Cleyde Murillo, Jr.
CIS 635 - 01
Final Project Report

Introduction

Our data is supplied by a client within the medical industry on their employee base. Attributes such as ID, health metrics (temperature, blood pressure, head/body ache, etc.), and a Risk classification were taken on each of their employees. The client is requesting a pipeline that conducts a periodic analysis on two data sources and outputs predictions to counteract the effects of employees that can potentially test positive for the virus. This process will begin with data cleansing and merging, which is then preceded by exploratory data analysis, and furthermore testing classifiers to ultimately deduce which model will work best for this case.

Data Cleansing & Merging

We performed a quick summary of the merged training sets which hold a combined 12 unique attributes with a mix of numerical and categorical data types. This told us which attributes had missing or outlier type data, which was then isolated using the ranges of acceptable values provided. We found 8 observations with NA values and 11 with outliers outside acceptable ranges. Given we have a large sample size of 6700, we removed the observations missing data. For the outliers, instead of removal we substituted the anomaly values with the median value of their appropriate attribute.

```
       id              temp            bpSys             vo2            throat          atRisk.x          headA            bodyA            cough       
 Min.   :   0   Min.   : 15.00   Min.   : 20.0   Min.   : 10.00   Min.   : 81   Min.   :0.0000   Min.   :  0.000   Min.   :1.000   Min.   :0.0000  
 1st Qu.:1673   1st Qu.: 97.79   1st Qu.:119.0   1st Qu.: 34.00   1st Qu.: 97   1st Qu.:0.0000   1st Qu.:  3.000   1st Qu.:4.000   1st Qu.:0.0000  
 Median :3352   Median : 98.19   Median :124.0   Median : 39.00   Median :100   Median :0.0000   Median :  3.000   Median :4.000   Median :0.0000  
 Mean   :3376   Mean   : 98.47   Mean   :124.6   Mean   : 37.76   Mean   :100   Mean   :0.4652   Mean   :  3.461   Mean   :4.016   Mean   :0.3418  
 3rd Qu.:5084   3rd Qu.: 98.93   3rd Qu.:130.0   3rd Qu.: 42.00   3rd Qu.:103   3rd Qu.:1.0000   3rd Qu.:  4.000   3rd Qu.:4.000   3rd Qu.:1.0000  
 Max.   :6780   Max.   :198.83   Max.   :501.0   Max.   :150.00   Max.   :122   Max.   :1.0000   Max.   :100.000   Max.   :7.000   Max.   :1.0000  
                NA's   :1        NA's   :1       NA's   :2        NA's   :1                      NA's   :1                                                
     runny           nausea          diarrhea      
 Min.   :0.0000   Min.   :0.0000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000  
 Median :0.0000   Median :0.0000   Median :0.000  
 Mean   :0.1986   Mean   :0.2367   Mean   :0.102  
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000  
 Max.   :1.0000   Max.   :5.0000   Max.   :1.000  
 NA's   :1                         NA's   :1
```

Figure 1: Data Summary

Exploratory Data Analysis

After cleansing we wanted to get an idea of what each variable looked like. To visualize this, we created a combined density table for each attribute. From this visual we see that the numeric variables typically follow a normal distribution with exception of temperature. For categorical variables we also notice a fairly even distribution of 'yes' and 'no' answers (shown as 0 or 1). This tells us that the data source is reliable for further analysis.

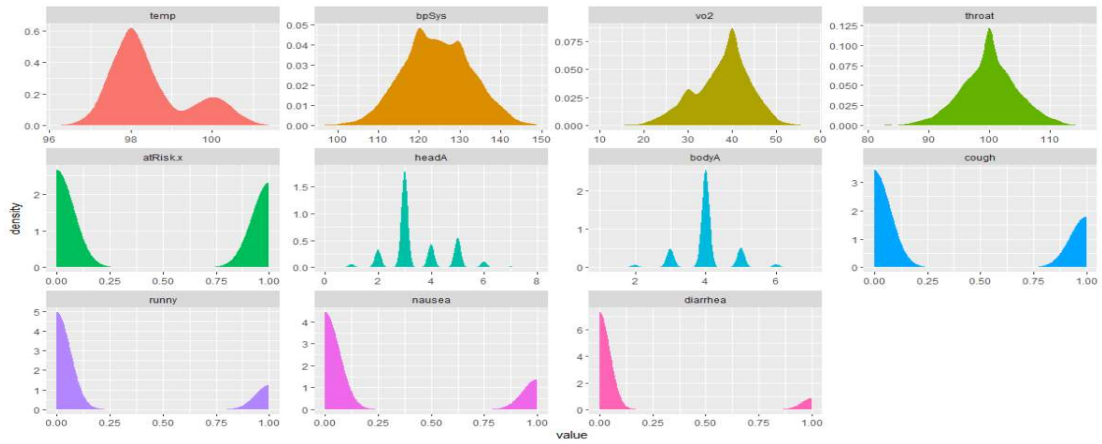


Figure 2: Density tables each attribute

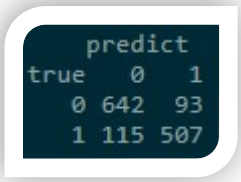
After the distributions we wanted to determine any underlying relationships between our dependent and independent attributes. This meant determining correlations between Risk classification and the other 10 variables (not including ID). We calculated and visualized Pearson's correlation shown on Figure 3. We notice distinct positive correlations between Risk classification, temperature, headache & nausea. So, moving forward those are variables we will be paying attention to since they typically correlate to a positive Risk classification.

Figure 3: Correlations between attributes

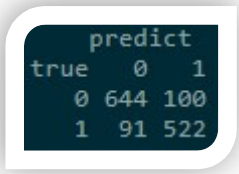
Model Testing

Since we are classifying data, we decided to leverage testing on 3 algorithms: Naïve Bayes, Decision Trees, and Support Vector Machines (Linear/Polynomial). To fully evaluate each model we extracted the Accuracy, Recall, and Precision from each model tested.

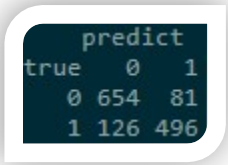
Naïve Bayes



Decision Trees



SVM (Polynomial)



Accuracy	0.8467	Accuracy	0.8592	Accuracy	0.8467
Recall	**0.8481**	**Recall**	**0.8762**	**Recall**	**0.8481**
Precision	0.8735	Precision	0.8656	Precision	0.8735

In this case we are assessing a medical result to people of potential risk to having a viral disease. This means that we want to absolutely mitigate the possibility of False Negatives. A false negative interprets to an employee that has been predicted to **not** have the virus but ends up having a **positive viral presence**. An outcome that is much more detrimental than a False Positive, where an employee is predicted to have it but ends up with a negative viral presence. Given this information we will take the 'recall' metric into heavy consideration when evaluating the models. In overview, they all deem similar results, with Support Vector Machines being the lowest performing. Naïve Bayes excels in precision (better with False Positives) to Decision Trees but as iterated we want to mitigate False Negatives as much as possible. Decision Trees has the highest recall potential; thus, we selected Decision Trees for production.

Weekly Pipeline

We developed and finalized a weekly script that requires 4 parameters; 2 training & 2 testing data sources (given that was the shape of how we received the initial data files). Within the script we aggregated multiple procedures from merging, cleaning, data conversion, then ultimately to training the decision tree. The script produces 3 complex plots and 1 text file containing the data and predictions. The figures below illustrate the plots produced for the user to receive overview of the procedures.

Figure 4: Visualized Confusion Matrix

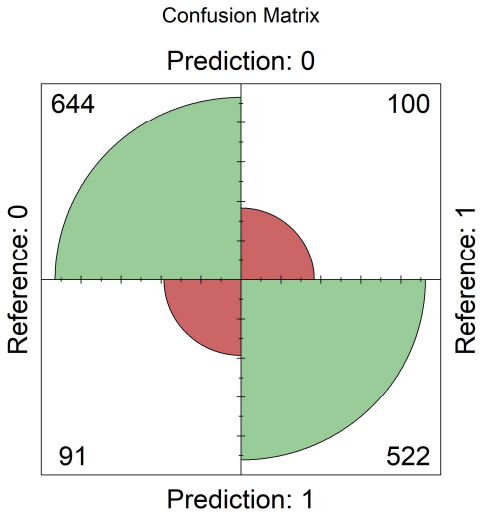
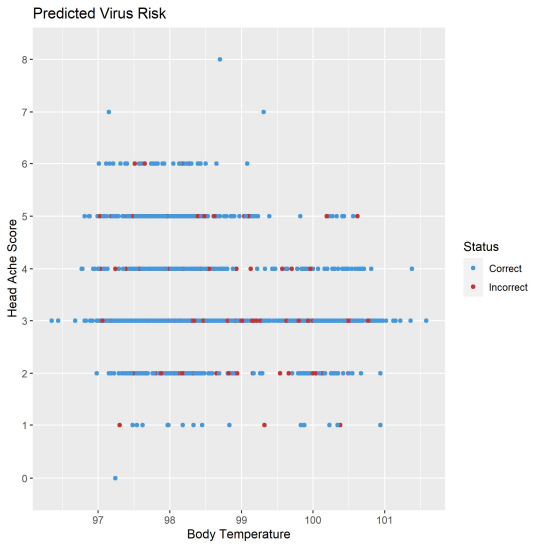


Figure 5: Predictions with Temperature vs. Headache



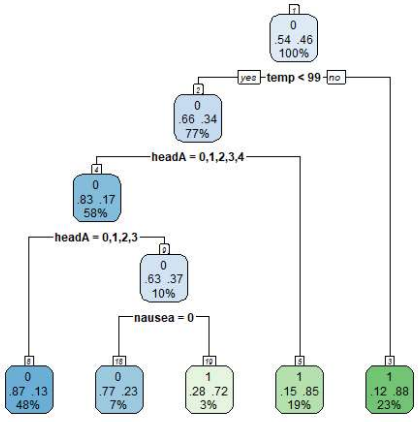


Figure 6: Final Decision Tree Plot

After Analysis & Concluding Thoughts

On the test set we were able to successfully predict 1166 employees correctly. Of the remaining 191 employees, 100 received False Positives where they ended up testing negative for viral presence, and 91 with False Negatives. While we tested for the best model, there can always continuously be improvements whether that be with the data or conducting more efficiency tests on another algorithm. Ultimately the results bear significant positive predictions in determine virus risk potential for the client's employees. With the resulting analysis and predictions finalized, we followed through with a cluster analysis to visualize specific groups with most relatable attributes. The following provides information on these 4 clusters.

Cluster 1: Size (1476) Lowest average Blood Pressure (113.5(

Cluster 2: Size (1558) Highest average Risk Potential (1.801)

Cluster 3: Size (2397) Lowest average Temperature (98.06)

Cluster 4: Size (1339) Highest average Headache score (5.232)

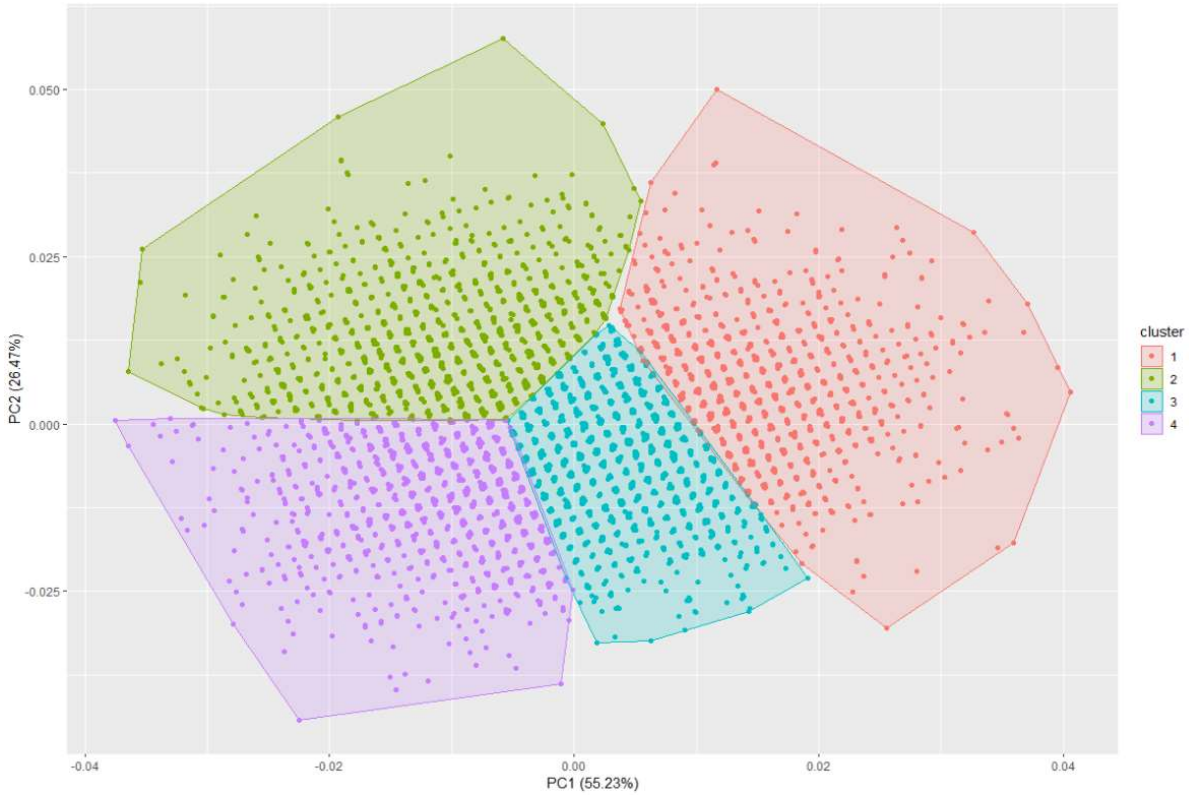


Figure 7: Clustering of Groups