Smart Agricultural Practices using Machine Learning techniques For Rainfall Prediction: A case Study of Valkenburg station, Netherlands.

Dr. Razeef Mohd[1] Lone Faisal[2] Madiha zahoor[3] Dr. Juneed Iqbal[4]

[1]AI&ML Expert,Sher-e-Kashmir University of Agricultural Sciences and Technology Kashmir, (SKUAST-K)Shalimar, JK, India

[2] Assistant Professor, Department of Electrical Engineering, Mewar University Rajasthan India,

[3] Assistant Professor (C), Islamia College of Science and Commerce, Srinagar, J&K, India

[4]Assistant Professor (C), Islamia College of Science and Commerce, Srinagar, J&K, India

Article Info
Page Number: 8451-8462
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022

Abstract
Forecasting weather in general and rainwater or downpour in particular is significant for cropping-plan decision-making, water resource management which helps to determine impending irrigation potentials and other important aspects that help the farmer to have better and quality yield. Meteorological conditions and the precise prediction of weather patterns are important for agriculture and allied sectors. Predicting weather conditions like rainfall which is a type of weather pattern that is influenced by various weather parameters like wind direction, wind speed, humidity, geographical location, temperature etc. and have inherent connection with the agricultural activities. Adoption of disruptive technologies like artificial intelligence & machine learning for rainfall predictive analytics have achieved better results with appreciable performance and accuracy in predicting rainfall as compared to the traditional statistical methods. In this work, we have implemented Naive Bayes technique for the rainfall prediction. The historical weather data is collected from Royal Netherlands Meteorological Institute (KNMI) which is available on http://www.sciamachy-validation.org/climatology/daily_data/selection.cgi.Out of 39 available weather attributes, 5 most relevant attributes are selected for rainfall prediction using genetic algorithm and are more relevant for better rainfall prediction. The weather data set of 7670 daily weather instances of Valkenburg station from 1990 to 2010 is used to build the model using Naïve Bayes approach and its accuracy is tested on a test data set having 1826 daily weather samples from 2011 to 2105. The experimental results using Naïve Bayes algorithm show 71.2% accuracy rate for rainfall prediction which is appreciable and significant for predictive analytics.
Keywords: Smart Agricultural, disruptive technologies, artificial intelligence, machine learning, Climate change, weather forecasting, rainfall prediction, , Naive Bayes Classifier, Bayes theorem.

1. Introduction

In today's world of information technology and communication (ITC), weather prediction has become the most demanding and important task which aids in predicting the climatic condition of a location with reasonable amount of certainty. Forecasting precise weather

patterns has good significance in the field of meteorology and its allied sectors that influence directly or indirectly on agriculture field [1]. The rapid evolution and advancement in science and technology enable scientists to make better and precise weather prognosis. Next-gen computing techniques and technologies are used by the scientists and researchers to make more accurate weather predictions. Automated surface-observing systems, super computers, Doppler radar, wired and wireless sensors (Radiosondes), meteorological satellites and are some of the tools used to collect the weather data for weather forecasting [2]. Better climate prognosis helps to be alert and ready for any climatic threats like flash- floods, famine, flood etc. Data is the most indispensible unit for any predictive analytics process. Prognostic analytics is the use of data, statistical techniques and machine learning algorithms to identify the likelihood of future outcomes based on historical data [3]. The objective is to go beyond knowing what has happened in past to providing a best assessment of what will happen in the future. An enormous volume of weather data is available which is rich in information and can be used to predict weather conditions by implementing artificial intelligence and machine learning techniques to forecast atmospheric patters like temperature, wind-speed, rainfall, floods etc. [5]. Some commonly used Deep Learning and Machine Learning techniques for weather prognosis are Decision Trees, Artificial Neural Networks (ANN), Naive Bayes Networks, Support Vector Machines (SVM), Fuzzy Logic, Rule-based Techniques which includes Memory based reasoning Techniques and Genetic Algorithms [5].

Climate change has led to global warming and ozone depletion which will have harmful effects on humans and environment across the globe including agriculture sector. Crop quality and production will also reduce significantly with climate change. As mercury rises and the air becomes warmer, more water content will evaporate from earth and water bodies into the atmosphere which will lead to more rainfall and snow. In some areas the global warming will cause decreasing runoff and stream-flow causing an overall increase in famine hazard and severity [6]. Rainfall is important for crop production, water resource management, humidifying the atmosphere, producing streams and rivers, replenishing the water table and redistribution of fresh water in the water cycle [7]. The occurrence of prolonged dry period or heavy rain at the critical stages of the crop growth and development may lead to significant reduction in crop yield [8]. Predicting precipitation is very essential because incessant and irregular rainfall can have many effects that cause destruction of crops and farms, livestock, orchards etc. and can also decimate residential and commercial infrastructure, so an adaptive predictive model is needed for timely warning that can decrease risks to life and property and also help in giving advisory to farmers time to time in a better way for smart agricultural practices [9]. Prior information about climate helps farmers to know when to apply the pesticides and other chemicals to evade the crop wastage and increase the production [10].

2. Related work

Machine Learning and Artificial Intelligence techniques have significantly revolutionized weather forecasting system. New weather mobile applications that work on underlying disruptive technologies issue alerts and information regarding weather updates with good

prediction accuracy. In this section we will review briefly some of the existing rainfall prediction methods based on machine learning and artificial neural networks.

Liu et al. [11], proposed rainfall prediction model using improved Naïve Bayes Classifier. The authors used genetic algorithms (GAs) for feature selection and Naïve Bayes Algorithm for prediction process. The weather data of Hong Kong was collected and pre-processed, the experimental results showed that the proposed prediction model has better accuracy rate of 90% with compared with other classifiers. Furthermore, the results also exhibited that improved Naïve Bayes Classifier has considerable prognostic ability to forecast rainfall range levels with 65%-70% of correctness.

Yim, S.Y et al. [12], build a model for rainfall prediction constructed on a physical–empirical model. For this work, the weather dataset was collected from Taiwan Meiyu. The result shows that the proposed prediction model has appreciable accuracy that can forecast the quantity of Meiyu precipitation which supports in water resource management and utilization.

Nikam, V.B. and Meshram, B.B [13], proposed prediction data-intensive model based on Bayesian method of machine learning technique with appreciable prediction accuracy. The model offers enhanced precision and uses reasonable computing resources for forecasting rainfall. The exact prediction model is achieved by working on the issues in the hybrid model of multiple data mining approaches or even uniting compute based models with the data driven models. The historical weather data is collected from Indian Meteorological Department (IMD) Pune which consists of 36 attributes out of which only 7 attributes which are most relevant were considered for prediction of rainfall. The experimental results showed that proposed prediction model using Bayesian method of machine learning shows good accuracy figures with moderate compute resources utilization to predict rainfall.

F Nhita et al., [14] developed Evolving Neural Network (ENN) method which ensembles Artificial Neural Network (ANN) and Genetic Algorithm to optimize and tune the neural network in finding the best weights and biases. The method consists of one hidden layer in ANN which illustrated enough for providing enhanced performance for various datasets. Weather data for this research work was collected from the Indonesian Agency for Meteorology Climatology and Geophysics (BMKG). The data comprised of air temperature, rainfall, air humidity, length of sun radiation, airpressure, and wind speed in Kemayoran Jakarta within five years from 2007 to 2012. The proposed model provided prediction with good accuracy rate.

Marjanović, M et al., [15] proposed a prediction model for evaluating rainfall-induced massive land sliding in Western Serbia from year 2001 to 2014. In May 2014, there where floods and land sliding that happened due to heavy rainfall in Serbia, Bosnia and Herzegovina that had shocking effects, such as human losses, and devastation of the natural and metropolitan cities. The authors implemented Decision Tree algorithm to classify rainfall situations that caused landslides in the definite period.

Rivero, C.R [16] designed a technique for modeling the short rainfall time-series based on Bayesian enhanced modified combined approach (BEMCA) which uses permutation and relative entropy using Bayesian inference. The ANN disagrees according to diverse data models selected and can be pooled with entropic information for the time series. The experimental observations and results found after the computation are accessed on time series and rainfall series. This technique produces complex problems and lacks ensemble forecasting methods. BEMCA filter demonstrations an rise of accuracy in 3-6 prediction horizon examining the dynamic behavior of chaotic series for short series predictions.

3. Data collection and preprocessing

Following data mining process steps have been applied to pre-process and clean the collected raw weather data set as shown in figure 1. Understanding how the data is collected, stored, transformed, reported, and used is vital for the data mining process [17].

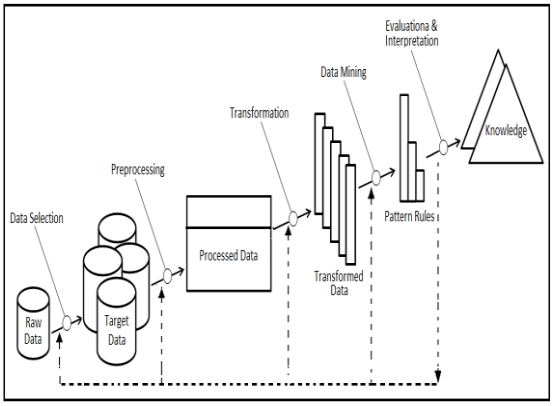

Figure 1: Data Mining Process Model

a) Study area and Data collection

Historical weather data is collected from Royal Netherlands Meteorological Institute (KNMI) Ministry of Infrastructure and Environment and is available on http://www.sciamachy-validation.org/climatology/daily_data/selection.cgi [18]. The complete description of area of study is given in table 1 along with the location and station number. Out of 50 available stations we have chosen only one for this research work.

Credentials of location	**Values**
Study Area	Valkenburg
Station Number	210
LON(east)	4.430
LAT(north)	52.171
ALT(m)	-0.20

Table 1: Location Description

Weather dataset of 20 years from 01-01-1990 to 31-12-2010 is collected for training the model and weather dataset from 01-01-2011 to 31-12-2015 is used for testing the model. The website KNMI provides 39 measured attributes, out of which we have used five most relevant and appropriate weather attributes/features are selected using the Genetic Algorithm (GA) for feature selection[19]. The attributes that we have chosen include temperature in 0c, humidity in %age, sea level pressure hPa, windy in Km/h and rainfall as shown in table 2. Less relevant features are left in the dataset for better model computation and prediction.

Attribute	**Type**	**Description**
Temperatur e	Numerical	Temp is in deg. C
Humidity	Numerical	Humidity in Percentage
Sea Level Pressure	Numerical	Sea Level Pressure in hpa
Windy	Numerical	Wind Speed in Kmph
Rainfall	Numerical	Rainfall in mm

Table 2: Weather data description

b) Data Preprocessing and Data Cleaning

The main challenge in weather prediction is the poor data quality and selection. For this reason we try to preprocess data carefully to obtain accurate and correct prediction results. In this phase unwanted data or noise is removed from the collected data set which is done by removing the unwanted attributes and keeping the most relevant attributes that help in better prediction [20]. Another major issue that is to be rectified is the missing values in the collected dataset. Missing values in the data set is filled by using various techniques. In this work, the missing values for attributes in the dataset are replaced with the modes and means based on existing data. Adding the missing values provides a more complete dataset for the classifiers to be trained on [21].

c) Data transformation.

In this phase the data set is transformed and converted into appropriate forms that helps us in better data mining and prediction. Smoothing, Aggregation, Generalization and Normalization are the various techniques that are implemented in this phase. In our collected data set, Rainfall attribute is selected as class variable or target class. Being a binary classification problem, we converted the rainfall data, i.e. numerical values for both data sets in inches to two values (YES/NO); 'YES' for Rainfall and 'NO' for No-Rainfall.

d) Data Discretization

Data discretization also known as binning is the process of converting numerical (continuous) data variables into categorical (discrete) counterparts [22]. Binning is usually done in modeling methods or in machine learning algorithms like Naive Bayes, Decision tree and so on. Predictive accuracy of the models can be improved by binning methods that reduces noise or non-linearity. Since our attribute values are in numerical form; however our model requires categorical values for computation. So we have used equal width binning (discretization) method for converting our numerical data variable into categorical counterparts [23].

e) Data Mining

Data mining is the process of extracting the useful information from a large collection of data which has been previously unknown [24]. For extracting useful information we need to follow data mining process model that will give us clean valuable dataset for model computation and better prediction. Very rarely data are available in the form required by the data mining algorithms. Most of the data mining algorithms would require data to be structured in a tabular format with records in rows and attributes in columns. The methodological discovery of useful relationships and patterns in data is enabled by a set of iterative activities known as data mining process. Not all discovered patterns leads to knowledge. It is up to the practitioner to invalidate the irrelevant patterns and identify meaningful information [25].

4. Experimental Study and Results:

Naive Bayes classifier is probabilistic classifier based on Bayes' theorem with an assumption of independence among the predictors. The Naive Bayesian classifier was first described in [26] in 1973 and then in [27] in 1992. It shows the independence assumption among all features in a data instance. A Naive Bayesian Model is easy to build as no complicated parameter estimation is needed which makes it useful for large data sets. Naive Bayes often performs better in terms of prediction than sophisticated classification algorithms. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables.

In the model building process (training), the Naïve Bayes algorithm finds relationships between the values of the predictors and the values of the target. These relationships are summarized in a model called prediction model which is then applied to the test weather data set in which the class assignments are unknown. Naïve Bayes algorithm is a supervised machine learning algorithm which accurately predicts the target class for each case in the data with good accuracy rate [28]. The collected and pre-processed historical data set in typically divided into two data sets; one for training the model known as training dataset and other for testing the model known as testing dataset.

In this research work, the model trained by entering the weather data from 01-01-1990 to 31-12-2010, which consists of 7670 daily weather data instances with rainfall data values as YES/NO. For testing phase, we have used weather data set from 01-01-2011 to 31-12-2015

which has 1826 weather data instances and contains all the weather attribute values except the rainfall data which the prediction model is supposed to predict. In both weather data sets, the attributes and their types remained the same, as listed in the Table 2. We deliberately deleted the data of rainfall attribute column in test weather data set, to enable the model to predict the rainfall so that we can validate the prediction accuracy of the proposed model. The actual performance of the proposed machine learning classification prediction model can be evaluated by comparing actual rainfall (target variable) data set values with the predicted rainfall values. First, we developed a model using the pre-processed and cleaned training weather data set, and then this trained model is provided with test weather data set to predict the rainfall.

We have used WEKA (Waikato Environment for Knowledge Analysis), a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand [29]. The Naïve Bayes algorithm when applied to the weather data set produces a classifier (model). The model produces 5795 correctly classified instances which are 75.55% and 1875 incorrectly classified instances which are 24.45%. The proposed rainfall prediction model is used to predict the future events when applied on test weather data set. The prediction accuracy of the proposed rainfall prediction model is evaluated by comparing the actual rainfall data with the predicted data and is found to be 71.2%.

a) Performance Measure and Model Evaluation l

Performance evaluation of a model (classifier) is an integral part of model development process. Model evaluation helps to find the better model for our data and also reveals how well the chosen model will perform in future. In order to evaluate the performance of developed model, a number of performance measure are used which are mainly based on confusion matrix [30].

b) Confusion Matrix

Confusion matrix is the matrix visualization of outcome of machine learning model. A basic confusion matrix is traditionally arranged as a 2 x 2 matrix for binary classification problem which contains information about actual and predicted classifications done by a classification algorithm as shown in table 3. Performance of such systems is commonly evaluated using the data in the matrix [31].

Predicted by Model	**Actual Labels**		
		YES	**NO**
	YES	True-Positive**(TP)**	False-Positive**(FP)**
	NO	False-Negative **(FN)**	True-Negative**(TN)**

Table 3: A sample confusion matrix

Table 4 shows the confusion matrix that is produced after applying Naive Bayes algorithm to weather data set.

Predicted by Model	**Actual Labels**		
		YES	**NO**
	YES	3418(TP)	1134(FP)
	NO	741 (FN)	2377 (TN)

Table 4: Confusion Matrix of weather data set

There are many performance measures for machine learning classification algorithms, we have implemented following performance measures: Accuracy, Precision, Recall, F-measure, Receiver Operating Characteristic (ROC), root mean square error (RMSE), and mean absolute error (MAE) [32]. These performance measures are based on the values of confusion matrix as shown in table 5.

S.No	**Performance Measures**	**Values**
1	Sensitivity or Recall	0.756
2	False-Positive Rate (FPR)	0.257
3	Precision	0.756
4	F-measure	0.754
5	True Negative rate/Specificity	0.677
6	Overall Accuracy	0.80
7	ROC	0.835

Table 5: Performance Measures of the Developed model.

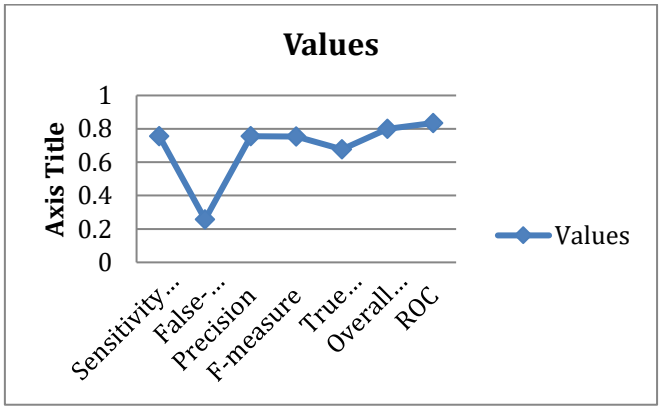

Figure 2: Performance Measures

A ROC curve is produced by plotting TP rate on y-axis versus FP rate on x-axis as shown in figure 2. The FP can also be expressed as (1 – specificity) or TN rate. A Precession- Recall curve is also plotted as shown in figure 3.

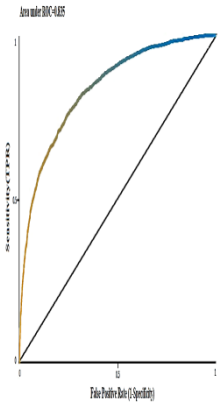

Figure 3: ROC Curve

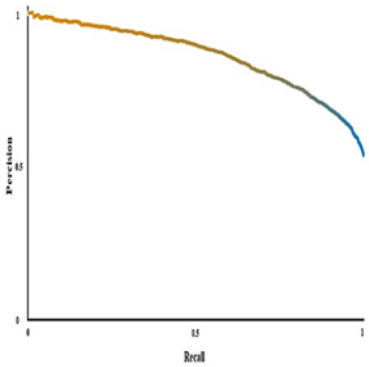

Figure 4: Precision-Recall Curve

As the ROC curve climbs quickly towards top left corner, which means that our rainfall prediction model correctly predicts the classes with appreciable accuracy. Area under ROC curve is often used as a measure of quality of the machine learning classification model. A random classifier has an area under the curve of 0.5, while area under ROC for a perfect classifier is 1. In our experimental work, Area under ROC curve (AUC) is 0.835 (83.5%) which implies that our prediction model has good level of prediction accuracy and is quite reasonable and acceptable.

5. Conclusion and Future Scope

In this work we have proposed a rainfall prediction model which predicts the rainfall using Naive Bayes approach. The performance of the proposed model is calculated using various performance measures like precision, F-measure, Recall, Accuracy, ROC etc. The model produces 5795correctly classified instances which are 75.55% and 1875 incorrectly classified instances which are 24.45%. The proposed rainfall prediction model produces 5795 correctly classified instances which are 75.55% and 1875 incorrectly classified instances which are

24.45% with ROC of 83.5%. The experimental observations showed that Naïve Bayes approach to develop rainfall prediction model has good level of prediction and is quite satisfactory in rainfall prediction. The developed machine learning model showed the prediction accuracy of 71.2% when applied on the test weather data set to forecast the rainfall. Every machine learning algorithm has its advantages and limitations; it is difficult to settle for the best algorithm. The prediction accuracy of the model can be increased by developing a hybrid prediction model where multiple machine learning algorithms are amalgamated in a single model to provide more accurate and correct predictions.

References:

1. Molua EL. The economic impact of climate change on agriculture in Cameroon. World Bank Policy Research Working Paper. 2007 Sep 1(4364).
2. Dorman CE. Early and recent observational techniques for fog. InMarine fog: Challenges and advancements in observations, modeling, and forecasting 2017 (pp. 153-244). Springer, Cham.
3. Diez-Olivan A, Del Ser J, Galar D, Sierra B. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. Information Fusion. 2019 Oct 1;50:92-111.
4. Ji L, Wang Z, Chen M, Fan S, Wang Y, Shen Z. How much can AI techniques improve surface air temperature forecast?—A report from AI Challenger 2018 Global Weather Forecast Contest.
5. Pham BT, Luu C, Van Phong T, Trinh PT, Shirzadi A, Renoud S, Asadi S, Van Le H, von Meding J, Clague JJ. Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling?. Journal of hydrology. 2021 Jan 1;592:125615.
6. Schindler DW. The cumulative effects of climate warming and other human stresses on Canadian freshwaters in the new millennium. InWaters in Peril 2001 (pp. 165-186). Springer, Boston, MA.
7. L′vovich MI. World water resources and their future. American Geophysical Union; 1979.
8. Lansigan FP, De Los Santos WL, Coladilla JO. Agronomic impacts of climate variability on rice production in the Philippines. Agriculture, ecosystems & environment. 2000 Dec 1;82(1-3):129-37.
9. Tobin D, Janowiak M, Hollinger DY, Skinner RH, Steele R, Radhakrishna R. Northeast regional climate hub assessment of climate change vulnerability and adaptation and mitigation strategies.
10. Idoje G, Dagiuklas T, Iqbal M. Survey for smart farming technologies: Challenges and issues. Computers & Electrical Engineering. 2021 Jun 1; 92:107104.
11. Liu, James NK, Bavy NL Li, and Tharam S. Dillon. "An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 31.2 (2001): 249-256.

12. Yim, S.Y., Wang, B., Xing, W and Lu, M.M, "Prediction of Meiyu rainfall in Taiwan by multi-lead physical–empirical models," Climate Dynamics, vol.44, no.11, pp.3033-3042, 2015.
13. Nikam VB, Meshram BB. "Modeling rainfall prediction using data mining method: A Bayesian approach". In2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation 2013 Sep 24 (pp. 132-136). IEEE.
14. Nhita F, Adiwijaya UN. Forecasting Indonesian Weather through Evolving Neural Network (ENN) based on Genetic Algorithm. InThe Second International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE2014) 2014 Mar 18 (pp. 78-82).
15. Marjanović M, Krautblatter M, Abolmasov B, Đurić U, Sandić C, Nikolić V. The rainfall-induced landsliding in Western Serbia: A temporal prediction approach using Decision Tree technique. Engineering Geology. 2018 Jan 8;232:147-59.
16. Rivero CR, Tupac Y, Pucheta J, Juarez G, Franco L, Otaño P. Time-series prediction with BEMCA approach: application to short rainfall series. In2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI) 2017 Nov 8 (pp. 1-6). IEEE.
17. Kotu, Vijay, and Bala Deshpande. Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.
18. http://www.sciamachy-validation.org/climatology/daily_data/selection.cgi
19. Molajou A, Nourani V, Afshar A, Khosravi M, Brysiewicz A. Optimal design and feature selection by genetic algorithm for emotional artificial neural network (EANN) in rainfall-runoff modeling. Water Resources Management. 2021 Jun;35(8):2369-84.
20. Elhoseny M, Shankar K, Uthayakumar J. Intelligent diagnostic prediction and classification system for chronic kidney disease. Scientific reports. 2019 Jul 3; 9(1):1-4.
21. Ahmed, Bilal. "Predictive capacity of meteorological data: Will it rain tomorrow?" Science and Information Conference (SAI), 2015. IEEE, 2015.
22. Chandola V, Banerjee A, Kumar V. Outlier detection: A survey. ACM Computing Surveys. 2007 Aug 15;14:15.
23. Sayad S. Real time data mining. Cambridge: Self-Help Publishers; 2011 Jan 5.
24. D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, (2001).
25. Kotu, Vijay, and Bala Deshpande. Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann, 2014.
26. R. O. Duda and P. E. Hart, Pattern classification and scene analysis, John Wiley and Sons, 1973.
27. P. Langley, W. Iba and K. Thompson, "An analysis of Bayesian Classifiers.," in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992.
28. Kesavaraj, G., and S. Sukumaran. "A study on classification techniques in data mining." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.
29. Frank E, Hall M, Trigg L. Weka: Waikato environment for knowledge analysis. The University of Waikato, Hamilton, New Zealand. 1999.

30. Xu J, Zhang Y, Miao D. Three-way confusion matrix for classification: A measure driven view. Information sciences. 2020 Jan 1;507:772-94.
31. N. Friedman, D. Geiger and M. Goldszmidt, "Bayesian Network Classifiers.," Machine Learning, vol. 29, pp. 131-163, 1997.
32. Liu Y, Zhou Y, Wen S, Tang C. A strategy on selecting performance metrics for classifier evaluation. International Journal of Mobile Computing and Multimedia Communications (IJMCMC). 2014 Oct 1;6(4):20-35.