Rainfall Prediction Using Machine Learning

Neha Jha[1] Ruchi Sanghvi[2] Dhaval Patel[3] Divyesh Rana[4] Anali Shah[5]
[1,2,3,4]Student [5]Assistant Professor
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]S S Agrawal Institute of Engineering & Technology, Navsari(Gujarat) -396445, India

Abstract— This paper introduces current supervised learning model which is based on machine learning algorithm for Rainfall prediction in India. Rainfall is always being a major issue across the world as it affects all the major factor on which the human being is depended. In current, Unpredictable and accurate rainfall prediction is very challenging task [1]. We applied linear regression algorithm of machine learning with the help of Python through Spyder in Anaconda Navigator to compare the accuracy of between training and test data. Our motive is to get the most accurate and a better rainfall prediction in result.

Keywords: Machine Learning, Linear Regression Algorithm, Spyder in Anaconda Navigator, Python

I. INTRODUCTION

Rainfall Prediction is the application of science and technology to predict the amount of rainfall over a region. It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre-planning of water structures. In this article, we have used Linear Regression to predict the amount of rainfall. Linear Regression tells us how many inches of rainfall we can expect. [2]

In Today's era, we see that global warming is affecting all over the world at large scale which majorly effect on mankind and cause the expedite change in climate. Due to this iceberg is meltion,air and oceans are warming, sea level is rising and flooding and drought etc. One of the serious consequences due to this climate change is on the Rainfall. Rainfall prediction now a days is an arduous task which is taking into the consideration of most of the major world-wide authorities. Rainfall is a climatic factors that affects several human activities on which they are depended on for e.g.:- agricultural production, construction, power generation and tourism, among others. This makes the rainfall serious concern and that's why there is requirement of better rainfall prediction. Rainfall is a complex atmospheric process, and due to the climate changes, it becomes more difficult to predict it. [3]

Due to such type of heavy rainfall, roads and bridges become completely destroyed and 100,000 pilgrims and tourists are always trapped which are on their "Char Dham Yaatra"[3] due to this hazardous and this disaster could not be predicted by the government, big industries or risk management entitles, as well as the scientific community before the incident. These are also may lead to the land slide which is also a most serious geo-hazard causing the loss of life and property all over the world.

II. PROBLEM STATEMENT:

As we know that climate is an important aspect of human life. So, the Prediction should be accurate as much as possible. In this paper we tried to deal with the prediction of the rainfall which is also a major aspect of human life and which provide the major resource of human life which is "Fresh Water". Fresh water has always been a crucial resource of human survival – not only for the drinking purposes but also for farming, washing and many other purposes.

A bad rainfall prediction can affect the agriculture mostly the framers because their whole crop is depend on the rainfall and agriculture has always been an important part for everyone's. So, for making an accurate prediction of the rainfall we need to use big dataset. There are number of techniques are used of machine learning but accuracy is always a matter of concern in prediction made in rainfall. There are number of causes made by rainfall affecting the world ex. Drought, Flood and intense summer heat etc. And it will also affect water resources around the world.

Making prediction on rainfall cannot be done by the traditional way, so scientists are using machine learning and deep learning to find out the pattern for rainfall prediction. Here we have used ML for the prediction.

A. About Dataset:

Whenever we think of Machine Learning, the very first thing that comes to our mind is a dataset. While there are many datasets that we can find on websites such as:- Kaggle, sometimes it is useful to extract data on our own and generate our own dataset. Generating our own dataset gives us more control over the data and allows us to train our machine learning model. Data comes in all forms, most of it being very messy and unstructured.

Datasets, large and small, come with a variety of issues which are:-

(1) invalid fields, (2)missing and additional values, (3)and values that are in forms different from the one we require. In order to bring it to workable or structured form, we need to "clean" our data, and make it ready to use. Some common cleaning includes parsing, converting to one-hot, removing unnecessary data, etc. In our case, our data has some days where some factors weren't recorded. Our algorithm requires numbers, so we can't work with alphabets popping up in our data. so we need to clean the data before applying it on our model. Once the data is cleaned, it can be used as an input to our Linear regression model.

This data set contains monthly rainfall detail of 36 meteorological sub-divisions of India.

B. Library Used:-

Following are the libraries used in this model:

1) Numpy
2) Pandas
3) Matplotlib
4) Sci-Kit learn
5) Seaborn

III. LINEAR REGRESSION MODEL:

Linear regression is a linear approach to form a relationship between a dependent variable and many independent explanatory variables. This is done only by plotting a line that fits our scatter plot the best, i.e. with the least errors. This gives value predictions, i.e. how much, by substituting the independent values in the line equation. We have used Sci-kit learn's linear regression model to train our dataset. Once the model get trained, we can give our own inputs for the various columns such as temperature, dew point, pressure, etc. to predict the weather based on these attributes.[lr-1]

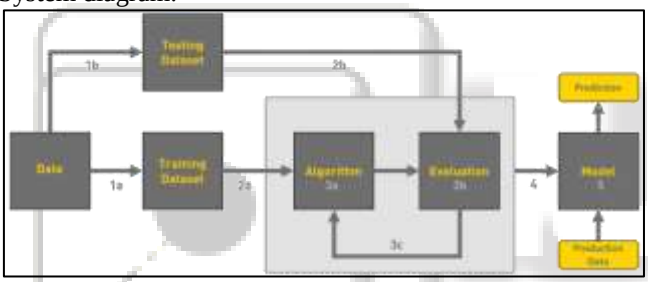
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.[lr-2]A linear regression line has an equation of the form : [lr-2]

$$Y = a + bX$$

Where,

X = explanatory variable
Y = dependent variable
b = slope of line
a = intercept (the value of y when x = 0).

System diagram:



IV. WORK OVERFLOW:

1) Gathering data
2) Data pre-processing
3) Researching the model that will be best for the type of data
4) Training and testing the model
5) Evaluation

V. ANALYSIS:

A. Total number and names of subdivisions:

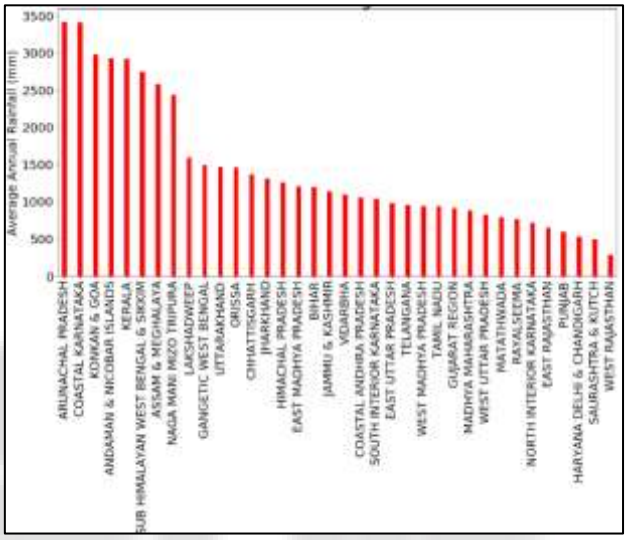
```
Total # of Subdivs: 36

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 & CHANDIGAR
H',
       'PUNJAB', 'HIMACHAL PRADESH', 'JAMMU & KASHMIR', 'WEST RAJASTHA
N',
       '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 NAD
U',
       'COASTAL KARNATAKA', 'NORTH INTERIOR KARNATAKA',
       'SOUTH INTERIOR KARNATAKA', 'KERALA', 'LAKSHADWEEP'], dtype=objec
t)
```

B. Average annual rainfall in each subdivision:

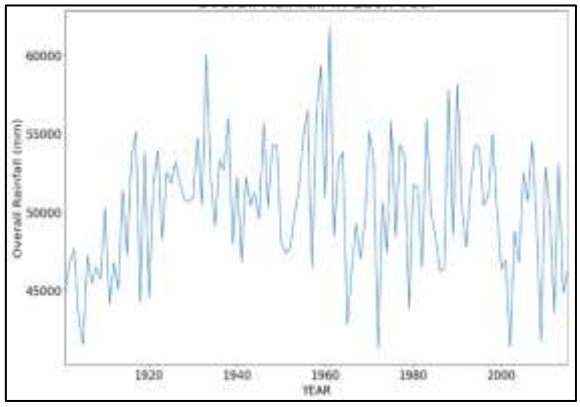
```
SUBDIVISION
ARUNACHAL PRADESH    3418.857143
COASTAL KARNATAKA    3408.409649
KONKAN & GOA         2977.686087
Name: ANNUAL, dtype: float64
SUBDIVISION
HARYANA DELHI & CHANDIGARH    530.496522
SAURASHTRA & KUTCH            495.161739
WEST RAJASTHAN                292.673043
Name: ANNUAL, dtype: float64
```

C. # Through graph:



- Subdivisions with highest annual rainfall are "Arunachal Pradesh", "Coastal Karnataka" and "Konkan & Goa" with approximate annual rainfall of 3418mm, 3408mm and 2977mm respectively.
- Subdivisions with lowest annual rainfall are "West Rajasthan", "Saurashtra & Kutch" and "Haryana Delhi & Chandigarh" with approximate annual rainfall of 292mm, 495mm and 530mm respectively

D. Overall rainfall in each year

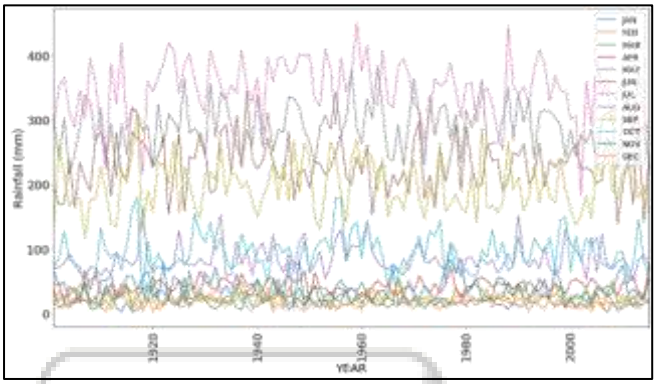


Table


```
Max: 61815.6 ocurred in [1961]
Max: 41273.6 ocurred in [1972]
Mean: 50182.83826086957
```

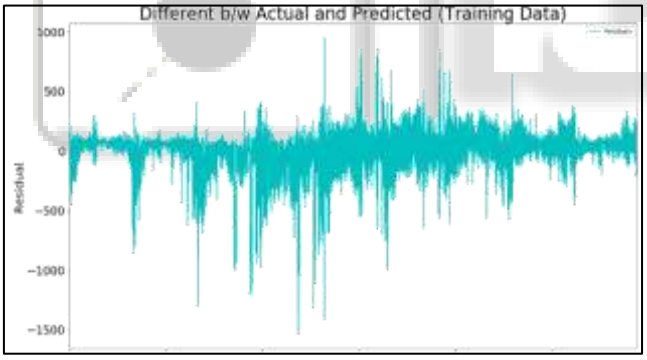
- Maximum overall rainfall (sum of all 36 subdivision) of 61815mm occurred in the year 1961.
- Minimum overall rainfall (sum of all 36 subdivision) of 41273mm occurred in the year 1972.
- Average (of all 36 subdivs.) overall rainfall (sum of all 36 subdivision) is 50182mm.

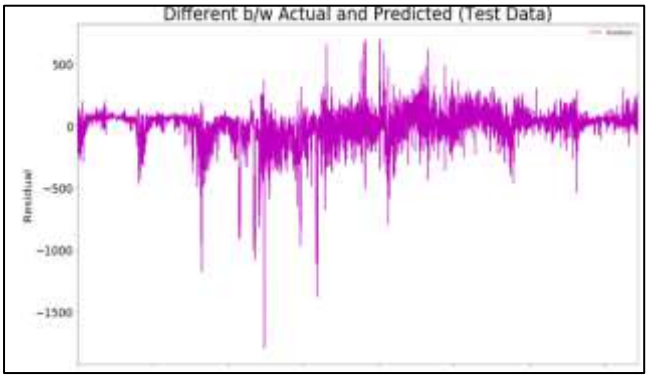
E. Overall rainfall in each months:



VI. On Applying Linear Regression:

A. Difference between actual and predicted Training and Test data:



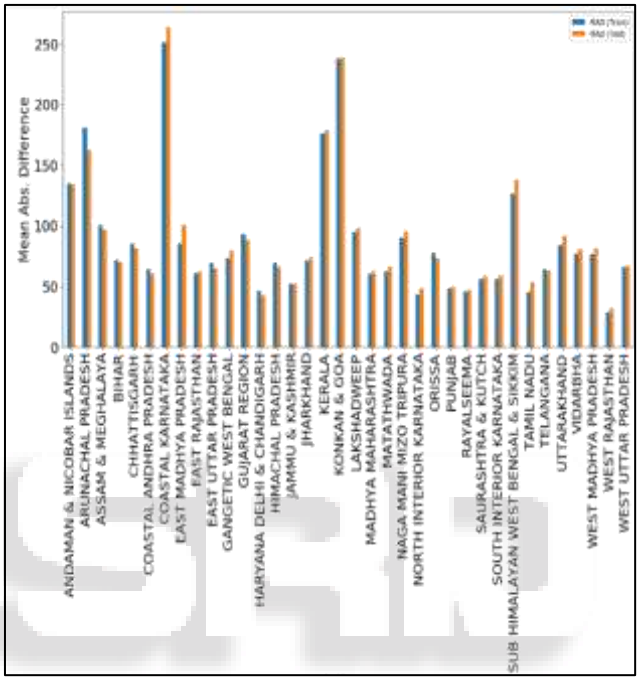


Table

```
MAD (Training Data): 88.9348725705
MAD (Test Data): 86.5573033039
```

There is 36 different linear models for each category corresponding to each subdivision. Firstly Data which have extracted for each subdivisions. For each month from April to December, data has appended to end of the data from. There have been four columns (one dependent variable and three independent variables). And for each subdivision the total number rows in the data come out to be 115*9 (approximately. 9 months (april to dec) appended at the bottom). Training/Testing is split in the ratio 80:20 randomly.

B. Mean absolute difference between Training and Test Data (Final Prediction):



Table

```
Overall MAD (Training): 85.954531315
Overall MAD (Testing): 88.67399873
```

VII. Literature Review:

Steve Oberlin, et.al (2012) proposed various Machine Learning strategies for the Big Data processing. He applied Machine Learning and various techniques from Artificial Intelligence to the complex and powerful data sets. Recommendation engines used by Netflix to see the rating and preferences of audience are one of the applications of Machine Learning. Informatics and Data Mining in which IBM's "Watson" uses different Machine Learning approach to process and depict human language and answer the queries [1]. Linear regression, massaging the data, Perception, k- means are the few strategies used by him for uncovering the relationships and finding patterns in data. The choice of Machine Learning algorithm basically depends on the nature of prediction. The prediction can be estimate type or classification. He also discussed how increasing features can make the algorithm complex and increasing computational requirements.

Jainender singh, et.al (2014) proposed machine learning technique that would be providing promising results to security issues faced in applications, its technologies and theories. He emphasized on mining from sparse, incomplete and uncertain data that would give optimized results when hidden patterns are discovered from the data sets using machine learning algorithms like Support Vector Machine (SVM), Naïve Bays classifiers, clustering techniques which are used to create supervised learning [4]. It would give insight knowledge in health, education, trade and many more fields.

Junfei Qiu, et.al (2017) proposed some of the latest advances of Machine Learning for processing Big Data. Representation Learning, a new advanced learning method in which data representation is useful and meaningful by extracting helpful information while constructing classifiers and predictors. It aims to capture vast input which would give computation as well as statistical efficiency. Feature selection, Feature extraction and Metric learning are the subtopic of Representation learning. Active learning is another advanced Machine learning method applied for big data.

Yasir Safeer, et.al (2010) presented Machine learning Algorithm i.e. k-means clustering for finding a document from a vast collection of unstructured text documents. He proposed a technique to portray documents that would be improving clustering result [3]. He discussed about the stream of document clustering, implemented k-means and devised an algorithm for better representation of documents and proposed how systematic domain dictionary would be used to get better similarity results of documents.
Roheet Bhatnagar, et.al (2018) presented about role of Machine Learning and Big Data Processing and Analytics (BDA). The development of Machine Learning and Big Data Analytics is complementary to each other. He discussed various future tends of Machine learning for Big data. Data Meaning implies how Machine Learning can be made more intelligent to acquire text or data awareness [5]. Technique Integration, another trend used to integrate data and process it. Classification, regression, cluster analysis are some of the techniques of Machine Learning which are used to perform analytics and predict future from existing patterns find correlation among the given data sets.

VIII. Summary and Conclusion:

This Paper has presented a supervised rainfall learning model which used machine learning algorithms to classify rainfall data. We used different machine learning algorithm to check the accuracy of rainfall prediction.

We have applied Linear Regression Technique from supervised Learning to the analysis over the Dataset for the prediction. From the above figure 3 we can conclude that Random forest is the Machine learning algorithm which is suitable for rainfall prediction in India.

Currently machine learning used in no. of industries. As the data increases the complexity of that data will increase and for that we are using machine for the better understanding of that data. In Weather predictions its pretty helpful with good accuracy score and in rainfall also its gives pretty good predictions.

In future, we are planning to increase our work in Storm predictions and Crop prediction with the rainfall prediction by using Forcasting Algorithm.

References

[1] https://ukdiss.com/examples/rainfall-prediction-machine-learning.php
[2] https://www.geeksforgeeks.org/machine-learning/
[3] http://www.stat.yale.edu/Courses/199798/101/linreg.htm
[4] https://www.kaggle.com/nasirmeh/prediction-of-rainfall
[5] https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94
[6] https://www.kaggle.com/

