Machine learning based forecasting model for rainfall prediction

Suvashisa Dash and Answeta Jaiswal *

Department of Mathematics, Centurion University of Technology and Management, Odisha, India.

World Journal of Advanced Research and Reviews, 2024, 21(01), 1678–1686

Publication history: Received on 05 December 2023; revised on 14 January 2024; accepted on 17 January 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.1.0180

Abstract

Weather forecasts can be performed by collecting large amount of data about the present condition of the environment. The state of atmosphere or environment can be the temperature, humidity and wind. The future can be determined by the atmosphere evolution and this evolution can be done through meteorology with understanding of atmospherically processes. The foundation stone of Indian economy is the agriculture, which depends on the rain. Hence, early prediction of rain is highly necessary in India for agriculture. In worldwide, rainfall prediction is one among the challenging issues. Here, varied and most popular machine learning models are utilized for forecasting the rainfall. The outcome from the model gives the centre of the weather forecasts for predicting that whether tomorrow will rain or not. The experimental results will compare and show that good level of accuracy with the help of machine learning algorithms.

Keywords: Weather Forecasts; Rain Fall Prediction; Machine Learning Model; Logistic Regression; Decision Trees; Random Forest; Support Vector Machine

1. Introduction

Weather prediction is the application of science and technology in meteorology. It is getting progressively very important for disaster management, global food security, farmers, agriculturists, scientists and associated organizations for understanding the weather events to be planned and also ready for the upcoming phenomenon [1]. During the past century, it became the highest technologically as well as scientifically challenging issues in all over the world. Forecasting the weather is required to help, getting ready for the worst as well as the best of the weather. The most important challenge issue across the world is, to predict the weather accurately. From the various weather predictions such as predicting cloud conditions, thunderstorm prediction, and rainfall prediction with the varied parameters of weather are major challenges for atmospheric research.

Rainfall prediction contains an important role to work in farming and agriculture field. Rainfall prediction helps in planning of crop production, water resource management and other different things, which have more concern for all mankind [2]. The weather forecast declares that has been observed "30% chance of rain," is a quantitative statement, what it implies, we know it [3]. The above statement is not an ambiguous. Instead, it, we can pass more information such as "It might rain tomorrow," which is a qualitative statement. The quantitative statement does not indicate the category of events due to the single-event probability forecast declaration. Based on rain probability, we have surveyed randomly for finding whether the same statement has been suggested varied interpretations or not [4].

Global warming presents a danger to Australia’s atmosphere and our life style. High temperature is probably to have worsened the effect of recent droughts [5]. Rainfall changes have significant strategy suggestions. An unnatural weather change might be influencing occasional examples of rainfall, which leads to reduce the rain as well as overflow in significant food production areas of Australia, for example, the Western Australia and Murray-Darling Basin. Australia

* Corresponding author: Answeta Jaiswal

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0.
has the bottom average annually rainfall of all populous continents. Nearly one half of the country gets an average of below the 300 mm for every year. Rainfall pattern is strongly seasonal, with large year-to-year variations. The initial steady rainfall observation data originate to 1832. Especially in southern Australia, it demonstrates yearly changeability just as inconstancy over longer timescales [5].

This article illustrates the Australian rainfall prediction, including patterns of weather inconstancy, regional changes, historical trends and projections for the rest of this century. With the help of Bureau of Meteorology, it draws on broad information and data gathered over the previous century. The remaining part of this article is structured in the following way. Section II is narrating the literature survey of weather prediction by using varied ML models. Section III discusses a detailed explanation of the proposed model comprised of data collection and pre-processing of data is given in Section III. Section IV provides the proposed ML models and the comparison of models done in section V. At the end, the paper is concluded with future work in section VI and VII.

2. Literature Survey

There are numerous research results that help the appropriateness of Machine Learning (ML) models for weather prediction.

Jena, T et al. [6], have been built dataset regarding Jordanian precipitation and weather-related data, which is gathered with the help of web and local resources. In light of historical data, Artificial Intelligence (AI) methodologies and data mining methods are utilized for prediction of rainfall and probable weather change. Nayak, S. K. [7] considered the varied regression models for prediction. This comprises the structure of regression, Multiple Linear Regression (MLR), Regression Equation using Covariance matrix method, Matrix formulation of MLR, Steps to calculate the coefficients (parameters), and Multi Structure Regression Equation.

Nayak, S. K [8], researchers presented a model named as Artificial Neural Network (ANN) for rainfall prediction based on the weather data of Pondicherry, India. Three distinctive training models, such as, feed-forward distributed time delay model, layer recurrent model and feed-forward back propagation model were utilized to construct ANN models with fixed 20 numbers of neurons. From the considerable number of models, the outcomes demonstrated that feed-forward distributed time delay model has best MSE and accuracy value as low as 0.0083. Several researchers in computer science have compared several forecasting methods in the prediction of rainfall and weather forecasting. A few of them are outlined here:

<table>
<thead>
<tr>
<th>Authors</th>
<th>Applications</th>
<th>Models</th>
<th>Algorithms</th>
<th>Accuracy</th>
<th>Merits</th>
<th>Demerits</th>
</tr>
</thead>
<tbody>
<tr>
<td>M Kannan et al.[10]</td>
<td>Short Term Rainfall prediction</td>
<td>Regression</td>
<td>MLR</td>
<td>52%</td>
<td>Applicable for small dataset also.</td>
<td>An approximate value is accessed, rather than the accurate value.</td>
</tr>
<tr>
<td>M. A. Kalyankar, S. J. Alaspurkar [12]</td>
<td>Meteorological data analysis</td>
<td>Clustering</td>
<td>K- mean clustering</td>
<td>-</td>
<td>Prediction accuracy is better as compared to others work</td>
<td>Need Dynamic data mining methodologies.</td>
</tr>
<tr>
<td>Reference</td>
<td>Methodology</td>
<td>Model(s)</td>
<td>Prediction Accuracy</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z Jan et al. [13]</td>
<td>Inter annual climate prediction</td>
<td>Lazy learning</td>
<td>96%</td>
<td>Huge number of attributes provides longer term accurate outcomes. Cannot be incorporated for reflecting global changes.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G J Sawale [14]</td>
<td>Weather prediction general</td>
<td>ANN</td>
<td>-</td>
<td>Combining both gives better prediction accuracy. Attribute normalization is required.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F Oliya, A B Adeyemo [4]</td>
<td>Weather prediction and Studies of Climate Changes</td>
<td>Decision tree, ANN, BPN, Hopfield networks</td>
<td>82%</td>
<td>For prediction, best network is chosen. Depending on the training dataset size, the accuracy varies extremely.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3. Collection and Pre-processing of Data

Machine learning (ML) models have been traditionally applied to large, highly dimensional databases [9]. ML is a subset of computer science, through which an algorithm learns from its past experience [9]. The steps of machine learning model as can be seen in Figure 1. Collection and pre-processing of data is the most fundamental part of machine learning model. This model has been applied to clean, normalize and pre-process the collected data called as raw data of weather dataset. It is a time-consuming process but also an essential for the model is need to understand the process of collecting, storing, transforming, reporting the data [2], [10].

#### 3.1. Dataset collection

![Figure 1 Steps of Machine Learning Model](image)
In this article, we have collected the weather dataset of Australia from the website https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. It has around 10 years (from 2008 to 2019) of daily weather observations from several Australian weather stations. In total, the raw data consists of 24 columns or attributes and 169299 numbers of instances. From that, we have taken 18 measured attributes for the prediction and rest attributes have ignored as they are not essential for prediction. Table II shows the description of weather data.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Type of Attribute</th>
<th>Description of Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinTemp</td>
<td>Numerical</td>
<td>Minimum temperature in degree celsius</td>
</tr>
<tr>
<td>MaxTemp</td>
<td>Numerical</td>
<td>Maximum temperature in degree celsius</td>
</tr>
<tr>
<td>Rainfall</td>
<td>Numerical</td>
<td>Amount of rainfall recorded for the day in mm</td>
</tr>
<tr>
<td>Evaporation</td>
<td>Numerical</td>
<td>Evaporation (mm) in the 24 hours</td>
</tr>
<tr>
<td>Sunshine</td>
<td>Numerical</td>
<td>Number of hours of bright sunshine in the day</td>
</tr>
<tr>
<td>Pressure9am</td>
<td>Numerical</td>
<td>Atmospheric pressure (hpa) reduced to mean sea level at 9am</td>
</tr>
<tr>
<td>Pressure3pm</td>
<td>Numerical</td>
<td>Atmospheric pressure (hpa) reduced to mean sea level at 3pm</td>
</tr>
<tr>
<td>Humidity9am</td>
<td>Numerical</td>
<td>Humidity (percent) at 9am</td>
</tr>
<tr>
<td>Humidity3pm</td>
<td>Numerical</td>
<td>Humidity (percent) at 3pm</td>
</tr>
<tr>
<td>Cloud9am</td>
<td>Numerical</td>
<td>Fraction of sky obscured by cloud (in “oktas”: eighths) at 9am.</td>
</tr>
<tr>
<td>Cloud3pm</td>
<td>Numerical</td>
<td>Fraction of sky obscured by cloud (in “oktas”: eighths) at 3pm.</td>
</tr>
<tr>
<td>WindGustSpeed</td>
<td>Numerical</td>
<td>Speed (km/h) of the strongest wind gust in the 24 hours to midnight.</td>
</tr>
<tr>
<td>WindSpeed9am</td>
<td>Numerical</td>
<td>Wind speed (km/hr) averaged over 10 minutes prior to 9am.</td>
</tr>
<tr>
<td>WindSpeed3pm</td>
<td>Numerical</td>
<td>Wind speed (km/hr) averaged over 10 minutes prior to 3pm.</td>
</tr>
<tr>
<td>RainToday</td>
<td>String</td>
<td>Boolean: Yes if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise No</td>
</tr>
<tr>
<td>RainTomorrow</td>
<td>String</td>
<td>Boolean: Yes if precipitation, otherwise No</td>
</tr>
</tbody>
</table>

4. Data Pre-processing

The first step of machine learning model is the data pre-processing. In this case, we will clean, transform and normalise the data. Data pre-processing is required to get the better accuracy for rain fall prediction. In the data cleaning phase, remove the unwanted attributes and keep the essential attributes used to get the better accuracy for the prediction. Another step for cleaning phase is to find the missing values from raw data and replace with mean value technique. In transformation phase, data will be converted into categorical type data value. Here we have used binary categorization type of value i.e. either 0 or 1.

5. Proposed Machine Learning (ML) Models

ML, a computer vision algorithm, is an automatic learning of machines to identify and classify the data, for example, objects, videos, scenes, images, etc. ML utilises models to do decision making, learn, as well as analyze from the raw data [11], [12]. In this paper, we have given more emphasis to machine learning model i.e. conventional machine learning models. This model is of two kinds of learning models, such as, Supervised and Unsupervised ML model. In supervised learning the predicted value of an attribute or attributes is provided to train the data and it is utilised for solving the classification and regression problems. Whereas in unsupervised learning the predicted variable is not assigned. This approach is utilised for making decision of clustering and association problems [11], [13].

In this article, we have considered the following machine learning models to predict the tomorrow's rainfall. They are Logistic regression (LR), Decision Tree, Random Forest (RF), and Support vector Machine (SVM) models. These four
machine learning models are supervised machine learning models. We have implemented the above four models for classification experimentally and compared with each other to find the highest accuracy.

5.1. Logistic Regression (LR)
Logistic regression (LR) [9] is a basic and traditional statistical model utilised for binary classification. LR is utilised for estimation of probability of two events such as “success” and “failure”. One has to go for LR once the predictor variable is binary in nature [16]. This classification scheme uses mathematical LR functions. LR measures the correlation between the categorical target variable and one or more predictor variables by the estimation of probabilities with the help of a logistic function are referred to as cumulative logistic distribution [14]. LR is a linear model, uses a sigmoid function as shown in equation (1). This equation is utilised to conduct the classification whose outcome lies in between 0 to 1 that permits the LR to possess the capability of probabilistic interpretation [4], [3].

\[ f(x) = \frac{1}{1+e^{-x}} \]  

The regression coefficients are normally assessed on the basis of maximum probability estimation. Most of the authors have utilised statistical models, for example, uni-variate or multivariate binary logistic regression for prediction of rainfall [15].

5.2. Decision Trees (DCT)
Decision Trees (DCT) [9] model split data into binary classifications on the basis of progressive iterations. The reason behind this model implementation is utilised for the prediction of predicted variable with the help of input variable through the learning decision rules [16]. The main purpose of ML models is utilised for finding the optimum features and making a branching tree structured diagram. Each node represents a point at which the data are splitting, and the leaves are the output variables. In other words, DCT have terminal nodes, branches, and internal nodes. Terminal nodes are representing the classification outcome, branches are representing the possible values of the attributes and each attribute are represented by internal nodes [2].

5.3. Random Forest (RF)
RF [17] model is an ensemble classifier. It is based on DCT and introduced by Siddique et al. that shows high performance in computer science [18]. RF model is a fast model and it handles the huge number of feature variables or attributes. An arbitrary selection of attributes is included for the development of tens or hundreds of decision tree at a time. These trees are generated with the help of the subsequent strategy [19]:

- For each tree, there is a distinct bootstrap sample data. This sample bootstrap data is equivalent to the actual data contained in each root node of the tree.
- With the help of best split technique, subset of variables is arbitrarily chosen from the predictor variables.
- Each tree is then developed to the most extreme degree conceivable for fully grown.
- When all trees are constructed in the forest, new instances are put together to all the trees. At that point voting process is used to choose the classification with highest votes as the new instance(s) prediction.

5.4. Support vector Machine (SVM)
Cortes and Vapnik introduced the SVM model in 1995. SVM is a well-known model, which separates two classes with the help of a hyperplane [16]. Kernel function is the best part of SVM model that transforms from primary data into high-dimensional data for separating the groups by finding a hyperplane. Hence, for classification, SVM is a powerful tool [12]. In addition, the determination of bit significantly influences the execution of classification. The RBF and other non-linear kernel are not easy to analyze. Similarly, for non-separable datasets, the linear kernel also doesn’t give high predictability. However, its prediction may be better in non-separable cases [20-57]. In our weather dataset, the SVM is taking more time to get the accuracy as compared to the LR, Decision Tree and RF.

6. Result Analysis
This experiment is conducted on weather data of Australia from 01-12-2008 to 29-11-2019. In this dataset, contains 169299 numbers of instances or records with 24 attributes of each. In this experiment, we have chosen only 18 attributes for prediction. Hence, we have applied all the steps of data pre-processing through the implementations of ML models. All these ML models are behaved as classifiers. Classification accuracy is the level of perfectly classified instances by the classifier model, which provides the performance measure of the classifier model [17]. At the end,
results are compared among all the ML models, such as, LR, Decision Tree, RF and SVM for the prediction of tomorrow’s rainfall. From the dataset of Australia’s weather, Rain Tomorrow is the target variable which represents that whether tomorrow will rain or not. The outcomes of varied machine learning models are contrasted based on accuracy prediction with the Australia weather dataset. The graphical observation and its prediction accuracy values are shown in figure 2 and table III respectively.

![Figure 2 Accuracy of Machine Learning Models](image)

From the outcomes, it has been found that Random Forest (RF) is the best prediction model as compared to the other classification algorithms.

7. Conclusion

In this work we carried out an experimental work to compare popular machine learning models for rainfall prediction using various accuracies over weather data of Australia. We have observed that Random Forest results best rainfall prediction with an accuracy of 86%. For this rainfall prediction, Random Forest model shows a proficient as well as an acceptable model. The percentage of accuracy as well as prediction is highly determined by the data being utilized as input for prediction and classification. All models have its own benefits along with limitations and the hardest part is to decide the best model. After analyzing all above-mentioned models of supervised learning, the Random Forest (RF) classification model has considerable level of accuracy and acceptance of our used weather dataset.

Future Work

The accuracy of the prediction for the model may be better by implementing a hybrid prediction model where various machine learning models are assembled to work. In our future work, we are planning for implementing the hybrid prediction model to get the better and higher accuracy.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References


[14] Gaurav J. Sawale, Dr. Sunil R. Gupta, April 2013, Use of Artificial Neural Network in Data Mining For Weather Forecasting, International Journal Of Computer Science And Applications, 6(2).


[27] Siddique, M., Samantara, T., Mishra, S.P., A hybrid prediction model of kernel principal component analysis, support vector regression and teaching learning based optimization techniques, Current Journal of Applied Science and Technology, 40(20), 17-25


