Will the Weather Tomorrow be Warmer?

Name: Ryan Li Jian Tang

Table of Contents

Analysis of Data

```
> table(WAUS$WarmerTomorrow)

   0    1
 940 1042
```

In the data set of 2000 days that I have, there are 1042 days where it is warmer than the previous day and there are 940 days where it was colder than the previous days. In other words, this mean that 52.1% of the days were warmer than the previous and 47.9% where it was colder. However, the day does not sum up to 2000 days, meaning that there are some rows that do not have a record of the 'WarmerTomorrow'.

```
> summary(WAUS[1:23])
      Day             Month             Year          Location      MinTemp          MaxTemp         Rainfall        Evaporation
 Min.   : 1.0   Min.   : 1.00   Min.   :2008   Min.   : 5.0   Min.   :-7.5   Min.   :-3.2   Min.   :  0.0   Min.   : 0
 1st Qu.: 8.0   1st Qu.: 4.00   1st Qu.:2011   1st Qu.: 8.0   1st Qu.: 5.3   1st Qu.:16.6   1st Qu.:  0.0   1st Qu.: 3
 Median :15.0   Median : 7.00   Median :2014   Median :29.0   Median :10.3   Median :21.5   Median :  0.0   Median : 4
 Mean   :15.5   Mean   : 6.53   Mean   :2014   Mean   :25.6   Mean   :10.2   Mean   :21.5   Mean   :  1.8   Mean   : 5
 3rd Qu.:23.0   3rd Qu.: 9.00   3rd Qu.:2017   3rd Qu.:36.0   3rd Qu.:15.4   3rd Qu.:26.3   3rd Qu.:  0.6   3rd Qu.: 7
 Max.   :31.0   Max.   :12.00   Max.   :2019   Max.   :43.0   Max.   :27.4   Max.   :42.9   Max.   :130.4   Max.   :21
 NA's   :20     NA's   :13      NA's   :19                    NA's   :35     NA's   :16     NA's   :43      NA's   :946
    Sunshine      WindGustDir  WindGustSpeed     WindDir9am     WindDir3pm    WindSpeed9am   WindSpeed3pm   Humidity9am
 Min.   : 0     W      :208   Min.   :  9.0   E      : 166   W      : 226   Min.   : 0.0   Min.   : 0.0   Min.   :  8.0
 1st Qu.: 5     E      :207   1st Qu.: 30.0   SW     : 164   WSW    : 206   1st Qu.: 7.0   1st Qu.:11.0   1st Qu.: 59.0
 Median : 9     ENE    :173   Median : 37.0   ENE    : 163   WNW    : 144   Median :13.0   Median :17.0   Median : 71.0
 Mean   : 8     WSW    :169   Mean   : 39.3   WSW    : 152   E      : 142   Mean   :13.6   Mean   :17.5   Mean   : 70.9
 3rd Qu.:11     WNW    :164   3rd Qu.: 48.0   W      : 137   SW     : 125   3rd Qu.:19.0   3rd Qu.:23.0   3rd Qu.: 86.0
 Max.   :14     (Other):983   Max.   :100.0   (Other):1051   (Other):1077   Max.   :48.0   Max.   :54.0   Max.   :100.0
 NA's   :1025   NA's   : 96   NA's   :101     NA's   : 167   NA's   :  80   NA's   :74     NA's   :77     NA's   :56
  Humidity3pm     Pressure9am     Pressure3pm      Cloud9am        Cloud3pm        Temp9am         Temp3pm
 Min.   :  7.0   Min.   : 993   Min.   : 993   Min.   :0      Min.   :0      Min.   :-5.2   Min.   :-3.6
 1st Qu.: 38.0   1st Qu.:1014   1st Qu.:1012   1st Qu.:2      1st Qu.:2      1st Qu.:10.4   1st Qu.:15.2
 Median : 51.0   Median :1018   Median :1016   Median :5      Median :5      Median :15.7   Median :20.0
 Mean   : 52.3   Mean   :1018   Mean   :1016   Mean   :5      Mean   :4      Mean   :15.2   Mean   :20.0
 3rd Qu.: 65.0   3rd Qu.:1023   3rd Qu.:1020   3rd Qu.:7      3rd Qu.:7      3rd Qu.:20.1   3rd Qu.:24.6
 Max.   :100.0   Max.   :1038   Max.   :1036   Max.   :8      Max.   :8      Max.   :36.0   Max.   :42.0
 NA's   :40      NA's   :264    NA's   :275    NA's   :757    NA's   :837    NA's   :40     NA's   :43
```

Looking at the summary of the data, we can see that the results range from the year 2008-2019. There are a couple of variables where there is a large amount of NA's (reaching 1000), most notable in 'Evaporation', 'Sunshine', 'Cloud9am' and 'Cloud3pm'. This leads to a concern for the lack of data for these columns, showing that the data needs to be pre-processed. At this point, I cannot consider omitting any attributes as there is not enough information to judge which attributes are important in the modelling phase and at this point they all seem equally important.

Initial Pre-processing Done:

There were a lot of NAs in the initial data set, and the Boosting and Random Forest model does not take kindly to NAs in the data. Hence, the data needed to be pre-processed. At first I decided to remove all rows containing NAs, however, by doing this I reduced the data set by approximately 65% to 707 rows, which in my opinion was not enough data to create robust models in this assignment. Furthermore, removal of that much data might lead to loss of information which results in models not giving expected results when predicting. Thus, instead I opted to fill in each NAs with a suitable replacement. For numerical data, I replaced the NAs with their respective median value for the column. For categorical data, I have replaced the NAs with their respective mode for the column. By doing this, I have introduced variance and bias to my data set however in my opinion it is better than removing all rows with NAs, which might cause data loss. Furthermore, this method of imputation is more computationally efficient, and hassle-free compared to other methods like using machine learning algorithms to predict the values.

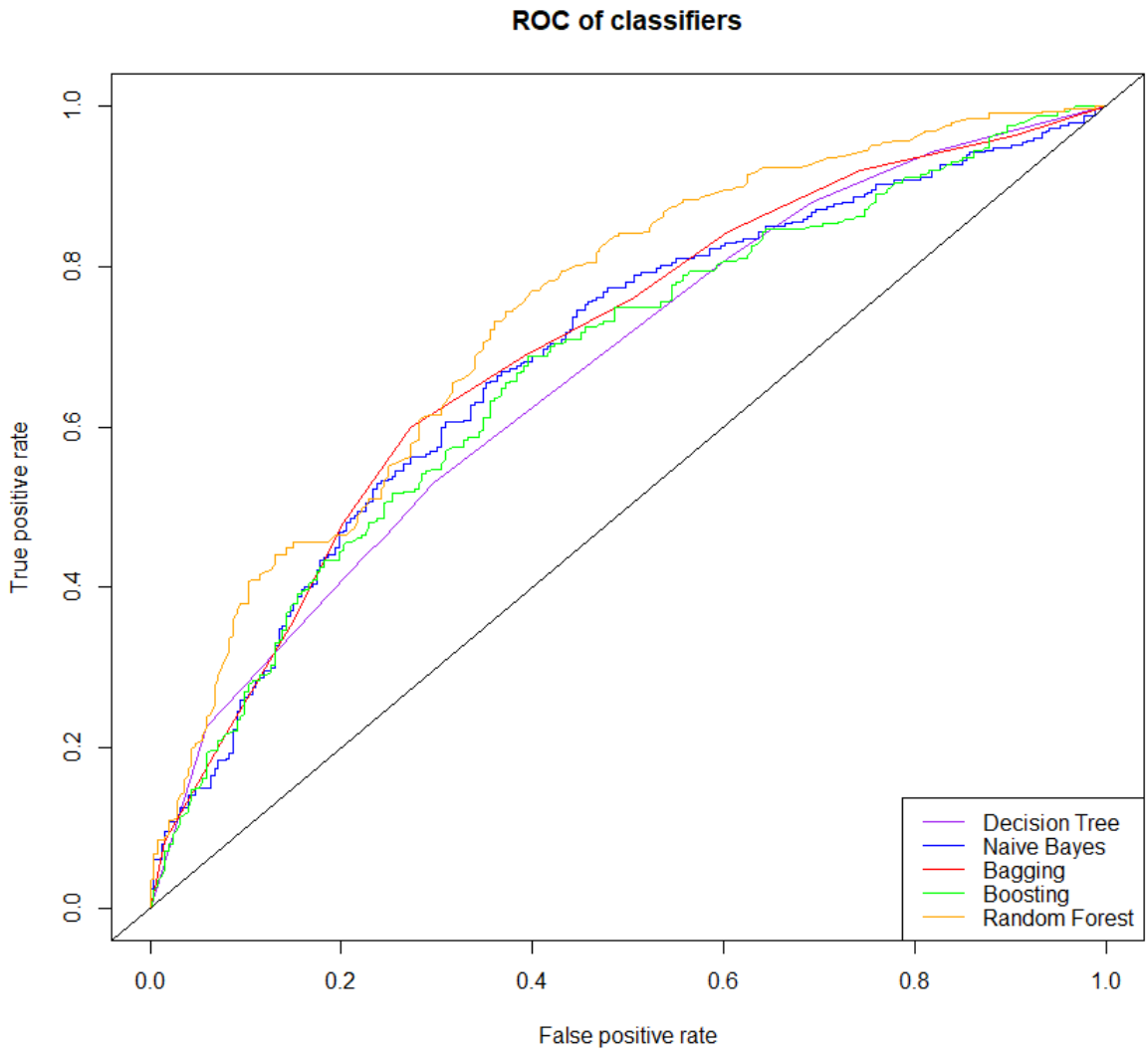
After the pre-processing, I have decided to remove the remaining few rows that contains NAs for the variable 'WarmerTomorrow'. Ultimately, there are no more NAs in the data set that is to be used.

Results of Initial Model

Classifier	TP	TN	FP	FN	Accuracy
Decision	173	178	75	154	0.605
Naïve Bayes	232	143	110	95	0.647
Bagging	208	162	91	119	0.638
Boosting	210	166	87	117	0.648
Random Forest	242	155	98	85	0.684

Between the models, it seems that the Random Forest model has the highest accuracy in predicting if the next day is warmer than the current day from given variables. The second highest accuracy model is then Boosting Model.

ROC of Models



Results of AUC can be found in the table under Question 7.

Comparison of Random Forest

Classifier	TP	TN	FP	FN	Accuracy	Precision	Recall	AUC
Decision	173	178	75	154	0.605	0.698	0.529	0.668
Naïve Bayes	232	143	110	95	0.647	0.678	0.709	0.686
Bagging	208	162	91	119	0.638	0.696	0.636	0.694
Boosting	210	166	87	117	0.648	0.707	0.642	0.689
Random Forest	242	155	98	85	0.684	0.712	0.740	0.737

Looking at the results above, the Random Forest Classifier performs the best and has the best results for all evaluation metrics used. Having the highest accuracy, precision, recall and AUC.

Analysis of Variables

```
> summary(WAUS.decision)

Classification tree:
tree(formula = WarmTmrw ~ . - WarmerTomorrow, data = WAUS.train)
Variables actually used in tree construction:
[1] "Humidity3pm" "WindDir9am"  "Pressure9am" "Temp3pm"     "Sunshine"    "MaxTemp"     "WindDir3pm"
Number of terminal nodes:  8
Residual mean deviance:  1.24 = 1660 / 1340
Misclassification error rate: 0.352 = 476 / 1351
```

```
> sort(WAUS.bagging$importance, decreasing=TRUE)
   WindDir3pm    WindDir9am   WindGustDir   Humidity3pm       MaxTemp   Pressure9am  WindSpeed3pm       MinTemp
       15.168        14.426        12.918        12.498         8.056         5.147         5.030         3.863
WindGustSpeed      Sunshine       Temp3pm   Pressure3pm       Temp9am      Cloud3pm   Humidity9am      Location
        3.479         3.185         2.918         2.734         2.177         2.152         1.933         1.002
 WindSpeed9am           Day         Month          Year   Evaporation      Cloud9am      Rainfall
        0.925         0.832         0.613         0.568         0.375         0.000         0.000
> sort(WAUS.boosting$importance, decreasing=TRUE)
   WindDir9am    WindDir3pm   WindGustDir       MaxTemp   Pressure9am   Humidity3pm       MinTemp      Sunshine
       15.067        14.648        13.264         8.760         6.070         5.358         4.459         4.363
      Temp9am       Temp3pm   Evaporation   Humidity9am      Location  WindSpeed3pm   Pressure3pm           Day
        2.949         2.926         2.806         2.345         2.329         2.265         2.132         2.089
        Month      Rainfall  WindSpeed9am      Cloud3pm          Year WindGustSpeed      Cloud9am
        2.059         1.683         1.662         1.490         0.870         0.406         0.000
> sort(WAUS.forest$importance[,1], decreasing = TRUE)
   WindDir9am    WindDir3pm   WindGustDir       MaxTemp   Humidity3pm       Temp3pm       MinTemp   Pressure9am
         65.9          63.5          62.2          41.6          40.8          36.2          32.0          30.0
      Temp9am   Pressure3pm   Humidity9am WindGustSpeed           Day  WindSpeed3pm      Sunshine   Evaporation
         29.0          27.0          25.0          23.2          23.0          22.2          21.8          19.5
 WindSpeed9am         Month          Year      Cloud9am      Location      Cloud3pm      Rainfall
         18.9          18.9          17.4          15.2          14.8          14.0          12.5
```

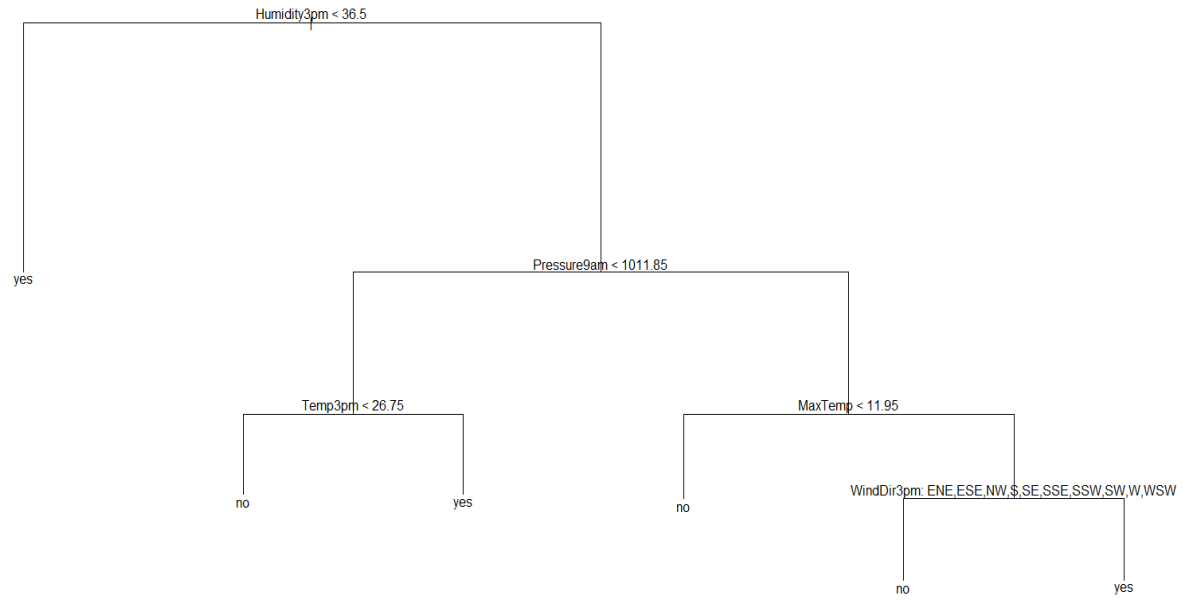
Important Variables

In the decision tree model, it has decided only 7 variables were needed amongst all the variables, where some of the variables include “WindDir3pm”, “WindDir9pm”. Since the decision tree has only chosen the top 7 variables, I have decided to pick the top 7 most important variables from the other models and place them as votes for the variables. Then from those votes, I proceeded to choose the top 7 most voted variables. Thus, from the models I have filtered out 8 variables where 2 variables had the same number of votes. These variables are ‘Humidity3pm’, ‘WindDir9am’, ‘Pressure9am’, ‘Temp3pm’, ‘MaxTemp’, ‘WindDir3pm’, ‘WindGustDir’, and ‘MinTemp’. These are the most important variables in my opinion.

Omittable Variables

On the other hand, variables like ‘Rainfall’, ‘Evaporation’, ‘Year’, ‘Month’, ‘Windspeed9am’, ‘Location’, ‘Cloud9am’, and ‘Cloud3pm’ can be removed from the data while having very little effect on the performance of the models. As these variables all generally have low importance in each model as shown in the above figure. There are some variables that are considered relatively important in one model and considered unimportant in other models. An example is ‘WindGustSpeed’, where it is relatively important in the bagging and random forest model but not in the boosting model.

Trimmed Decision Tree



Factors in Choosing

To create this simplified decision tree, I first proceeded to extract only the top 8 variables from the data set that was discovered in question 8 of the assignment. Therefore, reducing the amount of data that the model must go through and allowing the model to place more weight on the more important variables instead. Furthermore, it helps to reduce the overall complexity of the model and reduce the number of nodes. Allowing people to classify using the model by hand easily. This produced a tree with 7 leaf nodes. Afterwards, I proceeded to perform a cross validation on the completed model, to see if the decision tree could be pruned further. Looking at the results in the table below, I was able to reduce the tree to 6 leaf nodes while maintaining the same performance it had before.

Reading the Tree

If a day had a humidity reading below 36.5% at 3 pm, then the model predicts that the next day will be warmer. Else, the model will check if the pressure at 9 am is less than 1011.85 hpa. If it is then the model will then check if the temp is below 26.75 degrees at 3pm. If the temperature is below, then it will not be warmer the next day. Else if the temperature is above, then it will be warmer the next day. On the other hand, if the pressure at 9am is more than 1011.85 hpa, the model will move on to the next node. At the next node, if the max temperature of the day is below 11.95, then the next day will not be warmer. Otherwise, it moves on the final node and checks if the Wind Direction at 3pm is in one of ‘ENE’, ‘ESE’, ‘NW’, ‘S’, ‘SE’, ‘SSE’, ‘SSW’, ‘SW’, ‘W’, ‘WSW’. If it is, then the next day will not be warmer. Else the next day will be warmer.

Improved Results

Classifier	TP	TN	FP	FN	Accuracy	Precision	Recall	AUC
Decision	173	178	75	154	0.605	0.698	0.529	0.668
Naïve Bayes	232	143	110	95	0.647	0.678	0.709	0.686
Bagging	208	162	91	119	0.638	0.696	0.636	0.694
Boosting	210	166	87	117	0.648	0.707	0.642	0.689
Random Forest	242	155	98	85	0.684	0.712	0.740	0.737
Simplified Decision	176	166	87	151	0.590	0.669	0.538	0.643

Looking at the results of the simplified decision tree compared to the other models, it performs the worse with the lowest accuracy, precision, recall rate and followed by the lowest AUC. Therefore, showing that it is the worst at classifying the 'WarmerTomorrow' class.

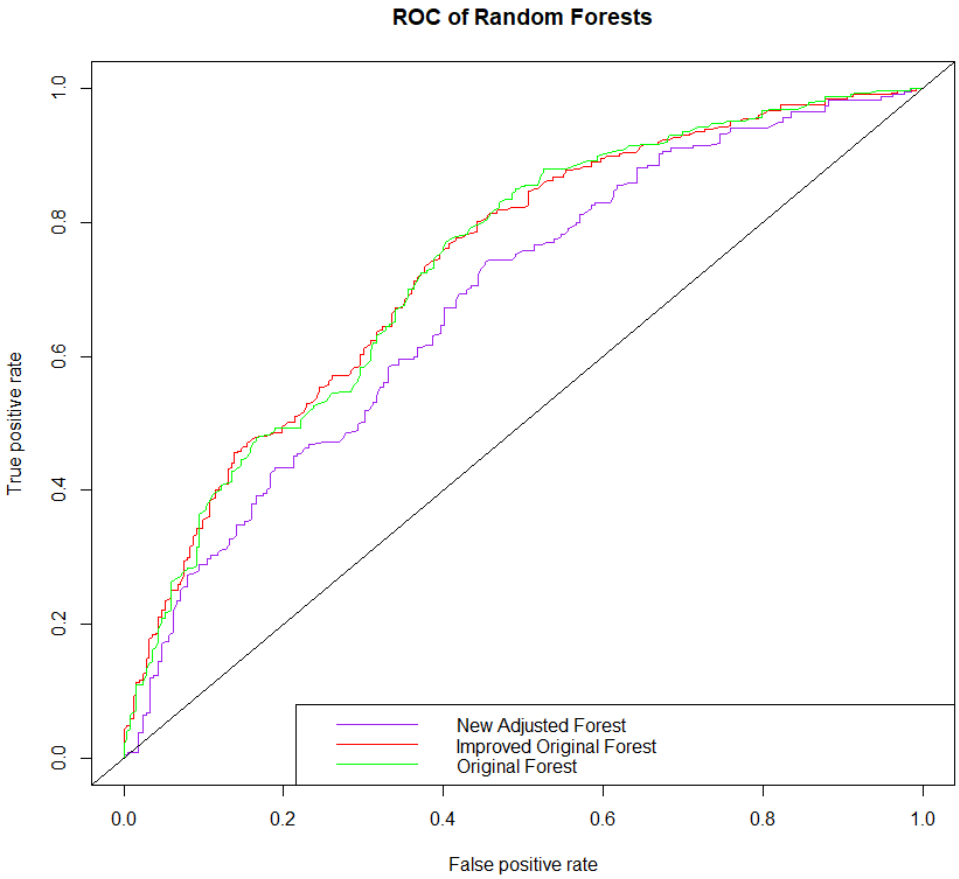
Improving Random Forest Model

I chose to improve the Random Forest model, as it was the best model in terms of all evaluation metrics (accuracy, precision, recall, and auc). The first step I took was to take only the 8 important variables, noted from question 8, from the original un-preprocessed data set. Then filtered out all rows with NAs, I was then left with 1488 rows from the original dataset which is good as I maintained most of the data. I did this to reduce the amount of variance and bias that was added from me previously imputing replacement values for NAs. Thus, there will only be minimal data loss as the data set still maintains approximately 75% of the rows.

To improve upon the random forest more, I tweaked the parameters in the training phase. I looked for the most optimum 'mtry' parameter. 'mtry' is the number of variables randomly sampled as candidates at each split and having the optimum reduces the out-of-bag (OOB) error rate of the tree, which in turn decreases the overall forest error rate. On the other hand, I have set the 'ntree' parameter to 1000 trees, to reduce the amount of overfitting while allowing it to stay computable by my local machine. Then I proceeded to build a Random Forest upon this. However, the results for the new Random Forest were worse than the original random forest model which is shown below.

Knowing this, I decided to try to improve on the original Random Forest from Part 4. Thus, proceeded to tweak the training parameters like 'mtry' and 'ntree' for the original random forest instead with the data set used with it before, with all attributes. It resulted in better performance as shown in the figures below.

Classifier	TP	TN	FP	FN	Accuracy	Precision	Recall	AUC
Random Forest	242	155	98	85	0.684	0.712	0.740	0.737
New Random	178	103	109	57	0.629	0.629	0.757	0.638
Improved Random	240	158	95	87	0.686	0.716	0.734	0.738



Looking at the table above, it shows that the new Random Forest that I initially created performed the worst out the 3 random forest models, having the lowest accuracy, precision, and AUC. Even though, it has the highest recall rate, it is still worse. On the other hand, by just tweaking the training parameters slightly, the extended Random Forest is slightly better than the original Random Forest in all accuracy, precision, recall and AUC. Thus, this new model has performed the best out of all the models so far created.

The main factor that I was considering, was the overall performance of the model. Thus, I chose Random Forest which had the best performance in the record. Furthermore, I have chosen to use all attributes in my final improved model instead of using a certain number of variables only. As the results gathered was worse when using less results, shown by the New Adjusted Forest.

Comparison of ANN Network

Classifier	TP	TN	FP	FN	Accuracy	Precision	Recall	AUC
Decision	173	178	75	154	0.605	0.698	0.529	0.668
Naïve Bayes	232	143	110	95	0.647	0.678	0.709	0.686
Bagging	208	162	91	119	0.638	0.696	0.636	0.694
Boosting	210	166	87	117	0.648	0.707	0.642	0.689
Random Forest	242	155	98	85	0.684	0.712	0.740	0.737
ANN	208	184	69	119	0.676	0.751	0.636	0.733

For the pre-processing phase, I have recoded all categorical data to have indicator columns. Then proceeded to bind the new indicator columns with the numerical dataset. For the numerical data, I proceeded to perform linear scaling to ensure that the values they hold are in a range between 0-1, allowing them to be comparable between each other. This helps to improve the processing time of the model and reduces the weight that variables with high values have on the model itself. If the scaling was not done, all the predicted values will rely heavily on the variables with large magnitudes.

I used all the attributes in the data set apart from the 'day', 'month', and 'location' as they did not bear as much importance as the other variables. Furthermore, 'day', 'month' and 'location' cannot be scaled logically as they are dates.

Looking at the results, the ANN has the second highest accuracy, falling short of the Random Forest by approximately 0.8%. This is same story for the AUC, showing that the ANN is strong at classifying the output class. On the other hand, the ANN has the highest precision rate out of all the models, higher by 3.9% than the second highest precision rate and tied for second last place for the recall rate.