Andrew Updegrove
661410127

1 Exploratory Data Analysis

Wine Data set

The first data set that I will be working with is the Wine data set from the UCI Machine Learning repository. To start, the white wine and red wine data sets were imported separately. In order to perform analysis, the data sets were combined. A new column called wine_type was created. The value "red" was assigned to every observation in the red wine data set and the value "white" was assigned to every observation in the white wine data set. The data sets were then vertically merged without issue because all columns names and types were identical.

In order to prepare for later analysis the data sets were split into training and testing data sets. An 80-20 heuristic split was used to separate the data. My ideas for analysis were generated before looking at the data to prevent data snooping.

The data was cleaned by first checking for missing values. No missing values were found so analysis could be carried on without worrying about insufficient data. Next I looked at histograms for each variable to see if there was anything unusual about the distributions that would alert me to issues with the data sets. No issues were found and two sample distributions have been provided.

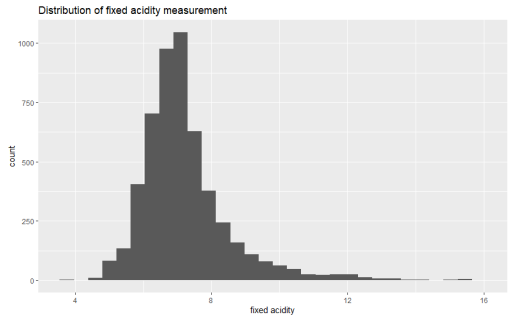

Figure 1: Fixed Acidity Distribution for UCI Wine Data set

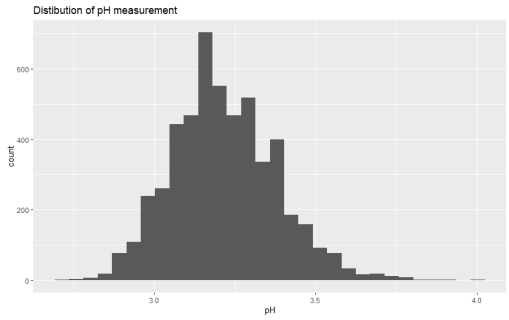

Figure 2: pH Distribution for UCI Wine Data set

Upon seeing the columns "fixed acidity", "volatile acidity", and "citric acid", I thought that maybe the dimensions of the problem could be reduced using Principal Component Analysis. Unfortunately upon inspection the variables were not correlated and therefore could not be reduced using this method. Two versions of the data set were prepared. Due to the discrete nature of the "quality" column, one version has the "quality" column as integers and the other has it as a factor.

Absenteeism at work

The other data set that I have chosen to work with is the Absenteeism at work data set from the UCI Machine Learning repository. The first thing I explored was the age distribution of the absence reports which can be seen in Figure 3. There were large spikes at 28 and 38.

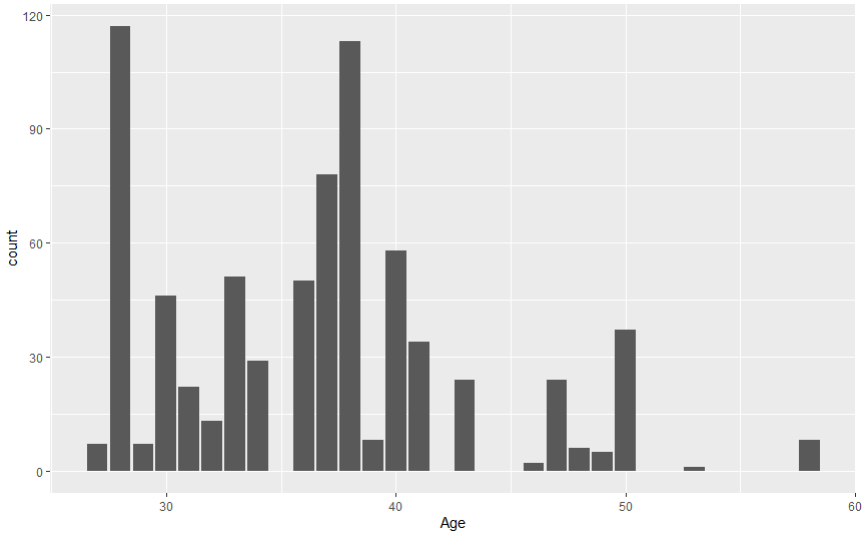

Figure 3: Absenteeism Age Distribution for all reports

I thought this was odd, and then I realized that each report was not an individual employees. Each observation in the data set was all the information about the incident and the employee. This meant that employee data was repeated for each observation. I then split the data set into two tables, one that had the unique employee ID with all the employees personal information and another that had all the individual incidents. The incidents table had an ID column that can be thought of as a foreign key to the unique employee ID in the other table. The data is therefore the absence reports for 37 different employees. I then repeated my plot of the age distribution of the 37 employees which can be seen in Figure 4, but this time including coloring for the education the employee had.

Figure 4: Age distribution for company employees

While doing all this, I also noticed when creating the employee information data set that there were two entirely different employees, one was 29 years old and the other was 41. Using the different ages I was able to change the ID of the 41 year old employee to 37. Also during my analysis I noticed that the employee with the ID of 3 had 113 absentee claims. Based on the reason for absence it appears as though this employee had some sort of large accident and thus was gone for a third of the year. For some of the analysis that was performed, this employee was left out due t the nature of the claims.

2 Model Development

Wine Data set

For this data set, there was two questions I wanted answers, could you use the data to classify between red and white wine, and is quality rating affected by the recorded data. To accomplish this I will be performing a classification using support vector machines to separate red and white wine. To predict quality I will be using random forest and the k-nearest neighbor algorithm. These three analyses are presented below.

Research was done on the chemical difference between red and white wine in order to pick what features to use to classify the data. An article at spoonuniversity.com attributes the difference between wines to the wine skins that are fermented with red wine. This leads to higher antioxidants in red wine and a higher acidity in white wine. Therefore I have chosen to use the pH and citric acid values to try and classify the data as red and white wine. The pH column is being used because the specific antioxidant has a higher pH than what is average

for wine. Simplifying the classification down to two variables was also for the purpose of being able to visualize at least one of the analyses.

Upon viewing the distributions I knew that the data would not be separable. Furthermore, I realized that I would not be able to come up with a good classifier based on how mixed the data was. Nevertheless I continued on with the analysis. A support vector classification was ran with a linear kernel and a cost value of 0.1. The cost was chosen to be so low because I saw before hand the data was heavily overlapped so there needed to be very little penalty for points penetrating the boundary. The results of the classification can be seen in Figure 5 and the points without the classification can be seen in Figure 6

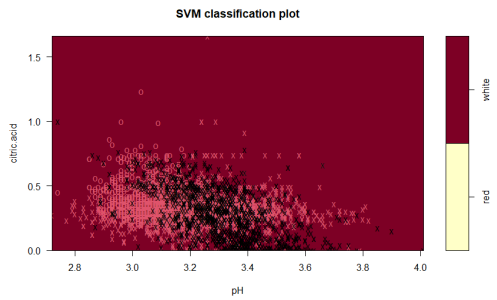

Figure 5: Support Vector Machine Classifier for pH and Citric Acid

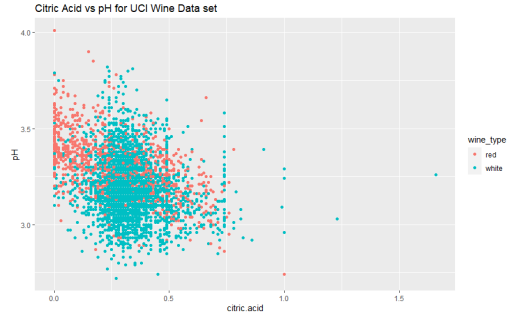

Figure 6: pH vs Citric Acid for UCI Wine Data set

The support vector machine uses 2537 support vectors to create its classification. As can be seen, the classification does not work at all. The classifier has chosen to label every point as white wine. This is simply because the data set is not separable enough and there are significantly more white wine data points than red wine data points. In Figure 6, you can see that the red wine data points are more to upper left than the white wine data points. In fact, this confirms the research I had done concerning acidity and the pH of wines. Some of the averages over the entire training set can be seen in the following table.

	Citric Acid Content	pH
Red Wine	0.2699	3.310
White Wine	0.3324	3.186

Table 1: Average Citric acid and pH values for red and white wine

As can be seen from the table of values, on average red wine has a higher pH and a lower citric acid content than white wine. This is what initial research showed, unfortunately the differences of those values is not enough to classify the data as red and white wine. Due to the fact that the classifier only predicts white wine, the error on the test data set is the number of red wine data points in the set, therefore $E_{test} = 25.6\%$.

Next, I attempted to predict the rating given by wine critics using the given data. I decided to do this using two different methods, k-nearest neighbor(k-nn) and random forests. From pre-existing notions, I chose not to include the alcohol column in consideration. First, we'll look at k-nearest neighbor.

Because k-nn uses distances to calculate the nearest neighbor, the data in each column was linearly normalized by the range of the minimum and maximum values so that certain values weren't biased based on scale. Although this form of normalization is not preferable because you are using the data to modify the data, it was the best option because the hypothetical range of the categories was not known.

The biggest part of k-nn is selecting the number of neighbors to use. In order to select the best number of neighbors to use, leave-one-out cross-validation was performed to select the best value. A range of odd numbers from 1 to the square root of the number of training samples were tested. The cross validation was done on the training data set in order to save the test sample for a final evaluation. The plot of the cross validation errors can be found in Figure 7.

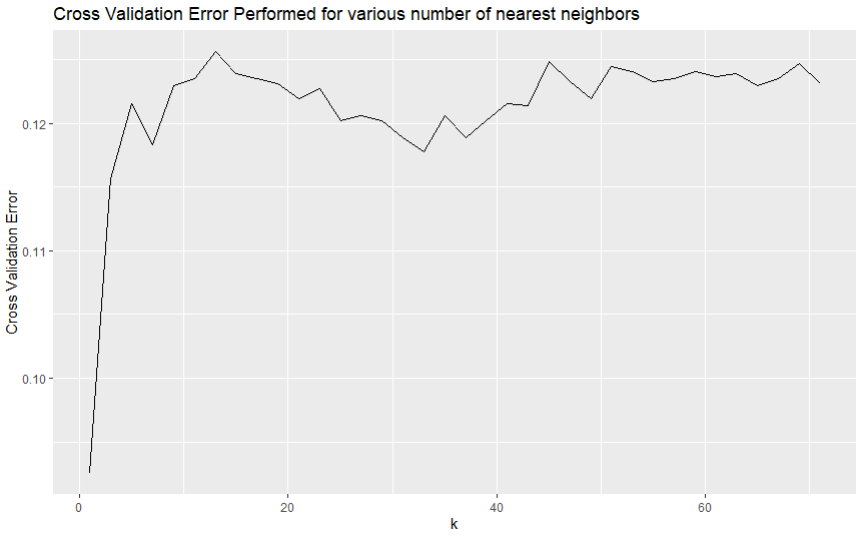

Figure 7: Cross Validation Error vs number of nearest neighbors

Although the lowest cross validation error was when $k = 1$, I chose to use $k = 4$ to avoid over-fitting the data. I then ran the k-nn algorithm using the whole training data set to predict the values for the test set. The test error was found to be 56.78%.

Now we can look at the random forest prediction. Due to the population voting methods of random forest, only one pass of random forest was performed. The random forest function in R was used with its default values. The value 'mtry' which determines the number of variables to randomly select for candidates at each level was chosen to be the square root of

the number of independent variables. 500 trees were generate for the forest. Once the forest was generated the forest was used to predict the test data set. The table of predictions can be found in Table 2.

		Predicted Values						
		3	4	5	6	7	8	9
Actual Values	3	0	1	3	2	0	0	0
	4	0	4	30	16	1	0	0
	5	0	2	287	132	6	0	0
	6	0	3	85	458	28	0	0
	7	0	0	5	87	116	0	0
	8	0	0	0	13	9	10	0
	9	0	0	0	0	1	0	0

Table 2: Wine Test Set Predictions by Random Forest

The forest had a prediction error of 32.6%. This is significantly better than the k-nearest neighbor predictions. This is because random forest take into account the importance of certain factors. The mean accuracy and gini ratings can be found in Figures 8 and 9.

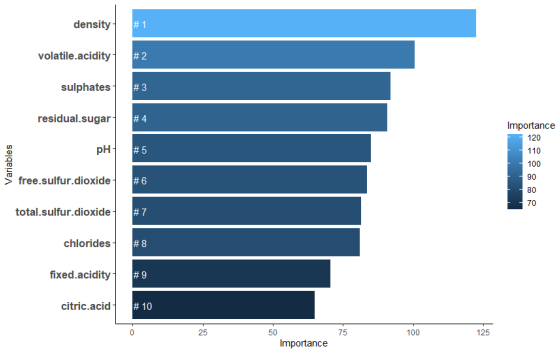

Figure 8: Wine Data Set Mean Decrease in Accuracy

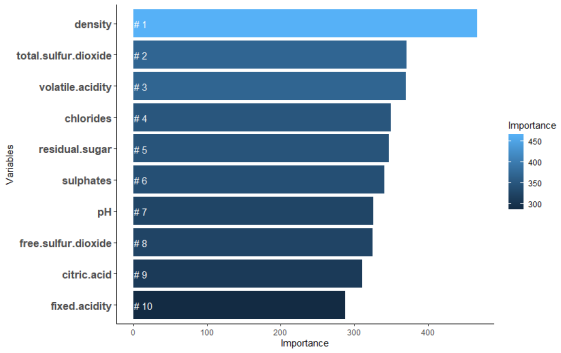

Figure 9: Wine Data Set Mean Decrease in Gini

The importance diagrams show that density was the most important variable when making building the trees and choosing splits.

Absenteeism at work

The first analysis I preformed was a simple linear regression to see if an employee's body mass index was correlated with the average time they were gone over all their absences. First, the

average number of hours and employee was gone was aggregated and added as a value to the employee's information table. The average time the employee was gone and the body mass index were then correlated with a simple linear model. The graphical representation can be seen in Figure 10.

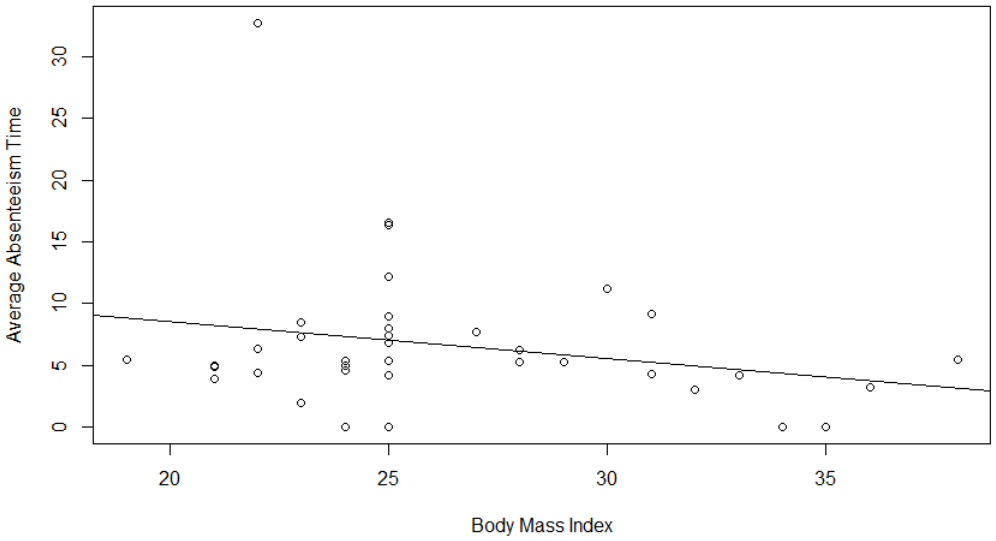

Figure 10: Linear Regression of Absenteeism Time vs Body Mass Index

As can bee seen in the graph, the slope of the line is nearly flat, meaning that the average time an employee is gone has no relation with their body mass index. Furthermore, the p-value test was .155 meaning that the data cannot even be considered correlated. The R-square value was .03 further emphasizing the fact that these variables have no statistically significant relationship.

The second analysis that was performed was to try and predict whether an absence would be a disciplinary infraction based on chosen factors. The chosen factors were ID, Reason for Absence, Month of absence, Day of the week, Age, Education, and Social drinker. These were chosen strictly on personal mental heuristics on what variables might influence someone's absence.

I split the data set into a training and testing set using an 80-20 split while excluding employee 3 due to the reasons I mentioned in the EDA. This resulted in a training set size of 502 and a test set size of 125. A radial kernel was used to in the support vector machine model. The support vector machine package in R has a built in tune function that allows for 10-fold cross-validation. The tune function was used to find the best cost value to use. The costs of 0.1, 0.5, 1, 2, 3, 4, 5, 8, 10 were considered. A plot of the cross-validation error versus the

cost value is shown in Figure 11

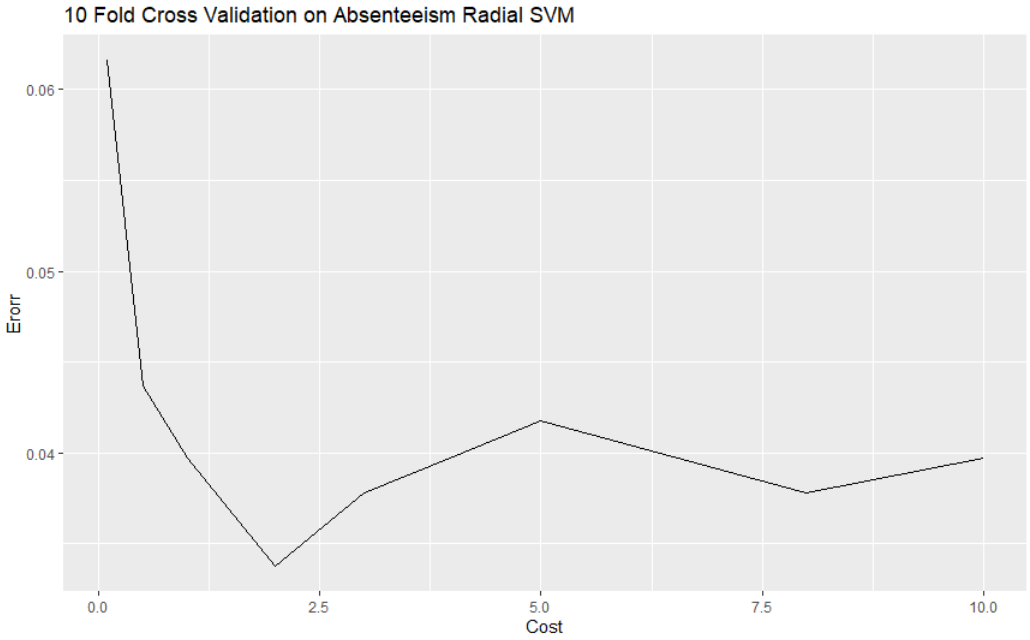

Figure 11: Cross-Validation Results to Select SVM Cost

The best model, the one with a cost of 2, was then used to predict the values of the test set. The matrix of the predicted values versus the actual values can be seen in Table **??**.The support vector machine has an accuracy of 97.6%.

		Actual	
		0	1
Predicted	0	116	1
	1	2	6

Table 3: Confusion Matrix for SVM Prediction

The final model was a random forest that was made to predict the reason for absence from given parameters. The parameters that were chosen for the analysis were Month of absence, Day of the week, Age, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, Body mass index, and Absenteeism time in hours. These were chosen personally just by conjecturing what variables might affect the reason for absence.

The data was once again split into a training set and test set; however, this time there was an issue. Some categories had so few data points that the training set ended up not having any

points for those factors. Therefore, I chose to use only to consider reasons for absence that had more than 6 data points associated with them. As before, I did not include employee 3's data due to the high frequency of their absences. Due to these circumstances, the training set had 429 observations and the test set had 119 observations.

The random forest was run with the default parameters and a model was obtained. This model was used to predict the values of the test set and then compared to the actual reason for absence. The confusion matrix has not been reproduced here due to the large number of factors. The model had an accuracy of 40.3%. The importance graphs have no provided due to how poor the model performed.

3 Decisions

Wine Data set

Overall, none of the models performed what could be considered well enough for practical use. The support vector machine classifier had an error rate of 25.6%. This is a very high error rate, and furthermore it isn't really making a decision. The model is classifying every value it gets as white wine which is in no way helpful. Although the model does not help with predicting wine types, it is okay that it failed. Failure has shown that you cannot use just citric acid content and pH to figure out if a wine is white or red.

The models for predicting a wine's rating also did not perform too well. The k-nn prediction was terrible and got less than half of the wine ratings right for the test set. The random forest did much better by predicting about two-thirds of the ratings correctly for the test set. Although the models overall performed poorly it does not mean that there is a way to get better predictions. The input parameters for this data set are strictly objective measures and the output value is a subjective measure. The metadata for this data set said that the ratings were created by averaging the scores given by three wine "experts". I put experts in quotation because there has been a number of studies done that have proven that rating wine and determining quality are arbitrary. Wine quality is based off of someone's preferences so trying to predict a subjective quality measure based on objective measurements might not be possible.

Absenteeism at work

The model's performed well over all. The worst model was certainly the random forest. I strongly believe that predicting the absence is incredibly difficult due to the very limited data points for some of the categories. It could also be very difficult due to the randomness of the reason of why someone would miss work.

The linear model was much more successful and showing discovered results. The linear regression model showed that the there is no relation between the time gone from work and

a person's body mass index.

The support vector machine classifier worked quite well getting a very impressive accuracy of 97.6%. However, there was only 40 disciplinary failures out of the 740 observations. When splitting up the data set, this left 8 disciplinary failures in the test set. So the model only got 75% of the disciplinary failures correct. The model had an almost equal likelihood of predicting one when the actual value is a zero and a zero when the actual value is a one. It would be preferable if the model would be more likely to give false negatives that way someone who should not have a disciplinary failure would not accidentally receive one. Overall, this data set is pretty poorly put together with lots of errors and anomalies.