Rainfall prediction using Machine Learning

R Hemanth Kumar
Computer Science and Engineering (AI)
Amrita School of Computing
Bengaluru, India
bl.en.u4aie22138@bl.students.amrita.edu

Vishnu Prasad B
Computer Science and Engineering (AI)
Amrita School of Computing
Bengaluru, India
bl.en.u4aie22164@bl.students.amrita.edu

Himanshu Yadav
Computer Science and Engineering (AI)
Amrita School of Computing
Bengaluru, India
bl.en.u4aie22167@bl.students.amrita.edu

***Abstract*—Rainfall will be difficult to predict using old methods, there are various ways for predicting the rainfall up until now, the prediction of the rainfall is accurate enough it will help many people in planning their days, farmers for their crops, and fishermen for their daily job basis. To solve this problem we developed a model using Machine learning Algorithms, which can predict the Rainfall based on the atmospheric factors to conclude whether it will rain or not. This will be extremely helpful for the meteorological department in forecasting rainfall.**

***Index Terms*—Machine learning, rainfall, prediction**

I. Introduction

To implement this model we used python programming language and some necessary libraries such as Numpy, Pandas, matplotlib, sklearn. We fetched the existing dataset of rainfall parameters such as temperature, humidity, windspeed. First we will pre-process the dataset, pre-processing includes outliers removal, null value replacing. The important factors of rainfalls are :- Max-Temp : lower humidity : High Sunshine : Less wind speed : High further we train the models using algorithms such as ,KNN classifiers,logistic regression, XGBClassifiers, SVC. We can add even more classification algorithms to increase the chances of evaluating the most efficient and accurate model. To evaluate the performance of the classifiers, the evaluation metrics are included such as precision, f1-score and recall with the results.

II. Literature Review

[1] This paper analyses various parameters in the atmosphere. Different weather features such as Temperature, Humidity, Dew Point are identified, and a detailed study is presented. While all features played an important role in rainfall prediction, only a few of them. Based on these, an efficient set of features are used in a data-driven machine learning algorithm for rainfall prediction [2] In this paper, ML and DL rainfall forecasting approaches including a hybrid optimized-by-PSO support vector regression (PSO-SVR), long-short term memory (LSTM), and convolutional neural network (CNN).This paper pointed that PSO-SVR and LSTM approaches performed almost the same and better than CNN. [3] In this paper, machine learning methods are evaluated, driven by atmospheric patterns, for long-term daily rainfall prediction in a semi-arid climate..The performance of the models is evaluated using several metrics and statistics related with rainfall intensity and occurrence at daily, monthly and annual aggregation scales. These analysis of variance are used to evaluate the differences among models. [4] This paper presents a comparative analysis using simplified rainfall estimation models based on conventional Machine Learning algorithms that are efficient for these applications. Models based on XGBoost, and an ensemble of XGBR, Linear SVR, were compared in the task of forecasting rainfall using time-series data.[5] This paper talks about the various methods and models utilized in rainfall prediction: These models focuses on machine learning algorithms remote sensing techniques and hybrid approaches and highlights the significance of accurate rainfall forecasting for various sectors. Machine learning algorithms used are decision trees, random forests, support vector machines and recurrent neural networks. The paper uses remote sensing techniques, utilizing satellites and radar systems provide valuable data for rainfall estimation over large areas and in real time, enhancing the precision and timeliness of forecasts.

Hybrid models combining machine learning and remote sensing offer promising avenues for improving prediction accuracy. The paper talks about using advanced technologies for more accurate and reliable rainfall forecasting. [6] This paper basically compared 24 machine learning models for day ahead photovoltaic (PV) power forecasting using numerical weather predictions (NWP). They found that the kernel ridge (KR) model performed best in terms of accuracy, although it required extensive training time and memory usage. The multilayer perceptron (MLP) model showed identical accuracy but with lower training time The paper showed that including additional predictors beyond basic NWP outputs led to a reduction in root mean square error (RMSE), Hyperparameter tuning was also found to be usefull for optimizing model performance with tuned models achieving lower RMSE compared to default settings. Their findings provide valuable insights into selecting the most suitable ML models and input features for operational PV forecasting. [7] This paper is on rainfall prediction across various ecological zones in Ghana: In this paper they have used five classification algorithms Decision Tree, Random Forest, Multilayer Perceptron, Extreme Gradient Boosting, and K-Nearest Neighbour. They have used evaluation metrics such as precision, f1-score, recall, overall accuracy and execution times to assess the performance of the classifiers. The results reveal distinct classification characteristics for rain and no rain classes in different zones. Decision Tree had the fastest execution time across all zones while Random Forest, Extreme Gradient Boosting and Multilayer Perceptron demonstrated consistent performance suggesting their suitability for rainfall prediction. In future research they have told to explore additional classification algorithms and hybrid models. [8] The paper demonstrates two methods for forecasting rainfall: One based on Autocorrelation Function (ACF) and the other on projected error. they have also used various regression models like Bayesian Linear Regression (BLR), Boosted Decision Tree Regression (BDTR), Decision Forest Regression (DFR), and Neural Network Regression (NNR). They have observed that BDTR performs best in predicting rainfall using ACF with noticeable improvements achieved through cross validation and hyper parameter tuning. BDTR and DFR with LogNormal normalization demonstrated superior performance particularly in weekly error prediction. The study concludes that method 1 which is based on ACF provides the best rainfall prediction while acknowledging the potential for further enhancement. [9] The paper is a study on rainfall prediction in India: Various machine learning algorithms such as ARIMA, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Self Organizing Map (SOM) are used in this paper. Deep learning techniques like Multilayer Perceptron (MLP) and Auto Encoders are introduced for rainfall prediction which outperforms the traditional methods in terms of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The study highlights the significance of accurate rainfall prediction for agricultural planning and disaster prevention and shows the potential of Artificial Neural Networks in handling nonlinear relationships in rainfall datasets. [10] The paper introduces a study focused on improving weather prediction methods to address the challenges posed by climate change: The paper particularly focuses on sectors like agriculture heavily reliant on weather conditions. The paper uses data analytics and machine learning like the random forest classification algorithm This method demonstrates results with an accuracy of 87.90 percent in predicting rain occurrences.

III. Design Methodology

The first part of our methodology involves the pre-processing the data which includes handling missing data, removing outliers, normalization. We after categorize the data into 2, training data and testing data. After that we will present the results based on the performance of the models. To evaluate accurate results, EDA Exploratory Data Analysis is used.

A. understanding the data

In this section, we discuss about the dataset. We have used a dataset that contains the rainfall factors details categorised according to the location in Austrailia. (weatherAUS.csv). This dataset contains features such as Date, Location, MinTemp, MaxTemp, Humidity, WindGustSpeed, evaporation, sunshine, windDir, pressure some are categorised according to the time, for example cloud at 9 am, cloud at 3 pm. Finally there is target label which declares whether it rained on that particular day or the day after or not also the risk factor is mapped.

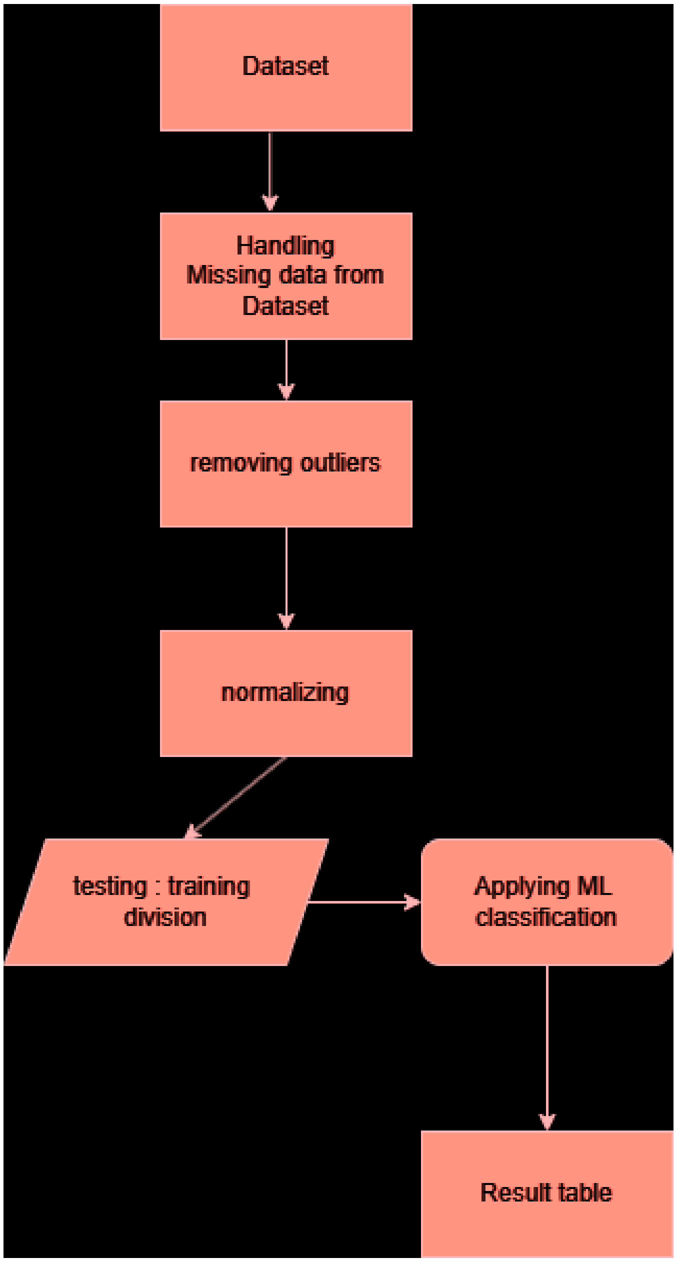

Dataset
Handling Missing data from Dataset
removing outliers
normalizing
testing : training division
Applying ML classification
Result table

Fig. 1. Design Flowchart

B. Pre-processing the Data

Pre-processing or cleaning the data is necessary, this will increase the accuracy of the calculation and also provide clarity to each segment of the data. We start of with the process of removing unwanted features such as the date, location, Rainfall, cloud9am, cloud3pm, risk-MM. Winddir. some of these features contains many null values, too much categorical data, irrelevant data. Next, we go with the process of replacing the null values with either mean/median for numerical data, or most-frequent for categorical data. So now we have a clean accurate data with relevant features. Next we will replace the null values of each columns. All these changes will be saved into the dataframe which will then be used for the next section that is Plotting the data.

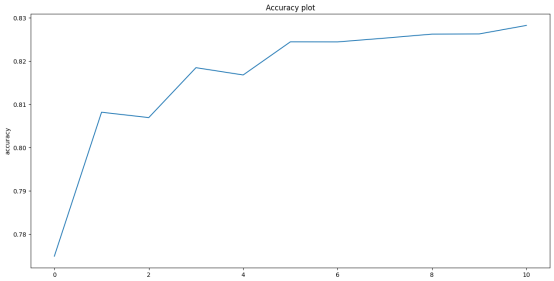
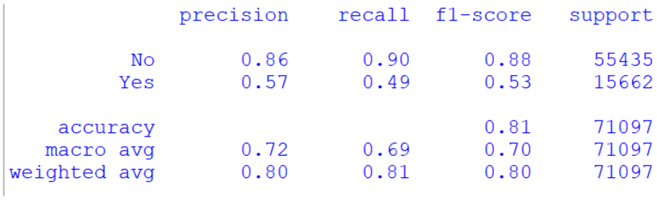


Fig. 2. Accuracy Plot



```
              precision    recall  f1-score   support

          No       0.86      0.90      0.88     55435
         Yes       0.57      0.49      0.53     15662

    accuracy                           0.81     71097
   macro avg       0.72      0.69      0.70     71097
weighted avg       0.80      0.81      0.80     71097
```

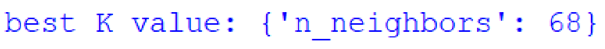
Fig. 3. Performance score

C. Applying KNN classifiers

In this section we first classify the data into X-test,y-test,X-train and y-train with the test size of 0.5, that means 50percent will be split into test and 50percent to train dataset. After this we apply the KNNclassifiers() with k value of 3 and train the dataset. Further we predict the test-vector. we shall experiment the above prediction from K ranging from 1 to 11. For K=3 we calculate the accuracy and the performance metrics such as F1-score,recall and accuracy. Finally we store the accuracy that we got from K=1 to K=11 and plot it.

D. Tuning Hyper-parameter

In This section we will be finding the best K value to insert in KNN classifier using method called hyper-parameter tuning.Initially we take the K value as 3 and use the method RandomSearchCV(), we use this method instead of Grid-searchCV() is that, because the frequency of our dataset is huge in number. So we give the SearchCV values from 1 to 100 to give us the best K value out of which we are going to be using in our model. For now we just find the best value for n-neighbors. In this case we got the best K value as 68. refer fig. [5].



```
best K value: {'n_neighbors': 68}
```

Fig. 4. Fitting

IV. Results

Lab - 3 progress :- Understood to topic, Structured the introductions and abstract, conducted a literature survey and designed methodology.

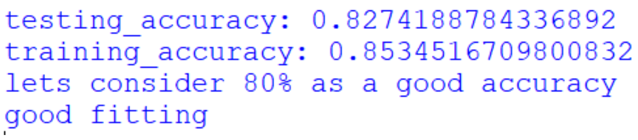
Lab - 4 progress :- data pre-processing.

Lab - 5 progress :- pre-processed data more effciently and implemented used KNNclassifiers, compared the K values and plotted accuracy and checked the fitting of our data set.

Lab - 6 progress :- Hyper parameter tuning for KNNclassifiers, tuned n-neighbors. some minor changes in the report.

V. Conclusion

Do you think the classes you have in your dataset are well separated? Justify your answer. Yes, partially though, even when we get accuracy above 80 we still had to drop some of the columns to remove the outliers. Explain the behavior of the kNN classifier with increase in value of k. Explain the scenarios of over-fitting and under-fitting in kNN classifier. As we plotted the accuracy plot we can see that we get a curve shaped line structure this means upto k=11 the accuracy is varying with slight variance. Underfitting occurs when the train accuracy is less and Overfitting occurs when the testing accuracy is less. In our case we get the both testing and training accuracy more that 80percent which is a Good fitting dataset. Do you think the kNN classifier is a good classifier based on the results obtained on various metrics? although we get 80% accuracy by using KNNclassifiers, we can train the model with other classification or regression algorithm. Also the accuracy can be increased in KNNclassifier by label encoding, since we did not do label encoding in our progress, the more we add the features the more accuracy will increase. Do you think the model has regular fit situation?Use train and testset performances to arrive at this inference. Yes as you can see in Fig. 4. we get the testing accuracy and training accuracy ¿ 80%. Overfitting happens when the testing accuracy is less. which in our case, is not.



```
testing_accuracy: 0.8274188784336892
training_accuracy: 0.8534516709800832
lets consider 80% as a good accuracy
good fitting
```

Fig. 5. Fitting

References

[1] Shilpa Manandhar et al. "A Data-Driven Approach for Accurate Rainfall Prediction",IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 57, NO. 11, NOVEMBER 2019, 0196-2892 © 2019 IEEE

[2] Fatemeh Rezaei Aderyani et al., "Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN", Journal of Hydrology 614 (2022), 0022-1694/© 2022

[3] Javier Diez-Sierra et al., "Long-term rainfall prediction using atmospheric synoptic patterns in semiarid climates with statistical and machine learning methods", Journal of Hydrology 586 (2020) 0022-1694/ © 2020

[4] Ari Yair Barrera-Animas et al., "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting ", 2666-8270/© 2022

[5] S. D. Latif, N. A. B. Hazrin, C. H. Koo, J. L. Ng, B. Chaplot, Y. F. Huang, A. El-Shafie, and A. N. Ahmed, ”Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches,” Alexandria Engineering Journal, vol. 63, no. 3, pp. 1325-1335, Sep. 2023. [Online]. Available: https://doi.org/10.1016/j.aej.2023.09.009.

[6] D. Markovics and M. J. Mayer, ”Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction,” Renewable Energy, vol. 188, pp. 812-824, Mar. 2022. [Online]. Available: https://doi.org/10.1016/j.renene.2022.03.093.

[7] N. K. A. Appiah-Badu et al., ”Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana,” IEEE Access, vol. 9, pp. 19425-19439, 2021. DOI: 10.1109/ACCESS.2021.3139312.

[8] W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, ”Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia,” Ain Shams Engineering Journal, vol. 12, pp. 1651-1663, Nov. 2020.

[9] C. Zeelan Basha, N. Bhavana, P. Bhavya, and S. V, ”Rainfall Prediction Using Machine Learning and Deep Learning Techniques,” in Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020), ISBN: 978-1-7281-4108-4, IEEE Xplore Part Number: CFP20V66-ART.

[10] N. Singh, S. Chaturvedi, and S. Akhter, ”Weather Forecasting Using Machine Learning Algorithm,” Noida, India, 2019, pp. 1-5, ISBN: 978-1-5386-9436-7.