Final Report

Overview

The "Rain in Australia" dataset available on Kaggle contains daily weather observations from numerous weather stations across Australia, spanning over a period of 10 years from 2008 to 2018. The dataset consists of 145460 rows and 23 columns with various data types including numerical and categorical, providing various details about the weather conditions.

Features in the dataset include Date, which indicates the date of the recorded weather data, and Location, which specifies the location of the weather station. Other attributes describe weather-related information, such as Rainfall and Evaporation in millimetres, Sunshine (in hours), WindGustDir (direction of the strongest wind gust in compass points) and WindSpeed9am/WindSpeed3pm (wind speed in killometres per hour at 9am and 3pm respectively). The dataset also includes attributes related to temperature, humidity, pressure and cloud cover. These attributes provide further insight into the weather conditions, allowing analysis of temperature variations, air pressure levels and cloudiness.

The objective of this project is to examine the provided dataset and construct a model capable of accurately forecasting whether it will rain the following day in a specific location in Australia, taking into account the prevailing weather conditions. Several classification methods, such as Logistic Regression, K-Nearest Neighbours (kNN), Decision Trees, Random Forest, Support Vector Machines (SVM), and Naive Bayes techniques will be employed to ascertain the target variable. Performance metrics such as Accuracy, Precision, Recall, and F1 Score will be utilised to assess the effectiveness of these techniques. By utilising these approaches and evaluation measures, the project seeks to develop a reliable model that can provide accurate predictions regarding rainfall occurrence based on the given weather parameters.

Exploratory Data Analysis

We are working on a classification task, determining if tomorrow will rain or not for a particular location in Australia. Let's first investigate the various attributes and patterns within the dataset, examining the distribution of the variables and calculating descriptive statistics.

```r
weather_dat <- read.csv("weatherAUS.csv", header = TRUE, stringsAsFactors = FALSE)
```

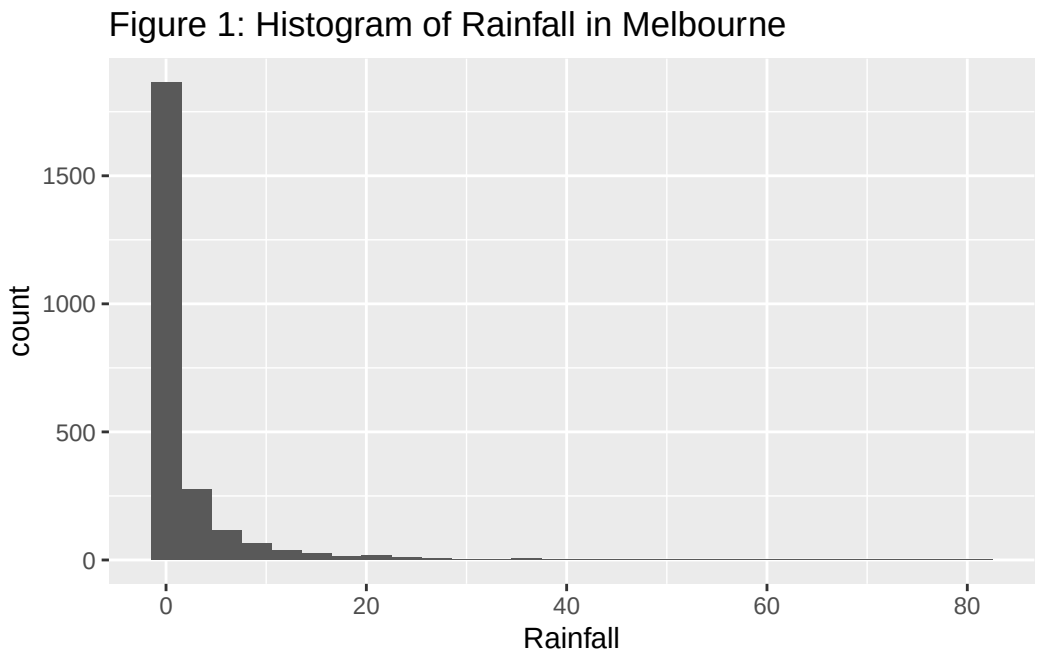
Due to the large dataset, for the sake of simplicity we have completed our initial data analysis and visualisation for Melbourne only.

```r
weather.melb <- weather_dat %>%
  filter(Location == "Melbourne")

weather.melb$Date <- as.Date(weather.melb$Date)
weather.melb$RainTomorrow <- as.factor(weather.melb$RainTomorrow)
weather.melb$RainTomorrow[is.na(weather.melb$RainTomorrow)] <- as.factor("No")
```

```r
weather.melb %>%
  ggplot(aes(x = Rainfall)) + geom_histogram(binwidth = 3) + ggtitle("Figure 1: Histogram
```

```
Warning: Removed 758 rows containing non-finite values (`stat_bin()`).
```



We begin with an initial inspection of figure 1 containing the rainfall in Melbourne. The right tailed distribution especially indicates there is minimal to no rain in Melbourne.