Using Machine Learning to Analyze Diabetics Disease for Pregnant Women

Anchal Kumari*[a], Rahul Deo Sah [b], Rajendra Kumar Mahto[c]

[a] Radha Govind University, Ramgarh
[ab] Dr. Shyama Prasad Mukherjee University, Ranchi
Corresponding author email id:
anchal4kumari@gmail.com

ABSTRACT

Data mining allows for the prediction of a variety of diseases, including but not limited to hepatitis, lung cancer, liver disorders, breast cancer, thyroid disease, diabetes, and so on. In this research paper, diabetes prognoses are investigated. Diabetes is a condition that develops when the body does not produce enough insulin. Diabetes mellitus can be broken down into these four distinct subtypes. Diabetes mellitus type 1, type 2 diabetes, diabetes caused by pregnancy, and diabetes that is present at birth are the most prevalent forms of diabetes. When individuals talk about having diabetes, they almost always imply that they have diabetes mellitus (DM). Diabetes mellitus (DM) patients are referred to as "diabetics." Congenital diabetes is caused by genetic insulin secretion defects, cystic fibrosis-related diabetes, and steroid diabetes caused by high glucocorticoid doses. Diabetes, if left untreated or poorly managed, can lead to a number of complications, including heart attack, stroke, kidney failure, blindness, erection problems (impotence), and amputation. In this paper, we focus on diabetic disease in pregnant women. Using the Dataset analyzes with the help of different machine learning techniques. To find out the best accuracy rate and less error rate provided by the Machine Learning, i.e., Support Vector Machine. Discovered confusion metrics and standard deviation in the algorithms.

Keywords: - *Pima, support vector machine, pca, boosted trees*

1. Introduction

The use of data mining in the field of medical science for the purpose of disease prediction is expanding on a daily basis, both in terms of its breadth and its potential applications. In this research paper applies classifiers for selection-based classification to medical disease data and presents a classifier selection approach based on clustering. The classification technique is based on classifier selection.

Within the framework of the approach, a large number of clusters are selected for an ensemble process. After that, the standard presentation of each classifier on each cluster is determined, and the classifier that has the best average performance is picked to categorical data that has been provided. In order to calculate the normal average, the technique known as the weighted average is utilized. The values of the weights that are assigned to each cluster are determined by using the distances that separate the given data from each cluster. There are two different approaches that can be taken when combining multiple classifiers: selection and fusion. The use of many classifiers is predicated on the assumption that each classifier possesses knowledge in some local regions of the feature space.

The performance of a classifier is typically the most essential part of its value, and it may be quantified using a variety of well-known methodologies and matrices. On the other hand, having knowledge of a classifier is typically considered secondary, or even ignored. Yet, it is essential for users of a classifier to understand how the classifier works since it might extract more knowledge about the relationships in the observed data. As a consequence of this, some of the older methods center on the acquisition of knowledge regarding classifiers through learning or the transformation of classifiers based on non-knowledge into structures based on human knowledge. The field of classifier construction has been transformed into a heuristic optimization crisis as a result of the lack of algorithms that treat the accuracy of classifiers and the knowledge they possess as being of equal significance. These algorithms are especially significant in fields where some regions of the attribute space can be classified with a high level of accuracy. Using knowledgeable classifiers and sections that require non- knowledgeable classifiers are both necessary in order to reach the requisite level of classification accuracy.

2. The Value of Data Mining in Health Care

Generally, all healthcare organizations around the world keep healthcare data in electronic format. The majority of information that constitutes healthcare data is made up of details about patients as well as organisations and individuals involved in the provision of healthcare. The amount of information that can be stored is growing at a rapid rate. As the size of the electronic healthcare data continues to expand over time, a level of complexity is introduced into the system. In a nutshell, the data associated with healthcare has become quite difficult to understand. It becomes extremely challenging to get useful information from it using the approaches that have traditionally been used. But, as a result of developments in areas such as statistics, mathematics, and other fields, it is now possible to extract significant patterns from them. Recently, researchers have been using data mining tools in a distributed medical environment in order to improve the quality of medical services provided to a large proportion of the population at a lower cost, to improve customer relationship management, to manage healthcare resources more effectively, and so on. It gives significant information in the field of healthcare that management can use to make decisions such as medical staff estimation, health insurance policy decisions, therapy selection, disease prognosis, so on.

3. Review of Current work

In the study "Cluster-Oriented Ensemble Classifier: Effect of Multi-cluster Characterization on Ensemble Classifier Learning," [1] a novel cluster-oriented ensemble classifier is introduced. This cluster-oriented ensemble classifier was developed using novel ideas. The base classifier is responsible for discovering the borders of the clusters, while the fusion classifier is in charge of mapping the confidences of the clusters onto the class decision. According to the study, an ensemble classifier is constructed using a group of basic classifiers that learn the class boundaries individually across the pattern. These classifiers are used to build the ensemble classifier. Because this difficulty is inherent in all basic classifiers, learning the borders between overlapping classes is a challenging challenge to solve. The idea of clustering develops as a direct consequence of this. Clustering is the process of grouping a set of objects into groups that each contain a number of individual items. The process of producing several iterations of a predictor and then compiling all of those into a single model is referred to as "bagging" [3] [4]. While attempting to forecast a class, the aggregate takes into account a plurality of votes but uses an average calculated across all versions to forecast a numerical outcome. The learning set is bootstrapped, and the findings are used to produce new learning sets, which allows for the creation of many versions. Bagging can yield significant accuracy gains in tests using real and simulated data sets, classification and regression trees, and subset selection in linear regression analyses. In the research paper titled "Analysis of Bagging [5] as a Linear Combination of Classifiers," the author discusses how an analytical framework for analyzing linearly combined classifiers may be applied to bagging ensembles. This leads to the development of an original analytical model of the chance of bagging misclassification as a function of the size of the ensemble. Experimentation performed on real-world data sets provides evidence that supports the theoretical expectations. An Implication of Data Diversity for a Classifier-Free Ensemble Selection in Random Subspaces [6], also known as an Ensemble of Classifiers (EOC), has been found to improve the performance of single classifiers by combining their outputs. Pattern recognition systems should strive for the highest feasible level of classification accuracy in order to fulfill their purpose. The fact that the projections are a linear combination of all the original features or variables is one of the major drawbacks of linear dimensionality reduction algorithms [7] like principal component analysis (PCA) and linear discriminate analysis (LDA). Another major drawback is that all of the weights in the linear combination, also known as loadings, are typically not zero. In a variety of application domains, one of the

most critical challenges is the representation of high-dimensional data in lower dimensions. Boosting [8] is a strategy that is used for the construction of classifier ensembles. The term "boosting" refers to a collection of several strategies for the construction of classifier ensembles. These approaches have the distinct advantage of being able to produce a powerful classifier from a group of less effective classifiers, which is their defining characteristic. As a direct consequence of this, boosting algorithms are suitable for use with relatively straightforward fundamental classifiers. One of the most straightforward classifiers is a decision stump, sometimes known as a decision tree with a single decision node. The musical groups that have been named As a novel binary discriminative learning technique, Toward a Structural Characterization of the Classification Boundary [9] [12] is based on a piecewise linear smooth additive model approximating the nonlinear decision boundary. This model is used to approximate the boundary of the decision. The decision border is geometrically defined by the characterizing boundary points that belong to the optimal boundary in accordance with a given idea of robustness. Because it is able to forecast the number of classifier training epochs that are necessary to attain optimal performance in an ensemble of MLP classifiers, the title an is used to define an MLP Classifier Design measure [10] [11]. The measure is computed between pairs of patterns on the training data and is derived from a spectral representation of a Boolean function. [14][15] This format allows for both accuracy and variety to be incorporated into a single measure as it characterizes the mapping from classifier decisions to the target label. MLPs are powerful classifiers that may surpass other classifiers in terms of performance, yet they are commonly criticized for the huge number of free parameters.

4. Proposed Prediction System

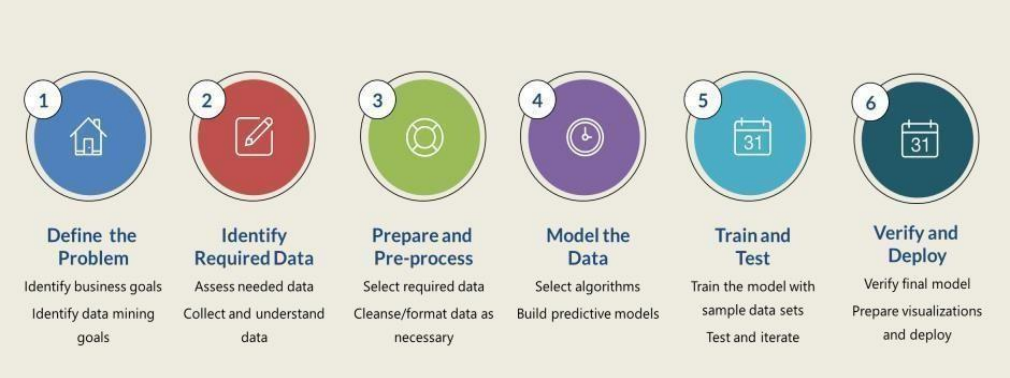

Fig. 1. Data Mining Approaches for System Building

We developed a data mining tool with the purpose of doing data analysis. Because of its straightforward user interface (depicted in Fig. 1), it simplifies the data mining process and makes it more accessible to users. This piece of equipment has been included in the ecosystem. First things first, you need to define the challenge, which is figuring out the goal of the data mining project. The second step is to determine the data that will be needed, which will then be evaluated, gathered, and analyzed. In the third steps Following that, we choose the necessary data for the data to be cleansed and prepared in accordance with the requirements for the data preprocessing.Fourth steps subsequent procedure resulted in the creation of some models for picking algorithms for the purpose of developing predictive models. Using the same user interface, it is possible to both develop and maintain courses as well as carry out all operations related to data mining. Fifth steps for training and testing the models with sample data sets for testing and iterating, which is a framework for developing data mining models such as classification, regression, clustering, pattern mining, and so on. Six teps for verifiying and deploying on dataset.

4.1 Machine Algorithm for Training and Testing

(a)-To retrain a classifier trained with T, Enter: [trained Classifier, validation Accuracy] = train Classifier(T) (b) yfit = trained Classifier to make predictions on new data T2. predictFcn(T2). T2 must have the same predictor columns as training.(Predict and react This code prepares data for model training. input Table = training Data predictor Names

='Pregnancies,' 'Glucose,' 'Blood Pressure,' 'Skin Thickness,' 'Insulin,' BMI,'DiabetesPedigreeFunction,' 'Age'; (d) predictors = input Table (predictor Names); (e) response = inputTable.Outcome; (f) isCategoricalPredictor = Classifier-training This code trains and specifies classifier options. Classification through SVM SVM = fitcsvm (predictors, response, 'Kernel Function', 'linear', 'Polynomial Order', [], 'Kernel Scale', 'auto', 'Box Constraint', 1,
'Standardize', true, 'Class Names', [0; 1]);
Create the result struct with predict function- (1) trainedClassifier.predictFcn = @(x) svmPredictFcn(predictorExtractionFcn(x)); Addto result struct
Classifier-trained- (2)
Age, BMI, Blood Pressure, DiabetesPedigreeFunction, Glucose, Insulin, Pregnancies, Skin Thickness;
Classification Trained Classifier = SVM (3)
How To Predict = sprintf ('To make predictions on a new table, T, use: yfit = c. predictFcn(T), where c is thename of this struct's variable, e.g., trainedModel. n nTABLE
Cross-validate trainedClassifier.ClassificationSVM, 'KFold,' 5; Validation predictions = kfoldPredict(partitionedModel); validationAccuracy = 1 - kfoldLoss (partitionedModel, 'LossFun', 'ClassifError');

4.2 Results Analysis

In the research paper, fitcsvm trains or cross-validates an SVM model for one-class and two-class (binary) classification on a low-dimensional or moderate-dimensional predictor data set. After training the dataset, three models were predicted. Cross validation for kfold-5 SVM true positive and false positive results

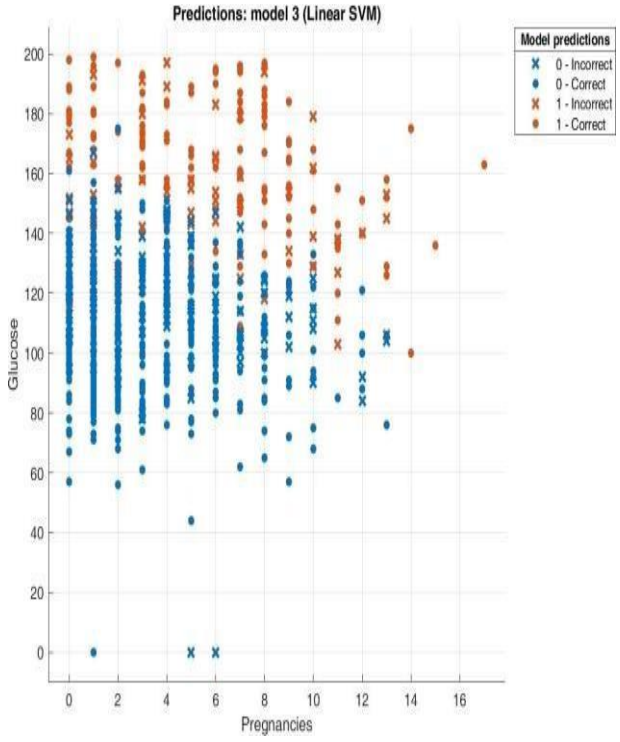

Fig. 2: Gaussian Naïve Bayes model for Pregnant Women

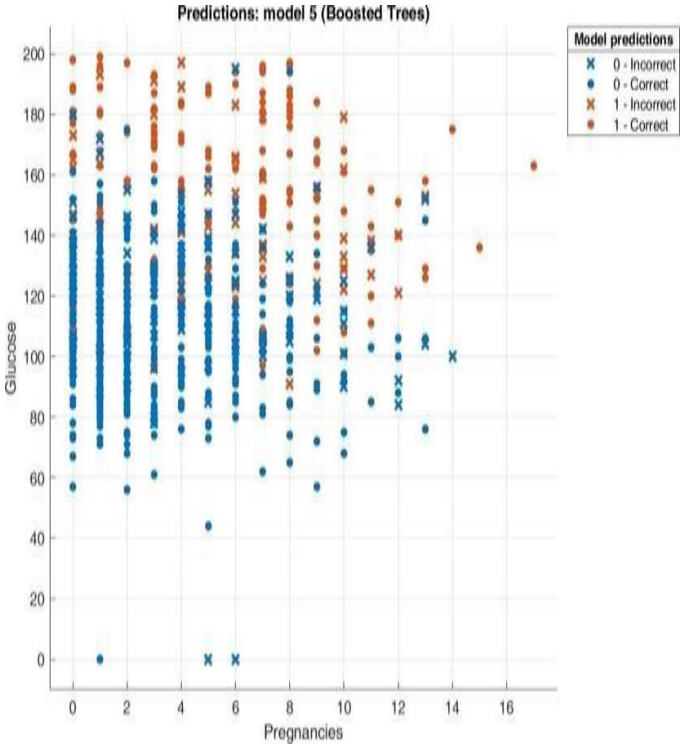

Fig. 3 Linear SVM model for Pregnant Women

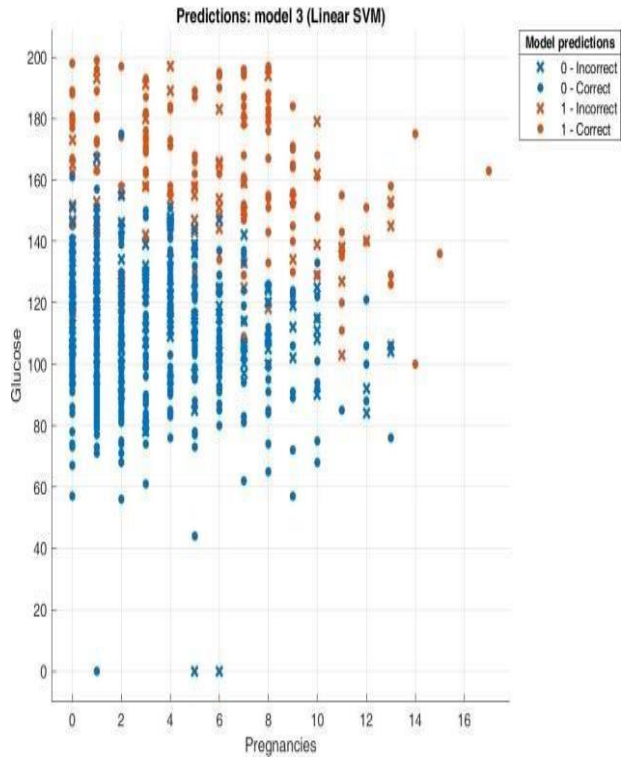

Fig. 4 coordinates for Gaussian naïve bayes model

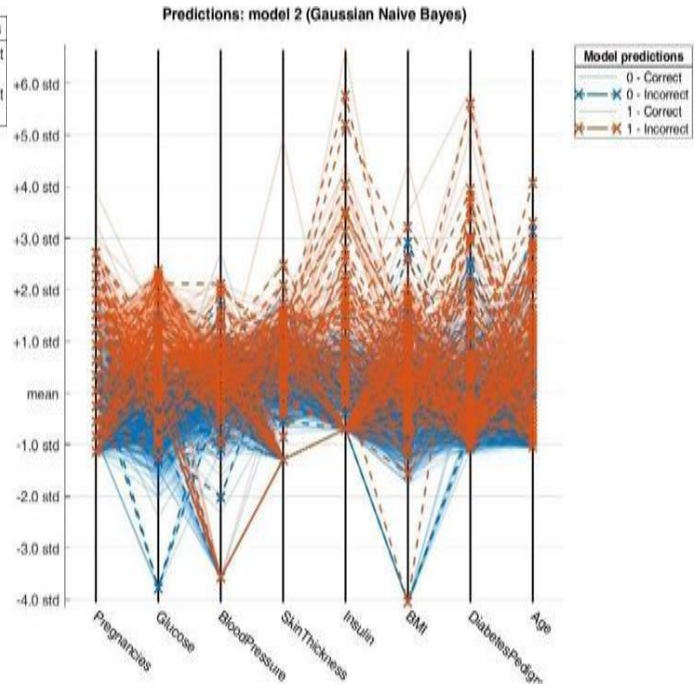

Fig. 5 coordinates for Gaussian naïve bayes model

The training step in naive Bayes classification is based on estimating $P(X|Y)$, the probability or probability density of predictors X given class Y. The naive Bayes classification model Classification Naïve Bayes and training function fitcnb provide support for normal (Gaussian), kernel, multinomial, and multivariate, multinomial predictor conditional distributions. To specify distributions for the predictors, use the Distribution Names name-value pair argument of fitcnb. It can specify one type of distribution for all predictors by supplying the character vector or string scalar corresponding to the distribution name, or specify different distributions for the predictors by supplying a length *D* string array or cell array of character vectors, where *D* is the number of predictors (that is, the number of columns of *X)*

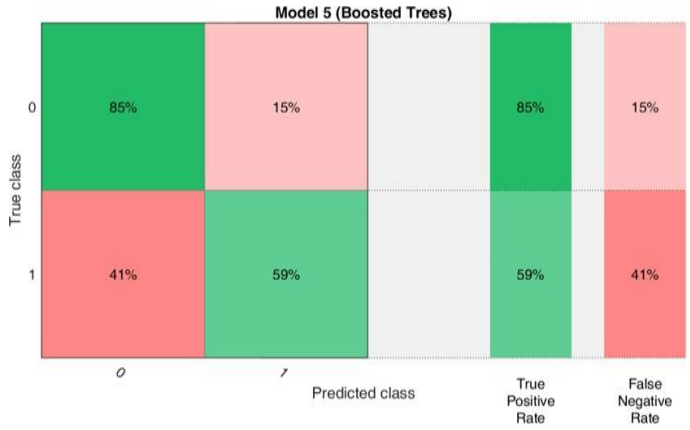

Fig.6 confusion matrices for naïve bayes

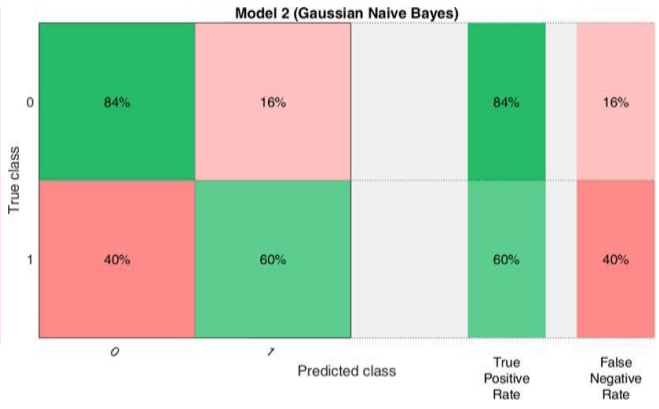

Fig.7 confusion matrices for boosted trees

In the confusion matrices true positive rate and false negative rate shown in thenaïve bayes.

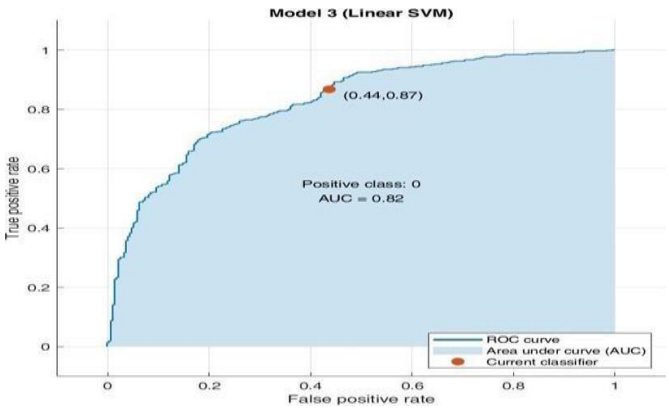

Fig. 8 Model-3 represents true positive rate and false positive rate is better than other that is positive class start from 0 to 0.44.0.87 and false positive rate is 1. Area under curve that is 0.82.

5. Conclusion

It is usual practice to refer to those who have been diagnosed with diabetes mellitus (DM) as "diabetics." Congenital diabetes is the term used to describe diabetes that is present at birth. Additionalconditions that can lead to diabetes include cystic fibrosis and steroid diabetes, which is brought on by taking an excessive amount of glucocorticoids. Diabetes can also be inherited. If diabetes is not treated or managed correctly, it can result in a number of serious complications, such as an increased risk of heart attack, stroke, kidney failure, blindness, problems with erections (impotence), and even amputation. The topic of diabetic sickness in pregnant women is the principal focus of the investigation that is being carried out here. Perform an investigation of the Pima dataset using machine learning methods in order to determine which type of supervised machine learning, specifically the support vector machine, provides the highest accuracy rate. The three machine learning algorithms that were formerly utilised have been improved as a result of the discovery of confusion metrics and standard deviation inside the algorithms

References

1. *Brijesh Verma and Ashfaqur Rahman (2012) "Cluster- Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning" in IEEE Transactions on knowledge and data engineering.*
2. *Nayer M.Wanas, Rozita A. Dara and Mohamed S. Kamel (2006) "Adaptive fusion and co-operative training for classifier ensembles" in Pattern Analysis and Machine Intelligence Lab, University of Waterloo.*
3. *Yoshua Bengio (2009) "Learning Deep Architectures for AI" in Foundations and Trends in Machine Learning.*
4. *Giorgio Fumera, Fabio Roli and Alessandra Serrau (2008)"A Theoretical Analysis of Bagging as a Linear Combination ofClassifiers" in IEEE Transactions.*
5. *Albert Hung-Ren Ko and Robert Sabourin (2017) "The Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces" in IEEE Transactions.*
6. *H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S.Chen, (2012) "Application of classification techniques ondevelopment an early-warning systemfor chronic illnesses", Expert Systems with Applications, vol. 39, pp. 8852-8858.*
7. *R. Bhuvaneswari and K. Kalaiselvi, (2012) "Naive Bayesian Classification Approach in Healthcare Applications",*
8. *Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, (2012) "A data mining approach for prediction of heart disease usingneural networks, international journal of computer engineering and technology".*

9. *M.A.Nishara Banu and B.Gomathy, (2014) " Disease Forecasting System Using Data Mining Methods".*

10. *Aqueel Ahmed, Shaikh Abdul Hannan, (2012) "Data Mining Techniques to Find Out Heart Diseases", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.*

11. *Ms. Ishtake S.H, Prof. Sanap S.A., (2013) "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research.*

12. *Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, (2012) "A data mining approach for prediction of heart disease usingneural networks, international journal of computer engineering and technology".*

13. *M.A.Nishara Banu and B.Gomathy," (2014) Disease Forecasting System Using Data Mining Methods".*

14. *Aqueel Ahmed, Shaikh Abdul Hannan, (2012)"Data Mining Techniques to Find Out Heart Diseases", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012.*

15. *R. D. Sah and D. J. Sheetalani, (2017) "Review of Medical Disease Symptoms Prediction Using Data Mining Technique," IOSR J. Comput. Eng., vol. 19, no. 03, pp. 59–70, May 2017, doi: 10.9790/0661-190301*

Received 18 October 2023
Revised 30 December 2023
Accepted 20 February 2024