INTERNATIONAL JOURNAL OF
TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

HEART DISEASE PREDICTION

I. vasantha kumari[1], V.manisha[2], CH. muralidar[3],K. prathiksha[4],
G. Sumeeth giri[5]
Computer Science Engineering, Siddhartha Institute of Technology and Sciences, Narapally
vangamanisha28@gmail.com

ABSTRACT

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. We are working on developing machine learning algorithms to predict potential heart diseases in people. In this project, we compare the performance of various classifiers, including decision trees, Naive Bayes, logistic regression, SVM, and random forests. We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples. In this project, we compare the performance of various classifiers, including decision trees, Naive Bayes, logistic regression, SVM, and random forests. We also suggest an ensemble classifier that performs hybrid classification by using both strong and weak classifiers because it can have a large number of training and validation samples. Finally, we compare the performance of the proposed classifiers, such as Ada-boost and XG-boost, which can provide better accuracy.

KEYWORD

SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Adaboost, XG-boost, Python programming, confusion matrix, and correlation matrix are some of the terms to remember.

1. INTRODUCTION

The World Health Organisation estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Numerous studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. In high-risk individuals, an early diagnosis of heart disease is crucial for helping them decide whether to change their lifestyle, which lowers consequences.

2. RELATED WORK

The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Additionally, the goal of this research is to determine the optimum classification method for detecting cardiac disease in a patient. Using three comparative studies and analyses, this work is supported by at various stages of assessments , the categorization algorithms Naive Bayes, Decision Tree, and Random Forest are applied. Although these machine learning methods are widely utilised, predicting cardiac disease is a crucial task requiring the highest level Of accuracy.

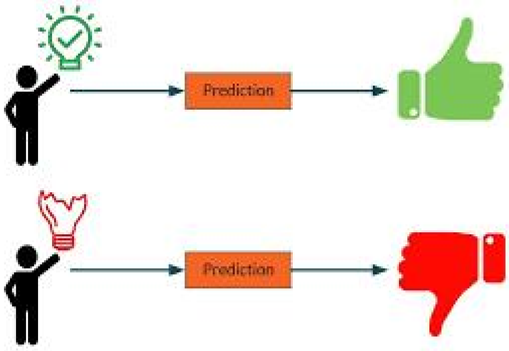
Consequently, a variety of levels and assessment strategy types are used to evaluate the three algorithms. This will enable scientists and medical professionals to create a better.

3. METHODOLOGY

Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease. The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre -processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

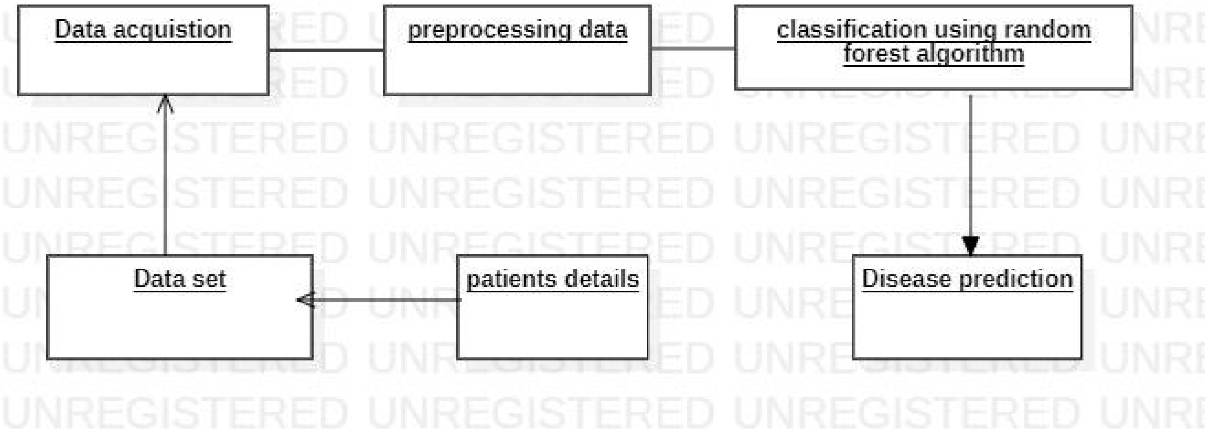
3.1 PREDICTION DISEASE

Diagnosis of Illness For classification, a variety of machine learning algorithms are employed, including SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Ada-boost, and Xg-boost. Algorithms are compared, and the one that predicts heart disease with the best degree of accuracy is chosen.



3.1.1 SYSTEM ARCHITECTURE

Dataset collection is the act of gathering information containing patient specifics. The method of selecting attributes chooses the relevant attributes for heart disease prediction. The available data resources are located, then further chosen, cleansed, and transformed into the required form. To accurately forecast cardiac disease, various classification approaches will be applied to pre-processed data. The accuracy of several classifiers is compared using the accuracy measure.



COMPUTER LEARNING A classification problem in machine learning is one in which a class label is anticipated for a specific example of input data.

● Supervised education Supervised learning is a sort of machine learning in which computers are trained with properly "labelled" training data, and on the basis of such data, machines predict the output. The term "labelled data" refers to input data that has already been assigned the appropriate output. In supervised learning, the training data that is given to the computers serves as the supervisor, instructing them on how to correctly predict the output. It employs the same idea that a pupil would learn under a teacher's guidance. The method of supervised learning involves giving the machine learning model the right input data as well as the output data. Finding a mapping function to link the input variable (x) with the output variable (y) is the goal of a supervised learning algorithm.

- Unsupervised education Because unlike supervised learning, we have the input data but no corresponding output data, unsupervised learning cannot be used to solve a regression or classification problem directly. Finding the underlying structure of a dataset, classifying the data into groups based on similarities, and representing the dataset in a compressed format are the objectives of unsupervised learning.• Finding relevant insights from the data can be aided through unsupervised learning. • Unsupervised learning is far more like how a human learns to think through personal experience, which brings it closer to the true AI. • Unsupervised learning is more significant because it operates on un labeled and uncategorized data.

• In the actual world, we don't always have input data that corresponds to the matching output, hence in these situations, unsupervised learning Reward-based learning Machine learning includes the discipline of reinforcement learning. It involves acting appropriately to maximise reward in a certain circumstance. Different programmes and machines use it to determine the optimal course of action to take in a given circumstance. While there is no answer in reinforcement learning, the reinforcement agent decides what to do to complete the task, in contrast to supervised learning where the training data includes the answer key and the model is thus trained with the correct answer itself. It is obligated to gain knowledge from its experience in the absence of a training dataset.

4. HARDWARE REQUIREMENTS

- Any Update as a Processor Processer
- RAM: 4GB minimum
- Hard Disc: 100GB minimum

5. SOFTWARE REQUIREMENTS

- Windows-based operating system
- Technology: Jupiter notebook with Python 3.7 IDE

6. WORKING

In a Decision Tree, the algorithm begins at the root node and works its way up to forecast the class of the given dataset. This algorithm follows the branch and jumps to the following node based on a comparison of the values of the root attribute and the record (real dataset) attribute. The process continues by comparing the attribute value for the subsequent node with those of the other sub-nodes once more. It keeps doing this until it reaches the tree's leaf node. The following algorithm can help you comprehend the entire procedure:

Step 1: The S start is to start the tree from the root node, which holds the entire dataset.

Step 2: Using the Attribute Selection Measure (ASM), identify the dataset's top attribute.

Step 3 is to split the S into groups that include potential values for the best qualities.

Step 4: Create the Decision Tree node that has the best attribute

Step-5: Using the subsets of the dataset generated in step three, develop new decision trees iteratively in step five. Continue along this path until you can no longer categorise the nodes any further and may designate the last node as a leaf node.

RANDOM FOREST ALGORITHM: A supervised learning algorithm is Random Forest. It is a development of machine learning classifiers that incorporates bagging to boost Decision Tree performance. It mixes tree predictors, and trees depend on an individually sampled random vector. All trees are distributed in the same way. Instead of splitting nodes based on variables, Random Forests uses the best among a prediction subset that is randomly selected from the node itself. The worst case of learning with Random Forests has a temporal complexity of O(M(dnlogn)), where M is the number of growing trees, n is the number of occurrences, and d is the data dimension. Both classification and regression can be done with it. Additionally, it is the most user-friendly and adaptable algorithm. Trees make up a forest. A forest is supposed to be stronger the more trees it has. On randomly chosen data samples, Random Forests build Decision Trees, obtain predictions from each tree, and then vote on the best answer. Additionally, it offers a fairly accurate indicator of the feature's relevance. Applications for Random Forests include recommendation systems, picture classification, and feature extraction. It can be used to categorise dependable loan candidates, spot fraud, and forecast sickness. The Boruta algorithm, which chooses significant features in a dataset, is built around it. Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is predicated on the idea of ensemble learning, which is the act of integrating various classifiers to address a complicated issue and enhance the model's performance. According to what its name implies, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

7. CONCLUSION

Application of promising technology, such machine learning, to the first prediction of heart problems would have a significant social impact because heart diseases are a leading cause of death in India and around the world. Early detection of cardiac disease can help high-risk patients make decisions about lifestyle modifications that will lessen problems, which can be a significant advancement in the field of medicine. Each year, more people are diagnosed with cardiac illnesses. This calls for an early diagnosis and course of action. The medical community as well as patients may benefit greatly from the use of appropriate technology support in this area. SVM, Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting are just a few of the seven machine learning algorithms employed in this study to evaluate performance. The dataset, which has 76 features, contains the anticipated characteristics that lead to heart disease in individuals, and 14 significant features are chosen from them to help evaluate the system. When all the features are taken into account, the efficiency of the system that the author obtains is lower. Attribute selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides greater accuracy. Some dataset features have virtually equal correlations, so they are eliminated. The efficiency significantly declines if all the attributes in the dataset are taken into account. A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. Thus, the objective is to employ a variety of evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The extreme gradient boosting classifier has the highest accuracy (81%), when all seven are compared.

8. REFERENCES

[1] Soni J., Ansari U., Sharma D., and Soni S. (2011). An overview of the use of predictive data mining for medical diagnostics is provided here. 17(8), 43–8 International Journal of Computer Applications [2] Dangare C. S. & Apte, S. S.improved research into the classification methods used in data mining for heart disease.47(10), 44–8, International Journal of Computer Applications. [3] See C. Ordonez (2006). Discovering association rules for heart disease prediction using the train and test method.10(2), 334–43, IEEE Transactions on Information Technology in Biomedicine. [4] According to Shinde R, Arjun S, Patil P, and Wagmare J (2015). using the Naive Bayes algorithm and K-means clustering, an intelligent system for predicting cardiac disease.The 6(1) issue of the International Journal of Computer Science and Information Technologies, p. 637-9. [5] Bashir S., Qamar U., and Javed M. Y. (2014). A paradigm for group-based decision-making for intelligent heart disease

diagnosis. i-Society 2014: International Conference on the Information Society (pp. 259–64). Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 IEEE. ICCRDA 2020 IOP Conf. 9 [6] Jee, S. H., Jang, Y., Oh, D. J., Oh, B. H., Lee, S. H., Park, S. W., & Yun, Y. D. (2014). The Korean Heart Study is a model for predicting coronary heart disease. e005025 in BMJ Open, 4(5). Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, rnlöv J, and Ingelsson E (2013). For the prediction of coronary heart disease, multilocus genetic risk scores. 33(9), 2267–72. Arteriosclerosis, thrombosis, and vascular biology. [8] (2013) Jabbar M A, Deekshatulu B L, and Chandra P. predicting heart disease with lazy associative classification. International MultiConference on Automation, Computing, Communication, Control, and Compressed Sensing (iMac4s), 2013, pp. 40–6. IEEE. [9] In 1997, Brown N., Young T., Grey D., Skene A. M., and Hampton J. R. Analysis of data from the Nottingham heart attack register on inpatient fatalities from acute myocardial infarction between 1982 and 1992. BMJ, 315(7101), 159-64.