FIT3152 Data Analytics

Assignment 2

Ritvik Mahadeshwar

29677068

Objective:

First things first the weather data explored in this report is sourced from the Kaggle competition data: Predict rain tomorrow in Australia. The data contains several metrological observations as attributed, and the class attribute "Rain Tomorrow" is our focus in the report as we are looking to predict if it will in fact rain tomorrow across 10 randomly selected Australian locations. We have taken a sample of 2000 record rows and this new synthetically created data set is called WAUS. This report will look to determine what variables within WAUS are best for predicting future rainfall patterns but also what classification model is best for purpose. We will explore many classification models and look at some techniques of how to alter them to increase their accuracy. Accuracy testing models will also be further examined.

Pre-processing Data:

The WAUS data set has been modified and cleansed by removing all N/A records or data values. This is to reduce the effect of random errors when conducting analysis within the report. Having N/A values may detrimentally impact the accuracy of our models and falsely provide us with information regarding the importance of each predictor variable.

Note: after removing N/A values our original sampling of 2000 records has now been reduced to 664 showing how important it is to filter out invalid data. We have a much more qualified data set now.

Question 1 – Exploring the data:

Variables indicating date such as day, month, year, have been removed from the data summary. When predicting if it will rain tomorrow, the predictor variable "Rain Tomorrow" is independent of the what the actual date is. This was testified beforehand with a simple linear regression model finding date variables to have little or no impact. Other variables specific to conditions such as Humidity, Temperature and Pressure are instead the factors that influence whether it will rain tomorrow. This report has a focus to analyse real-value attributes as well, focusing on numerical data. So, predictor attributes WindDir9am, WindDir3pm and WindGusDir have also been removed as they indicate direction (values are not numerical).

Proportion of rainy days to fine days:

	No	*Yes*	*Total*
Rain Today	1424	446	2000
Rain Tomorrow	1406	457	2000

Mean values of predictor variables:

Predictor Variables	Mean Value
Location	28.85
MinTemp	11.34
MaxTemp	22.99
Rainfall	1.71
Evaporation	5.30
Sunshine	7.76
WindGustSpeed	40.79
WindSpeed9am	15.31
WindSpeed3pm	19.34
Humiditiy9am	65.38

Humidity3pm	47.50
Pressure9am	1018.20
Pressure3pm	1015.92
Coud9am	4.16
Cloud3pm	4.24
Temp9am	16.40
Temp3pm	21.59

Standard Deviation of predictor variables:

Predictor Variables	Standard Deviation Value
Location	8.54
MinTemp	5.22
MaxTemp	7.00
Rainfall	5.46
Evaporation	3.64
Sunshine	3.97
WindGustSpeed	14.37
WindSpeed9am	9.59
WindSpeed3pm	9.10
Humiditiy9am	18.28
Humidity3pm	18.45
Pressure9am	7.19
Pressure3pm	7.16
Coud9am	2.86
Cloud3pm	2.67
Temp9am	5.90
Temp3pm	6.72

Question 4:

These are the classification models being used. Our refined WAUS data that contains 664 observations has been split up into 70% training (464) and 30% testing (200). A training data set to learn the new model and the testing data set to determine how well it will do on unseen data.

- Decision Tree
- Naïve Bayes
- Bagging
- Boosting
- Random Forest

Each classifier model has been built using R functions at their default setting.

Question 5:

Confusion Matrix

One way of testing the accuracy of our classifier models is to create a confusion matrix. A confusion matrix is created by using the test data and it works by adding the amount of true negatives (TN) and the amount of true positives (TP) and subsequently dividing that amount by the total amount of

predictions made. Remember confusion matrixes provides comparison at a specified confidence threshold, usually 50%.

Here is an example:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	2 (TP)	1 (FN)
	Class=No	3 (FP)	0 (TN)

Most widely-used metric:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{2+0}{6} = 33.3\%$$

Decision Tree Matrix:

		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	145	16
	Rain Tomorrow=Yes	17	22

Accuracy = 83.5%

Naïve Bayes Matrix:

		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	137	24
	Rain Tomorrow=Yes	14	25

Accuracy = 81%

Bagging Model Matrix:

		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	152	9
	Rain Tomorrow=Yes	23	16

Accuracy = 84%

Boosting Model Matrix:

		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	148	13
	Rain Tomorrow=Yes	18	21

Accuracy = 84.5%

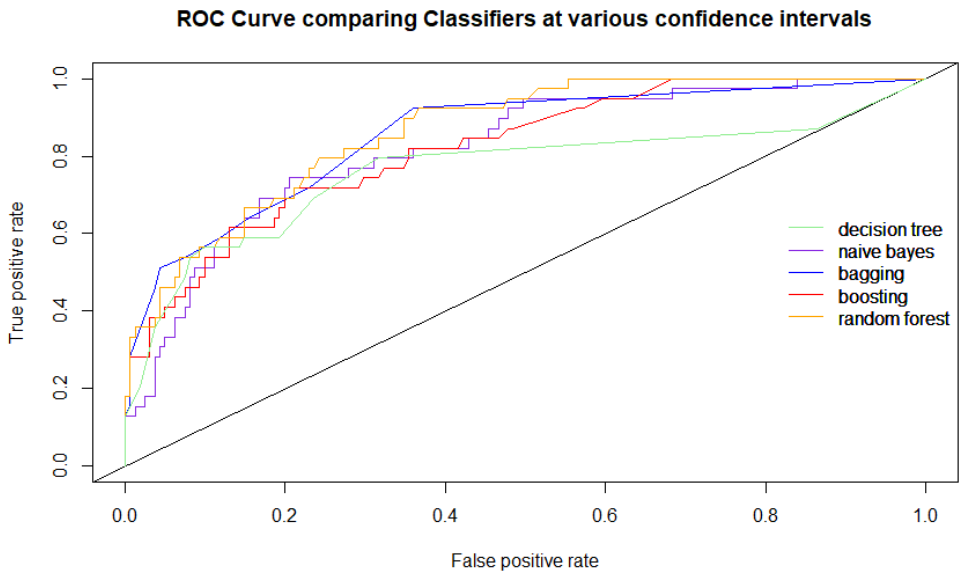
Random Forest Matrix:

		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	153	8
	Rain Tomorrow=Yes	21	18

Accuracy = 85.5%

Question 6 and 7 – Comparing classifiers through various performance measurement tools

The ROC Curve plots True Positive Rate, TPR, (on y axis) against False Positive Rate, FPR, (on x axis). The beauty with ROC curves is that they can be graphed at varying confidence 'threshold' levels for each classifier, while also giving a more comprehensive comparison of several classifiers. As we can see from the graph below with every different confidence threshold the correct classifications and 'false alarm" rates vary.



So far, we have discussed several methods on how to effectively evaluate our classifier models. From confusion matrices to the ROC Curve constructed above. However, there is one more that we are going to explore called AUC (Area under the curve). An AUC essentially can be calculated from the graph above and is an overall single measure of a test performance. Like its name the test performance is found by calculating the area under the graph in the ROC for each classifier. A classifier plot with a greater AUC is one that has a greater true positive to false positive ratio. The table below gives a nice summary of the results we achieved with each classifier. Classifiers are being compared based on two factors, their ability to predict rainfall accurately (accuracy – confusion matrix) and the area under the curve (AUC) when plotted in a ROC curve.

Classifier	Accuracy	AUC
Decision Tree	83.5%	0.7626
Naïve Bayes	81%	0.8207
Bagging	84%	0.8344
Boosting	84.5%	0.8446
Random Forest	85.5%	0.8646

After conducting accuracy calculations through confusion matrices and graphing ROC Curves to calculate AUC for each classifier, the data supports that 'Random Forest is the best classifier when looking to predict whether it will rain tomorrow.

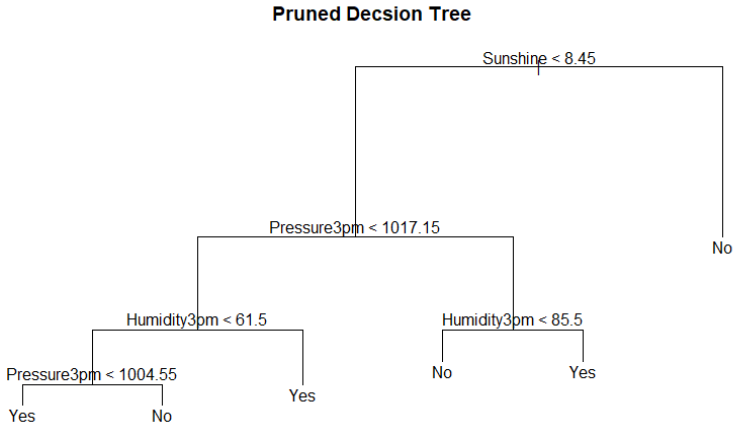
Question 8 – Important Variables

From the WAUS data set, it is quite evident there are some variables that give a greater indication of whether or not it is likely to rain in the following day. The graphic below uses the variable importance from Random Forest to tell the relative contribution of each predictor attribute to the classification. Random Forest has been selected as the classifier to analyse the most important variable because it was our most accurate classifier. However, it should be noted this report has done analysis on the other models as well, which have told a very similar story. The diagram uses a gini index to effectively show how important a variable is for estimating the value of the target variable across all the trees that make up the forest. A higher Mean Decrease in Gini indicates higher variable importance. The single greatest indicator of whether or not it will rain tomorrow is the Humidity at 3pm followed by hours of sunshine recorded over the previous day.

	MeanDecreaseGini
Location	2.6488
MinTemp	8.5453
MaxTemp	6.1626
Rainfall	9.3650
Evaporation	8.4117
Sunshine	19.0675
WindGustSpeed	8.5402
WindSpeed9am	4.6362
WindSpeed3pm	5.0318
Humidity9am	6.3645
Humidity3pm	20.0109
Pressure9am	14.0079
Pressure3pm	18.9723
Cloud9am	3.9707
Cloud3pm	5.6221
Temp9am	6.6458
Temp3pm	7.5570
RainToday	0.8705

Some findings that may seem surprising are the low Mean Decrease Gini values for Cloud9am and Cloud3pm, indicating that they have a relatively small impact on whether or not it will rain tomorrow. The cloud variable looks at what fraction of the sky is obscured by cloud. Everyone has this conception that overcast conditions generally lead to rain, but our Random Forest classifier says so otherwise and instead highlights that attributes like humidity, atmospheric pressure and hours of sunshine contribute more towards the outcomes of rain for the subsequent day. In addition, the variable of 'Location' is shown to of least importance and can likely be omitted from this study.

Question 9 – Improving the "Decision Tree" classifier:



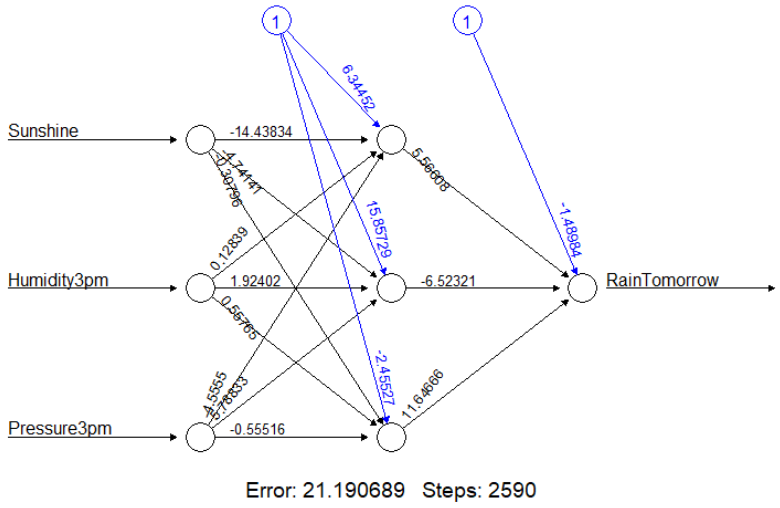
While the Random Forest is the best classifier and it can be furthered improved. The Decision Tree model is the most widely used model in the machine learning and AI space today. It is used by many developers because of its simplicity but also effectiveness. So, for sake of user cases and relative usefulness to real-life application we have decided to cross-validate and then prune the decision tree constructed in question 4 to give a much more refined and more accurate tree as shown above (the original decision tree had 21 terminal nodes, whereas the pruned one has 6). The original decision tree is presented in the appendix for those who want to further investigate. To achieve this pruned decision tree, the original decision tree was first cross-validated using the cv.tree function. This allows us to see at what size of the tree (at how many leaf-nodes) does our decision tree give the least counts of misclassifications. As you can see from the diagram above, in our case it was at 6 leaf-nodes. Then a prune.misclass function was applied to reduce the size of the tree to having 6 leaf nodes. In this example the post-pruning method was utilised. What is interesting to also consider is our decision tree uses predictor variables: Sunshine, Pressure3pm and Humidity3pm, the most important variables in our Random Forest classifier, to construct the decision tree in order to increase the homogeneity in the resulting datasets. These variables seem to consistently be the most weighted predictors in affecting the outcome of whether or not it will rain tomorrow in other classifiers too.

Performance Evaluation:

	Accuracy	AUC
Decision Tree	83.5%	0.7626
Pruned Decision Tree	85%	0.7626

In terms of results we can see post-pruning our original decision tree has resulted in an increase in accuracy of the model. Even with no change observed in the AUC a 1.5% increase is significant enough for this model to be considered a clearly better predictor of future rainfall.

Question 10 – Artificial Neural Networks



The above graph is a depiction of an Artificial Neural Network classifier for our WAUS data.

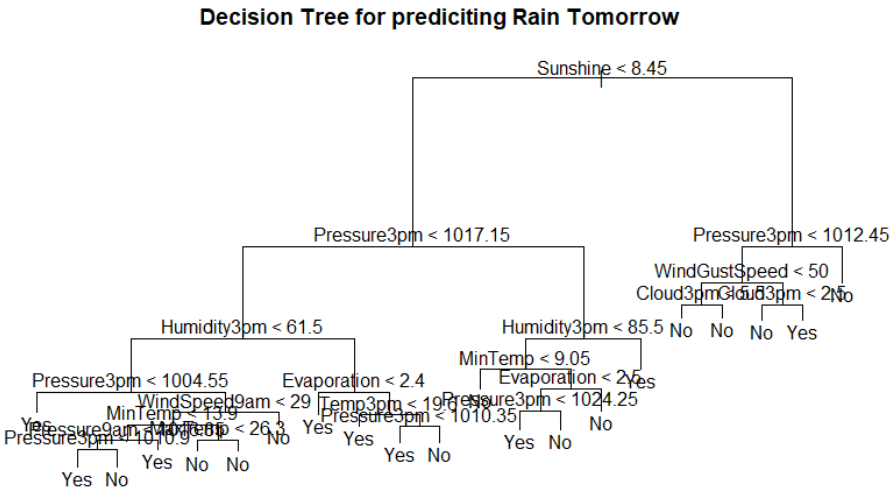
		Predicted	
		Rain Tomorrow=No	Rain Tomorrow=Yes
Observed	Rain Tomorrow=No	141	3
	Rain Tomorrow=Yes	27	15

Accuracy = 83.9%

Once again, the attributes used in this classifier are Sunshine, Humidity3pm and Pressure3pm. Like mentioned before these predictor variables are consistently identified as the most important variables in each classifier model we have examined in this report. They seem contribute the most in determining whether or not it will rain tomorrow so it makes sense to utilise them in our artificial neural network. There were a few data-processing processes conducted. Firstly, removing N/A's a habit that we all should be accustomed to. Sampling 2,000 observations from our WAUS data set and splitting it into 70% for a training set and 30% for a testing set. Finally recoding our Rain Tomorrow column into "No" = 0 and "Yes" = 1. Our Artificial Neural Network records an accuracy of 83.9%, meaning it hasn't performed as well as compared to our other classification models. In fact, it is just better than our original decision tree model and Naïve Bayes model from question 4 and falls behind the other 3 we have examined. Factors that could have led to this rather lower accuracy could be because we have only accounted 3 predictor variables in our network analysis and that we have only sampled 2,000 observations. As a thumb rule, a bigger sample size generally always leads to better results or in our case a higher accuracy.

Appendix:

Original Decision Tree:




```
Classification tree:
tree(formula = RainTomorrow ~ ., data = WAUSComplete.train)
Variables actually used in tree construction:
 [1] "Sunshine"      "Pressure3pm"   "Humidity3pm"   "WindSpeed9am"  "MinT
emp"        "Pressure9am"
 [7] "MaxTemp"       "Evaporation"   "Temp3pm"       "WindGustSpeed" "Clou
d3pm"
Number of terminal nodes:  21
Residual mean deviance:  0.347 = 154 / 443
Misclassification error rate: 0.0603 = 28 / 464
```