**Project Report Week 11**

**Project Report Draft**

1. **Project title**
   1. Predicting Red Wine Quality
2. **Problem statement**
   1. 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 (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol).
3. **Datasets used in the project**
   1. This dataset contains physicochemical and sensory data on red wine variants of the Portuguese “Vinho Verde” wine.
4. **Tools used in the project**
   1. We used Excel for basic data cleaning, Orange 3 to visualize the data, and Weka to apply discretization and classification.
5. **Data acquisition**
   1. We downloaded the data as a CSV file from Kaggle.
6. **Data preprocessing**
   1. The dataset has already been cleaned, so there was no need to apply additional data cleaning. Only discretization was performed. The dataset was automatically discretized from a range of numeric attributes into nominal by using the first-last method in Weka.
   2. Then, we created a new categorical attribute, Quality Category, using the Wine Quality numeric inputs. I used an IF statement in Excel to split the numerical quality values into three categories: Poor, Good, and Excellent.
7. **Data analysis and results**
   1. **correlations**

* 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 attribute with the strongest positive correlation towards quality was alcohol, with an r value of 0.476.
  1. **classification**
* Now that we have an understanding of which attributes are correlated with higher wine quality, the next step is to conduct classification experiments to get the bigger picture. We applied 4 different classification methods using Weka.
* In Weka, the datasets are separated into training and testing sets by using 10-fold cross validation. The training data is portioned into 10 sets of equal size and the algorithms executed 10 times.

1. **Your timeline for completion**
   1. **Week 12:** Continue testing the models. Remove less relevant attributes, and re-run the Random Forest algorithm to try and get a better result. Receive feedback from the Professor.
   2. **Week 13:** Begin preparing for the final presentation. Meet over zoom to discuss final results, Professor’s feedback, and plan data visualization.
   3. **Week 14:** Present final project.
2. **Team workload and roles**
   1. Our team uses a group chat and zoom meetings to break up the worldload for each week. When our schedule allows, we meet on Zoom and work together. Although we all help out wherever is needed; Fabio does most of the report writing, Tyler does most of the data preprocessing/data analysis, Mehmet does most of the Weka modeling.
3. **Citations**
   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

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

Our group presented on Thursday and received feedback from the Professor. Since Mehmet and Tyler were both battling the Flu this week, not a whole lot of progress got made unfortunately.

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

Our group focused on working individually on their assigned parts throughout the week. We communicated through our group text message and informed other members of the progress and also to comment on each other's work.

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

As of right now we have presented whatever we achieved during our mid term presentation and currently don't have much more to show. We are working on different aspects of the project and will include those results in later reports.

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

The only challenge as of right now is just to make sure every member is on top of their work and ensure everyone contributes to the project. This has been successful up until now and all members are contributing well to the project.

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

Next week we plan to analyze the data furthermore and write down our results in the weekly report.