e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

REAL TIME MACHINE LEANING DETECTION OF HEART DISEASE USING BIG DATA APPROACH

Ramya S*1, Mr. M. Vijayaraj*2

*1Master Of Engineering, CSI Engineering College, Tamil Nadu, India.

*2BE ME M.DIV, Assistant Professor, Computer Science Engineering, CSI Engineering College, Tamil Nadu, India.

ABSTRACT

According to a recent report by the World Health Organization, heart-related disorders cause 17.9 million deaths per year and are on the rise. The purpose of this study is to analyses different data mining approaches, particularly Naive Thomas Bayes, Random Forest Classification, call trees, and Support Vector Machines are used to predict cardiopathy using a qualified dataset that includes various parameters such as gender, age, type of pain, blood sugar, and pressure level. Finding connections between the dataset's many properties using high-quality data processing techniques is a key component of the research, which also involves treating the attributes appropriately to forecast the likelihood of a cardiopathy. These machine learning approaches forecast illnesses with a high degree of accuracy in a shorter amount of time, which can save the loss of important lives around the globe.

Keywords: Naive Bayes, Random Forest, And Support Vector.

I. INTRODUCTION

Biggest problems facing humanity on the planet is health. Based on the report of the World Health Organization (WHO), having one's own health is a basic human right .Thirty-one percent of all deaths worldwide are related to heart disease. Due to the lack of diagnostic tools, physicians, and alternative resources capable of accurately diagnosing and treating organ-affected patients, detecting and treating cardiovascular diseases is extremely difficult, especially in underdeveloped countries. In light of this problem, engineering and machine learning methods are now being used to create software that helps doctors detect cardiovascular diseases early**.** Mortality rates will be reduced through early detection of disease and identification of people at risk of cardiovascular disease. To extract important patterns and data, medical knowledge uses medical data processing techniques. Redundancy, diversity, consistency, and a strong relationship with time are all characteristics of medical data. Unfortunately, the problem of mishandling large amounts of information is becoming a serious problem. The methods and technologies for turning these bodies of knowledge into data useful for decision making are provided by data processing .This cardiovascular disease identification system will help cardiologists work faster The main goal of this study was to use several machine learning algorithms on a validated dataset to predict whether a patient has heart disease or not .perfecting the relationship between completely different characteristics. The following sections provide instructions for installing Ettercap, Wireshark, and Urlsnarf on Linux operating systems.

II. METHODOLOGY

Method and analysis which is performed in your research work should be written in this section. A simple strategy to follow is to use keywords from your title in first few sentences.

Hardware

1. Windows 7,8,10 64 bit
2. RAM 4GB

Software

1. Data Set
2. Python 2.7
3. Anaconda Navigator

Python's standard library

- Pandas

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

- Numpy
- Sklearn
- Seaborn
- Matplotlib

Existing system

Remote mobile health monitoring is already acknowledged as more than just a possibility. Healthcare big data (HBD) presents numerous obstacles for applications at every stage, including data collecting, data management, data integration, data analysis, and pattern interpretation. In the parallel processing computing model, there are numerous issues with analytical complexity and data scalability. They are ineffective at accurately predicting cardiac disease.

Drawbacks

The technology and methods to turn mountains of information into useful data for decision making are provided by data mining. This analysis compares different machine learning methods such as Support Vector Machines, Decision Trees, Random Forests, and Naive Bayes for predicting heart problems. To predict heart problems, naive scientists used chance; SVM is used for classification and regression; Random forests work with different calling trees.

Limitation

Some methods are only suitable for small data. Some classification combinations fit the data set better than others, depending on the combination. Some methods cannot be used to capture real-time database implementations.

Proposed System

The proposed system detects heart disease using Big Data method. As a concept-driven element in the IT field, machine learning helps humans predict data. The challenge is that creating a predictive model requires the use of multiple tools and programming languages. However, choosing the right technology or tool always helps customers create effective prediction models, delivering more accurate results. Therefore, choosing the right tool is an important responsibility. Finding the right tool for the right business is challenging. One of the problems that arises is complexity. The same can be said about the multifaceted nature of information that too complex mapping, as well as ease of use, can cause unacceptable reactions. The necessary resources may not be enough.

Block diagram

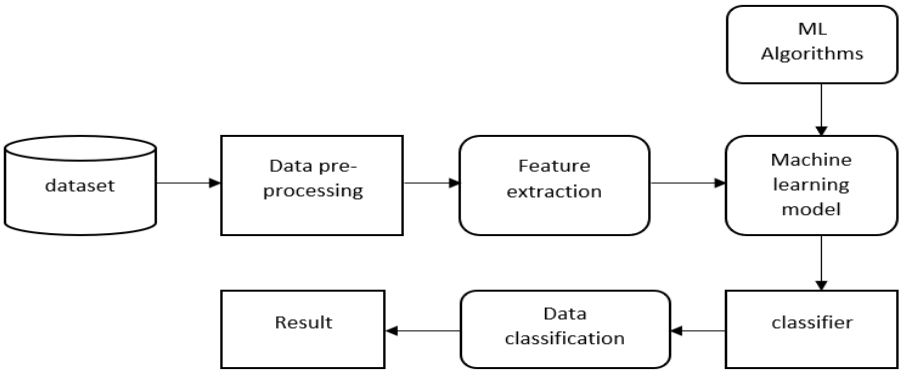

Figure 1: Block diagram

Features

The primary objective of this study was to construct a predictive model for anticipating the onset of heart disease. Furthermore, the study aimed to identify the most effective patient triage approach for heart disease detection. The research employed a comparative analysis employing diverse classification algorithms across various evaluation stages, making it a well-justified endeavor. Despite the widespread utilization of machine learning techniques, the significance of predicting heart disease necessitates the utmost precision in this endeavor.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Advantages

- Significant strides have been taken in categorizing the behavior of cardiac disorders.
- Multiple machine learning models have been employed to assess time complexity and precision, yielding a variety of metrics.
- A range of machine learning techniques has demonstrated exceptional accuracy in their outcomes
- Machine learning models have the capability to forecast high-risk factors in advance

III. MODELING AND ANALYSIS

DATA FLOW DIAGRAM

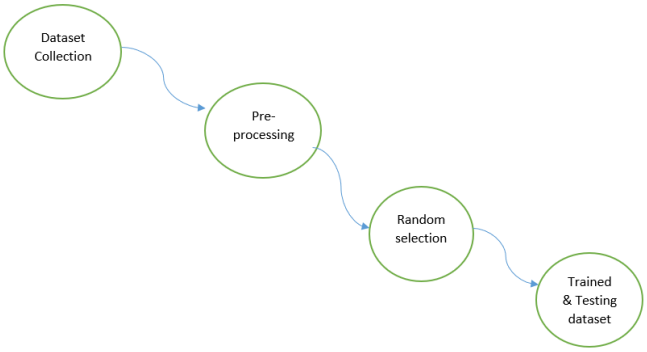

Figure 2.1: Data Flow Diagram (Level 0)

DATA COLLECTION

The practice of obtaining information from numerous sources is known as data collection. Developing machine learning models. Information needs to be stored in a way that is appropriate to the problem. At this stage, the dataset is converted into an understandable format so that machine learning models can use it.

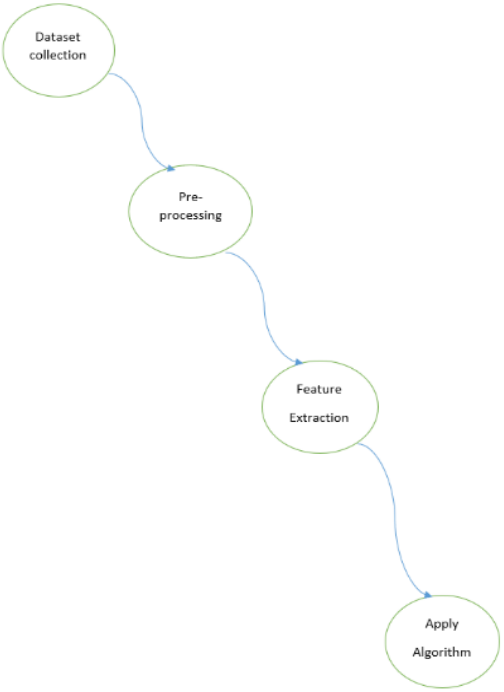

Figure 2.2: Data Flow Diagram (Level 1)

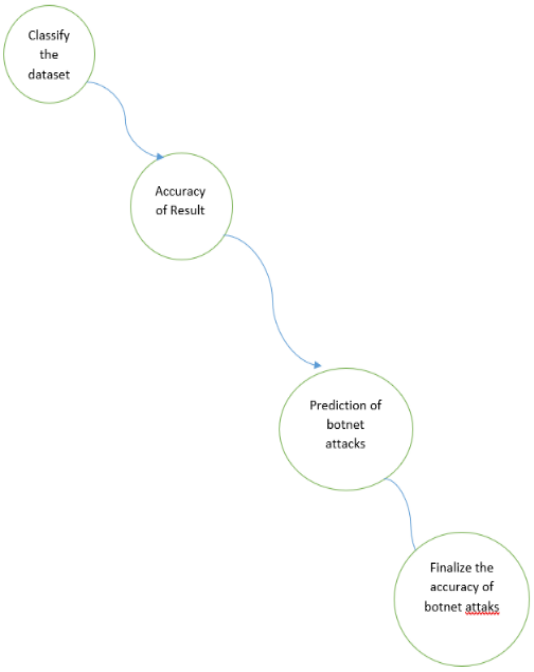

Figure 2.3: Data Flow Diagram (Level 2)

DATA PRE-PROCESSING:

The information could be in a relational database, but you prefer a flat file Cleaning is the process of removing or replacing missing data. In some cases, the data may not be sufficient, and the information you think you need to solve the problem may not be available. These events may need to be remove You can have access to much more carefully selected data than you need. Algorithms may take longer to run on larger amounts of data, and their computational and memory requirements may also increase.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

FEATURE EXTRACTION

The Attribute Reduction Process which is feature extraction .Feature extraction actually modifies the attributes, as opposed to feature selection, which ranks existing attributes based on their predictable relevance. The original attributes are linearly combined to produce modified attributes or features. Finally, the Classifier algorithm is used to train our models .We use the classification module from the Python Natural Language Toolkit library .We use the obtained dataset .The models will be evaluated using the remaining labeled data we have. The pre-processed data is classified using several machine learning methods.

EVALUATION MODEL

The process of model development encompasses the crucial step of evaluation. This step helps identify the model that most accurately represents our data and predicts its future performance. It's considered unacceptable to assess a model's performance solely using training data. Relying on training data alone can lead to over-optimistic and overfit models, which might not generalize well to new data. Data scientists employ two common techniques, namely holdout and cross-validation, to evaluate models effectively. Both approaches involve the use of a separate test set that is not seen by the model during training. This helps prevent overfitting. To gauge the performance of each classification model, the average performance across multiple evaluations is calculated. This average performance provides a more robust assessment of the model's capabilities. The results are often represented graphically to provide a clear and intuitive understanding of how data is classified. Accuracy, in this context, refers to the rate of correct predictions for the test data. It is calculated by dividing the number of correct predictions by the total number of predictions made, offering a straightforward measure of model performance.

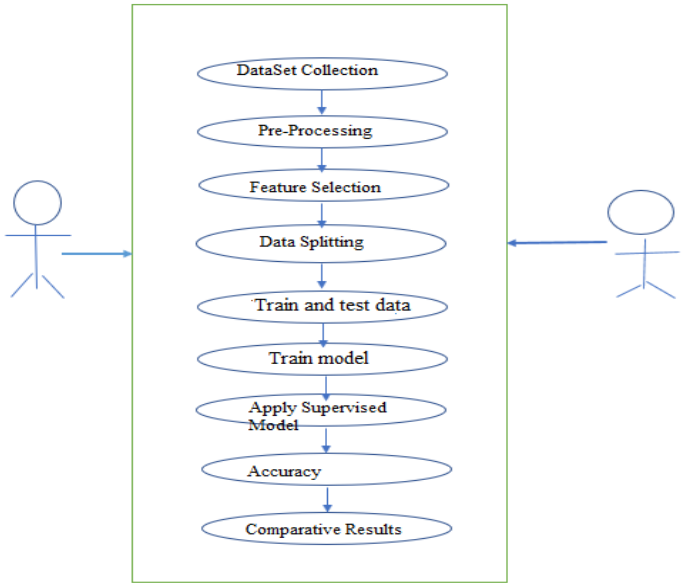

Figure 3: Use case diagram

• A standard language called UML is used to define, visualize, construct, and describe the artifacts of software systems.

• The Object Management Group (OMG) is the organization that developed UML, and OMG received the draft UML 1.0 definition in January 1997.

Figure 4: Class diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

The basic building block of the object-oriented model is the class diagram.

It is used to model application systems in a largely conceptual manner and translate the models into programming code for detailed modelling

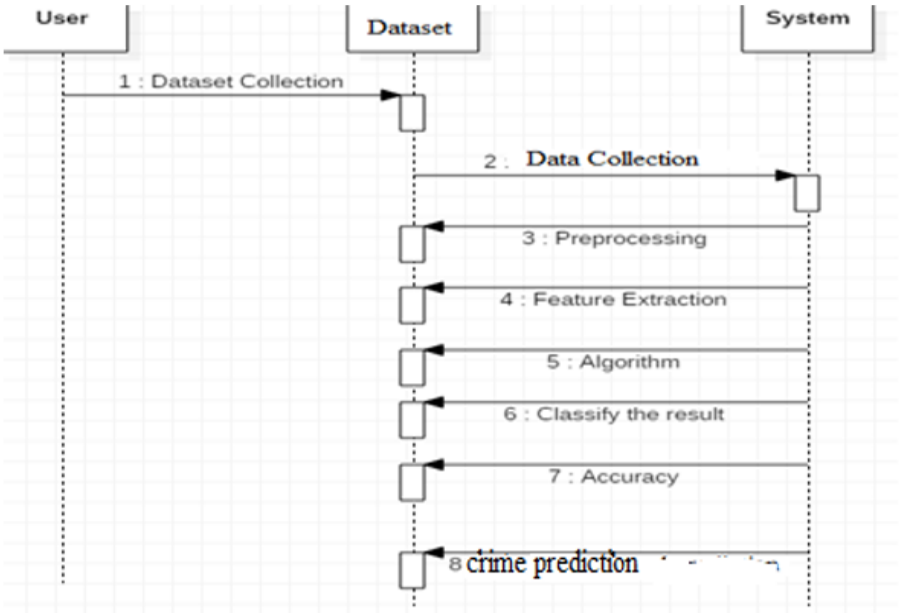

Figure 5: Sequence diagram

Sequence diagram Represents time vertically and objects participating in interaction horizontally. Each use case describes a specific behaviour, possibly with variations, that the subject can perform in conjunction with one or more actors.

Use cases describe the proposed behaviour for the subject without mentioning its internal structure.

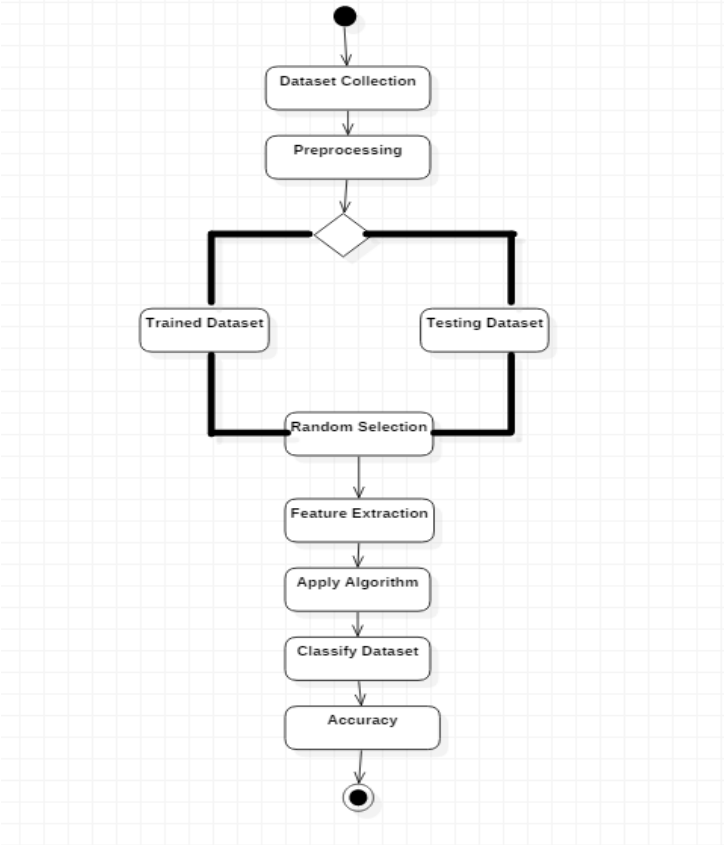

Figure 6: Activity diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Activity diagrams serve as graphical representations of workflows involving sequential activities and actions within a system. These diagrams offer the flexibility to incorporate choices, iterations, and concurrency in a visual manner. In the context of the Unified Modeling Language (UML), activity diagrams are a valuable tool for illustrating the operational and business workflows of various system components. They provide a comprehensive depiction of the overall control flow within a system or process.

DATA ATTRIBUTES

The data we used came from records of the presence of heart disease in patients. Dataset We found the dataset used in our book from Kaggle. The data set we use in the thesis has a total of 14 columns and 303 rows.

Row

It has a total of 303 rows which means we have with us data of over

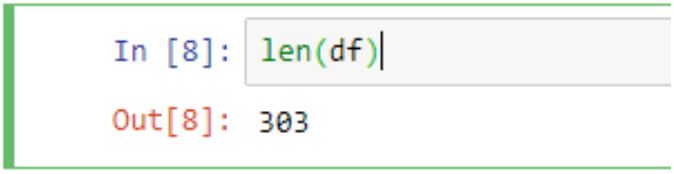

Figure 7: Row

Column

You have a dataset or a data table with 14 columns, each column representing a different dimension or variable.

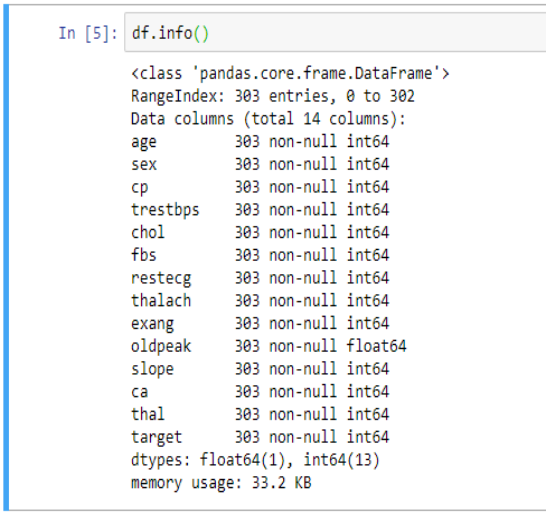

Figure 8: Column

IV. RESULTS AND DISCUSSION

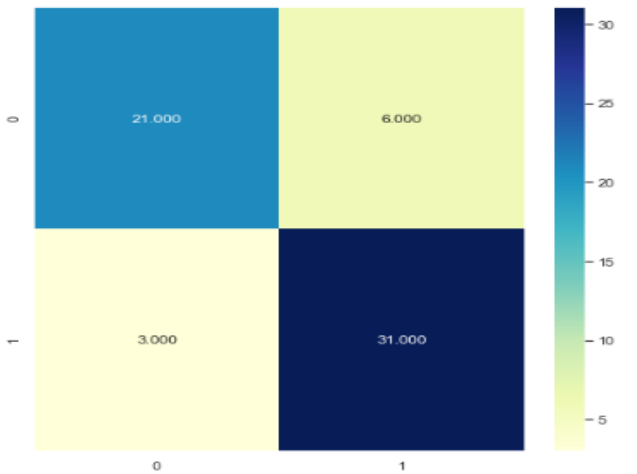

Figure 9.1: Naïve Bayes

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

In this figure 9.1 We compare the confusion matrix with the naive Bayes method

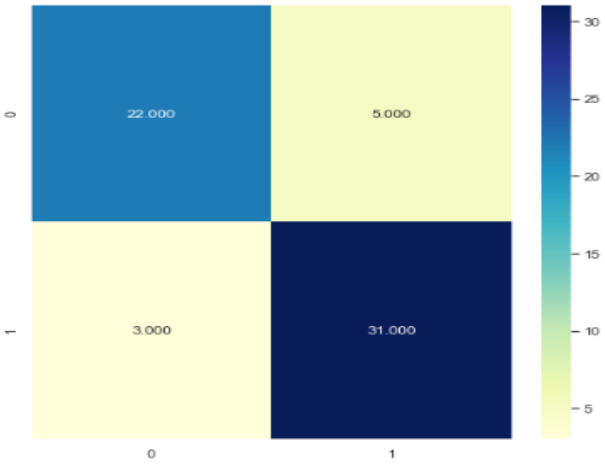

Figure 9.2: Random Forest Classifier

In figure 9.2 we are comparing the confusion matrix with random forest classifier

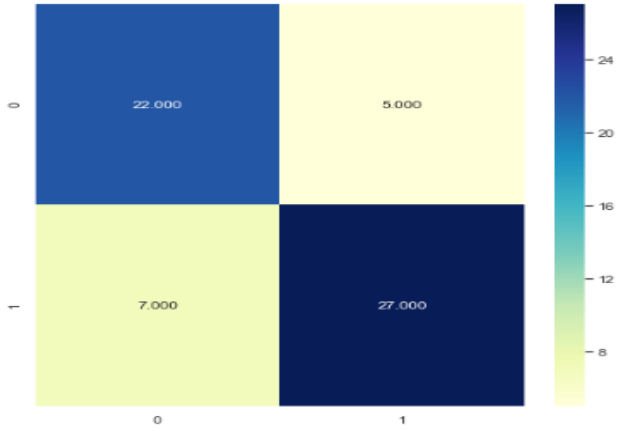

Figure 9.3: Decision Tree Classifier

In figure 9.3 we are comparing confusion matrix with decision tree classifier

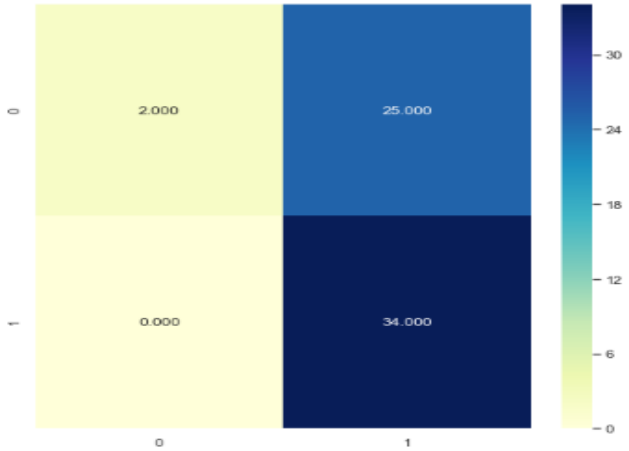

Figure 9.4: SVM

In figure 9.3 we are comparing confusion matrix with svm.

Comparative Result with Different Algorithm :

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

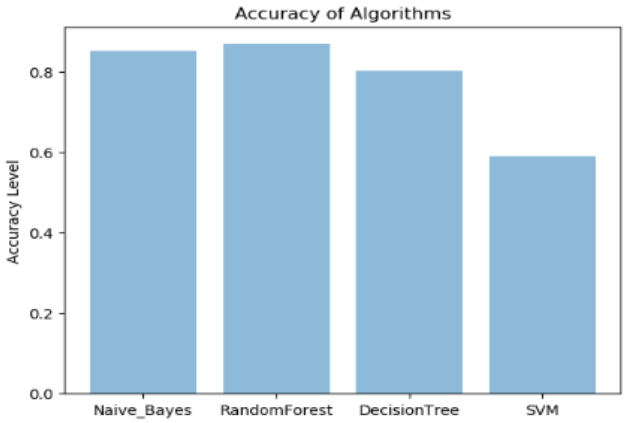

Figure 9.5 Accuracy of all Algorithm

In figure 9.5 is overall result of all the 4 comparison

Table 1. Comparison of Algorithm of all 4 cases

No of runs	Run 1	Run 2	Run 3	Run 4	Run 5
Naïve Bayes	0.85	0.85	0.85	0.85	0.85
Random forest	0.83	0.81	0.83	0.81	0.88
Decision tree	0.78	0.78	0.75	0.77	0.77
Svm	0.59	0.59	0.59	0.59	0.59

V. CONCLUSION

We aimed to assess the predictive capabilities of various machine learning algorithms in determining whether an individual is at risk of developing heart disease based on a combination of unique factors and symptoms. The primary objective was to compare the accuracy of these algorithms and gain insights into the factors contributing to variations in their performance. conduct our analysis, we partitioned the dataset into training and testing subsets using a 10-fold cross-validation approach. Our dataset consisted of 303 instances from the Cleveland dataset on cardiac disorders, and it comprised 13 distinct attributes. We evaluated the accuracy of our models using four different machine learning algorithms. Upon completing the implementation phase, our findings revealed that the Gaussian Naive Bayes and Random Forest algorithms achieved the highest accuracy, reaching an impressive 91.21%. Conversely, the Decision Tree algorithm yielded the lowest accuracy at 84.62%. It's important to note that these results were specific to our dataset and conditions .We acknowledge that different algorithms may perform better in alternative scenarios and with diverse datasets. Additionally, there is potential for further enhancing accuracy, but this would come at the cost of increased processing time and system complexity due to the need to handle more data. After carefully considering these factors, we made a pragmatic decision regarding the most suitable course of action for our particular circumstances .

VI. REFERENCES

[1] Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01

[2] Prerana THM, Shivaprakash NC et al “Prediction of Heart Disease Using Machine Learning Algorithms-Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS”, Vol 3, PP: 90-99 ©IJSE, 2015.

[3] Salam Ismaeel, Ali Miri et al “Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis”, IEEE Canada, DOI:10.1109/IHTC.2015.7238043, 03 September 2015.

[4] F Brain Boudi’Risk Factors for Coronary Artery Disease’, 2016. [Online]Available: https://emedicine.medscape.com/articl e/164163-overview.

[5] National Health Council,’ Heart Health Screenings’,2017.[Online]Available: http://www.heart.org/HEARTORG/Conditions/Heart-

[6] ScikitLearn,'MLPCLassifier',Available:http://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifi er.html

[7] Prediction System for heart disease using Naïve Bayes *Shadab Adam Pattekari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering.

[8] Comak E, Arslan A (2012) A biomedical decision support system using LS-SVM classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases. J Med Syst 36:549–556.

[9] Ahmed Fawzi Otoom, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye and Mohammad Ashour(2015) Effective Diagnosis and Monitoring of Heart Disease ISSN: 1738-9984 IJSEIA.