Predicting Wine Acidity for Enhanced Quality

Data Analysis and Machine Learning Approach

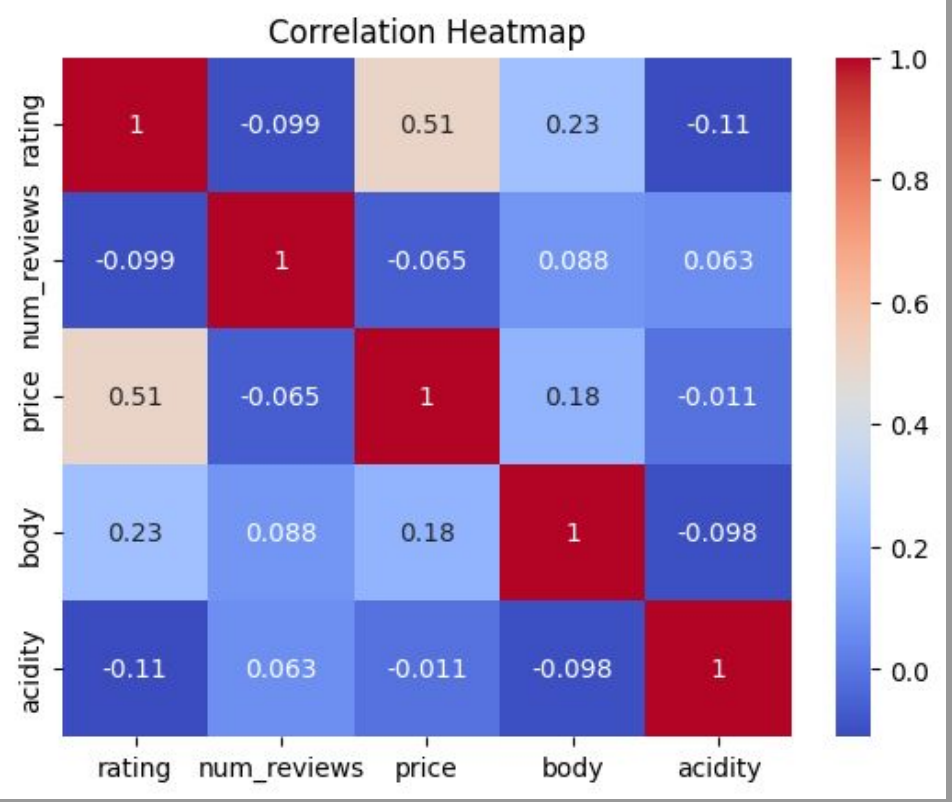
Welcome

Overview

- Our dataset consists of wine information, including variables like wine type, year, rating, number of reviews, country, region, price, body, and acidity.
- _____
- Our stakeholder is a renowned wine distributor in Spain.
- _____
- They are facing challenges in understanding the market demand and making informed decisions about the wines they should stock.

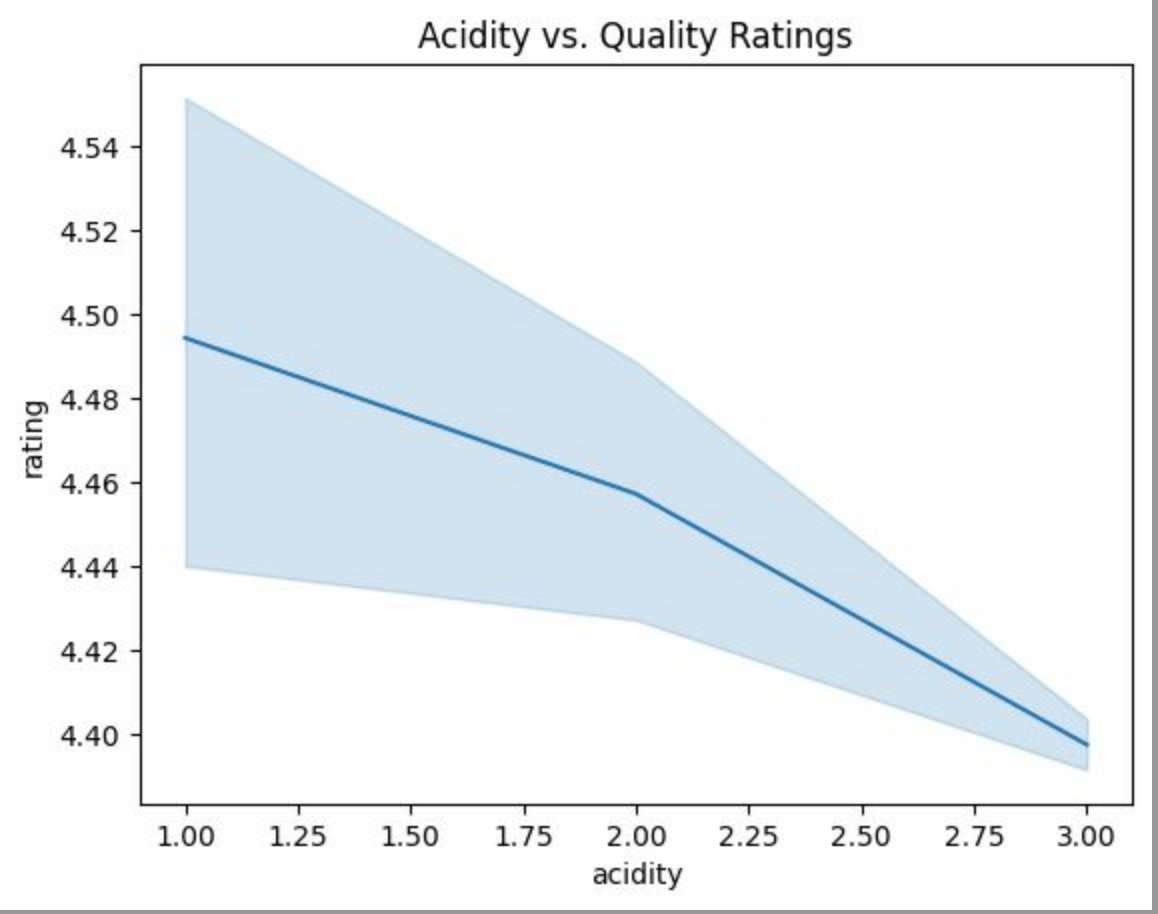
Exploratory Data Analysis

- To gain insights into the dataset, we performed exploratory data analysis.



Heat Map

- The correlation heatmap provides a visual representation of the relationships between different features in the Wine dataset. Darker colors indicate stronger correlations. This plot helps us identify any significant correlations that may exist between variables.



- **Line Plot**

  From the line plot, we can see that there is a moderate negative correlation between acidity and quality ratings. As the acidity level increases, the quality ratings tend to decrease

Let's evaluate the performance of our models and discuss their strengths and limitations.

We used various models to predict wine acidity: Linear Regression, Lasso, Ridge, Decision Tree, and Random Forest.

For the regression problem, the RMSE measures the average difference between predicted and actual acidity levels. The Random Forest model achieved the lowest RMSE of 0.2278, indicating its superior fit to the data. The R^2 score measures the proportion of variance explained by the models, and the Random Forest model achieved the highest score of 0.5140, capturing more variation in acidity levels.

In classification problems, false positive and false negative rates are important. These rates represent the misclassification of wines based on acidity levels. Balancing these rates is crucial to avoid misleading recommendations.

Conclusión

"Based on our analysis, we recommend using the Random Forest model for predicting wine acidity due to its superior performance in terms of RMSE and R^2 scores. This model shows the most accurate predictions and captures a significant amount of variation in acidity levels.