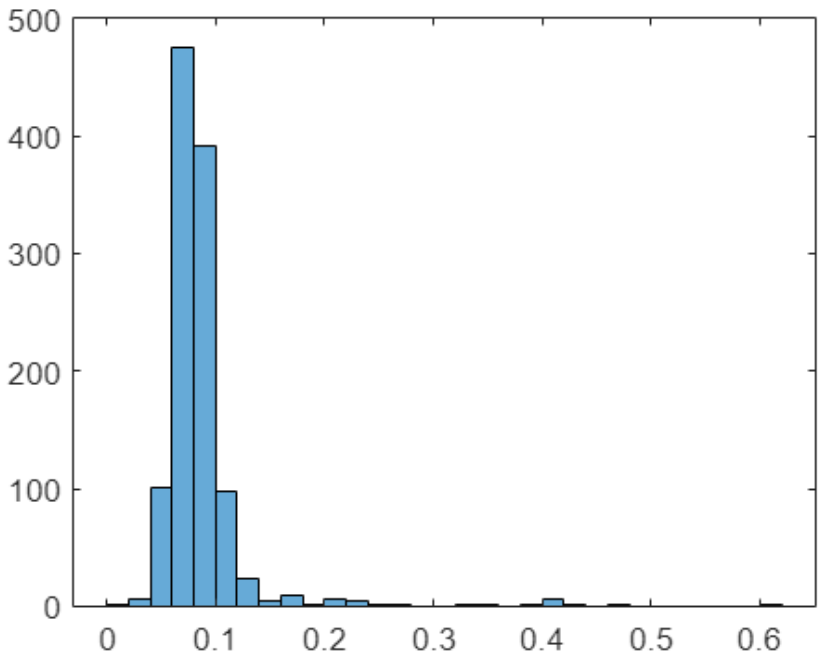
Hipótesis: Esta distribución al observar su distribución podemos ver que en general se trata de una distribución normal, además de que se encuentra bastante centrada como podemos observar por su simetría.

Interpretaciones:

La con la media podemos observar que el valor esperado de nuestros datos se encuentra en 0.9967, con esto y con la varianza y desviación estándar la cual es extremadamente baja, podemos interpretar que la gran mayoría de datos se encuentran bastante pegados a la media, además, con la simetría podemos observar que es bastante simétrica nuestra distribución de datos. Debido a que la curtosis es baja, tenemos nuestros datos mayormente concentrados cerca de la distribución principal.

Cloruro

Gráfica 2.4 Histograma de Cloruro



Media: 0.0869

Varianza: 0.0022

Desviación Estándar: 0.0473

Simetría: 6.0184

Curtosis: 49.8674

Hipótesis: En este caso tenemos una distribución normal, y al ver el sesgo tan grande que se observa a la izquierda, asumimos que la gran mayoría de datos de nuestra distribución se encuentran en este sector, con unas muy mínimas excepciones que podrían llegar a ser considerados datos atípicos.

Interpretaciones:

La con la media podemos observar que el valor esperado de nuestros datos se encuentra en 0.0869, con esto y con la varianza y desviación estándar la cual es extremadamente baja, podemos interpretar que la gran mayoría de datos se encuentran bastante pegados a la media, además, con la simetría podemos observar que no es perfectamente simétrica nuestra distribución de datos. Debido a que la curtosis es muy alta, tenemos varios datos atípicos alejados de nuestra gráfica estándar, estos probablemente no se vayan a tomar en cuenta en un futuro.

Determinación de la normalidad de los datos

Las hipótesis para la prueba de Anderson-Darling son:

- H0: Los datos siguen una distribución especificada
- H1: Los datos no siguen una distribución especificada

P determina el nivel de riesgo ya sea si puede rechazar la null hypothesis; datos pequeños significan que se puede rechazar. También para probar si los datos provienen de la distribución elegida

adstat: Si la distribución hipotética es un objeto de distribución de probabilidad completamente especificado, adtest calcula adstat usando parámetros específicos.

cv: determina cv mediante la interpolación en una tabla basada en el nivel de significación alfa especificado.

- Sulfatos
    - h: 1 logical
    - adstat: 37.1861
    - p: 0.0005
    - cv: 0.7514
- Alcohol
    - h: 1 logical
    - adstat: 25.1559
    - p: 5.0000e-04
    - cv: 0.7514
- Density
    - h: 1 logical
    - adstat: 2.8040
    - p: 5.0000e-04
    - cv: 0.7514
- Cloruro
    - h: 1 logical
    - adstat: inf
    - p: 5.0000e-04
    - cv: 0.7514

Al ver todos nuestras variables podemos observar que todas son lógicas por el resultado que mostraron en h, y como la p salió muy baja podemos confirmar que el riesgo de rechazar la null hypothesis es correcto.

Interpretación: Se puede observar que nuestras variables tienen una correlación, y vienen de una distribución elegida, lo cual rechaza una hipótesis nula, nos percatamos de eso mediante el valor de h y también de nuestro valor de p, los cuales son los que nos indican dicho parámetro.

Ajuste de distribuciones de probabilidad

Sulfatos

Gráfica 2.5 Distribuciones de probabilidad para sulfatos

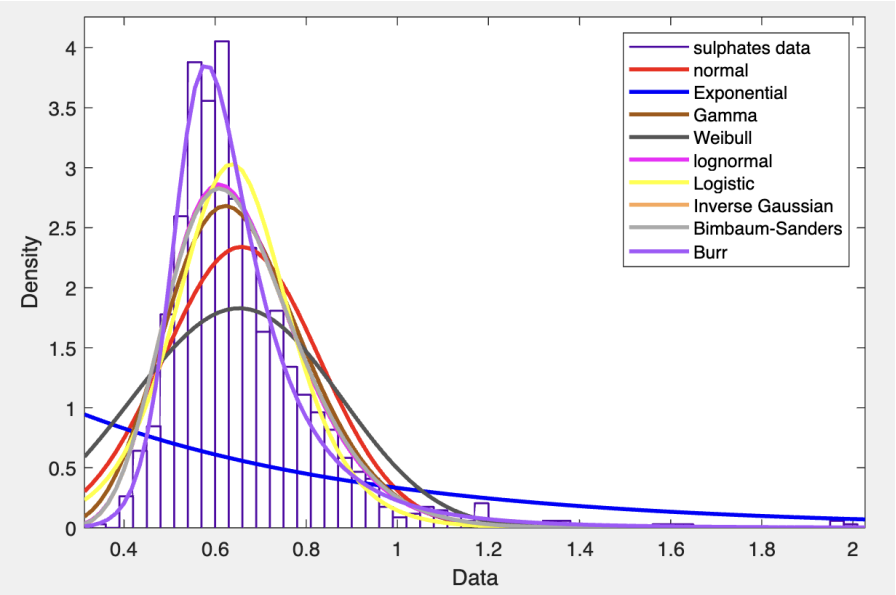


Tabla 2.6 Verosimilitudes de las distribuciones de sulfatos

Distribución	**Verosimilitud**
Normal	401.322
Exponencial	-664.089
Gamma	551.107
Weibull	303.274
Lognormal	601.975
Logística	533.769
Inverse Gaussian	597.528
Bimbaum-Saunders	596.38
Burr	673.612

Para esta variable la distribución Burr fue la más exacta. Su PDF es $\frac{ck}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}\left[1+\left(\frac{x}{\alpha}\right)^{c}\right]^{-k-1}$ en donde MATLAB nos ajustó los valores de c, k y alfa quedando $\frac{12.8724 \cdot 0.412618}{0.5525}\left(\frac{x}{0.5525}\right)^{12.8724-1}\left[1+\left(\frac{x}{0.5525}\right)^{12.8724}\right]^{-0.412618-1}$ sustituyendo estos valores.

Alcohol

Gráfica 2.7 Distribuciones de probabilidad para alcohol

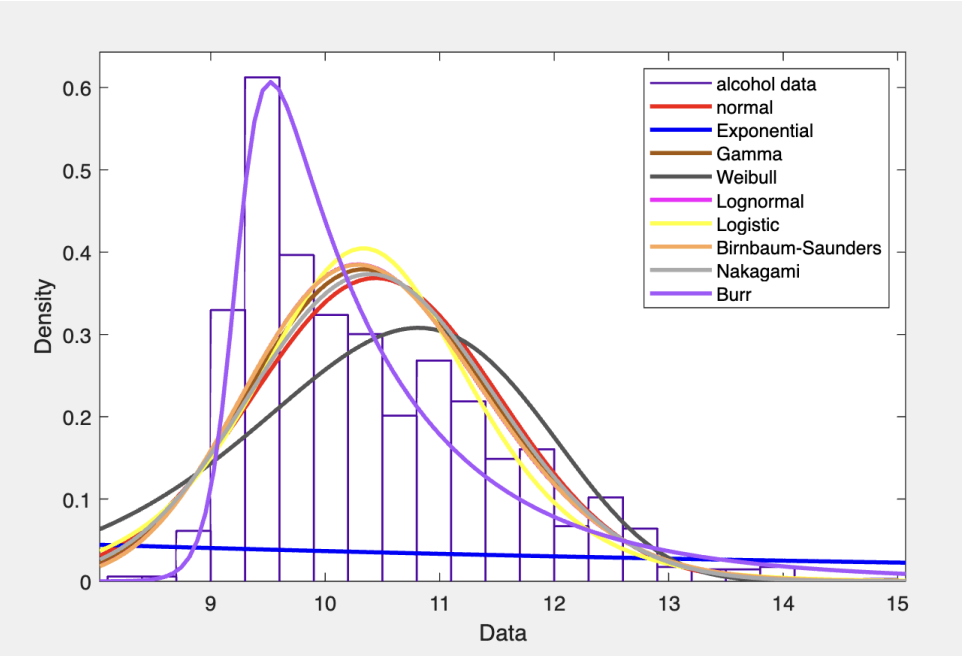


Tabla 2.8 Verosimilitudes de las distribuciones de alcohol

Distribución	**Verosimilitud**
Normal	-1711.63
Exponencial	-3824.3
Gamma	-1680.63
Weibull	-1842.86
Lognormal	-1667.36
Logística	-1719.88
Birnbaum-Saunders	-1667.23
Nakagami	-1695.28
Burr	-1585.25

Para esta variable la distribución Burr también fue la más exacta. Su PDF es de

$$\frac{70.4891 \cdot 0.120175}{9.25594}\left(\frac{x}{9.25594}\right)^{70.4891-1}\left[1+\left(\frac{x}{9.25594}\right)^{70.4891}\right]^{-0.120175-1}.$$

Densidad

Gráfica 2.9 Distribuciones probabilidad para densidad

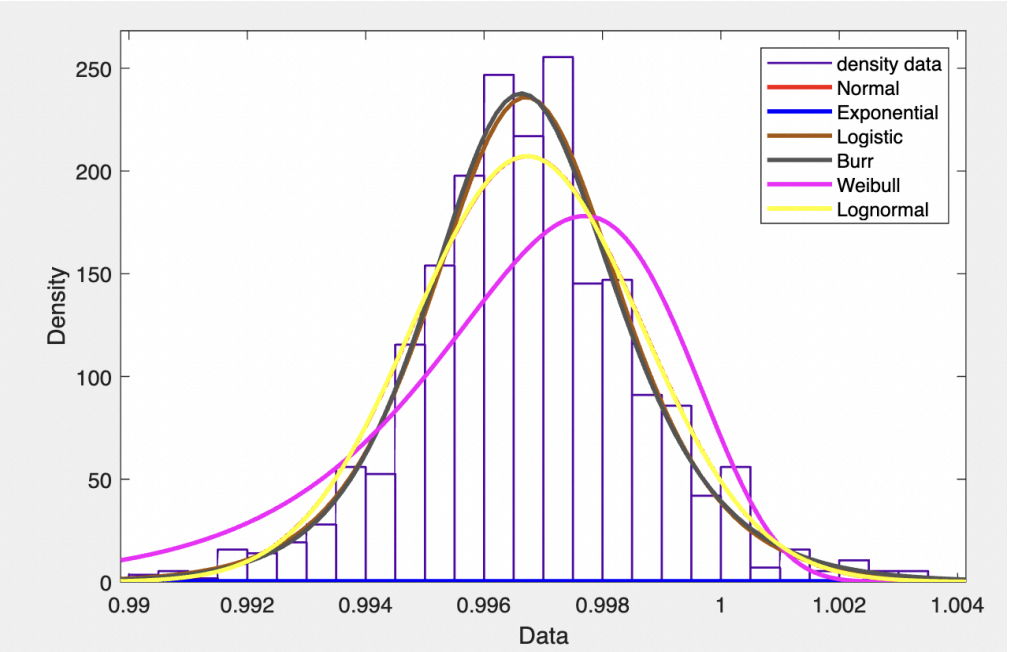


Tabla 2.10 Verosimilitudes de las distribuciones de densidad

Distribución	**Verosimilitud**
Normal	5525.6
Exponencial	-1139.26
Logística	5541.48
Burr	5542.4
Weibull	5384.99
Lognormal	5525.71

Esta variable sigue el mismo patrón, la distribución Burr es la más alta con un PDF de $\frac{990.217 \cdot 0.872332}{0.996496}\left(\frac{x}{0.996496}\right)^{990.217-1}\left[1+\left(\frac{x}{0.996496}\right)^{990.217}\right]^{-0.872332-1}$.

Cloruro

Gráfica 2.11 Distribuciones de probabilidad para cloruro

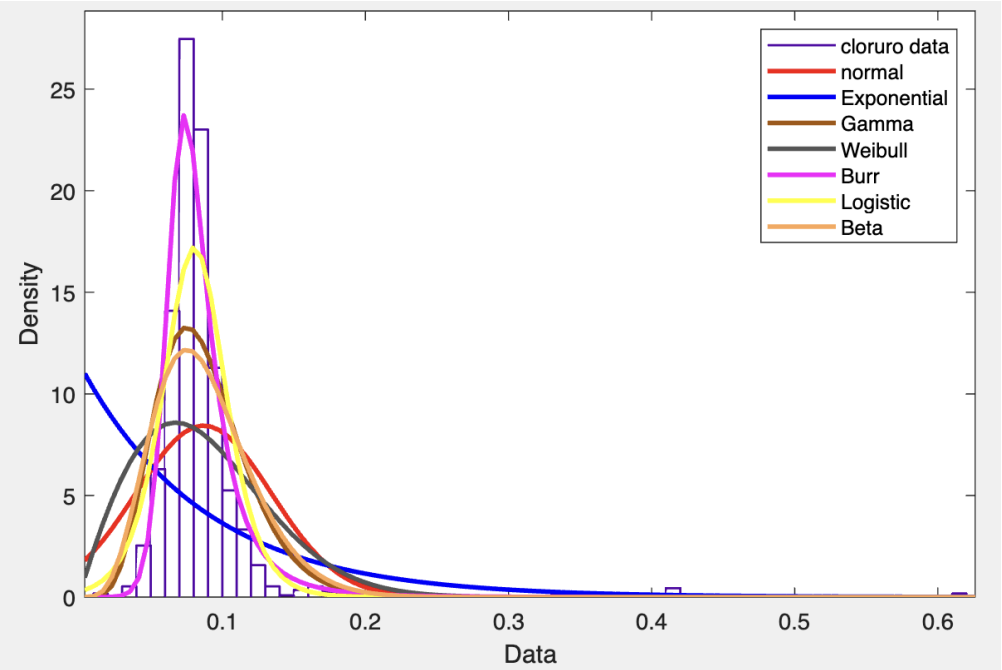


Tabla 2.12 Verosimilitudes de las distribuciones de cloruro

Distribución	**Verosimilitud**
Normal	1867.02
Exponencial	1648.92
Gamma	2369.79
Weibull	2068.54
Burr	2744.26
Logística	2410.55

Como era de esperarse, el PDF también es Burr aquí:

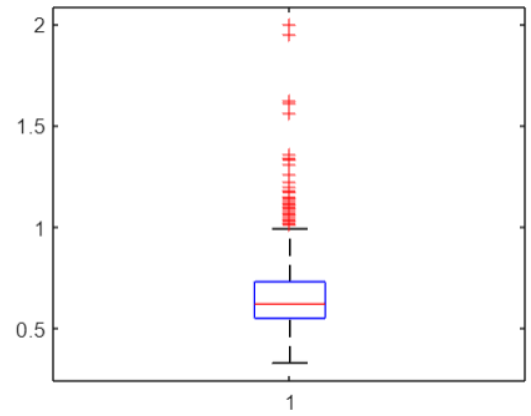
$$\frac{8.95624 \cdot 0.536565}{0.0711124}\left(\frac{x}{0.0711124}\right)^{8.95624-1}\left[1+\left(\frac{x}{0.0711124}\right)^{8.95624}\right]^{-0.536565\ -1}.$$

En todas las variables la verosimilitud más alta ha sido de la distribución Burr, por lo que esta es la que mejor se ajusta a todas las variables para modelar. También al observar las gráficas esta es la que se acerca más a los datos mostrados por el histograma lo cual refuerza lo mencionado anteriormente.

Datos atípicos

Utilizamos box plots para identificar datos atípicos en las cuatro variables elegidas.

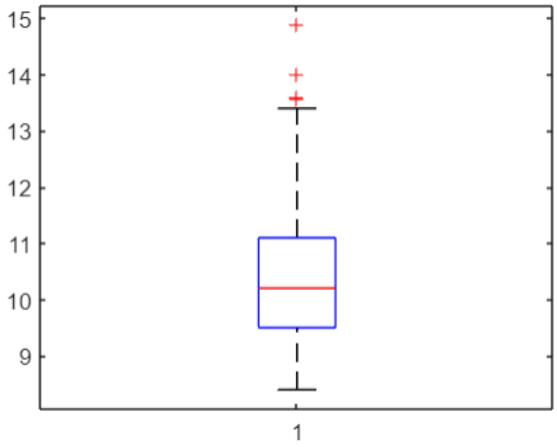
Sulfatos



Gráfica 2.13 Boxplot sulfatos

Hay gran cantidad de datos que están muy alejados de la media como se ve en el boxplot.

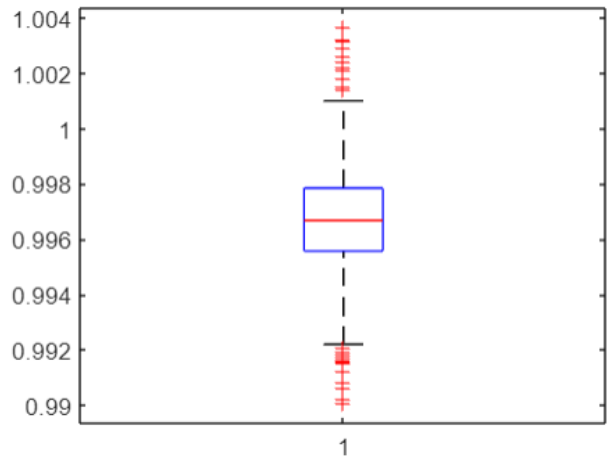
Alcohol



Gráfica 2.14 Boxplot alcohol

Los datos atípicos del alcohol son muy pocos ya que la media mantiene un rango muy amplio de los datos, los datos atípicos están a una distancia muy pequeña de la media.

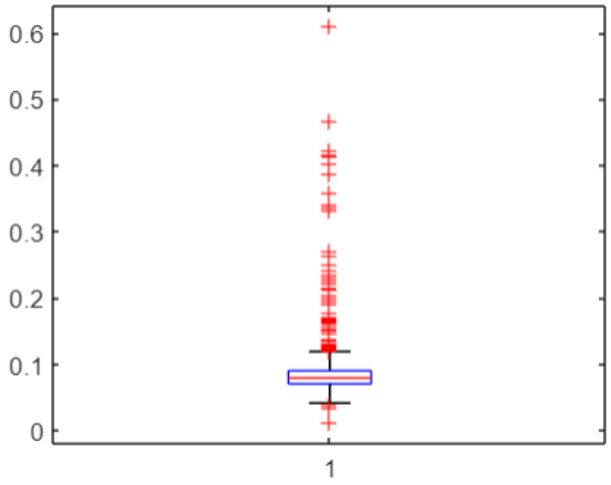
Densidad



Gráfica 2.15 Boxplot densidad

Los datos atípicos de la densidad están muy pegados a los bigotes de la gráfica, por lo que no son tan extremos.

Cloruro

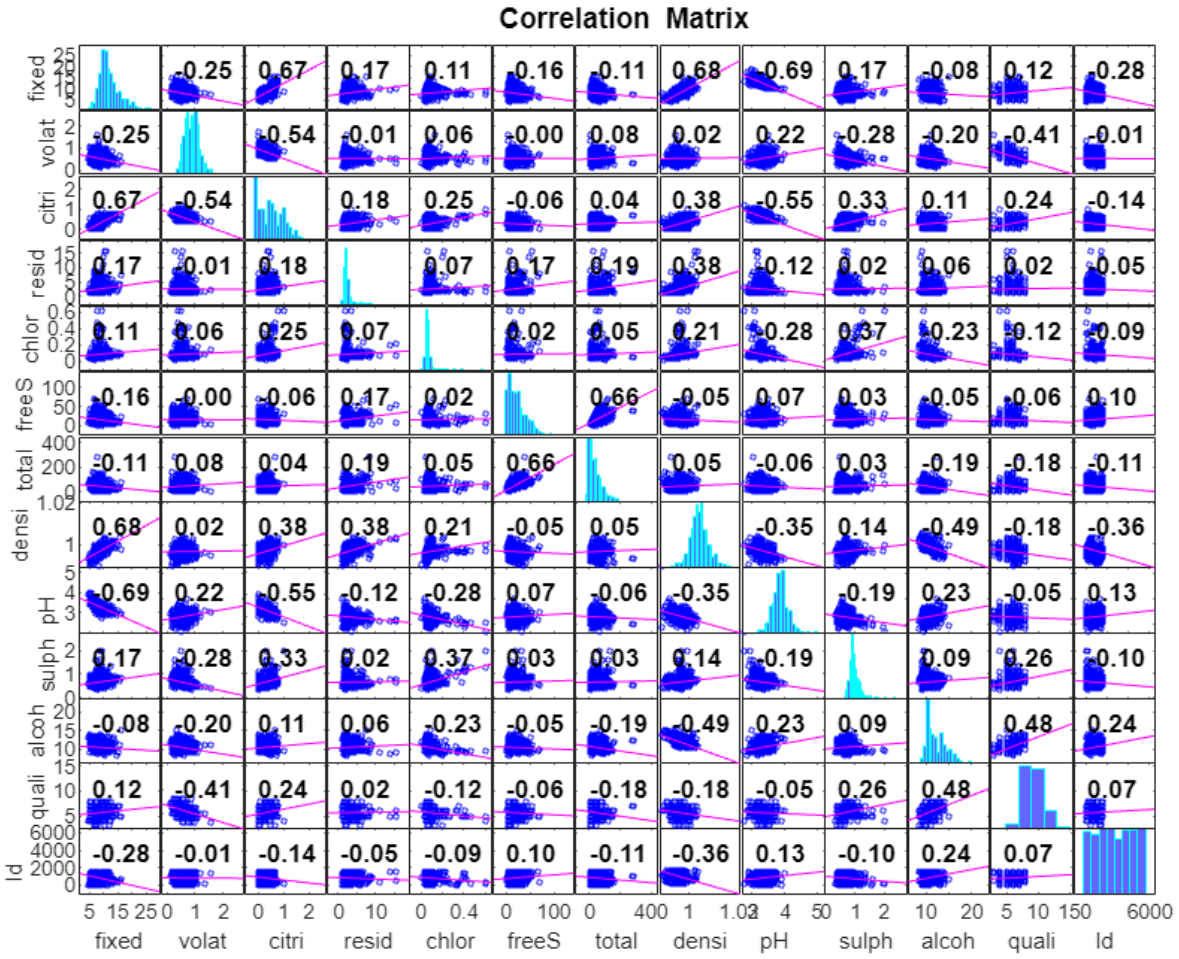


Gráfica 2.16 Boxplot cloruro

Se considera que hay muchos datos atípicos en la variable de cloruro debido a que los datos progresivamente se alejan de la media.

Matriz de correlación

Gráfica 2.17 Matriz de correlación de todas las variables



Variables de mayor correlación:

- **Alcohol:** Alcohol-densidad, Alcohol-Calidad

  En este podemos observar que el grado de alcohol afecta bastante la calidad del vino y la densidad de este.

- **CítricAcid:** Citri-Acidez, Citri-pH

  En esta relación podemos observar que el ácido cítrico afecta directamente la acidez del vino y el pH, esto lo consideramos bastante obvio así que no lo utilizaremos como variable.

- **Fixed Acidity:** Acidez-pH, Acidez-Citri, Acidez-Densidad

  En esta relación podemos observar que la acidez afecta directamente al pH y al ácido cítrico, lo cual también lo podemos obviar.

Ajuste del alcance

Hemos decidido seleccionar al Alcohol como variable crítica, debido a que está bastante relacionado con varios datos relevantes (densidad y calidad) como podemos observar en la matriz de correlación, además de que al analizar los datos como equipo llegamos a la conclusión de que afecta en gran parte la calidad y percepción del vino a largo plazo. También es una variable muy fácil de percibir para el consumidor.

Etapa 3: Modelación

Técnica 1: Redes Neuronales

Importamos librerías necesarias y tomamos los valores de WineQT.csv (nuestra base de datos). Después almacenamos los 4 principales que vamos a utilizar los cuales son densidad, cloruros, acidez (inputs) y alcohol (output y variable crítica).

Imagen 3.1 Código de inicio para red neuronal

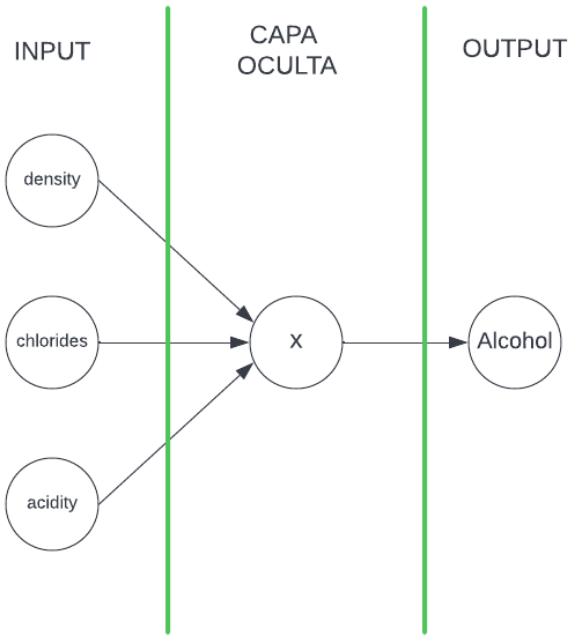
```
[45] import tensorflow as tf
     import numpy as np
     import pandas as pd
     from tensorflow.keras.layers import *
     from tensorflow.keras.models import Sequential, Model
     from tensorflow.keras.optimizers import Adam, RMSprop
```

```
[46] df = pd.read_csv (r'/content/sample_data/WineQT.csv')
     alcohol = np.array(df.alcohol, dtype=float)
     density = np.array(df.density, dtype=float)
     chlorides = np.array(df.chlorides, dtype=float)
     volAcidity = np.array(df.volatileacidity, dtype=float)

     print (alcohol)
     print(density)
     print(chlorides)
     print(volAcidity)
```

```
[ 9.4  9.8  9.8 ... 10.5 11.2 10.2]
[0.9978  0.9968  0.997   ... 0.9949  0.99512 0.99547]
[0.076 0.098 0.092 ... 0.09  0.062 0.075]
[0.7   0.88  0.76  ... 0.6   0.55  0.645]
```

La idea es crear una red neuronal como la siguiente:

Imagen 3.2 Ejemplo de red neuronal



Creamos red neuronal de 3 entradas y 1 salida estableciendo la capa de tipo keras, e inicializando el modelo. Además, creamos un merged array de los 3 arreglos que vamos a utilizar de input.

Imagen 3.3 Código para crear red neuronal

```
[61] #### Red neuronal de 3 entradas y 1 salida#######

     capa = tf.keras.layers.Dense(units=1, input_shape=[3])
     modelo = tf.keras.Sequential([capa])

     merged_array = np.stack([density, chlorides, volAcidity], axis=1)
```

Compilamos el modelo y declaramos el optimizador, el cual será de tipo Adam y utilizaremos como parámetro un learning rate de 0.1. Con esos valores entrenamos a nuestro modelo.

Imagen 3.4 Código para compilar y entrenar red

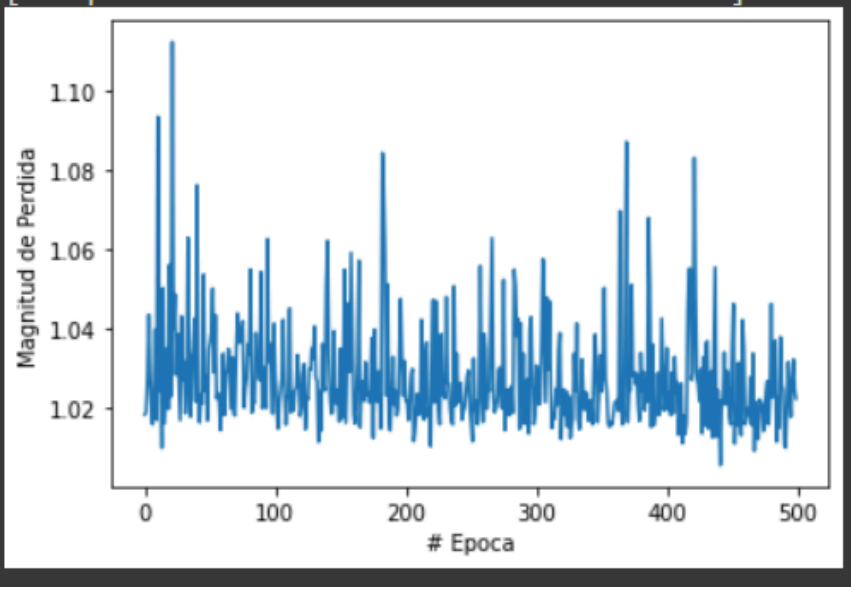
```
[99] modelo.compile(
         optimizer = tf.keras.optimizers.Adam(0.1),
         loss='mean_squared_error'
     )

[118] print("Comenzando entrenamiento...")
      historial = modelo.fit(merged_array, alcohol, epochs=500, verbose=False)
      print("Modelo entrenado!")

      Comenzando entrenamiento...
      Modelo entrenado!
```

Después, calculamos la Loss function de nuestro modelo.

Imagen 3.5 Código y gráfica de la función de pérdida

```
import matplotlib.pyplot as plt
plt.xlabel("# Epoca")
plt.ylabel("Magnitud de Perdida")
plt.plot(historial.history["loss"])
```

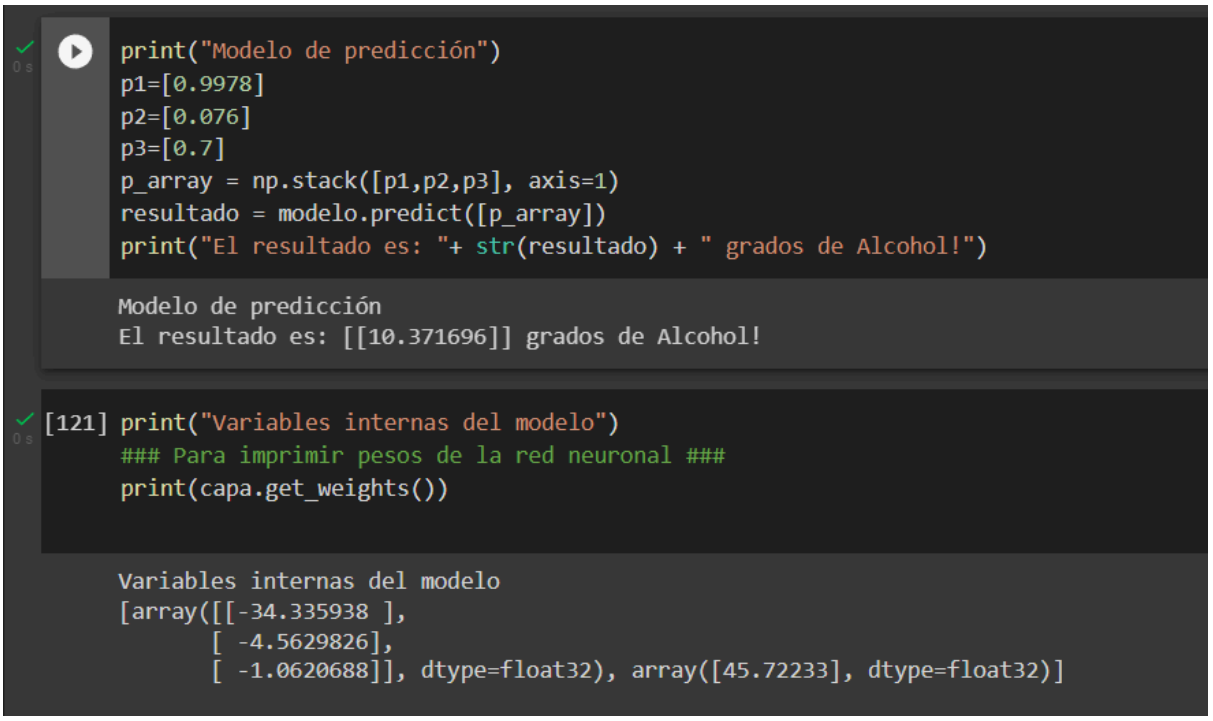
```
[<matplotlib.lines.Line2D at 0x7ffb16a8edd0>]
```



Esto lo podemos interpretar como que se realizaron muchos ajustes a nuestro modelo, y que fluctuaron mucho los datos y los valores de las variables internas del sistema antes de llegar al ajuste que tenemos actualmente.

Llevamos a cabo un modelo de predicción para probarlo, recibiendo como resultado la predicción en grados de alcohol. e imprimimos nuestras variables internas del modelo para analizarlas.

Imagen 3.6 Código para predecir



```
print("Modelo de predicción")
p1=[0.9978]
p2=[0.076]
p3=[0.7]
p_array = np.stack([p1,p2,p3], axis=1)
resultado = modelo.predict([p_array])
print("El resultado es: "+ str(resultado) + " grados de Alcohol!")
```

```
Modelo de predicción
El resultado es: [[10.371696]] grados de Alcohol!
```

```
[121] print("Variables internas del modelo")
      ### Para imprimir pesos de la red neuronal ###
      print(capa.get_weights())
```

```
Variables internas del modelo
[array([[-34.335938 ],
        [ -4.5629826],
        [ -1.0620688]], dtype=float32), array([45.72233], dtype=float32)]
```

Esto lo podemos interpretar como que al realizar nuestro modelo de predicción utilizando nuestra red neuronal ya entrenada, podemos predecir los grados de alcohol a partir de 3 valores: densidad, cloruros y acidez, los cuales utilizará para darnos un resultado predictivo de cual sería los grados de alcohol en ese caso. Finalmente imprimimos las variables internas del modelo, la cual nos muestra los valores de los nodos del diagrama de la red neuronal (que se encuentra arriba).

Técnica 2: Bosques Aleatorios

A continuación utilizaremos la técnica de modelación de bosques aleatorios, para explicar nuestra variable crítica (Alcohol) con los datos de nuestra base de datos y comprobar que funcione como mínimo un 65% del tiempo

Primero importamos nuestra base de datos wineQT y creamos un objeto Dataframe utilizando la librería de Pandas.

Imagen 3.7 Código de inicio segunda técnica

```
df = pd.read_csv("WineQT.csv")
df.head()
```

Python

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

Checamos los tipos de datos para que no hubiera complicaciones al crear el modelo y de esa manera tener un estándar.

Imagen 3.8 Tipos de datos del dataframe

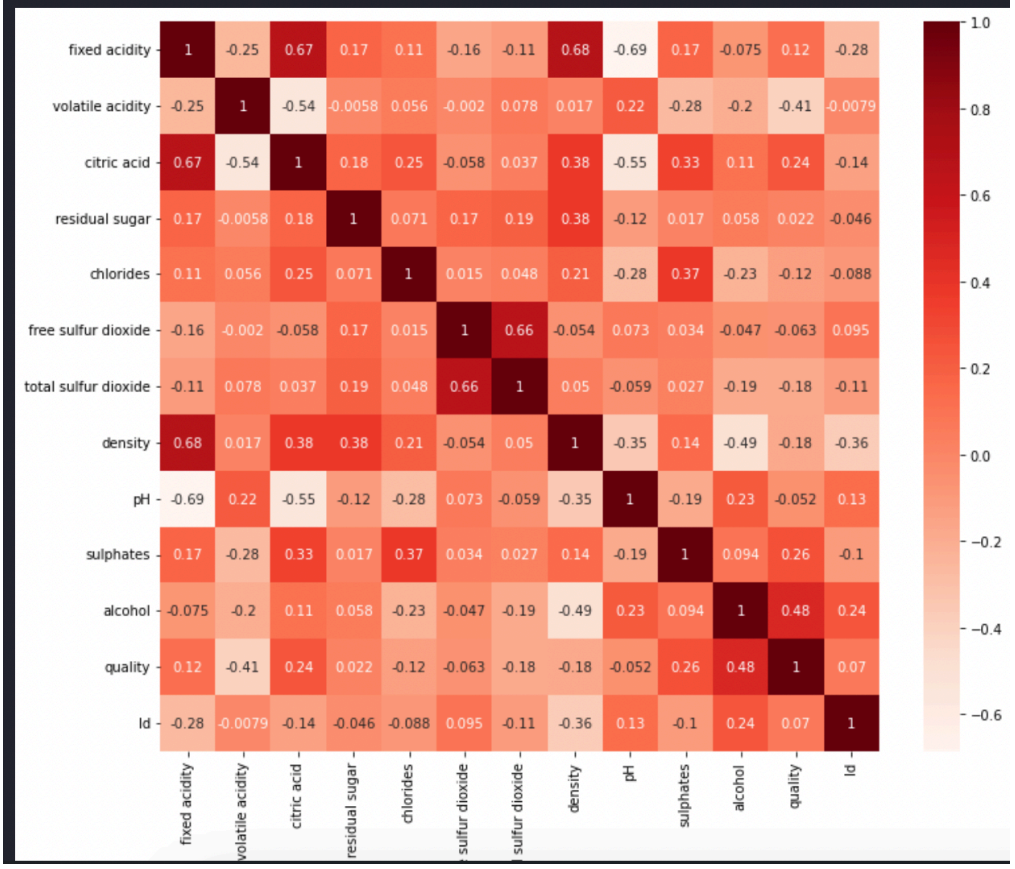
```
df.dtypes
```

```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
Id                        int64
dtype: object
```

Verificamos de nuevo la matriz de correlación utilizando la función corr.

Imagen 3.9 Matriz de correlación y código

```
plt.figure(figsize=(12,10))
cor = df.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)
plt.show()
```



Después de varias pruebas, nos dimos cuenta de que el modelo tenía mejor desempeño utilizando todas las variables. El Id es irrelevante así que sólo removimos esa variable.

Imagen 3.10 Código para quitar columna id

```
df.drop("Id", axis=1, inplace=True)
df.head()
```

Finalmente, comprobamos que no haya nulos en nuestros datos.

Imagen 3.11 Código para checar nulos

```
df.isnull().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Creamos sets de entrenamiento y prueba, utilizamos la función de train_test_split con un tamaño de 0.2, es decir, un quinto de nuestros datos serán utilizados como prueba y el resto para entrenamiento.

Imagen 3.12 Código para crear sets de prueba y entrenamiento

```
X_data = df.drop('alcohol', axis=1)
Y_data = df['alcohol']

x_train, x_test, y_train, y_test = train_test_split(X_data, Y_data, test_size=0.2)
print(x_train.shape, x_test.shape)
print(y_train.shape, y_test.shape)
```

```
(914, 11) (229, 11)
(914,) (229,)
```

Creamos un modelo usando RandomForestRegressor, lo creamos con los datos de entrenamiento y lo probamos con nuestros datos de prueba.

Imagen 3.13 Código para crear modelo

```
m = RandomForestRegressor(n_jobs=-1, oob_score=True)
m.fit(x_train, y_train)
m.score(x_train, y_train)
```

```
0.9642152542569961
```

```
print(m.score(x_test, y_test))
print("Oob Score: ", m.oob_score_)
```

```
0.7330265866407675
Oob Score:  0.7408571642382702
```

Tenemos un score de .73 y un OOB score de 0.74. Intentamos mejorarlos un poco cambiando los parámetros de nuestro regresor. Utilizamos ahora 100 árboles cambiando n_estimators.

Imagen 3.14 Código para crear segundo modelo

```
m2 = RandomForestRegressor(n_estimators=100,min_samples_leaf=1, n_jobs=-1, oob_score=True)
m2.fit(x_train, y_train)
m2.score(x_train, y_train)
```

```
0.9645528097802956
```

```
print(m2.score(x_test, y_test))
print("Oob Score: ", m2.oob_score_)
```

```
0.7382230873648195
Oob Score:  0.7444672692980303
```

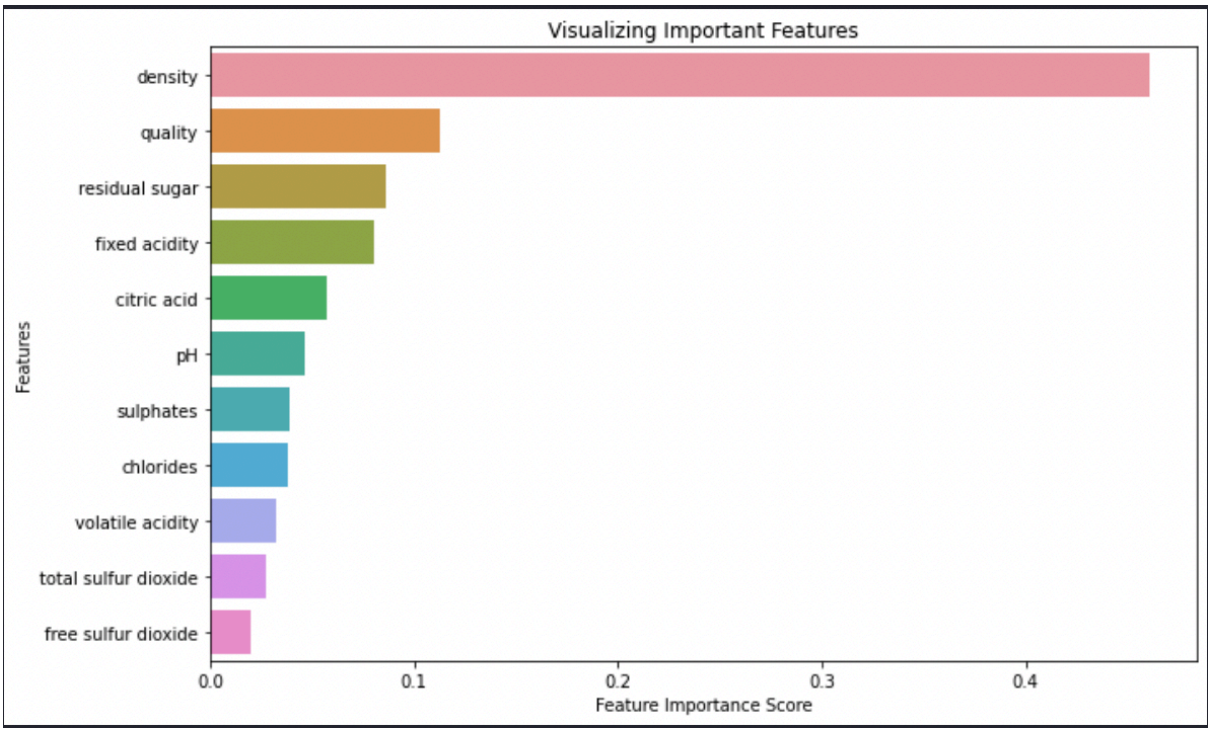
Ahora tenemos un score de 0.738 y OOB 0.744 en nuestro set de prueba, una pequeña mejora.

Finalmente ploteamos nuestras variables más importantes,

Imagen 3.15 Código para obtener gráfica

```
feature_imp = pd.Series(m2.feature_importances_, index=x_train.columns).sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=feature_imp, y=feature_imp.index)
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title("Visualizing Important Features")
plt.tight_layout()
```

Gráfica 3.16 Variables más importantes para el modelo



Algo interesante es que la variable de residual sugar no estaba en nuestras variables de correlación pero terminó siendo relevante para nuestro modelo.

Etapa 4: Simulación

Objetivo

Como objetivo de la simulación, tenemos la intención de utilizar los datos para observar sus repercusiones sobre la variable crítica. Para esto, analizaremos las variables que potencialmente podrían generar un impacto significativo en nuestra variable crítica, y analizaremos distintos escenarios que iremos planteando.

El primer escenario a analizar será el de la densidad, para el cual utilizaremos nuestro modelo bosques aleatorios para realizar un análisis y simulación de qué pasaría con la variable crítica (alcohol) en el escenario en el que nuestros valores de densidad se encuentren en el percentil más alto, y qué pasaría si se encuentran en el más bajo.

Además, realizaremos un análisis similar utilizando como referencia el valor de nuestra variable “calidad”, con la cual realizaremos un análisis de sus dos escenarios más extremos, y con esto podremos formular hipótesis sobre el impacto de estas variables sobre el comportamiento del sistema y el cambio en nuestra variable crítica.

Definición

Los datos nuevos generados a partir de nuestros objetivos son los siguientes escenarios y sus repercusiones sobre el sistema y la media de la variable crítica:

- **Escenarios de media de densidad baja o alta en el vino:** La hipótesis que tenemos sobre este escenario es que debido a que la densidad tiende a ser inversamente proporcional, cuando hagamos la predicción sobre la densidad baja, la media de el alcohol será mayor. En cambio, cuando llevemos a cabo la predicción sobre la densidad alta, la media del alcohol será más baja.

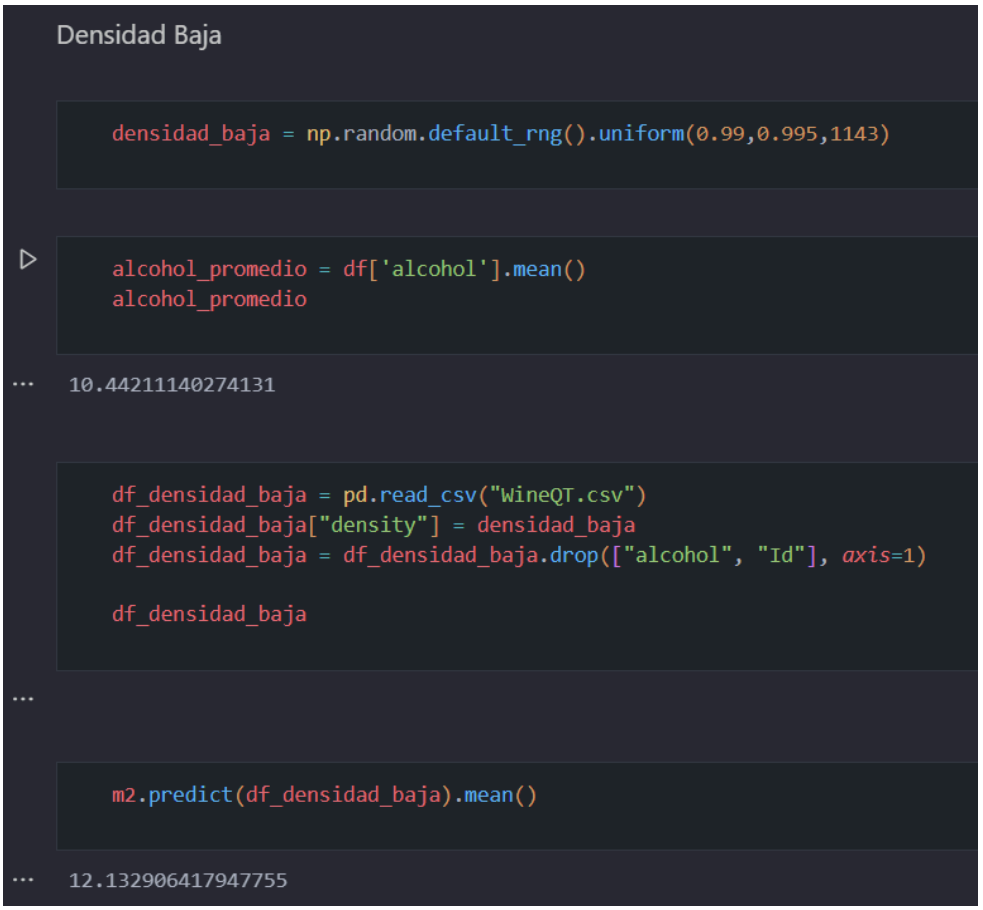
- **Escenarios de media de calidad baja o alta en el vino:** La hipótesis que tenemos sobre este escenario es que la calidad no tiene una relación tan simple como directamente o inversamente proporcional frente al grado de alcohol, así que creemos que van a fluctuar muy poco los resultados cuando hagamos la predicción sobre la calidad baja y alta.

- **Escenarios compuestos de calidad alta/baja y densidad alta/baja:** La hipótesis que tenemos sobre este escenario es que debido a que la calidad no tiende a tener un impacto tan significativo sobre la media de nuestra variable crítica, el resultado se va a inclinar más dependiendo de la variable de densidad, la cual tiende a ser inversamente proporcional, cuando hagamos la predicción sobre la densidad baja (con calidad alta o baja), la media de el alcohol será mayor. En cambio, cuando llevemos a cabo la predicción sobre la densidad alta (con calidad alta o baja), la media del alcohol será más baja.

Resultados

Generamos datos aleatorios con numpy entre 0.99 y 0.995 que es un rango bajo para la densidad. Sustituimos estos valores en nuestro dataframe original, creamos uno nuevo con esto y realizamos predicciones.

Imagen 4.1 Código de simulación con densidades bajas



Observamos que la media original del alcohol es de 10.44, mientras que con datos de densidad baja esta aumenta hasta 12.132. Un cambio significativo.

Realizamos el mismo procedimiento anterior ahora con datos en un rango alto, de 0.097 hasta 1.004. El comportamiento ahora debería ser a la inversa, es decir, el alcohol debería disminuir.

Imagen 4.2 Código de simulación con densidad alta

Densidad Alta

```
densidad_alta = np.random.default_rng().uniform(0.997,1.004,1143)
```

```
df_densidad_alta = pd.read_csv("WineQT.csv")
df_densidad_alta["density"] = densidad_alta
df_densidad_alta = df_densidad_alta.drop(["alcohol", "Id"], axis=1)

df_densidad_alta
```

```
m2.predict(df_densidad_alta).mean()
```

```
9.94844098619617
```

Podemos confirmar nuestra suposición, pues observamos que el alcohol con datos de densidad alta dan como resultado 9.9, un número por debajo de la media original.

Ahora realizaremos simulaciones modificando la variable de calidad, para ver cómo esta afecta al alcohol. Al igual que el proceso anterior, generamos datos aleatorios en rangos altos y después bajos.

Primero simularemos una calidad alta de vino, con valores de entre 7 y 9. Recordemos que la calidad del vino tiene valores del 1 al 10 enteros.

Imagen 4.3 Código de simulación de calidad alta

Podemos ver que la predicción arroja un resultado justo por encima de la media. Ahora realizamos lo mismo con valores bajos.

Imagen 4.4 Código de simulación con calidad baja

Las predicciones con calidad alta y baja son bastante similares, con variación de tan solo 0.6 grados. Podríamos decir que el alcohol no afecta mucho la calidad.

Ahora generamos simulaciones con combinaciones de las dos variables anteriores, utilizando los mismos valores generados anteriormente.

Imagen 4.5 Código de simulación con combinaciones de densidad y calidad

```
df_calidad_baja_densidad_alta = pd.read_csv("WineQT.csv")
df_calidad_baja_densidad_alta["quality"] = calidad_baja
df_calidad_baja_densidad_alta["density"] = densidad_alta

df_calidad_baja_densidad_alta = df_calidad_baja_densidad_alta.drop(["alcohol", "Id"], axis=1)

df_calidad_baja_densidad_baja = pd.read_csv("WineQT.csv")
df_calidad_baja_densidad_baja["quality"] = calidad_baja
df_calidad_baja_densidad_baja["density"] = densidad_baja

df_calidad_baja_densidad_baja = df_calidad_baja_densidad_baja.drop(["alcohol", "Id"], axis=1)

df_calidad_alta_densidad_alta = pd.read_csv("WineQT.csv")
df_calidad_alta_densidad_alta["quality"] = calidad_alta
df_calidad_alta_densidad_alta["density"] = densidad_alta

df_calidad_alta_densidad_alta = df_calidad_alta_densidad_alta.drop(["alcohol", "Id"], axis=1)

df_calidad_alta_densidad_baja = pd.read_csv("WineQT.csv")
df_calidad_alta_densidad_baja["quality"] = calidad_alta
df_calidad_alta_densidad_baja["density"] = densidad_baja

df_calidad_alta_densidad_baja = df_calidad_alta_densidad_baja.drop(["alcohol", "Id"], axis=1)
```

La simulación con calidad baja y densidad alta nos da un valor por debajo de la media.

Imagen 4.6 Predicción con calidad baja y densidad alta

```
m2.predict(df_calidad_baja_densidad_alta).mean()
```

```
9.65833796053271
```

Por otra parte, simulando calidad alta con densidad alta nos da un valor de alcohol casi igual a la media.

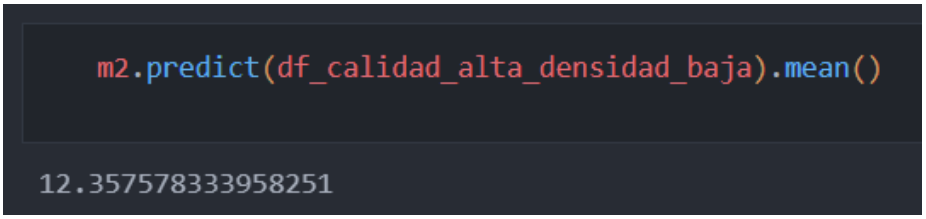
Imagen 4.6 Predicción con calidad alta y densidad alta

```
m2.predict(df_calidad_alta_densidad_alta).mean()
```

```
10.432670141926705
```

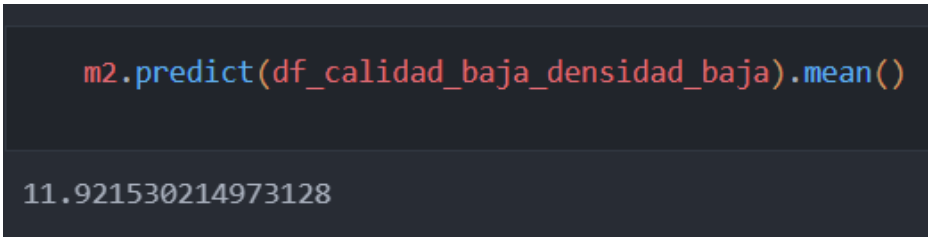
Simulando con calidad alta y densidad baja, tenemos un resultado bastante alto.

Imagen 4.6 Predicción con calidad alta y densidad baja



Finalmente, con una calidad baja y densidad baja, seguimos teniendo un valor alto de alcohol.

Imagen 4.6 Predicción con calidad baja y densidad baja



Etapa 5: Conclusiones

Insights

Al momento de analizar la información de los vinos, se pudo identificar las variables que hacen que el vino sea de más agrado al consumidor, mediante la cantidad de alcohol y la calidad.

Durante la investigación tuvimos que ser muy cautelosos al ver la relación entre todas las variables, ya que al no saber que tanto un sulfato o algún elemento puede afectar a la calidad del vino, no se podían descartar dichas variables. Nuestra hipótesis fue aceptada, se observó que la calidad no tenía un gran impacto dentro de nuestra variable crítica, pero sí depende un poco más de la densidad, y se llegó a la conclusión de que con una menor densidad se tiene un mayor nivel de alcohol y viceversa. Así que la calidad de un vino no está determinada por su grado de alcohol.

Este tipo de análisis son muy importantes para una empresa para así identificar qué variables de sus productos son la clave para tener una venta alta o de igual manera una venta baja, para que puedan seguir produciendo productos que le benefician pero que reduzcan la cantidad de vinos que no cumplen con los requisitos de la calidad.

Ya para finalizar, se pudo visualizar a detalle las variables que ayudan a un vino a ser mejor, tal como los azúcares, la densidad, los sulfatos entre otros. Dicho eso, las que tuvieron más impacto dentro de nuestra investigación fueron la densidad y el alcohol, ya que tenían una relación más directa con la calidad. Pero también no todo el vino depende de una sola variable, se tienen diferentes combinaciones de variables que la conllevan a tener su aspecto clave. Ya sea una relación con baja densidad y alto alcohol puede resultar más favorable a una que tenga alta densidad y bajo alcohol. Como el vino es una mezcla de todos estos factores, todos deben tener su medida específica para en conjunto llevar a un mejor producto.

Diagnóstico

Al escoger la base de datos nos aseguramos que esta estuviera conformada por todos los datos necesarios para poder analizar de manera más profunda la calidad y propiedades del vino, viendo cómo afectan las variables directamente con el vino y como es que baja o disminuye al aumentar o disminuir ciertas variables. La base de datos conformaba todos los elementos que buscábamos pero tuvimos que excluir el id porque no era relevante para nuestro proyecto. Al momento de ingresar los datos

de la base de datos para calibrar el programa ya realizado, excluimos los datos nulos y consideramos realizar la matriz de correlación con 4 variables y que relacion tenian pero al realizar más pruebas nos dimos cuenta que el modelo tenía mejor desempeño al utilizar todas las variables que solamente usando unas cuantas.

El modelo que generamos si sirvió para poder determinar si al utilizar diferentes variables podría afectar nuestra variable crítica(alcohol) ya que los resultados obtenidos en los escenarios que propusimos afectan directamente al alcohol cambiando su valor y afectando el vino en general, además al realizar las pruebas en los escenarios puestos pudimos comprobar que nuestras hipótesis eran correctas; a menor densidad el alcohol aumenta mientras que a mayor densidad el alcohol sube, también se determinó que si se modifica el valor de la calidad este no tiene un cambio significativo con el alcohol y que si cambiamos el valor de la calidad y de la densidad el resultado de la calidad no cambiará mucho al alcohol entonces el resultado será más inclinado a la variable de la densidad teniendo un resultado similar a la primera hipótesis.

Tuvimos la intención de utilizar los datos para observar sus repercusiones con la variable crítica (alcohol) pero esta misma información se puede utilizar para poder analizar qué cambios puede haber en todas las variables de las bases de datos al cambiar su valor, qué variables son afectadas por el cambio de otras variables y cómo esto cambia el vino y diferentes aspectos de este ya sea su azúcar, su densidad, su salinidad, el sabor, etc. se puede observar las repercusiones con otras variables no solo con el alcohol sino con todas las variables, de este modo se hace otro tipo de análisis en donde se cambie la variable que se quiere observar.

Con la misma información, el mismo análisis pudimos haber obtenido otros modelos que determinan diferentes aspectos de nuestra base de datos como por ejemplo haber determinado si varias variables al combinarlas afectan otras variables únicas o conjuntas del vino. Nosotros lo hicimos de únicamente con pocas variables pero es posible que se haga un análisis con cada una de las variables y cómo son afectadas por cambiar una o más variables.

Otros datos o variables que podríamos agregar a nuestra base de datos del vino son la edad del viñedo, la temperatura, la humedad, la elaboración. Estas son variables que llegan afectar a otras variables del vino y que podrían ser agregadas para determinar qué tanto afectan las variables anteriormente mencionadas con las variables que ya tenemos en la base de datos.

Para que utilicemos los datos hay que analizarlos y transformarlos para poder ver las relaciones que tienen en las variables. Con nuestra base de datos vimos que la mayoría de las variables eran importantes y que afectan a la calidad del vino, solo determinamos que una variable no era necesaria para nuestro análisis y la descartamos siendo esta la del id. Además verificamos que las variables que tenían

null sean porque no hay un dato y no porque ese dato sea cero, ya que un dato que tenga null no significa que sea cero.

Etapa 6: Reflexiones

Participante 1: Juan Camilo Granados

Primero, todo lo que me llamó la atención de la realización del proyecto, inicialmente fué el poder elegir una base de datos que nos llamara la atención, y que además conociéramos, ya que aprendimos sobre la importancia de conocer los datos antes de empezar a trabajar con ellos. Otra cosa que me llamó mucho la atención fueron los pasos a seguir para poder filtrar una buena base de datos trabajable, de una que no es tan buena y le faltan datos o sus variables no nos funcionan, además fue muy interesante también aprender a trabajar con los datos, sacar toda la información que podemos extraer de ellos y llegar a conclusiones útiles a partir de nuestro trabajo.
En lo personal, considero que se tiene que llegar a un balance frente a lo que quieren los usuarios contra lo que buscan los inversionistas, ya que siempre es importante mantener a ambos satisfechos. Pienso que existen y se deben utilizar todas las alternativas (donde apliquen) que ayudan a que todo sea más eficiente, para poder cumplir las expectativas de ambos.
Desde el punto de vista de ciudadanía, considero que una correcta modelización del proyecto ayuda a tener una mejor comunidad, ciudad y país en el sentido en el que siempre que se trabaja con datos de forma tan profunda, se pueden llegar a conclusiones que a simple vista no se ven y son beneficiosas para el usuario en caso de ser aplicadas. Además, considero que una modelización correcta puede ser muy beneficiosa para el medio ambiente en el contexto de la industrialización de los procesos, ya que cada vez es más valioso optimizar recursos en ese sentido y tiene un impacto muy grande hacerlo.

Participante 2: Alberto J Treviño Gonzalez

Me pareció muy interesante la manera en la que se llevó a cabo el proyecto, desde la oportunidad de elegir una base de datos de nuestro interés, hasta la parte de analizar los datos para encontrar predicciones y ver cómo las variables se relacionan con la variable crítica, y darte cuenta que quizá y aspectos que veías importante o pensabas que podían influir mucho dentro de un resultado siempre no tenían dicha magnitud y afectaba más otra variable.
Pensando como ingeniero yo creo que sí es importante cumplir con lo que el cliente pide, pero también si tenemos conocimiento propio, se puede brindar algún consejo o algo que veamos más oportuno hacer, pero solo como recomendación. Se debe intentar cumplir con

ambos papeles y también satisfacer al cliente tanto como al programador, debido a que un buen trato conlleva a un buen resultado.
Desde el punto de vista de ciudadanía, claro que tener una correcta modelización ayuda a tener un mejor país, con tanta falta de información, infraestructura, y datos, tener una herramienta de modelación puede optimizar el trabajo, y se pueden sacar datos que un simple humano puede dar por alto o simplemente no identificarlos.
Y finalmente como un punto de vista de desarrollo sustentable se puede buscar la manera para apoyar a empresas para no gastar tanto recursos y contaminar de dicha forma, ya que hay mucho desgaste de material y muy probablemente es por la falta de organización y optimización de sus sistemas.

Participante 3: David Sanchez Ortiz

Considero que es un buen proyecto que se puede realizar, ya que para los que disfrutan el análisis y la programación es una buena oportunidad para poder adentrarse a una nueva área de trabajo. Yo al inicio la verdad no sabia que era lo que teníamos que realizar en el proyecto y todo lo que nos decian que ibamos a hacer se me hacia muy interesante, aunque no sabia que hacer, pero mientras avanzamos el curso me di cuenta que paso a paso si se podía hacer el proyecto. Disfruté el transcurso del proyecto aprendiendo de nuestra base de datos y que es lo que buscábamos con ella, también me di cuenta el potencial que se podía sacar con hacer un análisis en alguna base de datos, porque con eso te das cuenta de relaciones que normalmente no se podrían ver a la vista.

Desde el punto de vista de la ciudadanía, considero que es importante realizar un buen análisis a cualquier proyecto que realize algun profesionista ya que con esto cada una de las empresas pueden poder progresar al hacer diferentes predicciones o checar correlaciones de estudios que quieran realizar en sus propios mercados, esto mejorando la empresa que ellos tienen, por consecuencia al país.

Participante 4: Luis Yerik Arámbula Barrera

Este proyecto fue una gran oportunidad para practicar muchos de los aspectos que conforman la ciencia de datos. Desde el análisis, modelación y simulaciones, pude realizar diferentes actividades que reforzaron y probaron mis conocimientos como programador. No tenía un interés muy particular en los vinos, pero a lo largo del proyecto me vi más interesado y entusiasmado de seguir descubriendo y aprendiendo cosas a partir de datos. Con este proyecto pude darme cuenta de todo el poder que tienen los datos, y tuvo mucho sentido el por qué la ciencia de datos y el análisis de estos son habilidades tan demandadas hoy en día.

Respecto a la pregunta sobre las inversiones, considero que se debe encontrar un punto medio entre inversionista y usuario, para que ambos puedan estar satisfechos. Un buen producto tendrá muchos clientes aunque signifique una mayor inversión al inicio. Por otro

lado, también puede haber mucha inversión para usuarios que no aprovecharán o notarán todas las mejoras de un sistema.

Una correcta modelización de un proyecto tendrá beneficios tanto para usuarios como para inversionistas. Esto aplica también si los afectados por este son la ciudadanía, pues estarán satisfechos con el uso de los recursos y el producto cumplirá a la perfección su propósito. Si hay un desbalance de esto podemos tener personas insatisfechas o un mayor uso de recursos que resulta en un desperdicio si no son utilizados. Esto último puede tener un impacto negativo en el medio ambiente, por lo que es muy importante analizar bien el proyecto a realizar.

Hablando de proyectos con datos también hay que tener mucho cuidado, pues aunque no sean un recurso físico como tal, almacenar grandes cantidades de ellos puede resultar sumamente costoso, al igual que su procesamiento. Tenemos que tener en cuenta cuántos datos utilizaremos y también cuáles, pues también caemos en riesgos de seguridad y privacidad.

En general, los proyectos de datos creo que tienen un potencial inmenso para un impacto positivo en nuestra sociedad. Creo que debemos invertir más en ellos para nuestro desarrollo, tomando en cuenta todos los riesgos y cuidados que esto conlleva.