PREDICTION OF CARDIOVASCULAR DISEASE

Project Reference No.: *45S_BE_1815*

College ***:*** *Vidyavardhaka College of Engineering, Mysuru*
Branch ***:*** *Department of Computer Science Engineering*
Guide(s) ***:*** *Dr. Ramakrishna Hegde*
Student(S) ***:*** *Ms. Shree Raksha G M*
Ms. Shivani M N
Ms. Shrinidhi P S
Ms. Thashwin Monnappa M M

Keywords:

Cardiovascular Disease, Database, Analysis, Prediction, Machine learning, X-radient Boost.

Introduction:

Firstly, let's understand what is cardiovascular disease, a cardiovascular health refers to the health of heart, blood vessels and arteries. It includes coronary heart disease, stroke, heart arrhythmias and heart failure. As understood above about cardiovascular disease, the symptoms include angina, chest pain, chest tightness, chest pressure. Also, there are several other reasons like numbness, pain, weakness or coldness in the legs or arms if the blood vessels in those parts of the body are constricted. Cardiovascular diseases are being the unusual cause behind a huge number of youth deaths for last two decades. It was observed that 16.3 million Americans aged 20 and older have coronary heart disease, a prevalence of 7 percentage. The percentage men in it are 8.3 percent and women have 6.1 percent. By looking at this can we say age is the risk factor, well traditionally it is a yes. But if we look at past 2 years the statics is not the same. Now 1 in 5 heart attack patients are younger than 40 years of age. Now a days there are some percentages of patients fall under 20s and 30s Basically, Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels, including coronary heart disease, cerebrovascular disease (Stroke), peripheral arterial disease, Cardiac arrest and Heart failure. In a study, it is observed that cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.

The emergence of machine learning techniques has demonstrated their effectiveness in disease prediction from massive amounts of healthcare data. Cardiovascular disease is difficult to recognize due to a variety of risk factors such as high blood pressure, cholesterol, and abnormal pulse rate. Because of the disease's complexity, it must be handled with care. Otherwise, the effects of heart or death may occur. With computer-aided decision-support/prediction systems, technological advancements have aided the field of medicine. In the healthcare industry, machine learning techniques have demonstrated accurate disease prediction in less time.

Objective:

Machine learning (ML) is a branch of artificial intelligence (AI) that is increasingly utilized within the field of cardiovascular medicine. It is essentially how computers make sense of data and decide or classify a task with or without human supervision. The conceptual framework of ML is based on models that receive input data (e.g., images or text) and through a combination of mathematical optimization and statistical analysis predict outcomes (e.g., favorable, unfavorable, or neutral). Several ML algorithms have been applied to daily activities.

Machine learning algorithms have emerged as highly effective methods for prediction in cardiovascular research. They can capture the complex interactions between predictors and nonlinear relationships between predictors and outcomes, producing better predictive performance than traditional statistical models. Studies suggested that random forest, support vector machine, outperformed traditional models. However, results are still inconsistent, a recently published meta-analysis showed that ML-based predictive models do not perform better than logistic regression.

Hence the gap between technology and medicine can be narrowed using these techniques to predict cardiovascular disease in the earlier stage and get proper medicine. Therefore using the core technical concepts of algorithms and machine learning it is possible to address the above problem.

System Requirements and Methodology:

System Requirements

1. Software Requirements

Table 1: Software Requirements

Operating System	Windows 7 and above
Tools	Jupiter Notebook
Runtime Platform	Python 3.9 and above
Architecture	64 – bit

2. Hardware Requirements

Processor	64 bit(x64) i3 processor (1.7GHz)
Memory	2GB RAM
GPU	NVidia 600s and above, AMD 5400s and

Methodology:

The purpose of the design phase is to plan a solution of the problem specified by the requirements document. This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed, design takes us toward how to

satisfy the needs. The design of a system is perhaps the most critical factor affecting the quality of the software; it has a major impact on the later phases particularly testing and maintenance.

This design basically explains how steps are performed one by one in order to achieve the desired outcome that is a machine learning model that is capable or predicting cardiovascular disease. The focus is on designing the logic and operation performed in each step. In other words, how analysis is done and can be implemented in software that can be used in laboratories in order to predict cardiovascular disease. It can also be termed as a design methodology is a systematic approach in creating a design with set of techniques and guidelines followed.

The main steps in the System Design mainly include the following steps:

The purpose of the design phase is to plan a solution of the problem specified by the requirements document.
This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed, design takes us toward how to satisfy the needs. The design of a system is perhaps the most critical factor affecting the quality of the software; it has a major impact on the later phases particularly testing and maintenance.

This design basically explains how steps are performed one by one in order to achieve the desired outcome that is a machine learning model that is capable or predicting cardiovascular disease. The focus is on designing the logic and operation performed in each step. In other words, how analysis is done and can be implemented in software that can be used in laboratories in order to predict cardiovascular disease. It can also be termed as a design methodology is a systematic approach in creating a design with set of techniques and guidelines followed.

The main steps in the System Design mainly include the following steps:

- Data collection
- Data Pre-processing
- Exploratory Data analysis
- Applying different Machine Learning Algorithms
- Testing of the Model
- Desired output

Collection of data is a phase in which data is collected from the sources where enough data is available. In this case we have collected the data from the Kaggle website in which there were 70000 rows and 13 columns.

Data pre-processing is all about making use of data mining techniques which is used to transform the raw data in a useful and efficient format. The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

Missing data is a situation arises when some data is missing in the data. It can be handled in various ways.

1. Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
2. Fill the Missing values: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

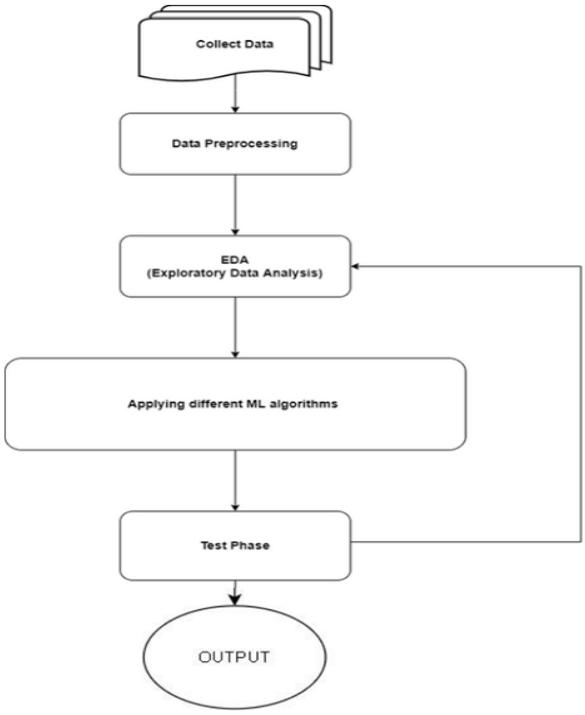


Fig.1 System Design

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Supervised learning uses classification and regression techniques.

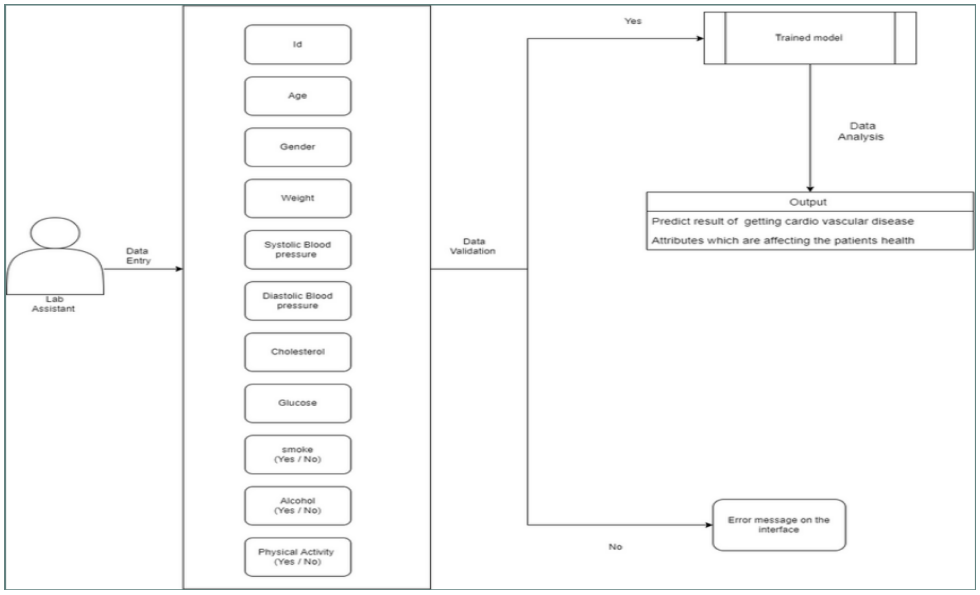
Classification techniques predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.

Regression techniques predict continuous responses—for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading.

Testing is a phase is which the ML model is evaluated on the basis of trained and test data it gives the number or predictions that the model has predicted correctly gives the accuracy of the model.

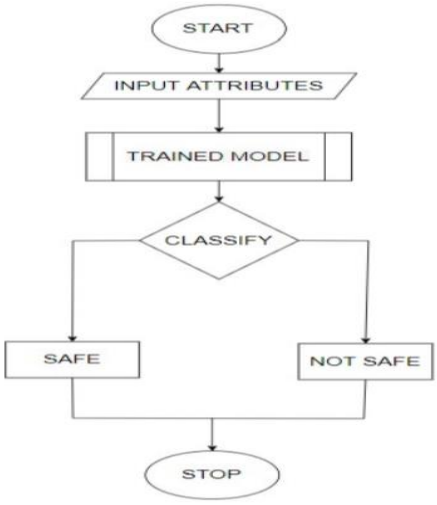
Then the training and evaluation phases are repeatedly performed to make the model better can reduce type one errors, type two errors resulting in better model at every stage. Now the desired output is available.

Use case diagram:



This is a use case diagram where the user performs his task. Our target users are laboratory assistants where they enter the details of the patients which should include id, age, gender, weight, height, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose levels, smoking and alcohol. The data is validated to check if the entered data is in the scale of normal values if found to be true the data is fed to the model else in the interface error message is displayed. If the data is right the model gives the result if the patient has chances of getting cardiovascular disease or not.

Flow of the Designed Model:



Flow of the entire design is like the entered details of the patients which will include id, age, gender, weight, height, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Glucose levels, smoking and alcohol. The data is validated to check if the entered data is in the scale of normal values if found to be true the data is fed to the model else in the interface error message is displayed. If the data is right the model gives the result if the patient has chances of getting cardiovascular disease or not. Then the process stops.

Result and Conclusion:

Result:

The following section gives the results observed for machine learning techniques like

1. Random Forest Classifier
2. Support Vector Classifier
3. K Neighbors Classifier
4. X Gradient Boost Classifier

The dataset which was fed to the above classifiers had 13 input columns which includes age, Female, Male, Weight, BMI (Body mass Index), ap_hi (Systolic), ap_lo (Diastolic), Cholesterol, Glucose level, Smoke, Alcohol, Physical activities and one target variable that is heart disease.

Random Forest

Random forest classifier performed quite well having the testing accuracy of 73.25% and average testing accuracy of 73.44%. Confusion Matrix that was plotted as mentioned above figure gave a Type 1 error of 2921 and Type 2 error of 1837. The model gave the correct prediction for 12,393 out of 17,151.

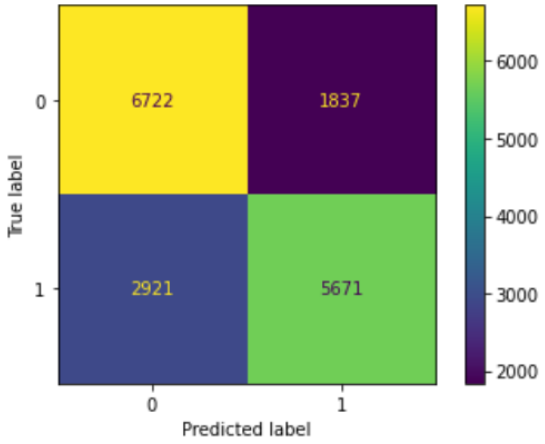


Figure 1: Confusion Matrix for Random Forest

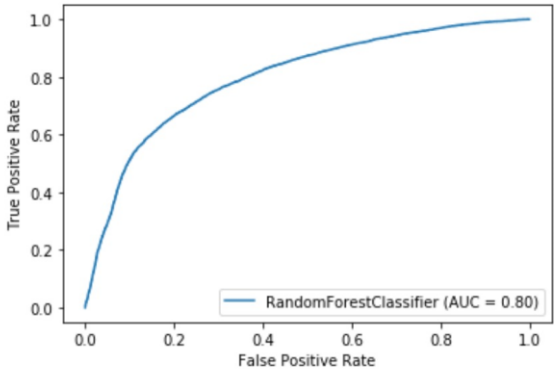
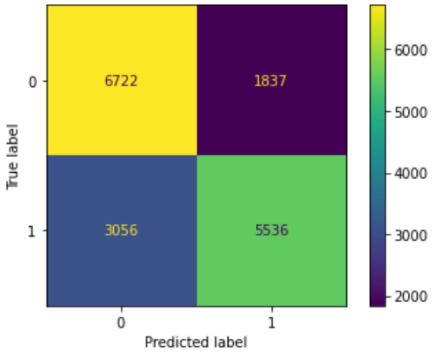


Figure 2: AUC curve for Random Forest

AUC curve is a measure i.e., it's the ability of a classifier to distinguish between the classes. Higher the AUC then much better the model performs so that it's easy to distinguish between the positive and the negative classes If the value of the AUC is between 0.5 and 1 then there are higher chances of the classifier to distinguish the positive values from the negative class values. From the above figure we found out that the AUC curve drawn for the Random Forest classifier is 0.80. So, there is high chances of the model to provide positive result.

K-Neighbors Classifier

The K Neighbors classifier having the neighbors value as 300 gave a mean cross validation score of 0.722 and F1 score of 72.35%. Confusion Matrix that was plotted for the same as mentioned above figure gave a Type 1 error of 3056 and Type 2 error of 1837. The model gave the correct prediction for 12,258 out of 17,151.



XGBoost

The XGBoost classifier is the best fitted model for the above problem giving a testing accuracy of 73.4% and Average Testing Accuracy of 73.6%.

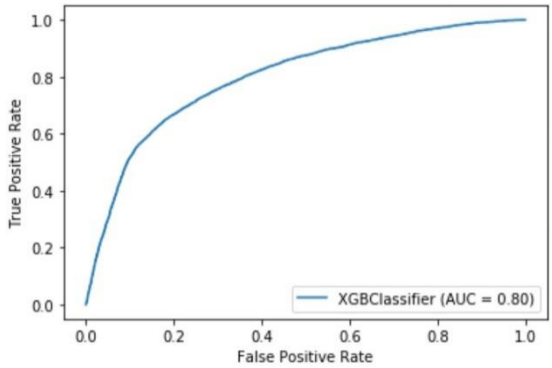


Figure 4: AUC curve for XGBoost Classifier

From the above figure we found out that the AUC curve drawn for the XGboost classifier is 0.80. So, there is high chances of the model to provide positive result.

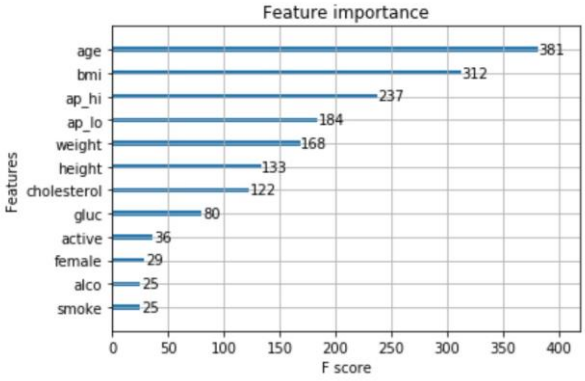


Figure 5: Feature importance

The Feature Importance graph was plotted for the XGBoost model as mentioned in the figure represented how F score acted against features/ attributes and the following results were observed where Age had highest feature importance and BMI, Systolic, Diastolic, Height,

Weight and Cholesterol levels showed reasonable F score whereas features such as Smoke, Alcohol, Physical activities and Glucose levels showed very less feature importance.

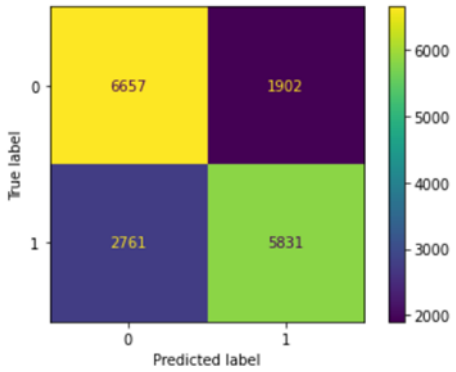


Figure 6: Confusion Matrix for XGBoost Classifier

Confusion Matrix that was plotted as mentioned in above figure gave a Type 1 error of 2921 and Type 2 error of 1837. The model gave the correct prediction for 12,393 out of 17,151.

	precision	recall	f1-score	support
not_Cardic	0.72	0.77	0.75	8710
Cardic	0.75	0.69	0.72	8441
accuracy			0.73	17151
macro avg	0.74	0.73	0.73	17151
weighted avg	0.73	0.73	0.73	17151

Figure 7: Classification of XGBoost Classifier

Classification report is one of the ways that is used to evaluate the performance of machine learning model using the values of precision, recall, F1 score and support.

Precision and Recall are given by the following formulas

precision = TP/(TP+FP)

recall = TP/(TP+FN)

In figure it is observed that ratio of number of correct positive prediction to total number of positive predictions is 0.75 for positive target and 0.72 for negative target. The ratio of positive and true predictions to that of total actual positives of the model is found to be 0.77 and 0.69 for non-cardiac and cardiac outputs. The observed values of F1 score are 0.75 and 0.72 for non-cardiac and cardiac outputs.

Conclusion:

In this project we have developed a cardiovascular predicting model which takes 11 inputs and reverts the predicted out according to the trained model. The 11 inputs are considered to be very prominent and required input attributes to consider whether the individual falls under the category of cardio vascular positive or not. Therefore, we can consider that these attributes play a very significant role in prediction.

We consider that structure of the data predicting model should be carried out very well, therefore we have followed steps like:

1. Data preprocessing and cleaning
2. Data Analysis
3. Probability and statistics
4. Predicting using Machine learning

Each step carries respective importance.

Here in our model, we have implemented various algorithms like Linear Regression, K-Nearest Neighbor, Random Forest, and X-Gradient Boost. We have also used feature selection in the predicting model in order to highlight which feature is affecting the most. It has been observed that feature selection gave us a very in-depth analysis regarding consideration of the features.

Lastly, we claim to have the least type one error that is 0.8. This will determine how accurately the model is predicting. Though we have 73.59% as accuracy the AUC curve draws a very least and negligible amount of type one error. By this we conclude that this model will serve the lab technicians, doctors and users to understand their/patients health due to their lifestyle adaptations and their blood profile.

Scope For Future Work:

We have made the maximum utilization of our potential and zest in developing this model. But gaining knowledge is a continuous process and so is this new technology. Therefore, in this section we present some of the ideas which can be used to enhance the functionalities of our project to widen its applications.

Using of deep learning algorithms

Machine learning requires less computing power whereas deep learning typically needs less ongoing human intervention. Deep learning can analyze unstructured data in ways machine learning can't easily do. Every industry will have career paths that involve machine and deep learning

Building a mobile application

Building a mobile app will definitely help the general public users. Initially we have not accomplished the mobile application because all the data which is asked for as an input will not be known to layman, therefore we have created a web application that will be used only by lab technicians.

Providing the list of features effecting the patient

With this feature, doctors will be able to analyse what exactly might be the root cause of that patient and can treat accordingly. Also, when this feature is shown to patients, they will also understand the severity behind their life style changes.