IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

RAINFALL PREDICTION FOR CROP PRODUCTION USING MACHINE LEARNING ALGORITHMS

Amar Suryawanshi[1], Shubham Argade[2], Sandip Pawar[3], Amit Bhosale[4], Prof. T.B. Tambe[5]
[1]B.E. Student, [2]B.E. Student, [3]B.E. Student, [4]B.E. Student, [5]Assistant Professor
[1]Department of CSE,
[1]P K Technical Campus, Pune, Maharashtra, India

Abstract: **The monthly rainfall projections are compared to actual data after training and testing to confirm the model's accuracy. When particular parameters are employed, the model can accurately anticipate monthly rainfall data, according to the findings of this study. Rainfall prediction is one of the most studied areas since it affects the lives and property of many people. Previous rainfall prediction models used a complicated combination of mathematical instruments, but these were insufficient to reach a higher categorization rate. In this study, we propose novel ways for calculating monthly rainfall using Machine Learning Algorithms. Rainfall forecasts are based on quantitative information about the current state of the atmosphere. Complex input-to-output mappings can be learned using a variety of machine learning approaches. The dynamic structure of the atmosphere makes precise rainfall prediction difficult. To forecast future rainfall conditions, the variation in previous year's circumstances must be used. We supported the use of machine Learning Algorithms based on a variety of factors such as temperature, humidity, and wind. Because the suggested algorithm forecasts rainfall based on prior data for a specific geographic area, this prediction will be extremely accurate. When compared to standard rainfall prediction techniques, the model's performance is more accurate.**

Index Terms **- Machine Learning, Linear Regression, Random Forest, Ridge Regression, Lasso Regression, Support Vector Machine.**

1. INTRODUCTION

Rainfall has an important role in the development of natural life's fauna and flora. It is critical for people, as well as animals, plants, and other living things. Water is one of the world's most valuable natural resources, and it is widely used in agriculture and farming. Changing climatic conditions and rising greenhouse gas emissions have made it more difficult for humanity and the earth to obtain the required amount of rainfall and continue to utilize it in daily life. As a result, it's become critical to track changing rainfall patterns and try to predict rain not just for human needs but also for natural disasters that could be generated by unusually heavy rains. The purpose of this study is to predict rainfall using Machine Learning. Rainfall forecasting can help with not only the analysis of changing rainfall patterns, but also the preparation of preventative actions in the case of a disaster and the management of the disaster. The rainfall forecast would also be useful in developing policies and strategies to combat the growing global problem of ozone depletion. Changes in rainfall patterns are linked to global warming, which is described as an increase in the earth's temperature caused by increased Chlorofluorocarbon emissions from ubiquitous products such as refrigerators, air conditioners, deodorants, and printers. Climate change is being exacerbated by rising temperatures. Rainfall forecasts and weather updates, meanwhile, not only assist in the management of macro-level concerns such as floods and agricultural challenges caused by insufficient or severe rainfall, but also in the management of micro-level concerns such as drought. Rainfall prediction utilizing Neuro Fuzzy and Artificial Neural Networks could potentially improve people's well-being and comfort by keeping them informed about rainfall trends and anticipating rainfall. Rainfall forecasts can help people cope with hot and humid weather. The technological improvements of the modern world have made greater opportunity for innovation and creativity. Although the current problems are most likely due to technology advancements 2, it is vital to consider the wide range of possibilities and opportunities that technological evolution has brought to humans. Furthermore, erroneous or insufficient rainfall prediction is a source of concern in the management of water reserves. A precise and accurate rainfall forecast may aid not only in the effective and economical use of this natural resource, but also in the management of power generation projects and plans. It is vital to develop and operate a system that can aid in accurate prediction and give consumers easy access to it. One of the most appropriate and reliable systems for rainfall prediction that has previously aided operators is the Artificial Neural Network for Rainfall Prediction.

1.1 RELEVANCE

Rainfall forecasting is critical not just at the local level, but also at the national level. The work is significant because it contributes to agriculture, water reserve management, flood prediction, and management, with the purpose of making people's lives easier by providing weather and rainfall forecasts. Agricultural businesses must also rely on current rainfall forecasts to maintain their crops healthy and ensure the production of seasonal fruits and vegetables. The study will also be useful to flood control authorities, since more precise and accurate forecasting of heavy monsoon rains will keep them alert and focused on an impending disaster that may be avoided by taking preventative measures. Because water is a scarce resource that must be protected for the purpose of humans, the rainfall forecast will be immensely beneficial in solving the developing issue of water resource management. It will also help people organize and manage their social activities.

1.2 MOTIVATION

Machine Learning (ML) is a technique for resolving problems where the link between the input and output variables is unknown or difficult to define. The automatic acquisition of structural descriptions from instances of the described item is referred to as "learning." Unlike traditional statistical methodologies, ML does not make assumptions about the correct structure of the data model that characterizes the data. This characteristic is very useful when modelling complex non-linear systems like agricultural yield prediction functions. One of the most effective applications of machine learning algorithms is crop yield prediction (CYP). In a supervised learning process, a goal / outcome variable (or dependent variable) must be predicted given a set of predictors (independent variables). Using this set of variables, we design a function that maps inputs to desired outputs. On the training data, the model is trained until it achieves the desired level of accuracy. Supervised learning includes techniques such as regression, decision trees, random forests, KNN, and logistic regression, among others.

2. PROPOSED SYSTEM

The recommended predictive algorithm is used to anticipate rainfall. The prediction model is built using historical rainfall data, mathematical formulae, data mining, machine learning, and other techniques. In the first stage, the dataset is preprocessed to remove superfluous data, noise, and missing values. Following preprocessing, the dataset is divided into two partitions: training data and testing data, with 80% of the dataset being used for training and 20% for testing the prediction build model. Following successful validation of the developed model, i.e., the model is working effectively with suitable output, the model is deployed for future applications.

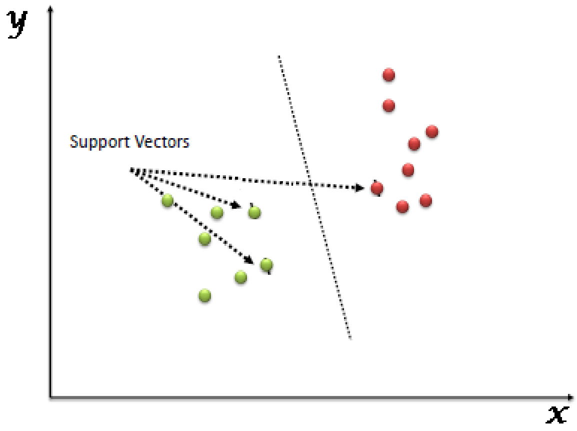


Fig: Training & Testing Diagram

Forecasting rainfall is important for water resource management, human life and the environment, and agricultural crop management. The statistical model gives misleading findings since rainfall is nonlinear in nature and its values do not remain constant. In order to anticipate rainfall, the survey article investigates a number of well-known Machine Learning and algorithm. These techniques would aid in the accurate forecasting of rainfall.
Other meteorological parameters, including as evaporation, mean temperature, humidity, and soil temperature, could also be used as inputs to improve SVM prediction accuracy.

2.1 ALGORITHMS

Linear regression and logistic regression are the two types of regression analysis methodologies used to tackle the regression problem utilizing machine learning. They are the most popular regression techniques. However, in machine learning, there are many distinct types of regression analysis methodologies, and their application differs based on the data.
The many regression approaches are as follows:

1. Support Vector Machine
2. Random Forest
3. Linear Regression
4. Ridge Regression
5. Lasso Regression

1. SUPPORT VECTOR MACHINE

The Support Vector Machine (SVM) is a supervised machine learning approach for classification and regression issues. However, it is usually used to tackle categorization problems. Each data item is represented as a point in n-dimensional space (where n is the number of features), with the value of each feature being the SVM algorithm's value for a certain coordinate. Then we locate the hyper-plane that clearly separates the two classes to complete categorization.

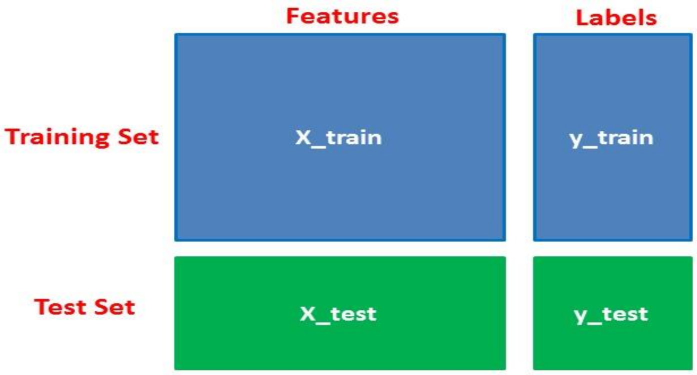


Fig -1: support vector machine

2. RANDOM FOREST

The Random Forest algorithm is a supervised learning technique that can be used to classify and predict data. The random forest technique, on the other hand, creates decision trees from data samples, extracts predictions from each, and then votes on the best alternative.

How the random forest algorithm works:

The random forest algorithm is performed in the following basic steps:

1. Select N records at random from the dataset.
2. Create a decision tree using these N entries as input.
3. Repeat steps 1 and 2 with the number of trees you want in your algorithm.
4. In the case of a regression problem, each tree in the forest predicts a value for Y for a new record (output). The ultimate value can be computed by averaging all of the anticipated values from all of the trees in the forest. Alternatively, each tree in the forest forecasts the category to which the new record belongs in the case of a classification challenge. Finally, the new record is given to the category that receives the greatest number of votes.

Why we Choose random forest (Advantages):

1. Because there are several trees and each tree is educated on a portion of data, the random forest technique is not biased. The random forest algorithm, in essence, relies on the "crowd's" power; as a result, the algorithm's overall bias is decreased.
2. This algorithm is quite dependable. Even if a new data point is added to the dataset, the overall process is unaffected because while new data may change one tree, it is extremely unlikely to affect all trees.
3. When you have both categorical and numerical variables, the random forest technique performs effectively.
4. When data has missing values or has not been scaled properly, the random forest technique performs well (although we have performed feature scaling in this article just for the purpose of demonstration).

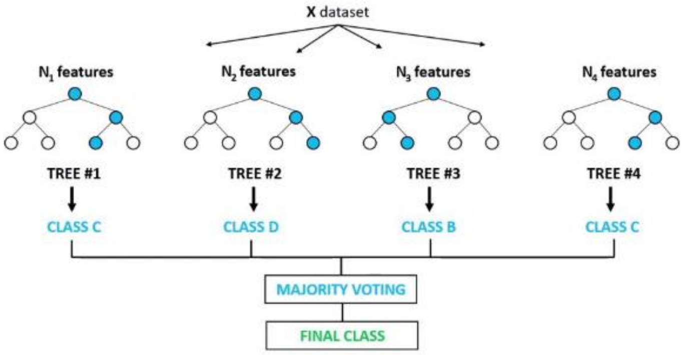


Fig -2: Random Forest

Use Random Forest for Regression:

1. Import Libraries
2. Importing Dataset
3. Preparing Data for Training
4. Feature Scaling
5. Training the Algorithm
6. Evaluating the Algorithm

1.Import Libraries

Execute the following code to import the necessary libraries:

```
import pandas as pd
import numpy as np
```

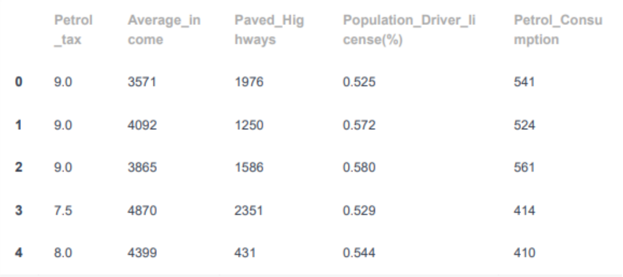
2. Importing Dataset

Execute the following command to import the dataset:

```
dataset =pd.read_csv('D:\Datasets\petrol_consumption.csv')
```

To get a high-level view of what the dataset looks like, execute the following command:

```
 dataset.head()
```



	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Consumption
0	9.0	3571	1976	0.525	541
1	9.0	4092	1250	0.572	524
2	9.0	3865	1586	0.580	561
3	7.5	4870	2351	0.529	414
4	8.0	4399	431	0.544	410

3. Preparing Data For Training :

The first task is to divide data into 'attributes' and 'label' sets. The resultant data is then divided into training and test sets.
The following script divides data into attributes and labels:

```
X = dataset.iloc[:, 0:4].values y = dataset.iloc[:, 4].value
```

Divide the data into training and testing sets:

```
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, random_state=0)
```

4.Feature Scaling

We recognize that our dataset is not yet scaled; for example, the Average Income field contains values in the hundreds, but Petrol tax has values in the tens. As a result, we should scale our data (albeit, as previously said, this step isn't as critical for the random forests approach). To accomplish so, we'll utilize the StandardScaler class from Scikit-Learn. To do so, run the following code:

```
# Feature Scaling

from sklearn.preprocessing import StandardScaler sc = StandardScaler() X_train = sc.fit_transform(X_train) X_test = sc.transform(X_test)
```

5. Training The Algorithm

we have scaled our dataset, it is time to train our random forest algorithm to solve this regression problem. Execute the following code:

```
fromsklearn.ensemble import RandomForestRegressor regressor = RandomForestRegressor(n_estimators=20, random_state=0)
Random Forest Algorithm with Python and Scikit-Learn

regressor.fit(X_train,y_train)
y_pred =regressor.predict(X_test)
```

The RandomForestRegressor class of the sklearn.ensemble library is used to solve regression problems via random forest. The most important parameter of the RandomForestRegressor class is the n_estimators parameter. This parameter defines the number of trees in the random forest. We will start with n_estimator=20 to see how our algorithm performs.

5. Evaluation the Algorithm

The last and final step of solving a machine learning problem is to evaluate the performance of the algorithm. For regression problems the metrics used to evaluate an algorithm are mean absolute error, mean squared error, and root mean squared error. Execute the following code to find these values:

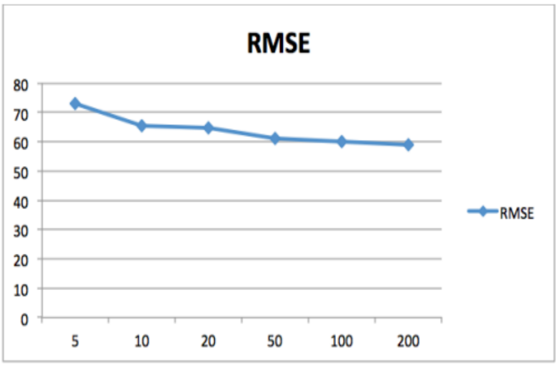
```
from sklearn import metrics

print('MeanAbsoluteError:',metrics.mean_absolute_error(y_test, y_pred))

print('MeanSquaredError:',metrics.mean_squared_error(y_test, y_pred))
print('RootMeanSquaredError:', np.sqrt(metrics.mean_squared_error(y_test, y_pred
```

The output will look something like this:

```
Mean Absolute Error: 51.765
Mean Squared Error: 4216.16675
Root Mean Squared Error: 64.932016371
```



3. LINEAR REGRESSION

Linear regression is one of the most basic types of regression in machine learning. In a linear regression model, a predictor variable and a dependent variable are connected linearly. Multiple linear regression models are used when there are several independent variables in the data. The equation below represents the linear regression model:

Y=mx + c +e where m denotes the line's slope, can intercept, and e the model's error.

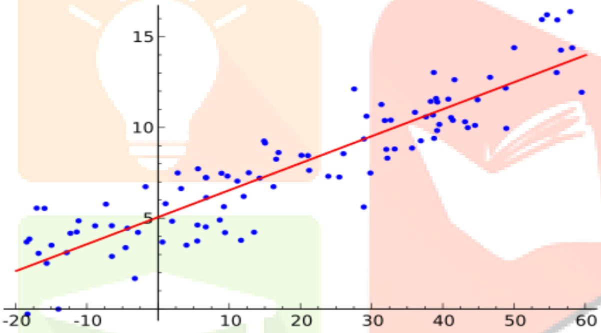


Fig -3: Linear regression

4. RIDGE REGRESSION

This is a different type of regression used in machine learning when the independent variables have a strong connection. This is because in the case of multi collinear data, least square estimations yield unbiased findings. However, there may be some bias if the collinearity is very great. As a result, the Ridge Regression equation now includes a bias matrix. This is a powerful regression approach that minimizes the risk of overfitting.

The Ridge Regression is represented by the equation below, where the inclusion of (lambda) solves the multicollinearity problem: (XTX + *I)-1XTy = (XTX + *I)-1XTy

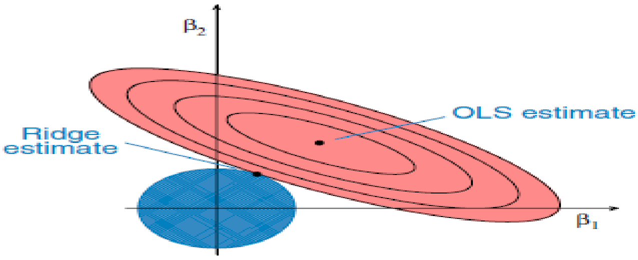


Fig -4: Ridge regression

5. LASSO REGRESSION

Regularization and feature selection are both performed using Lasso Regression, a type of machine learning regression. It prohibits the absolute size of the regression coefficient. As a result, the coefficient value approaches 0 in a way that Ridge Regression does not. As a result, feature selection is used in Lasso Regression to build the model, allowing for the selection of a set of features from the dataset. In Lasso Regression, just the necessary attributes are used, while the rest are assigned to zero. This keeps the model from becoming too tight. Lasso regression picks only one independent variable and compresses the rest to zero when the independent variables are significantly collinear.

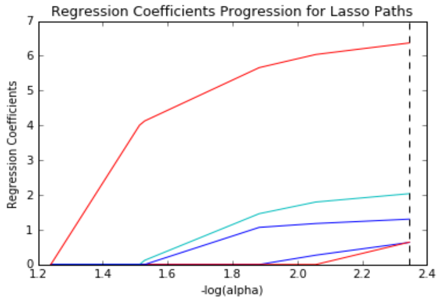


Fig -5: Lasso regression

The Lasso Regression approach is represented by the equation below:

$$\beta = (X^{T}X + \lambda*I)^{-1}X^{T}y$$

3. CONCLUSIONS

In this research, we'd like to investigate several existing machine learning and statistical methodologies for predicting rainfall throughout a region. We'll also investigate the possibility of combining ensembles of multiple machine learning algorithms to achieve high prediction accuracy.

In this research, we are aiming to deal with rainfall forecasting, which is a vital component of human life and provides the most significant resource of human life, Fresh Water.

REFERENCES

[1]. Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." In INFORMS International Conference on Service Science, pp. 139-147. Springer, Cham, 2019.

[2]. Zhao, Yi, Noemi Vergopolan, Kathy Baylis, Jordan Blekking, Kelly Caylor, Tom Evans, Stacey Giroux, Justin Sheffield, and Lyndon Estes. "Comparing empirical and survey-based yield forecasts in a dryland agro-ecosystem."Agricultural and Forest Meteorology 262 (2018): 147-156.

[3]. Majumdar J, Ankalaki S. Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves. In: Paper presented at international conferenceon computational science and engineering. 2016

[4]. Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity— A review article. In: IJCST. 2011;