Rainfall Prediction Using Machine Learning

Rosebell Paul
Department of Computer Science and Engineering
SCMS School of Engineering and Technology, karukutty, Ernakulam, India

Anjali R
Department of Computer Science and Engineering
SCMS School of Engineering and Technology, Karukutty, Ernakulam,India

Abstract- Rainfall may cause a variety of tragedies, thus forecasting rainfall is crucial. The forecast aids individuals in taking precautionary precautions, and it should also be correct. Short-term rainfall forecasting and long-term rainfall forecasting are the two forms of forecasting. Predictions, particularly those for the near term, can provide us with precise results. The Heavy most difficult task is to create a model that can predict long-term rainfall. Because it is strongly related to the economy and human lifespan, heavy precipitation forecast might be a serious negative for earth science departments. It's the reason for annual natural disasters like floods and droughts that affect people all over the planet. For nations like India, whose economy is largely based on agriculture, the accuracy of rainfall estimates is critical. Applied mathematics approaches fail to give reasonable accuracy for precipitation statements due to the dynamic nature of the atmosphere. Regression may be used in the prediction of precipitation utilizing machine learning approaches. This project aims to give non-experts simple access to the techniques and approaches used in the field of precipitation prediction, as well as a comparison of different machine learning algorithms.

I. INTRODUCTION

Global warming is harming people all across the globe now, and it is hastening climate change. As a result of this, the air and oceans are warming, sea levels are rising, and flooding and droughts are becoming more common. Rainfall is one of the significant repercussions of climate change. Rainfall forecast is a difficult issue nowadays, but it is one that is taken into account by the majority of the world's main authorities. Rainfall is a climatic phenomenon that has an impact on a number of human activities, including agricultural output, building, power generation, and tourism, to name a few. As a result, rainfall is a major problem, necessitating better rainfall forecasting. Rainfall is a complicated atmospheric phenomenon that is becoming more difficult to anticipate as the climate changes. Rainfall series are frequently labeled by a stochastic process due to their arbitrary properties. Floods and droughts are becoming increasingly regular, as the Indian state of Uttarakhand had its greatest natural calamity in June 2013. In comparison to normal monsoon rainfall, there was around 400 percent more rainfall. Roads and bridges were entirely wrecked by such strong rains, trapping 100,000 pilgrims and tourists on their "Char Dham Yatra." This calamity could not have been foreseen by the government, huge companies, risk management experts, or the scientific community prior to the occurrence. These factors may also contribute to landslides, a significant geohazard that has resulted in the loss of lives and property across the world. We employed a variety of machine learning techniques and models to create accurate and timely predictions in order to resolve this ambiguity. This study intends to give an end-to-end machine learning life cycle, starting with data preparation and ending with model implementation and evaluation. Imputing missing values, feature transformation, encoding categorical features, feature scaling, and feature selection are all processes in the data preprocessing process. Models including Logistic Regression, DecisionTree, K Nearest Neighbour, Rule-based, and Ensembles were applied. Accuracy, Precision, Recall, F-Score, and Area Under Curve were employed as assessment criteria for this study. We used Australian weather data from numerous weather stations in Australia to train our classifiers for our studies.

II. PROPOSED ALGORITHM

Machine Learning

We're approaching the artificial intelligence time zone, in which the machine controls and manages everything. Computer learning is a branch of artificial intelligence in which we train a machine to learn on its own, without the assistance of a human. In machine learning, we educate the computer to learn from its prior data and aim to improve its results in the future by applying what it has learned. The use of tools, methods, and procedures that assist machine learning in producing better outcomes is a part of machine learning. These algorithms and approaches provide machines and humans a new way to discover new information from existing data or by utilizing standard datasets. We aim to capture and then mimic certain behaviours in particular scenarios. As a result, modelling encourages individuals to learn more about the circumstance. Statistics has a little history with machine learning approaches. It's useful for figuring out how to extract the genuine message from a big quantity of data using a more advanced learning model. Although both machine learning and classical statistics techniques may be used to analyse data, their underlying concepts and properties differ significantly. In comparison to statistical data analysis, machine learning has several distinct advantages, including the ability to analyse large amounts of data and real-time data streams with mixed values, the ability to choose from a variety of learning models, and the ability to manage the results. We can also detect difficult patterns that cannot be stated in various mathematical terms, visualize the data for generating a forecast, and link the learning models with other databases management systems. Learning is a never-ending process, but machines, like people, try to encourage it. Learning is divided into three categories for machines: supervised, unsupervised, and reinforced. All of these factors influence how a learning model is designed. In supervised learning, the system specifies the number of labels to be used for data division as well as the intended output. The algorithm may "learn" by comparing the real output with the learned outputs, identifying faults, and adjusting the model as needed. Unsupervised learning data, on the other hand, is presented unlabelled, thus it is separated according to similarity across input data. The algorithm's learning task is to detect similarities among its input data. Because unlabelled data is large, machine learning algorithms that promote unsupervised learning are very useful.

SVM (Support Vector Machine)

SVM is used to solve classification issues by determining the best fit line between classes, often known as the hyperplane. The prediction's confidence is proportional to its distance from the hyperplane. As a result, we must increase the buffer between the hyperplane and the closest data point as much as feasible. SVM works by applying a kernel technique to low-dimensional input data in order to locate the best-fit line or decision boundaries in a high-dimensional feature space. There are no hyperplanes between the classes, however, the hyperplane that leaves the most margin from both classes is the best choice. SVM stands for "systematic variable management." It's an iterative approach for solving the optimization problem that emerges during SVM training. It's often utilized to divide an issue into no more than a few subproblems. Lin et al. forecasted rainfall using SVM-based models that included and excluded typhoon features. The classes are separated using the feature of two SVM class ideas on a plane termed the separating hyperplane, which is comparable to a dimension space. The hyperplane is chosen in such a way that both classes are distanced from it by a maximum distance; this plane is known as the maximum margin hyperplane. The separating hyperplane's equation is stated in the equation below:

$$w^{*}X + b = 0$$

where Xi is a d-dimensional feature matrix containing features of the classes to be separated, b is the bias, w is the hyperplane's normal, |b|/ ||w|| is the perpendicular distance between the hyperplane and the origin, and ||w||2 is the Euclidean norm of w.

Naive Bayes

The Bayes theorem, which is based on the probability theory, is used to create the Naive Bayes classifier. It's a sophisticated predictive modelling method. For categorizing high-dimensional training datasets, Naive Bayes is commonly employed. It's a probabilistic classifier since it employs the probability theorem to classify data and focuses on assuming that the existence of one characteristic is independent of the existence of other features, which is why it's named Naive.

The Bayes component originates from the Bayes theorem or rule, which offers us a way of calculating conditional probability, which is the likelihood of an occurrence dependent on some prior information about it. The Naive Bayes approach is based on Bayes' theorem, which yields P(C|A) from P(C), P(A), and P(A|C):

P(A|C) = P(A|C) = P(A|C) = P(A|C) = P(A|C P(C)
P(A)
The probability (Conditional probability) of event C occurring if event A is true is equal to the probability (Conditional probability) of event A occurring if event C is true multiplied by the likelihood of C upon the probability of A.

Random Forest

A classification tree is constructed using the Random Forest supervised learning approach. To classify a new item from an input feature vector, this technique inserts an input data vector into each forest tree. Each tree in the forest gives each other "votes," and the tree with the most votes is classified. Pick "k" characteristics at random from a total of "m." Using the best split, break the element into many elements.
Both classification and regression tasks may be performed by algorithms. It's a collection of decision trees; the greater the number of trees in the forest, the more reliable and accurate the findings.

MLP (Multiple perceptrons)

MLP refers to a type of feed-forward ANN. There are at least three layers of nodes in an MLP. Each node, with the exception of the input nodes, is a neuron with a nonlinear activation function. For training, MLP employs a supervised learning approach known as backpropagation. A multilayer perceptron is a neural network that connects many layers in a directed graph, meaning that the signal route via the nodes is one-way only. A nonlinear activation function exists for every node except the input nodes. The supervised learning approach used by an MLP is backpropagation. MLP is a deep learning approach since there are numerous layers of neurons. It tries to mimic how the human brain functions. In an artificial neural network made up of hardware and GPUs neural networks, a neural network design is encouraged by biological neural networks and can be made up of numerous layers. One layer's output is used as the input for the next layer. Deep learning methods can be either supervised or unsupervised, depending on whether they are used to categorize data or to do pattern analysis.
We used rainfall figures from the data.gov.in website of the Indian government in our work. To anticipate rainfall, we're utilizing machine learning algorithms and trying to figure out which one is the best. This procedure leads to the following steps:

Step 1: Collect the rainfall dataset from the open repository data.gov.in with no. of multiple features.

Step 2: Data Cleaning, Data Pre-processing, and feature selection.

Step 3: Output will be an algorithm with the optimized result.

The data was gathered from the Indian government's official website, data.gov.in.
Data pre-processing is used to prepare the data. The information we utilize comes from the Indian government's official website. This gives us the rainfall volume of data in millimetres from 1901 to 2013. These data sets include rainfall data from January to December for each month of the year. The information is skewed, partial, and lacking. For more relevant inputs to process and analyse, we must undertake feature selection. We utilized monthly rainfall volume in millimetres in the study to conduct the experiment.
The information is skewed, partial, and absent. For more relevant inputs for processing and analysis, we must undertake feature selection. We utilized monthly rainfall volume in millimetres in the article for the experiment.
To normalize the range of independent variables, we use data normalization or feature scaling. We used the conventional scalar formula as follows to rescale data between [0,1]:
Xi – mean(x) / standard deviation (x)
The standard deviation is denoted by stdev. The official website shows the total rainfall in mm over India over the last 23 years. We employ a variety of machine learning algorithms to forecast the following month's rainfall based on train data from prior months.

We mostly used the early months to train the data.

```
In [3]: file.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113 entries, 0 to 112
Data columns (total 18 columns):
YEAR        113 non-null int64
JAN         113 non-null float64
FEB         113 non-null float64
MAR         113 non-null float64
APR         113 non-null float64
MAY         113 non-null float64
JUN         113 non-null float64
JUL         113 non-null float64
AUG         113 non-null float64
SEP         113 non-null float64
OCT         113 non-null float64
NOV         113 non-null float64
DEC         113 non-null float64
ANN         113 non-null float64
Jan-Feb     113 non-null float64
Mar-May     113 non-null float64
Jun-Sep     113 non-null float64
Oct-Dec     113 non-null float64
dtypes: float64(17), int64(1)
memory usage: 16.0 KB
```

```
In [4]:
```

Out[2]:

	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	\
0	1901	34.7	38.6	17.8	38.9	50.6	113.2	241.4	271.6	124.7	52.4	
1	1902	7.4	4.2	19.0	44.1	48.8	111.7	284.9	201.0	200.2	62.5	
2	1903	16.7	8.0	31.1	17.1	59.5	120.3	293.2	274.0	198.1	119.5	
3	1904	14.9	9.7	31.4	33.7	73.8	165.5	260.3	207.7	130.8	69.8	
4	1905	24.7	20.3	41.8	33.8	55.8	93.7	253.0	201.7	178.1	54.9	

	NOV	DEC	ANN	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	38.7	8.2	1030.8	73.2	107.3	751.0	99.3
1	29.4	25.2	1038.4	11.6	111.9	797.8	117.2
2	40.3	18.0	1195.9	24.7	107.7	885.6	177.8
3	11.2	16.4	1025.1	24.5	138.8	764.3	97.4
4	9.6	10.1	977.5	45.0	131.4	726.4	74.7

We use a month as a feature in our work and try to estimate the rainfall for the next month. The rainfall in each month is depicted in the graph below.

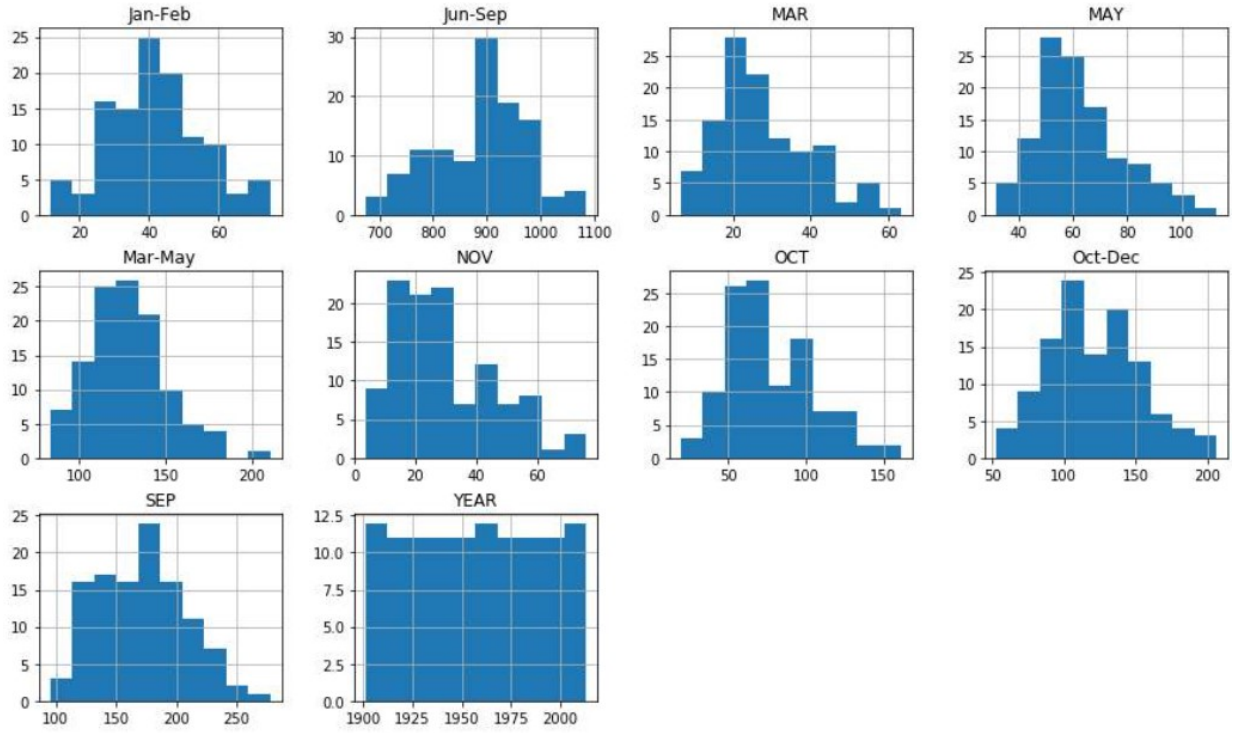

Figure 1. Monthly Rainfall

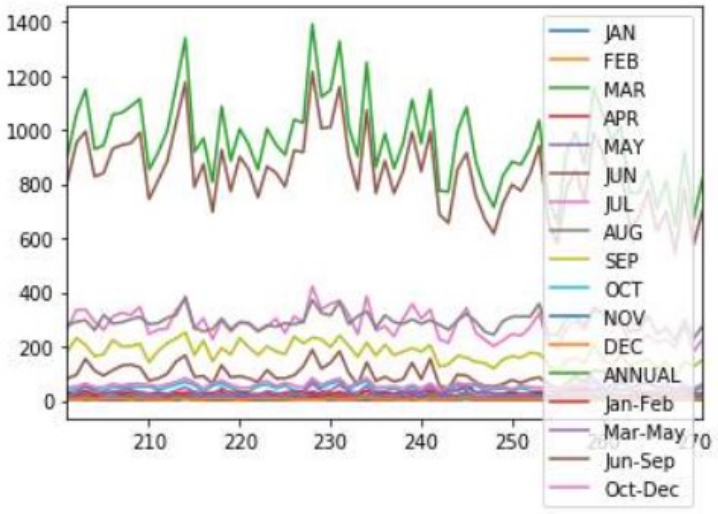

Figure 2. Each month's rainfall (mm)

III. Experiment and Result

For all the experiments and development of classifiers, we used Python 3 and Google collab's Jupyter Notebook. We used libraries such as Sckit Learn, Mat-Plot lib, Seaborn, Pandas, Numpy and Imblearn. We used weka for implementing Decision Table.

We carried out experiments with different input data; one with the original dataset, then with the undersampled dataset and the last one with the oversaw-pled dataset. We split out the dataset in the ratio of 75:25 for training and testing purpose.

Experiment 1 - Original Dataset:- Post all the preprocessing steps (as mentioned above in the Methodology section), we ran all the implemented classifiers each one with the same input data (Shape: 92037 x 4). Figure 12 depicts two considered metrics (10-skinfold Accuracy and Area Under Curve) for all the classifiers. Accuracy wise Gradient Boosting with a learning rate of 0.25 performed best, coverage wise Random Forest and Decision Tree performed worsts.

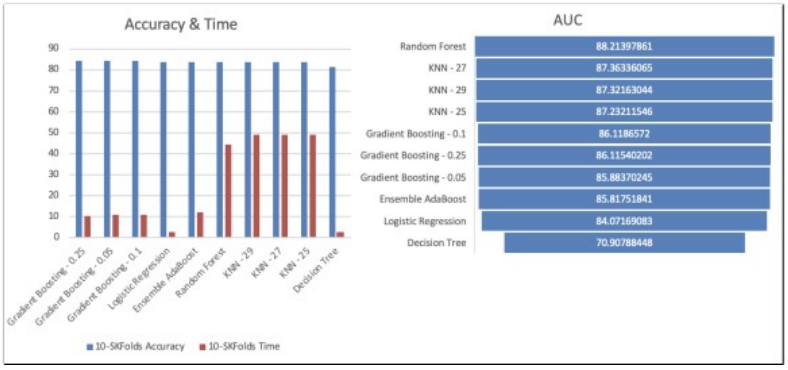

Figure 3. Experiment 1

Experiment 2 - Under sampled Dataset: Post all the preprocessing steps (as mentioned above in the Methodology section) including the under sampling step, we ran all the implemented classifiers each one with the same input data (Shape: 54274 x 4). Figure 13 depicts two considered metrics (10-skinfold Accuracy and Area Under Curve) for all the classifiers.

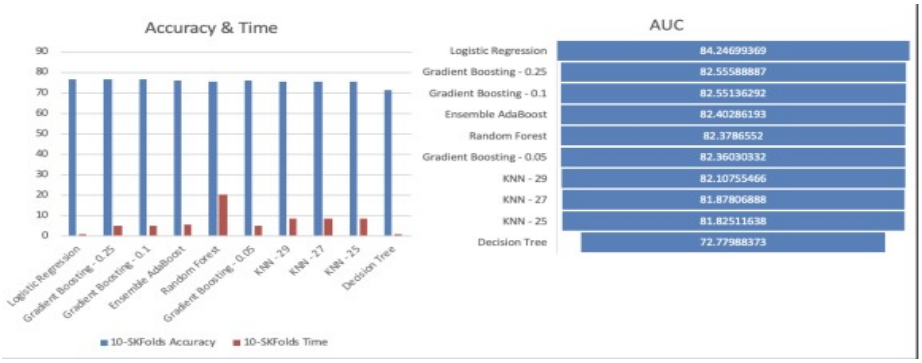

Figure 4 Experiment 2

Accuracy and coverage-wide logistic Regression performed best and Decision Tree performed worsts.

Experiment 3 - Oversampled Dataset: Post all the preprocessing steps (as mentioned above in the Methodology section) including the oversampling step, we ran all the implemented classifiers each one with the same input data (Shape: 191160 x 4). Figure 14 depicts two considered metrics (10-skfold Accuracy and Area Under Curve) for all the classifiers.

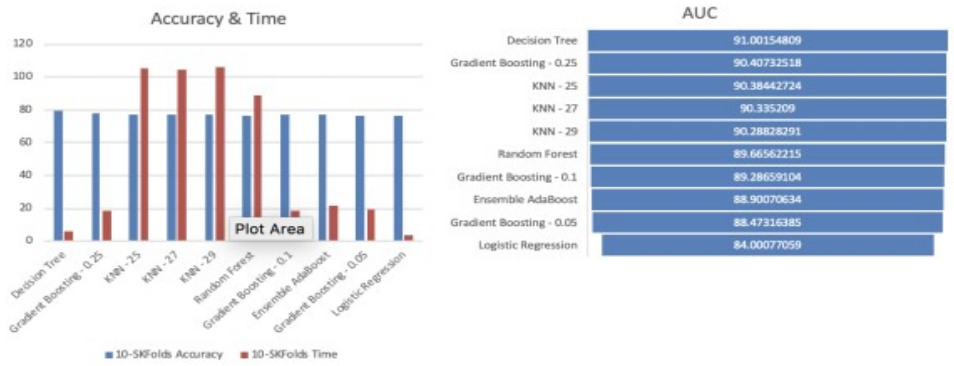

Figure 5. Experiment 3.

Accuracy and coverage wise Decision Tree performed best and Logistic Regression performed worsts. We have a varying range of results with respect to different input data and different classifiers. Other metrics are followed in the appendix.

IV.CONCLUSION

We investigated and implemented numerous pre-processing techniques in this article to see how they affected the overall performance of our classifiers. We also compared all of the classifiers with various input data to see how the input data influenced model predictions. We may infer that Australian weather is unpredictable and that there is no such relationship between rainfall and location or time in Australia. We discovered certain patterns and linkages in the data, which aided in the identification of key characteristics. See the appendix for more information. We may use Deep Learning models like Multilayer Perceptron, Convolutional Neural Network, and others since we have such a large amount of data. A comparison between machine learning classifiers with deep learning models would be quite useful.

REFERENCES

[1] "Heuristic rainfall prediction using machine learning approaches," by Chandrasegar Thirumalai and colleagues. International Conference on Electronics and Informatics Trends 2017 (ICEI). 2017 IEEE.
[2] "Data mining for meteorological applications: Decision trees for modeling rainfall prediction," by A. Geetha and G. M. Nasira. The 2014 IEEE International Conference on Computational Intelligence and Computing Research is a conference on computational intelligence and computing research. 2014, IEEE
[3] "Machine learning approaches for rainfall prediction: A review," say, Aakash Parmar, Kinjal Mistree, and Mithila Sompura. 2017 International Conference on Information Embedded and Communication Systems.
[4] "Rainfall forecast for Kerala state of India using artificial intelligence methodologies," by Yajnaseni Dash, Saroj K. Mishra, and Bijaya K. Panigrahi. 66-73 in Computers and Electrical Engineering, vol. 70, no.
[5] Deepak Kumar, Gurpreet Singh, and Singh "Forecasting Rainfall using Hybrid Prediction Models." The 9th International Conference on Cloud Computing, Data Science, and Engineering (Cloud Computing, Data Science, and Engineering) will take place in 2019. (Confluence). 2019 IEEE International Conference on.
[6] "Prediction of Rainfall Using Fuzzy Dataset," by Kaveri Kar, Neelima Thakur, and Prerika Sanghvi. (2019).
[7] "Rainfall Prediction: A Comparative Study of Neural Network Architectures," by Kaushik D. Sardeshpande and Vijaya R. School. Data Mining and Information Security: Emerging Technologies Singapore: Springer, 2019. 19-28.
[8] "Non-Linear Machine Learning Approach to Short-Term Precipitation Forecasting," by Binghong Chen and colleagues. (2018).
[9] "Application of machine learning to an early warning system for very short-term heavy rainfall." Journal of Hydrology 568 (2019): 1042-1054. Moon, Seung-Hyun, et al., "Application of machine learning to an early warning system for extremely short-term heavy rainfall."
[10] https://data.gov.in/resources/subdivision-wise-rainfall-andits-departure-1901-2015