|  |  |
| --- | --- |
|  |  |
| Australia Weather Prediction |  |
|  |  |
|  | DATECOURSE TITLE |
|  | STUDENT’S NAME |

Table Of Contents

[1. Problem Statement: 3](#_Toc1367555294)

[2. Introduction 3](#_Toc2004931131)

[3. Data Source 3](#_Toc875451726)

[4. Dataset Description : 3](#_Toc624628373)

[5. Importing Necessary Library: 5](#_Toc2121633728)

[6. Loading The Dataset: 5](#_Toc2103098313)

[I. Loading Dataset using pd.read\_csv function 5](#_Toc1407993943)

[II. Printing First 5 rows of the dataset using head function 5](#_Toc1570928641)

[III. Printing First 5 rows of the dataset using head function 5](#_Toc1681246139)

[7.Data Preprocessing: 6](#_Toc1627808428)

[I. Finding Null Values 6](#_Toc1477746718)

[II.Filling Null Values 7](#_Toc1070491761)

[III.Result of Data Preprosing 7](#_Toc2109877595)

[8.Data Visualization: 8](#_Toc2018469218)

[I. Plotting Different Plots 9](#_Toc657573748)

[9. Observations 11](#_Toc44614060)

[10.Creating Machine Learning Model: 12](#_Toc46849225)

[I. Splitting data into Training and Testing data 12](#_Toc1276840452)

[II. Printing shape and size of data 13](#_Toc84523495)

[10. Model Prediction Algorithms 13](#_Toc92862579)

[I. Logistic Regression 13](#_Toc733817766)

[1. Properties of Logistic regression 13](#_Toc857079085)

[2. Various kinds of logistic regression 13](#_Toc806800092)

[3. Comparison between Logistic and Linear Regression 13](#_Toc836747402)

[II. Cross-validation 14](#_Toc1968403195)

[1. Various kinds of Cross validation 14](#_Toc1354462785)

[2. Properties of Cross Validation 14](#_Toc1396045358)

[III. Random Forest 14](#_Toc1144589523)

[1. Properties of Random Forest 14](#_Toc1018198706)

[11. Comparing Different Models 15](#_Toc601185783)

[12. Summary 16](#_Toc1558909240)

[13. Conclusion 17](#_Toc903538590)

[14. References 17](#_Toc1811648329)

# 1. Problem Statement:

Design a predictive model with the use of machine learning algorithms to forecast whether or not it will rain tomorrow in Australia.

# 2. Introduction

In this Project we are going to implement a predictive model on the Rain Dataset to predict whether it will rain in Australia tomorrow. The dataset contains about 10 years of daily weather observations at various locations in Australia.

# 3. Data Source

Observations were drawn from numerous weather stations. The daily observations are available from <http://www.bom.gov.au/climate/data>.  
An example of latest weather observations in Canberra: <http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml>

Definitions adapted from <http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>Data source: <http://www.bom.gov.au/climate/dwo/> and <http://www.bom.gov.au/climate/data>.

Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

# 4. Dataset Description :

This dataset contains about 10 years of daily weather observations from numerous Australian weather stations.

RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No?  
This column is Yes if the rain for that day was 1mm or more.

Number of columns: 23

Number of rows: 145460

Number of Independent Columns: 22

Number of Dependent Column: 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Table 1. Descriptive table of the dataset.** |  |  |  |
| Feature Name | Description | Missing Values | Available Data | Type |
| Date | Day on which the measurement is  carried out | 0 | 142,193 | string/date |
| Location | Station location name meteorological. | 0 | 142,193 | string |
| MinTemp | Minimum temperature in degrees Celsius. | 637 | 141,556 | float |
| MaxTemp | Maximum temperature in  degrees Celsius. | 322 | 141,871 | float |
| Rainfall | Amount of rain recorded during the day  in mm. | 1406 | 140,787 | float |
| Evaporation | “Class A pan evaporation” (mm) in 24 h  until 9 a.m. | 60,843 | 81,350 | float |
| Sunshine | Number of hours of radiant sun during  the day. | 67,816 | 74,377 | float |
| WindGustDir | Direction of the strongest wind gust in the  24 h to midnight. | 9330 | 132,863 | string |
| WindGustSpeed | Speed (km/h) of the strongest wind gust  in the 24 h to midnight. | 9270 | 132,923 | float |
| WindDir9am | Wind direction at 9 a.m. | 10,013 | 132,180 | string |
| WindDir3pm | Wind direction at 3 p.m. | 3778 | 138,415 | string |
| WindSpeed9am | Average wind speed (km/h) in the 10 min  before 9 a.m. | 1348 | 140,845 | float |
| WindSpeed3pm | Average wind speed (km/h) in the 10 min  before 3 p.m. | 2630 | 139,563 | float |
| Humidity9am | Humidity (%) at 9 a.m. | 1774 | 140,419 | float |
| Humidity3pm | Humidity (%) at 3 p.m. | 3610 | 138,583 | float |
| Pressure9am | Atmospheric pressure (hpa) at the level of  evil, at 9 a.m. | 14,014 | 128,179 | float |
| Pressure3pm | Atmospheric pressure (hpa) at the level of  evil, at 3 p.m. | 13,981 | 128,212 | float |
|  | Fraction of sky obscured by clouds at |  |  |  |
|  | 9 a.m. The unit of measurement is “oktas”, |  |  |  |
|  | which is equal to a unit of eighths. It |  |  |  |
|  | refers to how many eighths of the sky are |  |  |  |
| Cloud9am | obscured by clouds. A value of 0 indicates | 53,657 | 88,536 | float |
|  | a completely clear sky, while a value of |  |  |  |
|  | 8 indicates that it is completely obscured. |  |  |  |
|  | Temperature at 3 p.m., in degrees Celsius. |  |  |  |
| Cloud3pm | Fraction of sky obscured by clouds at  3 p.m. The unit of measurements is the  same as in Cloud9am measurements. | 57,094 | 85,099 | float |
| Temp9am | Temperature at 9 a.m., in degrees Celsius. | 904 | 141,289 | float |
| Temp3pm | Temperature at 3 p.m., in degrees Celsius. | 2726 | 139,467 | float |
| RainToday | Boolean: 1 if precipitation exceeds 1 mm  in the 24 h to 9 a.m., if not 0. | 1406 | 140,787 | string |
| RISK\_MM | The amount of rain for the next day  in mm. | 0 | 142,193 | float |
|  | Variable created from variable RISK\_MM. |  |  |  |
| RainTomorrow |  | 0 | 142,193 | String |

# 5. Importing Necessary Library:

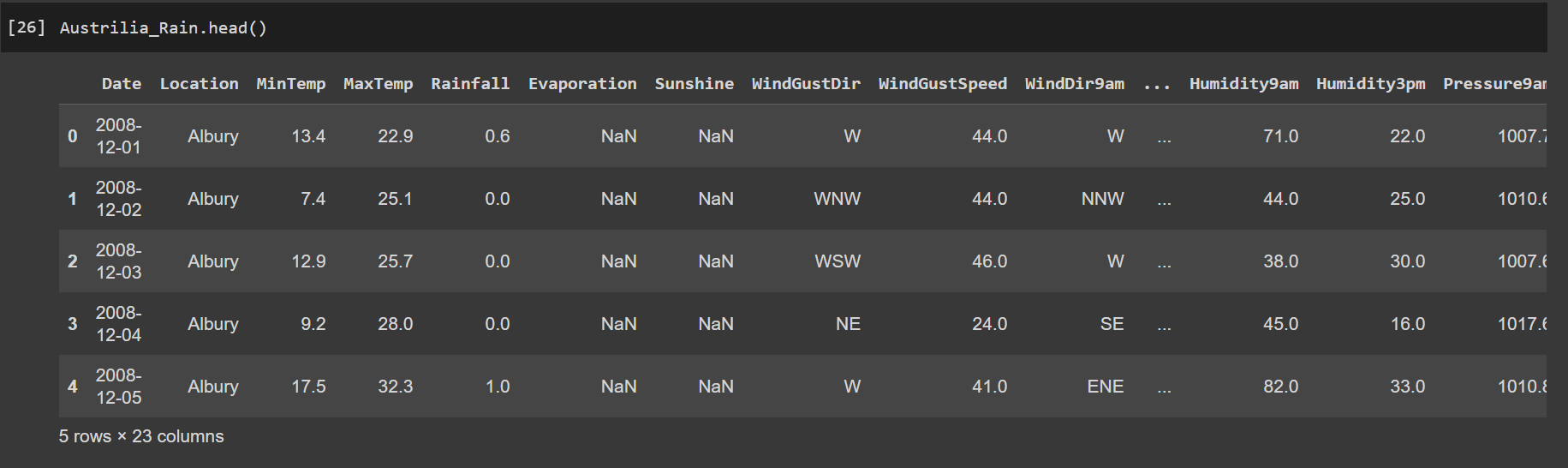


# 6. Loading The Dataset:

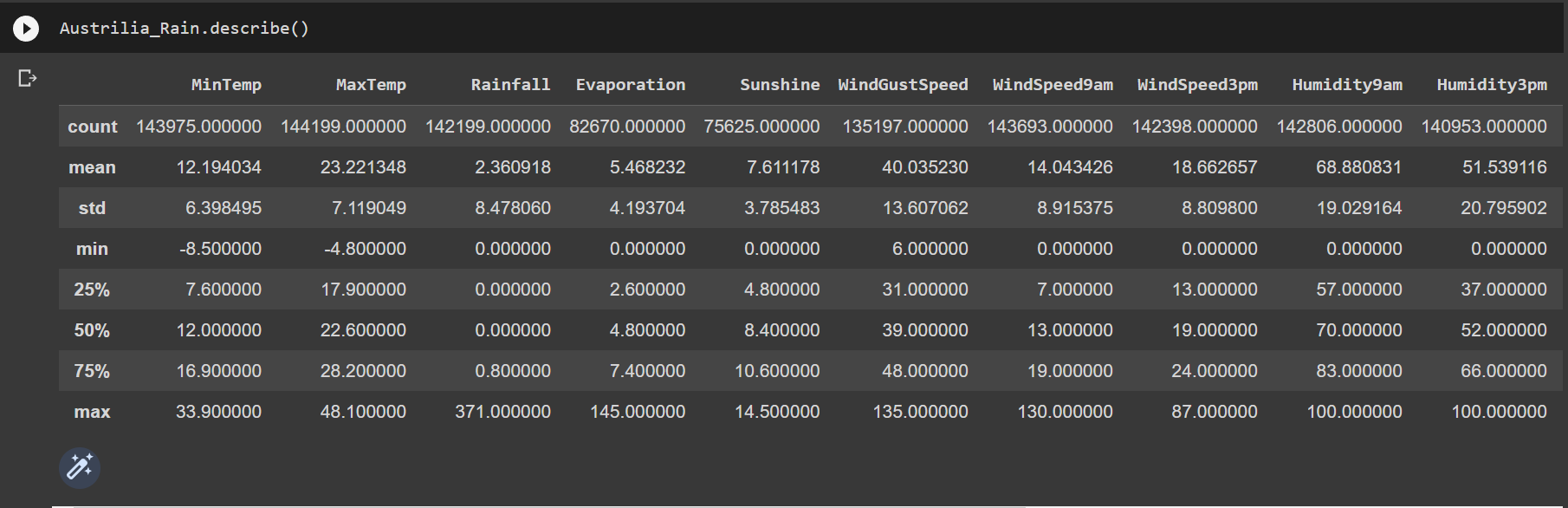
### Loading Dataset using pd.read\_csv function

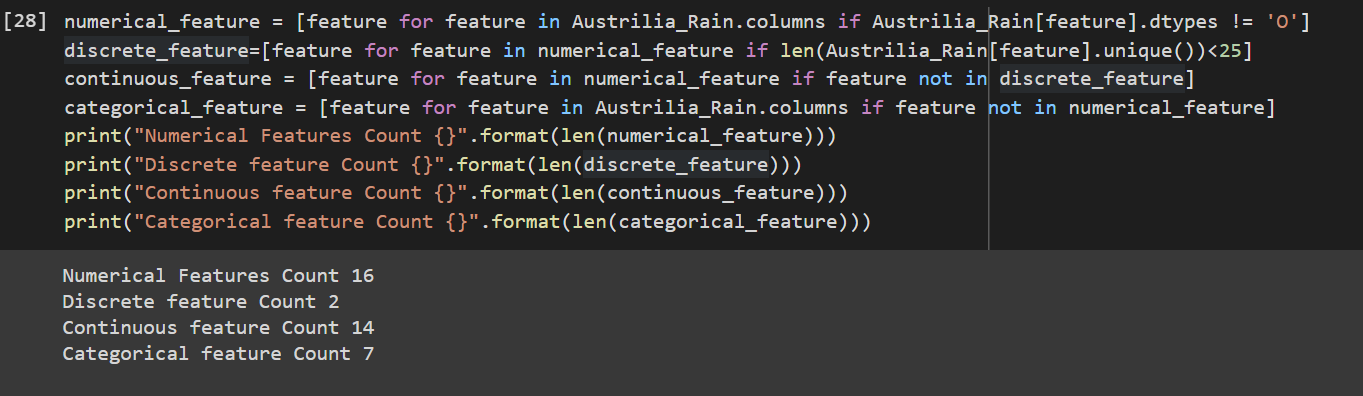


### II. Printing First 5 rows of the dataset using head function



### III. Printing First 5 rows of the dataset using head function



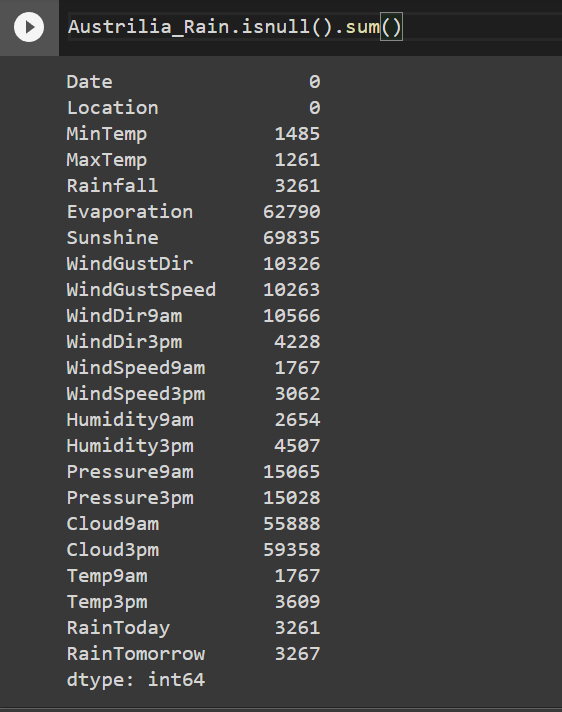


# 7.Data Preprocessing:

**Missing Data:**

Below Figure represents the number of samples of each of the variables for which there are no data. Thus, the total missing data correspond to 10% of the analyzed data (considering the total of 140,787 samples × 22 variables). Thus, if the samples that do not have data in any of their variables are eliminated, then approximately 50% of the samples would have to be eliminated.

### I. Finding Null Values



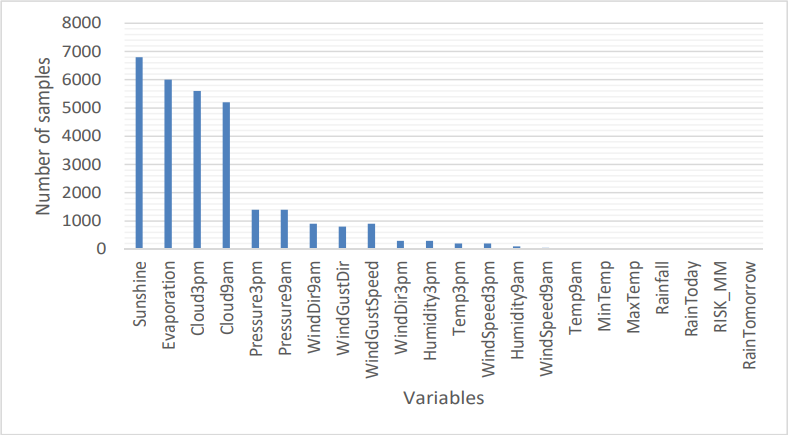
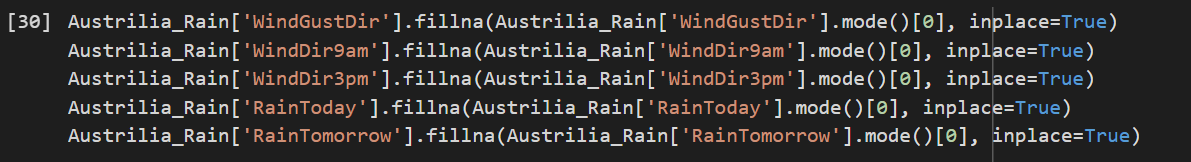
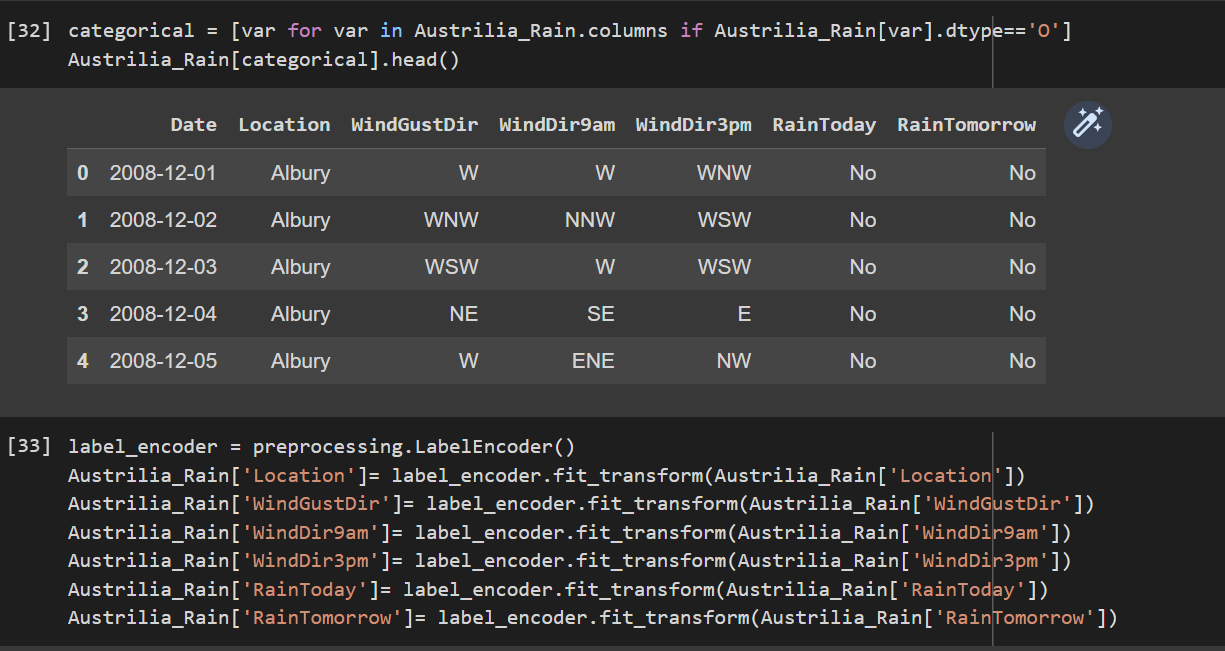


Figure : Number of variables with value “NA”.

### II. Filling Null Values

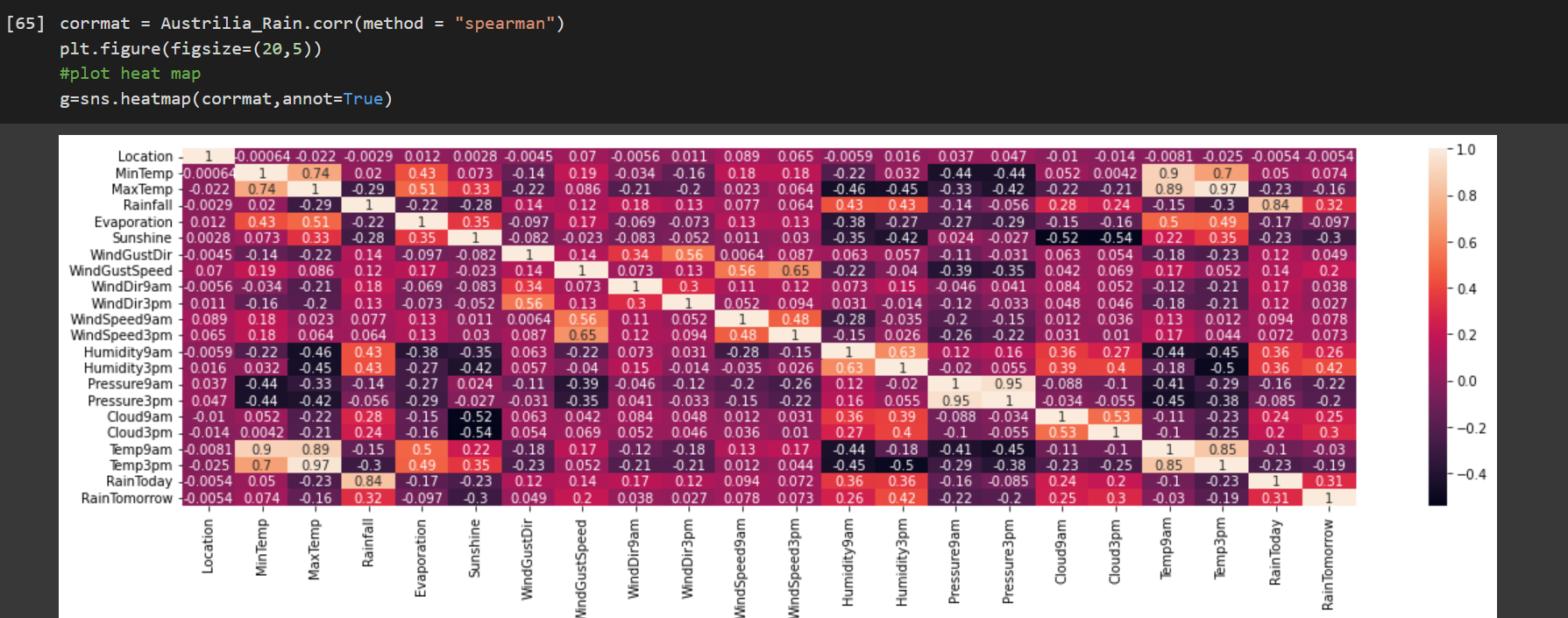
### III. Result of Data Preposing

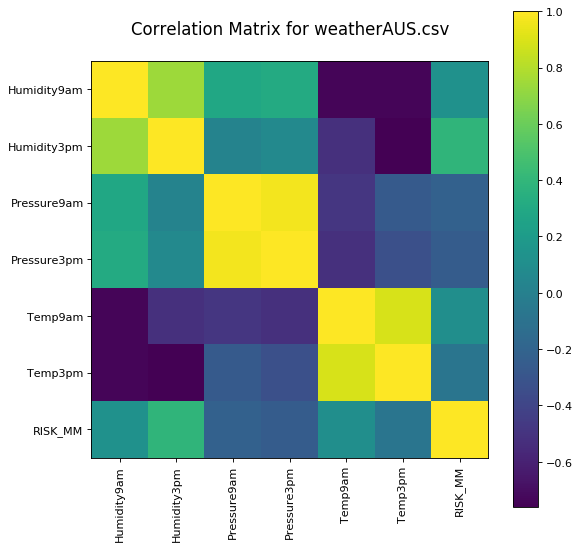
The following variables were eliminated:

* Date (eliminated because they are string variables)
* Rainfall (eliminated because they are highly related to the RainToday variable)
* RISK\_MM (artificial variable obtained to predict the rain)
* the variables WindGustDir, WindDir9am and WindDir3pm (each is split into two variables containing the cosines and sines of the wind direction angles).

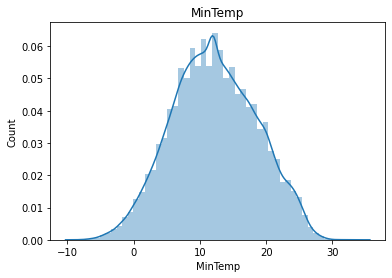
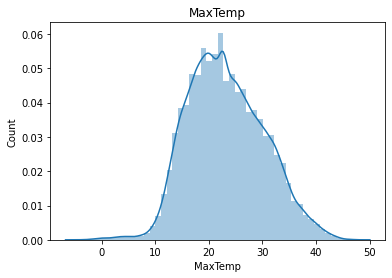
The samples that had “NA” were replaced by their mode.

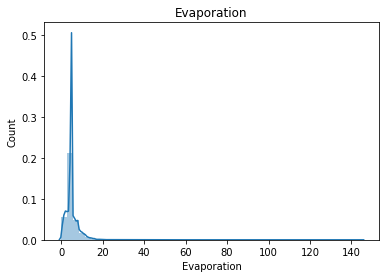
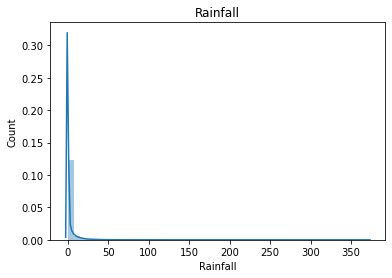
# 8.Data Visualization:

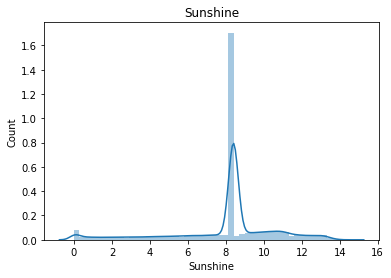
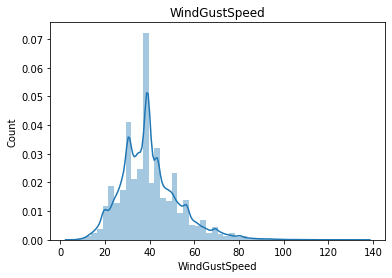


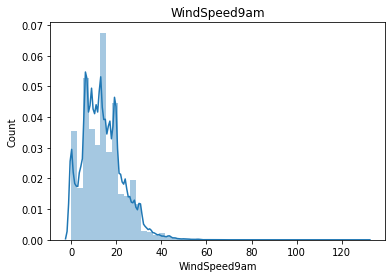
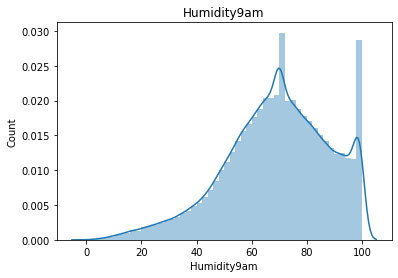


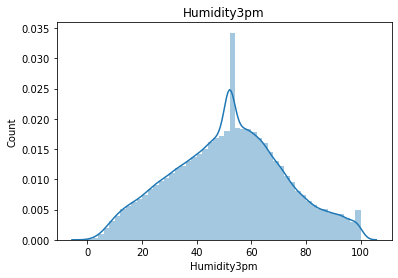
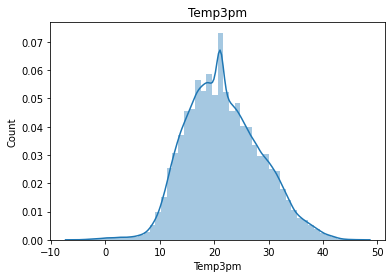
## I. Plotting Different Plots



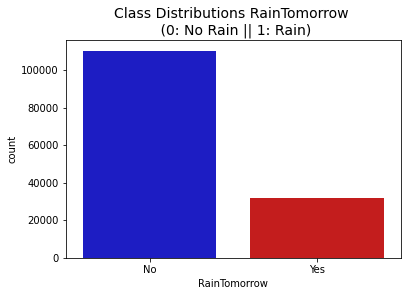


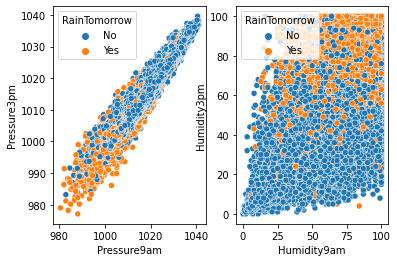
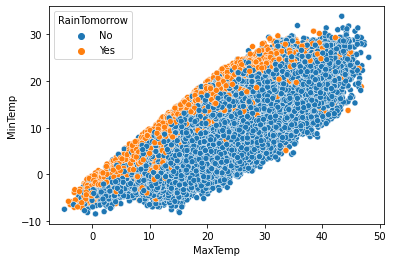


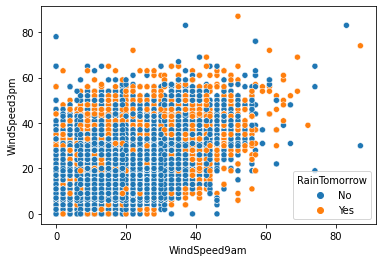




## 9. Observations





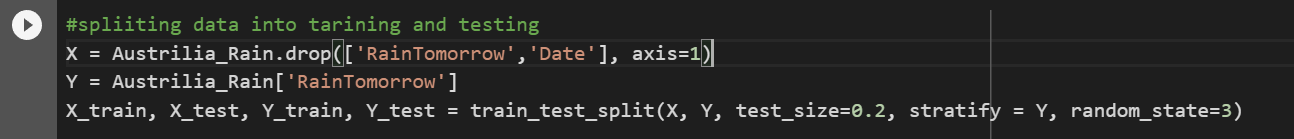


There are high chance to rain tomorrow when: -

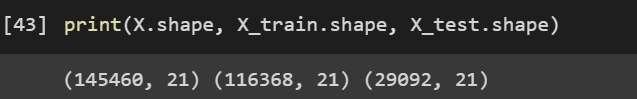
* it rained today
* the presure is low
* the humidity is high
* wind speed is high
* Max temp and min temp are close to each other

# 10.Creating Machine Learning Model:

### I. Splitting data into Training and Testing data



### II. Printing shape and size of data



# 10. Model Prediction Algorithms

## I. Logistic Regression

One of the most straightforward and widely used Machine Learning techniques for two-class classification is logistic regression. It may serve as the starting point for any binary classification issue and is simple to apply. Its foundational ideas are helpful for deep learning as well. The link between one dependent binary variable and independent variables is described and estimated through logistic regression.

A statistical technique for forecasting binary classes is logistic regression. The result or goal variable has a binary nature. There are just two conceivable classifications when something is dichotomous. It may be used, for instance, to issues with cancer detection. It determines the likelihood that an event will occur. When the target variable is categorical, linear regression is used in a specific way. A log of the odds is used as the dependent variable. Using a logit function, logistic regression makes predictions about the likelihood that a binary event will occur.

### 1. Properties of Logistic regression

* Logistic regression's characteristics include the dependent variable's conformance to the Bernoulli distribution.
* A maximum likelihood approach is used for estimation.
* Model fitness is determined using Concordance and KS-Statistics; there is no R Square.

### 2. Various kinds of logistic regression

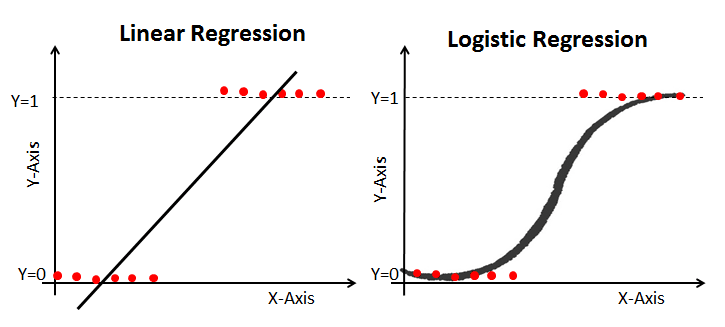
Binary logistic regression :The target variable in a binary logistic regression has just two potential outcomes, such as spam or no spam, cancer or no cancer.

Multinomial Logistic Regression: This method is used to predict the kind of wine when the target variable contains three or more nominal categories.

Ordinal Logistic Regression: The target variable, such as a restaurant or product rating from 1 to 5, has three or more ordinal categories.

### 3. Comparison between Logistic and Linear Regression

While logistic regression produces a constant result, linear regression produces a continuous output. The prices of homes and stocks are two examples of continual output. Examples of discrete output include determining if a patient has cancer or not and forecasting customer attrition. Ordinary Least Squares (OLS) is used to estimate linear regression, whereas Maximum Likelihood Estimation (MLE) is used to estimate logistic regression.



## II. Cross-validation

If you're computing R squared on your test set, the R squared returned is dependent on the way that you split up the data! The data points in the test set may have some peculiarities that mean the R squared computed on it is not representative of the model's ability to generalize to unseen data. To combat this dependence on what is essentially an arbitrary split, we use a technique called cross-validation.

### 1. Various kinds of Cross validation

K Fold Cross Validation : K-fold When the dataset is divided into K folds, cross-validation is performed to assess the model's performance when presented with fresh data.

Stratified K-Fold Cross Validation: An improvement on the cross-validation method used for classification issues is the stratified k fold cross-validation. The class ratio stays the same across the K folds as it did in the first dataset.

Leave P Cross Validation: If there are n data points in the original sample, then n-p samples are used to train the model, and p points are used as the validation set. This method excludes p data points from the training data. The error is averaged over all trials to determine overall efficacy after being carried out for all combinations where the original sample may be divided in this manner.

### 2. Properties of Cross Validation

* A powerful tool is cross-validation.
* We are able to use our data more effectively and learn considerably more about the effectiveness of our algorithm thanks to it.
* It might be simple to overlook something and apply the same data across multiple pipeline phases in sophisticated machine learning models.

## III. Random Forest

A supervised learning algorithm is random forests. Both classification and regression may be done with it. Additionally, it is the most user-friendly and adaptable algorithm. There are trees in a forest. A forest is supposed to be stronger the more trees it has. On randomly chosen data samples, random forests generate decision trees, obtain predictions from each tree, and vote for the best option. Additionally, it offers a fairly accurate measure of the feature's relevance.

### 1. Properties of Random Forest

Because so many decision trees are involved in the random forest approach, it is thought to be extremely precise and reliable.

It is not affected by the overfitting issue. The biases are eliminated mostly because it takes the average of all the forecasts.

The technique may be applied to issues involving classification and regression.

Missing values can also be handled using random forests. These can be dealt with in one of two ways: by computing the proximity-weighted average of missing values or by substituting continuous variables with their median values.

You may obtain the relative feature relevance, which is useful when choosing the characteristics that will aid the classifier the most.

# 11. Comparing Different Models

1. The challenge of model selection

Any machine learning algorithm or model has a number of characteristics that utilise the data in various ways. The data that is provided to these algorithms is frequently altered based on the results of earlier phases of the experiment. However, as developers and machine learning teams frequently document their trials, there is a wealth of data accessible for comparison.

Understanding which characteristics, data, and metadata must be taken into account to make the final decision is the difficult part. It's the age-old contradiction of having an absurd quantity of information but no understanding.

Even more difficult, we need to determine if a high value for a parameter, such as a higher metric score, genuinely indicates that the model is superior to one with a lower score, or if it is only due to statistical bias or misdirected effort.

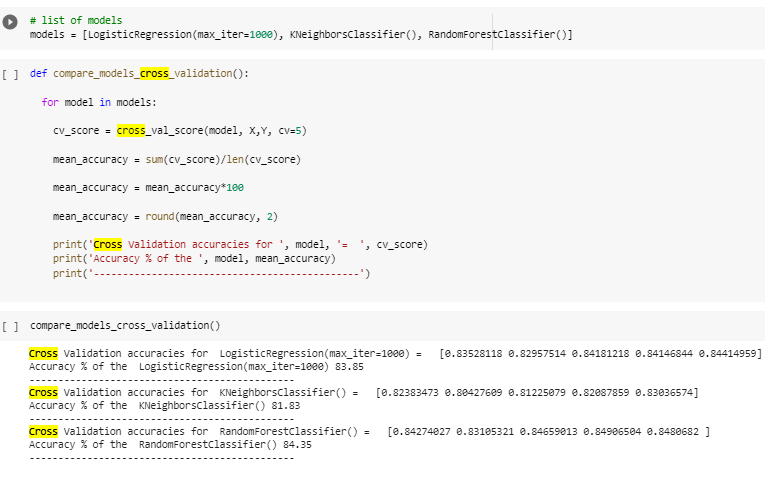
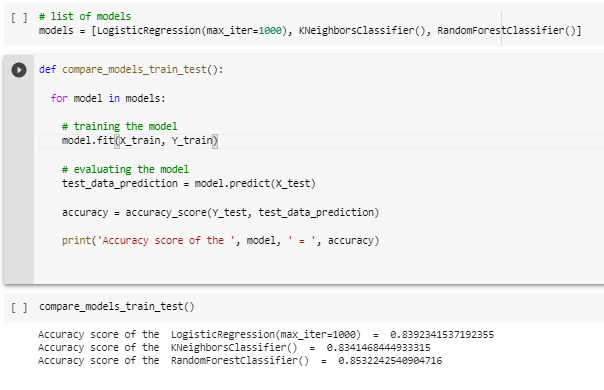
1. The goal of comparing machine learning algorithms

**Better Performance**:- Better performance of the machine learning software or solution is unquestionably the main goal of model comparison and selection. The goal is to focus on the algorithms that are optimal for the data and the business needs.

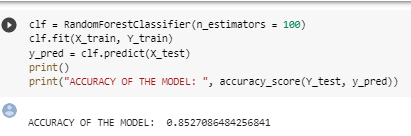
**Easier Retraining**:- Minute details and metadata are recorded when models are reviewed and prepared for comparisons, and they are useful during retraining. For instance, if a developer can clearly recall the factors that went into selecting a model, the reasons for model failure will be obvious right away, and retraining can begin with the same speed.

**Speedy production**:- With the model information at hand, it is simple to focus on models that can provide fast processing and make efficient use of memory resources. A number of factors are needed to customise the machine learning solutions throughout production as well. Having production-level data can be important for readily aligning with the production engineers. Additionally, it will be simpler to assess alternative algorithms' compliance and viability in relation to the organization's assigned resources if you are aware of their resource requirements.

**Longer Life**:- If the chosen model fails to interpret unknown input and is tightly associated with the training data, high performance may be short-lived. Finding a model that comprehends underlying data patterns is crucial in order to ensure that predictions are accurate over time and that little retraining is required.

1. 

Random Forest Highest Accuracy Score So we will use it to classify points.



# 12. Summary

* Extraction data from dataset and converting into pandas dataframe.
* Print the shape and size of the converted dataframe.
* Describe various features of dataset using describe function.
* Clean and tranform the dataframe using various machine learning library like sklearn,numpy,pandas.
* Plot various plots to understand relationship between various features of dataset.
* After applying all the pre-processing steps in our dataset we have to create our machine learning column.
* Split data into training and testing data.
* Choose an approprite machine learning model using K-fold cross validation Algorithm (k-fold cross validation is a procedure used to estimate the skill of the model on new data).
* Identify the potential features in our dataset.
* The model that perform best in hyperparameter tunning will be chosen as our machine learning model.

# 13. Conclusion

In this research, it was examined whether machine learning techniques could be used to solve the issue of rainfall forecasting in the particular situation of Australia. This kind of technique has been used in earlier research with various kinds of monthly, annual, and other time period datasets in locations other than Australia. The regions of Victoria and Sydney were the study venues. The potential benefits of using machine learning techniques as tools to replace traditional rain forecasting techniques (they also have some advantages over classical forecasting methods, such as the possibility of estimating the reliability of the results using the Indicators, Performance Key, or the possibility of adjusting the performance of the algorithms by manipulating their input parameters, which allows them to be adapted to particular cases).

# 14. References

[1]. V. Rao and J. Sachdev, "A machine learning approach to classify news articles based on location", 2017 International Conference on Intelligent Sustainable Systems (ICISS), pp. 863-867, 2017.

[2]. V. Kumar and S. Minz, "Poem classification using machine learning approach", Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), pp. 675-682, December 28–30, 2012.

[3]. B.Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques" in Language Processing (EMNLP), Philadelphia, pp. 79-86, July 2002.

[4]. B. Stehman, "Selecting and interpreting measures of thematic classification accuracy", Remote Sensing of Environment, 1997.