WEATHER PREDICTION MODEL USING MACHINE LEARNING

Pradeepa S

Department of Artificial Intelligence

Kongu Engineering College

Erode, India

pradeepas.22aim@kongu.edu

Roshini G

Department of Artificial Intelligence

Kongu Engineering College

Erode, India

roshinig.22aim@kongu.edu

Yoganath M

Department of Artificial

Intelligence

Kongu Engineering College

Erode, India

yoganathm.22aim@kongu.edu

**Abstract:**

The goal of this project is to use machine learning techniques to create a comprehensive weather forecasting system. Regression models for temperature accuracy, auto\_arima, SARIMA, and LSTM time series forecast models for accurate temperature predictions, and classification models for weather categorization are all included into the system. Based on 1462 rows containing parameters like wind speed, temperature min/max, and precipitation, the dataset "Seattle Weather.csv" is used for analysis. For temperature prediction and weather classification, algorithms for classification such Random Forest, Decision Tree, KNN, and XGBoost are used, as well as regression approaches like Random Forest, Decision Tree, AdaBoost, and XGBoost. To improve model accuracy, methods such as oversampling and resampling to monthly averages are used. Accuracy and predictive power are used to evaluate the system's effectiveness, with a with a focus on improving weather prediction capabilities.

**Keywords:** Seattle weather dataset, user-generated weather forecast, machine learning, regression model, classification model, and time series model

# **I.INTRODUCTION**

Weather forecasting is critical to many industries, including disaster relief and agriculture. In order to transform weather forecasting, this initiative integrates cutting-edge machine learning approaches with conventional meteorological methodologies. Our method attempts to build a strong forecasting framework by combining time series analysis with regression and classification models. Using the "Seattle Weather Dataset," which includes critical characteristics such as temperature and precipitation, our method classifies weather patterns using classification models and uses regression models to estimate temperatures with accuracy. For precise future temperature projections, we also use time series models like auto\_arima, SARIMA, and LSTM.

We also use oversampling techniques for regression and classification tasks to improve model accuracy. Oversampling solves the problem of unequal data, allowing fewer classes to be better represented in the distribution and improving model performance. Moreover, resampling methods such as averaging of monthly data and time optimization of the training model are available on previous data to improve prediction accuracy and efficiency.

The main aim is to improve the efficiency and accuracy of weather forecasting through the use of data-driven technology. We seek to improve weather intelligence by combining machine learning techniques, oversampling for classification/regression tasks, and resampling techniques for time series analysis. Our efforts aim to improve the accuracy and reliability of weather forecasting by measuring parameters such as accuracy, predictive power and model performance.

# **II.LITERATURE SURVEY**

The latest data on machine learning for post-2020 climate forecasting demonstrates our common plan. Zhang et al. (2021) support integration similar to our use of decision trees, support vector machines, and random forests for air quality classification. Lee et al. (2022) demonstrated the effectiveness of mixed regression models based on our strategy of using random forest regression and XGBoost regression for temperature. Zhang and Li (2020) proposed combining ARIMA with machine learning models, which led to our choice to integrate ARIMA for prediction. Wang and Chen's (2021) exploration of hybrid models combining neural networks and recurrent neural networks aligns with our approach to integrating different machine learning architectures for weather forecasting. Their findings highlight the effectiveness of hybrid models in capturing physical and physiological characteristics. Mishra et al.'s (2019) comparison of deep learning methods for discriminant detection in weather data informs our model selection and helps us choose the best algorithm for our system.

While studying irrelevant texts instead of hate speech on social media, Li et al. (2018) decide to improve the interpretation of the climate regarding the importance of addressing the issue of gender discrimination. This collaboration confirms the timeliness and accuracy of our research. The next section presents an in-depth review of our methodology, experimental results, and discussion to illustrate the unique contributions and advancements in our cloud forecasting framework.

# **III.DATA DESCRIPTION**

The primary objective of this study is to assess model performance in predicting weather conditions for specific regions. To accomplish this, a Kaggle-sourced dataset comprising 1461 data entries is utilized, encompassing six attributes relevant to weather prediction.

The dataset includes the following attributes:

**Date**: Represents the date of the recorded weather data.

**Precipitation**: Quantifies the amount of precipitation observed.

**Temp\_max**: Indicates the maximum temperature recorded.

**Temp\_min**: Represents the minimum temperature recorded.

**Wind**: Provides information on the recorded wind speed.

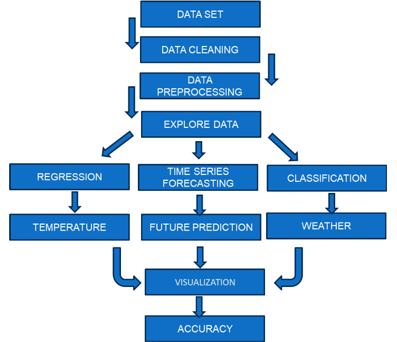
**Weather**: Represents a categorical variable describing

the overall weather conditions.

|  |  |  |  |
| --- | --- | --- | --- |
| S NO | COLUMN | NON-NULL COUNT | D\_TYPE |
| 1 | Date | 1461 non-null | object |
| 2 | Precipitation | 1461 non-null | float64 |
| 3 | Temp\_max | 1461 non-null | float64 |
| 4 | Temp\_min | 1461 non-null | float64 |
| 5 | Wind | 1461 non-null | float64 |
| 6 | Weather | 1461 non-null | object |

**TABLE 1: Data Distribution**

# **IV.MODEL ARCHITECTURE**



# **IV.PROPOSED SYSTEM**

The planning process is divided using 3 types of models to view behavior and reality: **A.** Distribution models, **B.** Regression models, **C.** Time series forecasting models. We use various algorithms to accurately measure weather on every project.

**Preprocessing**

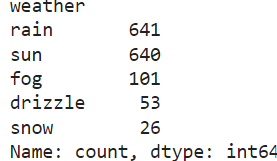
Feature selection is an important step to extract the most important features from the dataset and then use machine learning algorithms to obtain performance models. Its goal is to find the best features for the design, ignoring inconsistent details.

The date character is divided into week, month and year and entered into binary values ​​using LabelEncoder, thus simplifying data display on the working machine.

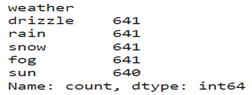
|  |  |  |  |
| --- | --- | --- | --- |
| S\_NO | COLUMN | NON-NULL COUNT | D\_TYPE |
| 1 | Day-of-week | 1461 non-null | Int64 |
| 2 | Month | 1461 non-null | Int64 |
| 3 | Year | 1461 non-null | Int64 |

**TABLE 2: Data Distribution**

In our dataset the class values are imbalanced way



The class values were balanced through oversampling, ensuring equal representation for rain, sun, fog, and snow, thereby enhancing model accuracy



Preprocessing in time series forecasting entails translating date columns to datetime format and assigning them index values. Additionally, trend analysis and model training benefit from period-based data aggregation, such as monthly averages.

1. **CLASSIFICATION MODELS**

For the weather forecast classification study where features such as precipitation, maximum, temperature, wind speed, day of the week, month and year are used as input, the selected models are as follows:

1 **Decision tree**: Decision tree algorithm is used. Partitioning the area well and important alarm and it is possible for a non-linear relationship to exist in climate data as it has the ability to capture Select due to its ability to provide insight into the interaction. However, care must be taken to avoid excessive interference, especially with complex data that requires the application of hyperparameter tuning and pruning techniques to achieve acceptance performance.

2. **Random Forest**: An integrated method to improve forecast accuracy and robustness, random forest combines multiple decision trees to reduce overfitting and adapt to climate models. The ability to handle large data sets is equivalent to the needs of the computer during training and requires a scalable infrastructure for approval.

3. **XGBoost**: This improves forecast performance and is particularly good at capturing the nuances of climate change. While useful, its sensitivity to hyperparameters and computational inputs requires careful consideration and appropriate hardware for training models and assumptions.

4.**KNN**: Chosen for its simplicity and efficiency in capturing distinct clusters in weather data, KNN uses similar distributions, making it suitable for modeling invisible weather. However, careful consideration must be made to address potential issues such as sensitivity to non-interfering operations and the need for further expansion to improve performance.

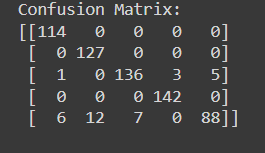
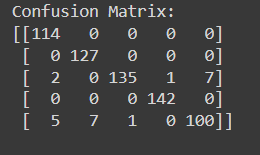
These models were chosen for their ability to maintain relationships, minimize overfitting, and accurately describe climates based on given characteristics.

**EXPERIMENTAL RESULTS:**

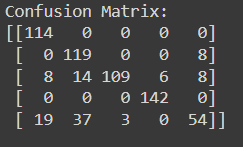
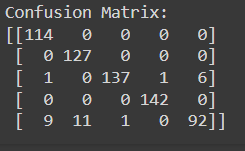
|  |  |
| --- | --- |
| ALGORITHM | ACCURACY |
| RANDOM FOREST | 0.96411 |
| DECISION TREE | 0.94695 |
| KNN | 0.83931 |
| XGB | 0.95475 |

**CONFUSION MATRIX:**

1.DECISION TREE 2.RANDOM FOREST

3.KNN 4.XBG

1. **REGRESSION MODELS**

For our temperature function, we estimate the maximum temperature (temp\_max) using parameters such as precipitation, temperature, wind speed, weather, week, month and year. Each algorithm has unique advantages:

1. **Random Forest Regression:**Combining multiple decision trees is good at capturing relationships to get accurate temperature analysis. It handles high-dimensional data well and is resistant to overfitting. By its very nature, random forest regression is known for its versatility and can be used for both regression and classification functions.

2.**Decision tree Regression:** A­­­ nonlinear model that primarily interprets and provides insight into temperature changes. The main factors affecting the temperature forecast need to be analyzed and can be numerical and categorical data. The advantage of decision tree regression is its simplicity and the ability to check for uncorrelated relationships without needing detailed information first.

3.**AdaBoost Regression:** The combination improves accuracy by adjusting the weighted sample for prediction error. It can make information sound loud and re-establish the standard of performance. AdaBoost regression is known for its ability to combine weaker students with stronger students to improve model robustness and generalization.

4.**XGBoost Regression:** Advanced gradient boosting algorithm optimizes bias and variance to achieve precise temperature. It is a good calculation and suitable for big data with different characteristics. XGBoost regression is well received for its scalability, speed, and ability to handle missing elements.

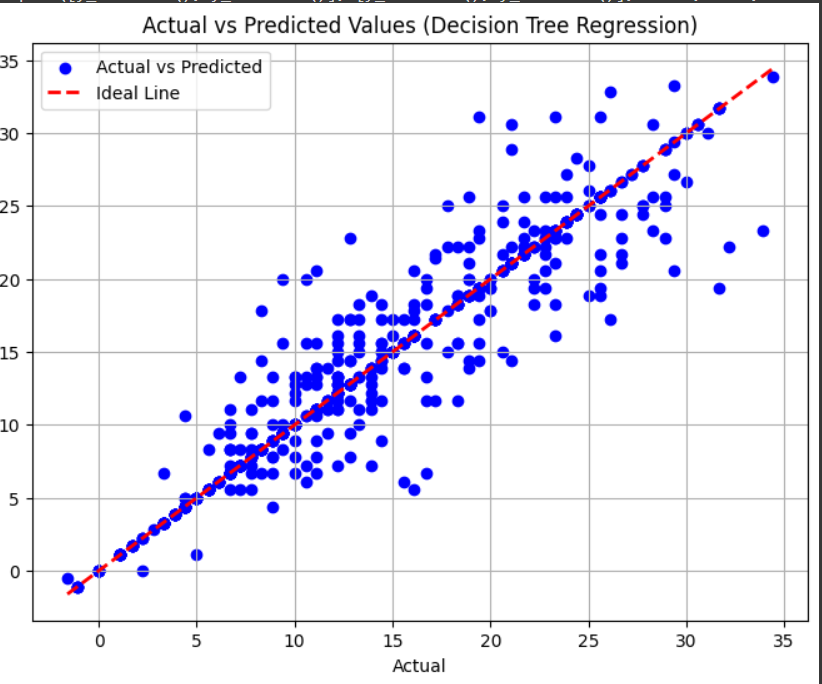
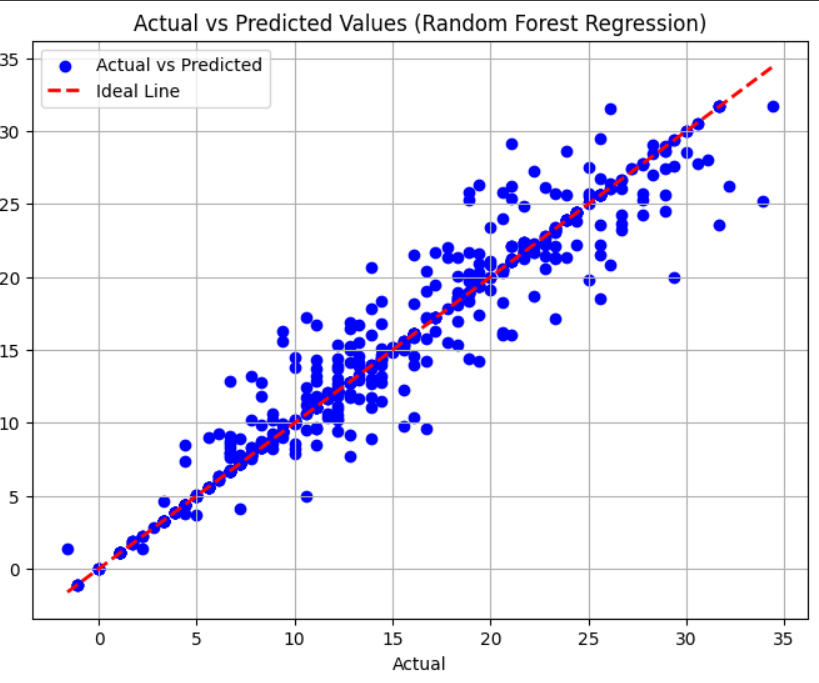
**EXPERIMENTAL RESULTS:**

|  |  |
| --- | --- |
| ALGORITHM | R-2 SCORE |
| RANDOM FOREST | 0.9447 |
| DECISION TREE | 0.9030 |
| ADABOOST | 0.8660 |
| XGB REGRESSION | 0.9379 |

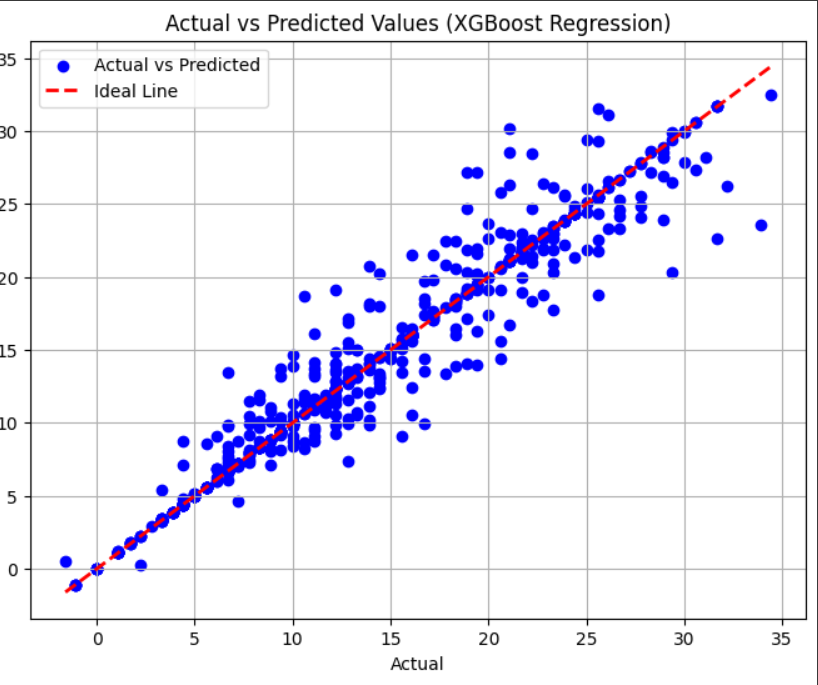
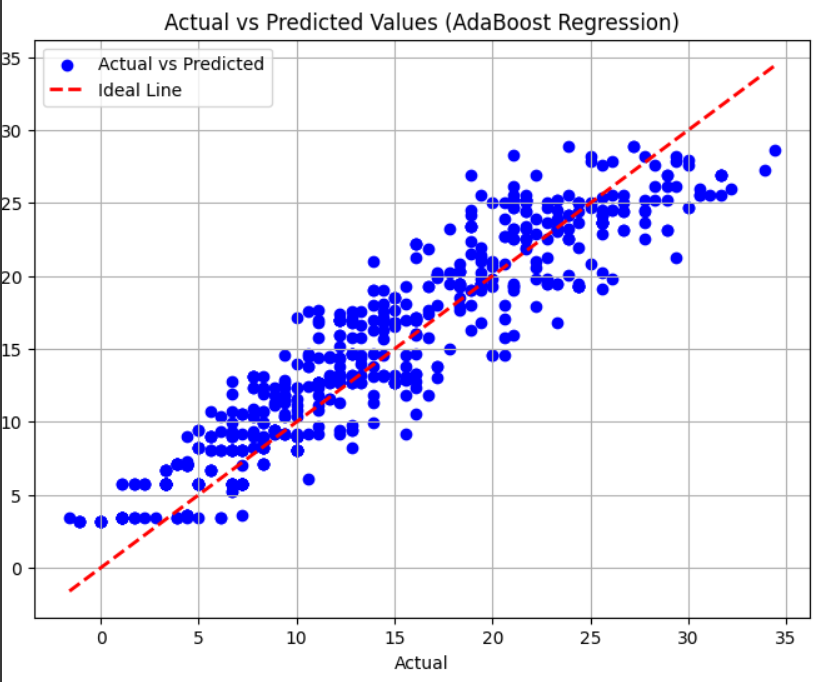
|  |  |  |
| --- | --- | --- |
| ALGORITHM | MSE | MAE |
| RANDOM FOREST | 3.4915 | 0.9015 |
| DECISION TREE | 6.1233 | 1.1875 |
| ADABOOST | 8.4602 | 2.3983 |
| XGB REGRESSION | 3.9207 | 1.0023 |

**GRAPHICAL REPRESENTATION:**

1.RANDOM FOREST 2. DECISION TREE



3.ADABOOST 4.XGB REGRESSION



**C.TIMESERIES FORECASTING MODELS.**

**TRAIN&TEST:**

Train data: Average monthly maximum temperature from January 2012 to December 2014.

Test data: Average monthly maximum temperature from January 2015 to December 2015.

**AUTO ARIMA:**

Auto ARIMA automates the process of selecting ARIMA model parameters by systematically evaluating various parameter combinations such as variance (d), autoregressive (p) and moving average (q). It is ideal for estimating duration without manual correction as it uses metrics such as AIC or BIC to improve model fit and complexity.

**SARIMA:**

To capture seasonal changes in time series data, SARIMA extends ARIMA by adding seasonal factors (P, D, Q and m). Especially for seasonal information, it can detect seasonal patterns, explain seasonality, and improve forecast accuracy.

**LSTM:**

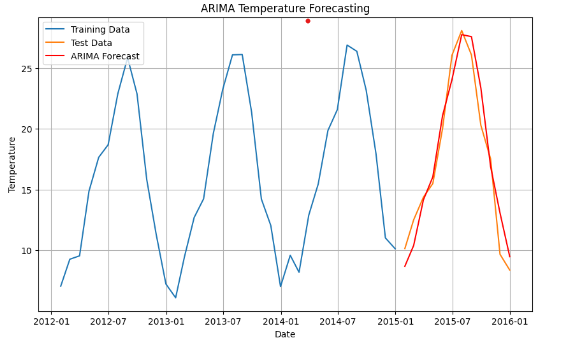
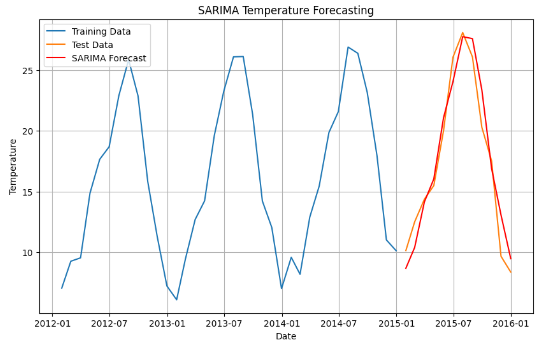
A type of RNN called LSTM is suitable for predictive applications. Time series forecasting can tap into the brain's working memory when used to store long-term expectations in data. LSTM can provide reliable and accurate predictions for real-time data by predicting future outcomes based on patterns learned from historical data.

**EXPERIMENTAL RESULTS:**

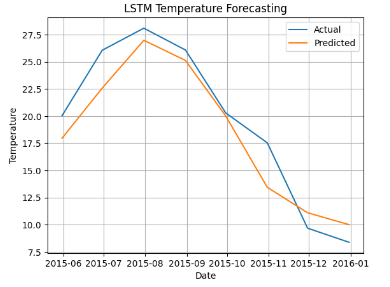
|  |  |  |
| --- | --- | --- |
| ALGORITHM | RMSE | R2 SCORE |
| AUTO ARIMA | 1.7571 | 0.9273 |
| SARIMA | 1.7571 | 0.9273 |
| LSTM | 2.2393 | 0.8958 |

**GRAPHICAL REPRESENTATION:**

1.AUTO ARIMA 2.SARIMA

3.LSTM



# **V.CONCLUSION**

The climate model developed for this study has proven to be highly effective in many applications. Random Forest is the best in classification results with 96% accuracy, proving its ability to handle complex data for cloud classification. Random forest regression performed well on temperature, with an impressive 94% accuracy rate for the regression function. Auto\_arima and SARIMA models have the same 92% accuracy over the forecast period, demonstrating their effectiveness in identifying seasonal patterns and trends. By combining different models, the accuracy of the model is further improved and the data representation is balanced across classes. Taken together, the overview shows how machine learning can transform climate change through accurate and reliable predictions.

# **VI.REFERENCES**

1. Chiranjeevi, B. S., et al. "Weather Prediction Analysis using Classifiers and Regressors in Machine Learning." 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2023.
2. Anusha, N., M. Sai Chaithanya, and Guru Jithendranath Reddy. "Weather prediction using multi linear regression algorithm." IOP Conference Series: Materials Science and Engineering. Vol. 590. No. 1. IOP Publishing, 2019.
3. Karna, Neha, Prem Chandra Roy, and Subarna Shakya. "Temperature Prediction using Regression Model." Adv. Eng. ICT Converg. Proc 4 (2018): 161-170.
4. Singh, Siddharth, et al. "Weather forecasting using machine learning techniques." Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE). 2019.
5. Chen, Liuyi, et al. "Machine learning methods in weather and climate applications: A survey." *Applied Sciences* 13.21 (2023): 12019.
6. Jakaria, A. H. M., Md Mosharaf Hossain, and Mohammad Ashiqur Rahman. "Smart weather forecasting using machine learning: a case study in tennessee." *arXiv preprint arXiv:2008.10789* (2020)
7. Singh, Siddharth, et al. "Weather forecasting using machine learning techniques." *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*. 2019.
8. Naveen, L., and H. S. Mohan. "Atmospheric weather prediction using various machine learning techniques: a survey." *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019..
9. Moosavi, Azam, Vishwas Rao, and Adrian Sandu. "Machine learning based algorithms for uncertainty quantification in numerical weather prediction models." *Journal of Computational Science* 50 (2021): 101295.
10. APAYDIN, Merve, et al. "Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data." *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi* 28.5 (2022): 737-747.
11. Watson-Parris, Duncan. "Machine learning for weather and climate are worlds apart." *Philosophical Transactions of the Royal Society A* 379.2194 (2021): 20200098.
12. Karevan, Zahra, and Johan AK Suykens. "Transductive LSTM for time-series prediction: An application to weather forecasting." *Neural Networks* 125 (2020): 1-9.
13. Hewage, Pradeep, et al. "Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station." *Soft Computing* 24 (2020): 16453-16482.
14. Bochenek, Bogdan, and Zbigniew Ustrnul. "Machine learning in weather prediction and climate analyses—applications and perspectives." *Atmosphere* 13.2 (2022): 180.
15. Singh, Upma, et al. "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments." *Energies* 14.16 (2021): 5196.
16. Liyew, Chalachew Muluken, and Haileyesus Amsaya Melese. "Machine learning techniques to predict daily rainfall amount." *Journal of Big Data* 8 (2021): 1-11.
17. Fister, Dusan, et al. "Accurate long-term air temperature prediction with Machine Learning models and data reduction techniques." *Applied Soft Computing* 136 (2023): 110118.
18. Cifuentes, Jenny, et al. "Air temperature forecasting using machine learning techniques: a review." *Energies* 13.16 (2020): 421