UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Predicting Wine Quality Using Machine Learning

Manuel Guerrero, Andres Ruiz, Juan Oviedo
Department of Systems Engineering, Universidad Distrital Francisco Jose de Caldas
Emails: mrguerreroc@udistrital.edu.co, afruizv@udistrital.edu.co,
jeoviedos@udistrital.edu.co

Background

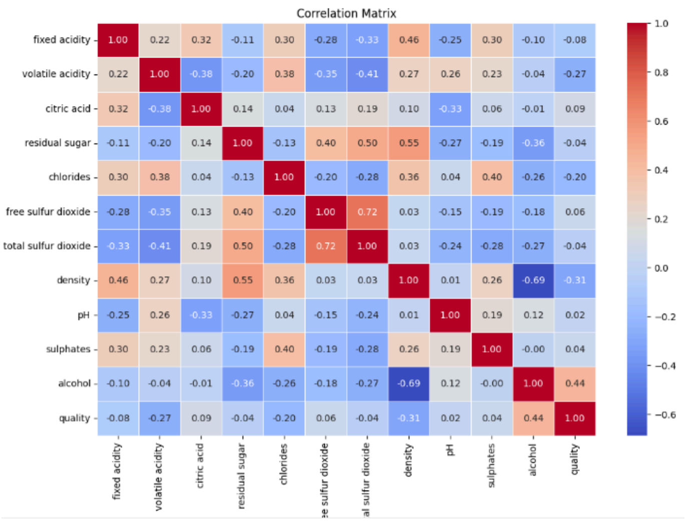
Wine is a liquor that has been used for several purposes, cooking, tasting, general consumption, etc. His main component is the grape, has several chemical factors during its manufacture that allow attributing a quality standard, the objective is to determine with a predictive model the quality of the wine from its chemical factors such as: PH, Citric Acid, Alcohol, etc.

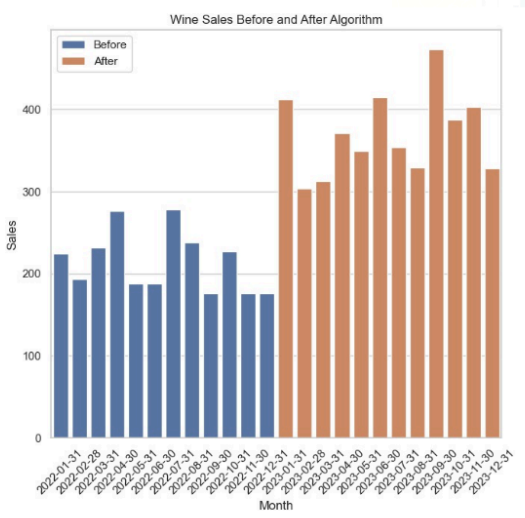
Previous solutions

This problem has been previously covered by a variety of people, expressing unique solution models, with xgboost, most taking as a single reference, the dataset provided by Kaggle, and from this performed an exploratory analysis, from which they looked for correlations and outliers to treat the data appropriately competence.

Our solution

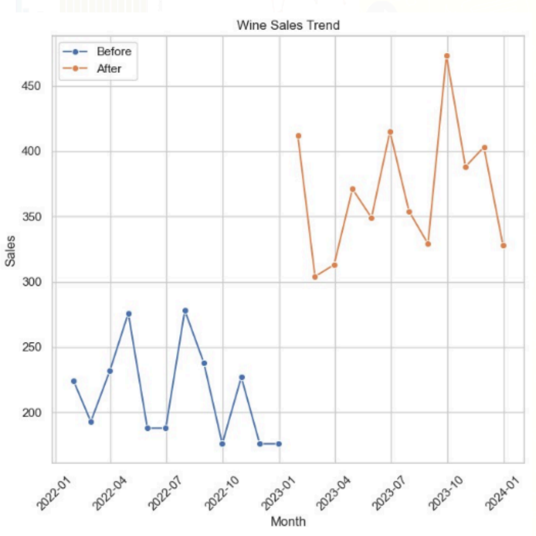
After inspecting the dataset, we decided to augment it with other datasets in order to increase the amount of data and improve the predictive model, data that were not in the original dataset, We also aimed to incorporate data not present in the original dataset. We analyzed variable correlations and their importance in training the machine learning model. The chosen algorithm was RandomForest, with a 70% training and 30% testing split, achieving an accuracy of 87%.



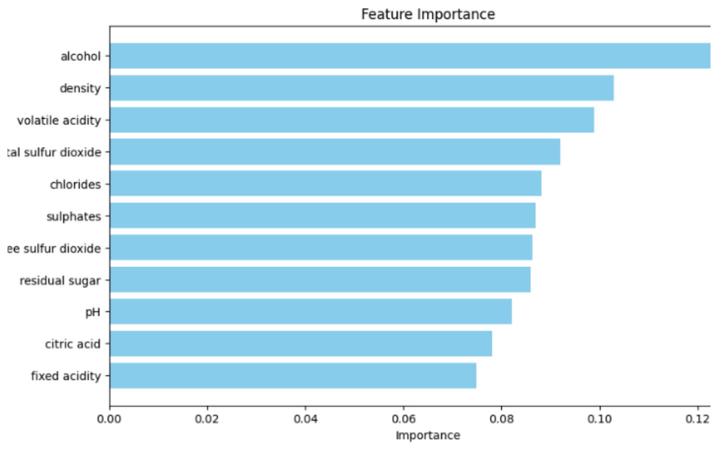


Methods

- Data Collection:
- Data Cleaning and Preprocessing
  - Identifying null values
  - Data transformation
- Exploratory Data Analysis
- Model Selection:
  - The RandomForest algorithm



Conclusion



The main challenge encountered was accessing data with specific chemical specifications of wines. Despite this, the project demonstrates significant opportunities to enhance productivity, quality, and production in the wine industry. The identified correlations provide a reliable approximation of wine quality, and the predictive model offers a precise margin for improving the production of high-quality wines. This, in turn, can enhance the industry's reputation, reviews, and sales.

References

[1] M. A. Cruz-de Aquino, R. A. Martínez-Peniche, A. BecerrilRoman, and M. S. Chávaro-Ortiz, ”Caracterización física y química de vinos tintos producidos en Queretaro,” Revista fitotecnia mexicana, vol. 35, no. spe5, pp. 61-67, 2012. [Online]. Available: http://www.scielo.org.mx/scielo.php?script=sci arttext&pid=S0187-73802012000500013&lng=es&tlng=es.

[2] M. Lopez Vanegas, ”Analisis de correlaciones entre variables,” Pragma, 2015. [Online]. Available: https://www.pragma.co/es/blog/analisis-de-correlaciones-entre-variables.

[3] C. Sierra, ”Notas de clase de Ciencia de Datos,” unpublished, Universidad Distrital Francisco José de Caldas, Bogotá, 2024.

[4] ”Playground Series S3E5,” Kaggle. [Online]. Available: https://www.kaggle.com/competitions/playground-series-s3e5/leaderboard.

[5] ”Wine Quality,” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/rajyellow46/wine-quality