It rains today, will it rain again tomorrow?

Contents

Tables

Figures

1. Introduction

1.1. Project aim

Predicting real-time and accurate rainfall remains challenging for many decades due to its non-linear nature. Weather forecast can be divided into various terms such as real-time, short-term, middle-term, long-term. In our project, we focus on short term rainfall prediction whose forecasting range is 0-72 hours and able to predict next day rain. This report evaluates the effectiveness of different outlier detection, classification, and machine learning methods to predict rainfall in terms of weather forecast. There are various factors such as Humidity and Cloud are influencing the weather patterns and affects different parts of the world. Our purpose is to find out those factors, which can play important part in rainfall prediction. So, our team is going to forecast rainfall in a part of Australia such as Canberra and Adelaide. In addition to that, our goal is to use machine learning methods for predicting accurate next day rainfall depends on today's rain.

1.2. Project background

Australia is home to some of driest regions in the world. It is gone through one of the worst droughts in its history between 2003 and 2012. however, has since seen an increase in average rainfall. Moreover, it has suffered from devastating bushfires and severe flooding recently. So, it can be said that forecasting rainfall is a key element of climate system to overcome natural disasters Australia and all over the world. Moreover, predecessors' work has collected much data recording Australian rainfall parameters over many years, but in the meantime, the performance of the long-term prediction results could be merely less than 80%. With short term prediction, the model performed better with approximately 90% accuracy in five-day forecast (https://scijinks.gov/forecast-reliability/). Hence, to tackle this issue in this Project, our main goal is to build a prediction model for short term rainfall forecast, which will predict rainfall on the next day in Australia. The data set covers 49 locations of Australia. But because of time constrain, we have just considered two areas for our project: Adelaide and Canberra.

1.3. Report structure

The report is being divided into different sections. Each section explains our project analysis in detail. In first section, we have discussed about introduction of the project. Literature review has been done in second section, which includes three types of machine learning methods. It is important to uncovered initial patterns and characteristics of dataset, so data exploration has been done as a first step of data analysis in third section. It consists of feature selection, data preprocessing and outlier detections. The fourth section is a heart of our projects, which builds machine learning model such as logistic regression, Artificial neural network, and support vector machine to predict rain from source data by considering with outliers and without outliers. At the end we will conclude which machine learning model is giving best result to forecast next day rainfall.

2. Literature review

2.1. Logistic Regression

Logistic regression model is to study the relationship between two or more variables. The outcome variable in logistic regression is Y (known as dependent variable) is expressed by variable X which is known as explanatory variable in a linear function. The basic model to express the relationship between two variables Y and X could be show as: $Y_i = c + \alpha X_i$ where c is a constant value for linear interception and α is a regression weight equivalent to expect change in Y_i per unit change in X_i (also known as a slope of the regression equation) (Cheung, Hart & Peart 2015; Ozechowski 2010; Tolles & Meurer 2016). Since the outcome rainfall is either 'yes' or 'no', it is reasonable to apply logistics regression model to predict whether it is going to rain tomorrow.

Imon et al. (2012) have built logistic regression models to predict the rainfall. This rainfall prediction has been done for the data of Giridhi, Bihar and India. Their approach is to formulate the model for fourteen years of data and then two years of data has been used as a future data for the cross-validation model. In our research we will build logistic regression model for rainfall forecast in Australia, and we will use existing resources to understand the application of logistic regression into rainfall prediction.

Adjusted R2 can be used to measure Goodness of fit in logistic regression model. While predicting rainfall, it has been found that logistic regression model suffers from the Masking (false negative) and swamping (false positive) problems when outliers present in the data.so outlier should be omitted from the data and then Logistic regression model will be fitted again to get accurate predictions. According to Imon et al. (2012), there are some climatic variables that important predictors for the rainfall such as evaporation, maximum temperature, minimum temperature, and humidity at certain time. In our project we have a similar predictors variable like humidity and evaporation at certain time which will be helpful to us for predicting a rainfall.

Moreover, the combination of logistic regression and generalized linear model has been used to predict the rainfall in Hong Kong. The data which has been used for this study are historical observations, gridded NCEP/NCAR reanalysis data and GCM projections of future climate scenarios. Logistic regression model has been used to determine rain occurrence and as a result, binary variables were created (rain=1 ,non-rain=0) and then using them liner model predicted the rainfall amount in terms of monthly rain days and monthly rainfall volume by Cheung, Hart & Peart (2015).The data set rainfall has also same variable "Rain Tomorrow" to predict rainfall.

From above all, it has been proved that Logistic regression model is a multi-predictor by predicting the rainfall in all India and its subdivisions such

Orrisa and Gujarat. Where DEMETER retrospective forecasts with a lead time of 1 month. The selection of variable has been done using Brier score where Brier score has been obtained by fitting the logistic regression by a cross-validation approach Prasad, Dash & Mohanty (2010).

As the result, the performance of logistic regression models from other researchers are at satisfactory status. This model is also very successful in rainfall forecast problem. Hence, logistic regression method can be applied to our data.

2.2. Artificial Neural Network (ANN)

During the process of review of the literature, the team noticed that the ANN technique has been widely used for predicting rainfall and other time series data. A neural network is a group of neurons interconnected coherently with one output neuron, which reports the result by Mandal & Jothiprakash (2012). Based on the observation of various papers analyzing the ANN technique, the team found that Feed forward back propagation neural network (BPN) is an extremely popular neural network model (Devi et al. 2016).

Devi et al. (2016) has proved this point by indicating that almost 90% of applications use this algorithm, which is a systematic method of training multilayer artificial neural networks. Their research also compared approaches and performances on predicting short-term rainfall among BPN and other ANN models. It is found that all ANN models have capability of capturing low and high intensity of rainfall pattern and the performance of BPN can be noticeably improved when there is one more added hidden layer of neurons. Mandal & Jothiprakash (2012) also proved the excellent capability of ANN in the experiment predicting the very next-day rainfall, even though the dataset only contains antecedent rainfall information without any other weather data.

Samhitha & Srikanth (2017) has predicted the rainfall using ANN (Artificial neural network) and IDW (Inverse Distance weighting) method for the PONNIYAR river basin in terms of five year of data from 2012 to 2016.They have used ANN using back propagation for prediction because back propagation works well and one of the widely used algorithm in Artificial neural network. ANN predict new outcomes with the help of input data as training data. At the result of ANN model and IDW were compared. The result shows that ANN gives better rainfall prediction as compared to the IDW.

Nanda et al. (2013) propose to build a new technique based in ARIMA and others ANN models to predict annually rainfall. Multi-Layer Perceptron (MLP) which is one of the implementations of BNP, Functional-link Artificial Neural Network (FLANN) and Legendre Polynomial Equation (LPE) has been used to predict timeseries data. The researchers compared all proposed ANN and ARIMA models to get better result. Consequently, it was found that generally ANN models give better result in terms of Rainfall forecasting.

2.3. Support Vector Machine (SVM)

SVM is a machine learning tool. It can be used as a supervised machine learning algorithm for classification and regression (Nayak, Ghosh 2013; Seo, Lee & Kim 2014). SVM belongs to category of kernel methods. This classifier widely used in bioinformatics because of its high accuracy. Also, SVM has an ability to deal with high dimensional data and in modelling diverse source of data for example gene expression.

Nayak, Ghosh (2013) predicted extreme rainfall in Mumbai using SVM method. They have used two weather patterns related to Mumbai rainfall. For example, two phase SVM model training set and Testing/Prediction set which has given significant improvements to predict the extreme rainfall in Mumbai. It also has been found that Model building of SVM is a hit and trial process. Which means finding a best kernel and its parameters, weights, and bias. Result shows large number of SVM affects to the generalization and small number of SVM increase the computational cost. To get the best prediction there should be 50 and 30 no extreme instances for training two model of SVM.

Most of the time the data cannot be easily classified because of errors in some features of instances. SVM introduce a margin called soft margin which solves the errors in data and misclassified such instances. So, it has been proved that it allows outlier to be misclassified without affecting the result by Nayak, Ghosh (2013). This can be a great advantage of SVM to our project.

Seo, Lee & Kim (2014) have predicted a heavy rainfall in a south Korea with lead time of one to 6 hours. They have used KNN, K-VNN and SVM algorithms in their discriminant analysis by dividing the data set in three parts training set, validation set and the test set. Also, Wrapper method has been used to solve the problem of the feature selection. Overall, it has been found that Prediction using SVM kernel method was slightly better as compared to other methods.

2.4. Literature Review Summary

From the Literature review we realized that every model works better in their own way. For example, Logistic regression model is a multi-predictor by predicting rainfall in different parts of the world together. Also, ANN models work well even with short term rainfall prediction data and has an ability to capture low and high intensity of rainfall pattern. While SVM can misclassified outliers without affecting the data. So, at this stage it is hard to give a name of best model, but we can analyze that after our practical data analysis on our rainfall data.

3. Data exploration

3.1. Data Overview

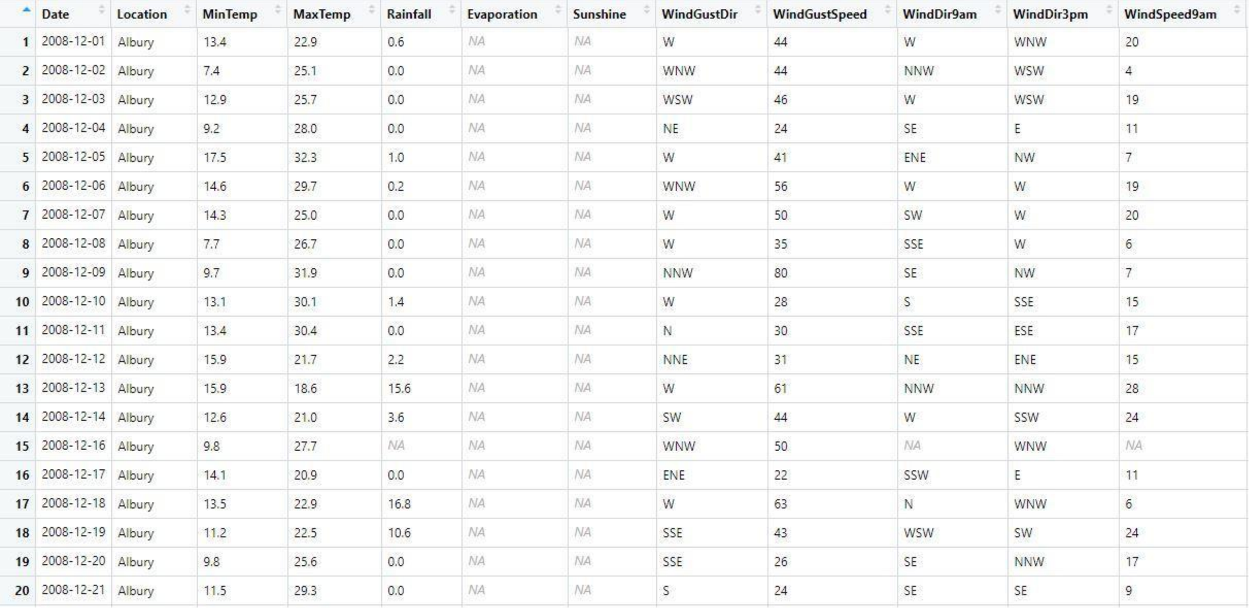
Dataset is used in this research is “weatherAUS.csv” and is published on Kaggle.com (source: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package). The dataset consists of 142,193 observations and 24 columns or also known as attributes. There is also large amount of missing value in the dataset, which is 316,559 missing value over 3,412,632 total values, take up to 9.28% of the whole dataset. Data was collected from all the states and territories in Australia, and it covers from 01/11/2007 to 25/06/2017. Table 1 shows the name of individual attributes in the dataset as well as its description.

Attributes	Format	Description	Data Type
Date	yyyy/mm/dd	The date of observation.	Nominal
Location	String	The common name of the location of the weather station.	Nominal
MinTemp	Decimal number	The minimum temperature in degrees Celsius.	Interval
MaxTemp	Decimal number	The maximum temperature in degrees Celsius.	Interval
Rainfall	Decimal number	The amount of rainfall recorded for the day in mm.	Interval
Evaporation	Decimal number	The so-called Class A pan evaporation (mm) in the 24 hours to 9am.	Interval
Sunshine	Decimal number	The number of hours of bright sunshine in the day.	Interval
WindGustDir	String	The direction of the strongest wind gust in the 24 hours to midnight.	Nominal
WindGustSpeed	Integer	The speed (km/h) of the strongest wind gust in the 24 hours to midnight.	Interval
WindDir9am	String	Direction of the wind at 9am.	Nominal
WindDir3pm	String	Direction of the wind at 3pm.	Nominal
WindSpeed9am	Integer	Wind speed (km/hr) averaged over 10 minutes prior to 9am.	Interval
WindSpeed3pm	Integer	Wind speed (km/hr) averaged over 10 minutes prior to 3pm.	Interval
Humidity9am	Integer	Humidity (percent) at 9am.	Interval
Humidity3pm	Integer	Humidity (percent) at 3pm.	Interval
Pressure9am	Decimal number	Atmospheric pressure (hpa) reduced to mean sea level at 9am.	Interval
Pressure3pm	Decimal number	Atmospheric pressure (hpa) reduced to mean sea level at 3pm.	Interval

Cloud9am	Integer	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.	Interval
Cloud3pm	Integer	Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values.	Interval
Temp9am	Decimal number	Temperature (degrees C) at 9am.	Interval
Temp3pm	Decimal number	Temperature (degrees C) at 3pm.	Interval
RainToday	String	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.	Binary
RISK_MM	Decimal number	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".	Interval
RainTomorrow	String	The target variable. Will it rain tomorrow?	Binary (Target)

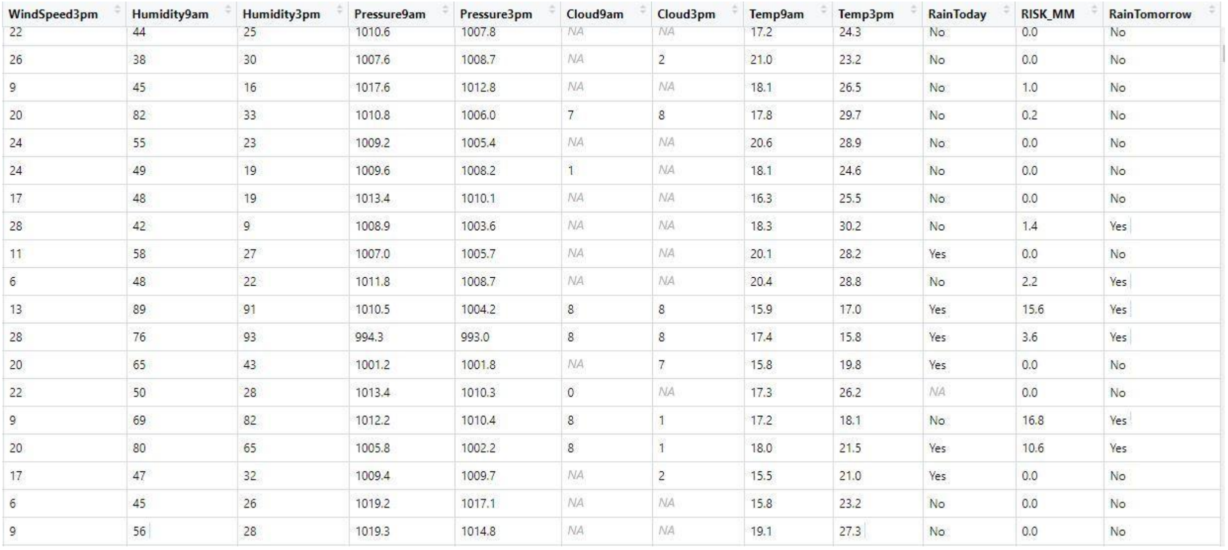
Table 1. Data Dictionary

As shown in Figure 1 and Figure 2 is the visualization of the dataset. The variable that we are interested in is “RainTomorrow” which is also a target variable for this research. It is labeled with the value of “Yes” or “No” to represent whether it is rain in the next day.



	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am
1	2008-12-01	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW	20
2	2008-12-02	Albury	7.4	25.1	0.0	NA	NA	WNW	44	NNW	WSW	4
3	2008-12-03	Albury	12.9	25.7	0.0	NA	NA	WSW	46	W	WSW	19
4	2008-12-04	Albury	9.2	28.0	0.0	NA	NA	NE	24	SE	E	11
5	2008-12-05	Albury	17.5	32.3	1.0	NA	NA	W	41	ENE	NW	7
6	2008-12-06	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W	19
7	2008-12-07	Albury	14.3	25.0	0.0	NA	NA	W	50	SW	W	20
8	2008-12-08	Albury	7.7	26.7	0.0	NA	NA	W	35	SSE	W	6
9	2008-12-09	Albury	9.7	31.9	0.0	NA	NA	NNW	80	SE	NW	7
10	2008-12-10	Albury	13.1	30.1	1.4	NA	NA	W	28	S	SSE	15
11	2008-12-11	Albury	13.4	30.4	0.0	NA	NA	N	30	SSE	ESE	17
12	2008-12-12	Albury	15.9	21.7	2.2	NA	NA	NNE	31	NE	ENE	15
13	2008-12-13	Albury	15.9	18.6	15.6	NA	NA	W	61	NNW	NNW	28
14	2008-12-14	Albury	12.6	21.0	3.6	NA	NA	SW	44	W	SSW	24
15	2008-12-16	Albury	9.8	27.7	NA	NA	NA	WNW	50	NA	WNW	NA
16	2008-12-17	Albury	14.1	20.9	0.0	NA	NA	ENE	22	SSW	E	11
17	2008-12-18	Albury	13.5	22.9	16.8	NA	NA	W	63	N	WNW	6
18	2008-12-19	Albury	11.2	22.5	10.6	NA	NA	SSE	43	WSW	SW	24
19	2008-12-20	Albury	9.8	25.6	0.0	NA	NA	SSE	26	SE	NNW	17
20	2008-12-21	Albury	11.5	29.3	0.0	NA	NA	S	24	SE	SE	9

Figure 1. Data Overview 1 (data_overview_1.JPG)



WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RISK_MM	RainTomorrow
22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	0.0	No
26	38	30	1007.6	1008.7	NA	2	21.0	23.2	No	0.0	No
9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	1.0	No
20	82	33	1010.8	1006.0	7	8	17.8	29.7	No	0.2	No
24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	0.0	No
24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	0.0	No
17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	0.0	No
28	42	9	1008.9	1003.6	NA	NA	18.3	30.2	No	1.4	Yes
11	58	27	1007.0	1005.7	NA	NA	20.1	28.2	Yes	0.0	No
6	48	22	1011.8	1008.7	NA	NA	20.4	28.8	No	2.2	Yes
13	89	91	1010.5	1004.2	8	8	15.9	17.0	Yes	15.6	Yes
28	76	93	994.3	993.0	8	8	17.4	15.8	Yes	3.6	Yes
20	65	43	1001.2	1001.8	NA	7	15.8	19.8	Yes	0.0	No
22	50	28	1013.4	1010.3	0	NA	17.3	26.2	NA	0.0	No
9	69	82	1012.2	1010.4	8	1	17.2	18.1	No	16.8	Yes
20	80	65	1005.8	1002.2	8	1	18.0	21.5	Yes	10.6	Yes
17	47	32	1009.4	1009.7	NA	2	15.5	21.0	Yes	0.0	No
6	45	26	1019.2	1017.1	NA	NA	15.8	23.2	No	0.0	No
9	56	28	1019.3	1014.8	NA	NA	19.1	27.3	No	0.0	No

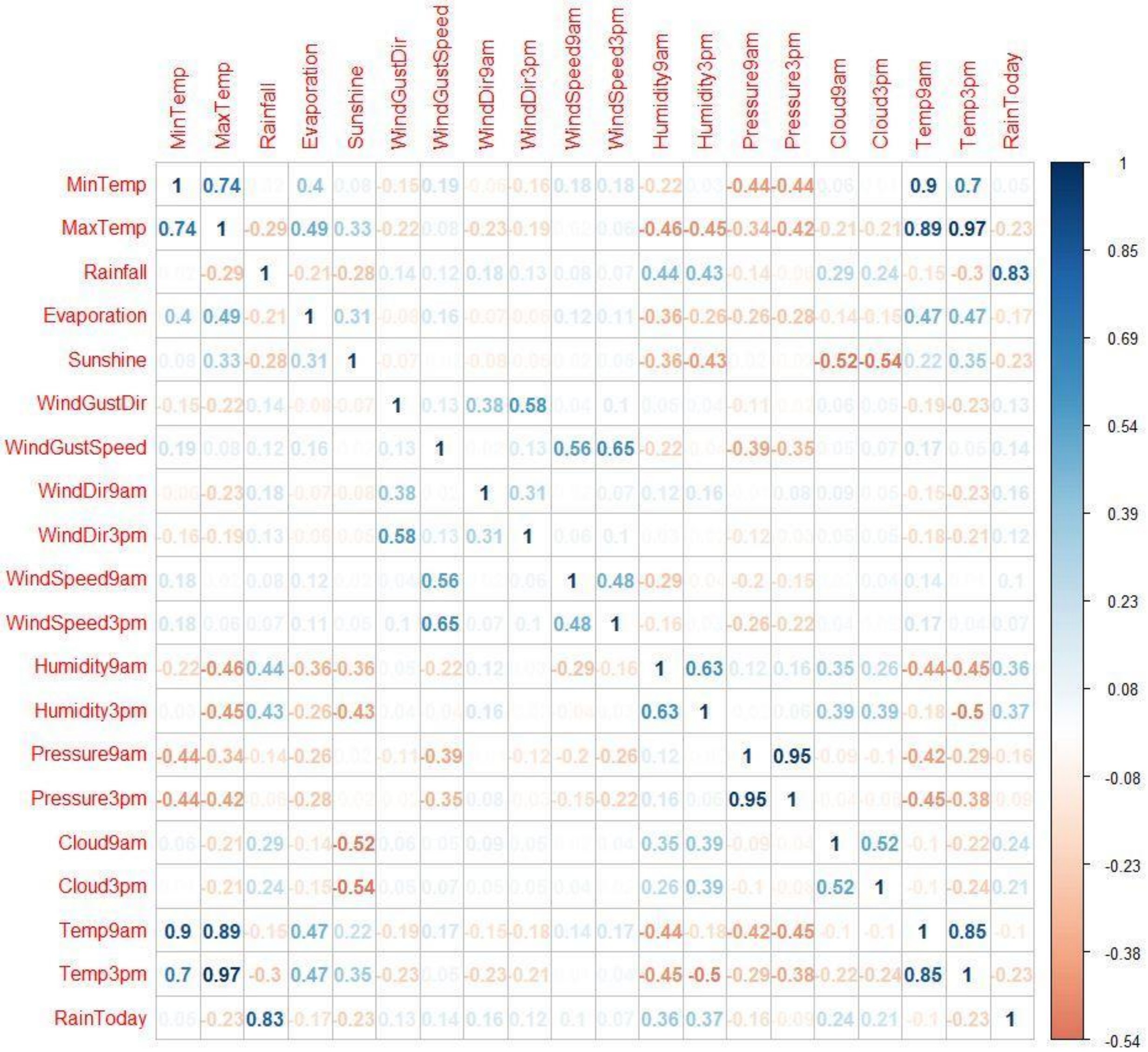
Figure 2. Data Overview 2 (data_overview_2.JPG)

3.2. Features Selection

As results of correlation matrix shown in Figure 3, the desire cutoff point for highly correlated variables will be 0.7. Because the variables are highly correlated with each other, they will have the same effect on the target variable. Hence, all the highly correlated variables are suggested to be removed from the analysis such as “Temp9am”, “Temp3pm”, “Rainfall”. However, with the case of 2 pairs “MinTemp” and “MaxTemp”, “Pressure9am” and “Pressure3pm”, instead of removing one of the attributes in each pair, new variables will be created to preserve the full meaning of the variables. The 2 new variables will be:

- $DailyTempMean = \frac{MaxTemp+MinTemp}{2}$
- $DailyPressureMean = \frac{Pressure9am+Pressure3pm}{2}$

According to the data dictionary, variable “RISK_MM” is the amount of next day rain in mm and used to create response variable RainTomorrow. A kind of measure of the "risk". Hence, “RISK_MM” is also be removed from the analysis.



	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday
MinTemp	1	0.74	0.02	0.4	0.08	-0.15	0.19	-0.06	-0.16	0.18	0.18	-0.22	0.03	-0.44	-0.44	0.06	0.04	0.9	0.7	0.05
MaxTemp	0.74	1	-0.29	0.49	0.33	-0.22	0.08	-0.23	-0.19	0.02	0.06	-0.46	-0.45	-0.34	-0.42	-0.21	-0.21	0.89	0.97	-0.23
Rainfall	0.02	-0.29	1	-0.21	-0.28	0.14	0.12	0.18	0.13	0.08	0.07	0.44	0.43	-0.14	-0.06	0.29	0.24	-0.15	-0.3	0.83
Evaporation	0.4	0.49	-0.21	1	0.31	-0.08	0.16	-0.07	-0.06	0.12	0.11	-0.36	-0.26	-0.26	-0.28	-0.14	-0.15	0.47	0.47	-0.17
Sunshine	0.08	0.33	-0.28	0.31	1	-0.07	0.01	-0.08	0.05	0.02	0.06	-0.36	-0.43	0.02	0.02	-0.52	-0.54	0.22	0.35	-0.23
WindGustDir	-0.15	-0.22	0.14	-0.08	-0.07	1	0.13	0.38	0.58	0.04	0.1	0.05	0.04	-0.11	0.02	0.06	0.05	-0.19	-0.23	0.13
WindGustSpeed	0.19	0.08	0.12	0.16	0.01	0.13	1	0.02	0.13	0.56	0.65	-0.22	0.04	-0.39	-0.35	0.05	0.07	0.17	0.05	0.14
WindDir9am	-0.06	-0.23	0.18	-0.07	-0.08	0.38	0.02	1	0.31	0.02	0.07	0.12	0.16	0.01	0.08	0.09	0.05	-0.15	-0.23	0.16
WindDir3pm	-0.16	-0.19	0.13	-0.06	0.05	0.58	0.13	0.31	1	0.06	0.1	0.03	0.02	-0.12	0.03	0.05	0.05	-0.18	-0.21	0.12
WindSpeed9am	0.18	0.02	0.08	0.12	0.02	0.04	0.56	0.02	0.06	1	0.48	-0.29	0.04	-0.2	-0.15	0.01	0.04	0.14	0.01	0.1
WindSpeed3pm	0.18	0.06	0.07	0.11	0.05	0.1	0.65	0.07	0.1	0.48	1	-0.16	0.03	-0.26	-0.22	0.04	0.01	0.17	0.04	0.07
Humidity9am	-0.22	-0.46	0.44	-0.36	-0.36	0.05	-0.22	0.12	0.03	-0.29	-0.16	1	0.63	0.12	0.16	0.35	0.26	-0.44	-0.45	0.36
Humidity3pm	0.03	-0.45	0.43	-0.26	-0.43	0.04	0.04	0.16	0.02	0.04	0.03	0.63	1	0.02	0.06	0.39	0.39	-0.18	-0.5	0.37
Pressure9am	-0.44	-0.34	-0.14	-0.26	0.02	-0.11	-0.39	0.01	-0.12	-0.2	-0.26	0.12	0.02	1	0.95	-0.09	-0.1	-0.42	-0.29	-0.16
Pressure3pm	-0.44	-0.42	-0.06	-0.28	0.02	0.02	-0.35	0.08	0.03	-0.15	-0.22	0.16	0.06	0.95	1	-0.04	-0.08	-0.45	-0.38	-0.09
Cloud9am	0.06	-0.21	0.29	-0.14	-0.52	0.06	0.05	0.09	0.05	0.01	0.04	0.35	0.39	-0.09	-0.04	1	0.52	-0.1	-0.22	0.24
Cloud3pm	0.04	-0.21	0.24	-0.15	-0.54	0.05	0.07	0.05	0.05	0.04	0.01	0.26	0.39	-0.1	-0.08	0.52	1	-0.1	-0.24	0.21
Temp9am	0.9	0.89	-0.15	0.47	0.22	-0.19	0.17	-0.15	-0.18	0.14	0.17	-0.44	-0.18	-0.42	-0.45	-0.1	-0.1	1	0.85	-0.1
Temp3pm	0.7	0.97	-0.3	0.47	0.35	-0.23	0.05	-0.23	-0.21	0.01	0.04	-0.45	-0.5	-0.29	-0.38	-0.22	-0.24	0.85	1	-0.23
RainToday	0.05	-0.23	0.83	-0.17	-0.23	0.13	0.14	0.16	0.12	0.1	0.07	0.36	0.37	-0.16	-0.09	0.24	0.21	-0.1	-0.23	1

Figure 3. Correlation matrix (correlation_matrix.JPG)

3.3. Outlier detection

3.3.1. K-mean Clustering

Several partition clustering methods were used to detect potential outliers for the dataset. However, it did not perform well in this dataset. For K-mean clustering, it took significant amount of time to compute the distance and produced a cluster plot from the dataset. implement principle components analysis (PCA) to reduce the dimensionality of the dataset and partition data into 3 clusters. Table 2 shows 2 principle components of the dataset, component 1 can be named as humidity since all the variables in component 1 are relevant to humidity factor, and component 2 can be named as wind since all the variables in component 2 are relevant to wind factor.

Attributes	Principle Component 1	Principle Component 2

Humidity9am	0.77	
Humidity3pm	0.75	
Sunshine	-0.69	
Cloud9am	0.63	
Cloud3pm	0.61	
Evaporation	-0.54	
DailyTempMean	-0.53	
RainToday	0.47	0.31
WindDir9am		
WindGustSpeed		0.83
WindSpeed3pm		0.74
WindSpeed9am		0.69
DailyPressMean		-0.58
WindGustDir		0.33
WindDir3pm		0.32

Table 2. PCA Results

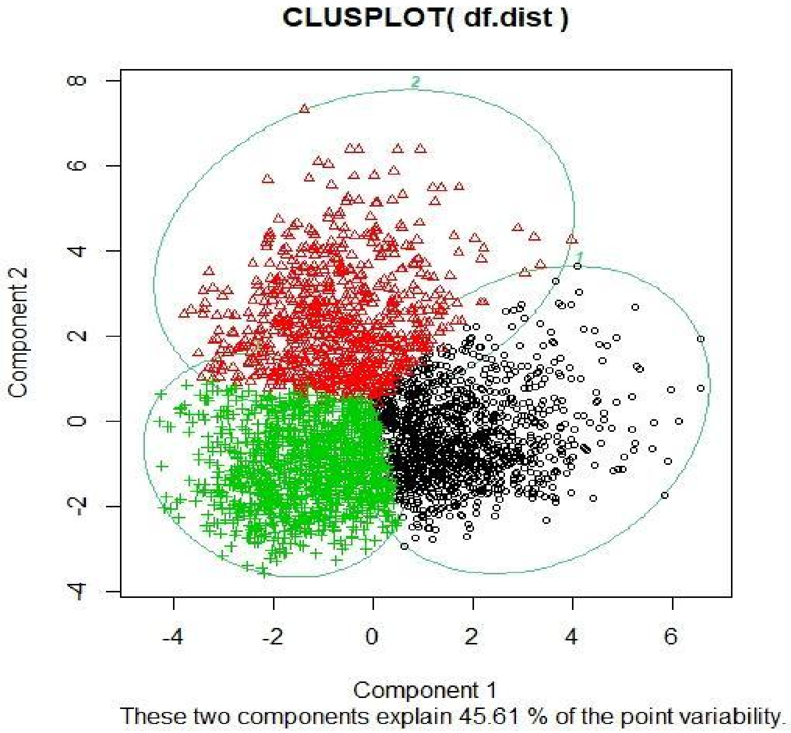


Figure 4. K-mean clustering result

As the result shown in Figure 4, 3 was suggested to be the best number of clusters for this dataset. However, there is high density data points in the middle of the plot and the clusters are overlapped on each other and the same results were found after trying other number of clusters for this dataset, suggesting K-mean is not a suitable clustering method for this dataset in order to find outlier. Hence, DBScan method will be performed to identify cluster and potential outliers for this dataset.

3.3.2. DBScan clustering

Density-based spatial clustering of application with noise (DBScan) is one of the methods used to partition the cluster that has high density from cluster that has low density data points. For each point in the cluster, it will scan for the minimum number of neighbor points with a specified radius. If there are enough points to satisfy the minimum points within the given radius condition, those points will form a cluster.

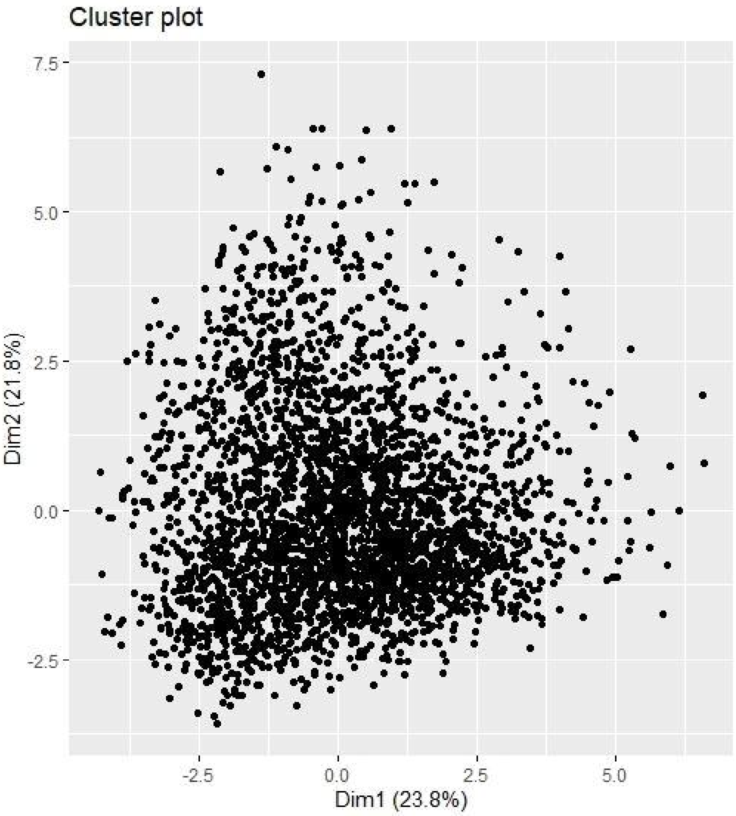


Figure 5. DBScan clustering result

As the result shown in Figure 5 is the DBScan clustering result for the dataset, different configuration of minimum neighbor points as well as the epsilon value have been tried in the DBScan method. However, with different configurations for number of minimum points and radius was performed in the experiment, the result only shows 1 cluster. Hence, it is suggested that DBScan is not suitable to identify potential outliers in this dataset as well.

3.3.3. Histogram method

As partition clustering method is not suitable to detect potential outliers in this dataset, we will go with histogram method to identify potential outliers. Histogram method is appropriate for weather dataset where it measures the frequency of certain factor. Using Figure 6 as an example, it shows the distribution of the average daily pressure in Adelaide, we can identify what value of average daily pressure occurs the most in Adelaide and what value is the least, the histogram is also help us to identify some of the unusual value which appeared on the left of the histogram, there is only one day that the average pressure fall below 990hpa in Adelaide.

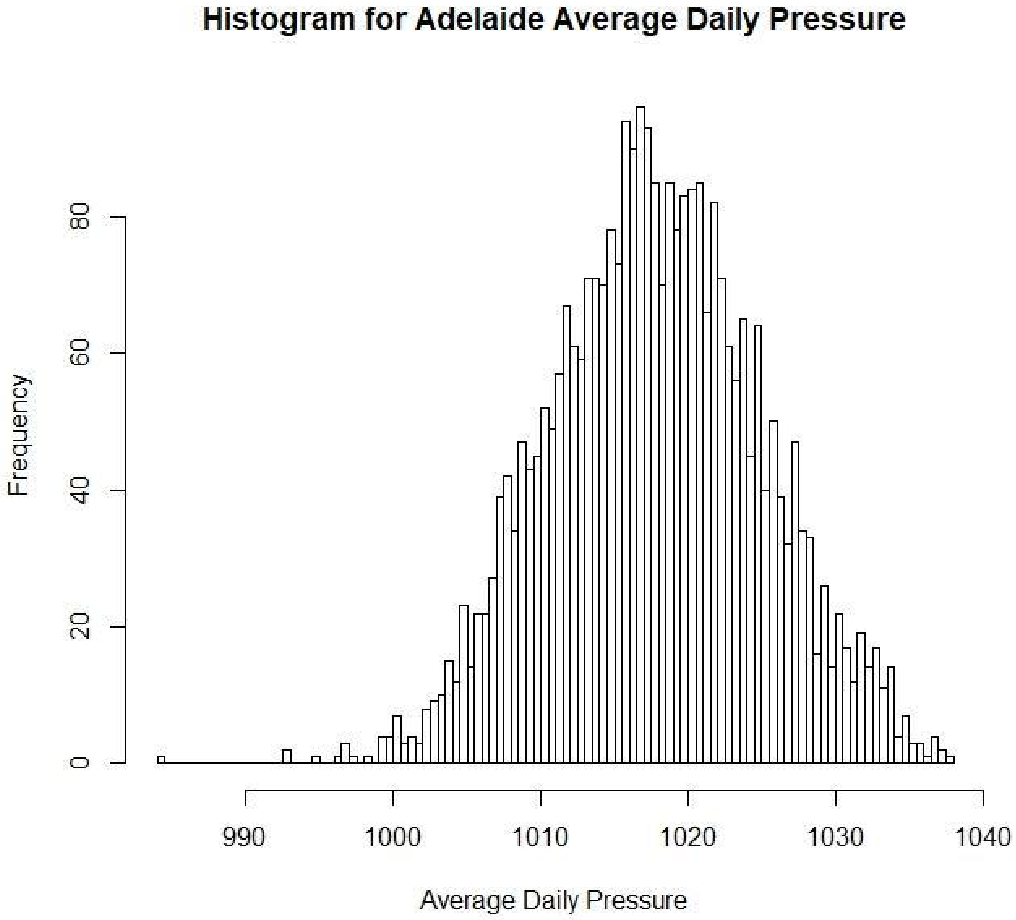


Figure 6. Average Daily Pressure in Adelaide

As shown in Figure 6 is the distribution of the average daily pressure in Adelaide, there is an observation on the left of histogram that is significant further from majority of the observations in the dataset. It is suggested for the observation that has average daily pressure below 990hpa to be potential outliers.

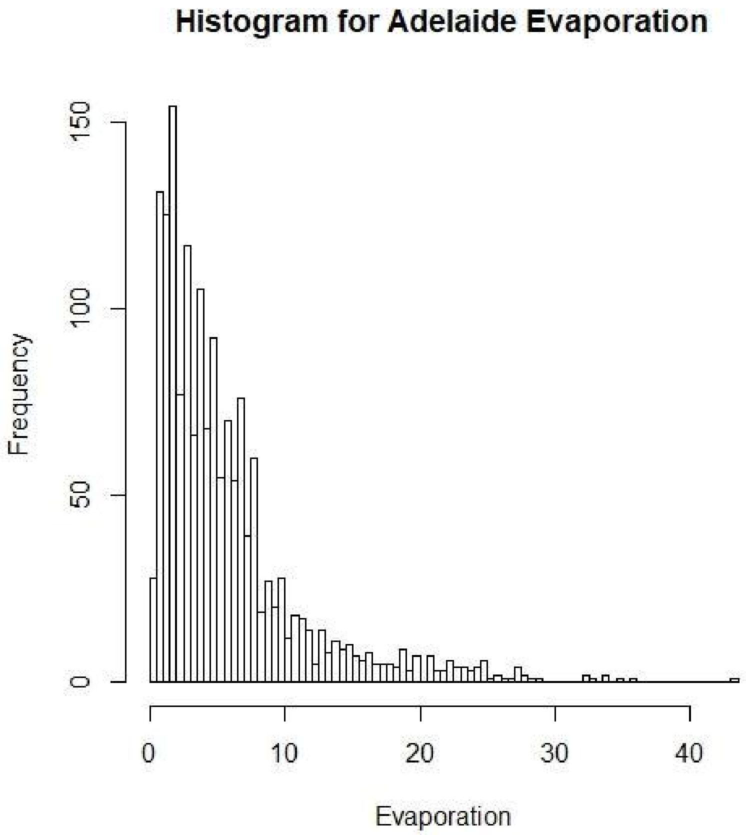


Figure 7. Evaporation in Adelaide

As result shown in Figure 7 is the distribution of the evaporation in Adelaide, the histogram does not follow normal distribution and severely right skewed. Majority of the evaporation in Adelaide are in the range from around 0 mm to nearly 30 mm in a day. Most days in Adelaide have the evaporation rates around 0 mm to 3 mm a day, which mean Adelaide has dry weather. However, there are some observations on the right of the histogram that are significant further from majority of the observations in the dataset that has more than 30 mm evaporate per day, this is unusual in Adelaide as it has dry weather. Hence, it is suggested for the observations that have evaporation greater than 30 mm to be potential outliers.

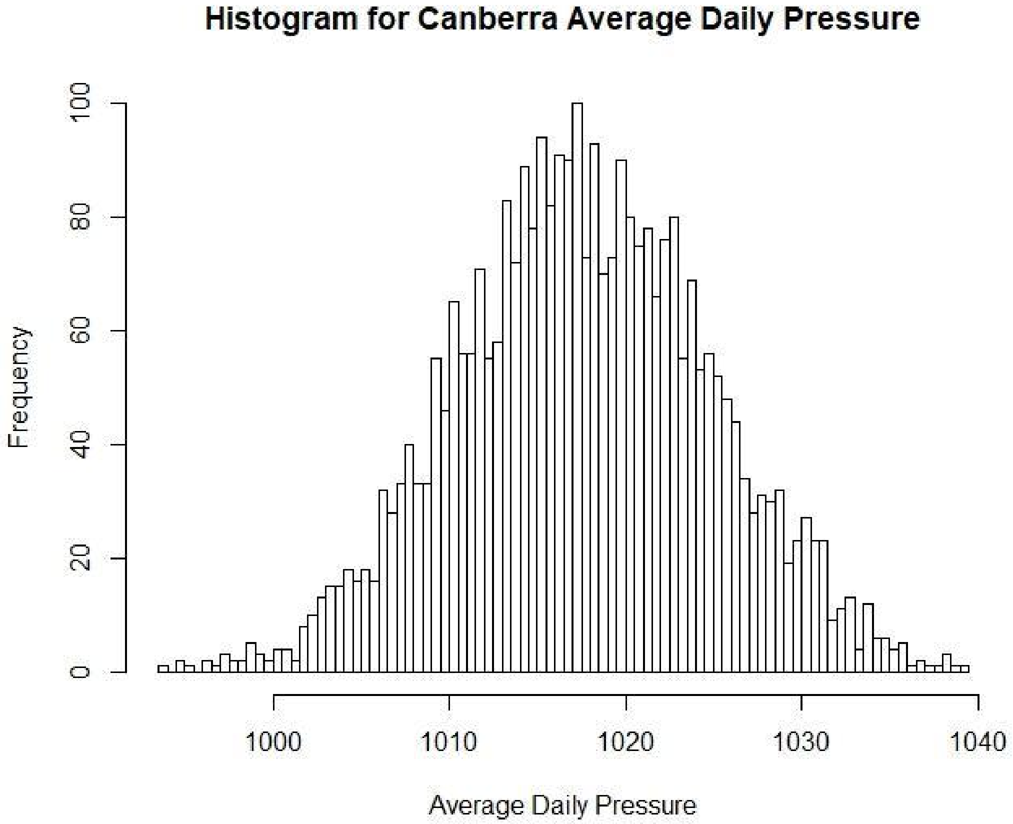


Figure 8. Average Daily Pressure in Canberra

Figure 8 shows the distribution of the average daily pressure in Canberra, the histogram seems to be following normal distribution. Most days, the air pressure in Canberra are in the range from around 1000 – 1040 hpa. However, there are some observations on the left of histogram that is significant further from majority of the observations in the dataset. It is suggested for the observation that has average daily pressure below 993hpa to be potential outliers.

Lastly, Figure 9 is the distribution of the evaporation in Canberra, majority of the evaporation are in the range from around 0 mm to nearly 15 mm in a day. Most days in Canberra have the evaporation rates around 0 mm to 5 mm a day, which mean Canberra has quite dry weather too, but it has more humidity compared with Adelaide weather. However, there are some observations on the right of the histogram at more than 15 mm a day, those observations are significant further from majority of the observations in the dataset. It is also unusual for Canberra where most of the time the evaporation rates are from 0 – 10 mm a day. It is suggested for the observations that have evaporation greater than 15 to be potential outliers.

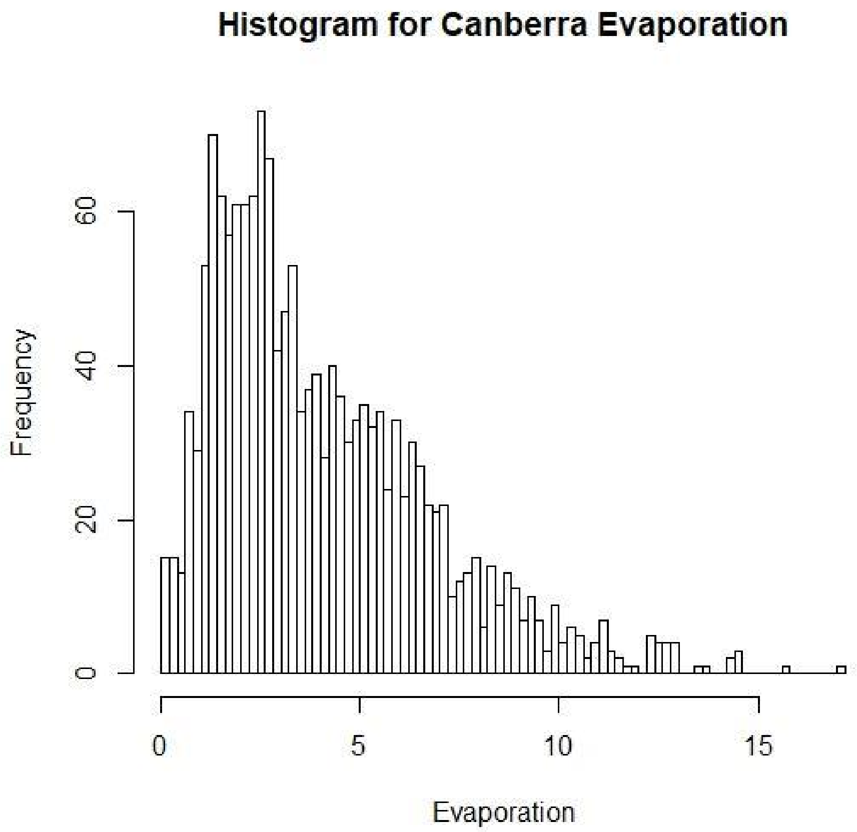


Figure 9. Evaporation in Canberra

3.3.4. Summary of outlier detection

Among all the performed methods to identify potential outliers in the dataset such as K – mean and DBScan for partition clustering and histogram. Histogram methods seem to be work best to identify potential outliers in this dataset, it could be the nature of the weather dataset is frequency observations which is better to visualize using histogram than other visualization methods. Therefore, the histogram is chosen to identify potential outliers for this dataset. Additionally, all attributes of the dataset were performed to visualize using histogram method, but only 2 attributes appeared to have potential outliers which are “Evaporation” and “DailyPressMean”. Hence, there are only histogram of these 2 attributes are selected to be shown in the report.

4. Model and Discussion

4.1. Logistic Regression

Logistic regression is an extension of a regression analysis where statistical model uses logistic function from binary dependent variable. Mathematically, that dependent variable in the model has two possible outcomes such as yes/no, pass/fail, 0/1 etc. on the other side, the independent variables in the model can be categorical variables or numeric variables.

The main task of logistic regression is to estimate the log odds (logarithm of odds) of an event which is called logistic unit or logit. Log odds are used for calculating probability of an event. In addition, logistic regression can be defined as:

$$Logit\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (1)$$

regression coefficients β are generated from maximum likelihood estimation.

To get optimal results from a logistic model some important considerations must be considered. The presence of outliers and highly correlated variables in the data might produce very poor estimation. On the other hand, overfitting is another important factor to be considered. Though adding an independent variable in the model will increase variance or R^2, more and more independent variables in the model might cause overfitting which decreases the generalizability of the logistic model. But there are some procedures to reduce overfitting such as stepwise regression. Stepwise regression is the process of fitting a logistic regression model where it eliminates candidate variables to get an optimal model. Three types of Stepwise regression are used to reduce overfitting such as forward selection, backward elimination, and bidirectional elimination. In forward selection, selection process starts with no variable and then best model produced by adding only most significant variables. Additionally, backward elimination starts with all variables and model produced by removing all insignificant variables. And Bidirectional elimination is a process of both forward selection and backward elimination.

4.1.1. Data Preprocessing

The first step of data pre-processing is to import libraries. As we have used R for our analysis, *InformationValue* library has been used for info matrix, *imputeTS* for filling missing values and *ggplot2* for visualization. Secondly, we have dealt with the categorical variables. As classifiers only recognize numbers, we have changed categorical into numbers such as for RainToday, RainTomorrow into 0,1 as well as WindDir3pm, WindDir9am and WindGustDir changed based on ranking (1 to 16). Thirdly, missing values have been filled with average values. Finally, for feature extraction, we have used correlation matrix to remove highly correlated values. In addition, we have created two extra variables such as DailyTempMean (daily mean temperature) and DailyPressMean (daily mean pressure).

4.1.2. Adelaide analysis

4.1.2.1. Model creation

The input variables for this model are selected by investigating the cross-correlation of the variables in the dataset. For this project, we have selected the variables according to 0.7 cut-off since the highly correlated variables have the same effect on target variable.

At first, we have divided the dataset into training and testing dataset where 75% data is used as training and 25% is used as testing data. Secondly, all selected variables have been applied to build the model using glm function in R. As we have already known that using all variables might cause overfitting, we have used a backward elimination process to get the most significant model. For Adelaide data, firstly we have used all the variables such as Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, RainToday, DailyTempMean, DailyPressMean. After that we have applied step function in R to find important variables. Backward elimination has created an optimal formula.

$$Logit(RainTomorrow) = 138.1 - 0.12Sunshine + 0.07WindGustSpeed - 0.08WindDir9am -$$

Finally, we have found that WindDir3pm is not statistically significant. So, after removing WindDir3pm we have found the best model for Adelaide data. According to logistic regression model (2) for Adelaide area, there are 2 variables WindGustSpeed and Humidity3pm that have positive effect to the variable RainTomorrow which mean the stronger wind speed and higher humidity, the higher chance of raining on the next day. In the other hand, variables like Sunshine, WindDir9am, WindSpeed3pm, DailyTempMean, DailyPressMean have negative effect on the variable RainTomorrow.

4.1.2.2. Performance for Adelaide Area

As results shown in Table 3, with the total of 773 observations in Adelaide area. There are 602 days observed as no rain, the model can correctly predict 567 days, and wrongly predict 35 days as rain. There are 171 days observed as rain, and the model can correctly predict 94 days, and wrongly predict 77 days as no rain.

Table 3. Confusion matrix with outliers

		Actual Observations	
		No Rain	Rain
Predicted	No Rain	**567 (94.19%)**	77 (45.03%)
	Rain	35 (5.81%)	**94 (54.97%)**

Total Percentage	100%	100%

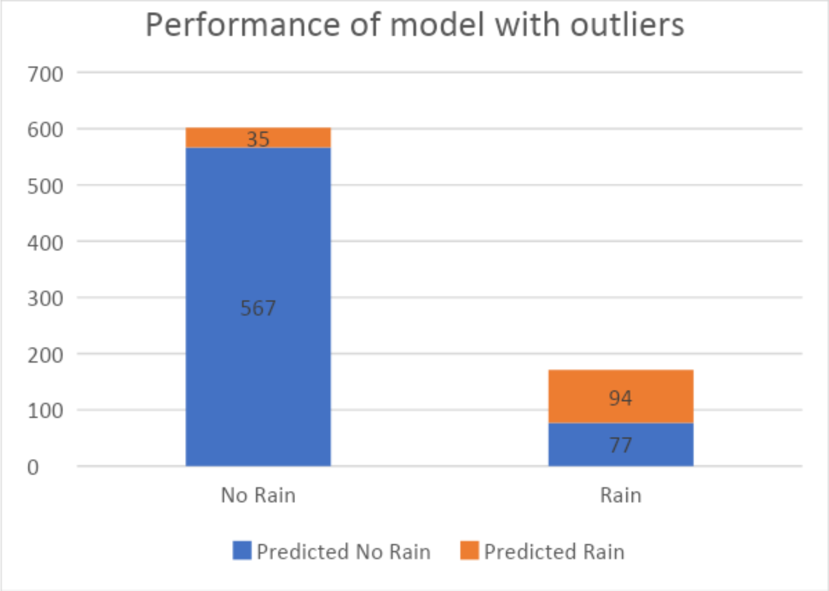


Figure 10. Visualization of Confusion Matrix for model with outliers

As shown in Table 4 the performance of model without outliers, similarly with total of 773 observations in Adelaide area. For 602 days observed as no rain, the model can correctly predict 570 days compare to 567 days for the model with outliers, and wrongly predict 32 days as rain. There are 171 days observed as rain, and the model can correctly predict 93 days compare to 94 days of for the model with outliers, and wrongly predict 78 days as no rain.

Overall, the model was built without outliers performed slightly better than the model with outliers in predicting there will be no rain in the next day with 3 more correct observations. However, the predicted observations of the next day will rain is 93 which is less than 1 observation compare to the model with outliers.

Table 4. Confusion matrix without outliers

	Actual Observations		
		No Rain	Rain
Predicted	No Rain	**570 (94.68%)**	78 (45.61%)
	Rain	32 (5.32%)	**93 (54.39%)**
Total Percentage		100%	100%

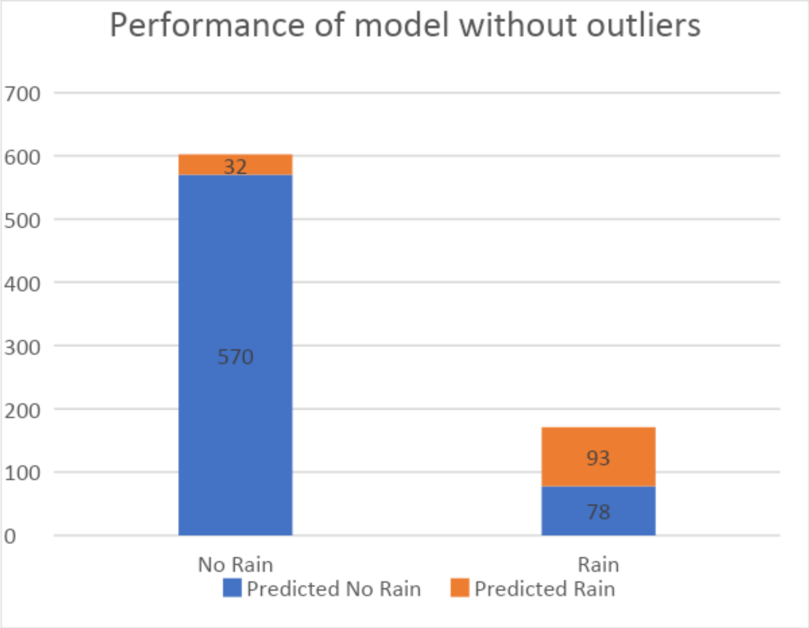


Figure 11. Visualization of Confusion Matrix for Model without outliers

As results shown in Table 5, we can see that the model was built without outliers performs slightly better than the model was built with outliers. The prediction accuracy is impressive at 85.77%, however, the model does not perform as good as the expectation as there is still 14.23% chance of predicting wrong about the rainfall in the next day, which is reasonably high chance of failure in predicting.

The reason for this could come from the perfection of data which is quite significantly large amount of missing data was found in the dataset. In fact, there are no data for 2 attributes “cloud9am” and “cloud3pm” in Adelaide Area. The missing of cloud data may could be the factor that affect the performance of the model as there is lack of data for analysis. As cloud is formed from the water vapor evaporates into the air then accumulate enough to become droplets, but the droplets are not heavy enough to fall back to the ground and it stay up in the sky to form cloud. Once the condensation of the droplets is long enough which makes the droplets of water stick together to create heavier and larger water drops, with the force from gravity, it then falls and become rain. Hence, cloud could be an important feature in the prediction model for Adelaide Area.

Table 5. Model performance comparison

	With Outliers	**Without Outliers**
Accuracy	85.51%	85.77%
Precision	88.04%	87.96%
Recall	94.19%	94.68%

F – score	91.01%	91.2%
Misclassification Error rate	14.49%	14.23%

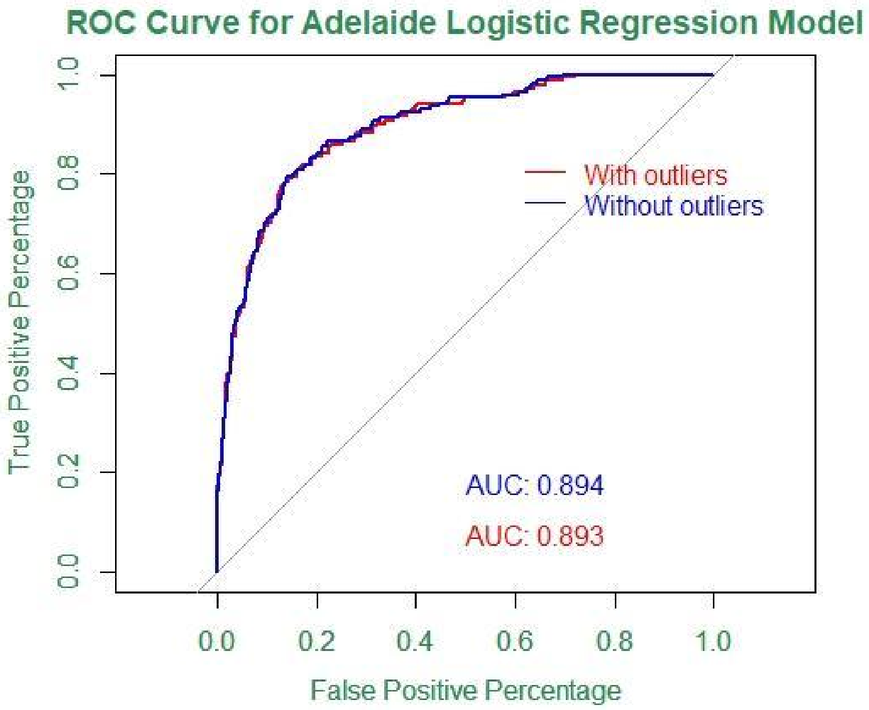


Figure 12. ROC Curve for Adelaide Model

Figure 12 shows the ROC curves of both logistic regression models were built with and without outliers for Adelaide area. We aim to build the model that has the curve closer to the upper left corner which also mean to maximize the area under the curve (AUC), this also explain how good the model performs. Since the larger AUC, the better model is, we can see that the model was built without outliers (AUC = 0.894) is slightly better than the model with outliers (AUC = 0.893). Additionally, 2 ROC curves are too close to each other which nearly make no different in performance, the model without outliers also has slightly better accuracy at 85.77%. Hence, we will use the model that is built without outliers.

4.1.2.3. Final Decision for Adelaide Model

From the performance analysis, we have found that the model without outliers performs better not only in the confusion matrix but also in the ROC curve. Therefore, we should choose model using data without outliers for Adelaide rainfall prediction.

4.1.3. Canberra analysis

4.1.3.1. Model Creation

For Canberra, we have divided the dataset as the same way we did for Adelaide data where 75% data is used as training and 25% is used as testing data. After that we have used all the variables such as Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, RainToday, DailyTempMean, DailyPressMean for Regression model. Next, using backward elimination process the final model has been generated which is given bellow.

$$Logit(RainTomorrow) = 73.79 - 0.09Sunshine + 0.06WindGustSpeed - 0.07WindSpeed9am$$

According to logistic regression model (3) for Canberra area, there are 2 variables WindGustSpeed, Humidity3pm, Cloud3pm and DailyTempMean that have positive effect to the variable RainTomorrow which mean the stronger wind speed, higher humidity with more cloud and higher in daily average temperature, the higher chance of raining on the next day. In the other hand, variables like Sunshine, WindSpeed9am, WindSpeed3pm, DailyTempMean, DailyPressMean have negative effect on the variable RainTomorrow.

Based on the model (2) and (3), it is suggested that Canberra and Adelaide have similar weather condition because most of the variables that contribute to the chance of raining in the next day are the same. The only different is model (3) have 2 additional variables that contribute significantly on the chance of raining are Cloud3pm and WindSpeed9am. Where model (2) have WindGustSpeed and no Cloud3pm variable.

4.1.3.2. Performance for Canberra Area

As results shown in Table 6, with the total 855 observations in Canberra area. There are 708 days observed as no rain, the model can correctly predict 688 days, and wrongly predict 20 days as rain. There are 147 days observed as rain, and the model can correctly predict 69 days, and wrongly predict 78 days as no rain.

Table 6. Confusion matrix with outliers

		Actual Observations	
Predicted		No Rain	Rain
	No Rain	**688 (97.18%)**	78 (53.06%)
	Rain	20 (2.82%)	**69 (46.94%)**
Total Percentage		100%	100%

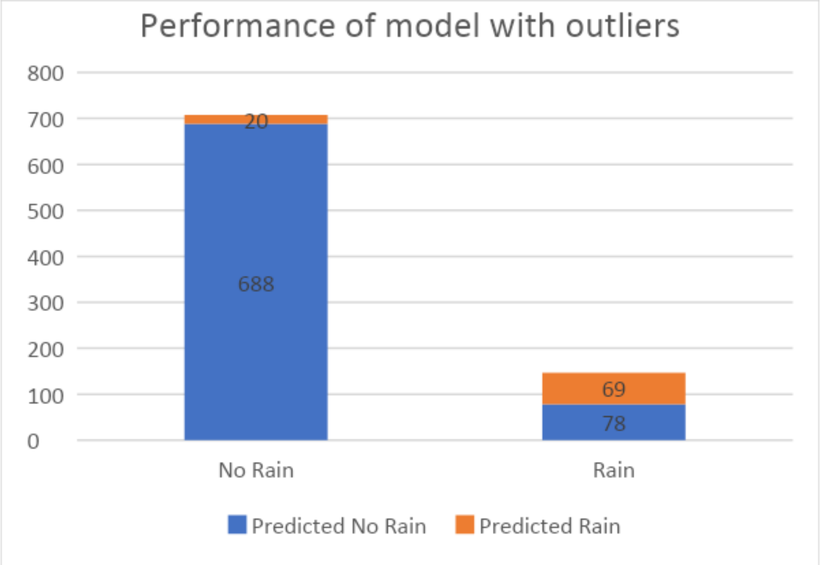


Figure 13. Visualization of Confusion Matrix for Model with outliers

As shown in Table 7 the performance of model without outliers, similarly with total of 855 observations in Canberra area. For 708 days observed as no rain, the model can correctly predict 690 days compare to 688 days for the model with outliers, and wrongly predict 18 days as rain. There are 147 days observed as rain, and the model can correctly predict 71 days compare to 69 days of for the model with outliers, and wrongly predict 76 days as no rain.

Overall, the model was built without outliers performed slightly better than the model with outliers in predicting there will be no rain in the next day with 2 more correct observations. Moreover, the predicted observations of the next day will rain also perform better which is predicted as 71 times, more than 2 times compared with the model with outliers.

Table 7. Confusion matrix without outliers:

		Actual Observations	
		No Rain	**Rain**
Predicted	**No Rain**	**690 (97.46%)**	76 (51.7%)
	Rain	18 (2.54%)	**71 (48.3)**
Total Percentage		100%	100%

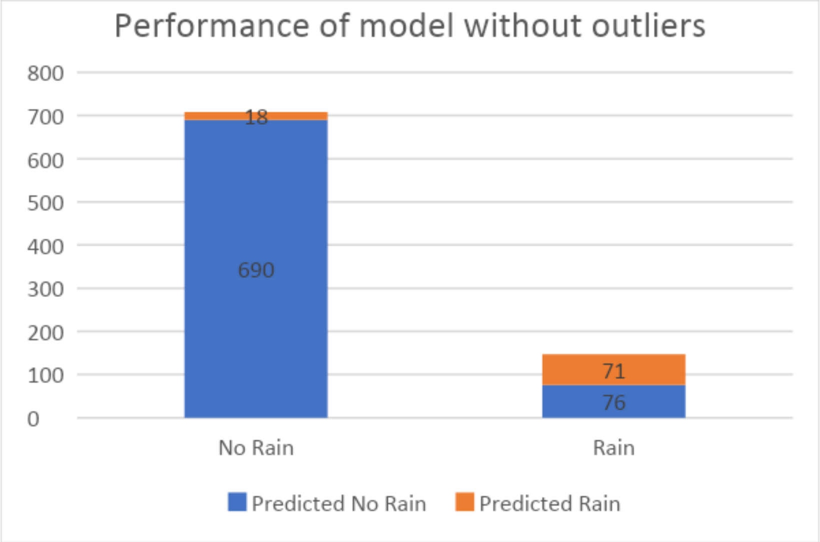


Figure 14. Visualization of Confusion Matrix for model without outliers

As results shown in Table 8, we can see that the model was built without outliers performs slightly better than the model was built with outliers. The prediction accuracy is impressive at 89.01%, however, the model barely perform as the expectation as there is 10.99% chance of predicting wrong about the rainfall in the next day, which is reasonably low chance of failure in predicting.

Table 8. Model performance comparison

	With Outliers	Without Outliers
Accuracy	88.54%	89.01%
Precision	89.82%	90.08%
Recall	97.18%	97.46%
F – score	93.35%	93.62%
Misclassification Error rate	11.46%	10.99%

As shown in Figure 15 are the ROC curves of both logistic regression models were built with and without outliers for Canberra area. We aim to build the model that has the curve closer to the upper left corner which also mean to maximize the area under the curve (AUC), this also explain how good the model performs. Since the larger AUC, the better model is, we can see that the model was built with outliers (AUC = 0.862) is slightly better than the model without outliers (AUC = 0.859). Additionally, 2 ROC curves are too close to each other which nearly make no different in performance, the model without outliers also has slightly better accuracy at 89.01%. Hence, we will use the model that is built without outliers.

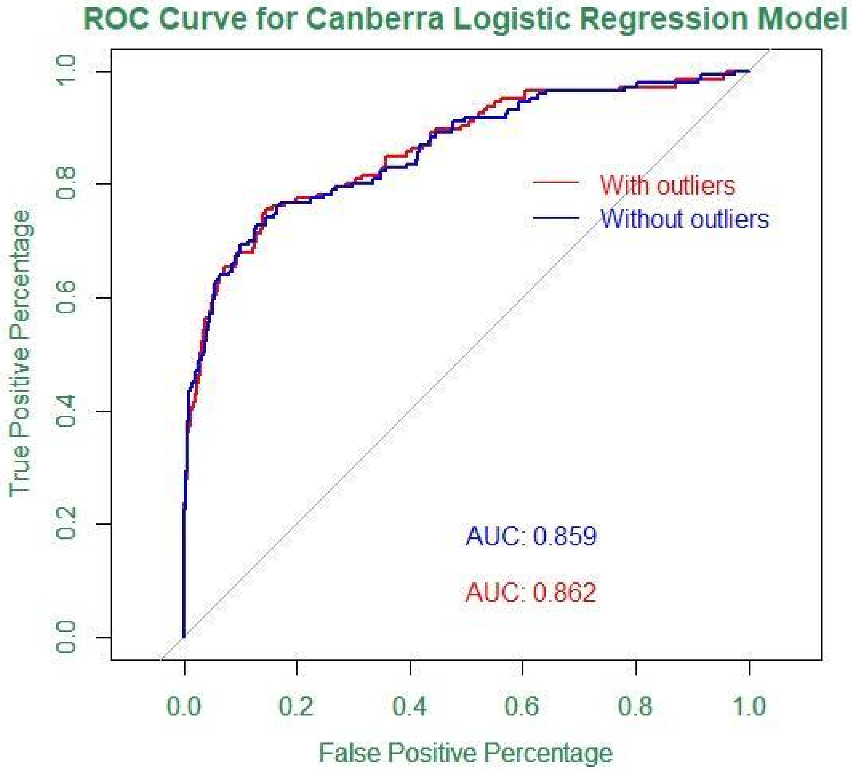


Figure 15. ROC Curve for Canberra Model

4.1.3.3. Final Decision for Canberra Model

From the performance analysis, we have found that the model without outliers has maximum accuracy whereas model with outliers has the better AUC rate. Form our understanding and the data nature we have found in Adelaide area; we have already known that data without outliers performs better. Besides, data without outliers for Canberra has reduced error rate which is better for our prediction analysis and the AUC rate of both models is very close. Therefore, we should choose model using data without outliers for Canberra area.

4.1.4. Summary of logistic regression

Overall, there are impressive performance for the logistic regression model to predict rain in Adelaide and Canberra are. However, each of the area will have its own type of weather characteristics, they will have different attributes to predict the rain in the logistic regression model. Additionally, as the number of rainy-day observations is significantly smaller than the number of observations for no rainy day for both Adelaide and Canberra, there are little training data for the rainy-day compare to no rainy-day. Hence, the model performed better when predicting no rainy-day and suggesting these 2 cities are quite similar in weather condition which is dry weather.

As the performance of the model without outliers is better for Adelaide area, it seems like the outliers do not contribute much into predicting power of the model. In contrast, for Canberra, the model with outlier is shown as the better option to predict rain in the next day. It may come different reasons, which is possible for the outliers present more for Canberra data than Adelaide.

In conclusion, logistic regression model barely meets the expectation of the team even though the misclassification rate is slightly greater than 10%. It could be the imperfection of data as there are a lot of missing data is reported in the previous section, or it could be the logistic regression method is not as powerful as other machine learning model for weather forecast problem.

4.2. Artificial Neural Network

4.2.1. Rationale

The input variables for this model are selected by investigating the cross-correlation of the variables in the dataset. For this project, we selected the variables according to 0.7 cut-off since the highly correlated variables have the same effect on target variable “RainTomorrow”. The variable selected for the ANN model included Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, RainToday, DailyTempMean, and DailyPressMean.

Caret package in R adopted feedforward ANN model. The number of input variables represented the number of input nodes. One hidden layer was used in the model and the number was determined using trial and error with decay rate between 0 and 0.5. This was to ensure that the model is smooth when a different dataset is introduced. K-fold cross-validation was used to evaluate the model performance (k was chosen to be 5) with ROC being used to compare the models(van Buuren & Groothuis-Oudshoorn 2011).

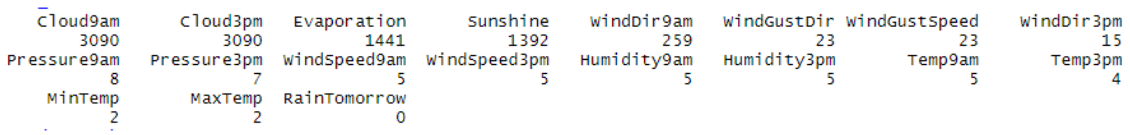
There is no specific formula for feedforward neural network, and thus we present the step function used in each neuron, where x is the input weight.

$$f(x) = \frac{1}{1+e^{-x}} \qquad (4)$$

This function is called logistic regression, and it is confirmed to be used in Caret. With this choice, the single-layer network is identical to the logistic regression model, widely used in statistical modeling. The logistic function is one of the family of functions called sigmoid functions because their S-shaped graphs resemble the final-letter lower case of the Greek letter Sigma.

4.2.2. Data Preprocessing For ANN

The first step we did, apart from data inspection, was data imputation. Although the original dataset contains around 140 thousand rows of data, the extracted volume for data in Adelaide is merely 3090 rows. As shown by the table below which summarizes the amount of missing values for each attribute in Adelaide dataset, all gauging on clouds was missing and almost half of evaporation and sunshine data was gone as well. This suggests that simply deleting rows containing missing values is not applicable in our case, as the performance of ANN model is proved to increase if more rows of data are fed into it.



```
     Cloud9am      Cloud3pm   Evaporation      Sunshine    WindDir9am   WindGustDir WindGustSpeed    WindDir3pm 
         3090          3090          1441          1392           259            23            23            15 
  Pressure9am   Pressure3pm  WindSpeed9am  WindSpeed3pm   Humidity9am   Humidity3pm       Temp9am       Temp3pm 
            8             7             5             5             5             5             5             4 
      MinTemp       MaxTemp  RainTomorrow 
            2             2             0 
```

Figure 16. Summary of missing values for Adelaide dataset

An alternative way to conduct the data imputation is replacing the missing values with other deduced meaningful measurements. The cloud attributes, of course, are not feasible to be used in the Adelaide dataset, as all of them are missing and there is no base for deduction. We adopted the method called Multivariate Imputation by Chained Equations (MICE) proposed by (van Buuren & Groothuis-Oudshoorn 2011). Unlike single variate imputation, replacing the missing values with means of that attribute, multivariate imputation takes all other variables into consideration and imputes data on a variable by variable basis by specifying an imputation model per variable. The imputed data contains no missing values and is ready to be partitioned.

The team then singled out outliers for attributes Pressure and Evaporation based on previous research. Two different datasets about Adelaide weather parameters were prepared, one with outliers and one without outliers. The team is interested in finding out the differences between models established from different datasets. For both datasets, 70% of them was extracted to form training set and 30% rest was the testing set. Both partitioned datasets were remained similar proportions of the target attribute, RainTomorrow, inherited from the original set. The output below proves this point.

	RainTomorrow = No	**RainTomorrow = Yes**
Original Set	0.777	0.223
Training Set	0.777	0.223

Table 9. Proportion of target class

Furthermore, the team adapted k-fold stratified cross validation process, also known as leave-one-out method, to establish validation sets. Cross-validation is a resampling procedure used to validate machine learning models on a limited data sample. It does not require us to separate the original dataset to another validation set, while instead, it repeatedly chooses one proportion of the dataset as validation set and calculate the performance until K times. Here the K equals 5, which means 5 iterations will be run and the training set will be validated 5 times.

4.2.3. ANN For Adelaide

This section describes the selection process and shape of produced ANN model established from Adelaide data, along with the performance evaluation for each model. There are totally 2 models to show, one with and one without outliers. The team employed the *nnet* package on R to build the model. Notice that the cutoff for correlated variables were done automatically in this approach. Our team has previously proved that 70% correlation is a significant threshold and any attributes with above 0.7 correlation coefficient shall be pruned to one attribute.

4.2.3.1. Model for Adelaide data with outliers

The output below shows the iterated ROC performances against decay and size of the final model. According to a passage written by (Venables & Ripley 2002) decay, specific to neural networks, uses as penalty the sum of squares of the weights Wij. In an essence, it is a regularization parameter to avoid over-fitting. Higher the decay is less overfitting a model would be. It illustrates how quickly it decreases in gradient descent. This is one way to ensure that the model is smooth when it is deployed with a totally different dataset. Another way is to restrict the class of estimates, for example, by using a limited number of neurons.

The size describes the number of nodes that will be used in the hidden layer. We also defined the maximum iterations to be carried out, which is 100. This is the reason why each time the results may vary.

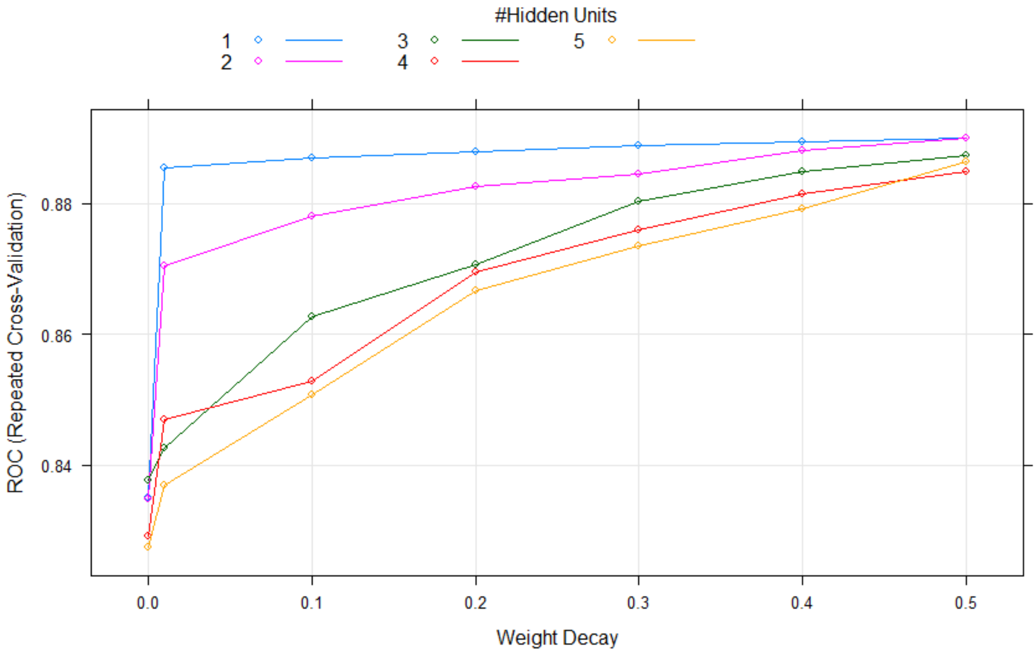


Figure 17. ROC comparisons for Adelaide ANN Model

The plot above suggests that the best model is the one with one hidden unit in one hidden layer which generally outperforms the rest models with more hidden units. The model was built from the Adelaide dataset without any outlier elimination process. It automatically scaled and centered the data and ignored three high correlated attributes. This is an essential step to employ the regularization parameter, as weight decay only makes sense if the inputs are rescaled to range about [0, 1] to be comparable with the outputs of internal units.

The shape and some important parameters are provided below. Notice that neural networks resemble black boxes a lot: explaining their outcome is much more difficult than explaining the outcome of simpler model such as a linear regression.

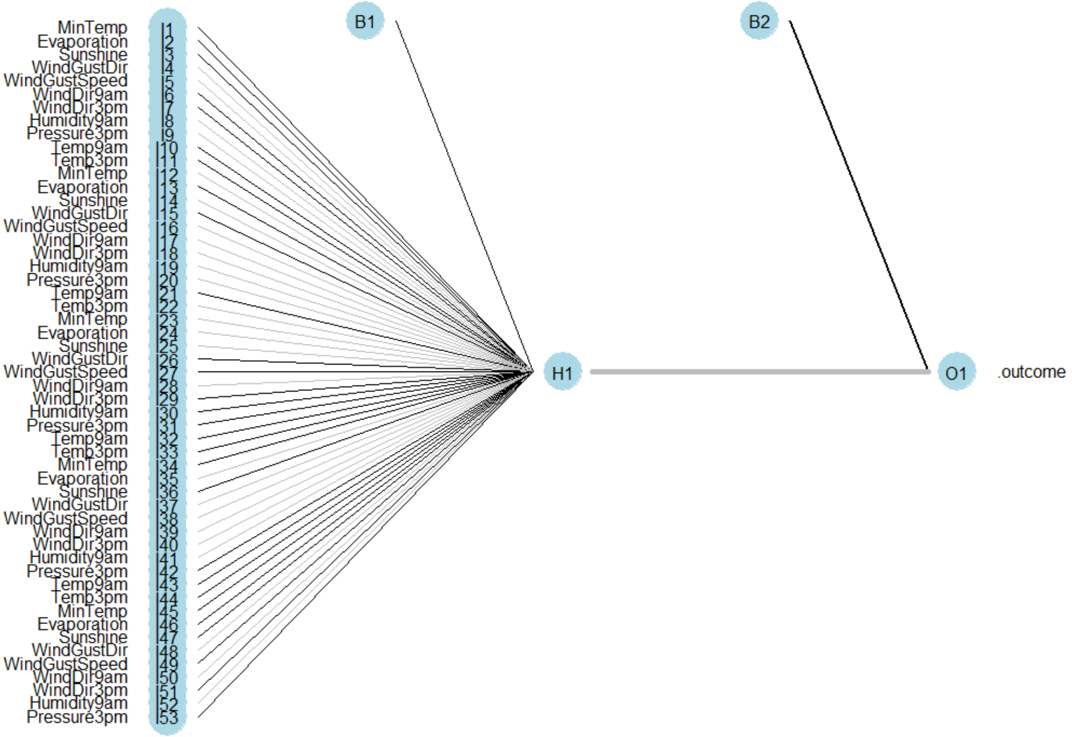
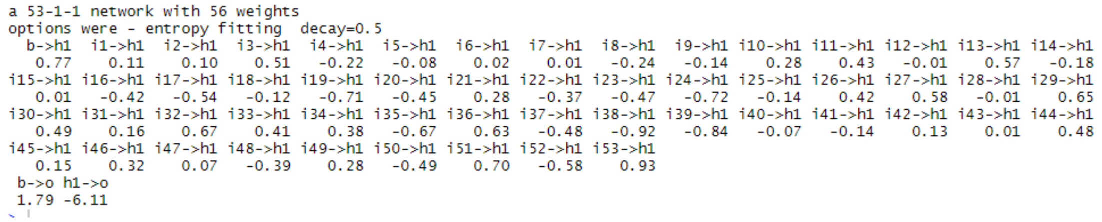


Figure 18. Final ANN model for Adelaide with outliers



```
a 53-1-1 network with 56 weights
options were - entropy fitting  decay=0.5
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1 i11->h1 i12->h1 i13->h1 i14->h1
   0.77    0.11    0.10    0.51   -0.22   -0.08    0.02    0.01   -0.24   -0.14    0.28    0.43   -0.01    0.57   -0.18
i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1
   0.01   -0.42   -0.54   -0.12   -0.71   -0.45    0.28   -0.37   -0.47   -0.72   -0.14    0.42    0.58   -0.01    0.65
i30->h1 i31->h1 i32->h1 i33->h1 i34->h1 i35->h1 i36->h1 i37->h1 i38->h1 i39->h1 i40->h1 i41->h1 i42->h1 i43->h1 i44->h1
   0.49    0.16    0.67    0.41    0.38   -0.67    0.63   -0.48   -0.92   -0.84   -0.07   -0.14    0.13    0.01    0.48
i45->h1 i46->h1 i47->h1 i48->h1 i49->h1 i50->h1 i51->h1 i52->h1 i53->h1
   0.15    0.32    0.07   -0.39    0.28   -0.49    0.70   -0.58    0.93
 b->o h1->o
 1.79 -6.11
```

Figure 19. Weights for model with outliers

Figure 18 shows the shapes of the model. There are 53 inputs where each input node represents the variables used, 1 hidden unit represented by H1 and one output unit represented by O1. The neurons shown by B1 and B2 above are biases that have a weight on each neuron, which are respectively 0.77 and 1.79 suggested by the table of weights above (second chart). The lines connecting each neuron are synapse which has its own weight. These weights, except the biases, are multiplied with corresponding input and added with the biases to feed forward to next layer. The weights, shown by Figure 19, are recursively updated by each row of the dataset and it finally gives us the final set of best weights.

4.2.3.2. Performance evaluation for Adelaide data with outliers

The team then tested the one-neuron model on the testing set and calculated following parameters. It is shown that the accuracy of the model

(84.77%) is better than the no information rate, proven by the p-value (5.364*10^-8 << 0.05). The p-value rejects the null hypothesis that the model is not better than no information rate. P-value is a one-sided test to see if the accuracy is better than the "no information rate”, which is taken to be the largest class percentage in the data. The No Information Rate is the best guess given no information beyond the overall distribution of the target classes. In this case, we know from the output below that most days (77.75%) did not rain. Therefore, our best guess with no other information is to pick the majority class and it will give us 77.75% accuracy, theoretically. In a word, it is easy to predict negative cases but difficult to predict positive cases (RainTomorrow = No). Hence, the sensitivity, also known as the true positive rate, is more important, as it shows how good a model is in verifying correct positive cases. However, the results below show a relatively low specificity, 61.17%, merely greater than random guess. This suggests us to explore more models regarding this dataset.

		Actual Observations	
Predicted		**No Rain**	**Rain**
	No Rain	**659**	80
	Rain	61	**126**

Table 10. Confusion matrix for model with outliers

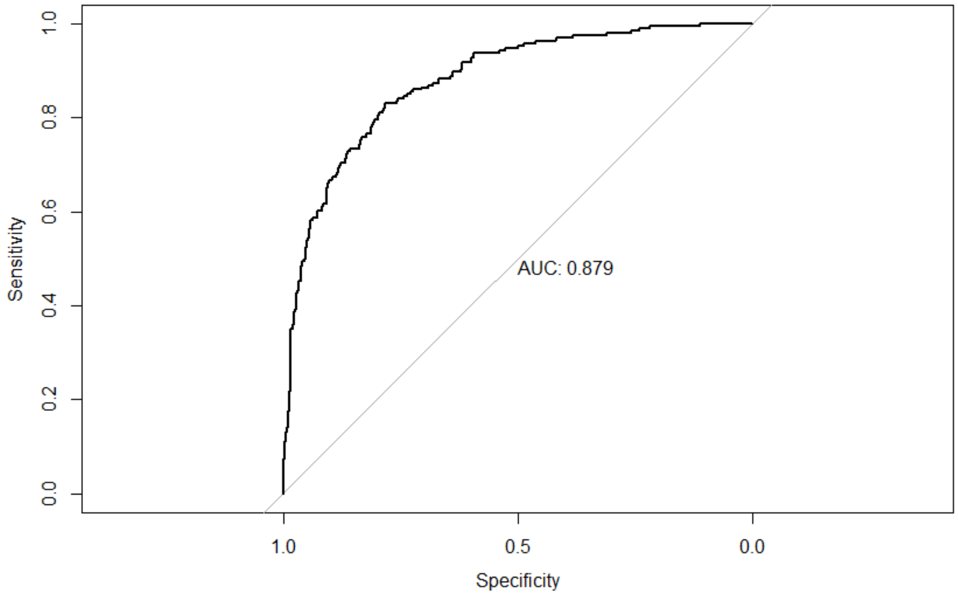


Figure 20. ROC for ANN model for Adelaide with outliers

Overall, the model is promising and satisfying. It gives a consistent accuracy in predicting on training (89.05%) and testing set (84.77%). There is no evidence showing that the model is overfitting. However, the model is less capable of predicting correct positive cases.

4.2.3.3. Model for Adelaide data without outliers

The next model, followed by the same pattern, is for the Adelaide dataset without outliers. The thresholds of the outliers were provided before in the logistic regression analysis. Based on that research, the team suspects that the model built from no-outlier version shall outstrip the one with outliers. The line chart below shows the comparison among models with different size of neurons and decays. The best goes to the one with one hidden unit and 0.5 decay.

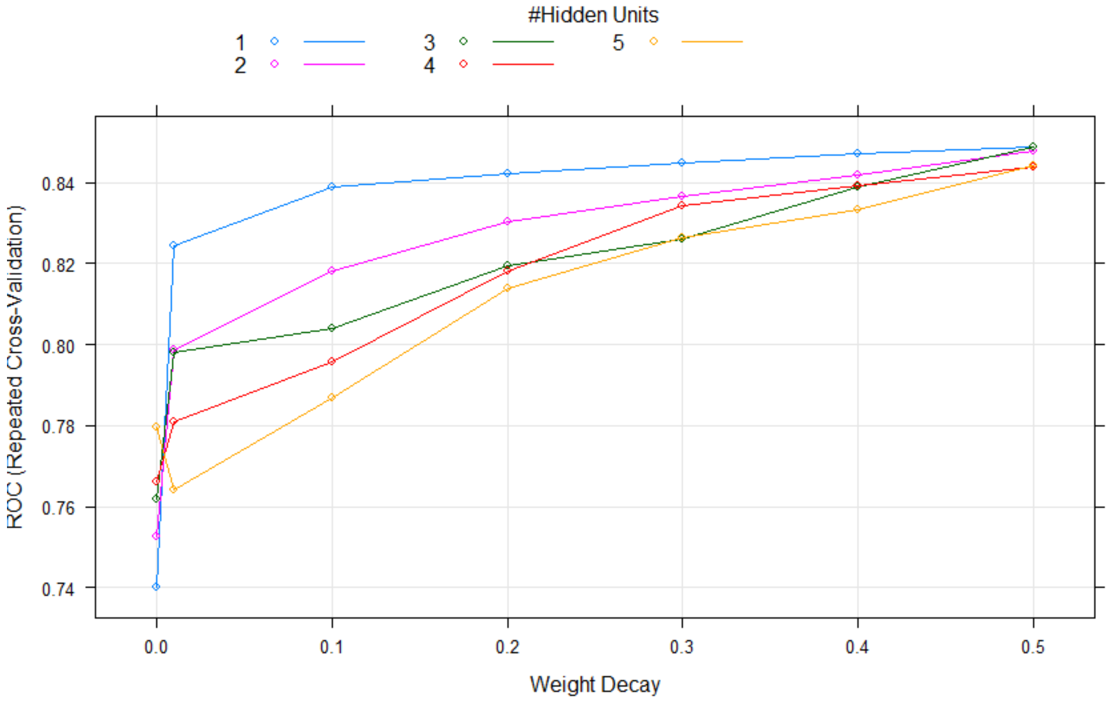


Figure 21. ROC selections for Model without outliers

The results of the model are shown below.

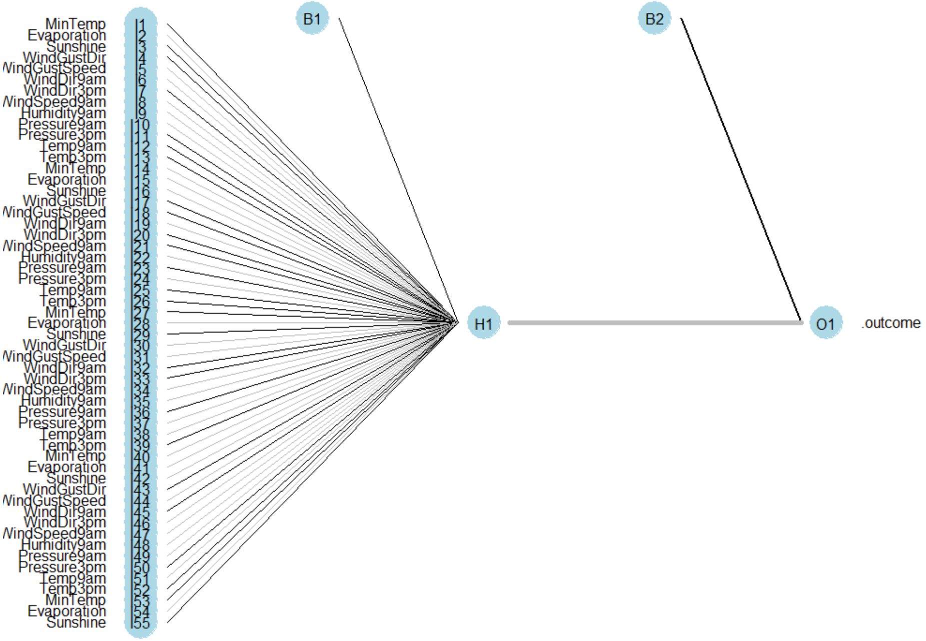
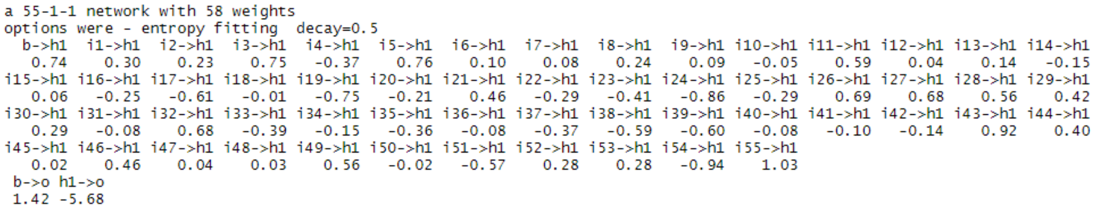


Figure 22. Model without outliers



```
a 55-1-1 network with 58 weights
options were - entropy fitting  decay=0.5
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1 i11->h1 i12->h1 i13->h1 i14->h1 
   0.74    0.30    0.23    0.75   -0.37    0.76    0.10    0.08    0.24    0.09   -0.05    0.59    0.04    0.14   -0.15 
i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 
   0.06   -0.25   -0.61   -0.01   -0.75   -0.21    0.46   -0.29   -0.41   -0.86   -0.29    0.69    0.68    0.56    0.42 
i30->h1 i31->h1 i32->h1 i33->h1 i34->h1 i35->h1 i36->h1 i37->h1 i38->h1 i39->h1 i40->h1 i41->h1 i42->h1 i43->h1 i44->h1 
   0.29   -0.08    0.68   -0.39   -0.15   -0.36   -0.08   -0.37   -0.59   -0.60   -0.08   -0.10   -0.14    0.92    0.40 
i45->h1 i46->h1 i47->h1 i48->h1 i49->h1 i50->h1 i51->h1 i52->h1 i53->h1 i54->h1 i55->h1 
   0.02    0.46    0.04    0.03    0.56   -0.02   -0.57    0.28    0.28   -0.94    1.03 
 b->o h1->o 
 1.42 -5.68 
```

Figure 23. Weights of model without outliers

The shape of this one is much similar with the one with outliers, except that there are two more input neurons in this one, Sunshine and Evaporation. It is normal that these variables were repetitively used as the *nnet* process need to continuously adjust the weight by adding new inputs when some weights have reached convergence. There are no other obvious differences apart from that.

4.2.3.4. Performance evaluation for Adelaide data without outliers

Next, the team tested the model built from data without outliers. It is expected there will be a slight improvement on the accuracy and specificity as not much noises were deleted. But surprisingly, from the results below we found that there is a great improvement on the specificity (68.18%) which matters most, although there is just a slight improvement on the overall accuracy (86.09%).

	Actual Observations		
Predicted		No Rain	Rain

	No Rain	**346**	35
	Rain	33	**75**

Table 11. Confusion matrix for model without outliers

4.2.3.5. Comparison of Adelaide Models

Overall, both models are promising and satisfying. It gives a consistent accuracy in predicting on training and testing set. Therefore, there is no evidence showing that the model is overfitting. The diagram below shows that ROC comparisons of these two models. AUC means the area under curves. The area measures discrimination, that is, the ability of the test to correctly classify whether it will rain tomorrow.

The ROC shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The comparison below shows that both gives promising accuracy but the one without outliers is slightly better. This is not only because the area is larger, but also it gives a much better specificity when holding a good sensitivity when specificity is about 0.75.

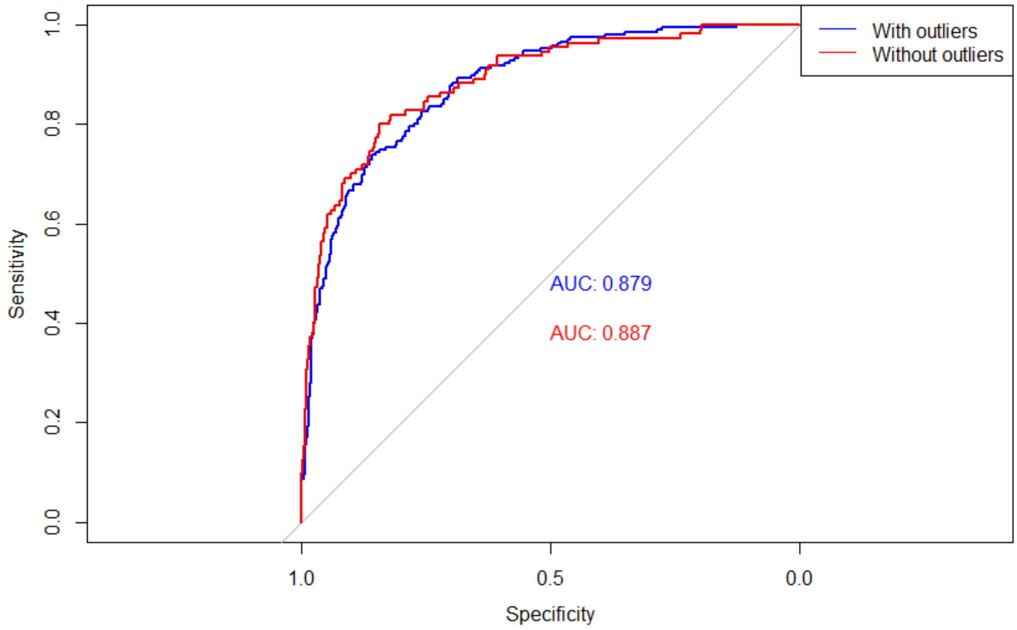


Figure 24. ROC comparison of Adelaide models

4.2.4. ANN for Canberra

In this section, two ANN models were created to predict the possibility of raining the following day; one without removal of outliers and the other with the outliers removed.

4.2.4.1. Model for Canberra data with outliers

Here, the model was created without elimination of outliers from Canberra dataset. Three of the highly correlated variables were ignored in the modelling of the ANN model. After 5-fold cross-validation was employed with decay of 0 to 0.5 and maximum iterations set to 100, the best model obtained based on the area under the ROC plot below.

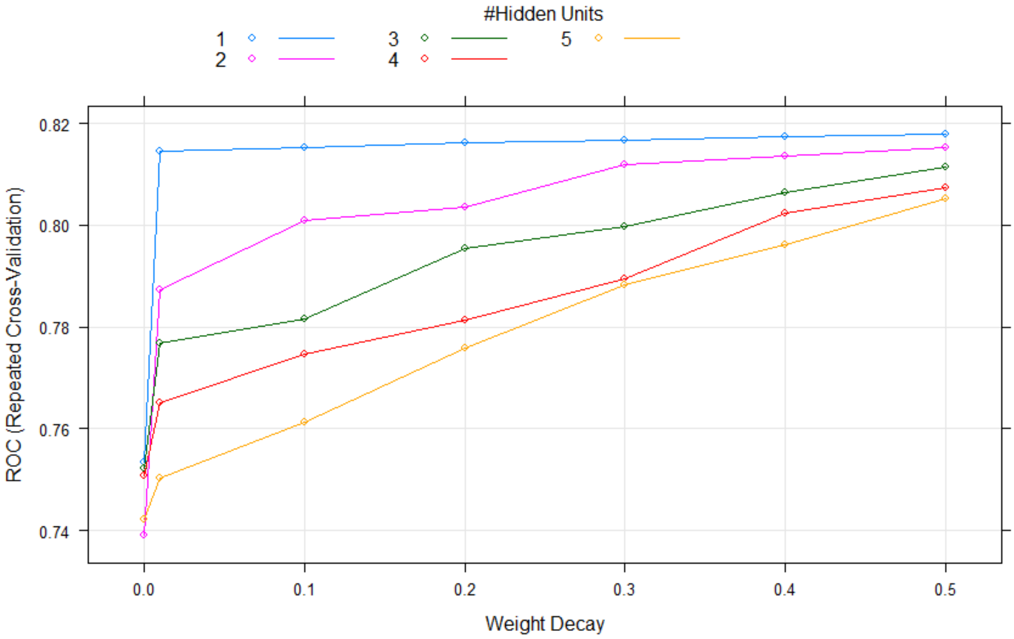


Figure 25. ROC selection for Canberra with outliers

The resulting model, shown below, had 54 input nodes representing the variables used, 1 hidden layer H1, and output node O1.

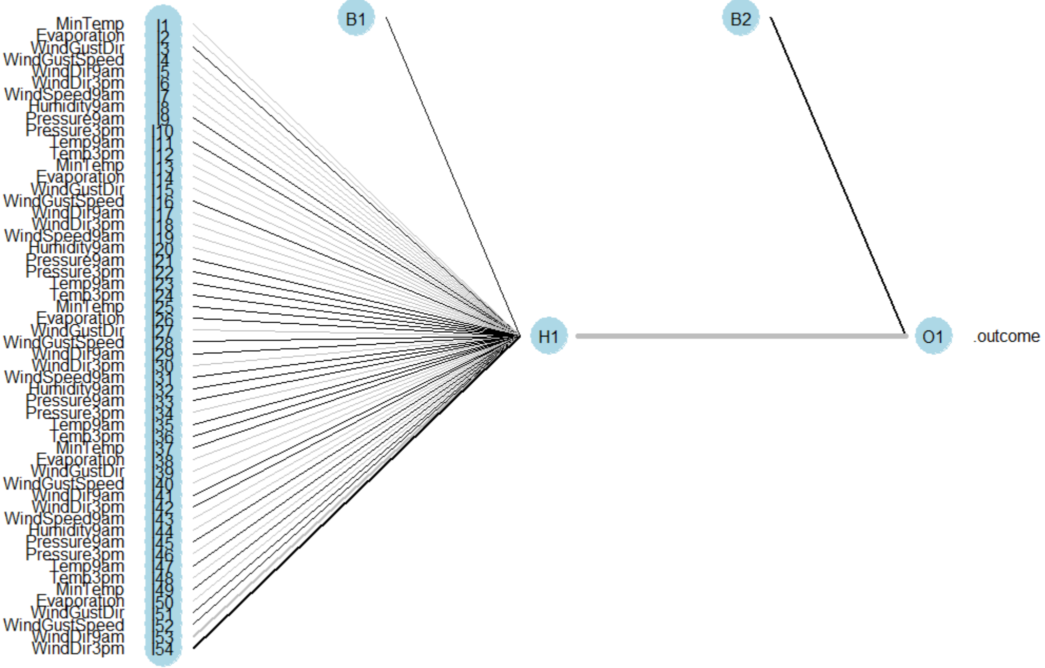
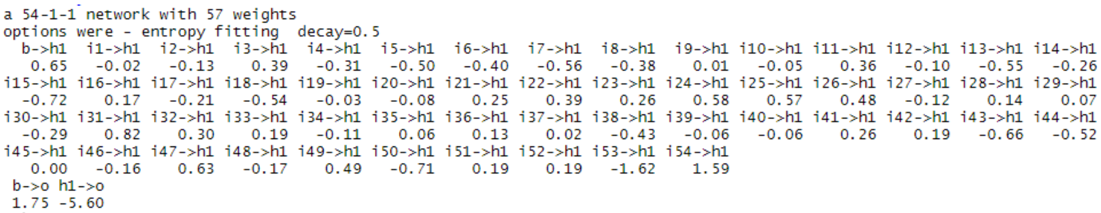


Figure 26. Model for Canberra with outliers

The weights (i1 to i54) for each line connecting the neurons and biases B1 and B2 are also shown below. These weights, except the biases, are multiplied with corresponding input and added with the biases to feed forward to next layer. The weights are recursively updated by each row of the dataset and it finally gives us the final set of best weights.



```
a 54-1-1 network with 57 weights
options were - entropy fitting  decay=0.5
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1 i11->h1 i12->h1 i13->h1 i14->h1 
   0.65   -0.02   -0.13    0.39   -0.31   -0.50   -0.40   -0.56   -0.38    0.01   -0.05    0.36   -0.10   -0.55   -0.26 
i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 
  -0.72    0.17   -0.21   -0.54   -0.03   -0.08    0.25    0.39    0.26    0.58    0.57    0.48   -0.12    0.14    0.07 
i30->h1 i31->h1 i32->h1 i33->h1 i34->h1 i35->h1 i36->h1 i37->h1 i38->h1 i39->h1 i40->h1 i41->h1 i42->h1 i43->h1 i44->h1 
  -0.29    0.82    0.30    0.19   -0.11    0.06    0.13    0.02   -0.43   -0.06   -0.06    0.26    0.19   -0.66   -0.52 
i45->h1 i46->h1 i47->h1 i48->h1 i49->h1 i50->h1 i51->h1 i52->h1 i53->h1 i54->h1 
   0.00   -0.16    0.63   -0.17    0.49   -0.71    0.19    0.19   -1.62    1.59 
 b->o h1->o 
 1.75 -5.60 
```

Figure 27. Weights for Canberra with outliers

4.2.4.2. Performance evaluation for Canberra data with outliers

The team tested the ANN model on the testing set and the following performance indicators as shown in the output below were obtained. The accuracy of the model is 84.86% with a 95% confidence of 82.52% and 87%. This would be a very good predictive model if only the sample was an equal split between the “yes” and “No” for the RainTomorrow variable. In our dataset there were 81% “No” and 19% “Yes”. In this case, we must investigate the “No information rate”. The ‘No information rate’ 81.64%

indicate that accuracy of predicting the majority class (“No Rain”) label. In this case if asked to predict whether it will rain or not, by choosing “No rain” only we can achieve a 81.64% accuracy on the test data. The model doesn’t look so good, but its accuracy is enough to give it a better performance over the “no information rate” as indicated by the p-value (0.0037) <0.05. The 0.81 AUC for the model, as shown in the ROC curve below, indicates that the model has a quite higher power of rainfall prediction than a random guess.

The sensitivity (proportion of predicting RainTomorrow = No correctly) for the model was 96.17% which was quite higher than the Specificity (proportion of predicting RainTomorrow = Yes correctly) of 34.57%, which is really poor. This means that the model has a chance of predicting the RainTomorrow = No rain class label than RainTomorrow = Yes. The model hardly has power in predicting affirmative rainfall, which reflect the weakness of ANN model in Canberra dataset.

	Actual Observations		
		No Rain	Rain
Predicted	No Rain	**804**	123
	Rain	32	**65**

Table 12. Confusion matrix for Canberra model with outliers

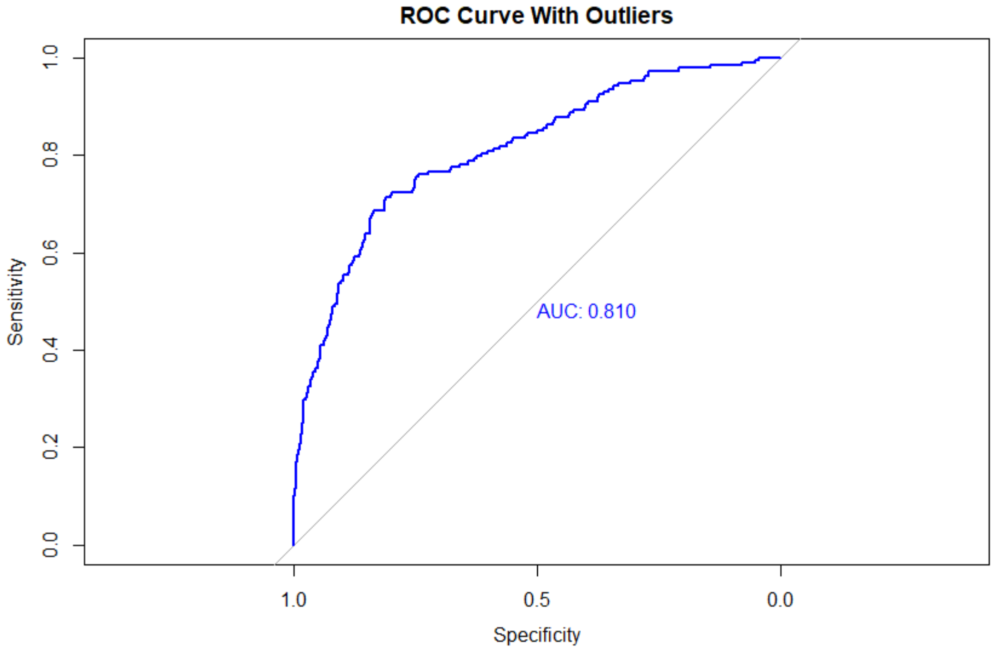


Figure 28. ROC for Canberra model with outliers

4.2.4.3. Model for Canberra data without outliers

Here, the model was created after elimination of outliers from Canberra dataset. After 5-fold cross-validation was employed with decay of 0 to 0.5

and maximum iterations set to 100, the best model obtained based on the area under the ROC plot below, Figure 29.

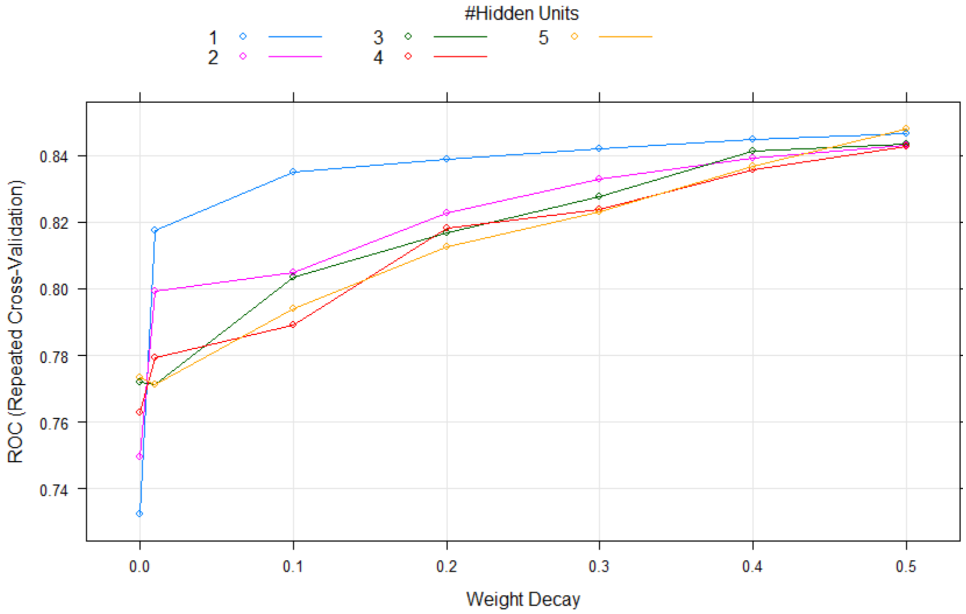


Figure 29. ROC selections for model without outliers

The best model produced had the 55 input nodes representing the input variables, 5 hidden layers (H1, H2..., H5) and 1 output node. The weights of the lines joining the Input nodes to Hidden layers, hidden layers to output nodes and the biases are also shown below, Figure 30.

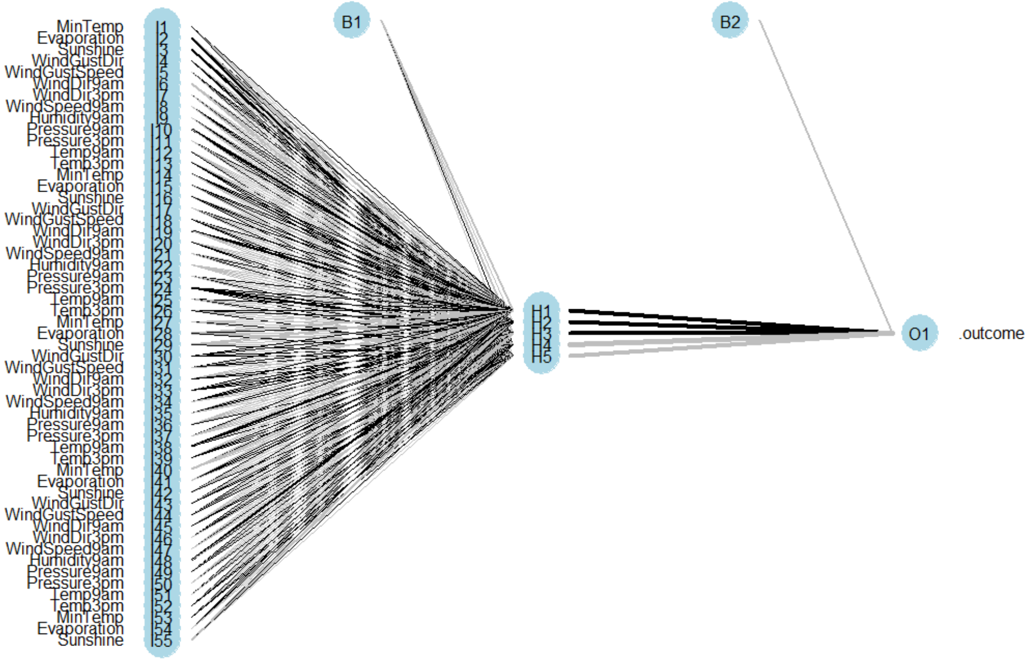
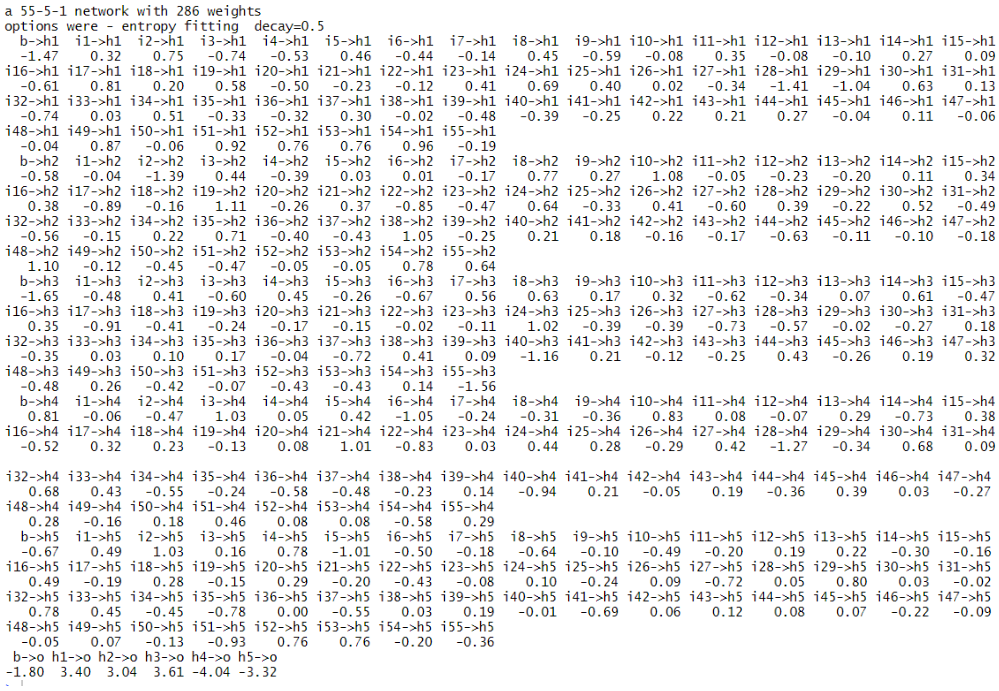


Figure 30. Model for Canberra without outliers



```
a 55-5-1 network with 286 weights
options were - entropy fitting  decay=0.5
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1 i11->h1 i12->h1 i13->h1 i14->h1 i15->h1 
  -1.47    0.32    0.75   -0.74   -0.53    0.46   -0.44   -0.14    0.45   -0.59   -0.08    0.35   -0.08   -0.10    0.27    0.09 
i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1 i30->h1 i31->h1 
  -0.61    0.81    0.20    0.58   -0.50   -0.23   -0.12    0.41    0.69    0.40    0.02   -0.34   -1.41   -1.04    0.63    0.13 
i32->h1 i33->h1 i34->h1 i35->h1 i36->h1 i37->h1 i38->h1 i39->h1 i40->h1 i41->h1 i42->h1 i43->h1 i44->h1 i45->h1 i46->h1 i47->h1 
  -0.74    0.03    0.51   -0.33   -0.32    0.30   -0.02   -0.48   -0.39   -0.25    0.22    0.21    0.27   -0.04    0.11   -0.06 
i48->h1 i49->h1 i50->h1 i51->h1 i52->h1 i53->h1 i54->h1 i55->h1 
  -0.04    0.87   -0.06    0.92    0.76    0.76    0.96   -0.19 
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2  i9->h2 i10->h2 i11->h2 i12->h2 i13->h2 i14->h2 i15->h2 
  -0.58   -0.04   -1.39    0.44   -0.39    0.03    0.01   -0.17    0.77    0.27    1.08   -0.05   -0.23   -0.20    0.11    0.34 
i16->h2 i17->h2 i18->h2 i19->h2 i20->h2 i21->h2 i22->h2 i23->h2 i24->h2 i25->h2 i26->h2 i27->h2 i28->h2 i29->h2 i30->h2 i31->h2 
   0.38   -0.89   -0.16    1.11   -0.26    0.37   -0.85   -0.47    0.64   -0.33    0.41   -0.60    0.39   -0.22    0.52   -0.49 
i32->h2 i33->h2 i34->h2 i35->h2 i36->h2 i37->h2 i38->h2 i39->h2 i40->h2 i41->h2 i42->h2 i43->h2 i44->h2 i45->h2 i46->h2 i47->h2 
  -0.56   -0.15    0.22    0.71   -0.40   -0.43    1.05   -0.25    0.21    0.18   -0.16   -0.17   -0.63   -0.11   -0.10   -0.18 
i48->h2 i49->h2 i50->h2 i51->h2 i52->h2 i53->h2 i54->h2 i55->h2 
   1.10   -0.12   -0.45   -0.47   -0.05   -0.05    0.78    0.64 
  b->h3  i1->h3  i2->h3  i3->h3  i4->h3  i5->h3  i6->h3  i7->h3  i8->h3  i9->h3 i10->h3 i11->h3 i12->h3 i13->h3 i14->h3 i15->h3 
  -1.65   -0.48    0.41   -0.60    0.45   -0.26   -0.67    0.56    0.63    0.17    0.32   -0.62   -0.34    0.07    0.61   -0.47 
i16->h3 i17->h3 i18->h3 i19->h3 i20->h3 i21->h3 i22->h3 i23->h3 i24->h3 i25->h3 i26->h3 i27->h3 i28->h3 i29->h3 i30->h3 i31->h3 
   0.35   -0.91   -0.41   -0.24   -0.17   -0.15   -0.02   -0.11    1.02   -0.39   -0.39   -0.73   -0.57   -0.02   -0.27    0.18 
i32->h3 i33->h3 i34->h3 i35->h3 i36->h3 i37->h3 i38->h3 i39->h3 i40->h3 i41->h3 i42->h3 i43->h3 i44->h3 i45->h3 i46->h3 i47->h3 
  -0.35    0.03    0.10    0.17   -0.04   -0.72    0.41    0.09   -1.16    0.21   -0.12   -0.25    0.43   -0.26    0.19    0.32 
i48->h3 i49->h3 i50->h3 i51->h3 i52->h3 i53->h3 i54->h3 i55->h3 
  -0.48    0.26   -0.42   -0.07   -0.43   -0.43    0.14   -1.56 
  b->h4  i1->h4  i2->h4  i3->h4  i4->h4  i5->h4  i6->h4  i7->h4  i8->h4  i9->h4 i10->h4 i11->h4 i12->h4 i13->h4 i14->h4 i15->h4 
   0.81   -0.06   -0.47    1.03    0.05    0.42   -1.05   -0.24   -0.31   -0.36    0.83    0.08   -0.07    0.29   -0.73    0.38 
i16->h4 i17->h4 i18->h4 i19->h4 i20->h4 i21->h4 i22->h4 i23->h4 i24->h4 i25->h4 i26->h4 i27->h4 i28->h4 i29->h4 i30->h4 i31->h4 
  -0.52    0.32    0.23   -0.13    0.08    1.01   -0.83    0.03    0.44    0.28   -0.29    0.42   -1.27   -0.34    0.68    0.09 

i32->h4 i33->h4 i34->h4 i35->h4 i36->h4 i37->h4 i38->h4 i39->h4 i40->h4 i41->h4 i42->h4 i43->h4 i44->h4 i45->h4 i46->h4 i47->h4 
   0.68    0.43   -0.55   -0.24   -0.58   -0.48   -0.23    0.14   -0.94    0.21   -0.05    0.19   -0.36    0.39    0.03   -0.27 
i48->h4 i49->h4 i50->h4 i51->h4 i52->h4 i53->h4 i54->h4 i55->h4 
   0.28   -0.16    0.18    0.46    0.08    0.08   -0.58    0.29 
  b->h5  i1->h5  i2->h5  i3->h5  i4->h5  i5->h5  i6->h5  i7->h5  i8->h5  i9->h5 i10->h5 i11->h5 i12->h5 i13->h5 i14->h5 i15->h5 
  -0.67    0.49    1.03    0.16    0.78   -1.01   -0.50   -0.18   -0.64   -0.10   -0.49   -0.20    0.19    0.22   -0.30   -0.16 
i16->h5 i17->h5 i18->h5 i19->h5 i20->h5 i21->h5 i22->h5 i23->h5 i24->h5 i25->h5 i26->h5 i27->h5 i28->h5 i29->h5 i30->h5 i31->h5 
   0.49   -0.19    0.28   -0.15    0.29   -0.20   -0.43   -0.08    0.10   -0.24    0.09   -0.72    0.05    0.80    0.03   -0.02 
i32->h5 i33->h5 i34->h5 i35->h5 i36->h5 i37->h5 i38->h5 i39->h5 i40->h5 i41->h5 i42->h5 i43->h5 i44->h5 i45->h5 i46->h5 i47->h5 
   0.78    0.45   -0.45   -0.78    0.00   -0.55    0.03    0.19   -0.01   -0.69    0.06    0.12    0.08    0.07   -0.22   -0.09 
i48->h5 i49->h5 i50->h5 i51->h5 i52->h5 i53->h5 i54->h5 i55->h5 
  -0.05    0.07   -0.13   -0.93    0.76    0.76   -0.20   -0.36 
 b->o h1->o h2->o h3->o h4->o h5->o 
-1.80  3.40  3.04  3.61 -4.04 -3.32 
```

Figure 31. Weights for model without outliers

4.2.4.4. Performance evaluation for Canberra data without outliers

The team tested the ANN model on the testing set and the following performance indicators as shown in the output below were obtained. The accuracy of the model is 84.31% with a 95% confidence of 80.73% and 87.45%. Assuming an equal split between “Yes” and “No” Rain Tomorrow in the sample, this would be a very good predictive model. However, in our dataset there were 81% “No” and 19% “Yes”. In this case, we must investigate the “No information rate”. The ‘No information rate’ indicates that if asked to predict whether it will rain or not, by choosing “No rain” only, we can achieve an 82.22% accuracy on the test data. The model does not look so good as its accuracy is not enough to give it a better performance over the “no information rate” as indicated by the p-value (0.127) > 0.05. The 0.848 AUC for the model, as shown in the ROC curve below, indicates that the model has a quite higher power distinguishing between the “Yes” and “No” class as compared to random guess.

The sensitivity (proportion of predicting RainTomorrow = No correctly) for the model was 94.91% which was quite higher than the Specificity (proportion of predicting RainTomorrow = Yes correctly) of 35.29%. Again, this proves that the weakness of ANN model in predicting affirmative result of rainfall for Canberra.

		Actual Observations	
Predicted		**No Rain**	**Rain**
	No Rain	**373**	55
	Rain	20	**30**

Table 13. Confusion matrix for model without outliers

The low specificity is also reflected on the ROC curve below, where the curve has a leaning-upward tendency. This means that the model is capable of satisfying high sensitivity but not specificity.

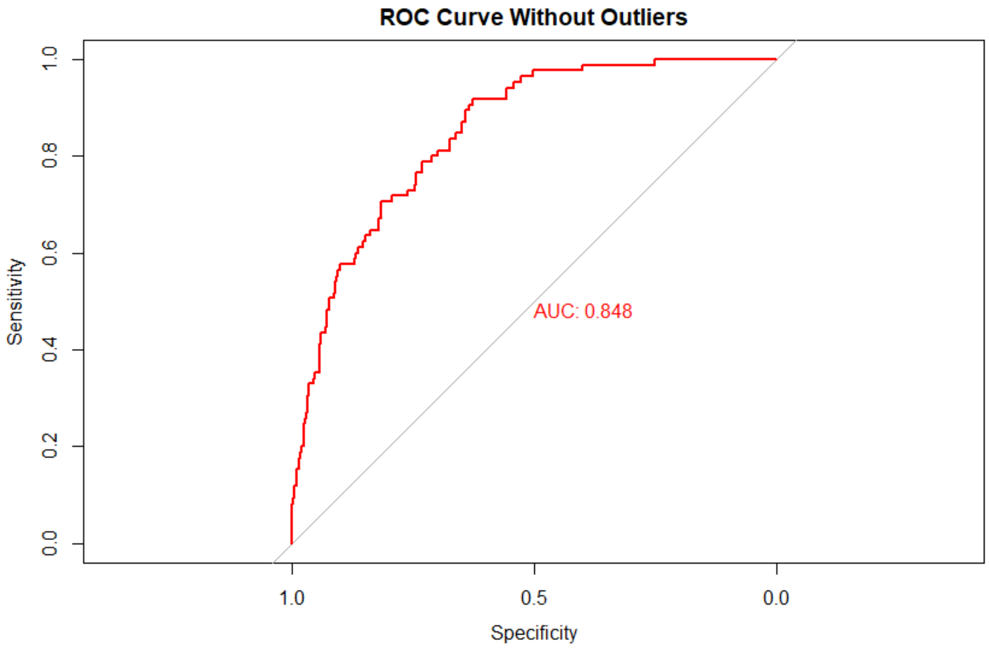


Figure 32. ROC for model without outliers

4.2.4.5. Comparison of Canberra Models

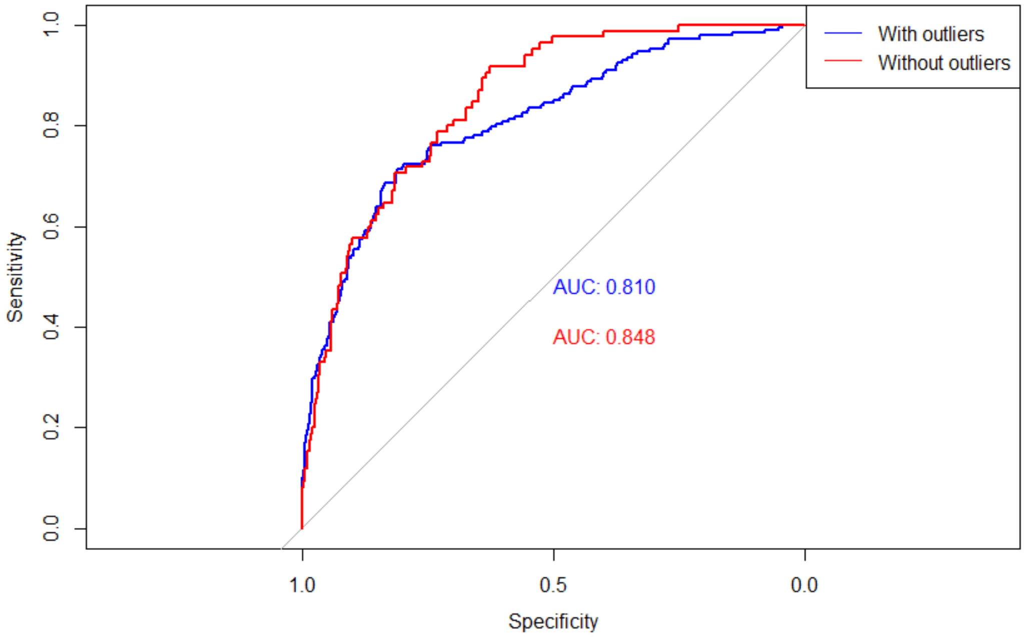


Figure 33. ROC comparisons for Canberra models

After building the model on data with and without outliers. We can deduce that the accuracy of the two models are close (84.86% and 84.31% respectively), however, the accuracy for the model without outliers was no better than no information rate. This indicates that the model on data with outliers would perform better. In terms of capability of the models to distinguish between the “Yes” and “No” class labels, shown below, the model on data without outliers outperforms the model on data with outliers.

4.3. Support Vector Machine

As discussed in literature review, support vector machine (SVM) are supervised learning models which analyze data used for classification and regression analysis. It is mostly used in classification problems. It is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In this project, each data item is plotted as a point in n-dimensional space (where n is number of features), with the value of each feature being the value of a coordinate.

4.3.1. Data Preprocessing

The first step of data pre-processing is to import libraries. As we have used R for our analysis, *InformationValue* library has been used for info matrix, *imputeTS* for filling missing values and *ggplot2* for visualization. Secondly, we have dealt with the categorical variables. As classifiers only recognize numbers, we have changed categorical into numbers such as for RainToday, RainTomorrow into 0,1 as well as WindDir3pm, WindDir9am and WindGustDir changed based on ranking (1 to 16). Thirdly, missing values have been filled with average values. Finally, for feature extraction, we have used correlation matrix to remove highly correlated values. In addition, we have created two extra variables such as DailyTempMean (daily mean temperature) and DailyPressMean (daily mean pressure).

4.3.2. Adelaide analysis

4.3.2.1. Model creation

The selection of input variable depends on the cross-correlation of the variables in the dataset. For this project, we have selected the variables according to 0.7 cut-off since the highly correlated variables have the same effect on target variable. Package *e1071* has been installed for model creation, which is a perfect guide on SVM Training & Testing models. At first, we have removed two variables Cloud9am and Cloud3pm.Then we have divided the dataset into training and testing dataset where 75% data is used as training and 25% is used as testing data. Mainly, linear and non-linear kernel both has been used to fit the hyperplane or data. As a result, we found linear kernel is a best to fit weather data, so we have used that Also, this model has generated below formula. Those variables have been written in formula helps to create a best model for area Adelaide.

$$RainTomorrow = Evaporation + Sunshine + WindGustDir + WindGustSpeed + WindDir9am$$

These variables play a significant role to create best model for Adelaide area. Here, model has made 746 support vector machines with two classes 376 and 370.

4.3.2.2. Performance for Adelaide Area

As results shown in Table 14, with the total of 773 observations in Adelaide area. There are 602 days observed as no rain, the model can correctly predict 569 days, and wrongly predict 33 days as rain. There are 171 days observed as rain, and the model can correctly predict 93 days, and wrongly predict 78 days as no rain.

Table 14. Confusion matrix with outliers

		Actual Observations	
		No Rain	**Rain**
Predicted	**No Rain**	**569 (73.61%)**	78 (45.61%)
	Rain	33 (26.39%)	**93 (54.39%)**
Total Percentage		100%	100%

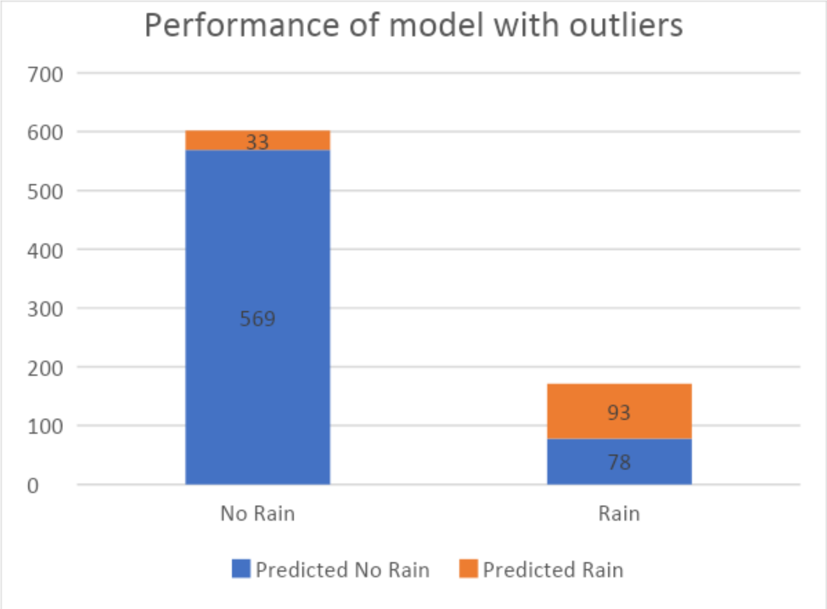


Figure 34. Visualization of Confusion Matrix for Model with outliers

As shown in Table 15 the performance of model without outliers, similarly with total of 773 observations in Adelaide area. For 602 days observed as no rain, the model has the same prediction to the model with outliers for correctly predict 569 days, and wrongly predict 33 days. There are 171 days observed as rain, and the model can correctly predict 95 days compare to 93 days of for the model with outliers, and wrongly predict 76 days as no rain.

Overall, the model was built without outliers performed slightly better than the model with outliers in predicting there will be rain in the next day with 2 more correct observations.

Table 15. Confusion matrix without outliers

		Actual Observations	
		No Rain	**Rain**
Predicted	**No Rain**	**569 (94.52%)**	76 (44.44%)
	Rain	33 (5.48%)	**95 (55.56%)**
Total Percentage		100%	100%

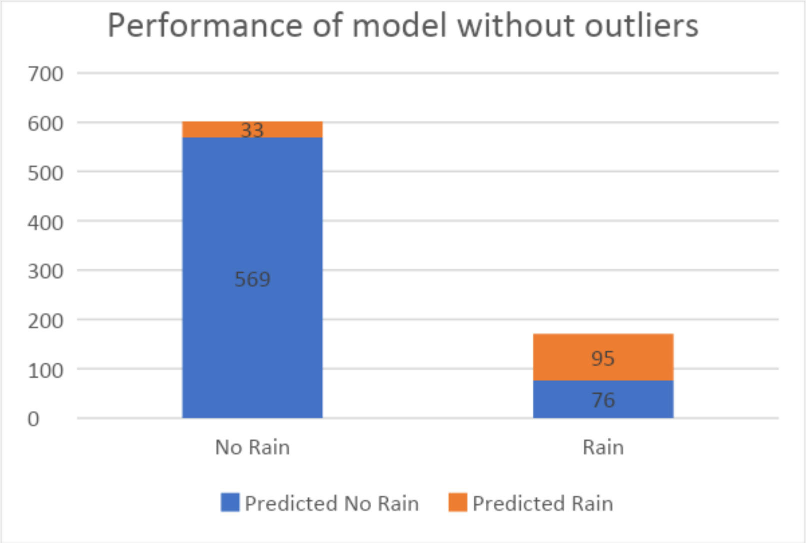


Figure 35. Visualization of Confusion Matrix for model without outliers

As results shown in Table 16, we can see that the model was built without outliers performs slightly better than the model was built with outliers. The prediction accuracy is impressive at 85.9%, however, the model does not perform as good as the expectation as there is still 14.1% chance of predicting wrong about the rainfall in the next day, which is reasonably high chance of failure in predicting.

The reason for this could come from the perfection of data which is quite significantly large amount of missing data was found in the dataset. In fact, there are no data for 2 attributes “cloud9am” and “cloud3pm” in Adelaide Area. The missing of cloud data may could be the factor that affect the performance of the model as there is lack of data for analysis. As cloud is formed from the water vapor evaporates into the air then accumulate enough to become droplets, but the droplets are not heavy enough to fall back to the ground and it stay up in the sky to form cloud. Once the condensation of the droplets is long enough which makes the droplets of water stick together to create heavier and larger water drops, with the force from gravity, it then falls and become rain. Hence, cloud could be an important feature in the prediction model for Adelaide Area.

Table 16. Model performance comparison

	With Outliers	Without Outliers
Accuracy	85.64%	85.9%
Precision	94.52%	94.52%
Recall	87.94%	88.22%
F – score	91.11%	91.26%
Misclassification Error rate	14.36%	14.1%

As shown in Figure 36, the ROC curves of both SVM models were built with and without outliers for Adelaide area. We aim to build the model that has the curve closer to the upper left corner which also mean to maximize the area under the curve (AUC), this also explain how good the model performs. Since the larger AUC, the better model is, we can see that the model was built without outliers (AUC = 0.75) is slightly better than the model with outliers (AUC = 0.745). Additionally, 2 ROC curves are too close to each other which nearly make no different in performance, the model without outliers also has slightly better accuracy at 85.9%. Hence, we will use the model that is built without outliers.

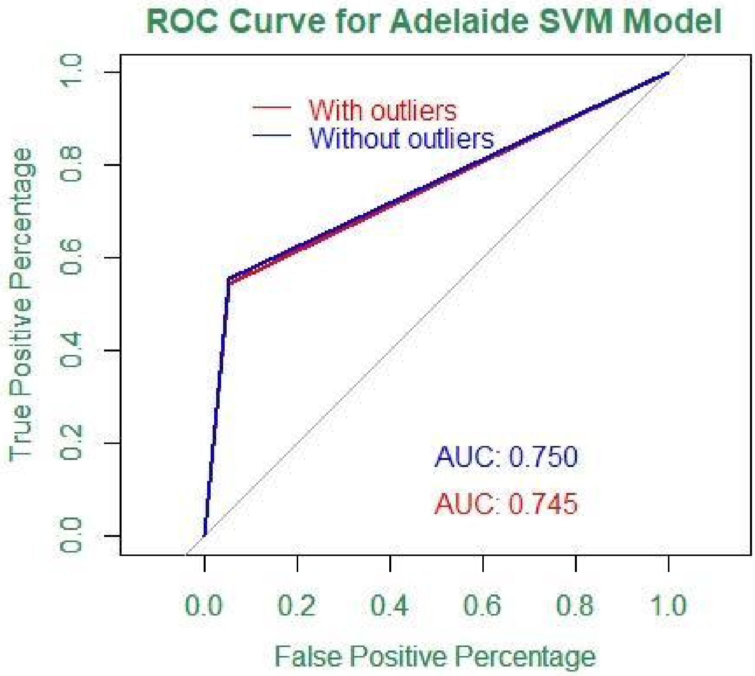


Figure 36. ROC Curve for Adelaide Model

4.3.2.3. Final Decision for Adelaide Model

From the performance analysis, even though the ROC values are not good enough, the model without outliers performs better not only in the confusion matrix but also in the ROC curve. ROC value represents the predictive power for a model where around 75% predictive power is not able to illustrate reliable prediction. However, data without outliers should be considered the best model for Adelaide rainfall prediction.

4.3.3. Canberra analysis

4.3.3.1. Model Creation

For Canberra, we have divided the dataset as the same way we did for Adelaide data, where 75% data is used as training and 25% is used as testing data. SVM model has created below formula for Canberra analysis.

$$RainTomorrow = Evaporation + Sunshine + WindGustDir + WindGustSpeed + WindDir9am$$

It has included two variables Cloud9am and Cloud3pm as compared to SVM for Adelaide. So, we can say that these two variables play a significant role to create a better model for Canberra analysis. Model creation has made 836 support vector machines with two classes 422 and 414.

4.3.3.2. Performance for Canberra Area

As results shown in Table 17, with the total of 855 observations in Canberra area. There are 708 days observed as no rain, the model can correctly predict 697 days, and wrongly predict 11 days as rain. There are 147 days observed as rain, and the model can correctly predict 65 days, and wrongly predict 82 days as no rain.

Table 17. Confusion matrix with outliers

		Actual Observations	
Predicted		**No Rain**	**Rain**
	No Rain	**697 (81.52%)**	82 (55.78%)
	Rain	11 (18.48%)	**65 (44.22%)**
Total Percentage		100%	100%

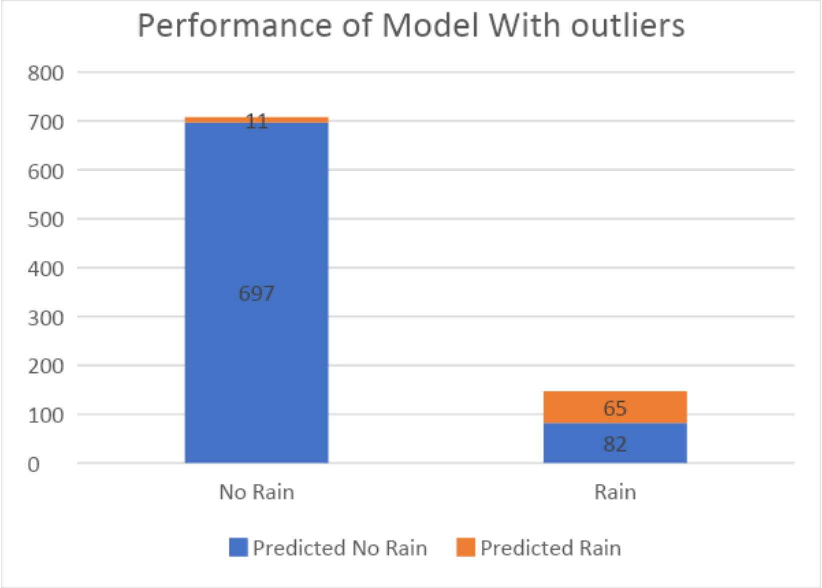


Figure 37. Visualization of Confusion Matrix for Model with outliers

As shown in Table 18 the performance of model without outliers, similarly with total of 855 observations in Canberra area. For 708 days observed as no rain, the model can correctly predict 698 days compare to 697 days for the model with outliers, and wrongly predict 10 days as rain. There are 147 days observed as rain, and the model can correctly predict 66 days compare to 65 days of for the model with outliers, and wrongly predict 81 days as no rain.

Overall, the model was built without outliers performed slightly better than the model with outliers in predicting there will be no rain in the next day with 1 more correct observation. Moreover, the predicted observations of the next day will rain also perform better which is predicted as 66 times, more than 1 time compared with the model with outliers.

Table 18. Confusion matrix without outliers

		Actual Observations	
		No Rain	**Rain**
Predicted	**No Rain**	**698 (81.64%)**	81 (55.1%)
	Rain	10 (18.36%)	**66 (44.9)**
Total Percentage		100%	100%

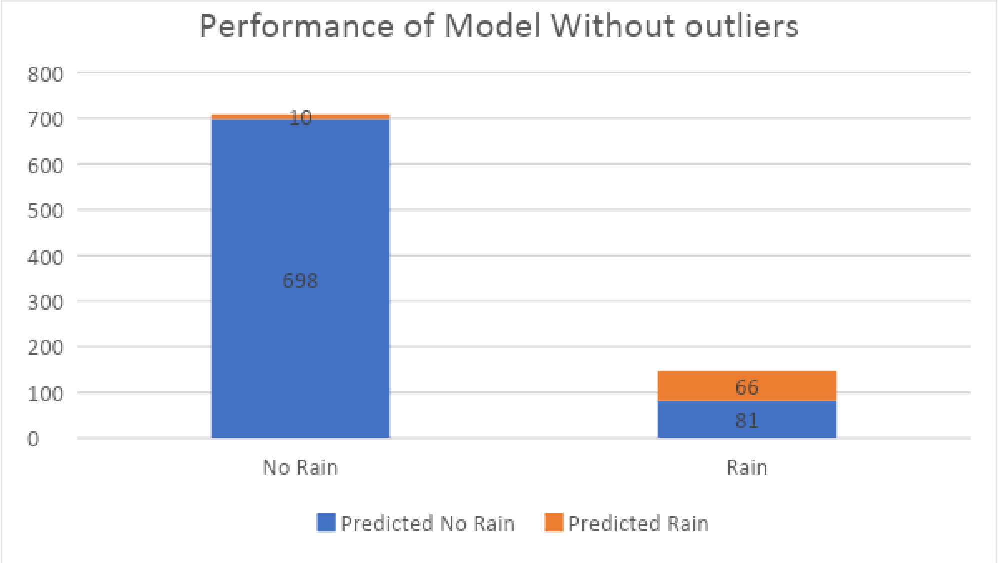


Figure 38. Visualization of Confusion Matrix for Model without outliers

As results shown in Table 19, we can see that the model was built without outliers performs slightly better than the model was built with outliers. The prediction accuracy is impressive at 89.36%, however, the model barely perform as the expectation as there is 10.64% chance of predicting wrong about the rainfall in the next day, which is reasonably low chance of failure in predicting.

Table 19. Model performance comparison

	With Outliers	Without Outliers
Accuracy	89.12%	89.36%
Precision	98.45%	98.59%
Recall	89.47%	86.84%
F – score	93.75%	93.88%
Misclassification Error rate	10.88%	10.64%

As shown in Figure 39, the ROC curves of both SVM models were built with and without outliers for Canberra area. We aim to build the model that has the curve closer to the upper left corner which also mean to maximize the area under the curve (AUC), this also explain how good the model performs. Since the larger AUC, the better model is, we can see that the model was built without outliers (AUC = 0.717) is slightly better than the model with outliers (AUC = 0.713). Additionally, 2 ROC curves are too close to each other which nearly make no different in performance, the model without outliers also has slightly better accuracy at 89.36%. Hence, we will use the model that is built without outliers.

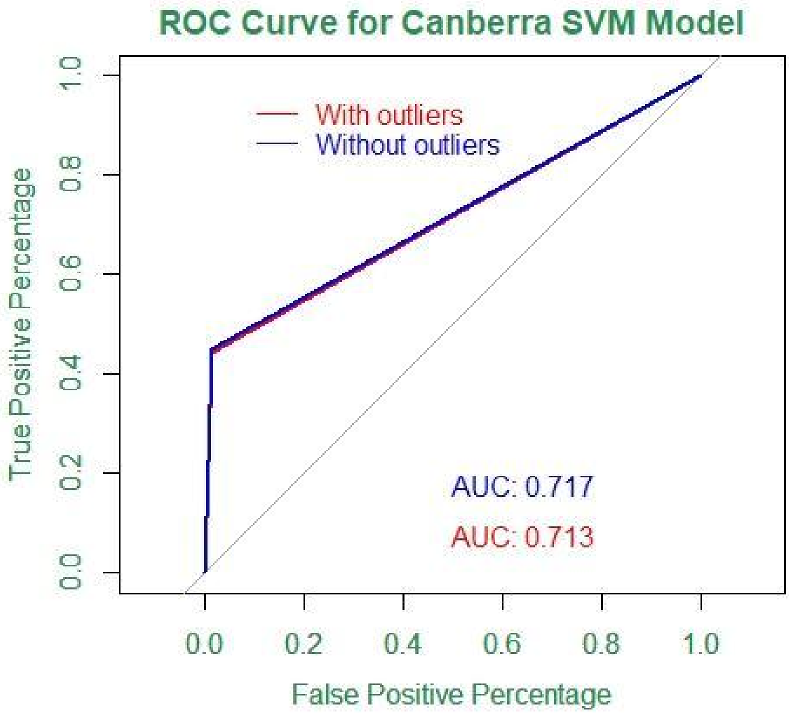


Figure 39. ROC Curve for Canberra Model

4.3.3.3. Final Decision for Canberra Model

From the performance analysis, we have found that the model without outliers performs better not only in the confusion matrix but also in the ROC curve. Therefore, we should choose model using data without outliers for Adelaide Canberra prediction.

4.3.4. Summary of SVM

Overall, there are impressive performance for the SVM model to predict rain in Adelaide and Canberra area. Even though each of the area will have its own type of weather characteristics, attributes are pretty much same for two models that has been selected by the model itself. On the other hand, performance of both Adelaide and Canberra area models without outliers is better compared to the models with outliers.

Additionally, as the number of rainy-day observations is significantly smaller than the number of observations for no rainy day for both Adelaide and Canberra, there are little training data for the rainy-day compare to no rainy-day. Hence, same as Logistic Regression model, SVM model also performed better when predicting no rainy-day and suggesting these 2 cities are quite similar in weather condition.

In conclusion, SVM model almost meets the expectation of the team even though the misclassification rate is around 10%. It could be the imperfection of data as there are a lot of missing data is reported in the previous section, or it could be the SVM method is not as powerful as other machine learning model for weather forecast problem.

4.4. Discussion

Overall, logistic regression model has an impressive performance in predicting rain in Adelaide and Canberra with around 14.23% of misclassification rate for Adelaide and 10.99% misclassification rate for Canberra. The performance could be affected by different reasons such as the perfection of data, especially for Adelaide area where there is no cloud data to train logistic regression model. Also, type of collected attributes in the data or the logistic regression is not as powerful as other machine learning model. Additionally, different cities and regions will have its weather characteristics, so the attributes that are used to build logistic regression model will be differently depends on the cities and regions.

For the ANN models, the performances are promising and satisfying from many aspects. The overall accuracy for ANN model in Adelaide area reaches 86.09%, while the accuracy for Canberra reaches 84.86%. Both show a significant evidence that the models can predict rainfall based on climate parameters which will always outstrip the random guess. It is also proven that a refined dataset without outliers can result in a better model with better performance. However, accuracy is just one angle we are reviewing the model. As discussed before, the true negative rate, which represents the ability of a model correctly predicting that there will be rain tomorrow, matters a lot. Viewing the specificity of these ANN models, ANN model from Adelaide data without outliers have a 68.18% specificity, which is better than the logistic model, while the ANN model from Canberra data merely has 35.29% of specificity, which is really poor. Hence, we can conclude that the ANN model is not suitable for Canberra data. It is a complicated model with many hidden neurons, which increases the likelihood of overfitting. The accuracy and specificity are also outweighed by other models. Therefore, the proposed model for Canberra prediction shall fall into SVM or logistic regression.

For SVM model, it has slightly better performance compared to logistic regression model with around 14.1% misclassification rate for Adelaide and around 10.64% misclassification rate for Canberra. Though it has better accuracy and reduced misclassification rate, ROC value is very low that is around 72%. As ROC value represents the predictive power of a machine learning classifier, model with low ROC value cannot produce better prediction. For example, ROC value 50% means the model cannot separate the positive and negative classes. In our case, ROC value of SVM is slightly better than 50% which is not that good enough to predict outcomes. Therefore, we should focus on some other machine learning models to build a better rainfall predictive model.

All 3 models have impressive performance with the data of 2 cities Canberra and Adelaide, especially highly accurate prediction on no rain day. As Adelaide and Canberra have similar dry weather, there are more training data for no rain day than rainy day. Hence, we cannot tell how well the models performed on the location with rainy weather. Due to the time constraint, the team could not implement the models on other cities which are expected to have different type weather compare to Adelaide and Canberra such Darwin, Brisbane to compare the performance of 3 models on location with rainy weather.

5. Conclusion

This project aimed to build rainfall predictive models that can be applied in predicting whether it would rain based on environmental conditions the previous day in Adelaide and Canberra. The contribution of attributes to the predictive models built to be varied based on the cities. Three machine learning models; Artificial neural network, Logistic regression and Support vector machine models were built for this purpose. Each model was built first on a dataset containing outliers and again built on data without outliers. K-mean clustering, DBScan clustering and Histogram methods were used to detect outliers in the dataset, at the end we settled on using Histogram method. The performance of each model was evaluated using well-known performance indicators. The results, summarized in table 16, showed the following.

- For Adelaide, the logistic regression model built on data without outliers performed better with an accuracy of 85.77% as compared to the model built on data with outliers which showed 85.51% accuracy. This was also the case for Canberra with Logistic regression models showing an accuracy of 88.54% and 89.01% on data with outliers and on data without outliers respectively. Overall, the logistic regression models performed better for data without outliers.
- Artificial neural network model on Adelaide Data with outliers showed an accuracy of 84.77% while on data without outliers showed 86.09% accuracy. On the other hand, Artificial neural network model on Canberra Data with outliers showed an accuracy of 84.86% while on data without outliers showed 84.31% accuracy.
- Support vector machine model on Adelaide Data without outliers showed 85.9% accuracy while one on data with outliers showed an accuracy of 85.64%. The SVM model on Canberra Data with outliers had an accuracy of 89.12% while the one on data without outliers showed 89.36% accuracy.
- Overall, for Adelaide, the best model among the three was ANN without outliers with an accuracy of 86.09%, while for Canberra, SVM model on data without outliers was the best as it showed an accuracy of 89.36%

Table 20. Comparison table for the three models built for Adelaide and Canberra

	Model	With Outliers	Without Outliers
Adelaide	Logit	85.51%	85.77%
	ANN	84.77%	86.09%
	SVM	85.64%	85.90%
Canberra	Logit	88.54%	89.01%
	ANN	84.86%	84.31%
	SVM	89.12%	89.36%

In conclusion, the performance of the three models built performs well, but they can be improved by collecting more training data. Also, it would be important to explore other machine learning models to see if they would perform better.

References

Cheung, Hart, Peart , 2015, 'Projection of future rainfall in Hong Kong using logistic regression and generalized linear model', *5th International Workshop on Climate Informatics.*

Devi, SR, Arulmozhivarman, P, Venkatesh, P & Agarwal, C 2016, 'Performance comparison of artificial neural network models for daily rainfall prediction' *International Journal of Automation and Computing*, vol. 13, no. 5, pp. 417-427.

Imon, AR, Roy, MC, Bhattacharjee, O 2012, 'Prediction of rainfall using logistic regression', *Pakistan Journal of Statistics and Operation Research*, vol. 8, no. 3, pp. 655-667.

Kim, T-W & Valdés, JB 2003, 'Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks', *Journal of Hydrologic Engineering* , vol. 8, no. 6, pp. 319-328.

Jeongwoo Lee, Chul-Gyum Kim, Jeong Eun Lee, Nam Won Kim, & Hyeonjun Kim 2018, 'Application of artificial neural networks to rainfall forecasting in the Geum River basin, Korea', *Water (Basel),* vol. 10, no. 10.

Mandal, T & Jothiprakash, V 2012, 'Short-term rainfall prediction using ANN and MT techniques', *ISH Journal of Hydraulic Engineering,* vol. 18, no. 1, pp. 20-26.

Nanda, S, Nayak, S & Mohapatra, S 2013, 'Prediction of rainfall in India using Artificial Neural Network (ANN) Models', *International Journal of Intelligent Systems and Applications* , vol. 5, no. 12, p. 1-22.

Nayak, M & Ghosh, A 2013, 'Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier', *Theoretical and Applied Climatology,* vol. 114, no. 3-4, pp. 583-603.

Ozechowski ,T 2010, *Logistic Regression*, pp.940-942.

Prasad, K, Dash, SK & Mohanty, UC 2010, 'A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts', *International Journal of Climatology,* vol. 30, no. 10, pp. 1577-1588.

Samhitha, SV & Srikanth, PG 2017, 'PREDICTION OF RAINFALL USING INVERSE DISTANCE WEIGHTING METHOD AND ARTIFICIAL NEURAL NETWORKS IN PONNAIYAR RIVER BASIN', *Indian Journal of Scientific Research*, pp. 144-149.

Seo, J-H, Lee, YH & Kim, Y-H 2014, 'Feature Selection for Very Short-Term Heavy Rainfall Prediction Using Evolutionary Computation', *Indian Journal of Scientific Research*, vol. 2014 , no. 2014, p. 15.

Tolles, J & Meurer, WJ 2016, 'Logistic Regression: Relating Patient Characteristics to Outcomes', *JAMA,* vol. 316, no. 5, Aug 2, pp. 533-534.

van Buuren, Stef, Groothuis-Oudshoorn, Catharina Gerarda Maria & Faculty of Behavioural, Management Social Sciences 2011, 'mice: Multivariate Imputation by Chained Equations in R', *Journal of statistical software,* vol. 45, no. 3, pp. urn:issn:1548-7660.

Venables, W, Ripley & Springerlink 2002, *Modern Applied Statistics with S*, Fourth Edition., Springer New York, New York, NY.