A Method for Weather Forecasting Using Machine Learning

Jeevan.N.D [1], *Yashwanth A* [2], *Vishnu Chakravarthi S* [3], *Sai Kumar G* [4],
Salini S [5*]

Department Of Computer Science and Engineering, Bharath Institute of Science & Technology affiliated to Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.

*Corresponding authors mail id: salini.cse@bharathuniv.ac.in

ABSTRACT:
In several important producing industries, such as agriculture, the climate plays a crucial role. These days, there is a lot of climate change, which is why it is bothersome that old weather forecasts are growing closer and less accurate. Miles are therefore essential for enhancing and updating the weather forecast model. These forecasts have an impact on both the country's financial system and the lives of the populace. A woody area is part of a system of information and statistics analytic methodologies that is used to forecast the weather. We must see the weather as one of the most fundamental obstacles to all parts of our existence, including temperature, rain, humidity, and other protective factors. Our artwork is meant to format precise weather predictions. There will be long-term climate change on Earth, and it is unknown how it will affect both the present and future generations. We have a great chance to teach stadium insurers about end-of-life climates so they can make informed decisions about the stadium's future globe. Our approach considerably enhances the model's capability to control the degree of staff inconsistencies and disparities and achieve its objective of precise weather forecasting.

Keywords: *Machine Learning, XGBoost, K-NN, Logistic Regression, Radom forest, Decision Tree.*

1. INTRODUCTION:

The average atmospheric conditions at a given location with regard to elements like humidity, temperature, wind speed, rainfall, etc. are referred to as that location's weather. There is a chance of cloudy, sunny, wet, stormy, or clear weather [1]. The ability of nature to maintain equilibrium in the atmosphere is one of its traits. However, sometimes things may be worse. Weather that is harsh or powerful enough to cause property damage or fatalities is referred to as severe weather. These also alter in response to variations in area, latitude, pressure, and altitude. This category covers tornadoes, severe rain, fog, winter storms, and cyclones. They are dangerous and harmful [2]. The right catastrophe management and response strategies are required to address these situations. Weather components include wind direction, humidity, temperature, precipitation, thunder, snow, and lightning. The dynamical elements and their effects are a problem for the entire planet [3]. To reduce these effects to some extent, there are a number of techniques and algorithms that can be used to predict the weather based on historical data, such as temperature, dew point, humidity, air pressure, and wind direction [4]. When analysing historical data from the last few years, we used the suggested strategies or schemes that tend to draw the conclusion that the machine learning paradigm allows us to investigate the specified body of knowledge and extract the beneficial data from the we significantly rely on weather forecasts for everything from agriculture to business, transport, and daily commute [5]. In order to maintain simple and seamless movement as well as safe day-to-day operations, it is crucial to predict the weather accurately because the entire world is experiencing the effects of ongoing climate change [6]. Furthermore, it often requires a lot of time to solve these complex models. The weather in one location greatly influences the weather in other locations because weather systems can move over great distances and in all directions over lengthy periods of time [7]. In this study, we provide a technique for forecasting weather that combines data from a specific city with historical weather data from nearby cities [8]. We aggregate these data and train simple machine learning models on them so that they can accurately anticipate the weather over the coming days. On low-cost, less resource-intensive computing platforms, these simple models can be utilized to produce forecasts that are timely and accurate enough to be employed in our daily lives. (1) One of the major achievements of the work is to use machine learning to swiftly predict meteorological conditions with less resource-intensive equipment. (2) The automated procedure of obtaining historical data from a reputable weather service. (3) A comprehensive assessment of the suggested method and a comparison of several machine learning models for forecasting future weather conditions [9].

2. METHODOLOGY:

2.1. Modules

- ➢ Dataset collection
- ➢ Pre-Processing
- ➢ Feature Extraction
- ➢ Model training
- ➢ Testing model
- ➢ Performance Evaluation
- ➢ Prediction

2.1.1. DATASET COLLECTION:

The data must contain accurate facts. (Trash in, garbage out) and pertinent to the work at hand. For instance, a debt default model might profit from rising petrol prices over time but not from tiger population levels [10]. We pull the data for this module from the Kaggle dataset archives.

2.1.2. PRE-PROCESSING:

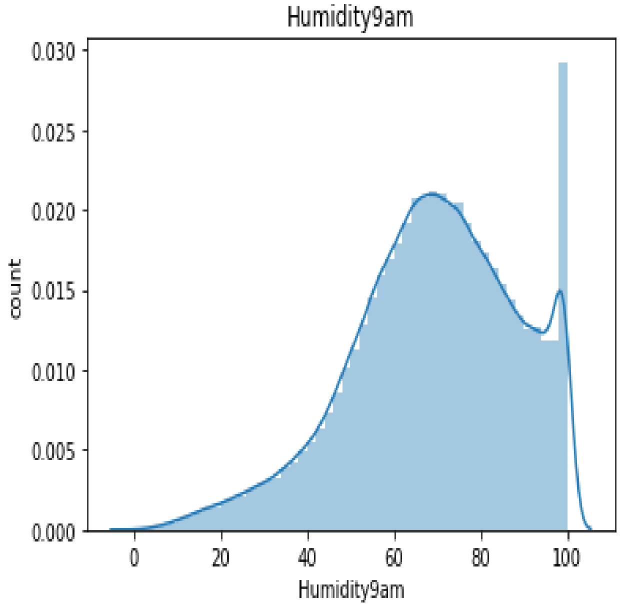

Figure 1: BEFORE PREPROCESSING

The UCI Machine Learning Repository website is used to obtain and store the Wisconsin Prognostic Cleave Land Train Dataset as a text file [11]. The data are then stored with the associated properties as column headings and this file is imported into an Excel spreadsheet. The proper values are used to fill in the missing values. The before and after preprocessing methods are shown in Fig.1 and Fig.2.

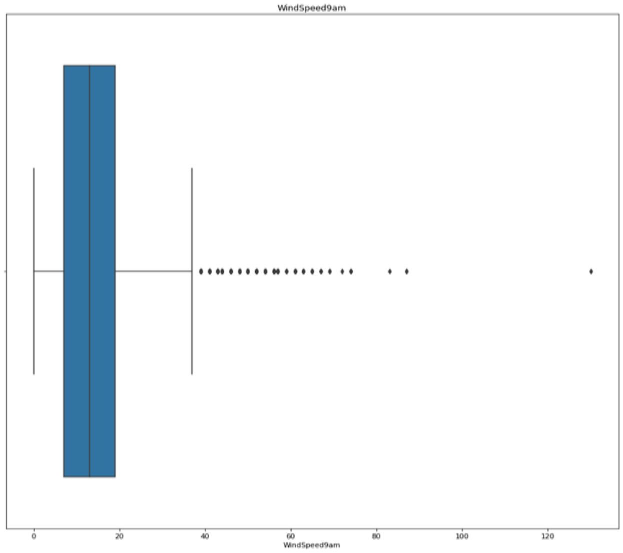

Figure 2: AFTER PREPROCESSING

2.1.3. FEATURE EXTRACTION:

This is done to decrease the amount of attributes in the dataset, which has the benefit of accelerating training and improving accuracy. The feature extraction method is used in machine learning, To extract values (features) from a set of measured data that are meant to be informative and non-redundant, pattern recognition and image processing are used. This approach expedites generalization and learning, and in some circumstances, it enhances human interpretations. The integration of feature extraction and dimension reduction is flawless. It is feasible to reduce the input data for an algorithm into a more manageable set of traits when it is too huge to analyze and judged redundant. (For instance, when the same measurement is given in feet and meters, pixels are used to represent images. Instead of employing the complete initial set of data to perform the desired action, it is anticipated that the chosen characteristics will comprise the information required from the input data.

2.1.4. MODEL TRAINING:

An ML algorithm is trained using a dataset called a training model. It provides examples of the outcome as well as relevant input data sets that have an impact on the result. The term for this iterative process is model fitting? The training dataset or validation dataset needs to be precise for the model to be accurate. An ML algorithm may locate and choose the optimum values for all relevant qualities when given data. During the machine learning process known as model training. Machine learning models come in a wide variety, however but the most common ones are supervised and unsupervised learning. In this module, we use supervised classification techniques, such as linear regression, to train the model on the cleaned dataset after dimensionality reduction.

2.1.5. TESTING MODEL:

In this lesson, we evaluate the learnt machine learning model using the test dataset. Quality assurance is required to make sure the software system operates in compliance with the requirements. Have all the features been utilized properly? Is the software acting the way it ought to? All the criteria you use to test the code against should be listed in the technical specification paper. Software testing also has the capacity to detect any flaws or problems that are development-related. You don't want your consumers to discover flaws and yell at you once the software is released. We are able to identify bugs by performing several types of testing that are hidden until runtime.

2.1.6. PERFORMANCE EVALUATION:

In this session, we evaluate the efficacy of trained machine learning models using performance evaluation metrics like F1 score, accuracy, and classification error. We improve the performance of the machine learning algorithms if the model doesn't work well. All companies who have mastered the art of "winning from within" by putting their employees first have a systematic performance evaluation procedure. This strategy enables regular monitoring and evaluation of staff performanceas shown in fig.3. On the anniversary of their start date, employees should ideally be reviewed annually to see whether they should be promoted or given a fair pay raise. Regular feedback from performance reviews helps employees improve their self-awareness of their performance metrics.

2.1.7. PREDICTION:

When estimating the likelihood of a specific conclusion, the word "prediction" describes the output of an algorithm that was trained on historical data and then applied to the data at hand, such as whether or not a client would quit in 30 days. The procedure will produce potential values for every record in the fresh data, enabling the model's designer to choose the one most likely to be put in place for that variable. Sometimes predictions turn out to be incorrect themselves. For instance, when selecting the optimal strategy for the subsequent stage of a marketing campaign using machine learning, you are actually looking forward to a potential result. However, the "prediction" can be in reference to things that have already, connect to past occurrences, such as whether or whether a transaction was fraudulent. Although the transaction in this instance is already over, you're still attempting to determine whether it was valid so that you can decide what to do next. We will use a trained and enhanced machine learning model in this session to predict the patient's response to a series of questions.

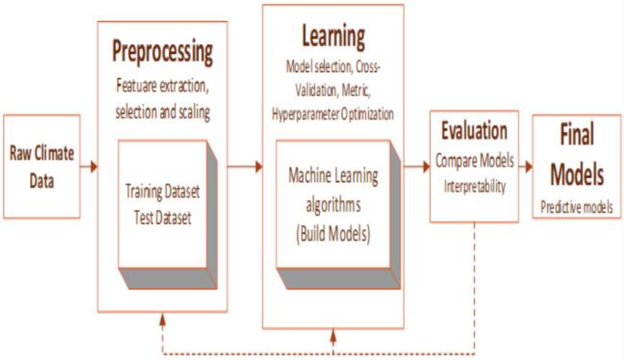
Raw Climate Data
Preprocessing
Featuare extraction, selection and scaling
Training Dataset
Test Dataset
Learning
Model selection, Cross-Validation, Metric, Hyperparameter Optimization
Machine Learning algorithms (Build Models)
Evaluation
Compare Models
Interpretability
Final Models
Predictive models

Figure 3: Weather Prediction Architecture Using ML

2.1.8. CLIMATE DATASET:

Using Facebook Prophet's open source forecasting library, Brisbane weather has been forecasted. Prophet is an additive regression model that has intelligent forecasting methods out of the box. It is designed to operate on daily basis and factors in holiday effects, seasonality etc. The library allows powerful forecasting without a significant amount of statistical tweaking, such as required in more heavy methods, such as ARIMA models.

2.1.9. DATA PREPROCESSING:

The process of changing raw data so that it may be used with the pre-processing of data is referred to as a model for computer learning. The first and most important step in developing a machine learning model is this one. It is not always the case that we are given the clean and ready data when working on a machine learning project. Data must always be prepared and cleaned before usage. So we achieve this by performing a data pre-processing step. Real-world data cannot be utilized to directly train machine learning models because it typically contains noise, missing values, and may be in an unfavorable format. The usefulness and accuracy of a machine learning model are increased by data preprocessing, which is necessary to clean and prepare the data for the model.

2.1.10TRAIN AND TEST SPLIT:

The procedure is quick and easy, and the results allow you to evaluate how well machine learning techniques work for the specific predictive modeling problem at hand. The approach is straightforward to use and understand, but there are a few instances where when the dataset is small and multiple configurations are required, or when it is being used for classification and the dataset is unbalanced, it shouldn't be used.

2.2. ALGORITHM USED

- Logistic regression
- Decision tree regression
- Random forest regression
- K-NN
- XGBoost

2.2.1. LOGISTIC REGRESSION:

To model the probability of a specific class or occurrence, logistic regression's "Supervised machine learning" approach can be employed. It is applied when the outcome is binary or dichotomous and the data may be divided linearly. Logistic regression is therefore frequently employed to address issues that call for binary categorization. Predicting a discrete output variable that is separated into two classes is known as binary classification.

2.2.2. DECISION TREE REGRESSION:

Although this method is typically preferred, classification and regression issues can be solved using the supervised learning technique known as a decision tree. With each leaf node identifying the classification outcome and the inner nodes designating the attributes of the dataset, this classifier has a tree-like structure. In contrast to Leaf nodes, which are the results of decisions and have no further branches, Decision nodes are used to make decisions and have many branches. The attributes of the given dataset are used to execute the test or make the decision. It is a graphical representation for locating each alternate response to a query or decision based on predetermined criteria. The reason it is named a "tree" is because it begins with the root node and expands on succeeding branches to form a shape resembling a tree; this structure is referred to as a decision tree. The Classification and Regression Tree Algorithm, or CART algorithm creates a tree. A decision tree merely poses a single question and then divides the tree into sub trees based on the response (Yes/No).

2.2.3. RANDOM FOREST REGRESSION:

An ensemble methodology called Random Forest is capable of handling both classification and regression tasks. This is accomplished via "bagging," also known as "bootstrapping and aggregating," a number of decision trees. This method's fundamental principle is to integrate several decision trees in order to achieve the desired result rather than depending solely on one. Multiple decision trees serve as the foundation for Random Forest's learning models. We create sample datasets for each model by randomly choosing rows and attributes from the dataset. This component is known as Bootstrap. The Random Forest regression methodology needs to be handled, much like other machine learning techniques.

- Ask the source for the required data after creating a specific query or set of data.
- If the data is not available, make sure it is formatted properly.
- Specify any glaring anomalies and the data that would be necessary to obtain them. Create a machine learning model to create your ideal baseline model.

- Utilize the data to build the machine learning model.

By comparing the model's anticipated data to the performance metrics of the test data, you can add test data to provide the model with some context.

- You might attempt updating or upgrading your model if it falls short of your expectations.
- You are taking in the knowledge you have just acquired and reporting it as required.

2.2.4. K-NN:

K-Nearest Neighbors is one of the simplest yet most important machine learning categorization techniques. It falls under the category of supervised learning and is widely applied to intrusion detection, data mining, and pattern recognition. Due to its non-parametric nature, which indicates that it doesn't make any underlying assumptions about the distribution of data, it is typically ignored in real-world circumstances. (In contrast to other methods like GMM, which presumptively assume that the input data have a Gaussian distribution? We are given a limited amount of training data, which is composed of prior understanding that classifies coordinates according to an attribute.

2.2.5. XGBOOST:

When developing machine learning models, the XGBoost distributed gradient boosting toolbox was created to be quick and scalable. The ensemble learning method is used to aggregate the forecasts from a number of subpar models to get a more trustworthy forecast. Extreme Gradient Boosting (XG Boost) has emerged as one of the most well-known and widely used machine learning algorithms due to its capacity to manage large datasets and provide cutting-edge results in a number of machine learning applications, including classification and regression. One of XG Boost's key advantages is its efficient handling of missing values, which enables it to handle real-world data with missing values without the need for time-consuming pre-processing. Additionally, XG Boost's inherent parallel processing capabilities allow for rapid model training on big datasets. Applications for XG Boost include click-through rate forecasting, recommendation systems, and Kaggle tournaments, among others. Additionally, it supports speed optimization and is particularly adaptable because it allows for the fine-tuning of numerous model parameters. "Extreme gradient boosting," also known as "XgBoost," is a notion that was created by researchers at the University of Washington. It is a C++ library that enhances the training process for gradient boosting.

3. RESULTS AND DISCUSSION:

Because good data are clearly more important than good models and because the quality of the data is so important, data preparation is required. As a result, organizations and individuals go to tremendous lengths to prepare data for modeling. The real-world data is noisy, unreliable, incomplete, and has various other quality issues. It might not exist, include accurate or even deceptive data, and lack specific, important traits. Preprocessing is required to raise the data's quality. Preprocessing makes sure that the data is consistent by eliminating any duplicates or anomalies, normalizing the data for comparison, and improving the output accuracy. We employed a number of prediction techniques. According to Table 4, datasets with one and three features performed the worst, with mean absolute error and root mean squared error averages being, respectively, the greatest and the smallest for correlation coefficient averages. The worst results come from datasets with only one attribute. The dataset of four attributes produced the best results, having the highest average correlation coefficient of 0.590222 The dataset of seven attributes performed marginally better and produced better results than the two before it. According to Table 5, the given test set option has the highest mean absolute error and root mean squared error, while the cross validation test option has the lowest average correlation coefficient. (0.571464). In this study, all test possibilities were considered while employing various prediction techniques. The mean absolute error and root mean squared error averages produced by the % split test option were the lowest (0.245464 and 0.245464) and However, we also found that the average correlation coefficient was 0.596612, greater than the cross validation test choice but lower than the given test set option. Of the experimental results provided in Table 6, the K Star technique has the highest correlation coefficient (0.8901), the lowest root mean squared error (0.2285), and the third lowest mean absolute error (0.1091). The M5P strategy trails K Star in terms of correlation coefficient with a value of 0.8863 and is inferior than KStar in terms of mean absolute error and root mean squared error, both of which are 0.1047 and 0.2322, respectively (01047).

3.1. EXISTING SYSTEM

- In addition to being utilized for short-term weather forecasting, a variety of climate forecasting approaches are also being used in studies on air pollution and the impact of greenhouse gases on global climate change.
- Weather figures that can be predicted, weather that can be predicted using a numerical solution to statistics that govern movement and climatic change.

3.1.1. DISADVANTAGES OF EXISTING SYSTEM

- Our model collects historical weather data that considers several important factors that influence weather change, such as temperature, including both maximum and lowest temperatures atmospheric moisture and humidity, precipitation, the atmosphere's UV Index, and atmospheric mean pressure.
- The acquired dataset is split into sections that the machine learning model can use and sections that it cannot in our suggested model.
- The dataset is then put through a process of data preparation, which involves replacing any missing or incorrect numbers with the mean value or the value that appears most frequently in that field.
- Decision trees, logistic regression, random forests, Xgboost, and K-NN are just a few of the machine learning techniques used to predict the weather. The best model is chosen for the app's implementation.

3.2. ADVANTAGES OF PROPOSED SYSTEM:

- The information from the Weather Channel and our trained models are thoroughly summarized here.

- Highlights how much the usage of training data from surrounding cities improves the visual performance of our models.

4. CONCLUSION

Our model gathers historical weather information that takes into account a number of significant variables that affect weather change, including temperature, both maximum and lowest temperatures, atmospheric moisture and humidity, precipitation, the atmosphere's UV Index, and atmospheric mean pressure. The collected dataset is used to train and test machine learning algorithms including Logistic regression, K-NN, Random forest regression, Decision tree, and Xgboost.In which Xgboost perform well with maximum accuracy. Hence the model with Xgboost is created and implemented with the app to predict the accurate weather.

REFERENCES:

[1] F. R. Leal and W. L. N. Matos, "Short-term lightning prediction in the Amazon region using ground-based weather station data and machine learning techniques," 2022 36th International Conference on Lightning Protection (ICLP), Cape Town, South Africa, 2022, pp. 400-405, doi: 10.1109/ICLP56858.2022.9942500.

[2] Prathyusha, Zakiya, Savya, Tejaswi, N. Alex and S. C C, "A Method for Weather Forecasting Using Machine Learning," 2021 5th Conference on Information and Communication Technology (CICT), Kurnool, India, 2021, pp. 1-6, doi: 10.1109/CICT53865.2020.9672403.

[3] N. Singh, S. Chaturvedi and S. Akhter, "Weather Forecasting Using Machine Learning Algorithm," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 171-174, doi: 10.1109/ICSC45622.2019.8938211.

[4] S. Kothapalli and S. G. Totad, "A real-time weather forecasting and analysis," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India, 2017, pp. 1567-1570, doi: 10.1109/ICPCSI.2017.8391974.

[5] S. Madan, P. Kumar, S. Rawat and T. Choudhury, "Analysis of Weather Prediction using Machine Learning & Big Data," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), Paris, France, 2018, pp. 259-264, doi: 10.1109/ICACCE.2018.8441679.

[6] Shivanshu, P. Nagwanshi and A. Chauhan, "Smart Real Time Weather Forecasting System," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 558-562, doi: 10.1109/ICAC3N53548.2021.9725697.

[7] K. G. Rani, D. C. J. W. Wise, S. S. Begum and S. Nirosha, "Designing A Model for Weather Forecasting Using Machine Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 385-388, doi: 10.1109/ICESC48915.2020.9155571.

[8] D. Mishra and P. Joshi, "A Comprehensive Study on Weather Forecasting using Machine Learning," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-5, doi: 10.1109/ICRITO51393.2021.9596117.

[9] N. L. and M. H.S., "Atmospheric Weather Prediction Using various machine learning Techniques: A Survey," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 422-428, doi: 10.1109/ICCMC.2019.8819643.

[10] A. Omary, A. Wedyan, A. Zghoul, A. Banihani and I. Alsmadi, "An interactive predictive system for weather forecasting," 2012 International Conference on Computer, Information and Telecommunication Systems (CITS), Amman, Jordan, 2012, pp. 1-4, doi: 10.1109/CITS.2012.6220375.

[11] S. E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic and S. Alessandrini, "Machine Learning for Applied Weather Prediction," 2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam, Netherlands, 2018, pp. 276-277, doi: 10.1109/eScience.2018.00047.