Trabajo práctico especial Fundamentos de las Ciencias de Datos

Análisis de datos sobre variedades de vinos

Grupo: 6

Integrantes: Luciano Montes, Cortavarria Cristian, Agustin Baldassarri

Introducción:

El presente documento es parte del trabajo práctico especial de la cátedra Fundamentos de las Ciencias de Datos. En el mismo se realiza un informe detallado de lo que resultó del análisis del dataset “winequality_br.csv” que fue provisto por la cátedra.

El objetivo del trabajo fue plasmar los conocimientos obtenidos sobre los contenidos dados utilizando las herramientas de análisis de datos provistas para interpretar y evaluar adecuadamente los resultados obtenidos.

En el dataset tenemos información acerca de dos tipos de variedad de vino la Garnacha que es un vino tinto que se caracteriza por su sabor afrutado, cuerpo medio y notas de frutos rojos y el Riesling que es un vino blanco se caracterizan por su frescura, acidez y aromas frutales. De estas variedades tenemos datos de su calidad como producto y varias variables cuantitativas
que describen la composición de los vinos, siendo estas características las siguientes:

- type: tipo de uva con la que se elabora el vino.

- fixed acidity: cantidad de ácidos no volátiles presentes en el vino, medida en gramos por litro.
  La variable fixed acidity contiene el conjunto de ácidos naturales presentes en el vino (incluyendo al ácido cítrico). Estos no son ácidos volátiles, sino que son ácidos naturales que aportan al balance de acidez y el gusto del vino. medida en gramos por litro.

- volatile acidity: La acidez volátil, son aquellos ácidos volátiles o gaseosos presentes en el vino. Los valores normales los sitúan entre 0.3 g/l y 0.6 g/l. En general si este valor sobrepasa 1 g/l se suele indicar que el vino está "picado", es decir presenta ciertos aromas y gustos avinagrados desagradables. Medida en gramos por litro.

- citric acid: El ácido cítrico es uno de los ácidos naturales que se pueden encontrar en el vino y es también parte de los ácidos que componen la fixed acidity. Este le aporta al vino sabores cítricos. Medido en gramos por litro.

- residual sugar: El azúcar residual es azúcar libre en el vino que no se descompuso para formar alcohol. Esta es responsable del dulzor del vino. Esta pueden clasificar el vino en seco (0-9), semiseco (9-18), semidulce (18-50) y dulce (mayor 50). Medida en gramos por litro.

- chlorides: Los cloruros son la concentración de sales minerales presentes en el vino y que pueden afectar a su calidad y sabor, medida en gramos por litro.

- free sulfur dioxide: cantidad de dióxido de azufre que no está ligado químicamente en el vino. Este componente actúa como un agente antimicrobiano y antioxidante, ayudando a preservar el color, el aroma y el sabor del vino al prevenir la oxidación y el crecimiento de microorganismos indeseados.

  total sulfur dioxide: Este valor representa la cantidad total de dióxido de azufre presente en el vino, tanto libre como combinado con otros compuestos (por ejemplo, con el azúcar y otros elementos del vino). El SO_2 total es importante porque contribuye a la estabilidad y longevidad del vino, medida en miligramos por litro.

- density: La densidad de un vino es percibida como la estructura del vino o espesor en boca. Este valor suele ser similar a la densidad del agua que ronda los 1 g/ml, medida en gramos por mililitro.

- pH: medida de la acidez o alcalinidad del vino.

- sulphates: Los sulfatos son compuestos de azufre presentes en el vino y afectan a su estabilidad y longevidad. Además, los sulfatos contribuyen a la

preservación y ayudan a controlar el crecimiento de bacterias durante el proceso de fermentación. Este valor se mide en gramos por litro.

- alcohol: El contenido alcohólico es la cantidad de etanol en el vino, y es una medida de su fuerza y cuerpo. Un vino con mayor contenido de alcohol generalmente se percibe con más cuerpo y calidez en boca. Los valores en esta columna están medidos en porcentaje de volumen (% vol).

- quality: puntuación del vino, con una escala que va de 0 a 10.

Como objetivo de estudio se pusieron las siguientes hipótesis que conforman el conjunto de resultados obtenidos:

- hipótesis 1: La calidad de un vino está dada de manera independiente al volumen alcohólico del mismo, se dice que este es un valor objetivo y no se utiliza como un factor relevante a la hora de asignar la calidad

- hipótesis 2: Los factores que caracterizan a un vino deberían ser distintos en dependencia del tipo de vino.

    - hipótesis 2.1: Las variables total sulfur dioxide, residual sugar, y citric acid aumentan la probabilidad de que el vino sea Riesling.

    - hipótesis 2.2: Mientras que las variables alcohol y density favorecen la clasificación de Garnacha.

- hipótesis 3: El ácido cítrico impacta en la calidad de los vinos rieslings pero no así en los vinos garnacha.

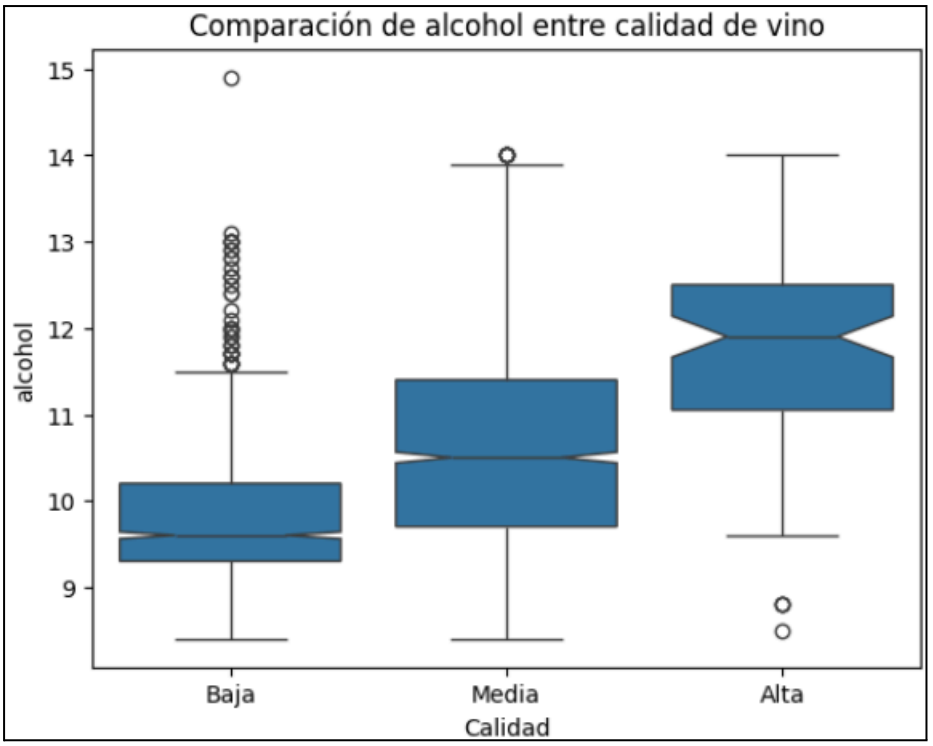
- hipótesis 4: El ácido volátil no impacta en la calidad del vino de la misma manera para los distintos tipos

Resultados:

-Hipótesis 1:

Para la hipótesis 1 utilizamos gráficos como el diagrama de cajas, comúnmente conocido como boxplot, que es una herramienta estadística, que representa

gráficamente la distribución de un conjunto de datos numéricos del estilo Boxplot combinando la variable Alcohol y los grupos de calidad. Luego utilizamos métodos para comparar los grupos. Estos métodos son conocidos como: Shapiro-Wilk, Levene, Mann-Whitney U y Kruskal-Wallis, los cuales son test paramétricos y no paramétricos que nos sirven para evidenciar o no la información que deducimos de los gráficos.



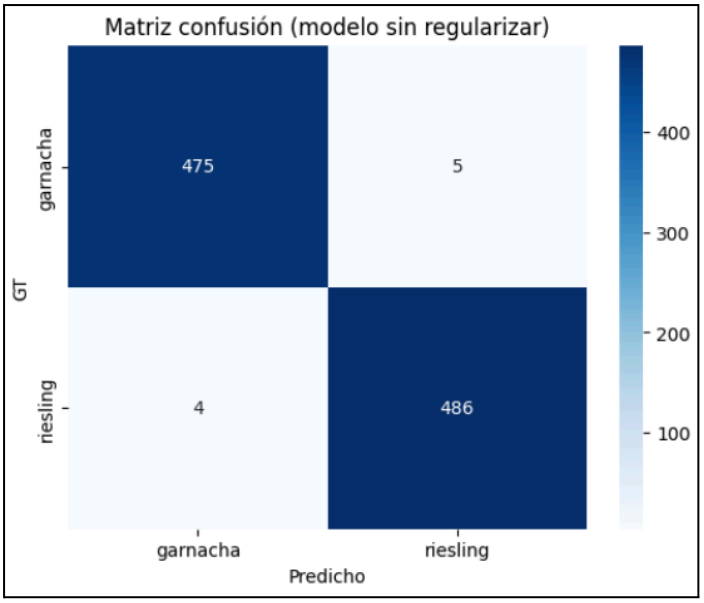
Boxplot de la variable alcohol para cada grupo de calidad

-Hipótesis 2:

Utilizamos un algoritmo de machine learning, en nuestro caso, conocido como Regresión Logística. Un modelo predictivo el cual está diseñado para problemas de clasificación, donde el objetivo es predecir la probabilidad de que una observación pertenezca a una categoría específica. Mediante esta herramienta obtendremos un modelo, al que luego le pasamos datos de ejemplo y nos va a predecir a qué grupo pertenece, por ejemplo.

Analizando el modelo que obtuvimos, logramos hallar las variables que caracterizan a un grupo y a otro, logrando resultados interesantes.

En conjunto con esto utilizamos gráficos de barras, matrices de confusión y gráficos Box Plots, test paramétricos y no paramétricos para hallar evidencias para verificar si se cumple o no la hipótesis.



Matriz de confusión

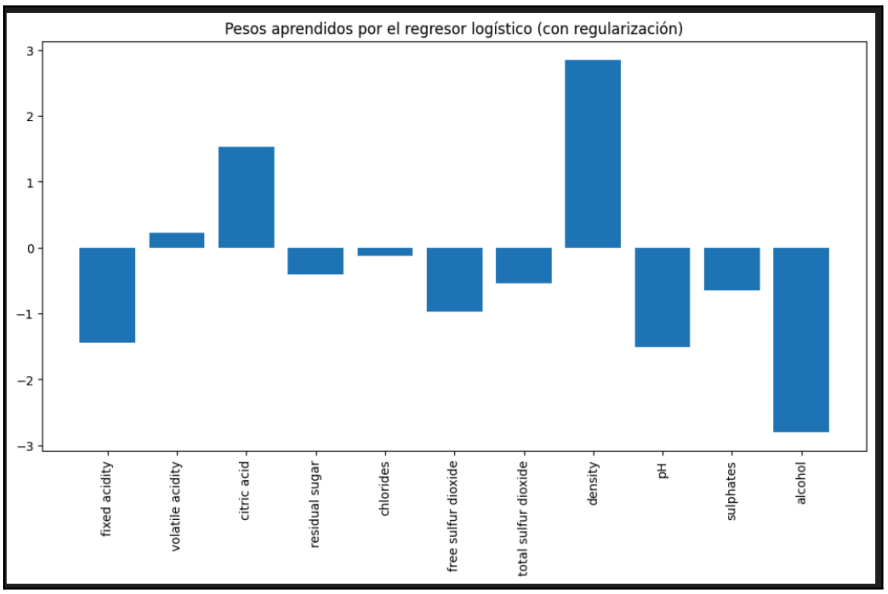


Gráfico de barras de los coeficientes aprendidos por el regresor logistico

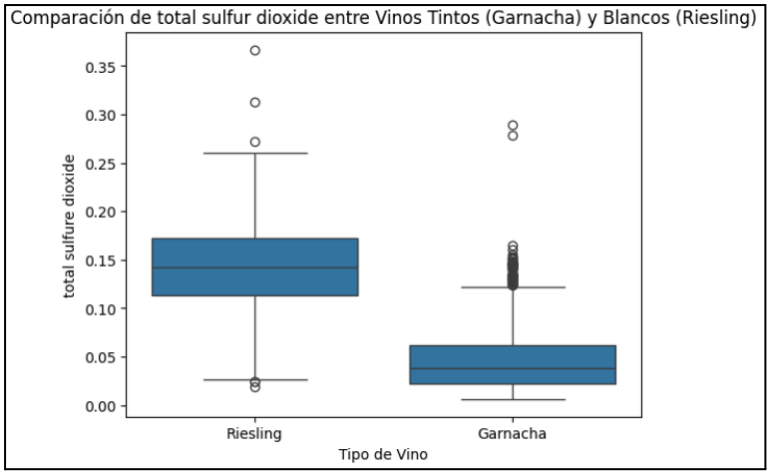
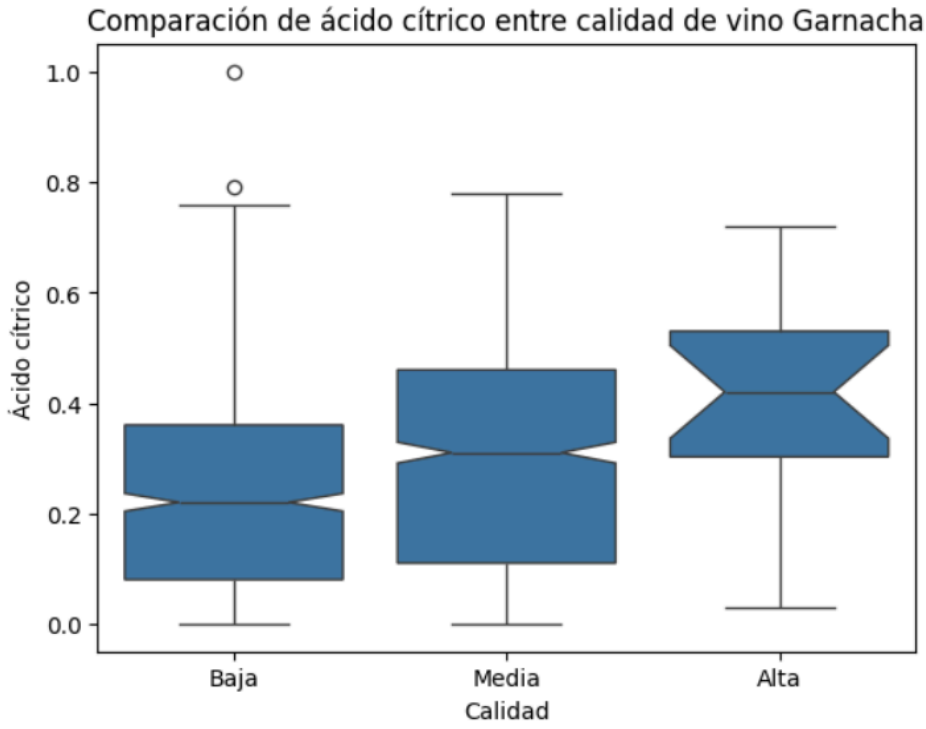


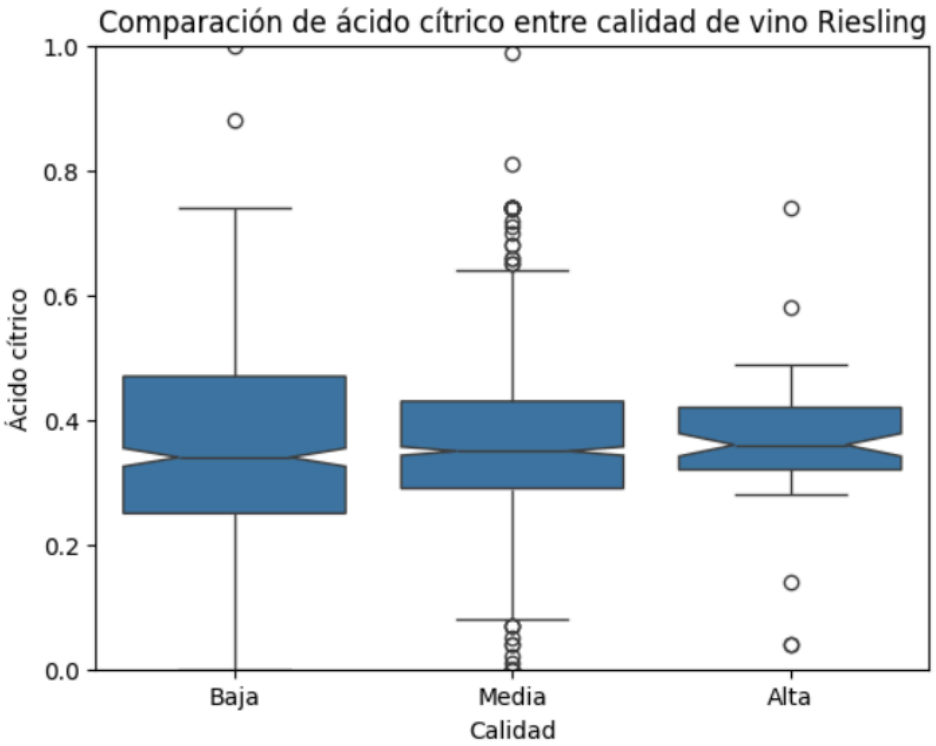
Gráfico Boxplot del total sulfur dioxide dividido por vinos riesling y garnacha

-Hipótesis 3:

“El ácido cítrico no impacta en la calidad del vino de la misma manera para los distintos tipos.”

En el desarrollo de la hipótesis 3 lo que se hizo fue separar el set de datos entre los tipos de vino para ver si el comportamiento es distinto. Para esto se generaron dos gráficos boxplot donde se observa una clara diferencia en la distribución del ácido cítrico cuando se trata de vinos de tipo Garnacha, no así cuando vemos el gráfico de Riesling.



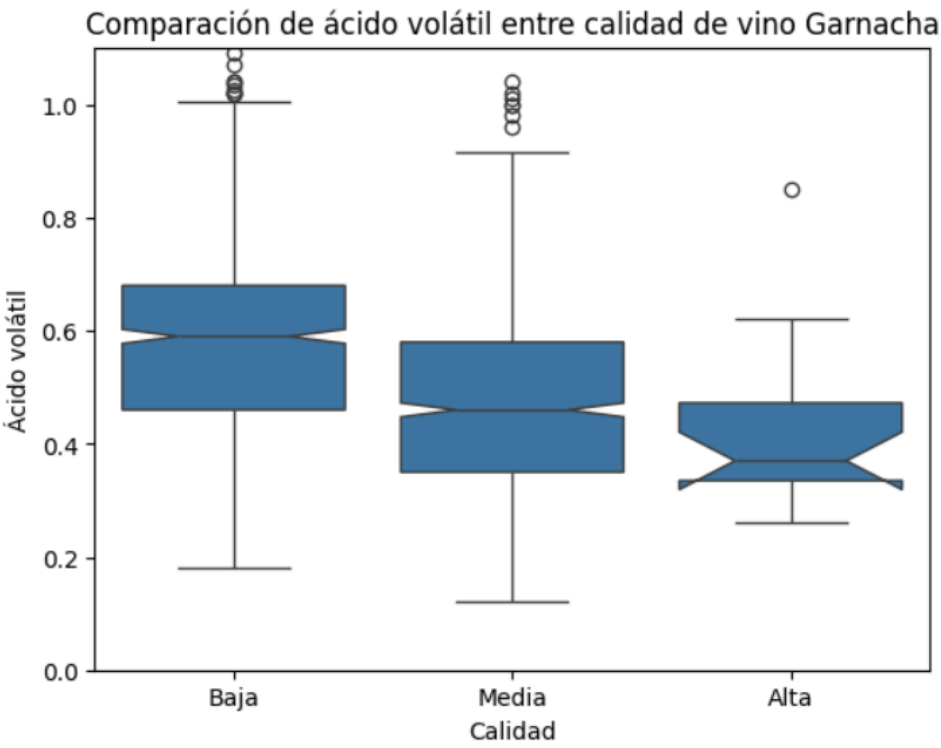


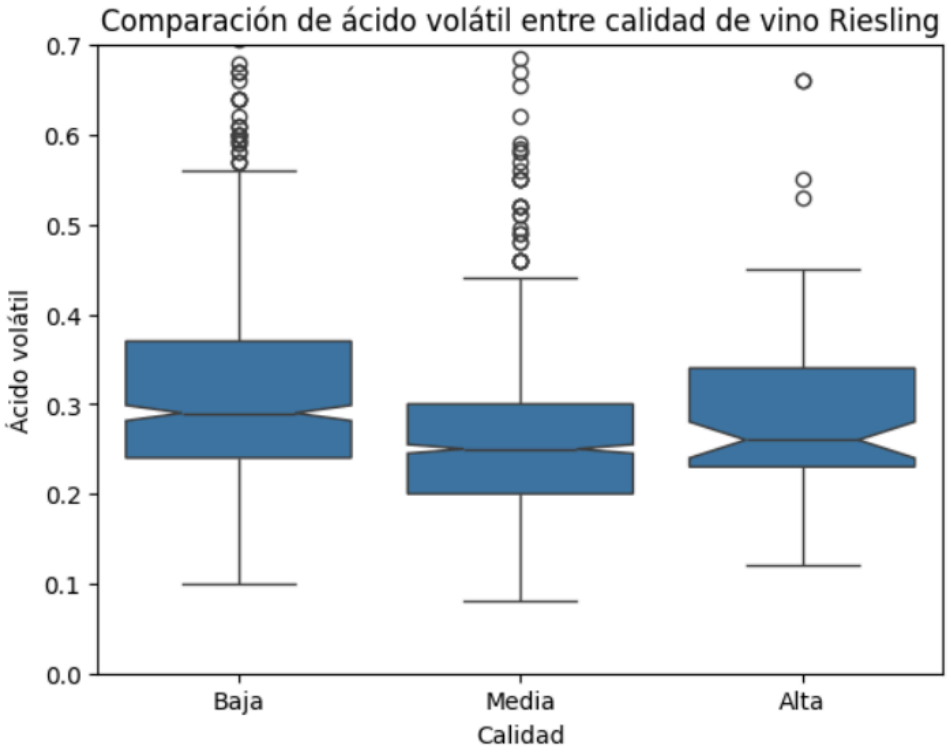
Para conseguir evidencia de la hipótesis se implementaron las herramientas estadísticas para validar si la distribución presentaba normalidad y homocedasticidad. Con estos datos se decidió usar test de Mann-Whitney U o Kruskal-Wallis según corresponda, confirmando nuestra hipótesis y logrando obtener evidencia que en los vinos garnacha un mayor contenido de ácido cítrico influye para tener una calidad alta. Mientras que en los vinos Riesling no varía el ácido cítrico con respecto a la calidad.

-Hipótesis 4:

“El ácido volátil no impacta en la calidad del vino de la misma manera para los distintos tipos.”

Para esta hipótesis, similar a lo planteado en la hipótesis 3, se generó dos gráficos boxplot, uno para cada tipo de vino. Donde se observa nuevamente una tendencia en los vinos de tipo Garnacha.





Se volvieron a usar test de Mann-Whitney U o Kruskal-Wallis según la distribución y homocedasticidad. En los vinos de tipo Garnacha se obtuvo evidencia para decir que la cantidad de ácido volátil es inversamente proporcional a la calidad del vino, mientras que en los de tipo Riesling, si bien se consiguió evidencia para decir la

cantidad de ácido volátil es significativamente diferente entre las calidades, no se observa la misma tendencia que en Garnacha, es decir, la cantidad no es inversamente proporcional a calidad del vino.

REFERENCIAS Y BIBLIOGRAFÍA

- https://rpubs.com/Cardroba/643591#:~:text=Cloruros,-hist(dataB%24Cloruros&text=Una%20de%20los%20principales%20componentes,(0.04577)%20mg%2FL.
- https://www.kaggle.com/code/mgmarques/wines-type-and-quality-classification-exercises/notebook#Data-Engineering---Cleaning,-Transforming,-Selection-and-Reduction
- https://www.bodegastrespiedras.com/acidez-del-vino/
- IAs como chat GPT, consensus.app.