International Journal of Clinical & Medical Informatics

ISSN: 2582-2268 **Review Article** | Vol 2 Iss 2

Rule-based Naïve Bayes Classifier for Heart Disease Risk Prediction and Therapy Recommendation

Fasidi FO[1] and Adebayo OT[2*]

[1]Department of Computer Science, The Federal University of Technology Akure, Nigeria

[2]Department of Information Technology, The Federal University of Technology Akure, Nigeria

[*]**Corresponding author:** Adebayo OT, Department of Information Technology, The Federal University of Technology Akure, PMP 704, Akure, Nigeria; Email: adebayoot@futa.edu.ng

Abstract

With machine learning, healthcare service providers can make better decisions on patient's diagnoses and treatment options for overall improvement of healthcare services but, it is important to solve the issue of disparate origins of data collected. In this paper, Rule-based Naïve Bayesian Classifier (RNBC) is adopted for prediction and therapy recommendation of heart disease risks. Dataset is scanned only once, off-line, at the time of building the classification rule set therefore subsequent scanning of dataset is avoided as it is with conventional Naïve Bayes Classifier. Variables such as, gender, age, chest pain discomfort type, blood sugar level, resting electrographic results, maximum heart rate achieved, exercise induced angina, depression induced by exercise, patient medication, irregular heartbeat, systolic blood pressure, diastolic blood pressure, dark spot under nails, skin color changes, bleeding gum, swelling, vaccination, fatigue, heart burn, diabetes, indigestion were used for heart diseases prediction and therapy recommendation. The system is implemented using Java programming language with My Structured Query language (MySQL) as the Database management system. Standard metrics such as Precision, Recall, ease of understanding, user friendliness of the system was used to evaluate the performance of the proposed system and the result of evaluation shows that the system is adequate in Heart Disease prediction and therapy recommendation. The system could be used by Cardiologists and medical students to assist them in Heart Disease prediction and therapy recommendation for patients.

Keywords: ***Data mining; Heart diseases risks; Recommender system; Rule-based naïve bayesian classifier***

Received Date: May 15, 2019; **Accepted Date:** May 25, 2019; **Published Date:** June 01, 2019

Introduction

A model for predicting heart disease risk level of a patient using Rule-based Naïve Bayesian Classifier and recommend the necessary therapy is presented. A set of classification rules were constructed on top of Naïve Bayes classifier. Learned model using Naïve Bayes Classification is represented as a set of IF-THEN rules, in Rule-based Naïve Bayesian Classifier, rules are a very good way of representing information or bits of knowledge [1]. Classification algorithm was applied on patients' records collected from hospital. The diagnosis of heart disease is a significant and tedious task in medicine the detection of heart disease

Citation: Adebayo OT, Rule-based Naïve Bayes Classifier for Heart Disease Risk Prediction and Therapy Recommendation. Int J Clin Med Info 2019: 2(2) 51-59.

from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects, thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option [2]. The healthcare industry gathers enormous amounts of heart disease data that regrettably, are not "mined" to determine concealed information for effective decision making by healthcare practitioners [3].

In medical data mining, Naïve Bayes classification has played an indispensable role it is one of the most effective and efficient classification algorithms which has been successfully applied to many medical problems [4] , such as in diagnosis and treatment of breast and lungs cancer, in prediction of dementia, and so on. Naive Bayes is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions [5]. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable [6]. Depending on the precise nature of the probabilistic model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting, in many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods [7].

Related Works

Tremendous works in literature related with heart disease diagnosis using data mining techniques have motivated the research. The researchers in the medical field diagnose and predict the diseases in addition to providing effective care for patients [8] by employing the data mining techniques. The data mining techniques have been employed by numerous works in the literature to diagnose diverse diseases, for instance: Diabetes, Hypertension, Heart diseases and more [4]. A Prototype Knowledge based System for Anxiety and Mental Disorder Diagnosis with the aid of data mining techniques such as rule-based technique and back ward chaining mechanism was proposed in [7]. The study recommended full implementation of the functionality of the prototype system and integration of rule-based system with case-based techniques for a better result [9]. Presented heart disease diagnosis system using classification trees, the performance was compared with classification of different types of heart failures such as preserved ejection fraction and reduced ejection fraction. Comparative analysis was done using WEKA tool. [10] Presented an intelligent and effective heart attack prediction system. For the extraction of significant patterns from Cleveland dataset repository for heart attack prediction. In the work, K-means clustering algorithm was adopted. Afterwards, the Maximal Frequent Items et. (MAFIA) algorithm was used to mine the frequent patterns applicable to heart disease [9]. Presented a prototype Naive Bayes algorithm to find the chances of occurrence of eight skin diseases such as chicken pox, Kawasaki disease, measles, acne, rosacea, groin rash and callus on the basis of input variables. Analysis of data from different surveys was used to decide whether it is suitable to be analyzed with the use of the data mining methods [11]. Proposed feature selection in ischemic heart disease. The proposed method was to develop a system with efficiency in terms of improved cost, time and accuracy. It used an Artificial Neural Network to select important features from the input layer of the network and to select important features from an Ischemic Heart Disease (IHD) data base with 712 patients, the Multi-Layer Perception Neural Network was used.

Heart Disease and Risk Factors

The term Heart disease encompasses the diverse diseases that affect the heart. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood Cardiovascular disease (CVD) results in severe illness, disability, and death [12].

No matter what kind of heart disease a patient may have, there are risk factors that can cause or exacerbate it. Risk factors for developing heart disease include; age which increases patient's risk of damaged and narrowed arteries and weakened or thickened heart muscle, sex or gender; men are generally at greater risk of heart disease [4]. However, women's risk increases after menopause, family history of heart disease also increases patient risk of heart disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as mother or sister), smoking nicotine constricts blood vessels, and carbon monoxide can damage the inner lining, making them more susceptible to atherosclerosis [13]. Heart attacks are more common in smokers than in non-smokers, poor diet also contributes to risk of heart disease; a diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease, Uncontrolled high blood pressure can result in hardening and thickening of the arteries, narrowing the vessels through which blood flows, high levels of cholesterol in the blood can increase the risk of formation of plaques and atherosclerosis, diabetes increases risk of heart disease [14]. Both conditions share similar risk factors, such as obesity and high blood pressure, obesity and excess weight typically worsens other risk factors, physical inactivity also is associated with many forms of heart disease and some of its other risk factors, as well, unrelieved stress may damage the arteries and worsen other risk factors for heart disease, poor hygiene; not regularly washing hands and not establishing other habits that can help prevent viral or bacterial infections that can put a patient at risk of heart infections, especially if the patients already have an underlying heart condition. Poor dental health also may contribute to heart disease [12].

Research Methodology

MySQL a relational database that stands as the data store for the whole application [15]. Java programming language is suitable for this application because it can hold huge amount of data and seamlessly integrates its framework. The proposed models for prediction and recommendation identify three main heart diseases: Ischemic, cerebrovascular, and inflammatory heart diseases. Ischemic heart disease is caused by narrowing of the coronary arteries and therefore a decreased blood supply to the heart; cerebrovascular heart disease, caused by cerebrovascular accident or stroke is the result of an impeded blood supply to the heart and some part of the brain and inflammatory heart disease is caused by inflammation of the heart muscle (myocarditis), the membrane sac (pericarditis) which surrounds the heart, the inner lining of the heart (endocarditis) or the myocardium (heart muscle) [16]. Inflammation may be caused by known toxic infectious agents or by an unknown origin [17]. Java programming language was used to develop the model and the processing operation of the model.

Naïve Bayes Classifier for Heart Disease Risk Prediction and Recommendation

Naïve Bayesian classification is based on Bayesian Theorem. If a new record D is to be classified, Bayesian Theorem can be used to find the probability that it belongs to class Ci by using Equation. Naïve Bayes model was adopted for the heart disease risks prediction and therapy recommendation. The Naïve Bayes model is given as:

$$p(c_j|d_n) = \frac{p(c_j)\, p(d_n \mid c_j)}{p\,(d_n)} \tag{1}$$

$p(c_j|d_n)$ represents the conditional probability of C given that D has occurred. Ci is one of a set of classes {C_1, C_2, C_3 …} that are used to classify the data. The class whose P (C_i |D) is highest is selected as the record's class. When computing *P (*C_i |D*)* for every C_i to determine the class with the highest probability, the denominator *P* (D*)* is constant across all classes. Therefore, it can be removed from the computations. Therefore Equation (2) below can be used to find the class with the highest probability.

$$p(c_j|\mathrm{D}) \sim p(\mathrm{D}|c_j)\, p(c_j) \qquad (2)$$

where the symbol "~" indicates that the LHS is proportional to the RHS. Further, Naïve Bayesian classification assumes class-conditional independence (that is why it is called "naïve"). This assumption basically states that attribute values of the record D are independent of each other. In other words, if D is the n-record<d_1, d_2… d_n>, then

P (D | C_j) in Equation (2) can be computed as:

$$p(\mathrm{D}|c_j) = \prod_{k=1}^{n} p(d_k|C_j) = p(d_1|C_j) * p(d_2|C_j) * \ldots * p(d_n|C_j) \qquad (3)$$

This is because, from Probability Theory, the probability of the conjunction of independent events can be obtained by multiplying the probabilities of the individual events. In summary, to compute P (Ci | D) based on Equation (2), $\mathrm{P}(\mathrm{D}|c_j)$ is computed based on Equation (3) and compute P(Cj) then multiply the two results, this is repeated for each class Cj.

Rule-Based Naïve Bayes Classifier for Heart Disease Risk Prediction and Therapy Recommendation

A three-step methodology for building a rule-based classification system based on Bayesian classification is descried. In this approach, there is a learning phase in which the system follows the three steps to extract classification rules. In this approach and in classification in general, the attributes are assumed to be discretized (or categorized). If the values in these attributes are continuous, a pre-processing phase is performed to discretize them. The steps of the methodology used in RNBC are outlined below:

Step 1: Generate all possible combinations of attribute values that exist in the dataset.

Step 2: For each combination of attribute values found in step 1, compute the probability of each class

Step 3: Generate the classification rules, one rule for each combination of attribute values found in Step 1.

Learned model using Naïve Bayes Classification is represented as a set of IF-THEN rules. Rules are a very good way of representing information or bits of knowledge. An IF-THEN rule is an expression of the form:

IF condition THEN conclusion (4)

Set of classification rules constructed on top of Naïve Bayes classifier has shown great performance in terms of accuracy because it takes only some time to calculate the accuracy than other algorithms such as k-NN (k-Nearest Neighbor's algorithm) and decision list so if attributes are independent with each other we can use it in medical fields [1].

Experimental Results and Discussion

This section introduces the data sets and methods used in designing the experiment.

Data Description: Experimental evaluation of the system was done with data of 1600 patients alongside attributes (twenty-one risk factors), data preprocessing was done by transforming the heart disease data into an understandable format. The data collected was checked for the presence of error in data entry including misspellings and missing data. Following this process, there was no error in misspellings but there were missing data in the cells describing some records for the attributes chest discomfort type and exercise induced angina (ten cells), data reduction of the original twenty-one attribute was carried out to select the most important feature of attributes for heart disease risk prediction (ten most important feature were selected). The data was transformed into the attribute file format (.csv) for the purpose of the development of the predictive model for heart disease risk. The dataset collected for the purpose of the development of the predictive model for the diagnosis of heart disease risk was stored in (.csv) in the name heart disease Data (.csv). Rule-based Naïve Bayes model was used for heart disease risk and therapy recommendation. Rule-Based Naïve Bayes model was adopted for the heart disease risks prediction. The set of rules was constructed on top of Naïve Bayes Classifier.

Heart disease risk levels which can either be severe, high average or low. The system presented recommendation based on two therapies: life style therapy, which has to do with life style changes and pharmacology therapy which has to do with drug prescription.

A test set is used to determine the accuracy precision and recall of the model. The resulting model applied to the condition and testing instances, after completing the training.

The True Positive Rate, False Positive Rate, True Negative Rate and False Negative Rate determine the predictive data mining efficiency [18]. In this study, we used two predictive and recommender model performance evaluation metrics; precision and recall for the models.

S/N	Age	Gender	Chest Distype	Blood_sugar	Resting electrographic result	Max heartrate	Exang	IHB	BP1	BP2
1	>40	f	typical type1 angina	>120 mg/dl	normal	>125m/s	yes	yes	> 120mg/dl	>80 mg/dl
2	>40	f	typical type2 angina	>120 mg/dl	normal	>125m/s	yes	yes	>120mg/dl	>80 mg/dl
3	>40	f	non angina pain	>120 mg/dl	normal	>125m/s	yes	yes	> 120mg/dl	>80 mg/dl
4	>40	f	asymptomatic	>120 mg/dl	normal	>125m/s	yes	yes	>120mg/dl	>80 mg/dl
5	>40	f	typical type1 angina	<120 mg/dl	normal	>125m/s	yes	yes	> 120mg/dl	>80 mg/dl

Table 1: Sample dataset showing attribute value.

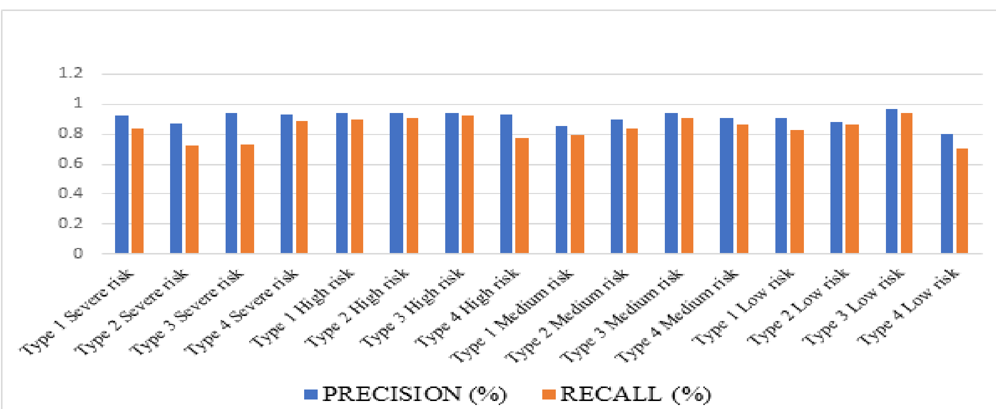

Figure 1: Result of precision and recall of Rule-based Naïve Bayesian Classifier.

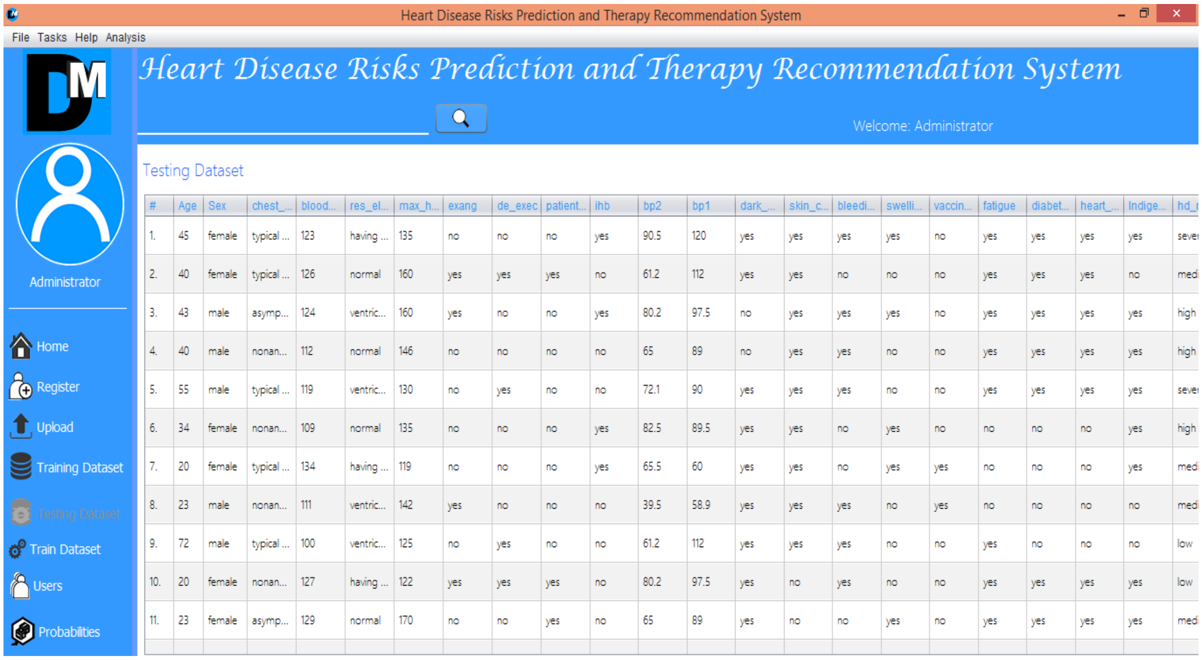

Figure 2: Interface for heart disease risks prediction.

Figure 3: Interface for therapy.

System Evaluation

To compare the performance of the system of Rule-based Naïve Bayesian Classifier, and ruled based recommendation, Heart Disease Dataset is divided into two parts for example, training data 70% and testing data 30%. The evaluation of the system was carried out using Precision, Recall, F-measure, Overall Precision, Average Precision, Overall Recall, and Average Recall. Precision shown in equation (4) (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances [16].

Both Precision and Recall are therefore based on an understanding and measure of relevance [20]. Overall and Average Precision are shown in equation (8) and (9) respectively. Overall and Average Recall are shown in equation (10) and (11).

$$Precision=\frac{TP}{TP+FP} \quad (4)$$

$$Recall= \frac{TP}{TP+FN} \quad (5)$$

$$F-measure = \frac{2 * Precision\ x\ Recall}{Precision\ x\ Recall} \quad (6)$$

$$Average\ F-measure = \frac{\sum_{i=1}^{n} F-measure_i}{n} \quad (7)$$

$$Overall\ Precision = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i}^{n} TP_i + FP_i} \quad (8)$$

$$Average\ Precision = \frac{\sum_{i=1}^{n} Precision_i}{n} \quad (9)$$

$$Overall\ Recall = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i}^{n} TP_i + FN_i} \quad (10)$$

$$Average\ Recall = \frac{\sum_{i=1}^{n} Recall_i}{n} \quad (11)$$

TP = (True Positive): Number of correct predictions of an instance that is relevant.

FP = (False Positive): Numbers of incorrect predictions of an instance that is relevant.

FN = (False Negative): Number of incorrect predictions of an instance that is irrelevant.

The model gives average precision of 91% and recall of 83% and average F-measure of 69% the precision and recall values for the classes are given in (Figure 1). The interface for heart disease risks prediction ad therapy is shown in (Figure 2 and Figure 3). Experimental tests were conducted based on the two models: Rule-Based Naïve Bayes and the conventional Naïve Bayes model. The Rule-based Naïve model was tested against 480 data point after being trained with 1,120 data points. This exact same sample was also executed by the conventional Naïve Bayes model. The results from the two model were compared.

Model	AP	AR	AF
Naïve Bayes	0.86	0.79	0.50
RBNC	0.91	0.83	0.69

Table 4: Result of Comparison of current work using conventional naïve bayes model in terms of average precision, recall and f-measure.

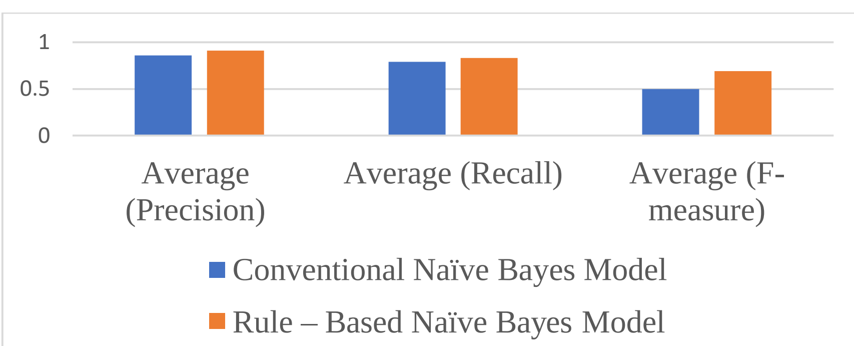

Figure 4: Comparison of current work using conventional naïve bayes model in terms of average precision, recall and f-measure.

The Rule-based Naïve Bayes model achieved higher Average Recall (91%), Precision (83%) and F-Measure (69%) as against the conventional Naïve Bayes model with Average Recall (86%), Precision (79%) and F-Measure (50%). (Table 4) and (Figure 4) shows the result of comparison. The result show that the Rule-Based Naïve Bayes model have higher precision, recall and f-measure than the conventional Naïve Bayes model, hence, higher performance.

Conclusion

Machine Learning techniques are widely used in various fields and primarily health care industry has benefitted a lot from machine learning prediction techniques. It offers a variety of decision support tools targeted at improving patients' healthcare quality. Most of the existing systems make predictions without recommending therapies. Hence, in this thesis, the possibility of adopting a model that can serve the purpose of risk prediction and therapy recommendation was explored. A Rule-based Naïve Bayes model is proposed for heart disease risks and therapy recommendation. The system allows record entry of new patient which provides categories of heart disease risks and give lifestyle ad pharmacology therapy. In the conventional Naïve Bayes model, whenever a new data record is to be classified, the entire dataset must be scanned in order to apply a set of equations that perform classification. This is time consuming especially for big data. The proposed Rule-Based Naïve Bayes model in this work helps to solve this problem.

Furthermore, the system was evaluated by twenty-five users consisting of ten experts and fifteen patients from hospital. The users rated the proposed system in terms of ease of usage, satisfaction, acceptability, user friendliness and speed. From the two test groups; 87% users were overall satisfied with the recommendation. They were pleased with the application's ability to give drugs and lifestyle therapy recommendation, 93% users indicated that the system is quite easy to use, 98% users indicated that the system is user friendly, 90% of the users love the system, and 98% of the users indicated that it has decent speed in recommendation. Therefore, the proposed system classifies the given data into different categories and also predicts the risk of the heart disease if unknown sample is given as an input thereafter gives therapy recommendations.

Future research work could incorporate other classification techniques, such as decision tree, KNN, and many other classification techniques can be conducted and compared to find the most valid one. Data associated with cardiovascular disease analysis, such as ECG, echocardiography or coronary angiography can be used for further research. In addition, improvement can be done in this area by involving a large number of experts through crowdsourcing or health social network.

References

1. Chitra R, Seenivasagam V (2013) Heart disease prediction system using supervised learning classifier. Bonfring International Journal of Software Engineering and Soft Computing 3(1): 01-07.
2. Rafiah H, Ashrafuzzaman H, Mazaharul, et al. (2008) Heart attack detection using data mining techniques. International Journal of Technology Enhancements and Emerging Engineering Research 1(5): 2447-4679.
3. Shouman M, Turner T, Stocker R (2013) Integrating clustering with different data mining techniques in the diagnosis of heart disease. Journal of Computer Science and Engineering 20(1).
4. Rish P (2010) Gene Selection Approach for Classifying Diseases Based on Microarray Datasets. 2nd International Conference on Computer Technology and Development 5(2): 40-55.
5. Austin PC, Tu JV, Ho JE, et al. (2013) Using methods from the data-mining and machine-learning literature for disease

classification and prediction: a case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology 66(4): 398-407.

6. Curiac DI, Vasile G, Banias O, et al. (2009) Bayesian network model for diagnosis of psychiatric diseases. In Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces IEEE 61-66.
7. Seblewongel Y (2011) Prototype knowledge-based system for anxiety mental disorder diagnosis. Ethiopia, Addis Ababa: Addis Ababa University 45-50.
8. Chaurasia, V. and Pal, S., 2017. A novel approach for breast cancer detection using data mining techniques. International Journal of Innovative Research in Computer and Communication Engineering.
9. Jack E (2013) Clinical Decision Support System for Heart Disease Diagnosis. Journal of Healthcare Information Management Archives 19(8): 155-164.
10. Manjusha KK, Sankaranarayanan K, Seena P (2014) Prediction of different dermatological conditions using naive bayesian classification. International Journal of Advanced Research in Computer Science and Software Engineering 4(1).
11. Manjusha C (2014) Heart Attack Prediction System. ACM Compute, Surveys.
12. Salama GI, Abdelhalim M, Zeid MAE (2012) Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 32(569): 2.
13. Ahmed A, Hannan SA (2012) Data Mining Techniques to Find Out Heart Diseases: An Overview. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 1(4): 18-23.
14. Tan K, Tech E (2009) A hybrid evolutionary algorithm for heart disease prediction. Journal of Expert System with Applications 2(5): 8616-8630.
15. Ilayaraja M and Meyyappan T (2013) Mining medical data to identify frequent diseases using Apriori algorithm. In 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering IEEE: 194-199.
16. Andreeva P (2006) Data modelling and specific rule generation via data mining techniques. In International Conference on Computer Systems and Technologies-CompSysTech 8(3): 118-120.
17. Hand DJ, Mannila H, Smyth P (2001) Principles of data mining (adaptive computation and machine learning). Cambridge, MA: MIT Press: 361-452.
18. Lavanya D, Rani KU (2012) Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services 2(1): 17.
19. Han J (2015) Data mining techniques (3rd Edn). The Morgan Kaufmann Series 4(5): 34-46.
20. Mining WID (2006) Data mining: Concepts and techniques. Morgan Kaufinann.