WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Rainfall Prediction Using Machine Learning

Shrutika Kadam[a], Neha Jadhav[a],Vaishanvi Mardane [a], Naresh Kamble[b]

[a]Students,Sanjay Ghodawat Polytechnic, Atigre, kolhapur, Maharashtra, India.
[b]FacultyCSE,Sanjay Ghodawat Polytechnic, Atigre, kolhapur, Maharashtra, India.
.

A B S T R A C T

The rainfall prediction project is based on machine learning.Genetic algorithm is used in this project. Numerous research works are being carried out by various people to predict the occurrence of rainfall before it actually comes. Use of genetic algorithms has seen in intense use for various research purposes and prediction of weather is not an exception. In this project, we have proposed a system that could predict the rain beforehand using genetic algorithm.Predicting the amount of rainfall improves agricultural productivity and secures food and water supply to keep citizens healthy. To predict rainfall, several types of research have been conducted using machine learning techniques of different environmental datasets. Use of rainfall water should be planned and practiced in the country to minimize the problem of the drought and food occurred in the country. The main objective of this study is to identify the relevant atmospheric features that cause rainfall and predict the intensity of rainfall using machine learning techniques. Select relevant environmental variables which were used as an input for the machine learning model. The dataset was collected from the Google websites to measure the performance of three machine learning techniques. The result of the study revealed that the Extreme Gradient Boosting machine learning algorithm performed better than others.

Keywords: ML, Rainfall Prediction, Genetic Algorithm, Accuracy, CSV data, SVM, Naivy bias, RBFN, Dataset, Humidity, Temperature, Visibility, Sunshine, Pressure, Evaporation.

1.INTRODUCTION

Rainfall prediction is crucial for increasing agricultural productivity which provide food and quality water supply. To quality water supply, and agricultural productivity. Agriculture and water quality depend on the rainfall and water amount on annual basis. Therefore, accurate prediction of daily rainfall is a challenging task to manage the rainfall water for agriculture and water supply.

Various researchers conducted studies to improve the prediction of daily, monthly and annual rainfall amounts using different countries' meteorology data. We use daily rainfall formula on bases of mathematical calculation. According to the results of the studies, the prediction process is now shifted from data mining techniques to machine learning techniques.

Consequently, this paper analysed different machine learning algorithms to identify the better machine learning algorithms for accurate rainfall prediction. Several environmental factors affect the existence of rainfall and its intensity.

To temperature, relative humidity, sunshine, pressure, evaporation, etc. are some of the factors that affect the existence of rainfall and its intensity directly or indirectly.We have proposed a model that makes use of a genetic algorithm for predicting rainfall using previous dataset. The genetic algorithm is also used for finding the accuracy and performance of the entire proposed model.

There are many hardware devices for predicting rainfall by using the weather conditions like temperature, humidity, pressure. In this project classifiers that we used SVM, Naivy-bias, and RBFN monetary unit are additional appropriate than alternatively applied math and numerical techniques. We have used CSV dataset of various regions of country for rainfall prediction.

To raw data is collected from regional meteorology and pre-processed to make it suitable for the experiment. Each feature of the pre-processed data is

correlated with the rainfall variable to identify the relevant features using Pearson correlation. To study then experimented Genetic machine learning algorithms.

2. LITERATURE REVIEW

Rainfall prediction is crucial for increasing agricultural productivity which in turn secures food and quality water supply for citizens of one's country. To scarcity of rainfall has a negative influence on the aquatic ecosystem, quality water supply, and agricultural productivity. Agriculture and water quality depend on the rainfall and water amount on a daily and annual basis[1]. Therefore, most researchers did not show the prediction of the daily rainfall amount rather conducting experiments on environmental data to predict whether rain or not rain and predict average annual rainfall amount that is the prediction of daily rainfall amount is a challenging task.

Data mining is now used in various domains, including time series data. Time series data analysis is used for weather forecasting or rainfall prediction with the help of data mining techniques. For time series data analysis, intelligent forecasting models perform better than methods that are traditionally used in forecasting[2]. Genetic algorithm (GA) arethe most popular techniques based on computational intelligence. In the literature, hybrid methods, which consist of combining more than one technique, are also commonly found.

The proposed system predicts rainfall for the approach which is more accurate. The data set is collected. There are two techniques to predict rainfall. The first one is machine learning approach, which includes LASSO regression. The second one is neural network approach. This system first compares both the process and then accordingly gives result with the best algorithm.[3].

We reduce the number of input parameters in the ESN/DeepESN model from seven to three, which include rainfall, pressure and humidity. Ten, we repeat the whole procedures which include the data training and data testing. In the case, the RMSE, NRMSE, and γ are 1.51, 0.02 and 0.518, respectively. Tis outcome indicates that removing the irrelevant parameters can improve the performance of rainfall prediction. In this study, although a small number of parameters can improve performance of prediction, it might be a special case. For the deep learning, more input parameters might get better performance for model training and testing[4].

In this paper, the rainfall was predicted using a machine learning technique. Tree machine learning algorithms such as Genetic algorithm(GA) were analysed which took input variables having moderately and strongly related environmental variables with rainfall.

3.SCOPE OF PROJECT

The goal of Rainfall Prediction is to provide information people and organizations can use to reduce weather-related losses and enhance societal benefits, including protection of life and property, public health and safety, and support of economic prosperity and quality of life. Rainfall Prediction Model has a main objective in prediction of the amount of rain in a specific well or division in advance by using various regression technique and find out which one is best for rainfall prediction.

This model also helps the farmer for agriculture to decide the crop, helping the watershed department for water storage and also helps to analyses the ground water level. The prediction of the state of the atmosphere for a given location using the application of science and technology. This includes temperature, rain, cloudiness, wind speed, and humidity. Weather warnings are a special kind of short-range forecast carried out for the protection of human life.

Rainfall Prediction Model has a main objective in prediction of the amount of rain in a specific well or division in advance by using various regression technique and find out which one is best for rainfall prediction. This model also helps the farmer for agriculture to decide the crop, helping the watershed department for water storage and also helps to analyze the ground water level. 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

4. METHODOLOGY

There are numerous works that have been proposed by various researches for predicting rainfall in given geographic area in their papers.
In daily life we measured rain using one mathematic formula which is A 1,000 square foot roof is bombarded by 625 gallons of water per inch of rain. As you probably know, rainfall amounts in the United States are typically measured in inches.

Actually, although we usually just say "inches," we really mean "inches in the storm" or "inches in the last 24 hours" or "inches in some time period."

Why does that matter? Well, obviously 1 inch of rain in a 15 minute period is a lot more water than 1 inch of rain in the last month. So we really don't know how much it's rained if we don't know the time period we're talking about.

Day	Mon	Tue	Wed	Thu	Fir	Sat	Sun
Rainfall (in mm)	0.0	12.2	2.1	0.0	20.5	5.5	1.0

Maximum rainfall = 20.5

Minimum rainfall = 0.0

Range = Maximum rainfall - Minimum rainfall

= 20.5 – 0.0

= 20.5

So what does 1 inch, 2 inches, or whatever number of inches of rain in some time period mean? Well, if all the rain that falls stays right where it lands—meaning it doesn't run off and accumulate in streams and rivers and eventually in lakes and oceans, and it isn't absorbed into the ground—then 1 inch of rain in an area is enough to evenly cover the ground in that area with a layer of water 1 inch deep. Of course, water typically does run into streams and is absorbed into the ground, so 1 inch of rain rarely means an inch of standing water.

But whether or not we actually see 1 inch of water on the ground for each inch of rain, we can use this definition to construct a device that measures how much it has rained—a so-called rain gauge. The math is actually pretty straightforward—although there are some *unit conversions* that we need to be careful about along the way.

4.1 Data collection

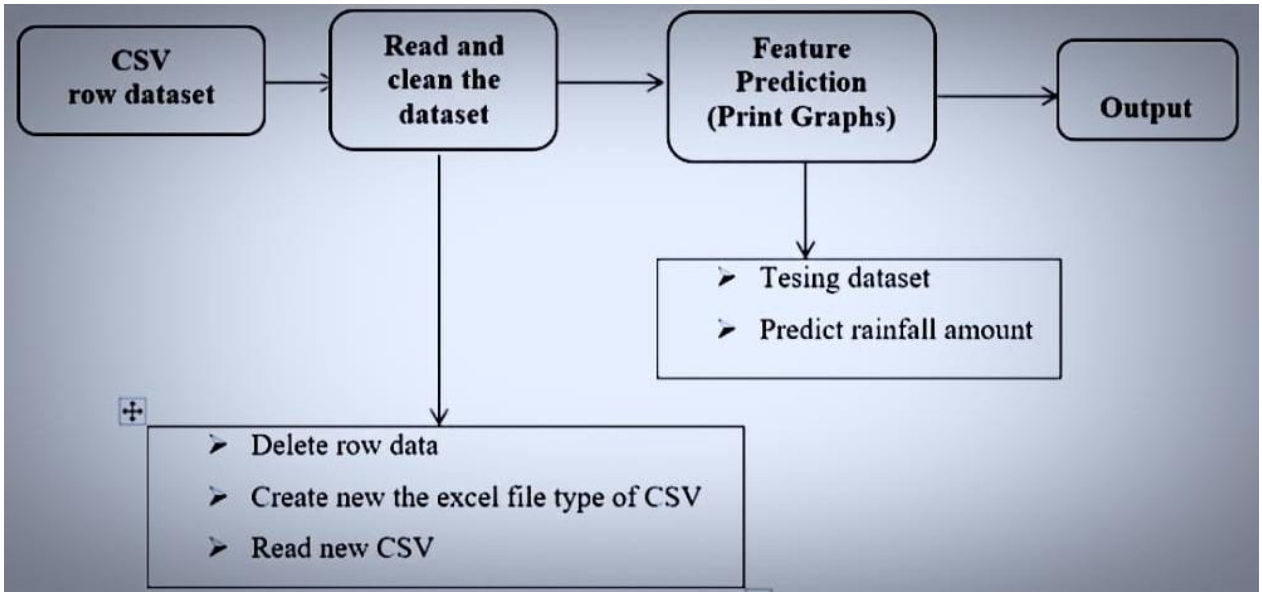
The raw data were collected of month, date, evaporation, sunshine, maximum temperature, minimum temperature, humidity, wind speed, and rainfall were included. To meteorology station records the values of the environmental variable every day for each year directly from the devices in the station. The data were recorded in the Microsoft Excel CSV file format.

To year and the days of the month were arranged in the row of tables related to environmental variables in the column of the table. To raw data recorded were used for the study. Work on uncorrupted dataset which is based on overall India's rainfall. After work on dataset we got the useful dataset get ready for the creating graphs.

4.2 Data pre-processing

To data pre-processing step included the data conversion, manage missing values, categorical encoding, and splitting dataset for training and testing dataset. Since the data were raw, they contained missing values, and wrongly encoded values so that the missing values of the target variable were removed and the other features were filled using the mean of the data.

In the meteorology once, the raw data were also arranged in a year based and the attributes in rows that need to combine and rearrange features in columns. Data were converted from excel data to CSV data. Encoding the dataset was performed and then the dataset was prepared for the project.



4.3 Genetic algorithm

The genetic algorithm that is intensely used in feature selection. The feature selected from the input dataset is used with the selected feature.The model consists of three modules. The first one is the Preprocessing module where the input is taken and then converted for extracting the features.

The second one is the Genetic Algorithm. Data processing is done for the given input dataset. We have a flow of genetic algorithm.This algorithm was used to predict the rainfall dataset values. For choosing best from the selected region or area.

4.4 Classifiers and Dataset

We have used India's CSV dataset for analysis. SVM, Navy bias and RBFN Classifiers. SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs Navy bias Naive Bayes uses a similar method to predict the probability of different class based on various attributes.

This algorithm is mostly used in text classification and with problems having multiple classes. RBFN an RBFN performs classification by measuring the input's similarity to examples from the training set.

Each RBFN neuron stores a "prototype", which is just one of the examples from the training set. When we want to classify a new input, each neuron computes the Euclidean distance between the input and its prototype Project objectives vs. project goals Genetic algorithm.

This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation Classifiers- Svm In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class.

Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values([-1,1]) which acts as margin Navy bias Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely RBFN Activation functions in RBFNs are conventionally implemented as Gaussian functions.

5. SYSTEM CONFIGURATION

5.1 Requirements for analysis

- Algorithm
- Classifiers
- Dataset

5.2 Hardware and Software Requirement

- Operating System -Windows 10, Android
- Languages – Python
- RAM : 4GB MINIMUM
- HARD-DISK : 5GB MINIMUM
- Compiler: VS Code
- Browser – Chrome, Firefox, Opera, UC-Browser, etc.

5.3 Database Requirement

CSV Dataset (Excel sheet)

1. Corrupted dataset
2. Final cleaned dataset

6. ADVANTAGES AND DISADVANTAGES

6.1 Advantages-

- The amount of time saved may be very high because of the ML program. And we can all use more energy in everyday life today.
- Easy to predict rainfall using ML rather than other platforms.
- We can schedule the system to irrigate a piece of rainfall efficiency.

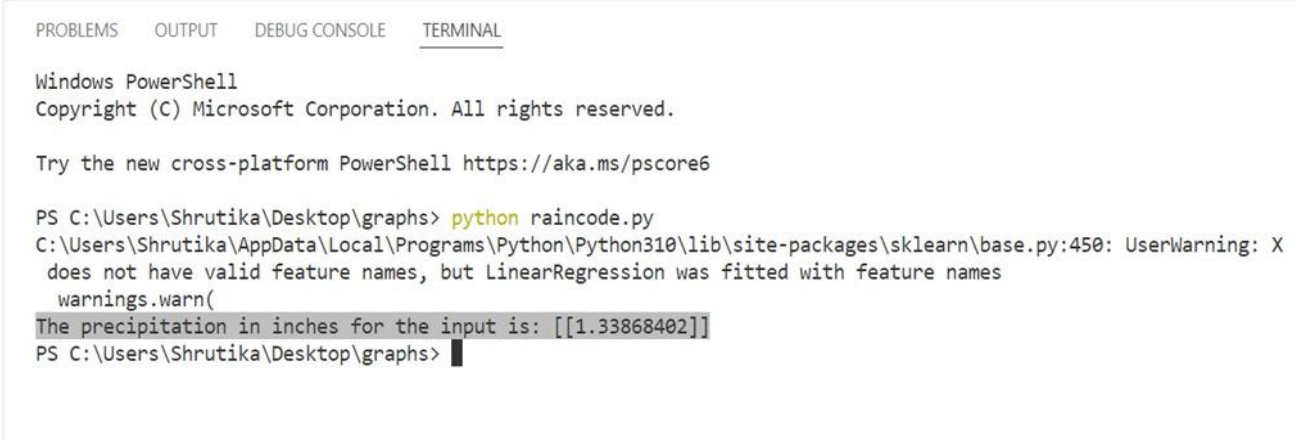
6.2 Disadvantages-

- High cost of maintenance.
- Lil bite difficult to work on raw dataset.
- Weather is extremely difficult to forecast correctly.
- Expensive resources.
- The computers needed to perform code by using multiple extensions.

7. IMPLEMENTATION AND RESULT

The result will be received in the form of graphs and excel sheets. For pre-process , all the result will be received in the form of different graphs and for machine learning.

User gives the dataset as input in the system. Along with that, there will be more buttons present. First one is for pre-process which represent the dataset in the form of graphs, and the second one is neural network which gives the neural network ‘s accuracy. So, to have a better understanding of the dataset and for better comparison, first pre-process should be done. Before going for the prediction, pre-process can be done.



It is representation of the dataset in form of graph. It eases the process of comparison and along with that it also gives a better understanding of the dataset

present. Dataset should be split in two parts, the first part deals with training the algorithm used and the rest part used to predict the amount of rainfall. Rainfall is predicted only with the algorithm with more accuracy.The algorithm used should undergo training before it does prediction. So, in this part of the system, the training is been done.

This step gives a proper idea of which algorithm is more accurate among the two. Then the remaining dataset (which is not used in training = Raw data) is being used and rainfall prediction is been done. This part is also done in both the approaches. Finally, after the all the process is completed.
The accuracy is received in the form of Metrics and excel sheet(csv dataset). In Metrics along with the accuracy different types of errors are also shown and the same is represented in the excel sheet(Corrupted data which is called raw data).

We work on CSV dataset for analyses of rainfall and for achieve the graphs also. We work on steps of coding to find ratio of rainfall.

7.1 Import Libraries

1. Import pandas as pd for powerful and flexible quantitative analysis tool, pandas has grown into one of the most popular Python libraries. It has an extremely active community of contributors.

2. Import numpy as np for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

3. Import sklearn as sk for providing a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

4. Import LinearRegression for performing a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

5. Import matplotlib.pyploat as plt for cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

6. Import XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text)

After cleaning the dataset we will understand the dataset for the learning and executing next result that's why we read all dataset and create it in proper manner. Precipitation level means the amount of rain, snow, hail, etc. that has fallen at a given place within a given period, usually expressed in inches or centimetres of water.(After reading the dataset we get result of Precipitation level rainfall.

On a next step we collect the some other levels of output such as date, evaporation, sunshine, maximum temperature, minimum temperature, humidity, wind speed, and rainfall were included.

The result is received in form of graph and table which shows the future rainfall and the accuracy of the algorithm. After pre-process the graphs which are received are precipitation, Temperature Average,DewPointAvgF, HumidityAvgPercent, SeaLevelPressureAvgInches, VisibilityAvgMiles, WindAvgMPHfor counting daily rainfall on bases of mathematic formula.

7.2 Process of reading CSV

We work on another Csv dataset which is based on Australia rainfall. We calculate the rainfall Evaporation, Sunshine, Location, Date etc. in this dataset

```
df = pd.read_csv('weatherAUS.csv')
df.head()
df.describe()
df.shape
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
PS C:\Users\Shrutika\Desktop\py>
```

- Calculate some points related to the rainfall Evaporation, Sunshine, Location, Date etc.

```
df=df.drop(["Evaporation","Sunshine","Cloud9am","Cloud3pm","Location", "Date"],
axis =1)
df.head()
```

```
PS C:\Users\Shrutika\Desktop\py>
```

```
importxgboostasxgb
xgb = xgb.XGBClassifier()
xgb.fit(x_train, y_train)
pred = xgb.predict(x_test)
print('acc',accuracy_score(y_test,pred))
print('f1',classification_report(y_test,pred))
print('matrix',confusion_matrix(y_test,pred))
```

```
Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
[[16723   866]
 [ 2616  2380]]
              precision    recall  f1-score   support

           0       0.86      0.95      0.91     17589
           1       0.73      0.48      0.58      4996

    accuracy                           0.85     22585
   macro avg       0.80      0.71      0.74     22585
weighted avg       0.84      0.85      0.83     22585

0.8458268762452955
[[15125  2464]
 [ 2275  2721]]
              precision    recall  f1-score   support

           0       0.87      0.86      0.86     17589
           1       0.52      0.54      0.53      4996

    accuracy                           0.79     22585
   macro avg       0.70      0.70      0.70     22585

0.7901704671241975
[[16782   807]
 [ 2451  2545]]
              precision    recall  f1-score   support

           0       0.87      0.95      0.91     17589
           1       0.76      0.51      0.61      4996

    accuracy                           0.86     22585
   macro avg       0.82      0.73      0.76     22585
weighted avg       0.85      0.86      0.84     22585

0.8557449634713306
acc 0.8550808058445871
f1               precision    recall  f1-score   suppo

           0       0.88      0.95      0.91     17589
           1       0.74      0.54      0.62      4996

    accuracy                           0.86     22585
   macro avg       0.81      0.74      0.77     22585
weighted avg       0.85      0.86      0.85     22585

matrix [[16628   961]
 [ 2312  2684]]
PS C:\Users\Shrutika\Desktop\py>
```

8. CONCLUSION

Rainfall Prediction is the project area of data science and machine learning to predict the state of the atmosphere. It is important to predict the rainfall intensity for efective use of water resources and crop production to reduce mortality due to food and any disease caused by rain.

First we resurch on our daily mathematical formula for measuring rainfall. Then we compare that with our predicted result of our ML project.

This paper analyzed various machine learning algorithms for rainfall prediction. The machine learning algorithms such as Genetic, MLR, FR were presented and tested using the data collected from the meteorological station. To selected features were used as the input variables for the machine learning model used in this paper.

A comparison of results among the three algorithms (Genetic, MLR, RF) was made and the results showed that the Genetic was a better-suited machine learning algorithm for rainfall prediction. Genetic algorithms play a vital role in classification techniques and it also plays a vital role in predicting the weather of a specific area or a country for a particular period of time.

The Rainfall prediction accuracy can be improved using meteorological datasets with additional diferent environmental features. Hence, in future work, big data analysis can be used for rainfall prediction if the meteorological datasets are used for the rainfall prediction study.

REFERENCE

1. Liyew and Melese Journal of Big Data (2021) https://doi.org/10.1186/s40537-021-00545-4
2. ICT Research and Applications(2017).https://www.researchgate.net/publication/319702656
3. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020.
4. Scientific Reports | (2019).https://www.nature.com/sre