International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

ANALYSIS AND PREDICTION OF RAINFALL USING MACHINE LEARNING TECHNIQUES

Anurag Kumar[1], Lalsingh Chouhan[2]

[1]*Assistant professor, CSE, Dr. A.P.J. Abdul Kalam UIT Jhabua, Madhya Pradesh, India*
[2]*Assistant professor, CSE, Dr. A.P.J. Abdul Kalam UIT Jhabua, Madhya Pradesh, India*

Abstract - *: Weather and climate prediction are dominated by high dimensionality, interactions on many different spatial and temporal scales, and chaotic dynamics. Machine learning techniques can predict rainfall by extracting hidden patterns from historical weather data. In this technique apply the Multiple Linear regression (MLR) and Support vector regression (SVR) model for rainfall prediction. To design and implement the system, we have gathered 115 years of data from 1901 to 2017 from Kaggle. Our proposed model has been tested and validated with respect to Multiple Linear regression and Support Vector regression. Compared results reveal the satisfactory performance, the SVR had provided maximum accuracy*

Key Words: **Rainfall prediction, Machine Learning, Linear Regression, Support Vector Regression, Accuracy**

1. INTRODUCTION

Weather forecasting on the basis of historical data is a complex but very helpful task. Which comes with several problems that require to be solved in order to achieve optimal result. Rainfall prediction is important all over the world and it play a key role in human life. It's difficult to predict rainfall precisely with varying atmosphere conditions. Accurate rainfall predictions are crucial for several areas of society specially in agriculture. India is an agricultural country and therefore the success of agriculture depends of rainfall. There are several recourses for water but in India agriculture is usually dependent on rainfall. The weather has a significant impact on the agricultural industry and because of that, having the ability to predict it helps farmers in their day-to-day decisions such as how to plan efficiently, minimize costs and maximize yields. The concept of machine learning is getting used in every sector to reduce the labour cost and increase the productivity. Every Machine learning algorithm has three steps: Depiction, judgment, development. Depiction guides us to represent the discovered knowledge done from the data mining. Here we have used the two most popular machine learning techniques to predict the rainfall. Those techniques are Support Vector regression and Multiple Linear Regression Linear Regression [1][12] is very useful for finding relationship between two continuous variables, one is independent variable and another is dependent variable. In Statistics, Linear regression refers to a model which show relationship between two variables and how one can impact the other. In Linear Regression, it shows how the variation in the "dependent variable" can be captured by change in the "independent variables". Linear Regression is statistical technique which used to generate insights on consumer behaviour, understanding business and factors influencing profitability. Linear regressions can be used in business to evaluate the trends and make decision for future. For example, if an organisation's sales have increased regularly every month for the last few years, by conducting linear analysis on the sales data with monthly sales, the company could forecast sales in future months. We have used Multiple linear regression model, unlike simple linear regression MLR has multiple independent variables. SVR is a regression algorithm, so we can use SVR for working with continuous Values instead of Classification which is SVM [2]. In regression technique we try to minimise the error rate while in SVR we try to fit the error within a certain threshold.

2. RELATED WORK

There are many works in the literature for the prediction of rain fall. This section discusses some of the work related to our proposed methodology.

Kumar Abhishek et al. have proposed a rainfall prediction technique using neural network in [3]. The proposed model in [3] predicts the rainfall of Udupi district from Karnataka state of India. BPNN with feed forward, layer recurrent and BPNN with cascade feed forward neural networks are experimented. The proposed model takes 70% of the data for training and 30% for testing. The recurrent network gives better accuracy when compared to BPNN. The MSE is high in BPNN

Nasimul Hasan, Nayan Nath (2015) this paper exhibits a robust rainfall prediction technique in view of recent rainfall data of Bangladesh using Support Vector Regression (SVR), a relapse methodology of Support Vector Machine (SVM). It was challenging to make a 100 percent perfect prediction and the data was preprocessed manually

to suit the algorithm [5]. The evaluation results of the study conducted on the data shows that the projected technique performs higher than the conventional frameworks in term of accuracy and process running time [5]. Approach yielded the utmost prediction of almost 99.92%.

G.Mahalakshmi and S.Sridevi (2016) presented a paper which gives detailed survey of the various techniques applied for forecasting different types of time series. This survey covers the overall forecasting models, the algorithms used within the model and other optimization techniques used for better performance and accuracy [6]. The various performance evaluation parameters used for evaluating the forecasting models are also discussed in this paper [6]. This study gives the reader an idea about the various researches that take place within forecasting using the time series data.

Paper proposed by [7] introduced rainfall prediction system using deep mining KNN technique. A single K value is given which is used to find the total number of nearest neighbours that helps to determine the class label for unknown data. Similar parameters are clustered into same type of cluster and thus with the help of KNN we determine the category of a specific datasets. This algorithm does not require time for training of classification or regression. This system may not lead to good accuracy if the incorrect value of K is picked.

Sandeep Mohpatra and Animaka Upadhyay (2017) presented a paper that focuses on use of data mining techniques for predicting rainfall of an area on basis of some dependent features like precipitation and wet day frequency. They have collected data for years ranging from 1901 to 2002 of Bangalore, India [8]. The regression model developed has been trained and validated against the actual rainfall of that area. The performance of the algorithm was further boosted using Ensemble techniques using k-fold [8].

Chandrasegar, K S Harsha (2017) carried experiment on a heuristic prediction of rainfall using machine learning techniques. This paper discusses the rate of rainfall in previous years according to various crop seasons like Rabi, Kharif and Zaid and predicts the rainfall in future seasons [9]. Also, it measures the different categories of data by linear regression method. Results help farmers to make correct decision to harvest a particular crop according to crop seasons. Linear regression method suggests the lower correlation between various crop seasons [9].

3. METHODOLOGY

In this paper we have used Multiple Linear regression and Support Vector regression to predict the amount of rainfall.

3.1 Machine Learning Model

The proposed method is based on the multiple linear regression and support vector regression. The data for the prediction is collected from the publicly available sources and the 70 percentage of the data is for training and the 30 percentage of the data is used for testing. Figure 1 describes the block diagram of the proposed methodology. Multiple regression is used to predict the values with the help of descriptive variables and is a statistical method. It is having a linear relationship between the descriptive variable and the output values. The following is the equation for multiple linear regression:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon$$

Here we are using "k" for the number of predictor variables and we have k+1 regression parameters Where, β_0 is constant term, β_1 variable is coefficient of x_1, β_2 variable is coefficient for x_2, β_k is x_k coefficient variable and ε is error associated with predicted value. Support Vector Regression (SVR) uses the same principle as SVM,

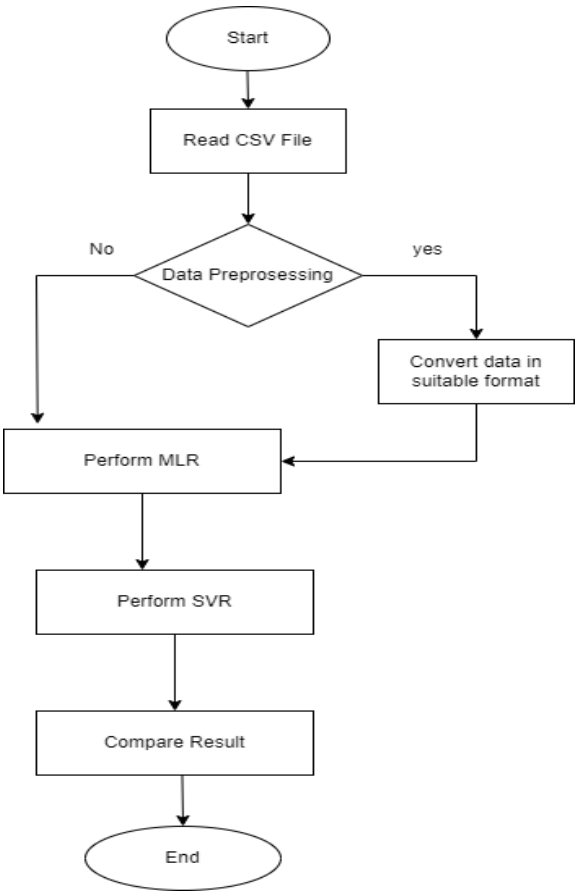


Figure 1. block diagram of the proposed methodology

but only for regression problems. SVR works on the principle of structural risk minimization from statistical learning theory [2] and establishes a hyperplane that can predict the distribution of data. The principle of the SVR algorithm has a given set of input training data set *{(Ai, Bi), i = 1, 2,...,k}*, $x_i \in R^M$, where A_i is the input 3-D vector, $B_i \in R$ is the response output data, and *k* is the number of samplings. The optimal linear decision function in the high-dimensional feature space is expressed as follows:

$$f(x_I) = \omega A + b$$

where ω refers to weight vectors and *b* denotes the bias.

3.2 Data and Sources of Data

For this study data has been collected from the publically available source Kaggle [11], it contains the monthly rainfall of each state of India form Jan 1901 to Dec 2017.

3.3 Performance metrics

Mean Absolute Error:

MAE [10] is the average of the absolute differences between the actual value and the model's predicted value. The bigger the MAE, the more serious the error is.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| y_i - \hat{y}_i|$$

Mean Squared Error:

MSE or Mean Squared Error[10] is one of the most popular metrics for regression algorithms. It is simply the average of the real value's squared difference with the regression model's predicted value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Absolute Percentage Error:

MAPE or Mean Absolute Percentage Error [10] is the average absolute difference between the actual value and the value predicted by the model divided by the real value

$$MAEP = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Root Mean Squared Error:

RMSE or Root Mean Squared Error[10] is similar to MSE, just the final value is square rooted and calculated the square of errors in MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R^2 Error:

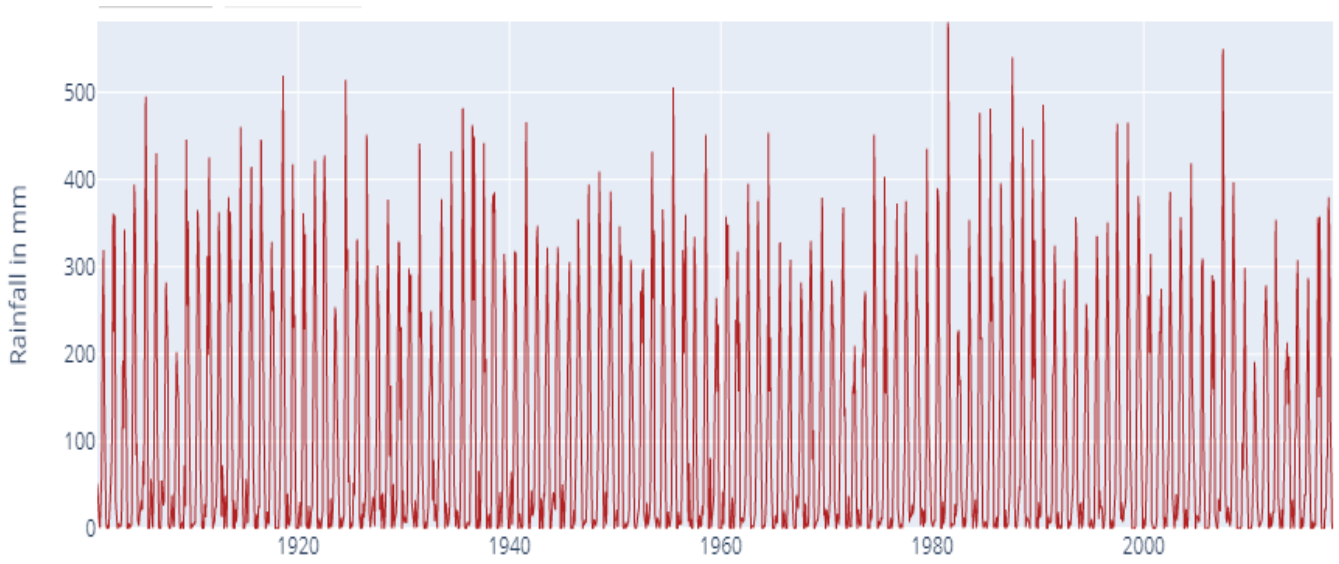
R^2 or Coefficient of Determination is a prevalent metric [10]. R^2 uses two mean squared error calculations. While the first is the mean square of each real value versus the average of observations, the second is the mean squared error of the actual value versus the predicted one.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

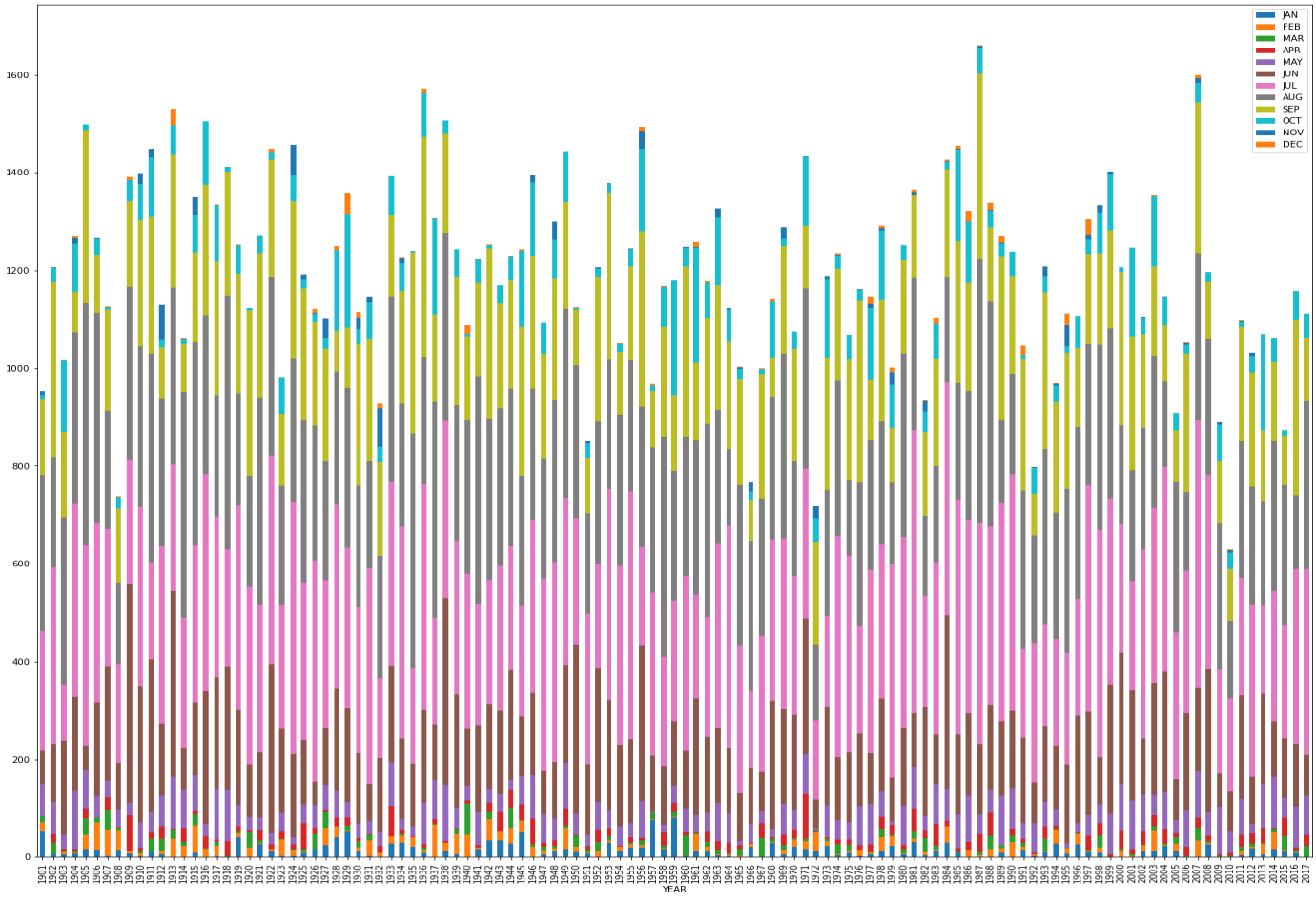
4. RESULTS AND DISCUSSION

We have considered dataset from 1901 to 2017 of Bihar state, India. Visualization from different graphs help us to understand more about the data and drives us to decide the next step to taken. It provides important perceptions.

Forecasting gives appropriate and reliable input regarding to present, past and future activities with definite numerical and scientific methods. There are some steps involved in predicting the numerical values for a specific task. Initial step is to recognize the problem with complete analysis and second is collecting the appropriate data to analyze the problem for further estimation. After estimation, compare the actual and estimated values with necessary actions. The data is arranged in such a way that rainfall is plotted according to year i.e., yearly counts of rainfall shown in graph 1.



Graph 1. Overall monthly data plot of rainfall from 1901 to 2017



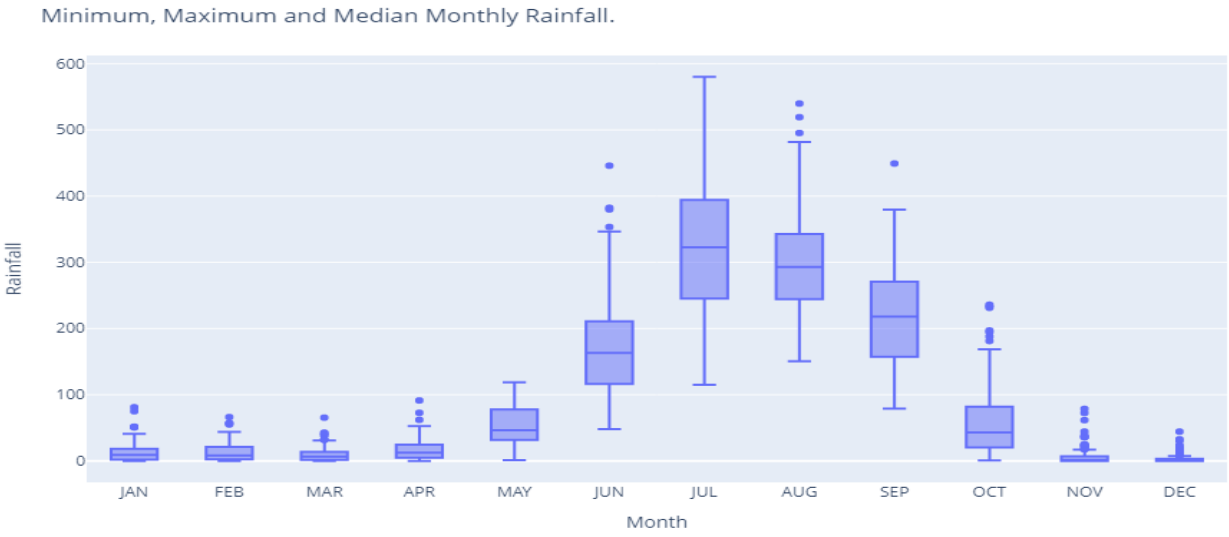
Graph 2: Stacked bar chart of each year

The stacked bar chart above depicts the rainfall of each month for a particular year, across each month. we can see from the sorted overall bar heights that year 1987 has the highest rainfall and 2010 has lowest.

Graph 3 show the minimum, maximum and madian rainfall in each month using box plot. It clearly indicates that,

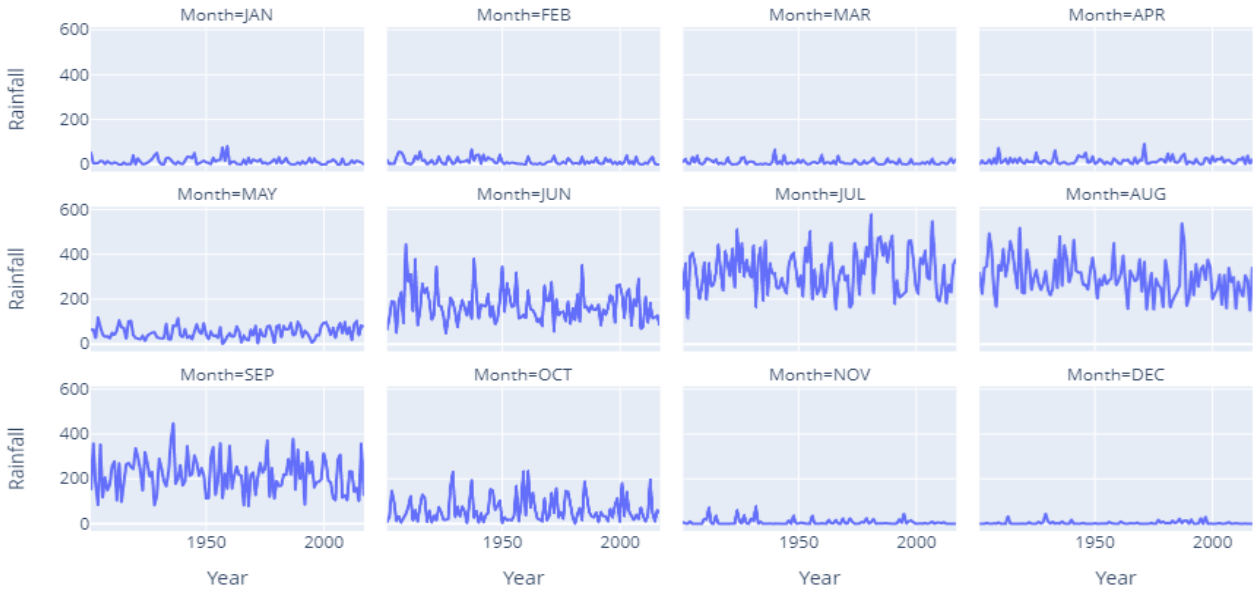
- The rainfall in the months January, February, March, April, November and December is very less.
- The rainfall in the months May and October is average.
- The rainfall in the months June, July, August, and September are high compared to rainfall in other months of the year.

We can see a seasonal effect with a cycle of 12 months.



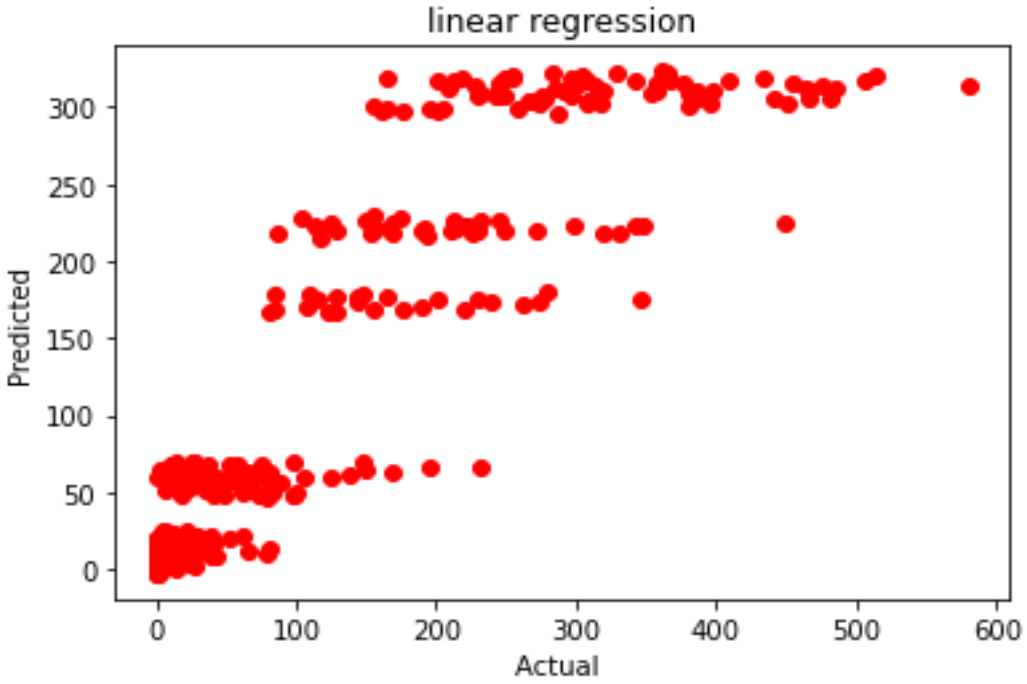
Graph 3: Box Plot graph describing the rainfall in each month.

Graph 4. shows rainfall in each month from 1901 to 2017.

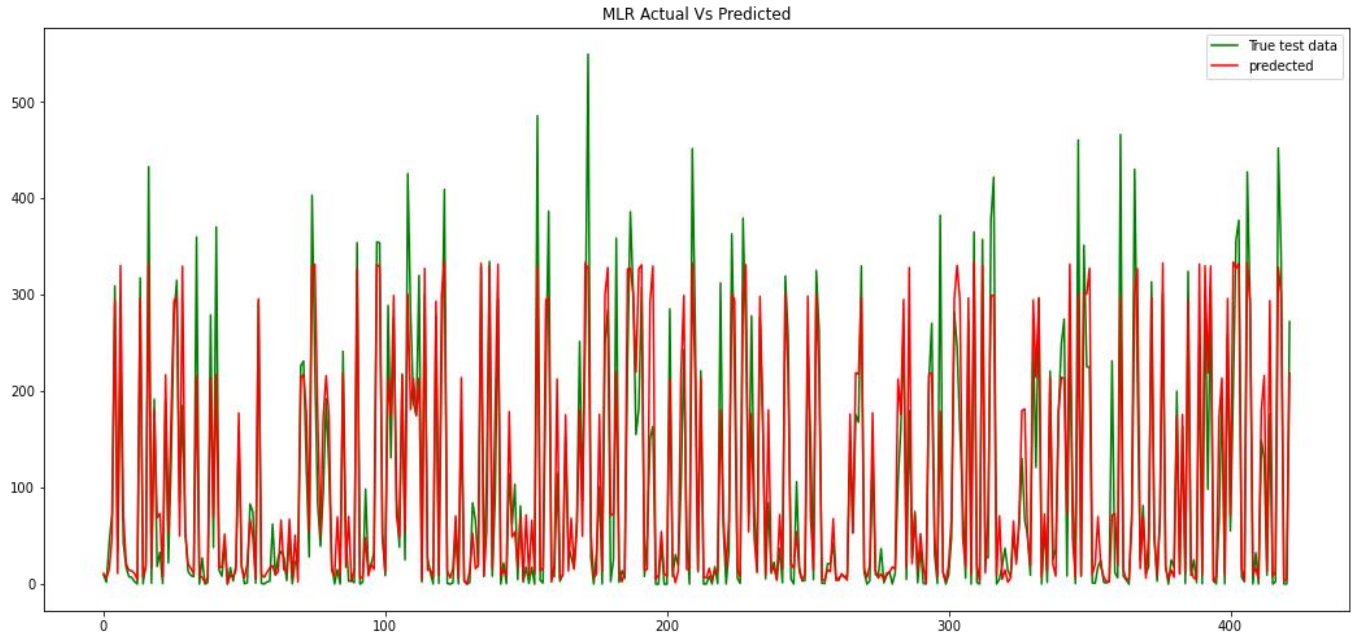


Graph 4: Monthly rainfall through history.

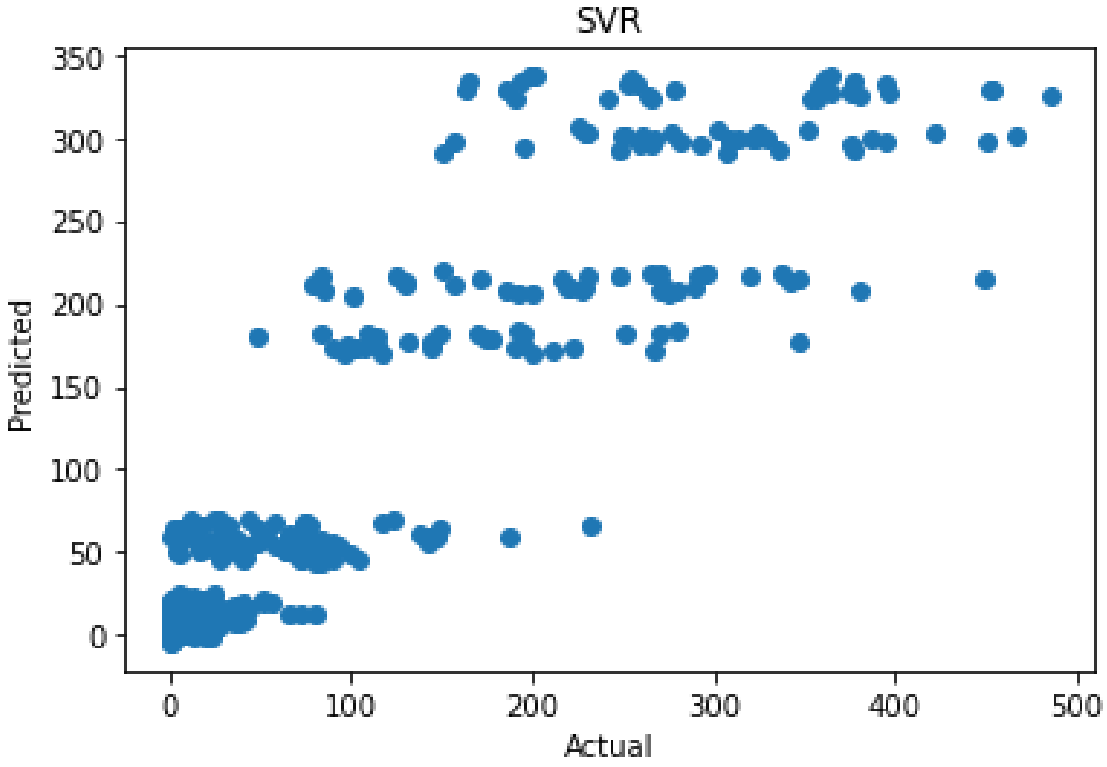
we divide the data into Train and Test Sets: Number of entries (training set, test set): (982, 422) Now we compare the MLR and SVR model to understand which model gives better result. Splitting the dataset into train and test data we have taken 70% for training and 30 % for testing the model. A total of 982 train data and 422 test data is used. Plotting the scatter plot of actual and predicted rainfall we get the following graphs. Graph 5 shows the scatter plot of actual vs predicted rainfall using MLR model. In Graph 6 we can clearly see the comparison between each actual and predicted value.



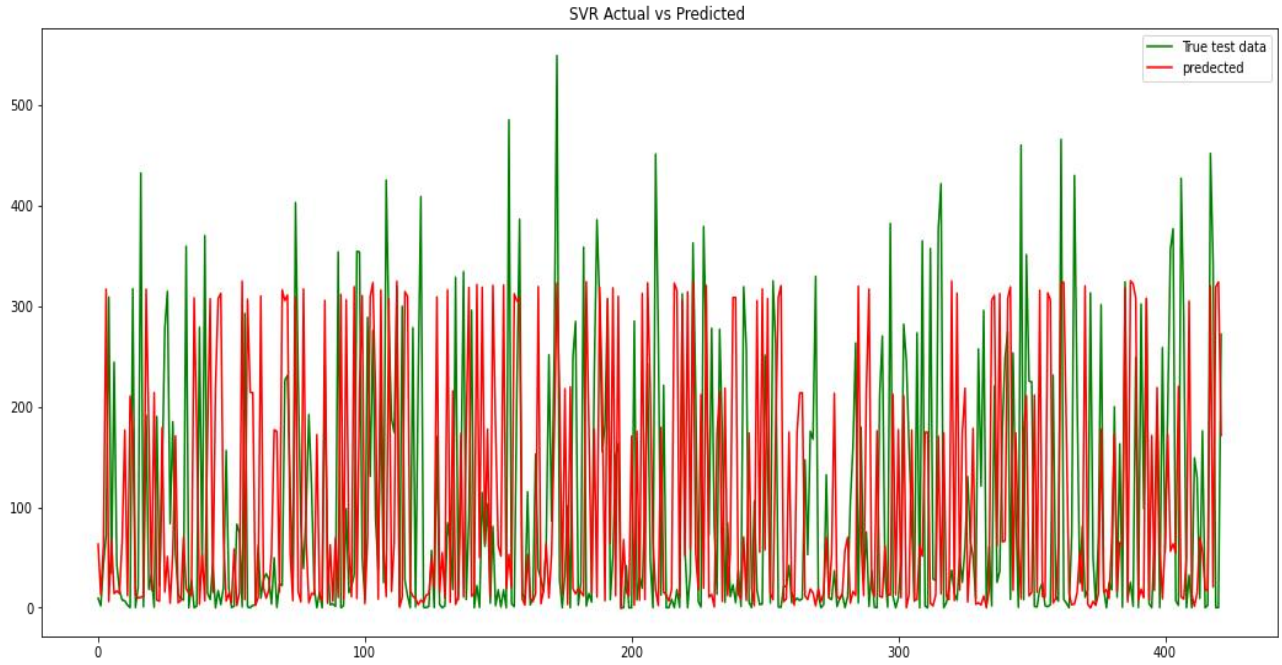
Graph 5: Scatter Plot of Actual vs Predicted rainfall using MLR



Graph 6: Actual vs Predicted rainfall using MLR



Graph 7: Scatter Plot of Actual vs Predicted using SVR



Graph 8: Actual vs Predicted rainfall using SVR

Graph 7 shows the scatter plot of actual vs predicted rainfall using SVR model.

In Graph 8 we compared each actual and predicted value.

Table 1 Comparison of MLR and SVR Performance

S. No.	matrices	LR	SVR
1	Train score	0.827149236683498	0.828585087676299
2	Test score	0.811790903215551	0.836907038442013
3	MAE	0.249229855446679	0.248383868743224
4	MSE	0.154636268612835	0.151737104047022
5	RMSE	0.393238183055556	0.389534470935528
6	MAPE	0.674927107614465	0.614009940253406
7	R2 Score	0.833790903215551	0.836907038442013

5. CONCLUSION

Here, we are using time series analysis to predict the rainfall using monthly rainfall from year 1901 to 2017. For yield to accuracy, machine learning algorithms such as MLR and SVR, were implemented and tested on the given datasets from the Bihar states. Both algorithms are compared with their accuracy. Comparing the different performance matrices, we can conclude that SVR accuracy is better than MLR.

REFERENCES

[1] Amanpreet Singh , Narina Thakur , Aakanksha Sharma "A review of supervised machine learning algorithms", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) IEEE Oct 2016.

[2] Mariette Awad, Rahul Khanna, "Support Vector Regression, "Efficient Learning Machines Theories, Concepts, and Applications for Engineers and System Designers" Apress (pp.67-80), January 2015.

[3] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, Sarthak Kumar, "A Rainfall Prediction Model using Artificial Neural Network", 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC 2012), pp. 82-87, 2012.

[4] H. M. Meighani, C. Ghotbi, T. J. Behbahani, and K. Sharifi, "Evaluation of PC-SAFT model and support vector regression (SVR) approach in prediction of asphaltene precipitation using the titration data," Fluid Phase Equilibria, vol. 456, pp. 171–183, Jan. 2018.

[5] Nasimul Hasan, Nayan Chandra Nath, Risul Islam Rasel, "A Support Vector Regression Model for Forecasting Rainfall", Proceeding of International Conference on Electrical Information and Communication Technology (EICT 2015), IEEE, 554 -559.

[6] G.Mahalakshmi, Dr. S. Sridevi, Dr. S. Rajaram, "A Survey on Forecasting of Time Series Data", IEEE, 2016.

[7] Zahoor Jan, Muhammad Abrar, Shariq Bashir and Anwar M Mirza, "Seasonal to interannual climate prediction using data mining KNN technique", International Multi-Topic Conference, pp. 40-51, 2008.

[8] Sandeep Kumar Mohpatra, Anamika Upadhyay, Channabasava Gola, "Rainfall Prediction Based on 100 years of Meteorological Data", International Conference on Computing and Communication Technologies for Smart Nation(IC3TSN),IEEE,2017,162-166.

[9] Chandreshekhar Thirumalai , M. Laxmi Deepak, K Sri Harsha, K Chaitanya Krishna, " Heuristic Prediction of Rainfall using Machine Learning Techniques", International

[10] Ravish Raj "Evaluation Metrics for Regression Models in Machine Learning" https://www.enjoyalgorithms.com/blog/evaluation-metrics-regression-models

[11] https://www.kaggle.com/datasets/saisaran2/rainfall-data-from-1901-to-2017-for-india

[12] Thirumalai, C., Harsha, K. S., Deepak, M. L., & Krishna, K. C. (2017). Heuristic prediction of rainfall using machine learning techniques. 2017 International Conference on Trends in Electronics and Informatics (ICEI)