RAINFALL PREDICTION

A BTP REPORT
B21SSR02

By
G.SAI DIVYA(S20180010059)
G.SAI MANI(S20180010057)
R.BHAVANI(S20180010024)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY
09/05/2021
1st Semester Report

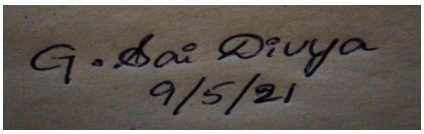
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the BTP entitled **"RAINFALL PREDICTION"** in the partial fulfillment of the requirements for the award of the degree of B.Tech and submitted in the Indian Institute of Information Technology SriCity, is an authentic record of my own work carried out during the time period from January 2021 to May 2021 under the supervision of Prof. SREEJA SR, Indian Institute of Information Technology SriCity, India.

The matter presented in this report has not been submitted by me for the award of any any other degree of this or any other institute.

Signature of the student with date



This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of BTP Supervisor with date

(Prof. Sreeja SR 16/5/21)

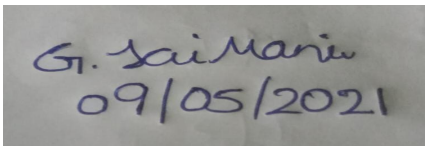
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the BTP entitled **"RAINFALL PREDICTION"** in the partial fulfillment of the requirements for the award of the degree of B.Tech and submitted in the Indian Institute of Information Technology SriCity, is an authentic record of my own work carried out during the time period from January 2021 to May 2021 under the supervision of Prof. SREEJA SR, Indian Institute of Information Technology SriCity, India.

The matter presented in this report has not been submitted by me for the award of any any other degree of this or any other institute.

Signature of the student with date



This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of BTP Supervisor with date

(Prof. Sreeja SR 16/5/21)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the BTP entitled **"RAINFALL PREDICTION"** in the partial fulfillment of the requirements for the award of the degree of B.Tech and submitted in the Indian Institute of Information Technology SriCity, is an authentic record of my own work carried out during the time period from January 2021 to May 2021 under the supervision of Prof. SREEJA SR, Indian Institute of Information Technology SriCity, India.

The matter presented in this report has not been submitted by me for the award of any any other degree of this or any other institute.

Signature of the student with date

R. Bhavani
9/5/21

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Signature of BTP Supervisor with date

(Prof. Sreeja SR 16/5/21)

ABSTRACT

Rainfall prediction could be very essential in numerous factors of our financial system and might assist us save you severe herbal disasters. Some regions in India are economically dependent on rainfall as agriculture is the number one career of many states. In India, Agriculture is an important factor for survival. For agriculture, rainfall is essential. Prediction of rainfall offers focus to humans and that they recognize earlier approximately rainfall to take sure precautions to shield their crop from rainfall. Many strategies got here into life to be expecting rainfall like many linear and non-linear algorithms that are typically used for seasonal rainfall prediction. In this paper, we applied different regression models to predict the rainfall, a few algorithms like Support Vector Regression, Random Forest regressor, ElasticNet, Lasso, Ridge regressions and neural networks. Dataset used for this utility is taken from Indian Meteorological Department which includes rainfall of each month, seasonal months, annual over 36 divisions throughout India from 1901-2015. Overall, we examine that set of rules that is viable for use with the intention to qualitatively are expecting rainfall.

Contents

INTRODUCTION

Rainfall prediction is one of the hard and unsure obligations which has a massive effect on human society. Timely and correct predictions can assist to proactively lessen human and monetary loss. This study presents a set of experiments which involve the use of prevalent machine learning techniques to build models to predict the amount of rainfall in upcoming months. This prediction particularly facilitates farmers and additionally water assets may be applied efficiently.

Rainfall prediction is a hard challenge and the effects must be correct. There are many hardware gadgets for predicting rainfall via means of the use of climate situations like temperature, humidity, pressure. These conventional techniques can't paint in a green manner so via means of the system getting to know strategies we will produce correct effects. We can simply do it via means of having the historic information evaluation of rainfall and might be expecting the rainfall for destiny seasons. We can practice many strategies like classification, regression in keeping with the necessities and additionally we will calculate the mistake among the real and prediction and additionally the accuracy. Different strategies produce one-of-a-kind accuracies so it's far vital to choose the right algorithm and model it according to the requirements.

This study takes a look at is specializing in predicting rainfall the use of Machine Learning Algorithms and Artificial Neural Network. The rainfall prediction will now no longer simply help in reading the converting styles of rainfall however it will additionally assist in organizing the precautionary measures in case of catastrophe and its management. The rainfall prediction might additionally help in making plans, regulations and techniques to address the growing worldwide trouble of ozone depletion. The converting styles of rainfall are related much with the worldwide warming; this is the growth of the earth’s temperature because of increased Chlorofluorocarbons emitted from the refrigerators, air conditioners, deodorants and printers etc. which might be a massive part of regular life. The growing temperature is certainly affecting the weather considerably.

Similarly, rainfall prediction and climate updates now no longer simply assist in coping with macro degree issues like flood and agricultural problems due to terrible or severe rainfall. The rainfall prediction can also make a contribution to the health and luxury of the humans via means of retaining them knowledgeable via means of monitoring the rainfall styles and predicting the rainfall via means of Machine getting to know algorithms and Artificial Neural Network. The rainfall predictions assist humans to address warm and humid climate. The technological improvement withinside the contemporary-day global has elevated the distance for innovation and revolution. Although the problems involved are in all likelihood related to those technological advancements ,one wishes to don't forget the variety of opportunities and possibilities that this technological Evolution has opened to human beings.

Regression analysis: Regression analysis deals with the dependence of one variable (called as dependent variable) on one or more other variables, (called as independent variables) which is useful for estimating and / or predicting the mean or average value of the former in terms of known or fixed values of the latter. For example, the salary of a person is based on his / her experience here, the experience attribute is independent variable salary is dependent variable. Simple linear regression defines the relationship between a single dependent variable and a single independent variable. The below equation is the general form of regression.

$y = \beta0 + \beta1x + \varepsilon$ where $\beta0$ and $\beta1$ are parameters, and ε is a probabilistic error term. Regression analysis is a vital tool for modeling and analyzing information. It is used for predictive analysis that is forecasting of rainfall or weather, predicting trends in business, finance, and marketing. It can also be used for correcting errors and also provide quantitative support.

The advantages of regression analysis are:

1. It is a powerful technique for testing the relationship between one dependent variable and many independent variables.
2. It lets in researchers to manipulate extraneous factors.
3. Regression assesses the cumulative impact of more than one factor.
4.It additionally enables to obtain the degree of blunders the usage of the regression line as a base for estimations.

LITERATURE SURVEY

Bushra Praveen et al., IIT INDORE [1] analyzes and forecasts the long-term Spatio-temporal changes in rainfall using the data from 1901 to 2015 across India at meteorological divisional level. They used the Pettitt test, the Mann-Kendall test and Sen's Innovative trend analysis to analyze the rainfall trend. The Artificial Neural Network-Multilayer Perceptron was employed to forecast the upcoming 15 years rainfall across India. Results show that the most of the meteorological divisions exhibited a significant negative trend of rainfall in annual and seasonal scales, except seven divisions during. This implementation can be the full package and should be helpful to the Indian planners proposing plans for small and large scale regions.

Nikhil Oswal., University of Ottawa, Canada [2] discusses various ways of exploring data analyses, data preprocessing, modelling and evaluation on australian dataset. He carried experiments with different input data; one with the original dataset, then with the undersampled dataset and last one with the oversampled dataset. Results show that in a few cases they achieved higher accuracy (Decision Tree) clearly implying the classic case of overtting. The performance of classifiers varied with different input data. To count a few, Logistic Regression performed best with undersampled data whereas it performed worst with over-sampled data; same goes with KNN, it performed best with oversampled data and worst with undersampled data. Hence Finally the input data has a very important role here. Ensembles to be precise Gradient Boosting performed pretty consistently in all the experiments.

Emilcy Hernandez et al., Coventry University UK [3] In This paper they presented a deep learning approach based on the use of autoencoders and neural networks to predict the accumulated precipitation for the next day. The approach forecasts the daily accumulated rainfall in a specific meteorological station located in a central area of Manizales city (Colombia). The proposed architecture has been compared with other state of the art methods. The results suggest that our proposed architecture outperforms other approaches in terms of the MSE and the RMSE.

Pritpal Singh et al., Tezpur University [4] They develop a model of ANN using BPNN with MLFF for prediction of ISMR on monthly and seasonal time scales, along with that they demonstrate its applicability for advance prediction of the seasonal rainfall amounts. But, the non-stationary nature of ISMR makes it's deterministic prediction more complex. Results from time consumption point of view, the model is found to be very

efficient for simulation, training, testing, and analyzing the data, which is important from the perspective of prediction studies which involves the prediction of dynamic variables of the environment.

Kumar Abhishek et al., NIT Patna [5] In this paper, the possibility of predicting average rainfall over Udupi district of Karnataka has been analyzed through artificial neural network models. In formulating artificial neural network based predictive models three layered networks have been constructed. The models under study are different in the number of hidden neurons. Results show that the CBP algorithm shows a high MSE in almost every case as compared to BPA or LRN. Also, its results were deviating from the fact that as the number of neurons increases, the MSE decreases.

To summarize the machine learning techniques, deep learning techniques can be used to achieve very high quality forecasting data. Data also plays a very important role to predict and to choose the best model.

DATASET DESCRIPTION

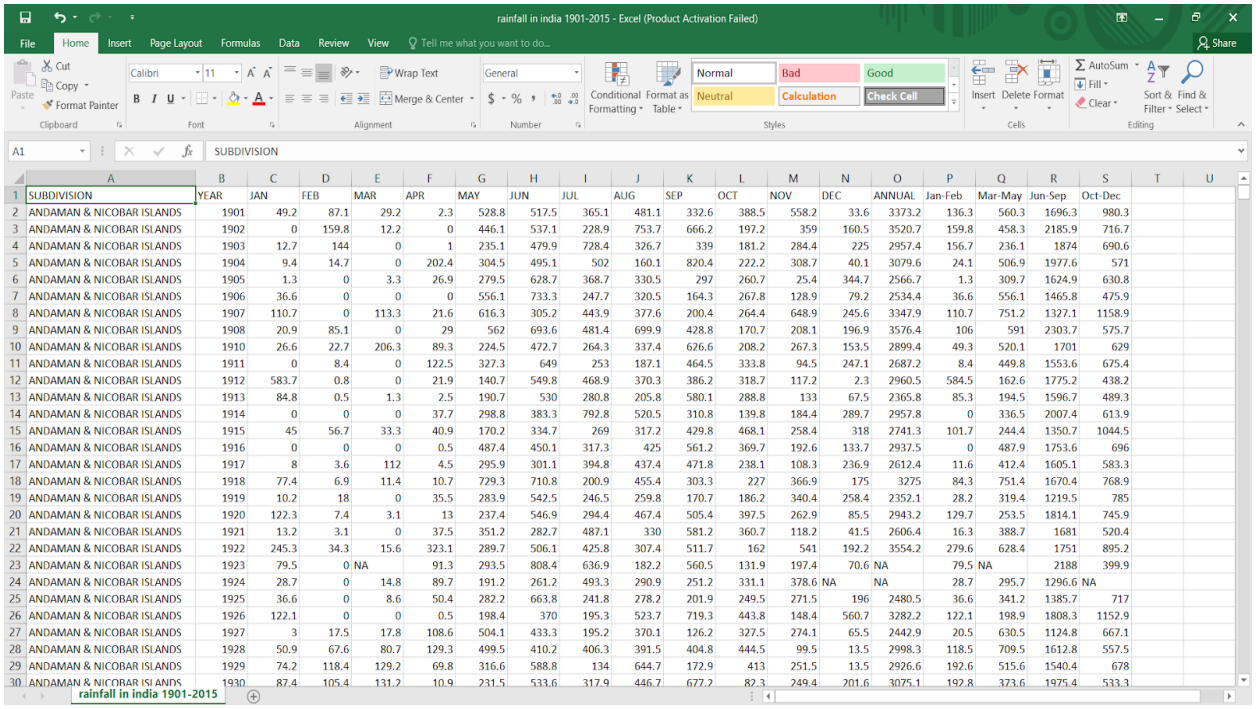
The data set consists of the measurement of rainfall from 1901-2015 for each state.

- Data consists of 19 attributes (individual months, annual, and seasonal months) for 36 sub divisions.

  [subdivision, year, jan, feb, mar, apr, may, june, july, aug, sep, oct, nov, dec, annual, jan-feb, mar-may, jun-sep, oct-dec]

- The attributes are amount of rainfall measured in mm

This dataset was provided by IMD.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec		
2	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3		
3	ANDAMAN & NICOBAR ISLANDS	1902	0	159.8	12.2	0	446.1	537.1	228.9	753.7	666.2	197.2	359	160.5	3520.7	159.8	458.3	2185.9	716.7		
4	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144	0	1	235.1	479.9	728.4	326.7	339	181.2	284.4	225	2957.4	156.7	236.1	1874	690.6		
5	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0	202.4	304.5	495.1	502	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571		
6	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0	3.3	26.9	279.5	628.7	368.7	330.5	297	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8		
7	ANDAMAN & NICOBAR ISLANDS	1906	36.6	0	0	0	556.1	733.3	247.7	320.5	164.3	267.8	128.9	79.2	2534.4	36.6	556.1	1465.8	475.9		
8	ANDAMAN & NICOBAR ISLANDS	1907	110.7	0	113.3	21.6	616.3	305.2	443.9	377.6	200.4	264.4	648.9	245.6	3347.9	110.7	751.2	1327.1	1158.9		
9	ANDAMAN & NICOBAR ISLANDS	1908	20.9	85.1	0	29	562	693.6	481.4	699.9	428.8	170.7	208.1	196.9	3576.4	106	591	2303.7	575.7		
10	ANDAMAN & NICOBAR ISLANDS	1910	26.6	22.7	206.3	89.3	224.5	472.7	264.3	337.4	626.6	208.2	267.3	153.5	2899.4	49.3	520.1	1701	629		
11	ANDAMAN & NICOBAR ISLANDS	1911	0	8.4	0	122.5	327.3	649	253	187.1	464.5	333.8	94.5	247.1	2687.2	8.4	449.8	1553.6	675.4		
12	ANDAMAN & NICOBAR ISLANDS	1912	583.7	0.8	0	21.9	140.7	549.8	468.9	370.3	386.2	318.7	117.2	2.3	2960.5	584.5	162.6	1775.2	438.2		
13	ANDAMAN & NICOBAR ISLANDS	1913	84.8	0.5	1.3	2.5	190.7	530	280.8	205.8	580.1	288.8	133	67.5	2365.8	85.3	194.5	1596.7	489.3		
14	ANDAMAN & NICOBAR ISLANDS	1914	0	0	0	37.7	298.8	383.3	792.8	520.5	310.8	139.8	184.4	289.7	2957.8	0	336.5	2007.4	613.9		
15	ANDAMAN & NICOBAR ISLANDS	1915	45	56.7	33.3	40.9	170.2	334.7	269	317.2	429.8	468.1	258.4	318	2741.3	101.7	244.4	1350.7	1044.5		
16	ANDAMAN & NICOBAR ISLANDS	1916	0	0	0	0.5	487.4	450.1	317.3	425	561.2	369.7	192.6	133.7	2937.5	0	487.9	1753.6	696		
17	ANDAMAN & NICOBAR ISLANDS	1917	8	3.6	112	4.5	295.9	301.1	394.8	437.4	471.8	238.1	108.3	236.9	2612.4	11.6	412.4	1605.1	583.3		
18	ANDAMAN & NICOBAR ISLANDS	1918	77.4	6.9	11.4	10.7	729.3	710.8	200.9	455.4	303.3	227	366.9	175	3275	84.3	751.4	1670.4	768.9		
19	ANDAMAN & NICOBAR ISLANDS	1919	10.2	18	0	35.5	283.9	542.5	246.5	259.8	170.7	186.2	340.4	258.4	2352.1	28.2	319.4	1219.5	785		
20	ANDAMAN & NICOBAR ISLANDS	1920	122.3	7.4	3.1	13	237.4	546.9	294.4	467.4	505.4	397.5	262.9	85.5	2943.2	129.7	253.5	1814.1	745.9		
21	ANDAMAN & NICOBAR ISLANDS	1921	13.2	3.1	0	37.5	351.2	282.7	487.1	330	581.2	360.7	118.2	41.5	2606.4	16.3	388.7	1681	520.4		
22	ANDAMAN & NICOBAR ISLANDS	1922	245.3	34.3	15.6	323.1	289.7	506.1	425.8	307.4	511.7	162	541	192.2	3554.2	279.6	628.4	1751	895.2		
23	ANDAMAN & NICOBAR ISLANDS	1923	79.5	0	NA	91.3	293.5	808.4	636.9	182.2	560.5	131.9	197.4	70.6	NA	79.5	NA	2188	399.9		
24	ANDAMAN & NICOBAR ISLANDS	1924	28.7	0	14.8	89.7	191.2	261.2	493.3	290.9	251.2	331.1	378.6	NA	NA	28.7	295.7	1296.6	NA		
25	ANDAMAN & NICOBAR ISLANDS	1925	36.6	0	8.6	50.4	282.2	663.8	241.8	278.2	201.9	249.5	271.5	196	2480.5	36.6	341.2	1385.7	717		
26	ANDAMAN & NICOBAR ISLANDS	1926	122.1	0	0	0.5	198.4	370	195.3	523.7	719.3	443.8	148.4	560.7	3282.2	122.1	198.9	1808.3	1152.9		
27	ANDAMAN & NICOBAR ISLANDS	1927	3	17.5	17.8	108.6	504.1	433.3	195.2	370.1	126.2	327.5	274.1	65.5	2442.9	20.5	630.5	1124.8	667.1		
28	ANDAMAN & NICOBAR ISLANDS	1928	50.9	67.6	80.7	129.3	499.5	410.2	406.3	391.5	404.8	444.5	99.5	13.5	2998.3	118.5	709.5	1612.8	557.5		
29	ANDAMAN & NICOBAR ISLANDS	1929	74.2	118.4	129.2	69.8	316.6	588.8	134	644.7	172.9	413	251.5	13.5	2926.6	192.6	515.6	1540.4	678		
30	ANDAMAN & NICOBAR ISLANDS	1930	87.4	105.4	131.2	10.9	231.5	533.6	317.9	446.7	677.2	82.3	249.4	201.6	3075.1	192.8	373.6	1975.4	533.3		

Figure 1. Dataset overview

METHODOLOGY

In this paper the overall architecture includes four major components: EDA, Data preprocessing, Model Implementation and testing each model with MAE and MSE.

Data Visualization:

Data visualization gives us a clear idea on data before making any assumptions. We can understand more about the data when we put the visual context through maps or graphs. This makes the data more natural for the human to identify trends, patterns within large data sets.Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and researchers.

1)Heat Map:

Heat map is used for a 2dimensional correlation matrix. Where Correlation ranges from -1 to +1. Heat Map shows the correlation(dependency) between the amounts of rainfall over months. Heat maps are basically used to know the relationship between variables.

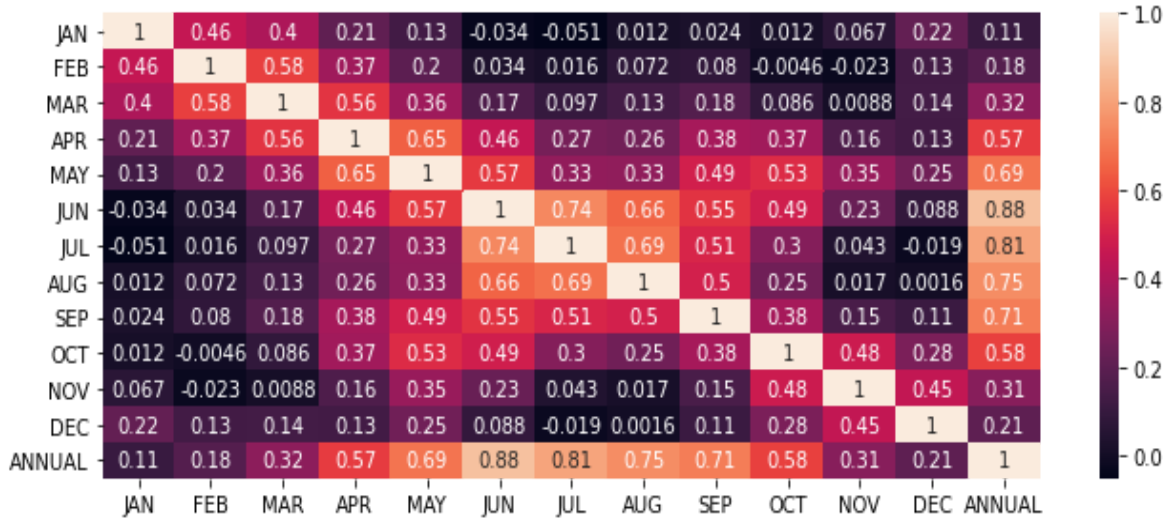


Figure 2. Correlation matrix for all months

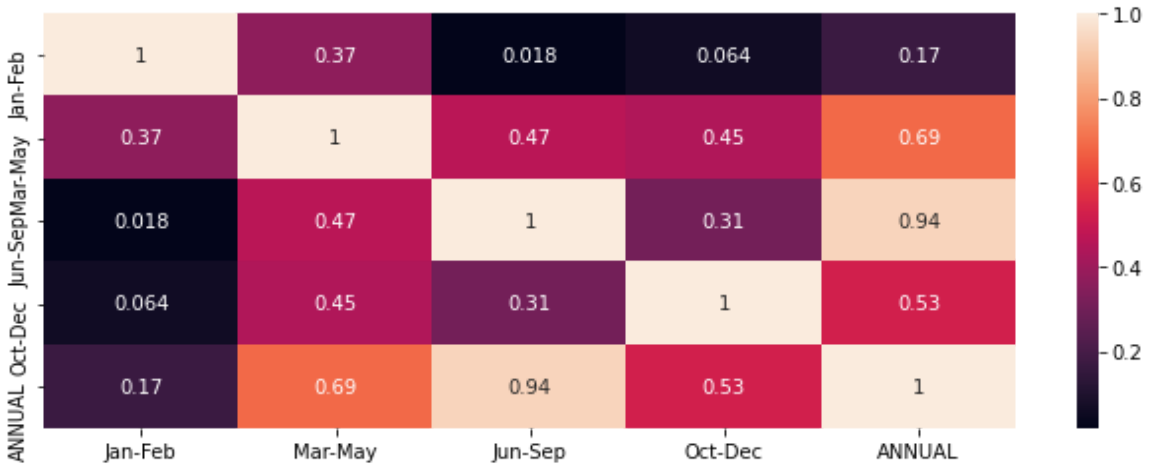


Figure 3. Correlation matrix for all seasonal months

Observations: From this heatmap we can observe the months are correlated with each other. Values closer to zero means there is no linear trend between the two variables. The closer to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is.

2)Bar Plot: Bar graphs are used to compare things between different groups.

a) Here is the graph of month wise rainfall in each subdivision.

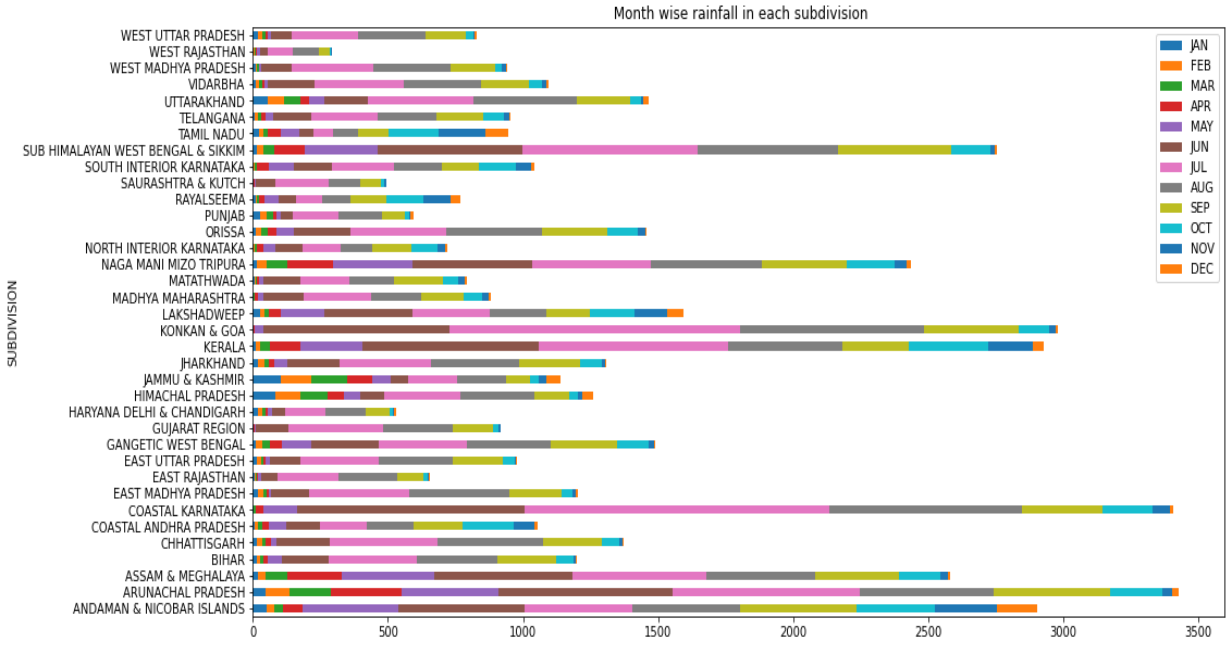


Figure 4. Barplot for monthwise rainfall for each division

Observations: We observe that in each sub division the amount of rainfall is high in jun, july, aug, sep months.

b) Here is the graph of seasonal wise rainfall in each subdivision.

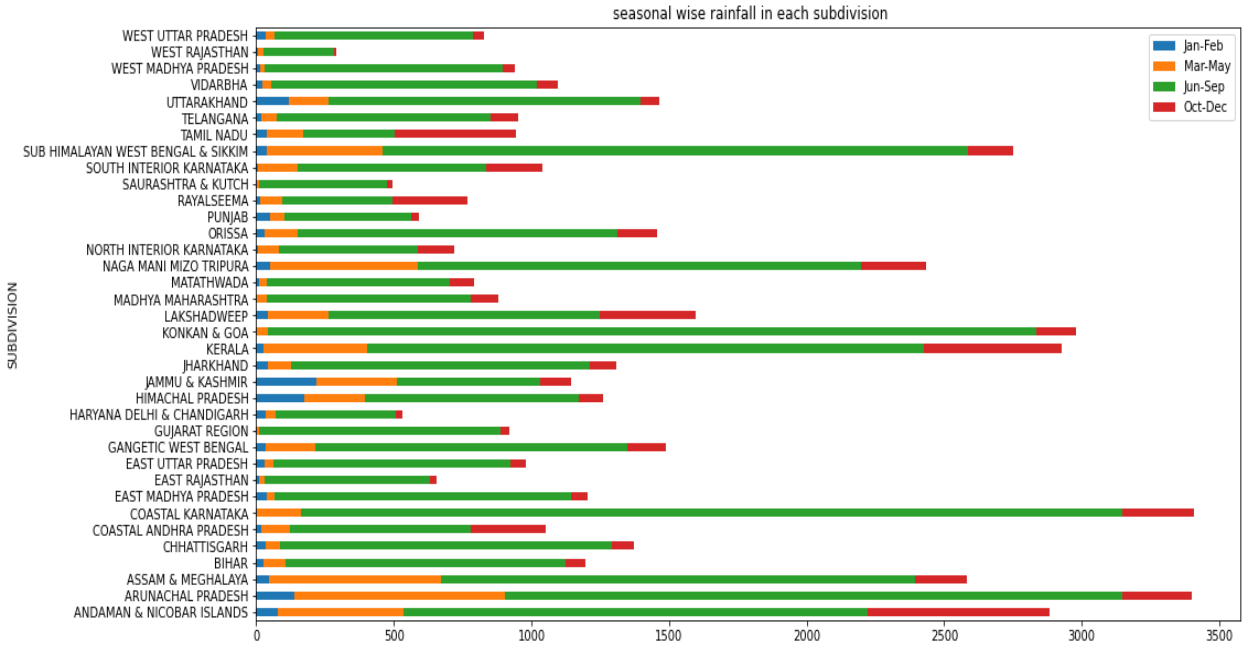


Figure 5.Barplot for seasonal months rainfall for each division.

Observation: We observe that in each sub division the amount of rainfall is high in june-sept season.

3)Line Plot: A line plot is a graph that displays data using a number line.

a) Here is the graph of overall rainfall in each month of the year.

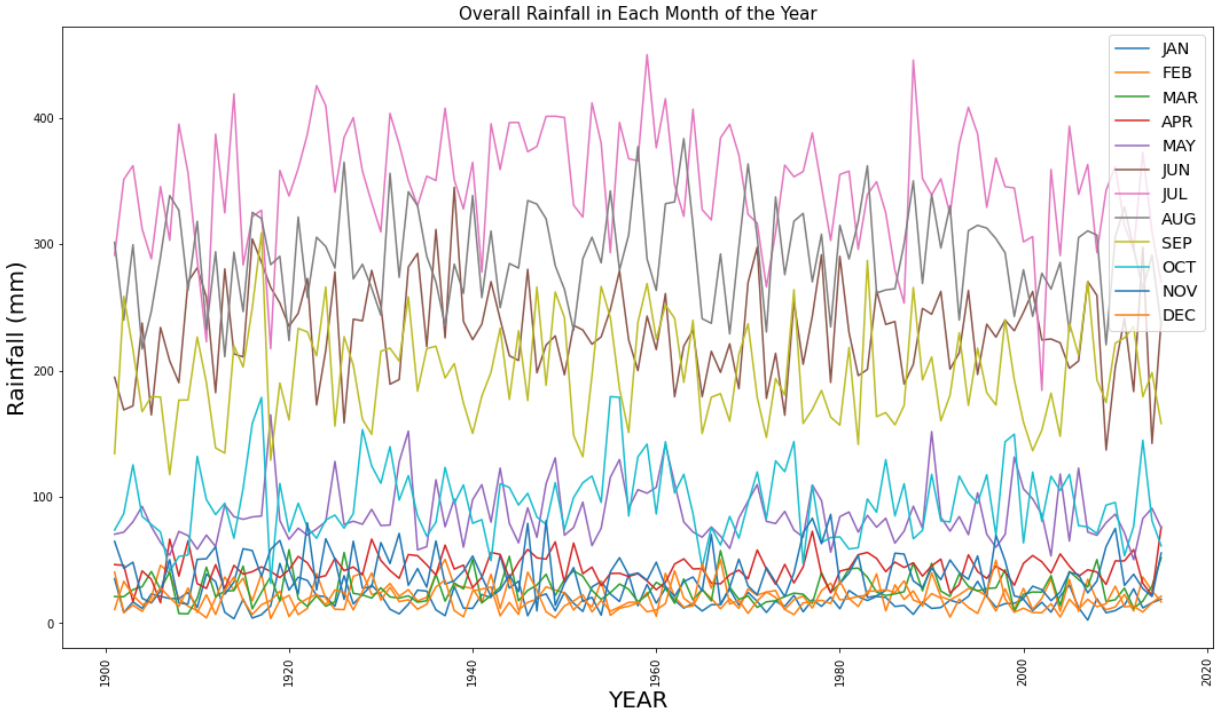


Figure 6. Line graph for month wise rainfall for every year.

Observation: From this we can observe that the amount of rainfall is high in the july month(i.e., pink line) .

b) Here is the graph of overall rainfall in seasonal months of the year.

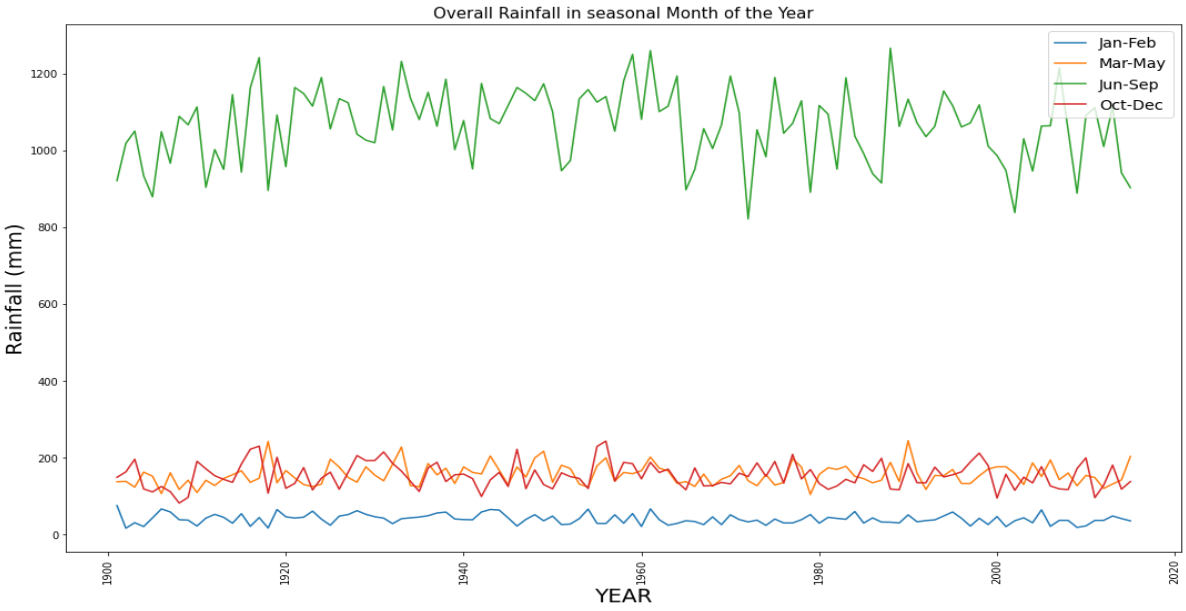


Figure 7. Line graph for seasonal month wise rainfall for every year

Observation: From this we can observe that the amount of rainfall is high in the jun-sept season.

c) Here is the graph of the annual trend of rainfall in india and it shows the distribution of rainfall over years.

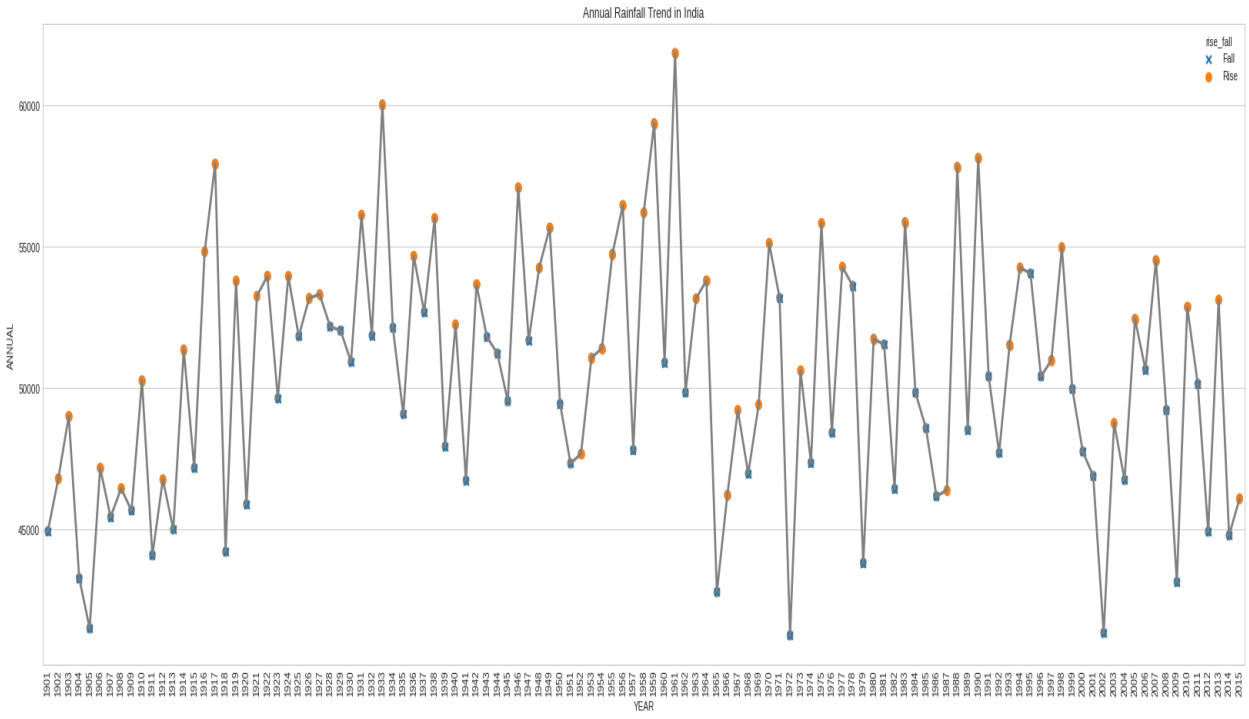


Figure 8. Line graph for annual rainfall for every year

Observation: From here we can observe that the year 1961 had recorded the highest rainfall and the year 1972 had recorded the lowest rainfall in india.

4)Box Plot: It is a simple way of representing statistical data on a plot in which a rectangle is drawn.

Here is the graph of the annual rainfall in subdivisions of india.

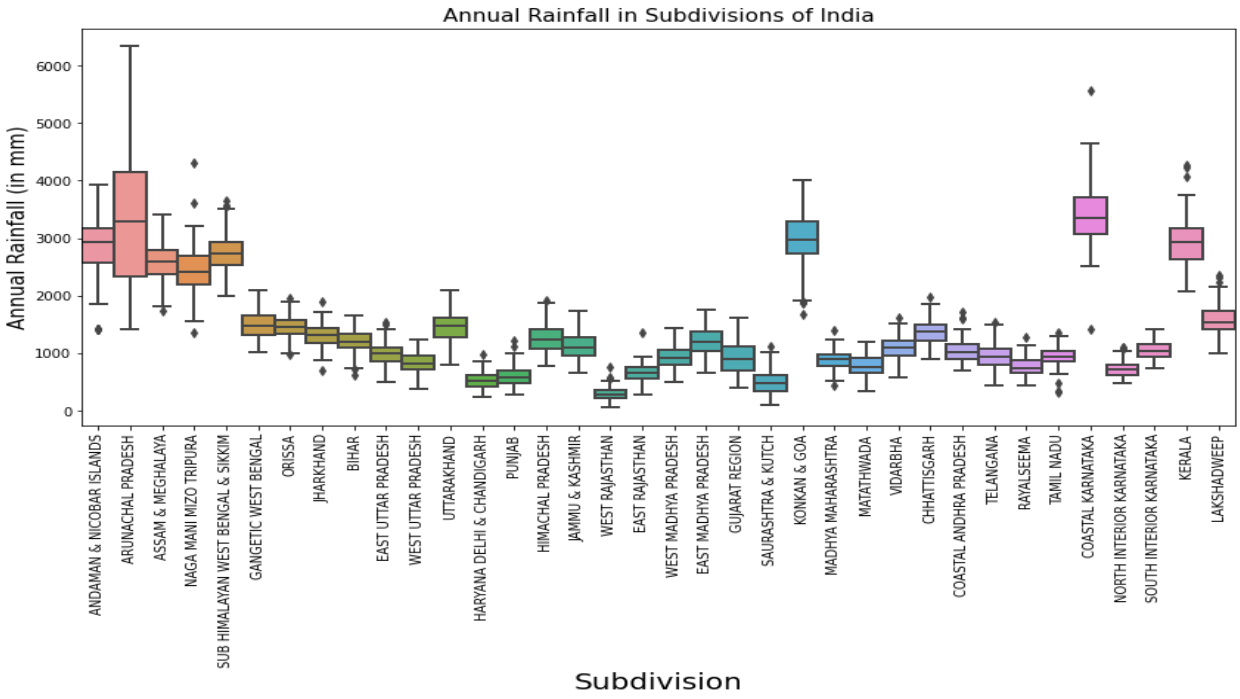


Figure 9. Boxplot for Annual rainfall for each subdivision

Observation: From here we can observe that Arunachal pradesh had recorded the highest rainfall and west rajasthan had recorded the lowest rainfall in india.

2. Missing values:

Missing Values: As per our EDA step, we learned that we have few instances with null values. Hence, this becomes one of the important steps. So first we have to remove missing values and apply machine learning and deep learning models.

There so many methods to fill the missing values

Ignore the missing values :

Missing data under 10% for an individual case or observation can generally be ignored, except when the missing data is a MAR or MNAR.

Drop the missing values (Dropping a variable) :

The number of missing values in a feature is very high in the data, then that feature should be left out of the analysis.

Imputation by Mean/Mode/Median:

Imputation is the process of substituting the missing data by some statistical methods. If the missing values in a column or feature are numerical, the values can be imputed by the mean of the complete cases of the variable. Mean can be replaced by median if the feature is suspected to have outliers. For a categorical feature, the missing values could be replaced by the mode of the column.

Regression Methods:

The variables with missing values are treated as dependent variables and variables with complete cases are taken as predictors or independent variables. The independent variables are used to fit a linear equation for the observed values of the dependent variable. This equation is then used to predict values for the missing data points.

K-Nearest Neighbour Imputation (KNN) :

This method uses k-nearest neighbour algorithms to estimate and replace missing data. The k-neighbours are chosen using some distance measure and their average is used as an imputation estimate. This could be used for estimating both qualitative attributes (The most frequent value among the k nearest neighbours) and quantitative attributes (the mean of the k nearest neighbours) .

So From all of the above methods, as per our data (completely numerical), we want to proceed with the imputation method by means.

Applying Missing Values method:

In any data, None and NaN are used for indicating missing or null values. NaN is used to indicate numerical and None to object data. To facilitate this convention between missing values, there are several useful methods. They are:

1. isnull() : Generate a boolean mask indicating missing values
2. notnull() : Opposite of isnull()
3. dropna() : Return a filtered version of the data
4. fillna() : Return a copy of the data with missing values filled or imputed

So first we will total missing values sum using :

dataset.isnull().sum()

Then we apply the missing values method using fillna method Again we check the missing value sum to confirm that there are no missing values in our dataset.

Mean: Mean (commonly known as average) is equal to the sum of all values in the column divided by the number of values present in the column. In Excel, you can use the AVERAGE() function to compute the mean.

dataset = dataset.fillna(dataset.mean())

When we apply fillna of mean to the dataset the missing values will be replaced by mean of the column.

Before applying mean method

```
SUBDIVISION     0
YEAR            0
JAN             4
FEB             3
MAR             6
APR             4
MAY             3
JUN             5
JUL             7
AUG             4
SEP             6
OCT             7
NOV            11
DEC            10
ANNUAL         26
Jan-Feb         6
Mar-May         9
Jun-Sep        10
Oct-Dec        13
dtype: int64
```

After applying mean method

```
SUBDIVISION     0
YEAR            0
JAN             0
FEB             0
MAR             0
APR             0
MAY             0
JUN             0
JUL             0
AUG             0
SEP             0
OCT             0
NOV             0
DEC             0
ANNUAL          0
Jan-Feb         0
Mar-May         0
Jun-Sep         0
Oct-Dec         0
dtype: int64
```

```
[27] data = data.fillna(data.mean())
     data.head()
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
1	ANDAMAN & NICOBAR ISLANDS	1902	0.0	159.8	12.2	0.0	446.1	537.1	228.9	753.7	666.2	197.2	359.0	160.5	3520.7	159.8	458.3	2185.9	716.7
2	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144.0	0.0	1.0	235.1	479.9	728.4	326.7	339.0	181.2	284.4	225.0	2957.4	156.7	236.1	1874.0	690.6
3	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0.0	202.4	304.5	495.1	502.0	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571.0
4	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0.0	3.3	26.9	279.5	628.7	368.7	330.5	297.0	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8

Figure 10. Missing values sum before and after applying mean method

To summarise we are filling the missing values with mean. As our data is numerical data and only contains rainfall measurements in mm.

3. Modelling:

We chose different regression models like neural networks, SVR, Random forest regression, ElasticNet, Ridge, Lasso regressions.

The following Regression algorithms have been used to build prediction models to perform the experiments:

Support Vector Regression (SVR) is a regression algorithm, so we can use SVR for working with the continuous Values instead of Classification which is SVM. Support Vector Machine maintains all the core features that describe the characteristics of the algorithm. Support Vector Machine (SVR) our main focus is to fix the error within a certain threshold.

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees. 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.

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. Another biased regression technique, principal components regression. Ridge regression is the more popular of the two methods.

Lasso, or Least Absolute Shrinkage and Selection Operator,is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the

so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. Elastic Net aims at minimizing the following loss function.

Neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. We have to train the neural network in that way to fit the model for our data. Supervised training involves a mechanism of providing the network with the desired output either by manually "grading" the network's performance or by providing the desired outputs with the inputs. Unsupervised training is where the network has to make sense of the inputs without outside help.

In order to evaluate the performance of the algorithms, we use the Mean Square Error (MSE) and the Mean Absolute Error (MAE) as measurement errors.

Finally, first we imported the data, using visualization we made some assumptions, then applied imputation by mean method to fill missing values, then for prediction we formatted data in this way, given the rainfall in the last three months we try to predict the rainfall in the next consecutive month. We divided the data into training size and test size with ratio 80:20 then applied the Machine learning and deep learning algorithms. Found MAE and MSE to evaluate models.

Results

MAE and MSE :

1. Results of all models MAE, MSE error

Model	Mean Absolute Error	Mean Squared Error
Elastic Net Regression	93.3734	142.5711
Ridge Regression	93.9519	141.8863
Lasso Regression	93.9520	141.8863
SVR	125.1810	202.5503
Random Forest Regressor	91.7085	145.7345
Neural Network	85.1830	135.8926

We Observe that all the linear models have almost the same MAE and MSE.
The Neural Network got less errors among all other methods, the second better algorithm was Random Forest Regression.
So we want to proceed using a neural network model for upcoming work.

We applied a neural network for every division in our data and calculated mean absolute error. From this we observed that for some of the divisions the neural network model got some error and for a few divisions the neural network model got high error.

So for the regions who got high error we will modify the model that fits that data and predict future year rainfall values using that model.

Here is the results of error for every division and division names:

```
[10] subdivs

array(['ANDAMAN & NICOBAR ISLANDS', 'ARUNACHAL PRADESH',
       'ASSAM & MEGHALAYA', 'NAGA MANI MIZO TRIPURA',
       'SUB HIMALAYAN WEST BENGAL & SIKKIM', 'GANGETIC WEST BENGAL',
       'ORISSA', 'JHARKHAND', 'BIHAR', 'EAST UTTAR PRADESH',
       'WEST UTTAR PRADESH', 'UTTARAKHAND', 'HARYANA DELHI & CHANDIGARH',
       'PUNJAB', 'HIMACHAL PRADESH', 'JAMMU & KASHMIR', 'WEST RAJASTHAN',
       'EAST RAJASTHAN', 'WEST MADHYA PRADESH', 'EAST MADHYA PRADESH',
       'GUJARAT REGION', 'SAURASHTRA & KUTCH', 'KONKAN & GOA',
       'MADHYA MAHARASHTRA', 'MATATHWADA', 'VIDARBHA', 'CHHATTISGARH',
       'COASTAL ANDHRA PRADESH', 'TELANGANA', 'RAYALSEEMA', 'TAMIL NADU',
       'COASTAL KARNATAKA', 'NORTH INTERIOR KARNATAKA',
       'SOUTH INTERIOR KARNATAKA', 'KERALA', 'LAKSHADWEEP'], dtype=object)
```

```
meanerror

array([140.4678507 , 199.28770968, 109.12406223,  94.3462306 ,
       128.72484457,  77.49932806,  81.35422742,  67.86587383,
        77.17628137,  61.79250835,  62.64865462,  88.31259251,
        40.20679799,  46.87334736,  66.61765539,  57.08613747,
        28.17995848,  41.28195253,  63.45461393,  78.16995787,
        80.97948572,  50.93588132, 200.61164137,  54.52411525,
        60.60954418,  64.22243438,  75.99868872,  64.00701032,
        60.46937416,  50.05342925,  54.359456  , 185.69336385,
        49.22969688,  61.56358943, 189.81332926, 104.37353324])
```

11.MAE for each subdivision

CONCLUSION

In this paper, we explored the data using various visualization ways and missing value techniques and learned their impact on overall performance of various regression models. We also carried a comparative study of all the regressors with input data and observed how the input data can affect the model predictions.

Using comparative study we choose the best model as Neural networks by MAE, MSE. So we want to proceed with the neural network model.

For future work, we want to predict the future rainfall values of each division using a neural network model. We want to display the results through an application.

We want to predict the rainfall values for upcoming 5/10 years month-to-month and need to reveal the outcomes both via software program or via an internet site wherein you could take a look at for a specific month or specific year(s) for special regions.

This will assist many human beings to take precautions beforehand.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SVR	Support Vector Regression
IMD	Indian Meteorological Department
MAE	Mean Absolute Error
MSE	Mean Square Error
KNN	K-Nearest Neighbour
NAN	Not a number
EDA	Exploratory Data Analysis
ANN	Artificial Neural Network
BPNN	Backpropagation Neural Network
MLFF	Multi Layer Feed Forward
RMSE	Root mean Square error
BPA	Back Propagation Algorithm
LRN	Layer Recurrent Network
CBP	Cascaded Back-Propagation
MAR	Missing at Random
ISMR	Ionospheric Scintillation Monitoring Records
MNAR	Missing not at random

REFERENCES

1. Praveen, Bushra, et al. "Analyzing trend and forecasting of rainfall changes in India using non-parametric and machine learning approaches." *Scientific reports* 10.1 (2020): 1-21.

2. Oswal, Nikhil. "Predicting rainfall using machine learning techniques." *arXiv preprint arXiv:1910.13827* (2019).

3. Hernández, Emilcy, et al. "Rainfall prediction: A deep learning approach." *International Conference on Hybrid Artificial Intelligence Systems*. Springer, Cham, 2016.

4. Singh, Pritpal, and Bhogeswar Borah. "Indian summer monsoon rainfall prediction using artificial neural network." *Stochastic environmental research and risk assessment* 27.7 (2013): 1585-1599.

5. Abhishek, Kumar, et al. "A rainfall prediction model using an artificial neural network." *2012 IEEE Control and System Graduate Research Colloquium*. IEEE, 2012.

6. Kang, Hyun. "The prevention and handling of the missing data." *Korean journal of anesthesiology* 64.5 (2013): 402.

7. Mostafa, Samih M. "Imputing missing values using cumulative linear regression." *CAAI Transactions on Intelligence Technology* 4.3 (2019): 182-200.

8. De Silva, Hiroshi, and A. Shehan Perera. "Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data." *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2016.