Predicting Wine Quality Using Machine Learning

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

Abstract—**Problem: Wine is a beverage used for various purposes such as cooking, tasting, and general consumption. Its main component, grapes, undergoes several chemical processes during production that determine its quality. The goal of this project is to predict wine quality using a model based on chemical properties such as pH, citric acid, and alcohol content.**

Previous Approaches: This problem has been tackled by various researchers, each proposing unique models from different perspectives. Most commonly, the Kaggle dataset has been used as a reference for exploratory analysis to identify correlations and outliers, facilitating data preparation for AI model training. Approaches have varied, with some using Random Forest and others XGBoost, achieving high ranks in competitions.

Our Approach: After examining the dataset, we decided to augment it with additional datasets to increase data volume and improve the predictive model. We also aimed to incorporate data not present in the original dataset. Following the enhancement, we conducted an Exploratory Data Analysis (EDA) to understand data distribution, properties, and values. 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%.

I. Introduction

The world of wine is often perceived as straightforward, involving the fermentation of grapes and the addition of some ingredients before it is ready for sale. However, it involves a complex chemical process, where different factors and outcomes can define the wine's quality from a scientific and chemical standpoint. The main properties considered when evaluating a wine's chemical attributes are:

- **Fixed Acidity**: Various acids contribute to the fresh taste and longevity of wine, ideally present in a proportion of 4-10 g/L.
- **Volatile Acidity**: Acids that can evaporate, indicating the wine's state; a high value suggests fermentation issues, ideally 0.2-1.2 g/L.
- **Citric Acid**: A minor acid contributing to the wine's freshness and balance, with high values possibly indicating wine adulteration, ideally less than 1 g/L.
- **Residual Sugar**: The amount of sugar left after fermentation, varying by wine type, typically 0.1 to 150 g/L.
- **Chlorides**: Contribute to the wine's salty taste, with excess indicating contamination or production issues, ideally 0.01-0.2 g/L.
- **Free Sulfur Dioxide**: Acts as a preservative and antioxidant; excess affects the wine's taste, ideally 10-50 mg/L.
- **Density**: Relates to sugar and alcohol content, decreasing as fermentation progresses, ideally 0.990-1.005 g/cm³.
- **pH**: Reflects color, microbial stability, and taste, ideally 2.9-4.0.
- **Sulfates**: Contribute to stability and preservation, with high values making the wine bitter, ideally 0.4-2 g/L.
- **Alcohol**: Essential for the wine's harmony and quality, ideally 8-15%.

These values are present in the dataset, allowing for exploratory data analysis and correlation analysis to provide insights into potential relationships and patterns. The correlations guided the feature importance analysis for training the RandomForest model, achieving an 87% prediction accuracy.

II. Methods and Materials

A. Problem Definition

The objective is to utilize a dataset containing information on the chemical properties of various wines to develop an effective predictive model that can determine wine quality based on metrics such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, sulfates, pH, and alcohol content. The following questions guided our process: Is there a correlation between the data? What benefits can we derive from predictive analysis?

B. Data Collection

The data was sourced from [4], a competition designed for data science learning through challenges. The dataset includes a training set and a test set, each with 13 columns and 1371 records, which was insufficient for robust model training. Therefore, we integrated two additional datasets containing similar information, such as [5], which had 6497 records and the same columns. All datasets were stored in CSV format for convenience.

C. Data Cleaning and Preprocessing

Data cleaning involved:

- **Identifying null values**: Records with three or more empty columns were deleted, while those with one empty value were filled with the average of similar records.
- **Data transformation**: Addressing inconsistent values and formats between datasets using basic mathematical operations and data type transformations (e.g., from float to int or category) for easier analysis and visualization. Anomalous values were removed to reduce model imprecision.

D. Exploratory Data Analysis

Statistical analysis generated descriptive graphs using functions like mean, median, and standard deviation. Various types of graphs, such as boxplots and heatmaps, were used to identify data patterns. A heatmap revealed positive correlations between chemical components like alcohol and sulfates, while a high negative correlation between volatile acidity and wine quality was noted.

E. Model Selection and Preparation

The RandomForest algorithm was chosen due to its ensemble learning method, which combines multiple decision trees to create a robust predictive model by averaging their predictions to reduce overfitting. The data was split 70/30 for training and testing. Feature selection and engineering were employed to improve model performance, achieving an estimated accuracy of 87%.

III. 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. Becerril-Román, and M. S. Chávaro-Ortiz, ”Caracterización física y química de vinos tintos producidos en Querétaro,” 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, ”Análisis 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.