Wine Quality Analysis and Machine Learning

Machine learning utilizes statistics and data visualization to detect patterns in data much faster and more accurately than humans. This concept will be used to analyze datasets for bottles of red wine and white wine. The members of the group contributing to this analysis are Yousra Samman, Katie Stallings, Gladys Kuria, Elianny Herrera Montas, Thanh Tran, and Mariah Toy.

From the initial visual analysis performed by the group, an hypothesis was formed from the group’s knowledge of wine that alcohol content, volatile acidity, sulphates, and residual sugar can be used to determine whether a wine is of good or bad quality. Red and white wines have a quality range from three, the lowest quality, to eight, the highest quality. Wine quality is categorized as good being 7 to 8 and bad being 3 to 6. The machine learning models used are decision tree, random forest, and support vector classifiers. The decision tree model has a 0.92 precision for determining if a wine is bad quality, but only 0.58 precision for determining if a wine is good quality. The random forest model has a 0.90 precision for determining if a wine is bad and a 0.78 precision if a wine is good. The support vector has a 0.85 precision for bad wine and a 0.61 precision for good wine. Out of the three models used to determine quality, the random forest model provided the most accurate results. According to the random forest model, the most distinct features indicating a high-quality wine are alcohol content and density, with the next two features being volatile acidity and chlorides. A separate support vector model was tested to determine whether a wine is white or red. This test yielded a 0.98 precision for red wine and a 0.99 precision for white wine. Although determining if a wine is white or red is calculated with higher accuracy, more applications can be made with the models determining the quality of the wines.

The group dynamic supported asking questions, open communication, brainstorming, and sharing resources. Each members’ questions and opinions were considered seriously, and decisions were agreed upon prior to implementation. The group displayed excellent communication skills both inside and outside of class using the Slack group chat and Github to share any updates. The reading, cleaning, initial visual analysis, and statistics for the datasets were completed as a group. The different machine learning models were completed individually and discussed as a group. Mariah completed the decision tree model, Thanh completed the random forest, and Katie completed the support vectors for wine quality and for wine type. The information from the group was compiled and organized by Thanh.