Journal of Engineering Sciences
Vol 12, Issue 06, JUNE /2021
ISSN NO:0377-9254

RAINFALL ANALYSIS AND PREDICTION USING MACHINE LEARNING TECHNIQUES

K.Sarvani[1], Y.Sai Priya[2], Ch.Teja[3], T.Lokesh[4] and E.Bala Bhaskara Rao[5]

[1]*Student, Department of Computer Science and Engineering, Gudlavalleru Engineering College, India*

[2]*Student, Department of Computer Science and Engineering, Gudlavalleru Engineering College, India*

[3]*Student, Department of Computer Science and Engineering, Gudlavalleru Engineering College, India*

[4]*Student, Department of Computer Science and Engineering, Gudlavalleru Engineering College, India*

[5]*Professor, Department of Computer Science and Engineering, Gudlavalleru Engineering College, India*

kallurisarvani123@gmail.com, saipriya4479@gmail.com, tejachelamala123@gmail.com, Lokesh.jiyyar@gmail.com, balabhaskar605@gmail.com

***Abstract*— Rainfall has been a major concern these days. Weather conditions have been changing for the last 10 years. This has paved the way for drastic changes in patterns of rainfall. The factors that have been affecting rainfall are temperature, humidity, wind speed, pressure, and precipitation. These are primary factors that affect rainfall. It is highly important to study the behavior of rainfall against the factors that have been affecting it. Only then we will be able to predict the rainfall accurately. Machine Learning has made our work easier. There are lots of supervised and unsupervised algorithms that are very useful for predictions. The prediction of rainfall using machine learning techniques may use regression. This project intends to provide non-experts easy access to the techniques, approaches utilized in the sector of rainfall prediction and provide a comparative study among the various machine learning techniques.**

***Keywords*— Machine Learning, Regression, Rainfall, Supervised Learning**

1. INTRODUCTION

Rainfall forecasting is important otherwise, it may lead to many disasters. Irregular heavy rainfall may lead to the destruction of crops, heavy floods, that can cause harm to human life. It is important to exactly determine the rainfall for effective use of water resources, crop productivity, and pre-planning of water structures. We know that Agriculture is the primary source of the Indian economy. During the last 10 years, there have been vast improvements in technology and this has increased the rate of global warming, pollution of air, water, noise, dust, etc. This resulted in drastic changes in climate and weather conditions. Rainfall is a key part of the hydrological cycle and alteration of its patterns directly affects the water resources. Changes in the pattern have become a major issue for harvesting crops. Hence, the research on changes in rainfall occurrences is the most sustainable water resource management...

Technology is much more advanced now. Machine Learning has become trending for predictions. It contains various algorithms that can help us in predicting our required value. One major concern is the selection of algorithms. We have to select the algorithms based on our problem statement. Supervised Algorithms are classified as classification and regression algorithms. Regression algorithms are perfect for predicting when a dataset has a single variable(dependent variable) or more variables (independent variables).

Regression analysis is an important tool for modeling and analyzing information. It is used for predictive analysis, for instance, forecasting rainfall or weather, predicting trends in business, finance, and marketing. It can also be used for correcting errors and also provide quantitative support.

2. PROBLEM STATEMENT

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.

3. REVIEW OF LITERATURE

There are various algorithms used for the prediction of rainfall. Fundamentally, there are two approaches to predict Rainfall. They are Empirical and Dynamic methods. The Empirical approach is based on the Analysis of past data of weather and its relationship to different atmospheric variables. In a dynamical approach, predictions are generated based on numerical methods, like using mathematical equations. They have surveyed different algorithms that are helpful for rainfall prediction like, Multiple Linear Regression [MLR], AutoRegressive Integrated Moving Average [ARIMA], Adaptive Splines Threshold Autoregressive [ASTAR], Support Vector Machine [SVM], Back Propagation Neural Network[BPNN]. These are some of the algorithms that are reviewed for their applicability to predicting rainfall [1]. The authors have selected a supervised machine learning algorithm, the Naïve Bayes for rainfall prediction. The algorithm analyzes the previous data related to temperature, region, area, and year, It takes historical data as input and output as a result. The algorithm has proved to simplify and reduce manual work and also provides smooth workflow [2]. In this paper, the authors have used the Linear Regression algorithm along with Neural Network, SVM, and Random Forest algorithms for predicting rainfall. The highest accuracy is attained by the LR algorithm [3]. The authors have used time-series data for prediction. The time-series data is temporal data as it is generated from scientific data, financial applications, GPS, weather data, etc. Artificial Neural Network (ANN) is an extensively established technique for modeling non-linear and dynamic systems. The model has been developed to predict one-month and two-month rainfall [4]. In this paper, the authors have made their analysis using the correlation data and then prediction using the Multiple Linear Regression [MLR] model. The model has selected four parameters like cloud cover, precipitation, vapor pressure, average temperature. They have considered the Udaipur rainfall dataset for the prediction of rainfall.[5] The authors have provided a complete analysis of non-parametric methods like the Mann-Kendal test, Pettitt Test, it's a rank-based test, Method for change point detection, a Method for innovative trend analysis, Method for analyzing rainfall changes, likewise certain methods have been used for analyzing the pattern of rainfall. It provides a complete analysis of the weather dataset of India [6]. Shreekanth Parashar and Tanveer Hurra

made an experiment using Data Mining Techniques for the prediction of rainfall. They have used the Naïve Bayes algorithm, ANN, Decision Trees, Random Forest, and k-nearest neighbor Algorithm to find out which algorithm fit(s) the situation best. They have further concluded that decision trees and Random forest-based models are best for predicting rainfall[8]. Deepak Ranjan Nayak and his team have surveyed rainfall prediction using ANN, this method is more suitable than traditional and numerical methods. They have also surveyed some of the commonly used neural networks like Back Propagation Network [BPN], Radial Basis Function Network [RBFN], Support Vector Machine [SVM], Self Organizing Map [SOM], this is a special class of artificial neural network. The paper can be more helpful for people who are using ANN for their predictions and the rainfall predicting algorithms that use MLR, SVM, BPN, and SOM are more beneficial [9]. Nawaraj Paundel and Tekendra Nath Yogi conducted a comparative study of machine learning algorithms for rainfall prediction in Nepal. They have used classification algorithms like Decision-tree, Random Forest, and SVM. They have calculated precision, accuracy, recall and F-measure values for all the algorithms, Out of all, Random Forest has given 80% accuracy and it is higher [10].

4. METHODOLOGY

The proposed method is represented with the help of a simple flowchart.

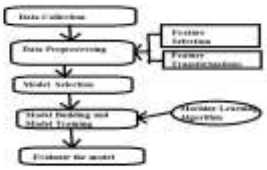

Fig: Methodology

The above flowchart depicts the working of our proposed method. The project uses three algorithms namely, Decision Tree Regression, Random Forest Regression, and SVM. The project is a case study of how the factors vary the rainfall pattern. The prediction is completely based on the independent variables that cause rainfall.

4.1. IMPLEMENTATION

The implementation of the project is divided into seven sections. In the first section, we are going to import the required libraries and then study them. Next, we are going to prepare the dataset with required attributes, then transformations on data are performed, and then data analysis can be made using correlation, followed by splitting of a dataset into train and test sets, finally, model training is done to know the best model that fit(s) our data for predicting rainfall.

Step-1: Import Libraries:

Import Libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
import scipy.stats as stats
from sklearn.model_selection import train_test_split
from sklearn import linear_model as lm
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import RandomForestRegressor
from sklearn.tree import ExtraTreesRegressor
from sklearn import SVM
```

We have imported Numpy, Pandas, Seaborn, Matplot, libraries for evaluating the dataset. Numpy is an open-source module that provides fast mathematical computation on arrays and matrices. We know that Arrays are an integral part of the Machine Learning Ecosystem. Pandas will be useful for performing operations on the data frame. Seaborn and Matplot lib are visualization tools that help us to visualize data in a better way.

We have also imported the required algorithms, Random Forest, Decision Tree, and SVM. The Label Encoder is used to convert the categorical variables

into numerical variables. The data is trained after it is split into train set and test set.

Step-2: Prepare Dataset

We have prepared our dataset from various datasets taking the required attributes that are useful for our case study. We should have a basic understanding of our dataset before moving further.

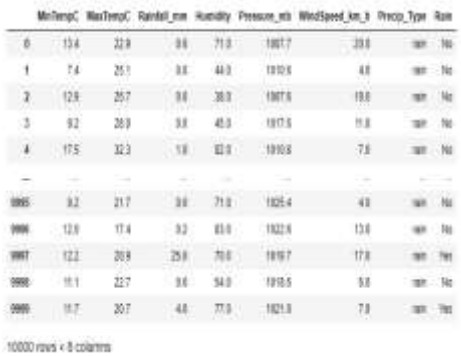

Fig1. Dataset

Step-3: Data Preprocessing

Data Preprocessing is the most vital step while preparing our dataset for model training. Data is often inconsistent, incomplete, and consists of noise or unwanted data. So, preprocessing is required. It involves certain steps like handling missing values, handling outliers, encoding techniques, scaling.

Removing null values is most important because the presence of null values will disturb the distribution of data, and may lead to false predictions. There is very less percent of null values in the dataset.

Fig2: Null values

Missing values:

Imputation is used for replacing missing values. There are few kinds of imputation techniques like Mean imputation, Median imputation, Mode Imputation, Random Sampling imputation, etc. Based on the type of data we have, we can use the required imputation. We have used median imputation to handle missing values.

Handling Outliers:

Outliers are nothing but an extreme value that deviates from the other observations in the dataset. These outliers are either removed or replaced with their nearest boundary value, either upper boundary or lower boundary value.

Label Encoding:

Label Encoding is one of the kinds of encoding techniques that will change categorical variables into numerical variables. It is important to convert the labels because our model can only understand numeric data.

Step-4: Visualization using the technique of Correlation

We need to understand our data. Data visualization is a powerful technique that helps us to know about the trends, patterns that our data follows. There are different techniques to visualize data, one such method is a correlation. Correlation tells us how one or more are related. If two variables are correlated, then we can tell that both are strongly dependent on each other. The variables that are strongly correlated to the target variable, are said to have more influence on the target variable.

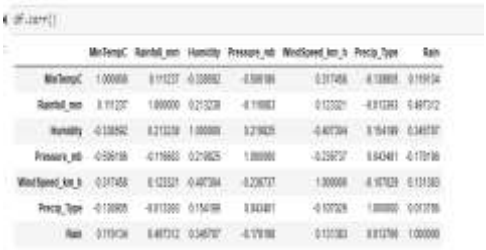

Fig3. Correlation Matrix

Correlation can also be visualized using Heatmap. Heatmap is one of the visualizing graphs like Histograms, Boxplots that help us to know our data easily. As human minds are so complex to understand data from numbers, they can easily understand using pictures.

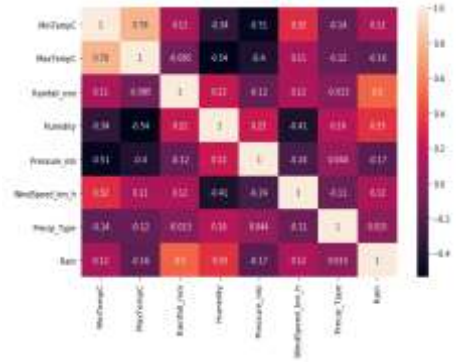

Fig4: Heatmap

Based on the color, we can easily distinguish how the variables are correlated.

Light color- Less correlated
Medium-light color- correlated
Medium color-less strongly correlated
Dark color- strongly correlated

Based on the above heatmap, we can draw few inferences,

1) Humidity is strongly correlated with the target variable.
2) Next to Humidity, Precipitation also correlated with the target variable.
3) Atmospheric Pressure variable is negatively correlated with a target, which means that the increase in one value may decrease the other value. In short, we can tell that both variables are inversely proportional.

Correlation also helps us to remove certain values, as this is one of the feature extraction techniques. If two independent variables are strongly correlated with each other, we can remove any one of the variables. This may not cause any disruption to the dataset.

Step-6: Splitting Dataset

Dividing the dataset into two sets should be done precisely. The dataset can be divided into the ratio of 80% train set, 20% test set or 70% train set, 30% test set, or any other way. The division of the dataset also affects the accuracy of the training model. A slicing operation can be performed to separate the dataset.

We've take care while splitting the dataset, assure that the test set must hold an equivalent features as the train set and also the datasets must be statistically meaningful.

Considering the independent variables into 'x' and therefore the target variable into 'y',

x = df.iloc [:,:-1].values

y = df.iloc [:, 7].values

```
#Splitting the dataset into Training set and Test set

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)
```

Fig5. Dividing dataset

In the above figure, we have divided the dataset into,
80% training dataset
20% testing dataset

Ensure that the dataset is split into train and test sets before training the model. Sometimes we may find more accuracy for our models if we involve both datasets.

Step-7: Model Training

There are several algorithms in machine learning, but we have chosen only three from them to train our model. In Regression, accuracy can be measured by using R2-Score or Mean Squared Error [MSE] or Root Mean Squared Error [RMSE]. The model should be imported from the Sklearn package and then trained.

We have to install the Sklearn package and then import it. We have defined a

model method to call various models. The model method contains the training and validation statements that will train and test the dataset. MSE value, RMSE value, R2-score are also calculated after training and testing the model. The test dataset will check whether our trained model is efficient for real-time data or not.

In Regression, to know the accuracy of the model, we can simply go through R2-score, RMSE, and MSE values. The higher the R2-Score, the efficient the model. The lesser the RMSE and MSE values, the efficient the model. The error value must be less so that our model is more efficient.

Fig6.decision tree

Fig7.RandomForest

Fig8.SVM

5. RESULTS

The models have been trained successfully. All the values can be drafted into tabular form as shown,

Regressor	MAE difference	RMSE difference	R2 score
Decision Tree Regressor	-0.06012	-0.2418	98.133
Random Forest Regressor	-0.03011	-0.1000	97.181

Fig9.Result

The R2 score value, RMSE value difference, MSE value differences are displayed in the table. The RMSE value difference is calculated by subtracting the test data RMSE value from train data RMSE value. Similarly, the MSE difference value is also calculated.

The difference is calculated to know whether our testing dataset is accurate or not. If the testing RMSE value is greater than the training RMSE value then the result will be a negative value. This shows us that the testing dataset is trained better than the training dataset.

The R2-score value provides us with the accuracy of each model. The random Forest Regression algorithm shows us the highest accuracy. The difference between errored values is also almost less. Random Forest Algorithm is strictly trained well when compared with other algorithms.

The accuracy of SVM is 80% after training. The accuracy is good but is less when compared to other algorithms, this is because of categorical values that are present in the dataset. As we know that, Classification algorithms are best suited for numerical data, this has resulted in a slight decrease in the accuracy of SVM.

6. CONCLUSION

The project aims at the selection of a definite algorithm to predict rainfall concerning the factors that affect rainfall. It is proved that Random Forest Regression Algorithm can be an adaptable strategy for prediction. In this project, we have studied various algorithms and their reaction to each variable for the target variable.

Machine Learning can provide us with intelligent models rather than traditional models. The computational power required is also less and manual effort is also reduced. Regression is best

for prediction, So, We have considered regression algorithms and one classification algorithm.

We have learned different preprocessing techniques that are required in preparing the dataset. The dataset must be free from all kinds of noise, inconsistency, overfitting, and other odds that may affect the performance of the model.

We also explored few regression algorithms that can change the predictions with disturbances in input data. The comparative study has made us understand various algorithms effectively.

REFERENCES

[1] Mr. Dhawal Hirani, Dr. Nitin Mishra, "A survey on Rainfall Prediction Techniques", International Journal of Computer Application (IJCA), Volume 6-No.2, 2016, 2250-1797.

[2] Mrs. P. Chaya, I. Kavya, S. Likitha, M. Tanushri, C. Roshni Poovanna, "A Prognostic Rainfall Using Machine Learning Technique", International Journal For Research In Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC value: 45.98, Volume 8 Issue VIII July 2020, Research Paper Available Online at: www.ijraset.com

[3] Mrunmay Jalgaonkar, Dr. Umesh Kulkarni, "Rainfall Prediction using Regressions and Multiple Algorithms", International Research Journal of Computer Science (IRJCS), Volume 8, Issue 4, April 2021. Research Paper Available online at: https://www.irjcs.com/archives

[4] Neelam Mishra, Hemanth Kumar Soni, Sanjiv Sharma, A. K. Upadhyay, "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction Using Time-Series Data", Intelligent systems and Applications (I.J), 2018, 1, 16-23.

[5] Nikhil Sethi, Dr. Kanwal Garg, "Exploring Data Mining Technique for Rainfall Prediction", Vol. 5(3), 2014, ISSN: 0975-9646.

[6] Bushra Praveen, Swapan Talukdar, Shahfahad, Susanta Mahato, Jayanta Mondal, Pritee Sharma, Abu Reza Md.Towfiqul Islam, Atiqur Rahman, "Analyzing Trend and Forecasting of rainfall changes in India using non-parametrical and machine learning approaches", Scientific Report, 2020.

[7] Aakash Parmar, Kinjal Mistree, Mithila Sompura, "Machine Learning Techniques for Rainfall Prediction: A Review", International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), March 2017.

[8] Shreekanth Parashar, Tanveer Hurra, "A Study of Rainfall Using different Data Mining Techniques", Research Gate, Article-May 2020.

[9]Deepak Ranjan Nayak, Amitav Mahapatra, Pranati Mishra, "A Survey on Rainfall Prediction using Artificial Neural Networks", International Journal of Computer Applications, volume 72-No.16, June 2013.

[10] Nawaraj Paudel, Tekendra Nath Yogi, "Comparative study of Machine Learning Algorithms for Rainfall Prediction- A case study in Nepal", International Journal of Advanced Research in Engineering and Technology (IJARET), Volume 11, Issue 10, 2020.