**RAINFALL PREDICTION**

**DATE OF SUBMISSION- 08-01-2022**

**VISUALISATION OF RAINFALL IN INDIA AND BIHAR RAINFALL ANALYSIS**

**MODEL 1**

**1.1 Motivation and description:**

Monsoon prediction is clearly of great importance for India.Two types of rainfall predictions we have done are- i) predict rainfall of a particular month of a year given the data of predecessive three months. ii) predict the rainfall of particular year with the help of data of predecessive three years.

Indian meteorological department provides forecasting data required for project. In this project we are planning to work on long term predictions of rainfall. The main motive of the project is to predict the amount of rainfall in a particular division or state well in advance. We predict the amount of rainfall usingpast data.

**1.2 Dataset**

Dataset(dataset) – This dataset has average rainfall for every year from 1901-2015 for each state.

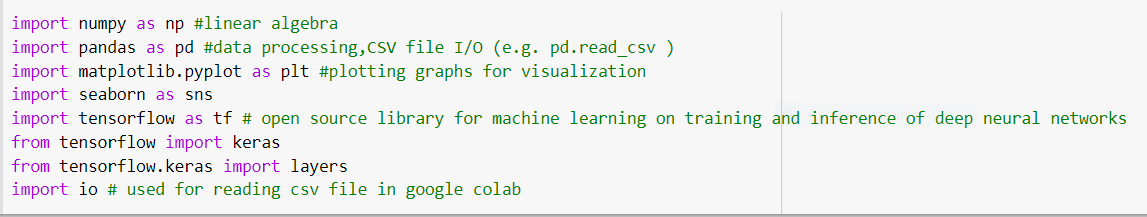
<https://data.gov.in/resources/subdivision-wise-rainfall-and-its-departure-1901-2015>

**1.3 Methodology**

• Converting data in to the correct format to conduct experiments.

• Make a good analysis of data and observe variation in the patterns of rainfall.

• Finally, we try to predict the average rainfall by separating data into training and testing. We apply various statistical and machine learning approaches(SVM, etc) in prediction and make analysis over various approaches. By using various approaches we try to minimize the error.



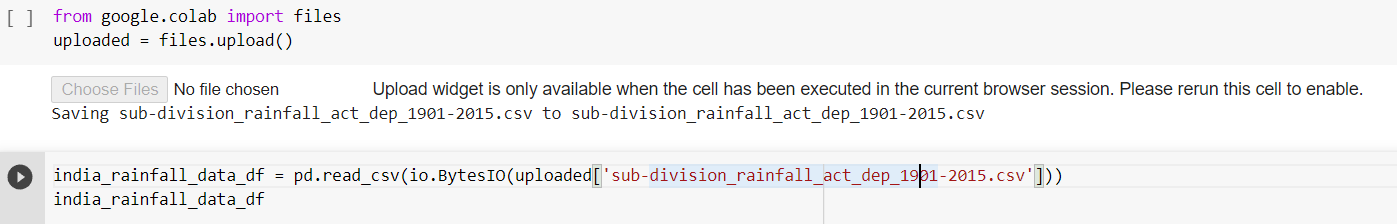
**1.4 Types of Graphs**

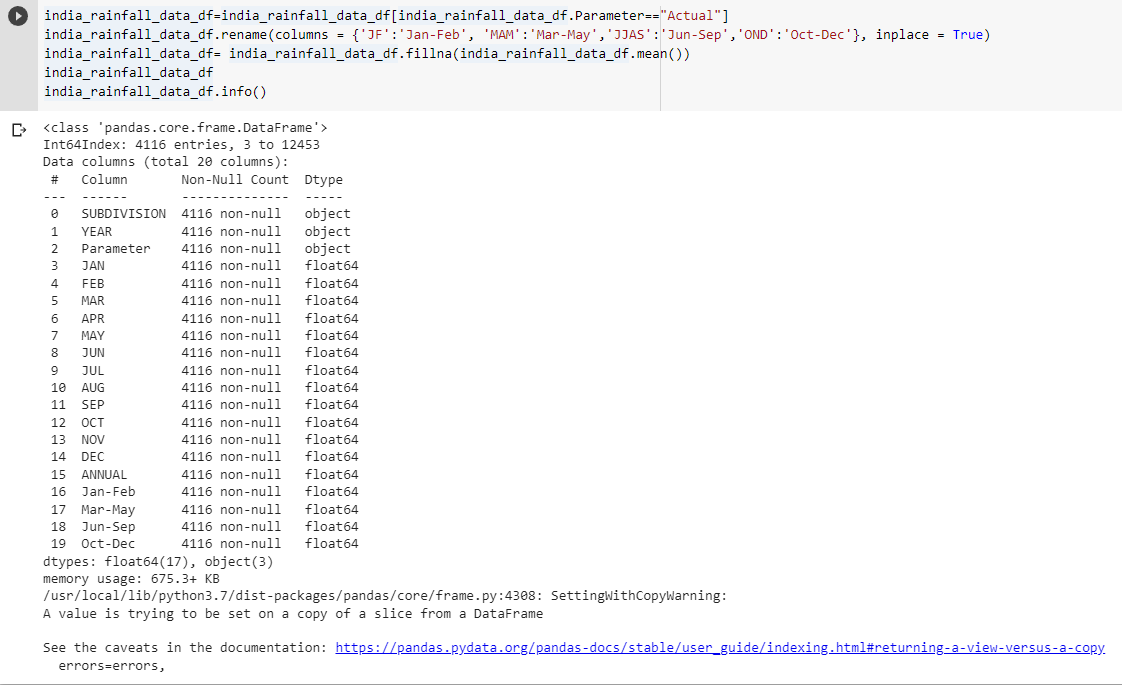
• Bar graphs showing distribution of amount of rainfall.

• Distribution of amount of rainfall yearly, monthly, groups of months.

• Distribution of rainfall in subdivisions, districts form each month, groups of months.

• Heat maps showing correlation between amount of rainfall between months.



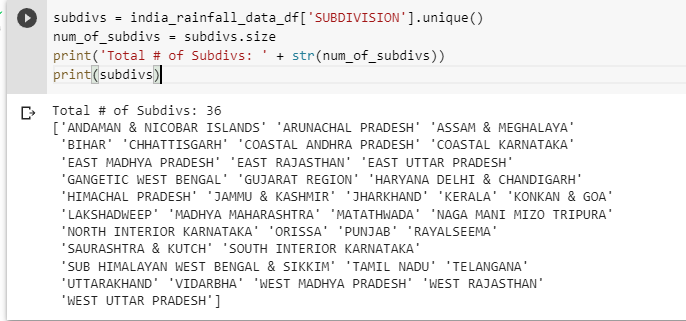


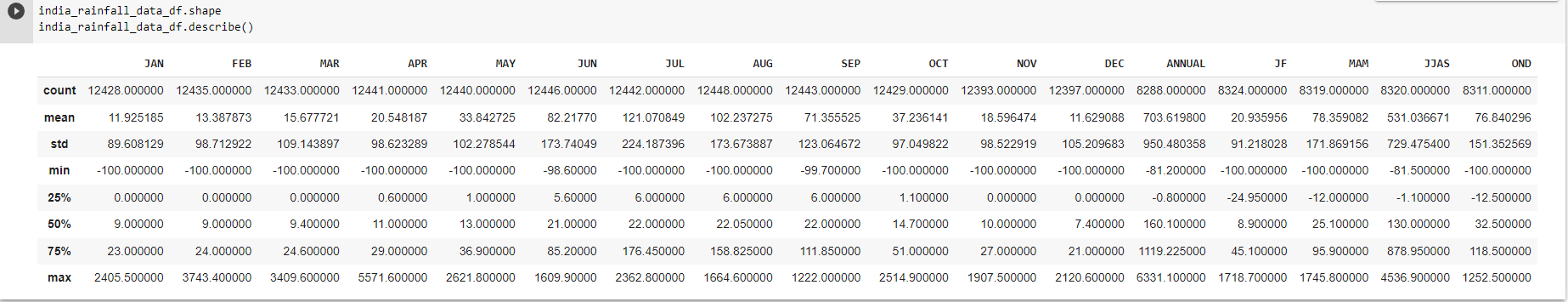
**1.5 Dataset Description**

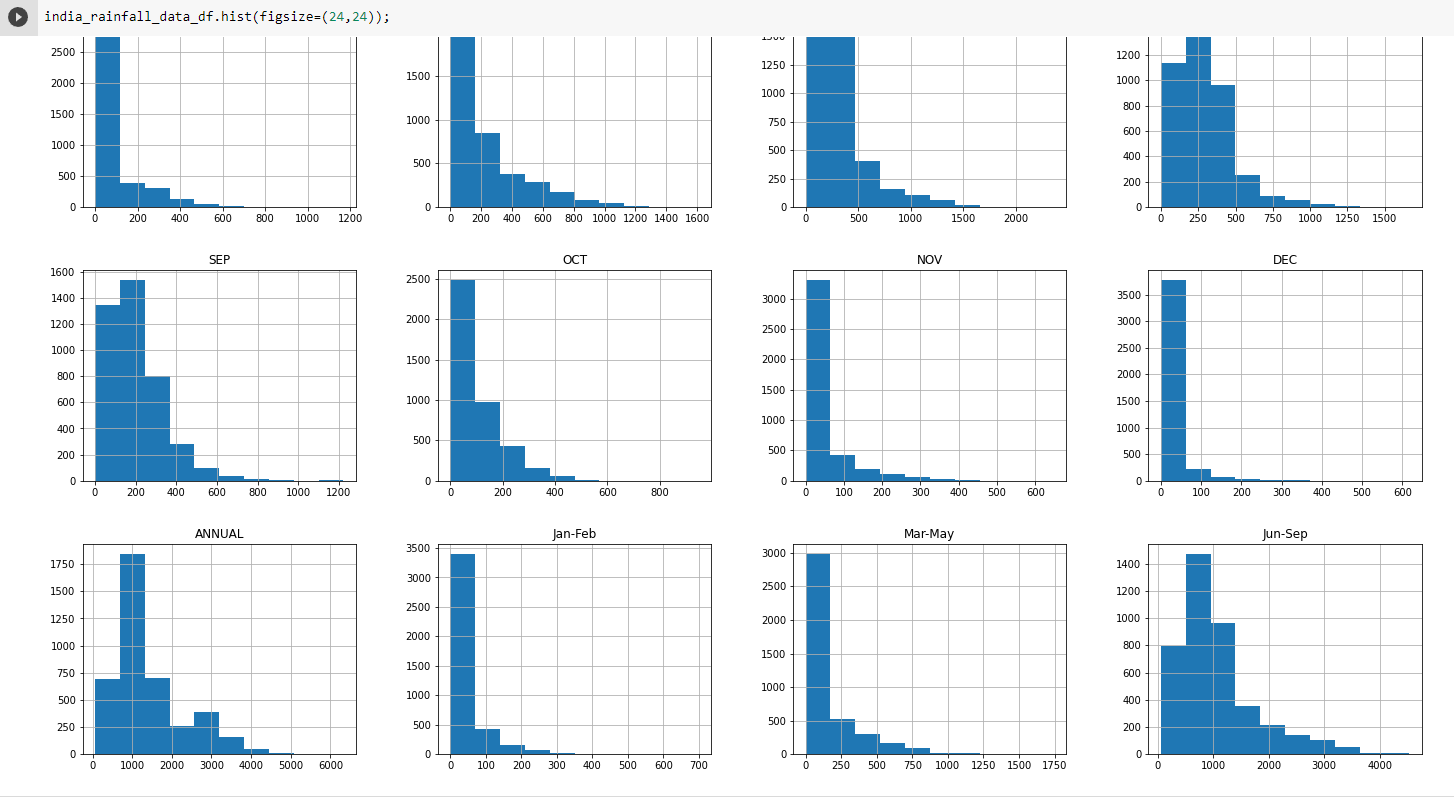
• Data has 36 sub divisions and 19 attributes (individual months, annual, combinations of 3 consecutive months).

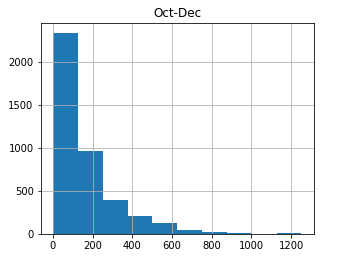
• For some of the subdivisions data is from 1950 to 2015.

• All the attributes has the sum of amount of rainfall in mm.





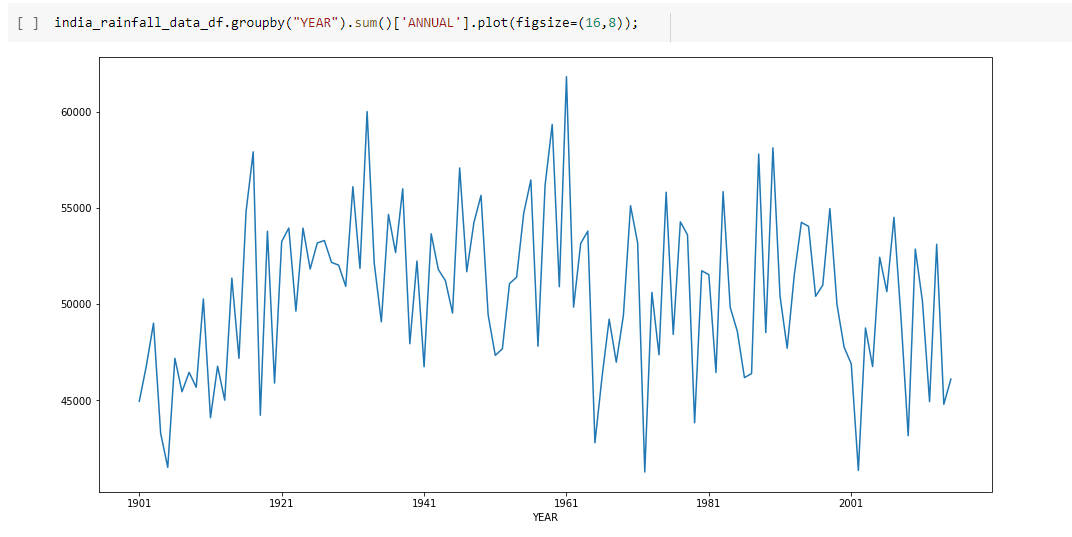




**1.6 Observation**

• Above histograms show the distribution of rainfall over months.

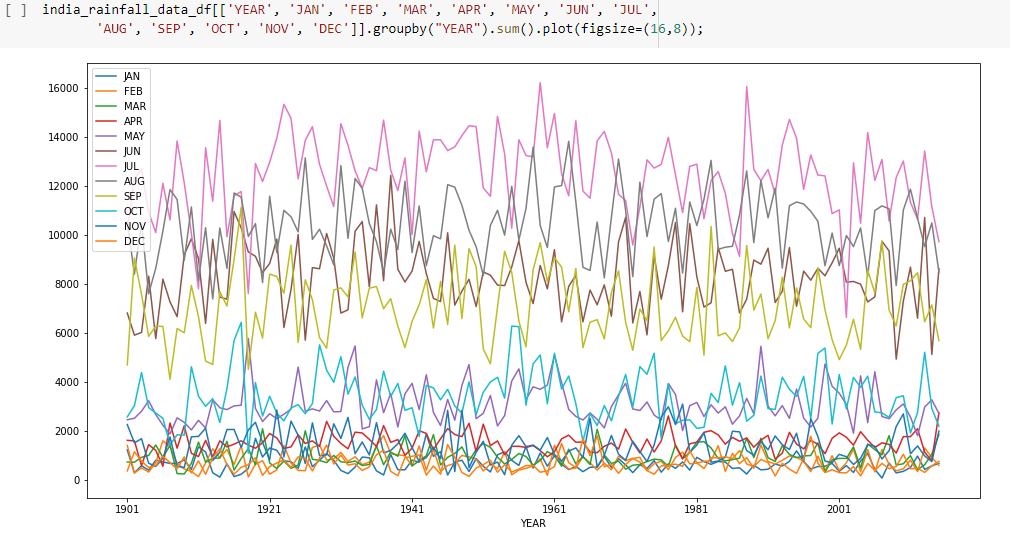
• Observed increase in amount of rainfall over months July, August, September.

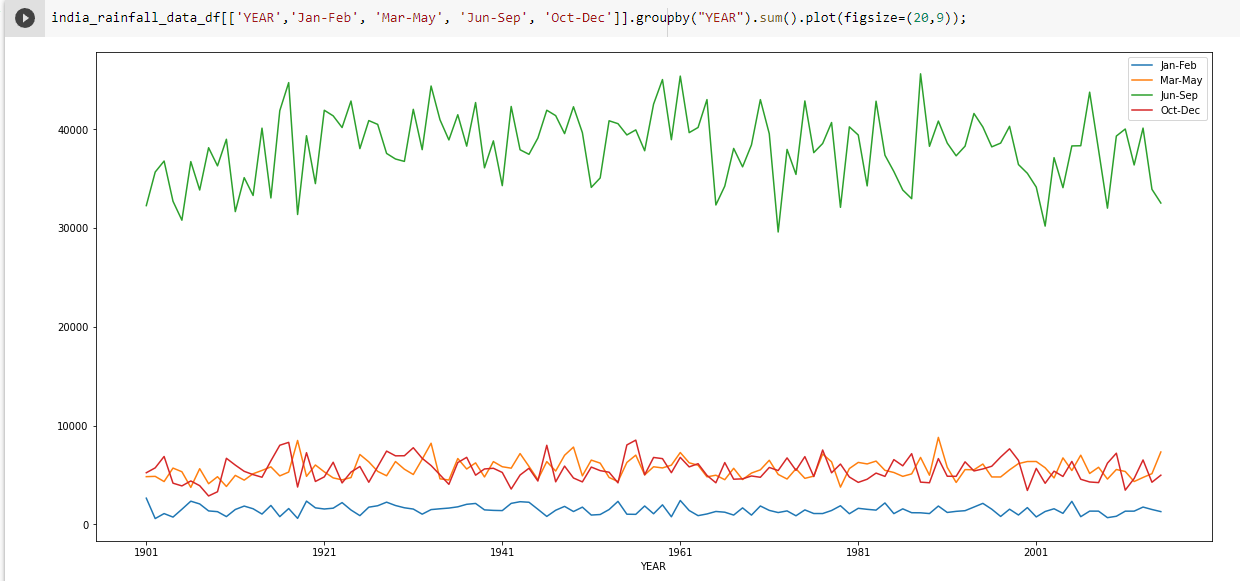


**1.7 Observation**

• Shows distribution of rainfall over years.

• Observed high amount of rainfall in 1950s to 1961**.**

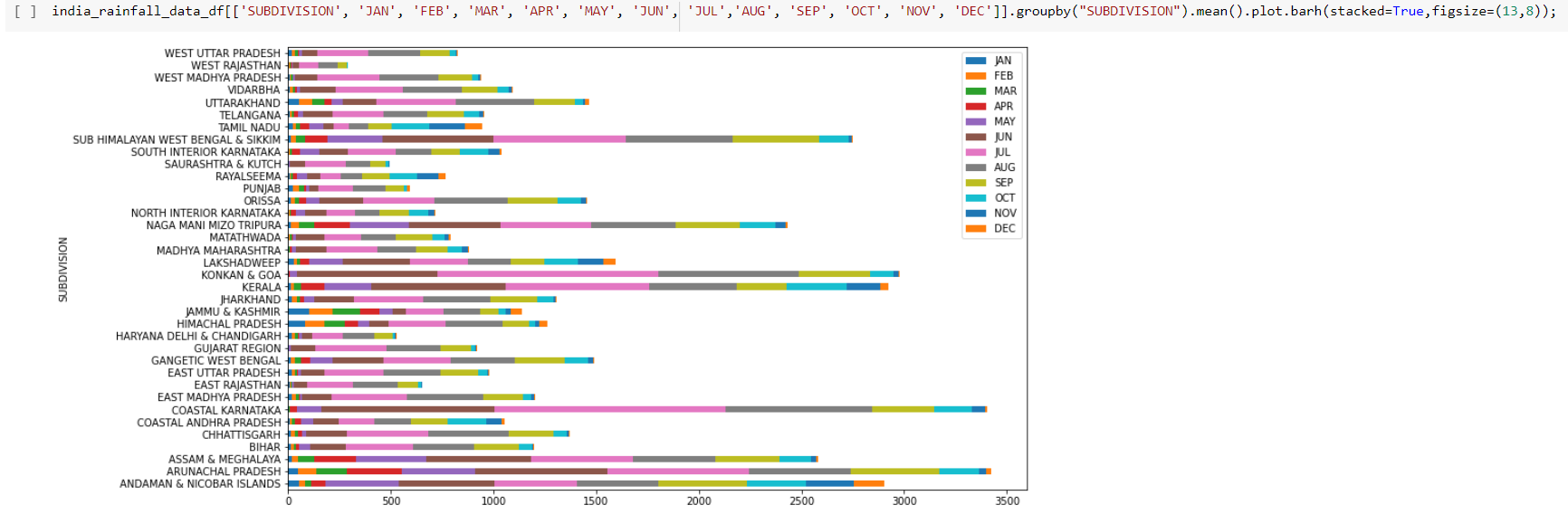
****

****

**1.8 Observations**

• The above two graphs show the distribution of rainfall over months.

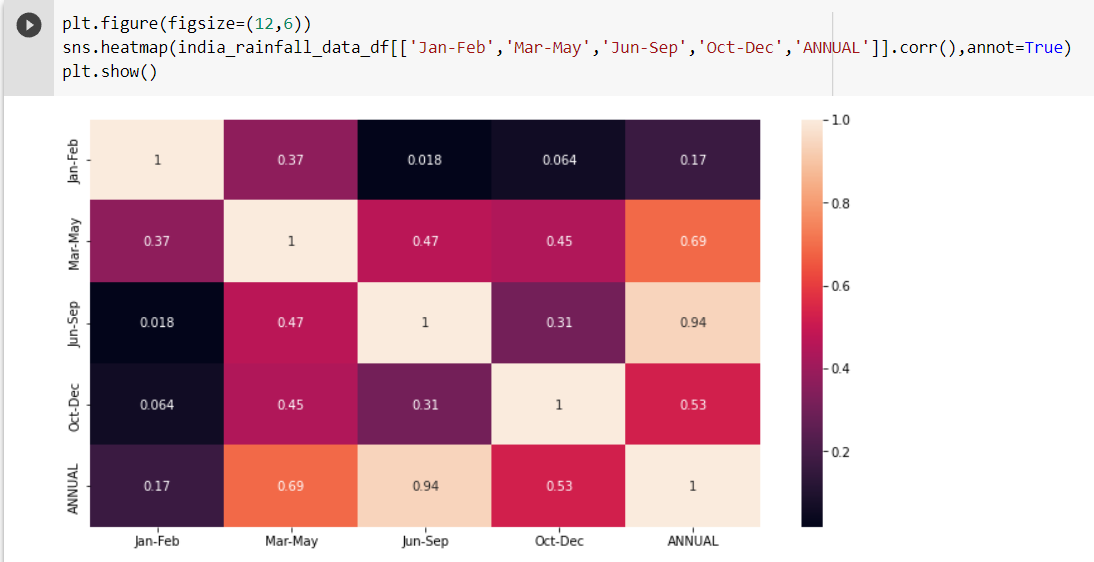
• The graphs clearly shows that amount of rainfall is high in the months july, aug, sep which is monsoon season in India.





**1.9 Observation**

• Above two graphs shows that the amount of rainfall is reasonably good in the months of march, april, may in eastern India.



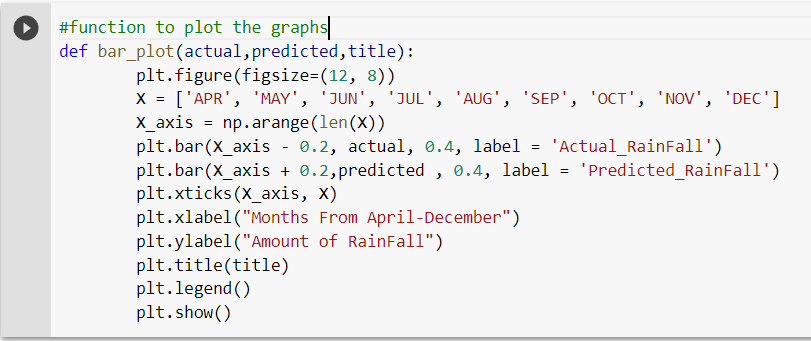


**1.10 Observation**

• Heat Map shows the co-relation(dependency) betwenn the amounts of rainfall over months.

• From above it is clear that if amount of rainfall is high in the months of july, august, september then the amount of rainfall will be high annually.

• It is also observed that if amount of rainfall is good in the months of october, november, december then the rainfall is going to be good in the overall year.



**1.11 Prediction**

• For prediction we formatted data in the way, given the rainfall in the last three months we try to predict the rainfall in the next consecutive month.

• For all the experiments we used 80:20 training and test ratio. – Linear regression

– SVR

– Artificial neural nets

- Random forest

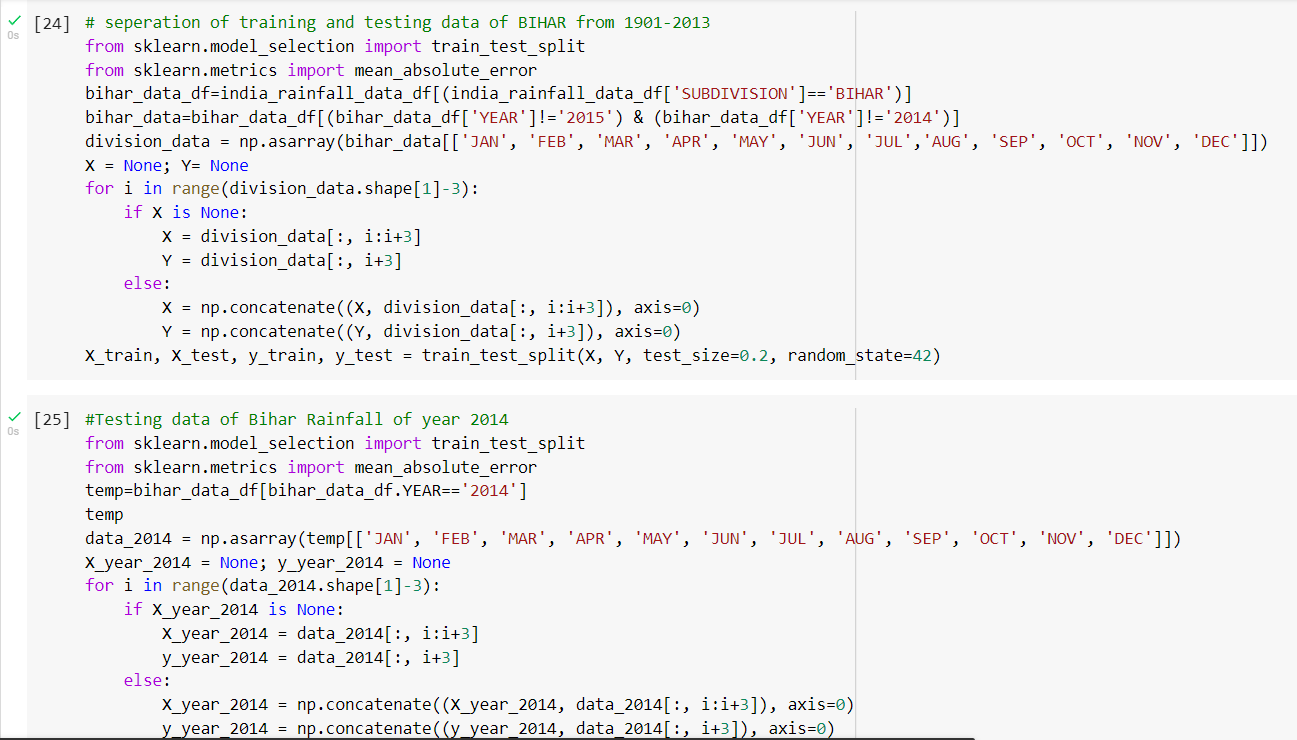
- Multilayered perceptron

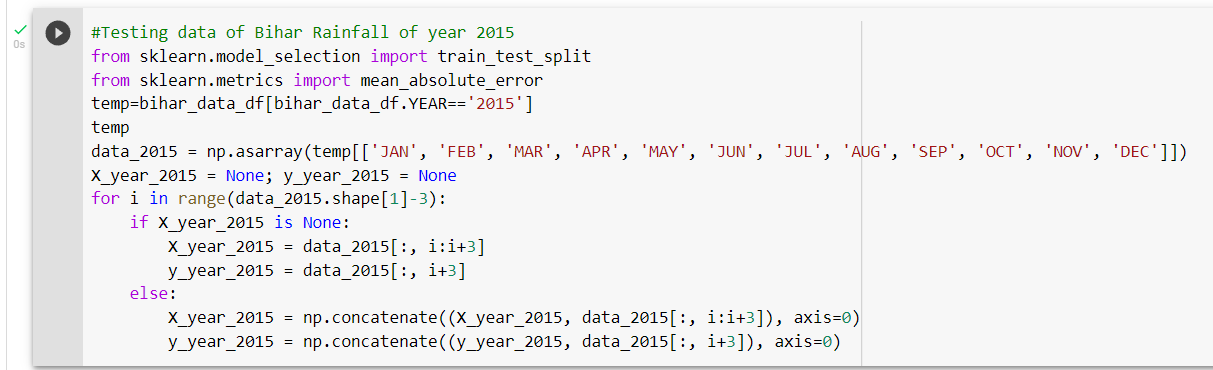
• Tersting metrics: We used Mean absolute error to train the models.

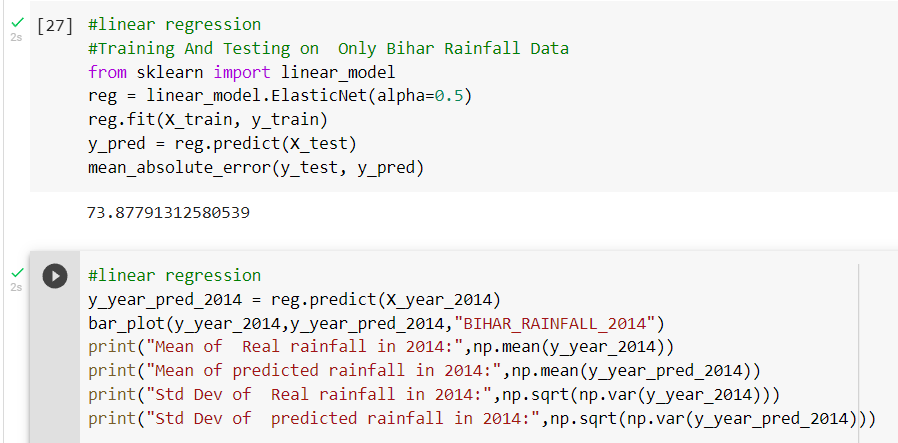
• We also shown the amount of rainfall actually and predicted with the histogram plots.

• We did two types of trainings once training on complete dataset and other with training with only telangana data

• All means are standard deviation observations are written, first one represents ground truth, second one represents predictions.







**Predictive Modelling Using Linear Regression**

Regression analysis comprises of the entire process of identifying the target and predictors,finding the relationship, estimating the coefficients, finding the predicted values of target, and finally evaluating the accuracy of the fitted relationship.

Regression analysis estimates the relationship between two or more variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

# Benefits of using Regression Analysis?

1. Regression explores significant relationships between dependent variable and independent variable.

2. Indicates the strength of impact of multiple independent variables on a dependent variable.

3. Allows us to compare the effect of variable measures on different scales and can consider nominal, interval, or categorical variables for analysis.

## **Equation with one dependent and one independent variable is defined by the formula:**

# ****y = c + b \* x****

## **where y = estimated dependent score**

## **c = constant**

## **b = regression coefficient,**

## **x = independent variable.**

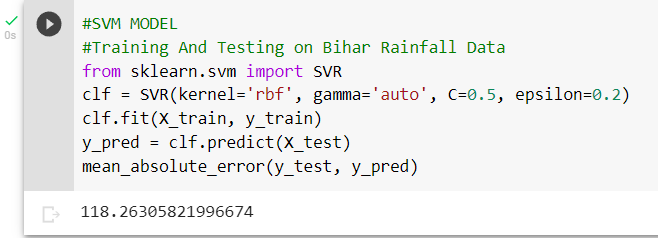
**Predective modelling using support vector regression**

We have to train a SVR model with the data to understand the correlation between the given rainfall data of three consecutive months and rainfall data of next month and be able to predict the rainfall of next month based on their given previous three months rainfall data.

Support Vector Regression is similar to Linear Regression in that the equation of the line is y=wx+ b In SVR, this straight line is referred to as hyperplane. The data points on either side of the hyperplane that are closest to the hyperplane are called support vector which is used to plot the boundary line.

Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value (Distance between hyperplane and boundary line), **a.** Thus, we can say that SVR model tries satisfy the condition -a < y-wx+b < a. It used the points with this boundary to predict the value.

In this, the function SVMis imported and is assigned to the variable regressor. The kernel “rbf” Radial Basis Function is used. RBF kernel is used to introduce a non-linearity to the SVR model. This is done because our data is non-linear. The regressor.fit is used to fit the variables X\_train and y\_train by reshaping the data accordingly.



**Predective modelling using Deep neural multilayer perceptron**

In the **Scikit-Learn**package, MLPRegressor is implemented in neural\_network module. Independent train and test dataset are further scaled to make sure that the input data is standard normally distributed are centred around zero and have variance in the same order. Now we will scale the data.

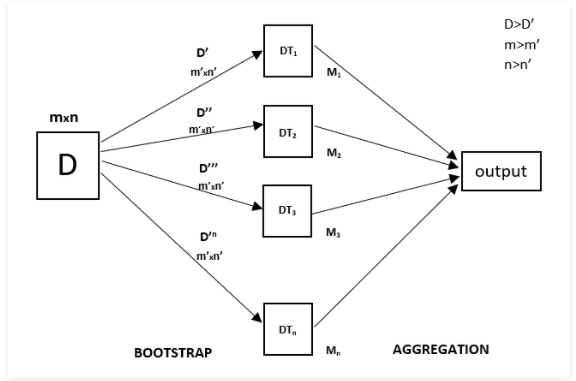
In the below code, three hidden layers are modelled, with 64 neurons in each layer. Considering the input and output layer, we have a total of 5 layers in the model. In case any optimiser is not mentioned then “Adam” is the default optimiser and it can manage pretty large dataset.

In addition to “RELU” activation, MLPRegressor supports the “sigmoid” and “hyperbolic tan” function.

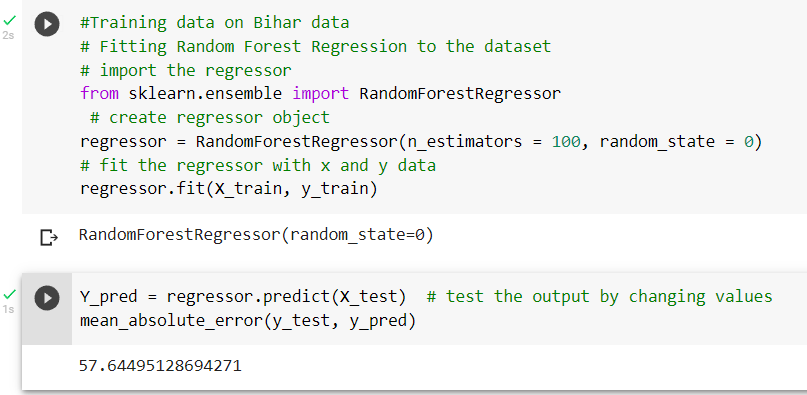


**Predective modelling using random forest regression**

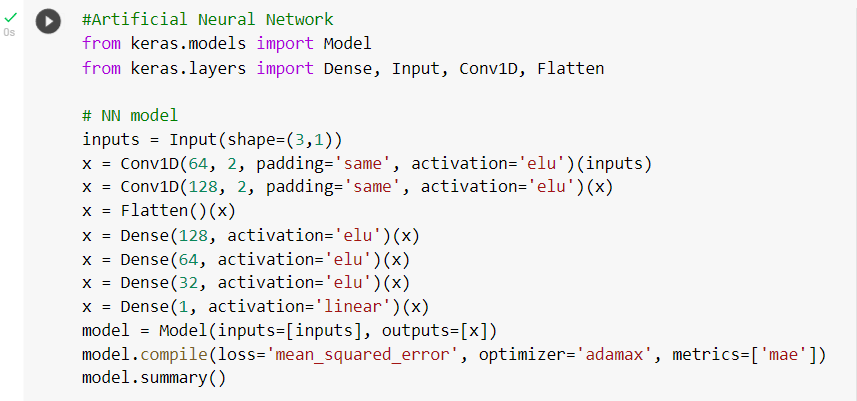
Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn’t depend on one decision tree but multiple decision trees. In the case of a regression problem, the final output is the mean of all the outputs. This part is Aggregation.

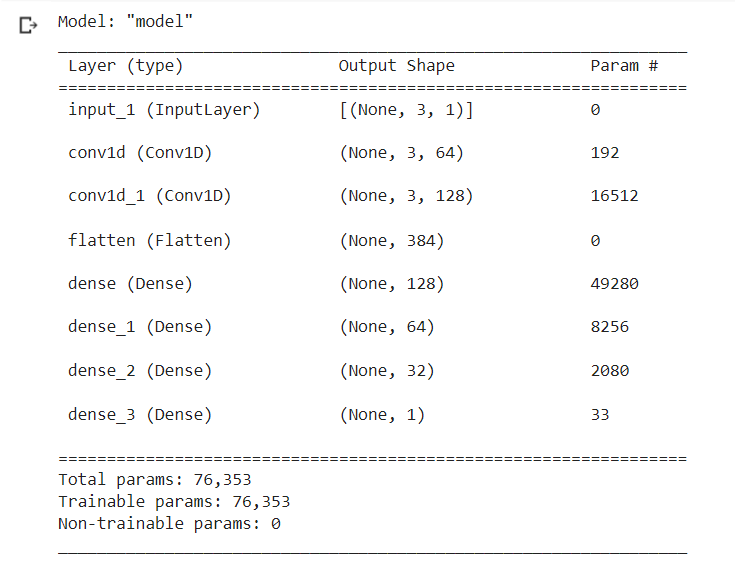
**** reference[**https://media.geeksforgeeks.org/wp-content/uploads/20200516180708/Capture482.png**](https://media.geeksforgeeks.org/wp-content/uploads/20200516180708/Capture482.png)

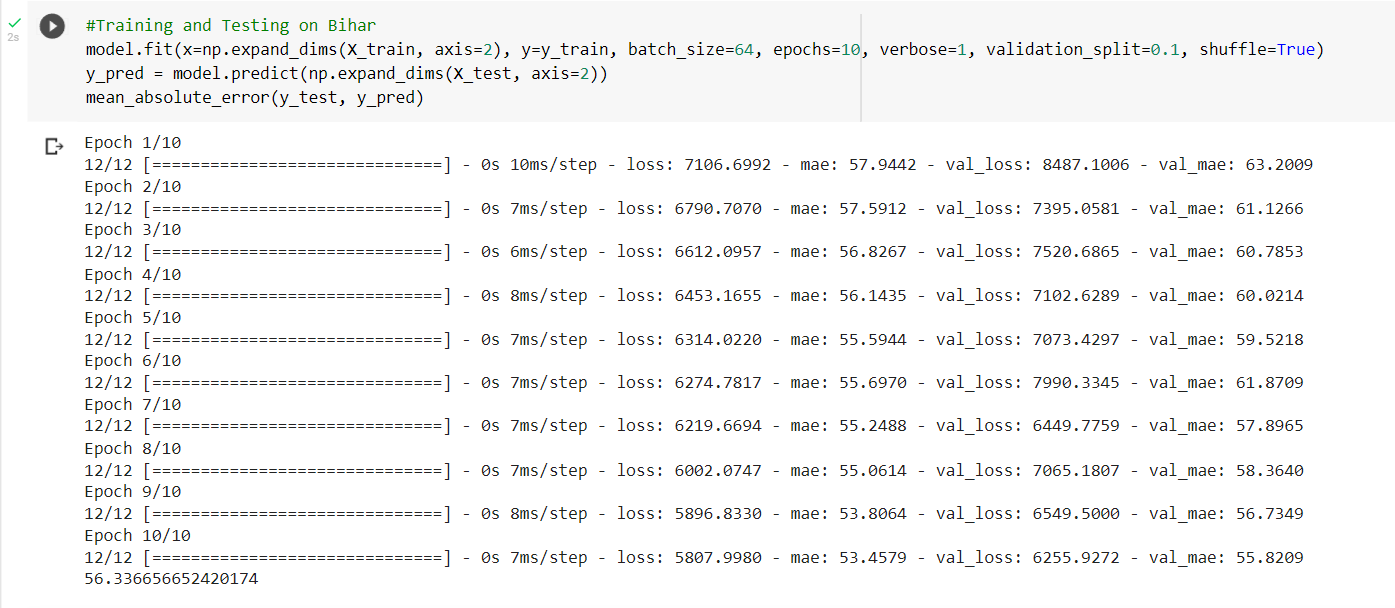
The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.  
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.



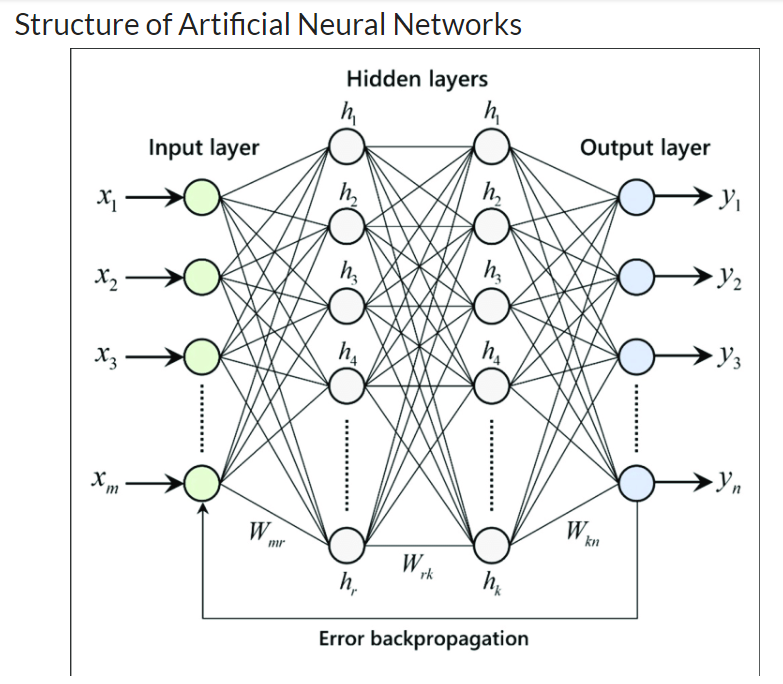
**Predective modelling using artificial neural network**

****

****

****

The purpose of using Artificial Neural Networks for Regression over Linear Regression is that the linear regression can only learn the linear relationship between the features and target and therefore cannot learn the complex non-linear relationship. In order to learn the complex non-linear relationship between the features and target, we are in need of other techniques. One of those techniques is to use Artificial Neural Networks. Artificial Neural Networks have the ability to learn the complex relationship between the features and target due to the presence of activation function in each layer.



Reference-<https://editor.analyticsvidhya.com/uploads/44122Architecture-of-multilayer-artificial-neural-network-with-error-backpropagation.png>

The Artificial Neural Networks consists of the Input layer, Hidden layers, Output layer. The hidden layer can be more than one in number. Each layer consists of n number of neurons. Each layer will be having an Activation Function associated with each of the neurons. The activation function is the function that is responsible for introducing non-linearity in the relationship. In our case, the output layer must contain a linear activation function. Each layer can also have regularizers associated with it. Regularizers are responsible for preventing overfitting.

Artificial Neural Networks consists of two phases,

* Forward Propagation
* Backward Propagation

Forward propagation is the process of multiplying weights with each feature and adding them. The bias is also added to the result. Backward propagation is the process of updating the weights in the model. Backward propagation requires an optimization function and a loss function.

**Prediction Observation**

**Training on bihar dataset**

|  |  |
| --- | --- |
| **Algorithms** | **Mean Absolute Error (MAE)** |
| Linear Regression | **73.877** |
| Support Vector Regression | **118.26** |
| Multilayer Perceptron | **57.644** |
| Random Forest Regression | **57.644** |
| Artificial Neural Network | **56.366** |

* Artificial neural network, random forest regression, multilayer perceptron performs better than linear regression and support vector regression.
* Observed MAE is high which indicates machine learning models won’t work well for prediction of rainfall due to drought in some particular year and very high fluctuation in rainfall
* Analysed individual year rainfall pattern for years 2014,2015.
* Approximately close means, noticed less standard deviation .

**MODEL 2**

This model is used for predicting the rainfall in a particular month of a particular year with the help of data of that month of previous consecutive three years. We can calculate for each month of that particular year.

X\_input=[];

Y=[]

We are storing every three consecutive years in X\_input and storing next of those three consecutive year Rainfall data in Y

Eg. X\_input=[[1901\_data,1902\_data,1903\_data],

[1902\_data,1903\_data,1904\_data],

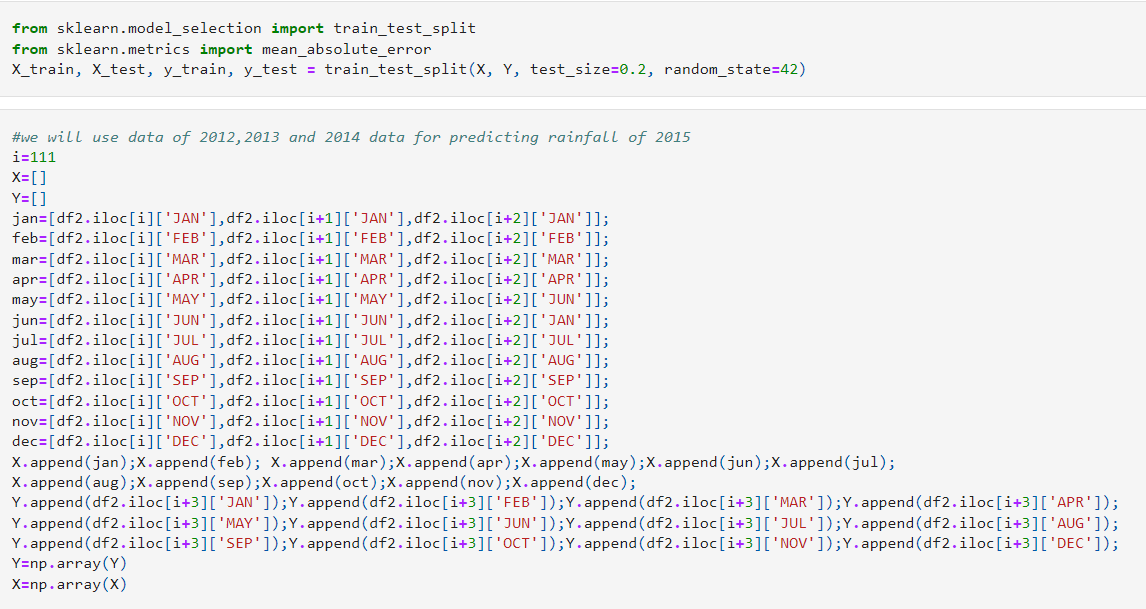
[1903\_data,1904\_data,1905\_data],

…….[2011\_data,2012\_data,2013\_data]]

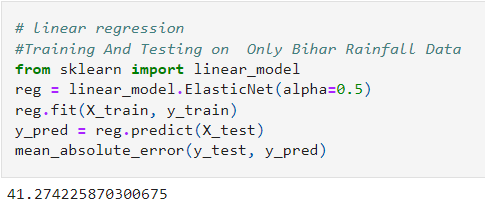
Y=[1904\_data,1905\_data,1906\_data,……2014\_data]



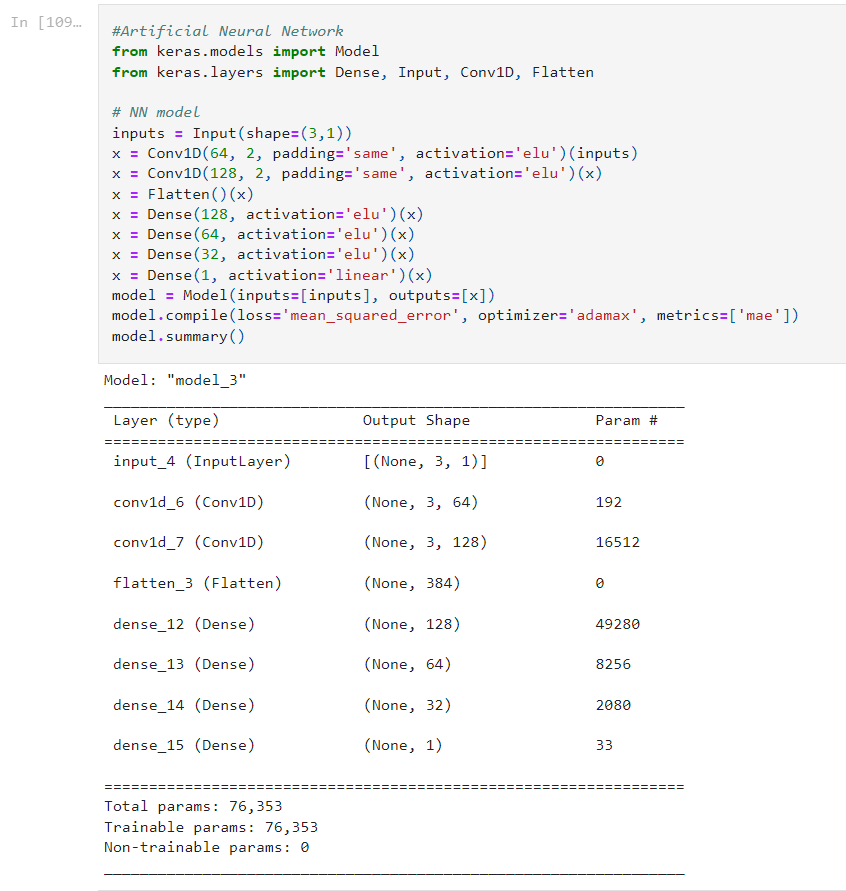
* Now we will separate rainfall data for training and testing in 80:20 ratio(80% rainfall data for training the machine learning algorithms and 20% data for testing).
* For final testing we gonna use year 2012,2013 and 2014 as input rainfall data and predict the 2015 rainfall data using five different machine learning algorithms and finally choose the model with lowest MAE(mean absolute error).

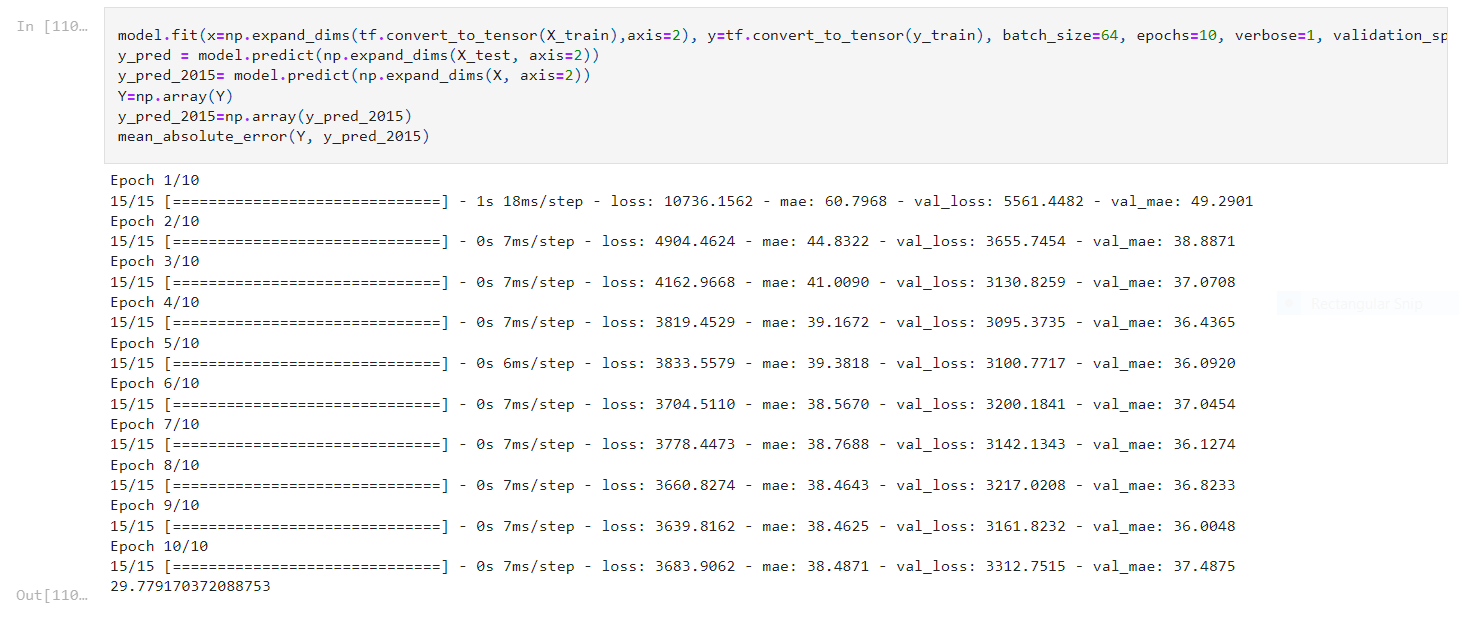


**Predective modelling using linear regression**

****

**Predective modelling using artificial neural network**

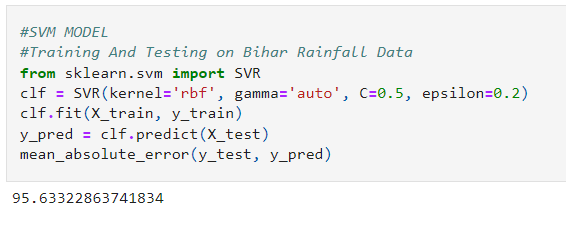
****

****

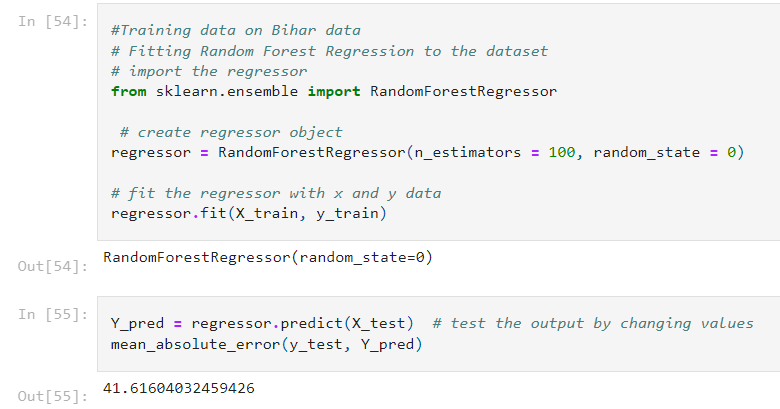
**Predective modelling using deep neural mutlilayer perceptron**

****

**Predective modelling using svr**

****

**Predective modelling using random forest regression**

****

**Prediction Observation**

**Training on bihar dataset**

|  |  |
| --- | --- |
| **Algorithms** | **Mean Absolute Error (MAE)** |
| Linear Regression | **41.27** |
| Support Vector Regression | **95.633** |
| Multilayer Perceptron | **39.227** |
| Random Forest Regression | **41.616** |
| Artificial Neural Network | **29.779** |

* Artificial neural network performs better than random forest regression, multilayer perceptron linear regression and support vector regression.
* Analysed individual year rainfall pattern for years 2015.
* Approximately close means, noticed less standard deviation .

**THANK YOU**