OSTA5003 Project plan and EDA - Group1

Semester 1B, 2023

Problem Overview:

The dataset we have chosen is “Rain in Australia” from kaggle which contains 145460 weather observations recorded in various weather stations around Australia covering a period of 10 years from end of 2007 to mid 2017. This dataset includes a combination of categorical and numerical weather attributes like temperature, rainfall amount, cloud, humidity, pressure, wind components with the binary target variable to predict rain for today and tomorrow which is therefore a classification problem.This will be an interesting problem as predicting weather patterns is a crucial driving factor for planning purposes in several fields like agriculture, aviation etc.

```
suppressWarnings(library(readr))
suppressWarnings(library(dplyr))
suppressWarnings(library(reshape2))
suppressWarnings(library(caret))
suppressWarnings(library(lattice))
## Reading "Rain in Australia" Dataset
library(readr)
library(dplyr)
library(ggplot2)
AUS_rain_data <- read.csv("weatherAUS.csv", header = TRUE,
                          stringsAsFactors = TRUE)

# Calculate the percentage of missing values
missing_percentage <- sum(is.na(AUS_rain_data)) /
  (nrow(AUS_rain_data) * ncol(AUS_rain_data)) * 100

#Remove NA values
AUS_rain_data <- na.omit(AUS_rain_data)

#Quick summary of the dataset
```

```
#summary(AUS_rain_data) #commenting to save pages
```

There are quite a few challenges with this dataset as the data is imbalanced for target variable-RainTomorrow as there are more no-rain observations than rain. Also, we can see that around 10.2597457% of missing values in this dataset which could have certain level of impact in the accuracy of classification models. Further to this, there could also be some outliers (eg: heavy rains and floods, may be once in a century) which can lead to overfitting. Some of the independent variables could be highly correlated, so we will have to be careful in choosing the attributes while training the models by applying PCA for feature selection.

```
# Load the caret package
library(caret)
library(lattice)

# Scale the numerical variables to be within the min-max range
num_vars <- c("MinTemp", "MaxTemp", "Rainfall", "Evaporation", "Sunshine",
              "WindGustSpeed", "WindSpeed9am", "WindSpeed3pm", "Humidity9am",
              "Humidity3pm", "Pressure9am", "Pressure3pm", "Cloud9am",
              "Cloud3pm", "Temp9am", "Temp3pm")

AUS_rain_data[num_vars] <- scale(AUS_rain_data[num_vars])

# Determine correlation matrix
data_num <- AUS_rain_data[, num_vars]
cor_mat <- cor(data_num)
#print(cor_mat)
```

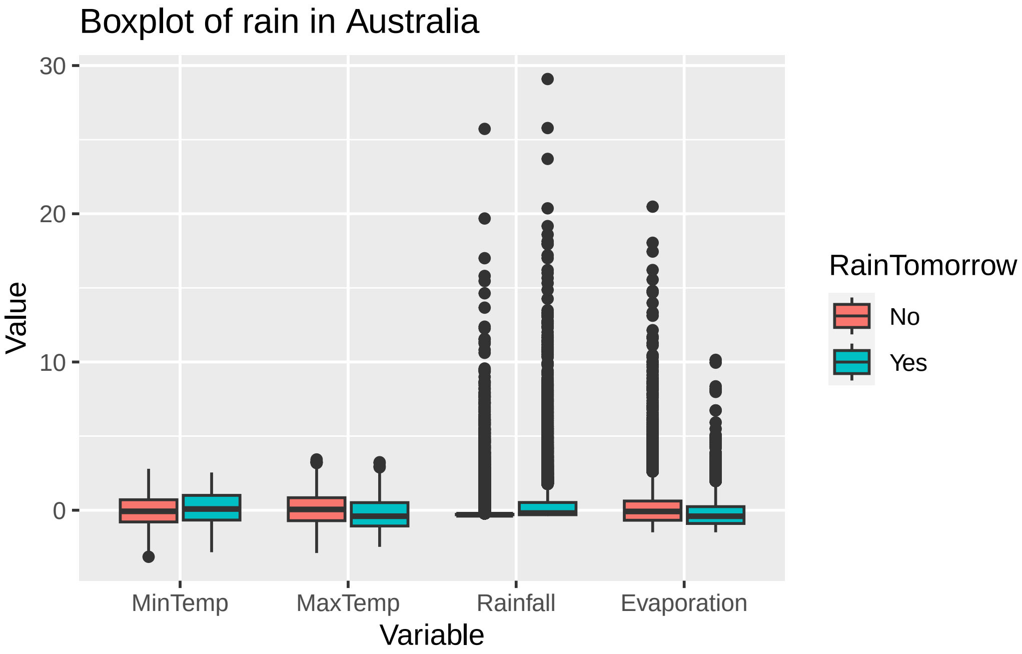
From correlation matrix, it is clear that certain attributes are highly correlated. For example, MinTemp and Temp9am/Temp3pm features are highly correlated so it is better to exclude Temp9am/Temp3pm during feature selection process. This leads to performing PCA for dimensionality reduction in the feature selection process.

```
# Perform PCA
pca <- prcomp(AUS_rain_data[, num_vars], scale. = TRUE)
#print(pca)
```

Based on the outcome from PCA, four features with highest variances are {MinTemp, MaxTemp, Rainfall, Evaporation} which will be chosen for training the models. There is a potential for using various classification models but we will be using Logistic Regression, SVM, Decision trees and Random Forest techniques by splitting the data into 75:25 for training and test sets. Models will be evaluated using test data and classification metrics will be represented in the form of confusion matrix and accuracy measure. With this exploratory data analysis, we are planning to train the classification models and prepare the final report and presentation in the coming days and targeting for completion prior to the due date.

```
library(reshape2)
data_long <- melt(AUS_rain_data, id.vars = "RainTomorrow",
                  measure.vars = c("MinTemp", "MaxTemp", "Rainfall", "Evaporation"),
                  variable.name = "Variable", value.name = "Value")

# Create box plot
ggplot(data_long, aes(x = Variable, y = Value, fill = RainTomorrow)) +
  geom_boxplot() +
  labs(title = "Boxplot of rain in Australia")
```



There are few outliers in the key features. Given the number of records in the dataset, we will be focusing to remove the records with extreme outliers which in this case is the observation with evaporation rate of 145.