Exploratory Analysis of RainFall Data in India for Agriculture

A PROJECT REPORT

TEAM ID:PN2022TMID30456

Submitted by
In partial fulfillment for the award of the degree
Of
SAMAYA.S(611419106049)
ANITHA.M(611419106007)
DHANALAKSHMI.S(611419106021)
ARTHI.D(61141910612)

–

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
MAHENDRA ENGINEERING COLLEGE FOR WOMEN
KUMARAMANGALAM

CONTENTS

1. INTRODUCTION

1.1. Project Overview

India is an agricultural country and secondary agro based market will be steady with a good monsoon. The economic growth of each year depends on the amount of duration of monsoon rain, bad monsoon can lead to destruction of some crops, which may result in scarcity of some agricultural products which in turn can cause food inflation, insecurity and public unrest. In our analysis we are trying to understand the behavior of rainfall in India over the years, by months and different subdivisions.

Agriculture is the backbone of the Indian economy. For agriculture, the most important thing is water source, i.e ., rainfall. The prediction of the amount of rainfall gives alertness to farmers by knowing early they can protect their crops from rain. So, it is important to predict the rainfall accurately as much as possible. Exploration and analysis of data on rainfall over various regions of India and especially the regions where agricultural works have been done persistently in a wide range. With the help of analysis and the resultant data, future rainfall prediction for those regions using various machine learning techniques such as Logistic Regression, Linear Regression, Catboost Classifier etc.

PRE-REQUISTIES

Anaconda Installation:

Anaconda is a distribution of the Python and R programming languages for scientific computing that aims to simplify package management and deployment. The distribution includes datascience packages suitable for Windows, Linux, and macOS. Developed and maintained byAnaconda.

Founded

in 2012 by Peter Wang and Travis Olyphant. As Anaconda, also known as Anaconda Distribution or Anaconda Individual Edition, the company's other products include hisAnaconda Team Edition and Anaconda Enterprise Edition, neither of which are free.

WAY TO INSTALL ANACONDA:

STEP 1: Download and Anaconda
STEP 2: Install the Anaconda
STEP 3: Click I Agree
STEP 4: Choose the Installation Location
STEP 5: Installing the Requiring packages
STEP 6: Setting up the base environment
STEP 7: Successfully Installed and check the Anaconda Navigator working or not

Python packages installation:

Step 1: Open the anaconda navigator in the start menu

Step 2: Open the CMD.exe prompt
Step 3: Install the NUMPY package
To enter the numpy package enter the command in the CMD.exe Command: Pip install numpy
Numpy:
This package is used to perform numerical computations. This package comes pre-installed with Anaconda. NumPy is used for manipulating arrays. NumPy stands for Numerical Python.
Step 4: Install the pandas package.
To enter the pandas package enter the command in the CMD.exe Command: Pip install pandas
Pandas:
Pandas is one of the most widely used Python libraries for data science. It provides powerful and easy-to-use structure and data analysis tools. This package comes pre- installed with Anaconda. An open source library built on top of the NumPy library. A Python package that provides various data structures and operations for working with numerical data and time series. Mainly, it's common for data to be imported and analyzed much easier. Pandas is fast, providing users with high performance and productivity.
Step 5: Install the Matplotlib package.
To enter the Matplotlib package enter the command in the CMD.exe Command: Pip install Matplotlib
Matplotlib:
Matplotlib is a comprehensive library for creating static, animated and interactive visualizations in Python. This package comes pre-installed with Anaconda. Matplotlib is a nice visualization library in Python for 2D plotting of arrays. Matplotlib is a cross-platform data visualization library based on NumPy arrays and designed to work with the wider SciPy stack. Introduced by John Hunter in 2002.
Step 6: Install the Scikit-learn package.
To enter the Scikit- learn package enter the command in the CMD.exe Command: Pip install Scikit-learn
Scikit-learn:
This is a machine learning library for the Python programming language. This package comes pre-installed with Anaconda. Scikit Learn in Python is primarily used to focus on modeling in Python. It was only focused on modeling, not loading data.
Step 7: Install the Flask package.
To enter the Flask package enter the command in the CMD.exe Command: Pip install Flask
Flask:
Flask is a lightweight WSGI web application framework Flask is a web application framework written in Python. It is developed by Armin Ronacher, who leads an internationalgroup of Python enthusiasts called Pocco. Flask is based on the WSGI toolkit tools and the Jinja2 template engine. Both are Pocco projects.
1.2 Purpose
The main aim of objective is to find the

- Rainfall Prediction is the application of science and technology to predict the amount of

rainfall over a region.

• It is important to exactly determine the rainfall for effective use of water resources, crop productivity and pre-planning of water structures.

LITERATURE SURVEY

1.2. Existing Problem

Climate is important aspect of human life. So, the Prediction should accurate as much as possible. In this paper we try to deal with the prediction of the rainfall which is also a major aspect of human life, and which provide the major resource of human life which is Fresh Water. Fresh water is always a crucial resource of human survival – not only for the drinking purposes but also for farming, washing and many other purposes. Making a good prediction of climate is always a major task because of the climate change.

Now climate change is the biggest issue all over the world. Peoples are working on to detect the patterns in climate change as it affects the economy in production to infrastructure. So as in rainfall also making prediction of rainfall is a challenging task with a good accuracy rate. Making prediction on rainfall cannot be done by the traditional way, so scientist is using machine learning and deep learning to find out the pattern for rainfall prediction.

A bad rainfall prediction can affect the agriculture mostly framers as their whole crop is dependent on the rainfall and agriculture. It is always an important part of every economy. So, making an accurate prediction on the rainfall. There are number of techniques are used of machine learning, but

accuracy is always a matter of concern in prediction made in rainfall.

There are number of causes made by rainfall affecting the world ex. Drought, Flood, and intense summer heat etc. And it will also affect water resources around the world.

1.3. References

Spatial analysis of Indian
Markand
Oza
Understanding the variability in
Summer monsoon Rainfall
C.M.Kishtawal
rainfall, analysis of IndianSummer
(Mar 26,2014)
monsoon rainfall using Spatial
resolution.
Climate impacts on Indian
K.Krishna
kumar
Presents about the analysis of
Agriculture.
K.Rupa
Kumar

Crop-climate relationships for
(16 June,2004)
R.G.Ashrit
India, using historical predictions.
N.R.Deshpande
J.W.Hansen
Exploratory data Analysis of Indian Rainfall Data
Anusha Gajinkar
This Study shows that, India has two monsoon rainfall season one
is northwest monsoon and
second
one is southeast monsoon.
1.4. Problem Statement Definition
 Climate is a important aspect of human life. So, the Prediction should accurate as much as possible. In
this paper we try to deal with the prediction of the rainfall which is also a major aspect of human life
and which provide the major resource of human life which is Fresh Water. Fresh water is always a
crucial resource of human survival – not only for the drinking purposes but also for farming,
❖ Making a good prediction of climate is always a major task now a day because of the climate change.
❖ Now climate change is the biggest issue all over the world. Peoples are working on to detect the patterns
in climate change as it affects the economy in production to infrastructure. So as in rainfall also making prediction of rainfall is a challenging task with a good accuracy rate. Making prediction on
rainfall cannot be done by the traditional way, so scientist is using machine learning and deep learning to
find out the pattern for rainfall prediction.
❖ A bad rainfall prediction can affect the agriculture mostly framers as their whole crop is depend on
the rainfall and agriculture is always an important part of every economy. So, making an accurate prediction
of the rainfall.somewhat good
2. IDEATION AND PROPOSED SOLUTION
2.1. Empathy Map Canvas
2.2. Ideation and Brainstorming
2.3. Proposed Solution
S.No.
Parameter

Description

1.

Problem Statement

Climate is a important aspect of human life. So, the Prediction should accurate as much as possible. In this paper we try to deal with the prediction of the rainfall which is also a major aspect of human life and which provide the major resource of human life which is Fresh Water.

• Now climate change is the biggest issue all over the world. Peoples are working on to detect the patterns in climate change as it affects the economy in production to infrastructure.

2.

Proposed Solution

Analyzing the previous 10 years data can give us a rough idea about Rainfall pattern. Using Data Science, we can predict the Rainfall up to some good extent.

3.

Uniqueness

• This application is useful for the beginners in agriculture.

• Seed maturity selection features are available.

4.

Social Impact

• Different types of crops can be planted for good health. • Helps in producing healthy crops and good fields.

5.

Business Model

This comparative study is conducted concentrating on the following aspects: modeling inputs, Visualizing the data, modeling methods, and pre-processing techniques. The results provide a comparison of various evaluation metrics of these machine learning techniques and their reliability to predict rainfall by analyzing the weather data. We will be using classification algorithms such as Decision tree, Random forest, KNN, and xgboost

6.

Scalability

• When we predict rainfall correctly,it helps growth of crop and yielding will be better.

2.4. Proposed Solution Fit

3. REQUIREMENT ANALYSIS

3.1. Functional Requirements

FR No.

Functional Requirement (Epic)

Sub Requirement (Story / Sub-Task)

FR-1

Import necessary packages

Import necessary packages Importing packages like NumPy, pandas,seaborn, etc

FR-2

Download and load dataset

Download the dataset Load the Appropriate dataset
FR-3
Pre-processing of data
Making data suitable for building a good model
FR-4
Building Machine learning model
Choose the best algorithm. Check for the best optimised result.
FR-5
Train the data
Train the model using training data.
FR-6
Test the mode
Test the model for the best evaluation and analysing..
3.2. Non-Functional Requirements
FR No.
Non-Functional Requirement
Description
NFR-1
Usability
The usability of the website is to make all users will be satisfied with our requirements of the product.
The user should reach the summarized text or result with one button press if possible
NFR-2
Security
The security of the project is to develop the website that prevents SQL injection attack, XSS attack and DOS attack
NFR-3
Reliability
The reliability of the system is to make sure the website does not go offline.
The users can be reach and use program at any time, so maintenance should not be big issue.
NFR-4
Performance
The performance of the website isto provide data to all users without unnecessary delay and provide 24*7
availability.
NFR-5
Availability
The availability of the website is that the website will be active on
The Internet and people will be able to browse to it.
NFR-6
Scalability

The scalability of the system is we have limited our project to Indian cities
We have plans to scale it to continent's level in coming updates.

4. PROJECT DESIGN

4.1. Data Flow Diagrams

4.2. Solution and Technical Architecture

SOLUTION ARCHITECTURE

TECHNICAL ARCHITECTURE

S.No	Component	Description	Technology
1.	Website	User interacts with the prediction model through website to predict the rainfall data	HTML, CSS, JavaScript
	Model	ModelModelModelModelModel	
2.	Cloud Database	The model is provided with data from IBM cloud database	IBM Cloud DB, ibm_db(python package)
3.	API	Used to extend the service to other applications	Flask Application
4.	JWT & Sessions	It is used for Handling JSON web tokens (signing, verifying, decoding)	PyJWT, Flask-Sessions
5.	Machine Learning Model	This model is developed to predict the rainfall using ML algorithms.	Sklearn, Algorithms - DT & MLR
6.	Data processing	Data is pre-processed and	

then used for prediction.
Pandas, Numpy, Matplotlib
7.
File Storage
File storage requirements
IBM Block Storage or Other Storage Service
or Local Filesystem
4.3. User Stories

6.1 Sprint Planning & Estimation
Sprint
Functional Requirement (Epic)
User Story Number
User Story / Task
Story Points
Priority
Team Members
Sprint-1
Rainfall Prediction ML Model (Dataset)
USN-1
Weather Dataset Collection, Data preprocessing, Data Visualization.
5
High
J.Murugavasan , B.Rohith
Sprint-1
USN-2
Train Model using Different machine learning Algorithms
5
High
S.Sakthivel , M.Suresh
Sprint-1
USN-3
Test the model and give best
10
High
J.Murugavasan , R.Mohamed Yousuf
Sprint-2
Registration
USN-4
As a user, they can register for the application through Gmail. Password is set up.
5

Medium
S.Sakthivel , R.Mohamed Yousuf
Sprint-2
Login
USN-5
As a user, they can log into the application by entering email & password
5
Medium
B.Rohith , M.Suresh
Sprint-2
USN-6
Credentials should be used for multiple systems and verified
4
Medium
S.Sakthivel , J.Murugavasan
Sprint-2
Dashboard
USN-7
Attractive dashboard forecasting live weather
6
Low
R.Mohamed Yousuf , B.Rohith
Sprint-3
Rainfall Prediction
USN-8
User enter the location, temperature,
10
High
M.Suresh , R.Mohamed Yousuf
humidity
Sprint-3
USN-9
Predict the rainfall and display the result
10
High
J.Murugavasan , B.Rohith
6.2 Sprint Delivery Schedule
Sprint
Total Story Points
Duration
Sprint Start Date

Sprint End Date (Planned)
Story Points Completed (as on
Planned End Date)
Sprint Release Date (Actual)
Sprint-1
20
6 Days
31Oct 2022
05 Nov 2022
20
05 Nov 2022
Sprint-2
20
6 Days
05 Nov 2022
10 Nov 2022
20
10 Nov 2022
Sprint-3
20
6 Days
10 Nov 2022
15 Nov 2022
20
15 Nov 2022
Sprint-4
20
6 Days
15 Nov 2022
21 Nov 2022
20
21 Nov 2022

7.CODING AND SOLUTIONING

7.1Feature-1: Model Building

For this feature we have made use of Jupyter notebook which uses Python programming language. To use Jupyter Notebook install Anaconda, which is a desktop graphical user interface (GUI)
included in Anaconda® Distribution that allows you to launch applications and manage conda packages, environments, and channels without using command line interface (CLI) commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, macOS, and Linux. It provides all basic necessary python libraries which

are needed for Data Analysis and Visualizations.
Below images are source code for this feature:

In the above image, we import all necessary libraries needed for data exploration, preprocessing, model building and saving it. The below image specifies the values present in the dataset.
The below image specifies types of features and its count along with number of missing values in the dataset.
The lines 6 is used to drop rows which have high count missing values.
The above code displays the correlation between the columns present in the dataset.
The above code shows the distance plot and box plot of continuous features.
The above code removes null values from continuous features.
The above code removes null values by replacing it with Mode value.
The above code makes use of Label Encoding technique, which is used to convert labels into machine readable numeric values.
The above image is used to remove the remaining null values.
The above image is used to find values which lies outside the Inter-Quartile Range of each continuous feature. After finding the lower and higher bound, we remove the outliers from each continuous feature.
The above image shows the boxplot of each continuous feature after removing the outliers.
We split the dataset into independent and dependent variables. Here we must predict 'RainTomorrow', hence it will be the dependent variable and Date columns are unnecessary columns hence we drop it. And all other columns are independent variables. Using RobustScaler, we perform feature scaling to normalize the independent variables such that the standard distribution results to zero and standard deviation to one. This also removes remaining outliers in the independent
variables.
Now using 'train_test_split', we split the variables into train and test variables for each variable.
SMOTE (Synthetic Minority Oversampling Technique) is used to increase the number of test cases in a balanced way to avoid overfit cases.
The algorithm chosen here to build the model is CatBoostClassifier. CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters.
The above image shows the Confusion Matrix, Accuracy Score and Classification report.
The above image shows the roc curve and roc accuracy score for the built model.
The above image shows the Hyperparameter and Cross Validation score of the model.
Finally save the model using joblib library.
4.4. Feature-2:
4.5. User Interface
4.6. Index.html:

```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
<title>Weather App using Flask in Python</title>
<link rel="stylesheet"
href="https://cdn.jsdelivr.net/npm/bootstrap@4.6.1/dist/css/bootstrap.min.css">
<style>
body {
background-image: url('https://www.worldatlas.com/r/w768/upload/7e/2e/5a/untitled-design-
79.jpg');
background-repeat: no-repeat;
background-attachment: fixed;
background-size: cover;
}
</style>
</head>
<body>
<div class="container">
<br><br><br>
<div class="row"><h2 style="color:Blue;">Weather Prediction App</h2></div>
<br>
<div class="row">
<b style="color:Tomato;">Get weather details of any city around the world.</b>
</div>
<div class="row">
{% block content %}
<form action="{{ url_for("index")}}" method="post">
<div class="form-group">
<label style="color:Red;" for="Email">Email:</label><br>
<input type="email" id="Email" name="Email" value="{{Email}}" placeholder="Email"
required><br>
<label style="color:blue;" for="cityName"><b>Password:</b></label><br>
<input type="password" id="password" name="password" value="{{password}}"
placeholder="password" required><br>
<label for="cityName"><b style="color:Yellow;">City Name:</b></label><br>
<input type="text" id="cityName" name="cityName" value="{{cityName}}" placeholder="City
Name" required><br>
<br>
<button class="submit">Find</button>
```

```
{% if error is defined and error %}
<br><br><span class="alert alert-danger">Error: Please enter valid city name.</span></br>
{% endif %}
</div>
{% endblock %}
{% if data is defined and data %}
<table class="table table-bordered">
<thead>
<tr>
<th>Country Code</th>
<th>Coordinate</th>
<th>temperature</th>
<th>Pressure</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td class="bg-success">{{ data.sys.country }}</td>
<td class="bg-info">{{data.coord.lon }} {{data.coord.lat}}</td>
<td class="bg-danger">{{data.main.temp }} k</td>
<td class="bg-warning">{{data.main.pressure}}</td>
<td class="bg-primary">{{data.main.humidity}}</td>
</tr>
</tbody>
</table>
{% endif %}
</div>
</div>
</body>
</html>
```

App.py

```
from flask import Flask, request, render_template
import requests
from flask import Flask, request, render_template
import requests
app = Flask(_name_)
@app.route('/', methods=["GET", "POST"])
def index():
weatherData = ''
error = 0
```

```
cityName = ''
if request.method == "POST":
cityName = request.form.get("cityName")
if cityName:
weatherApiKey = '3f5d38932ad9ae0caa0302a35fbc8496'
url = "https://api.openweathermap.org/data/2.5/weather?q=" + cityName + "&appid=" +
weatherApiKey
weatherData = requests.get(url).json()
else:
error = 1
return render_template('index.html', data=weatherData, cityName=cityName, error=error)
if _name_ == "_main_":
app.run()
app = Flask(_name_)
@app.route('/', methods=["GET", "POST"])
def index():
weatherData = ''
error = 0
cityName = ''
if request.method == "POST":
cityName = request.form.get("cityName")
if cityName:
weatherApiKey = '3f5d38932ad9ae0caa0302a35fbc8496'
url = "https://api.openweathermap.org/data/2.5/weather?q=" + cityName + "&appid=" +
weatherApiKey
weatherData = requests.get(url).json()
else:
error = 1
return render_template('index.html', data=weatherData, cityName=cityName, error=error)
if _name_ == "_main_":
app.run()
```

TESTING

4.7. Test Cases

4.8. User Acceptance Testing

8.2.1. Defect Analysis

8.2.2. Testcase Analysis

5. RESULTS

5.1. Performance Metrics

9.1.1. Machine Learning

S.No.

Parameter
Values
Screenshot
1.
Metrics
Classification Model: Confusion Matrix - Accuracy Score- Classification Report -
2.
Tune the Model
Hyperparameter Tuning –
Validation Method -
9.1.2. Artificial Intelligence
S.No.
Parameter
Values
Screenshot
1.
Model Summary
-
2.
Accuracy
Training Accuracy -
Validation Accuracy -

6. ADVANTAGES AND DISADVANTAGES

6.1. Advantages

- Farmers can know when to plant or harvest their crops

People can choose where and when to take their holidays to take advantages of good weather

- Surfers known when large waves are expected

Regions can be evacuated if hurricanes or floods are expected

- Aircraft and shipping rely heavily on accurate weather forecasting
- It will help the farmers to take precautionary steps
- Technological solutions to improve their production

6.2. Disadvantages

- Weather is extremely difficult to forecast correctly
- It is expensive to monitor so many variables from so many sources
- The computers needed to perform the millions of calculations necessary are expensive
- The weather forecasters get blamed if the weather is different from the forecast

Leading to poor growth and overall health of crop

- Limited Foods Access

7. CONCLUSION

The weather prediction has become one of the most essential entities now a days. To improve the risk management systems and to know the weather in coming days in an automatic and in

scientific way, many models have been
emerging to assist in weather Prediction. In this paper, we have seen building a Weather Prediction Web Application from scratch by making use of 6 different ML algorithms namely CatBoost Classifier, RandomForset Classifier, Logistic Regression, GaussianNB, KNN and XGB Classifier. In the result section, the results from the all the six models and its results such as Accuracy, Error rate, mean absolute error, Root mean squared error, Relative squared error, Root relative squared error and time taken to build the model are tabulated. The results show that the CatBoost Classifier and XGB Classifier has output the results of high accuracy than all the other classifiers that were used. When coming to the time taken to build the model, The CatBoost Classifier outperforms all the other classifiers in solving the Problem under scrutiny.

8. FUTURE SCOPE

In upcoming future updates, the WEATHER FORECASTING application will have additional features such as:

- Live Location tracking
- News on Live Disasters
- Weather Forecast for next one week
- Will deploy as android app
- Help in predicting which crop will be best suited according to weather conditions

13.APPENDIX

1.1. Source Code

13.1.1. Ipynb file Link: RAINFALL PREDICTION

13.1.2. UI Link: FILE

1.2 Links

13.2.1. GITHUB

13.2.2. DEMO VIDEO