WEATHER PREDICTION USING MACHINE LEARNING ALGORITHMS

Aiswarya Shaji
Department of Computer Science and IT
Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India
Kochi, India
aiswaryashaji.98@gmail.com

Amritha A.R
Department of Computer Science and IT
Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India
Kochi, India
amrithavihar123@gmail.com

Rajalakshmi V.R
Department of Computer Science and IT
Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India
Kochi, India
rajiprithviraj@gmail.com

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) | 978-1-6654-7258-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCSP53532.2022.9862337

Abstract— ***Weather forecasts have grown increasingly significant in recent years since they can save us time, money, property, or even our lives. Despite the fact that India has a large number of weather stations, they are mainly located in inhabited regions such as cities, suburbs, or towns. This makes weather forecasting in isolated regions more imprecise, which can be inconvenient for individuals such as farmers who rely largely on weather reports in their daily work. In this paper, we are predicting the weather by analyzing features like temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, cloud cover with Random Forest, Decision Tree, MLP classifier, Linear regression, and Gaussian naive Bayes are examples of machine learning methods. Based on the results obtained a comparative study is done concerning the accuracy.***

Keywords— ***Machine Learning; Prediction; Weather Forecast; Classification.***

I. INTRODUCTION

Pattern extraction from data sets was done manually in the past. The collecting, manipulation, and storage of data sets have expanded tremendously in the modern computer era. Pattern recognition becomes extremely sophisticated as a result of this. The computerized application of specialized algorithms to detect specific patterns from massive data sets is known as data mining. Machine learning is a type of data mining in which a model is created by learning concepts with a computer. This model learns by training and testing for the supplied large data sets, and it predicts future data instances using the learning principle. Modeling refers to the practice of creating a classification model to anticipate an outcome Classification is the data mining process of creating a model based on one or more attributes of categorical variables to predict the value of a target categorical variable. Based on the training set and class labels, this can classify the given data. Since weather systems can extend a significant distance in all directions over time, the weather of one locality can have a massive effect on the weather of others. In this paper, we offer a method for predicting weather conditions by combining historical weather data from nearby cities with data from a single city. These data are combined and used to build simple and basic machine learning models that can correctly foretell weather conditions for the next several days. These simple models may be run on low-cost, constrained computational systems and provide quick and accurate forecasts that we can utilise in our everyday lives.

(1) One of the paper's major advances is that machine learning may be used to anticipate weather conditions over short periods of time using less resource-intensive equipment.

(2) A thorough assessment of the suggested technique, as well as a comparison of numerous machine learning models for forecasting future weather conditions.

II. LITERATURE REVIEW

Several researchers compared weather predictions using different methods and a few of them are mentioned below. In,[19] the author introduced a methodology of using the R tool, and a comparison of Decision Tree and Random Forest was conducted. The algorithm is executed using Rattle, an R GUI for analysis of data mining algorithms, that will predict whether it rains the next day or not. The redistribution error rate is compared; and found that a random forest has less error value than a decision tree since the decision tree is an ensemble of trees. [8] The author presented the approaches for improving the accuracy of random forest classifiers. To improve the accuracy, a weighted hybrid decision tree model is used. Disjoint partitions of the training dataset and ranking of training bootstrap samples are two more methods used. Both of these are leading effective learning and classification using the Random Forest classifier.[18] The objective of their project was to build a desktop application that predicts weather automatically and extracts data that is needed to capture global data. The author collected a dataset on the actual weather of Nashville city from wunderground.com. After pre-processing, some features with empty or invalid data are eliminated. After splitting the dataset into training and testing data. The first phase of results brings the accuracy of prediction by adding more features. Since the predicted results are continuous numerical values, the author used Random Forest Regression (RFR) which is considered a superior regressor. Several machine learning strategies with the RFR process are also considered in building the model. The model performance is measured using Mean Squared Error. And concluded that the system operates at a high level of efficiency without malfunctioning. As it is an application, the author points out that it will not consume much RAM and

978-1-6654-7258-6/22/$31.00 ©2022 IEEE

phone memory. [11] Several clustering algorithms were used in the research to propose a model for weather forecasting. K-means clustering, hierarchical clustering, density-based clustering, filtered clustering, and farthest first clustering are all examples of clustering techniques. This model takes the findings of the investigation and selects the best from them using various clustering algorithms. Formatted date, Summary, Precipitation type, Temperature, Humidity, Wind speed, Cloud coverage pressure, and Daily summary are only a few of the elements in the dataset. According to the findings, K-means and Filtered clustering are two successful methods, with K-means taking 0.01 seconds less time to develop the model. [5] Two machine learning algorithms were utilized by the author. There are two types of functional regression: linear regression and a variant of functional regression. The algorithms were trained using weather data from Stanford. The weather data from the previous two days was fed into this model, which included maximum and minimum temperatures, mean humidity, mean air pressure, and weather categorization for each day. Over the next seven days, the output was the lowest and highest temperatures. The conclusion for professional weather-predicting service contained RMS error for Linear Regression and a version on Functional Regression with RMS error. Professional weather-predicting services outperformed both linear regression and functional regression. [13] This paper provides an application support vector regression for atmospheric temperature prediction. Support Vector Machine (SVM) performance was compared with MLP, which results in the performance of SVM being better than MLP which is when trained with backpropagation. SVM can replace some Neural Network-based models for weather prediction if the correct parameters are used, according to the study's conclusion. [17] The author of this research proposed a Raspberry Pi-based weather forecasting program that uses data directly from humidity, temperature, and pressure sensors to forecast rainfall for the current day. Because the backend incorporates a machine learning model trained using Random Forest classification, it can forecast whether or not it will rain. [10] The author completed a work that included a comparison of classifiers such as Random Forest and Random Tree in the context of a microarray dataset. The Weka tool was utilized. The dataset was obtained from the machine learning repository at UC Irvine. To record classification accuracy, the author utilized 10-fold cross-validation as the test mode. Finding the classification performance of the classifier in the dataset and finding classification performance using an attribute filter were the only two phases in the experimental design. [3] The author of this paper compares the performance of three different risk classifiers: Random forest, REP tree, and J48 classifiers. An open-source machine learning application is used to evaluate the performance of tree-based classifiers. The performance is evaluated using the training set and cross-validation methods. For credit risk prediction, the random forest classifier outperforms the REP and J48 classifiers when many criteria such as classification accuracy, mean absolute error, and time spent to develop the model are taken into account. [4] This is a weather forecasting model that uses the combined effect of important weather parameters to produce forecasts. They used a gradient tree-based learner to do temporal analysis using short and long-term features. Future work will entail including weather forecasts for extended amounts of time into the feature. [6] The author created a machine learning and XML search model. The news resources available online, such as the news in the Google repository, were used as input to this model. This gathered information was then transformed into a term frequency matrix for future study. On a dataset containing 649 news items, three classifiers were used: REP Tree, Simple CART, and Random Tree. After processing, the output is presented in the form of a confusion matrix. The paper concludes by stating that the performance of the Indian news repository utilizing Random Tree yielded 100% TP and 0% FP. [12] In this article, a hybrid methodology for monitoring bank account credit risk is proposed. The approach presented in this survey can be a good way for banking institutions to identify the level of credit risk for their customers concerning their financial situation. As a result, future research could compare the performance of the provided hybrid model with those of existing credit evaluation methods. [9] The author proposed a simple technique based on McNemar's direct test to limit the number of classifiers merged in multiple classifier systems in this study. The method compares a batch of MCS with a lower proportion of classifiers against a batch of MCS with a larger number of classifiers. They claim that if the prediction sets for the McNemar test do not differ, the smallest number of classifiers is required to achieve the same level of certainty for the McNemar test. Experiments on four different MCSs applied to the C4.5 decision tree, as well as cross-validations on five large benchmark databases, showed that a small number of classifiers can be chosen a priori. [1] The authors are looking at sample data acquired from a college's Moodle database after data for course enrolment was collected from students. The paper then examines the five categorization algorithms to see which one is the best for proposing a course to students based on their preferences. The categorization methods employed are ADTree, Simple Cart, J48, ZeroR, Nave Bays, and Random Forest. The authors also discovered that the ADTree classification technique is more effective. [16] The researchers used WEKA, an intriguing data mining tool, to undertake tests to see which algorithm best anticipates if an email is spam or not. On the ground of varying proportions of successfully predicted examples, four algorithms were compared: ID3, J48, Simple CART, and ADTree. All four belong under the categorization methods of data mining, which map data points to establish a bond between a dependent (output) variable and an independent (input) variable. For the spam email datasets comprising 58 attributes with each 4601 attributes, the J48 classifier showed the best performance in terms of classification accuracy. Furthermore, Simple CART generated identical results to J48, with only minor variations. As compared to the previous two, the ADTree and ID3 classifiers were less precise. When classification accuracy is crucial in a spam email application, the J48 classification technique should be favoured above Simple CART, ADTree, and ID3 classifiers. They suggested that future study may include an extension of the WEKA simulation that compares the classification accuracy performance of the proposed methodologies by keeping the number of instances in a dataset the same but reducing the number of attributes. [15] The authors compared ten supervised learning methods: SVMs, neural nets, logistic regression, naive Bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps in a large-scale empirical comparison. They also look at how applying Platt Scaling and Isotonic Regression to calibrate the model's influences their performance. One of the most important aspects of their work was the requirement for a variety of performance indicators to evaluate the learning techniques. They concluded that learning approaches like as boosting, random forests, bagging, and SVMs perform remarkably well. [2] In this research, the researchers identified three data mining techniques: Nave Bayes, back-propagated neural networks, and the C4.5 decision tree algorithms. They used these algorithms to estimate the survivability rate of the SEER breast cancer data set, and they determined that these three categorization systems were the most effective at forecasting cancer survival rates.Three data mining strategies are reviewed for accuracy, with the goal of having high accuracy in addition to high precision and recall metrics. The C4.5 algorithm is more accurate. [20] The authors of this paper presented a novel use of NN techniques in severe numerical modeling of the environment. They created an HGCM, a complicated hybrid environmental numerical model that combines deterministic modeling and machine learning techniques in a synergetic way. This method uses neural networks as a statistical or machine learning tool to create highly accurate and quick simulations of the most time-consuming deterministic model components. Other complicated numerical models utilized outside of the realm of environmental modeling applications, such as advanced models in computational physics, chemistry, biology, and so on, can benefit from the established hybrid modeling paradigm and related NN emulation technology. [7] The paper analyses the performance of datasets using several classification techniques, with accuracy and execution time as an evaluation criteria. The performance of classification techniques is observed to vary with diverse datasets. Dataset, Number of instances and attributes, and Type of attributes are all factors that influence the classifier's performance. Other data sets used in the comparison yielded excellent performance with J48 and NaiveBayesUpdatable. [14] Anomaly Detection System (ADS) monitors a system's behavior and marks significant departures from expected behavior as anomalies. Anomaly detection is used to detect computer network assaults, malicious actions in computer systems, and Web-based system misuses. The paper discusses anomalies and many supervised and unsupervised anomaly detection strategies, as well as individual K-means and Id3 Decision Tree usage, comparative research, and the proposed system's combined approach. To summarise, the training instance is first partitioned into k different clusters using the k-Means technique. The ID3 decision tree in each cluster learns the cluster's sub-classifies and divides the decision space into classification sections.

III. ALGORITHMS TAKEN FOR COMPARISON

A. Random Forest Algorithm

The supervised learning method is used by Random Forest, a well-known machine learning algorithm. It can be used for both classification and regression issues in machine learning. According to Wikipedia, a Random Forest is a classifier that averages multi-criteria trees from multiple subsets of a dataset to improve the dataset's projected accuracy. The random forest, rather than relying on a single decision tree, incorporates inputs from each tree and forecasts the eventual output based on the majority votes of projections.

B. Decision Tree Algorithm

A sort of predictive modeling known as decision tree analysis can be used to a wide range of scenarios. An algorithmic technique can also be used to generate decision trees, which can segment data in a variety of ways based on particular parameters. Decision trees are the most powerful algorithm in the realm of supervised algorithms.

C. Gaussian Naïve Bayes

Continuous data is supported by the Gaussian Naive Bayes version, which follows the Gaussian normal distribution. The Naive Bayes classification methods, which are supervised machine learning classification algorithms, are based on the Bayes theorem. It's a simple categorization method with a lot of power. They're advantageous when the inputs' dimensionality is high. The Naive Bayes Classifier can also handle complex classification issues.

D. MLP Classifier

Backpropagation is used to train a multi-layer perceptron (MLP) technique in the MLP Classifier. A multilayer perceptron can have more than one linear layer (combinations of neurons). The input layer receives our data, and the output layer receives our output. By increasing the number of hidden layers, we may build the model as complex as we like.

E. Linear regression

In machine learning, linear regression lets you uncover patterns and relationships in data so you can make an informed choice or prediction. In machine learning, linear regression models a linear connection between data features. A linear relationship across continuous variables is modeled by linear regression. We examine two variables, one of which is a predictor and the other of which is a response.

IV. EXPERIMENTAL TECHNOLOGY

A. Dataset

For our proposed system weather data is collected from Kaggle.com and processed using python. We are considering the attributes and their summary for the prediction. We focused on eight factors in the dataset and they are temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, cloud cover, and pressure. There is a total of 27 summaries and they are: Partly cloudy, Mostly cloudy, Overcast, Foggy, Breezy and mostly cloudy, breezy and partly cloudy, Humid and mostly cloudy, Humid and partly cloudy, Clear, Breezy and overcast, Light rain, Breezy and foggy, Dry and partly cloudy, Windy and Foggy, windy, Drizzle, Dry, Windy and partly cloudy, Breezy, Humid and overcast, Windy and overcast, Dry and mostly cloudy, Windy and mostly cloudy, Rain, Dangerously windy and partly cloudy, Breezy and dry, Windy and dry.

B. Pre-processing

The pre-processing of data is the first phase in the process's commencement. After acquiring the dataset, the first step to do is pre-processing. Because the dataset will be comprised of data gathered from multiple sources, that can be incomplete, inconsistent, or inaccurate. Thus, pre-processing has a great role. Once the dataset is ready it must be put in CSV file format. As we use python, it has many libraries for pre-processing. Two libraries used here are NumPy and Pandas. NumPy is a Python library that allows you to perform scientific calculations. Using this we can also add large multidimensional arrays and matrices to our codes. Pandas is a data manipulation library written in Python that is open-source. It's a powerful platform for importing and managing datasets. During data pre-processing, it's vital to find and handle missing values correctly. We can remove a feature with a null value for a specific row or a column with more than 75% of the entries missing. Since we only require numbers, another option utilized is encoding category data in the equation, which can generate some complications. As a result, we'll turn it into numerical values. The next stage in data pre-processing is to split the dataset. The data should be split up into two: training and testing.

C. Proposed Model

This is an analysis paper using the weather dataset. We have a hybrid model, that has four analysis parts consisting of two phases of machine learning algorithms each. 70% of the original dataset is split into training data and 30% for test data.

i. The first analysis part consists of the Random Forest algorithm and Linear Regression.
ii. The second analysis part consists of a Decision Tree and Linear Regression.
iii. The third analysis part consists of Random Forest and MLP Classifier.
iv. The fourth analysis part consists of Gaussian Naïve and MLP Classifiers.

Initially, the feature importance score of each feature is calculated using the machine learning algorithms coming in the first phase of every analysis part (i.e., Random forest, Decision tree, and Gaussian Naïve respectively). The strategies that determine a score for all of the input features for a particular model are referred to as feature importance. The 'importance' of each characteristic is simply represented by the score. We are giving an input threshold frequency concerning the importance score of each feature in the first phase. The only features which are greater than the given input threshold frequency will be taken. These selected features will display the actual weather and final predicted weather which is received after training with the respective algorithms.

In the second phase of every analysis part, a confusion matrix is created using the machine learning algorithms (Linear regression and MLP classifiers) with the test dataset and final predicted value. It shows only the selected labels which are filtered from the first phase which gives us the accuracy of each analysis part. Comparing the four analysis part, Part 3 and 4 gives a better accuracy with 51% that is shown under Random forest, Gaussian Naïve, and MLP Classifiers.

D. Accuracy Table

TABLE 1 ACCURACY TABLE

Analysis part	Algorithms	Accuracy
1	Random Forest + Linear Regression	0.2354
2	Decision Tree + Linear Regression	0.2163
3	Random Forest + MLP Classifier	0.5138
4	Gaussian Naïve + MLP Classifier	0.5103

V. CONCLUSION AND FUTURE ENHANCEMENT

We performed hybrid comparative research in which we used machine learning approaches to offer weather forecasts in this publication. Intelligent models can be created using machine learning technologies that are far simpler than traditional physical models. They need limited resources and maybe run on nearly any computer, including mobile devices. The Random Forest and Gaussian Nave with MLP Classifier models predict weather features more correctly than the other hybrid models described here, according to our evaluation results. To forecast the weather in a specific place, we also analyze past data from surrounding locations. We show that focusing primarily on the location where weather forecasting is done is ineffective. The accuracy of a dataset can be improved by focusing on only two or three features.

VI. GRAPHICAL REPRESENTATION

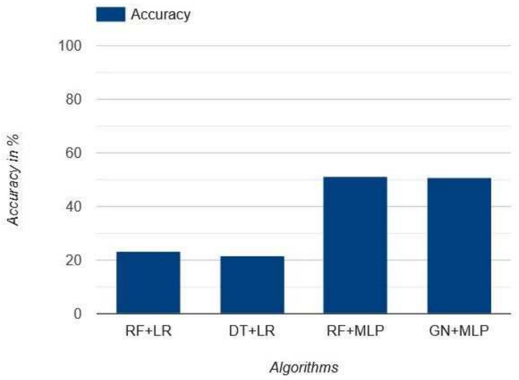

Fig. 1. Graphical Representation of accuracy obtained

VII. REFERENCES

[1] Aher, Sunita B., and L. M. R. J. Lobo. "Comparative study of classification algorithms." *International Journal of Information Technology* 5.2 (2012): 239-243.

[2] Bellaachia, Abdelghani, and Erhan Guven. "Predicting breast cancer survivability using data mining techniques." *Age* 58.13 (2006): 10-110.

[3] Devasena, C. Lakshmi. "Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction." *International Journal of Computer Applications* (2014): 0975-8887.

[4] Grover, Aditya; Kapoor, Ashish; Horvitz, Eric (2015). [ACM Press the 21th ACM SIGKDD International Conference - Sydney, NSW, Australia (2015.08.10-2015.08.13)] Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15 - A Deep Hybrid Model for Weather Forecasting. , (), 379–386.

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

[6] Kalmegh, Sushilkumar. "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news." International Journal of Innovative Science, Engineering & Technology 2.2 (2015): 438-446.

[7] Kharche, Deepali, K. Rajeswari, and Deepa Abin. "Comparison of different datasets using various classification techniques with weka." International Journal of Computer Science and Mobile Computing 3.4 (2014): 389-393.

[8] Kulkarni, Vrushali Y. and Pradeep K. Sinha. “Effective Learning and Classification using Random Forest Algorithm.” (2014).

[9] Latinne, P., Debeir, O., & Decaestecker, C. (2001). Limiting the Number of Trees in Random Forests. Lecture Notes in Computer Science, 178–187

[10] Mishra, Ajay Kumar, and Bikram Kesari Ratha. "Study of random tree and random forest data mining algorithms for microarray data analysis." *International Journal on Advanced Electrical and Computer Engineering* 3.4 (2016): 5-7.

[11] Nalluri, Sravani; Ramasubbareddy, Somula; Kannayaram, G (2019). Weather Prediction Using Clustering Strategies in Machine Learning. Journal of Computational and Theoretical Nanoscience, 16(5), 1977–1981

[12] Pourdarab, Sanaz, Ahmad Nadali, and Hamid Eslami Nosratabadi. "A hybrid method for credit risk assessment of bank customers." *International Journal of Trade, Economics and Finance* 2.2 (2011): 125-131.

[13] Radhika, Y., and M. Shashi. "Atmospheric temperature prediction using support vector machines." *International journal of computer theory and engineering* 1.1 (2009): 55.

[14] Rao, K. Hanumantha, et al. "Implementation of anomaly detection technique using machine learning algorithms." International journal of computer science and telecommunications 2.3 (2011): 25-31.

[15] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning. Association for Computing Machinery, New York, NY, USA, 161–168.

[16] Sharma, Aman Kumar, and Suruchi Sahni. "A comparative study of classification algorithms for spam email data analysis." *International Journal on Computer Science and Engineering* 3.5 (2011): 1890-1895.

[17] Singh, Nitin; Chaturvedi, Saurabh; Akhter, Shamim (2019). [IEEE 2019 International Conference on Signal Processing and Communication (ICSC) - NOIDA, India (2019.3.7-2019.3.9)] 2019 International Conference on Signal Processing and Communication (ICSC) - Weather Forecasting Using Machine Learning Algorithm.

[18] Singh, Shashank & Faraz, Ahmed & Nagrami, & Pillai, Aditya. (2020). WEATHER PREDICTION BY USING MACHINE LEARNING.

[19] T R, Prajwala. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. IJARCCE. 196-199. 10.17148/IJARCCE.2015.4142.

[20] Vladimir M. Krasnopolsky; Michael S. Fox-Rabinovitz (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. 19(2), 122–134.