

# Análisis de calidad de vinos

Data Science – Comisión 19145



Nicolás Boccardo

Facundo Cestari

German Grinberg

Alonzo David Linares Mora



# Tabla de Contenidos



- **Introducción**
  - Introducción.
  - Contexto.
  - Motivación.
- **Análisis Exploratorio de Datos.**
- **Análisis Bivariado y Multivariado.**
- **Algoritmo de Clasificación:**
  - Árbol de Decisión.
  - Random Forest.
  - KNN.
- **Futuras Líneas.**
- **Conclusión.**



# Introducción

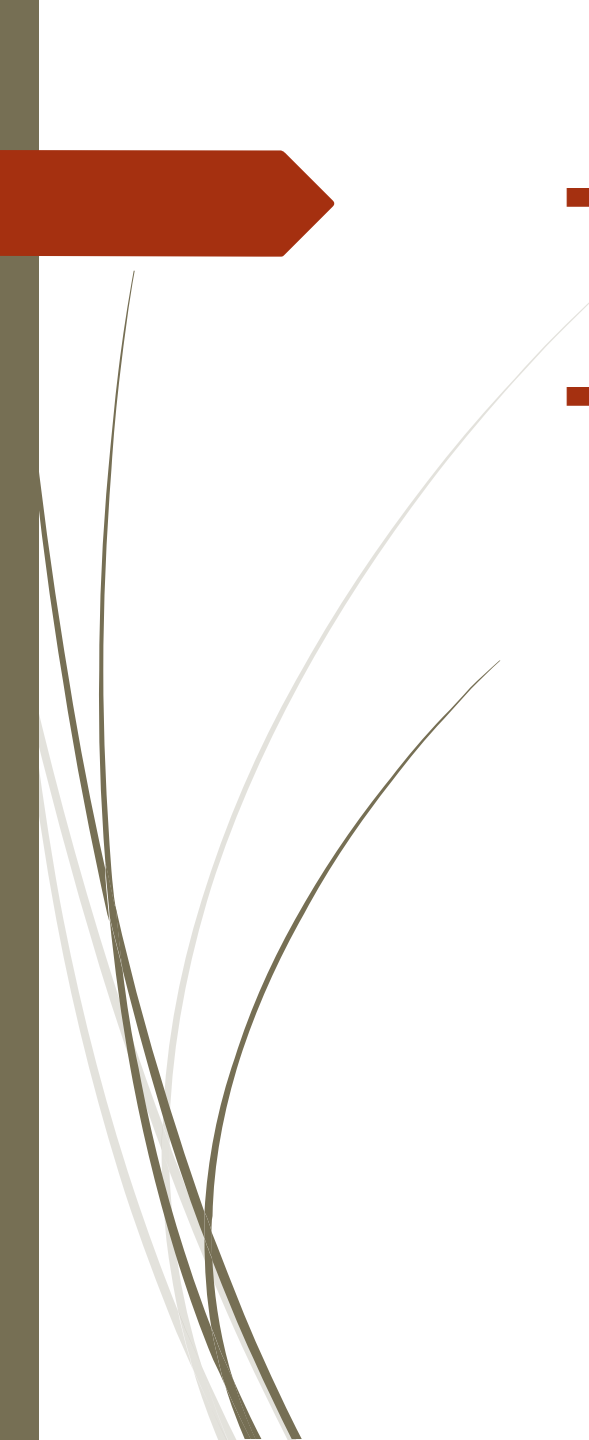


- En este trabajo de fin de curso, proponemos un enfoque de análisis de datos para predecir la calidad en el sabor del vino humano a partir de 11 características fisicoquímicas.
- Se considera un gran conjunto de datos con muestras de *vino rojo de Portugal* y se realiza un análisis exploratorio de datos.
- Luego se realiza un análisis bivariado y multivariado de los datos para saber la relación entre las distintas variables y nuestro target (la calidad del vino).
- Por ultimo, se realizan tres modelos de predicción(KNN, árbol de decisión y random forest) para predecir la calidad del vino de acuerdo a determinadas características fisicoquímicas y evaluar cual de los tres modelos es mas útil para este caso.



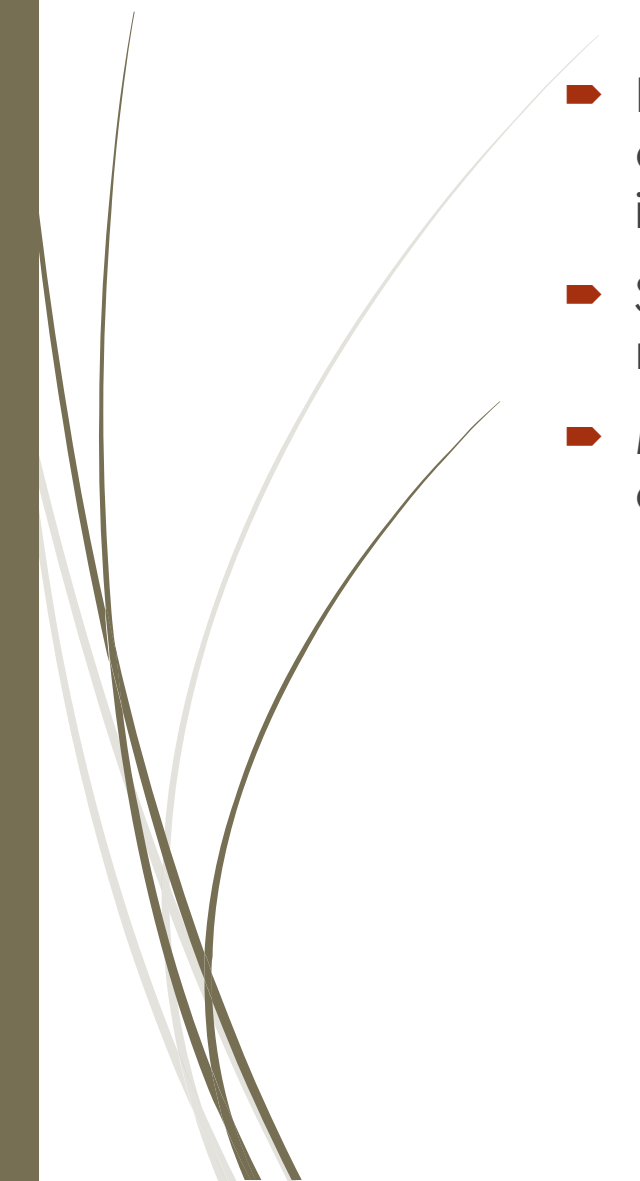
# Contexto

- Hoy en día el vino es disfrutado cada vez más por una amplia gama de consumidores.
- Portugal es uno de los diez principales países exportadores de vino, con una cuota de mercado del 3,17 % en 2020.
- Para respaldar su crecimiento, la industria del vino está invirtiendo en nuevas tecnologías tanto para la elaboración como para los procesos de venta.
- La certificación del vino y la evaluación de la calidad son elementos clave en este contexto.
- La evaluación de la calidad suele ser parte del proceso de certificación y se puede utilizar para mejorar la elaboración del vino (mediante la identificación de los factores más influyentes) y para estratificar vinos como marcas premium (útil para fijar precios).
- La certificación del vino se evalúa generalmente mediante pruebas fisicoquímicas y sensoriales.

- 
- Los avances en las tecnologías de la información han hecho posible recopilar, almacenar y procesar conjuntos de datos masivos, a menudo muy complejos.
  - Todos estos datos contienen información valiosa, como tendencias y patrones, que se pueden utilizar para mejorar la toma de decisiones y optimizar las posibilidades de éxito.



# Motivación

- La construcción del modelo de clasificación es valiosa no solo para las entidades de certificación, sino también para los productores de vino e incluso para los consumidores.
  - Se puede utilizar para apoyar las evaluaciones del vino del enólogo, mejorando potencialmente la calidad y la velocidad de sus decisiones.
  - Medir el impacto de las pruebas fisicoquímicas en la calidad final del vino es útil para mejorar el proceso de producción.
- 



# Análisis exploratorio de los datos



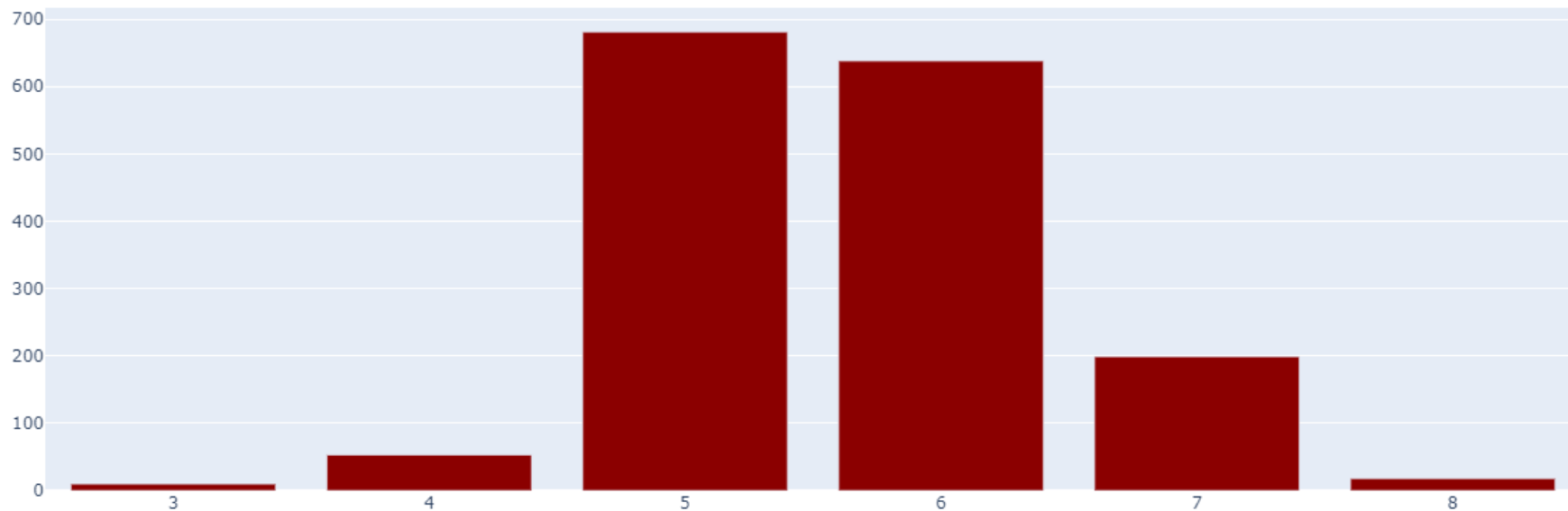
# Explicación de dataset



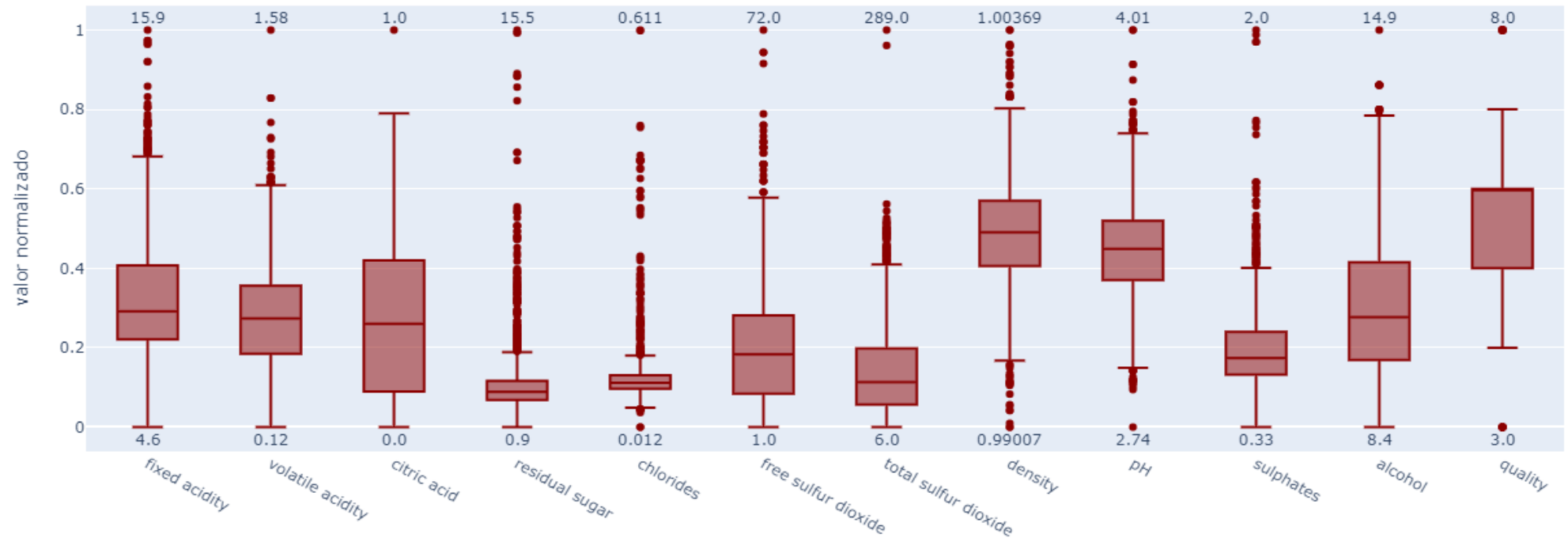
- **Fixed acidity:** La mayoría de los ácidos no volátiles involucrados con el vino no volátiles (no se evaporan fácilmente).
- **Volatile acidity:** La cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre.
- **Citric acidity:** Encontrado en pequeñas cantidades, el ácido cítrico puede agregar 'frescura' y sabor a los vinos.
- **Residual sugar:** La cantidad de azúcar que queda después de que se detiene la fermentación. Es raro encontrar vinos con menos de 1 gramo/litro. Los vinos con más de 45 gramos/litro se consideran dulces.
- **Chlorides:** La cantidad de sal en el vino.
- **Free sulfur dioxide:** La forma libre de SO<sub>2</sub> existe en equilibrio entre el SO<sub>2</sub> molecular (como gas disuelto) y el ion bisulfito; previene el crecimiento microbiano y la oxidación del vino.
- **total sulfur dioxide:** Cantidad de formas libres y ligadas de SO<sub>2</sub>; en bajas concentraciones, el SO<sub>2</sub> es mayormente indetectable en el vino, pero en concentraciones de SO<sub>2</sub> libres superiores a 50 ppm, el SO<sub>2</sub> se vuelve evidente en la nariz y el sabor del vino.
- **Density:** La densidad del agua es cercana a la del agua dependiendo del porcentaje de contenido de alcohol y azúcar.
- **pH:** Describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3 y 4 en la escala de pH.
- **Sulphates:** Un aditivo del vino que puede contribuir a los niveles de dióxido de azufre (SO<sub>2</sub>), que actúa como antimicrobiano y antioxidante.
- **Alcohol:** Porcentaje de contenido de alcohol del vino.
- **Quality:** Variable de salida (basada en datos sensoriales, puntuación entre 0 y 10).



Calidades de vinos analizados



- Contamos con una base de datos de **1599 diferentes vinos tintos** agrupados según su **calidad del 1 al 10**

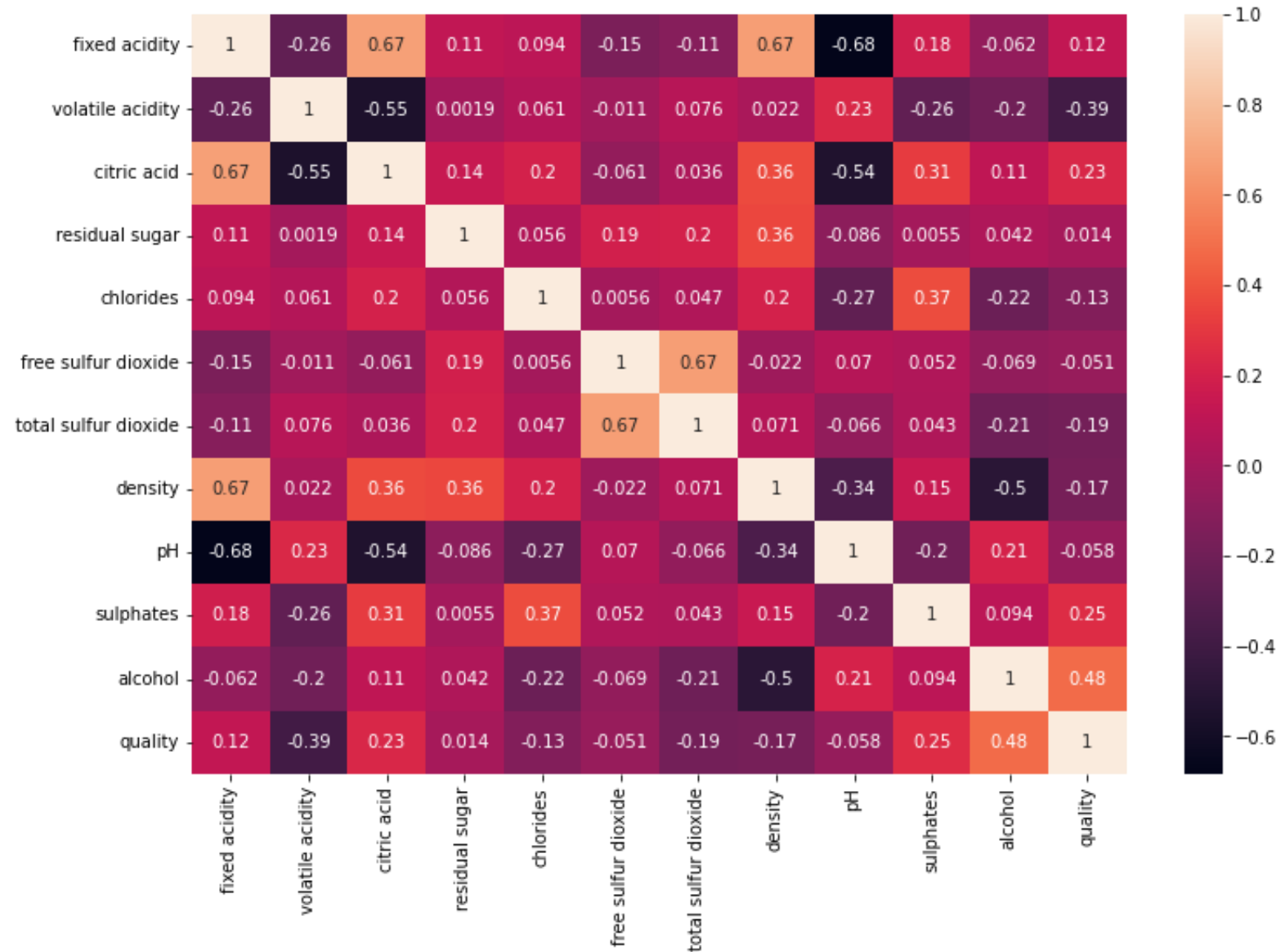


- Esta base de datos cuenta con **11 características analizadas** sobre cada uno de los vinos ensayados

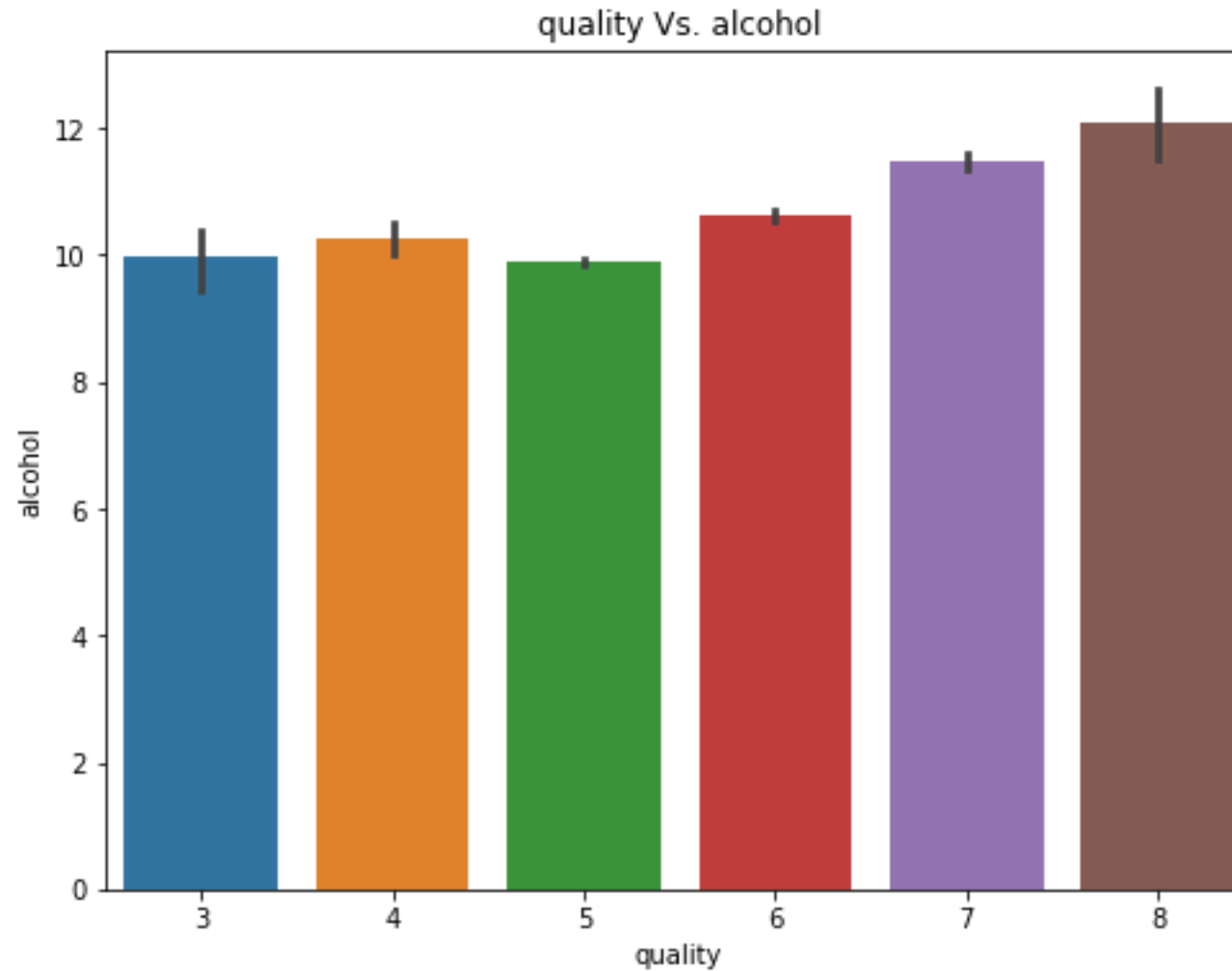


# Análisis bivariado/multivariado

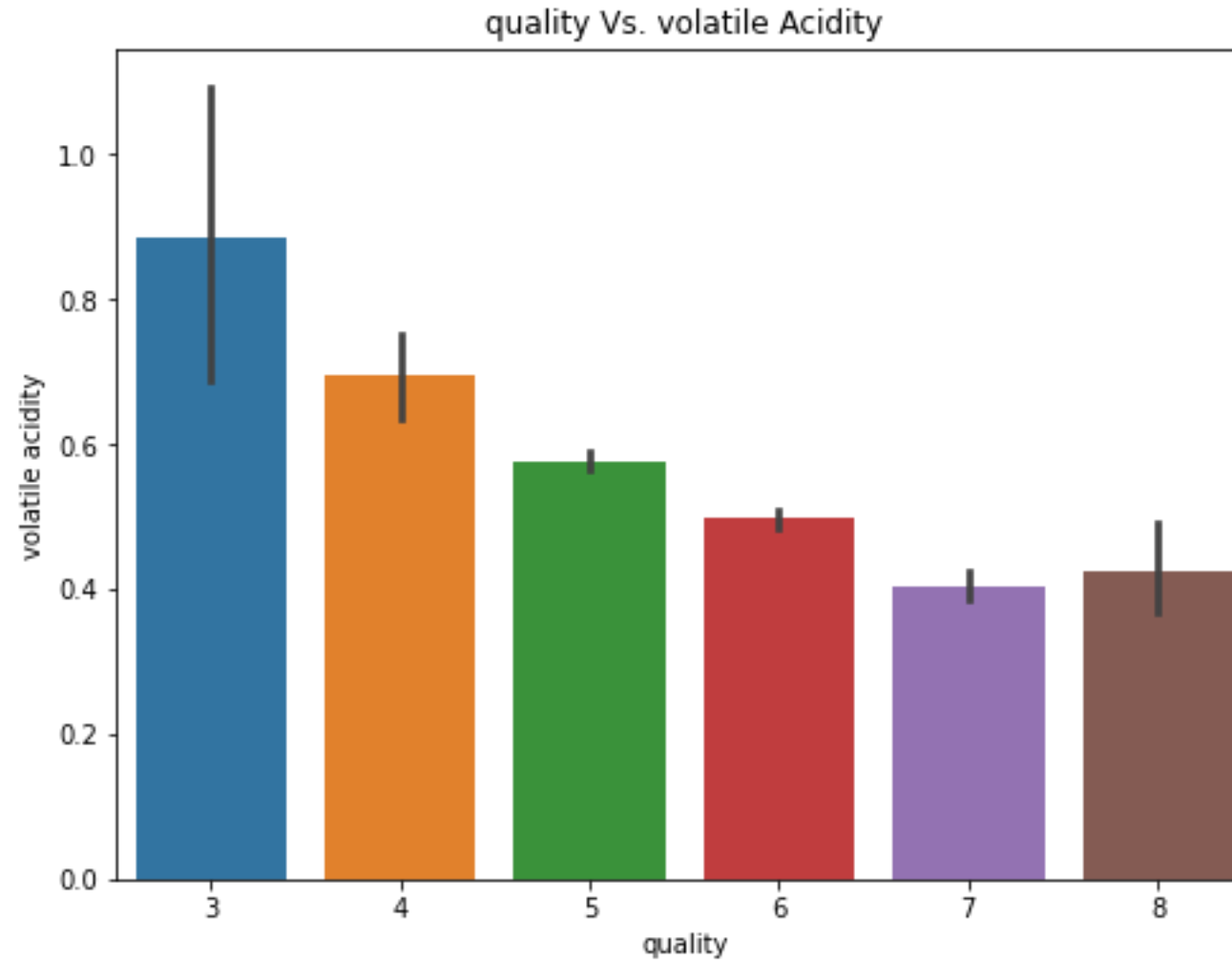
- Puede notarse en los datos una correlación significativa entre la **calidad** del vino y la cantidad de **sulfatos**, **acidez volátil** y **alcohol**



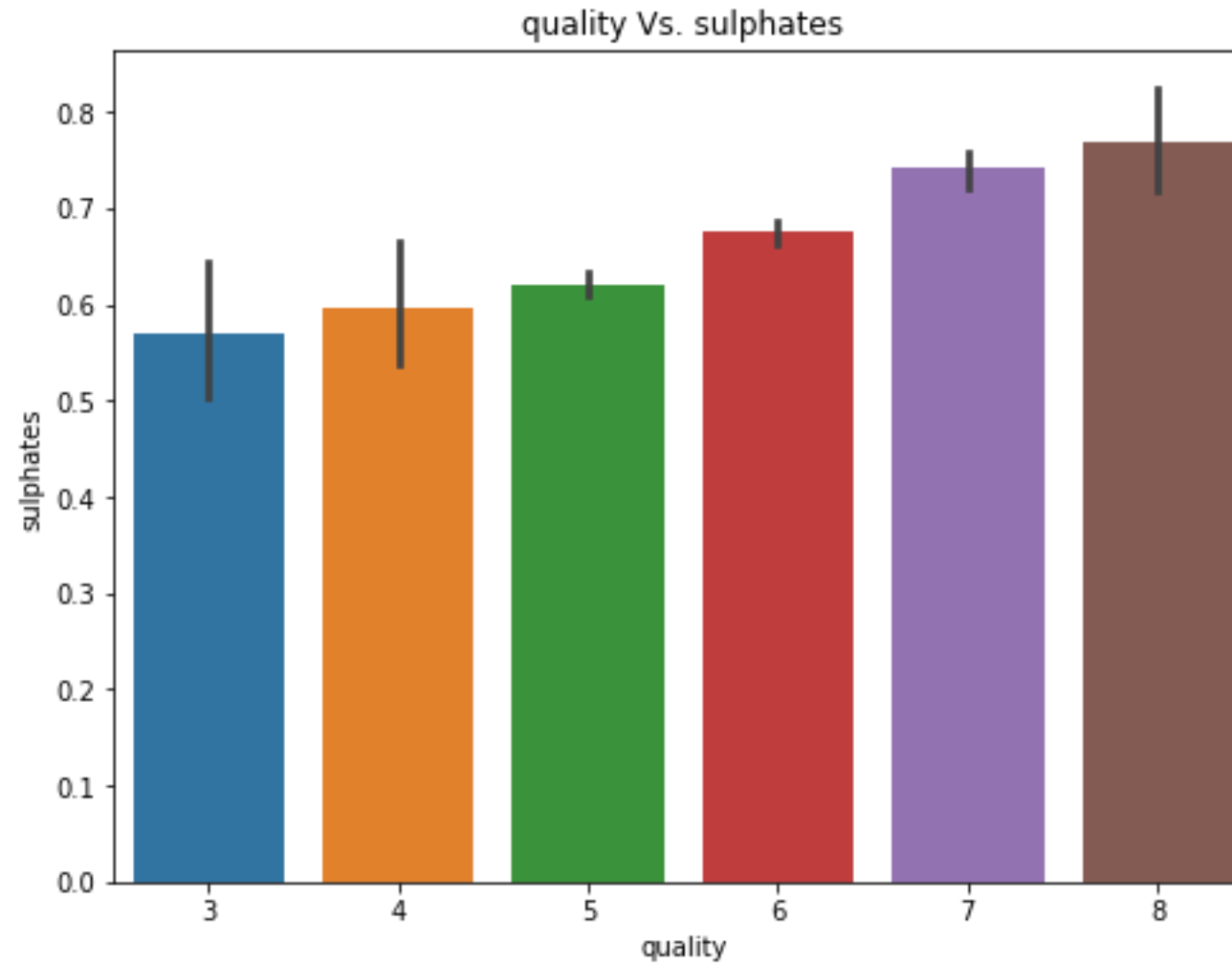
- Los vinos de mayor **calidad** cuentan con **mayor** cantidad de **alcohol**





- ➡ Los vinos de mayor **calidad** cuentan con **menor** cantidad de **ácidos volátiles**



- Los vinos de mayor **calidad** cuentan con **mayor** cantidad de **sulfatos**



- 
- 
- En base a este análisis, se toma la decisión de realizar un árbol de decisión para la variable 'quality', con énfasis en las tres variables 'sulphate', 'alcohol' y 'volatile acidity'.







# Árbol de decisión

Detección de casos positivos



# Motivación

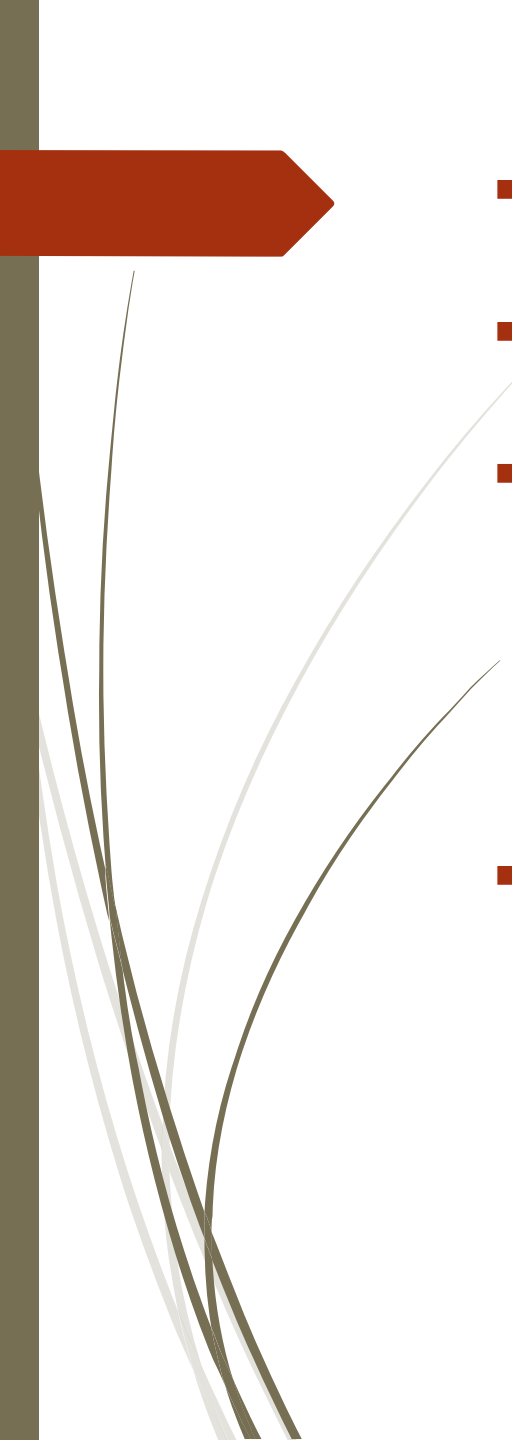
- Dadas las muestras de nuestra base de datos, se construye un predictor del tipo “árbol de decisiones” para identificar, en líneas generales, las principales características de un buen vino, detectables de entre los ensayos disponibles
- 

- 
- Primero se realiza un agrupamiento en los valores de nuestro target para hacer mas fácil el análisis y el modelo predictivo, siguiendo la relación:
    - Aquellos vinos con calidad entre 0 y 4 son 'malo'
    - Aquellos vinos con calidad entre 4 y 6 son 'regular'
    - Aquellos vinos con calidad entre 7 y 8 son 'bueno'
  - Luego del agrupamiento, las cantidades resultan:

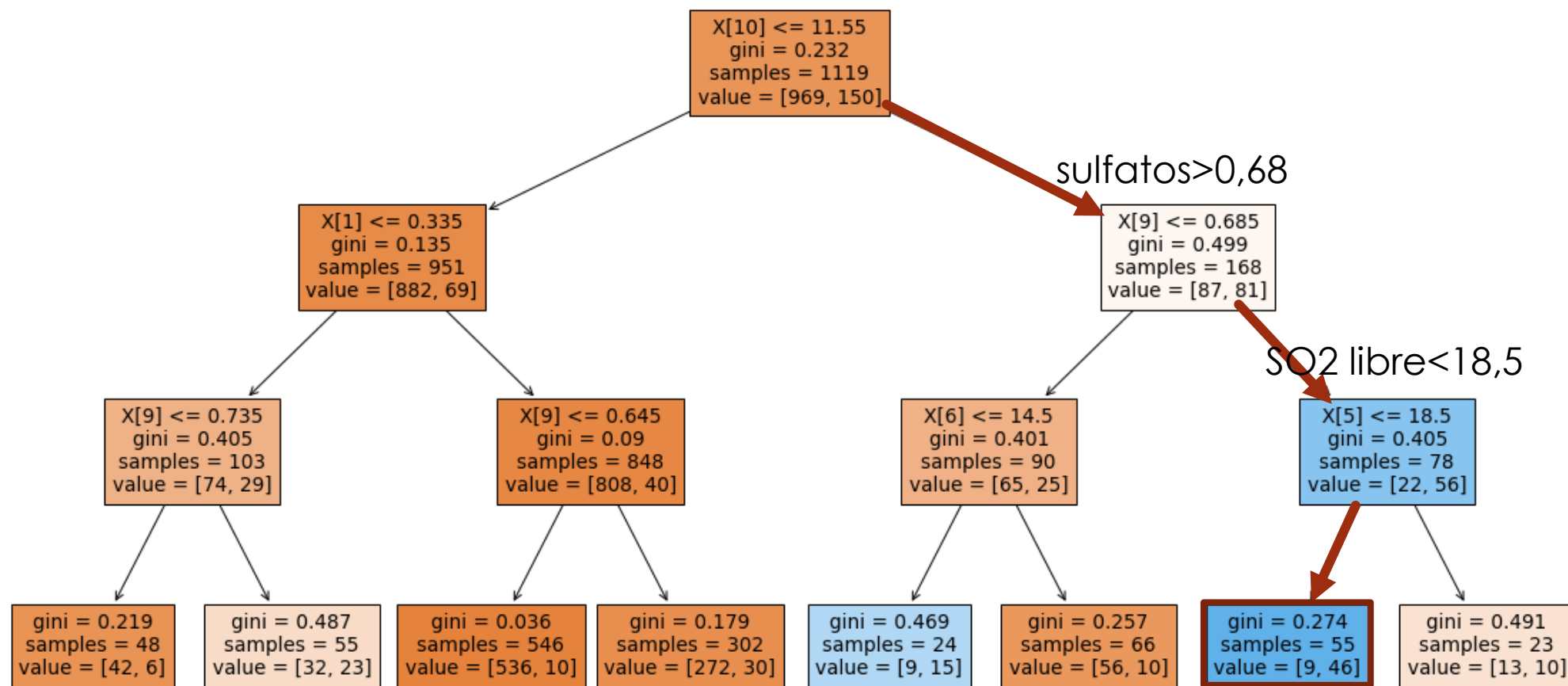
```
df['quality'].value_counts()
✓ 0.2s
```

regular	1319
bueno	217
malo	63

Name: quality, dtype: int64

- 
- A continuación, se separa el dataset en dos: 'X' e 'y'. 'y' contiene la columna de quality y 'X' contiene el resto del dataset original.
  - Luego definimos que un 70% de nuestro dataset será para entrenar el modelo y el 30% restante será para testearlo
  - Se crea un árbol de decisión con los siguientes parámetros:
    - Max\_Depth: 3.
    - random\_state: 42.
    - Min\_sample\_Split: 10.
  - Este modelo arroja los siguientes resultados de acierto:
    - 90% de acierto en el dataset de entrenamiento.
    - 85% de acierto en el dataset de testeo.

Alcohol>11.55





- Los **mejores vinos** se caracterizan por:
  - un contenido de **alcohol mayor a 11,5%**
  - Un contenido de **sulfatos mayor a 0,68**
  - Un contenido de **SO2 libre menor a 18,5**
- Si enviamos **un vino al azar** a calificar, la probabilidad de hallar un vino considerado **bueno** es del **14%**
- Si enviamos **un vino candidato** a calificar, la probabilidad de hallar un vino considerado **bueno** es del **54%**

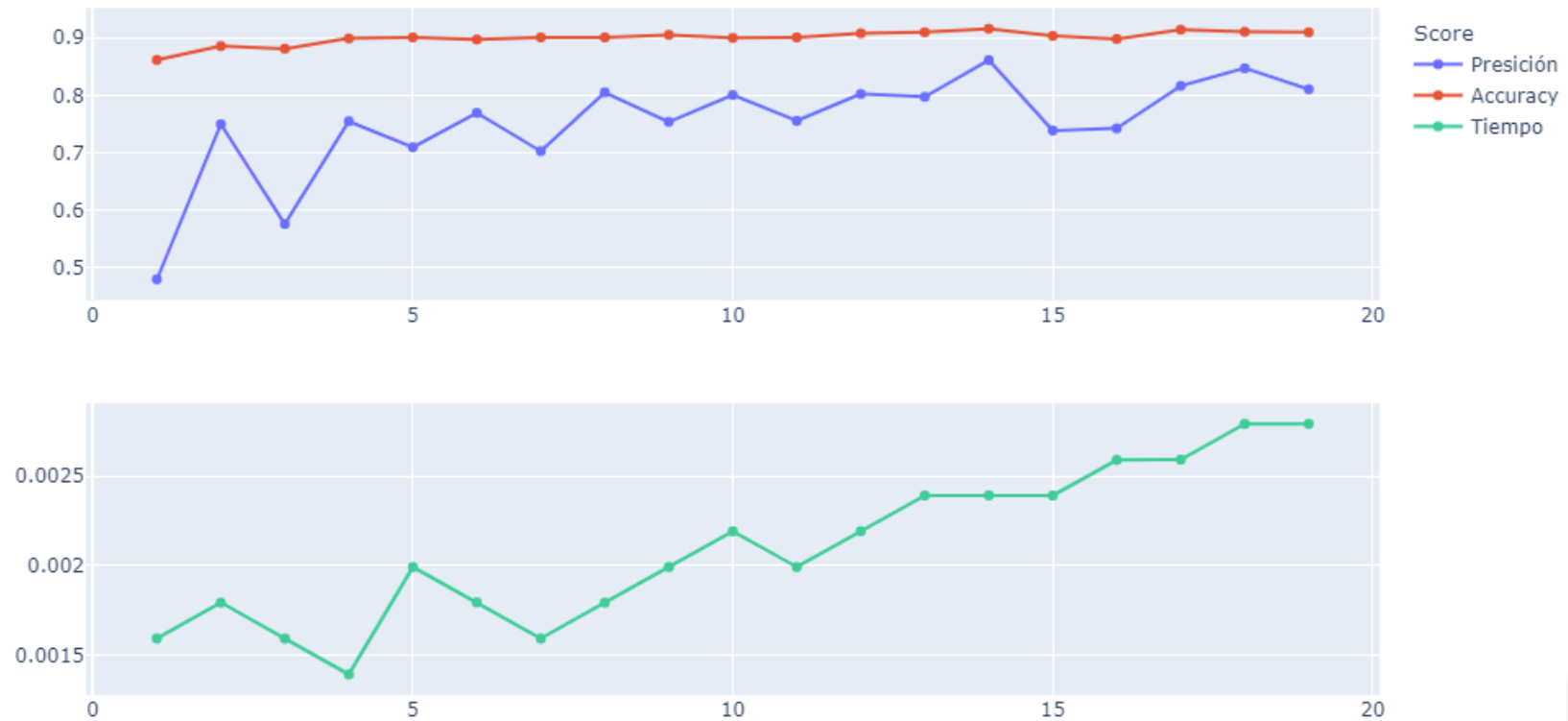


# Predictor – Random Forest

- Abandonando la necesidad de reconocer los funcionamientos internos del predictor en forma clara en base a las características del vino, es posible construir un mejor predictor de vinos buenos utilizando algoritmos de mayor complejidad.
- Estos algoritmos podrán utilizarse para detectar mejores candidatos para pasar a categorización a partir de los 11 ensayos fisicoquímicos utilizados, siendo la categorización de los vinos un proceso costoso y con demora
- Esta calificación funcionará como target para la producción, esto no implica que un vino potencialmente bueno será detectado sino que el vino que sea detectado, será bueno
- Se elabora un modelo de decisión más complejo utilizando la técnica de Random Forest

# Predictor Random Forest


Random Forest - scores vs estimadores





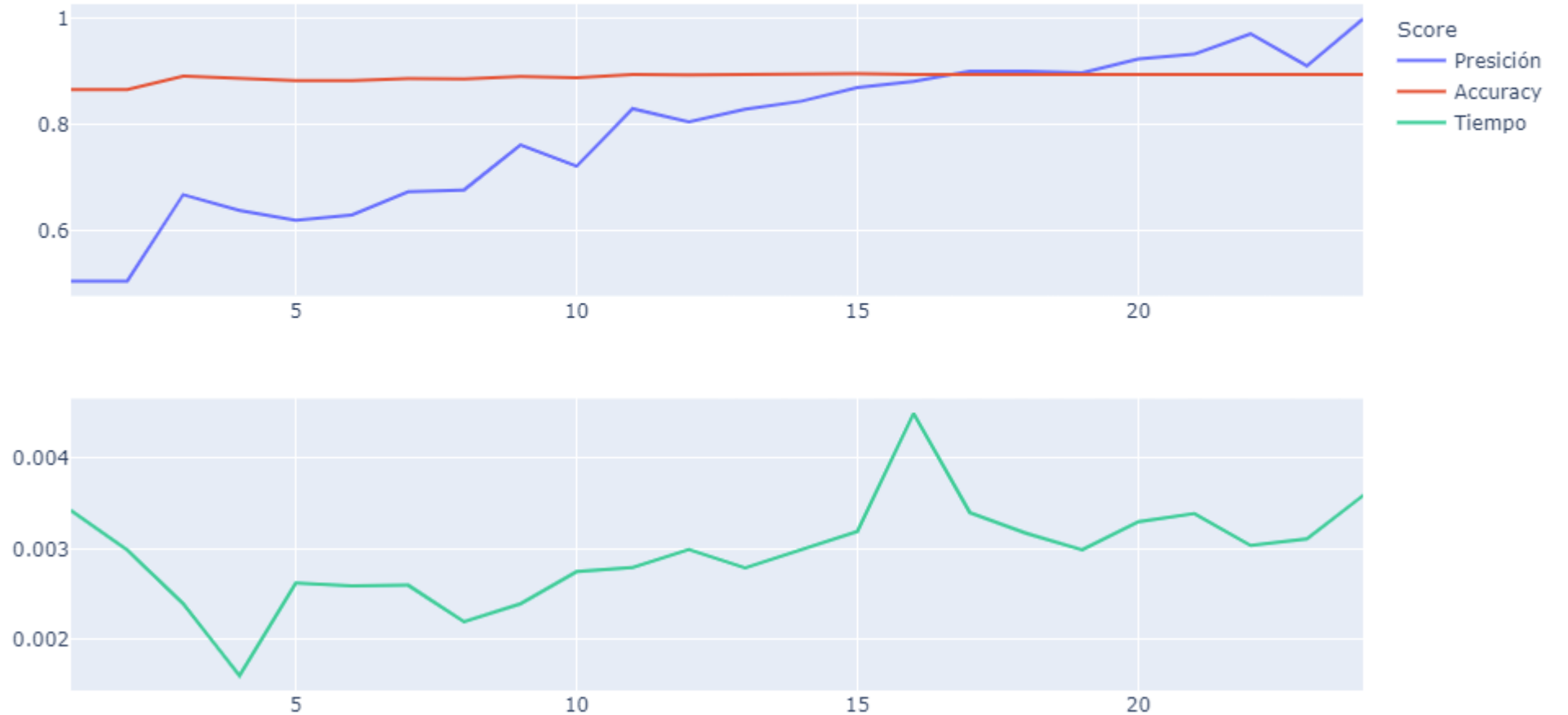


# Predictor - KNN

- Siendo la clasificación de los vinos un proceso subjetivo complejo, no es sencillo establecer reglas en base a sus características.
  - Si es razonable pensar que si la experiencia subjetiva de un vino es buena, la experiencia basada en los sentidos de un vino similar también lo será.
  - Basado en esto abandonaremos el enfoque de un predictor en base a reglas en función de uno basado en la similaridad de muestras.
- 

# Predictor - KNN

KNN - scores vs estimadores

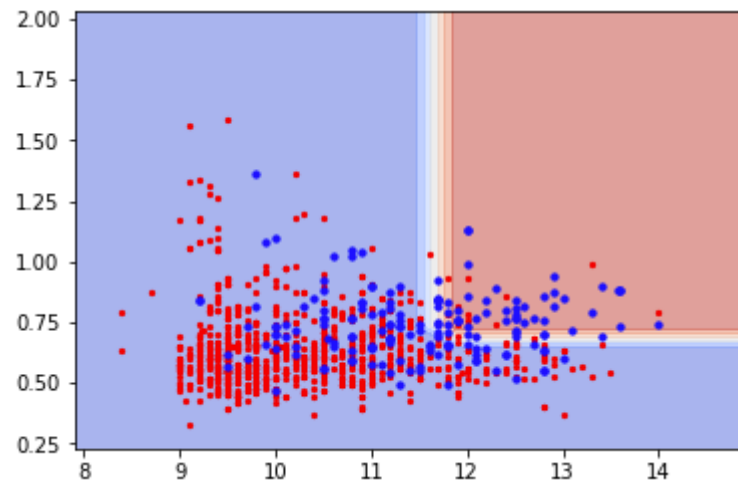


# Predictor - KNN

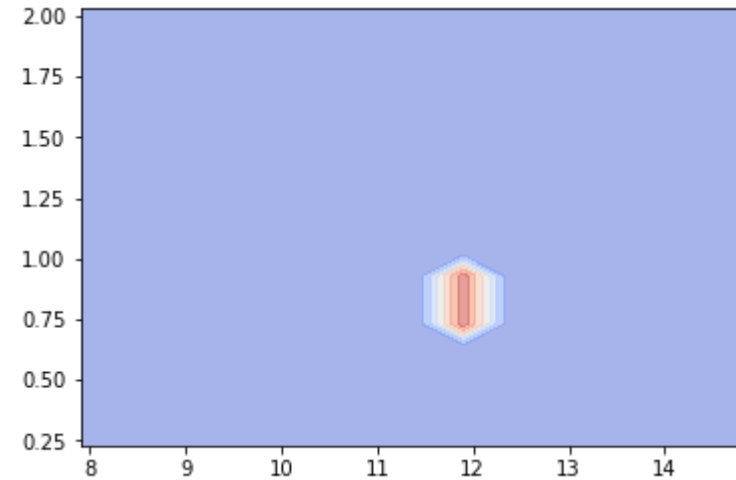
- Se cuenta con un predictor con una precisión sobre la muestra de ensayo es de un 95% y la exactitud de un 90%
- Pueden relevarse aquellos conjuntos de características que darán por resultado un buen vino

## Área de detección para alcohol y sulfatos

Corte para reglas básicas 55%




Corte para puntos objetivo 95%





# Futuras Líneas

- Como ampliación al trabajo realizado, se puede realizar una detección y clasificación de distintos tipos de vinos dentro del dataset (por ejemplo, Cabernet, Malbec, Bonarda, Merlot, Etc.)
  - Se podría realizar esta detección empleando grafico de silueta.
  - Luego, dentro de esos tipos de vinos, se podría implementar un detector para vinos de buena calidad.
- 



# Conclusiones

- En este trabajo se evaluaron 3 algoritmos distintos (Random Forest, Árbol de decisión y KNN).
  - Concluimos que para nuestro dataset el algoritmo que mejor se ajusta es el KNN, ya que la clasificación de los vinos es un proceso subjetivo y no es sencillo establecer reglas en base a sus características pero es razonable pensar que si la experiencia subjetiva de un vino es buena, la experiencia basada en los sentidos de un vino similar también lo será. Siguiendo esa línea, el algoritmo KNN resulta bueno en este caso ya que realiza agrupamiento de datos de acuerdo a la clase mas común entre sus “vecinos”.
  - Consideramos que a lo largo de este trabajo hemos abarcado los contenidos vistos a lo largo del curso de Data Science. El trabajo resultó de gran utilidad para poder terminar de comprender y reforzar los temas vistos en clase.
- 