https://doi.org/10.46335/IJIES.2022.7.9.5 **e-ISSN: 2456-3463**
Vol. 7, No. 9, 2022, PP. 21-27

International Journal of Innovations in Engineering and Science, www.ijies.net

An Analysis of Machine Learning Algorithms For Forecasting Rainfall

Kamna Mishra[1], Snehi Jaiswal[2], Prashant Mishra[3], Rhutik Giradkar[4], Shrikant Kalar[5], Ravindra Kale[6]

[1 2 3 4 5] Student, Department of Computer Science & Engineering,
[6]Assistant Professor, Department of Computer Science & Engineering,
G H Raisoni Institute of Engineering and Technology, Nagpur, India

kamna.mishra.cs@ghrietn.raisoni.net

***Received on**: 11 June,2022* ***Revised on**: 05 August ,2022,* ***Published on**: 09 August ,2022*

Abstract – *Rain forecasting has become a research topic of increasing relevance in recent years due to its high complexity and numerous applications, including forecasting floods and monitoring pollutant concentrations. The motive of this study was to investigate various algorithms for foreseeing rainfall in Australia. Because different algorithms have varying levels of accuracy, the choice of algorithm is dependent upon the specific requirements. The following algorithms are used in this study: Artificial Neural Networks (ANN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB).*

Keywords- ***Rainfall Prediction, Machine Learning, Logistic Regression, Support Vector Machine, Artificial Neural Network, Naive Bayes, Random Forest.***

I -INTRODUCTION

The forecasting of rain is a very important task, as heavy and irregular rainfall can cause damage to crops and farms as well as damage to property. In order to mitigate danger to life and property and also manage agricultural farms more effectively, a better forecasting model is essential. Farmers are primarily helped by this prediction and also waterworks can be used effectively. Rainfall prediction is a challenging task, so the results have to be accurate.

Hardware devices like weather stations that are traditionally used to predict rainfall, do not work well efficiently, so we are using machine learning methods to forecast rainfall more accurately. With the historical data analysis of rainfall, we can predict the future rainfall based on historical data.

In addition, we can also determine the error between the prediction and the actual data by applying classification and regression methods, according to requirements. The precision of different methods varies, so it is important to choose the appropriate algorithm based on the requirements.

The algorithms compared in this project are:
1. Random Forest
2. Support Vector machine
3. Artificial Neural Network
4. Naive Bayes
5. Logistic Regression

II - LITERATURE REVIEW

It is extremely important to predict rainfall for agriculture, and in turn for the whole living world. Unfortunately, an incorrect prediction can result in destruction and death. This is done by analysing and comparing various techniques to determine which is most suitable.

Urmay Shah (2018) [1], proposed a method for forecasting rainfall in which he used a blend of different machine learning and forecasting techniques. By adjusting a handful of parameters, we can get impressive classification accuracy, even though rainfall is affected by a wide range of factors. In addition, once we classify rain into eight categories, the accuracy remains acceptable. Forecasts are validated using RMSE. Based on empirical results, ARIMA works best for maximum temperature, minimum temperature is predicted using neural networks, and relative humidity and wind speed by SVR. Classification performance can be measured in terms of accuracy, precision and recall. Random forest is the best classifier for rainfall classification based on the ROC curves.

R .Senthil Kumar (2016) [3], presented in his paper a review of various popular data mining techniques for rainfall prediction. Along with conventional methods, Data Mining utilizes machine learning techniques. Additionally, these techniques can be used to generate simulations or decision models, based on historical data. This analysis also combines several other algorithms. Some of the algorithms discussed in this paper include Naive Bayes, K-Nearest Neighbor algorithms, Decision Trees, ANFIS, ARIMA, SLIQ, and neural networks. There is a comparison between decision trees and k-means clustering in this paper, which demonstrates their suitability for this application. After a certain point, the accuracy decreases as the training set becomes larger.

Moulana Mohammed (2020) [6], in his paper concentrated on the estimation of rainfall and concluded that SVR would be useful and adaptable in this context. SVR displaying is based on the selection of bit capacity.
Step 1: **Data Collection** - Rainfall Prediction begins with data collection, which involves gathering information from all relevant sources to find answers to research questions, test hypotheses, and assess outcomes.

Step 2: **Data Exploration** - During data exploration, users explore a large dataset in an unstructured manner to uncover patterns, characteristics, and points of interest.

We examined the dataset and took in the statistics and correlation between the features.

Step 3: **Data RedundancyCheck** - A variety of data stores are used during data integration in data mining. Data redundancy can arise as a result. Data set attributes (columns or features) formed from other attributes or sets of attributes is called redundant. It is also possible for attribute or dimension naming inconsistencies to cause redundancies.

We discovered that in our dataset there are several missing values and distinctive data types. To remove these redundancies, we have to perform data cleaning.

It instructs tenderfoots to use straight and RBF pieces separately for direct and non-straight relationships. Accordingly, SVR is preferable to MLR as an expectation strategy. In such cases, MLR cannot capture the non-linearity present in a data set, and SVR becomes useful. SVR gives the best result as expected.

III- METHODOLOGY

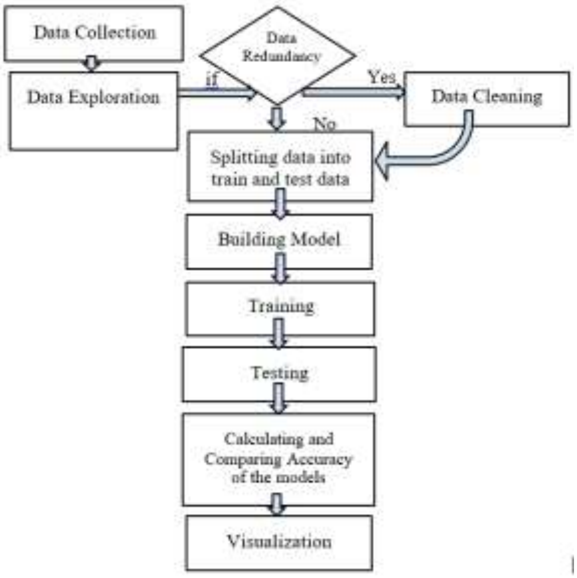


Fig. 1. Flow chart for Rainfall Prediction

Step 4: **Data Cleaning** - Model building relies heavily upon data cleaning. The process of data cleaning is necessary, but it is often overlooked. Good information management is dependent on good data. Problems relating to data quality can occur anywhere in the information system. They are solved through data cleansing. In data cleaning, incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data is fixed or removed from a dataset. The outcome and algorithms are unreliable if the data is incorrect, even though they appear to be correct. It is inevitable that some data will be duplicated or mislabelled when combining several data sources.

We utilized the simple imputer and most frequent method to cope with missing variables. Furthermore, we used the label encoder to convert non – numerical values to numerical values.

Step 5: **Separating data into train data and test data** - The train-test split is applied to an algorithm for estimating how well it performs when used for prediction-based applications. Using this method allows us to compare our own machine learning model results with those of a machine. Testing is done with 20 % of the actual data and training with 80 % of the actual data. Train set statistics are used to fit the model to the train set. In the second set of data, the only purpose is to make predictions.

We separated the training and testing sets of our modelling dataset by 0.2.

Step 6: **Calculating and Comparing Accuracy** - In this step, the accuracy of the algorithms is calculated and then they are compared based on their accuracy and the best analytical algorithm is selected.

Step 7: **Visualization** -

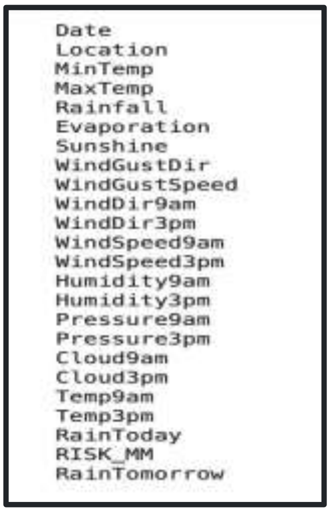
The definition of data visualization involves a graphical representation containing information and data. A data visualization technique makes trends, outliers, and patterns in data easier to see and understand thanks to visual elements such as charts, graphs, and maps.

We visualized the results using graphs and table

III- MODELS

A. Material

For this project, we have used the Rainfall in Australia dataset from kaggle. About 10 years of daily weather observations are included in this dataset from a number of Australian weather stations. In total, the dataset contains 142193 rows and 23 columns.
Below is the snapshot of Features in the dataset:



Date
Location
MinTemp
MaxTemp
Rainfall
Evaporation
Sunshine
WindGustDir
WindGustSpeed
WindDir9am
WindDir3pm
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Temp9am
Temp3pm
RainToday
RISK_MM
RainTomorrow

Fig. 2-. Features in dataset

Below is the snapshot of dataset:

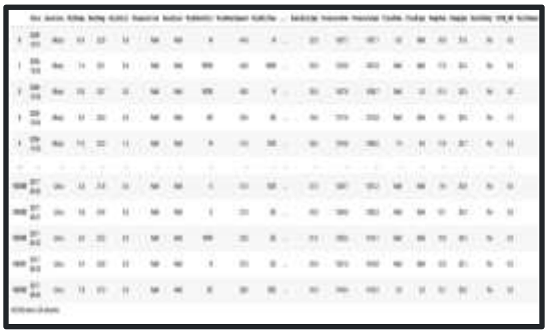


Fig. 3-Glimpse of the dataset

Our modeling dataset is split into training dataset and testing dataset and the test size is 0.2. To determine how well our machine learning model performs, we need to split a dataset into train and test sets. The statistics of the train set are known to fit the model. The second set is called the test data set and is only used to make predictions. Testing is done with 20 % of the actual data and training with 80 % of the actual data.

Therefore, Rain Tomorrow can be predicted utilizing these 23 features. As long as there is at least 1 mm of precipitation, the weather is classified as rainy. The dependent variable (Y) is whether it will rain tomorrow in Australia, and the model must make good predictions.

B. Machine Learning Models

There are many techniques that can be used including classifications, regressions depending on the requirement. We can calculate both the error between the actual and predicted values as well as their accuracy. Because different methods yield differing levels of accuracy, it is essential to pick the right algorithm and model it accordingly.

We are using following algorithms for our study:

1. Support Vector Machine (SVM):

A Support Vector Machine is a classifier and regression engine that uses supervised learning. Although we can refer to regression problems as well, classification problems are better suited. The hyper-plane that is generated in this algorithm is in an N-dimensional space that categorizes the data points clearly. Each feature determines the size of the hyper-plane. There are only two features, so the hyper-plane is a line. Three input features turn the hyper-plane into a two-dimensional plane. More features make it hard to imagine.

As compared to other machine learning algorithms, SVMs have their own approach to implementation, due to their capability to deal with multiple continuous as well as categorical variables, they are extremely popular right now.

Working of SVM:

The SVMs represent different classes in a multidimensional hyper-plane. SVM will generate hyper-planes iteratively to minimize error. By using SVM, you can analyse a dataset to determine the maximum marginal hyper-plane (MMH).

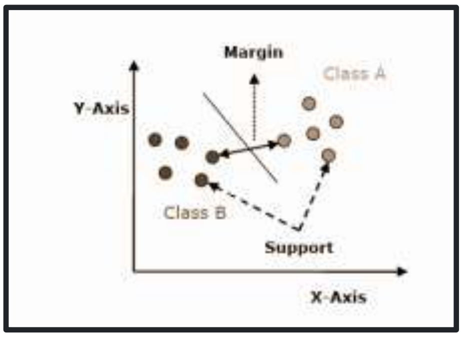


Fig.4. Support Vector Machine

Here are some important concepts in SVM: −

1. In a hyper-plane, support vectors are the nearest points. By using these points, separating lines can be defined.
2. The hyper-plane is a plane or open decision space that encloses objects of different classes, as shown in the diagram.
3. Margin refers to a distance between two lines drawn between points of different classes. A line is measured in terms of its perpendicular distance to the support vectors. A major objective of the SVM algorithm is to find a maximum marginal hyper-plane (MMH) by classifying datasets. SVM will create hyper-planes that distinguish the classes most effectively iteratively. Afterward, it will select the appropriate hyper-plane for separating the classes.

2. Artificial Neural Network (ANN):

The central idea of ANN is analogous to biological neural networks and is used as a means of efficient computing. Besides being referred to as a neural network, it is also called an artificial neural network, or a parallel distributed processing system. These systems acquire a large collection of units that are interconnected so that they can communicate with each other. The units, also called nodes or neurons, operate in parallel. A connection link connects every neuron to another neuron. An input signal's weight is assigned to every link that has a connection. Neurons use this information to solve a specific problem, since the weight is usually what excites or inhibits the signal that is being communicated. The activation signal is the internal state of each neuron. Other units may receive the output signals produced by using input signals and activation rules together.

ANN model: The following diagram illustrates the ANN model and its processing.

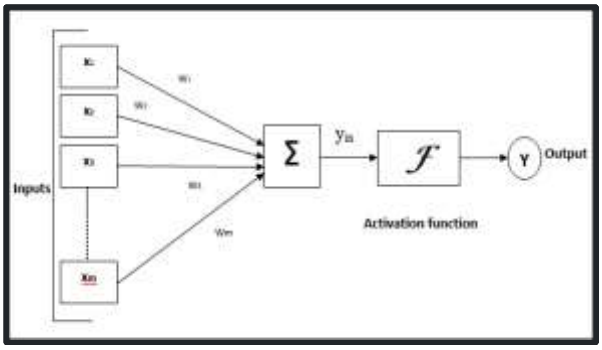


Fig.5. Artificial neural network

This general model of artificial neural network leads to the following net input calculation

Equation 1:

yin=x1.w1+x2.w2+x3.w3…xm.wm

Equation 2:

yin= ∑ mixi.wi

Activating the net input and calculating the output is possible.

Equation 3:

Y=F(yin)

Output = function netinputcalculated

3. Naive Bayes (NB):

In the Naive Bayes algorithm, Bayes' theorem is applied with the strong assumption that the predictors are independently related. An attribute's presence in one class is assumed to be independent of its presence in any other class. Bayes' classification concerns finding posterior probabilities, or the probability of a label given a set of observed features,P(L| features).

This can be expressed in quantitative form by means of Bayes theorem as follows −

Equation 4:

$$P(L|features) = P(L)P(features|L)/P\ (features)$$

Here, $P\ (L \mid features)$ - posterior probability of class.

$P(L)$- prior probability of class.

$P(features \mid L)$ - probability of predictor given class.

$P(features)$ - prior probability of predictor.

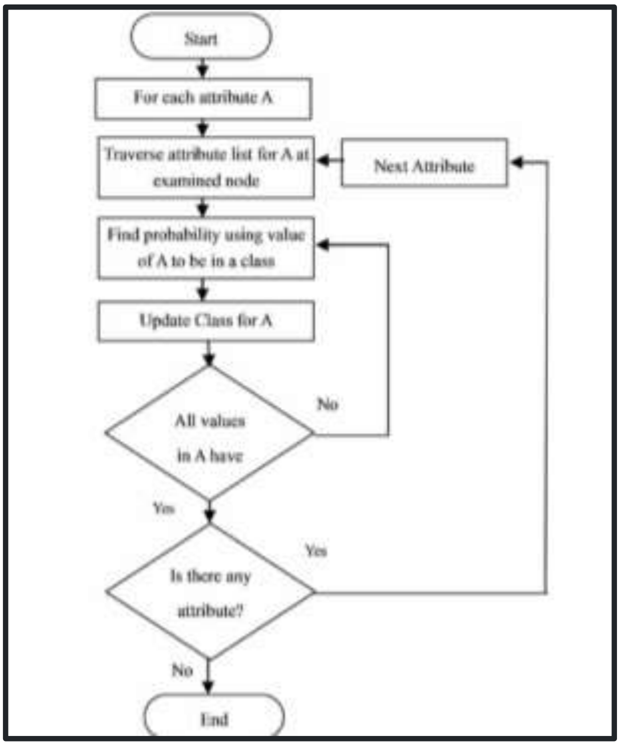


Fig.6. Naive Bayes

4. Logistic Regression (LR):

Logistic regression is a predictive algorithm designed for supervised learning. It is dichotomous in nature and can only be grouped into two classes. There are 2 classes i.e, failure or success (0 or 1) Probability (Y=1) is predicted mathematically With respect to X by a logistic regression model. It is one of the easiest ML algorithms, it is useful for a wide variety of classification problems. Prior to implementing LR we should be aware of the following assumptions –

1. Whenever binary LR is used, variables must always be binary and factor level 1 must represent the desired outcome.
2. Ideally, the independent variables should be independent of one another, so there should be no multi-collinearity in the model.
3. The variables we use in our model must be meaningful.
4. The sample size for logistic regression should be large.

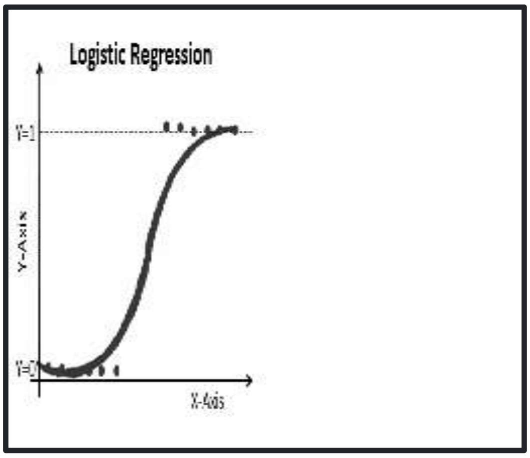


Fig.7. Logistic regression

5. Random Forest (RF):

Using a random forest, you can both perform classification and regression. A forest consists of trees, and a forest that has more trees will be more robust.

In a similar fashion, based on data samples, a random forest algorithm creates decision trees, then both obtains predictions and votes to determine the best outcome. As a result of its averaging, it reduces the effects of over-fitting more effectively than a single decision tree.

Implementation of Random Forest Algorithm:

Using the below mentioned steps, we can understand Random Forest's working −

1. Choosing arbitrary tests from a dataset is the primary step.
2. It then constructs a choice tree for each test. Each choice tree will at that point provide the prediction result.
3. During this step, every predicted result will be voted for.
4. In the fourth step, choose the prediction result that received the most votes.

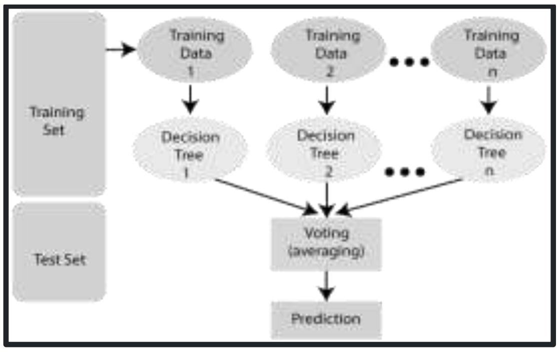


Fig.8. Random Forest

6. Accuracy:

Accuracy is the state of being correct or precise. Getting closer to a specific value is what accuracy is about. It is an execution assessment metric for classification that's broadly utilized in machine learning.

An accurate prediction of the positive class is considered as the true positive i.eTP.

True negatives are one of the components of a confusion matrix designed to illustrate how classification algorithms work.

False Positives are incorrectly predicted positive outcomes.

False negatives are the negative outcomes.

$Accuracy\ equals$:

Equation 5:

$$(TN + TP) / (FP + TP + FN + TN).$$

Precision is the ratio between True positives and all positives.

Equation 6: $Precision\ equals\ TP/(FP + TP)$

Recall is the True Positive rate and is calculated as

Equation 7: $TP/(FN + TP)$.

IV- CONCLUSION

Initially, data exploration was performed, followed by data redundancy check, data cleaning, data splitting, model building, and model evaluation, and then accuracy for different models was calculated and finally the best algorithm with higher accuracy was selected.

Results for Performance measures:

In this study, several supervised models were employed to predict rainfall in Australia and found that Different algorithms have varying accuracy. In the following table comparison between algorithms evaluated is shown.

Sr. No	Algorithms	Accuracy	Precision	Recall
1.	Random Forest	85.74%	0.77	0.51
2.	Logistic Regression	84%	0.73	0.48
3.	Naive Bayes	80%	0.55	0.59
4.	Support Vector Machine	84.2%	0.74	0.45
5.	Artificial Neural Network	85.20%	0.72	0.54

Fig.9. Comparison between algorithms

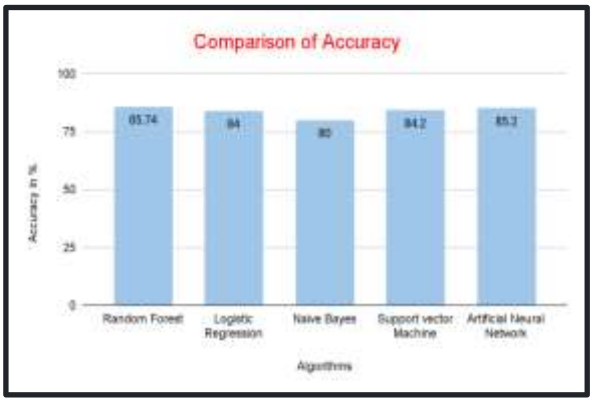


Fig.10. Graphical Representation of Comparison between algorithms

The first model we evaluated was Random Forest which was 85.74% accurate. Next was Logistic Regression which was found to be 84% accurate. After that Naïve bayes was evaluated and it was 80% accurate. Support Vector Machine and Artificial Neural network were found to be 84.2% and 85.2% accurate respectively. So, we deduced that Random Forest is the machine learning algorithm that Best Fits for rainfall prediction based on our dataset.

REFERENCES

[1] *Urmay Shah, Sanjay Garg, Neha Sisodiya, Nitant Dube, Shashikant Sharma (Nirma University, Ahmedabad, India). Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques at 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India.*

[2] *Mislan, Haviluddin, SigitHardwinarto, Sumaryono, Marlon Aipassa. Rainfall Monthly Prediction Based on Artificial Neural Network. International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).*

[3] *R.Senthil Kumar (Research Scholar), Dr.C.Ramesh (Research Supervisor), Department of Computer Science and Engineering Sathyabama University, Chennai (2016). A Study on Prediction of Rainfall Using Data Mining Technique.*

[4] *JinghaoNiu and Wei Zhang, School of Control Science and Engineering Shandong University 73 Jingshi Road Jinan, Shandong Province, China. Proceedings of the 2015 IEEE. International Conference on Information and Automation Lijiang, China, August (2015). Comparative Analysis of Statistical Models in Rainfall Prediction.*

[5] *Shilpa Manandhar, Soumyabrata Dev, Yee Hui Lee, Yu Song Meng and Stefan Winkler. IEEE Transactions On Geoscience And Remote Sensing (2019). A Data-Driven Approach for Accurate Rainfall Prediction.*

[6] *Moulana Mohammed, RoshithaKolapalli, Niharika Golla, Siva Sai Maturi. International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020. Prediction of Rainfall Using Machine Learning Techniques.*