Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Precision Rainfall Predictions: A Daily Weather Data Approach using Machine Learning

Sathesh. A

Research Scholar, Tianjin Key Laboratory of Process Measurement and Control, School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

E-mail: sathesh4you@gmail.com

Abstract

Rainfall prediction is an important task since a lot of individuals rely on it, especially in agriculture. The study attempts to predict rainfall using machine learning algorithms, taking into account the impact of shortages or excessive rainfall on rural and urban life. Several techniques and approaches for predicting rain have been developed; however, there is still a lack of precise outcomes. The comparative study focused on incorporating Machine Learning (ML) models, analyzing various situations and time horizons, and predicting rainfall by using three different approaches. This research uses data preprocessing, feature selection, and machine learning methods like Random Forest, K-nearest neighbor (KNN), and Logistic Regression. This study shows the usefulness of machine-learning approaches in forecasting rainfall. In comparison, Random Forest performs better when compared to other models with a high precision rate.

Keywords: Rainfall prediction, Weather Forecast, Machine learning, KNN, Logistic Regression, Random Forest, Performance Metrics

1. Introduction

Data from time series mining is a popular study subject in the field of data mining. The time series technique collects data across defined time periods, such as every day, every week,

DOI: https://doi.org/10.36548/jscp.2023.4.003
Received: 03.12.2023, received in revised form: 27.12.2023, accepted: 11.01.2023, published: 22.01.2024

every month, every quarter, or annual [1]. This data may be used to anticipate several fields, including banking, the stock market, and climate change. Predicting weather using time series statistics is a tough but important endeavor. Rainfall prediction remains one of the most difficult challenges in the weather forecasting process. Because of tremendous climatic changes, predicting rainfall is more challenging than ever before. The climate and weather are often characterized by a few meteorological factors; the most important is rainfall intensity or amount. The rainfall-runoff-soil interaction is one of the most difficult hydrological phenomena because it involves a complex, non-linear link between precipitation and runoff. This activity is very difficult to assimilate due to the large number of components involved in the testing of physical process [2].

Rainfall forecasting is crucial because heavy as well as irregular rainfall may result in a wide range of implications, including crop loss and property damage, necessitating an improved forecasting model for warning in advance that can reduce risk to life and assets while also improving agricultural farm management. This rainfall forecast primarily helps farmers, as the water supply could be used effectively. Water has a significant impact on a region's economic, social, and environmental growth. There are several hardware systems available for forecasting rainfall using meteorological characteristics such as humidity, temperature, and pressure. Since traditional methods were inefficient, we can obtain accurate results by using machine learning techniques.

The ever-changing science of predicting rainfall has made tremendous advances in the past few years. Machine learning (ML) plays a significant role in increasing prediction efficiency. ML is quite successful in improving the accuracy of rainfall prediction algorithms. By finding patterns in previous data, machine learning models can forecast rainfall occurrences, even in very complicated and unpredictable systems. Machine learning makes predictions using statistical models. Statistical models discover patterns and links in previous weather data and use that information to forecast future weather conditions. These models may also take into account a variety of parameters, including humidity, wind speed, temperature, amount of cloud cover, and data via satellites, radars, and meteorological observatories.

Advantages of Machine Learning in Rainfall Prediction:

- Machine learning techniques can detect complicated patterns and correlations in vast datasets, resulting in more accurate predictions.
- Traditional approaches are less effective at incorporating different information sources than ML models.
- ML excels at coping with faulty data. It can learn to recognize patterns in data despite noise as well as missing information.
- ML models can analyze previous data as well as quickly adapt to new situations, allowing forecasters to make forecasts with a longer lead time.
- Machine learning models can help design efficient early warning systems during extreme rainfall.

This study proposes a rainfall prediction system that employs a machine learning approach. Implementing different types of machine learning for weather forecasting has various benefits, including increased accuracy, efficiency, and overall prediction efficacy.

2. Literature Review

[3] investigates the ability of incorporating machine learning for enabling weather forecast assistance. Researchers have created regression models, LightGBM and XGBoost and further evaluated the effectiveness in forecasting rainfall. These findings have increased the potential of incorporating ML in weather forecasting, with the potential to help analysts make better informed decision.

The purpose of this research work carried out in [4] is to create a ML-based rainfall forecasting model. This model utilizes a dataset size of 2,391 records. Five different models (Decision tree, Logistic Regression, KNN, Multinominal NB, and Random Forest) were employed in the investigation. Each model was trained using eight input characteristics and then tested for its rainfall forecasts. Among all the models, random forest predictions on the dataset were most effective, making it the best for this investigation [4].

For most meteorological stations, the suggested prediction algorithm accurately identified the climatic zones considering their diverse geographic and climatic characteristics. This led to accurate precipitation projections, addressing a shortcoming observed by prior studies. In this study, a comprehensive approach is presented for obtaining reliable forecasts of meteorological information, such as precipitation.[5]

Experimental research was conducted to investigate the use of machine learning to forecast the rainy season. The most prominent machine learning algorithms like decision trees, multinomial NB, random forest, SGD and logistic regression were used in conjunction with a collection of datasets comprising of rainfall measurements gathered over preceding ten years from key cities in Northern Sumatra After analysis, the data indicated that logistic regression yielded the best fit for the predictions [6].

This research investigates the application of support vector machine (SVM) or artificial neural network (ANN) methods for binary classification for summer monsoon rainfall based on common meteorological indicators including relative humidity, temperature, and pressure. SVM and ANN methods obtained 82.1 and 82.8%, accuracy in classification respectively, on an unbalanced dataset [7]. Prior to classification, the dataset underwent pre-processing operations such as data cleaning and normalization. The findings show that the suggested machine learning fusion-based approach outperforms the other models [8].

The study [9] proposes the collection and analysis of rainfall data from the previous ten years to forecast future rainfall with the aim of improving outcomes. The study demonstrates that Back Propagation Neural Networks (BPNNs) may outperform other methods, yielding optimal inferences. This showcases the effectiveness of Artificial Neural Networks (ANNs) in forecasting rainfall and predicting the likelihood of landslides in the near future [10].

3. Proposed Methodology

The proposed block diagram is shown in Figure 1, illustrating the consideration of the rainfall prediction data. The dataset has undergone various pre-processing techniques and feature engineering processes. The preprocessed datasets are split to training and testing. The dataset is further subjected to different ML algorithms and the models are analyzed using various performance metrics.

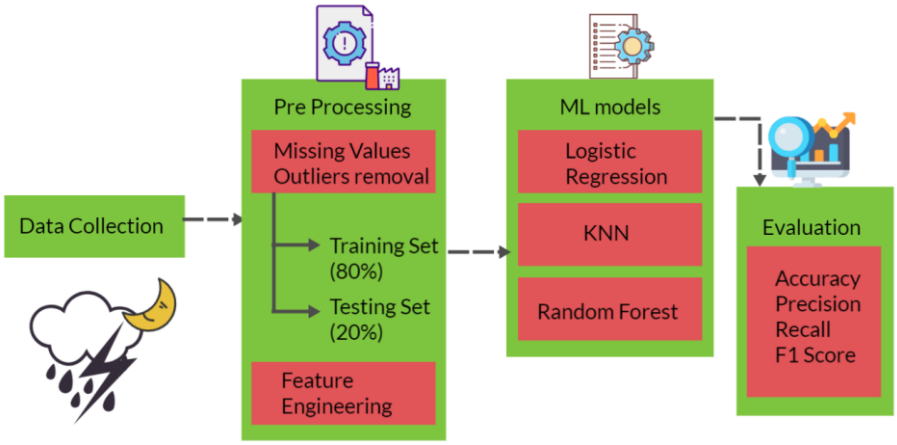

Figure 1. Block Diagram of the Proposed Method

Data Collection: This dataset comprises around ten years of daily weather records from several Australian weather stations with 'Tomorrow' being the variable to forecast. The dataset was sourced from https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package/data.

Date The date of observation	Location The common name of the location of the weather station	MinTemp The minimum temperature in degrees celsius	MaxTemp The maximum temperature in degrees celsius	Rainfall The amount of rainfall recorded for the day in mm	Evapor The so-ca evaporatic 24 hours t
2007-11-01 2017-06-25	Canberra 2% Sydney 2% Other (138680) 95%	NA 1% 11 1% Other (143076) 98%	506 unique values	0 63% 0.2 6% Other (45619) 31%	NA 4 Other (79
2008-12-01	Albury	13.4	22.9	0.6	NA
2008-12-02	Albury	7.4	25.1	0	NA
2008-12-03	Albury	12.9	25.7	0	NA
2008-12-04	Albury	9.2	28	0	NA
2008-12-05	Albury	17.5	32.3	1	NA
2008-12-06	Albury	14.6	29.7	0.2	NA
2008-12-07	Albury	14.3	25	0	NA

Figure 2. Dataset used (Rain in Australia from Kaggle)

Data Preprocessing: Clean up the dataset by removing outliers, missing values, and any discrepancies. Convert category variables to numerical representations as needed. Split the data set into sets for training and testing.

Feature Engineering: Extract important elements through the dataset that may affect rainfall estimates. The model's performance can be increased by adding new features or transforming the current ones. To find patterns within the data, the study developed a proprietary machine-learning model on 80% of the records and tested its effectiveness on the remaining 20%.

Predicting Models: In this study, three machine learning algorithms were considered. The dataset was trained to predict tomorrow's rainfall based on the following attributes: 'Rainfall', 'MinTemp', 'Humidity9am', 'Evaporation', 'WindGustSpeed', 'WindSpeed3pm', 'MaxTemp', 'Sunshine', 'WindSpeed9am', 'Humidity3pm', 'Pressure3pm', 'Cloud9am', Temp9am', 'Cloud3pm', "Pressure9am and 'Temp3pm'. These are the columns that were considered for prediction.

3.1 Logistic Regression

Logistic regression is a supervised learning classification technique that predicts observations into discrete classes. Its practical use is to categorize observations. Thus, the result is discrete in nature. The logistic regression technique uses linear equations using independent or explicable variables to forecast a response value. The study may utilize the knowledge of the sigmoid function and decision boundary to create a prediction function. In logistic regression, a prediction function yields the likelihood that the observation is positive, either Yes or True. The success of the logistic regression model is dependent on the sample size. Typically, great accuracy demands a large sample size.

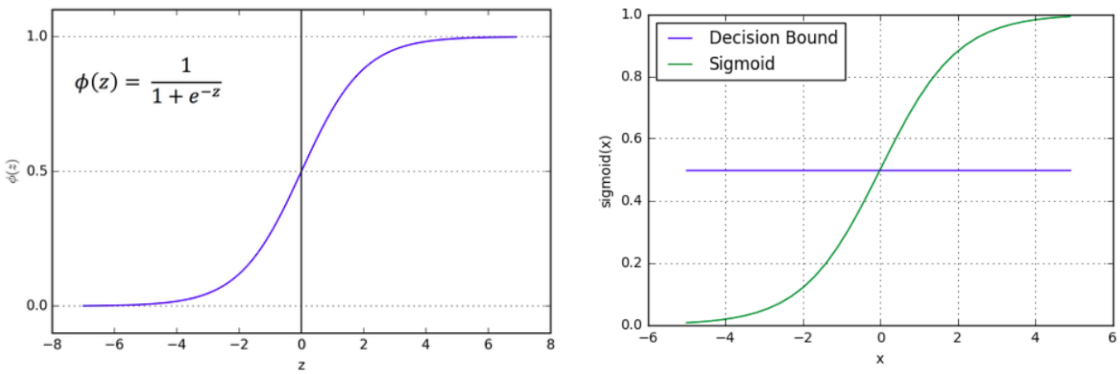

Figure 3. Sigmoid and Decision Bound [11]

3.2 KNN: K- Nearest Neighbor

KNN is a nonparametric approach. where the outcome is a class, and an object is assigned a classification based on the majority vote of its neighbors, The classification is determined by the item that is most prevalent among its neighbors.. A helpful strategy for classification and regression is to apply weights to the contributions of neighbors so that the closer neighbors contribute greater amounts to the average than the most distant ones.

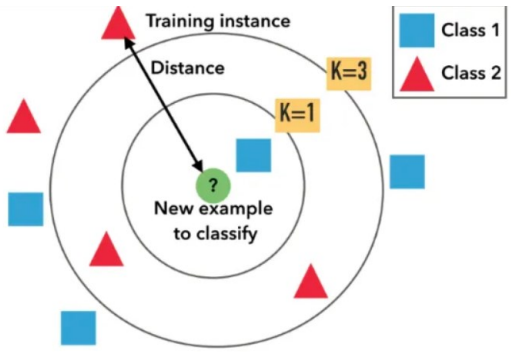

Figure 4. K- Nearest Neighbor Strategy [12]

3.3 Random Forest

Random Forest represents an ensemble method of learning that predicts rainfall by merging the outcomes from multiple decision trees. Each tree of decisions is trained individually using a chosen portion of the initial training data as well as features. This is especially effective for predicting rainfall since weather patterns are typically complex and non-linear.

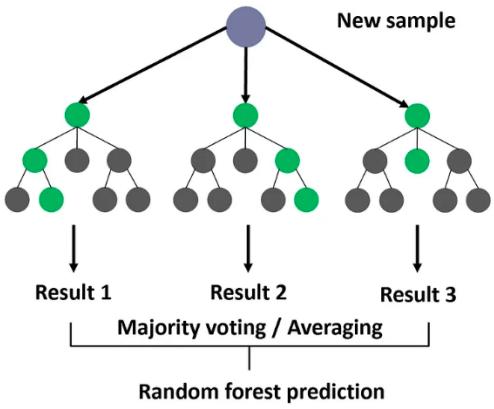

Figure 5. Random Forest

4. Results

Experiments are done on pre-processed and cleansed weather data from Australia. The tests attempt to examine different machine-learning algorithms for predicting rainfall. Various analyses of the considered features are shown below.

Wind Gust Speeds Histogram

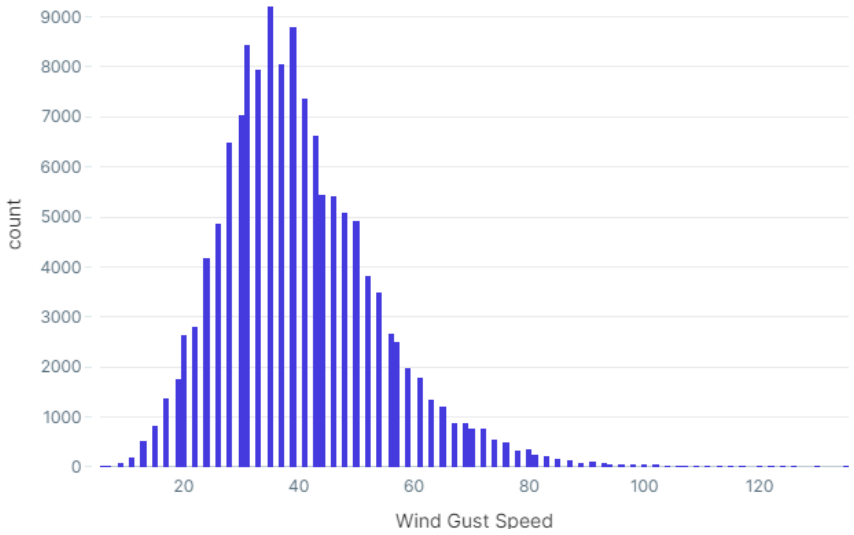

Figure 6. Wind Gust Speeds Histogram

Distribution of Rainfall for Tomorrow's Occurrence

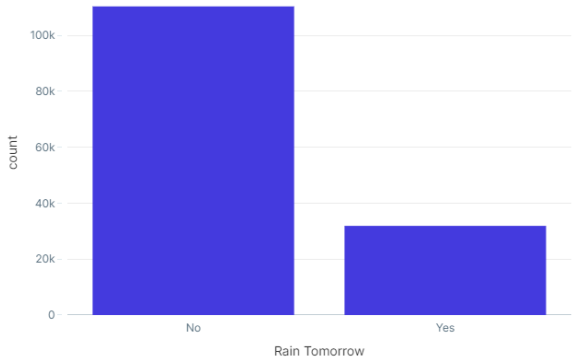

Figure 7. Count of Rain Tomorrow

Confusion Matrix

A confusion matrix seems to be a table utilized for classification to assess the effectiveness of a machine-learning model. The matrix summarizes a classification algorithm's performance by showing how many true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions the model made. The confusion matrix of all three ML algorithms is shown below:

Logistic Regression

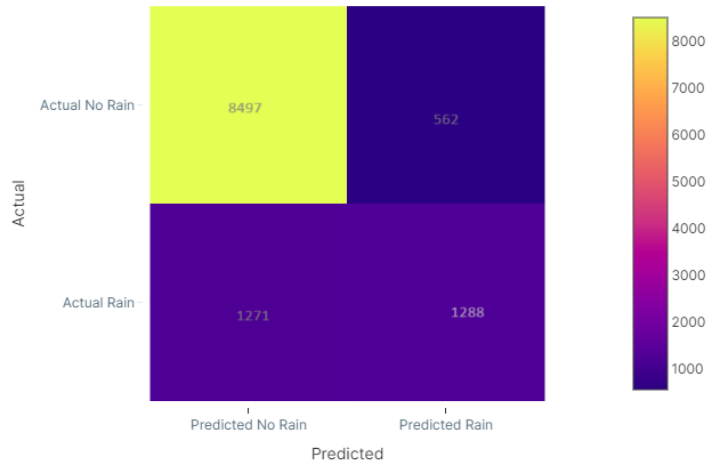

Figure 8. Confusion Matrix of Logistic Regression

K- Nearest Neighbor

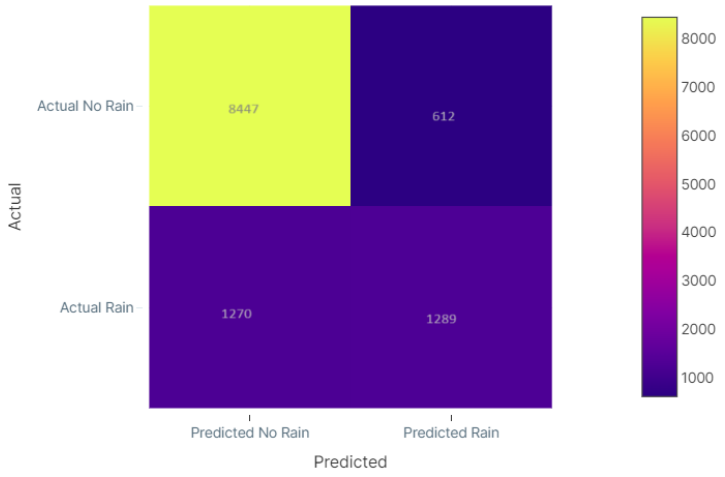

Figure 9. Confusion Matrix of the KNN Model

Random Forest

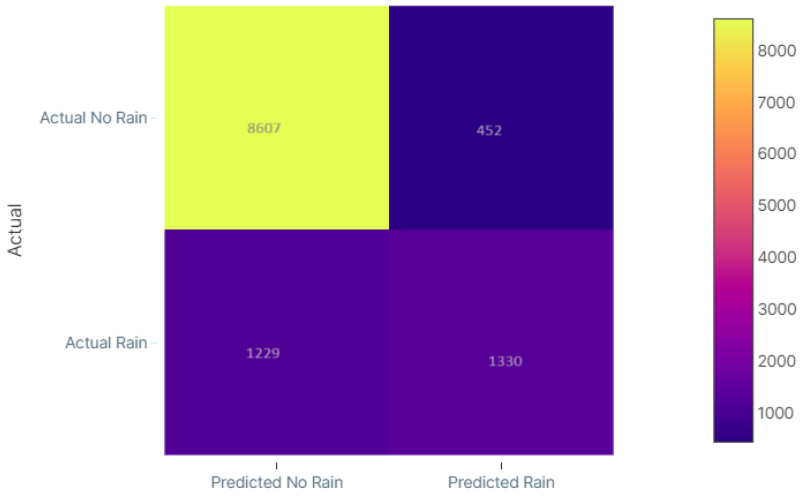

Figure 10. Confusion Matrix of Random Forest Model

Measures	Definitions	Formula
Accuracy	The algorithm's accuracy in predicting variables is calculated by its accuracy.	A = (TP+TN) / (Total no. of samples)
Precision	Precision assess the correctness.	P=TP/(TP+FP)
Recall	The recall is used to evaluate a classifier's completeness or sensitivity.	R=TP/(TP+FN)
F1-score	The average of Precision and Recall is known as F1-score.	F=2*(P*R)/(P+R)

Figure 11. Performance Measures

Table 1. Performance Metrics of the Considered ML Models

ML models	Accuracy	Precision	Recall	F1-Score
LR	0.842	0.696	0.503	0.584
KNN	0.838	0.678	0.503	0.578
RF	0.855	0.741	0.519	0.612

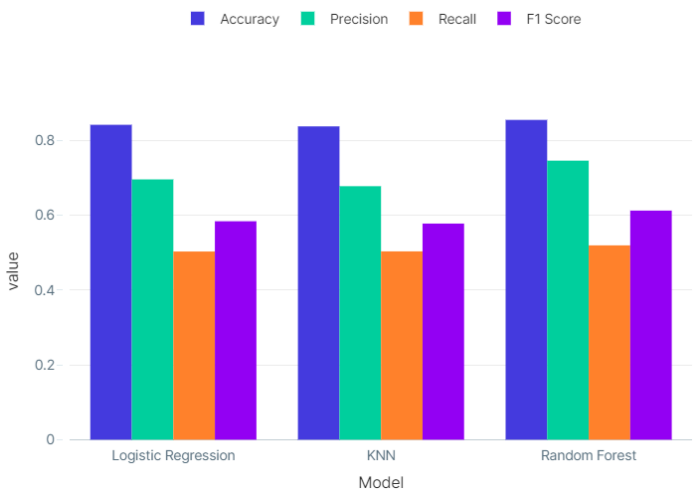

Figure 12. Performance Metrics Comparison

In the framework of rainfall prediction, all of the decision trees in the Random Forest might be trained to forecast the quantity of rain based on past meteorological data. Random forest plays a better role in terms of accuracy as well as precision when compared to other ML models.

5. Conclusion

In this research, machine learning algorithms are explored with preprocessing methods to learn about the overall performance of the classifier The study focuses on the unpredictability of Australian weather, highlighting the lack of a clear association between rainfall and specific locations and times. Machine learning models are applied to the Australian rainfall dataset to predict the daily rainfall. Here, K - Nearest Neighbor, Logistic Regression, and Random Forest algorithms are analyzed based on their performance metrics. Random Forest's ability to handle complex interactions, and large datasets, and avoid overfitting makes it an effective choice for rainfall prediction. Random Forest outperforms with an accuracy of 85% and the highest precision rate compared to the other two ML models which is a rate of 0.74. Furthermore, prediction accuracy can be increased by implementing various preprocessing techniques and training those datasets with multiple and fused ML algorithms.

References

[1] Shabib Aftab, Munir Ahmad, Noureen Hameed, Muhammad Salman Bashir, Iftikhar Ali and Zahid Nawaz, “Rainfall Prediction in Lahore City using Data Mining Techniques” International Journal of Advanced Computer Science and Applications(ijacsa), 9(4), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090439

[2] Ridwan, Wanie M., Michelle Sapitang, Awatif Aziz, Khairul Faizal Kushiar, Ali Najah Ahmed, and Ahmed El-Shafie. "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia." Ain Shams Engineering Journal 12, no. 2 (2021): 1651-1663.

[3] Triandini, Evi et al. “Regression Based Machine Learning Model for Rainfall Forecasting on Daily Weather Data.” 2023 Eighth International Conference on Informatics and Computing (ICIC) (2023): 1-6.

[4] Ria, Nushrat Jahan, Jannatul Ferdous Ani, Mirajul Islam and Abu Kaisar Mohammad Masum. “Standardization Of Rainfall Prediction In Bangladesh Using Machine Learning Approach.” 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (2021): 1-5.

[5] Skarlatos, Kyriakos, Eleni S. Bekri, Dimitrios Georgakellos, Polychronis Economou, and Sotirios Bersimis. "Projecting Annual Rainfall Timeseries Using Machine Learning Techniques." *Energies* 16, no. 3 (2023): 1459.

[6] Rumapea, Humuntal, Marzuki Sinambela, Indra Kelana Jaya and Indra M Sarkis. “Prediction of Rainfall in North Sumatera Using Machine Learning.” 2023 International Conference of Computer Science and Information Technology (ICOSNIKOM) (2023): 1-4.

[7] Hudnurkar, Shilpa, and Neela Rayavarapu. "Binary classification of rainfall time-series using machine learning algorithms." International Journal of Electrical and Computer Engineering 12, no. 2 (2022): 1945-1954.

[8] Rahman, Atta-ur, Sagheer Abbas, Mohammed Gollapalli, Rashad Ahmed, Shabib Aftab, Munir Ahmad, Muhammad Adnan Khan, and Amir Mosavi. "Rainfall prediction system using machine learning fusion for smart cities." Sensors 22, no. 9 (2022): 3504.

[9] Prashanthi, Vempaty, Srinivas Kanakala, Deepika Borgaonkar and D. Suresh Babu. “Rainfall Prediction Using Catboost Machine Learning Algorithm.” 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) (2023): 1-5.

[10] Srivastava, Shikha, Nishchay Anand, Sumit Sharma, Sunil Dhar and Lokesh K. Sinha. “Monthly Rainfall Prediction Using Various Machine Learning Algorithms for Early Warning of Landslide Occurrence.” 2020 International Conference for Emerging Technology (INCET) (2020): 1-7.

[11] https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial

[12] https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336

[13] https://medium.com/@roiyeho/random-forests-98892261dc49