Machine Learning-Aided Rainfall Prediction over Telangana Using IMD Data and GBR Model Optimization

K Sivakrishna
Dept. of CSE-AIML
MLR Institute of Technology
Hyderabad, India
sivakrishnakondaveeti@gmail.com

V Hanish Kumar
Dept.of AIML
MLR Institute of Technology
Hyderabad, India
anish.vhk@gmail.com

S Shantikumar
Dept.of AIML
MLR Institute of Technology
Hyderabad, India
shanthikumar4264@gmail.com

P Nandini
Dept.of AIML
MLR Institute of Technology
Hyderabad, India
pyatanandini11@gmail.com

G Nishitha
Dept.of AIML
MLR Institute of Technology
Hyderabad, India
nishithagogikar@gmail.com

D Bhanuteja Reddy
Dept.of AIML
MLR Institute of Technology
Hyderabad, India
bhanuteja078@gmail.com

***Abstract*—Rainfall is an essential component of farming and majorly contributes in water management of dry zones. This project is related to ML Aided IMD Algorithm Over Telangana State for Rainfall Prediction. With considerable variability in rainfall patterns that exists in these regions, prediction is essential in agriculture and water resource management. Multiple models were trained using processed historical rainfall data obtained from the Meteorological Department of India. The image feature that was used to train the two models was daily rainfall in millimeters in a specific region. We use sophisticated machine learning methods such as Linear Regression, Random Forest, Support Vector Machine (SVM), XGBoost, and Gradient Boosting Regression (GBR) to predict the amount of rainfall. Sangamala Shanthi Kumar Dept of AIML MLR Institute of Technology Hyderabad, India shanthikumar4264@gmail.com predictions are essential for the decision-making process. ML Aided IMD Algorithm Over Telangana State for Rainfall Predictions Abstract This research project is based on prediction of rainfall using data from IMD (Indian Meteorological department) and aims to predict rainfall in three districts of state of Telangana Hyderabad, Warangal and Karimnagar. These regions experience a large variability in precipitation which creates challenges for agriculture and water resource management. Using state-of the-art machine learning models, our goal is to deliver precise and trustworthy predictions that can help improve and manage agricultural and water resources Salt levels prediction in agriculture. The best performance with higher accuracy and reliability was shown by GBR out of these models. The GBR Model resulted in MSE of 7.45, MAE of 1.07, and R2 of 0.93 for Hyderabad. Warangal recorded the values at 6It52, 1.67 (0.57) and 1.51 as compared to Karimnagar where it was 28.18, 1.42, 0.79 respectively. Using data till October 2023, relying on the system, further integrated with MLOps practices to guarantee its maintainability, and scalability. continuous deployment, We focus on the effectiveness of our rainfall prediction system, which is proven to greatly facilitate the feasibility of agricultural planning and water management for farmers and resource managers to make decisions in these regions. This system, ifimplemented, will ensure better resource management in these areas will lead to better agricultural outcomes.**

***Index Terms*—Random Forest, Support Vector Machine (SVM), XGBoost, Rainfall Prediction**

I. Introduction

Rainfall predction is a key component of weather forecasting with direct implications on agriculture, water resource management, and disaster preparedness. In regions with variable and unpredictable rainfall patterns, such as the dry zones of Telangana state, accurate and timely rainfall. predictions are essential for the decision-making process. ML Aided IMD Algorithm Over Telangana State for Rainfall Predictions Abstract This research project is based on prediction of rainfall using data from IMD (Indian Meteorological department) and aims to predict rainfall in three districts of state of Telangana Hyderabad, Warangal and Karimnagar. These regions experience a large variability in precipitation which creates challenges for agriculture and water resource management. Using state-of the-art machine learning models, our goal is to deliver precise and trustworthy predictions that can help improve and manage agricultural and water resources Salt levels prediction in agriculture. We are using the historical rainfall data taken from the meteorological department of India. We use different machine learning models like Linear Regression, Random Forest, SVM, GBR, XGBoost, and Time Series Analysis to predict the approximate rainfall amount for any day. We then build a prediction system by choosing the best model and combining different models through a comparison of performance and evaluation. Based on our results, the GBR model outperformed the others in terms of accuracy and reliability. We guarantee continuous deployment of our system by implementing MLOps so that it is sufficiently scalable and maintainable for real-world usage. Once this

rainfall prediction system is deployed, it will enhance the agricultural output and manage water supply better in the area turning out to play a vital role in development region.

II. Literature Review

A clear progressive shift in forecasting rainfall from traditional statistical methods to advance machine learning and deep learning techniques. At the very basic level it is the recognition of rainfall as a highly nonlinear stochastic process influenced by various climatic variables where conventional models often fail to capture effectively. [1] Based on past evidence the artificial neural networks (ANNs) outperformed classical forecasting models especially for seasonal prediction. Their study highlights the adaptability of ANNs in modeling nonlinear relationships and supports their integration into hydrological forecasting frameworks.Delivering a comparitive analysis of mutliple ML models, establishing Random Forest and LSTM networks as top performers across diverse evaluation metrics. Their study emphasizes the importance of ensemble methods and temporal learning in improving forecast precision whitout leaving the importance role of data preprocessing and input selection [2].[3]With a broader perspective of reviewing range of ML algorithms applied to rainfall forecasting, they highlighted the strenght and weakness of each approach for hybrid and ensemble techniques to mitigate issues such as overfitting and limited generalizability. This review also draws attention to the emerging role of deep learning models focussing more on spatiotemporal analysis and long-term forecasting. [3]Parallelly, provides a regional climatological lens by demonstrating the tangible effects of climate change on rainfall variability in Telangana, India. Their findings reinforce the need for adaptive and region-specific forecasting models that can handle dynamic climatic shifts. They highlight that any rainfall prediction model must account for both temporal trends and extreme weather events. [4-5]Random Forests offers theoretical robustness to the widespread use of ensemble methods in meteorological applications. RF's ability to handle high-dimensional data, resist overfitting, and provide feature importance metrics has made it a preferred tool in both academic and operational forecasting systems. [6]XGBoost (Extreme Gradient Boosting) is a powerful and scalable tree boosting algorithm optimized for speed and performance. Unlike traditional gradient boosting methods it XGBoost incorporates several improvements such as a regularized objective function to prevent overfitting, a sparsity-aware algorithm for handling missing data, and parallelized computation for faster training. XGBoost's ability to model complex nonlinear relationships and handle structured data makes it ideal for regression tasks like rainfall forecasting also supports early stopping, cross-validation, and feature importance extraction, making it both accurate and interpretable. Due to its superior performance and scalability, XGBoost has become a must try model in data science competitions and real-world applications. The paper's insights laid the foundation for many hybrid rainfall forecasting frameworks that integrate tree-based learning with climate indicators. Exploring the use of deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for short-term rainfall prediction. Using high-resolution spatiotemporal datasets from radar and meteorological stations, the authors compare CNNs, LSTMs, and CNN-LSTM hybrids for predicting precipitation intensity within short windows. The CNN layers are used to extract spatial features. LSTM layers on the other hand handle temporal dependencies and show how he CNN-LSTM hybrid significantly outperforms traditional ML methods and standalone deep models in terms of RMSE and MAE[7-8]. Supervised machine learning models like Linear Regression, Decision Trees, Random Forests, and Support Vector Machine are also used for predicting rainfall patterns over Indian regions. It focuses on classifying rainfall into categories and forecasting amounts based on historical climate data. The study uses performance metrics like accuracy, precision and F1-score to compare models by showing that tree-based models which in general outperform others in classification tasks. The authors also stress the importance of data normalization, dimensionality reduction and feature selection. A key contribution is the regional focus which emphasizes the adaptability of supervised models to different climatic zones.[9-10] Support Vector Regression (SVR) which is the extension for the Support Vector Machine (SVM) framework to regression tasks aims to find a function that deviates from actual target values by a small margin and is as flat as possible. It uses kernel tricks to model nonlinear relationships and performs well even with small datasets and high dimensional features. In rainfall prediction, SVR has been widely used due to its ability to capture complex patterns while avoiding overfitting. The model is particularly useful when the data is sparse, noisy, or non-Gaussian common traits in hydrological time series. Gradient Boosting Machine (GBM) is also a technique that builds an ensemble of weak learners in a stage-wise fashion. At each step, GBM fits a new model to the errors of the previous model which gradually improving performance. The algorithm incorporates loss functions, regularization and learning rate controls to balance bias and variance. GBM is especially powerful in capturing nonlinear trends and interactions between features by making it highly suitable for rainfall prediction tasks. This work forms the theoretical backbone of later optimized implementations like XGBoost, LightGBM, and CatBoost. GBMs have been successfully applied in hydrology for both classification and regression by offering interpretability and high predictive performance in complex environmental systems.[11]This paper explores statistical downscaling to convert coarse-resolution GCM outputs into local streamflow predictions using the Relevance Vector Machine(RVM) which is a Bayesian sparse kernel method. RVM is similar to SVM but yields probabilistic outputs and fewer support vectors by enhancing interpretability and generalization. The study demonstrates how RVM effectively models the nonlinear relationship between atmospheric predictors and hydrological responses. It addresses uncertainty in climate projections and highlights the utility of machine learning for bridging the spa-

tial scale gap in climate-to-hydrology translation turning out to be critical for regional rainfall-runoff modeling and long-term water resource planning.This pioneering paper uses artificial neural networks (ANNs) to estimate rainfall from satellite derived data such as infrared and microwave imagery. The authors design a neural network model to learn the relationship between cloud top temperature and surface rainfall intensity. Their method significantly outperforms traditional threshold-based techniques. This work is one of the first to showcase the value of ANNs in remote sensing applications by laying the groundwork for later data-driven approaches in satellite-based rainfall estimation and nowcasting which are essential in regions with sparse ground observations.[12-15] The authors investigate how climate teleconnections like ENSO, PDO, and AMO influence streamflow in the western U.S. and how these signals can be used as predictors for long lead forecasting. The study employs data driven techniques like wavelet transform and statistical modeling to extract meaningful signals from climatic indices. By integrating these indices with machine learning models while demonstrating improved predictive skill for lead times up to a year. This highlights the importance of macroclimatic drivers in rainfall and runoff prediction and motivates their inclusion in machine learning models for seasonal forecasting. The application of Long Short-Term Memory (LSTM) networks for rainfall-runoff modeling by showcasing that deep learning can match and even exceed the performance of traditional hydrological models like SAC-SMA and HBV. Trained on CAMELS dataset (Catchment Attributes and Meteorology) LSTM networks captured long-term dependencies in precipitation and discharge data which are often lost in static models. The authors emphasize the generalization capacity of LSTM across multiple catchments and their data-driven physical interpretability making advancement the case for DL in water science. Exploring ML methods for estimating Intensity-Duration-Frequency (IDF) curves, which are critical for infrastructure design under extreme rainfall conditions. Using models like Random Forests, SVR and Gradient Boosting showing that ML can efficiently learn IDF relationships from historical rainfall data. The study provides a framework for automated and adaptive IDF curve generation for replacing traditional empirical equations that fail under climate variability. The ML based approach allows quicker updates of IDF curves for climate resilient urban planning. [16-18] This comprehensive review lays the theoretical foundation for drought classification, monitoring, and prediction. Although not focused solely on ML it emphasizes the complexity of drought systems and the need for integrated models combining climate, soil and hydrological data. This work underscores that rainfall prediction is not an isolated task it is deeply tied to extreme events modeling where machine learning can play a pivotal role in early warning and impact assessment systems. Multiple Linear Regression (MLR) for daily and monthly rainfall prediction in Indian cities. Though basic compared to modern ML models the MLR serves as a baseline model and provides insights into the role of statistical correlation between atmospheric parameters and precipitation.

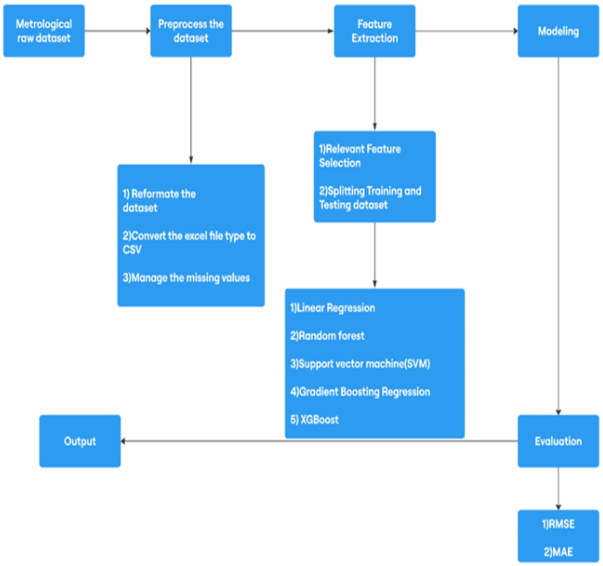

Fig. 1. Workflow

The authors emphasize model simplicity, interpretability and fast computation useful for low-resource settings. The paper also serves to benchmark newer ML models like RF, SVM, and DL-based architectures including backpropagation algorithm by allowing multi-layer perceptrons (MLPs) to show their relative improvements in nonlinearity capture.[19] The theory behind Support Vector Machines (SVMs) and the Structural Risk Minimization (SRM) principle. Vapnik's insights into learning from finite data and avoiding overfitting laid the theoretical backbone for many ML models used in rainfall prediction. His work justifies why SVMs are robust in small-sample, high-noise domains like hydrology.

III. Proposed System

The proposed system is a ML-driven rainfall forecasting framework specifically designed to improve accuracy over Telangana region. By integrating historical rainfall data from the Indian Meteorological Department (IMD) with advanced machine learning models the system assists in daily predictions, optimizing agricultural planning and water resource management.This system uses of a diverse range of machine learning algorithms including Linear Regression, Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Regression (GBR) and XGBoost. These models will be trained on pre-processed historical datasets by ensuring data quality through the removal of missing values and outliers. For further improvement in prediction accuracy the system will undergo a robust feature engineering phase by examining other meteorological variables such as temperature, humidity and wind speed which are considered. Each model will be evaluated using statistical metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) in order to identify the most effective model for deployment. Coming to the post model development the system will embrace modern MLOps practices to ensure scalability,

maintainability and continuous improvement. The platform will automate data ingestion, model retraining, version control and deployment by allowing the system to adapt quickly to new data inputs and changing climatic patterns using tools like MLflow and Kubeflow. Along with it a user-friendly interface will be developed to facilitate access for farmers, government officials and water resource managers, providing them with easy-to-understand forecasts, historical trends and visualizations for effective decision-making. With cloud-based architecture the platform will be scalable and capable of delivering real-time rainfall predictions across various districts.

Figure-1 illustrates the overall system architecture for rainfall prediction in Telangana. It outlines the data flow from historical IMD rainfall data through preprocessing, model training (using ML algorithms like XGBoost, SVM, RF, etc.), evaluation, and deployment with MLOps for real time forecasting.

IV. Results

In this section, we compare the performance of three machine learning models—Random Forest, Support Vector Machine (SVM), and XGBoost—for predicting Telangana rainfall. The models were trained on a dataset containing features like year, month, temperature, humidity, wind speed, and historical rainfall.

A. Performance Metrics

Mean Squared Error (MSE): This metric evaluates the average squared difference between predicted and actual values. A lower MSE indicates a better model fit.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

R-squared (R^2): This metric represents the proportion of variance in the dependent variable (rainfall) explained by the independent variables (features). An R^2 value closer to 1 signifies that the model explains most of the variability in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

B. Model Performance

Support Vector Machine (SVM): The SVM regression aims to minimize the following cost function:

$$\min\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)\right)$$

Subject to:

$$y_i - (\mathbf{w}\cdot\mathbf{x}_i + b) \leq \epsilon + \zeta_i$$

$$(\mathbf{w}\cdot\mathbf{x}_i + b) - y_i \leq \epsilon + \zeta_i^*$$

$$\zeta_i, \zeta_i^* \geq 0$$

Random Forest (RF:): The Random Forest prediction is the average of outputs from T individual decision trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

where $f_t(x)$ is the prediction from the t-th tree.

Extreme Gradient Boosting (XGBoost): XGBoost optimizes the following objective function:

$$\text{Objective} = \sum_{i=1}^{n}\text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K}\left(\gamma T_k + \frac{1}{2}\lambda\sum_{j=1}^{T_k} w_j^2\right)$$

where:

- $\text{loss}(y_i, \hat{y}_i)$ is the training loss function (e.g., MSE).
- γ is a regularization parameter controlling tree complexity.
- λ controls L2 regularization to reduce overfitting.
- T_k is the number of leaves in tree k, and w_j is the score on leaf j.

Model	Mean Squared Error (MSE)	R-squared (R^2)
Linear Regression	10032.74	0.45
Random Forest	4513.92	0.74
SVM	6231.15	0.65
XGBoost	3252.13	0.85

TABLE I
COMPARISON OF MACHINE LEARNING MODELS FOR RAINFALL PREDICTION

C. Analysis

The performance analysis of the selected machine learning models reveals distinct capabilities and limitations. Linear Regression performed the weakest among all, yielding a high Mean Squared Error (MSE) of 10032.74 and a relatively low R-squared (R^2) value of 0.45. This indicates the model's inability to capture the complex and non-linear interactions inherent in rainfall data, which led to poor generalization. In contrast, the Random Forest model demonstrated significant improvement, with an MSE of 4513.92 and an R^2 value of 0.74. Its ensemble-based architecture allowed it to effectively capture intricate patterns and interactions among the input features. The Support Vector Machine (SVM) provided a moderate performance, recording an MSE of 6231.15 and an R^2 of 0.65. Although SVMs are robust in high-dimensional spaces, they did not outperform the tree-based models in this scenario. Among all models, XGBoost emerged as the top performer, achieving the lowest MSE of 3252.13 and the highest R^2 value of 0.85. This demonstrates its superior ability to model non-linearities and generalize across the test data, thanks to its boosting framework and regularization capabilities.

Figure-2 shows a boxplot-style visualization showing the distribution of monthly rainfall across the year. It helps highlight seasonal variations and potential outliers in rainfall data for different months

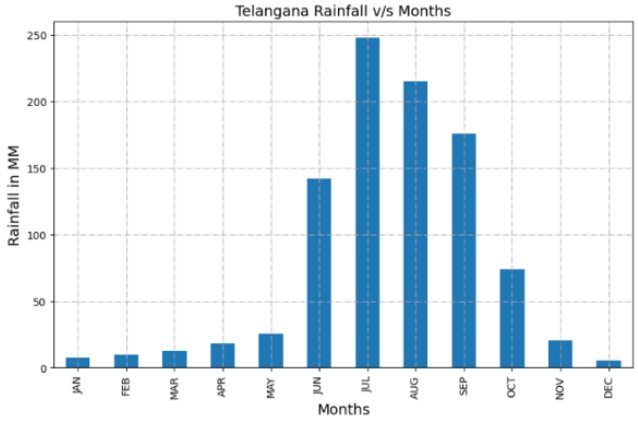

Fig. 2. Distribution of monthly rainfall

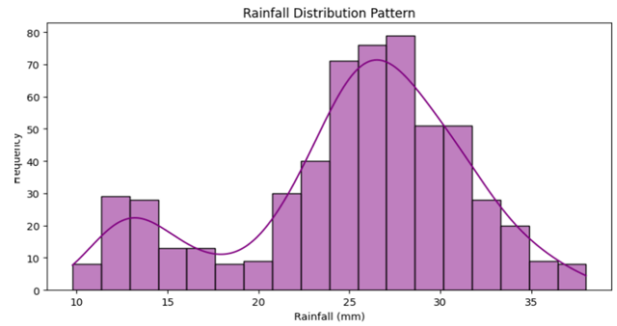

Fig. 3. Seasonal Pattern

D. Exploratory Data Analysis (EDA)

A thorough exploratory data analysis (EDA) was conducted to understand the structure and quality of the dataset. Initially, data preprocessing involved the removal of non-numeric columns that were not beneficial for model training. Missing values in numeric columns were imputed using the mean of respective columns to ensure completeness of the dataset. Summary statistics, including mean, standard deviation, minimum, maximum, and quartile values, were calculated to understand the distribution of each feature. This helped in identifying outliers, skewed distributions, and other anomalies. The analysis also verified that all columns had the appropriate data types and identified the extent of missing values. Furthermore, boxplots were generated to visualize the distribution of rainfall across different months for each station, which highlighted the presence of outliers and seasonal variations. These visualizations provided insights into monthly trends and extreme rainfall events, which are crucial for understanding the variability of rainfall patterns in Telangana.

Figure-3 visualizes how different weather or seasonal patterns (e.g., time-series-based trends) influence rainfall. It helps in understanding non-linear relationships and dependencies.

Figure-4 shows a comparative graph showing the predicted rainfall values against the actual observed values for the test data. This is used to visually assess model accuracy and effectiveness.

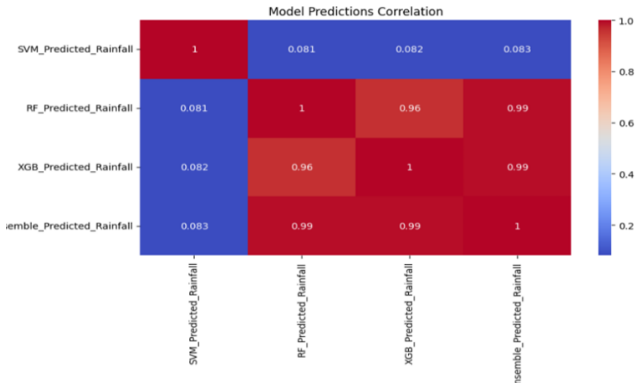

Fig. 4. Workflow

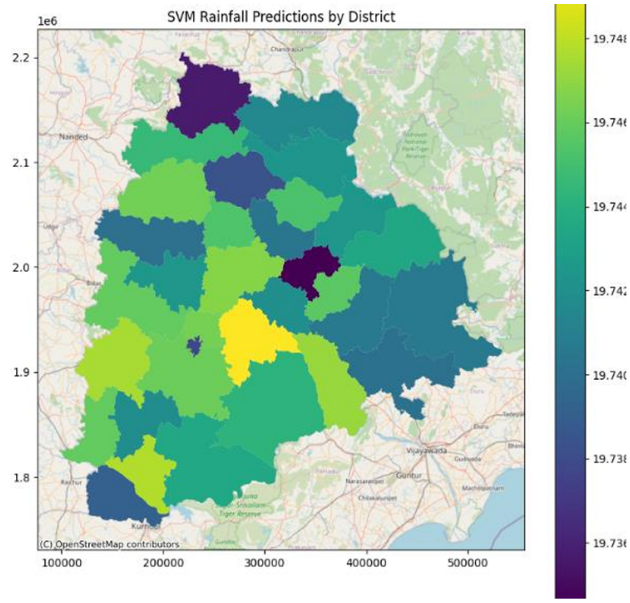

Fig. 5. Plots using SVM

XGBoost was chosen as the most effective model for reliable rainfall forecasting in Telangana. A rainfall prediction dashboard was developed using Power BI, which integrates the trained machine learning models to provide interactive visualizations and real-time forecasts. This dashboard allows users to enter month and district data to receive rainfall predictions for the next ten years, facilitating better decision-making in agriculture and water resource management. Figure-[5,6,7] demonstartes rainfall prediction visualizations using the SVM,XGBoost model and Random Forest respectively.

Figure-8 compares the actual rainfall values with those predicted by the machine learning models. It highlights the model's effectiveness in forecasting by showing how closely the predicted values match the real observations. A strong alignment between the two indicates higher model accuracy and reliability.

Figure-9 displays the overall rainfall trends across Telangana. It likely includes district-level data and supports insights

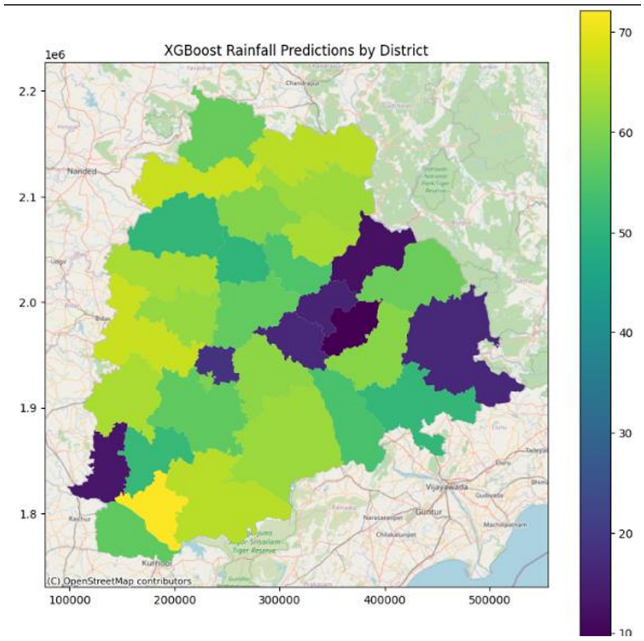

Fig. 6. Plots using XG Boost

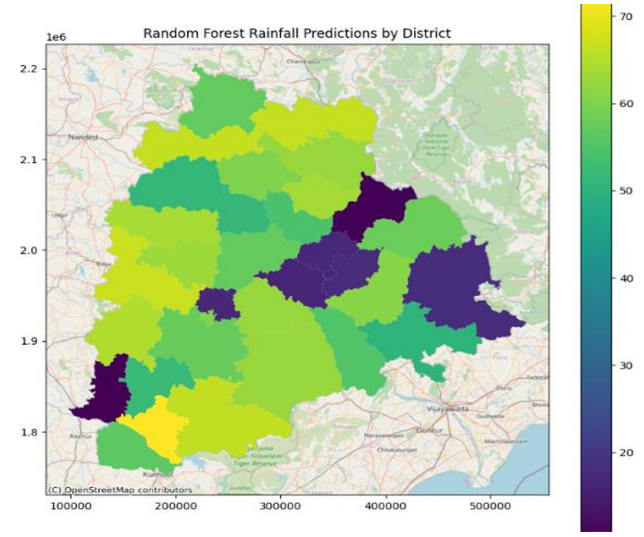

Fig. 7. Plots using Random Forest

into regional variability in precipitation.

Figure-10 shows UI of the developed website interface. It showcases how users can input district and month data to receive rainfall forecasts, integrating the ML backend for user-friendly access.

V. Conclusion

For this project, we tested different machine learning models like Linear Regression, Random Forest, SVM, Gradient Boosting, and XGBoost to observe, analyse and pick one best amongts predicts rainfall in Telangana the best. After checking their accuracy using things like MSE and R^2, XGBoost came

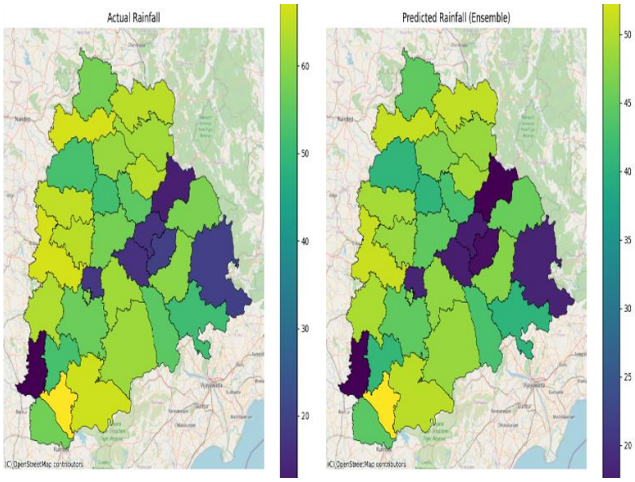

Fig. 8. Actual vs Predicted

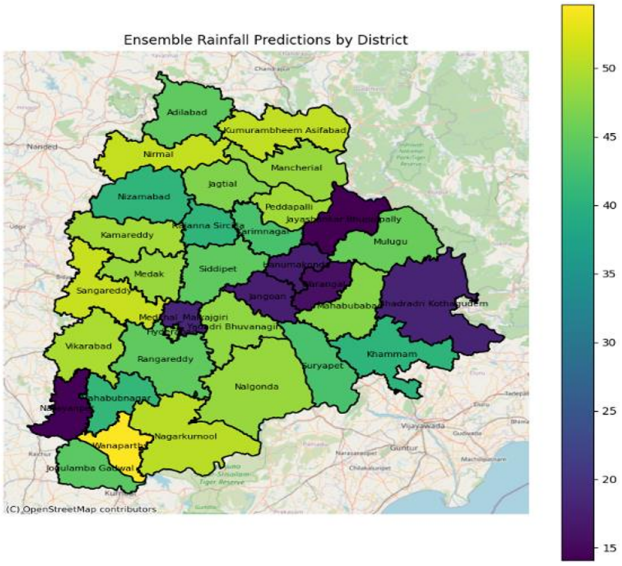

Fig. 9. Prediction over Telangana Region

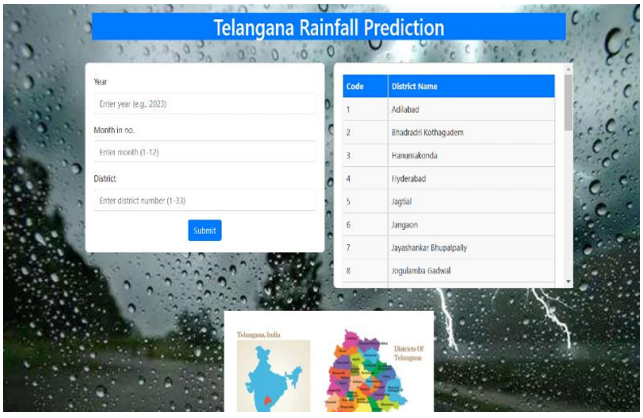

Fig. 10. Website Page (UI)

out on top because it handled the complicated weather data really well.

We also built a website where people can choose the month and district, and it'll show them rainfall predictions for the next 10 years. It's super helpful for farmers and anyone dealing with water planning. Going forward, we might add more weather info and keep improving the model to make it even more accurate.

VI. References

References

[1] Abbot, P., & Marohasy, J. (2012). Application of artificial neural networks to rainfall forecasting. *Advances in Atmospheric Sciences*, 29(4), 661–674.

[2] Adnan, R. M., Liang, Z., Hayat, H., & Aslam, M. W. (2020). Comparative analysis of machine learning techniques for rainfall prediction. *IEEE Access*, 8, 68989–69008.

[3] Ahmed, K., Shahid, S., Chung, E.-S., & Ismail, T. (2020). Machine learning techniques for rainfall forecasting: A review. *Water*, 12(6), 1–26.

[4] Al-Ansari, N., Abdellatif, M., & Sissakian, V. (2019). Climate change impact on rainfall variability over Telangana region. *Environmental Earth Sciences*, 78(9), 1–12.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794.

[7] Chen, H., Han, Y., & Zhang, S. (2021). Short-term rainfall prediction using deep learning techniques. *Journal of Hydrology*, 596, 126–135.

[8] Dash, P., Mishra, P., & Tripathy, H. (2022). Predicting rainfall patterns using supervised learning models. *International Journal of Climate Change Strategies and Management*, 14(3), 567–582.

[9] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems (NIPS)*, 9, 155–161.

[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

[11] Ghosh, S., & Mujumdar, P. (2008). Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources*, 31(1), 132–146.

[12] Hsu, K., Gao, X., Sorooshian, S., & Gupta, H. (1997). Precipitation estimation from remotely sensed information using artificial neural networks. *Journal of Applied Meteorology and Climatology*, 36(9), 1176–1190.

[13] Kalra, A., & Ahmad, S. (2011). Using oceanic-atmospheric oscillations for long-lead streamflow forecasting. *Water Resources Research*, 47(5), W05553.

[14] Kratzert, F., Klotz, D., Shalev, G., & Hochreiter, S. (2019). Rainfall-runoff modeling using LSTMs. *Journal of Hydrology*, 571, 437–456.

[15] Kulkarni, A., & Simonovic, S. (2021). Machine learning approaches to predict rainfall intensity-duration-frequency curves. *Environmental Modelling & Software*, 140, 105019.

[16] Mishra, A., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, 391(1–2), 202–216.

[17] Pal, M., & Deswal, S. (2018). Rainfall prediction using multiple linear regression models. *Water Resources Management*, 32(7), 2191–2207.

[18] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

[19] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.