ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 15, Issue, 05, pp. 68437-68441, May, 2025
https://doi.org/10.37118/ijdr.29564.05.2025

RESEARCH ARTICLE OPEN ACCESS

SHORT-TERM WEATHER PREDICTION USING HISTORICAL DATA: EVALUATING SUPERVISED MACHINE LEARNING TECHNIQUES

Sneha Kandacharam* and Akanksha Kaushik and Tushar

Department of Computer Science and Engineering, The NorthCap University, Gurugram, India

ARTICLE INFO

Article History:
Received 15th February, 2025
Received in revised form
17th March, 2025
Accepted 26th April, 2025
Published online 30th May, 2025

Key Words:
Weather prediction, supervised learning, machine learning classifiers, random forest, model evaluation metrics, machine learning algorithms.

****Corresponding author:***
Sneha Kandacharam,

ABSTRACT

Weather Forecasting or Predictions plays an important role in human life as it impacts agriculture, transportation and natural disasters. Therefore, the development of new techniques in machine learning is continuing to enable better and accurate predictions of weather using historical data such as temperature, dew point temperature, relative humidity, visibility, pressure, and wind speed over the years. The primary goal of my system is to develop a model that will be precise and accurate for short-term predictions. This proposed system will first do data preprocessing which involves dealing with missing or null values. Then the architecture applies several machine learning algorithms such as logistic regression, Decision tree classifier, Random forest classifier, SVC, K neighbor classifiers and Gaussian Naive Bayes. Now each model is evaluated based on the prediction accuracy by MAE and RMSE. After comparison, the proposed system confirms that the random forest model indeed outperformed all other models.

Citation: Sneha Kandacharam and Akanksha Kaushik and Tushar, 2025. *"*Short-term Weather prediction using historical data: evaluating supervised machine learning techniques*". International Journal of Development Research*, 15, (05), 68437-68441.

INTRODUCTION

For centuries, modern humans have relied heavily on weather forecasting due to its profound impact on agriculture, transportation, and disaster management. Traditional forecasting approaches relied on empirical observations and numerical models aimed at simulating atmospheric processes—these methods used to consume an overbearing amount of computational resources. Machine learning has emerged as an alternate approach that aims to improve the speed and accuracy of overall weather prediction. A formidable asset of machine learning is its ability to analyze voluminous amount of meteorological data which, through traditional methods, would not be able to identify any patterns. Because of the historical data available, machine learning algorithms are able to vastly enhance their predictive capabilities, which makes them more suitable for weather forecasting. Our proposed system explores the multitude of applications in which varying machine learning techniques can be implemented to predict weather conditions while focusing on complex datasets. Today many repositories of datasets are available from satellites, weather stations and other resources. The repositories include an array of dataset such as temperature, humidity, precipitation, wind speed, and snowfall as well as the atmospheric pressure over time. Trends and correlations useful for short term weather prediction can be formulated by machine learning algorithms using these datasets.

The proposed system has two levels, first is to build a machine learning model capable of accurate weather forecasting and the second level is to compare the performance of these models with traditional weather prediction models. For achieving this we will pre-process the weather datasets and remove the null values to ensure all inputs fed into our models are clean. The machine learning models that will be used in this study include linear regression, decision tree, random forests, support vector machines and many more. Each algorithm in the proposed system has its own unique advantages and disadvantages, and their accuracy in weather forecasting will be measured by inter-model competitions with MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) serving as metrics. the proposed system is aimed to demonstrate that machine learning can serve as a powerful tool for improving weather prediction.

The proposed system follows 2 -Level approach:

- Develop and train multiple Machine Learning Models for Classification
- Evaluate the performance of the models using metrics and compare the result to identify the most effective model.

Related Work

The use of machine learning techniques for weather forecasting has been studied quite extensively with numerous approaches proposed to

further improve the accuracy of predictions. As noted in the literature, Patkar (2022) showed how important data cleaning and feature extraction is to weather forecasting, using an example of Bayes and logistic Regression And their Ensemble Models to show how multiple machine learning approaches improve the outcomes of the model. In the pursuit of improving prediction accuracy, Jakaria et al. (2018) analyzed weather data from different regions using Random Forest Regression (RFR). They also explored the impact of preprocessing on the model, showing that one-hot encoding and mean scaling helped the model perform better. Various methods for weather prediction using linear and functional regression were also studied by Holmstrom et al. (2016). Although initial attempts at using professional forecasting services outperformed models, with the aid of machine learning, the accuracy of long-term forecasting improved, thus validating the logic of incorporating historical data. Sudarshan et al. (2020) used Artificial Neural Networks (ANN) for weather forecasting, obtaining better than 94% accuracy on temperature and rainfall prediction. They stressed the importance of feature selection and model tuning with regards to other domains such as agriculture. Random Forest was also implemented by Singh et al (2019) usingby applying a Random Forest algorithm classification on 20 years of weather data, the rainfall prediction accuracy achieved was 87.9%. As pointed out in their study, one of the primary focus areas of the research was humidity. They also suggested implementing sensor-based systems which are fairly inexpensive, thus making machine learning more practical for real-world forecasts on weather. There are other studies that further explored the use of big data combined with ensemble learning for weather forecasting. Madan et al. (2018) analyzed large meteorological data sets through a combination of SVM, linear regression, and decision tree methods, proving machine learning techniques provided more accurate results than traditional forecasting models. Recently, other researchers tackled the concept of probabilistic forecasting. For instance, Price et al. (2024) developed Gencast, a global forecasting model focused on machine learning that was reported to outperform predictive numerical weather prediction (NWP) for forecasting severe weather events. To build on prior works, our research uses several supervised learning algorithms for weather classification and forecasting such as Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes. The performance of these models is evaluated in precision, recall, F1-score, and ROC-AUC. Our findings revealed that for short-term weather condition predictions, the Random Forest model surpassed all other classifiers in predictive accuracy.

MATERIALS AND DISCUSSION

Dataset: The dataset used in this study is sourced from Kaggle, a publicly available repository for machine learning datasets. It contains 8784 rows and 8 columns, representing historical weather conditions such as temperature, humidity, pressure, wind speed, and visibility. This dataset is crucial for training machine learning models in weather prediction.

Attributes

The dataset includes the following key attributes:

1. **Temperature (Temp_C)** – Measures the temperature in degrees Celsius.
2. **Dew Point Temperature (Dew Point Temp_C)** – Indicates the temperature at which air becomes saturated with moisture.
3. **Relative Humidity (Rel Hum_%)** – Represents the percentage of atmospheric moisture.
4. **Wind Speed (Wind Speed_km/h)** – Measures wind speed in kilometers per hour.
5. **Visibility (Visibility_km)** – Defines the distance at which objects can be clearly seen.
6. **Atmospheric Pressure (Press_kPa)** – Measures atmospheric pressure in kilopascals.
7. **Weather Condition (Weather)** – Categorical variable describing weather conditions.
8. **Standardized Weather Condition (Std_Weather)** – Derived feature to categorize weather patterns into simplified labels such as Rain, Fog, Snow, etc.

Data Preprocessing: Data preprocessing is a crucial step to ensure data quality and improve model accuracy. The following preprocessing techniques were applied:

- **Handling Missing Values**: Removing or imputing missing data to maintain data consistency.
- **Feature Scaling**: Standardizing numerical features using **StandardScaler** to ensure uniformity.
- **Encoding Categorical Variables**: Converting categorical data, such as weather conditions, into numerical labels using **Label Encoding**.
- **Feature Engineering**: Creating a new feature, **Std_Weather**, to group similar weather patterns for better classification.

Model Training: The dataset was split into 80% training and 20% testing. Multiple machine learning models were implemented for weather classification and prediction:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier (SVC)
- K-Nearest Neighbors (KNN)
- Gaussian Naïve Bayes

Each model was trained and evaluated using the standardized dataset, with hyperparameter tuning was performed using Research applied to optimize model performance.

Why random Forest was Chosen: Random forest was chosen for hyperparameter tuning because random forest provides

- High accuracy with the noisy datasets
- Robustness to overfitting
- Ability of the model for the complex interactions
- Feature importance capability

Evaluation: The models were assessed based on multiple performance metrics:

- **Accuracy Score**: Measures the overall correctness of predictions.
- **Precision**: Evaluates the proportion of true positive predictions.
- **Recall**: Measures the ability of the model to detect relevant instances.
- **F1-Score**: Harmonic mean of precision and recall, balancing both metrics.
- **ROC-AUC Score**: Assesses the ability of models to distinguish between different weather categories.
- **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)**: Evaluates the error between predicted and actual values, particularly for regression-based forecasting.
- **Confusion Matrix:** Provides a visual representation of model performance by showingcorrect and incorrect classifications.
- **Heatmap Analysis:** Used to visualize correlations between different weather attributes and identify key influencing factors.

Through this structured approach, the study aims to enhance short-term weather prediction accuracy by integrating machine learning techniques with historical meteorological data, is shown in Figure 1.

Methodology and Model Specification: The model takes historical data from the Kaggle with weather data as input, which includes the parameter like temperature, Dew point temperature, humidity,

pressure, and wind speed. These are the features which are preprocessed and scaled to ensure for the uniform and efficient model training and testing then the proposed system do data preprocessingthen do feature engineering, sampling and balancing, feature selection, label encoding. The proposed systemthen does EDA (Exploratory Data Analysis) which includes data visualization (histogram and boxplots, heatmap. The proposed system then does the model training and proposed system trained a variety of models like Decision tree classifier, Random Forest Classifier, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Logistic regression, Gaussian Naïve Bayes. The proposed system thenevaluates our models by using Precision, Recall, F1-Score and ROC-AUC. The proposed system then does cross validation. Then we perform hyperparameter tuning for the optimized result.

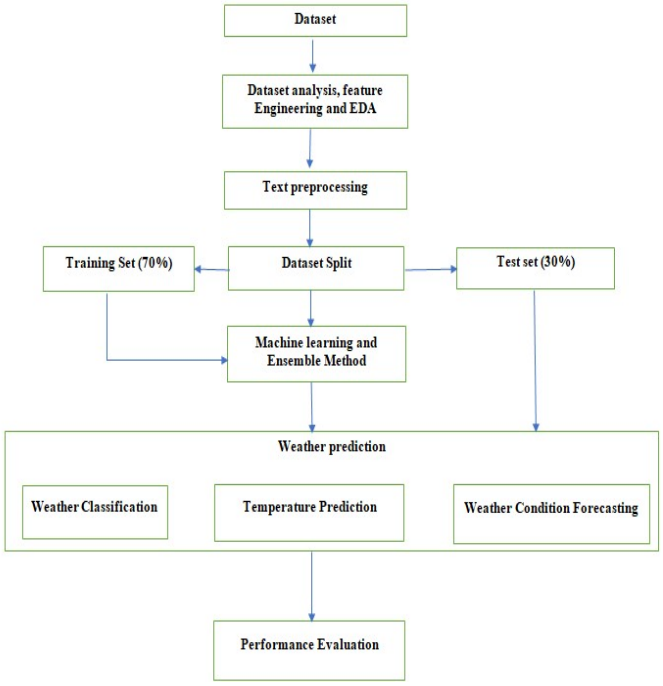

Figure 1. Shows the overall performance of the proposed system

Models and Algorithms

We trained six machine learning models

- **Decision tree Classifier:** The decision tree is a supervised machine learning algorithm that is used for the classification tasks. The decision tree classifier works by splitting the data into the subsets based on the features values which forms a tree like structure with the decision nodes and leaf nodes. The decision nodes represent tests on an attribute, and the leaf nodes represent the final output class.

$$A^* = arg \overset{max}{A} [Criterion(S, A)]$$

Where:
- A* is the feature selected for the split.
- Criterion (S, A) is the measure of the split.
  - Information gain:

$$Criterion(S, A) = H(S) - \sum_{v \epsilon Values(A)} \frac{|Sv|}{|S|} H(Sv)$$

  - Negative Gini Impurity (to Maximize purity):

$$Criterion(S, A) = - \sum_{v \epsilon Values(A)} \frac{|Sv|}{|S|} G(Sv)$$

Random Forest Classifier: The random forest classifier is an ensemble machine learning algorithm that built multiple decision tree and combines their predictions for the improved and accurate and robustness of the result. Then each tree us trained on a random subset of the data and features, and then the final output is determined by aggregating the outputs of all the trees.

$$\hat{y} = Aggregate(T1(x), T2(x), \ldots., Tk(x))$$

Where:
- Ti(x) is the prediction of the i-th decision tree in the forest for input x.
- K is the total number of trees in the forest.
- Aggregate is typically:
  - Majority Voting for Classification:

$$\hat{y} = arg \max_{y} \sum_{i=1}^{k} 1(Ti(x) = y)$$

Where 1 is an indicator function that equals 1 if Ti(x) = y and 0 otherwise.

  - Average for Regression tasks:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} Ti(x)$$

- Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for classification and regression tasks. It works by finding a hyperplane that will best separate the data points of different class in high dimensional space.

The Decision boundary is defined as: $f(x) = w.x + b$

Where:

- W is the weight vector.
- X is the input feature vector.
- b is the bias term.

K-Nearest Neighbors (KNN): The -Nearest Neighbors (KNN) is a non-parametric instance-based machine learning algorithm method which is used for classification and regression. It will predict the class of a data point by considering the k coolest training points in the feature space.

$$\hat{y} = arg \max_{c} \sum_{i \epsilon Nk(x)}^{k} 1(yi = c)$$

Where:

- Nk(x) is the set of the k-nearest neighbors of x.
- yi is the class label of the i-th neighbour.
- C is a candidate class.
- 1(.) is the indicator function which is equal to 1 if yi = c and 0 otherwise.

Logistic Regression: The logistic regression is a supervised learning algorithm used for binary classification. It models the probability that a given input belongs to a particular class using a logistic (sigmoid) function.

Formula for logistic regression:

$$\hat{P}(y = 1|x) = \sigma(w.x + b)$$

Where:
- σ(z)=1/1+e^−z is the sigmoid function.
- z = w.x + b
- w is the weight vector.
- x is the input feature vector.
- b is the bias term.

Naïve Bayes: The Naïve Bayes is a probabilistic machine learning algorithm which is based on the Bayes Theorem. It assumes the independence between features and this algorithm is commonly used for the classification tasks.

Formula for Naïve Bayes:

$$\hat{y} = \arg\max_{c \epsilon C} P(c) \prod_{i=1}^{n} P(xi|c)$$

Where:
- C is the set of all possible classes.
- P(c) is the prior probability of class c.
- P(xi|c) is the likelihood of features xi given class c.
- n is the number of the features in the input x = {x1, x2, ..., xn}.

RESULTS AND EVALUATION

The proposed system divided the dataset into training and testing sets in the ratio 80:20. Then each model was trained using the training set and then evaluated using metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Then the Machine learning algorithm Random Forest underwent through the hyperparameter tuning using GridSearchCV to optimize the performance of the Random Forest.

Visualization: The proposed system then visualized our results using visualization tools like:

- Heatmap for feature correlations.

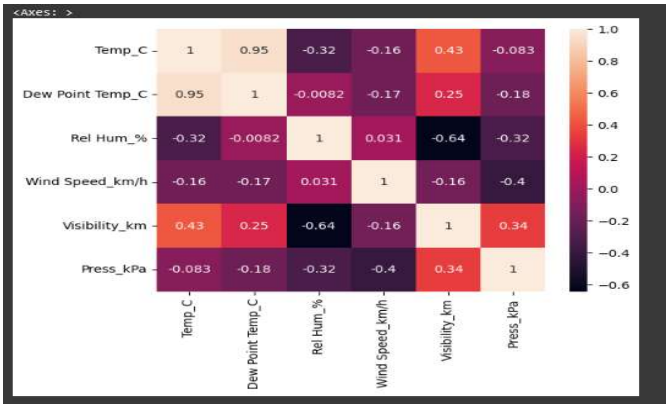

Figure 2. Showing the Heatmap for the feature's correlations

Heatmap represents the correlation coefficients between different meteorological features like:

- Temperature (Temp_C)
- Dew Point Temperature
- Relative Humidity
- Wind Speed
- Visibility
- Pressure

Key Insights:

- Temperature and Dew Point Temperature has the strongest positive correlation (0.95), indicating they move together.
- Visibility and Relative Humidity shows a strongest negative correlation (-0.64), which shows that higher humidity means lower Visibility.
- Temperature and Visibility also shows moderate positive correlation (0.43).
- Overall, most variables shows weak to moderate correlations, which will be important for feature selection in Machine Learning.

- Confusion Metrices to evaluate the predictions.

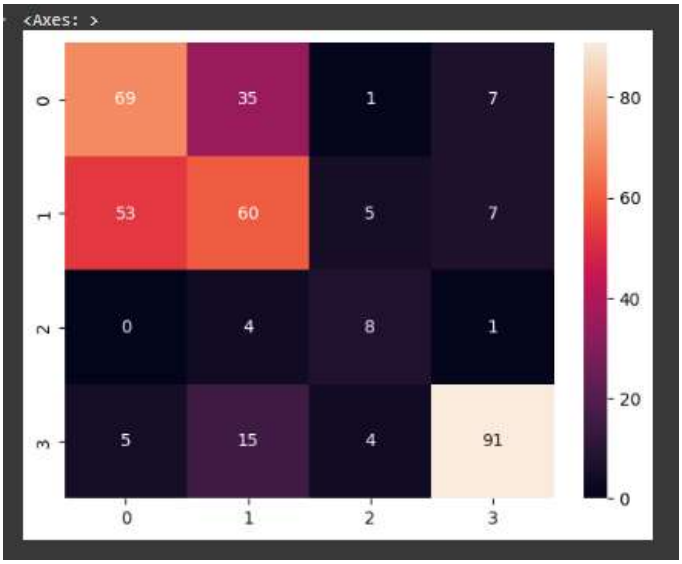

Figure 3. Showing the confusion matrices to evaluate the predictions

Confusion Matrix which is the key evaluation metrics for the classification problems. confusion matrix provides a visual representation of the performance of the classification model by comparing actual and predicted values.

- Diagonal Values (like – 69,60,8,91) represents the correct prediction for each class.
- Off-diagonal values indicate misclassification, showing where the model is confused on one class with another.

Conclusion From the matrix

- Class 0 and Class 1 have relatively highest misclassification rates between each other (35 and 53), suggesting that they shared the similar features and the model finds it challenging to differentiate between them.
- Class 3 demonstrated the best performance with the 91 correct predictions and relatively fewer misclassification.
- Class 2 has the fewer samples but shows the moderate performance with the 8 correct predictions.

Confusion matrix helps us understand that where the model is struggling and which class require more focused feature engineering or data balancing.

- Line Charts for all the model performance comparisons.

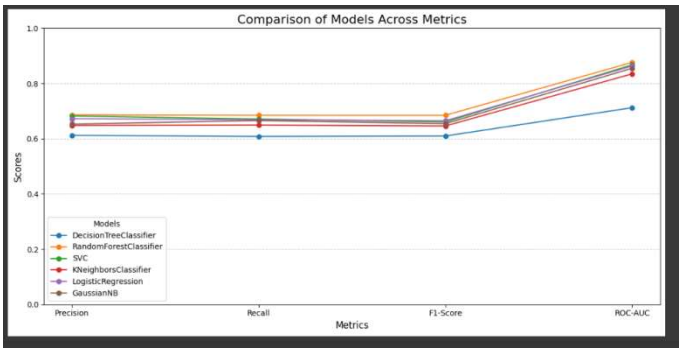

Figure 4. Showing the line chart comparison of all the model performance

Key Findings of the Visualizations

- The Heatmap revealed a strong correlation between the features such as temperature, humidity, and atmospheric pressure.

RESULT AND ANALYSIS

The performance of each model is summarized in the below table.

S.no	Model	Precision	Recall	F1-score	ROC-AUC
1	Decision Tree Classifier	0.591297	0.586301	0.587366	0.699047
2	Random Forest Classifier	0.695677	0.687671	0.682538	0.868579
3	SVC (Support Vector Machine)	0.668947	0.654795	0.642666	0.841049
4	K Neighbors Classifier	0.632358	0.624658	0.621763	0.808877
5	Logistic Regression	0.624998	0.630137	0.623284	0.840700
6	Gaussian Naïve Bayes	0.636733	0.643836	0.624577	0.840327

- The Confusion metrices identified as the misclassification scenarios, particularly in overlapping of the weather categories.
- The line charts illustrated the comparative performance of all models we used across the evaluation metrices.

Key Findings of performance table

- Random forest outperformed from all the models that we choose for our project in the terms of accuracy, F1- Score, ROC-AUC.
- The Machine Learning algorithms SVM (Support Vector Machine) and KNN (K-Nearest Neighbors) demonstrated a moderate performance but lacked the robustness in handling the overlapping weather condition.
- The Machine learning algorithm Logistic Regression and Naïve Bayes served as effective baseline models, offering simplicity and efficiency.

Conclusion and Future Work: The project successfully implemented multiple machine learning models to predict weather conditions based on historical data. The Random Forest model emerged as the best performer, showcasing its suitability for handling complex and noisy datasets. The developed system provides a foundation for accurate weather prediction and can be scaled further for real-time applications.

Future Work Could Involve:

- The Future work can involve additional features such as satellite data and real time sensor reading for the weather prediction.
- The Future work can include the dataset to include diverse geographical locations and seasons.
- The Future work can also include advanced deep learning models like LSTMs and CNNs for the enhanced temporal and spatial analysis.

REFERENCES

Analysis of Weather Prediction using Machine Learning by Shubham Madan, Parveen Kumar, Seema Rawat, Tanpuriya Choudhary on 2018

Designing a Model for Weather Forecasting Using Machine Learning by K. Geetha Rani, Dr. D.C. Joy Winnie Wise, S. Sufiyah Begum, S. Nirosha

Developing a Hybrid Weather Prediction Model Using Machine Learning and Traditional Forecasting Techniques by Priya Singh, Isabella Cruz, Rafael Mendoza, Noah Kim, Sofia Patel and Aria Martinez.

Evaluation of Empirical Equations and Machine Learning Models for Daily Reference Evapotranspiration Prediction Using Public Weather Forecasts by Yunfeng Liang, Dongpu Feng, Zhaojun Sun and Yongning Zhu.

Hyper Localized Weather Prediction Using Machine Learning and AI by Khurram Waris, Hadnall and Shropshire

Machine Learning Applied to Weather Forecasting by Mark Holmstrom, Dylan Liu, Christopher Vo in Stanford Universityon December15, 2016

Machine Learning Models for Prediction of Meteorological Variables for Weather Forecasting by Omodara E. Obisesan

Probabilistic Weather Forecasting with Machine Learning by Ilan price, Alvaro sachez-Gonzalez, Ferran alet, Tom R. Anderson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam and Maththew willson

Seasonal Weather Pattern Prediction Using Machine Learning by Mohammad Mohsin, Fahima Akhtar, Tanima Ghosh and Sanupa Sarkar.

Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee by AHM Jakaria, MD Mosharaf Hossain, Mohammad Ashiqur Rahman on November 2018

Smart Weather Prediction Using Machine Learning by Suvendra Kumar Jayasingh, Jibendu Kumar Mantri and Sipali Pradhan.

The study by Karthik Sudarshan, Vishnu Soman, Kiran K, Shreenidhi S Deshpande (2020)

Weather Based Wheat Yield Prediction Using Machine Learning by Ananta Vashisth, Achal Lama, P. Krishnan and Mausam IMD

Weather Event Severity Prediction Using Buoy Data and Machine Learning by Vikas Ramachandra.

Weather Forecasting Prediction Using Ensemble Machine Learning for Big Data Applications by Hadil Shaiba, Radwa Mazrouk, Mohamed K Nour, Noha Negm, Anwer Mustafa Hilal, Abdullah Mohamed, Abdelwahed Motwakel, Ishfaq Yaseen, Abu Sarwar Zamani and Mohammed Rizwanullah

Weather Forecasting Using Machine Learning Algorithms by Nitin Singh, Saurabh Chaturvedi, Shamim Akhtar in 2019

Weather Forecasting using Machine Learning Techniques by Siddharth Singh, Mayank Kaushik, Ambuj Gupta, Anil Kumar Malviya

Weather Forecasting Using Machine Learning Techniques: Rainfall and Temperature Analysis by Adil Hussain, Ayesha Aslam, Sajib Tripura and Vineet Dhanawat.

Weather prediction using machine learning by Dr Uday chandarkant Patkar.

Weather Prediction Using Random Forest Machine Learning Model by R. Meenal, Prawin angel Michael, D. Pamela, E. Rajasekaran.

****