JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Weather Prediction Using Machine Learning Algorithms

[1]Lokesh V S, [2]Latha Narayanan Valli, [3]N.Sujatha,[4]Mukul Mech

[1]Student, M.S Data Science, School of Engineering and Applied Sciences,
University at Buffalo, The State University of New York, United States of America.

[2]Vice President, Standard Chartered Global Business Services Sdn Bhd.,
Kuala Lumpur, Malaysia,

[3]Associate Professor, PG and Research Department of Computer Science,
Sri Meenakshi Government Arts College for Women, Madurai, Tamil Nadu, India,

[4] Student, M.Sc., Cyber Security, School of Computer Science,
University of Birmingham, Birmingham, United Kingdom,

Abstract: Predicting the classification of data into a suitable class is a challenging task. It depends on various factors to predict the dependent variables. Since Random Forest and K-nearest neighbor evaluation can be quantified and is simple to use, here proposed a model using Random Forest and K-nearest neighbor to predict events like expected temperature by inputting maximum temperature, minimum temperature, dew point, humidity, and pressure. Which can be used by farmers or by people from all walks of life to make intelligent decisions. This model can be used in machine learning, and further, the proposed model has scope for improvement as more and more relevant attributes can be used in predicting the dependent variable. This system will take this parameter and predict the weather after analyzing the input information. The outcome demonstrated that these algorithmic procedures can be sufficient for weather forecasting. This project is one of the first to compare the performance of K-nearest neighbor and Random forest methods in weather forecasting using machine learning in Python.

IndexTerms **- K Nearest Neighbour, Machine Learning, Random Forest, Weather Forecasting.**

I.INTRODUCTION

CONCEPT DESCRIPTION

MACHINE LEARNING

Weather forecasts are made by collecting quantitative data about the current state of the atmosphere and using a scientific understanding of atmospheric processes to project how the atmosphere will evolve. The chaotic nature of the atmosphere, the massive computational power required to solve the equations that describe the atmosphere, the error involved in measuring the initial conditions, and an incomplete understanding of atmospheric processes mean that forecasts become less accurate as the difference between the current time and the time for which the forecast is being made increases.

Forecasts dependent on maximum temperature, minimum temperature, dew point, humidity, and pressure are important to predict the weather. Utility companies utilize temperature forecasts to assess demand over the coming days. There is a variety of end uses for weather forecasts. Weather warnings are important forecasts because they are used to protect life and property. On an everyday basis, people use weather forecasts to determine what to wear on a given day. Since outdoor activities are severely curtailed by heavy rain, snow, and wind chill, forecasts can be used to plan activities around these events and to plan and survive them. To predict the weather in a very effective way and to help overcome all such problems, we have proposed a weather forecasting model using a machine learning algorithm. One of the biggest advantages of machine learning algorithms is their ability to improve over time. Machine learning technology typically improves efficiency and accuracy. Using various algorithms, it gives us a predicted value that is nearly equal to the actual value.

CLASSIFICATION

Classification is a type of supervised machine learning task where the goal is to predict the class or category of a new observation based on the characteristics or features of that observation. The input data is a set of labeled examples, where each example is a tuple consisting of a set of features and the corresponding label or class.The machine learning algorithm learns to map the input features to the correct class by analyzing he relationships between the input features and the class labels. There are many different algorithms that can be used for classification, including decision trees, logistic regression, support vector machines, and neural networks. Each algorithm has its own strengths and weaknesses, and the choice of algorithm depends on the nature of the problem

being solved and the characteristics of the data being analyzed. It is a powerful tool for automating decision-making processes and can lead to significant improvements in efficiency and accuracy in many applications.

DATA IMPORT AND EXPORT

Python provide several libraries for importing and exporting data in various file format.

- CSV files
- EXCEL files
- JSON files

EASILY AVAILABILITY OF DATA ANALYSIS TECHNIQUES

Python incorporates all the data manipulation, a statistical model that is frequently required by modern researchers. The researcher can easily build a prediction model at no cost. Design and Analysis of Experiments with Python presents a unified treatment of experimental designs and design concepts commonly used in practice. It connects the objectives of research to the type of exploratory data analysis, feature engineering, and visualization of the data, shows how to perform the proper analysis of the data, and illustrates the interpretation of results.

BASIC STATISTICS

The basics of statistics include the measure of central tendency and the measure of dispersion. The central tendencies are mean, median, and mode, and the dispersion comprises variance and standard deviation. The mean is the average of the observations. The median is the central value when observations are arranged in order.

ADVANCED STATISTICS

This section describes more advanced statistical methods. This includes the discovery and exploration of complex multivariate relationships among variables. Links to appropriate graphical methods are also provided throughout.

DATA VISUALIZATION

Python offers several libraries for data visualization, including Matplotlib, Seaborn, Plotly, Bokeh, and more. Matplotlib provides a wide range of tools for creating different types of charts, graphs, and other visualizations. Some of the most common types of plots created with Matplotlib include line charts, scatter plots, bar charts, histograms, and heatmaps. Seaborn is provides higher-level functions for creating more complex visualizations, such as heatmaps with annotated values, facet grids, and violin plots. Plotly is a library for creating interactive visualizations, such as interactive line charts, scatter plots, and 3D plots. Plotly provides tools for creating interactive dashboards and web applications that can be shared online. Bokeh is another library for creating interactive visualizations. It provides a range of tools for creating interactive line charts, scatter plots, and other types of visualizations. Bokeh is designed to work well with large datasets and supports streaming data. In addition to these libraries, Python also offers range of other visualization tools and libraries, such as ggplot.

1.2 PROJECT DESCRIPTION

This is an important application used in the analysis and prediction of the weather's temperature. This is used to analyze previous weather data sets, and the system will calculate weather based on this data. This analysis is used in many countries for predicting the weather. The proposed works deal with a k-nearest neighbor and random forest; these are two methods in statistical techniques. This analysis is suitable for predicting the weather. Regression and classification can also be used in the prediction. The evaluation metric (accuracy, f1 score, precision, recall) that measures a model’s accuracy.

II. SYSTEM ANALYSIS

EXISTING SYSTEM

On an everyday basis, people use weather forecasts to determine what to wear on a given day. Since outdoor activities are severely curtailed by heavy rain, snow, and the wind chill, forecasts can be used to plan activities around these events and to plan ahead and survive them. In order to predict weather in a very effective way and to help overcome all such problems, we have proposed a weather forecasting model using a machine learning algorithm. The existing system uses k- nearest neighbor and random forests for predicting temperature using machine learning. However, the accuracy, precision, and recall are not better by model. To address these problems, the present work has some preprocessing steps for better accuracy.

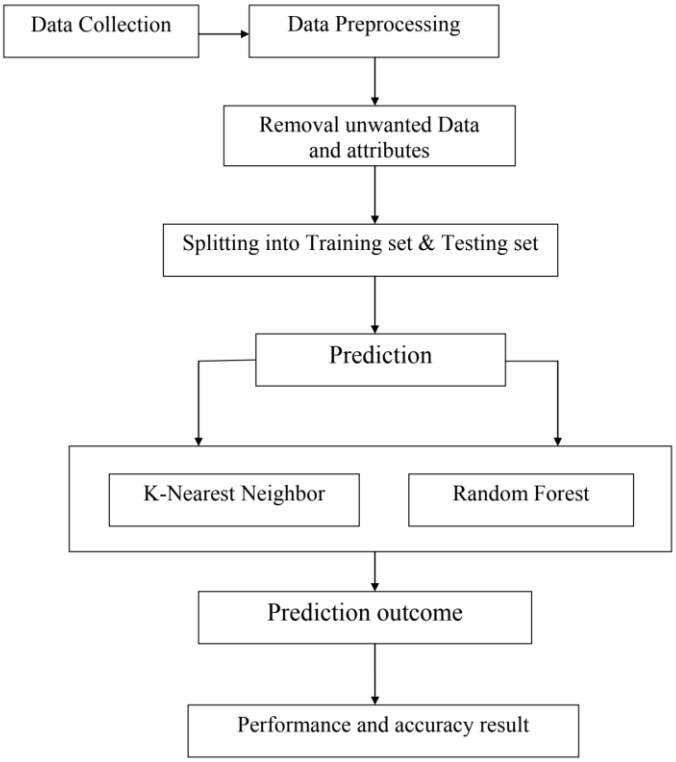
PROPOSED SYSTEM

The user will enter the current maximum temperature, minimum temperature, humidity, dew point, and pressure. The system will take these parameters and predict the weather based on previous data in the database. Add previous weather data to the database so that the system can calculate weather based on this data. The weather forecasting system takes parameters such as maximum temperature, minimum temperature, humidity, dew point, and pressure and forecasts the weather based on the previous record at Madurai from January 1981 to February 2023. The evaluation matrix can be analyzed to process data in the K-Nearest Neighbor and Random Forest. These are some of the specific processes that can be analyzed.

ADVANTAGES

- It can easily predict the weather.
- The predicted temperature will have better accuracy.
- The error rate is reduced.

III. METHODOLOGY



COLLECTING DATA

Machines initially learn from the data that you give them. It is of the utmost importance to collect reliable data so that your machine learning model can find the correct patterns. The quality of the data that you feed to the machine will determine how accurate your model is. If you have incorrect or outdated data, you will have wrong outcomes or predictions that are not relevant. Good data is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories and classes present.

PREPARING THE DATA

Putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process. Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. You might even have to restructure the dataset and change the rows and columns or index of rows and columns. Visualize the data to understand how it is structured and understand the relationship between various variables and classes present. Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

CHOOSING A MODEL

A machine learning model determines the output you get after running a machinelearning algorithm on the collected data.. It's engineers have developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, you also have to see if your model is suited for numerical or categorical data and choose accordingly.

TRAINING THE MODEL

Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

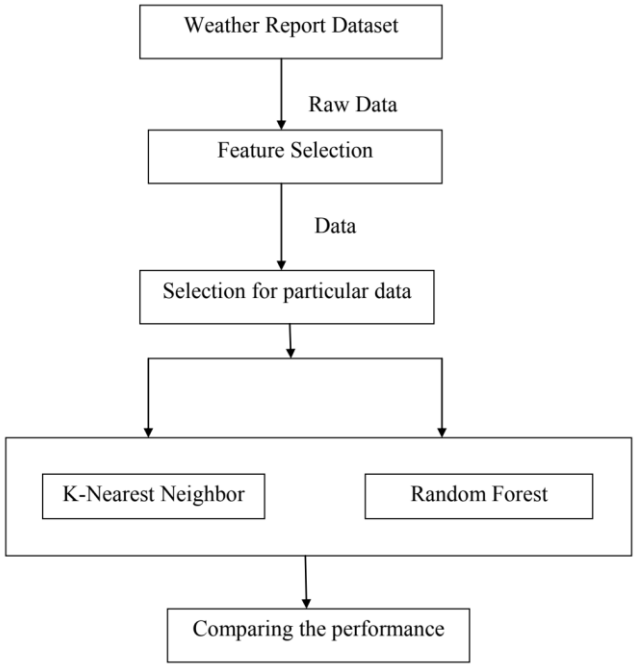
EVALUATING THE MODEL

After training your model, you have to check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that you split our data into earlier. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data. This will give you disproportionately high accuracy. When used on testing data, you get an accurate measure of how your model will perform and its speed.

MAKING PREDICTIONS
In the end, you can use your model on unseen data to make predictions accurately.

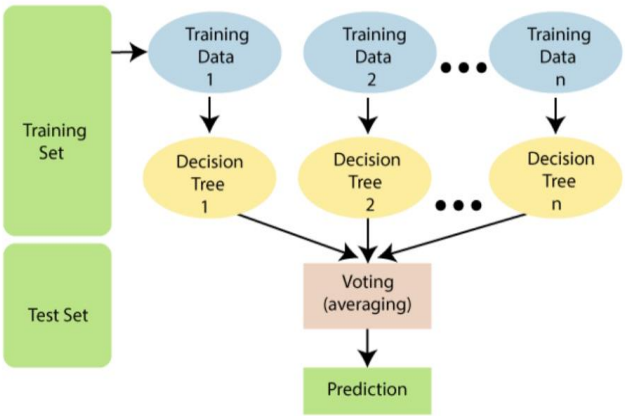
IV. SYSTEM DESIGN

DATA FLOW DIAGRAM



MODULE DESIGN

RANDOM FOREST
Random Forest is an algorithm for classification and regression. Summarily, it is acollection of decision tree classifiers. The random forest has an advantage over the decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each tree and then a decision tree is built; each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances, training is extremely fast in the random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting. Random Forest is a collection of decision tree classifiers. Random Forest works on the same weak learners. The random forest has an advantage over the decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances, training is extremely fast in the random forest and because each tree is trained independently of the others. It combines the output of multiple decision trees and then finally come up with its output. Random Forest works on the same principle as Decision Trees. Random forest ranks the importance of variables in a regression or classification problem in a natural way that can be done by Random Forest.



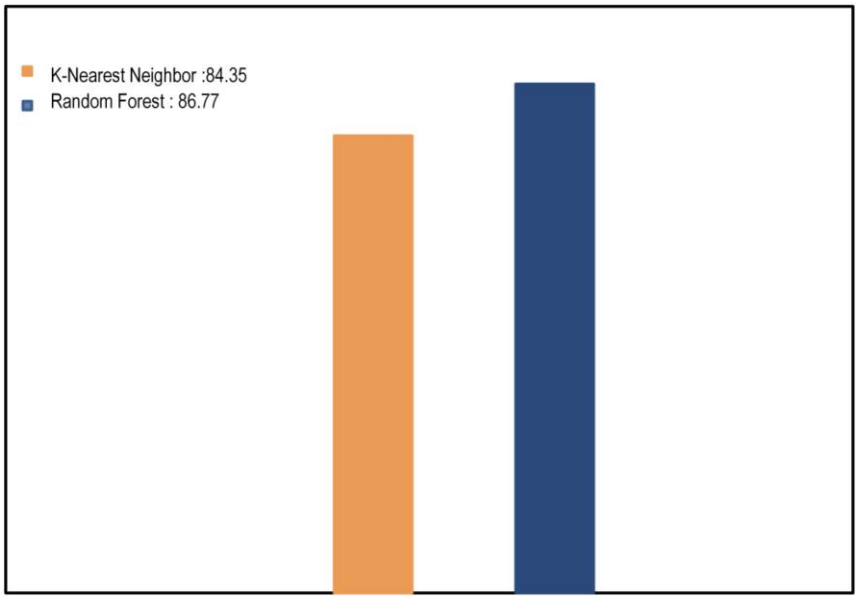
K-NEAREST NEIGHBORS

The k-nearest neighbors (KNN) algorithm is a non-parametric machine learning algorithm that is commonly used for classification and regression tasks. It is a simple algorithm that can be easily understood and implemented. In this essay, I will explain the KNN algorithm, its working, and its various applications.KNN Algorithm Working: The KNN algorithm works by finding theK nearest neighbors of a new data point based on some distance metric (usually Euclidean distance) and assigning the class of the majority of the K neighbors to the new data point. In the case of regression tasks, the algorithm returns the average of the K nearest neighbors as the predicted value. The KNN algorithm is a lazy algorithm because it does not make any assumptions about the distribution of

the data and does not perform any training. Instead, it stores the training data and uses it to classify new data points. Therefore, it is computationally expensive for large datasets because it has to calculate the distance between the new data point and all the training data points. However, the KNN algorithm has some limitations. One of the main limitations is that it can be sensitive to the choice of distance metric and the value of K. The choice of K is important because if K is too small, the algorithm may be sensitive to noise and outliers, while if K is too large, the algorithm may oversimplify the problem. In conclusion, the KNN algorithm is a simple and effective machine learning algorithm that is widely used for classification and regression tasks. It can be used with any type of data and is easy to implement. However, the choice of distance metric and the value of K can have a significant impact on the performance of the algorithm. Therefore, it is important to carefully select these parameters based on the problem at hand.

V. RESULTS

PERFORMANCE MEASURES	K-Nearest Neighbor	Random Forest
Accuracy	84.35	86.77
Precision	59.38	69.05
F1_ Score	68.67	62.15
Recall	57.48	56.53
Misclassification Error	16.08	13.22



VI. CONCLUSION

The weather prediction done using the K-nearest neighbor algorithm and the Random forest algorithm is very essential for improving the future performance of the people. For predicting the weather, the K-nearest neighbor algorithm and the Random forest algorithm were applied to the datasets of the weather. We made a model to predict the weather using some selected input variables collected from NASA (.gov). This model yields performance metrics such as accuracy, precision, F-measure, r2 score, and error rate. The best accuracy score is 84.35, attained by Random Forest on these data sets. The results of our experiments show that Random Forest is a promising algorithm. This project couldn't reach our goal of 100% accuracy in weather prediction. This model can be further improved with the addition of more algorithms, such as Naive Bayesian and Support Vector Machine. Hence, to know the weather scenario with high accuracy considering every factor that affects the weather scenario, this model is created.

REERENCES

[1] Maqsood, I., M. R. Khan, and A. Abraham. "An ensemble of neural networks for weather forecasting." Neural Computing & Applications.

[2] Ghosh et al., "Weather Data Mining using Artificial Neural Network," 2011 IEEE Recent Advances in Intelligent Computational Systems.

[3] Singh, Nitin, Saurabh Chaturvedi, and Shamim Akhter. "Weather forecasting using machine learning algorithm." 2019 International Conference on Signal Processing and Communication (ICSC). IEEE, 2019.

[4] Holmstrom, Mark, Dylan Liu, and Christopher Vo. "Machine learning applied to weather forecasting." Meteorol. Appl 10 (2016): 1-5.

[5] Jakaria, A. H. M., Md Mosharaf Hossain, and Mohammad Ashiqur Rahman. "Smart weather forecasting using machine learning: a case study in tennessee." arXiv preprint arXiv:2008.10789 (2020).

[6] Singh, Siddharth, et al. "Weather forecasting using machine learning techniques." Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE). 2019.
[7] Dadhich, Shruti, et al. "Machine learning for weather forecasting." Machine Learning for Sustainable Development 10 (2021): 9783110702514-010.
[8] Dadhich, Shruti, et al. "Machine learning for weather forecasting." Machine Learning for Sustainable Development 10 (2021): 9783110702514-010.
[9] Bhawsar, Mihir, Vandan Tewari, and Preeti Khare. "A survey of weather forecasting based on machine learning and deep learning techniques." International Journal of Emerging Trends in Engineering Research 9.7 (2021).
[10] Purwandari, Kartika, et al. "Multi-class weather forecasting from twitter using machine learning aprroaches." Procedia Computer Science 179 (2021): 47-54.
[11] Valli, Latha Narayanan, N. Sujatha, and D. Divya. "A Novel Approach for Credit Card Fraud Detection Using LR Method-Comparative Studies." *Eduvest-Journal of Universal Studies* 2, no. 12 (2022): 2611-2614.