**Article**

**Rainfall Prediction: Weather Forecasting**

**This article contains the following sub-topics:**

**Problem definition**

**Data analysis**

**EDA concluding remark**

**Pre-processing pipeline**

**Building machine learning models**

**Concluding remarks**

**Problem definition:**

**Rainfall prediction is the prediction of rain on the basis of the previous records collected and the weather conditions present. Weather forecasting is the analyzing or estimating of the weather conditions so that it is easy to determine when the rainfall will be there. In other words, the weather or climate conditions are different from the normal climate conditions when there is possibility of rainfall in the environment therefore to do the prediction of rain is done by analyzing the weather conditions present before the rainfall in the environment. The data collected of the climate temperature before rainfall, wind speed, humidity, clouds, the previous records, etc. plays a major role in the prediction of rain therefore the data collected should be accurate. With the help of climate conditions and the previous data, the prediction of how much the rainfall will be there can also be done.**

**The rainfall prediction is necessary because when the rainy season comes there are some places where the amount of rainfall is very high and it is necessary to predict the rainfall by the government so that the places can be put in alert. There are so many incidents in which people got safe by the predicting of rainfall as the amount of rainfall give occurrence to flood, landslide, other natural disasters. The estimating of weather condition helps in the predicting of rainfall.**

**The problem statement is by estimating the weather conditions in the environment and analyzing the previous data to know whether there is possibility of rainfall at a place on the given time or not and also predicting the amount of rainfall to be there.**

**Data analysis:**

**The data has 8425 rows and 23 columns. The 23 columns are divided into 2 dependent and 21 independent variables according to the problem statement. The dependent columns of the data are to be predicted with the help of independent variables. The independent columns of data contain necessary information about the weather conditions and collected data of previous records. The data is big as the rows are so many. The feature columns of the data are: Date, Location, MinTemp, MaxTemp, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainToday, and the target columns of data are: Rainfall, RainTomorrow. The dataset is mostly float datatype as the most of the columns are float. The data has the null values present in the columns of both float and object datatype. The target column which is raintomorrow has the class imbalance. The data has the information about the change of climate condition in the different days.**

**EDA concluding remark:**

**The dataset is going under the process of EDA so that the data can be studied and necessary information can be analyzed. To perform EDA, there are important libraries that to be imported to perform EDA. The dataset has so many columns so each column can be seen. The first 5 rows and last 5 rows of the data are to be seen. The shape of the dataset is 8425 rows and 23 columns. The columns are of the dataset are to be studied. The datatype of the columns of the dataset are shown as there are more float datatype as compared to object datatype. The info of the dataset contains every information about the columns such as datatype, non-null values count, columns name, entries, no of datatype, etc. Finding out the null values if present in the dataset, sum of the null values. The null values present in the dataset are shown with the help of heatmap for better understanding. The columns which have null values present in it and are of float datatype, being replaced by mean of that column so that accuracy can be maintained. The datatype of columns which are object has also null values but they can’t be removed as the accuracy of dataset will be reduced and it can’t be replaced by the mean of the column as it is object datatype. The dataset has the columns of object datatype too those columns are to be converted into numeric value with the help of label encoder so that it is easy for machine to learn and the accuracy of data can be maintained. Checking the correlation between independent columns and dependent columns to see that which column is how much correlated with the target column. plotting the correlation values in the heatmap for better understanding of correlation between the columns. The statistical view of the dataset is shown for knowing the mean, median, count, standard deviation, minimum and maximum value, etc. To find out the outliers present in the dataset if any, the dataset is visualized with boxplot. The dataset has outliers present in it so to remove the outliers, using the zscore method. Checking skewness of the columns of the dataset and also plotting the values of the columns in the kdeplot and histplot for better visualizing of the skewness and also for the visualizing of the dataset. The skewness present in the columns is: Evaporation, Sunshine, WindSpeed9am, are removed with the help of power transform technique.**

**Pre-processing pipeline:**

**The pipeline used in the first prediction are making the dataset scaler in the format for making easy to learn by machine, it is done with the help of standard scaler. The dataset is divided into feature columns and target columns which represents as x and y respectively .**

**In second prediction, before performing pipeline, the dataset is divided into x and y variable which represents the feature columns and the target column respectively. The pipeline for preprocessing used are smote technique and standard scaler. The target column has the class imbalance so that is why Smote technique is used to balance the dataset and for scaling the dataset, the standard scaler is used.**

**Building machine learning models:**

**After performing the EDA, the data is splitting for training and testing for machine learning models. In the first prediction, the machine learning models that are used to achieve better score and best prediction that are: linear regression, lasso, ridge, support vector regressor, random forest regressor.**

**In second prediction, the machine learning models used for the dataset to achieve better prediction and best accuracy score are: logistic regression, decision tree classifier, kneighbors classifier, support vector classifier, cross validation score for kneigbors classifier, random forest classifier, adaboost classifier, gradient boosting classifier.**

**Concluding remarks:**

**In the first prediction, the conclusion is that the best model is random forest regressor which is achieving the score of 94%. The best fit line under random forest regressor is also shown.**

**In the second prediction, the best model that stands is the random forest classifier which is showing the accuracy score of 92.6%. The roc plot is also shown under random forest classifier.**

**According to me, the random forest regressor and random forest classifier is the best machine learning model for the machine to learn the data and thus provide the best prediction of the target column after learning the data.**

**After doing the studies about the dataset I come up to the conclusion that the machine learning model helps the government to predict that which place is in the danger of high volume of rainfall. I came to know that the weather forecasting also helps as it tells the climate condition of the place by estimating it on the previous records running on the previous days, how the climate changes, wind speed, humidity in the environment and also direction of the wind, etc. All these estimation helps the prediction of rainfall and also the amount of rainfall. There are places which are affected by the heavy rainfall, these places are put in alert about rainfall only by estimating the weather conditions going on the day to day.**

**Author’s name: Mohit Meel**