



# Proyecto Integrador: Análisis de datos de vinos

Equipo:

Santiago Francisco Robles Tomayo

Reyna Yanet Hernández Mada

María Elena Martínez Manzanares

Martín José Vega Noriega

1 de octubre del 2022

# Resumen ejecutivo

---

Base de datos: Calidad de vino, Vinho verde

---

Metodología: CRISP – DM

---

Modelo: Random Forest

---

Variables: Calidad del vino(Dependiente),  
Características (Independientes)

---

Objetivo : Brindar una herramienta de toma de decisiones sobre que variables enfocarse para aumentar la calidad de la bebida.

---

Resultados: Streamlit /Github

# Outline

- 1. Comprensión del negocio**
- 2. Comprensión de los datos**
- 3. Modelado**
- 4. Evaluación**
- 5. Implantación**





# Comprensión del negocio



# Contexto

- Un precedente del uso de datos vinculados a la producción de vino para esta investigación es el trabajo de Cortez et al (2009), quien realizó un experimento de minería de datos usando datos recolectados de mayo 2004 a febrero 2007 con valores atribuidos a vinos tintos y blancos.

# Objetivos


---

## **Objetivos generales:**

- Determinar la clasificación del vino a través de sus características.

## **Objetivos específicos:**

- Realizar un análisis multivariado para identificar relaciones de dependencia entre las características del vino.
- Efectuar una implantación en un servicio de host gratuito a través de framework de alto nivel programable en Python.
- Identificar el algoritmo de clasificación o regresión adecuado para discernir la calidad del vino a través de sus características.



# Comprensión de los datos

# Descripción de los datos

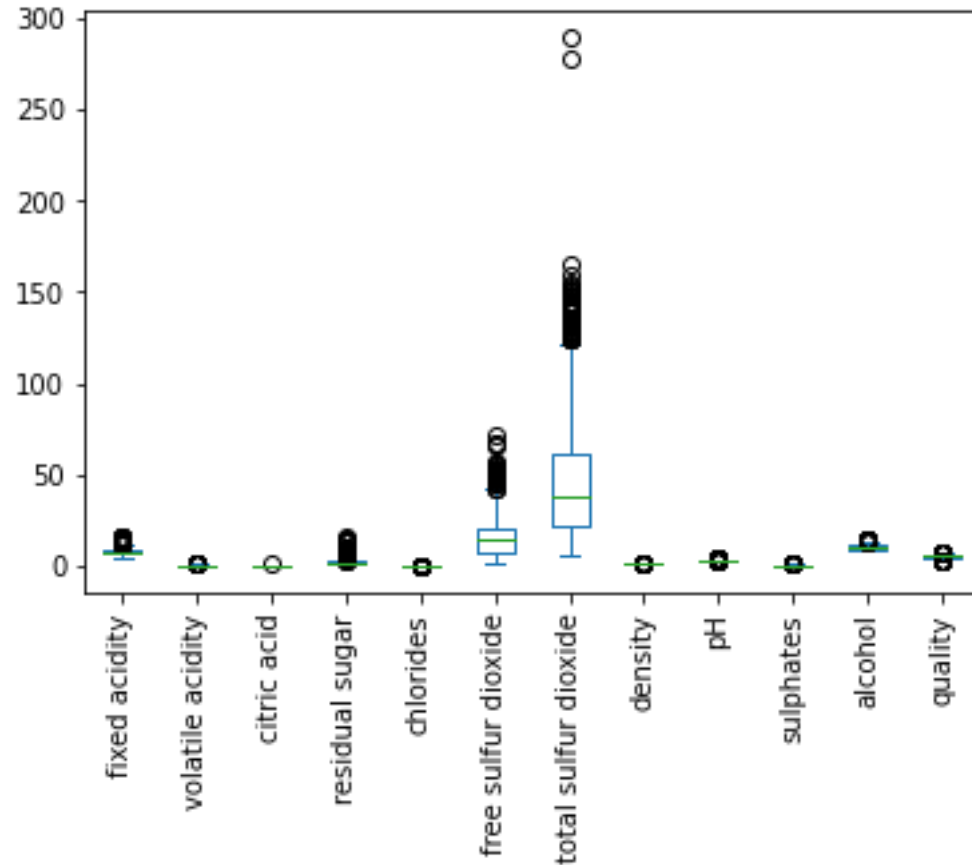
Las doce variables que componen la tabla son, básicamente, componentes que definen la calidad, sabor y estructura de un vino.

Las entradas del dataset incluyen pruebas objetivas, mientras que las salidas se basan en datos sensoriales (mediana de al menos tres evaluaciones realizadas por expertos en vino).

Cada experto calificó la calidad del vino entre 0 (muy malo) y 10 (excelente).



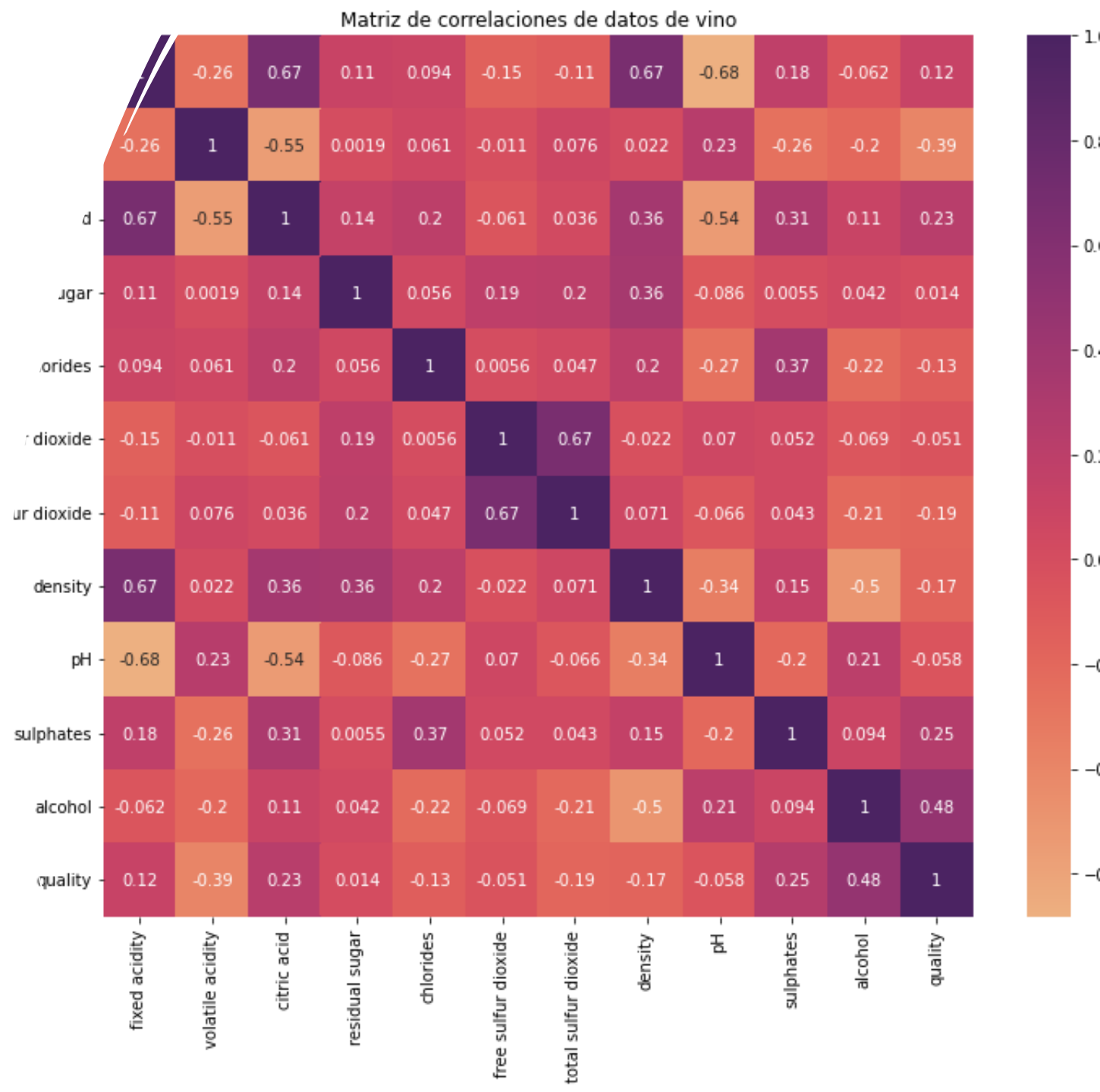
# Análisis exploratorio de datos




Se determinó que la mayoría de las columnas de datos presentan una **desviación estándar pequeña y con rangos de valores** no muy amplios.

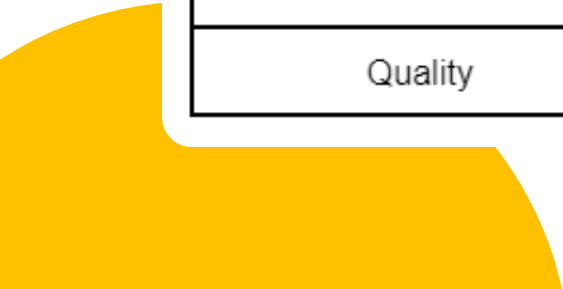
Se detectó una **presencia de una cantidad considerable de datos atípicos**.

- Un análisis bivariado y multivariado a la muestra fue realizado, por medio del cálculo de correlaciones y del Factor de Inflación de Varianza , respectivamente.





Variable	VIF
Fixed acidity	74.45
Volatile acidity	17.96
Citric acid	9.19
Residual sugar	4.66
Chlorides	6.64
Free sulfur dioxide	6.46
Total sulfur dioxide	6.60
Density	1528.15
pH	1078.17
Sulphates	22.46
Alcohol	147.61
Quality	77.72



- Se consideró que una variable no presentaba multicolinealidad cuando su VIF resultaba menor o igual que 10.
- Solamente **5 variables obtuvieron VIF aceptable** para el análisis.

---

La tabla de datos presenta un **desbalance** importante con respecto a la variable objetivo.

La mayoría de los registros corresponden con los niveles de calidad 5 y 6, representando el **85% de la muestra**.

Calidad	Incidencias	Porcentaje (%)
3	10	0.65
4	53	3.31
5	681	42.58
6	638	39.89
7	199	12.44
8	18	1.12



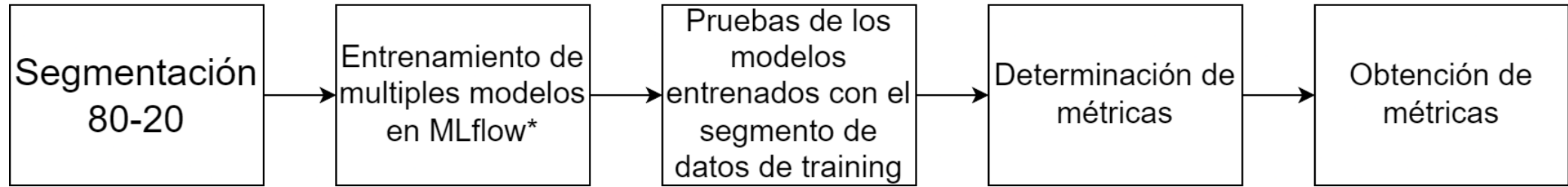
Modelado

A large orange circle is positioned on the left side of the slide, partially cut off by the edge.

## Escoger técnica de modelado

Se consideraron para entrenamiento los modelos Random Forest, SVC, SVC lineal y Naive Bayes con el objetivo de discernir cuál era el que mejor se ajustaba a la problemática.



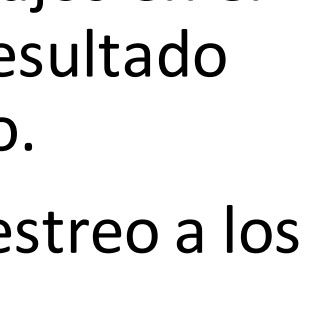


# Construcción del modelo

Iteración 1 (sprint 1)

- \*Random Forest, SVC, SVC lineal y Naive Bayes

Method name	Training accuracy (exactitud de entrenamiento)	Training precision (precisión de entrenamiento)	Training F1 Score (Valor-F de entrenamiento)
Linear SVC	0.557	0.544	0.543
Gaussian NB	0.573	0.579	0.573
SVC	0.513	0.577	0.464
Random Forest	1	1	1

- Se conjeturó que los resultados bajos en el desempeño de los modelos fue resultado del imbalance de datos detectado.
  - Se aplicó la técnica de sobre muestreo a los datos de la tabla.
- 



## Iteración 2 (sprint 2)



Calidad	Incidencias	Porcentaje (%)
3	551	16.6%
4	551	16.6%
5	551	16.6%
6	551	16.6%
7	551	16.6%
8	551	16.6%

Se tuvieron un total de 3306 registros en el cual cada clase se componía de 551 registros, contrastado con los 1599 registros originales.

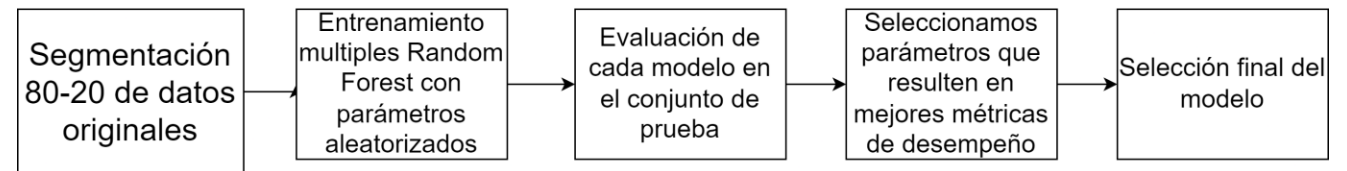
\*Random Forest, SVC, SVC lineal y Naive Bayes

Method name	Training accuracy (exactitud de entrenamiento)	Training precision (precisión de entrenamiento)	Training F1 Score (Valor-F de entrenamiento)
Linear SVC	0.384	0.359	0.333
Gaussian NB	0.528	0.501	0.498
SVC	0.412	0.422	0.388
Random Forest	1	1	1

- El imbalance de datos no presentó ser un factor determinante para el desempeño de los modelos, donde de nueva cuenta el modelo con los mejores resultados fue el Random Forest con valor de 1 en todas sus métricas.

## Iteración 3 (sprint 3)

	Testing accuracy (exactitud de prueba)	Testing precision (precisión de prueba)	Testing F1 Score (Valor-F de prueba)
Muestra original	0.666	0.635	0.648





Evaluación

# Revisar proceso

---



Se conjetura que el EDA se puede expandir si se estudian las correlaciones multivariadas detectadas por el VIF.



Valorar el entrenamiento de un modelo clasificador a través de características derivadas por medio de un método de agregación de las variables multicolineales, o por PCA podrían llevar a nuevos resultados en las evaluaciones de los modelos entrenados.



Implantación

La implantación del modelo entrenado se compone principalmente de dos secciones: **programación y diseño del tablero** de uso del modelo, y **puesta a disposición de uso (host)** para el usuario final.

Se hace uso de la programación del tablero a través del framework de alto nivel disponible denominado ***Streamlit***, siendo utilizado el sistema de host de la nube comunitaria de Streamlit.

[Liga](#)

# Revisar proyecto

---



En el presente trabajo se reconoce como área de oportunidad la incorporación de técnicas de **tratamiento para datos atípicos**.



Es posible realizar mayores pruebas en una segunda versión de este análisis en el área de creación de **variables agregadas** o de **creación sintética** para la evaluación de modelos clasificatorios alternativos.