Rainfall Prediction: Linear Regression

Harshal Shirke, Malhar Padval, Shantanu Bokey

Abstract – Rainfall prediction is important especially for countries like India where agriculture is the backbone of the economy. Since precipitation in India has been uneven over course of time and changes in the seasonality have resulted into loss of economy and livelihood.

The research deals with the study of various machine learning approaches to predict rainfall in accurate manner. Basically two rainfall prediction approaches are considered that is linear and non-linear models. Orthodox machine learning behave perfectly in case of linear trend however as the complexity of dataset increases the accuracy of linear models degrade. Thus to outgrow this ambiguity and identify non-linear patterns present in the data, we use classification algorithm like Logistic regression, Artificial Neural Network(ANN), Linear regression, Support vector machine(SVM). Dataset used for research is taken from Meteorological department.

Thus, this research helps to analyse rainfall patterns in the past and predict the precipitation.

Keyword- Machine learning, Data visualization, Linear Regression, Prediction.

I. INTRODUCTION

Rainfall has always been an important aspect for survival. In India, rainfall carries utmost importance since agriculture in India is predominantly dependent on rainfall. It accounts almost 55% arable land which is directly dependent on precipitation. Droughts are also major hazards faced by India due to uneven rainfall leading to the shortage of water resources. In order to practice rain water harvesting in such region it becomes necessary to predict rainfall. Also India has 7516 km long coastline which supports tourism and livelihood of half a billion population.

Thus the research can help farmers and governmental bodies to analyse the rainfall trend and manage water resources also it becomes necessary to know the amount of rainfall in coastal areas to take precautionary measures to minimize the loss of life and property in case of natural calamities.

Machine learning approach is used to solve this complex and intricate procedure to predict the rainfall. Machine learning employs gathering of historical data and predicting the future results by training a mathematical model depending upon types and trends in the dataset. Different machine learning models produce different results with corresponding accuracy hence it is important to choose right machine learning model and process the data accordingly.

There are two types of learning approaches i.e, a) Supervised Learning b) Unsupervised Learning. Supervised learning accounts labelled input and output data unlike unsupervised learning where human intervention is needed to label data. Thus supervised learning models are more accurate than unsupervised learning models. Supervised learning is further categorized into classification and regression. For large datasets Artificial Neural Networks (ANNs) is preferred since it produces more accurate results. It is popularly known as deep learning where each layer consists layer neural network and output of one layer is used as input to another layer.

In this research we have employed linear regression for predicting the rainfall depending upon the trends in recorded data. The data consists of annual & monthly distribution of rainfall from 1901-2015 across various subdivisions of India.

II. RELATED WORKS

There are numerous researches undertaken in the field of machine learning and data science to deal with complex data.

Advancement in this field includes wide range of technology. However a standard machine learning pipeline typically includes:

Algorithm selection

Data pre-processing

Feature engineering

extraction Feature

Hyper-parameter tuning

III. PROPOSED METHODOLOGY

The research employs linear regression for predicting the rainfall in more accurate manner.

Dataset:

The dataset consists of annual & monthly distribution of rainfall from 1901-2015 for every state in India. The dataset in total has 19 attributes namely subdivision, year, month, annual. Rainfall in India is classified into 36 subdivisions.

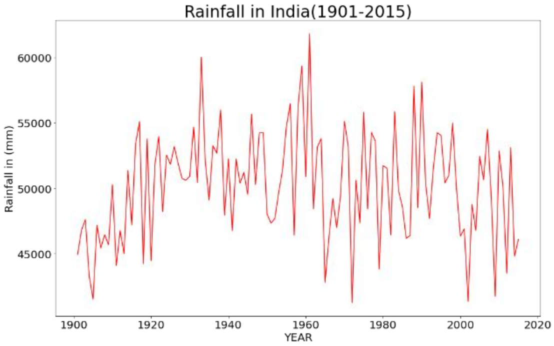

Fig. 1. Rainfall from 1901-2015

The above graph gives an overview about the annual rainfall in India from 1901-2015.

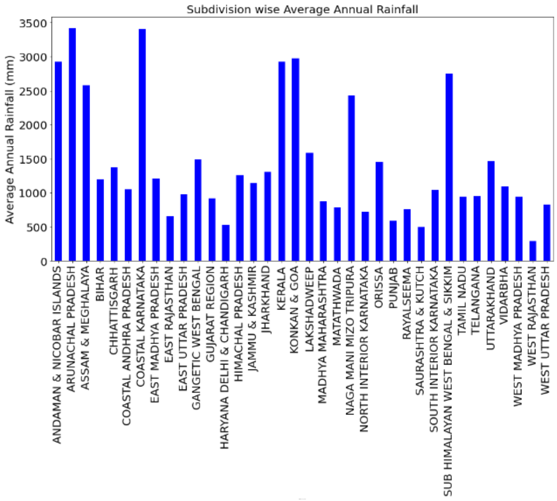

Fig. 2. Subdivision wise average annual rainfall

Data Processing :

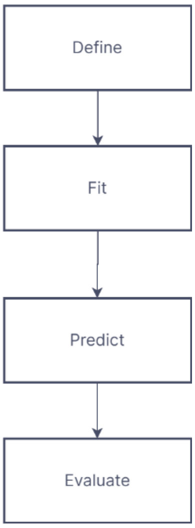

Fig. 3. Machine Learning approach

Data processing is important since the accuracy of the model depends upon the data. The data can be categorized into four types : numerical data, categorical data, time-series data and text.[1] Also we need to deal with missing values in order to increase the effectiveness of the model. In this research we have used Simple Imputer (i.e. Imputation) to replace missing values with mean value. We can also replace missing values with zero. [2]

Categorical values are the values that take only finite number of distinct values. There are various approaches to eliminate categorical values, In this research one hot encoding is used.[3] Label Encoder is used to convert string to numerical values. [4] Prediction target(y) and prediction features(X) are specified.[5] The data is split into training and testing data with the help of sklearn library. The data is split in the ratio 8:2 (80% training data & 20% testing data) increasing accuracy.[6] Finally, the linear regression model is employed for prediction. Model is fit with training data and predictions are made with input data.

Following is the heatmap that represents the relations between the features.

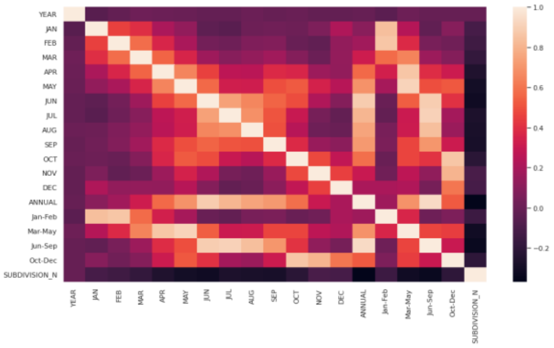

Fig. 4. Correlation heatmap

Linear Regression:

Linear regression is a model that assumes linear relationship between the input variables and single output variable i.e, it establishes relationship between dependent and independent variable. It shows how the dependent variable changes according to the value of independent variable. Mathematically Linear Regression can be represented as

$$y= a_0+a_1x+ \varepsilon$$

where y = Dependent Variable

X = Independent Variable

a_0 = intercept of line

a_1 = Linear regression coefficient

Linear Regression is further divided in two types :

Simple Linear Regression: Single independent variable is used to predict the value of numerical dependent variable.

Multiple Linear Regression: One or more independent variable is used to predict the value of numerical dependent variable.

In this model we have employed Multiple Linear Regression where multiple inputs (Subdivision, rainfall in months) are given to the model to predict the output (Annual Rainfall).

Model Validation:

Model validation is the last step that evaluates the model and determines the accuracy. There are many metrics for summarizing model quality such as confusion matrix, F1 Score, Mean absolute error, Mean squared error, Root mean squared error. In this research we have considered R2 Score for model validation. Mathematically it is represented as

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

R2 score is a statistical measure that represents the proportion of the variance for a dependent variable that’s explained by an independent variable.

Architecture of the proposed model :

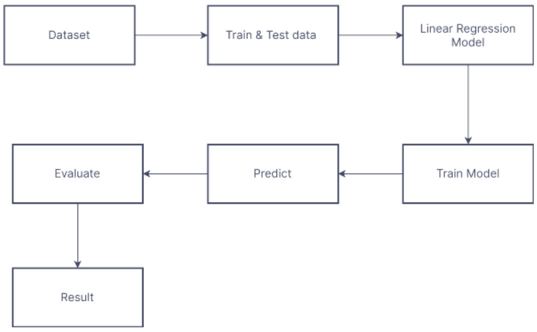

Fig. 5. Architecture of the model

IV. EXPERIMENTAL RESULTS

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. It is mostly used for finding out the relationship between variables and forecasting.

User gives the required data as an input to the system. Linear regression model will be initiated and accordingly the results will be generated

Following are the that represent annual rainfall for corresponding months.

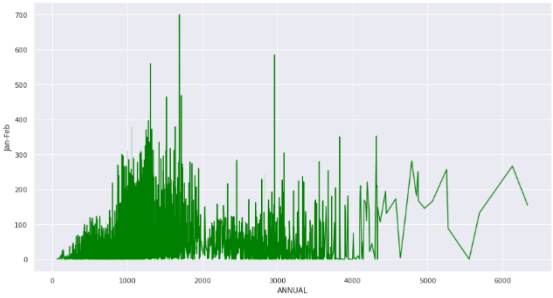

Fig. 6. Annual rainfall in Jan-Feb

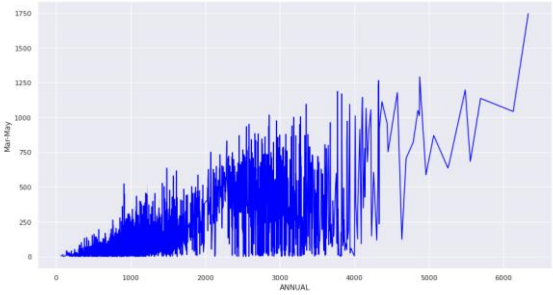

Fig. 7. Annual rainfall in Mar-May

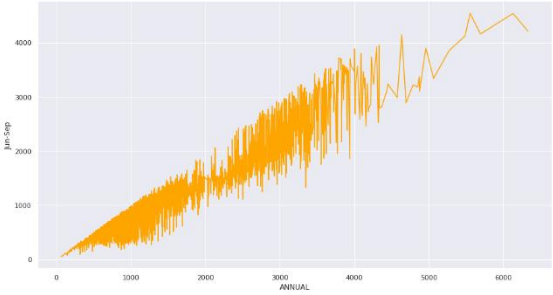

Fig. 8. Annual rainfall in Jun-Sep

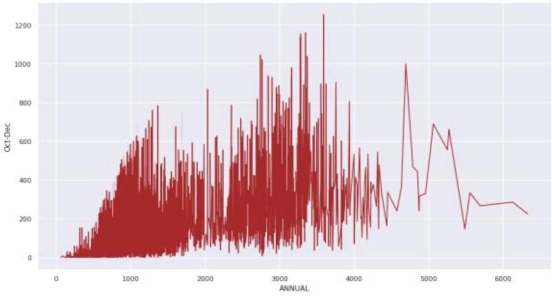

Fig. 9. Annual rainfall in Oct-Dec

. Following table represents the model validation parameters which indicates the accuracy of the model.

Table 1. Validation Parameters

Parameter	value
R2 score	0.99
Mean square error	0.84318
Mean absolute error	0.14836

The R2 score value tends to 1 which indicates the correctness of the model. Also the other parameters such as MSE & MAE whose value is close to zero validate the correctness.

V. CONCLUSION

Thus the research is an attempt to forecast precipitation using linear regression. Data visualization of dataset is performed to analyse the trends and pattern of the data. Prediction target and features were determined. Dataset is split into train and test data. Further model is fit with Linear Regression model We employed limited features still the results have acceptable accuracy. After prediction the model is validated and thus confirmed accuracy. Though rainfall depends on various factors still we achieved satisfactory results.

Further enhancement of the project can be weekly report of rainfall prediction. This would give beforehand idea about rainfall with adequate amount of time in hand to take actions. In order to increase the accuracy more errors can be calculated to test accuracy of linear regression model. Having more accuracy would help in better management of resources.

REFRENCES

1. Mathew, S., Saravanan, M. Improvising an automation reference tool GRASP using active data handling approach , Proceedings of 2015 IEEE 9th International Conference on Intelligent Systems and Control, ISCO.
2. Ireland, G., Volpi, M., & Petropoulos, G. P. (2015). Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: a case study from a Mediterranean flood. Remote sensing, 7(3), 3372-3399.
3. Narasimha Prasad, Prudhvi Kumar, and Naidu Mm. An approach to prediction of precipitation using gini index in sliq decision tree. In Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on, pages 56{60. IEEE, 2013.
4. Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203{224, 2007.
5. Rob J Hyndman. Moving averages. In International Encyclopedia of Statistical Science, pages 866{869. Springer, 2011.
6. Sarah N Kohail and Alaa M El-Halees. Implementation of data mining techniques for meteorological data analysis. Intl. Journal of Information and Communication Technology Research (JICT), 1(3), 2011.
7. Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, and Dong Hyun Jeong. Designing a rule-based hourly rainfall prediction model. In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pages 303{308. IEEE, 2012.

AUTHORS

Shantanu Bokey : He is currently pursuing his second year of B.Tech in Department of Computer Science & information Technology from Rajarambapu Institute of Technology, Sangli. His area of interest includes programming and computer vision.

Malhar Padval : He is currently pursuing his second year of B.Tech in Department of Computer Science & information Technology from Rajarambapu Institute of Technology, Sangli. His area of interest includes programming and cloud computing.

Harshal Shirke: He is currently pursuing his second year of B.Tech in Department of Computer Science & information Technology from Rajarambapu Institute of Technology, Sangli. His area of interest includes programming and Machine Learning.