Rainfall Prediction

Almas Banu
PES1UG20CS535
B Tech Computer Science and Engineering
PES University
Email: almas.banu082002@gmail.com

Brinda N
PES1UG20CS542
B Tech Computer Science and Engineering
PES University
Email:brinda.n162@gmail.com

***Abstract*—Rainfall Prediction is one of the difficult and uncertain tasks that have a significant impact on human society. In India,Agriculture contributes major role to Indian economy. For agriculture, Rainfall is important but during these days' rainfall prediction has become a major challenging problem. Good prediction of rainfall provides knowledge and know in advance to take precautions and have better strategy about theirs crops. Global warming is also having severe effect on nature as well as mankind and it accelerates the change in climatic conditions. Because of its air is getting warmer and level of ocean is rising, leads to flood and cultivated field is changing into drought. Due to adverse climatic change leads to unseasonable and unreasonable amount of rainfall. To predict Rainfall is one of the best techniques to know about rainfall and climate.**

I. Introduction

Rainfall prediction is utmost necessary all over world and it plays a key role in human life. It's cumbersome responsibility of meteorological department to analyze the frequency of rainfall with precariousness. It is difficult to forecast the rainfall precisely with varying atmospheric condition. It is conjectured to predict the rainfall for both summer and rainy seasons. Rainfall prediction remains a serious concern and has attracted the attentionof governments, industries, risk management entities, as well as the scientificcommunity. Rainfall is a climatic factor that affects many human activities likeagricultural production, construction, power generation, forestry and tourism,among others [1]. To this extent, rainfall prediction is essential since this vari-able is the one with the highest correlation with adverse natural events suchas landslides, flooding, mass movements and avalanches. These incidents haveaffected society for years [2]. Therefore, having an appropriate approach for rain-fall prediction makes it possible to take preventive and mitigation measures forthese natural phenomena [3].

An erratic rainfall distribution in the country affects the agriculture on which the economy of the country depends on. Wise use of rainfall water should be planned and practiced in the country to minimize the problem of the drought and flood occurred in the country.

In order to predict the rainfall,there are many techniques,one of such skilled and effective technologies is Machine Learning. Machine learning covers various classifiers of Supervised, Unsupervised and Ensemble Learning which are used to predict and find the accuracy of the given dataset.We can use this knowledge of Machine Learning to predict the rainfall which will help a lot.

There are many Machine Learning algorithms like Logistic Regression, Decision Tree, K-Nearest Neighbor, Random Forest are compared to find the most accurate model.This study presents a set of experiments that involve the use of common machine learning techniques to create models that can predict whether it will rain tomorrow or not based on the weather data for that day in major cities in Australia.

II. Feature Selection

Feature Selection Feature Selection is the process where you automatically or manually select those features which contribute most to our prediction variable or output. Having irrelevant features in data can decrease the accuracy of the models and make the model learn based on irrelevant features. Feature selection helps to reduce over fitting, improves accuracy and reduces training time. We used two techniques to perform this activity and got the same results. 1)Univariate Selection: Statistical tests can be used to select those features that have the strongest relationship with the output variable. The scikitlearn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features. We used chi-squared statistical test for non-negative features to select 5 of the best features from our data set. 2)Correlation states how the features are related to each other or the target variable. Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable). Heatmap makes it easy to identify which features are most related to the target variable, we plotted heatmap of correlated features using the seaborn library.

III. Models

1)Logistic Regression is a classification algorithm used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. We can also think of logistic regression as a special case of linear regression when the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit

function. Hence, this makes Logistic Regression a better fit as ours is a binary classification problem.

2)Decision Tree have a natural if then else construction that makes it fit easily into a programmatic structure. They also are well suited to categorization problems where attributes or features are systematically checked to determine a final category. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables. This characteristics of Decision Tree makes it a good fit for our problem as our target variable is binary categorical variable.

3)K - Nearest Neighbour is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure is determined from the dataset. Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. KNN performs better with a lower number of features than a large number of features. We can say that when the number of features increases than it requires more data. Increase in dimension also leads to the problem of overfitting. However, we have performed feature selection which helps to reduce dimension and hence KNN looks a good candidate for our problem.

4) Random Forest is a supervised ensemble learning algorithm. Ensemble means that it takes a bunch of weak learners and have them work together to form one strong predictor. Here, we have a collection of decision trees, known as Forest. To classify a new object based on attributes, each tree gives a classification and we say the tree votes for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

5)Neural Network - Artificial Neural Networks is one of the most popular machine learning and deep learning algorithms. They are machine learning and deep learning algorithms. They are inspired by human neurons which are capable of making human like decisions with the help of computations. For example, in our case we trained the Neural Networks with different features like humidity, temperature, pressure, etc and they learn to identify and analyze the rainfall based on these features using the results of training dataset.

6)XGBoost gradient descent - XGBoost stands for eXtreme Gradient Boosting; it is a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model. XGBoost is implemented for the supervised machine learning problem that has data with multiple features of xi to predict a target variable yi. Most authors use XGBoost for different regression and classification problems due to the speed and prediction accuracy of the algorithm.

IV. Evaluation

For evaluating our classifiers we used below evaluation metrics: 1)Accuracy is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class. As we have, imbalanced data, we will also consider other metrics.

2)Precision is the number of correct positive results divided by the number of positive results predicted by the classifier.

3)F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells how precise our classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model.

4) Confusion Matrix - Confusion Matrix gives us a matrix as output and describes the complete performance of the model. It focuses on True Positives - the cases in which we predicted YES and the actual output was also YES; True Negatives - the cases in which we predicted NO and the actual output was NO; False Positives - the cases in which we predicted YES and the actual output was NO; False Negatives - the cases in which we predicted NO and the actual output was YES.

V. Related Works

The study by Arnav Garg and Kanchipuram [4] shows three machine learning algorithm experiments such as support vector machine (SVM), support vector regression (SVR), and K-nearest neighbor (KNN) using the patterns of rainfall in the year. The SVM algorithm performs best among the three machine learning algorithms. This research did not show the experiment result that which environmental features impact the intensity of rainfall. This paper shows the environmental features that have a positive and negative impact on rainfall and predicts the daily rainfall amount using those features.

VI. Future Scope

Various ML-driven technologies are sought to automate data discovery, reducing the gap in information ingestion across both spatial and temporal scales. Decision trees are ideal for multiple variable analyses, it is particularly important in current problem-solving task like weather forecasting.

Acknowledgment

The authors would like to thank...

References

[1] upta D,Ghose U. A Comparative Study of Classification Algorithms for Forecasting Rainfall. IEEE. 2015.

[2] lcntara-Ayala, I.: Geomorphology, natural hazards, vulnerability and preventionof natural disasters in developing countries. Geomorphology 47(24), 107124 (2002).

[3] icholls, N.: Atmospheric and climatic hazards: Improved monitoring and predictionfor disaster mitigation. Natural Hazards 23(23), 137155 (2001).

[4] rnav G, Kanchipuram Tamil Nadu. Rainfall prediction using machine learning. Int J Innovative Sci Res Technol. 2019.