WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Rainfall Prediction Using Machine Learning

Bikan Kushal Prasad [a], Mohd. Zoeib Uddinl[b], Pendem Nithin[c], Thoka Anil Goud[d], H. Venkata Subbaiah[e,*]

[a,b,c,d] Student, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Assistant Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT –

India is an agrarian nation, and its economy mostly depends on agricultural production and rainfall. Precipitation forecasting is crucial for evaluating agricultural productivity and is indispensable for all farmers. Rainfall prediction employs scientific and technical methods to anticipate atmospheric conditions. Precise monitoring of precipitation is crucial for the effective utilisation of water resources, improvement of agricultural output, and proactive planning of water infrastructure. Diverse data mining methodologies can be utilised to predict precipitation. Data mining approaches are utilised to quantitatively evaluate precipitation. This paper analyses various prominent data mining techniques for rainfall prediction. Algorithms including Random Forest, K-Nearest Neighbours, Logistic Regression, Support Vector Machine, and Decision Tree have been employed. This comparison facilitates the evaluation of whether technique produces greater accuracy in rainfall forecasting.

Keywords: Machine Learning, Feature Engineering, Data Transformation, Weather Forecasting, Data Selection, Supervised Learning, Hybrid Models, Meteorological Data, Artificial Neural Networks (ANN), Support Vector Machines (SVM).

INTRODUCTION

The aim is to develop a machine learning model for Rainfall Prediction that might replace the existing updatable supervised machine learning classification models by attaining superior accuracy through comparisons with supervised methods. This forecast facilitates the prediction of precipitation, improves agricultural yield, and evaluates climatic conditions in farming countries. This application is really user-friendly. It can operate accurately and effortlessly in a different context. It reduces the workload and improves task efficiency. It possesses temporal value. Diverse software applications utilise different approaches and frameworks, such as the Waterfall model, Iterative model, Spiral model, V model, and Big Bang model. I utilised the waterfall model in this application. I endeavoured to employ test case and case software approaches. This documentation begins with a formal introduction. Following the introduction, the project's analysis and design are outlined. The project's analysis and design include several elements: the proposal, mission, objectives, target audience, and environment. Use cases are presented in chapter 2, whereas test cases are detailed in chapter 3. This documentation culminates in the results and conclusion section. Precipitation is essential for agricultural planning, water resource management, and ecological operations. Prolonged droughts or excessive precipitation at vital stages of crop growth may lead to a significant decline in agricultural yield. India is an agrarian nation, and its economy primarily depends on agricultural output. Thus, precipitation forecasting is essential in agricultural countries like India. Predicting precipitation has been one of the most substantial scientific and technological difficulties worldwide during the past century.

We first obtain the dataset from our source, then do data preprocessing and visualisation algorithms for cleansing and presenting the dataset, respectively. We subsequently apply machine learning algorithms to the dataset and create a pickle file for the ideal algorithm, employing Flask as the user interface to display the findings.

RELATED WORK

Empirical research demonstrates that the notion of ISMR cannot be precisely forecasted by insights or empirical evidence. This review illustrates the utilisation of three methodologies: object generation, entropy, and artificial neural networks (ANN). In response to this innovation, an alternate approach for forecasting ISMR timeframes has been established to address the notion of ISMR. This model has garnered endorsement and support from the studio and exploration data. Analytical evaluation of various data and associated enquiries illustrating the execution of the typical method. This movement primarily illustrates the advantages of AI algorithms and the enhanced capabilities of intelligent systems relative to current rainfall predicting methods. We analyse the momentum execution (Markov chain augmented by rainfall study) in conjunction with the forecasts of the six least proficient AI systems: Genetic Programming, Support Vector Regression, radio networks, M5 networks, M5 models, and Happy models. We conducted a comprehensive review by analysing rainfall data from 42 metropolitan cities. Random Forest (RF) was utilised to predict the probability of rain within a day, while Support Vector Machine (SVM) was employed to forecast precipitation on a rainy day. The constraints of the Hybrid model were underscored by the decrease in daily precipitation across three sites at the rainfall level in the eastern area of Malaysia. Crossover models have demonstrated the ability to mimic total

variance, consecutive day counts, 95% of monthly precipitation, and the distribution of observed rainfall. Agriculture is essential in India. Downpour is a significant plant. At present, climate is a major problem. Climate measurement yields data on precipitation assessment and agricultural viability. A variety of methods has been established to detect precipitation. Machine learning methods are essential for forecasting precipitation. Climate will inevitably undergo change in the near future. Climatic circumstances are uncertain, shaped by various causes; the key elements are utilised in weather predictions. The choosing of such an item is greatly influenced by the timing of your decision. The foundational display is utilised to amalgamate the future of modelling, artificial intelligence applications, data interchange, and identification analysis. Unlike other areas lacking rainfall data, considerable time is necessary to conduct a thorough water evaluation over prolonged durations. Improving complex neural networks is intended to serve as an efficient instrument for predicting severe weather phenomena. This sequence of precipitation events was confirmed with an advanced perceptron neural network. Estimations like MSE (Mean Squared Error), NMSE (Normalised Mean Squared Error), and the structuring of datasets for transitional planning are apparent in the examination of numerous entities, including Adanaive. AdaSVM

PROPOSED METHODOLOGY

The design of the proposed rainfall prediction system serves as the basis for efficiently processing meteorological data and generating accurate forecasts. It includes several components, including as data collecting, preprocessing, feature extraction, model training, and prediction generation. The system employs machine learning techniques, big data analytics, and cloud computing to ensure scalability, efficiency, and accuracy in weather forecasting. The architecture employs a modular design, wherein each component performs a specific role. Meteorological data is collected from several sources, including satellites, weather stations, and IoT sensors. The raw data undergoes preprocessing to remove noise and inconsistencies before being transformed into a structured format suitable for analysis. Machine learning models are trained on past meteorological data to identify patterns and trends, which are employed for predicting future rainfall. The system architecture diagram clearly represents the components and their linkages, clarifying the data flow inside the system—from collection to final forecasts. This methodical methodology enhances the precision of precipitation forecasts, facilitating disaster management, agricultural planning, and water resource management.

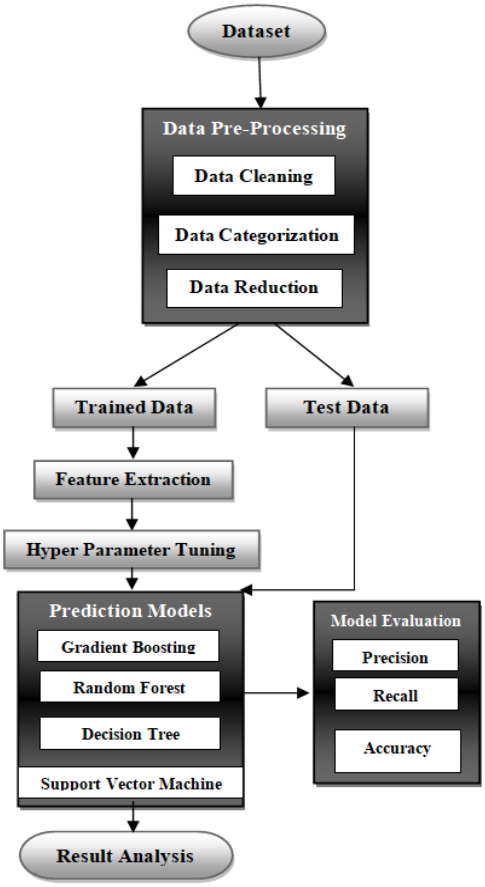


Figure 1: Demonstration of Proposed System

3.1 DATA CLEANING

A standardised framework for the data model was built, encompassing the identification of missing data, detection of duplicate data, and elimination of erroneous data. Ultimately, the sanitised data was converted into a format suitable for data mining. .

3.2 DATA PRE-PROCESSING

At this point, relevant data for the study, including the decision tree, was identified and extracted from the dataset. The Meteorological dataset had eleven attributes, of which two were employed for future forecasts. The Cloud Form data, marked by consistent values, and the sunlight data, which displays a considerable amount of missing values, were omitted from the analysis. The data mining procedure was divided into three stages. All methodologies were utilised to analyse the meteorological datasets at each stage. The study utilised a percentage split testing methodology, training on a segment of the dataset, cross-validating on that segment, and evaluating on the residual fraction. Subsequently, compelling patterns indicative of knowledge were found. .

K-NEAREST NEIGHBOR

At this point, relevant data for the study, including the decision tree, was identified and extracted from the dataset. The Meteorological dataset had eleven attributes, of which two were employed for future forecasts. The Cloud Form data, marked by consistent values, and the sunlight data, which displays a considerable amount of missing values, were omitted from the analysis. The data mining procedure was divided into three stages. All methodologies were utilised to analyse the meteorological datasets at each stage. The study utilised a percentage split testing methodology, training on a segment of the dataset, cross-validating on that segment, and evaluating on the residual fraction. Subsequently, compelling patterns indicative of knowledge were found.

RANDOM FOREST

Random forests, also known as random choice forests, are an ensemble learning method employed for classification, regression, and other tasks. They operate by creating numerous decision trees during the training phase and yielding the class that signifies the mode of the classes (for classification) or the average prediction (for regression) of the individual trees. Random decision forests reduce the tendency of decision trees to overfit their training datasets. Random forest is a supervised machine learning method based on ensemble learning. Ensemble learning is a technique that integrates many algorithms or multiple instances of the same algorithm to develop a more resilient predictive model. The random forest technique amalgamates various methods of the same type, specifically numerous decision trees, resulting in a "forest" of trees. This methodology is suitable for both regression and classification tasks.

LOGISTIC REGRESSION

This is a statistical method for analysing a dataset with one or more independent variables that affect an outcome. The outcome is evaluated using a dichotomous variable, which offers only two potential results. The aim of logistic regression is to determine the ideal model that clarifies the relationship between the binary feature of interest (dependent variable) and a set of independent components (predictors or explanatory variables). Logistic regression is a machine learning classification method used to predict the likelihood of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable denoted as 1 (affirmative, success, etc.) or 0 (negative, failure, etc.). The logistic regression model predicts P(Y=1) as a function of X.

DECISION TREE

It is one of the most powerful and extensively employed algorithms. The decision tree algorithm is categorised as a supervised learning method. It is relevant to both continuous and categorical output variables. It disaggregates a dataset into increasingly smaller segments while simultaneously incrementally developing a corresponding decision tree. A decision node contains two or more branches, whereas a leaf node represents a categorisation or decision. The apex decision node in a tree, signifying the most potent predictor, is referred to as the root node. Decision trees can handle both categorical and numerical inputs. A decision tree creates classification or regression models depicted as a tree structure. It utilises a mutually exclusive and exhaustive if-then rule set for classification. The rules are assimilated incrementally from the training data, one at a time. With each newly acquired rule, the tuples governed by the rules are discarded. This method continues on the training set until a termination criterion is met. It is constructed via a top-down recursive divide-and-conquer methodology. All attributes must be categorical.

EVALUATION METRICS

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table. The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$

F1-score	$2 * \frac{R * P}{R + P}$
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y - y\hat{}i)^2$
RMSE	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{(y - y\hat{}i)2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\lvert(y - y\hat{}i)^2\rvert$

RESULT AND DISCUSSION

Images and graphs are crucial for visualising the results and effectiveness of the rainfall forecasting system. They enhance the understanding of complex data patterns, evaluate model effectiveness, and support informed decision-making. This study use various graphical representations to analyse rainfall trends, prediction accuracy, and categorisation results. The Confusion Matrix Heatmap is a commonly employed visualisation that distinctly illustrates True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) in a color-coded matrix. This facilitates the assessment of the machine learning model's effectiveness in predicting precipitation. Additionally, Precision-Recall Curves and Receiver Operating Characteristic (ROC) Curves are utilised to evaluate the model's classification performance, demonstrating its efficacy in distinguishing between rainy and non-rainy days.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindS
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	NaN	E	31.0	SE	ENE	13.0	
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	NaN	NNW	22.0	SE	N	13.0	
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	NaN	N	37.0	SE	WNW	9.0	
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	NaN	SE	28.0	SSE	N	13.0	
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	NaN	NaN	NaN	ESE	ESE	17.0	

145460 rows × 23 columns

Figure 1: Weather Dataset

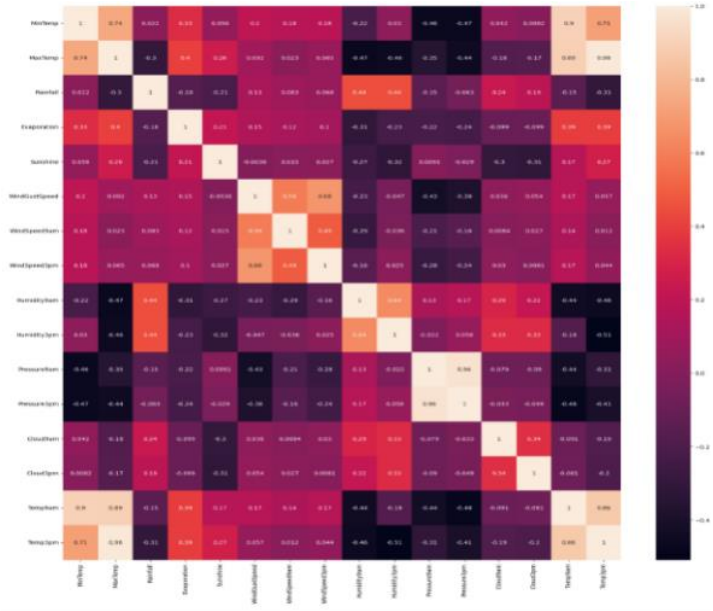


Figure 2: Confusion Matrix

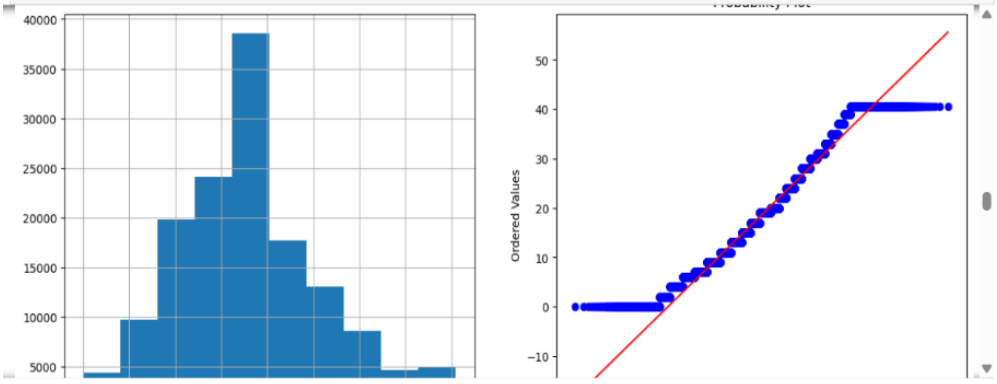
```
y_pred = cat.predict(X_test)
print(confusion_matrix(y_test,y_pred))
print(accuracy_score(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

```
[[21521  1196]
 [ 2788  3587]]
0.8630551354324213
              precision    recall  f1-score   support

           0       0.89      0.95      0.92     22717
           1       0.75      0.56      0.64      6375

    accuracy                           0.86     29092
   macro avg       0.82      0.76      0.78     29092
weighted avg       0.86      0.86      0.86     29092
```

Figure 3 : Classification Report

The Random Forest approach achieved an accuracy of 86.30%, a macro-average F1-score of 92%, and a weighted-average F1-score of 91%. The Decision Tree algorithm exhibited notable performance, with an accuracy of 78.64%. The macro-average F1-score is 72%, while the weighted-average F1-score is 87%. The Logistic Regression attained an accuracy of 83.2%. The macro-average and weighted-average F1-scores were notably low, indicating challenges in generalisation with this dataset. The KNN algorithm achieved an accuracy of 84.27%, a macro-average F1-score of 41%, and a weighted-average F1-score of 69%. Its performance, however modest, indicates challenges with datasets marked by imbalanced or non-linear relationships, as evidenced by its insufficient ability to appropriately represent all performance classes. The system classifies individual students according to the instructional models employed. The outcome predicts either Good Performance (y=1) or a Reason for Poor Performance (y=0). For example: Anticipated: "Reason for Inferior Performance: Dropout" and Identified Features: 0 (indicating the primary indicators for dropout). For each sample, the model predicts whether the cause of subpar performance is a "dropout" (Extracted Feature: 0) or associated with good performance (Extracted Feature: 4).

```
for feature in continuous_feature:
    print(feature)
    plt.figure(figsize=(15,6))
    plt.subplot(1, 2, 1)
    df[feature].hist()
    plt.subplot(1, 2, 2)
    stats.probplot(df[feature], dist="norm", plot=plt)
    plt.show()
```




```
df.to_csv("preprocessed 1.csv", index=False)
```

Figure .7. Accuracy of various Prediction models

Figure.5. By using these images and graphs, the rainfall prediction system becomes more interpretable and user-friendly, allowing stakeholders such as farmers, meteorologists, and disaster management teams to make data-driven decisions with ease.

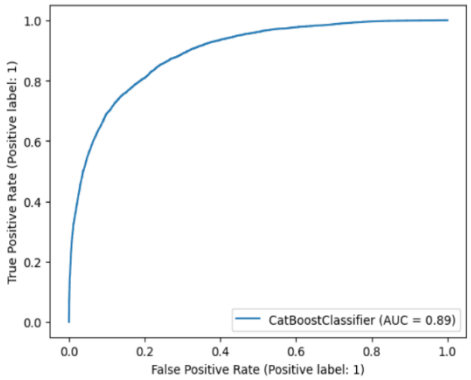


Figure 5 : True Positive and False Positive rate

Algorithm	Accuracy
Logistic Regression	0.83
RandomForest	0.86
KNeighborsClassifier	0.84
DecisionTreeClassifier	0.78

Figure 6: Algorithm Accuracy

The results validate the effectiveness of feature selection in improving model performance, with the Random Forest method recognised as the most reliable model for accurate predictions in this context. The results are evaluated to uncover substantial insights into the impact of numerous variables on student performance, identifying potential areas for model improvement and visualising the outputs for enhanced interpretability. .

CONCLUSION

Weather forecasting is a meteorological activity that enhances research via the use of numerical weather prediction techniques. Weather forecasts generated using various data mining techniques, specifically classification, clustering, neural networks, and decision trees. This study aims to improve the classification and predictive accuracy of the traditional weather forecasting model. Nonetheless, specific limits of the model have been discovered, requiring a modification of the proposed technique prior to its imminent deployment. Moreover, specific issues and challenges in soil require the enhanced utilisation of data mining tools in the field of weather forecasting. Moreover, real-time data integration and advancements in artificial intelligence can substantially enhance the effectiveness of weather forecasting. The application of deep learning techniques, ensemble models, and hybrid approaches can improve weather forecasting, which is essential for industries such as agriculture, disaster management, and water resource planning. Accurate weather forecasting allows farmers to make informed decisions on crop production, irrigation, and harvesting, therefore improving productivity. This study has analysed various data mining approaches, including classification, clustering, neural networks, and decision trees, to enhance the accuracy of rainfall forecasting. Utilising these advanced approaches, weather forecasting models can efficiently manage large datasets and provide more precise and reliable predictions. This study primarily sought to improve a weather forecast model by combining numerical weather prediction approaches with machine learning procedures. The proposed strategy, albeit demonstrating enhanced accuracy compared to traditional methods, nonetheless faces certain limitations. Unanticipated climatic changes, insufficient historical data, and variations in meteorological parameters may impact prediction accuracy. Therefore, continuous research and improvement in data collection, preprocessing, and model training are essential for system optimisation. .

REFERENCES

1. Xiong, Lihua, and Kieran M. OConnor. “An empirical method to improve the prediction limits of the GLUE methodology in rainfallrunoff modeling.” Journal of Hydrology 349.1-2 (2008): 115-124.
2. Schmitz, G. H., and J. Cullmann. “PAI-OFF: A new proposal for online flood forecasting in flash flood prone catchments.” Journal of hydrology 360.1-4 (2008): 1-14.
3. Riordan, Denis, and Bjarne K. Hansen. “A fuzzy casebased system for weather prediction.” Engineering Intelligent Systems for Electrical Engineering and Communications (2002): 139-146.
4. Guhathakurta, P. “Long-range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network.” Current Science 90.6 (2006): 773-779.
5. Pilgrim, D. H., T. G. Chapman, and D. G. Doran. “Problems of rainfallrunoff modelling in arid and semiarid regions.” Hydrological Sciences Journal 33.4 (1988): 379-400.