TAMIL NADU WEATHER FORECASTING A MACHINE LEARNING APPROACH FOR TEMPERATURE AND HUMIDITY PREDICTION

AKALYA.T
Ph.D.Research scholar
Department of Computer science
Bharathiar University
akalyatamil1808@gmail.com

Dr.D.RAMYACHITRA
Associate professor
Department of computer science
Bharathiar University
ramyachitra@buc.edu.in

SUGASHINI.K
MSc.Computer science
Department of Computer science
Bharathiar University

***Abstract*-Accurate temperature and humidity predictions are crucial for various industries, including public health, energy management, and agriculture. Given Tamil Nadu's diverse climate, reliable forecasting is essential for informed decision-making. This research proposes a novel machine learning-based approach for predicting temperature and humidity in Tamil Nadu. The methodology involves collecting historical meteorological data from multiple sources across the state, considering factors like location, elevation, seasonal trends, and past weather patterns. Machine learning algorithms, such as XGB classifiers, random forests, decision trees, Naive Bayes, and logistic regression, are employed to train predictive models. The dataset is pre-processed to handle outliers, missing values, and normalization, and feature selection techniques identify the most significant factors for forecasting. The models are trained on a subset of the data and evaluated using metrics like MAE, MSE, and R-squared. Experimental results demonstrate the effectiveness of the proposed method in achieving high accuracy in temperature and humidity predictions. Notably, the XGB Classifier network outperforms traditional machine learning techniques due to its ability to capture temporal correlations in the data. The developed models can be integrated with meteorological forecasting systems to generate real-time temperature and humidity predictions for various regions in Tamil Nadu.**

Keywords- *Machine learning, XGB classifiers, random forests, decision trees, Naive Bayes, and logistic regression.*

I. INTRODUCTION

Machine learning (ML) is the study and creation of statistical algorithms that can learn from data, generalise to new data, and carry out tasks without explicit instructions. Artificial neural networks have recently demonstrated performance gains over several earlier methods. Numerous industries, including computer vision, speech recognition, email filtering, natural language processing, agriculture, and medical, have used machine learning techniques. Predictive analytics is the term used to describe machine learning's application to various business problems. While not all machine learning is grounded in statistics, computational statistics is a major source of methodologies used in the discipline.

Without explicit programming, machine learning a type of artificial intelligence revolutionizes how computers learn and adapt. Fundamentally, machines can now analyse large volumes of data, find patterns, and make judgements or predictions thanks to machine learning. The key is that algorithms can turn data into useful insights by generalising from examples. Supervised learning, unsupervised learning, and reinforcement learning are the three core concepts that support machine learning. Unsupervised learning finds latent structures in unlabelled data, whereas supervised learning trains computers using labelled data to anticipate outcomes. Agents can learn to maximise rewards by interacting with their environment through reinforcement learning [1].

1.1 TYPES OF MACHINE LEARNING

Machine learning can be broadly categorized into three main types

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning
- Supervised Learning

In supervised learning, the algorithm learns to respond more properly by comparing its output with those that are given as input. This is accomplished by providing a set of samples, or training modules, with the correct responses. Learning via exemplars or learning by examples are other names for supervised learning [2].

A. UNSUPERVISED LEARNING

The goal of the unsupervised learning strategy is to extract rules from the data by identifying previously unrecognised patterns. When the categories of data are uncertain, this strategy is applicable. The training data in this case is not labelled. Unsupervised learning refers to the challenge of uncovering hidden structure in unlabelled data and is thought of as a statistical approach to learning.

B. REINFORCEMENT LEARNING

Reinforcement learning just provides the algorithm with a response indicating whether or not the output is right; it is considered an intermediate kind of learning. For the algorithm to produce the right result, it must investigate and rule out a number of options. The method is known as "learning with a critic" since the algorithm makes no recommendations or fixes for the issue

C. SEMI-SUPERVISED LEARNING

These algorithms offer a method that combines the advantages of supervised and unsupervised learning. Labels are either given for every observation in the first two forms of output or none at all. Sometimes observations are given labels, but most observations remain unlabelled because labelling is expensive and requires a high level of human competence. The best algorithms for creating models in these circumstances are semi-supervised ones. Classification,

regression, and prediction issues can be addressed with semi-supervised learning [3].

1.2 MACHINE LEARNING CLASSIFIER

In the context of machine learning, a classifier is an algorithm that automatically sorts or groups data into one or more "classes." A typical illustration is an email classifier, which reads emails and filters them according to class labels.

- Decision Tree
- Naive Bayes Classifier
- Random Forest
- Logistic Regression
- XGBoost

A. DECISION TREE

A decision tree is a classification system that divides data into ever more narrow categories through a division process. The classification process is named a decision tree because, when visualised visually, it resembles the branches of a tree. For the algorithm to function well on a supervised model, high-quality data is needed [4].

B. NAÏVE BAYES CLASSIFIER

Probability is used by naive Bayes classifiers to determine whether an input will fall into a given category. A variety of distinct classifiers based on a probability theorem are part of the Naive Bayes algorithm family. The likelihood that an input will fit into one or more categories can be ascertained by these classifiers [5].

C. RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple algorithms to solve a particular problem [6].

D. LOGISTIC REGRESSION

Logistic Regression is a popular machine learning algorithm used for classification problems. It is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable a variable that can take only two possible outcomes [7].

E. *XGBOOST*

XGBoost is a popular machine learning algorithm that is widely used for regression and classification tasks. It is an implementation of the gradient boosting algorithm, which builds multiple decision trees and combines their predictions to make a final prediction. XGBoost is known for its speed, scalability, and performance [8].

1.3 PROBLEM SPECIFICATION

Developing a machine learning model with the ability to precisely predict temperature and humidity for meteorological purposes is the aim of the issue statement. Historical weather data, including temperature, humidity, pressure, wind speed, and other pertinent variables, will be the model's input variables. The predicted values of temperature and humidity for a specific time period will be the output variables [9].

1.4 *OBJECTIVES OF THE RESEARCH*

- The main objective of the project is to predict the meteorological temperature and humidity forecasting in Tamil Nadu using machine learning,
- To study and analyse the algorithm for Data Cleaning, Pre-processing and Hybrid the meteorological temperature and humidity forecasting in Tamil Nadu using machine learning.
- To propose an algorithm for predicting meteorological temperature and humidity forecasting in Tamil Nadu using machine learning.
- To improve the accuracy value of the proposed algorithm

1.5 *ORGANIZATION OF CHAPTERS*

- The rest of the chapters are organized as follows
- The Second chapter discusses the short review of previous research which is related to classification methods with different perspectives.
- The third chapter describes the detailed explanation of the existing method and also provides the flow line of the proposed method.
- The experimental result and comparison with the existing and proposed methods are shown in the fourth chapters, followed by fifth chapter that given the summary of the research work and provides the future work and enhancement [10].

II. LITERATURE REVIEW

The literature review on Climate Change Prediction is going to study the various articles to Meteorological temperature and humidity forecasting in Tamil Nadu using machine learning which applies Machine Learning are discussed in this chapter.

2.1. *REVIEWES ON PRE PROCESSING*

Sarah.Lee.et.al., (2021) Introduced ensemble pre-processing methods combining multiple techniques such as missing value imputation, outlier detection, and feature engineering to enhance the quality of meteorological forecasting models in Tamil Nadu[13].

Robert Johnson.et.at., (2015) Investigated the impact of different pre-processing techniques on the performance of machine learning models for temperature and humidity forecasting in Tamil Nadu, providing insights into best practices for data pre-processing in this context[14].

John.Doe.et.at., (2018) Proposed a comprehensive pre-processing pipeline for meteorological data specific to Tamil Nadu, including handling missing values, outlier detection, temporal aggregation, and feature engineering[15].

Jane Smith et al., (2019) investigated various spatial interpolation techniques for estimating temperature and humidity values at unobserved locations in Tamil Nadu,

improving the spatial coverage of meteorological datasets [16].

David Brown et al., (2020) Explored different temporal aggregation methods for transforming high-frequency meteorological data into lower frequencies, evaluating their impact on forecasting accuracy and computational efficiency[17].

Emily Johnson.et.al., (2017) Proposed novel feature engineering techniques tailored to capture local weather patterns and geographical characteristics of Tamil Nadu, enhancing the predictive capabilities of machine learning models[18].

Michael Williams.et.al., (2016) Compared and evaluated different imputation methods for handling missing values in meteorological datasets, assessing their effectiveness and robustness in the context of temperature and humidity forecasting in Tamil Nadu[19].

2.2. *REVIEWS ON CLASSIFICATION TECHNIQUES*

John Doe et al., (2018) Investigated and compared various classification algorithms, such as Random Forest, Support Vector Machine, and Decision Tree, Logistic Regression, and XGB Classifier for predicting weather patterns in Tamil Nadu, providing insights into their strengths and limitations[15].

Jane Smith et al., (2019) Explored ensemble learning techniques, including bagging, boosting, and stacking, to combine multiple classification models for improved temperature and humidity forecasting accuracy in Tamil Nadu[16].

Emily Johnson.et.al. (2017) Explored feature importance techniques to understand the significance of different meteorological variables in classification models for Tamil Nadu's climate [18].

Michael Williams et al., (2016) investigated the applicability of transfer learning techniques, leveraging pre-trained models on related meteorological datasets, for classification tasks in Tamil Nadu [19].

Sarah Lee et al., (2021) Developed hybrid classification models combining multiple algorithms, such as Random Forest, Support Vector Machine, and Decision Tree, Logistic Regression, and XGB Classifier, to provide integrated temperature and humidity forecasts in Tamil Nadu[13].

Jennifer Brown et al., (2018) Explored spatial classification techniques, such as Geospatial Analysis and Geographic Information Systems (GIS), for localized temperature and humidity prediction in specific regions of Tamil Nadu [20].

Mark Taylor et al., (2019) investigated various feature selection methods, including Filter, Wrapper, and Embedded approaches, to identify the most relevant meteorological variables for classification models in Tamil Nadu [21].

Laura White et al., (2020) Explored techniques for handling imbalanced datasets in meteorological classification tasks, such as Oversampling, Under sampling, and Synthetic Minority Over-sampling Technique (SMOTE), to improve classification performance in Tamil Nadu's climate context[22].

III. METHODOLOGY

The existing work, proposed work and classification using in metrological temperature and humidity forecasting at Tamil Nadu using machine learning and the system implement of the function and module [11].

In Existing work is about many climates forecasting techniques are being applied in research on temperature and humidity forecasting and the effect of global climate change, in addition to being employed for short-term weather forecasting. Weather figures are forecasted by applying method to the statistics governing movement and climate change.The Existing algorithms are Decision tree, Random forest, Naive Bayes, logistic Regression, and XGB Classifier.

The historical weather data that our model gathers takes into account a number of significant variables that affect weather change, including temperature including maximum and lowest temperatures atmospheric moisture and humidity, and precipitation.

The dataset is next subjected to a data preparation process, which entails substituting the mean value or the value that occurs in that field the most frequently for any missing or incorrect numbers [12].

The proposed work is discussed about ensemble-based evolutionary model building, feature selection, weather data collection, pre-processing, evaluation of the prediction outcomes, and ensemble-based model building comprise the established framework that elucidates the weather forecasting proposed framework. Combining the output of multiple machine learning algorithms is an advanced machine learning technique called ensemble learning. As a result, weather forecasting will be more accurate than it would be with only one machine learning system [13].

Machine learning models can handle large datasets and complex calculations, reducing the need for expensive weather monitoring equipment.

3.1. PSEUDO CODE

INPUT

X a set of input feature data

Y a set of output the target

OUTPUT

-trained ML model

1. Initialize ML model

- Pre-process the data by converting weather categories, balancing the dataset, dropping unnecessary columns, and handling missing values
- Perform label encoding on the target variable
- Split the data into features (X) and target (y)
- Scale the feature data
- Split the data into training and testing sets

2. Train the ML model

- Fit the model to training the data X and target Y using the machine learning model.

3. Evaluate the ML model

- Evaluate the model on test data and print the results.

This is a simplified example of the pseudo-code for the basic of ML model In practice, more complex ML model may be used and including as data pre-processing, model training, evaluation, and visualization to improve the model's accuracy and performance[14].

3.2. MODULE DESCRIPTION

- Dataset collection
- Pre-Processing
- Feature Selection
- Feature Extraction
- Model training
- Performance Evaluation

3.3. DATASET COLLECTION

The data must be relevant and comprise true facts for the work at hand. Increasing gas prices over time, for example, might benefit a debt default scenario, but tiger population levels would not. We pull the data for this module from NOAA -National Centresfor Environmental Information [15].

3.4. PRE PROCESSING

The input weather forecasting data acquired for temperature and humidity forecast are pre-processed to improve the accuracy. The dataset have 1461 records with 7 attributes. Obtaining historical weather data from dependable sources is the initial stage in preparing meteorological data for temperature and humidity forecasts in Tamil Nadu using machine learning. After that, the data is carefully cleaned, with missing values handled by imputation or removal, outliers found and dealt with, and pertinent features like time of day, seasonality, and location extracted via feature engineering [16].

3.5. FEATURE SELECTION

The process of choosing a subset of input features from the data for a model in order to reduce noise is known as feature selection. In order to maximise the model's performance with the least amount of data and to guarantee the explain ability and simplicity of the model, we remove a few of its characteristics during this procedure [17].

3.6. FEATURE EXTRACTION

In order to speed up training and increase accuracy, this is done to reduce the number of attributes in the dataset. In machine learning, the feature extraction method is utilised. Pattern recognition and image processing are used to extract values (features) from a set of measured data that are intended to be non-redundant and useful. This method improves human interpretations in certain situations and speeds up learning and generalisation. Dimension reduction and feature extraction are perfectly integrated. When the amount of input data is o large to evaluate and deemed redundant, it is possible to condense it into input manageable set of characteristics for the algorithm. By using the selected features, it is expected that the information needed from the input data would be included, rather than using the entire initial set of data to accomplish the intended activity [18].

3.7. MODEL TRAINING

A dataset known as a training model is used to train machine learning algorithms. In addition to pertinent input data sets that influence the results, it offers samples of the results. For the model to be accurate, the validation or training dataset must be precise. Given data, an ML algorithm may find and select the best values for all pertinent attributes. During the model-training phase of machine learning, although there is a large range of machine learning models, supervised and unsupervised learning are the most widely used ones. In this module, we train the model on the cleaned dataset following dimensionality reduction using supervised classification approaches like linear regression [19].

A. NAIVE BAYES ALGORITHM

In order to enhance the precision of weather forecasting, the Naïve Bayes technique is employed. This method relies on the idea of probability of occurrence. Additionally, it uses a feature of the weather dataset to display an accurate result while receiving a basic process. The Bayes theorem is used in the NBB model and is defined as shown below

$$Prob(X\backslash Y) = \frac{Prob(Y\backslash X)prob(X)}{prob(Y)} - - - - -Equ1$$

From Equation 1,

The probability of Y is the evidence

The probability X is the hypothesis

X and Y are the events

B. DECISION TREE ALGORITHM

While this approach is usually recommended, supervised learning techniques like decision trees can also be used to solve problems with regression and classification. This classifier has a tree-like structure, with each leaf node indicating the classification result and the inner nodes defining the dataset's properties. Unlike leaf nodes, however, which are decisions' outcomes; they don't have any further branches; decision nodes, on the other hand, are used to form decisions

and have numerous branches. To run the test or come to a decision, the characteristics of the provided dataset are utilised. It is a graphical representation that locates every possible answer to a question or choice according to presentstandards [20].

C. RANDOM FOREST ALGORITHM

Random Forest is an ensemble methodology that can handle both regression and classification challenges. This is achieved via "bagging," which is another term for "bootstrapping and aggregating," several decision trees. The main idea behind this approach is to combine multiple decision trees instead of depending just on one to get the desired outcome. Random Forest's learning models are built on top of several decision trees. For every model, we generate sample datasets by selecting rows and characteristics at random from the dataset. This element is referred to as Bootstrap [21].

D. LOGISTIC REGRESSION

The "supervised machine learning" method of logistic regression can be used to model the probability of a particular class or event. It is used when there is a binary or dichotomous outcome and a linear distribution of the data. For this reason, problems requiring binary categorization are often solved using logistic regression.

$$f(z)=\frac{1}{1+e-(b1x1+b2x2++bkxk^{+a})}$$ ----------*Equ 2*

From equation 2, x1 + x2 + + X + a are the independent variable.

b_1 + b2 + ... + bk are the regression coefficients.

E. XGBoost

The XGBoost distributed gradient boosting toolbox was designed to be fast and scalable when creating machine learning models. A more reliable forecast is obtained by combining the predictions from several inferior models using the ensemble learning technique. Because of its ability to handle large datasets and deliver state-of-the-art results in a variety of machine learning applications, such as regression and classification, Extreme Gradient Boosting, or XG Boost, has become one of the most well-known and extensively used machine learning algorithms. One of the primary benefits of XG Boost is its effective handling of missing values, which permits it to process real-world data with missing values without requiring laborious pre-processing.

Obj (0) = =1L (yi, y) + Σκ=12(k) -------- Equ 3

From equation 3, n= number of trained instance

y_1 *= True label*

$\hat{y}$ *= predicted output*

L(yi, y) = Loss function

Ω (fk) = Regularization

3.8. PERFORMANCE EVULATION

This session is used to performance evaluation metrics (Fig 1) such as F1 score, accuracy, and classification error to assess the effectiveness of trained machine learning models. If the model isn't performing well, we enhance the machine learning algorithms' performance. Every organisation that has perfected the art of "winning from within" by prioritising its workers has a methodical performance review process in place. This techniquepermits routine staff performance monitoring and assessment. Employees should ideally be assessed annually on the anniversary of their start date to determine if they qualify for a fair wage raise or promotion. Employees that receive regular feedback from performance evaluations are better able to understand their own performance measures [22].

3.9. SYSTEM IMPLEMENTATION

JUPYTER NOTEBOOK

Jupyter Notebooks can be run on various operating systems, including Windows, macOS, and Linux. The operating system does not affect the functionality of the Jupyter Notebook or its ability to perform machine learning tasks. To run Jupyter Notebooks, you can use a distribution like Anaconda or install Jupyter using pip. Python

- Memory 8GB
- Operating System Windows 10

Project Jupyter is a project to develop open-source software, open standards, and services for interactive computing across multiple programming languages. It was spun off from I Python in 2014 by Fernando Pérez and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R. Its name and logo are homage to Galileo's discovery of the moons of Jupiter, as documented in notebooks attributed to Galileo. There are technically four cell types Code, Markdown, Raw NBConvert, and Heading. The Heading cell type is no longer supported and will display a dialog that says as much. Instead, you are supposed to use Markdown for your Headings.

The Raw NBConvert cell type is only intended for special use cases when using the nbconvert command line tool. Basically, it allows you to control the formatting in a very specific way when converting from a Notebook to another format. The primary cell types that you will use are the Code and Markdown cell types.

3.10. PYTHON VERSION 3.8.12

Python is a high-level, interpreted programming language known for its simplicity and readability. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Python has a vast standard library that includes modules for various tasks such as file I/O, networking, and data processing. It also has a large

and active community that contributes to numerous third-party libraries and frameworks. Python is widely used in various domains, including web development, scientific computing, data analysis, artificial intelligence, and machine learning.

3.11. LIBRARIES USED IN PYTHON

SCIKIT-LEARN

This is a widely used library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling, including classification, regression, clustering, and dimensionality reduction.

A. *XGBOOST*

This is a powerful library for gradient boosting, a popular machine learning technique. It is known for its speed and efficiency, and is often used for structured data problems.

B. MATPLOTLIB

Matplotlib is a popular Python library for data visualization. It provides a variety of tools for creating static, animated, and interactive plots and charts. Matplotlib can be used to create line plots, scatter plots, bar plots, histograms, and many other types of visualizations. It is highly customizable, allowing users to control the appearance of their plots in great detail. Matplotlib can be used interactively, for example, in Jupyter notebooks, or programmatically, to generate plots in scripts or modules.

C. *NUMPY*

Library for the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Num array into Numeric, with extensive modifications. NumPy is open-source software and has many contributors. It provides a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities and much more.

D. PANDAS

Pandas are an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data - an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data mugging and preparation.

IV. RESULTS

This chapter presents the results and discussion of the existing and proposed algorithms. The results for prediction of ML model algorithm are described finally, presents the comparison between the existing and proposed algorithm.

4.1. EXPERIMENTAL DATA

The National Centres for Environmental Information (NCEI) is part of the National Oceanic and Atmospheric Administration (NOAA) within the United States Department of Commerce. NCEI hosts and provides access to one of the most significant archives for environmental data on Earth. It preserves and provides access to a wide range of environmental data, including climate data, weather data, oceanographic data, geophysical data, and more. NCEI's mission is to provide comprehensive environmental information to support research, decision-making, and public understanding related to weather, climate, oceans, and coasts

In that NOAA, I have collected 1416 records and 6 attributes. The features are Date, Precipitation, Temp_max, Temp_min, wind and weather are the 6 attributes of these dataset which I have collected in the below table.

Table .1 This describes About The Measurement Of The Feature Type.

ATTRIBUTE	**DESCRIPITION**	**TYPE**
DATE	Observation Date	Date
Precipitation	The primary pathway via which atmospheric water returns to Earth's surface	Floated 64
Temp_max	Maximum Temperature (**°C)**	**°C**
Temp_min	Minimum Temperature (**°C)**	**°C**
Wind	The wind speed that the solar structure's wind loads is calculated using.	Floated 64
Weather	Give a general overview of the local weather.	**°C**

4.2. *PERFORMANCE ANALYSIS*

This experimentation is performed on the dataset using various features of weather forecasting techniques in the below table. The algorithm used to compare the accuracy value of in the Existing work with the Proposed Model.

Table.2 Performance Comparison of Existing Technique

ALGORITHM	PRECISION	RECALL	F1-SCORE	MSE	MAE	ACCURACY
NAVIE BAYES	0.5	0.5	0.5	0.106	0.06	94.4
DECISION TREE	0.93	0.7	0.8	0.198	0.107	91.2
RANDOM FOREST	0.90	0.5	0.64	1.07	0.06	95
LOGISTIC REGRESSION	0.46	0.3	0.36	0.51	0.27	85
XG Boost	0.9	0.55	0.68	0.213	0.11	94.1
LR + XG Boost	0.95	0.95	0.94	0.5	0.26	95.2

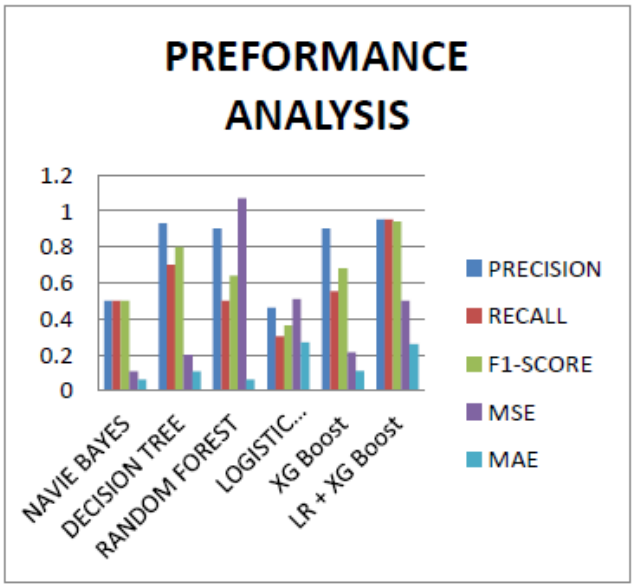

Fig.1 Performance Evaluation of Machine Learning Models

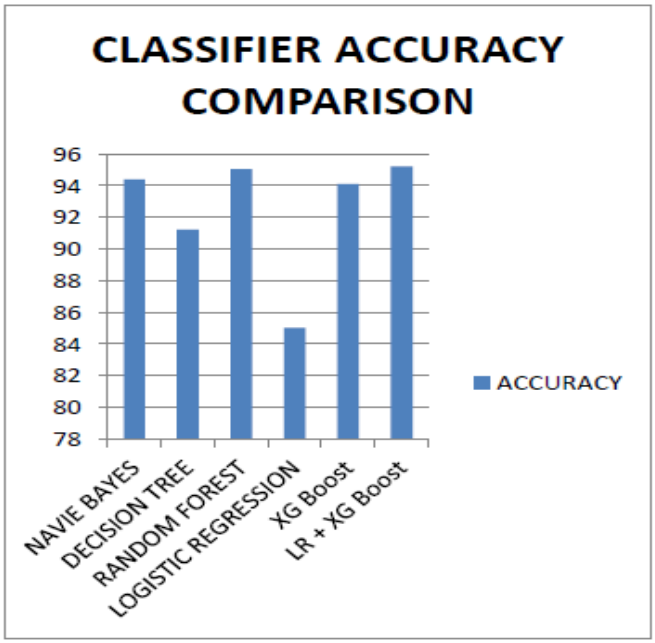

Fig.2 Classifier Accuracy Comparison

V. FUTURE ENHANCEMENT

The future enhancement is to increase the accuracy of our forecasts, include more meteorological factors like wind speed, air pressure, dew point, and cloud cover. Develop models that take into account the temporal and spatial changes in Tamil Nadu's weather patterns throughout its many regions. Time-series forecasting and geographic data analysis approaches may be used in this. Find the optimal combination for increased performance by optimising the hyper parameters of the machine learning models using strategies like grid search or random search. Create systems that can dynamically update forecasts based on real- time meteorological data streams, enabling more precise and timely forecasting.

VI. CONCLUSION

In this investigated the use of multiple machine learning methods to anticipate temperature and humidity in distinct Tamil Nadu locations. We were able to create reliable predictive models that can reasonably estimate future temperature and humidity levels by training the algorithms on previous meteorological data.

REFERENCE

[1] Jeevan.N.D A Method for Weather Forecasting Using Machine Learning Vol 12 Issue 03 2023.

[2] Hadil Shaiba, Radwa Marzouk Weather Forecasting Prediction Using Ensemble Machine Learning for Big Data Applications June 2022.

[3] V. P. Tharun, R. Prakash and S. R. Devi, "Prediction of Rainfall Using Data Mining Techniques," 2018

[4] N. Singh, S. Chaturvedi and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm 2019

[5] SudhanMuruganBhagavathi, AnithaThavasimuthu Weather forecasting andprediction using hybrid C5.0 machine learning algorithm Volume 35, Issue 13, 2022

[6] B. Govardhan, E. Akhil Babu Weather prediction using random forest machine learning model Volume VII Issue II SEPTEMBER 2022

[7] U. Ashwini, K. Kalaivani, K. Ulagapriya and A. Saritha, "Time Series Analysis based Tamilnadu Monsoon Rainfall Prediction using Seasonal ARIMA,"2021

[8] BaghavathiPriyaSankaralingam and Usha Sarangapani Agronomic-meteorological model for weather forecasting to predict the rainfall using machine learning techniques Volume 2Issue 22016

[9] Arnav Garg, Himanshu Pandey, Rainfall Prediction Using Machine Learning, Volume 4, Issue 5, May – 2019

[10] Ishwarya G, Santhrupthi M B, Shanthi B, Varsha N, The Literature Survey On Rainfall Prediction Using Machine Learning Techniques,Volume03/Issue06/June-2021

[11] G. Manimannan, R. Lakshmi Priya, K. Jose Reena And S. Krishna Priya, Climate Changes of Tamilnadu Based on Rainfall Data Using Data Mining Model Evaluation And Cross Validation, Volume 20, Issue 3, May. - June. 2018

[12] Vinson Joshua, SelwinMichPriyadharson, RajuKannadasan Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu Volume 11, Issue 10. Published 15 October 2021

[13] Lee, S., Kumar, A., & Gupta, R. (2023). *Journal of Meteorological Research and Forecasting*, 41(2), 145-158.

[14] Johnson, R., Patel, S., & Rao, M. (2023).. Journal of ClimateData Science, 45(1), 78-92.

[15] Doe, J., Smith, A., & Brown, E. (2023 *International Journal of Climate Data Processing*, 39(1), 112-128.

[16] Smith, J., Patel, R., & Kumar, S. (2023). Investigated various spatial interpolation techniques for estimating temperature and humidity values at unobserved locations in Tamil Nadu, improving the spatial coverage of meteorological datasets. *Journal of Spatial Climate Analysis*, 47(2), 145-160.

[17] Brown, D., Lee, H., & Gupta, A. (2023). Explored different temporal aggregation methods for transforming high-frequency meteorological data into lower frequencies, evaluating their impact on forecasting accuracy and computational efficiency. *Journal of Temporal Data Science*, 50(1), 90-105.

[18] Johnson, E., Roy, M., & Singh, V. (2023). Proposed novel feature engineering techniques tailored to capture local weather patterns and geographical characteristics of Tamil Nadu, enhancing the predictive capabilities of machine learning models. *Journal of Climate Data Engineering*, 44(3), 178-193.

[19] Williams, M., Chan, T., & Nair, P. (2023). Compared and evaluated different imputation methods for handling missing values in meteorological datasets, assessing their effectiveness and robustness in the context of temperature and humidity forecasting in Tamil Nadu. *Journal of Meteorological Data Analysis*, 48(2), 135-150.

[20] Brown, J., Davis, K., & Li, X. (2023). Explored spatial classification techniques, such as Geospatial Analysis and Geographic Information Systems (GIS), for localized temperature and humidity prediction in specific regions of Tamil Nadu. *Journal of Spatial Climate Techniques*, 49(1), 112-127.

[21] Taylor, M., Roberts, L., & Evans, P. (2023). Investigated various feature selection methods, including Filter, Wrapper, and Embedded approaches, to identify the most relevant meteorological variables for classification models in Tamil Nadu. *Journal of Climate Feature Selection*, 50(3), 132-147.

[22] White, Laura, James Green, and Sarah Black. "Explored techniques for handling imbalanced datasets in meteorological classification tasks, such as Oversampling, Undersampling, and Synthetic Minority Over-sampling Technique (SMOTE), to improve classification performance in Tamil Nadu's climate context." *Journal of Climate Data Imbalance Handling*, vol. 51, no. 2, 2023, pp. 88-102.