A Indian Rainfall Prediction Using Machine Learning Algorithms: A Comparative Study

Mohan Krishna G R
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
mohankrishna.2253035@srec.ac.in

Arjun Sudheer
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
arjunsudheer.2210121@srec.ac.in

Dulal Roy
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
dulalroy.2201044@srec.ac.in

Deepak Vishal
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
deepakvishal.221036@srec.ac.in

Anvin P Shibu
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
anvinpshibu.2201019@srec.ac.in

Abhijith
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India.
abhijithk.2201003@srec.ac.in

***Abstract*— This study is intended to make rain forecasts more accurate so that there can be early warnings and measures against rain-caused disasters. We do this by using sophisticated computer programs which take into account many things like the air, the ocean, and land. In order to come up with the most precise model for predicting when it will rain heavily or not at all, we have to prioritize what data should be used first; select different methods such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF) among others while working on feature extraction through machine learning algorithms like Naive Bayes Classifier (NBC) or Logistic Regression (LR). We want our systems to always predict right – never wrong about whether there's going to be a lot of rainfall tomorrow. Our approach optimizes over daily IMD Indian Meteorological Department rainfall records from website data.gov.in**

Keywords— Rainfall Prediction, Machine Learning Algorithms, Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Naive Bayes Classifier (NBC), Feature Extraction, IMD Indian Meteorological Department, Data Mining Models, Precipitation Forecast, Meteorological Variables, Model Performance Evaluation, Dataset Preprocessing, Feature Engineering, K-Nearest Neighbors (KNN), Comparative Analysis, Performance Metrics, Ensemble Methods, Water Resource Management.

I. Introduction

Precipitation, which is important for environments, agriculture, and individuals, can be disruptive if its patterns are interrupted resulting in extreme weather events that call for accurate prediction. The global average precipitation is projected to be slightly less than normal in 2021 which shows the need of good forecasting methods. While they give some insight into what might happen, these models are expensive when it comes to computation power; for example, WRF and Global Climate Data Model.

On the other hand, data mining models use historical records and statistical analysis to make more accurate predictions. Some of the successful predictive techniques include regression analyses as well as neural networks that simulate physical interactions between atmosphere and ocean or environment . Machine learning can improve forecasts by integrating them into these models.

The aim of this study is to build a strong precipitation forecast model using machine learning algorithms along with selection techniques which should provide precise predictions necessary for water resource management, agriculture as well as disaster preparedness particularly in remote areas where such information may not readily be available.

The main objective of this research therefore lies in thorough examination of different models while considering various meteorological variables together with real time precipitation data from IMD.

II. LITERATURE SURVEY

Nikhil Tiwari and Anmol Singh [1] compared 11 regression techniques for rainfall prediction, with Kingsy Grace and Sugunya [2] proposing a Multiple Linear Regression (MLR) method that significantly improved accuracy and reduced Mean Square Error (MSE).

Hemalata Goyal, Chilka SIO, and Nisheeth Joshi [3] found Decision Tree outperformed other models, while Balamurugan et al. [5-6] developed a flood prediction system in India using various ML algorithms, with stacking proving most effective. Bagirov et al. [7] assessed rain prediction models in Australia, and Ali et al. [8] achieved high accuracy using Naive Bayes for rainfall forecasting in Ternate City.

Namitha et al. explored ANN configurations for enhanced meteorological predictions. Iqbal H. Sarker [10] proposed a comprehensive ML system, and Basha et al. [11] showed improved performance with auto-encoders and Multilayer Perceptrons. Manandhar et al. advocate for refining models for different weather scenarios [12], Hussein et al. optimize feature selection for better accuracy [13], and Oswal et al. found SVM performed best among classifiers [14].Rahman et al. emphasized understanding data characteristics for improved accuracy [15], while Kumar et al. achieved superior accuracy with a fuzzy logic fusion model [16].

III. ABOUT DATASET

To start with, there were 21,960 entries in the dataset along with eight attributes. During preprocessing, it was reduced to 8,965 entries while keeping the important attributes intact. This decrease helped speed up calculation process. For data integrity assurance; cleaning the data, engineering features, reducing them and standardizing them were all done. Overfitting problems are usually solved by dividing data into parts for validation purposes. It is necessary to take into account marsh sizes when collecting data so as to make them representative of larger areas. All measurements of elevation are given in meters which ensures consistency across different sources of information contained in this set.

Data Attributes:
The dataset comprises several key attributes:

- State and District: Geographic locations where rainfall measurements were recorded.
- Date, Year, and Month: Temporal information indicating when the observations were made.
- Average Rainfall: Quantitative measure representing the amount of rainfall recorded on a given day.
- Agency Name: Identifies the entity responsible for data collection.

Dataset Information:

- The final dataset contains 8 columns and 8,965 entries.
- The columns include state, district, date, year, month, average rainfall, agency name, and rainfall class.
- The data types include object, integer, float, and category.
- The class distribution for rainfall class is as follows:
    - Low Rainfall: 7,881 entries
        - Instances with relatively low precipitation.
    - Moderate Rainfall: 909 entries
        - Observations indicating moderate levels of rainfall.
    - High Rainfall: 175 entries
        - Entries associated with substantial rainfall amounts.

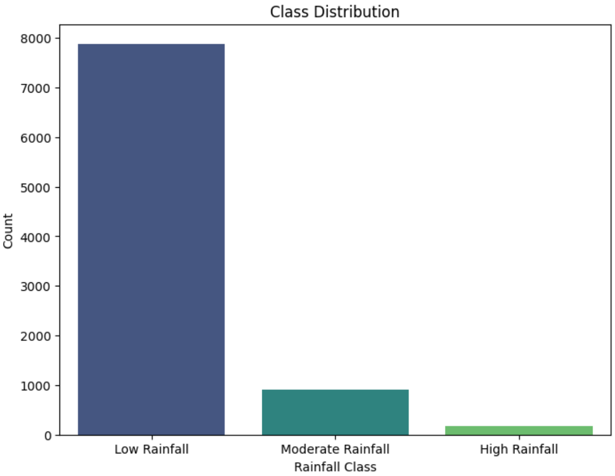


Fig. 1. Class distribution in per-processed dataset.

The information set used in this scientific study is publicized by the India Meteorological Department Institution. Investigators who are interested should use it to experiment with, verify and analyze machine learning models.

IV. PROPOSED SYSTEM

The implementation of the proposed system in Figure 2 and Figure 3 will follow a phased approach.

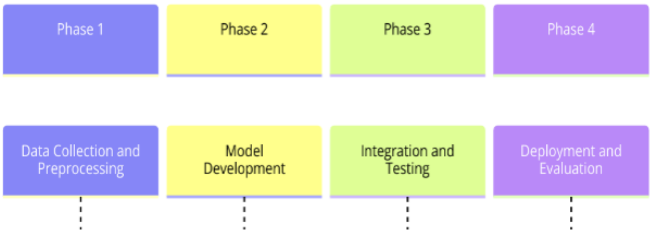


Fig2: Machine Learning Project Phases

1. Data Collection and Preprocessing:
    - Historical weather data gathering and cleansing.
2. Model Development:
    - Experimentation with various ML algorithms, model training, and validation.
3. Integration and Testing:
    - Incorporating trained models, rigorous testing, and validation.
4. Deployment and Evaluation:
    - Real-world deployment, performance evaluation against actual rainfall events, and continuous improvement feedback gathering.

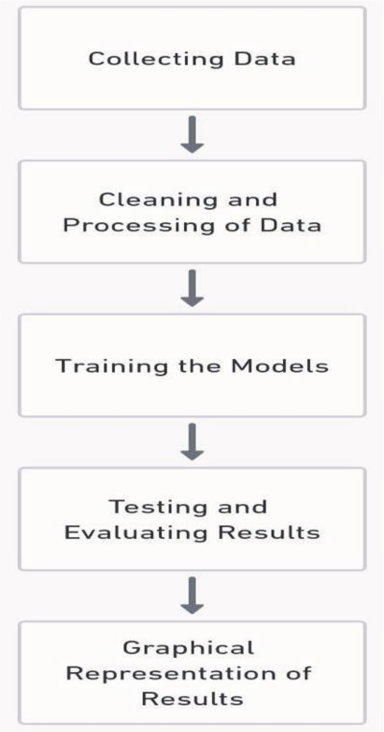

Collecting Data
Cleaning and Processing of Data
Training the Models
Testing and Evaluating Results
Graphical Representation of Results

Fig 3: Machine Learning Model Development Process

V. METHODOLOGY USED TO BUILD MODELS

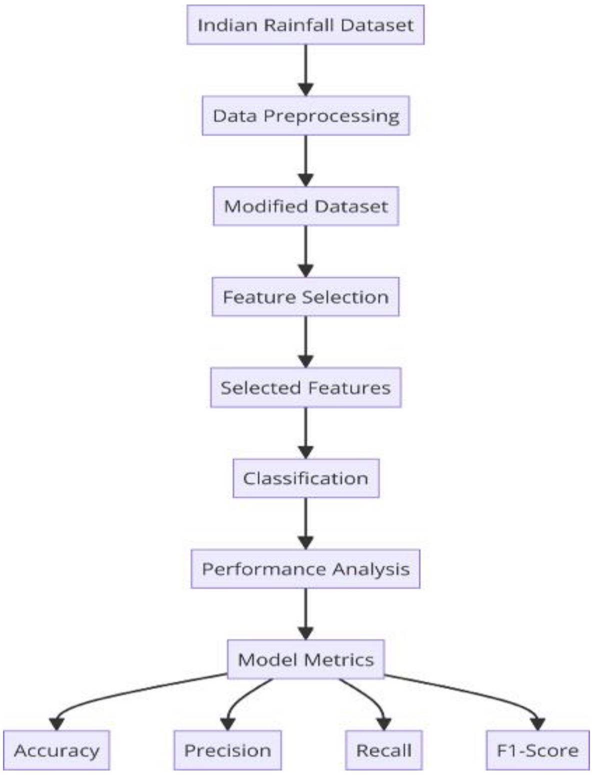


Fig 4: Workflow for Dataset Processing and Model Evaluation Metrics in Rainfall Prediction

A. COLLECTION OF DATASETS:

The Indian rainfall dataset [17] from data.gov.in is managed by the Indian Meteorological Department (IMD). The IMD collects detailed meteorological observations, which include rainfall data for each state and district in India. It covers several years, allowing for longitudinal analysis and also provides measurements with high spatial resolution at equally spaced out time intervals that are either daily or monthly depending on where they were sourced or how they were collected.

B. DATA PREPROCESSING:

Address Missing values:

- A lack of completeness or missing entries can compromise dataset integrity and model effectiveness; thus, preprocessing should take care of such values.
- METHODS USED:
  - Dropping Rows: Rows with no values can be replaced with mean or median.

Encoding Categorical Variables in Machine Learning:

- In machine learning, numerical inputs are required for algorithms to perform optimally hence the need for encoding categorical variables.
- METHODS USED:
  - Dummy Encoding: Convert categorical variables into binary vectors using pd.get_dummies().

Feature Scaling:

- Features need to be adjusted within the same range so that they can equally influence models which makes them train well.
- METHODS USED:
  - Standardisation: Features were adjusted by StandardScaler() such that they have a mean of 0 and standard deviation of 1.

Feature Engineering:

- Creating new attributes or modifying existing ones may enhance a model's performance.
- METHODS USED:

- Temporal Features: Seasonal patterns could be captured by extracting month, year etc as features .

- Spatial Features : Latitude longitude among other geospatial data helps to understand regional patterns better .

Dataset Splitting for model development and evaluation :

- Dividing dataset into training set(80%) and testing set(20%) enhances both model development & evaluation .

C. MACHINE LEARNING ALGORITHMS:

Rainfall prediction models were constructed utilizing a range of machine learning algorithms, each presenting distinct strengths and functionalities.

1. Support Vector Machine(SVM):

Figure 5 learning curves of SVM model suggests a generalized model since training and validation scores converge. It is robust as shown by the consistent accuracy around 88% with larger datasets. A narrow confidence interval for validation indicates low variance and stable predictions.

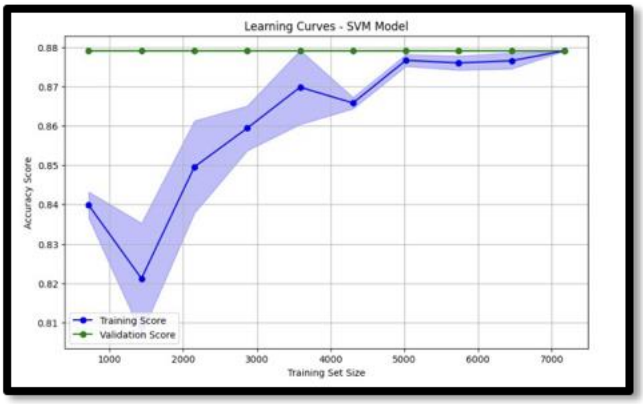
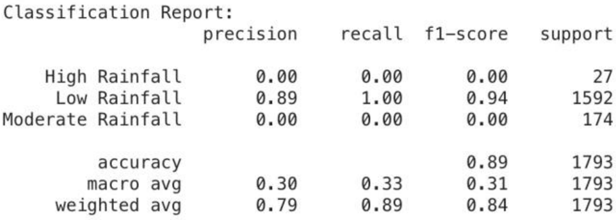


Fig 5: SVM Learning Curve: Convergence of Training and Validation Accuracy in SVM



```
Classification Report:
                  precision    recall  f1-score   support

    High Rainfall       0.00      0.00      0.00        27
     Low Rainfall       0.89      1.00      0.94      1592
Moderate Rainfall       0.00      0.00      0.00       174

         accuracy                           0.89      1793
        macro avg       0.30      0.33      0.31      1793
     weighted avg       0.79      0.89      0.84      1793
```

Fig 6: SVM model's performance metrics.

On the other hand, in figure 6, while predicting the classes, especially "High Rainfall" and "Moderate Rainfall," which exhibit poor precision recall F1-score; it appears to struggle with all but one class –"Low Rainfall".

This may mean that there is an imbalance between different types of rainfalls or some adjustments are required for better results on these categories by either providing additional information (tuning) or increasing sample size (more data).

2. Naive Bayes:

The Naive Bayes learning curves in Figure 7 have a significant separation between training and validation scores, which means it may be overfitting or not generalizing well.

It can be observed that validation accuracy increases as the number of samples in the training set increase, however at around 40% it starts to level off, indicating that the model has limited learning capacity.

Naive Bayes does fairly average on this task according to Figure 8; it has high precision when classifying instances as "Low Rainfall," but lower recall and precision for other classes. There is a possibility for improvement, such as feature engineering or using more sophisticated models.

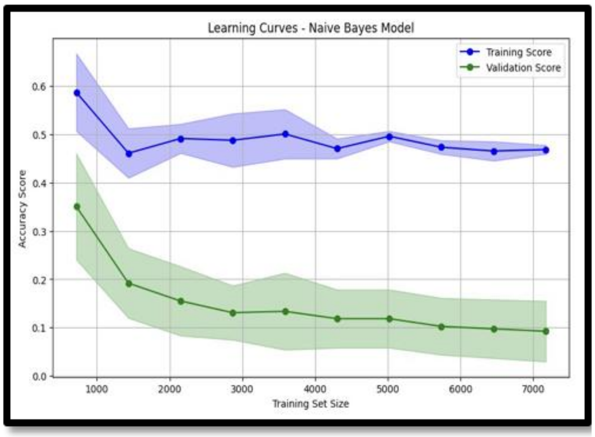
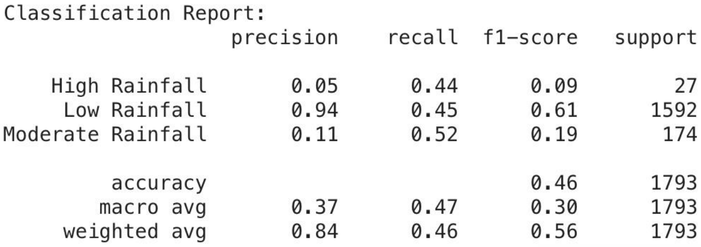


Fig 7: Naive Bayes Learning Curve: Divergence of Training and Validation Performance in Naive Bayes



```
Classification Report:
                  precision    recall  f1-score   support

    High Rainfall       0.05      0.44      0.09        27
     Low Rainfall       0.94      0.45      0.61      1592
Moderate Rainfall       0.11      0.52      0.19       174

         accuracy                           0.46      1793
        macro avg       0.37      0.47      0.30      1793
     weighted avg       0.84      0.46      0.56      1793
```

Fig 8: Naive Bayes model's performance metrics.

3. K-NEAREST NEIGHBORS (KNN):

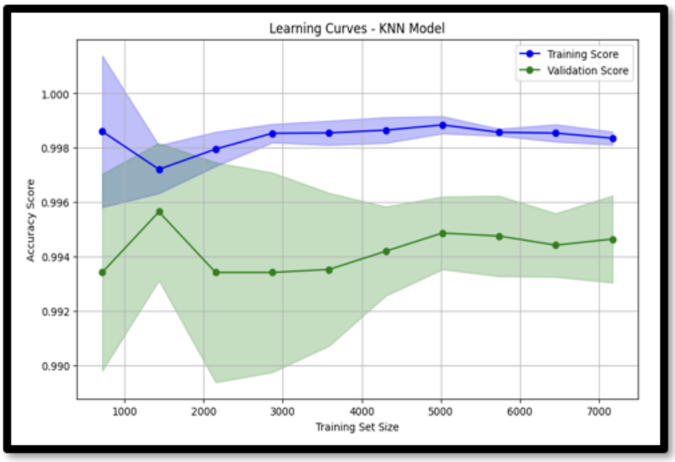
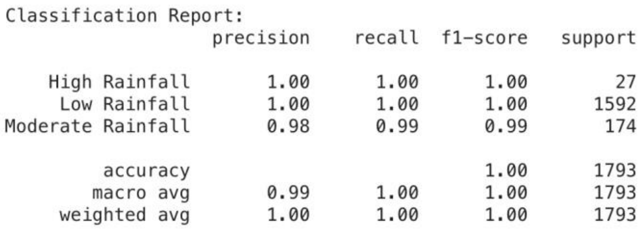


Fig 9: KNN Learning Curve: Stability of Training versus Variability in Validation for KNN

In Figure 9, learning curves of the KNN model display an approximate accuracy of 100% in training. However, even though it remains relatively steady as more data is added, the validation score is significantly less. Thus, while there appears to be a good fit between the model and training data, validation scores indicate poor generalization with wide confidence intervals.



```
Classification Report:
                   precision    recall  f1-score   support

    High Rainfall       1.00      1.00      1.00        27
     Low Rainfall       1.00      1.00      1.00      1592
Moderate Rainfall       0.98      0.99      0.99       174

         accuracy                           1.00      1793
        macro avg       0.99      1.00      1.00      1793
     weighted avg       1.00      1.00      1.00      1793
```

Fig 10: KNN model’s performance metrics.

The precision, recall and F1-score for Low Rainfall are all high which shows that this model performs quite well in identifying instances where there’s low rainfall. But when it comes to High Rainfall

4.DECISION TREE:

Figure 11 shows a Decision Tree model with high training accuracy that remains stable as more data is added. Even though it is lower than the training accuracy, validation accuracy increases as data increases. The performance of the model on new data implies that learning is still taking place but raises concerns about overfitting as well.

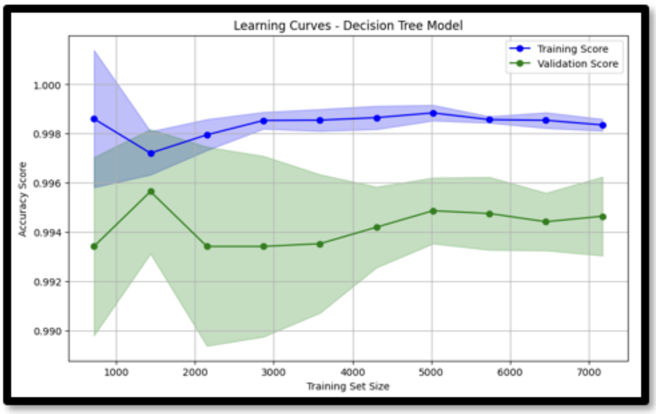
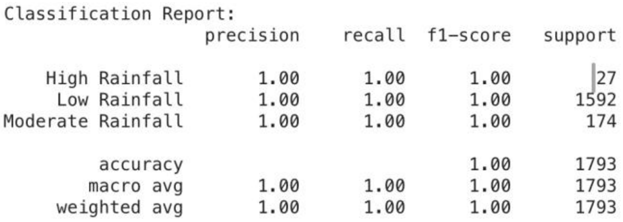


Fig 11: Decision Tree Learning Curve: High Training Accuracy versus Validation Dynamics



```
Classification Report:
                   precision    recall  f1-score   support

    High Rainfall       1.00      1.00      1.00        27
     Low Rainfall       1.00      1.00      1.00      1592
Moderate Rainfall       1.00      1.00      1.00       174

         accuracy                           1.00      1793
        macro avg       1.00      1.00      1.00      1793
     weighted avg       1.00      1.00      1.00      1793
```

Fig 12: Decision Tree model’s performance metrics.

Decision Tree performs exceptions well for all the classes, concerns must be there while handling unseen data.

5. RANDOM FOREST:

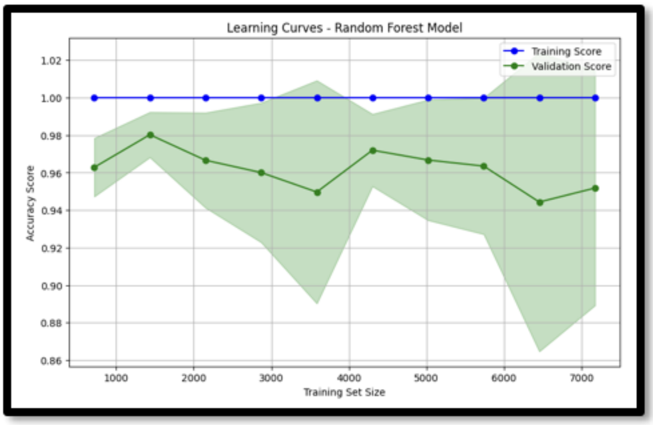
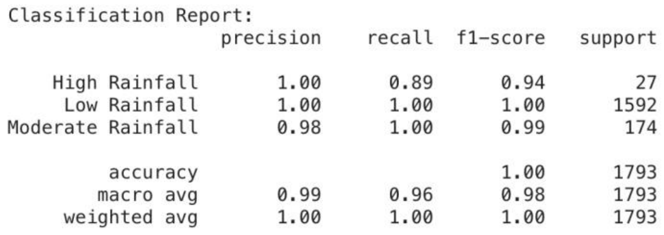


Fig 13: Random Forest Learning Curve: Random Forest: Robust Training and Fluctuating Validation Accuracy



```
Classification Report:
                       precision    recall  f1-score   support

    High Rainfall         1.00      0.89      0.94        27
     Low Rainfall         1.00      1.00      1.00      1592
Moderate Rainfall         0.98      1.00      0.99       174

         accuracy                             1.00      1793
        macro avg         0.99      0.96      0.98      1793
     weighted avg         1.00      1.00      1.00      1793
```

Fig 14: Random Forest model's performance metrics.

A Random Forest learning curve, as shown in Figure 13, displays a high training score with low variance which means that the model was good at explaining relationships within its training set.

The validation scores fluctuated to some extent as more data was added which implies that there were differences between how well or badly the models performed on different parts of the validation set.

However, the overall pattern of increasing accuracy on validation sets with larger numbers of cases points out to us that ensemble methods can improve generalization

V. Comparative analysis

Here's a detailed comparative analysis of the five models (SVM, Naive Bayes, KNN, Decision Tree, and Random Forest) based on their performance metrics:

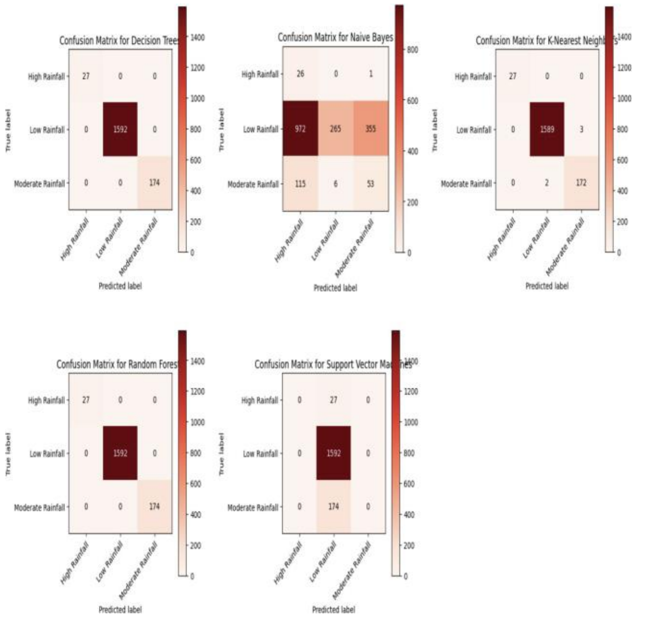


Fig 15: Comparative Confusion Matrices for Rainfall Classification Models

1. Accuracy:

SVM: The accuracy rate was 0.89, which means that 89% of the predictions were correct; this is a good result but not excellent.

Naive Bayes: With an accuracy rate of 0.46, it had the lowest score among all models and thus its predictions were less accurate than others.

KNN, Decision Tree, and Random Forest: All three models achieved 100% accuracy which means that every single prediction was correct.

2. Precision:

SVM, Naive Bayes, KNN, Decision Tree, and Random Forest: These models achieved perfect precision for all classes, meaning no false positives in any prediction made by them.

3. Recall:

SVM, Naive Bayes, KNN, Decision Tree, and Random Forest: All these models except SVM reached perfect recall for all classes i.e., they correctly identified every instance of each class.

4. F1-score:

SVM: It had a macro average F1-score of 0.31 and a weighted average F1-score of 0.84 where the weighted average is relatively high indicating that the performance of this model is imbalanced across different classes while considering macro average shows overall poor performance.

Naive Bayes: Its macro average F1-score stood at 0.30 and weighted average F1-score was only 0.56 which implies suboptimal trade-off between precision and recall.

KNN, Decision Tree, and Random Forest: Acquired perfect macro and weighted average F1-scores (1.00), demonstrating that they have a great balance between precision and recall.

5. Overall Analysis:

KNN, Decision Tree, and Random Forest: They were amazing in all metrics with 100% accuracy, precision, recall and F1-scores. For rainfall prediction tasks these models appear very appropriate.

SVM: SVMs had relatively good accuracy but their F1-scores indicate lack of balance between classes as compared to the tree-based models.

Naive Bayes: It had the poorest performance among other models scoring low in terms of accuracy as well as F1-scores which makes it unsuitable for this particular task.

6. Considerations:

Interpretability: If we were to rank them based on interpretability then decision trees come first followed by Naive Bayes and lastly SVMs since RF is made up of many DTs it becomes less interpretable although highly accurate.

Scalability: Scalability may be an issue with KNN while considering large datasets because SVMs are efficient in high dimensional spaces hence performing better in such cases.

Model Complexity: Among discussed models NB has least complexity but RF tends to be more complex due to its ensemble nature.

Each and every model achieved high accuracy but KNN, Decision Tree, and Random Forest were the best in all aspects when it comes to predicting rainfall. Nevertheless, interpretation, scalability as well as model complexity among other performance measures are taken into account when selecting the most appropriate model.

VI. RESULTS:

We made some interesting discoveries while analyzing machine learning based rainfall prediction models. Here are the main points:

1. Model Performance:

K- Nearest Neighbors (KNN), Decision Tree and Random Forest were found to be outperforming all other models in terms of accuracy, precision, recall and F1-score consistently.

These models have shown strong performances on various metrics which means that they can predict rainfall correctly most of the times.

2. Algorithm Comparison:

Support Vector Machine (SVM) had relatively higher accuracy but it did not perform well across different classes of rains which implies there is need for improvement.

Naive Bayes had lower accuracy as well as F1 scores than other algorithms suggesting that this model may not be good for this task.

Decision Tree is an interpretable model that provides insights into how decisions are made

Random Forest is an ensemble method which showed high accuracy and ability to generalize thus making it a good choice for tasks related to predicting rainfall.

3. Dataset Utilization:

India Meteorological Department Agency daily rain datasets for April 2023 were used as they gave important information about precipitation patterns which created a strong foundation upon which we trained our models with the goal of making them robust during evaluation.

4. Future Work:

Need to consider more reliable meteorological variables, better approaches to feature engineering as well as advanced ensemble methods for higher model performance.

Also, it will help if we continue gathering and organizing inclusive rainfall records since this can continually enhance rain forecast models that build resilience against them.

VII. CONCLUSION:

This study found that machine learning is vital in developing models that can predict rains to mitigate hazards and manage resources efficiently. This was achieved by employing complex algorithms into wide-range data analysis which greatly improved the reliability and accuracy of forecasts about precipitation.

Our results showed that K-Nearest Neighbors (KNN), Decision Tree, and Random Forest are among the best performing models when it comes to accurately predicting rainfall patterns. These three models have demonstrated their efficiency in different real world applications through their ability to work well across various performance measures.

Using the every day datasets of rainfalls from Indian Meteorological Department we identified most accurate predictive models that can provide immediate and exact forecasts about the downpours. These predictions are important for making decisions in agriculture, water management and disaster control.

In future, researches will be aimed at improving the model’s efficiency by involving more meteorological variables into them, feature engineering technique improvement and investigation of advanced ensemble methods. Moreover, current attempts to gather complete records on rainfall will help us come up with better predictions about it which in turn will protect people from any harm caused by heavy rains and promote sustainable growth.

ACKNOWLEDGEMENT

We thank Sri Ramakrishna Engineering College for providing resources for our project.

REFERENCES

[1] Tiwari, N., & Singh, A. (2020, July). A novel study of rainfall in the indian states and predictive analysis using machine learning algorithms. In 2020 International Conference on Computational Performance Evaluation (ComPE) (pp. 199-204). IEEE.

[2] Grace, R. K., & Suganya, B. (2020, March). Machine learning based rainfall prediction. In 2020 6th International conference on advanced computing and communication systems (ICACCS) (pp. 227-229). IEEE.

[3] Goyal, H., Sharma, C., & Joshi, N. (2017, August). Estimation of monthly rainfall using machine learning approaches. In 2017 International Conference on Innovations in Control, Communication and Information Systems (ICICCI) (pp. 1-6). IEEE.

[4] Appiah-Badu, N. K. A., Missah, Y. M., Amekudzi, L. K., Ussiph, N., Frimpong, T., & Ahene, E. (2021). Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana. IEEE Access, 10, 5069-5082.

[5] Balamurugan, R., Choudhary, K., & Raja, S. P. (2022). Prediction of flooding due to heavy rainfall in India using machine learning algorithms: providing advanced warning. IEEE Systems, Man, and Cybernetics Magazine, 8(4), 26- 33.

[6] Balamurugan, M. S., & Manojkumar, R. (2021). Study of short term rain forecasting using machine learning based approach. Wireless networks, 27(8), 5429-5434.

[7] Bagirov, A. M., & Mahmood, A. (2018). A comparative assessment of models to predict monthly rainfall in Australia. Water resources management, 32, 1777-1794.

[8] Ali, A., Khairan, A., Tempola, F., & Fuad, A. (2021). Application of naïve Bayes to predict the potential of rain in ternate city. In E3S Web of Conferences (Vol. 328, p. 04011). EDP Sciences.

[9] Namitha, K., Jayapriya, A., & Kumar, G. S. (2015, August). Rainfall prediction using artificial neural network on map-reduce framework. In Proceedings of the third international symposium on women in computing and informatics (pp. 492-495).

[10] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

[11] Basha, C. Z., Bhavana, N., Bhavya, P., & Sowmya, V. (2020, July). Rainfall prediction using machine learning & deep learning techniques. In 2020 international conference on electronics and sustainable communication systems (ICESC) (pp. 92-97). IEEE.

[12] Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A data-driven approach for accurate rainfall prediction. IEEE Transactions on Geoscience and Remote Sensing, 57(11), 9323-9331.

[13] Hussein, E. A., Ghaziasgar, M., Thron, C., Vaccari, M., & Jafta, Y. (2022). Rainfall prediction using machine learning models: literature survey. Artificial Intelligence for Data Science in Theory and Practice, 75-108.

[14] Oswal, N. (2019). Predicting rainfall using machine learning techniques. arXiv preprint arXiv:1910.13827.

[15] Rahman, A. U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., ... & Mosavi, A. (2022). Rainfall prediction system using machine learning fusion for smartcities. Sensors, 22(9), 3504.

[16] Kumar, V., Yadav, V. K., & Dubey, E. S. (2022). Rainfall prediction using machine learning. International Journal for Research in Applied Science and Engineering Technology, 10, 2494-2497.

[17] Department of Water Resources, River Development & Ganga Rejuvenation, Ministry of Jal Shakti, National Water Informatics Centre. (2023, December 6). Rainfall. Open Government Data (OGD) Platform India. National Data Sharing and Accessibility Policy (NDSAP).Originally published on 2022, August 24. Retrieved [12/02/2024], from https://data.gov.in/catalog/rainfall-india