Rainfall Prediction Using LASSO Regression

Jinashree P, Pooja D R, Meghana M V, Aishwarya G P, SIDDHARTH B K
Student, Student, Student, Student, Assistant Professor,
Computer science Department
B G S Institute of technology B G Nagara, Mandya(di), Karnataka, India.

Abstract: In Today's era global warming is affecting all over the world which majorly effect on mankind and cause the expedite the change in climate. Rainfall prediction model mainly based on artificial neural networks have been proposed in India until now. This research work does a comparative study of two rainfall prediction approaches and finds the more accurate one. The present technique to predict rainfall doesn't work well with the complex data present. The approaches which are being used now-a-days are statistical methods and numerical methods, which don't work accurately when there is any non-linear pattern. Existing system fails whenever the complexity of the datasets which contains past rainfall increases. To find the best way to predict rainfall, study of both machine learning and neural networks is performed. The paper investigates the performance of the various Machine Learning (ML) models, namely Lasso regression, Back propagation and Liner Regression. Those fashions performances had been calculated thru the assessment metrics which include R^2 score, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). Rainfall is considered the primary source of most of the economy of our country. Agriculture is considered the main economy driven source. To do a proper investment on agriculture, a proper estimation of rainfall is needed. Along with agriculture, rainfall prediction is needed for the people in coastal areas. People in coastal areas are in high risk of heavy rainfall and floods, so they should be aware of the rainfall much earlier so that they can plan their stay accordingly. For areas which have less rainfall and faces water scarcity should have rainwater harvesters, which can collect the rainwater. To establish a proper rainwater harvester, rainfall estimation is required. Weather forecasting is the easiest and fastest way to get a greater outreach. This research work can be used by all the weather forecasting channels, so that the prediction news can be more accurate and can spread to all parts of the country. The aim of this study is to compare different machine learning regression algorithms in rainfall dataset.

Keywords – Artificial neural network, Machine learning, Rainfall prediction.

I. INTRODUCTION

Rainfall is one the most significant atmospheric occurrence that is not only useful for the environment itself but for all the living beings on the earth. It affects everything directly or indirectly and because it is one of the most important natural phenomena. This project deals with the accuracy of rainfall using machine learning & neural networks. The project performs the comparative study of machine learning approaches and neural network approaches then accordingly portrays the efficient approach for rainfall prediction. First of all,

Pre-process is performed. Pre-process is the process of representing the dataset in the form of several graphs such as bar graph and histogram.

When this comes to machine learning, LASSO regression is being used and for neural network, ANN (Artificial neural network) approach is being used. After calculation, types of errors, accuracy of both LASSO and ANN has been compared and accordingly conclusion has been made. To reduce the systems complexity, the prediction has been done with the approach that has better accuracy. The prediction has been done using the dataset which contains rainfall data from year 1901 to 2015 for different regions across the country. It contains month wise data as well as annual rainfall data for the same.

Currently, rainfall prediction has become one of the key factors for most of the water conservation systems in and across country. One of the biggest challenges is the complexity present in rainfall data. Most of the rainfall prediction system, nowadays are unable to find the hidden layers or any non-linear patterns present in the system. This project will assist to find all the hidden layers as well as non-linear patterns, which is useful for performing the precise prediction of rainfall. Rainfall prediction is the application to predict the rainfall in a given region. It can be done in two types. The first is to analyze the physical law that affects rainfall and the second one is to make a system which will discover hidden patterns or the features that affects the physical factors and the process involved in achieving it. The second one is better because it doesn't include any type of mathematical calculations and can be useful for complex and non-linear data.

Due to presence of the system which doesn't locate the hidden layers and nonlinear styles accurately, the prediction effects to be incorrect for maximum of the instances and that could cause large losses. So, the main objective for this research work is to find a system that can resolve both the issues i.e. able to find complexity as well as hidden layers present, which will give proper and accurate prediction.

The remainder of the paper is structured as follows: Section 2 anlayze the state of the art while Section 3 explains the existing methodology. Section 4 introduce a new methodology and reports the performance of the proposed system, Section 5 includes the result part where we come across many graphical outputs, and finally Section 6 concludes and discusses future directions for further improvements to the research.

II. LITERATURE REVIEW

Machine learning technique offers with predicting rainfall the use of device learning technique. There are 3 predominant developments of device studying are getting used. The first one is known as hybridization, because of this that more than one device studying procedures are getting used collectively and consequently prediction is being executed. The 2d one offers with enhancing the first-class of dataset that's getting used. The 3d one is to apply of ensemble of approach for growing the cap potential of the approach. One of the predominant blessings of this machine is its to boom the first-class of set of rules and dataset.

A few vital findings on this vicinity of examine were posted during the last ten years or so on. The author's discussed that Accuracy of rainfall announcement has first-rate significance for nations like India whose financial system is largely depending on agriculture. The dynamic nature of environment, carried out arithmetic strategies fail to offer realistic accuracy for precipitation announcement. Landslides are one of the predominant geo risks chargeable for the massive lack of assets worldwide. It employs Artificial Neural Networks to expecting at some point improve rainfall depth after which assesses the hazard of landslide prevalence with the aid of using and evaluating it with rainfall threshold. Artificial strategies like Single layer Feed-Forward Neural Network (SLFN) and Extreme Learning Machine (ELM) have been used for expecting the summer time season monsoon rainfall as heavy rainfall takes place throughout state. The proposed SLFN community prediction version offers 6.3977% of suggest absolute blunders while ELM offers 3.8729% of blunders. Rainfall prediction is an vital and tough project in meteorology. Rainfall may be expected the use of numerous device studying strategies, right here Artificial Neural Network (ANN) including Feed Forward Neural Network (FFNN) version is constructed for predicting the rainfall. The prediction accuracy is measured the use of confusion matrix and RMSE. It recognize the importance of modifications in weather and environment parameters like precipitation, temperature, humidity. Precipitation estimate is one of the vital investigations in discipline of meteorological research. The prediction enables human beings to take preventive measures and furthermore the prediction must be correct. There are varieties of prediction quick time period rainfall prediction and long time rainfall. Prediction ordinarily quick time period prediction can offer us the correct result. The predominant task is to construct a version for long time rainfall prediction. The prediction of precipitation the use of device studying strategies can also additionally use regression. The specializes in the non-linear device studying procedures like gradient boosting choice tree version and deep neural networks for a quick time period prediction of rainfall and the effectiveness is calculated with the aid of using the use of category metrics AUC, F1 score, precision and accuracy and with the aid of using Regression metric RMSE, correlation and Data Visualisation .The statistics visualisation styles includes the highest, lowest and common rainfall within side the States/Union Territories the climate of India has been visualised. Intention of this mission is to provide non-specialists smooth get entry to to the strategies, procedures applied within side the quarter of precipitation prediction and offer a comparative examine most of the numerous device studying strategies. We determined that the accuracy may be improvised by using Lasso regression method instead of using older technology and methods.

III. METHODOLOGY

The predictive model is used to prediction of the precipitation. The first step is converting data in to the correct format to conduct experiments then make a good analysis of data and observe variation in the patterns of rainfall. We analyze the rainfall by separating the dataset into training set and testing set then we apply different machine learning approaches and statistical techniques and compare and draw analysis over various approaches used. With the help of numerous approaches we attempt to minimize the error.

3.1 LINEAR REGRESSION

Linear regression approach deals with finding a relation between dependent and nondependent variables. This approach can give a good estimation of rainfall for a certain period of time. It deals with collection of data set and set it for further processing. The data's collected are further processed and result is predicted. One of the advantage of using this approach is that it is better than data mining approaches. Data mining approach gives generalized value unlike linear regression which gives estimated value. The biggest disadvantage of this approach is that it fails when it comes for long term estimation.

- **Hypothesis function for Linear Regression**

$$Y = \theta_1 + \theta_2.X$$

x: input training data

y: labels to data

θ_1: intercept

θ_2: coefficient of x

Cost Function (J): By calculating the best-match regression line, the version targets to are expecting y fee such that the mistake distinction among expected value and authentic value is low.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

3.2 BACKPROPAGATION:

Backpropagation is a supervised gaining knowledge of algorithm, for training Multi-layer Perceptrons. The back-propagation technique works well with less complex system, but as the complexity of the system increases back propagation method's accuracy decreases. This process deals with four types of inputs and three types of outputs layers. The four-input layer used are: Air temperature, Air humidity, Wind speed and Sunshine duration.

The three-output layers used are: Rainfall, Medium rainfall and High rainfall

3.3 LASSO

The word "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression plays L1 regularization, which provides a penalty identical to absolutely the fee of the significance of coefficients. The goal of this algorithm is to minimize the number of errors.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where,

- λ denotes the quantity of shrinkage.
- λ = 0 implies all functions are taken into consideration and it's miles equal to the linear regression in which best the residual sum of squares is taken into consideration to construct a predictive version.
- λ = ∞ implies no characteristic is taken into consideration i.e, as λ closes to infinity it removes an increasing number of functions.
- The bias increases with increase in λ
- variance increases with decrease in λ

IV. IMPLEMENTATION

This system is developed using Python programming with Anaconda framework.. Datasets are collected from data.gov.in. Entities associated with the architecture are user input data, preprocess, LASSO, neural network, splitting of data, training of the algorithm, testing of the data and we obtain the result at the final step. User gives data to the system from the local system. After that User does the preprocess step and then compares the LASSO regression and neural network approach. The user finds the accuracy of all the algorithm and perform prediction with the most accurate one. For both LASSO and neural network, accuracy is calculated after algorithm is being trained. The accuracy is being calculated after calculating errors. The errors calculated are MAE, MSE, R-SQUARED and RMSE.

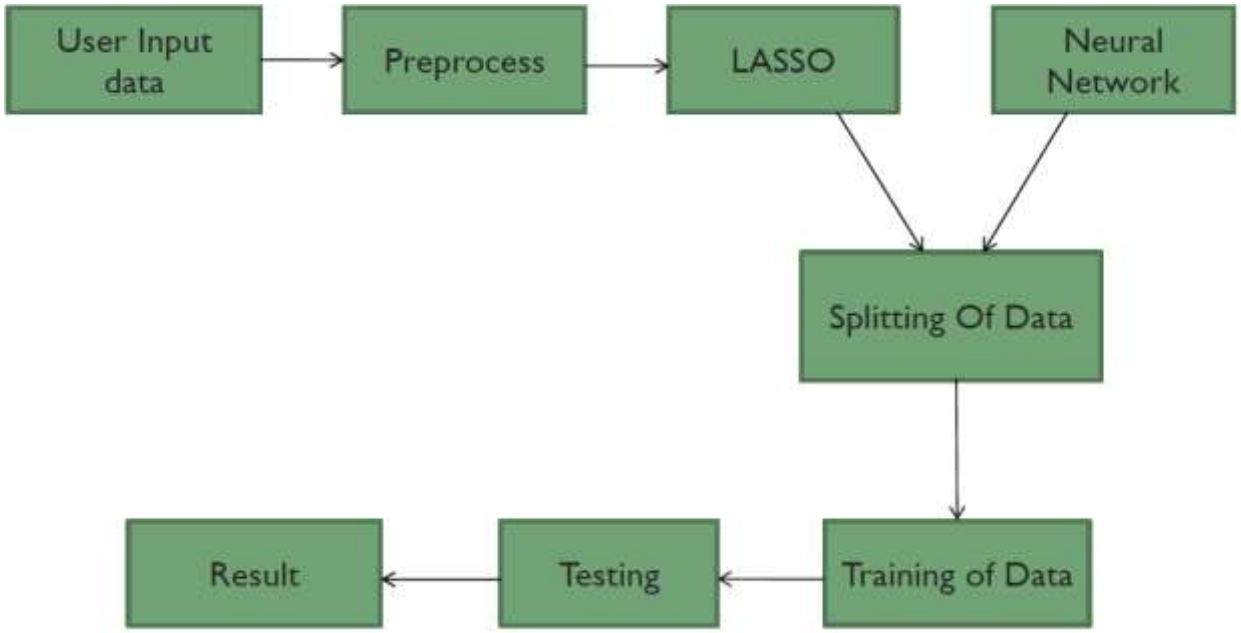

Fig.1: System Architecture

To reduce the complexity of the system, prediction will only be done for the most accurate algorithm because if both the algorithms are giving predicted value, then that can lead to unnecessary usage of data and that will increase the complexity of the system and make it slow. The result will be received in the form of graphs and metrics. The graphs received are Correlation Metrics, Scatter Metrics, Max value for month by month, Subdivision mean value for month by month, Sum of every quarter of subdivision, Sum of every quarterly, Sum of year by year and Sum of month year by year.

4.1 Errors Calculated

The accuracy of the approaches is being calculated against the types of errors that can produce negative effect on the algorithm. These errors can affect the algorithm's accuracy and hence are being calculated. The types of errors that is being calculated are MAE, MSE, RMSE and R-SQUARED.

MAE calculates all the absolute errors and then finds the mean value for all. It first calculated the mean of all the dataset present, then subtract the mean value with each data individually and add all the resultant value and finally divides it with the total number of dataset which we are having.

$$MAE = \frac{1}{n}\sum |xi - x| \qquad (1)$$

Next error is MSE. It is almost similar to mean absolute error.

$$MSE = \frac{1}{n}\sum (xi - x)^2 \qquad (2)$$

The only difference is, instead of adding the resultant (subtracted value of mean with each dataset), it finds the square of it and add them. RMSE error is being calculated by subtracting all the predicted and actual values with each other, finding all the squares of it and adding all the squared

$$RMSE = \sqrt{\frac{\sum (xi-yi)2}{N}} \qquad (3)$$

(xi = Predicted value yi = actual value)

The total value that we will receive is stored and this value is further divided by total values present. The resultant value is squared rooted.

4.2 Dataset

The dataset used in this system contains the rainfall of several regions in and across the country. It contains rainfall from 1900 – 2020 for the same. Along with that annual rainfall is also been used and the rainfall between the transition of two months. The dataset is been collected from data.gov.in.

Category – Rainfall in India
Released under – NDSAP
Contributor – Ministry of Earth Sciences, IMD
Group – Rainfall
Sectors – Atmosphere science, earth sciences, science & technology
Source: OGD

V. RESULTS & ANALYSIS

User gives the dataset as input in the system. First we pereform preprocess which represent the dataset in the form of graphs, the second one is LASSO which gives the accuracy of LASSO regression and the third one is neural network which gives the neural network ‘s accuracy. So, to have a better understanding of the dataset and for better comparison, Before going for the prediction, preprocess can be done. Dataset should be split in two parts, the first part deals with training the algorithm used and the rest part used to predict the amount of rainfall. Rainfall is predicted only with the algorithm with more accuracy. Training is done in both the approaches i.e. on LASSO and ANN. This step gives a proper idea of which algorithm is more accurate among the two. Then the remaining dataset is being used and rainfall prediction is been done. Finally, after the all the process is completed, the result is received in form of graph and table which shows the future rainfall and the accuracy of the algorithm. After preprocess the graphs which are received are Correlation Metrics, Max value for month by month, Subdivision mean value for month by month, Sum of every quarter of subdivision, Sum of every quarterly, Sum of year by year and Sum of month year by year. After that for each LASSO and neural network, the accuracy is acquired within side the shape of Metrics and excel sheet . In Metrics along with the accuracy different types of errors are also shown and the same is represented in the form of graphs.

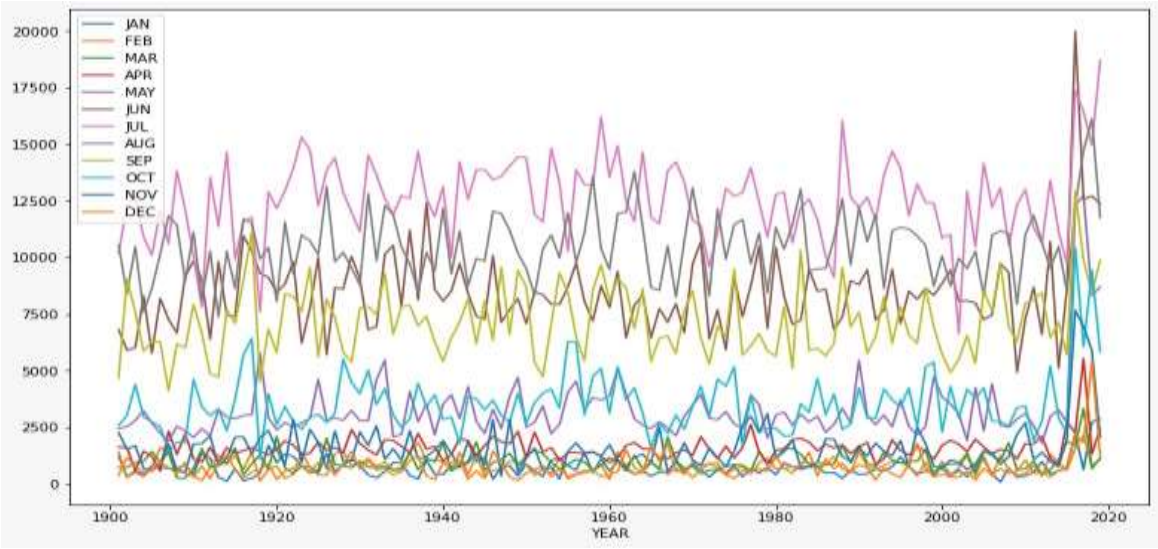

Fig 5.1. Sum of Month Year by Year

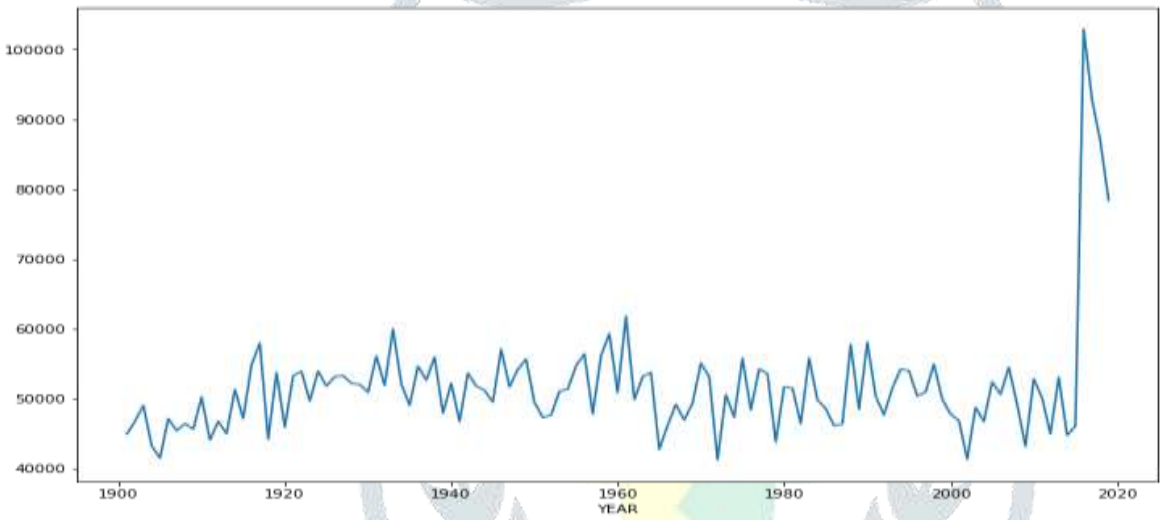

Fig 5.2. Sum of Year by Year

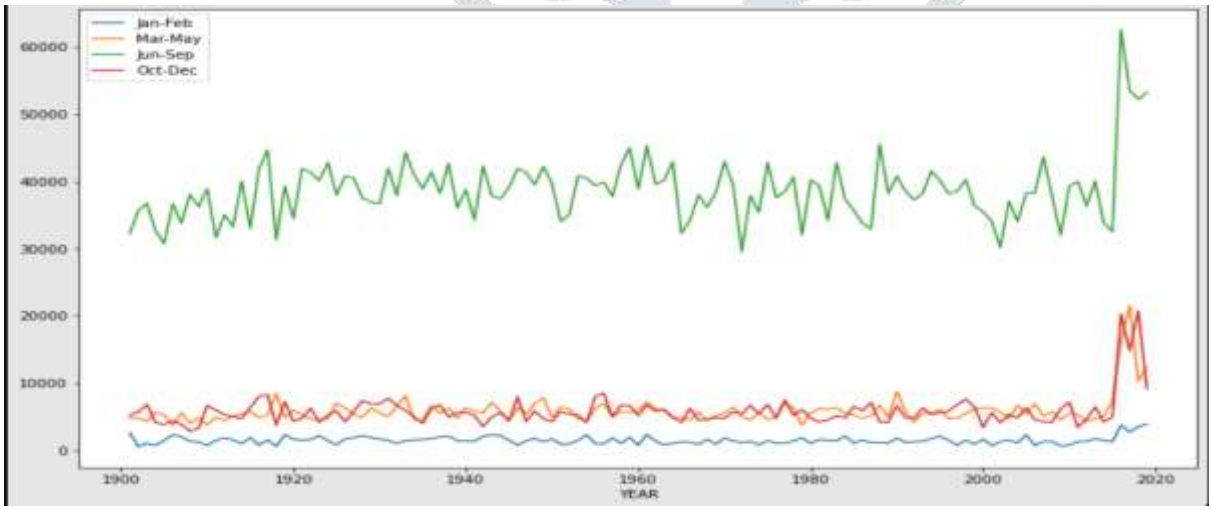

Fig 5.3. Sum of every quarterly

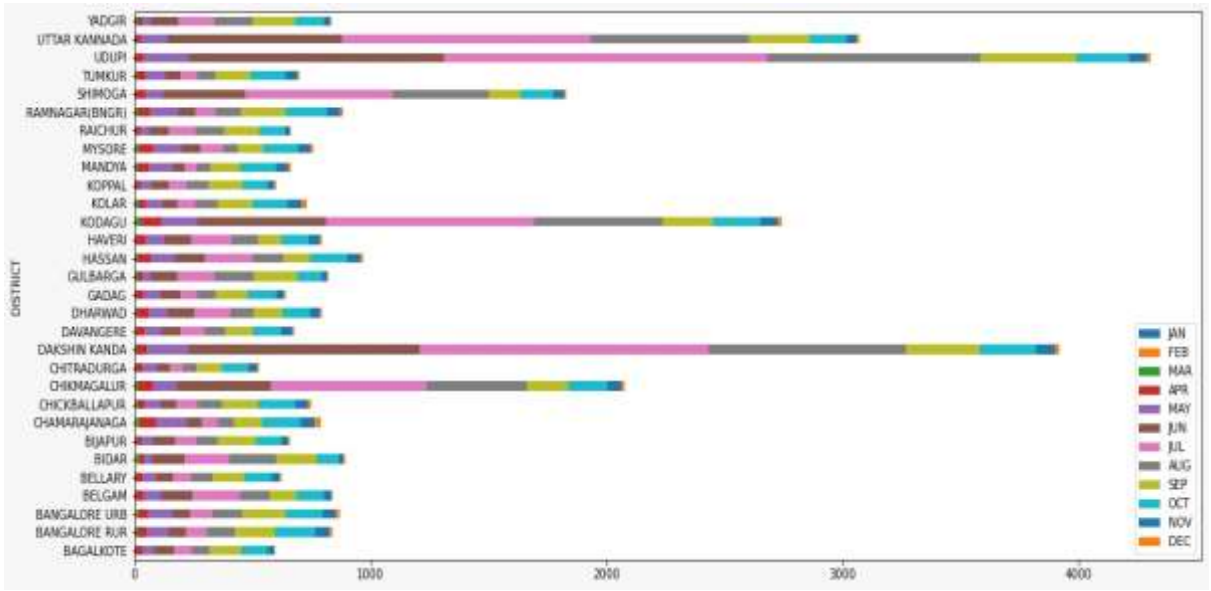

Fig 5.4. Predictions in Districts of Karnataka

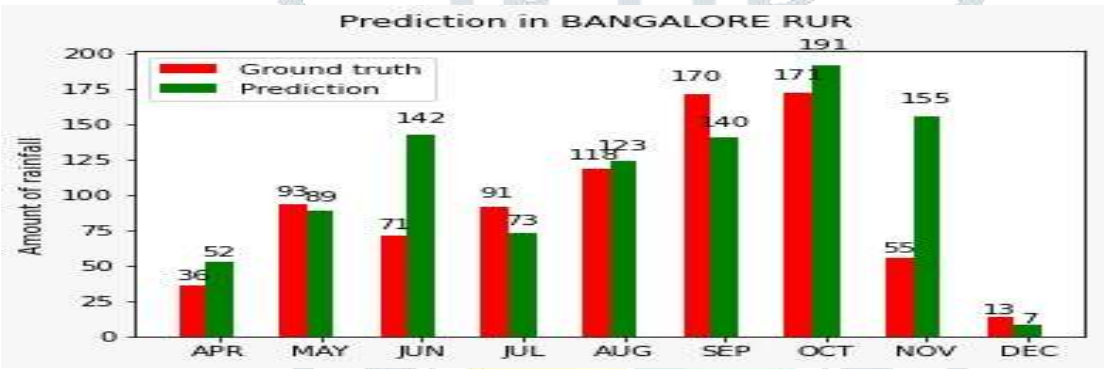

Fig 5.5. Prediction in Bangalore Rural

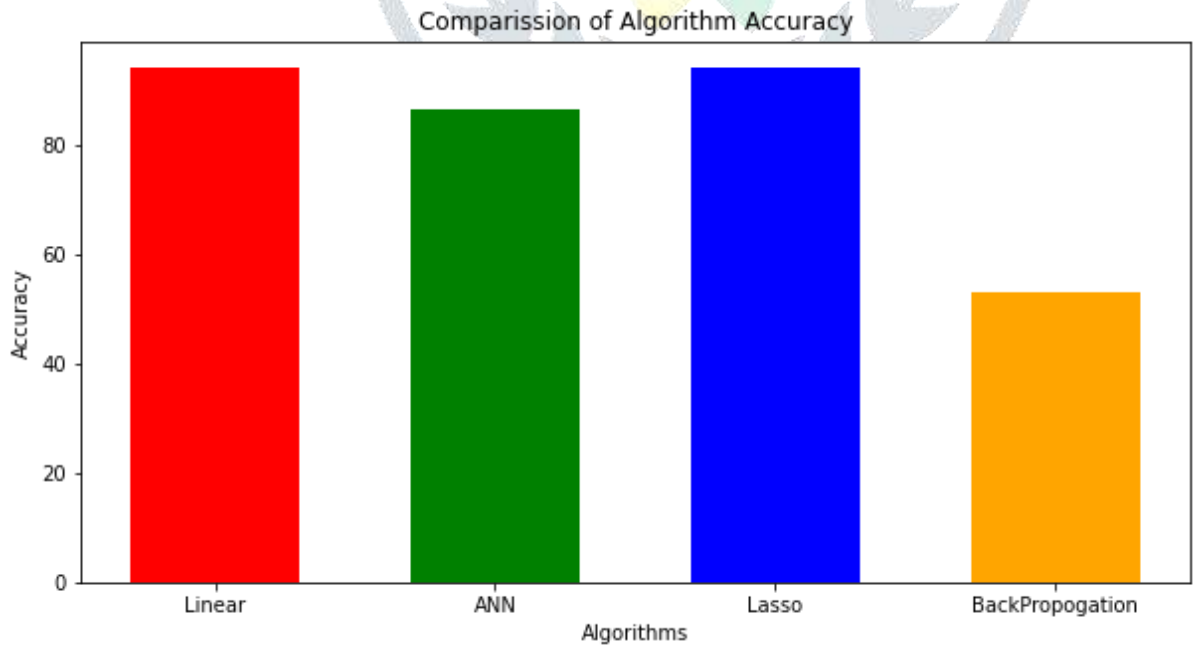

Fig 5.6. The Final Output in the form of Graph

VI.CONCLUSION AND FUTURE SCOPE

Rainfall being one of the sole responsibilities for maximum economy of India, it should be considered the primary concern for most of us. The current approach for rainfall prediction fails in most of the complex cases, as it is unable to predict the hidden layers present, which is yet to be recognized for performing the precise prediction. To achieve an effective way to predict rainfall, two ways are being compared. One is machine-learning approach and the second one is artificial neural networks approach. Initially, LASSO regression approach is being taken. Dataset is divided into two parts namely, train data and test data. Train data is for training the algorithm and test data is for doing the prediction. Both of these processes were compared based on their accuracy and along with that, error types such as MSE, MAE, R-SQUARED and RSME were considered. The one with more accurate was considered and prediction was performed with that approach itself. The rainfall was predicted from that data used for testing as part of the data being used to train the algorithm. After performing the comparison, the conclusion of the system is that LASSO regression process is more accurate than the artificial neural network process. After comparison, we got to understand that the accuracy for LASSO is around 94% whereas ANN is 77%. Therefore, LASSO is the best analytical algorithm for predicting the rainfall in any given region.

The future enhancement of this project can be an approach towards about how to reduce the percentage of errors present. Along with that one of the major enhancements will be to decrease the ratio for train data to test data, so that it will assist in improving the level of prediction within the available time and complexity. The accuracy of the algorithm can be additionally tested on increase in the complexity. Many other types of errors can be calculated in order to test the accuracy of any of the above algorithms. Henceforth, algorithm for testing daily basis dataset instead of accumulated dataset could be of paramount Importance for further research. Along with that, this will be an efficient tool for people in coastal areas of the country thereby making them well aware of the situation in advance.

REFERENCES

[1] Janani, B; Sebastian, P. Analysis of the rainfall prediction and techniques. International Journal of Advanced Research in Computer Engineering & Technology, 3(1), 59–61.

[2] Puneet Sharma and Nadim Chishty, "Machine Learning-Based Modelling of Human Panther Interactions in Aravalli Hills of Southern Rajasthan", Indian Journal of Ecology.

[3] A.El-shafie " Performance of artificial neural network and regression techniques for rainfall prediction", International Journal of the Physical Science vol 6(8).

[4] Aakash Parmar, Kinjal Mistree, M. S. (2020). Machine Learning Techniques for rainfall prediction: A Review. International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS).

[5] Chaudhari, M. S., & Choudhari, N. K. (2019). Study of Various Rainfall Estimation & Prediction Techniques Using Data Mining. American Journal of Engineering Research (AJER), 7, 137–139

[6] https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017.

[7] Aftab, S., Ahmad, M., Hameed, N., Bashir, M. S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction using data mining techniques: A systematic literature review. International Journal of Advanced Computer Science and Applications, 9(5), 143–150

[8] Kavitha Rani, B., & Govardhan, A. (2018). Effective Features of Rainfall Prediction. International Journal of Computational Intelligence Systems, 7(5), 937–951.

[9] Prabakaran, S., Naveen Kumar, P., & Sai Mani Tarun, P. (2017). Rainfall prediction using modified linear regression. ARPN Journal of Engineering and Applied Sciences, *12*(12), 3715–3718. Vol.51, Issue.1, pp.5-10, 2017.

[10] Etuk, E. H., & Mohamed, T. M. (2017). Time Series Analysis of Monthly Rainfall data for the Gadaref rainfall station, Sudan, by Sarima Methods. International Journal of Scientific Research in Knowledge, July, 320–327

[11] Kar, K., Thakur, N., & Sanghvi, P. (2019). Prediction of Rainfall Using Fuzzy Dataset. International Journal of Computer Science and Mobile Computing, 8(4), 182–186.

[12] Zeyi Chao, Fangling Pu, Yuke YinLing, B. and X. (2018). Research on real-time local rainfall prediction based on MEMS sensors.