Journal of Agriculture, Biology and Applied Statistics
Vol. 4, Nos. 1-2, 2025, pp. 27-34

URL : www.arfjournals.com
https://DOI:10.47509/JABAS.2025.v04i1-2.04

Monthly Rainfall Prediction Using Machine Learning Models: A Comparative Study of Decision Tree, SVM, and KNN in Rajnandgaon District

Devika Santhosh[1], Dr. Lakshmi Narsimhaiah[2] and Vijayakumar S[3]

[1]M.Sc. Research Scholar, Department of Agricultural Statistics and Social Science (L.), College of Agriculture, IGKV, Raipur, (C.G.), India.
[2]Assistant Professor, Department of Agricultural Statistics and Social Science (L.), College of Agriculture, IGKV, Raipur, (C.G.), India.
[3]M.Sc. Research Scholar, Department of Agricultural Statistics and Social Science (L.), College of Agriculture, IGKV, Raipur, (C.G.), India.

To cite this article

Devika Santhosh, Dr. Lakshmi Narsimhaiah and Vijayakumar S (2025). Monthly Rainfall Prediction Using Machine Learning Models: A Comparative Study of Decision Tree, SVM, and KNN in Rajnandgaon District. Vol. 4, Nos. 1-2, pp. 27-34.

Abstract: Accurate rainfall prediction is critical for agricultural planning and water resource management in agrarian regions. This study presents a comparative analysis of three machine learning models—Decision Tree (DT), SVM, and K-Nearest Neighbors (KNN)—for predicting monthly rainfall in the Rajnandgaon district of Chhattisgarh, India. Historical meteorological data was sourced from the NASA POWER repository and preprocessed to handle missing values, outliers, and feature scaling. The models were implemented in Python and evaluated using standard metrics: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Results demonstrated that the KNN model outperformed the others, achieving the highest R^2 value (0.81) and the lowest errors (RMSE: 61.30 mm, MAE: 38.16 mm). Based on its superior performance, the KNN model was employed to forecast monthly rainfall for the period 2023–2028, revealing significant inter-annual variability. The findings indicate that KNN is a highly effective tool for rainfall prediction in this region, offering valuable insights for stakeholders in agriculture and water management to enhance climate resilience and decision-making. And KNN is used to forecast Average monthly rainfall from 2023 to 2028.

Keywords: Rainfall Prediction, Rajnangaon, Machine Learning, KNN, Support Vector Machine, Decision Tree

Introduction

Rainfall is incredibly important for life and the economy, especially in farming regions like the Rajnandgaon district of India. The amount and timing of rain directly affects crop success,

Article History Received : 10 April 2025; Revised : 08 May 2025; Accepted : 19 May 2025; Published : 02 June 2025

income, and access to water for many people. Unpredictable rain can lead to droughts or floods, causing serious problems for the community. Therefore, finding ways to predict rainfall accurately is a crucial task for farmers and planners.

Traditional weather forecasting methods can be complex and are not always perfect for predicting local rain patterns. This is where modern technology can help. Machine Learning (ML) is a type of artificial intelligence that allows computers to learn from past weather data to find patterns and make predictions. Models like Decision Tree, SVM, and KNN are particularly useful for this job because they can handle complex relationships in nature and provide reliable forecasts.

This study focuses on using these three machine learning models—Decision Tree, SVM, and KNN—to predict monthly rainfall in Rajnandgaon district. The main goal is to test these models using historical weather data, compare their accuracy, and identify which one is the most accurate. This best-performing model will then be used to forecast monthly rainfall for the next six years. An accurate long-term forecast can serve as a powerful tool for the district, helping everyone from farmers to local officials make better decisions and prepare for the future.

Materials and Methods

Study Area and Data Collection

Description of the Study Area: This study is conducted in the Rajnandgaon district, located in the southern part of the state of Chhattisgarh, India. It lies between latitudes 21°10' N and longitudes 81°03' E. The district covers a total geographical area of approximately 8,70 square kilometers. The climate of Rajnandgaon is classified as tropical, characterized by a hot summer, a humid monsoon, and mild winter. The district receives the majority of its annual precipitation from the southwest monsoon, which typically occurs between June and September. The average annual rainfall is around 1,200 mm, but it exhibits high spatial and temporal variability. This variability makes accurate prediction crucial for the district's economy, which is predominantly agrarian. Key crops include paddy rice, maize, and pulses, all of which are heavily dependent on the timing and amount of monsoon rainfall. Therefore, Rajnandgaon serves as a highly relevant and critical site for a study on rainfall prediction.

Data Collection and Preprocessing

Data Collection: This study will utilize historical meteorological data to develop and train the machine learning models. The primary data source for this research is the NASA Prediction of Worldwide Energy Resources (POWER) project, accessible through its Data Access Viewer (https://power.larc.nasa.gov/data-access-viewer/).

Methodology

Data Preprocessing

Data preprocessing is a critical prerequisite in machine learning, encompassing the techniques used to organize, clean, and transform raw data into a structured format suitable for modeling. This study will employ a multi-step preprocessing pipeline:

1. Handling Missing Values: The dataset will be checked for missing or anomalous entries. Given the reliability of the NASA POWER dataset, significant gaps are not anticipated. However, any minor missing values will be addressed using appropriate imputation techniques, such as linear interpolation or time-series averaging, to preserve the integrity of the temporal sequence.
2. Outlier Detection and Treatment: Statistical methods and visualization tools (e.g., box plots, interquartile range (IQR) analysis) will be used to identify potential outliers. These outliers will be investigated to determine if they represent genuine extreme weather events or data errors. Legitimate extremes will be retained, while erroneous entries will be corrected or removed.
3. Data Splitting: The preprocessed dataset will be partitioned into two subsets to ensure robust evaluation:
    - Training Set: 80% of the data will be used to train the machine learning models. This allows the algorithms to learn the underlying patterns and relationships between the input features and the target variable (rainfall).
    - Testing Set: The remaining 20% of the data will be held out and used exclusively for testing the final models. This provides an unbiased evaluation of the models' performance and their ability to generalize to new, unseen data.

Analytical Methods and Model Implementation

All data preprocessing, analysis, model development, and visualization will be conducted using the Python programming language within a Jupyter Notebook environment. Python offers a powerful and comprehensive suite of open-source libraries for scientific computing, including pandas for data manipulation, scikit-learn (sklearn) for machine learning, and matplotlib and seaborn for visualization.

Three distinct machine learning algorithms will be implemented and compared for the task of monthly rainfall prediction:

1. Decision Tree (DT): A supervised learning algorithm that models decisions and their potential consequences as a tree-like structure. It will be employed for its interpretability and ability to capture non-linear relationships without requiring extensive data preprocessing.

2. Support Vector Machine (SVM): A powerful algorithm used for both classification and regression tasks. For this continuous prediction problem, Support Vector Regression (SVR) will be used. It works by finding a hyperplane in a high-dimensional space that best fits the data while maximizing the margin of error tolerance.
3. K-Nearest Neighbors (KNN): A simple, instance-based learning algorithm for regression. It predicts the target variable for a new data point by averaging the values of its 'k' most similar neighbors in the training set. The optimal value for 'k' will be determined empirically.

Evaluation criteria: To compare the model accuracy the following given in Table 1 evaluation criteria have been used:

Table 1: Evaluation Criteria used to determine the best model

S. No	*Evaluation Criteria*	*Formula*
1.	Root Mean Square Error (RMSE)	RMSE = $RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{yi})^2}$
2.	Mean squared Error (MSE)	$MSE = \frac{1}{N}\Sigma_{i=1}^{N} \mid y_i - \hat{yi} \mid$
3.	Coefficient of determination	$R^2 = 1 - \frac{\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{N} y_i^2}$
4.	Mean Absolute Error (MAE)	$MAE = \frac{1}{n}(y_i - \hat{yi})$

y_i and $\hat{yi}$ are the actual value and predicted value of the response variable and N is the number of data points. The best-performing model was then applied to forecast monthly rainfall for Rajnandgaon for the years 2023 to 2028 using historical rainfall time series data as input.

Result and Discussion

To assess the predictive accuracy of three machine learning algorithms-Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)-for monthly rainfall estimation, a comparative analysis was conducted across the district Rajnandgaon. The models were rigorously evaluated using a set of standard statistical indicators: the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). A summary of the performance metrics for each model across the three districts is presented in Table 2.

Table 2: Performance comparison of Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) models for monthly rainfall prediction in Rajnandgaon district of Chhattisgarh, evaluated using R^2, RMSE, MSE, and MAE

	Decision Tree	*KNN*	*SVM*
R2	0.74	0.81	0.76
RMSE	70.95	61.30	67.86
MSE	5031.13	3757.76	4606.24
MAE	41.95	38.16	48.47

Among the three models, the K-Nearest Neighbors (KNN) algorithm demonstrated superior performance across all evaluation metrics. It achieved the highest Coefficient of Determination (R^2) value of 0.81, indicating that it explains 81% of the variance in the monthly rainfall data, which is a significantly better fit to the observed data than the Decision Tree ($R^2 = 0.74$) and SVM ($R^2 = 0.76$) models. Furthermore, the KNN model registered the lowest error rates, with an RMSE of 61.30 mm, an MSE of 3757.76, and an MAE of 38.16 mm. These error metrics confirm that the KNN model's predictions are, on average, closer to the actual observed rainfall values, with a smaller average magnitude of error and fewer large outliers.

In contrast, the Support Vector Machine (SVM) model ranked second, with moderately strong results. While its R^2 value of 0.76 is respectable, its higher error scores, particularly an MAE of 48.47 mm, suggest its predictions have a larger average deviation from the actual values compared to the KNN model. The Decision Tree (DT) model performed the least effectively of the three, yielding the lowest R^2 value and the highest error metrics, including an RMSE of 70.95 mm and an MSE exceeding 5000.

In conclusion, based on this comprehensive evaluation, the K-Nearest Neighbors model is identified as the most accurate and reliable for predicting monthly rainfall in this study. Its optimal balance of high explanatory power and minimal prediction error makes it the recommended model for this application. The performance hierarchy of the models, from best to worst, is KNN, followed by SVM, and then Decision Tree.

Based on the superior performance of the K-Nearest Neighbors (KNN) model for rainfall prediction, projections for the Rajnandgaon district were generated for the forecast period spanning 2023 to 2028 in Table 3. Utilizing historical monthly rainfall data from 1991 to 2022, the model produced estimates that indicate noticeable year-to-year variability in precipitation. This variability aligns with the inherent climatic fluctuations characteristic of the region's tropical environment. The projections aim to provide valuable insights for water resource management and agricultural planning in the district.

Table 3: Forecasted Monthly Average Rainfall (mm) in Rajnandgaon District (2023-2028)

Month	*2023*	*2024*	*2025*	*2026*	*2027*	*2028*
January	13.20	33.60	12.70	17.50	11.60	32.00
February	14.80	11.10	18.80	14.50	13.40	19.00
March	10.60	5.80	6.50	6.00	10.50	6.10
April	14.90	19.30	21.40	19.40	7.80	16.50
May	22.10	73.30	18.80	18.80	46.40	21.10
June	265.50	188.90	122.00	255.30	196.10	210.10
July	442.60	403.90	334.60	268.50	318.30	298.30
August	259.70	310.90	326.20	363.40	309.20	320.40
September	147.60	161.90	226.40	195.20	213.60	213.60
October	17.50	31.60	99.40	72.60	62.20	54.70
November	1.00	9.90	25.10	4.10	1.00	1.10
December	7.60	8.00	1.70	10.50	1.80	9.30
Total Monthly average Rainfall	1217.10	1258.20	1213.60	1245.80	1191.90	1202.20

Conclusion

This research undertook a comparative analysis of three prominent machine learning algorithms-Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)-for the purpose of predicting monthly rainfall in the Rajnandgaon district. The objective was to identify the most effective and reliable data-driven model to address the critical need for accurate rainfall forecasting in this agriculturally dependent region.

The evaluation, based on a rigorous set of statistical metrics, yielded a clear hierarchy in model performance. The K-Nearest Neighbors (KNN) algorithm consistently outperformed its counterparts, registering the highest Coefficient of Determination ($R^2 = 0.81$). This indicates that the KNN model successfully explains approximately 81% of the variance present in the historical rainfall data, signifying an excellent fit. Furthermore, it achieved the lowest error values across all metrics: Root Mean Squared Error (RMSE = 61.30 mm), Mean Squared Error (MSE = 3757.76), and Mean Absolute Error (MAE = 38.16 mm). These error values, which are expressed in millimeters of rainfall, demonstrate that the KNN model's predictions have the smallest average deviation from the actual observed values, making it the most accurate predictor.

In contrast, the Support Vector Machine (SVM) model displayed moderate performance, while the Decision Tree (DT) model was found to be the least accurate for this specific dataset and prediction task. The superior performance of the KNN model can be attributed to its inherent ability to effectively capture localized patterns and non-linear relationships within the climate data without making strong underlying assumptions about the data's distribution.

Therefore, this study conclusively identifies the K-Nearest Neighbors algorithm as the optimal model for monthly rainfall prediction in the Rajnandgaon district. The implementation of this KNN-based forecasting tool can provide local farmers, water resource managers, and policymakers with valuable insights, enabling more informed decision-making for crop selection, irrigation planning, and drought mitigation strategies, thereby enhancing the region's climate resilience.

References

1. Shukla, R., Kumar, A., & Singh, P. (2024). Application of machine learning models in rainfall prediction over Indian monsoon regions. *Theoretical and Applied Climatology, 157*(3), 1345–1359. https://doi.org/10.1007/s00704-023-04325-1
2. Wani, O. A., Khan, F. A., & Bhat, M. S. (2023). Evaluation of machine learning approaches for rainfall prediction: A case study of Indian states. *Environmental Monitoring and Assessment, 195*(6), 733. https://doi.org/10.1007/s10661-023-11378-7
3. Patel, R., Jha, A., & Kumar, S. (2023). Performance comparison of statistical and machine learning techniques for rainfall forecasting in India. *Modeling Earth Systems and Environment, 9*(2), 1451–1464. https://doi.org/10.1007/s40808-022-01557-y
4. NASA POWER. (2024). NASA Prediction of Worldwide Energy Resources (POWER) Data Access Viewer. Retrieved from https://power.larc.nasa.gov/data-access-viewer/
5. India Meteorological Department (IMD). (2024). Mausam: Weather Information and Data Services. Retrieved from https://mausam.imd.gov.in
6. IBM. (2024). Decision tree and support vector machine explained. Retrieved from https://www.ibm.com/topics
7. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
9. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964
10. Ghosh, S., & Mujumdar, P. P. (2008). Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources, 31*(1), 132–146. https://doi.org/10.1016/j.advwatres.2007.07.005
11. Kannan, S., & Ghosh, S. (2013). A nonparametric kernel regression model for downscaling precipitation to river basin in India. *Hydrology and Earth System Sciences, 17*(3), 887–905. https://doi.org/10.5194/hess-17-887-2013
12. Mishra, A. K., & Desai, V. R. (2005). Drought forecasting using stochastic models. *Stochastic Environmental Research and Risk Assessment, 19*(5), 326–339. https://doi.org/10.1007/s00477-005-0238-4
13. Bhowmik, A. K., & Das, D. (2019). Rainfall variability and trend analysis in Chhattisgarh. *Journal of Agrometeorology, 21*(1), 29–36. https://doi.org/10.54386/jam.v21i1.526

14. Chattopadhyay, S., & Chattopadhyay, G. (2010). Identification of best predictor variables for Indian summer monsoon rainfall using statistical and neural network models. *Atmospheric Research, 98*(2–4), 281–289. https://doi.org/10.1016/j.atmosres.2010.07.011

15. Goyal, M. K., & Ojha, C. S. P. (2010). Evaluation of neural network models for runoff simulation. *Water Resources Management, 24*(6), 1065–1080. https://doi.org/10.1007/s11269-009-9489-9

16. Jain, S. K., & Kumar, V. (2012). Trend analysis of rainfall and temperature data for India. *Current Science, 102*(1), 37–49.

17. Sen, P. N. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

18. Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: A support vector machine approach. *Journal of Hydrology, 330*(3–4), 621–640. https://doi.org/10.1016/j.jhydrol.2006.04.030

19. Yadav, S. K., & Singh, R. (2021). Machine learning techniques for climate and weather prediction: A review. *Earth Science Informatics, 14*(4), 1687–1706. https://doi.org/10.1007/s12145-021-00663-2

20. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*(C), 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0