International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 479-483 www.ijaem.net ISSN: 2395-5252

Survey Paper on Rainfall Prediction using Machine Learning Approach

Diksha Pangare, Shayoni Laskar, Siddhi Pathak, Sonal Jain, Prof.S. A. Nalawade

Student,BE-IT Department of Information Technology , Dr. D. Y. Patil Institute of Technology , Pimpri , Pune, India
Guide Department of Information Technology , Dr. D. Y. Patil Institute of Technology , Pimpri , Pune, India

Submitted: 01-05-2022 Revised: 07-05-2022 Accepted: 10-05-2022

ABSTRACT- Rain prediction is an important topic that continues to gain attention throughout the world. The rain has a big impact on various aspects of human life both socially and economically, for example in agriculture, health, transportation, etc. Heavy Rainfall causesdisasters such as landslides and floods. The various impact of rain on human life prompts us to build a model to understand and predict rain to provide early warning in various fields such as agriculture, transportation, etc. This research aims to build a rain prediction model using a rule-based Machine Learning approach by utilizing historical meteorological data.The proposed study aims to predict future rainfall using linear regression technique as we also compared various models such as Support Vector Machine, Random Forest Regression, and Neural Network Model.

Keywords—Machine Learning, Neural Network, Multiple Linear Regression, Random Forest, Support Vector Machine

I. INTRODUCTION

Agriculture is the backbone of Indian economy. Irrigation facility is still not good in India and most of agriculture activities depend upon the rain. Less rainfall results in the occurrence of a dry period for a long time or heavy rain affects both the crop yield as well as the economy of country, so due to that early prediction of rainfall is very crucial. A wide range of rainfall forecast methods are employed in weather prediction at regional and national levels. Fundamentally, two approaches are used for predicting rainfall. One is empirical approach and the other is dynamicalapproach.

Today, most of the short-term monsoon predictions are based on numerical weather prediction or taking localmeteorological parameters as indicators. The forecasters use data generated by the satellites around cloud motion, cloud top temperature, water vapour content that help in rainfall estimation.

Apart from tracking satellite data, India Meteorological Department collaborates with ISRO for ground-based observations from the Automatic Weather Stations (AWS), the Global Telecommunication System (GTS) that measure temperature, sunshine, wind direction, speed and humidity.

Presently, Rainfall prediction is the most crucial factor for most water storage schemes worldwide. The uncertainty of rainfall data is one of the most complex problems. Today, most rainfall forecasting methods are incapable of detecting hidden patterns or non-linear trends in rainfall data. This research would help discover all hidden patterns and non-linear trends, which would be necessary for predicting accurate rainfall.

Due to the presence of complex issues in existing methods that cannot find the hidden patterns and non-linear trends efficiently the majority of the time, the forecast predictions were incorrect, resulting in massive losses. Thus, this research aims to find a rainfall prediction system that can solve these issues, find complexity and hidden patterns present, and provide proper and reliable predictions, therefore assisting the country in developing agriculture and the economy. Usually machine learning algorithms are classified into two major categories: (i) unsupervised learning (ii) supervised learning. All the clustering algorithms come under supervised machine learning. Even though many models have developed, but it is necessary for doing research using machine learning algorithms to get accurate predictions. The error free prediction provides better planning in the agriculture and other domains.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 479-483 www.ijaem.net ISSN: 2395-5252

Understanding and comparing various rainfall prediction techniques is important. The techniques are studied and presented in this paper. Their accuracies are analyzed and the most accurate technique is used so that the rainfall prediction is accurate and timely.

II. AIM

Rainfall prediction is important in Indian civilization and it plays major role in human life to a great extent.

It is demanding responsibility of meteorological department to predict the frequency of rainfall with uncertainty as irregular rainfall can have many impacts like destruction of crops and farms, damage of property and humans life.

It is complicated to predict the rainfall accurately with changing climatic conditions. Therefore, this study aimed to identify most accurate machine learning model among different models used.

III.METHODOLOGY:

The paper refers to four Machine Learning models for rainfall prediction and comparison amongst them.

1. Artificial Neural Networks:

Neural networks prediction is what the "artificial" component in them is, and how they are used in data science. Neural networks may be used for solving problems the human brain is very good at, such as recognizing sounds, pictures, or text. They can be used to extract features from algorithms for clustering and classification, essentially making them modules of larger Machine Learning applications. An Artificial Neural Network (ANN) is a predictive model designed to work the way a human brain does. In fact, ANNs are at the very heart of deep learning.

There are three layers to the structure of a neural-network algorithm:

Input Layer: It is the layer in which we give input to our model. The number of neurons in this layer is equal to total number of features in our data.

Hidden Layer: The output from Input layer is then feed into the hidden layer. There can be many hidden layers depending upon our model and data size. Each hidden layer can have different numbers of neurons which are generally greater than the number of features.

Output Layer: The output from the hidden layer is then fed into an activation function like sigmoid, SoftMax orReLU which then gives the final output. The data is then fed into the model and output from each layer is obtained. This step is called feedforward, we then calculate the error using an error function, some common error functions are cross entropy, square loss error etc. For example, in our case we are training the Neural Networks with different features like humidity, temperature, pressure etc. and they learn to identify and analyze the rainfall based on these features using the results of training dataset.

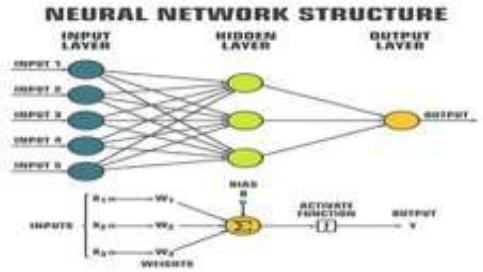

Fig. 3.1 Neural Network Structure

How Neural Network predict:

Each neurontakes into consideration a set of input values. Each of them gets linked to a "weight", which is a numerical value that can be derived using either supervised or unsupervised training such as data clustering, and a value called "bias". The network chooses from the answer produced by a neuron based on its' weight and bias.

Steps involved in Artificial Neural Network Model :

Step1: Import the necessary libraries.
Step2: Import the Dataset
Step3: Pre-Process the imported data.
Step4: Train and Test the Model.
Step5: Analysis of result.
Step6: Cross-validation with K-Fold.

2. Random Forest:

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. It can be used for both classification and regression tasks. Random forests are based on ensemble learning method in which multiple classifiers are combined to improve the overall performance of the model. In random forest model, number of decision trees are constructed and trained on various subsets of training data and predicts the output on the basis of majority votes i.e., the output of the random forest is the class selected by most trees.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 479-483 www.ijaem.net ISSN: 2395-5252

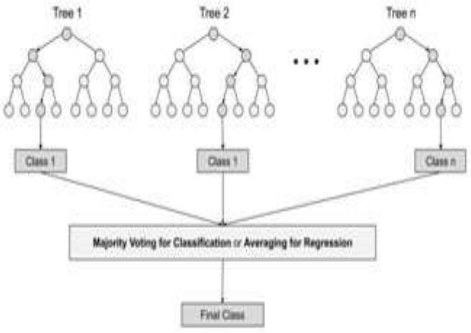

Fig. 3.2 Random Forest model

Steps in Random Forest Technique

Step 1: In Random Forest k number of random data points are chosen from the training data set having n number of data points.

Step 2: Individual decision trees are constructed and trained for selected data points.

Step 3: The test data is provided to all decision trees. Each decision tree will predict the outcome.

Step 4: The average of all outputs is taken as final result.

3. Multiple Liner Regression (MLR):

Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. In other words, multiple linear regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

Using MLR model, we can predict rainfall by using the predictor variables such as minimum temperature, maximum temperature, humidity etc.

Multiple regression models the connection between two or additional variables and a response by fitting an equation to determine information. The general form of multivariable linear regression model is: $y=\alpha+\beta_1x_1+ \beta_2x_2+\ldots+ \beta_kx_k+\varepsilon$ where y = dependent variable and x_1, $x_2\ldots$ x_k are independent variables, α,β_1, β_2, β_k are coefficients.

The process of this method is explained in the following steps:-

- Pre processing of the input rainfall data.
- Dividing the data into training and testing datasets.
- Finally, we will predict the rainfall by applying Multiple Linear Regression model.

4. Support Vector Regression (SVR):

Almost all principals of Support Vector Machine classification are followed by Support Vector Regression (SVR). As the output is real number and there are unlimited possibilities to get a real number.To handle this problem, a limitation to the tolerance (epsilon) is set to the SVM which would have effectively asked for from the problem. The model structure includes two vital stages, training and testing. Both the stages are continuous process as the main objective of the study was to build an efficient model for rainfall forecasting.

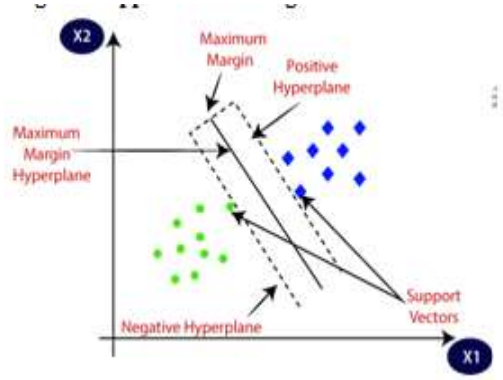

Fig. 3.3 Support Vector Regression

Training and testing stages are as follows:

1)Training stage:

The training stage starts with the retrieving of data from the repository. The data items were processed before entering this stage. After this, the role of the attributes in the dataset was assigned. Then one of the attributes (id)was selected as the id and another as the label. The selection of label was different for three different models. The next step was to set the windowing operator. The parameters of windowing operator were set and then windowing was performed. A special and efficient validation 'Sliding Window Validation' was performed for validation. Afterwards, the main processwas fed and the parameters of the Support Vector Regression. Then the model will be set to run.

2)Testing stage:

In this phase the testing data is retrieved from the repository and the role of the attributes of the data was set and then the id and label were selected as done in the training stage. Like the training stage, the parameters of windowing operator were set and run and then the main process was fed and the performance is compared.

Steps for SVRModel:

Step 1:Importing the libraries.

Step 2:Reading the dataset

Step 3:Feature Scaling

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 479-483 www.ijaem.net ISSN: 2395-5252

Step 4:Fitting SVR to the dataset Step
Step 5:Predicting a new result
Step 6:Visualizing the SVR results.

IV. SYSTEM ARCHITECTURE

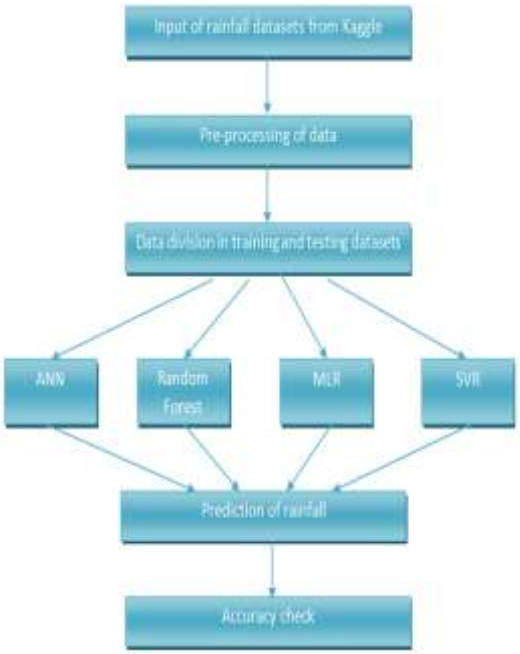

Fig 4.1: System Architecture

V.EXPERIMENTION

Libraries and Platforms suggested:

NUMPY is a numerical Python application that provides fast mathematical calculation functions for computation. It can be used to read data in arrays and for computing procedures. PANDAS library in python can read and write various files and directories. Furthermore, data processing of information frameworks makes the data source file highly performed and easy to use.

SEABORN Library is based on matplotlib, which is used for data visualization in python that provides a high-level platform for displaying, appealing and detailed statistical graphics.

MATPLOTLIB is a Python two-dimensional plotting library that produces high-quality reports in various in a graphical format. MATPLOTLIB is used in Python scripts, Jupyter notebooks, IPython shells, Web frameworks, and four graphical user interface toolkits. Matplotlib strives to make things easy, fast and also make complex things possible. Using a few code lines, you can generate graphs, histograms, bar charts, scatterplots, etc.

Anaconda Navigator is a user interface software application that allows you to quickly launch applications and access conda packages, configurations, and channels without using command-line functions. It is compatible with Linux and Windows OS.Jupyter notebook is available in anaconda navigator. It is an open-source computational notebook that allows researchers to combine source code, computational performance, descriptive language, and multimedia tools into a single document.

Collection of data:

The collection of data used in this system includes rainfall data from many regions.Along with that, average rainfall and rainfall between the transition of two months have been included. The dataset contains a total of 966 rows. The dataset was obtained from the Kaggle website, which is a data collection and publishing website. It includes attributes such as date, location, minimum temperature, maximum temperature, wind direction and speed, wind gust direction and speed, humidity, pressure, temperature, etc.

Data Pre-processing:

There are four different data pre-processing stages:

DataArrangement:Thematerialwe'vechosenisunlikelytobeinaformatthatallowsyou tointeractwithit.Thedetailscouldbeinasocialdatabase, or arestricted record configuration. So, it is formatted in a data framework.

Import dataset and libraries: The formatted data is imported into a CSV file So that Jupyternotebookcanreadthefileandcontinuetheprocess.All-importantlibrariesrequired like Numpy, pandas, matplotlib, seaborn are imported for reading, visualization, and manipulating thedata.

Removing null values: The information may sometimes be missing. In that case, we can performtwomethodsinremovingthenullvalueseither deletingtherowwhichcontainsthe null value or calculating the mean value of the particular column or a row, and themissing value is replaced with the mean value. Therefore, it gives better results than the previous method.

Splitting the data: Choose the independent variables or feature columns of the database, represent them as x, and define the target or dependent variable rain tomorrow as y. The database is separated into two separate sets - train data and test data using function train_test_split(). Typically, the dataset is divided into 7:3 or 8:2 ratios. That means wecan use 70-80% of the data for training the algorithm while leaving out the remaining 20-30% for test data. The splitting of data depends on the form and size of thedataset.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 479-483 www.ijaem.net ISSN: 2395-5252

VI. LIMITATIONS

Errors in weather model forecasts arise because we don't know what every molecule of air in the atmosphere is doing, and even if we did, we have an imperfect understanding of how these molecules interact with each other at various scales, molecules might be up to sometime from now. While our models are really good considering how immensely complex the task of weather prediction is, they'll never be perfect.

VII.CONCLUSION

The proposed work is based on rainfall prediction as in future it is useful for farmers for crop fertility.

Four machine learning models are studied. The ML models specified are Multiple Linear Regression,Artificial Neural Networks, Support Vector Machine and Random Forest. According our survey, we came to the conclusion that the Multiple Linear Regression model will give better accuracy as compared to the other three models.

REFERENCESs

[1] LihuaXiong, K. M. Connor, "An empirical method to improve the prediction limits of the GLUE methodology in rainfall– runoff modeling," Journal of Hydrology, Vol. 349, pp: 115–124, 2008.

[2] Gurpreet Singh and Deepak Kumar, "Hybrid Prediction Models for Rainfall Forecasting", IEEE, (2019), pp. 392-396.

[3] MasafumiGoto, Faizah Cheros, Nuzul Azam Haron, Nur-AdibMaspo, AizulNahar Bin, MohdNawiHarun, and MohdNasrun,(2020)," Evaluation of Machine Learning approach in flood prediction scenarios and its input parameters: A systematic review." IOP Science journal.

[4] J.Refonaa, M. Lakshmi, Raza Abbas, Mohammad Raziullha, "Rainfall Prediction using Regression Model", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019.

[5] Gowtham Sethupathi.Ma, Yenugudhati Sai Ganesh b, Mohammad Mansoor Alic , "Efficient Rainfall Prediction and Analysis using Machine Learning Techniques", Vol.12 No.6 (2021), 3467-3474