International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

RAINFALL FORECASTING USING WEKA DATA MINING TOOL

Abinaya P L[1], Janani N[2]

[1,2]*Department of Information Technology, Kumaraguru College of Technology, Tamil Nadu, India*

ABSTRACT - *Rain water, additionally called precipitation, is a characteristic element of the world's climate framework. India is a rural nation which to a great extent relies upon monsoon for irrigation purpose. Rainstorm have coordinated effect on agriculturists' life and harvest creation. The Indian economy is heavily dependent on agriculture and the job of the Indian agriculturist to a great extent relies upon the Monsoon downpours. Rainfall forecasting is essential and important for development of the nation. Weather factors including temperature, dew point temperature, humidity and pressure have been used to forecasts the rainfall. The dataset was collected from Kaggle and is classified using Naïve Bayes, K nearest Neighbour, Decision tree and Support vector machine. The Naïve Bayes has given 80.56% accuracy, K nearest Neighbour provides 93.96% accuracy, Decision tree gives 94.10% accuracy while Support vector machine has given 93.66% accuracy.*

Keywords: Rainfall Prediction, Decision tree, Naive Bayes, K Nearest Neighbour, SVM

1. INTRODUCTION

Rainfall forecasting is essential since it facilitates to spot potential floods in future and to set up for higher water management which decides future irrigation desires. Weather forecasts may be categorised as: current forecasts that is forecasts up to few hours, Short term forecasts that is downfall forecasts is one to three days forecast, Forecasts for four to ten days are Medium vary forecasts and future forecasts are for over ten days. Short vary and medium vary downfall forecasts are necessary for flood forecasting and water resource management. Classification has been applied in several fields like detection frauds by banks and firms, by numerous service suppliers to predict their performance in future, to classify patients on the premise of their symptoms. In this paper we have used Naïve Bayes, K nearest Neighbour, Decision tree and Support vector machine to forecast rainfall. Classification trees are created using the concept of Gini Index, Bayes theorem is used in Naïve Bayes for prediction, Euclidean distance is used to compute Nearest neighbours and Support vector machine use Kernel trick.

2. NAÏVE BAYES

Naive Bayes is a basic, yet viable and regularly utilized, machine learning classifier. The Naive Bayes algorithm is a basic probabilistic classifier. Naive Bayes classifier requires a number of parameters linear in the number of features in a learning problem and it is highly scalable. Classification may be Gaussian, Kernel, multi nominal. Here Bayes theorem is used which assumes all attributes to be independent where the class variables are given. Assuming distribution for weather information, Bayesian classifiers use Bayes theorem to search out posterior chances of occurrence of input data instance in all classes. This assumption is known as class conditional independence. The algorithm starts with computing prior probability, $P(C_i)$ for each class. Then for an input data instance q, compute posterior probabilities for each class

$$P(q|C_i) = \prod_{j=1}^{n} P(Fj|Ci)$$

Where F_j are attributes for input sample q. Finally compute

$P(C_i|q) = P(q|C_i) * P(C_i) / \sum_{i=1}^{m} P(C_i)$

3. DECISION TREE ALGORITHM

A decision tree could be a structure that has a root node, branches, and leaf nodes. It is used for both classification and regression problem. Every internal node denotes a test on the associate attribute, every branch denotes the end result of a test, and every leaf node holds a category label. The upmost node within the tree is the root node. Tree pruning is performed so as to get rid of anomalies due to noise or outliers. The cropped trees are smaller and fewer complicated. The Decision tree algorithm produces correct and explicable models with comparatively very little user intervention. This algorithm is often used for each binary and multiclass classification issues. The algorithm is quick, each at build time and apply time. The build method for decision tree is parallelized. Decision tree rating is particularly quick. The tree structure, created within the model build, is employed for a series of easy tests. Every check relies on one predictor. The algorithm is non-parametric and can efficiently manage vast, muddled datasets without forcing a convoluted parametric structure. At the point when the example measure is sufficiently huge, study about data can be partitioned into training and validation datasets. Utilizing the training dataset to assemble a decision tree model and a validation dataset to settle on the suitable tree measure expected to accomplish the optimal final display. Decision trees are now recently referred as Classification and Regression Trees (CART). The best attribute for splitting and constructing decision tree is found by using Gini index as

a selection measure. In CART the impurities in the training dataset S is measured using Gini index.

Gini(S)=1-$\sum_{i=1}^{m} p_I$

Where m denotes the class labels. The one with least impurity is considered to be the best option. This process is done for all the attributes. The best splitting attribute is one which has maximum Δ $Gini(A)$

Δ $Gini(A)$ = Gini(S) - $Gini_A(S)$

Where,

$Gini_A(S)$ = $|S_1|$ / |S| Gini (S_1) + $|S_2|$ / |S| Gini (S_2)

4. K NEAREST NEIGHBOUR

K nearest neighbours are referred as lazy learners where learning depends on relationship. The algorithm can't be utilized until the example for which neighbours are to be computed is accessible. KNN are processed for given instance t for which class name is to be anticipated. The features are wanted to be numeric in nature as neighbours are processed utilizing the distance metric. Euclidean distance is performed here for analysing,

$$d(Y,Z) = \sqrt{\sum_{i=1}^{n}(y_i - z_i)^2}$$

Where Y and Z represents n dimensional observations from training and test dataset respectively for which class is to be predicted. The decision for the instance is taken by majority voting. The input sample is considered based on the class which has maximum neighbours. Based on the size of the training dataset the value of K that is the number of neighbours is determined.

5. SUPPORT VECTOR MACHINE

Support Vector Machine is a regulated machine learning calculation which can be utilized for both classification or regression difficulties. In any case, it is for the most part utilized in grouping issues. In this calculation, we plot every datum thing as a point in n-dimensional space where n represents the number of features, we have with the estimation of each component being the estimation of a specific facilitate. At that point, we perform classification by finding the hyper-plane that separate the two classes exceptionally well. Support Vectors are basically the co-ordinates of individual perception. Support Vector Machine is a border which best isolates the two classes that is hyper-plane/line. The least complex approach to isolate two gatherings of information is with a straight line, level plane or a N-dimensional hyperplane. In any case, there are circumstances where a nonlinear locale can isolate the gatherings all the more productively. SVM handles this by utilizing a kernel function (nonlinear) to map the information into an alternate space where a hyperplane can't be utilized to do the detachment. It implies a non-linear function is found out by a linear learning machine in a high-dimensional component space while the limit of the framework is controlled by a parameter that does not rely upon the dimensionality of the space. This is called kernel trick which implies the kernel function change the data into a higher dimensional element space to make it conceivable to play out the linear separation. In actuality, we probably won't have the capacity to drive a straight line between the classes which makes Support vector machines somewhat progressively confused however it's possible to characterize the most extreme edge hyperplane with Gaussian kernel.

6.PROCESS AND OUTCOMES

A dataset is collected which provides the data from May 2016 to March 2018 of the specific city that is Jaipur in India. The dataset consists of 679 instances and 13 attributes.

Running the algorithms utilizing Naïve Bayes we break down the classifier yield with such a large number of statistic-based yield by utilizing 10 cross validation to make a prediction of each example of the dataset. After running the algorithm, we accomplished a classification precision of 80.55% and the mean absolute error is 0.0446, time taken for building model is 0.01 seconds.

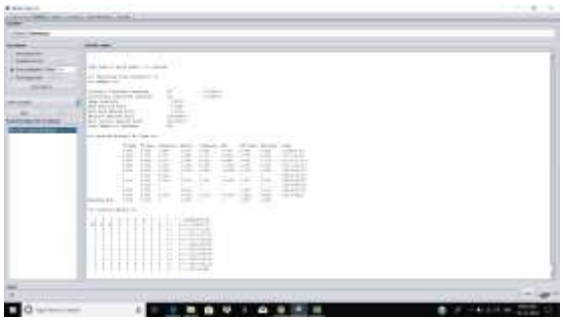

Fig.1.Screenshot view for Naïve Bayes Algorithm

J48 Tree has been utilized in this paper to choose the objective esteem dependent on different attributes of dataset and to characterize their accuracy. The output is based on 10 cross validation which produced an accuracy of 94.10% and the mean absolute error is 0.0227. The time taken to build the model is 0.01 seconds.

Fig.2.Screenshot view for J48 Algorithm

The K nearest neighbour is processed utilizing Euclidean distance. Here 10 cross validation is used to make

prediction for every instance in the dataset. The algorithm has shown an output of 93.96% of accuracy. The mean absolute error is 0.0181 and the time taken to build the model is 0 seconds.

Fig.3.Screenshot view for K Nearest Neighbour Algorithm

By running the Support vector machine algorithm, we broke down the classifier yield with various statistics dependent on yield by utilizing 10 cross validation. The accuracy obtained for this algorithm is 93.66% with a mean absolute error of 0.1598 and it took 0.69 seconds to build the model.

Fig.4.Screenshot view for Support Vector Machine Algorithm

7.CONCLUSION AND FUTURE WORK

The main aim of this paper is to forecast rainfall using WEKA data mining tool. Classification which is a part of data mining is extremely helpful in discovering obscure patterns like forecasting the future trends. Here in this paper we have used four algorithms namely Naïve Bayes, Decision tree, K Nearest Neighbour and Support Vector Machine and the outputs were compared based on the accuracy achieved. The result has shown that the decision tree algorithm has given the highest accuracy of 94.10% when compared to other three algorithms. So, we can conclude that the Decision tree algorithm is the best classification algorithm for forecasting rainfall on the basis of given features in the dataset.

REFERENCES:

1. Dhamodharan S, Liver Disease Prediction Using Bayesian Classification, Special Issues, 4th National Conference on Advance Computing, Application Technologies, May 2014

2. SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 58575860,2014.

3. Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.

4. David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.

5. Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

6.https://www.sciencedirect.com/science/article/pii/S0925231202005775

7.https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_decisiontree.htm#CACCJCEJ

8. Tina R. Patil, Mrs. S. S. Sherekar, Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal Of Computer Science And Applications, Apr 2013

9. Nicholas I.Sapankeyych, Ravi Sankar, Time series prediction using support vector machines: A survey, IEEE Computational Intelligence Magazine, April 2009

10.https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/