ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse106932020.pdf
https://doi.org/10.30534/ijatcse/2020/106932020

Decision Support System for Predicting Cardiovascular Diseases Using Naïve Bayesian Algorithm

Nilda N. dela Cruz[1], Ricardo Q. Camungao[2]
[1]Master in Information Technology, Isabela State University, Philippines
delacruznilda1315@gmail.com
[2] Doctor in Information Technology, Isabela State University, Philippines
camungaor@yahoo.com

ABSTRACT

The health care field provides enormous quantities of data that contain unseen pattern that can be useful for decisions. It is perplexing to orchestrate in an appropriate manner. Nature of the information association has been influenced because of improper administration of the information. Improvement in the measure of information needs some appropriate ways to concentrate and procedure information adequately and proficiently. This paper intended to develop a Decision Support System (DSS) for diagnosing cardiovascular diseases. The system used data mining technique, the Naïve Bayes Classification algorithm. This paper used Rapid Application Development (RAD) Software Life Cycle in designing and developing the system. The system was simulated, and its performance was evaluated in terms of accuracy using synthetic datasets namely, Cleveland and Statlog. Results showed that the system provided the adequate features for predicting heart diseases with an accuracy of 91% using the Cleveland dataset, 89% using the Statlog dataset and 90% using the combined instances of the two datasets.

.

Key words: Data Mining, Decision Support System, Heart Disease, Naïve Bayes.

1. INTRODUCTION

Every year, 170,000 Filipinos kick the bucket from cardiovascular infections, up from 85,000 over 20 years prior, as per a recent report by the Department of Health. However, without anyone noticing, cardiovascular sickness discreetly sneaks into our homes and medical clinic beds, snuffing the life out of 19 Filipinos every hour[1].

Data mining strategies plays a vital role in healthcare analysis. There were several mechanisms used for managing diagnostic results like decision support systems that are based upon computer may play a vital role. Health care field generates big data about clinical assessment, report regarding patient, cure, follow-ups, medication etc. It is perplexing to orchestrate in an appropriate manner. Nature of the information association has been influenced because of unseemly administration of the information. Improvement in the measure of information needs some appropriate way to concentrate and procedure information adequately and proficiently [2],[11].

Medical data mining has incredible potential for investigating the covered-up designs in the informational indexes of the clinical area. These examples can be used for clinical finding. In any case, the accessible crude clinical information is broadly dispersed, heterogeneous in nature, and voluminous. This information should be gathered in a sorted-out structure. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data [3],[12].

Medical diagnosis is viewed as a significant yet convoluted task that should be executed precisely and proficiently. The mechanization of this framework would be incredibly beneficial. Unfortunately, some doctors do not possess expertise in every subspecialty and moreover there is a shortage of resource persons at certain places. There is a lack of asset people at certain spots. In this way, an automatic medical diagnosis system would likely be exceedingly gainful by bringing every one of them together. Appropriate computer-based information and/or decision support systems can help in accomplishing clinical tests at a decreased cost [3].

This paper intended to develop a Decision Support System (DSS) for diagnosing cardiovascular disease. The system used data mining technique, the Naïve Bayes Classification algorithm. The accuracy, sensitivity, specificity and precision of the said classifier was measured for its efficiency in predicting cardiovascular diseases.

2. METHODOLOGY

2.1 Conceptual Framework

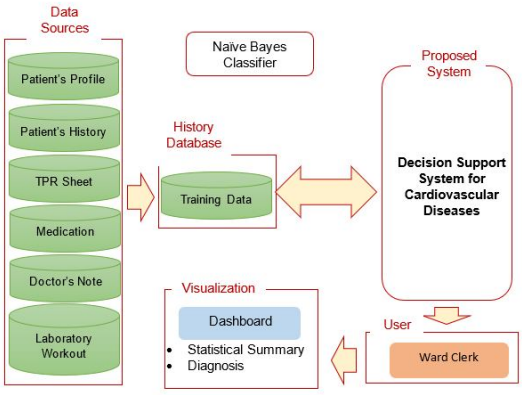

Figure 1: Conceptual Framework of the Research

Figure 1 shows the source of data for this research. The patients' profile and history, Temperature, Pulse, and Respiration(TPR) sheet, medication, Doctor's note and laboratory workout. Variables for the dataset were chosen from these data sources and this dataset served as the historical heart disease dataset for the decision support system. The user who is the ward clerk entered the patient's medical data, then these data would matched the records in the historical heart disease datasets, if the new patient's records would matched in any of the records in the database, then the system would give the predicted heart disease with the help of the Naïve Bayes classifier. The developed system also provides the statistical charts for a given query, like if the user wants to know, how many patients for a given age was diagnosed with a heart disease? What type of heart disease is the highest in occurrence in the patient's database?

2.2 Methods and Tools

This research used Rapid Application Development (RAD) Software Life Cycle in designing and developing the DSS. This model targets at developing the system in a short span of time [4]. The first phase of this model is Analysis and Quick Design. In this phase, finding related studies are done and hardware and software requirements are being identified. System functionalities' conceptualization are also done in this phase. The second phase of the model is the Build phase. In this phase, system user's designs are coded and designed. Chosen classifier algorithm was embedded in this phase. The third phase of the model is Demonstrate, Refine and Testing. In this phase system's functionalities are tested to ensure that every module of the system is working. The last phase of the model is Evaluation. The developed DSS is evaluated based on precision measures; the system's accuracy, sensitivity, specificity and precision. Three synthetic datasets were tested and evaluated.

2.3 Naïve Bayesian Algorithm

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm used conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes[5].

Steps**:**

1. Convert the dataset into frequency table
2. Create likelihood table by finding the probability.
3. Calculate the posterior probability of each class.
4. The class with highest priority probability is the outcome of prediction.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to beat even profoundly advanced characterization techniques. A preferred position of the Naive Bayes classifier is that it requires just a modest quantity of preparing information to appraise the parameters (means and variances of the variables) necessary for classification. Since independent variables are assumed, just the changes of the variables for each class need to be resolved. It very well may be utilized for both binary and multiclass classification issues [6].

Naïve bayes is one of the data mining techniques demonstrating impressive achievement compared to other data mining techniques over different heart disease datasets. Palaniappan and Awang explored contrasting various information mining strategies in the diagnosis of heart disease patients. These techniques included naïve bayes, decision tree, and neural network. The outcomes indicated that the naïve bayes achieved the best accuracy in the diagnosis of heart disease patients. Rajkumar and Reena investigated Naïve Bayes, k-nearest neighbour, and decision list in the diagnosis of heart disease patients. The results showed that the Naïve Bayes achieved the best accuracy in the diagnosis of heart disease patients [7].

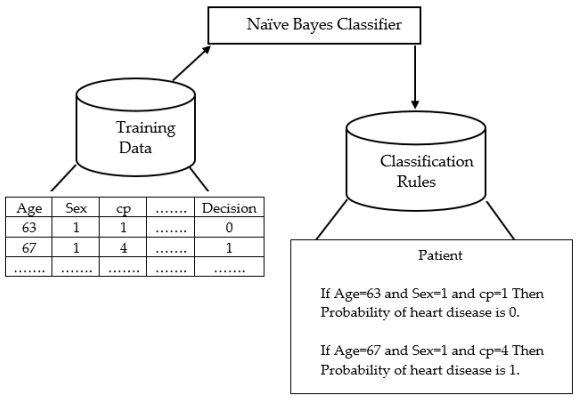

Figure 2: Using Naïve Bayes for Classification

Medical data are entered classification algorithm in order to be learned as shown in figure 2. Mostly, the connection between attributes needs to be found by the algorithm to forestall the result. At the point when another case is shown up the developed classification algorithm is used to classify it into one of the predefined classes. For example, the training set in the medical database would have a lot of important patient data recorded as of now, where the forecast result is whether the patient had a heart disease.

2.4 Data Source

The publicly available heart disease database was used, the Cleveland and Statlog datasets. Fourteen (14) medical data with categorical values are used in testing the system; the age, sex, chest pain type, resting blood sugar, cholesterol, fasting blood sugar, electrocardiographic result, maximum heart rate achieved, exercise induces angina, ST depression, slope, number of major vessels colored by fluoroscopy, defect type and the target value.

Table 1: Description of Datasets

Datasets	No. of records	No. off attributes	With Heart Disease	Without Heart Disease
Cleveland	303	14	139	164
Statlog	270	14	120	150
Synthethic_03	573	14	259	314

Table 1 describes the three datasets to be used in this research. Precision measures such as accuracy, specificity, sensitivity and

precision were calculated to test the efficiency of the chosen algorithm to the developed Decision Support System. Synthetic_03 dataset is the combined dataset of Cleveland and Statlog datasets.

2.4 Performance Measures

Four common performance measures [8] have been used to evaluate the accuracy of the Naïve Bayes Algorithm as shown in table below.

Table 2: Metrics of Measurement

Metrics	Meaning	Equation
Accuracy	Correct Predictions on total	(FP + FN)/ (TP + TN + FP + FN)
Precision	Correct Positive Predictions over Total Positive Predictions	(TP) / (TP + FP)
Sensitivity	Correct Positive Predictions over Actual Positive Values	(TP) / (TP + FP)
Specificity	Correct Negative Predictions over Actual Negative Values	(TN) / (FP + TN)

Where:

True Positives (TP) - the algorithm predicts that the patient has the disease and the patient has the disease.
True Negatives (TP*)* – the algorithm predicts that the patient does not have the disease and the patient is disease free.
False Positives (FP) – the algorithm predicts that the patient has the disease, but the patient is disease free.
False Negative (FN) – the algorithm predicts that the patient does not have the disease, but the patient has the disease.

2.5 Evaluation of Naïve Bayes Algorithm to the developed DSS

In measuring the performance of Naïve Bayes Classification algorithm to the developed Decision Support System of the three chosen synthetic datasets, the researcher adapted the metrics from the paper of Wen Zhu [8] as shown in table 3.

Table 3: Evaluation Measures for Data

Probability Range	Classification
90%-100%	Excellent
80%-89%	Good
70%-79%	Worthless
60%-69%	Not good

When the evaluation result's rate is from 90%-100% then the algorithm's performance is "Excellent" and it means, the algorithm is efficient to use as part of the developed system. 80%-89% means "Good" and it means, the algorithm serves its purpose but not perfectly. When the evaluation result's rate is 79% and below, then the algorithm is unusable for the developed Decision Support System. Since the system is for medical field, the algorithm's evaluation's result must be classified as excellent for accurate diagnosis or else other data mining techniques may considered.

3. RESULTS AND DISCUSSION

3.1DSS for Predicting Cardiovascular Diseases using Naïve Bayesian

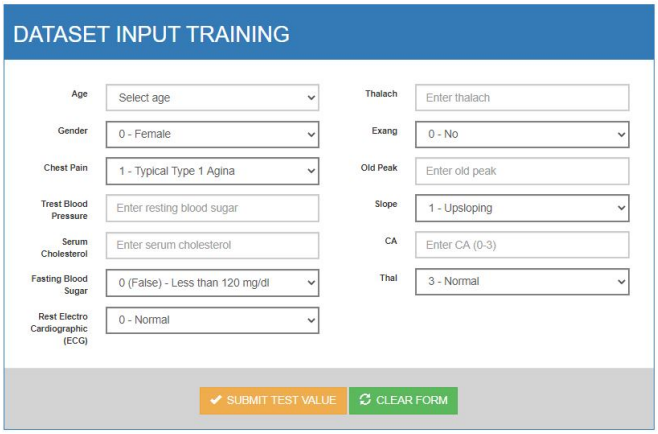

Figure 3: Dataset Input Training Page

Figure 3 shows the page where the ward can enter the medical data of a patient for data testing. Medical data was made categorical for better understanding. Ward can select the values in every combo box, the ward should carefully select values in order to get the accurate prediction result from the developed DSS.

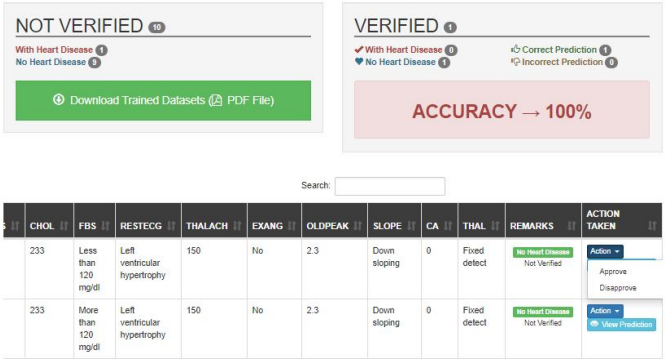

Figure 4: Data Verification Page

Figure 4 shows the page where Doctor's approval is needed if the patient's data did not match to any record in the historical database. If the doctor approved the prediction result, the patient's data will be added to the historical database, else the data will not be stored in the database. This feature of the DSS will let the system's data reserved its integrity since the all predictions are verified by a specialist.

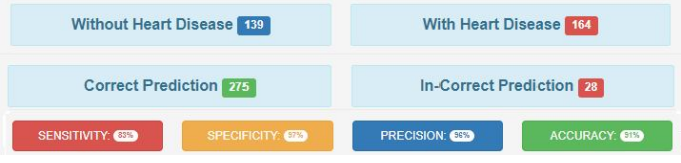

Figure 5: The Accuracy and Frequency Page

Figure5 shows the instances from the chosen dataset, the three synthetic datasets can be viewed in this page but not wholly meaning, the ward can only view the datasets separately. Also, in this page the ward can see the dataset's frequency. It also shows the confusion matrix, sensitivity, specificity, precision and accuracy values of the dataset. Similarly, in this page you can import CSV (Comma-Separated Values) file that stores the instances that can be served as the historical database that can be used for prediction, it should have the same column names and number of columns to be able to store in the database of the developed DSS.

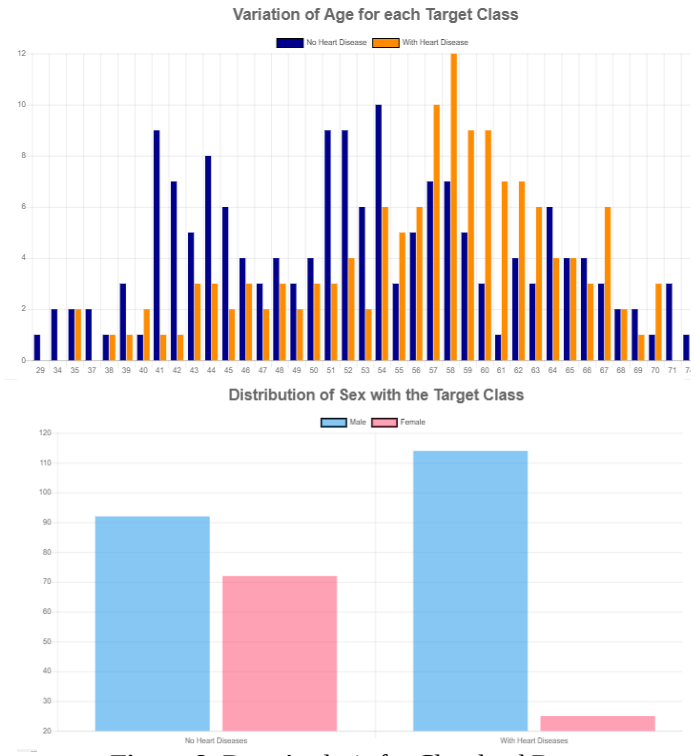

Figure 6: Data Analysis for Cleveland Dataset

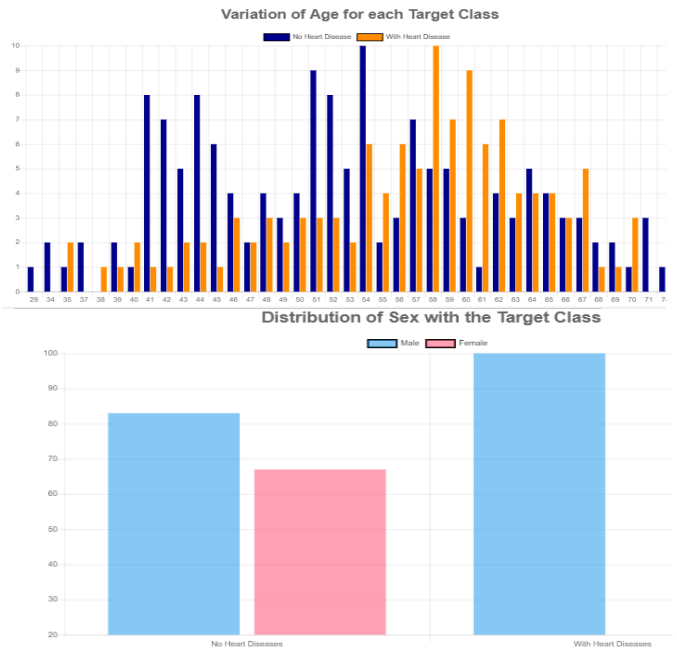

Figure 7: Data Analysis for Statlog Dataset

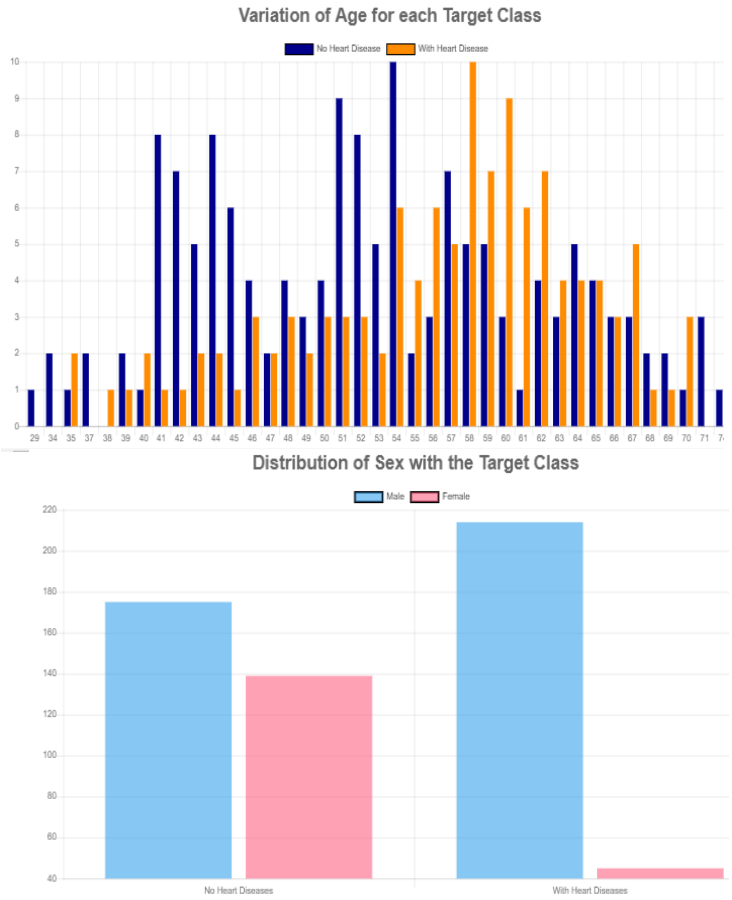

Figure 8: Data Analysis for Synthetic_03 Dataset

Figure 6,7 and 8 shows the data analysis of the three synthetic datasets. This is seen in the Doctor's page of the DSS. The data analysis includes;1) Variation of Age for each Target Class 2) Distribution of Sex with the Target Class 3

From the three synthetic datasets, most people who are suffering from heart disease are of the age 58,57 followed by 60. Majorly, people belonging to the age group of 50+ are suffering from heart disease. There is a greater number of males who are suffering from the disease than females.

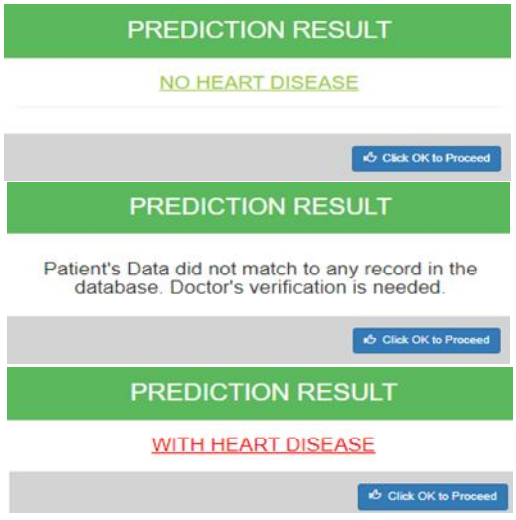

Figure 9: Prediction Result's Page

Figure 9 shows the prediction result's page. If the patient's data is 100% matched correlated to any record in the database, the prediction is with heart disease or no heart disease, else doctor's approval for the data is needed.

3.2 Evaluation Results

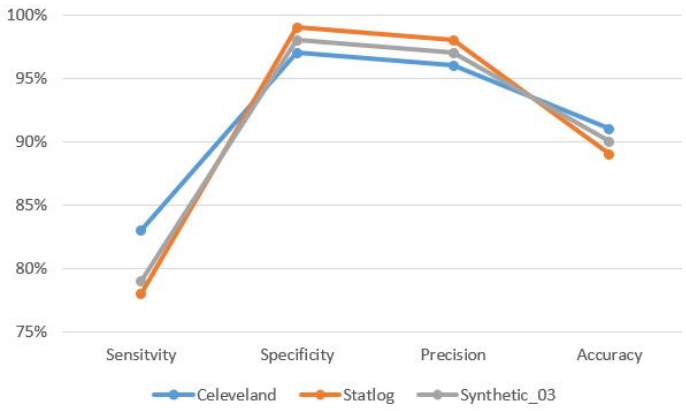

Figure 5: Precision Measures of the three datasets

Figure 10 shows the evaluation result obtained with the use of the developed decision support system and Naïve Bayes classification algorithm.

Cleveland dataset got the highest sensitivity of 83% which means, Cleveland got the highest predictions for patients diagnosed with heart disease over actual number of patients diagnosed with heart disease followed by Synthetic_03 (79%) then Statlog datasets (78%).

Statlog dataset got the highest specificity of 99% which means, Statlog dataset got the highest correct predictions for patients diagnosed without heart disease over actual number of patients diagnosed without heart disease followed by Synthetic_03 (98%) then Cleveland dataset (97%).

Statlog dataset got the highest precision of 98% which means, Statlog dataset got the highest correct predictions for patients diagnosed with heart disease over the total number of correct predictions both diagnosed with and without heart disease followed by Synthetic_03 (97%) then Cleveland dataset (96%).

Cleveland dataset got the highest accuracy of 91% which means, Cleveland got the highest the highest number of correct predictions for patients diagnosed with and without heart disease followed by Synthetic_03 (90%) then Statlog datasets (89%).

4. CONCLUSION

This paper presented the Decision Support System for Predicting Cardiovascular Diseases Using Naïve Bayesian Algorithm that provides functionalities in accepting health data inputs, predicting the presence of heart disease, Doctor's verification for uncertain prediction that caused by unmatched record from historical database and data analysis based on user's queries.

Integrating concept of the Naïve Bayesian Algorithm in the system provides decision support system that can be use as basis for interpretation and decision making.

The evaluation result of the system in terms of accuracy, sensitivity, specificity and precision using three (3) synthetic datasets is "Excellent" which implies that the developed Decision Support System is efficient to use for medical purposes especially for predicting the existence of cardiovascular diseases in patient's data. The developed system can be helpful for medical analysts or practitioners for accurate diagnosis.

REFERENCES

1. A.Jambora. **Cardiovascular disease is still the country's top killer**. Retrieved from http://www.pchrd.dost.gov.ph/index.php/news/library-health-news/4123 cardiovascular-disease-is-still-the-country-s-top-killer-read-more-http-lifestyle-inquirer-net-178609
2. V. Kirubha and S. Manju Priya, **Survey on Data Mining Algorithms in Disease Prediction**, InternationalJournal of Computer Trends & Technology (IJCTT), vol. 38, no. 3, pp. 124–128, 2016.
   https://doi.org/10.14445/22312803/IJCTT-V38P122
3. J.Soni,, Ansari, U., Sharma, D., & Soni, S. (2011). **Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction**. International Journal of Computer Applications, 17(8), 43–48. doi: 10.5120/2237-2860
4. What is RAD Model? Advantages & Disadvantages. (n.d.). available at https://www.guru99.com/what-is-rad-rapid-software-development-model-advantages-disadvantages.html
5. K.Vembandasamy, Sasipriya, R. and Deepa, E. (2015) **Heart Diseases Detection Using Naive Bayes Algorithm**. IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441-444.
6. M.Rambhajani, Deepanker, W. and Pathak, N. (2015) A **Survey on Implementation of Machine Learning Techniques for Dermatology Diseases Classification**. International Journal of Advances in Engineering & Technology, 8, 194-195.
7. M.Shouman, Turner, T., & Stocker, R. (n.d.). **Integrating naive bayes and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients**.
8. W. Zhu, **Sensitivity, Specificity, Accuracy**, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations,Health Care and Life Sciences.
9. R. Camungao, **Decision support system framework with k-means algorithm for faculty performance evaluation rating**, International Journal of Scientific and Technology Research, 2020.
10. R. Camungao, **Slim-tree clustering large application based on randomized search (STCLARANS) algorithm**

simulator, Journal of Advanced Research in Dynamical and Control Systems, 2019.
https://doi.org/10.5373/JARDCS/V11SP12/20193333

11. P. Santhi, Deeban N., Jeyapunitha N.,Muthukumaran B., Ravikumar R.,**Prediction of Diabetes using Neural Networks**, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, No. 02, 2020, pp 985-990.
https://doi.org/10.30534/ijatcse/2020/13922020
12. A.Rhagini, T. Dhanush Narayanan, M. Santhosh, C. Sharmila, G. Swetha. **An Unsupervised approach for Predicting the Breast Cancer using K-Means with Compound Feature Generation**. Vol. 9, No. 02, 2020, pp 930-934.
https://doi.org/10.30534/ijatcse/2020/04922020