COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 3, Issue 1, Jan-Jun 2023, pp. 12-22
DOI: 10.17492/computology.v3i1.2302
https://www.journalpressindia.com/cjacsit

Effective Detection of Heart Disease Symptoms Using Machine Learning

Gunji JaiSadhashiva, Shaik Mohammad Mohaboob Shareef**, Devarakonda Aditya***, Leela Venkat Muppavarapu****, Senthil Athithan***** and B. Suneetha*******

ABSTRACT

Cardiac disease is leading around the globe for deaths, although early detection and prevention can improve survival rates. Using machine learning, various Activation Functions create new models and make predictions using the data they collect. In previous studies, the detection of disease signs was accomplished through the application of machine learning algorithms. Several pieces of paper were examined. Dr. Mohan was able to predict heart disease by analyzing blood pressure. It was a supervised machine-learning technique that he called random forest. In this study, principal component analysis as well as five different techniques were used. Using the methods described above, we projected that the Random Forest technique would provide the highest accuracy.

Keywords: Feature bagging; Classifier; Supervised learning; Activation function; Training dataset.

1.0 Introduction

Cardiac disease affects people of all ages, from children to the elderly. Traditional ways of diagnosing cardiac disease first produce human repercussions. Nevertheless, these methods are the most often used. The purpose of this article will be to construct an application to treat heart disease effectively [1]. These algorithms can assess hidden patterns in addition to several medical areas associated with heart disease thanks to the application of Artificial Intelligence (AI) and data visualization techniques. These kinds of technological advancements might very well be the secret to the professional's enormous success as well as to improved outcomes in linked algorithms like navy bases and random forests [2]. The information processing about heart disease will be significantly aided by the well-accurate data sets' ability to develop an effective training model. All healthcare institutions and research grads must provide precise patient data to contribute the most effective robust algorithms, will be included in the data sets that will be created.

**Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: destroyershiva123@gmail.com)*
***Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: suhanshaik2717@gmail.com)*
****Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: adithyadevarakonda777@gmail.com)*
*****Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: leelavenkatmuppavarapu@gmail.com)*
******Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: senthilathithan@hotmail.com)*
*******Corresponding author; Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (E-mail: suneethabulla@kluniversity.in)*

Implementing algorithms for machine learning will also ensure that cardiovascular disease remains the most prevalent health problem the world is currently confronting. As we compared the conditions before and after COVID-19, we discovered that the epidemic claimed many people's lives. Therefore, this provides an early indication of cardiac disease [3]. When it comes to healing and making forecasts, companies must employ technologies like these [4].

The medical term for heart sickness is "cardiac disease." Many diagnostic and prognostic tools are available on our globe; however, these tools come at a very high price. Some people live in poverty and cannot afford such measures [5].

There is a pressing need for reform in the medical field due to the proliferation of clinics, hospitals, and other facilities equipped with various diagnostic tools for cardiac disease. The reason for this is that the generation has been upgraded and altered. Variety approaches to determine whether people have cardiac disease. Support vector machines and decision trees are examples of several kinds of algorithms. The priority that should be placed on a person's health must be balanced. One of these conditions is heart disease, which typically strikes between the ages of 60 and 65. However, given the current state of affairs, heart disease can manifest in its early phases [6]. There is a probability of developing heart disease earlier if certain risk factors are present, such as high blood pressure, diabetes, an unhealthy diet, etc. There are several distinct forms of cardiac disease.

The Random Forest, the Decision. Utilizing these methods allows for the early detection of cardiovascular disorders [7]. When we consider the rapid expansion of machine learning and the various methods it employs in the industry, every clinic needs to use, implement, and develop strategies for predicting heart illnesses [8].

Svm is a popular and significant code for machine training. Data linear and logistic regression [12], as well as the identification of data outliers, are the primary applications for this tool. This code can be used for many purposes, including categorizing images and spam titles [9]. High Efficiency in Applications is due to its ability to accommodate and manage multi-dimensional data through processing. An algorithm separates unlabeled data into distinct clusters with similar characteristics. This algorithm applies to a broad range of clusters.

1.1 Different heart diseases

Include Failure of the heart muscle to pump blood effectively, one of the conditions that fall under the umbrella of the category known as heart disease. The most common causes of heart failure are cardiovascular attacks and excessive blood pressure [6]. Valve disease is a form of heart disease that can manifest itself when one or more than one of the veins in the heart cannot function properly. Pain in the chest, lightheadedness, and low or high blood pressure are warning signs of this condition. Aneurysms of the Aorta: An aneurysm appears like a balloon in the afrothere. It is responsible for transporting and the aorta. This disease manifests If Heart fails to pump blood effectively or becomes clogged. The aortic is a collection of conditions that all work together to lower the blood's resistance to illness and contribute to high blood pressure [10].

1.1.1 Problem statement

Primary is to provide evidence that heart disease can be forecast using a variety of coding models mainly of models. Heart disease is becoming increasingly common in younger people as a result of an increase in the intake of alcohol and smoking, as well as an increase in the consumption of foods

high in cholesterol. Therefore, to address this issue, specialists in machine learning have developed several reliable algorithms that can accurately detect the presence of cardiac disease efficiently.

1.1.2 A survey of the relevant literature

A significant aspect of the Research process is the review of the relevant literature. Predicting Algorithms Authors who try to predict an accurate outcome frequently utilize prediction algorithms. Within the scope of this investigation, we went through approximately 20 publications and uncovered a variety of datasets [11]. Both of these approaches were combined. With the help of the Kaggle Dataset and the K-means Algorithm, Boujraf made an accurate prediction regarding heart disease. The additional accuracy that was generated was 97.24%. To diagnose heart disease, Hamida made use of an EMR dataset [12]. The convolutional neural network proved to be the most effective algorithm. A Gogoi performed a prediction on a Clinical Dataset by using X-rays and an algorithm known as a Decision Tree. The program accurately identified Heart disease 95.47% of the time.

2.0 Supplementary Materials and Method

In the course of the inquiry, information on coronary heart disease accessible to the general public is used [13]. This collection has nearly 500 entries, each containing an identifier: Age, Gender, Height, Weight, Cholesterol, Glucose, Smoking Status, Alcohol consumption, Cardiac Activity. PYTHON COMPILER, KERNEL, JUPYTER NOTEBOOK.

3.0 Data Set

Age, Height, Weight, and Cardio are All Included in the Description of the Dataset. The number of days represented the age in the dataset that was included. The unit of measurement for blood pressure is called mm Hg. Cholesterol is a factor in a dataset that determines whether a person is considered normal, above average, or considerably above normal for that particular factor. Smoking is a component in the dataset, and whether or not a person smokes can be used to decide between them. The statistics are included in the cardio. The "cardio" variable will determine whether or not the subject has an illness.

3.1 Feature selection

Image processing's feature selection variables entered into a model [Fig 1]. This method is called "feature selection." During this stage of the process, the model selects data that is accurate and relevant [14]. In this Feature Selection exercise, one of the primary focuses will be reducing noise, which is also an essential component process [15-16].

3.2 Algorithmic detection

Regarding procedure, we used Random Forest and a Decision Tree to predict heart disease. The program will gather the data for the input and then make an accurate prediction of what the output will be. When new data is input into these algorithms, they can learn and improve their performance by optimizing themselves [17-19].

4.0 Methodology

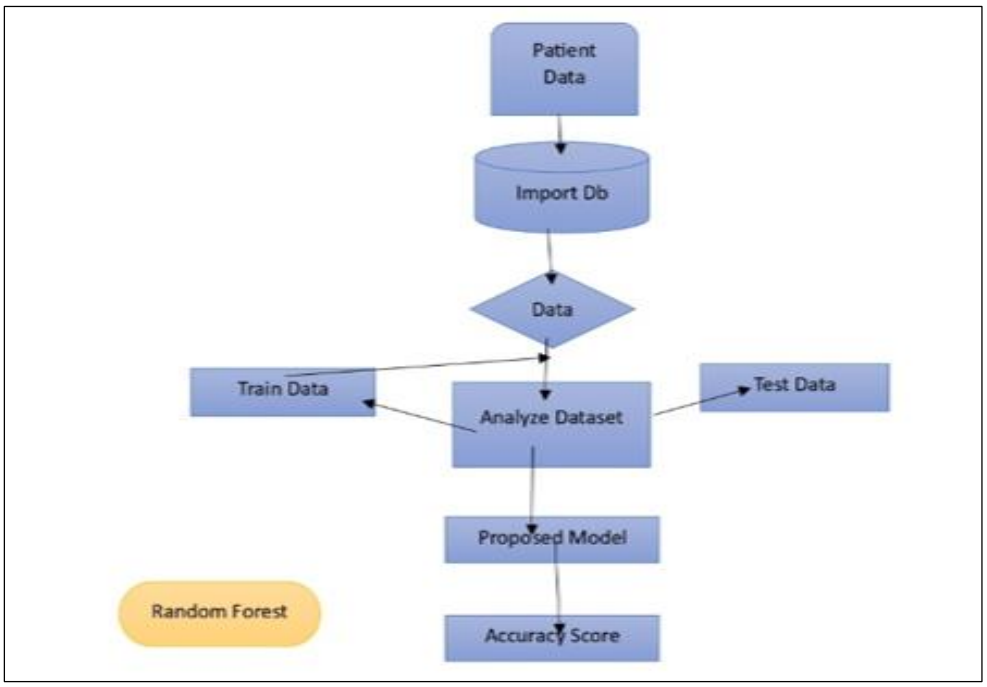
Random Forest and Decision Tree are two algorithms used in this study to demonstrate the analysis. For the Random Forest Mechanism to function correctly, many Decision Trees need to be

built. When all of the decision trees from the random forest are combined, a final prediction of the accuracy of the random forest will be created. On average, this accuracy will be determined. The decision tree will function by dividing the datasets into several smaller ones, which it will continue producing until the data is considered normal.

5.0 Comprehensive Study

During the course of the work that we did for our Research project, we carried out a study, and while we were doing it, we were aware of certain shifts in the manner in which various algorithms were applied to the dataset. During the process of carrying out the execution of the decision tree, we obtained the precision of the depth level as the output as well as the dimensions of the dataset. An accuracy score of 84.70 was determined to be achieved overall. When working on Random Forest, a complex algorithm called Sklearn was utilized to load and import the data set. This was done in parallel with the work on Random Forest. The utilization of this approach resulted in an accuracy of 86.82, which was generated. After we implemented it, it was 82.80 percent. It does this by working on the classifier. The Support Vector Machine is an improvement on the Classifier that can tackle a wide variety of issues in logistic and linear regression. We were able to produce a one-of-a-kind ID when dealing with K-means clustering by making use of the random state, the number of clusters, and the k-means Label. When we carried out the method, the level of accuracy that we achieved.

Figure 1: Proposed Model



We realized that over-fitting is an issue whenever a model is detected or not. The Random Forest technique utilized for classification, regression, while the decision tree is more effective as a classification tool. The following is a breakdown of the accuracy produced by the two algorithms: The first thing you need to do is import the Libraries into the new notebook or project. These libraries include

the math library as well as Seaborn, which is a library that is used for graph plots. NumPy is an essential module in the Python programming language that provides support for mathematical operations and matrices. The Python Software Foundation developed NumPy. As the [**Figure1**] describes The efficient library known as Panda can be utilized to manipulate data, which is a task that can be completed. Using Python, one can plot graphs and charts based on a specific dataset. This is possible thanks to the software. Check if any null values are present in the given dataset.

6.0 Analysis

Analyzing a variety of Medical Record Datasets and algorithms to predict heart disease was a team effort that included all of us—the most effective algorithms to reinforce the learned data and generate an accurate output from the dataset. The result of the performed code will be considered as scores for both the random forest and the decision tree [**Table 1**]. The more accurate and sophisticated the model, the greater the precision with which it can identify cardiac problems [**Figure 1,2**].

6.1 Confusion matrix

Confusion matrices are matrices that are developed to predict the classification and performance of models that are generated from datasets. It is possible to figure out by utilizing only the variables from the dataset that are accurate. The confusion matrix allows for the prediction of certain types of errors. The confusion matrix comprises two dimensions: one represents the expected values, and the other represents the actual values.

False Negative (FN), False Positive (FP), and True Negative (TN) parameters were inserted into the dataset. The Dataset received TN, TP, FN, and FP parameters [**Table1**].

True Negative: It establishes that the model does not have a forecast and that it has not found any real values.

True Positive: It is a comparison between the values that were expected and those that were actually obtained.

False Negative: This is an example of a type 2 error, and it demonstrates that the model does not have a prediction or a prediction with actual values.

False Positive: It establishes that the model was accurate in its predictions, despite the fact that real values are not observed. This is what's known as a kind 1 error.

Precision: It will define the right forecast as well as the total prediction that the model has made. It will carry out the operation on a metric of the dataset.

Precision:$\frac{TP}{TP+FP}$

Recall: The dataset model is used to construct almost two different classes for the calculation of recall: true positive and false negative.

Recall:$\frac{TP}{TP+FN}$,

F1Score Combines accuracy with recall, making it more useful than accuracy for classes with an uneven distribution of students.

F1 Score:$\frac{2TP}{2TP+FP+FN}$

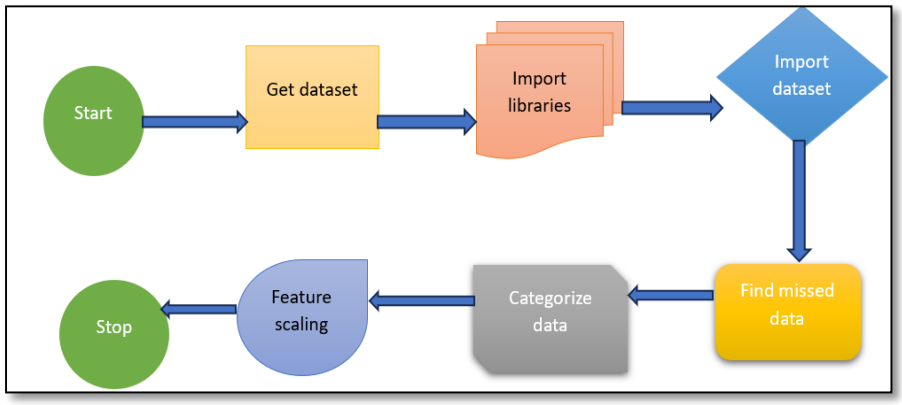
Accuracy: The concept of accuracy, which is determined by the total number of examples to be classified, is the driving force behind the confusion matrix.

Accuracy**:** $\frac{TP+FN}{TP+TN+FP+FN}$

As [**Figure 2**] shows the, A flowchart illustrates the procedural steps involved in the process of gathering data for feature scaling

Figure 2: Flow of Execution by Feature Scaling

.

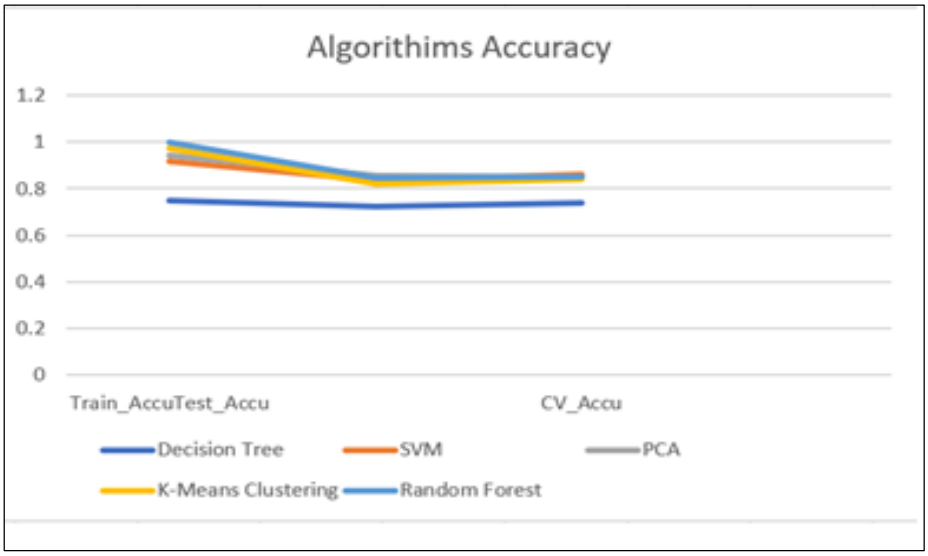


As [**Table 1**] shows It depicts the training, testing, and cross-validation accuracy metrics for evaluating the performance of machine learning algorithms on a dataset.

Table 1: Accuracy of Datasets

Algorithm	Train_Accu	Test_Accu	CV_Accu
Decision Tree	0.7512	0.7222	0.7372
SVM	0.9187	0.8333	0.8613
PCA	0.9426	0.8556	0.8518
K-Means Clustering	0.9761	0.8222	0.8423
random forest	0.9999	0.84444	0.8518

Figure 3: Algorithm Accuracy



Among all the algorithms considered, Random Forest achieved the highest accuracy, while Decision [**Figure 3**]Tree exhibited the lowest accuracy, with the remaining algorithms falling in between. The comparison showcases the original accuracy of the dataset and the accuracy achieved with PCA applied to it [**Table 2**].

Table 2: Principal Component Analysis

Algorithm	Original	With PCA
random forest	0.8662	0.8222
Decision Tree	0.847	0.8111
SVM	0.828	0.8444

This Plot is Relates to the Type of Train and Test and CV Accuracy and the Following is the Depth the Accuracy is High up to 1.00 Compared all other it is Better Fit Model [**Figure 4**]

Figure 4: Random Forest

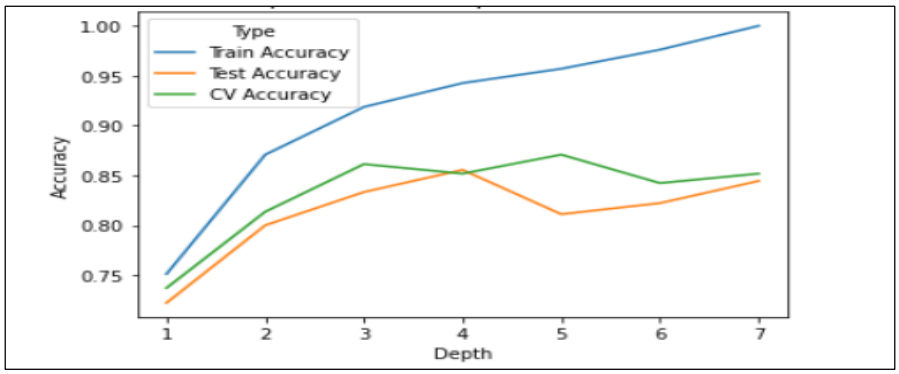
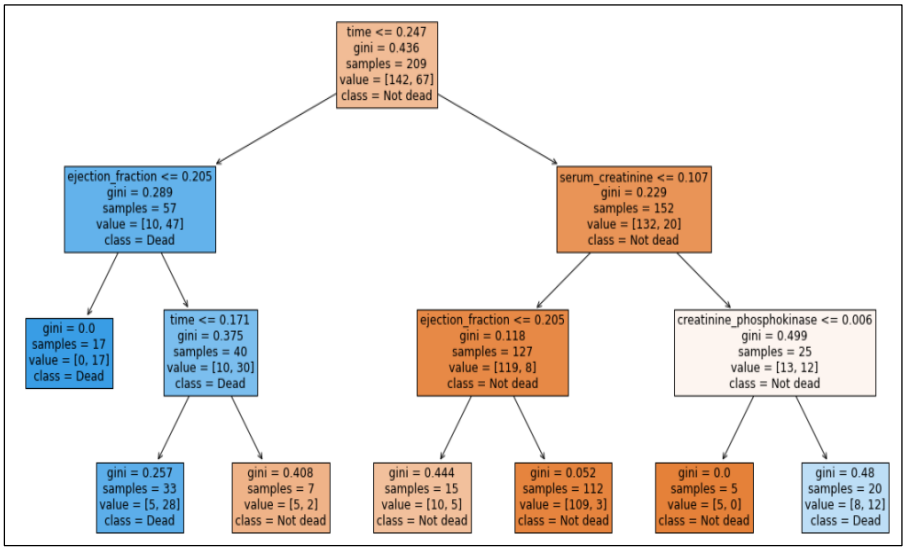


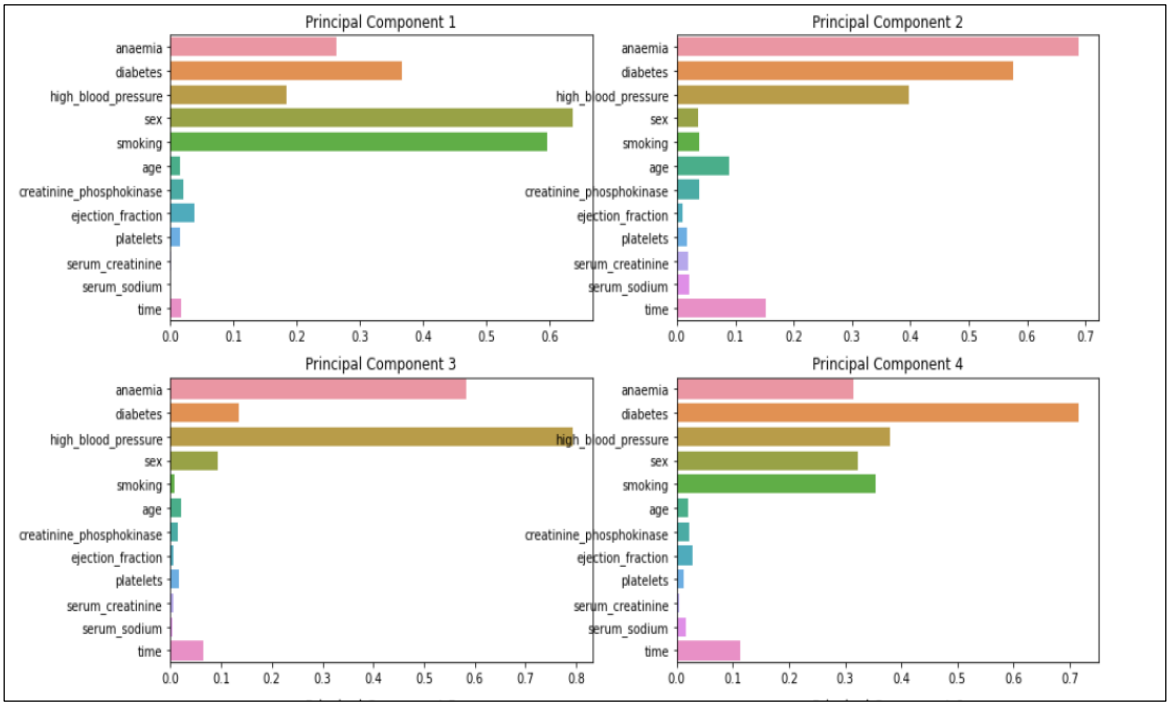
Figure 5: Decision Tree



This is a Random Generation of Decision Tree with the True or False it be Forming in the Form of the [**Figure 5**]Branches the Gini Index will be Including in This Model and Value Generated at the End of Tree is 8,12

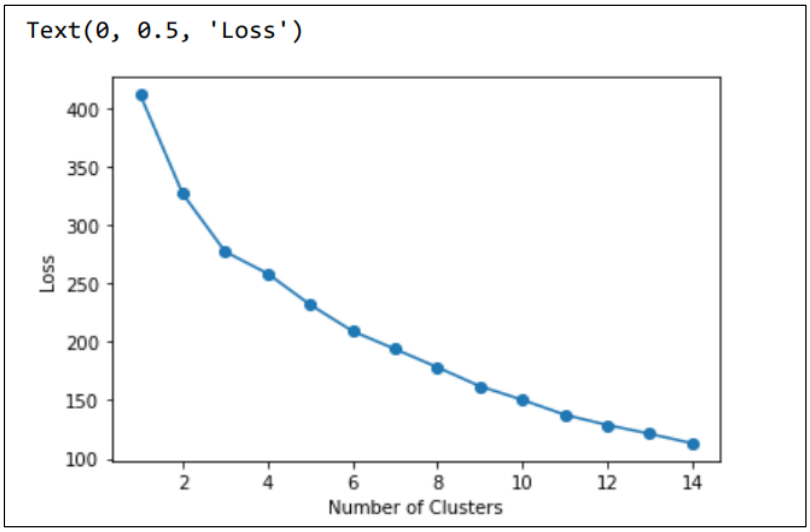
This Plot Relating the Principal Component Analysis Where the Dataset Prediction will be Dividing in to All Other Components the Smoking Rate is More in the Component 1 and Following by the Blood Pressure in the Next set of Data [**Figure 6**]

Figure 6: Principal Component Analysis



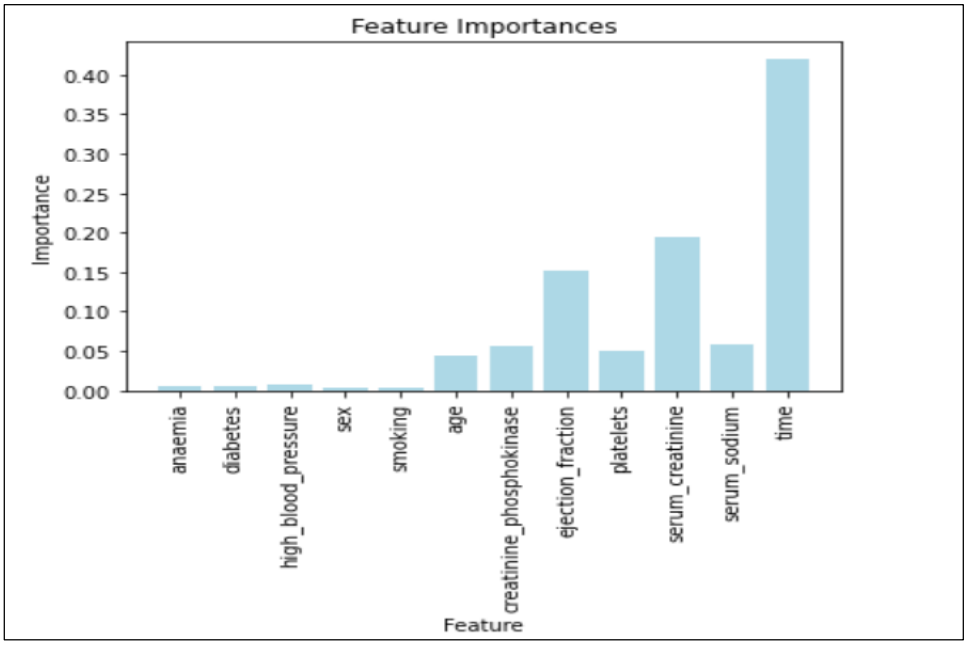
As [**Figure 7**] shows the It will Describes that the Number of Clusters are Divided on the Base of Dataset Parameters Like Age, Smoke, Alcohol Rate and the Loss is Generated was at Increase Level of 400

Figure 7: Clustering



This is an Better Learning Algorithm for the Prediction of Health Disease Dataset Because it Separates the Data by Logistic and in a Linear Regression Way [**Figure 8**]

Figure 8: Support Vector Machine



The algorithm for the random forest [**Figure 4**] will be determined by this plot. The accuracy of the training is the same as the accuracy of the decision tree, however the accuracy of the tests and the CVS varies. Despite this, the predictions made by CV Accuracy are quite accurate in comparison.

7.0 Results & Discussions

Following the completion of the prediction of the dataset using five different algorithms, we ultimately discovered that the approach known as random Python forest is the most effective one for the forecast. 0.8518 and 0.9999[**Table 1**] were the results that were obtained via CSV analysis for random forest. During generation of training dataset the dataset executed with on parameters and divided the dataset into two components such as x and y. After division of the training dataset the next result would be preprocessing the dataset and creating an instance for each algorithm which we are going to implement on a specific dataset. The algorithms took the dataset and pre-processed and given a dataset as a specified target and produced the best efficiency for generating plot graphs for a comparative study between types of algorithms. Finally we and our team got the best accuracy with the help of supervised and unsupervised machine learning algorithms.

8.0 Conclusion

Humans in their current environment are disproportionately affected by an illness known as heart disease. Our methodology makes use of the medical dataset on heart illness; the accuracy we achieved was the result of efficiently executing the algorithms.

References

1. World Health Organization. (2021). Cardiovascular diseases, key facts. Retrieved from https://www.who.int/news-room/factsheets/detail/cardiovascular-diseases-(cvds).
2. Choudhury, R. P. & Akbar, N. (2021). Beyond diabetes: A relationship between cardiovascular outcomes and Glycaemic Index. *Cardiovascular Research, 117*(8), E97–E98.
3. Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine, 10*(2), 334–343.
4. Magesh, G., & Swarnalatha, P. (2021). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence, 14*(2), 583–593.
5. Chowdary, R. K., Bhargav, P., Nikhil, N., Varun, K., & Jayanthi, D. (2022). Early heart disease prediction using ensemble learning techniques. *Journal of Physics: Conference Series, 2325*(1), 012051. Retrieved from DOI:10.1088/1742-6596/2325/1/012051.
6. Liu, J., Dong, X., Zhao, H., & Tian, Y. (2022). Predictive classifier for cardiovascular disease based on stacking model fusion. *Processes, 10*(4), 749.
7. Devi, A. G. (2021). A method of cardiovascular disease prediction using machine learning. *International Journal of Engineering Research and Technology, 9*(5), 243–246.
8. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making, 19*(1), 281.
9. Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Information in Medicine Unlocked, 26,* 100696. Retrieved from https://doi.org/10.1016/j.imu.2021.100696.
10. Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments, 7*(3), 263–275.
11. Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers accuracy improvement based on missing data imputation. *Journal of Artificial Intelligence and Soft Computing Research, 8*(1), 31–48.
12. Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings, 3*(1), 127–130.
13. Ananey-Obiri, D., & Sarku, E. (2020). Predicting the presence of heart diseases using comparative data mining and machine learning algorithms. *International Journal of Computer Applications, 176,* 17–21. Retrieved from doi:10.5120/ijca2020920034.
14. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access, 7,* 81542–81554. Retrieved from https://doi.org/10.1109/ACCESS.2019.2923707.
15. Kodati, S., & Vivekanandam, R. (2018). Analysis of heart disease using in data mining tools orange and weka. *Global Journal of Computer Science and Technology, 18*(1), 17–21.
16. Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and Its Applications, 482,* 796–807. Retrieved from DOI: 10.1016/j.physa.2017.04.113.

17. Jain, A., Dwivedi, R. K., Alshazly, H., Kumar, A., Bourouis, S., & Kaur, M. (2022). Design and simulation of ring network-on-chip for different configured nodes. *Computers, Materials & Continua, 71*(2), 4085-4100.
18. Kumar, A., & Jain, A. (2021). Image smog restoration using oblique gradient profile prior and energy minimization. *Frontiers of Computer Science, 15*(6), 1-7.
19. Jain, A., & Kumar, A. (2021). Desmogging of still smoggy images using a novel channel prior. *Journal of Ambient Intelligence and Humanized Computing, 12*(1), 1161-1177.