IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11657

Rainfall prediction using machine learning

S. Kannan[1], R. Dinesh[2], A. Sengodi[3]

Student, B.E. Computer Science and Engineering, Anand institute of higher technology, Chennai, India[1,2]

Assistant Professor, CSE, Anand institute of higher technology, Chennai, India[3]

Abstract: Drastic change in climatic conditions is a very big and challenging task for people around the globe. Most of the biological, constructional, transportation and agricultural sectors get affected due to uneven weather conditions, i.e. flood, rainfall, drought, etc. As part of the weather system, rainfall being most prominent phenomena, its rate is treated as one of the most important variables. Meteorological scientists try to identify the parameters of the atmosphere such as temperature, sunshine, cloudiness and humidity of the earth by applying conventional techniques and developing a prediction model. These days, Machine Learning (ML) techniques are more evolving and give more accurate results than the traditional approaches. ML is a subset of artificial intelligence (AI) which is used in this paper for predicting the next day's rainfall from the past 10 year's weather dataset . This paper presents the ML classifiers such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF),and Extreme Gradient Boost (XGB) to predict the rainfall of the next day. This is followed by sequential stages of data visualization, training, testing, modelling, and cross-validation. The evaluation metrics like Area under the Receiver Operating Characteristic (AUROC) curve, recall, accuracy, precision, and Cohen kappa are used to check the performance of ML algorithms.

I. INTRODUCTION

The prediction of weather is an indispensable requirement because it plays an essential role in agriculture, transportation industrial, commercial, and tourism sectors etc. Flood, precipitation, evapotranspiration, thunderstorm, hurricane, typhoons, lightning, and fog are the most salient atmospheric events for the sustainable weather forecast system. Changing climatic conditions each day has motivated the researchers and scientists to predict the next day's rainfall, which has a major impact on society As a rapidly changing behavior of weather patterns worldwide many researchers and scientists are trying to find a different way to forecast by using features such as temperature, humidity, pressure, wind speed, wind direction. ML algorithms use computational techniques to learn information from past data and extract useful data to improve performance . This study aims to address the rain uncertainty problem to predict whether the rain will be happening the next day or not based on Australian rainfall dataset obtained from Kaggle by using ML techniques by creating a model for accurate predictions. This work explores and compares the performance of different ML methods, including LR, DT, RF, XGB to predict weather uncertainty.

1.1 OBJECTIVE

The objective of this work is to carry out an exploratory data analysis on the use of machine learning algorithms to model the phenomenon of rain, taking as an example a dataset of precipitation measurements and atmospheric conditions, as well as other characteristics of the main cities . In addition, using this data, some of the most important machine learning algorithms were applied to evaluate its usability and efficiency

1.2 SCOPE

Using this approach, it is possible to predict rainfall with ML after running the application in a runtime environment. ML algorithms such as , Logistics Regression(LR) ,Decision Tree(DT) ,Random Forest(RF) and Extreme Gradient Boost(XGB) can be applied to all these approaches.

II. ANALYSIS

2.1 SYSTEM ANALYSIS

System Analysis is a combined process dissection the system responsibilities that are based on problem domain characteristics and user requirement.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11657

2.1.1 Problem Definition

Rainfall forecasting is very important because heavy and irregular rainfall can have many impacts like the destruction of crops and farms, damage of property so a better forecasting model is required for an early warning that can reduce the risks to life and property and also helps to manage the agricultural farms in a better way. Heavy rainfall is a cause for natural disasters like floods and drought that square measure encountered by individuals across the world each year. Many models are developed to evaluate the rainfall and for predicting the likeliness of rain. These models are based on both supervised and unsupervised machine learning algorithms. Taking into consideration of overall rainfall will not help us to know if it rains in specific conditions. Accuracy is the major concern in machine learning. We are going to understand the data and then train the model accordingly to predict whether if it rains under given conditions or not.

2.1.2 Existing System

In previous years, numerous methods were often used to predict the weather uncertainty such as Multiple Linear Regression, Naïve Bayes, Support Vector Machine, artificial neural network, and K-nearest neighbor models Since the ML model can do feature extraction as well as feature selection from the previous data and forecast the future outcomes depending on the weather parameters. Some of the published papers based on ML have been examined and summarized as follows. The variety of ML methods for classification tasks has been studied over the years related to the efficiency, accuracy, and performance of the algorithms while dealing with a huge number of predictor features designed a forecast model based on multiple linear regression techniques to collect three months' weather data and got the 52% accuracy.

2.1.3 Proposed system

Machine-learning algorithms used for rainfall prediction include linear regression, multiple linear regression, polynomial linear regression, logistic linear regression and rule-based methods .it can produce meaningful results for larger datasets. The primary goal of this research is to forecast rainfall using six basic rainfall parameters of maximum temperature, minimum temperature, relative humidity, solar radiation, wind speed and precipitation. The predicts rainfall for the approach which is more accurate. This system first compares both the process and then accordingly gives result with the best algorithm. Steps associated with the proposed system are input of data, preprocess of data, splitting of data, training of the algorithm, testing of the dataset, comparing both the algorithm, giving the best algorithm, prediction with the more accurate algorithm and result at the end

III.MODULES

3.1 Dataset collection
3.2 Data Preprocessing
3.3 Exploring Data Analysis
3.4 Over sampling

3.1 Dataset collection

Based on everyday observations and data availability of different weather stations, we have used the binary classification dataset collected by the Kaggle competitive website Platform . It contains the dataset (rows*columns) consisting of 145460 records and 23 features. Ten years of data for weather stations over the period from 2007 to 2017 is obtained for the prediction of target variable. The data set consists of 23 features, out of which the numerical features are Date (DT), Min Temperature (MINT), Max Temperature (MAXT), Rainfall (RAFL), Evaporation (EVPN), Sunshine (SS), Wind Speed9am (WS9), Wind Speed3pm (WS3), Humidity9 am (HM9), Humidity3pm (HM3), Pressure9am (PR9), Pressure3pm (PR3), Cloud9am (CLD9), Cloud3pm (CLD3), Temp 9 am (TEMP9), Temp 3 pm (TEMP3), Wind Gust Speed(WGS), and categorical features or variables are Locations (LOC), Wind Gust Dir (WGD), Wind Dir 9 am (WD9), Wind Dir 3 pm (WD3), Rain Today (RTDY) & Rain Tomorrow (RTMORO) . The following dataset contains float and object values in which 69.6% of data contains a float value, and 30.4% contain an object value. A dataset is divided into two sets: 75% is used for training, and 25% is used for testing. To obtain the aim of this work, twenty two columns are selected as input data for binary classification models, and one column is for the target variable. TEMP9 Predictor Observed Temperature (in degrees C) at 9am TEMP3 Predictor Observed Temperature (degrees C) at 3pm RTDY Predictor 1 if precipitation exceeds 1mm, otherwise 0 RTMORO Target / Response The target variable. The rain will happen next day or not. The entire data set contains two class labels in the weather classification task, namely Yes or No for Rain Today & Rain Tomorrow.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11657

3.2 Data Preprocessing

Data cleaning is a critically important step in any machine learning project. In this module data cleaning is done to prepare the data for analysis by removing or modifying the data that may be incorrect, incomplete, duplicated or improperly formatted. In tabular data, there are many different statistical analysis and data visualization techniques you can use to explore your data in order to identify data cleaning operations you may want to perform the rainfall prediction

3.3 Exploring Data Analysis

During this progression playing out some enlightening examination and deciding the objective variable. At that point investigating what number of classes were in the objective and a determination of other potentially hazardous factors and Visualizing the objective variable in a histogram which is a decent method for understanding the dissemination of the information to aid parameter tuning.

3.4 Over sampling

Next, we will check if the dataset is unbalanced or balanced. If the data set is unbalanced, we need to either down sample the majority or oversample the minority to balance it. We can observe that the presence of "0" and "1" is almost in the 78:22 ratio. So there is a class imbalance and we have to deal with it. To fight against the class imbalance, we will use here the oversampling of the minority class. Since the size of the dataset is quite small, majority class subsampling wouldn't make much sense here Obviously, "Evaporation", "Sunshine", "Cloud9am", "Cloud3pm" are the features with a high missing percentage. So we will check the details of the missing data for these 4 features. We observe that the 4 features have less than 50 per cent missing data. So instead of rejecting them completely, we'll consider them in our model with proper imputation.

IV.RESULTS AND DISCUSSION

This section describes the experimental results and comparison with other ML classifiers and the related works. The model assumptions must also be reviewed to ensure accuracy, and the necessary corrections must be made.In addition, important feature selection and correlation are used to reduce the number of factors that are taken into consideration for accelerating the prediction process. The 75% of the dataset is used for the training purpose of each ML algorithm and 25% of the testing dataset is used to check the model efficiency.

V. CONCLUSION

In this paper, a different stage of end to end ML life cycle has been envisaged, and the performance of all ML classifiers is observed meticulously. A comparative study is done of all the ML classifiers, and conscientious endeavour has been made to reveal the appropriate methodology of fitting data into the models for getting accurate target variables related to the weather forecast. This paper presents a novel predictive ML algorithm, namely XGB for binary classification imbalanced datasets, which supports the popular Scikit-learn package of Python. The experiments are performed for the proposed dataset based on two imbalanced classification labels, and performance is then observed by evaluation metrics.Based on highly imbalanced data of the Australian weather dataset, XGB is designed to improve the performance of traditional models to detect rain on the following day

REFERENCES

1. Xiao, Z., Liu, B., Liu, H., & Zhang, D. (2012). Progress in climate prediction and weather forecast operations in China. *Advances in Atmospheric Sciences*, *29*(5), 943-957.
2. Atmospheric Sciences, 29(5), 943-957.Bengtsson, L. (1980). The weather forecast. Pure and applied geophysics, 119(3), 515-537.
3. Jain, H., & Jain, R. (2017, March). Big data in weather forecasting: Applications and challenges. In *2017 International conference on big data analytics and computational intelligence (ICBDAC)* (pp. 138-142). IEEE.
4. Rodríguez-Mazahua, L., Rodríguez-Enríquez, C. A., Sánchez-Cervantes, J. L., Cervantes, J., García-Alcaraz, J. L., & Alor-Hernández, G. (2016). A general perspective of Big Data: applications, tools, challenges and trends. *The Journal of Supercomputing*, *72*(8), 3073-3113.
5. Talia, D. (2013). Clouds for scalable big data analytics. *Computer*, *46*(05), 98-101.
6. Li, K., & Liu, Y. S. (2005, August). A rough set based fuzzy neural network algorithm for weather prediction. In *2005 international conference on machine learning and cybernetics* (Vol. 3, pp. 1888-1892). IEEE.
7. Hewage, P., Trovati, M., Pereira, E., & Behera, A. (2021). Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, *24*(1), 343-366.