**Week 12 Project Report**

**Predicting Red Wine Quality**

**Group 4**

**Group Members: Tyler S. Perry, Mehmet Ozmen, Fabio Nasseh**

**Submitted for LIS4761**

**On 11/14/2021**

**1. What has been done in the past week?**

The group reconducted the classification experiments before and after removing three irrelevant attributes : residual sugar, free sulfur dioxide, and pH. Then, we began compiling our results into the final report. We finished a rough draft of the final report.

**2. Did you work individually or together?**

The group worked together to conduct the classification experiments, and to decide which irrelevant attributes to remove. Then, we worked individually to compile the final report.

**3. Do you have any intermediate results to show?**

We found that after removing irrelevant attributes, the Random Forest algorithm classified the instances into their appropriate class 82.49% of the time. By comparison, the Naive Bayes algorithm classified the instances correctly 82.61% of the time.

**4. What are the challenges you faced in this activity?**

It was a busy week due to our schedules and activities around the campus. We were challenged by the attributes in the data and business of our schedules. We were able to talk and meet privately using zoom and group chat software.

**5. What do you expect to do next week?**

We are looking to meet together as a team and work on our final report and create an outline if needed to spread the work in between the group members to satisfy performance and deadlines.

**Introduction**

We will attempt to predict the quality of different Vinho Verde red wine varieties based on some or all of the variables provided to us in the dataset **[1]** (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol).

**Dataset**

The dataset we used was *Red Wine Quality* by UCI Machine Learning available for download from Kaggle. This dataset contains physicochemical and sensory data on 1,599 different red wine variants of the Portuguese “Vinho Verde” wine. We downloaded the data from Kaggle as a CSV file.

**Tools Used**

The CSV file was imported into Google Sheets for basic data cleaning. Then, we used the “CSV import” function in Orange 3 to create data visualizations to better understand the data. We used Weka to preprocess the data and apply classification experiments.

**Data Acquisition**

The dataset is an Open Database, available for free on Kaggle. We downloaded the dataset as a CSV file and opened it in Google Sheets.

**Data Preprocessing**

The dataset was already cleaned, contained no missing data, and no outliers. There was no need to apply additional data cleaning. Only discretization was performed.

FIrst, we created a new categorical attribute, Quality Category, using the Wine Quality numeric inputs. I used an IF statement in Google Sheets to split the numerical quality values into three categories: Poor, Good, and Excellent. This process is shown below in Figure 1. Then, the “quality” attribute was removed.

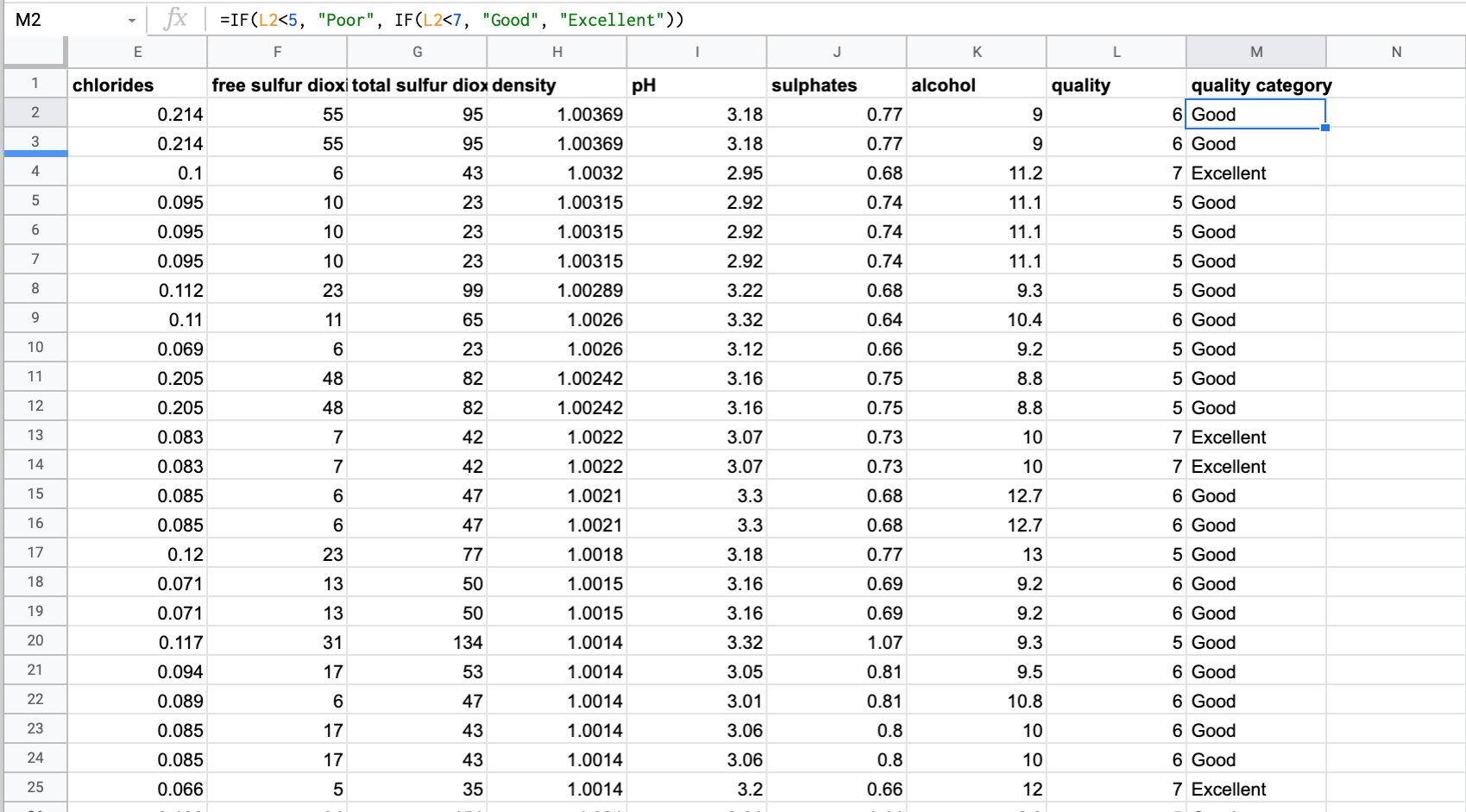


Figure 1. *We used an IF statement in google sheets to separate each red wine instance into a quality category based on its quality score. Wines scored at less than 5 were classified as “poor”, wines scored between 4 and 6 were classified as “good”, and wines scored higher than 6 were classified as “excellent.”*

Then, the dataset was automatically discretized from a range of numeric attributes into nominal by using the first-last method with a precision of 10 in Weka.

**Data Analysis and Results**

Our goal is to identify which attributes are responsible for producing excellent quality wine. We used Orange to calculate the r value of each attribute on wine quality. The correlation of each attribute on quality is shown below in Table 1.

Two classification algorithms, Random Forest Decision Tree and Naïve Bayes were then applied on the dataset. We applied the algorithms first with all attributes, and then after removing irrelevant attributes. In Weka, the datasets are separated into training and testing sets by using 10-fold cross validation. We applied the algorithms first with all attributes, and then after removing **residual sugar, free sulfur dioxide, and pH,** and compared the results. The results are shown below in Table 2.

The classification experiment is measured by the percent of correctly identified instances. We found that after removing irrelevant attributes, the Random Forest algorithm classified the instances into their appropriate class 82.49% of the time. By comparison, the Naive Bayes algorithm classified the instances correctly 82.61% of the time.

| **X** | **Y** | **Correlation** | **correlation coefficient (r)** |
| --- | --- | --- | --- |
| quality | volatile acidity | negative correlation | -0.390 |
| quality | total sulfur dioxide | negative correlation | -0.185 |
| quality | density | negative correlation | -0.175 |
| quality | chlorides | negative correlation | -0.129 |
| quality | pH | negative correlation | -0.058 |
| quality | free sulfur dioxide | negative correlation | -0.051 |
| quality | residual sugar | positive correlation | 0.014 |
| quality | fixed acidity | positive correlation | 0.120 |
| quality | citric acid | positive correlation | 0.230 |
| quality | sulphates | positive correlation | 0.251 |
| quality | alcohol | positive correlation | 0.476 |

Table 1. *The correlation coefficient for each attribute on wine quality was calculated in Orange 3. The attributes with the strongest correlation were alcohol, volatile acidity, sulphates, and citric acid.*

|  | Accuracy | Precision | Recall | F1 |
| --- | --- | --- | --- | --- |
| Random Forest, all attributes | 81.8637% |  |  |  |
| Naive Bayes, all attributes | 82.0513 |  |  |  |
| Random Forest, only relevant attributes | 82.4891% |  |  |  |
| Naive Bayes, only relevant attributes | 82.6141 |  |  |  |

Table 2. *We applied Random Forest and Naive Bayes classification algorithms on the data. First, we applied the algorithms using all attributes, and then we removed residual sugar, free sulfur dioxide, and pH, since those attributes were irrelevant.*

**Timeline for Completion**

**Week 1-3**

The first few weeks we spent familiarizing ourselves with the dataset. We read more in depth about the wine data set and the description of each property of the input variables. We created an account on Kaggle and also played around with Orange 3 and Weka. After that, we downloaded the dataset and began creating visualizations of the data.

**Week 4-6**

The next few weeks were spent on data preprocessing. The wine quality category attribute was created, and discretization was performed. The group met to discuss the correlations of each attribute on wine quality.

**Week 7-9**

After we got a better understanding of the data, we spent the next few weeks uncovering intermediate results and preparing for the midterm presentation. Many classification algorithms were applied, and we settled on Random Forest and Naive Bayes to include in the final report. The group met on Zoom to put together our midterm presentation.

**Week 10-12**

After receiving instructor’s feedback on the midterm presentation, the group met again to retest the dataset after removing irrelevant attributes. Our final classification results were collected, and we began working on the final report.

**Team Workload and Roles**

Our team utilized a text message group chat to communicate throughout the week. We also utilized Zoom to meet synchronously and break down the worldload for each week. All work is done through a shared google drive folder, so that we are all working on one version of each file.

Although we all contributed wherever is needed; Fabio does most of the report writing, revision, and formatting of each report, Tyler does most of the data preprocessing/data analysis, and Mehmet does most of the Weka modeling.

**References**

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