International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-10 Issue-4, November, 2021

Rainfall Prediction using Machine Learning and Deep Learning Algorithms

B.Meena Preethi, R.Gowtham, S.Aishvarya, S.Karthick, D.G.Sabareesh

***Abstract:** The project entitled as "Rainfall Prediction using Machine Learning & Deep Learning Algorithms" is a research project which is developed in Python Language and dataset is stored in Microsoft Excel. This prediction uses various machine learning and deep learning algorithms to find which algorithm predicts with most accurately. Rainfall prediction can be achieved by using binary classification under Data Mining. Predicting the rainfall is very important in several aspects of one's country and can help from preventing serious natural disasters. For this prediction, Artificial Neural Network using Forward and Backward Propagation, Ada Boost, Gradient Boosting and XGBoost algorithms are used in this model for predicting the rainfall. There are totally five modules used in this project. The Data Analysis Module will analyse the datasets and finding the missing values in the dataset. The Data Pre-processing includes Data Cleaning which is the process of filling the missing values in the dataset. The Feature Transformation Module is used to modify the features of the dataset. The Data Mining Module is used to train the dataset to models using any algorithm for learning the pattern. The Model Evaluation Module is used to measure the performance of the model and finalize the overall best accuracy for the prediction. Dataset used in this prediction is for the country Australia. This main aim of the project is to compare the various boosting algorithms with the neural network and find the best algorithm among them. This prediction can be major advantage to the farmers in order to plant the types of crops according to the needy of water. Overall, we analyse the algorithm which is feasible for qualitatively predicting the rainfall.*

***Keywords:** Machine learning, Deep learning, Adaboost, Gradient Boosting, XGBoost, ANN*

I. INTRODUCTION

The project is entitled as "Rainfall Prediction using Machine Learning and Deep Learning Algorithms" used to predict the rainfall in the 49 cities of Australia. The main drawback of these type of project is that the prediction is not accurate. The predictions using various algorithms should be improved. Predicting the rainfall is very important in several aspects of a country and can help from preventing serious natural disasters.

The main objective of this project is

- ➢ To help the farmers and citizens of the country from the natural calamities like Tsunami, Floods, landslides, tsunamis, storms.
- ➢ This project can able to predict the rainfall of the respective cities in Australia. Then, it can able to find the best model since we use more models.
- ➢ Finally we display all the models with the accuracy percentage using data visualization like graphs, histogram, etc.

II. PROBLEM DESCRIPTION

- ➢ **Data analysis**
  - ○ Analyse the dataset and find the missing values
- ➢ **Data Pre-processing**
  - ○ Data Cleaning
- ➢ **Feature Transformation Module**
  - ○ Drop columns
  - ○ One hot encoding
  - ○ Scaling the data
  - ○ Modifying dataset
  - ○ Feature selection
- ➢ **Data Mining**

Data mining is frequently used for the KDD. Data mining is a part of the KDD-process, in which appropriate approach and algorithm is chosen and applied to the data set.

- ➢ **ANN**
- ➢ **ADABOOST**
- ➢ **GRADIENT BOOST**
- ➢ **XGBOOST**

A. Model Evaluation

To evaluate the performance of models, evaluation metrics are used.

Accuracy

$$\text{Accuracy} = \frac{TN + TP}{NP + FP + TP + FN}$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 score

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Manuscript received on November 16, 2021.
Revised Manuscript received on November 22, 2021.
Manuscript published on November 30, 2021.

* Correspondence Author

B.Meena Preethi, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

R.Gowtham*, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

S.Aishvarya, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

S.Karthick, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

D.G.Sabareesh, PG Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore (Tamil Nadu), India.

Retrieval Number: 100.1/ijrte.D66111110421
DOI: 10.35940/ijrte.D6611.1110421
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

B. Confusion Matrix:

To calculate the performance of machine learning model is the confusion matrix. Confusion Matrix is also called contingency table. It distinguishes between true positive, false positive, true negative and false negative predictions. Fig 1 shows how the confusion matrix is classified.

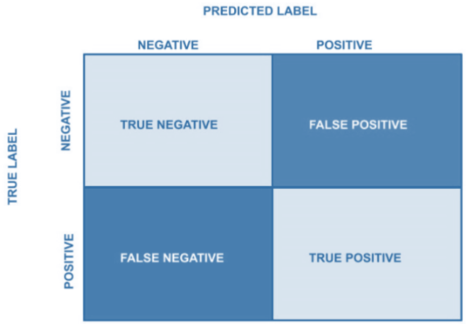


Fig 1 – Confusion matrix

III. PROPOSED METHODOLOGY

This project will be benefited for all the living people all over the country.

PHASE – I

In Phase – I, Ada Boost (or) Adaptive Boosting is used in the model. First the input data is checked and visualized. Then Data Pre-processing is done to ensure that there is no null values. If any null values is found in the dataset, it is replaced under three conditions which are mean, median and mode. After that process, Parameter Selection is done. In this process, the parameters which are not needed for the prediction is removed and one hot encoding is done for the categorical features. Then the dataset has to be Normalized using Standard Scalar. Finally the dataset is splitted into Train dataset and Test dataset. First the train dataset is trained to the model using Ada boosting algorithm. After the model is trained, test dataset is sent to the model after the RAIN TOMORROW feature is removed from the test dataset. Then using evaluation metrics, the model's accuracy is noted. If the accuracy metrics are satisfactory then stop the model after printing the scores.

WORKING OF ADABOOST ALGORITHM:

In Adaboost algorithm, initially the weights are assigned using 1/n formula. A decision tree with only one node and two leaves is called a stump. All the features are then converted into several individual stumps. Then Gini impurity is calculated for each and every stump and the execution start from the stump which has the least gini impurity. After that total error and performance of the selected stump is calculated. According to the performance of the stump, the weights are updated and normalized. Then find the bucket range of values and for each iteration a random value is selected from 0 to 1. The corresponding data is now transferred to another new dataset consists of same data type. This process is repeated until all the records in the dataset are passed to the stump and get trained.

PHASE – II

In Phase – II, Gradient Boosting Algorithm is used in this model. First the input data is checked and visualized. Then Data Pre-processing is done to ensure that there is no null values. If any null values is found in the dataset, it is replaced under three conditions which are mean, median and mode. After that process, Parameter Selection is done. In this process, the parameters which are not needed for the prediction is removed and one hot encoding is done for the categorical features. Then the dataset has to be Normalized using Standard Scalar. Finally the dataset is splitted into Train dataset and Test dataset. First the train dataset is trained to the model using gradient boosting algorithm. After the model is trained, test dataset is sent to the model after the RAIN TOMORROW feature is removed from the test dataset. Then using evaluation metrics, the model's accuracy is noted. If the accuracy metrics are satisfactory then stop the model after printing the scores.

WORKING OF GRADIENT BOOSTING:

But in Gradient Boosting algorithm, first it calculates the initial prediction. After finding the initial prediction, it is converted into probability. With the help of probability residual is calculated. Then decision tree is constructed using any feature as root node. After constructing the binary tree, similarity score is found for the reference of calculating the predicted score. Finally the predicted score (or) log(odds) prediction is calculated. Then the predicted score (or) log(odds) prediction is converted into probability. After that update the residual and build another decision tree. Repeat these steps until all data in the dataset passes into decision trees.

PHASE – III

In Phase – III, XGBoost is used in this model. First the input data is checked and visualized. Then Data Pre-processing is done to ensure that there is no null values. If any null values is found in the dataset, it is replaced under three conditions which are mean, median and mode. After that process, Parameter Selection is done. In this process, the parameters which are not needed for the prediction is removed and one hot encoding is done for the categorical features. Then the dataset has to be Normalized using Standard Scalar. Finally the dataset is splitted into Train dataset and Test dataset. First the train dataset is trained to the model using XGBoost algorithm. After the model is trained, test dataset is sent to the model after the RAIN TOMORROW feature is removed from the test dataset. Then using evaluation metrics, the model's accuracy is noted. If the accuracy metrics are satisfactory then stop the model after printing the scores.

WORKING OF XGBOOST ALGORITHM:

So in Extreme Gradient Boosting, first it calculates the initial prediction. After finding the initial prediction, it is converted into probability. With the help of probability residual is calculated. Then decision tree is constructed using any feature as root node. The calculation of output score and mark those values in the leaf.

After finding the output score, log(odds) for probability value is determined. Now log(odds) value for prediction is also calculated. Then convert the log(odds) for prediction into probability. Then plot the graph so that the residual is reduced. Similarly we predict for all values using the decision tree. We keep on building the trees until the residuals are super small. Then the final prediction is calculated.

PHASE – IV

In Phase – IV, Artificial Neural Network is used in the model. First the input data is checked and visualized. Then Data Pre-processing is done to ensure that there is no null values. If any null values is found in the dataset, it is replaced under three conditions which are mean, median and mode. After that process, Parameter Selection is done. In this process, the parameters which are not needed for the prediction is removed and one hot encoding is done for the categorical features. Then the dataset has to be Normalized using Standard Scalar. Finally the dataset is splitted into Train dataset and Test dataset. First we have to define the architecture of the neural network. Then the train dataset is trained to the model using Neural Network. After the model is trained, test dataset is sent to the model after the RAIN TOMORROW feature is removed from the test dataset. Then using evaluation metrics, the model's accuracy is noted. If the accuracy metrics are satisfactory then stop the model after printing the scores.

WORKING OF NEURAL NETWORK:

For Artificial Neural Network using Forward and Backward Propagation, first we need to define the architecture of the neural network. First the type of model is defined with the kernel initializer (or) initial weights are assigned to the neural network. Then activation function is also applied. The optimizer is initialized and loss function is chosen according to the problem whether it is classification (or) regression. Finally performance metrics is also specified. First the neural network uses forward propagation. Mostly Sigmoid activation function is used but relu activation function is used in the middle layers of the neural network. After the forward propagation, the loss is calculated using any one of the loss functions. Then using back propagation the weights are updated using chain rule of propagation. This process goes on till the loss is reduced to minimum.

ADVANTAGES

- This gives better performance to predict the rainfall.
- It helps farmers to adjust their farming according to the expected weather condition.
- It helps the air transportation.
- It can help to guide and encourage tourists to visit certain areas.
- Using neural networks, the performance of the model is more accurate.

TABLE – 1 COMPARISON OF ALGORITHM

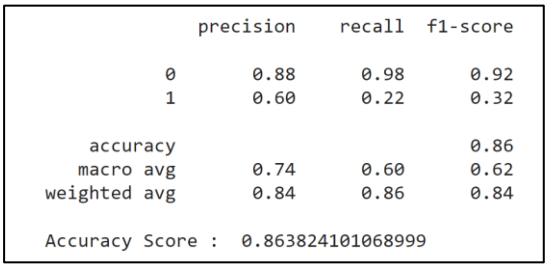
COMPARISON	ADA BOOST	GRADIENT BOOSTING	XGBOOST	ANN
INITIAL WEIGHTS OF THE MODEL	The weights are initially assigned	The initial prediction is calculated and converted to initial probability	The initial prediction is calculated and converted to initial probability	The kernel initializer (or) initial weights , activation function, loss function, optimizer is selected
RESIDUAL CALCULATION	A stump is created and gini impurity is calculated	Using probability, residual is calculated and decision tree is constructed	Using probability, residual is calculated and decision tree is constructed	The output of the final neuron is calculated with the bias
SIMILARITY SCORE OF THE MODEL	Total error of the stump is calculated	Similarity score is calculated for calculating the log(odds) prediction	Similarity score is calculated for calculating the log(odds) prediction	Then the activation function is Calculated
PERFORMANCE OF THE MODEL	Performance of the selected stump is calculated	Then the log(odds) prediction is converted to probability	Then the log(odds) prediction is converted to probability	After the forward propagation, the loss is calculated using any one of the loss function
UPDATE THE WEIGHTS	According to the performance of the stump, weights are updated, normalized and find the bucket range of values	Residuals gets updated and build another decision tree	Update the residuals using graphs and check whether the residual is getting smaller	Then using back propagation the weights are updated using chain rule of propagation
NUMBER OF ITRATIONS OF THE MODEL	For each iteration a random value is selected from 0 to 1. The data is now transferred to another new dataset. This process is repeated until all records are passed to the stump	Repeat these steps until all data is passed to the decision trees	Then another decision tree is built and this process goes until the residual is super small or max decision tree is attainted	For each epochs, one forward and backward propagation take place. This process goes on till the loss is reduced to minimum

Retrieval Number: 100.1/ijrte.D66111110421
DOI: 10.35940/ijrte.D6611.1110421
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

IV. RESULTS AND FINDINGS

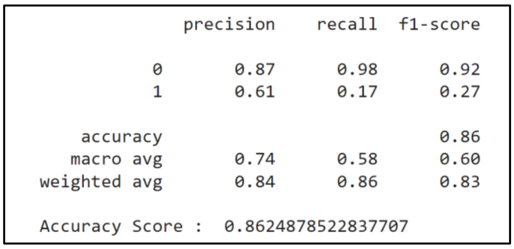
Ada Boost:



```
              precision    recall  f1-score

           0       0.88      0.98      0.92
           1       0.60      0.22      0.32

    accuracy                           0.86
   macro avg       0.74      0.60      0.62
weighted avg       0.84      0.86      0.84

Accuracy Score :  0.863824101068999
```

Fig 2. Performance of the model using AdaBoost

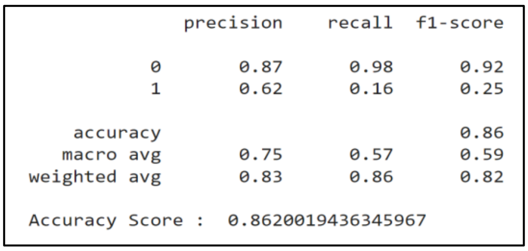
Gradient Boost:



```
              precision    recall  f1-score

           0       0.87      0.98      0.92
           1       0.61      0.17      0.27

    accuracy                           0.86
   macro avg       0.74      0.58      0.60
weighted avg       0.84      0.86      0.83

Accuracy Score :  0.8624878522837707
```

Fig 3. Performance of the model using Gradient Boosting

XGBoost:



```
              precision    recall  f1-score

           0       0.87      0.98      0.92
           1       0.62      0.16      0.25

    accuracy                           0.86
   macro avg       0.75      0.57      0.59
weighted avg       0.83      0.86      0.82

Accuracy Score :  0.8620019436345967
```

Fig 4. Performance of the model using XGBoost

Artificial Neural Network:

Table – II Artificial Neural Network

Loss	Accuracy	Val_loss	Val_accuracy
0.2830	0.8864	0.3237	0.8763

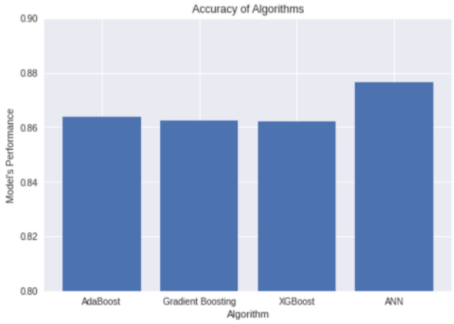


Fig 5. Comparison of Performance

Fig 2, Fig 3, Fig 4 and Table-II shows the performance of each model. From Fig 5, the accuracy percentage of AdaBoost algorithm is 86.3%, Gradient Boosting algorithm is 86.2%, XGBoost algorithm is 86.2% and Artificial Neural Network is 87.63%. Even with precision, recall and F1 score Artificial Neural Network using forward and backward propagation is leading with its scores. When compared to other all algorithms, deep learning algorithm performs much higher when compared to machine learning algorithms.

V. CONCLUSION

For predicting rainfall, here this project uses three boosting techniques and one artificial neural network using forward and back propagation. This project uses almost 1.45 lakh data of Australia in excel format. After the training dataset have been trained, they have tested it by predicting some unseen day's temperature and found accurate results. Finally artificial neural network using forward and backward propagation have performed well when compared to boosting algorithms. It is more suited to predict the tomorrow's rainfall.

FUTURE ENHANCEMENT

In the future, this project can be enhanced by

- Improving the model's accuracy by hyper parameter tuning
- By deploying into end-to-end application
- Predicting with live dataset

REFERENCE

1. M.J.C., Hu, Application of ADALINE system to weather forecasting, Technical Report, Stanford Electron, 1964.
2. S. Zhang, L. Lu, J. Yu, and H. Zhou, "Short-term water level prediction using different artificial intelligent models,"
3. S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall
4. Prediction,"
5. C. S. Thirumalai, "Heuristic Prediction of Rainfall Using Machine Learning Techniques,"
6. H. Vathsala and S. G. Koolagudi, "Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches,"
7. M. P. Darji, V. K. Dabhi, and H. B. Prajapati, "Rainfall forecasting using neural network: A survey,"
8. Mr. Dhawal Hirani, Dr. Nitin Mishra, "A Survey On Rainfall Prediction Techniques"
9. Karim Solaimani, "Rainfall-runoff Prediction Based on Artificial Neural Network (A Case Study: Jarahi Watershed)"
10. Kesheng Lu, Lingzhi Wang, "A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction"
11. G. Geetha, R Samuel Selvaraj, "Prediction of monthly rainfall in Chennai using back propagation neural network model"
12. Dr S. Santosh Baboo and I. Khadar Shareef, "An efficient Weather Forecasting Model using Artificial Neural Network"
13. Gwo-Fong Lin and Lu-Hsien Chen, 2005, "Application of an artificial neural network to typhoon rainfall forecasting"

AUTHORS PROFILE

Prof. B. Meena Preethi M.Sc., M.Phil., with 10 years of teaching and administrative experience is a university gold medalist at her Post Graduation from Bharathiar University. She is also awarded with the "Best Outgoing Student Award" for the year 2008. She is highly skilled in bringing innovative practices for evaluation process. As a committed individual, she has mentored many others in the same field. She has been continuously guiding other college faculty in implementing Outcome Based Education evaluation process. She is rigorously inspires and motivates other team members to do the same. To her credit she has presented more than 30 research papers and published 28 research articles in National and International journals. Her area of research is Data Mining in Medical Field. To add another feather to her cap she is an Oracle Certified Trainer.

R. Gowtham, currently pursuing 4th year in M.Sc(Software Systems) at Sri Krishna Arts and Science College, Affiliated to Bharathiyar University, Coimbatore. He has presented papers in international conference.

S.Aishvarya is currently pursuing 4th year in M.Sc(Software Systems) at Sri Krishna Arts and Science College, Affiliated to Bharathiyar University, Coimbatore. She has presented papers in international conference.

S.Karthick is currently pursuing 4th year in M.Sc(Software Systems) at Sri Krishna Arts and Science College, Affiliated to Bharathiyar University, Coimbatore. He has presented papers in international conference.

D. G. Sabareesh is currently pursuing 4th year in M.Sc(Software Systems) at Sri Krishna Arts and Science College, Affiliated to Bharathiyar University, Coimbatore. He has presented papers in international conference.