IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 3, Ver. II (May. - June. 2018), PP 32-38
www.iosrjournals.org

Climate Changes of Tamilnadu Based on Rainfall Data Using Data Mining Model Evaluation and Cross Validation

G. Manimannan[*], R. Lakshmi Priya[**], K. Jose Reena[***] and S. Krishna Priya[*]
[*] *Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai.*
[**] *Assistant Professor, Department of Statistics, Dr. Ambedkar Govt. Arts College, Chennai.*
[***] *Assistant Professor, Department of Computer Science, TMG College of Arts and Science, Chennai*
Corresponding Author: G. Manimannan

Abstract: *This research paper attempts to identify the climate changes based on rainfall data and to cross validate the changes of climate using various machine learning method. In recent and past years, climate changes due to various parameters like, pollution, real estate business, and shortages of lake and pond waters. The climate of Tamilnadu is generally tropical and features of fairly hot temperature over the year except monsoon seasons. Tamilnadu seasons and climate are classified into three categories viz, summer, winter and monsoon seasons. This research paper aims at analyzing rainfall model evaluation and cross validation to evaluate the district wise data of Tamilnadu and make possible the result in various seasons to understand the climate changes. The secondary database was collected from Indian Meteorological department during the year 2008 to 2012 and it is monthly rainfall data from various districts of Tamilnadu from January to December. The application of machine learning method of Logistic Regression, Support Vector Machine and Random Forest as data mining tools to explore the classification model and cross validate in the present dataset of the study period. All the three classification models were applied first and extracted. Area under the Curve, Classification Accuracy, F1 Score, Precision and Recall are all closer to unity. All the above measures show that five major categories of climate based rainfall. Finally, machine learning methods achieved best model and they are labeled as Very High Rainfall (VHR), High Rainfall (HR), Moderate Rainfall (MR), low Rainfall (LR) and Very Low Rainfall (VLR). The results of the present study indicate that the machine Learning Data Mining Tools can be used as a feasible tool for the analysis of large set of rainfall data. Finally, the five model classification is visualized using Silhouette plot.*

Keywords: *Rainfall, Machine Learning Methods, Logistic Regression, Random Forest Method, Support Vector Machine, Euclidean distance and Silhouette plot*

Date of Submission: 20-05-2018 Date of acceptance: 05-06-2018

I. Introduction

Meteorological data mining is a form of data mining concerned with finding hidden patterns inside the largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Weather is one of the meteorological data that is rich in knowledge. Weather Forecasting [1] is vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. Accurate prediction is one of the major challenges faced by meteorologist all over the world.

Air temperature Humidity, Rainfall is important property of the urban climate that has implications in areas related to human reassure and health. They are essential components of a comfortable environment. The aim of this study is to understand the relationship between air temperature and its moisture holding capacity and thus its effect on Relative Humidity[1]. From the study it has been statistically proved that the moisture holding capacity of air depends on the air's temperature. It increases with increase in temperature. Atmospheric dispersion models are employed for prediction of air quality, under different terrain and meteorological conditions.

In fact, geometrically, Tamil Nadu touches the acute southern tip of the Indian Peninsula. The climate of Tamilnadu is generally wet subtropical climate and features fairly hot temperature over the year except during the monsoon season. The state has three distinct monsoon periods of rainfall. The south west monsoon starts from the period of June to September with strong southwest winds. The north east monsoon starts from the period of October to December with dominant northeast winds. Finally, dry season starts from January to May. The normal annual rainfall of the state is about 945 mm (37.2 in) of which 48% is through the North East monsoon, and 32% through the South West monsoon. Since the state is fully dependent on rains for recharging its water resources, monsoon failures lead to acute water scarcity and severe drought. Moreover, factors like

climate change and urbanization have also had an impact on the variation in rainfall. Recent studies have stated that any analysis of hydro-climatic variables should be done at the local scale rather than at a larger or global scale [2].

II. Review of Literature

Climate changes due to various factors, like pollution, plastic bags, corporate and government occupies cultivated lands and failure of monsoon for the past years. In this connection the rainfall is a key factor for determining the sustainability and conservation of living species on the earth. In dry farming areas, where rainfall is the sources of water for crops, changes in both quantity and distribution of rainfall during the year could affect the economy of an area. Many researchers have applied MPL (Multi variables Polynomial regression) to implement the precipitation forecast model over Myanmar.

Nikhil Sethi [3]) discussed an artificial neural network based model with wavelet decomposition for prediction of monthly rainfall on account of the preceding events of rainfall data [4]. Wavelets transform an extraction of approximate and detail coefficient of the rainfall data series. Manimannan *et. al.* [5] developed a data mining model using classification and Geographical Information System (GIS) for annual rainfall data distribution of Tamilnadu. The results of the models are classified and visualized high, moderate and low categories rainfall in the districts of Tamilnadu.

All these research studies are attempted and achieved some pattern of rainfall and climate changes in different location of the world. The main objective of the present study is identifying the structural data mining model, classification, cross validation and visualization of all the season of rainfall data in Tamilnadu using different data mining techniques: (a) To develop data mining model and identify the pattern of rainfall data in the study period using Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) methods. (b) To identify the final test score, classification and cross validation of rainfall data using the above methods. (c) Finally, to visualize the rainfall data based on extracted test score using Silhouette plot.

III. Database

The district wise monthly rainfall database was recorded in various districts of Tamilnadu. The secondary sources of rainfall data were collected from Meteorological Department, India, during the period of 2008 to 2012 was considered as database. Three seasonal rainfall data were chosen for the present research that had been used in previous studies. The present research work analyses the rainfall information of all the monsoon, winter and summer seasons of Tamilnadu districts. The summer seasons begin from March to May, winter seasons starts in the month of January and February, Northeast monsoon from October, November and December, and Southwest monsoon from June to September.

IV. Methodology

Data Mining Techniques is an interdisciplinary field, the confluence of a set of disciplines in the following heads including database systems, Statistics, machine learning, visualization and information science. Although data mining is a new term, the technology is not. Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. In the present context data mining exhibits the patterns and cross validation by applying few techniques namely, logistic regression, random forest method and support vector machine rule. As such KDD is an iterative process, which mainly consist of the following steps; Step 1: Data cleaning; Step 2: Data Integration; Step 3: Data selection and transformation; Step 4: Data Mining and Step 5: Knowledge representation. Of the above iterative process Steps 4 and 5 are most important. If clever techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

In this research paper, the researcher uses orange data mining software. Orange is an open source machine learning and data visualization for learner as well as experts. Interactive data analysis work flows with a large toolbox is available in this package. The software is developed with python script. Python is an interpreted high-level programming language for general-purpose programming and it was created by Guido van Rossum [9].

4.1 Logistic Regression

Regression analysis always requires numeric data, when attributes are categorical, the researcher have to change to numerical values to apply regression analysis [6]. Regression and classification are data mining techniques used to solve similar problems, but they are frequently confused. Both are used in prediction analysis, but regression is used to predict a numeric or continuous data while classification assigns data into discrete data. Classification is a data mining technique that assigns categories to a collection of data in order to

aid in more accurate predictions and analysis. Classification is one of the several methods intended to make analysis of very large datasets effective.

The general model for logistic regression is $p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \Rightarrow \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$. The logistic distribution constrains the estimated probabilities to lie between 0 and 1.

4.2 Random Forest (RF)

Random forest is a group learning method used for classification, regression and other tasks. It was first proposed by Tin Kam Ho and further developed by Leo Breiman [7] and Adele Cutler. Random Forest builds a set of decision trees. Each tree is developed from a bootstrap sample from the training data. When developing individual trees, an arbitrary subset of attributes is drawn (hence the term "Random"), from which the best attribute for the split is selected. The final model is based on the majority vote from individually developed trees in the forest. Random Forest works for both classification and regression tasks.

4.3 Support Vector Machine (SVM)

Support vector machine (SVM) is a machine learning technique that separates the attribute space with a hyper plane, thus maximizing the margin between the instances of different classes or class values. The technique often yields supreme predictive performance results. Orange embeds a popular implementation of SVM from the LIBSVM package. Support Vector Machine (SVM) is primarily a classifier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1.

4.4 Cross Validation

The data set is split into training and test cases randomly. The centroids of the clusters from the training cases can be used to cluster the test cases. The centroids of the clusters formed by test data are then computed and compared with the training data. Comparable results validate the clustering that has achieved.

4.6 Proposed Algorithm for Cross Validation with various Data Mining models work Flow

Step 1: Initially, the district wise monthly rainfall database of Tamilnadu State from the year 2008 to 2012 given through the file widget and connect through the work flow with Test score widget with Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM).

Step 2: The Test score widget assign the training and testing data sets with 10 folds cross validation. The training data sets are 70 percent of the original database using random sampling method up to target class.

Step 3: The data table widget connected to file widget for checking the original data.

Step 4: Repeat the train or test database to reach best model or repeat from step 1, change the training data and cross validation folds. Figure 1 represents the data mining work flow for various data mining Techniques, Confusion Matrix and Visualization of the database.

4.7 Logistic Regression

Step 1: The logistic regression widget chooses from various machine learning method and connects to test score widget.

Step 2: Open logistic regression widget and select regularization type (Ridge L2 by default and widely used model).

Step 3: The logistic regression strength must always $C = 1$ is in the middle of the model and the remaining two extreme points of left and right side of the line are labeled as weak and strong strength.

Step 4: Repeat the step 3 with various folds and C value, to get the better model.

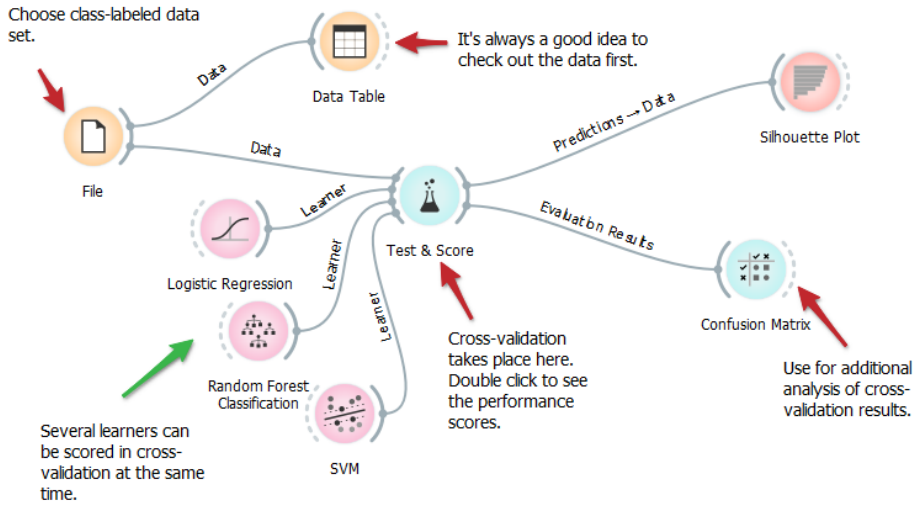


Figure 1. Work Flow diagram Models of classification, Cross Validation and Machine Learning Methods

4.8 Random Forest classification

Step 1: The Random Forest widget chooses from various machine learning method and will be connected to test score widget.

Step 2: Open Random Forest widget and select growth control model and number of trees are 10 from basic properties with number of attributes considered at each split at 5.

Step 3: Repeat the step 3 with various size of trees and splits to get a better model. .

4.9 Support Vector Machine

Step 1: The Support Vector Machine widget chooses from various machine learning method and will be connected to test score widget.

Step 2: Open Support Vector Machine (SVM) type and select SVM cost type $C = 1{,}00$, Regression loss epsilon $(\varepsilon) = 0{,}10$ and choose the menu of optimization parameter set to 0.0010 with iteration limit 100

Step 3: Repeat the step 2 with various cost, epsilon, optimization parameter and iteration limit, to get better model of SVM.

V. Result and Discussion

5.1 Test and Scoring methods

In general, the test and scoring methods are AUC, CA (Classification Accuracy), F1, Precision and Recall measures are displayed in the output window. The definition of each measure is given below.

5.2 Area under the Curve (AUC)

The Area under the curve is a performance metrics for a binary classification in data mining. By comparing the Receiver Operating Characteristic Curve with the area under the curve, or Area Under the Curve (AUC), it captures the extent to which the curve is up in the Northwest corner. The AUC score less than 0.5 is not a better random estimate. The measure of AUC with 0.9 would be a very good model, but a score of 0.999 would be the best model to be true and indicate correct model (Table 1).

5.3 Classification Accuracy (CA)

The classification accuracy with 1 is the model which is the best and 0 is the model as worst (Table 1), the following formula is used to calculate the CA measure based on Type I and Type II errors of statistics,

$$CA = \frac{true\ \ positive\ + true\ \ negative}{true\ \ positive\ + true\ \ negative\ + false\ pos\ \ itive\ + false\ nega\ \ tive} = \frac{tp + tn}{tp + tn + fp + fn}$$

5.4 F1 Score

In statistical study of binary classification, the F_1 score, alternatively named as (F -score or F -measure) a measure of a test's precision. The F_1 score considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier. The r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F_1 score is the harmonic mean of the precision and recall, where a F_1 score reaches its best model at 1 (that is perfect Precision and Recall score) and worst model at 0 (Table 1). The general formula for F_1 score is the harmonic average of p and r

$$F_1 = 2 * \frac{1}{\frac{1}{r} + \frac{1}{p}} = 2 * \frac{p * r}{p + r}$$

5.5 Precision of Test Score

The precision measure is the ratio of $P = \frac{true\ \ positive}{true\ \ positive\ + false\ \ positive} = \frac{tp}{tp + fp}$ where tp is the number of true positives and fp is the number of false positives. The precision is naturally the ability of classifier not to label as positive a sample that is negative. The precision value 1 is the best model and 0 is the worst model (Table 1)..

5.6 Recall of test Score

Recall measure is the ratio of $R = \frac{true\ \ positive}{true\ \ positive\ + false\ \ negative} = \frac{tp}{tp + fn}$ where tp is the number of true positives and fn is the number of false negatives [8]. The precision is naturally the ability of classifier not to label as positive a sample that is negative. The precision value is 1 (Table 1). .

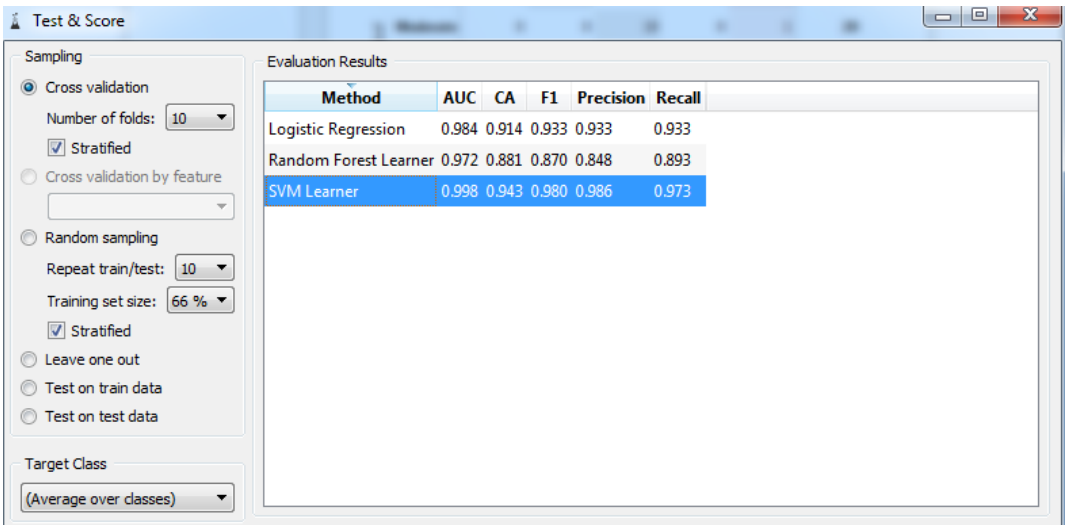


Table 1. Test Score Machine Learning Methods

The above table shows the test and their scores of various data mining techniques. All the three methods of AUC, CA, F1, Precision and Recall values are closer to unity. The results achieved best model and cross validation of rainfall database. The classification and confusion matrix of three models are classified as 95 percent and above and the remaining five percent are misclassified (*Table 2 to 5*) due to climate change and delayed monsoons. Based on this classification accuracy the rainfall database were classified and labeled as Very High, High, Moderate, Low and Very Low categories.

Table 2. Confusion Matrix for Logistic Regression

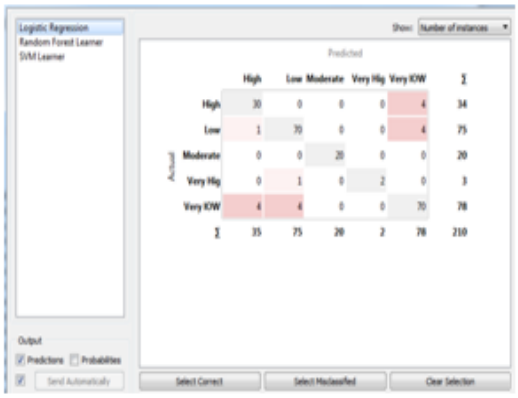


Table 3. Confusion Matrix for Support Vector Machine

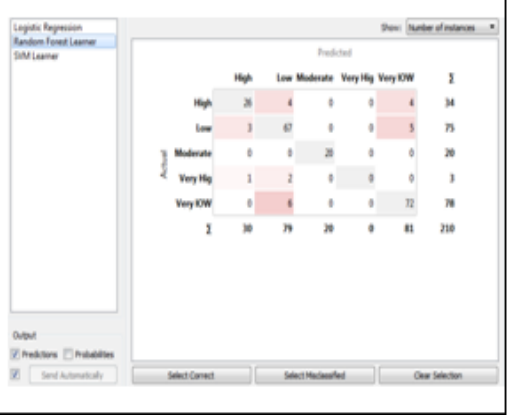


Table 4. Confusion Matrix for Random Forest Method

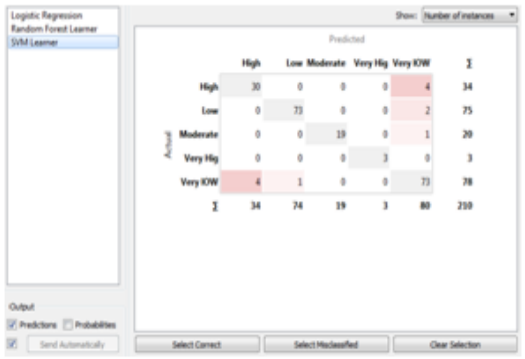


Figure 2. Silhouette plot for Logistic Regression

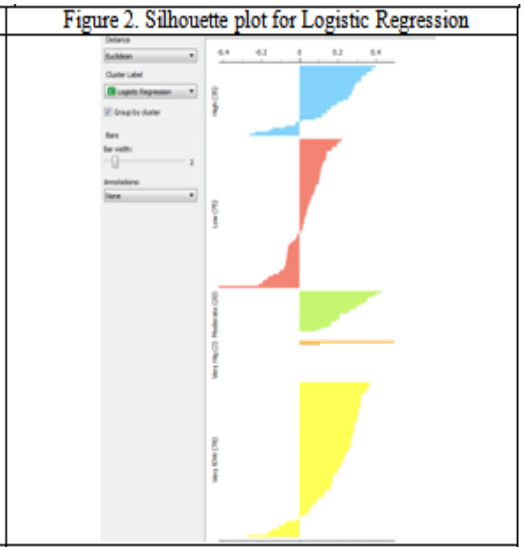
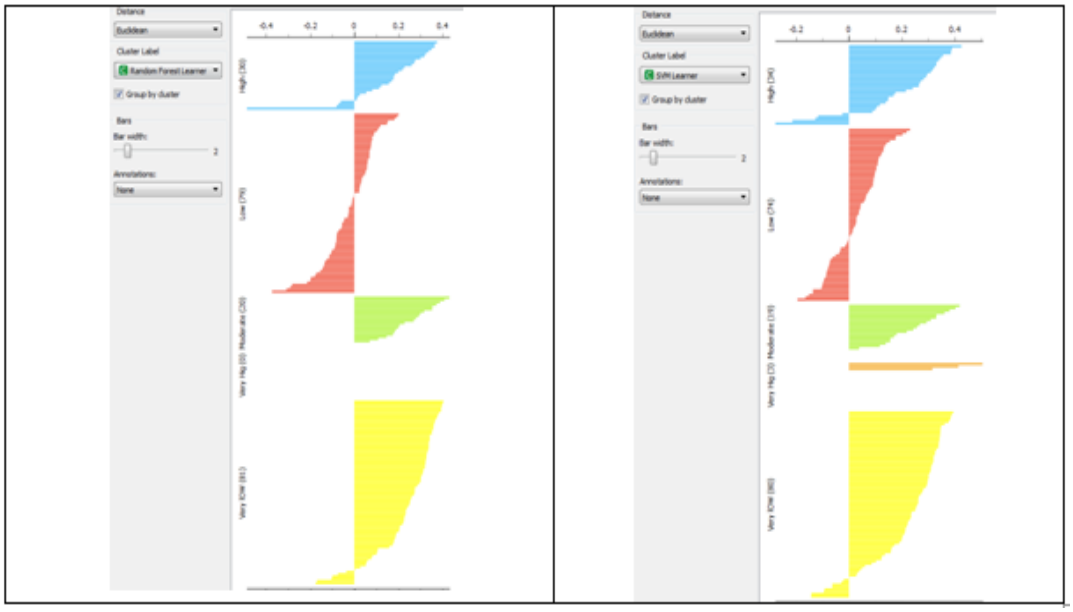


Figure 3. Silhouette plot for Support Vector Machine

Figure 4. Silhouette plot for Random Forest Method



The silhouette plot visualize (Figure 2 to 4) five categories of rainfall data during the study period from 2008 to 2012 and all the methods has less than five percent misclassification due to climate changes during the study period. After performing data mining techniques of logistic regression, random forest and support vector machine, the next stage is to assign initial classification based on Tamilnadu climate. Formations of classification are explored by considering 2- classification, 3- classification 4- classification and so on. Out of all the possible splits, 5- classification exhibited meaningful interpretation than two, four and higher splits. Having decided to consider only 5 classifications, it is possible to group the rainfall as Very High, High, Moderate, Low and Very Low classification depending on whether the rainfall belonged to classification 1, classification 2, classification 3, classification 4 and classification 5 respectively. Classification 1 and 2 is a group of rainfall that has excess rainfall of the study period. The years with poor rainfall are grouped into classification 4 and 5. Classification 2 is those years which received normal rainfall when compared with Classification (1, 2) and (4, 5). In spite of incorporating the results of data mining for the study period, only the summary statistics are reported in Table 2 to 5. All the data mining methods attained best model test and score.

VI. Conclusion

This research paper attempts to identify a meaningful classification, Cross validation and model evaluation of the district wise data of Tamilnadu and make possible the result in various seasons to understand the climate changes. The secondary database was collected from Indian Meteorological department during the year 2008 to 2012 which is a monthly wise rainfall data from various districts of Tamilnadu from January to December. The application of machine learning method of Logistic Regression, Support Vector Machine and Random Forest as data mining tools to explore the classification model and cross validate the present dataset of the study period. All the three classification models were applied first and extracted best model. Area under the Curve, Classification Accuracy, F1 Score, Precision and Recall are all closer to unity. All the above measures show that five major categories of climate based on rainfall. Finally, machine learning methods achieved best model test and score and they are labeled as Very High Rainfall (VHR), High Rainfall (HR), Moderate Rainfall (MR), Low Rainfall (LR) and Very Low Rainfall (VLR). The results of the present study indicate that the machine Learning Data Mining Tools can be used as a feasible tool for the analysis of large set of rainfall data. Finally, the five model classification is visualized using Silhouette plot. The negative portions of silhouette plot support the climate changes during the study period. The scope and future study is to identify the prediction and classification of rainfall data year wise and district wise with the help of other data mining techniques.

References

[1]. Olaiya, Folorunsho, and Adesesan Barnabas Adeyemo. "Application of data mining techniques in weather prediction and climate change studies."International Journal of Information Engineering and Electronic Business (IJIEEB) 4.1 (2012): 51.

[2]. Sharma RH, Shakya NM. 2006. Hydrological changes and its impact on water resources of Bagmati watershed, Nepal. Journal of Hydrology 327 (3-4): 315-322.

[3]. Nikhil Sethi, Dr.Kanwal Garg "Exploiting Data Mining Technique for Rainfall prediction" , International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3982-3984.

[4]. Pasanen, A-L., et al. "Laboratory studies on the relationship between fungal growth and atmospheric temperature and humidity." Environment International17.4 (1991): 225-228. Swinbank, W. CQJR. "Long-wave radiation from clear skies." Quarterly Journal of the Royal Meteorological Society 89.381 (1963): 339-348.

[5]. Manimannan G. and Lakshmi Priya R (2001), Rainfall Fluctuation and Classification over Tamilnadu Region: Using Data Mining Techniques, IOSR Journal of Mathematics (IOSR-JM) e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 10, Issue 5 Ver. IV (Sep-Oct. 2014), PP 05-12.

[6]. A.B.M. Shawahat Ali and Saleh A. Wasimi (2009), Data Mining: Methods and Techniques, Cenage Learning India Private Limitted, New Delhi.

[7]. Leo Breiman (2001), Random Forest, Statistics Department, University of California, Berkeley, CA 9472, USA, 1-33.

[8]. David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63.

[9]. Guido van Rossum (1991), Interpreted high level programming language, Python Software Foundation. .

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

* G. Manimannan. " Climate Changes of Tamilnadu Based on Rainfall Data Using Data Mining Model Evaluation And Cross Validation." IOSR Journal of Computer Engineering (IOSR-JCE) 20.3 (2018): 32-38.