STAT451 Project Proposal (Wine Quality Prediction)

Shihao Yang
syang536@wisc.edu

Yifan Zhang
yzhang2258@wisc.edu

Kexiao Zhu
kzhu52@wisc.edu

1. Introduction

Nowadays, people love to have a cup of wine when they are celebrating or simply resting. Wine is an alcoholic drink fermented from grape juice. It's welcome in many different areas around the world. As we can see from figure 1, the wine consumption continuously growing these years. And surprisingly, China is one of the fastest growing wine consumption countries.[3] Aside from it's good growth situation, Research suggests that it's good for your health if you drink a glass of wine occasionally. Because wine provides antioxidants which may promote longevity, and can help to protect you against heart disease.

Since we are investigating a specific wine type - Vinho Verde. We need to know something about its background. Vinho Verde is not a type of grape, it's a name of a region located in the north of Portugal. It's not expensive, and it pairs well with food. And it can be easily find in our local retail store.

Since there are so many good facts about drinking wine, one of the most important factor that affects people to buy wine is its taste. In a typical wine, there are so many physicochemical factors that can affect its taste.

In our project, we want to utilize those factors to predict the quality score of the wine which is given from 0(terrible) to 10(excellent). There were 11 factors measured: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates and alcohol.And in our dataset, it contains 6498 rows and 13 columns. We want to help the wine maker to find the most delicious formula to make Vinho Verde wine.

2. Resources

The Kaggle dataset obtained from the UCI machine learning repository(`https://archive.ics.uci.edu/ml/datasets/wine+quality`) consist of the following data fields: [2].

- **fixed acidity**: most acids involved with wine or fixed or nonvolatile(do not evaporate readily)
- **volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste)

Figure 1. the histgram of wine market growing.[3]

- **citric acid**: found in small quantities citric acid can add 'freshness' and flavor to wines)
- **residual suger**: the amount of suger remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter)
- **chlorides**: the amount of salt in the wine)
- **free sulfur dioxide**: the free sorm of SO2 exists in equiliburium between molecular SO2(as a dissolved gas) and bisulfite ion)
- **total sulfur dioxide**: amount of free and bound forms of SO2)
- **density**: the density of water is close to that if water depending on the percent alcohol and suger content)
- **pH**: describes how acidic or basic a wine is on a scale from 0(very acidic) to 14(very basic); most wines are between 3-4)
- **sulphates**: a wine additive which can contribute to sulfur dioxide gas(SO2)levels)

The computer hardware would be each group members' laptop(CPU) and the computational tools we expect to use are Python through jupyter notebook and we may use some packages like Scikit-Learn, XGBoost, and some decision tree model.

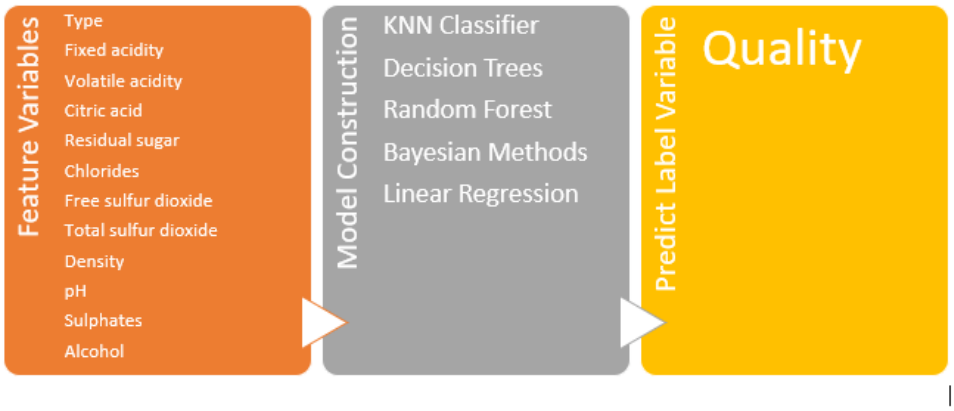

Figure 2. Example of our model construction and prediction process.

3. Motivation

'Vinho Verde' wine has been popular in the past decade due to its irresistible fruity flavor, festive fizziness, and freshness. 'Vinho Verde' take up about 16.8 percent of the wine made in the country of Portugal, and Germany and the United States imported over 53 percent of its wines production[1].

One important characteristic of 'Vinho Verde' wine which makes it attractive to customers is it goes well with most types of food. Pairing the wine with light dishes like salads and chicken, it can also be accompanied by more filling meals like potato and pork dishes. In addition, the white 'Vinho Verde' also goes great with seafood dishes, which are very popular in Portuguese cuisine like cod and monkfish.

'Vinho Verde' wine is always been known as 'cheep and cheerful', which retail price can lower to 4 dollar. But various 'Vinho Verde' wine do exist great differences in their qualities. So we hope our model can accurately present the quality of 'Vinho Verde' wine, using their physicochemical tests data. Having the ability to predict the wine quality based on physicochemical tests will not cause a qualitative change to the wine, but it will allow the manufacturers to clearly positioned 'Vinho Verde' with different brands or genres, set most suitable retail prices, and aim corresponding customers. This will even enable wine merchants to advertise accurately depending on the customers' favorites on wine quality. In the terminology of Economics, it may increase the total surplus of 'Vinho Verde' wine market.

4. Evaluation

Basically, we will consider our project as successful if our constructed model based on the training data set can label the wine's quality of our test data set accurately. At the same time, we intend to figure out which features are more significant and more decisive in our constructed model.

Our data set includes 12 feature variables, which are mostly quantitative, and more than 6000 examples. We plan to approximately split this whole data set into 80% training data set and 20% test data set, which are with the same distribution. Temporarily, we plan to construct multiple models with our training data set using different methods first, including linear regression, nearest neighbor methods, decision trees, random forest, and Bayesian methods. We may evaluate our model on the test data set with different test statistics, depending on the method we use, such as, KNN score (the mean accuracy) for nearest neighbor methods or Gini impurity and misclassification error for tree-based method. During this process, we must consider the problem of overfitting. We will respectively use our training data set and test data set to check the accuracy of the constructed model. If this problem of overfitting exists, we will use model selection methods to decide one model with a better fit for our test set. Finally, we expect to use some feature selection and feature extraction strategies, such as dimensionality reduction, to further optimize our model. Figure 2 above briefly demonstrates our model construction and prediction process.

Overall, we hope our model can have an accuracy around 90% of predicting the test data set. Ideally, we want to have a conclusion about which chemical components highly affect the wine quality, which may help the wine manufacturer to improve their product quality.

5. Contributions

Each member of our group will participate in the whole process of this project. For the computational tasks, Yifan will first clean the data and split the data set. After this, we will respectively construct different models. Kexiao will

mainly use linear regression and random forest methods to construct the models, Shihao will mainly use nearest neighbor and tree-based methods to construct the models, and Yifan will mainly use Bayesian methods to construct the models. If any of us needs help while constructing model, we will work together to solve the problem. Then, we will evaluate these models together and choose a better fit. All members will be responsible for feature selection and model optimization.

For the final report, each group member will write the part related to their used methods of constructing models. Besides, Kexiao will be responsible for results, discussions, and conclusions. Shihao will be responsible for abstract, introduction, and reference list. Yifan will be responsible for making some plots and figures to visualize our project. Before submitting, we will organize and review our report together and make some further improvements.

We have created a Github repository to share all materials related to this project, which ensures each of us clearly knows about the progress of our project. We will also meet each other usually either within the campus or online with Zoom to solve problems or develop our project.

References

[1] Vinho verde – grape, wine, wine style or wine region? 2019.

[2] S. Garg. Wine quality dataset. 2021.

[3] https://www.mordorintelligence.com/industry-reports/wine market. Wine market: 2021 - 26: Industry share, size, growth - mordor intelligence. *Website*, 2021.