BLOGS/ ARTICLES DETAILS

PROJECT NAME: RAINFALL WEATHER PREDICTION

AUTHOR: *SHIKHA KAMAL*

1. Problem Definition.

This system will predict rainfall forecasting based on different parameters that is included in datasets. The datasets is being executed from given site in jupyter notepad and then analysing it.

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

b) Design a predictive model with the use of machine learning algorithms to **predict how much rainfall could be there**.

2. Data Analysis.

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.

The datasets contains about 10 years of daily weather observations of different locations in Australia.

Number of columns = 23

Number of rows = 8425

Number of columns contains different factors included to predict and help in forecasting the rainfall today and tomorrow. The columns with their description is discussed below:

Date - The date of observation

Location -The common name of the location of the weather station

MinTemp -The minimum temperature in degrees celsius

MaxTemp -The maximum temperature in degrees celsius

Rainfall -The amount of rainfall recorded for the day in mm

Evaporation -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.

WindGustDi r- 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

WindSpeed9am -Wind speed (km/hr) averaged over 10 minutes prior to 9am

WindSpeed3pm -Wind speed (km/hr) averaged over 10 minutes prior to 3pm

Humidity9am -Humidity (percent) at 9am

Humidity3pm -Humidity (percent) at 3pm

Pressure9am -Atmospheric pressure (hpa) reduced to mean sea level at 9am

Pressure3pm -Atmospheric pressure (hpa) reduced to mean sea level at 3pm

Cloud9am - Fraction of sky obscured by cloud at 9am.

Cloud3pm -Fraction of sky obscured by cloud

Temp9am-Temperature (degrees C) at 9am

Temp3pm -Temperature (degrees C) at 3pm

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

RainTomorrow -The amount of next day rain in mm. Used to create response variable . A kind of measure of the "risk".

- The info function added to give the overall view of datasets by executing the number of rows, columns, null values, data types of each columns and memory usage information.

- Shape function indicate the shape or size of data sets with executing the number of rows and columns in braces . the first number is indexing to number of rows present while second number indexing to number of columns present .

- Columns function executing all the columns name as an output

- Dtypes function executing the data types of each columns which indicating the how the variable value stored (ie in numeric form or string form) . it is executing the results and showing the data types of float type and object type.

- Isnull function executing to verify whether the null values present in the data sets or not . Here showing many null values present in data sets and we need to fill that values by having mean and mode method which may effect the slight changes in accuracy score.

- Describe function executing the statistics of data sets which calculated the different values function type like mean, maximum values, minimum values, different stages percentile , standard deviation.

- Unique function execute the variety of outputs presents and indexing whether the column variable is continuous data type or categorical data type.

3. EDA Concluding Remark.

After executing the datasets basic analysis is being done like, checking for head, tail, sample, info, columns, isnull, shape, dtypes, unique, nunique.

The describe function used which include the statistics of datasets including different calculations like mean, standard deviation, different percentile percentage and maximum and minimum values also counted. The include function included in describe function execute the string data type with their count and frequency of first row data value.

The describe function includes variation in mean value and standard deviation , which show the spreading of data values in normal distribution curve.

The variations between 75 % percentile and max values also contains presence of outliers which lie outside the normal distribution curve. These outliers need to be removed from datasets to have accurate score or prediction.

The correlation calculate the correlated values which show how it includes or playing major role in resulting target variable. These values contains the negative and positive values. Positive values prefer the direct relation whereas negative values relates the inverse relation with each other.

Different plotting graphs included like univariant(taken only one column at a time), bivariant(taking 2 columns and comparing them), and trivariant /multivariant which compare more than 2 columns at a time.

In project we used the count plot (univariant) and scatter plot(bi-variant) and pair-plot also included.

- The count plot is being plotted to have the view and helps in counting the values so that we can compare the values which category in particular columns is dominating and playing the major role for target variable.

- While taking only one columns at a time , called as uni – variant type whereas while taking two columns at a time , called as bi – variant and helps in comparing the column whether they are dependent or independent.

- If dependent columns are high in count then we need to check the VIF (Variance inflation factor) . vif indicating and calculating the dependency factor , and which holding the higher factors of dependency and comparing to other columns, then we may remove any one columns to exclude the dependency factor from our data sets.

- Scatter plot is being used and taking the two columns through which we can compare the data points and check for density. Higher density region indicate the major role

for prediction for target variable. And if data points are in different or distant from density region then it indicate the outliers points.

- Different plotting can be used by using different graphs like box plot, distribution plot, violin plot, histogram plot, bar plot, regression plot, etc by our convenience
- Box plot is plotted to check for outliers present in data sets and if present then we need to remove it by z- score method. The z score range is containing the thresholds value which lies in less than 3 (zscore < 3) value.
- Distribution plot checks for the skewness factor , as it show the spreading of data in distribution curve, and for better accuracy it should be in range of -3 to +3 along x-axis.
- Scatter plot show the linear positive relation, linear negative relation, no relation as points lie overall plotting range.
- Pairplot is also plotted which makes all combinations of columns possible and make different plots with their convenience or predicting the columns relation.

4. Pre-Processing Pipeline.

Pre-processing pipeline is a basic analysis that being done while executing the data sets. Observing certain things so that we can proceed in a right path for accuracy.

The data types of columns is float and object type which contains numeric value as well as string value.

The datasets contains the object type /string type values which converted into numeric code by providing Label Encoder.

Null values also presents in datasets which we need to fill it while using mode (frequency) method and simple imputer .

Checking the correlated values, and analyse that which playing major role and which one is playing minor role, which to be excluded or included.

Plotting done for outliers analysis and hence remove it by putting z-score method.

Plotting is also done for skewness checking , and to have them in normal –distribution range by improving it by power transform method.

- Label encoder function used converts all string type data value into numeric code type which is easily understand by python.
- Different encoder is also there and used but most commonly used encoder is Label encoder
- Simple imputer used to fill nan or null values present in data sets. It can fill null values by having the mean of columns value or by having mode calculated.
- There is also different imputers used by our choice and preferences. Mean method, mode method, imputer type, simple imputer type, etc.

5. Building Machine Learning Models.

Splitting of datasets into x and y , where x holding all columns except target variable whereas y holding only target variable.

First prediction: To predict the rain tomorrow as it categorical data type, we run all classifier models.

- Logistic regression
- Gaussian NB
- Decision tree classifier
- Random forest classifier

- Kneighbors classifier

Steps:

- Find the maximum random state to find the best accuracy state to run the model
- Run all models in one loop , trained them and find their accuracy score, confusion matrix, classification report.
- In training and testing best model is random forest classifier , which best percentage is 90.27 %.
- Cross validify each model with their mean accuracy calculated from cross validation steps. During cross validation best model is again random forest whose score % coming 84.72 %.
- Tuning the parameter for best fit model using grid search cv. Find the best parameters of grid search and finally putting their best parameters in best fit model selected.
- Again find the accuracy score, confusion matrix , classification report
- Finally plot the roc auc curve and also find out the roc auc score. Roc auc score is 70.94 %

Second prediction: To predict the how much rainfall can be there? This contains the quantitative data type , we run all regressor models.

- Linear regression
- Decision tree classifier
- Random forest classifier
- SVR (support vector machines)
- Kneighbors regressor

Steps:

- Find the random state again to find the best accuracy state to run the model

- Train test split method applied, and then run all models in one loop and find their r2 score
- In training and testing, best model is random forest regressor and contains the maximum score of 66.15 %
- Cross validation done of each model and find their cross validation score by having the mean of cross validation score
- The best model in cross validation is random forest and contains the best score among all models of 58.32 %
- Tuning the best parameter of best fitted model(i.e random forest regressor) and fit the model
- Find the best parameters and best parameters is fit into best fitted model and then predicting the r2 score

6. Concluding Remarks.

We can change the few conditions to get different values percentages, by changing the cross validation stages , or we can also change the random state conditions. By altering the different conditions we can predict different score and accuracy stage.

The first prediction accuracy score is 85.77 % and best fitted model is random forest classifier

The second prediction r2 score is coming 65.24 % and best fitted model is random forest regressor.

The prediction score in both steps can be improved and make changes and save it.