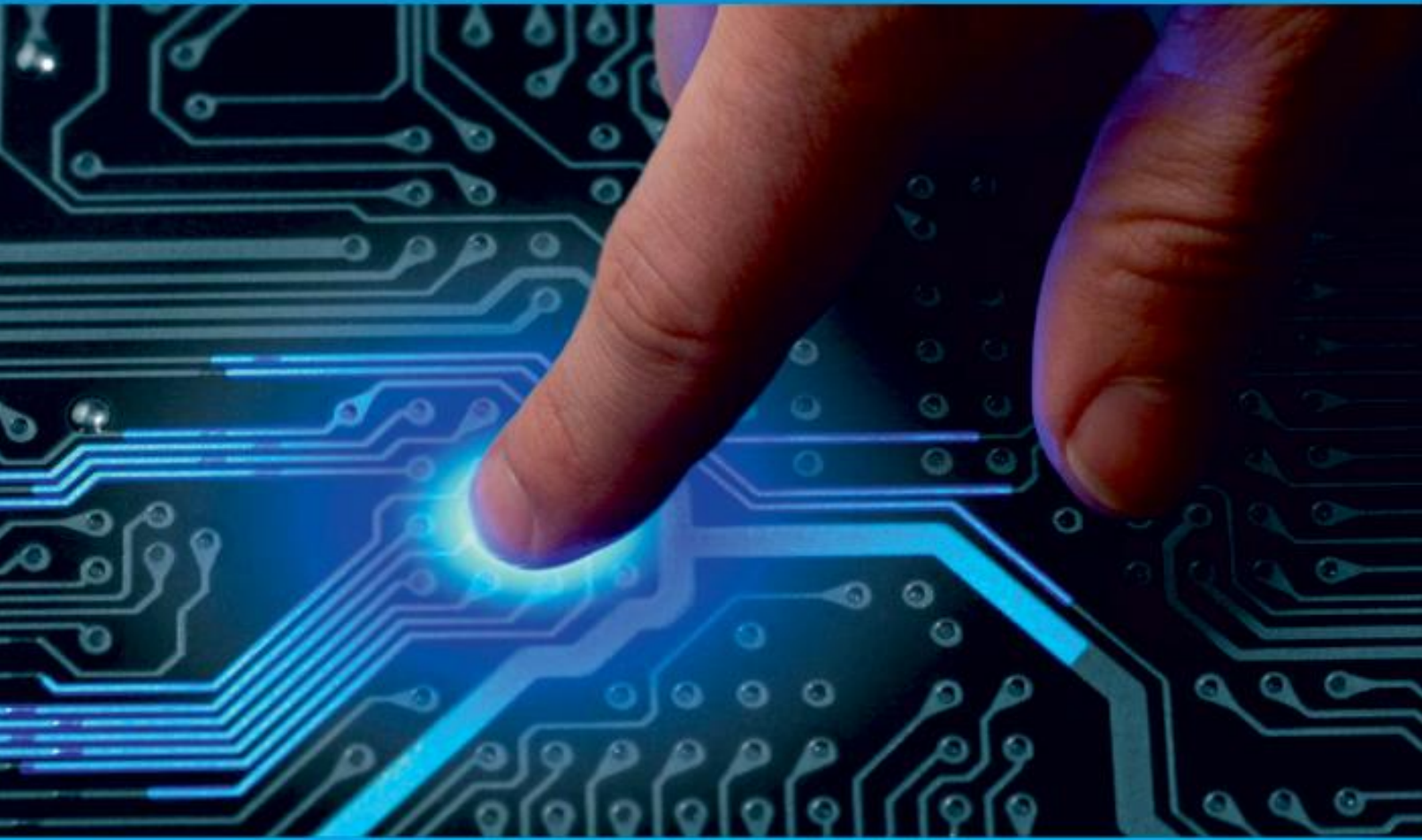




IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Special Issue 1, February 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165

 9940 572 462

 6381 907 438

 ijircce@gmail.com

 www.ijircce.com

Exploratory Analysis of Rainfall Data in India for Agriculture

Mr.M.Ashok kumar¹, B.Sanjukthaa², M.Shiva Ramya³, A.Soniya⁴, K.Sujeetha Bai⁵

Assistant Professor, Department of Electronics and Communication Engineering, Adhiyamaan College of Engineering,
Hosur, Krishnagiri, Tamilnadu, India¹

UG Scholars, Department of Electronics and Communication Engineering, Adhiyamaan College of Engineering,
Hosur, Krishnagiri, Tamilnadu, India^{2,3,4,5}

ABSTRACT- Agriculture is an backbone for the Indian economy. For agriculture, the foremost vital issue is water supply, i.e. rainfall. The prediction of the number of precipitation offers alertness to farmers by knowing early they will defend their crops from rain. So, it's vital to predict the precipitation accurately the maximum amount as attainable. Exploration and analysis of information on precipitation over numerous regions of India and particularly the regions wherever agricultural works are done persistently during a big selection. The study identified that the generated reliable rules with decision tree algorithms are important and efficient for future rainfall prediction with maintaining high accuracy. For agriculture purpose, the most important one is the water source, i.e. rainfall. With the help of analysis and the resultant data, future rainfall prediction for those regions using various machine learning techniques such as XGBoost classifier, SVM classifiers, multiple linear regression, lasso regression, Artificial neural networks, Decision tree, Naive bayes classifier, Logistic regression.

KEY WORDS: XGBoost classifier, SVM classifiers, Decision tree, Naive bayes classifier, Logistic regression.

I. INTRODUCTION

Rainfall prediction is vital as serious precipitation will lead to several disasters. The prediction helps folks to require preventive measures and what is more the prediction ought to be correct. There are 2 forms of prediction: short term precipitation prediction and future precipitation. Prediction, principally short term prediction, will offer United States correct results. The most challenge is to make a model for future precipitation prediction. Heavy precipitation prediction may be a major downside for the planet science department as a result of it is closely related to the economy and lifelong of humans. Rainfall prediction is essential since this variable is the one of the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches etc., These incidents have affected society for years. Rainfall also determines how fast a crop will grow from seed, including when it will be ready for harvesting. A proper balance of rainfall and proper irrigation can lead to faster-growing of plants and then cut down on germination time and the length between seeding and harvesting. Therefore, having an appropriate approach for rainfall prediction makes it possible to do agricultural processes in a calculated way. To solve this uncertainty, we used various machine learning techniques, artificial intelligence methods and models to make accurate rainfall timely predictions.

II. OBJECTIVE

The main objective is to protect the crop from the Rainfall and improves crop production which helps the Farmers to get more yield. Farmers are finding it difficult to cope with this change primarily because of the crops are seasonal and rainfall dependent.

III. LITERATURE REVIEW

Agriculture in India "gamble with monsoon", short term growth rate is linked to annual rainfall, Rainfall is one of the most complex and difficult elements of the hydrology cycle to understand the model due to complexity of the atmospheric processes that generate rainfall and the tremendous range of variation over a wide range of scales will be both in space and in time. Heavy rainfall prediction is a major problem for meteorological department as it is closely associated with the economy and life of human. Heavy rainfall may be a cause for natural disasters like flood and less rainfall leads to drought which are encountered by the people across the globe every year. Accuracy of rainfall data

forecasting has greatest importance for countries like India ,singapore etc.,whose economy is largely dependent only on agriculture. Due to dynamic nature of the earths atmosphere, Statistical techniques are failed to provide good accuracy for rainfall prediction forecasting.

IV. PLANNED TECHNIQUES

Accuracy/error of the prediction

Step1:

Importing the rainfall data sets to common source vector file.

Step2:

Fill the missing values with mean value of the rainfall data.

Step3:

scaling the rainfall data to a fixed mean value.

Step4:

Feature Reduction- PCA is used to minimize the data.

Step5:

The rainfall data is divided into two sets.one is training set (70%) and another is testing set (30%).

Step6:

Multiple Linear Regression algorithm, Artificial neural networks, Support Vector Regression and Lasso Regression is applied and the Mean Absolute Error, r2 score is calculated.

Step7:

The scatter plots are plotted between predicted and testing data for the applied models and the errors are compared and best model among them is selected.

Step 8:

Getting the output and displaying it as a result.

A.Data Collection

Rainfall is an important part of the hydrological cycle and metrological cycle. One of the first steps in any hydrological and meteorological study is accessing related quality of rainfall data. However, precipitation data is frequently incomplete .At first we collecting the data from metrological station .then the data will be consist of data ,location ,Min_term ,Max_ term etc., , achange in the measurement site, a change in data collectors, the irregularity of measurement in rainfall data

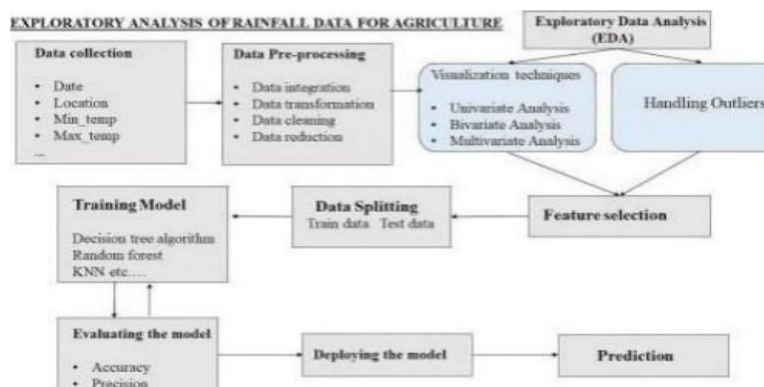


Fig 1:Block Diagram

B.Data Preprocessing

Rainfall data preprocessing .Data preprocessing to implementing models to evaluating them. Data pre-processing steps include imputing missing values, feature transformation, encoding categorical features, feature scaling and feature selection. We implemented models such as Logistic Regression, Decision Tree, K Nearest Neighbour and Random Forest Classifier. For evaluation purpoweatherses, we used Accuracy, Precision, Recall, F-Score and Area Under Curve as evaluation metrics. For our experiments, The smoothed data is then used to feed into each forecasting model.The process of transforming raw data into an understandable format. The targeted value of k corresponds to the optimal model performance in terms of RMSE.3.3.2. PCAPCA is employed in two ways: one for reduction of the dimensionality or preventing collinearity (depending on Eq. (2)); second for noise reduction by choosing leading

components (contributing most of the variance of the original rainfall data) to reconstruct rainfall series (depending on Eq. (7)). The major task involved in this Data preprocessing are

- a) Data cleaning
- b) Data integration
- c) Data reduction
- d) Data transformation

The percentage of total variance (see Eq. (6)) is set at three horizons, 85%, 90%, and 95% for principal component selection. SSA Data cleaning is nothing but process of removal of incorrect, incomplete data, inaccurate data, also replaces missing data. This approach of filtering a time series to retain desired modes of variability is based on the idea that the predictability of a system can be improved by forecasting the important oscillations in timeseries taken from the system. To ensure effective rainfall prediction, input datasets went through the exploratory data analysis by which chained equations algorithm was used to replace missing data, outliers were removed from the datasets and normalized before the classification stage. The general procedure is to filter the original record first and then to build the forecasting model based on the filtered series. De Silva et al. the arithmetic mean method, the normal ratio (NR) method, and the inverse distance method, linear and nonlinear methods to estimate missing rainfall data.

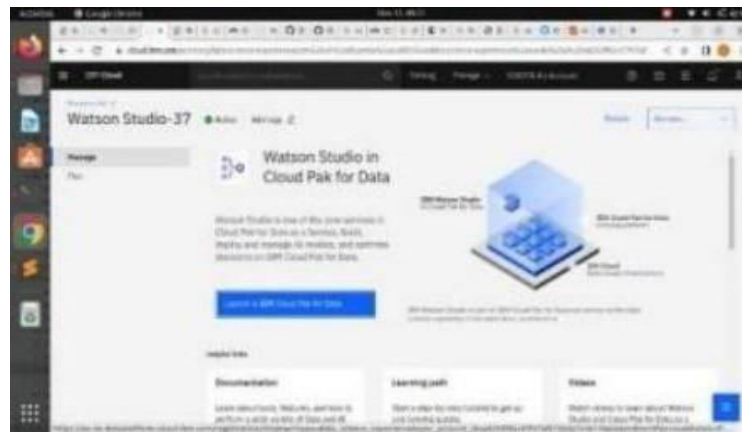


Fig 2: IBM Watson 37

C. Handling Missing Values

The estimation of missing data in hydrological and metrological studies is necessary for timely implementation of projects such as dam or canal construction. Placing of missing values, we can replace with "NA", even can replace with mean values and even with median values. sometimes replaced with most probable values. missing values can be filled into two ways. one is manual (small), and another one is automatic (more efficient, large data sets). missing data present various problems. they are

Reduces statistical power

The lost data can cause bias in estimation of parameters

It can reduce the representativeness of the samples

It may complicate the analysis of study

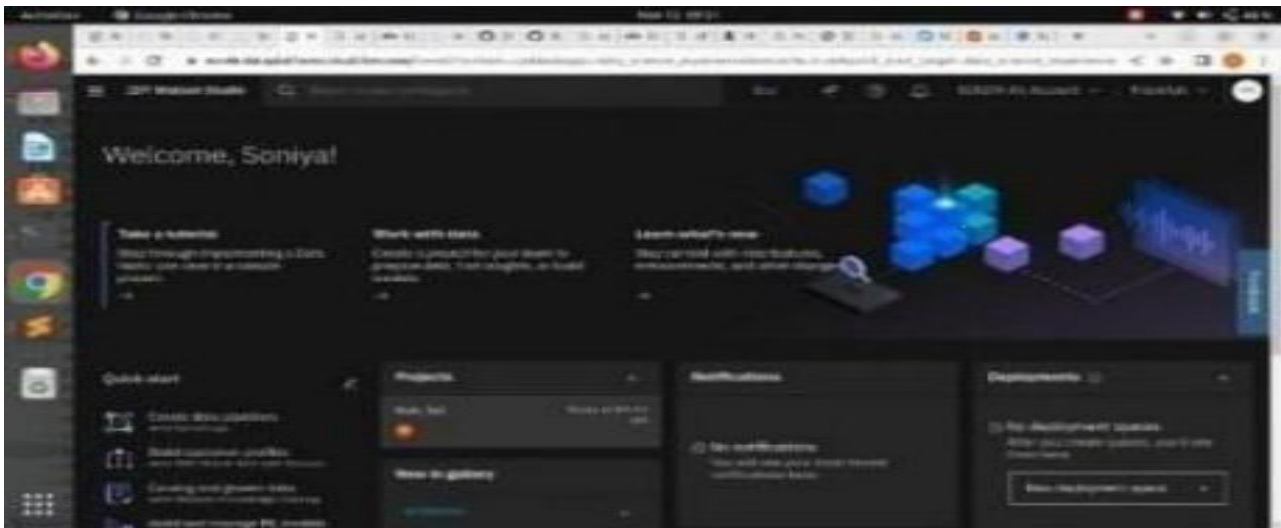


Fig 3: Data platform

D.Data Platform

The literature shows several methods for identifying traditional rainfall dataset approach trends. However, statistical exploratory data analysis using Kendall equation and graphical data trend analysis are the two widely used and simplest tests in data trend analysis.

Rainfall events StartfrneNumber (mm/h) Rainfallintensity

(yyyy mmdd)	lh:mm) of nages		Max	Avg.
10	13:00	22	107.4	9.14
20161230	04:20		70.1	
201 70105	4:30	10		10.4
20171008	16:40	9	1068	
201 71220	13:30	19	113.5	9.4
201 71226	6:30		114.5	147
201 80204	3:40	6	106.3	
201 8021 1	4:40	7	89.6	10.6
20130307 R9	1 3:30	14	117B	13.6

Fig 4: Rainfall events

Checking Data Homogeneity

For statistical analysis rainfall data from a single series should ideally possess property of homogeneity..The principle of homogeneity of dimensions states that the dimensions of all the terms in a physical expression should be the same. Rainfall dataset for multiple series at neighbouring stations should ideally possess spatial homogeneity. Tests of homogeneity are mainly required for validation purposes and there is a shared need for some of the tests with other climatic changing variables. Testing of rainfall data are therefore described in other Modules as follows:



.77	.11		.33	.55		.33
.11			.33	0.11	11	
.110,3	011,33	0.33	-0.33	oss,11	,33	.55,33
	0.11,11	11	-0.33		-0.11	
.0.11			,.33	.11	-0.11	

0.77	0.11	.11	.33	.55	0.11	33
-0.11	1.0	-0.11	.33	11	111	0.11
0.11	.11	1.0	-0.33	11	.11	.55
u33	.33	-0.33		033	33	.33
	.11	.11	.33	1.00	0.11	.11
-0.11	11	-0.11		011	1.00	-0.11

Fig 5: Intensity values

Module 9 Secondary validation of rainfall data

A. Spatial homogeneity testing

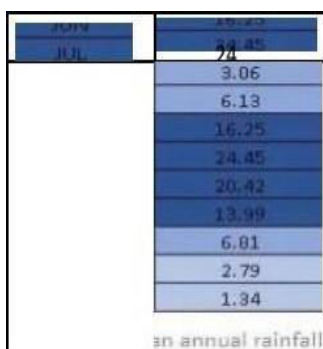
B.Consistency tests using double mass curves

Module 10 Correcting and completing rainfall data A. Adjusting rainfall data for long-term systematic shifts — double mass curves.

Module 17 Secondary validation of climatic data Single series tests of homogeneity, including trend analysis, mass curves, residual mass curves, Student's t and Wilcoxon W- test on the difference of means and Wilcoxon-Mann-Whitney U-test to investigate if the sample are from same population.

B. Multiple station validation including comparison plots, residual series, regression analysis and double mass curve one by one to make a brand new set of a similar variety of data for analysing through homogeneity.

Annual rainfall by subdivision



Month	% of Rainfall
	1.32
	1. so
MAR	1.93
1. APR	2. 3.06 6.13

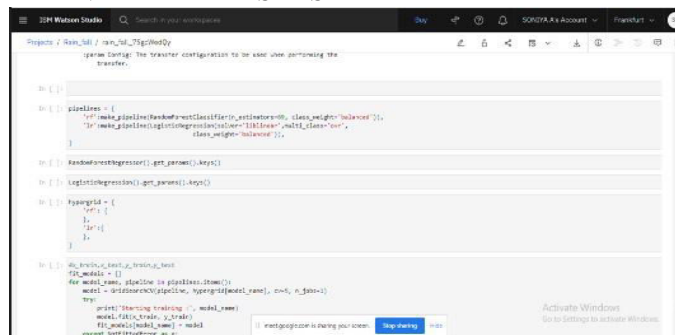
The following is a heat map plotted based on sum of rainfall received by each subdivision for all these years. The subdivisions with large area represents high rainfall and with small boxes represent less rainfall. of India have received more rainfall compared to central India.



Fig 7 : % of Rainfall in a month

V. RESULT

RAINFALL DATASETS



```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LogisticRegression

# Load the dataset
df = pd.read_csv('rainfall.csv')

# Check the shape of the dataset
print(df.shape)

# Check the first few rows of the dataset
print(df.head())

# Check the distribution of the target variable
print(df['rainfall'].describe())

# Split the data into training and testing sets
X = df[['temp', 'humidity', 'wind_speed']]
y = df['rainfall']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize the features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Train the Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict the rainfall for the test set
y_pred = model.predict(X_test)

# Calculate the R-squared value
r2 = r2_score(y_test, y_pred)
print('R-squared value: ', r2)

# Train the Random Forest Regressor model
model = RandomForestRegressor()
model.fit(X_train, y_train)

# Predict the rainfall for the test set
y_pred = model.predict(X_test)

# Calculate the R-squared value
r2 = r2_score(y_test, y_pred)
print('R-squared value: ', r2)

# Train the Logistic Regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict the rainfall for the test set
y_pred = model.predict(X_test)

# Calculate the R-squared value
r2 = r2_score(y_test, y_pred)
print('R-squared value: ', r2)
    
```

Fig 6: Annual rainfall

VI. CONCLUSION

The main purpose of this project work is to find the best prediction model i.e. The best machine learning technique which will accurately predict the rainfall which will be useful in the agricultural process. Determining the rainfall for effective use of water resources, crop productivity, and preplanning of water structures. Comparing with other algorithms and building the prediction model with the most efficient algorithm among them. To develop a web based application that may help in accurately predicting the rainfall which is useful in agricultural process. Currently machine learning is used in no. industries. As the data increases the complexity of that data will increase and for that we are using machines for the better understanding of that data .

REFERENCES

1. Bardossy and E. J. Plate. Spacetime model for daily rainfall using atmospheric circulation patterns, Water resources Research:1247_1259.
2. S. P.Chales. B. C. Bates, S. P. Charles, B. C. Bates, 1. N. Smith, and J. P. Hughes. Spacetime model for daily rainfall using atmospheric circulation patterns. Hydrological Processes, 18:13731394, 2004.
3. Brahmananda Rao ,K. Hada 1994:An experiment with linear regression in forecasting of spring rainfall over south Brazil K.Hrona Filzmoserb and K. Thompsonc 2009 Linear regression with compositional explanatory variables.

BIOGRAPHY



Mr.M.Ashokkumar,M.E.,
Assistant Professor,
Electronics and Communication Engineering Department,
Adhiyamaan college of Engineering,
Anna University.



A.Soniya,
Bachelor of Engineering(student),
Adhiyamaan college of Engineering,
Anna University.



B.Sanjukthaa,
Bachelor of Engineering(student),
Adhiyamaan college of Engineering,
Anna University.



M.Shiva Ramya,
Bachelor of Engineering(student),
Adhiyamaan college of Engineering,
Anna University.



K.Sujeetha Bai,
Bachelor of Engineering(student),
Adhiyamaan college of Engineering,
Anna University



INNO  SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

 **doi**[®]
CROSS **ref**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details