**Synopsis**

**on**

**Heart Problems Prediction**

Submitted in partial fulfillment of the requirements for the award of degree of

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE & ENGINEERING**



**Department of Computer Science and Engineering**

**Chandigarh University, Gharuan**

**August, 2022**

**Submitted By:**

Aji Santhosh Kumar Nair (19BCS1903)

Anil Kumar Pradhan (19BCS1898)

Keshav Mishra (19BCS1887)

Sahil Lamba (19BCS1922)

Sahil Mallick (19BCS1073)

**TABLE OF CONTENTS**

1. **Introduction…………………………………………………………..…….………1-2**

**1.1 Scope of the Project……………………………………………………………....1**

**1.2 Objectives………………………………………………………………………....1**

**1.3 Problem Statement……………………………………………………………….2**

1. **Methodology/Planning of Work…………………………………………………..2-4**
   1. **Data Wrangling…………………………………………………………………..2**
   2. **Exploratory Data Analysis………………………………………………………2**
   3. **Machine Learning Algorithms & Predictive Analytics………………………..3**
   4. **Finding Feature Score……………………………………………………………4**
2. **Timeline………………………………………………..……………………………...5**
3. **Software and Hardware Requirements……………………..………………………5**

**Bibliography……………………………………………………………………………..6**

1. **Introduction:**

According to the World Health Organization, 12 million people die every year due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. 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 analyzing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

This project aims to predict future Heart Disease by analyzing data of patients, which classifies whether they have heart disease or not using machine-learning algorithms. A dataset is selected from the UCI repository with patient’s medical history and attributes. By using this dataset, we predict whether the patient can have a heart disease or not. To predict this, we use various medical attributes of a patient and classify if the patient is likely to have a heart disease. These medical attributes are trained under seven algorithms: Logistic Regression, KNN, SVM, Naives Bayes Classifier, Decision Trees Classifier, Random Forest Classifier and XGBoost.

**1.1 Scope of the Project:**

The main idea behind our project is to create a heart disease prediction system based on the medical attributes from the dataset. The proposed system will classify patients as either positive or negative for heart disease by exploring the above mentioned classification algorithms and do performance analysis.

**1.2 Objectives:**

The main objective of developing this project are:

1. To develop a machine learning model to predict future possibility of heart disease by implementing various classification algorithms.
2. To determine significant risk factors based on medical dataset which may lead to heart disease.
3. To analyze feature selection methods and understand their working principle.

**1.3 Problem Statement:**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and over all complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1. **Methodology/Planning of Work:**

**2.1 Data Wrangling:**

We’ll first look at the dataset we are working with by converting it into a simpler and more understandable format. It would help us use the data more appropriately.

All attributes are listed below -

1. age (#)
2. sex : 1= Male, 0= Female (Binary)
3. (cp)chest pain type (4 values -Ordinal):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic (
4. (trestbps) resting blood pressure (#)
5. (chol) serum cholestoral in mg/dl (#)
6. (fbs)fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)
7. (restecg) resting electrocardiographic results(values 0,1,2)
8. (thalach) maximum heart rate achieved (#)
9. (exang) exercise induced angina (binary) (1 = yes; 0 = no)
10. (oldpeak) = ST depression induced by exercise relative to rest (#)
11. (slope) of the peak exercise ST segment (Ordinal) (Value 1: upsloping , Value 2: flat , Value 3: downsloping )
12. (ca) number of major vessels (0-3, Ordinal) colored by fluoroscopy
13. (thal) maximum heart rate achieved - (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversable defect.

**2.2 Exploratory Data Analysis:**

#### After completing data wrangling, we will perform exploratory data analysis. Here are the primary tasks we will perform in this stage of our heart disease prediction project:

#### ****Finding Correlations:****

We’ll create a correlation matrix that helps us see the correlations between different variables.

#### ****Using Violin and Box Plots:****

With Violin and Box plots we can see the basic statistics and distribution of our data. We can use it to compare the distribution of a specific variable across different categories. It will help us identify outliers in the data as well.

#### ****Filtering Data:****

Here, we’ll filter the data according to positive and negative heart disease patients. We’ll start with filtering data by Positive heart disease patients.

**2.3 Machine Learning Algorithms & Predictive Analytics:**

We’ll use multiple machine learning algorithms and find the one that offers the highest accuracy. The machine learning algorithms we’ll be using is listed below:

1. **Logistic Regression:**

Logistic regression is one of the most popular Machine Learning algorithms,which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. But instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

1. **KNN (K-Nearest Neighbors):**

K Nearest Neighbor (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

1. **SVM (Support Vector Machine):**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

1. **Naives Bayes Classifier:**

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is mainly used in text classification that includes a high-dimensional training dataset.

Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

1. **Decision Trees Algorithm:**

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node.

Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a Decision Tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A Decision Tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

1. **Random Forest Classifier:**

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. The time complexity of the worst case of learning with Random Forests is O(M(dnlogn)) , where M is the number of growing trees, n is the number of instances, and d is the data dimension. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

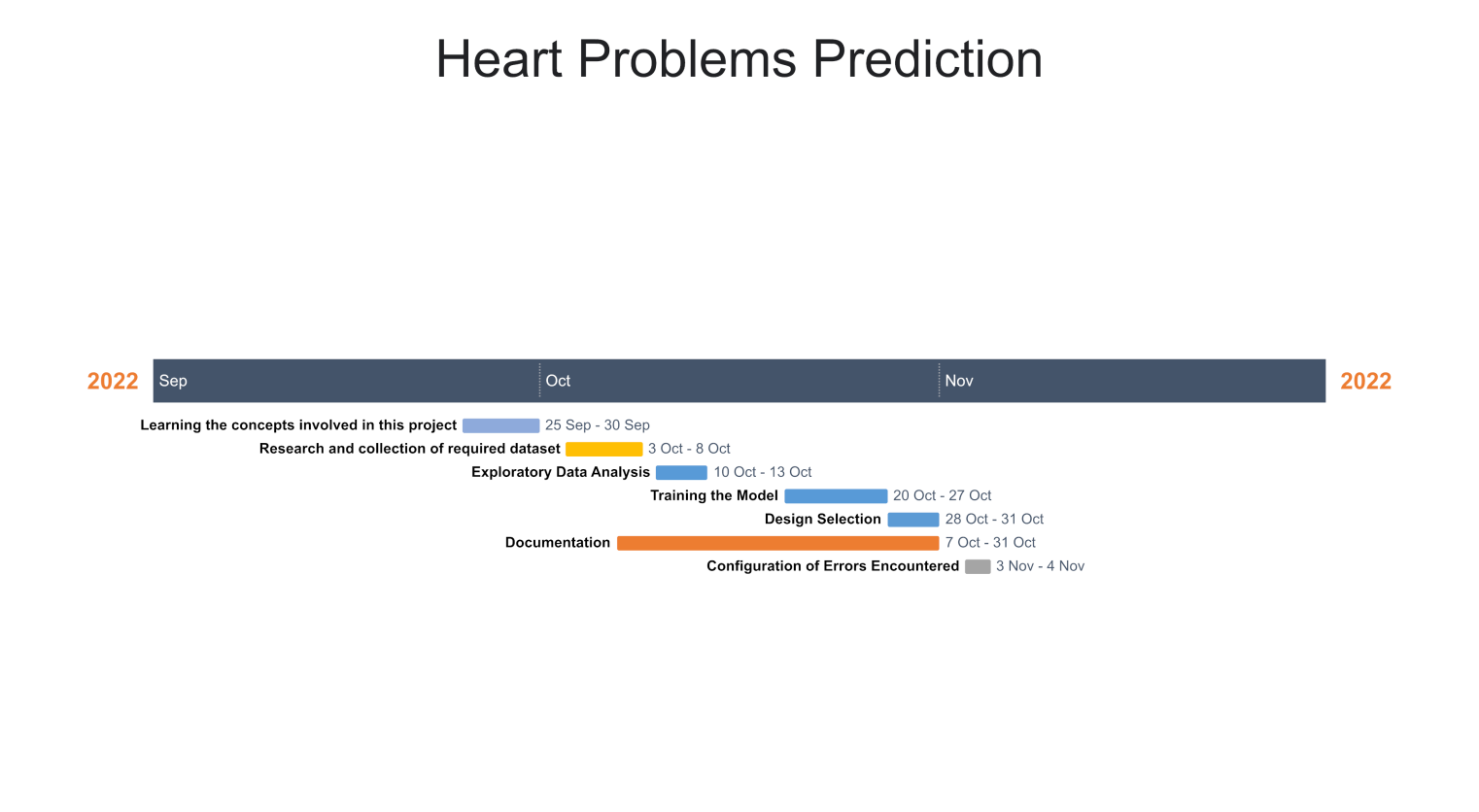
1. **XGBoost:**

XG-boost is an implementation of Gradient Boosted decision trees. It is a type of Software library that was designed basically to improve speed and model performance. In this algorithm, decision trees are created in sequential form. Weights play an important role in XG-boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these the variables are then fed to the second decision tree. These individual classifiers/predictors then assemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined predict.

**2.4 Finding Feature Score:**

Feature Importance provides a score that indicates how helpful each feature was in our model. The higher the Feature Score, the more that feature is used to make key decisions & thus the more important it is.

**3. Timeline**



**4. Software and Hardware Requirements:**

**Hardware Requirements:-**

• Processor: Minimum 1 GHz; Recommended 2GHz or more

• Hard Drive: Minimum 32 GB; Recommended 64 GB or more

• Memory (RAM): Minimum 1 GB; Recommended 4 GB or above

**Software Requirements:-**

• Python 3.10

• Jupyter Notebook

**Bibliography:**

• <https://www.google.com>

• <https://www.wikipedia.org>

• <https://www.kaggle.com/datasets/harikajds/heart-disease-prediction?resource=download&select=dataset.csv>

• World Health Organization, Cardiovascular Diseases, WHO, Geneva, Switzerland, 2020, [https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab\_1](https://www.who.int/health-topics/cardiovascular-diseases/" \l "tab=tab_1" \t "https://www.hindawi.com/journals/cin/2021/8387680/_blank)/

• R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, “Automatic methods to extract New York heart association classification from clinical notes,” in Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.

View at: [Publisher Site](https://doi.org/10.1109/bibm.2017.8217848" \t "https://www.hindawi.com/journals/cin/2021/8387680/_blank) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Automatic%20methods%20to%20extract%20New%20York%20heart%20association%20classification%20from%20clinical%20notes&author=R.%20Zhang,%20S.%20Ma,%20L.%20Shanahan,%20J.%20Munroe,%20S.%20Horn,%20and%20S.%20Speedie" \t "https://www.hindawi.com/journals/cin/2021/8387680/_blank)