PREDICTING RAINFALL

Because of climate change across the globe, untimely rainfall has become a major issue. This unnatural rainfall comes with many problems for farmers and their crops, postponing exams and events , floods, drainage block problems in cities , diseases etc. So it become necessary for predicting the rainfall for regions.

If we make a model which can predict the rainfall it can help people to be ready for the possible situation and to make a plan to deal with it before.

Aim of the project: The aim of the project is to develop a machine learning model for a real data of rainfall , project can be used to predict the rainfall and we can try to minimize errors in prediction.

We will compare different machine learning models and will identify the best method to predict the rainfall.

Stage 1: Data collection and Pre processing :- Here we are using real time data and this data has been retrieved from the Australia metrological department website.

This data set contains columns having Min Temp , Max Temp , Rainfall , Windspeed , Wind direction , Humidity , pressure etc.

But in this obtained data there are some values are missing or NA value , so for such case we will do data pre processing step and will remove the data set having any column with value NA. Then we will apply our algorithms for the remaining data sets. We will also encoding data into numericals wherever needed so that our model can become efficient.

The original data set before pre processing step contains 145460 rows and 23 columns.

We will use drop method to drop missing values, after processing our data set has 112925 rows and 17 columns. Now we will apply our operations on this filtered dataset.

Stage 2: Applying machine learning algorithm to data set :-

Applying relevant ML algorithms to predict the output. To create the confusion matrix and find the accuracy of algorithm for the given data set. We have applied Logistic Regression , Decision Tree and Random Forest Classifier. The algorithms are applied in Python applications using numpy , pandas, matpotlib libraries. There are 3 algorithms that we have used for this project.

1.) Logistic Regression Algorithm

2.) Decision Tree classifier

3.) Random Forest Classifier

AT FIRST WE WILL LOOK INTO BRIEF THEORY OF THESE ALGORITHMS

(A.) LOGISTIC REGRESSION :- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

(B) DECISION TREE CLASSIFIER:- Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.
The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data).
In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

(C) RANDOM FOREST CLASSIFIER :- Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. **A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.**

And now we will implement all these 3 algorithms one by one and will compare the result of accuracy

CONCLUSION :-

So after applying all 3 algorithms on the given data set, we have found that accuracy of the **Random Forest Classifier** Algorithm is found maximum out of all algorithms. So it is the best algorithm for the given data set.