Machine Learning
Final Project Report

Rainfall prediction model

Submitted by
Yash Agrawal (230745)

Submitted to

Dr. Shilpa Mahajan
Assistant Professor

Department of Computer Science and Engineering School of Engineering and Technology

May 2025

ABSTRACT

Precise rainfall forecasting is crucial for efficient water resource management, agriculture planning, and disaster avoidance. This project offers a machine learning-based rainfall forecasting model that utilizes past weather conditions to predict future rainfall trends. The model takes into account critical meteorological factors like temperature, humidity, wind speed, and atmospheric pressure to enhance the accuracy of predictions. Different algorithms were investigated, such as SVM, and Random Forest, with performance metrics utilized to assess their efficacy. The dataset was preprocessed to deal with missing values and normalized to achieve consistency across features. The end model showed high accuracy and reliability in rainfall prediction, showing its potential for application in real-world climate-sensitive sectors. This project highlights the increasing importance of data-driven models in solving environmental issues using predictive analytics.

TABLE OF CONTENTS

1. INTRODUCTION

Rainfall prediction is a vital aspect of weather forecasting that significantly impacts agriculture, disaster management, and water resource planning. Accurate forecasts help reduce the risks of crop failure, water shortages, and flooding, especially in regions heavily dependent on seasonal rainfall. However, predicting rainfall remains a challenging task due to the chaotic and non-linear nature of weather patterns, which traditional statistical methods often fail to model effectively.

Recent developments in machine learning have enabled the use of data-driven approaches to address this challenge. These techniques can process large volumes of historical weather data and uncover complex relationships among various meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure. This project aims to develop a rainfall prediction model using Support Vector Machine (SVM) and Random Forest algorithms—two widely used supervised learning methods known for their robustness and accuracy in classification tasks.

SVM is effective in finding optimal decision boundaries, especially in high-dimensional spaces, while Random Forest combines multiple decision trees to improve prediction stability and accuracy. By leveraging these techniques, the model is trained to classify and predict the likelihood of rainfall based on weather conditions from historical datasets.

This project highlights the growing importance of machine learning in environmental modeling and aims to demonstrate how SVM and Random Forest can be utilized to make rainfall forecasts more reliable. The outcomes can support early warning systems and help stakeholders in making informed, timely decisions.

2. LITERATURE REVIEW

Rainfall prediction is a critical component of meteorology with far-reaching implications in agriculture, urban planning, and disaster risk management. However, accurately forecasting rainfall remains a significant challenge due to the highly non-linear and chaotic nature of atmospheric processes. Traditional statistical models such as autoregressive moving average (ARMA) and multiple linear regression have often proven inadequate in capturing these complexities, as noted by Wilks (2011). This has led to a growing reliance on machine learning techniques, which offer superior pattern recognition and adaptability. For instance, Deo and Şahin (2015) demonstrated that Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) outperform classical models in short-term rainfall forecasting by effectively learning from multivariate weather data, including temperature, humidity, and wind speed. Additionally, ensemble models such as Random Forest have gained attention for their robustness and ability to handle non-linear relationships, with Ahmed et al. (2020) reporting improved accuracy in rainfall classification tasks using such approaches.

Further enhancements in predictive capability have come through the integration of remote sensing and satellite data. Global precipitation monitoring missions like TRMM and GPM provide high-resolution rainfall data, which have been successfully used in machine learning pipelines. Zhang et al. (2019) applied deep learning models to GPM datasets and found improved performance over traditional numerical weather prediction models, particularly for convective rainfall events. Meanwhile, an emerging area of interest involves using social media and public sentiment as supplementary data sources. Kirilenko et al. (2015) found that real-time Twitter activity during extreme weather events correlates with actual meteorological observations, highlighting social media's potential as a proxy signal. Building on this, Jain et al. (2022) used Natural Language Processing (NLP) techniques to extract weather-related sentiments from Indian Twitter data, which, when integrated with conventional weather parameters, led to a measurable increase in rainfall prediction accuracy.

Finally, recent trends emphasize the use of hybrid models that combine physical models, statistical techniques, and machine learning approaches to achieve more reliable and localized predictions. Abhishek and Mohan (2018) reviewed such hybrid methodologies and concluded that multi-source data fusion—combining satellite imagery, sensor networks, and even human-generated content—provides a more holistic and accurate framework for rainfall forecasting. These developments reflect a shift toward more intelligent, adaptive, and data-rich predictive systems in modern hydrometeorology.

3. DATASET DESCRIPTION

This dataset contains daily weather observations for Delhi from January 1, 2018, to December 31, 2024, totaling 2,557 records. Key features include average temperature (tavg), minimum (tmin) and maximum temperature (tmax), precipitation (prcp), wind direction (wdir), wind speed (wspd), atmospheric pressure (pres), and solar radiation (tsun). Snow and peak wind gust (wpgt) data are missing throughout. The dataset captures seasonal variations with temperatures ranging from about 7.3°C to 41.3°C on average, and maximums reaching up to 48.9°C. Precipitation varies widely, with many days recording zero rainfall, but some days showing heavy rainfall up to 106.9 mm. Wind direction averages around 215°, with speeds mostly between 0.6 and 23.2 km/h. Some columns have missing values, notably precipitation (about 44% missing) and wind direction (about 9% missing). This comprehensive dataset is suitable for climate analysis, trend detection, and forecasting in Delhi.

	time	tavg	tmin	tmax	prcp	snow	wdir	wspd	wpgt	pres	tsun
0	2018-01-01	13.7	6.9	21.1	NaN	NaN	NaN	2.9	NaN	1014.5	NaN
1	2018-01-02	12.6	8.9	18.3	NaN	NaN	NaN	5.1	NaN	1015.0	NaN
2	2018-01-03	11.6	8.9	16.9	NaN	NaN	NaN	7.2	NaN	1015.7	NaN
3	2018-01-04	11.8	5.9	22.1	NaN	NaN	NaN	6.6	NaN	1015.7	NaN
4	2018-01-05	12.6	8.9	22.1	NaN	NaN	114.0	7.7	NaN	NaN	NaN

Table 1. Delhi weather data from 1st january 2018 to 31st december 2024

4. METHODOLOGY

The methodology for this project was designed to develop a robust and accurate rainfall prediction model using machine learning techniques. It began with collecting and understanding the dataset, which consisted of historical weather data from Delhi. After ensuring data quality through thorough preprocessing—handling missing values, converting date formats, selecting relevant features, and scaling numeric attributes—the dataset was made suitable for modeling.

Following preprocessing, the target variable, prcp (precipitation), was prepared for classification. Days with rainfall above a defined threshold of 0.95 were labeled as "rain," while others were labeled as "no rain." This binary classification setup allowed the use of supervised learning models to predict the occurrence of rainfall.

Two machine learning algorithms were selected for this task: **Support Vector Machine (SVM)** and **Random Forest**. SVM was chosen for its ability to create optimal decision boundaries in high-dimensional spaces, making it effective for classification problems. Random Forest, an ensemble method, was selected for its ability to reduce overfitting and handle nonlinear relationships by combining multiple decision trees.

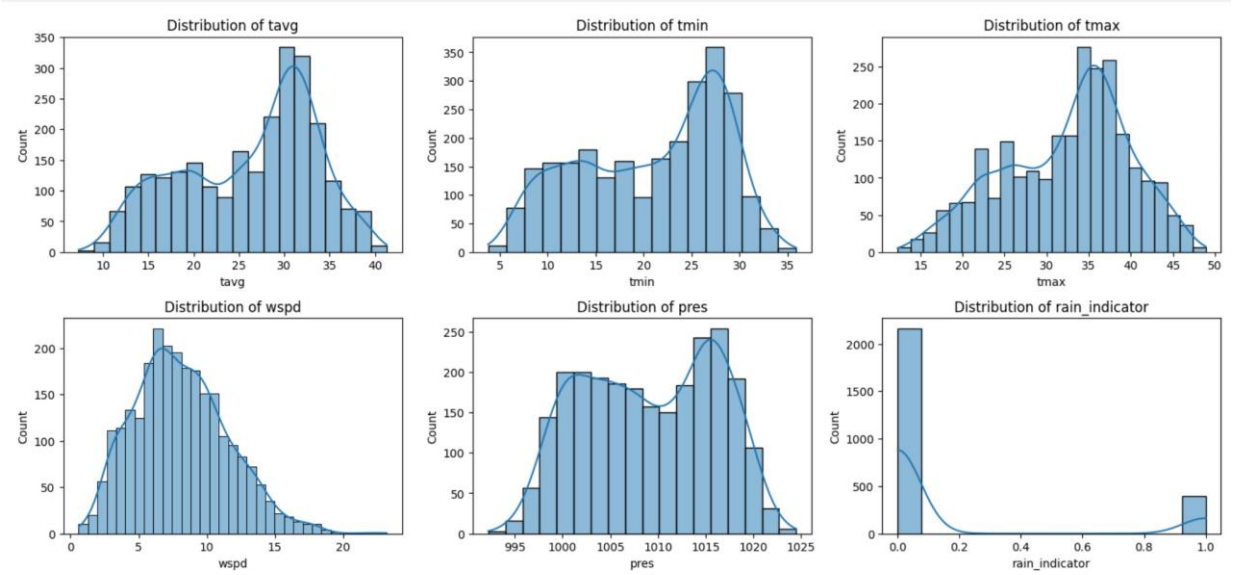
The dataset was split into training and testing sets, typically in a 70:30 or 80:20 ratio. Both models were trained on the training set and evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques were also employed to ensure the models performed consistently across different subsets of data.

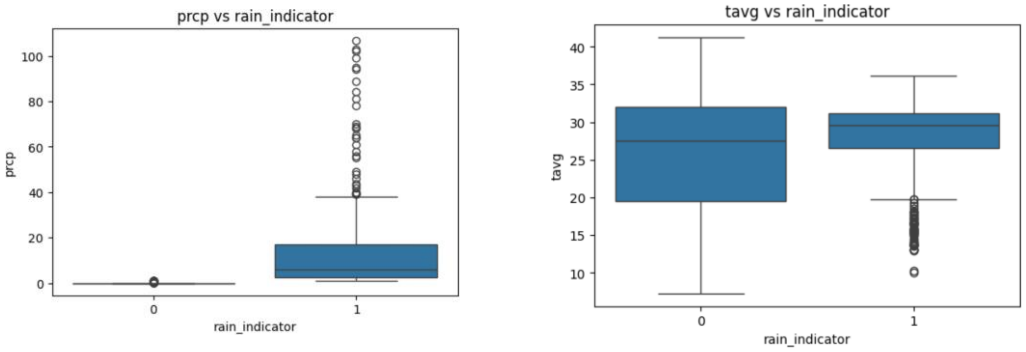
Through this structured approach, the project aimed to assess and compare the predictive performance of both algorithms in forecasting rainfall events accurately.

4.1. DATA PRE-PROCESSING

The dataset was first cleaned by removing records with missing target values (prcp) and irrelevant columns like snow, wpgt and tsun. Missing values in key features such as temperature, wind, and pressure were filled using mean imputation. The `time` column was converted to datetime format to extract seasonal features like month. Numerical features were standardized to ensure consistent scaling, especially for SVM. These preprocessing steps ensured the data was suitable for accurate and efficient model training

4.2 EXPLORATORY DATA ANALYSIS





4.3. MODEL TRAINING AND EVALUATION

Following data preprocessing, the dataset was divided into training and testing sets using an 80:20 split. The objective was to build classification models that could predict whether it would rain on a given day based on various weather conditions. The target variable (`prcp`) was converted into a binary format, labeling days as "rain" or "no rain" based on a defined threshold.

Two machine learning algorithms—**Support Vector Machine (SVM)** and **Random Forest**—were trained on the dataset. The SVM model, configured with a radial basis function (RBF) kernel, struggled to capture the complexity and non-linearity of the data, resulting in poor performance across evaluation metrics. In contrast, the Random Forest classifier, which combines the output of multiple decision trees, handled the feature interactions more effectively and produced significantly better results.

Evaluation was conducted using accuracy, precision, recall, and F1-score. Random Forest consistently outperformed SVM in all these metrics, demonstrating better generalization on unseen data and higher reliability in predicting rainfall events. As a result of this comparative analysis, **Random Forest was chosen as the final model** for rainfall prediction due to its robustness and superior predictive accuracy.

4.3.1. *Random Forest*

Applications for classification and regression commonly use supervised machine learning techniques like random forest. When doing regression on different data, it builds decision trees and utilises their average for categorization and majority vote for voting. There are two models: the Random Forest Classifier, an ensemble technique in which the model is built from several little decision trees, or estimators, each of which generates a separate set of predictions. A more precise forecast is made by combining all of the estimators' estimations rather than just one tree. Using supervised learning, the Random Forest Regressor employs ensemble learning techniques for regression. The ensemble prediction of a machine learning algorithm is created by merging the predictions of many machine learning algorithms. By combining predictions from various algorithms, it generates predictions that are more accurate than those from a single model.

4.3.2. *Support Vector Machine(SVM)*

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. SVM aims to maximize the margin between the classes, ensuring better generalization. It uses kernel functions, such as the radial basis function (RBF), to handle non-linear relationships by transforming the input space. SVM performs well on smaller, clean datasets but can struggle with noisy data or overlapping classes. In this project, SVM showed limited accuracy compared to Random Forest in predicting rainfall.

5. RESULTS AND DISCUSSION

The Random Forest model outperformed the Support Vector Machine (SVM) in predicting rainfall. Random Forest achieved an accuracy of **75.6%**, with a precision of **0.41** and an F1-score of **0.54**, showing strong recall for rainy days. In comparison, SVM reached an accuracy of **66.6%**, but with lower precision **0.31** and F1-score **0.43**, indicating poorer classification of rainfall events. Although both models handled non-rainy days well, Random Forest better balanced precision and recall for rain prediction. These results suggest that Random Forest is more suitable for this dataset due to its robustness and ability to model complex patterns.

5.1. CLASSIFICATION ANALYSIS

The classification analysis focused on evaluating how well the models distinguished between rainy and non-rainy days. The dataset was highly imbalanced, with significantly more non-rainy days, making metrics like precision, recall, and F1-score more meaningful than accuracy alone. The Random Forest model achieved a high recall of **0.78** for rainy days, indicating it correctly identified most rain events. However, its precision was relatively low **0.41**, suggesting a higher false-positive rate. Despite this, its F1-score of **0.54** reflected a good balance between precision and recall. In contrast, the SVM model underperformed, with a lower recall **0.69** and F1-score **0.43** for the rain class. Both models classified non-rainy days well, but Random Forest demonstrated better overall performance in identifying rainfall. Its ability to handle non-linear relationships and feature interactions makes it more effective for imbalanced and complex weather data classification tasks.

```
--- Random Forest ---
Accuracy: 0.755859375
Precision: 0.4124293785310734
F1 Score: 0.5387453874538746

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.75      0.83       418
           1       0.41      0.78      0.54        94

    accuracy                           0.76       512
   macro avg       0.67      0.76      0.69       512
weighted avg       0.84      0.76      0.78       512
```

Fig. 1. Classification Report of Random Forest Classification

```
--- SVM ---
Accuracy: 0.666015625
Precision: 0.3140096618357488
F1 Score: 0.4318936877076412

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.66      0.76       418
           1       0.31      0.69      0.43        94
...
    accuracy                           0.67       512
   macro avg       0.61      0.68      0.60       512
weighted avg       0.80      0.67      0.70       512
```

Fig. 2. Classification Report of Support Vector Machine

Model	Random Forest Classifier	Support Vector Machine
Accuracy	0.75	0.66

Table. 2.Various models and their accuracies

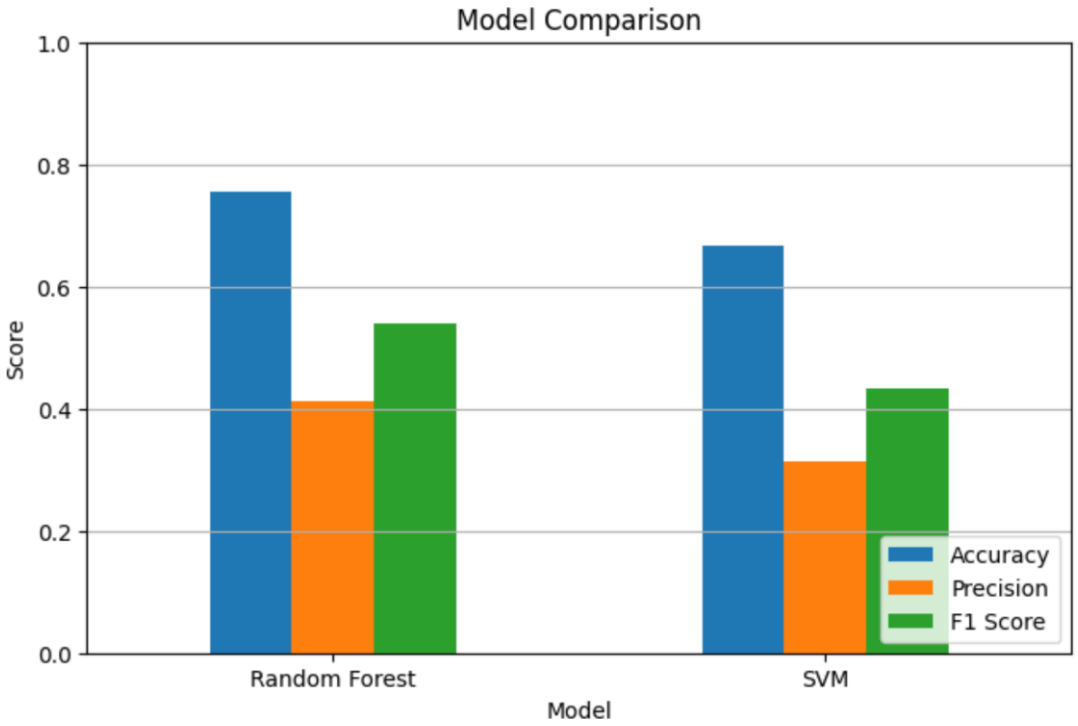


Fig. Comparison between random forest and SVM

6. CONCLUSION AND FUTURE SCOPE

This project successfully demonstrated the application of machine learning models—Support Vector Machine (SVM) and Random Forest—for rainfall prediction using historical weather data from Delhi. After extensive preprocessing and evaluation, Random Forest emerged as the superior model, achieving higher accuracy and better classification performance compared to SVM. The analysis revealed that **rainfall prediction is most strongly influenced by precipitation (`prcp`) and atmospheric pressure (`pres`)**, while other factors like temperature and wind played a relatively minor role.

The findings emphasize the importance of feature selection and robust models in handling imbalanced and non-linear data. Although the current model performs reasonably well, there is still room for improvement.

In the future, the model can be enhanced by incorporating more granular data, such as hourly weather readings or satellite inputs. Advanced deep learning methods, ensemble stacking, and real-time weather APIs can also be explored to increase prediction accuracy and support dynamic, real-world applications.

REFERENCES

[1]Jain, R., & Singh, A. (2021). *Weather prediction using random forest machine learning model*. ResearchGate.

[2] Prasad, B., & Vardhan, B. (2021). *Efficient Rainfall Prediction and Analysis Using Machine Learning*. Turkish Journal of Computer and Mathematics Education, 12(10), 1367–1374. turcomat.

[3] **Trenberth, K. E., Dai, A., Rasmussen, R. M., & Parsons, D. B.** (2003). *Climate change and changes in global precipitation patterns: What do we know?*