An Analysis of Modern Machine Learning Algorithms for Rainfall Prediction

Love Vishwakarma[1], Shiv Shankar Gupta[2]

[1,2]Assistant Professor, Department of Computer Science, LCIT College of Commerce and Science, Bilaspur (C.G.)

Abstract

Agriculture has a significant importance in the Indian economy. Rainfall is crucial for agriculture, however these days, predicting rainfall has become a very difficult subject. A good rainfall forecast enables people to plan ahead, take safety measures, and have better crop-related strategies. Both nature and humanity are being severely impacted by global warming, which also hastens the shift in climatic conditions. Due to the warming of the air and rising ocean level, floods are occurring and farmed fields are becoming drier. Unseasonable and excessive amounts of rainfall are a result of unfavourable climate change. One of the finest methods for learning about the rain and climate is to anticipate rain. The primary goal of this study is to accurately describe the climate to the clients from a variety of aspects, including agriculture, research, power generation, etc. to understand the requirement for changing the environment and the variables such as temperature, humidity, precipitation, and wind speed that ultimately lead to rainfall forecast. Predicting rain is difficult since it relies on geographic places as well. Machine learning, an expanding branch of artificial intelligence, aids in rainfall forecasting. For the purpose of forecasting the rainfall in this research study, we will use a dataset from the UCI repository that has many properties. The major goal of this work is to analyse a system for predicting rainfall and to do it more accurately by using machine learning classification algorithms.

Keywords: Rainfall Prediction system, Machine Learning, Dataset, Classification algorithms.

1. INTRODUCTION

Worldwide, rainfall forecasting is absolutely essential since it affects how people live their lives. The meteorological agency is burdened with the difficult task of analysing rainfall frequency. With changing atmospheric conditions, it is challenging to predict rainfall with accuracy. It is hypothesised that rainfall may be predicted for both the rainy and summer seasons. This is the main reason why it is necessary to examine the algorithms that may be used to predict rainfall. Machine learning is one of these adept and powerful technologies; it is a method of manipulating and extracting implicit, previously unknown and known, and potentially helpful information about data. Machine learning is a vast and in-depth topic, and its application and scope are expanding daily. The term "machine learning" refers to a variety of classifiers from supervised, unsupervised, and ensemble learning that are used to predict and assess the correctness of a given dataset. Since it will benefit many individuals, we may apply this information to our idea of a rainfall prediction system. To identify the best accurate model, different Machine Learning algorithms are examined, including Naive Bayes, Decision Trees, K-Nearest Neighbour, and Random Forest. Here, the UCI repository's rainfall dataset is utilised. The available classification approaches are discussed and compared in this study. The report also discusses the potential for development and the range of next study.

The goal of this research article is to forecast the amount of rain that will fall at a certain place using user-supplied input data. Date, location, maximum and lowest temperatures, humidity, wind speed and direction, evaporation, and other factors are among the characteristics. Four algorithms—KNN, Decision Tree, Random Forest, and Naïve Bayes—are used to train these variables related to rainfall. The most effective of these algorithms, Random Forest and KNN, provide accuracy of about 89%. And finally, we'll forecast whether or not it will rain there.

2. LITERATURE REVIEW

This paper's main goal is to examine the many strategies suggested by the authors in order to create a real-time rainfall prediction system that fixes the drawbacks of earlier techniques and provides the most reliable answer. The method [1] forecasts the rainfall in India's Karnataka state for the Udupi district. The technology utilised is BPNN with cascading feed forward neural networks. When compared to BPNN, the network exhibits more accuracy. It's possible that this system won't provide an accurate long-term rainfall prediction.

The device [2] For the Chennai region, G. Geetha and R. Selvaraj employed an ANN model to estimate monthly rainfall while taking into account a variety of meteorological factors, including maximum and lowest temperatures, relative humidity, wind speed, and wind direction. They evaluated the data and forecasted weekly rainfall for a few Chennai neighbourhoods. More accurate predictions are made using ANN than with multiple linear regression models. The forward pass and the reverse pass are the two passes that this algorithm uses. The forward layer receives input, which is then sent to the following layer over the network. After assessing the results of the preceding layer, the outcome is finally created at the backward layer.

Rainfall prediction system employing deep mining KNN approach was introduced in paper by [3]. The total number of nearest neighbours, which aids in determining the class label for unclassified data, is found using a single K value. We can identify the class or category of a certain dataset using KNN because similar parameters are grouped into the same sort of cluster. It doesn't take long to train this algorithm for classification or regression. If the wrong value of K is used, the accuracy of this system could suffer.

3. PROPOSED METHODOLOGY

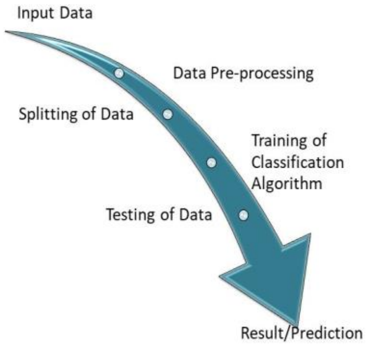

Figure 1: Proposed Model for Prediction

A. Exploration and Analysis of Data

Data analysis is done to ensure that future results will be near, allowing for reliable and accurate prediction. This assurance that the data was collected without error can only be obtained after the raw data has been validated and examined for anomalies. Additionally, it aids in identifying the data that have irrelevant characteristics for a prediction model.

B. Data Pre-processing

Data pre-processing is a data mining approach that transforms unstructured, erroneous input into a format that the model can utilise and comprehend. Raw data has many inaccuracies, missing features, and is inconsistent and incomplete. We discovered via data exploration and analysis that the raw data for our model has a significant number of null values that need to be updated with their mean values. We may also deal with the missing values by removing any unnecessary columns or rows. Due to the fact that the model is based on mathematical calculations and equations, it is important to encode categorical data into numeric form. The pre-processing step of feature selection involves choosing just those features that are relevant to our model for predicting rainfall. This shortens training time and improves model accuracy. The final stage of pre-processing is feature scaling, in which independent variables are brought within a predefined range such that no one variable dominates the others.

C. Modelling

The suggested model first cleans, then pre-processes, and finally arranges redeemed meteorological data. Finally, according to rules set out by the Indian Meteorological Department, rainfall data is classified into a number of categories. In this research, we provide a method for rainfall forecasting based on machine learning classification techniques. 30% of the pre-processed data are used for testing, and 70% are for training. The portioned data is applied to four distinct machine learning algorithms, and each result is examined before the final, precise result is shown. The next section provides an explanation of how each classifier operates.

Naïve Bayes Classifier It is a classifier based on probabilities that uses the Maximum Posteriori determination rule in a Bayesian framework to classify data. Due to the branching structure of the goal or dependent variable, there are only two possible outcomes: class 0 for failure and class 1 for success.

K-Nearest Neighbour One of the earliest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. The K-NN method takes into account how similar the new case and data are to existing instances and places the new case in the category that is most closely connected to the existing categories. It categorises items based on their nearest neighbour. It groups the designated points and applies them to the marking of additional marks. Similar data are clustered, and K-NN may be used to fill in the data's null values. We use machine learning (ML) algorithms on the data set as soon as the value gaps are filled in. By combining these algorithms in different ways, it is feasible to achieve greater precision.

Random Forest: Using decision trees built from data samples, Random Forest is a supervised learning approach that may be used for both classification and regression.

a. A given dataset contains a number of random samples.
b. For each data sample, a decision tree is built, and from each decision tree, a prediction is made.
c. Voting will then be conducted on each expected outcome.
d. Finally, choose the prediction result that received the most votes as the final forecast result.

Decision tree: It is a type of classification technique that may be used with both category and numerical data. It produces structures like trees, is simple to use, and analyses data in graphs with trees as nodes. Based on the most significant signs, this algorithm assists in separating the data into two or more similar groupings. We first determine the entropy of each characteristic, then partition the data into predictors with the highest information gain or lowest entropy: The outcomes are simpler to read and understand. As it evaluates the set of data in the tree-like graph, this method is more accurate than other algorithms.

D. Assessment

1. **Accuracy:** It measures the proportion of accurate outputs to all input samples.
2. **Precision:** It is calculated by dividing the total number of accurate positive outcomes by the total number of accurate positive outcomes that were predicted by the classifiers.

4. RESULT AND ANALYSIS

This study paper’s main goal is to develop a model, evaluate the effectiveness of several Machine Learning algorithms, and identify the most precise approach for predicting rainfall. On the dataset, this study used K-Nearest Neighbour, Random Forest, Decision Tree, and Logistic Regression algorithms. We provided the actual real-time figures of the highest and minimum temperatures, relative humidity, wind speed, etc. for the experimental purpose. After the models were trained, the dataset was divided into training and testing data, and the accuracy score was recorded and examined prior to making a final prediction. Below is an assessment of the algorithms’ results, along with a table that displays their accuracy ratings.

Table 1: Accuracy on test dataset

Method	Classification Accuracy	Precision
Random Forest	89.66	0.763
K-Nearest	87.63	0.799
Decision trees	86.12	0.808
Naïve Bayes	81.32	0.22

CONCLUSION

The general goal is to describe several machine learning (ML) algorithms that may be used to forecast rainfall. The purpose of this study is to develop an accurate and effective model using fewer features and tests. The data is first pre-processed before being utilised in the model. The most effective classification algorithms are Random Forest classifier with about 88% efficiency and K-Nearest Neighbour with 87% efficiency. Nevertheless, the Decision Tree classifier has the lowest accuracy (73%). This study may be expanded to include more ML methods, such as time series, clustering and association rules, and other ensemble methods. Given the limits of this study, more complicated and combined models must be developed in order to increase the accuracy of rainfall forecast systems. This type of model may be developed for huge datasets with stronger articulate monitoring for a specific region, which will enhance computation pace while improving precision and accuracy.

REFERENCE

1. Kumar Abhishek. Abhay Kumar, Rajeev Ranjan, Sarthak Kumar," A Rainfall Prediction Model using

Artificial Neural Network", 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC2012), pp. 82-87, 2012.
2. G. Geetha and R. S. Selvaraj, "Prediction of monthly rainfall in Chennai using Back Propagation Neural Network model," Int. J. of Eng. Sci. and Technology, vol. 3, no. 1, pp. 211 213, 2011.
3. 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.
4. Elia Georgiana Petre, "A decision tree for weather prediction", Seria Matematica - Informatica] – Fizic, no. 1, pp. 77-82, 2009.
5. Gupta D, Ghose U. A Comparative Study of Classification Algorithms for Forecasting Rainfall. IEEE. 2015.
6. Rajeevan, M., Pai, D. S., Anil Kumar, R. & Lal, B. New statistical models for long-range forecasting of southwest monsoon rainfall over India. Clim. Dyn. 28, 813–828 (2007).
7. Mishra, V., Smoliak, B. V., Lettenmaier, D. P. & Wallace, J. M. A prominent pattern of year-to-year variability in Indian Summer Monsoon Rainfall. Proc. Natl Acad. Sci. USA 109, 7213–7217 (2012).
8. 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).
9. Kumar, V., Yadav, K.V., & Dubey, S. (2022). Rainfall Prediction using Machine Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET) vol. 10, 2321-9653, pp. 2494-2497.