Timely and Accurately Predict Rainfall by using Ensemble Predictive Models

Veera Ankalu.Vuyyuru, Giduturi.Apparao, S. Anuradha

Research Scholar, CSE Department, GITAM Deemed To be University, Visakhapatnam, AP, India.
Professor, CSE Department, GITAM Deemed To be University, Visakhapatnam, AP, India.
Assistant Professor, CSE Department, GITAM Deemed To be University, Visakhapatnam, AP, India.

veeraankalu14@gmail.com, apparao.giduturi@gitam.edu, asesetti@gitam.edu

Abstract. Weather and rainfall are important factors for human life. By depend upon the rainfall agriculture, horticulture harvesting, and goods transportation. These all are goods and supply chain processes. If in supply chain process breaks any point, Ultimately, the farmer will get the loss. Timely predicting rainfall helps the farmers and agriculture and horticulture stock manage to maintain people require India’s coastal area. For these reasons, this paper proposes the Ensemble Models (Catboost, Boost). Most of the authors are working on rainfall prediction using statistical models. Using statistical models to analyze and predict a huge amount of data is very difficult, depending upon the features. But using Ensemble models is likely to boost up the elements, and apply the classification to prediction makes it easy. This paper discusses and Compares the statistical decision tree model with ensemble models to find out the difference between the characteristics of algorithms and how they impact the timely predict the rainfall.

1 Introduction

The weather conditions around the world are changing quickly and continuously. Right predictions are central in today's everyday life. From agriculture to industry, from travel to regular transport, we are heavily dependent on weather forecasting. As the world is suffering from continuous climate change and its side effects, it is very important to forecast weather without error to ensure simple and smooth mobility and healthy day-to-day operations. [1].

In India, 30-50 percent of wheat yield variation depends on climate variability, and yields are especially susceptible to adverse climate conditions. Seasonal rainfall and water scarcity, particularly during spring, are the main causes of India’s different crops. [2].

In this regard need to collect the data from the metrology department of rainfall to apply to the statistical and ensemble models. Here need to know the statistical model and ensemble model. The method of applied statistic analysis to the data set is statistical modeling. A gathering (or mathematical model) of observed data is a statistical model. When data analysts use different statistical models for their analysis, the information can be more strategically understood and interpreted. This practice helps them classify connexions between variables, make projections of the future datasets, and visualize rather than select the raw data[3].

An analyst must accumulate or restore data in a folder, cloud, social network or plain-excellence file before a statistical model is used. Analysts would need to have a sound interpretation of the data system and administration, including the processing, storage, and retention of data. Many employed in this sector should also express their enthusiasm for facts and proof and recognise the basics of data exploitation [4]. Once the time has come to analyze the data, analysts can choose to use a range of statistical models. According to Mello, the most common techniques fall into the following two groups: supervised learning, including regression and classification models, and unsupervised learning, including clustering algorithms and association laws[3][5].

The Ensemble Model integrates multiple models, which have been designed using the same or different algorithms, to create a single model for use. Created with methods such as bagging boosters
and Bayesian model averages, the ensembles seem to be incredibly complex yet tend to surpass their component models with new results. [3][7].

If uncertainty is measured by feature rather than structure, that location of the razor is restored rather than how it appears. The sophistication of a black-book modelling procedure calculates by re-sampling a GDF system rather than, say, counting parameters. A series of experiments on a two-dimensional decision tree topic have shown that the bagging of many trees has less GDF difficulty than a single tree part, avoiding the paradox of ensembles’ generalisation. [3].

2 Literature Survey

To study environmental changes on the earth depends upon the different attributes or parameters. These things technical termed as a feature of the object. Here the item is nothing but a thing that required hidden patterns and classified to determine the future. So in this section study, the rainfall predictions apply the statistical models and choose the models’ way and accuracy.

Ahmad et al. undertake research as a research study (1) developing satellite remote-sensing estimates of maize acreage in Pakistan’s typical maize area, (2) defining a statistical-empirical maize yield forecasting or (3) finally evaluating the impact of temperature on the inter-annual variability of maize yield over a decade. Classification models were validated around the region of the sample by 200 randomly chosen field check points. The effects of maize mapping work utilises Landsat’s multi-temporal uniform vegetation difference index (NDVI) and surface soil (LST) data to assess interannual maize yields. The yield predictors were selected through a primary variable screening and fed into the least absolute shrinkage and regression model (LASSO) [7][13].

Chul-Min Ko et al. The research aimed to enhance the predictive precision of intense rainfall, using a machine learning technique to predict the hydrological effect. In this analysis, machine learning with the XGBoost technique uses to correct the quantitative precipitation forecast (QPF) given by the Korea Meteorological Administration (KMA) for the production of a hydrological quantitative precipitation forecast (HQPF) for flood modelling [9][14].

Diez-Sierra et al. Eight mathematical and machine learning methods were used to estimate long-term daily precipitation in semi-arid climates (Tenerife, Spain). Cross-validation to recreate at 17 gauges 36 years of regular rainfall results. It’s independent for each indicator. Compared to the observed data, the reconstructed sequence chooses the optimum hyperparameters for each model unit. The model efficiency evaluation utilises multiple rainfall rate and regular, weekly and annual aggregate scales. [10][20].

Elder et al. Analyze a modelling community of models trained by various approaches or algorithms to generate the final predictions. We will see in this chapter that clusters will surpass single-algorithm models. This discovery, though, appears to refute Occam Razor’s theory that the most basic answer is usually the correct one. We would discuss this obvious contradiction and prove that such uncertainty is a positive thing.

Lazriet al. Discuss the precipitation forecasts of MSG (Meteosat Second Generation) using a multi-classifying machine-based learning model. Learn and test a multi-class model by comparing MSG satellite data and radar data. The outcomes of the classification calculated six certitude coefficients. Tier 2 classification is accomplished by the Random Forest Classifier (RF2), taking as input parameters the trust coefficients. Six types of precipitation strength gain exceptionally high intensity of precipitation, mild to high intensity of precipitation, mean heavy precipitation, low to medium intensity of precipitation, low intensity of precipitation and no precipitation. The comparisons between the findings of the multi-classifier and the results from independently used classification models indicate a substantial increase in classification accuracy [11].

Lei Xu et al. Discuss seasonal precipitation predictions in regional or local areas to help direct farming and urban water resource management. The North American Multi-Model Ensemble (NMME) is a seasonal forecasting method offering global precipitation forecasts. Spatial analysis
reveals that the corrected NMME precipitation forecasts display the best South China ability with an average RMSE of about 30 mm, whereas the worst in Central and Southwest China with an 80 mm RMSE. Despite the correction, uncertainties of summer seasonal precipitation forecasts and severe wet cases are still high. WSVM and WRF methods can serve as an important tool in bias correction of NMME precipitation forecasts[12].

Some of the authors also discuss the statistical model and ensemble models and predict the model. The following sections are data set description ensemble models and heuristics of ensemble models algorithm flow, and mechanism and results discuss.

### 3 Rainfall Data Set Description

In Indian atmospheric conditions, rainfall is significant. Most of India depends upon agriculture and horticulture. Suppose the rain will be a help in the growth of agriculture productivity. In the Indian Economy, a significant part is the Agriculture and Horticulture. So for this reason, the thesis mainly focused on the rainfall predictions system. So In the system need to predict the rainfall range. The system requires data. This section discusses mostly the dataset's data set and features—the dataset collecting from the Indian Metrology data(IMD) data source. And data Set contains around 150000 records of data it is approximately 1960 to 2019 years of the monthly rainfall data of the Vishakhapatnam region, and it includes attributes like Location, Minimum Temperature (Degrees), Maximum Temperature (Degrees), Rainfall (Milli Meters), Evaporation, Sunshine, Industry, Relative Humidity, Pressure, Cloud Coverage.

The Sample Data set will consist mainly of numerical and categorical data. This particular data required to apply to a model is complicated. Why because this specific data needs to generalize and remove the noise. Sections, subsections and subsubsections

3.1.1 shows the sample data set with a description of the data set attributes.

| Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustSpeed | reRelativeHumidity | Pressure | Cloud | RainT | rain tomorrow |
|----------|---------|---------|----------|-------------|----------|---------------|-------------------|----------|-------|-------|---------------|
| VSKP     | 13.4    | 22.9    | 0.6      | 5.8         | 10.6     | 44            | 71                | 1007.7   | 8     | No    | No             |
| VSKP     | 7.4     | 25.1    | 0        | 5.8         | 6        | 44            | 44                | 1010.6   | NA    | No    | No             |
| VSKP     | 12.9    | 25.7    | 0.4      | 9.4         | 46       | 38            | 1007.6            | NA       | No    | No    | No             |
| VSKP     | 9.2     | 28      | 0        | 3.2         | 10.4     | 24            | 45                | 1017.6   | NA    | No    | No             |
| VSKP     | 17.5    | 32.3    | 1        | 4.8         | 5.1      | 41            | 82                | 1010.8   | 7     | No    | No             |
| VSKP     | 14.6    | 29.7    | 0.2      | 2.6         | 5        | 56            | 55                | 1009.2   | NA    | No    | No             |
| VSKP     | 14.3    | 25      | 0        | 3.2         | 9.5      | 50            | 49                | 1009.6   | 1     | No    | No             |
| VSKP     | 7.7     | 26.7    | 0        | 7.2         | 9.7      | 35            | 48                | 1013.4   | NA    | No    | No             |
| VSKP     | 9.7     | 31.9    | 0        | 8.8         | 11       | 80            | 42                | 1008.9   | NA    | No    | Yes            |
| VSKP     | 13.1    | 30.1    | 1.4      | 7           | 8.3      | 28            | 58                | 1007     | NA    | Yes   | No             |
| VSKP     | 13.4    | 30.4    | 0        | 8           | 11.9     | 30            | 48                | 1011.8   | NA    | No    | Yes            |
| VSKP     | 15.9    | 21.7    | 2.2      | 6           | 5.9      | 31            | 89                | 1010.5   | 8     | Yes   | Yes            |
The above data, the table consists of the details of the attributes are mention below:

- **MinTemp & MaxTemp**: Mintemp define as minimum temperature, and max temp defines as maximum temperature, and both measure on a scale of the degree in Celsius.
- **Rainfall**: The quantity of rain which is measure in mm for a day.
- **Evaporation**: It is a measure in the form of a millimeter. Evaporation calculation per day.
- **Sunshine**: The number of hours of bright daylight sunlight.
- **WindGustSpeed**: The velocity (km / h) in the 24 hours to midnight of the highest wind gust.
- **RelHumid**: It measures the relative moisture on the scale of the percentage. It calculates in the form of relative humidity in the air.
- **Cloud**: A cloud-obsured fraction of the sky, measured in "oktas," is a unit of eighths. It records how the cloud covers many eighths of the sky. A 0 measure indicates an apparent atmosphere while an eight suggests that it is completely overcast.
- **Rain Today Integer**: One if precipitation (mm) in the 24 hours to 9 am exceeds 1mm, otherwise 0.

### 4 Ensemble Models Are Applied Dataset

In this section, discuss the Ensemble model's functionalities flow of the algorithms on rainfall datasets. Here two algorithms taken for the predictions, one is CATBOOST and XGBOOST.

#### 4.1 System Architecture

![System Architecture Diagram](image-url)

**Figure 4.1.1 System Architecture**
4.2 CATBOOST

CatBoost is a new decision boosting gradient (GBDT) algorithm that can manage categorical features correctly. This algorithm varies on the following aspects from conventional GBDT algorithms [15].

![CATBOOST Flow Diagram](image)

**Figure 4.2.1** CATBOOST Flow Diagram

### 4.2.1 Functionalities of CATBOOST

It works with categorical features rather than preprocessing during training. CatBoost requires the whole testing dataset to be included. Goal statistics (TS) are a highly effective means of handling categorical functionality and minimal data loss. In fact, CatBoost conducts a random permutation for each event, computing an average mark value with, for instance, the same category value in a permutation before the specified category value [16]. Role variations. All the categorical characteristics are new. CatBoost uses a gullible approach to perceive the combinations when constructing a new tree break. No mix thinks about a first break, but in the second and subsequent split data sets, Catboost integrates both presets and categorical functions. Both divisions selected from the tree are treated as a two-value category and merged. Categorical features are developing without discrimination. By using the TS process, the distribution can differ from the initial distribution to convert categorical characteristics into numerical values. The variance of this distribution would trigger the answer to depart, which for standard GBDT approaches is an inevitable problem and establishes a new methodology for solving the gradient variation by theoretical analysis; Quick scorer. Speedy scorer. CatBoost utilises forgotten trees as base predictors for the same dividing parameters in the entire stage of the tree. These trees are balanced and less prone to be overcrowded. In ignorance trees, each leaf index encodes as a binary vector with a length equivalent to the depth of the branch. This principle uses CatBoost model assessor model predictions as all binaries use float, statistic and single-hot encoded capabilities. [17]. These moves are nothing but CATBOOST algorithms to test the
rainfall data predictions. The XGBOOST algorithm is another Ensemble type algorithm, as follows in the flow diagram and in functionality [18].

### 4.3 XGBOOST

XGBoost was initially proposed in 2011 and continuously refined and improved in many study follow-ups studies by Tianqi Chen and Carlos Guestrin. The paradigm is a method of learning based on the Boosting Tree model. For traditional boosting trees [20], only the first derivative details are used. Due to the rest of the former n-1 trees, dispersed teaching is not easy when training the nth tree. XGBoost conducts Taylor's second-order extension of the loss feature and can utilize CPU multithreading to perform parallel loading. XGBoost employs a variety of techniques to avoid overfitting. [22].

![XGBOOST Algorithm Functionalities as a Flow Diagram](image)

**Figure 4.3.1.1XGBOOST Algorithm Functionalities as a Flow Diagram**

#### 4.3.1 XGBOOST Functionalities

In this experiment, the regression model XGBoost trains with each target gene separately, with an input mark of 943 genes, suggesting that the input dimension was 943, which is very high. However some XGBoost techniques to prevent overpasses will mitigate overpasses and increase predictive regression accuracy[21][23].

The XGBoost design of the experiment has been adapted to guarantee optimal model performance:

1. **New estimators n estimators evaluate the amount of training iterations.** Too limited estimators can overlook fitness and leave the model incomplete. The estimator, though is typically too large and not secure, so it overflows.
2. **Crete's Min Weight**
The weight of the smallest leaf nodes is determined above by the weight of the smallest boy, to avoid excess.

3. Limit Profiling
It's the forest's full distance. The larger the tree heart, the more complicated the tree model and the better the ability, but the model is much simpler to overwrite.

4. Under-review
This parameter means the sampling rate of all samples.

5. Example (s)
The last parameter setup is a tree coal sample. When each tree forms, it is the sampling rate. The sampling rate of the hallmark gene for this mission is equal.

6. Rate for learning
The learning rate is an essential parameter for several XGBoost-related algorithms. It significantly affects the performance of the model. Each move's weight reduces to make the model longer lasting.

5 Experimental Results
The above section discusses how predictive ensemble models are working their functionalities. These ensemble models practically working is required. In this section, algorithms implemented using python and using the library's sklearn, NumPy, and matplotlib libraries implement the system. Python to implement the algorithms like Logistic Regression, Decision Tree, Random Forest, CATBOOST, XGBOOST[24].

5.1 DATA SET PREPROCESSING

Table 5.1.1 shows the sample data set as a training data set with attributes.

| Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustSpeed | relative humidity | Pressure | Cloud | RainToday | rainTomorrow |
|----------|---------|---------|----------|-------------|----------|---------------|-------------------|----------|-------|------------|--------------|
| VSKP     | 13.4    | 22.9    | 0.6      | 5.8         | 10.6     | 44            | 71                | 1007.7   | 8     | No         | No           |
| VSKP     | 7.4     | 25.1    | 0        | 5.8         | 6        | 44            | 44                | 1010.6   | NA    | No         | No           |
| VSKP     | 12.9    | 25.7    | 0        | 4           | 9.4      | 46            | 38                | 1007.6   | NA    | No         | No           |
| VSKP     | 9.2     | 28      | 0        | 3.2         | 10.4     | 24            | 45                | 1017.6   | NA    | No         | No           |
| VSKP     | 17.5    | 32.3    | 1        | 4.8         | 5.1      | 41            | 82                | 1010.8   | 7     | No         | No           |
| VSKP     | 14.6    | 29.7    | 0.2      | 2.6         | 5        | 56            | 55                | 1009.2   | NA    | No         | No           |
| VSKP     | 14.3    | 25      | 0        | 3.2         | 9.5      | 50            | 49                | 1009.6   | 1     | No         | No           |
| VSKP     | 7.7     | 26.7    | 0        | 7.2         | 9.7      | 35            | 48                | 1013.4   | NA    | No         | No           |
| VSKP     | 9.7     | 31.9    | 0        | 8.8         | 11       | 80            | 42                | 1008.9   | NA    | No         | Yes          |
| VSKP     | 13.1    | 30.1    | 1.4      | 7           | 8.3      | 28            | 58                | 1007     | NA    | Yes        | No           |
| VSKP     | 13.4    | 30.4    | 0        | 8           | 11.9     | 30            | 48                | 1011.8   | NA    | No         | Yes          |
| VSKP     | 15.9    | 21.7    | 2.2      | 6           | 5.9      | 31            | 89                | 1010.5   | 8     | Yes        | Yes          |
Table 5.1.1.2 shows the sample data set as a training data set with attributes.

| Min temp | MaxTemp | Rainfall | Evaporation | Sunshine | Wind Gust Speed | relative humidity | Pressure | Cloud |
|----------|---------|----------|-------------|----------|----------------|------------------|----------|-------|
| 13.4     | 22.9    | 0.6      | 5.8         | 10.6     | 44             | 71               | 1007.7   | 8     |
| 7.4      | 25.1    | 0        | 5.8         | 6        | 44             | 44               | 1010.6   | 6     |
| 12.9     | 25.7    | 0        | 4           | 9.4      | 46             | 38               | 1007.6   | 5     |
| 9.2      | 28      | 0        | 3.2         | 10.4     | 24             | 45               | 1017.6   | 6     |
| 17.5     | 32.3    | 1        | 4.8         | 5.1      | 41             | 82               | 1010.8   | 7     |
| 14.6     | 29.7    | 0.2      | 2.6         | 5        | 56             | 56               | 1009.2   | 7     |
| 14.3     | 25      | 0        | 3.2         | 9.5      | 50             | 49               | 1009.6   | 1     |
| 7.7      | 26.7    | 0        | 7.2         | 9.7      | 35             | 48               | 1013.4   | 5     |
| 9.7      | 31.9    | 0        | 8.8         | 11       | 80             | 42               | 1008.9   | 1     |
| 13.1     | 30.1    | 1.4      | 7           | 8.3      | 28             | 58               | 1007     | 9     |
| 13.4     | 30.4    | 0        | 8           | 11.9     | 30             | 48               | 1011.8   | 4     |
| 15.9     | 21.7    | 2.2      | 6           | 5.9      | 31             | 89               | 1010.5   | 1     |

The above tables discuss the training data table in the training data table consists of the attributes with the labels. Second, without a data table as a test data table after getting the data set to apply, the pre-processing techniques analyze the data.

5.1.1 DATA EXPLORATION

Range Index: 94773 entries, 0 to 94772

5.1.1.1 Data columns (total of 24 columns):
The above represents the data exploration consists of the Column name object types represent a data set as trained data set representation. It consists of 94773 entries of records and 24 columns of the dataset.

![Figure 5.1.1.1](image1.png)  
**Figure 5.1.1.1** Rain Tomorrow Indicator No(0) and Yes(1) after Oversampling (Balanced Dataset)

![Figure 5.1.1.2](image2.png)  
**Figure 5.1.1.2** Heat Map for the Trained data set
The Below Table shows how to evaluate the MICE package (Multiple Imputation by Chained Equations). Afterward, we will detect outliers using the Inter-Quartile Range and remove them to get the final working data set.

**Table 5.1.1.** The Inter-Quartile Range and remove them to get the final working data set

| Date  | 1495 |
|-------|------|
| Location | 6    |
| MinTemp | 8.4  |
| MaxTemp | 7.8  |
| Rainfall | 2.6  |
| Evaporation | 3.6  |
| Sunshine | 6.001871 |
| WindGustDir | 7    |
| WindGustSped | 18   |
| WindDir9am | 8    |
| WindDir3pm | 6    |
| WindSpeed9am | 13   |
| WindSpeed3pm | 11.53649 |
| Humidity9am | 24   |
| Humidity3pm | 32   |
| Pressure9am | 6.9  |
| Pressure3pm | 6.9  |
| Cloud9am | 4.039764 |
| Cloud3pm | 3.943048 |
| Temp9am | 7.4  |
| Temp3pm | 7.4  |
| rain today | 1    |
| RISK_MM | 5.8  |
| rain tomorrow | 1    |
Figure 5.1.3 Correlation Heat Map between Attributes

Figure 5.1.4 Attributes Maps
5.2 FEATURE SELECTION

Feature selection and feature enhance mechanisms to apply to the rainfall data set.

| Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustDir | WindGustSpeed | WindDir9am | WindDir3pm | WindSpeed9am | WindSpeed3pm | Humidity9am | Humidity3pm | Pressure9am | Pressure3pm | Cloud9am | Cloud3pm | Temp9am | Temp3pm | RainToday |
|------|-----------|---------|---------|----------|-------------|----------|-------------|--------------|------------|------------|-------------|-------------|------------|-------------|------------|-------------|----------|----------|--------|--------|-----------|
| 0    | 0.115284  | 0.03125 | 0.571429| 0.518987 | 0.435873    | 0.565373 | 0.468635    | 0.866667    | 0.521127   | 0.866667   | 0.933333    | 0.516057    | 0.615385   | 0.539969    | 0.212121   | 0.223214 | 0.277286 | 0.858482 | 0.409095 | 0.515957 |
| 1    | 0.115575  | 0.03125 | 0.400000| 0.574684 | 0.373192    | 0.561494 | 0.727678    | 0.933333    | 0.521127   | 0.400000   | 1.000000    | 0.108527    | 0.564103   | 0.236236    | 0.242424   | 0.309524 | 0.297935 | 0.278259 | 0.223735 | 0.523936 |
| 2    | 0.115866  | 0.03125 | 0.557143| 0.589873 | 0.373192    | 0.690262 | 0.775584    | 1.000000    | 0.549296   | 0.866667   | 1.000000    | 0.490587    | 0.666667   | 0.168740    | 0.292929   | 0.220238 | 0.324484 | 0.277728 | 0.167728 | 0.625000 |
| 3    | 0.116157  | 0.03125 | 0.451429| 0.648101 | 0.373192    | 0.579307 | 0.756129    | 0.266667    | 0.239437   | 0.600000   | 0.000000    | 0.286821    | 0.230769   | 0.247486    | 0.151515   | 0.517857 | 0.445428 | 0.237588 | 0.173953 | 0.547872 |
| 4    | 0.116448  | 0.03125 | 0.688571| 0.756962 | 0.477660    | 0.637367 | 0.384110    | 0.866667    | 0.478873   | 0.066667   | 0.466667    | 0.184939    | 0.512821   | 0.663712    | 0.323232   | 0.315476 | 0.244838 | 0.769158 | 0.670914 | 0.539894 |
|      |           |         |         |          |             | 0.744161 | 0.174811    | 0.015385    | 0.0        |            |             |            |             |             |             |            |         |         |        |        |          |
5.3 LOGISTIC REGRESSION ALGORITHMS IMPLEMENTATION

Table 5.3.1 Shows the Logistic Regression Algorithm Performance applies to the rainfall data set

| Parameters     | Performances |
|----------------|--------------|
| Accuracy       | 0.77         |
| ROC Area       | 0.75         |
| Cohen's Kappa  | 0.51         |
| Time Taken     | 1.37         |
| Time Taken     | 1.37         |

Table 5.3.2 Shows the Logistic Regression Algorithm Confusion Matrix apply to the rainfall data set

|       | precision | recall | f1-score | support |
|-------|-----------|--------|----------|---------|
| 0     | 0.79388   | 0.85497| 0.82329  | 15866   |
| 1     | 0.73959   | 0.6498 | 0.69179  | 10057   |
| macro avg | 0.76673   | 0.75238| 0.75754  | 25923   |
| weighted avg | 0.77282   | 0.77537| 0.77228  | 25923   |
5.3.1 Logistic ROC Curve

5.3.2 Confusion Matrix for Logistic

5.4 DECISION TREE

Table 5.4.1 Shows the Decision Tree Algorithm Performance applies to the rainfall data set
Table 5.4.2 Shows the Decision Tree Algorithm Confusion Matrix apply to the rainfall data set

|   | precision | recall   | f1-score | support |
|---|-----------|----------|----------|---------|
| 0 | 0.88726   | 0.85466  | 0.87065  | 15866   |
| 1 | 0.78327   | 0.82868  | 0.80533  | 10057   |
| accuracy |          | 0.84458  |          | 25923   |
| macro avg | 0.83527   | 0.84167  | 0.83799  | 25923   |
| weighted avg | 0.84692   | 0.84458  | 0.84531  | 25923   |
Figure 5.4.1 Decision Trees ROC Curve

Figure 5.4.2 Confusion Matrix for Logistic
5.5 RANDOM FOREST

Table 5.5.1 shows the RANDOM FOREST Algorithm Performance applies to the rainfall data set.

| Parameters       | Performances |
|------------------|--------------|
| Accuracy         | 0.92         |
| ROC Area under Curve | 0.92       |
| Cohen’s Kappa    | 0.84         |
| Time Taken       | 20.0         |

Table 5.5.1 Shows the RANDOM FOREST Algorithm confusion Matrix apply to the rainfall data set.

|          | precision | recall   | f1-score | support |
|----------|-----------|----------|----------|---------|
| 0        | 0.94714   | 0.93502  | 0.94104  | 15866   |
| 1        | 0.89951   | 0.91767  | 0.90850  | 10057   |
| accuracy |           |          | 0.92829  | 25923   |
| macro avg| 0.92332   | 0.92634  | 0.92477  | 25923   |
| weighted avg | 0.92866 | 0.92829  | 0.92842  | 25923   |

**Figure 5.5.1** Decision Trees ROC Curve. **Figure 5.5.2** Confusion Matrix for Decision Trees.
5.6 XGBOOST

Table 5.6.1 Shows the XGBOOST Algorithm Performance applies to the rainfall data set

| Parameters       | Performances |
|------------------|--------------|
| Accuracy         | 0.95         |
| ROC Area under   | 0.95         |
| Curve            |              |
| Cohen's Kappa    | 0.91         |
| Time Taken       | 105.97       |

Table 5.6.2 Shows the XGBOOST Algorithm confusion Matrix apply to the rainfall data set

| precision | recall   | f1-score | support |
|-----------|----------|----------|---------|
| 0         | 0.98714  | 0.94743  | 0.96368 | 8541    |
| 1         | 0.92310  | 0.97100  | 0.94644 | 5551    |

| accuracy  |          |          |         |
|-----------|----------|----------|---------|
| macro avg | 0.95180  | 0.95921  | 0.95506 | 14092   |
| weighted avg | 0.95789  | 0.95671  | 0.95689 | 14092   |
Table 5.7.1 Shows the CATBOOST Algorithm Performance applies to the rainfall data set
Table 5.6.2 Shows the CATBOOST Algorithm confusion Matrix apply to the rainfall data set

| precision | recall   | f1-score | support |
|-----------|----------|----------|---------|
| 0         | 0.97714  | 0.93743  | 8541    |
| 1         | 0.91310  | 0.97100  | 5551    |

| accuracy  |          |          |         |
|-----------|----------|----------|---------|
|           | 0.95671  | 14092    |         |

| macro avg | weighted avg |
|-----------|--------------|
| 0.94180   | 0.94689      |
| 0.94921   | 0.946671     |
| 0.95506   | 0.95689      |
| 14092     | 14092        |

5.8 MODEL COMPARISONS
Figure 5.8.1 Bar charts Compare the Models Accuracy of all the algorithms

Figure 5.8.2 Bar charts Compare the Models ROC Curve and kappa Accuracy of all the algorithms
6. Conclusion
In this paper, the objective is to find out the accurate predictions. These days weather and changes and rainfall information are very important. By dependence on rainfall, horticulture, and agriculture, they are farming very crucial. In this regard, to find out the accurate rain, rain is possible next hour or the next day. This type of weather scenario consists of a huge amount of data. This historical data is very important for trained data, So here machine learning algorithms trained this huge amount of data to analyze and predict heuristic algorithms. This work mainly discusses the statistical algorithms logistic, random forest and decision tree algorithms, and Ensemble algorithms CATBOOST and XGBOOST algorithms apply the rainfall prediction using these algorithms and find out every algorithm's performances. To find out accurately predicting the rainfall to compare all models. In this context, XGBOOST is the best performing algorithm to compare with the remaining algorithms.

References
1. Jakaria, A. H. M., Hossain, M. M., & Rahman, M. A. (2020). Smart weather forecasting using machine learning: a case study in tennessee. arXiv preprint arXiv:2008.10789.
2. Wang, B., Feng, P., Waters, C., Cleverly, J., Li Liu, D., & Yu, Q. (2020). Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia. Agricultural and Forest Meteorology, 291, 108043.
3. John Elder. The Apparent Paradox of Complexity in Ensemble Modeling, 291, 108043.
4. Tufaner, F., &Özbeyaz, A. (2020). Estimation and easy calculation of the Palmer Drought Severity Index from the meteorological data by using the advanced machine learning algorithms. Environmental Monitoring and Assessment, 192(9), 1-14.
5. Das, A., Khan, M. N., & Ahmed, M. M. (2020). Detecting lane change maneuvers using SHRP2 naturalistic driving data: a comparative study machine learning techniques. Accident Analysis & Prevention, 142, 105578.
6. Fullhart, A. T., Nearing, M. A., McGehee, R. P., &Weltz, M. A. (2020). Temporally downscaling a precipitation intensity factor for soil erosion modeling using the NOAA-ASOS weather station network. CATENA, 194, 104709.
7. Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., ... & Chau, K. W. (2020). Predicting standardized streamflow index for hydrological drought using machine learning models. Engineering Applications of Computational Fluid Mechanics, 14(1), 339-350.
8. Ahmad, I., Singh, A., Fahad, M., & Waqas, M. M. (2020). Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. Computers and Electronics in Agriculture, 178, 105732.
9. Ko, C. M., Jeong, Y. Y., Lee, Y. M., & Kim, B. S. (2020). The Development of a Quantitative Precipitation Forecast Correction Technique Based on Machine Learning for Hydrological Applications. Atmosphere, 11(1), 111.

10. Diez-Sierra, J., & del Jesus, M. (2020). Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. Journal of Hydrology, 124789.

11. Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J., Yu, X., ... & Zhou, H. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. Journal of Hydrology, 574, 1029-1041.

12. Li, Y., & Qin, Y. (2019). The Response of Net Primary Production to Climate Change: A Case Study in the 400 mm Annual Precipitation Fluctuation Zone in China. International Journal of Environmental Research and Public Health, 16(9), 1497.

13. Kuradusenge, M., Kumaran, S., & Zennaro, M. (2020). Rainfall-Induced Landslide Prediction Using Machine Learning Models: The Case of Ngororero District, Rwanda. International Journal of Environmental Research and Public Health, 17(11), 4147.

14. Jakariya, M., Alam, M. S., Rahman, M. A., Ahmed, S., Elahi, M. L., Khan, A. M. S., ... & Ali, M. S. (2020). Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. Science of the Total Environment, 742, 140255.

15. Jakariya, M., Alam, M. S., Rahman, M. A., Ahmed, S., Elahi, M. L., Khan, A. M. S., ... & Ali, M. S. (2020). Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. Science of the Total Environment, 742, 140255.

16. Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G., & Dineva, A. A. (2020). Susceptibility mapping of soil water erosion using machine learning models. Water, 12(7), 1995.

17. Pham, B. T., Le, L. M., Le, T. T., Bui, K. T. T., Le, V. M., Ly, H. B., & Prakash, I. (2020). Development of advanced artificial intelligence models for daily rainfall prediction. Atmospheric Research, 237, 104845.

18. Raza, A. S., Shoaib, M., Faiz, M. A., Baig, F., Khan, M. M., Ullah, M. K., & Zubair, M. (2020). Comparative Assessment of Reference Evapotranspiration Estimation Using Conventional Method and Machine Learning Algorithms in Four Climatic Regions.(CF Alexander Rabinovich, Ed.). Journal of Pure and applied geophysics, 30.

19. Bushra, P., Swapan, T., Susanta, M., Jayanta, M., Pritee, S., Towfiqul, I. A. R. M., & Atiqur, R. (2020). Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Scientific Reports (Nature Publisher Group), 10(1).

20. Chen, Y. J., Nicholson, E., & Cheng, S. T. (2020). Using machine learning to understand the implications of meteorological conditions for fish kills. Scientific Reports, 10(1), 1-13.

21. Neelakandan, S., & Paulraj, D. (2020). An automated exploring and learning model for data prediction using balanced CA-SVM. Journal of Ambient Intelligence and Humanized Computing, 1-

22. Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene Expression Value Prediction Based on XGBoost Algorithm. Frontiers in genetics, 10, 1077.
