Jason Dean

DSC680-T301 (2241-1)

October 27, 2023

Project 3 – Milestone 1

Topic – Classifying Wine Quality: A Data-Driven Approach to Tasting

Business Problem – The wine industry thrives on quality assessments which often require expert tasters. However, can we predict the quality of wine based on its physicochemical properties? This project aims to classify wines into "low quality", "medium quality", and "high quality" using data-driven techniques, potentially reducing the need for extensive taste tests.

Datasets – The dataset will be sourced from the UCI Machine Learning Repository. It contains physicochemical properties of wines, such as acidity, sugar content, and alcohol percentage, along with a quality score given by experts.

Methods – The project will begin with exploratory data analysis to understand the distribution of wine qualities and their physicochemical properties. This will be followed by feature engineering and selection. For the classification task, algorithms such as Random Forests, Support Vector Machines, and Gradient Boosted Trees will be considered. Model performance will be evaluated using metrics like accuracy, precision, recall, and the F1-score.

Ethical Considerations – Over-reliance on models could potentially reduce the diversity of wines in the market if producers strictly follow the model's criteria for 'high quality'. Additionally, personal tastes vary, and a model based on expert opinions might not reflect the preferences of the general public. There is also a risk of overfitting the model to the dataset, which might not generalize well to wines outside the dataset.

Challenges/Issues

1.) The quality of wine can be subjective and varies from taster to taster.

2.) The dataset might not capture all factors influencing wine quality, such as storage conditions or age of the wine.

3.) Imbalance in data: Some quality categories might have fewer samples than others.

References:

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, 47(4), 547-553.

UCI Machine Learning Repository: Wine Quality Data Set.

https://archive.ics.uci.edu/dataset/186/wine+quality