

# Rainfall Prediction Using Data Mining Techniques: A Performance Analysis

Sadia Dilawaiz<sup>1</sup>, Uswa Farooq<sup>1</sup>, Erssa Arif<sup>1\*</sup>, Asma Tariq<sup>2,3</sup>, Shahrukh Hamayoun<sup>2</sup>, Mudasir Zaheer<sup>3</sup>, Naila Nawaz<sup>2</sup>, and Muhammad Amjad<sup>1</sup>

<sup>1</sup>College of Computing, Riphah International University, Faisalabad Campus, Pakistan.

<sup>2</sup>Computing Department, NUML University, Faisalabad Campus, Pakistan.

<sup>3</sup>Computer Science Department, Agriculture University, Faisalabad, Pakistan.

\*Corresponding Author: Erssa Arif. Email: [dr.erssa@riphahfsd.edu.pk](mailto:dr.erssa@riphahfsd.edu.pk)

Received: September 15, 2025 Accepted: November 25, 2025

**Abstract:** Rainfall prediction is a crucial aspect of weather forecasting, as accurate and timely predictions enable the implementation of effective precautionary measures across various sectors, including transportation, agriculture, construction, flight operations, and flood management. By leveraging historical data, data mining techniques offer a promising approach to predicting rainfall based on key features. This research presents a critical analysis of data mining techniques employed for rainfall prediction, highlighting their strengths and limitations. The analysis reveals that while these techniques show promising results in classifying non-rain conditions, they perform poorly in accurately predicting rainfall events. The Support Vector Machine algorithm, for instance, achieved an F-measure of 0.958 for non-rain classification, but failed to predict rainfall with an F-measure of 0. Similarly, the Random Forest Method exhibited strong performance in classifying non-rain events, with an F-measure of 0.946, but only managed an F-measure of 0.357 for rainfall prediction. These results suggest that the data mining models struggled with rainfall classification due to various factors, including missing values, the absence of critical climate attributes, and the dataset's lack of quantitative rainfall measurements. Moreover, the dataset's focus on categorical weather conditions, rather than specific rainfall amounts, further limited the patterns available for classification. This work provides valuable insights into the limitations of current data mining approaches and establishes a foundation for future research aimed at improving rainfall prediction accuracy.

**Keywords:** Rainfall; Data Mining; Knowledge Discovery in Database; Decision Tree Model; Random Forest Model

## 1. Introduction

Weather data encompasses a range of atmospheric features, such as wind speed, humidity, pressure, and temperature. Data mining techniques have emerged as powerful tools for extracting hidden patterns from historical weather data, enabling the prediction of future weather conditions based on these identified patterns. Among weather-related phenomena, rainfall stands out as a complex process influenced by numerous atmospheric variables. Accurate and timely rainfall prediction is vital, as it supports crucial decisions in water resource management, early flood warnings, flight operations, and the management of transportation and construction activities. However, due to increasing climate variability, achieving accurate rainfall predictions has become even more challenging in recent years [1].

One of the most significant aspects of modern research in knowledge discovery is the analysis of time series data. Time series data, which is collected over specific intervals—whether hourly, daily, weekly, monthly, or

yearly—enables the prediction of future conditions in various domains, including climate change, education, and finance. Data mining techniques are used to uncover hidden insights within this time series data, offering valuable knowledge for forecasting. Weather forecasting, in particular, represents a highly beneficial but complex task, especially when it comes to predicting rainfall [2].

In agricultural countries like Pakistan, where rainfall plays a critical role in economic and societal well-being, the importance of accurate rainfall prediction cannot be overstated. Given the country's reliance on rain-fed agriculture, predicting rainfall has become one of the most scientifically and technologically demanding challenges. Factors such as temperature, pressure, wind speed, humidity, and mean sea level are typically considered when making rainfall predictions. However, the increasing unpredictability of climate patterns has further complicated the task.

Recent research indicates that integrating multiple data mining techniques enhances the accuracy of rainfall prediction models. This study aims to provide a critical and systematic review of the latest research developments in rainfall prediction using data mining methods. The focus is on papers published between 2013 and 2017, with an emphasis on advancements in predictive accuracy [3].

### 1.1. Objectives

- To critically evaluate and compare data mining techniques for rainfall prediction.
- To explore the potential of integrated techniques in improving rainfall prediction accuracy.
- To provide insights into the latest advancements in data mining approaches for rainfall prediction in the context of climate variability.

### 1.2. Scope

This research focuses on the application of data mining techniques for rainfall prediction, particularly in the context of agricultural countries like Pakistan, where timely and accurate forecasts are critical for planning and decision-making. The study analyzes time series data and evaluates various algorithms to determine their effectiveness in predicting rainfall.

### 1.3. Justification

The increasing unpredictability of climate patterns, combined with the significant role rainfall plays in agriculture and infrastructure planning, makes it imperative to develop more reliable prediction models. By reviewing the latest data mining techniques, this research aims to contribute to improving rainfall forecasting, with practical implications for agriculture, disaster management, and other sectors reliant on accurate weather predictions [4].

## 2. Literature Review

The analysis of time series data has become a fundamental aspect of modern research in knowledge discovery. Time series data, which is collected over specific intervals—such as hourly, daily, weekly, monthly, or annually—can be utilized by data mining techniques to forecast future events in various fields, including climate change, education, and finance. These techniques are crucial for extracting hidden patterns from time series data, which can then be leveraged for accurate predictions. Weather forecasting, although highly beneficial, remains a challenging task due to the complexity of atmospheric variables and the dynamic nature of climate.

In [5], a comparative analysis of Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Adaptive Neuro Fuzzy Inference System (ANFIS) for rainfall prediction was conducted. The models were evaluated based on four criteria: (i) the use of different lags as modeling inputs; (ii) training with data from heavy rainfall events only; (iii) performance of forecasting over 1 to 6 hours; and (iv) Performance under peak and all values. The results indicated that ANN performed better when trained on heavy rainfall datasets, with previous two-hour data inputs being optimal for forecasting 1 to 4 hours ahead in all three techniques. ANFIS demonstrated greater robustness in filtering noise with varying input lags, while SVM showed resilience under extreme typhoon events.

In [6], a survey of various Neural Network architectures for rainfall prediction over the past 25 years highlighted that the Propagation Network delivered significant results in many cases. Researchers in [8]

applied Artificial Neural Networks (ANN) for rainfall prediction in Thailand, using Back Propagation Neural Networks, which yielded acceptable accuracy. It was recommended that additional features, such as Sea Surface Temperature, be included in future models, particularly for regions like Andhra Pradesh and Southern India.

Recent advancements in rainfall prediction using data mining techniques, which differ from conventional weather prediction methods, have garnered increasing attention. Machine learning models can exploit historical observational data to predict future rainfall more conveniently compared to traditional approaches. For example, Jae-Hyun Seo et al. compared the performance of SVM, k-nearest neighbor (k-NN), and variant k-NN (k-VNN) models, achieving high accuracy in predicting rain/no-rain events in South Korea [1]. Similarly, Nikam and Meshram developed a rainfall prediction model using data from the Indian Meteorological Department, with commendable accuracy [6]. Despite these successes, classification accuracy remains a challenge due to continuous attributes and variations in discretization methods. Moreover, predictions have primarily been validated on datasets from specific locations, limiting their generalizability to larger regions.

It has been observed that integrated techniques improve accuracy in rainfall prediction. For instance, [8] proposed a methodology to predict maximum daily temperatures using Support Vector Regression. The model utilized features from various European measuring stations, including temperature, precipitation, relative humidity, and air pressure, alongside synoptic situations and monthly cycles. In [9], researchers implemented a rainfall forecasting model using Focused Time-Delay Neural Networks (FTDNN). The daily rainfall data, collected from the Malaysia Meteorological Department, was aggregated into monthly, quarterly, and yearly datasets for training and testing. The most accurate forecasts were obtained from the yearly dataset, with suggestions for incorporating additional parameters like temperature, humidity, and sunshine data to enhance accuracy.

In [2], a Decision Tree model was developed for rainfall prediction using data from Trivandrum station for 2013, with test data from 2014. The model tested 12 out of 20 parameters, demonstrating that Decision Trees can be effective in prediction and that accuracy could be further enhanced by integrating Artificial Neural Networks, Fuzzy Logic, and Genetic Algorithms. Similarly, [1] presented a rainfall prediction model using Naive Bayes, SVM, and Back Propagation Neural Networks, using seven input attributes and comparing overall-data-rate (RO) with rainfall-data-rate (RR).

Further, in [2], an Artificial Neural Network (ANN) integrated with data mining techniques such as classification and clustering was employed to predict rainfall, using data from the National Oceanic and Atmospheric Administration. The model yielded accurate predictions. In [10], an ANN model was developed to predict average monthly rainfall using data collected from Udupi, Karnataka over 50 years. The model employed a three-layered ANN with backpropagation, and the attributes were normalized using mean and standard deviation. This approach resulted in highly accurate predictions.

In [11], a modified Back Propagation ANN model was introduced for predicting Indian monsoon rainfall. The model, trained on 80% of the dataset and validated on 20%, achieved prediction accuracy between 80-90%, demonstrating the suitability of neural networks for rainfall prediction. In another study [5], a hybrid model combining Moving Average with ANN was applied to 136 years of data from the Indian Institute of Tropical Meteorology, using 100 years for training and 36 years for testing. The results showed improved accuracy with hybrid models.

Lastly, in [6] and [7], various data mining algorithms such as Naive Bayes, k-Nearest Neighbor, Decision Trees, Neural Networks, and Fuzzy Logic were compared for rainfall prediction, with Neural Networks consistently outperforming others in accuracy. A similar conclusion was reached in [28], where a Decision Tree-based weather prediction model was applied to data from multiple cities over three years, proving Decision Trees to be effective for multivariable analysis in weather forecasting.

\* Deep learning models (CNN-LSTM, ConvGRU).

\* Hybrid prediction frameworks.

\* Studies using high-resolution satellite and radar data.

\* Recent comparative analyses of machine-learning rainfall forecasting models.

### 3. Methodology

The methodology adopted in this study is designed to systematically collect, process, and analyze the data, ensuring a critical evaluation of the strengths and reliability of the research. This approach provides a structured way to address the research problem and uncover the truth behind the findings. This section explains how the data was collected, processed, and analyzed.

The chapter is divided into four subsections. The first subsection focuses on the data collection methods employed in this research. The second subsection outlines the data preprocessing techniques used to prepare the data for analysis. The third subsection details the properties of the data, providing insight into its characteristics. Finally, the fourth subsection discusses the analysis of the data, offering a comprehensive understanding of how the results were derived.

#### 3.1. Data Collection

Data collection is often one of the most challenging phases for any data miner, as identifying and obtaining a relevant dataset can be a daunting task. This section addresses the selection or creation of the target dataset for this research, collected from various sources. While numerous datasets are available from different locations, the challenge lies in determining how useful they are for the specific research objectives. Finding the appropriate dataset for this study was not an easy task, as it was not well-organized in any singular source.

The data collect from different source such as:

1. OpenWeatherMap([https://openweathermap.org/](https://openweathermap.org/))
2. AccuWeather ([https://www.accuweather.com/](https://www.accuweather.com/))

This research is based on customer feedback gathered from multiple weather forecasting sites, compiled into a single, consolidated file. Our primary goal was to determine the type, format, and source of the target data. While the data was available in various formats, it was essential to convert it into CSV format so that it could later be analyzed using data mining tools.

##### 3.1.1. Data Understanding and Selection

The primary data for this research was collected from various weather forecasting sites. The data plays a critical role, as it directly impacts the accuracy of future predictions, depending on identified time series patterns. The dataset selected focused on several important dimensions, including:

- **Date/Time:** Timestamp of the forecast
- **District:** Name of the specific city
- **Humidity:** Measured humidity levels
- **Pressure:** Measured atmospheric pressure
- **Temperature:** Measured temperature
- **Wind Speed:** Measured wind speed
- **Weather Description:** Description of the weather conditions

##### 3.1.2. Data Collection Sources

The data for this research was primarily obtained in Excel/CSV format. Initially, we explored various online meteorological sites and tools to gather the required data, but many of these sources did not provide useful or relevant information. Consequently, we manually collected data from several meteorological websites. Although finding domain-specific data was challenging, we were able to compile datasets that could be utilized for our analysis.

##### 3.1.3. Data Format

A significant challenge in this research was converting the data into a usable format. The requirement was to organize the data in a spreadsheet format to enable analysis. Initially, the data was available in formats such as notepad or word documents. However, after obtaining the data in Excel format, some irrelevant objects were removed during the preprocessing phase, allowing the data to be used effectively for analysis.

The dataset only qualitatively labels events as rain/no-rain rather than providing rainfall intensity values, the models had insufficient granularity to learn rainfall-specific patterns.

#### 3.2. Data Properties

The properties of the dataset are crucial to the accuracy and relevance of the research. Choosing the right data properties is essential to avoid misinterpretation of the results. In this study, we focused on weather data with attributes such as humidity, temperature, pressure, and wind speed to analyze and predict weather conditions such as rainfall, clouds, fog, and clear skies. The dataset included various dimensions, including district name, date/time, humidity, temperature, pressure, wind speed, and weather description.

Our primary goal was to identify weather conditions in specific regions and perform a comparative analysis of various data mining techniques used for rainfall prediction. Each of these factors was thoroughly explored in this study.

### 3.3. Data Analysis

Data analysis involves systematically applying statistical or logical techniques to describe and evaluate the data. It is essential to perform accurate and appropriate analysis to derive meaningful insights. After preprocessing, the data was analyzed using data mining techniques to understand how various factors influence weather conditions, particularly rainfall.

The analysis focused on enhancing the accuracy of rainfall prediction by utilizing various methodologies. Factors such as humidity, temperature, pressure, and wind speed were analyzed to determine their influence on rainfall. The study also emphasized the use of data mining techniques like Artificial Neural Networks (ANN), Random Forest (RFOREST), and SVM to evaluate data patterns and identify trends.

The main attributes used in the analysis were City, Date/Time, Humidity, Pressure, Temperature, Wind Speed, and Weather Description. The analysis explored relationships between these variables to predict weather conditions, such as mist, fog, snowfall, rainfall, and thunderstorms. By identifying independent and dependent variables, the study was able to predict dependent variables (e.g., Weather Description) based on the values of independent variables (e.g., Humidity, Temperature, Wind Speed). The inclusion of covariate variables, such as City, Date/Time, Humidity, Pressure, Temperature, and Wind Speed, further increased the accuracy of the predictions.

#### 3.3.1. Tools

For this research, we utilized R Studio as the primary tool for generating results. R is a widely used programming language and software environment for statistical computing and data analysis, supported by the R Foundation for Statistical Computing. The Random Forest Model (RFM), SVM, and various data visualization techniques were employed to generate and visualize the desired results.

## 4. Results and Discussions

### 4.1. Experimentations

#### 4.1.1. Data Collection

As discussed in the Methodology section, obtaining a domain-specific dataset is often a challenging task. The difficulty lies in both locating the dataset and determining its suitability for the research. In this research, weather data was gathered from multiple meteorological sources and compiled into a single, comprehensive file. The data was available in various formats, and the goal was to standardize it in CSV format so it could be efficiently used in data mining tools for analysis.

#### 4.1.2. Data Environment: Tools & Techniques

R Studio was utilized for generating the results in this study. R is a well-known programming language and software environment used for statistical computing and data analysis, supported by the R Foundation for Statistical Computing. We employed several advanced techniques, including the Random Forest (RFOREST) Model and Support Vector Machines (SVM) for classification, as well as data visualization techniques to interpret the results.

These techniques were applied to analyze weather conditions across various cities in Pakistan. By classifying attributes such as humidity, temperature, pressure, wind speed, and others, we used statistical methods to extract insights from the data and identify patterns that could help predict weather conditions.

#### 4.1.3. Data Preprocessing

To prepare the dataset for analysis, the data was read from a CSV file and processed as described in the Methodology section. Key variables included:

- **Independent variables:**

- City
- Date/Time
- Humidity
- Pressure
- Temperature
- Wind Speed

- **Dependent variable:**

- Weather Description (either its rainfall or some other weather condition)

The independent variables were used to predict the weather description, which served as the dependent variable in the analysis. This structured approach allowed for the successful application of machine learning techniques and statistical models.

- **Factors:**

```
is.factor(WEATHER$DateTime)
WEATHER$DateTime = as.numeric(WEATHER$DateTime)
is.numeric(WEATHER$DateTime)
is.factor(WEATHER$City)
WEATHER$City = as.numeric(WEATHER$City)
is.numeric(WEATHER$City)
```

- **Train Set:**

```
train <- WEATHER[1:15000, ]
train %>%
  select(WeatherDescription) %>%
  group_by(WeatherDescription) %>%
  summarise(count = n()) %>%
```

- **Test Set:**

```
test <- WEATHER[15001:1040001, ]
test %>%
  select(WeatherDescription) %>%
  group_by(WeatherDescription) %>%
  summarise(count = n()) %>%
```

- **For Statistics:**

```
rfModel <- randomForest(weatherDescription~City+DateTime+Humidity+Pressure+Tempera
test$weatherDescription <- predict(rfModel, test)
library(caret)
confusionMatrix(test$weatherDescription, test$weatherDescription)
```

- **Output:**

```
Overall Statistics

              Accuracy : 1
              95% CI : (1, 1)
    No Information Rate : 0.5109
    P-Value [Acc > NIR] : < 2.2e-16

              Kappa : 1

    Mcnemar's Test P-Value : NA
```

Furthermore WeatherDescription contains different classes like:

- Brokenclouds
- fewclouds
- fog
- haze
- heavyintensityrain
- heavysnow
- lightintensitydrizzle
- lightrain
- lightsnow
- mist
- moderaterain
- overcastclouds
- proximitythunderstorm
- proximitythunderstormwithrain
- scatteredclouds
- skyisclear
- snow
- squalls
- thunderstorm
- thunderstormwithheavyrain
- thunderstormwithlightrain
- thunderstormwithrain
- veryheavyrain

• **Statistics by Class:**

E:/RAINFALL PREDICTION/ ↗				
Statistics by class:				
	class: brokenclouds	class: fewclouds	class: fog	class: haze
Sensitivity	1.0000	1.0000	1.0000000	1.00e+00
Specificity	1.0000	1.0000	1.0000000	1.00e+00
Pos Pred Value	1.0000	1.0000	1.0000000	1.00e+00
Neg Pred Value	1.0000	1.0000	1.0000000	1.00e+00
Prevalence	0.1121	0.0247	0.0007571	3.22e-05
Detection Rate	0.1121	0.0247	0.0007571	3.22e-05
Detection Prevalence	0.1121	0.0247	0.0007571	3.22e-05
Balanced Accuracy	1.0000	1.0000	1.0000000	1.00e+00
	class: heavyintensityrain		class: heavysnow	
Sensitivity	1.00000		1.000e+00	
Specificity	1.00000		1.000e+00	
Pos Pred Value	1.00000		1.000e+00	
Neg Pred Value	1.00000		1.000e+00	
Prevalence	0.00277		1.854e-05	
Detection Rate	0.00277		1.854e-05	
Detection Prevalence	0.00277		1.854e-05	
Balanced Accuracy	1.00000		1.000e+00	

	Class: lightintensitydrizzle	Class: lightrain	Class: lightsnow
Sensitivity	NA	1.00000	1.000e+00
Specificity	1	1.00000	1.000e+00
Pos Pred Value	NA	1.00000	1.000e+00
Neg Pred Value	NA	1.00000	1.000e+00
Prevalence	0	0.06456	1.951e-06
Detection Rate	0	0.06456	1.951e-06
Detection Prevalence	0	0.06456	1.951e-06
Balanced Accuracy	NA	1.00000	1.000e+00
	Class: mist	Class: moderaterain	Class: overcastclouds
Sensitivity	1.00000	1.00000	1.0000
Specificity	1.00000	1.00000	1.0000
Pos Pred Value	1.00000	1.00000	1.0000
Neg Pred Value	1.00000	1.00000	1.0000
Prevalence	0.07995	0.01195	0.1504
Detection Rate	0.07995	0.01195	0.1504
Detection Prevalence	0.07995	0.01195	0.1504
Balanced Accuracy	1.00000	1.00000	1.0000
	Class: skyisclear	Class: snow	Class: squalls
Sensitivity	1.0000	1.000e+00	1.000e+00
Specificity	1.0000	1.000e+00	1.000e+00
Pos Pred Value	1.0000	1.000e+00	1.000e+00
Neg Pred Value	1.0000	1.000e+00	1.000e+00
Prevalence	0.5109	2.927e-06	4.878e-06
Detection Rate	0.5109	2.927e-06	4.878e-06
Detection Prevalence	0.5109	2.927e-06	4.878e-06
Balanced Accuracy	1.0000	1.000e+00	1.000e+00
	Class: thunderstorm	Class: thunderstormwithheavyrain	
Sensitivity	1.0000000	1.000e+00	
Specificity	1.0000000	1.000e+00	
Pos Pred Value	1.0000000	1.000e+00	
Neg Pred Value	1.0000000	1.000e+00	
Prevalence	0.0004917	1.171e-05	
Detection Rate	0.0004917	1.171e-05	
Detection Prevalence	0.0004917	1.171e-05	
Balanced Accuracy	1.0000000	1.000e+00	
	Class: thunderstormwithlightrain	Class: thunderstormwithrain	
Sensitivity	1.000e+00	1.000e+00	
Specificity	1.000e+00	1.000e+00	
Pos Pred Value	1.000e+00	1.000e+00	
Neg Pred Value	1.000e+00	1.000e+00	
Prevalence	7.122e-05	5.756e-05	
Detection Rate	7.122e-05	5.756e-05	
Detection Prevalence	7.122e-05	5.756e-05	
Balanced Accuracy	1.000e+00	1.000e+00	



```

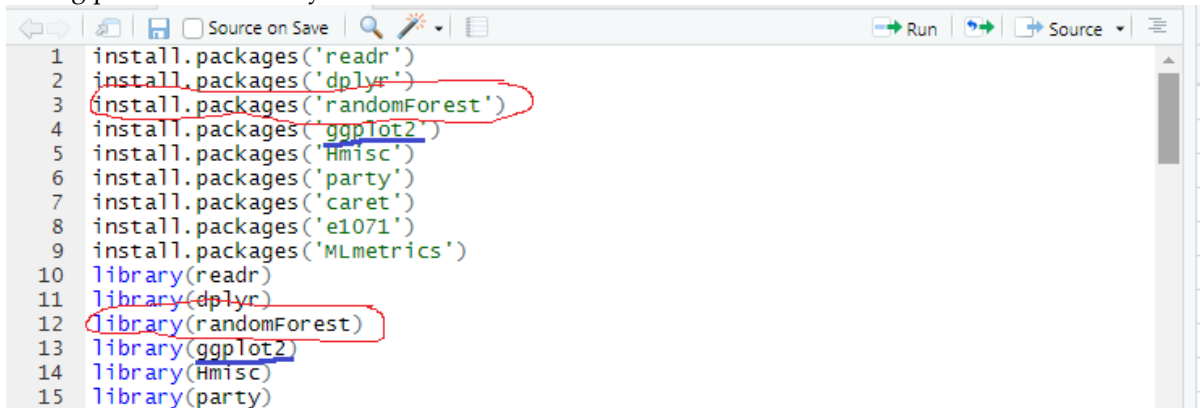
Sensitivity          class: veryheavyrain 1.000000
Specificity          1.000000
Pos Pred Value       1.000000
Neg Pred Value       1.000000
Prevalence           0.001112
Detection Rate       0.001112
Detection Prevalence 0.001112
Balanced Accuracy     1.000000
> |

```

## 4.2. Model Fitting

### 4.2.1. Random Forest

In RStudio, the Random Forest Model can be implemented using the appropriate package and library functions, which enable the model to efficiently handle both classification and regression tasks while improving predictive accuracy.



```

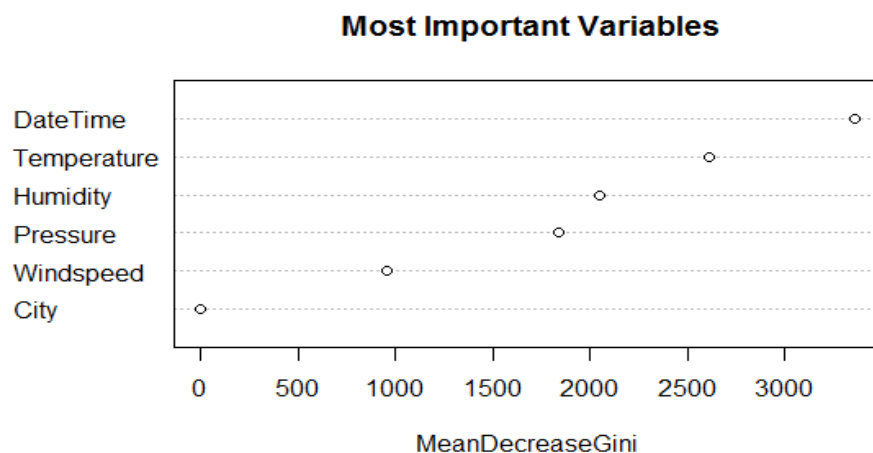
1 install.packages('readr')
2 install.packages('dplyr')
3 install.packages('randomForest')
4 install.packages('ggplot2')
5 install.packages('Hmisc')
6 install.packages('party')
7 install.packages('caret')
8 install.packages('e1071')
9 install.packages('MLmetrics')
10 library(readr)
11 library(dplyr)
12 library(randomForest)
13 library(ggplot2)
14 library(Hmisc)
15 library(party)

```

To apply RandomForest Model on all variables and to check out the dependency of WeatherDescription on them first we apply:

```
rfModel
```

```
<- randomForest(WeatherDescription~City+DateTime+Humidity+Pressure+Temperature+Windspeed,data=train)
```



**Figure 1.** Most Important Variables Graph Using RandomForest Model

And to checkout dependency of WeatherDescription for different combinations with dependent variables in light of above graph we performed:

- On DateTime

```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime, data = train)
```

The resultant value : 0.1889207

- On DateTime+Temperature

```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime+Temperature, data = train)
```

The resultant value : 0.2657955

- On DateTime+Temperature+Humidity

```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime+Temperature+Humidity, data = train)
```

The resultant value : 0.3563041

- On DateTime+Temperature+Humidity+Pressure

```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime+Temperature+Humidity+Pressure, data = train)
```

The resultant value : 0.3317387

- On DateTime+Temperature+Humidity+Pressure+Windspeed

```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime+Temperature+Humidity+Pressure+Windspeed, data = train)
```

The resultant value : 0.2851344

- On DateTime+Temperature+Humidity+Pressure+Windspeed+City

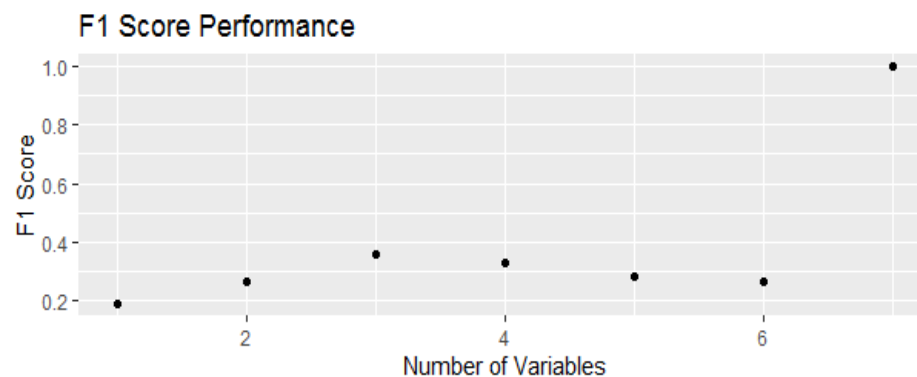
```
rfModelTrim1 <- randomForest(WeatherDescription ~ DateTime+Temperature+Humidity+Pressure+Windspeed+City, data = train)
```

The resultant value : 0.2640421

Now, for all :

```
F1_all = randomForest(WeatherDescription ~ DateTime+Temperature+Pressure+Humidity+Windspeed+City, ntree = 500, data = train)
```

Values	
F1_1	0.188920729370106
F1_2	0.265795450436978
F1_3	0.356304099357178
F1_4	0.331738677125403
F1_5	0.285134426720587
F1_6	0.26404221740742
F1_Score	num [1:7] 0.189 0.266 0.356 0.332 0.285...

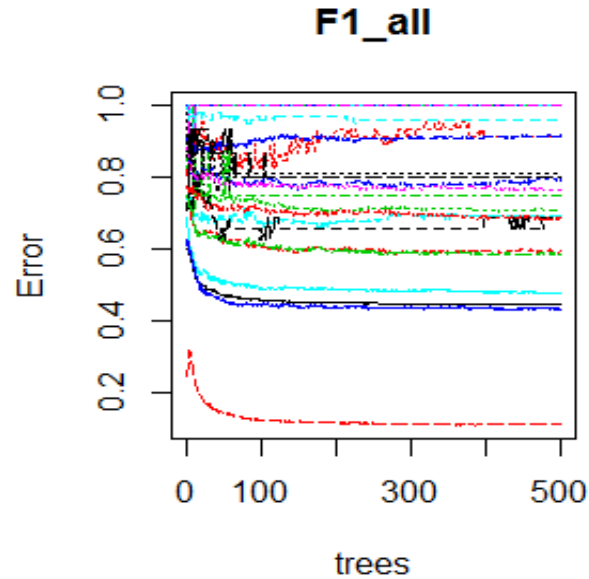


**Figure 2.** F1 Score Performance Graph Using Random Forest Model

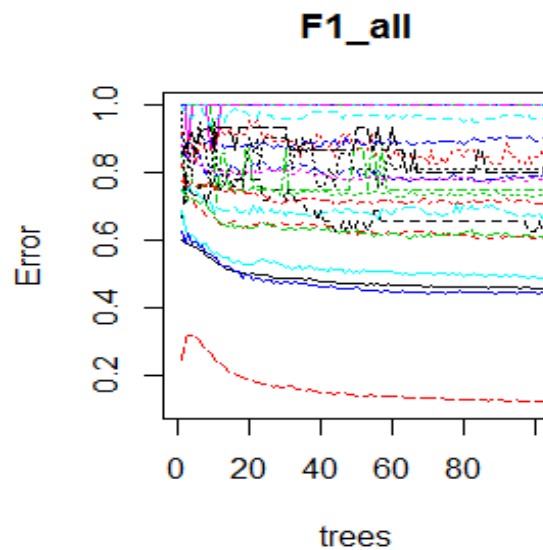
This plot shows the class error rates of the random forest model. As the number of trees increases, the error rate approaches zero.

Also; options(repr.plot.width=6, repr.plot.height=4)

plot(F1\_all, xlim=c(0,100))



**Figure 3.** F1\_all(i) Graph Using Random Forest Model

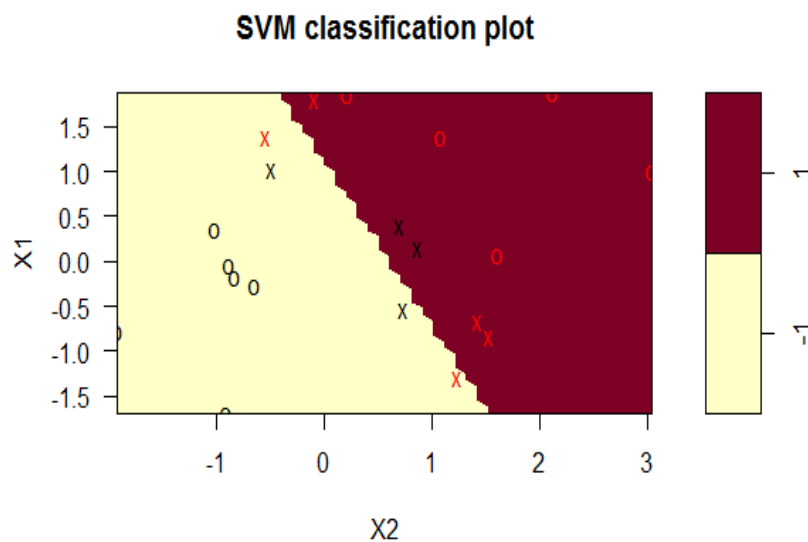


**Figure 4.** F1\_all(ii) Graph Using Random Forest Model

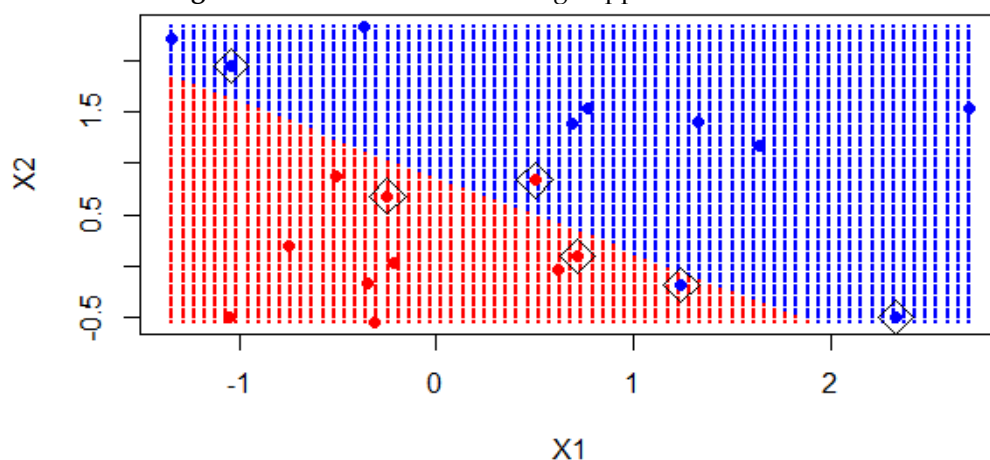
#### 4.2.2. Support Vector Machine (SVM)

SVM is a powerful classification technique that uses hyperplanes to separate data into distinct classes. The concept behind SVM is intuitive and easy to grasp. Given labeled data, SVM generates multiple hyperplanes to divide the data space into segments, where each segment contains data points belonging to a single class.

For rainfall analysis in Pakistan, SVM is employed to separate the data into two categories: rainfall and no rainfall. Initially, we generate data in two dimensions, representing variables such as Humidity and Temperature. After setting a random seed, a matrix  $x$  is created, which contains 20 observations distributed between the two classes (rainfall and no rainfall). The  $y$  variable is used to assign class labels, with values of either -1 or 1 for each class. For the rainfall class ( $y = 1$ ), the mean of each coordinate is shifted from 0 to 1. The data can then be plotted, with points color-coded to represent their corresponding class. By using the plotting character 19, large, clearly visible dots are generated—colored blue for *no rainfall* and red for *rainfall*—enabling easy visualization of the classification results.



**Figure 5.** Classification Plot Using Support Vector Machine



**Figure 6.** Output Graph Using Random Forest Model

#### 4.3. Results and Discussion

The results provided for Random Forest Model are:

SVM provides almost same results as shown in table below:

**Table 1.** RFM Results

Proportion	Class	Precision	Recall	F-Measure
10:30	No rain	0.935	.95	.942
	Rain	.358	.295	.324
20:80	No Rain	.935	.946	.941
	Rain	.35	.306	.326
30:70	No Rain	.936	.948	.942
	Rain	.365	.317	.339
40:60	No Rain	.937	.948	.943
	Rain	.368	.317	.343
50:50	No Rain	.936	.948	.942
	Rain	.363	.321	.337
60:40	No Rain	.937	.948	.942
	Rain	.364	.317	.339

70:30	No Rain	.936	.948	.942
	Rain	.361	.313	.335
80:20	No Rain	.942	.943	.942
	Rain	.35	.346	.348
80:10	No Rain	.942	.949	.946
	Rain	.373	.343	.357

**Table 2.** SVM Results

Proportion	Class	Precision	Recall	F-Measure
10:30	No rain	0.914	1	.955
	Rain	0	0	0
20:80	No Rain	.914	1	.955
	Rain	0	0	0
30:70	No Rain	0.914	1	.955
	Rain	0	0	0
40:60	No Rain	.914	1	.955
	Rain	0	0	0
50:50	No Rain	0.914	1	.955
	Rain	0	0	0
60:40	No Rain	.914	1	.955
	Rain	0	0	0
70:30	No Rain	0.914	1	.955
	Rain	0	0	0
80:20	No Rain	.919	1	.958
	Rain	0	0	0
80:10	No Rain	0.919	1	.958
	Rain	0	0	0

The data mining techniques employed in this research did not yield accurate results. The F-Measure serves as a robust metric for accuracy, as it provides the average of Precision and Recall. Several factors could contribute to the low accuracy observed, including missing values, as mean values may not accurately represent the dataset, the absence of one or more critical climate attributes, or generally low rainfall levels.

Moreover, due to climatic variations, the rainfall rate in many locations has significantly decreased compared to historical data. Additionally, the dataset does not quantify or measure rainfall; it merely indicates whether it is raining or describes other weather conditions. This limitation resulted in fewer patterns available for the classification algorithms, ultimately leading to weaker performance in predicting rainfall classes.

Nonetheless, the data mining techniques exhibiting a high F-Measure are as follows:

**Table 3.** DM Technique with high F-Measure

DM Algorithm	Class	Proportionality	F-Measure
SVM	Other	80:20	0.958
	Rain	90:10	0
RFM	Other	90:10	0.946
	Rain	90:10	0.357

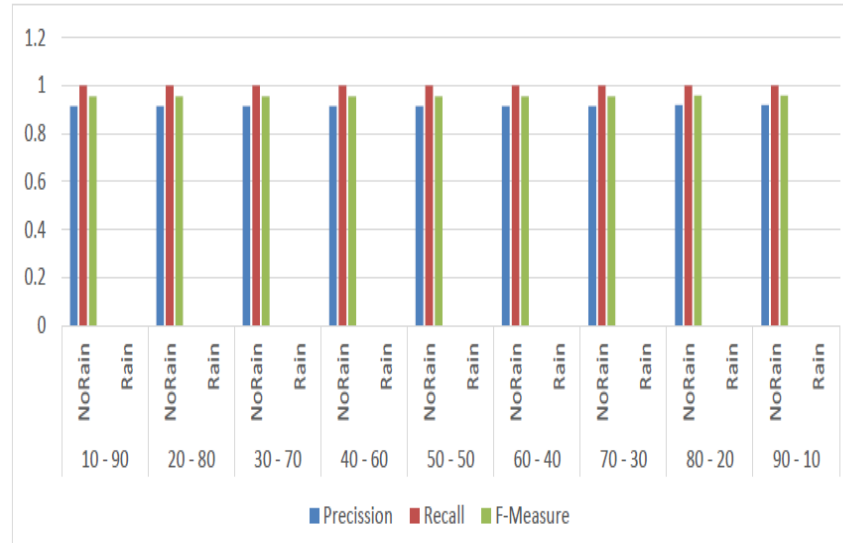


Figure 7. SVM Results

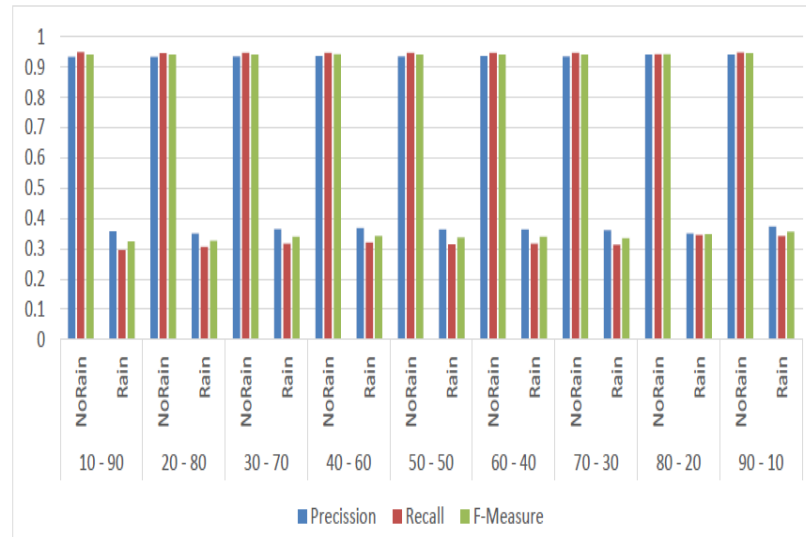


Figure 8. RFM Results

- Using SMOTE or oversampling for minority classes.
- Incorporating temporal features (lag variables, sequence models).
- Adding new climate attributes (cloud cover %, dew point, precipitation intensity).

## 5. Future Work

This research focused on rainfall prediction in Pakistan utilizing Random Forest Machine (RFM) and Support Vector Machine (SVM) algorithms. The analysis was based on data from the past five years. For effective predictions, a classification framework was implemented, where the input data underwent a preprocessing stage to ensure cleanliness before the classification process.

To examine the performance dependency of the classification techniques on training data, various ratios of training and test data were employed, ranging from 10:90 to 90:10.

The results indicated that while the classification techniques performed well for classes indicating no rain, their effectiveness diminished for classes indicating rainfall. This discrepancy may be attributed to missing values or the absence of crucial climate attributes. Therefore, it is recommended that future work explore additional classification techniques and consider a wider range of climate attributes for improved weather data analysis.

**References**

1. DATAMINING Introductory and Advanced Topics Part1 Margaret H. Dunham Department of Computer Science and Engineering Southern Methodist University). Fundamentals of Database System Ramez Elmasri, 1989.
2. (Wu, S. (2013), "A review on coarse warranty data and analysis", Reliability Engineering and System, 114: 1–11, doi:10.1016/j.res.2012.12.021). Introduction to Datamining Srinivasan Parthasarathy.
3. S. Zhang, L. Lu, J. Yu, and H. Zhou, "Short-term water level prediction using different artificial intelligent models," in 2016 5th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2016, 2016.
4. D. Nayak, A. Mahapatra, and P. Mishra, "A Survey on Rainfall Prediction using Artificial Neural Network," Int. J. Comput. ..., vol. 72, no. 16, pp. 32–40, 2013.
5. M. Hussain, W. Sharif, M. R. Faheem, Y. Alsarhan, and H. A. Elsalamony, "Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model", Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29779–29786, Dec. 2025.
6. B. K. Rani and A. Govardhan, "Rainfall Prediction Using Data Mining Techniques - A Survey," pp. 23–30, 2013.
7. Z. Awais et al., "ISCC: Intelligent Semantic Caching and Control for NDN-Enabled Industrial IoT Networks," in IEEE Access, vol. 13, pp. 169881–169898, 2025, doi: 10.1109/ACCESS.2025.3614984.
8. N. Solanki and G. P. B., "A Novel Machine Learning Based Approach for Rainfall Prediction," Inf. Commun. Technol. Intell. Syst. (ICTIS 2017) - Vol. 1, vol. 83, no. Ictis 2017, 2018.
9. C. S. Thirumalai, "Heuristic Prediction of Rainfall Using Machine Learning Techniques," no. May, 2017.
10. N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data," Int. J. Intell. Syst. Appl., vol. 10, no. 1, pp. 16–23, 2018.
11. Zubair, M.; Hussain, M.; Albashrawi, M.A.; Bendeche, M.; Owais, M. A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Computer Methods and Programs in Biomedicine. 2025, 272, 109014. <https://doi.org/10.1016/j.cmpb.2025.109014>.
12. Zubair, M., Owais, M., Hassan, T. et al. An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. Sci Rep 15, 13087 (2025). <https://doi.org/10.1038/s41598-025-97256-0>
13. Salahuddin, H., Imdad, K., Chaudhry, M. U., Nazarenko, D., Bolshev, V., & Yasir, M. (2022). Induction Machine-Based EV Vector Control Model Using Mamdani Fuzzy Logic Controller. *Applied Sciences*, 12(9), 4647. <https://doi.org/10.3390/app12094647>
14. M. Hussain, W. Sharif, M. R. Faheem, Y. Alsarhan, and H. A. Elsalamony, "Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model", Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29779–29786, Dec. 2025.
15. Salahuddin, H., Imdad, K., Chaudhry, M. U., Iqbal, M. M., Bolshev, V., Hussain, A., Flah, A., Panchenko, V., & Jasiński, M. (2022). Electric Vehicle Transient Speed Control Based on Vector Control FM-PI Speed Controller for Induction Motor. *Applied Sciences*, 12(17), 8694. <https://doi.org/10.3390/app12178694>