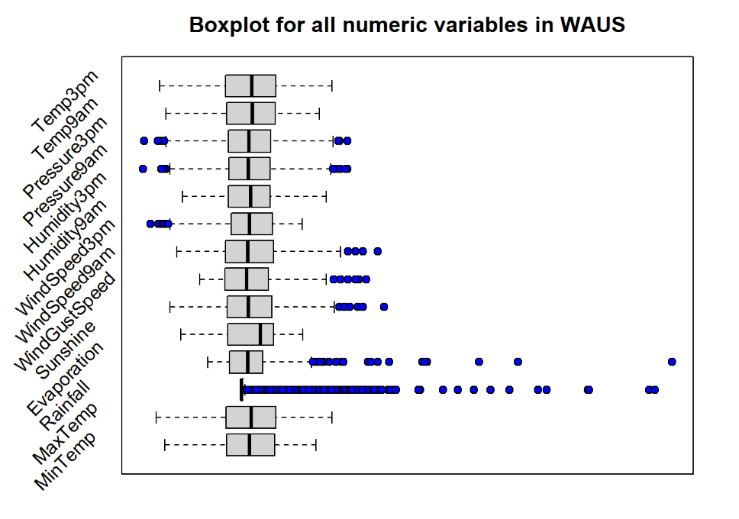
**A picture containing table

Description automatically generated1. Exploratory data analysis (half page)**

**Text

Description automatically generatedTable

Description automatically generated**

****

**Exploratory of variables :**

**Based on the statistical summary of the dataset (figure!!), there are few noteworthy insights explained below:**

* **The proportion of target variable is 30.65% (Not cloudy = 613, Cloudy = 1387).**
* **The mean of Pressure9am and Pressure3pm are notably high.**
* **Most of the variables are skewed. For example, Rainfall and Evaporation are extremely right-skewed. That means its median is larger than its mean. Contrariwise, Humidity9am and Sunshine are left-skewed.­**
* **Lots of variables contain NA values. Evaporation and Sunshine have 43% and 39% missing values.**
* **Since this analysis is predicting whether or not the following day will be cloudy, there are not needed to be predictor variables, namely Location, Day, Month, and Year; hence they will be omitted.**

**On top of the insights, based on (figure!!), Rainfall is the most extreme variable as it has the highest range and lots of outliers. Few variables like Temp3pm and MaxTemp are of similar distribution.**

**2. Data Cleaning and Pre-processing (half page)**

**Before modelling, there are few types of data pre-processing should be conducted which is gleaned from the part 1 summary.**

**2.1. dealing with Categorical values**

**The data type of few variables are Char which are meant to be categorical. I used as.factor () to categorise the variables of all wind direction, RainToday, and CloudTomorrow.**

**2.2. Handling Missing values**

* **Generally, it is assumed that these data points are missing at random (MAR), meaning that the tendency for a data point to be missing is not related to the missing data, but it relates to some of the observed data. Therefore, all numeric variables with NA values are imputed. Particularly, variables with lots of outliers like Rainfall are imputed with median otherwise with mean. Also, categorical variables with NA values omitted as it is arbitrary to impute them with mode.**
* **Another removing or imputing the values, instead of ignoring them, is that this analysis is to make classification models which is better to have binary values.**

RF cannot NA

Bagging can NA

Boosting can NA

ANN cannot NA

For the sake of comparing models, NA needs to be removed

Why AUC not accuracy in case

* do not care FP or TN, none of which is serious, so use AUC

**2.3. Data** Splicing

**At the stage, the dataset is cleaned to be ready to model. We split the dataset into 70% training and 30% test set, with random seed set as 31084222.**

**3. Original modelling**

**3.1 – 3.5 5 models**

**Evaluation:**

**3.6 confusion matrix**

**3.7.1 ROC 3.7.1 AUC**

**A picture containing text, scoreboard

Description automatically generatedTable

Description automatically generatedChart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**3.8 variable importance**

**Only Humidity3pm, MinTemp, Sunshine,** **WinDir3pm, WinDir9am and WinGustDir could be selected from the data to increase the performance of the model.**

**How handle and why**

**They are based on the Gini index, which is a measure of variance. The higher the** variance means the more misclassification exists, thus driving the values of Gini Index higher.

In this scenario, ensemble methods like Bagging, Boosting and Random Forrest are used to determine the variable importance. The reason I do not use normal decision tree and Naïve bayse is that their variables are of equal weighting. Whereas methods like bagging are suitable as their task is to **reduce the variance error. They analyse** **different subset of data and generate a collective output at the end. Since boosting weighs its variables from the bags, it is particularly useful. Therefore, variable importance is based on the ensemble methods.**

**The way features selected can be seen in figure!!. Three models follow the same patterns on weighing its variables. Therefore, I set the** threshold value to 0.3 (figure!!). Only features higher than 0.3 are selected, which yield better classification.

**4. model improved:**

**4.1. based on simplicity**

* Describe your model, either with a diagram or written explanation.
* How well does your model perorm, and how does it compare to those in Part 4?
* What factors were important in your decision and why you chose the attributes you used.

**4.2 based on performance**

* How to improve
  1. prune trees
  2. CV
  3. boosting
* You may do this by adjusting the parameters, and/or cross-validation of the basic models in Part 4, or using an alternative tree-based learning algorithm.
* Show that your model is better than the others using appropriate measures.
* Describe how you created your improved model, and why you chose that model.
  1. What factors were important in your decision and why you chose the attributes you used.

**5.ANN**

* Comment on attributes used and your data pre-processing required.
* **2.5 Feature Scaling?**
* but there could be some value in standardizing your variables if you're interested in predictor importance scores. RF will tend to favour highly variable continuous predictors because there are more opportunities to partition the data. A better way to deal with this issue, however, is to use particular approaches ([cite](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25))
* How does this classifier compare with the others? Can you give any reasons?