AML Final Project Proposal - Team 33

Twisha Jain (tj2481), Rakshith Kamath (rk3165), Rahulraj Singh (rs4211), Yue Zhang (yz4155), Kechengjie Zhu (kz2407)

Background and context

According to Statista.com [1], revenue from the global wine market stood at 340.8 billion US dollars in 2020 which is roughly 25% of the entire global beverages market revenue. From all wine varieties, the most ordered and consumed ones across the globe are Red and White. So, what makes wine such a loved drink through geographies? In this project, we will aim to study wine sourced from grapes in the northwestern region of Vinho Verde in Portugal, which is one of the world's largest wine producers, amounting to 85 million liters of wine each year [2]. From the wine formulation data available to us, we would try to investigate what impacts the quality of wine, specifically Red and White wine? Is there any noticeable difference in quality between red wine and white wine? What should be a good threshold to identify good wine? Lastly, based on the results from the analysis of physicochemical features that make a high-quality wine, we will attempt to build a recommender system for indicators that point manufacturers to using the best composition for their wine.

Description of the Dataset

The dataset is obtained from Wine Quality hosted by Kaggle, and it consists of 6497 rows, 11 input features and one output feature. The dataset consists of red and white variants of "Vinho Verde", a Portuguese wine. The dataset has a class imbalance as there are few excellent or poor wines, while most of them are just normal wines. Therefore, we foresee the possibility of using outlier detection to detect the excellent or poor wines. Another known problem is that we don't know which of these features will be relevant for classification or regression. Thus feature selection methods will need to be used. Furthermore, we need to check the dataset features for collinearity and remove features with a high correlation. The dataset does not consist of any missing values, however we must look out for noisy data.

Proposed ML Techniques

Predicting the wine quality score can be viewed as a regression problem. We can start with linear models for regression, such as Linear regression, Ridge regression, Lasso regression and Elastic-net regression. Aside from using regression modeling, the problem can also be framed as a classification one, where we define wine with quality above a certain threshold to be good. In this case, Logistic regression, SVM as well as Random Forest models can be used. Since the data is imbalanced, to improve the performance, we can consider ensemble methods such as bagging and boosting. Boosting-based techniques for imbalanced data including Adaboost, XGBoot and LightGBM are good candidates.

Resources

1. https://www.statista.com/statistics/922403/global-wine-market-size/
2. https://en.wikipedia.org/wiki/Vinho_Verde