**Rainfall Prediction - Weather Forecasting**

**Problem Statement:**

**Weather forecasting** is the application of science and technology to predict the **conditions of the atmosphere** for a given **location**and **time**. **Weather forecasts**are made by collecting **quantitative data**about the **current state of the atmosphere** at a given place and using meteorology to project how the atmosphere will change.

Rain Dataset is to predict whether or not it will rain tomorrow.

**Data Analysis:**

* First import all necessary libraries
* There are 145460 rows and 23 columns
* . Columns names are

Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm', 'RainToday', 'RainTomorrow'], dtype='object')

### Types of variables

In this section, I segregate the dataset into categorical and numerical variables. There is a mixture of categorical and numerical variables in the dataset. Categorical variables have data type object. Numerical variables have data type float64.

Summary of categorical variables

* There is a date variable. It is denoted by Date column.
* There are 6 categorical variables. These are given by Location, WindGustDir, WindDir9am, WindDir3pm, RainToday and RainTomorrow.
* There are two binary categorical variables - RainToday and RainTomorrow.
* RainTomorrow is the target variable.

Finding missing values of categorical value

* There are only 4 categorical variables in the dataset which contains missing values. These are WindGustDir, WindDir9am, WindDir3pm and RainToday

Summary of numerical variables

* There are 16 numerical variables.
* These are given by MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am and Temp3pm.
* All of the numerical variables are of continuous type.

### Number of labels: cardinality

The number of labels within a categorical variable is known as **cardinality**. A high number of labels within a variable is known as **high cardinality**. High cardinality may pose some serious problems in the machine learning model. So, I will check for high cardinality.

* The data type of Date variable is object. I will parse the date currently coded as object into datetime format..
* Drop date column fom the dataset
* let's do One Hot Encoding of WindDir3pm variable
* get k-1 dummy variables after One Hot Encoding
* also add an additional dummy variable to indicate there was missing data
* preview the dataset with head() method
* We need to find number of missing values then fill null values for both categorical and numerical
* I assume that the data are missing completely at random. There are two methods which can be used to impute missing values. One is mean or median imputation and other one is random sample imputation. When there are outliers in the dataset, we should use median imputation. So, I will use median imputation because median imputation is robust to outliers.
* I will impute missing values with the appropriate statistical measures of the data, in this case median. Imputation should be done over the training set, and then propagated to the test set. It means that the statistical measures to be used to fill missing values both in train and test set, should be extracted from the train set only. This is to avoid overfitting
* The Rainfall, Evaporation, WindSpeed9am and WindSpeed3pm columns contain outliers. I will use top-coding approach to cap maximum values and remove outliers from the above variables.
* The outliers in Rainfall, Evaporation, WindSpeed9am and WindSpeed3pm columns are capped.
* **Building Machine Learning Models**

# ****Logistic Regression****

* When we come across a new classification problem, the first algorithm that may come across their mind is **Logistic Regression**. It is a supervised learning classification algorithm which is used to predict observations to a discrete set of classes. Practically, it is used to classify observations into different categories. Hence, its output is discrete in nature. **Logistic Regression** is also called **Logit Regression**. It is one of the most simple, straightforward and versatile classification algorithms which is used to solve classification problems.
* Split the datas into input and output, input is all the columns except Rain Tomorrow and ouput is RainTomorrow column

X = df.drop(['RainTomorrow'], axis=1)

y = df['RainTomorrow']

### predict\_proba method

**predict\_proba** method gives the probabilities for the target variable(0 and 1) in this case, in array form.0 is for probability of no rain and 1 is for probability of rain.

* **y\_test** are the true class labels and **y\_pred\_test** are the predicted class labels in the test-set.
* , the model accuracy is 0.8501. But, we cannot say that our model is very good based on the above accuracy. We must compare it with the **null accuracy**. Null accuracy is the accuracy that could be achieved by always predicting the most frequent class.

A confusion matrix is a tool for summarizing the performance of a classification algorithm. A confusion matrix will give us a clear picture of classification model performance and the types of errors produced by the model. It gives us a summary of correct and incorrect predictions broken down by each category. The summary is represented in a tabular form.

Four types of outcomes are possible while evaluating a classification model performance. These four outcomes are described below:-

\*\*True Positives (TP)\*\* – True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class

\*\*True Negatives (TN)\*\* – True Negatives occur when we predict an observation does not belong to a certain class and the observation actually does not belong to that

\*\*False Positives (FP)\*\* – False Positives occur when we predict an observation belongs to a certain class but the observation actually does not belong to that class. This type of error is called \*\*Type I error

\*\*False Negatives (FN)\*\* – False Negatives occur when we predict an observation does not belong to a certain class but the observation actually belongs to that class. This is a very serious error and it is called \*\*Type II error.\*\*

The confusion matrix shows 20892 + 3285 = 24177 correct predictions and 3087 + 1175 = 4262 incorrect predictions.

In this case, we have

* True Positives (Actual Positive:1 and Predict Positive:1) - 20892
* True Negatives (Actual Negative:0 and Predict Negative:0) - 3285
* False Positives (Actual Negative:0 but Predict Positive:1) - 1175 (Type I error)
* False Nega;tives (Actual Positive:1 but Predict Negative:0) - 3087 (Type II error)

**Conclusion:**

I have implemented the five models Logistic Regression, Decision tree, SVM, random forest,KNN ., plotted roc curve for both train and test data

Among these five models random forest model gives better accuracy of 93 percent and train score and test score lso same