## Australia Rain Fall Prediction

## Introduction

Using science and technology, it is possible to forecast the quantity of rainfall that will occur in a certain location. It is essential to have precise knowledge of rainfall for accurate water resource allocation, agricultural production, and water structure design. Since the prediction of rainfall remains an important topic in meteorology, there is still much to be learned about it. Rainfall prediction is especially important in informing the community of a severe rain event before it occurs, which helps avoid damage and loss of life. This is most prevalent when it comes to both the economy and human lives. The primary need for a successful prediction system is the ability to detect the arrival of drought and flood ahead of time, allowing both the public and government to be prepared for any kind of natural catastrophe. Failure of current statistical methods for rainfall prediction may be largely attributed to the presence of atmospheric dynamics.

A major worry in the minds of all those involved in government, industry, and risk management is the difficulty of predicting rain, a problem that has attracted attention from many entities, including scientific organizations. Human activities are affected by a climatic factor such as rainfall crop, building, electricity, forestry, and touristic production among numerous others . The correlation of bad weather with rain helps us see how important forecasting rain is mass movements, avalanches, flooding, and landslides. These situations The years have passed since the attack has affected society . So, since planning ahead to predict rainfall is possible, planning is possible to carry out both preventive and reactive measures. Natural events like this We employed a range of machine learning methods to resolve this ambiguity. frameworks that will produce credible and up-to-date predictions The aims of this paper are to deliver endending the classification algorithm life cycle from beginning to end methods for testing them Data imputing missing in preprocessing.

For Rainfall Prediction, I will be using multiple techniques of Classification models that are available in Machine Learning Techniques such as, Support Vector Machines, Decision Tree , Xg-boost, etc., are the models that we will be using as part of this project. I will be using Python language and Jupyter IDE for the whole project. To successfully completing the project I will be using Python libraries such as NumPy, Pandas, Scikit Learn, etc.

## 1.1.Data

In this project, we will use Machine Learning to solve this problem of predicting rainfall with higher accuracy. The data that we are using in this project belongs to Australian weather conditions, which were made available by the Australian government. This data was collected from the Kaggle website, and this data includes a variety of useful features that can help us classify the rainfall prediction. Features such as Date, Location, Temperatures, Wind speed, Wind direction, Sunshine, etc., are few such features that are in the data. There are total 145460 rows and 23 columns in the datasets. These are the following features in the dataset:

Date : The Date of Observation

Location : This feature describes about the Location name of the weather station

MinTemp : The minimum temperature in degree Celsius

MaxTemp : This feature describes about the maximum temperature in degree Celsius

Rainfall : This feature describes about the amount of rainfall recorded for the day in mm.

Evaporation : This feature describes about the so-called Class A pan evaporation (mm) in the 24 hours to 9am

Sunshine : The number of hours of bright sunshine in the day

WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight

WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight

WindDir9am : Direction of the wind at 9am

WindDir3pm : Direction of the wind at 3pm

RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

Temp3pm: Temperature (degrees C) at 3pm

Temp9am: Temperature (degrees C) at 9am

Cloud9am: Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

Cloud3pm : Fraction of sky obscured by cloud (in eighths) at 3pm

Humidity3pm : Humidity (percent) at 3pm

Humidity9am : Humidity (percent) at 9am

RainTomorrow : The target variable. Did it rain tomorrow?

## 1.2. Research Questions

## A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study. The research questions that we are trying to address in this project are:

1. What are the feature engineering methods that can be used in the data to extract more useful information and how the feature engineering will help us achieve higher accuracies while training our models?
2. What type of machine learning algorithm will help us achieve the best accuracy among the ones we are using in this project?

## 1.3.Project Aim

The goal of this project is to develop a classifier that can predict the Rainfall in Australia. The outcome will be assessed using alternative machine learning models. Because the data contains both numerical and text elements and there is large number of outliers. So we have to remove them first. Feature Engineering is very important part of this project. The Machine learning models will take the features as an input and gives the prediction of Rainfall as an output.

## 1.4. Objectives

* Clean the data by removing the outliers and handling the missing values.
* Make interpretations from the data visualization during the data analysis.
* Use classification models to accurately predict the Rainfall in Australia.

## 

## 1.5. Tools

There are several tools employed as phase of this project to achieve its main objective. a few essential tools utilized during this Endeavour include:

NumPy: NumPy is a Library in python designed to help with array management. It has features for linear algebra, Fourier transforms, and matrices, as well.

Pandas: Pandas is a Python-based library that works with data analysis and manipulation. The focus is on the operations and data structures required to manipulate tables and time - series data.

Scikit-learn: Scikit-learn is a library for Python that helps with machine learning. It is open source and available to anyone. It includes several algorithms, such as support vector machines, among others.

Matplotlib: Matplotlib is a tremendous 2D plotting library in Python, perfect for visualizing array data. Matplotlib is a library built on the concept of NumPy arrays, and it is made to work with the other components of the Scipy stack. Matplotlib has numerous plots, including lines, bars, scatter plots, histograms, and more.

Seaborn: Seaborn is an example of a Python library that works with the matplotlib data visualization framework and integrates with pandas data structures. Seaborn is Seaborn's central visualization system, which is crucial in helping the exploration of data. See how the distribution is univariate and bivariate.

## 1.6. Ethical, Legal, Professional and Social issues

## 1.6.1. Ethical Issues

**Risk Zone 3 - Economic & Asset Inequalities -**

Automation has led to the loss of employment for people who belong to certain industries. With the growth in Machine Learning and Artificial Intelligence, companies are adapting to automation since it is more economical and efficient compared to human labour. This project has a high chance of falling in this ethical risk zone since the outcome of the project is to automate the accurate predictions of the rainfall. This might end up creating unemployment by replacing the human workforce in this sector.

**Risk Zone 4 - Machine Ethics & Algorithmic Biases**

With the growth in Machine Learning and Artificial Intelligence, the dependency of humans on technology has increased vastly. However, there are cases where Artificial Intelligence failed in many scenarios. Considering such cases, how reliable is Artificial Intelligence in predicting the rain, and to what extent can we depend on the predictions made by the algorithms.

## 1.6.2. Legal Issues

Legal issues arise when we are doing something illegal either knowingly or unknowingly which might end up facing legal charges from the government. Hence it is highly important to identify the legal threats prior and avoid them.

The possible legal issue identified in this project is the issue of copyrights of the data. Since the data was made available by the Australian government and it is available on the Kaggle for research purposes it can be considered safe to use this data.

## 1.6.3. Social Issues

As discussed in the ethical issues, unemployment can be considered a social issue that harms society and the economic functionality of the nation. The companies should consider the situations and take actions that can protect the jobs. Also, the people should get themselves skilled as per the requirements of the companies to protect their jobs from being axed by companies.

## 2. Methodology

## 2.1. Installing set-up

I have used Python 3.7, which I downloaded from the official Python website, for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries.

Example : Pandas, Numpy, Scikit-learn etc.

## 2.2.Data Exploration

The data for this project is available at Kaggle [- https://www.kaggle.com/jsphyg/weather-dataset-rattle-package](-%20https:/www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes).The dataset includes 1,45,460 records with 23 variables combination data kinds make up the information.

The dataset is further explored to identify Exploratory data Analysis.

## 2.3 Exploratory Data Visualization

Due to the greater number of useful features in the data, the data gives us a high score of exploratory data analysis on the data. Data Analysis will help us understand the patterns in each feature with respect to the dependent variable and helps us understand how each feature is contributing to the rainfall prediction.

It is planned to perform a univariate analysis on each of the important features in the data. Using Data Visualizations, we can interpret the data much more efficiently.

In Exploratory data Analysis we will perform the visualization part and we will gain insights from the data. Exploratory Data Visualization (EDA) is very important part of the data science pipeline or any data science project. Exploratory Data Analysis is a vital method that involves conducting initial investigations on data in order to identify trends, identify discrepancies, evaluate assumptions, and verify conclusions using summary statistics and data visualizations. Exploratory Data Analysis (EDA) is a computational data analysis technique focused on John Tukey's pioneering work. EDA offers a basis for a wide variety of data analytic activities and addresses the diverse types of data and architecture encountered by applied researchers. EDA's fundamental conceptual and computational tools include the use of graphics and interactive data visualization, a focus on model creation, diagnosis, and evaluation, addressing fundamental measurement issues associated with various distributions.

Although these methods serve as a foundation for all research, EDA places a high value on data - based learning from data to enhance standard hypothesis testing procedures that may neglect critical unanticipated aspects of data and their effect on modelling and estimation. The EDA, it is claimed, is critical both in the early stages of science, where hypotheses and model development must be well-informed.

## 

## 2.3.1. Univariate Analysis

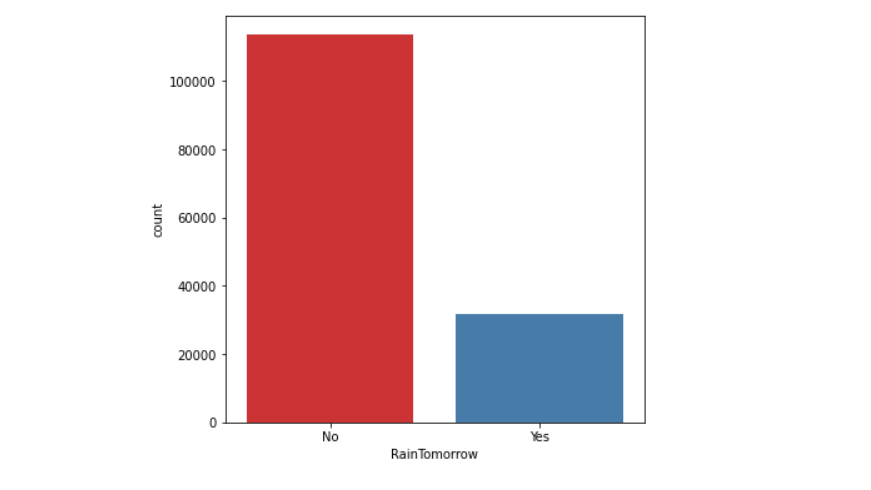
The technique of univariate analysis is for comparing and analyzing the relationship between a single feature and response variable. The prefix "uni" highlights the analysis only covering a single variable and its impact on a parameter.

For Example, the study can focus on a variable such as "gender," "height," or "weight."

However, only one variable is examined each time.

## 2.3.1.1 RainTomorrow

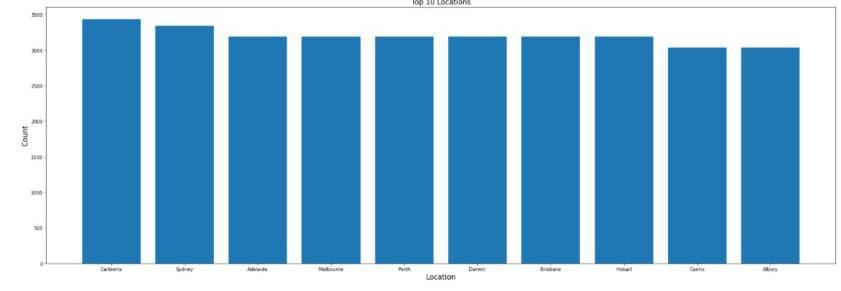
RainTomorrow is dependent variable in our dataset which is categorical feature. There are two categories which is yes and no.



The above plots show that most of the data points shows that there will no Rainfall tomorrow. Most of the data points (more than 100000) belongs to “No” category and around (3000) data points belongs to “yes” category.

## 2.3.1.2 Location

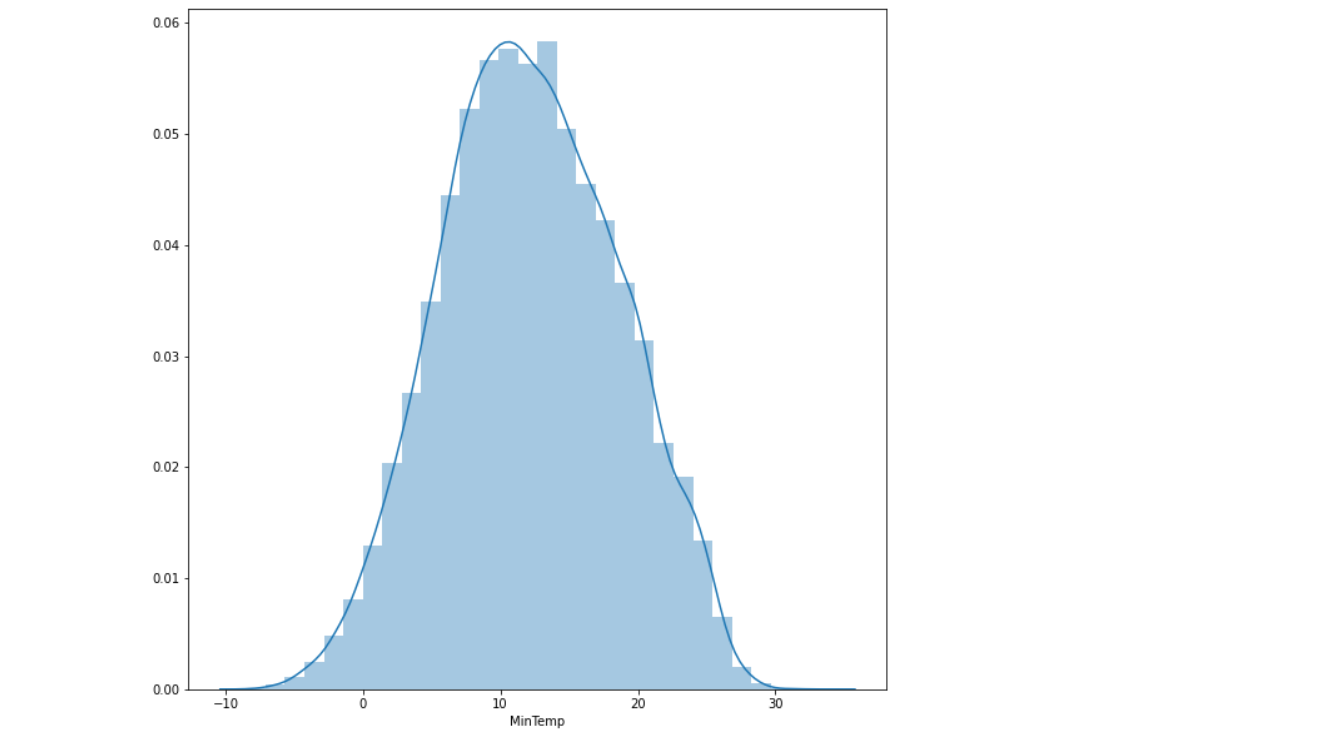
Location is a Categorical variable. This feature describes about the Location name of the weather station. Location feature contains 49 labels. In the below bar graph we can see that the top 10 Locations of Rainfall.



The topmost location for the Rainfall is Canberra.

## 2.3.1.3 MinTemp

MinTemp is a numerical feature. This feature describes about the minimum temperature in degrees Celsius.

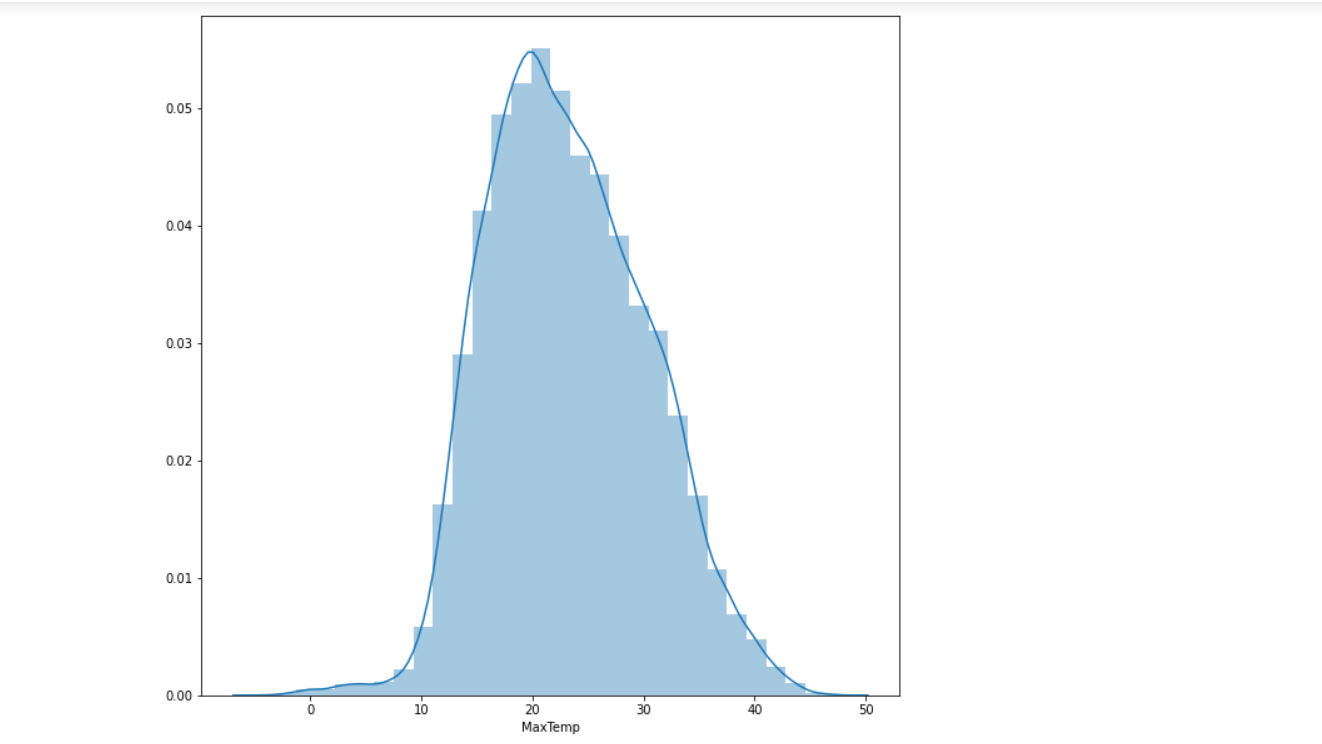


We can see that the graph is normally distributed there is no outliers in this feature. The min value of this temperature is -8.5 and maximum value is 33.900. The following table shows the description of the MinTemp variable:

|  |  |
| --- | --- |
| Count | 143975.000 |
| Mean | 12.1940 |
| Std | 6.398 |
| min | -8.5000 |
| 25% | 7.60000 |
| 50% | 12.0000 |
| 75% | 16.9000 |
| max | 33.9000 |

## 2.3.1.4 MaxTemp

MaxTemp feature describes about the maximum temperature in degrees Celsius. This is a integer variable.

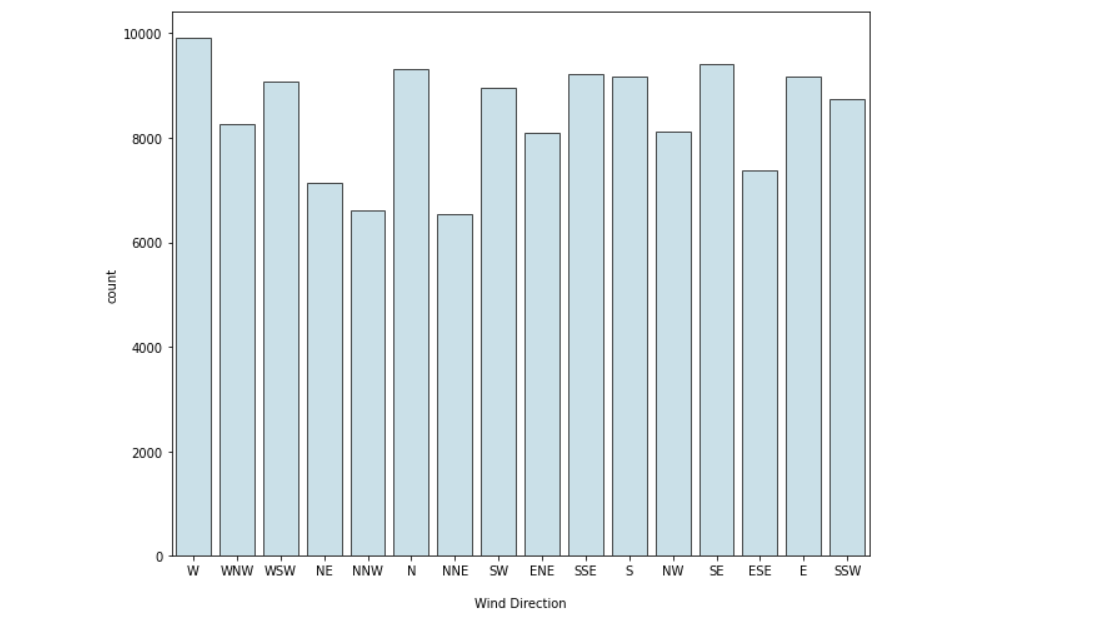


We can see that the graph is normally distributed there is no outliers in this feature. The min value of this temperature is -4.800 and maximum value is 48.1000. The following table shows the description of the MaxTemp variable:

|  |  |
| --- | --- |
| Count | 144199.000 |
| mean | 23.22 |
| std | 7.119 |
| min | -4.800 |
| 25% | 17.6000 |
| 50% | 22.6000 |
| 75% | 28.2000 |
| max | 48.10000 |

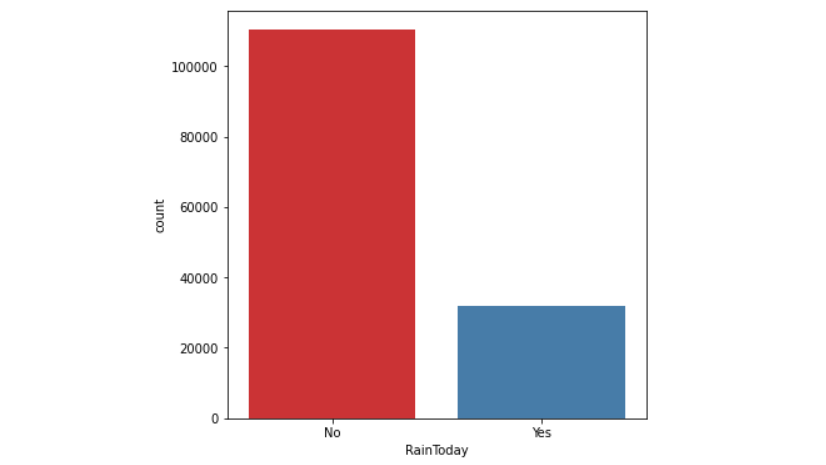
## 2.3.1.5 WindDir

WindDir is a Categorical variable which will give the information about the direction of wind. The direction of the strongest wind gust in the 24 hours to midnight. There are several categories in this feature.



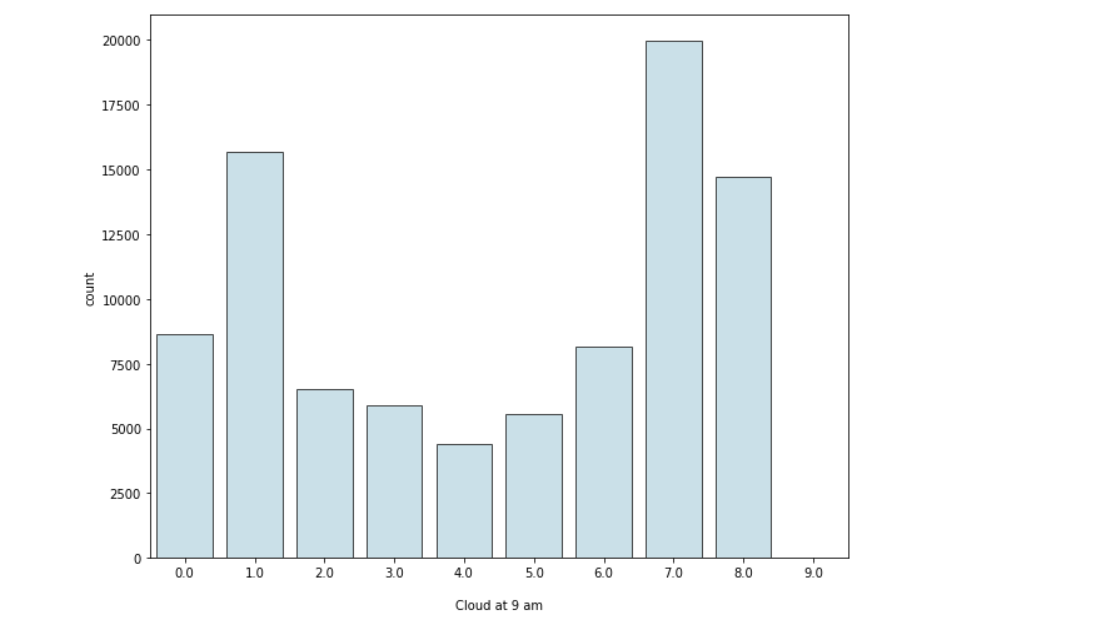
## 2.3.1.6 RainToday

RainToday is a Categorical feature. This is 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.



## 2.3.1.7. Cloud9am

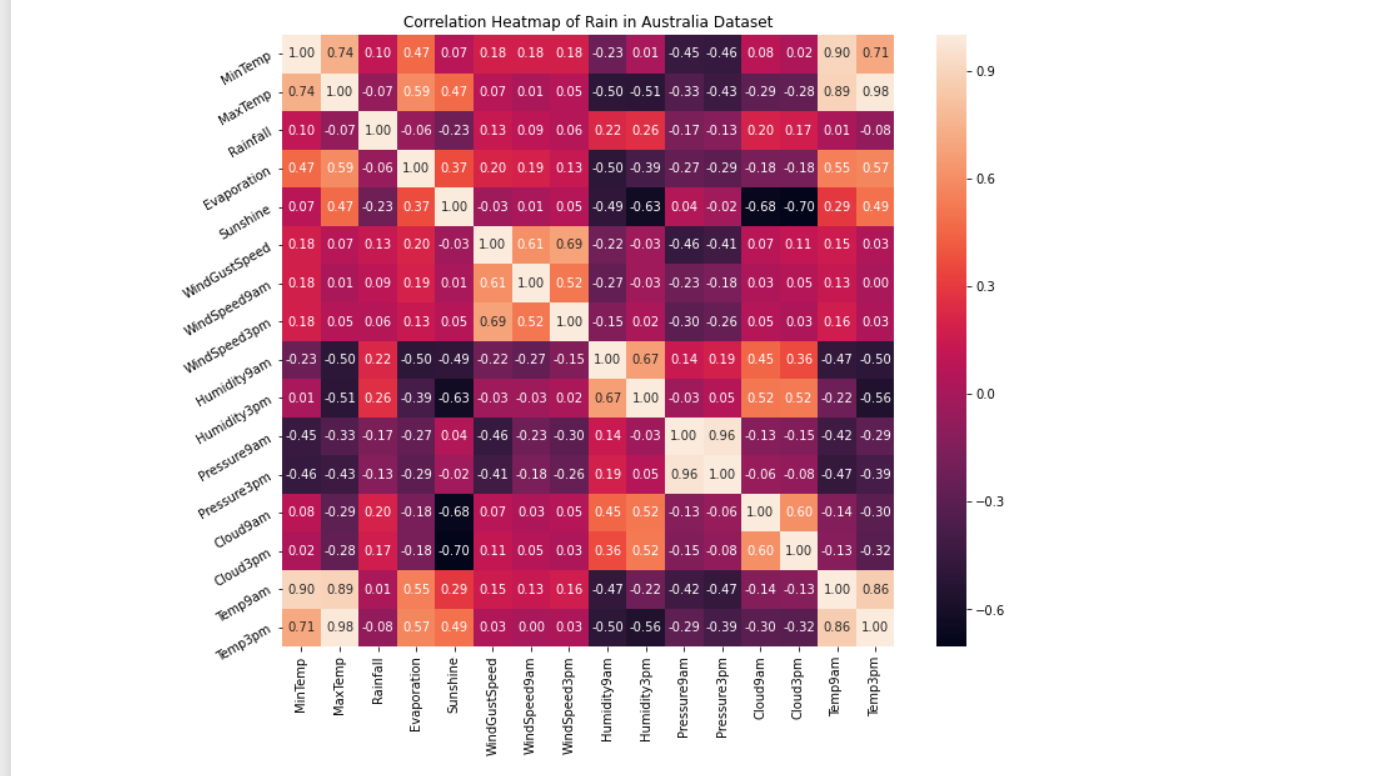
Cloud9am is a numerical feature. Which describes about the Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.



## 2.3.2 MultiVariate Analysis

The technique of Multivariate analysis is for comparing and analyzing the relationship between a Multiple features and response variable. The prefix "Multi" highlights the analysis only covering a multi variable and its impact on a parameter.

* An important step in EDA is to discover patterns and relationships between variables in the dataset.
* I will use heat map discover the patterns and relationships in the dataset.

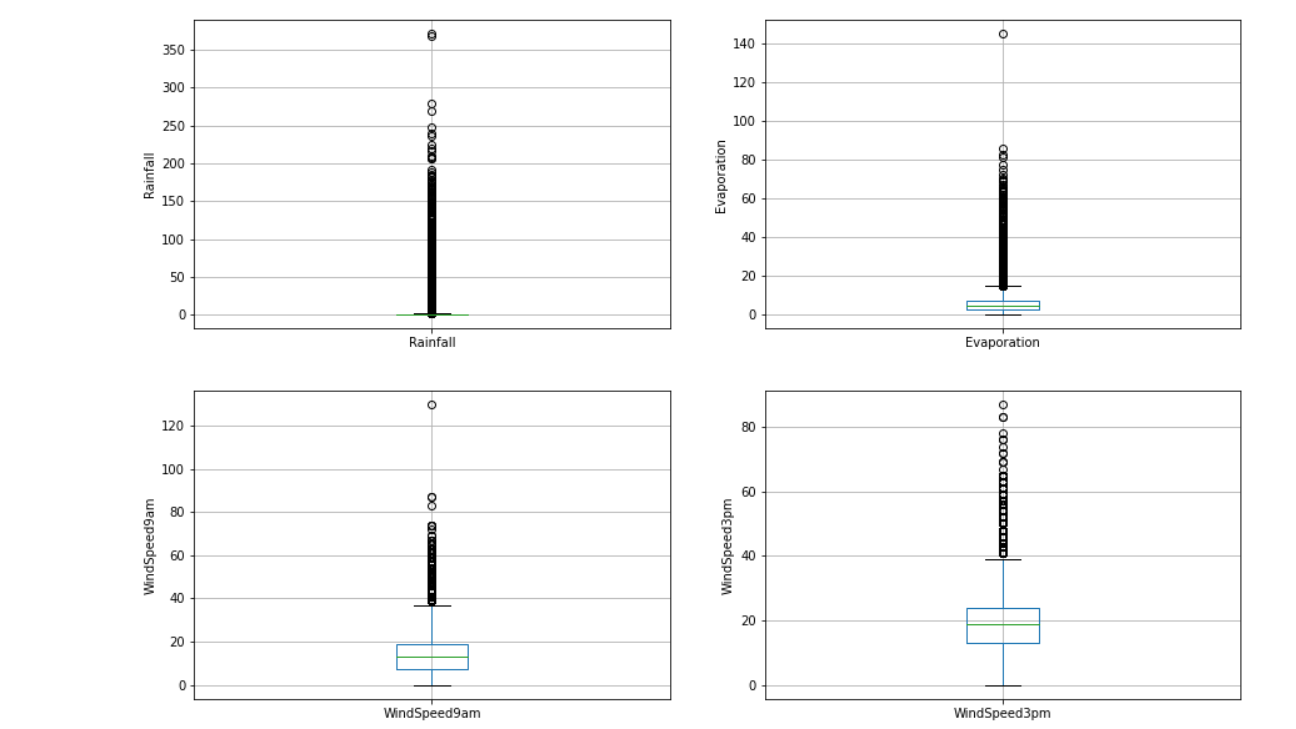


From the above correlation heat map, we can conclude that :-

* MinTemp and MaxTemp variables are highly positively correlated (correlation coefficient = 0.74).
* MinTemp and Temp3pm variables are also highly positively correlated (correlation coefficient = 0.71).
* MinTemp and Temp9am variables are strongly positively correlated (correlation coefficient = 0.90).
* MaxTemp and Temp9am variables are strongly positively correlated (correlation coefficient = 0.89).
* MaxTemp and Temp3pm variables are also strongly positively correlated (correlation coefficient = 0.98).
* WindGustSpeed and WindSpeed3pm variables are highly positively correlated (correlation coefficient = 0.69).
* Pressure9am and Pressure3pm variables are strongly positively correlated (correlation coefficient = 0.96).
* Temp9am and Temp3pm variables are strongly positively correlated (correlation coefficient = 0.86).

## 2.3.3. Outliers detection

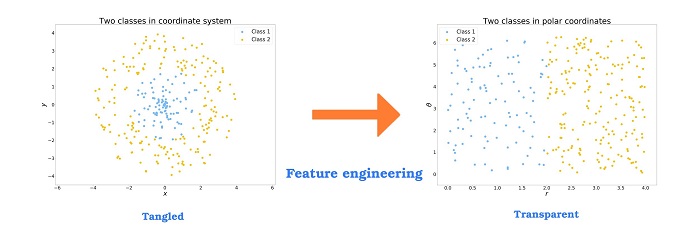
We can see the Outliers in the numerical columns with the help of boxplots. I have plotted the boxplots in numerical columns for outlier detection.



## 2.4.Feature Engineering

feature engineering makes data easier to examine. The world we live in is imperfect, therefore the data we use in it can be messy and unpredictable. Regardless of the type of data source (e.g., linear SQL database, Excel file, etc.), The data, which is normally formed as a table with a new row representing each sample and each column showing a trait, may be difficult to comprehend and handle.

To better understand our machine learning models' data and hence help them perform better, we need execute feature engineering. The task of converting data into a simpler form of comprehension belongs to feature engineering in learning algorithms. In this case, we are working to make the information clearer for a trained model, but feature can be generated such that data visualizations more approachable for non-data professionals can be created. But understanding the concept of clarity in ml algorithms is complex since the technique varies according on the type of data used.



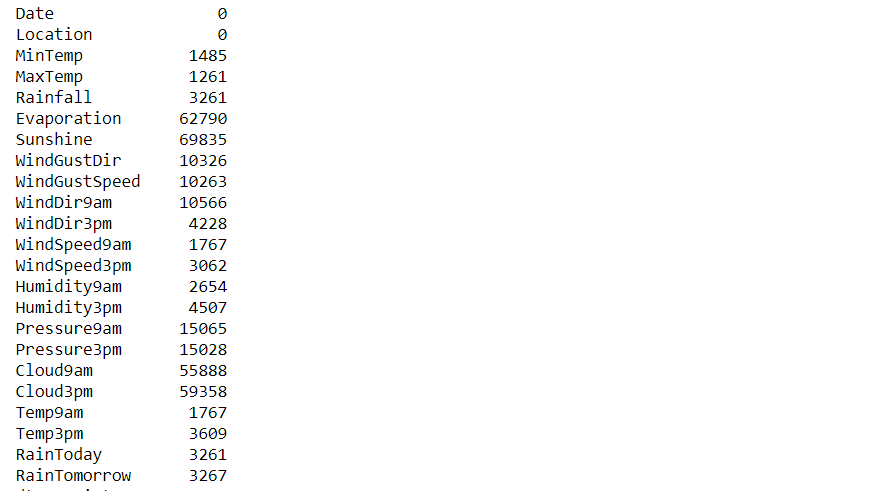
In this project feature Engineering is most important part because the dataset contains Outliers in Numerical features and also there is lots of missing values in the features and feature scaling is also require for this project because the feature ranges lies in different – different scales. In this project I have perform these following steps in feature Engineering:

## 2.4.1 Handling Missing Values

In a dataset, missing value are values that have not been recorded. Missing a full observation or only a single value from a single cell (row). Missing data can happen with both continuous (e.g. height) and categorical (e.g. gender) variables (e.g. gender of a population).Missing data can be hard to spot if you're using machine learning. Ignoring or erasing the lack of information isn't a viable option. It is imperative that they be handled delicately, as they can indicate an important matter. Two ways to handle missing data are widely used:

1. Dropping the rows containing missing values
2. It is not advisable to drop missing values because doing so reduces information.
   * Perhaps the missing value is telling us something.
   * In addition, you often must make predictions about new data even though some of the attributes are missing in the real world!. We can impute the missing values of Numerical columns with Median or Mean and for Categorical columns we can use mode of the feature.

In the given dataset there is a lots of missing values.



As we can see from the above diagram our dataset contains large number of missing values. Even Rain Tomorrow Variable which is dependent variable also contains 3267 missing values. I have impute the numerical columns with the help of median and for categorical variables I have used mode. For target variable also I have used mode because it is a categorical variable.

## 2.4.2 Outlier Treatment

The outliers may signify some type of anomaly or an experimental error. A person's age may be erroneously recorded as 200 instead of 20 years. An anomaly like that should absolutely be removed from the dataset.

But some outliers are, in fact, good. Some outliers imply data is clearly different from other measurements. It could indicate, for example, something as odd as banking fraud or maybe even a rare condition

Example : Assume the data 1, 2, 1, 6, 4, 3, 100. If these values represent the number of breads eaten in lunch, then 50 is clearly an outlier.

Significance of Outliers:

* Standard deviation and mean are both heavily influenced by outliers in the dataset. Erroneous findings are a possibility with these.
* When outlier is present, many machine learning techniques fail. To avoid getting rid of outliers, they must be discovered and eliminated.

Interquartile Range :

To quantify the inconsistency of a data collection, divide it into quarters and measure it with an IQR. The information has been divided into equal portions with ascending order. Q1, Q2, and Q3 are the names of the three numbers that define the boundaries between four equal divisions. The data is divided into quarters (four equal parts), with the 25th percentile marked as Q1.

The middle 50% of the data is found in Q2.

The 75% of the data is found in Q3.

The range between first and third quartiles is known as the Interquartile range (IQR), and it can be calculated as Q3 – Q1. The outliers are those values found to be beyond the bounds of Q1.5 IQR below Q1 or Q3 + 1.5 IQR above Q3.

These features contains Outliers in the dataset:

Rainfall, Evaporation, Windspeed9am, Windspeed3pm . I have used Interquartile range for Outlier detection and Outlier Treatment in these features.

## 2.4.3. Encoding of Categorical Features

All inputs and Output variables must be numeric for a machine learning model. To fit and evaluate a model, we have to encode your category data into numbers first. I have used One hot Encoding in this project.

One Hot Encoding

In categorical variables, it means translating those variables into binary vectors. In order to be used, these categorical data are first transformed to integer values. A binary vector, all 0s, is used to represent every integer value ( Except the integer which is marked as 1).

Example: For example, to specify a “colour”, three categories are necessary, requiring three binary variables. For the selected colour, we will set the variable to 1 (one) and 0 (zero) for the remaining colours.

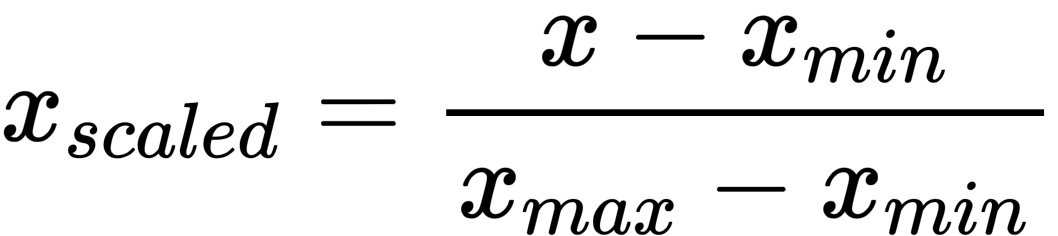


In this project I have used One hot encoding for converting the categorical features into numerical ones. These are the Categorical variables of the dataset

Location, WindGustDir, WindDir9am, WindDir3pm , RainToday .I have used One hot encoding in these features.

## 2.4.4 Feature Scaling

Feature Scaling is also very important step of Data Cleaning if our feature lies on different –different scale. It is better to apply feature Scaling. A common practice when preparing data for machine learning is normalization. To normalize, it helps to keep everything on the same scale, to keep different value ranges constant. To make the best use of machine learning, datasets do not have to be normalized.



In this project I have applied feature Scaling on all the features except dependent variable (price). It will convert all the features in between 0 to 1 range.

## 2.5. Machine Learning

In this project, I will be training a set of Machine Learning models with optimal hyper parameters to extract the high accuracy with each model. I will be experimenting with different models such as Linear Models – Logistic Regression, Support Vector Machines, Ensemble Models – Random Forest, Gradient Boosting Decision Trees, and Base Models – K-Nearest Neighbours.

Since this is a two-class classification model we are training, I will be using these models that are popularly used in the real world to experiment in this project.

## 2.5.1 Logistic Regression

Logistic Regression is a predictive model technique where, using pattern recognition, a previously recorded dataset is used to deduce the future likelihood of a certain event. One way to use the Logistic Regression algorithm is to find the independent variable(s) and dependent variable that you want to forecast, and then use the regression to make predictions.

In the presence of several explanatory variables, logistic regression is employed to obtain the odds ratio. Binomial regression has similarities to other linear regression models, except the dependent variables is binary. The effect of each variable is calculated to show the effect on the odd values of the observed event.

## Hyper Parameter Tuning in Logistic Regression

I have applied hyper parameter tuning in Logistic Regression Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting. For C=1 the Training Accuracy is 84.87% which is highest among all and for C=1 the accuracy on testing data is 84.84%.

## 2.5.2 Random Forest Algorithm

One of the most prominent algorithms for supervised learning is Random Forest, which uses an approach known as regression. ML applications can make use of it to solve classification and regression difficulties. Ensemble learning refers to the methodology of amalgamating several classifiers to deal with an elaborate problem and to better the model performance.

Its name is self-explanatory, as it uses a number of different decision trees on various subsets of the given dataset to take an average and improve the dataset's predictive accuracy. The random forest doesn't only rely on one decision tree. It instead bases its forecasts on majority voting amongst all the predictions and then predicts the ultimate result.

The below diagram explains the working of the Random Forest algorithm:



## Hyper parameter Tuning in Random Forest Classifier

After applying hyper parameter tuning in Random forest Algorithm we can see that for 100 number of estimators and for depth of 100 the algorithm gives higher accuracy The model gives around 85.59% accuracy on training data and 85.39% accuracy for testing data. So there is no overfitting in the model.

## 2.5.3 Support Vector Machine

SVM is a classification and regression learning technique that is popular because of its widespread use. Classifying is one of the most common uses of it in machine learning. The algorithm will aim to design the optimal lines or decision boundary for separating data points in multi-dimensional space, making it easy to place data points in their correct categories. This hyperplane, often known as the best choice boundary, is better than the previous one.

## Hyper parameter Tuning in SVM Classifier

I have applied hyper parameter tuning in Support Vector Machine Algorithm. For C=1 and gamma = 0.1 the algorithm gives higher accuracy. For training data the accuracy is 85.48% and for testing data the accuracy is 84.64%.

## 2.5.4 Xg- Boost

XGBoost is a variation of the ensemble learning technique. The findings of a single machine learning model may not be enough to rely on sometimes. Ensemble learning is a solution to combining the predictive potential of several learners, and it is more orderly than many other alternatives. The end result is a model that puts together all of the results from multiple models.

Base learners are generally split between multiple algorithms, though occasionally the same method will be replicated for simplicity. Two kinds of ensemble learners with a lot of traction are boosting and bagging. Decision trees are the most common statistical methodology to which these two methods have been applied.

## Hyperparameter Tuning in Xg-Boost Classifier

I have Applied Hyperparameter tuning on Xg-boost classifier for finding the best parameters. After Applying Hyperparameter tuning we can see that the best parameters are:

Base learner=50, Optimal depth=50

For these parameters the accuracy on training data is 85.25% and for testing data the accuracy is 85.21%. hence we can say that there is no overfitting in the model.

## 2.5.5 Decision Tree

Decisions trees are the most popular and efficient classifiers and predictors available. Decisions are created by dividing a group of information into "decision branches," which separate results into the appropriate categories based on attribute tests.

These are the following benefits of Decision Tree:

* The final outcome of a decision tree is usually well defined.
* A decision tree does not require significant processing for classification.
* For both continuous and categorical, decision trees can be used.
* Decision trees show the which variables to use when predicting or classifying.

## Hyperparameter Tuning in Decision Tree

I have applied Hyperparameter tuning on Decision Tree. After Appling Hyperparameter tuning we can see that

{'criterion': 'entropy', 'max\_depth': 3, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2}

These are the best parameters. For Training data the algorithm gives 83.006% accuracy and with testing data the algorithm gives 82.95% accuracy.

## 2.5.6 K-Nearest Neighbors

The Supervised Learning algorithm K-Nearest Neighbor is one of the easiest algorithms, with no prior assumptions on the properties of data. The K-NN method groups the new data/case based on resemblance to other, existing categories, assigning the new case to the category most similar to the current accessible cases. KNN method uses data from all the previous data points to compare similarities and classify new data points. The data can be categorized quickly with the help of K-NN, since the new information is easy to understand.

K-NN is best for Classification problems but may also be useful for Regression. Non-parametric algorithms like K-NN assume nothing about their data. It is sometimes referred to as a lazy learner algorithm since it takes a long time to learn from the training set, instead of doing so instantly. This method conducts an action on the dataset at the time of classification, rather than learning from of the training set immediately. When getting fresh data, KNN algorithms only classify it into a category with a similarity to the new data.

## Hyperparameter Tuning on KNN Algorithm

I have applied hyper parameter tuning in K Nearest Neighbors Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting.

For n\_neighbors=3 the algorithm gives higher accuracy with training and testing data. The accuracy with the training data is 84.87% and the accuracy with the testing data is 84.84%.

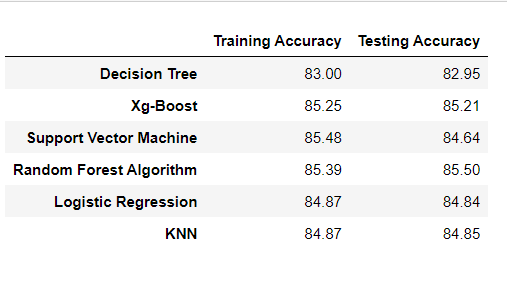
## 2.6 Performance Evaluation Metrics

While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

Accurately identifying the ratio of correctly predicted instances to the total amount of instances evaluated is a metric for measuring accuracy. While the accuracy might be adequate as a measurement for model performance, it may not be good enough because it fails to include incorrect predictions. If someone treats a fake post as a real one, it could cause a serious issue. False positives and false negative issues that accommodate for misclassification should be considered because of this.



The Accuracy of each model is :



## 3. Conclusion

Human daily activities are greatly affected by both daily happenings and major global occurrences like weather. On the one side, weather serves as a natural resource that aids in human survival. For example, weather contributes to farming as a source of water and provides wind power for energy. The devastating events that have brought hardship to so many lives have been extremely destructive, as floods and droughts ruin the lives of those in their paths. This is why people find forecasting the weather interesting.

In this paper, I have worked with different models of Machine Learning to predict the Rainfall in Australia. from the given data to identify the better performing models. Firstly I have applied Exploratory data Analysis on the given dataset to gain the insights of the dataset. After EDA I have applied feature Engineering methods because the contains large number of missing values and Outliers. So it is require to apply feature Engineering. After that I have some machine learning models (Random forest, Decision Tree, Logistic Regression, SVM) in order to gain the better accuracy.

Based on the analysis my findings from answering the two research questions:

These are some feature Engineering methods that can be used in the data to extract more useful information:

* Null Value Imputation
* Outlier detection and Removal
* One hot Encoding
* Feature Scaling

These Feature Engineering methods helps to increase the accuracy of the machine learning models. If our dataset contains null values then Machine learning models will not be able to perform well and if Our dataset contains outliers then the model will not generalized one and feature engineering is also required because distance based algorithms(like knn) with not be able to calculate the distance between the features lies between different-different scales. So it is required to use feature engineering before giving the data to machine learning model.

Random Forest and Xg-Boost gives better accuracy with testing data and also with the training data and the model is also not overfitted.

But there is still a long way to go before reliable weather forecasting can be done. Even more importantly, the weather is unstable and getting worse. because of this, better forecasts frequently needs to be updated to include them adaptations will need to be made to accommodate A rise in weather complexity. as well better weather prediction models are require a method for improving long-term forecast accuracy