ABSTRACT:

Cardiovascular disease is also called heart and blood vessel disease. In the last few decades the primary cause of global death is cardiovascular disease.
According to the World Health Organization (WHO), 17.9 million people die each year as a result of cardiovascular disease.It is not possible to predict heart disease by the medical practitioners that require higher knowledge for prediction. It's a heart or blood vascular disease. Recent development in medical technology based on machine learning and collection of crucial information from past documentation for future analysis(Data Mining) plays an essential role in predicting a heart or blood vessel disease. Usually data mining techniques are used for processing a large dataset. Researchers use various KDD(Knowledge Discovery in Data) and ML algorithms for predicting a cardiovascular disease. KNN,RF,Naive Bayes,DT,XGBoost,Neural network,support vector machine are the algorithms used to check which algorithm is giving maximum accuracy . Among that RF algorithm gives maximum accuracy.Dataset used in this project is Heart disease Dataset of Cleveland from UCI repositories. It consists of Three hundred three rows and seventy six columns in which only 14 columns have been taken for testing. This project aims to predict the heart disease caused in future for the patient.The accuracy score 91.8% is achieved by RF algorithms.

TECHNIQUES USED:

Based on the past information classification task is used for predicting a heart disease. K Nearest Neighbour (KNN) ,Random Forest (RF), Naive Bayes (NB), Decision Tree (DT) ,XGBoost,Support Vector Machine(SVM),Logistic Regression(LR)and Neural Network (NN) was used in our project to check which algorithm is giving maximum accuracy. The accuracy given by the various algorithms varies with the number of attributes. RF algorithm gives maximum accuracy.

STEPS:

- **Data Set Collection:** For machine learning related projects strictly there is a need for a dataset to train the model to predict the acquired result.
- **Splitting the Dataset:** Split the dataset into training and testing to fit the model to predict the attack.
- **Model Fitting:** Fit the model using training and testing dataset to predict the attack.
- **Save the model:** Save the model to predict the attack by using the user given data.

ALGORITHMS USED:

A. **LR:** One of the popular and easiest machine learning algorithms is linear regression. For predictive analysis a mathematical formula(statistical method) is used. LR predicts numeric values for example sales, salary, age, product price, etc. It displays the linear relationships between y(dependent) and one or more independent variables. So it is called linear regression. It finds out the values of a dependent variable and how it is changing according to the values of the independent variable. The model provides the straight line(slope) the straight line displays the relationship between variables. The accuracy we get through our model is 85.25%.

B. **NAIVE BAYES:** It is a supervised learning algorithm for addressing classification issues based on the Bayes theorem.It is the most basic and successful classification technique for generating machine learning models quickly and making accurate predictions.It makes predictions based on an object's probability. Naive bayes is commonly used for spam filtration, sentiment analysis, and classifying articles.It is made up of two terms, one of which is naïve and the other is bayes.The naive is described as an assumption in the occurrence of a specific characteristic that is independent of other features.Consider a fruit: if a fruit is identified by its colour, shape, and flavour, a red, spherical, and sweet fruit is recognised as an apple. As a result, we can say that each attribute helps to identify an apple without relying on the others.It is also known as bayes rule or bayes law because it is based on the premise of Bayes theorem. It's used to figure out how likely a hypothesis is based on prior knowledge. $P(A|B)=P(B|A)P(A)/P(B)$ is the naïve bayes formula. Our model provided us with an accuracy of 85.25 percent.

C. **SVM:** It is one of the most widely used supervised learning algorithms, with applications in classification and regression. SVM generates a best line or decision boundary that divides n-dimensional space into classes so that fresh data can be conveniently placed in the appropriate category in the future. SVM chooses the extreme vectors, which aids in the creation of a hyperplane, which is also known as the optimal decision boundary. Support vectors are the extreme vectors, which is why the machine is referred to as a support vector machine. Our model provided 81.97 percent accuracy.

D. **KNN:** Because of its simplicity, KNN (K-Nearest Neighbors) is the most often used machine learning method. It is a supervised learning technique that assumes similarities between new data and old data and places the new data in the most similar category to the existing categories. Typically, KNN maintains all existing data and classifies new data based on similarity. This implies that when new data arrives, the KNN algorithm can quickly categorise the category. It can be used for both regression and classification, however classification is the most typical application. Our model provided an accuracy of 81.25 percent.

E. DECISION TREE: It's a supervised learning technique that may be applied to both classification and regression issues, however it's most typically employed for classification. It has the appearance of a tree, with internal nodes representing dataset attributes, branches representing decision rules, and each leaf node representing the outcome. It all starts with the root node, which then spreads out to form tree-like structures. CARD is used to construct a tree (Classification and Regression Tree Algorithm). Our model provided us with an accuracy of 80.33 percent.

F. **Random Forest:** It's the most often used supervised learning method for classification and regression issues. It is the technique of merging numerous classifiers to solve a complex problem and improve a model's performance. Random Forest is a classifier that incorporates a number of decision trees, as the name suggests. It forecasts the ultimate output using the predictions from each tree based on the majority votes of prediction. The random forest's accuracy improves as the number of trees grows, and it also prevents overfitting. Our model provided us with a 91.8 percent accuracy.

G. **XGBoost:** XGBoost stands for Extreme Gradient Boosting. It is mainly used for supervised learning problems. It performs well as compared to all other machine learning algorithms. It performs well in competitions and hackathons. It is one of the excellent algorithms initially developed for structured data. It performs well in terms of speed and performance.The accuracy we got through our model is **85.25%** .

H. **Neural Network:** Dr.Robert Hecht-Nielsen he is the inventor of the first neurocomputer he says that ANN(Artificial Neural Network) is referred to as a neural network. In this paper we have used a sequential model and activation as relu in the first layer and sigmoid in the layer two .The accuracy we got in our model is 85.25%.