JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Hybrid Machine Learning Model for Rain fall Prediction

[1]Jeyadevan S, [2]Subha V, [3] Dr. Ganganagunta Srinivas
[1]Lecturer, Computer Engineering, University of Technology and Applied Sciences-Ibra, Oman
[3]Lecturer, Mechanical Engineering, University of Technology and Applied Sciences-Ibra, Oman

Abstract: Predicting the amount of rain is crucial since excessive downpours can trigger numerous disasters. In addition to assisting people in taking preventative action, the prediction should also be accurate. Short-term and long-term rainfall predictions are the two different categories of predictions. Most of the time, short-term predictions can provide us with accurate outcomes. Developing a model for long-term rainfall prediction is the primary obstacle. Because heavy precipitation is intimately related to the economy and human lifespan, it could be a significant disadvantage for the Earth Science Department. It is the reason behind the annual natural calamities like drought and flood that affect people all over the planet. It is the reason behind the annual natural calamities like drought and flood that affect people all over the planet. Since agriculture is the main driver of the economies of nations like India, the accuracy of the rainfall statement is very important. Because of the dynamic nature of the environment, precipitation statements cannot be reasonably accurately predicted using applied mathematics techniques. Regression may be used in machine learning approaches for precipitation prediction. The goal of this project is to provide a comparative analysis of the different machine learning algorithms and to make the techniques and approaches used in the field of precipitation prediction easily accessible to non-experts.
Key terms: Weather prediction, Rainfall prediction, Naïve bayes classifier.

I. INTRODUCTION

Rainfall forecasting is crucial because heavy and erratic rainfall can have a variety of negative effects, including damaging crops and farms and causing property damage. Therefore, an improved forecasting model is necessary for early warning systems that reduce the risk to people and property and improve agricultural farm management. This forecast primarily benefits farmers and allows for the effective use of water resources. Predicting when it will rain is a difficult process, and the outcomes need to be precise. Numerous hardware tools are available for forecasting rainfall based on meteorological factors including temperature, humidity, and pressure. Because these conventional approaches are inefficient, we can provide accurate results by applying machine learning techniques. We may simply do this by using past rainfall data analysis to forecast the amount of rainfall in next seasons. Numerous approaches, such as regression and classification, can be used based on the needs. The accuracy and error between the prediction and the real can also be calculated. Various methods yield varying degrees of accuracy, thus selecting the appropriate algorithm and modeling it in accordance with the specifications is crucial.

Regression analysis:

Regression analysis examines how one variable, referred to as the dependent variable, is dependent on one or more independent factors, referred to as the independent variables. This relationship can be used to estimate and/or predict the mean or average value of the dependent variable in terms of fixed or known values of the independent variables. For instance, a person's pay is determined by his or her experience; in this case, the experience attribute is the independent variable and the compensation is the dependent variable. The link between one dependent variable and one independent variable is defined by simple linear regression. An essential technique for information analysis and modeling is regression analysis. It is employed in predictive analysis, which includes anticipating weather patterns and rainfall as well as corporate, financial, and marketing trends. It can also be applied to give quantitative support and error correction.

II. LITERATURE SURVEY

Predicting the amount of rain is now crucial for both industry and agriculture. Accurate rainfall forecasting can identify heavy downpours and maybe provide alerts and details about impending tragedies. Due to inconsistent weather data, a number of methods were created and put into practice to forecast rainfall, but they were not very accurate.

In this study, we have attempted to construct a rainfall prediction model that will yield rainfall predictions with a notable degree of accuracy by implementing the Naïve Bayes technique. The Naïve Bayes method for rainfall prediction shows a notable degree of accuracy and acceptance, according to experimental data involving different performance indicators.[1]
Predicting when it will rain is a useful skill, but it can be difficult. By extracting and combining the hidden knowledge from the linear and non-linear patterns of historical meteorological data, machine learning algorithms can forecast rainfall using computational approaches. Although there are many tools and techniques available for rain prediction, accurate results are still lacking. Whenever large datasets are utilized to predict rainfall, current methods are failing. This study offers two machine learning algorithms that are good in predicting rainfall: logistic regression and random forest. Both strategies produce simple, accurate predictions, and it is possible to compare their relative effectiveness.[2]
Predicting when it will rain is one of the most challenging and uncertain tasks and has a significant impact on human civilization. Accurate and precise forecasts will aid in proactively mitigating escalating financial and human risks. This paper presents state-of-the-art supervised machine learning models with an emphasis on rainfall prediction. Because it affects every part of the earth that depends on humans, rainfall is also a major problem. Today, estimating rainfall in a dependable and unpredictable manner is a difficult task. This work uses logistic regression and the support vector machine (SVM) classifier to provide a maximum outcome and a stronger rainfall forecast for improved prediction.[3]
Predicting when it will rain is one of the most difficult challenges in weather forecasting. Predicting rainfall accurately and on time can be highly beneficial in taking preventative precautions against flooding, ongoing construction projects, transportation, agricultural work, flying operations, and other potential threats. It is obvious that monsoon forecasting is crucial for India. Rainfall projections can be made in two different ways: - Long-term forecasts: Estimate rainfall several weeks or months ahead of time. - Short-term forecasts: Estimate the amount of rainfall in a certain area a few days ahead of time. The forecasting data needed for the prediction is provided by the Indian Meteorological Department. The purpose of this technique is to forecast rainfall over an extended period of time. [4]
In India, agriculture is the most vital component for human survival. The most crucial element for agriculture is water, or rainfall. Forecasting rainfall is a big issue these days. Farmers can protect their crops and properties against rain by being aware of the quantity of rainfall in advance when it is predicted. There are many methods for forecasting rainfall. Rainfall prediction is best served by machine learning methods. The Auto Regressive Integrated Moving Average Model (ARIMA), Artificial Neural Network (ANN), Support Vector Machine, Logistic Regression, and Self-Organizing Map are some of the main machine learning methods that are employed frequently. Additionally, two models—linear and non-linear models—are frequently employed to forecast periodic rainfall.[5]

III. PROPOSED SYSTEM

Precipitation is predicted using the predictive model. First, data must be properly formatted in order to conduct experiments, do a thorough analysis of the data, and track variations in rainfall patterns. By splitting the dataset into training and testing sets, we are able to estimate the amount of rainfall. Next, we apply various machine learning algorithms (MLR, SVR, etc.) and statistical techniques, comparing and analyzing the various methodologies utilized. We try to reduce the mistake using a variety of techniques. If any null values is found in the dataset, it is replaced under three conditions which are mean, median and mode.[6]

Dataset Description:
The dataset consists of the measurement of rainfall from year 1901-2015 for each state.

- Data consists of 19 attributes (individual months, annual, and combinations of 3 consecutive months) for 36 sub divisions
- The data is available only from 1950 to 2015 for some of the subdivisions
- The attributes are the amount of rainfall measured in mm.

Due to the size of the dataset, feature reduction is carried out in order to enhance accuracy, decrease calculation time, and save storage. Using Principal Component Analysis (PCA), one can extract the variables that are needed from a large set of variables. The goal of extracting low dimensional sets is to get as much data as possible. The significance of visualization increases with fewer variables. The covariance matrix is used, and Eigen values are extracted from it. Our dataset's properties have been reduced by the use of PCA, which has only taken into account annual data from each subdivision and rainfall data from combinations of three consecutive months.
Methods employed: Multiple Linear Regression: Fitting an equation to determined data, multiple regressions attempt to model the relationship between two or more factors and a response. It's obvious that this is nothing more than a continuation of simple regression toward the mean. The multivariable linear regression model can be expressed in the generic form $y=\alpha+\beta 1x1+ \beta 2x2+\ldots+ \beta kxk+\varepsilon$, where y is the dependent variable, x1, x2,…, and xk are the independent variables, and α and β are the coefficients. When one explicit variable isn't clear enough to map the relationship between the independent and the variable quantity, multiple regression should be used to represent extra complex relationships arising from a variety of possibilities. The researcher encodes them with LabelEnoder from the Scikit-learn package.[7]
Support Vector Regression:
Support vector machine, or SVM, is a term used in machine learning and data science. However, SVR, or support vector regression, is not the same as SVM, as the name implies; SVR is an integration algorithm, which allows us to use SVR for working with continuous values rather than SVM, which is used for classification. Support Vector Regression, which we can refer to as supporting both linear and nonlinear regression, is supported by Support Vector Machines. Support Vector Regression attempts to fit as many cases on the street as possible with limiting margin violations, as opposed to fitting the greatest street between two classes. The hyperparameter Epsilon determines the lane's size.
Kernel- The function used to map a low dimensional data into higher dimensional data:

In SVM, a hyper plane is essentially The line that will assist us anticipate the continuous value or goal value is the separation line between the data classes, which we will also define in SVR. Boundary line: The SVM plane that divides two classes in the same notion and imagines that the support vector can be on or outside the boundary lines.
Random Forest is a popular machine learning algorithm which comes under supervised learning technique. Supervised learning is an approach where a computer algorithm is trained on input data that has been labeled for a particular output.[8]
Vectors are the data points with the smallest distances between them and the boundary. SVR does out linear regression in a space with more dimensions. With SVR, every training data point may be viewed as a separate dimension. The coordinate of your test point in that dimension is provided by the evaluation of the kernel between a test point and a point in the training set. The representation of the test point in the higher dimensional space is represented by the vector that results from evaluating the test point for each and every point in the training set, k. Wx+b=0 is the hyperplane's equation, while Wx+b=+e and Wx+b=-e are the boundary line equations formula that fulfills our SVR is e<=y-Wx-b<=+e. SVR has a different regression goal compared to linear regression in linear regression, we are trying to minimize the error between the prediction and data whereas in SVR a goal is to make sure that error do not exceed the threshold.
Lasso Regression:
Lasso is Least Absolute Shrinkage and Selection Operator Lasso regression works by introducing a bias term but instead of squaring the slope, the absolute value of the slope is added as a penalty term.

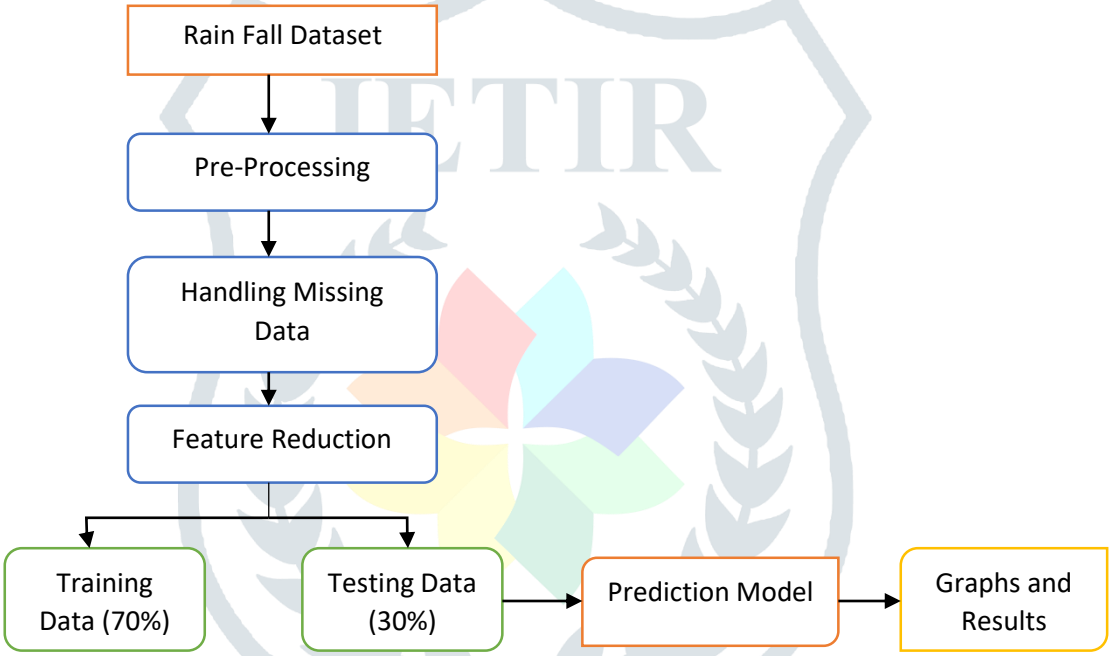

Figure-1: Proposed Model

Algorithm:
Rainfall prediction Input: Rainfall data set
Output: Accuracy/error of the prediction Step1: Import the rainfall data set csv file.
Step2: Fill the missing values with mean value of the data. Step3: Scaling the features- scaling the data to a fixed scale. Step4: Feature Reduction- PCA is used to minimize the data. Step5: The data is divided into training set (70%) and testing set (30%).
Step6: Multiple Linear Regression algorithm, Support Vector Regression and Lasso Regression is applied and the Mean Absolute Error, r2 score is calculated.
Step7: The scatter plots are plotted between predicted and testing data for the applied models and the errors are compared and best model among them is selected.
Step8: Display the results.

IV. RESULTS AND DISCUSSION

The data of rainfall from 1901-2015 is collected and data is studied and plotted to understand the rainfall in various regions. The below is the histograms plotted for the rainfall data monthly, annual and consecutive of three months. It is observed that there is a rise in volume of rainfall(Y-axis) in the months of July, August and September.

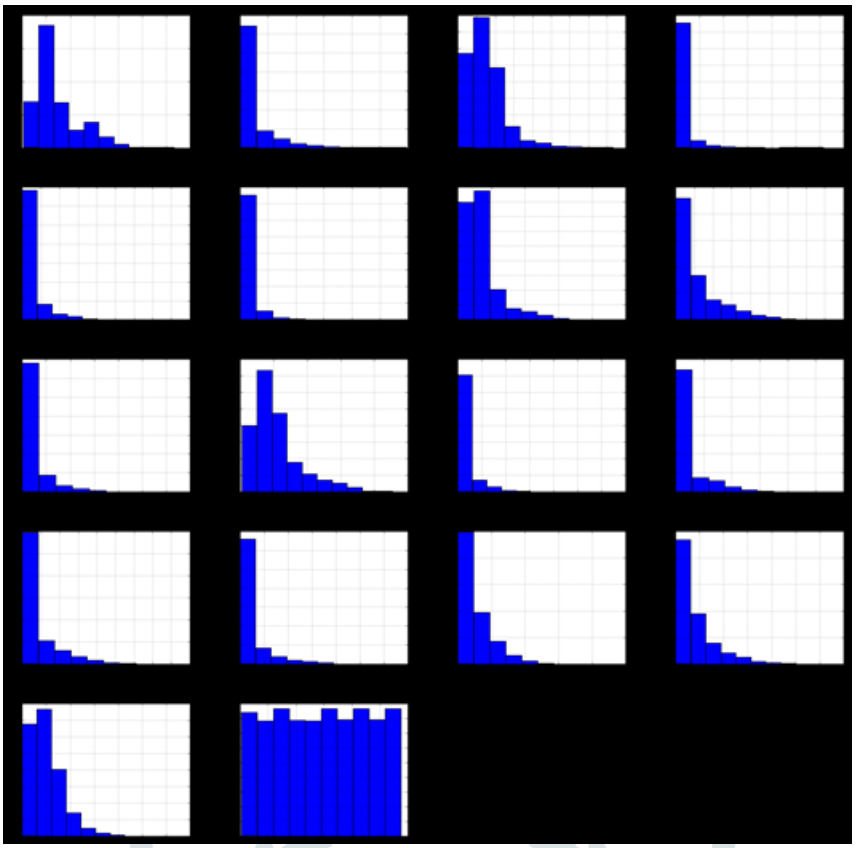

Figure-2: Histograms of the rainfall data monthly, annual and consecutive of three months

The below plot is the line graph for the amount of rainfall over the years and it is detected that there was a high volume of rainfall in 1950s

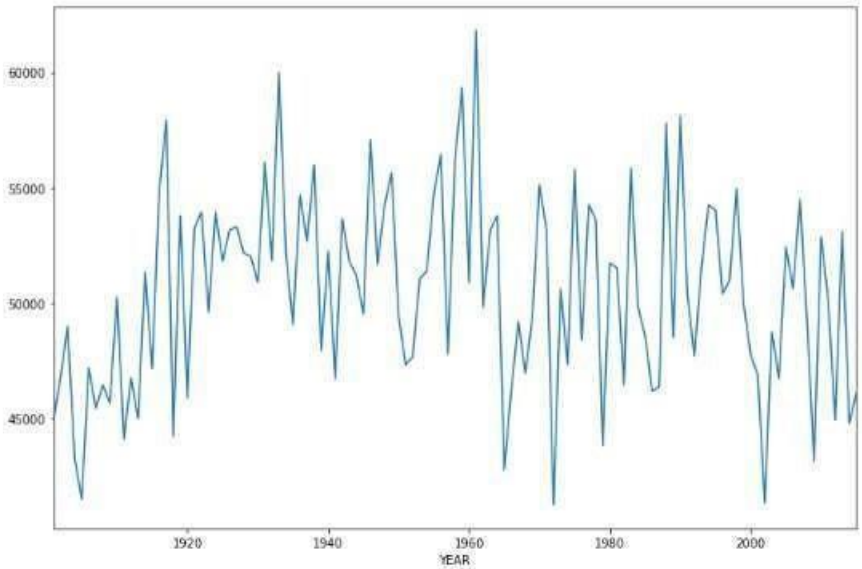

Figure-3: Line graph for distribution of rainfall from the year 1901- 2015.

The below bar graph shows the amount of rainfall for all months in the subdivisions and it is observed that the volume of rainfall is sensibly good in Eastern India in the months of March, April, May.

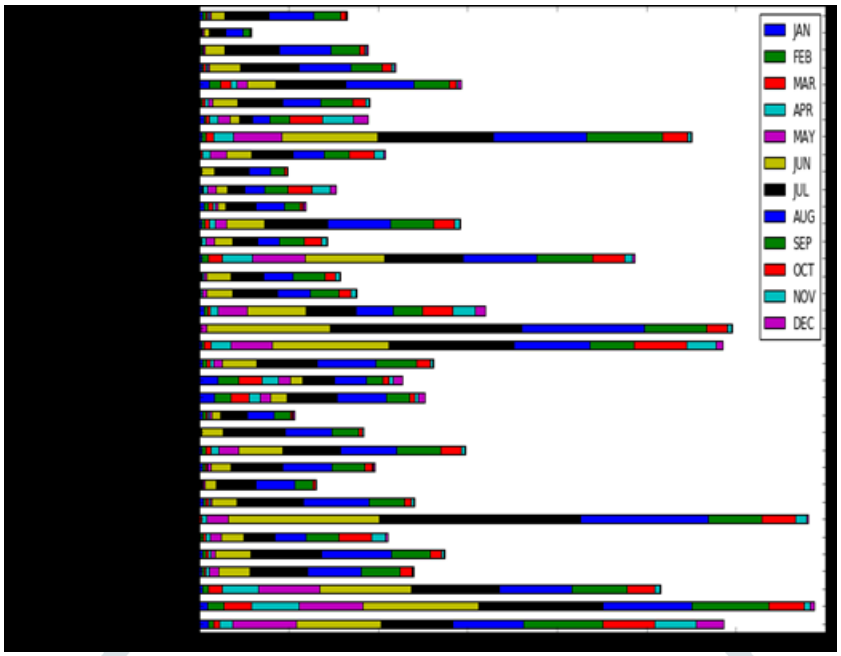

Figure-4: Bar graphs for the amount of rainfall in all subdivisions, monthly

After the analysis of data, pre-processing techniques are applied and regression models (MLR, SVR and Lasso) are applied and a scatter plot is plotted.

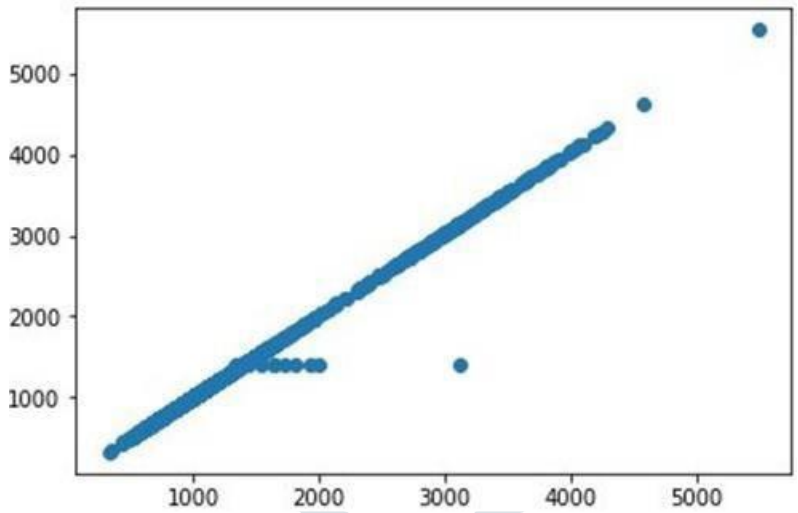

Figure-5: Scatter plot between the predictions and testing set

Then, for each regression model the MAE and r2 score are calculated and compared and a graph is plotted.

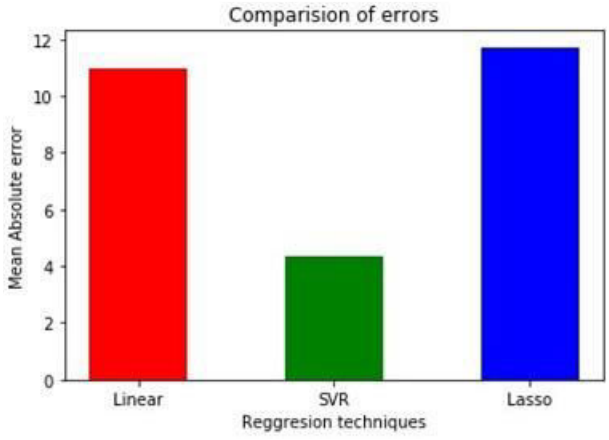

Figure-6. Comparison among applied models

V. CONCLUSION

The focus of this project was rainfall estimation, and it is predicted that SVR is a useful and flexible method that may assist the client manage the challenges associated with the geometry of the data, the typical problem of model overfitting, and the distributional features of fundamental components. For SVR display, the bit capacity choice is fundamental. We advise tenderfoots to use RBF and straight pieces for each direct and nonstraight interaction. It is evident that SVR performs better as an expectation technique than MLR. When a data collection has non-linearity, SVR becomes useful because MLR is unable to detect it. In order to evaluate the models' execution, we also process Mean Absolute Error (MAE) for the MLR and SVR models.

REFERENCES

[1] R. Mohd, M. Ahmed, and M. Zaman, "MODELING RAINFALL PREDICTION: A NAIVE BAYES APPROACH," 2018. [Online]. Available: http://iraj.in

[2] S. M. Gowtham, Y. S. Ganesh, and M. M. Ali, "Efficient Rainfall Prediction and Analysis using Machine Learning Techniques," 2021.

[3] R. Praveena, T. R. G. Babu, M. Birunda, G. Sudha, P. Sukumar, and J. Gnanasoundharam, "Prediction of Rainfall Analysis Using Logistic Regression and Support Vector Machine," in *Journal of Physics: Conference Series*, Institute of Physics, 2023. doi: 10.1088/1742-6596/2466/1/012032.

[4] S. B. Jahnavi KS Asst, "RAINFALL PREDICTION USING MACHINE LEARNING TECHNIQUES AND AN ANALYSIS OF THE OUTCOMES OF THESE TECHNIQUES," 2020. [Online]. Available: http://www.ijeast.com

[5] S. S. Kumar and M. B. R, "A Review on Rainfall Prediction using Machine Learning and Neural Network," *International Research Journal of Engineering and Technology*, p. 2763, 2008, [Online]. Available: www.irjet.net

[6] B. M. Preethi, R. Gowtham, S. Aishvarya, S. Karthick, and D. G. Sabareesh, "Rainfall Prediction using Machine Learning and Deep Learning Algorithms," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 10, no. 4, pp. 251–254, Nov. 2021, doi: 10.35940/ijrte.D6611.1110421.

[7] Y. Zhao, H. Shi, Y. Ma, M. He, H. Deng, and Z. Tong, "Rain Prediction Based on Machine Learning," 2022.

[8] D. Pangare, S. Laskar, S. Pathak, S. Jain, and P. A. Nalawade, "Survey Paper on Rainfall Prediction using Machine Learning Approach," *International Journal of Advances in Engineering and Management (IJAEM)*, vol. 4, p. 479, 2022, doi: 10.35629/5252-0405479483.