PREDICTION OF HEART DISEASE USING DATA MINING TECHNIQUES

Sindhu Maddineni, Meghana Bellamkonda, Pavitra Mekala, Reshma Madala, Rohith Kodali

SRM University, Andhra Pradesh

(sindhu_maddineni, meghana_sridhar, pavitra_mekala, reshmasri_madala, rohith_kodali)@srmap.edu.in

ABSTRACT:

Types of diseases are on the rise these days as a result of today's lifestyle as well as inevitable inheritance. Heart disease has become one of the most common of these ailments in recent years. Every year, over 659,000 people in the United States die from heart disease. Cardiovascular diseases (CVDs) are the leading cause of death worldwide, claiming the lives of an estimated 17.9 million people per year. Pulse rate, blood pressure and cholesterol are all varied for each person. However, medically confirmed data show that the normal blood pressure for adults is 120/90, total cholesterol for males aged 20 years and older is 40 mg/dL or higher, and for women aged 20 years and older is 50 mg/dL or higher, and pulse rate for adults is generally between 60 and 100 beats per minute. We employ various categorization techniques in this study to forecast each person's risk level based on age, gender, blood pressure, cholesterol, and pulse rate. Data mining classification techniques such as Naive Bayes, KNN, Decision Tree Algorithm, and Neural Network are used to classify the patient's risk level.

Introduction:

Data mining involves finding previously undiscovered patterns and trends in databases and using that information to build predictive models. The use of data mining in healthcare is growing increasingly popular. Nowadays, the healthcare industry generates a lot of complex information about patients, hospitals, disease diagnosis, electronic patient records, and devices. Healthcare professionals can use data mining to uncover hidden patterns in this processed data and subsequently make better decisions based on the knowledge they acquire. Providing quality healthcare implies diagnosing diseases correctly & providing effective treatments to patients. A misdiagnosis can have devastating repercussions, which are unacceptably dangerous.

Many factors, such as cholesterol level, blood pressure, smoking habits, pulse rate, age, diabetes, and more, contribute to heart disease. Machine learning, data mining, and neural network approaches are being used to identify the disease. Various Machine Learning models such as K-Nearest Neighbour, Nave Bayes, Decision Tree, and Support Vector Machines are introduced to determine the threat level these diseases possess, but data from the last few decades show that many people have these diseases at an early stage, and even some new born children suffer or die from heart disease. Machine Learning aspects can play a significant role in predicting these diseases and the threat level they pose.

The development of information technology, the coordination of frameworks, and additionally advancement in software have formed a multifaceted era of computer framework. Information technology specialists are now faced with a few challenges.
The healthcare services framework is one example of such a framework. Recently, there has been a heightened emphasis on utilizing data mining advances in healthcare frameworks. The present effort aims to determine the aspects of using healthcare data for the benefit of people by implementing methods of machine learning in combination with data mining. By analysing earlier data, we propose an automated system for diagnosing heart diseases.

Symptoms of Heart Attack:

- Discomfort, weight, swelling, or pain in the midsection, arm, or beneath the breastbone.
- Discomfort in the back, jaw, throat, or arm.
- Fullness, heartburn, or a suffocating sensation (may feel like indigestion).
- Sweating, nausea, heaving, or shakiness.
- Extreme inadequacy, nervousness, or shortness of breath.
- Heartbeats that are too fast or too slow.

Types of Heart Disease:

Heart disease is a broad term that encompasses a variety of illnesses that affect various parts of the heart. The heart represents "cardio." As a result, all heart diseases fall under the category of cardiovascular diseases.

A few sorts of Heart illnesses are:

I. **Coronary illness**: It is the most well-known type of coronary illness in the world and is also known as coronary supply route malady (CAD). It is a condition in which plaque deposits clog the coronary veins, resulting in a reduced supply of blood and oxygen to the heart.

II. **Angina pectoris:** It is a medical term for pain in the midsection caused by a lack of blood supply to the heart. It is a warning sign of a heart attack and is also known as angina. The midsection torment occurs in short bursts of a few seconds or minutes.

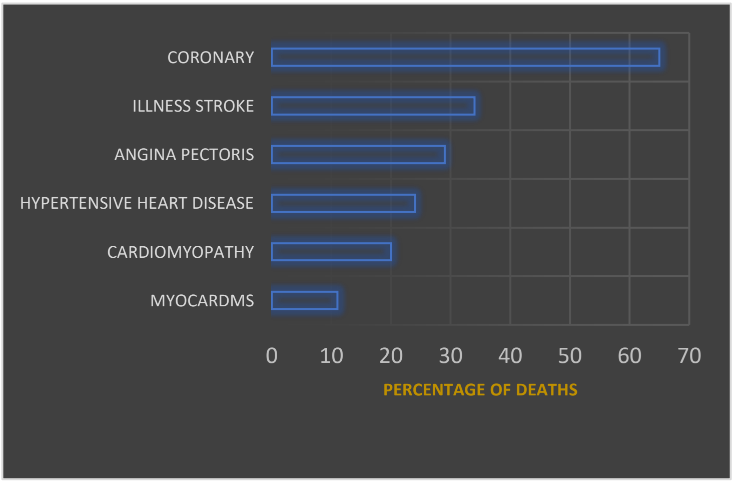
III. **Congestive heart disappointment:** It is a condition in which the heart is unable to pump enough blood to the rest of the body. It is commonly referred to as heartbreak.

IV. **Cardiomyopathy:** It is the weakening of the heart muscle or a change in the structure of the muscle as a result of insufficient heart pumping. Hypertension, alcohol use, viral diseases, and hereditary flaws are a few of the common causes of Cardiomyopathy.

V. **Innate coronary illness:** It refers to the development of an irregular heart as a result of a deformity in the structure or function of the heart. It is also a type of innate ailment that children are born with.

VI. **Arrhythmias:** It is related to a problem in the musical development of the pulse. The pulse can be slow, fast, or unpredictable. These unusual heartbeats are caused by a short in the electrical framework of the heart.

VII. **Myocarditis:** It is an aggravation of the heart muscle caused by common, parasitic, and bacterial contaminations affecting the heart. It is a rare disease with few symptoms such as joint pain, leg swelling, or fever that cannot be treated.



Literature Review:

Globally, heart disease is the leading cause of death. In 2019, 17.9 million individuals died from cardiovascular illnesses, accounting for 32 percent of all fatalities worldwide. Of these deaths, 85% were due to heart attack and stroke. It is important to detect heart disease as early as possible so that management with counselling and medicines can begin.

Changes in lifestyle, medicine, and sometimes surgeries can all help to manage heart disease. The symptoms of heart disease can be reduced and the heart's function can be improved with the right treatment. People can be protected from heart disease with the help of predicted outcomes, and the cost of surgical treatments will be reduced.

The primary goal of this study is to create a heart prediction system. A historical heart data set can be used to discover and extract hidden disease knowledge. The goal of the heart disease prediction system is to use data mining techniques on medical data sets to aid in the prediction of heart diseases. Data mining has enormous potential in the healthcare industry because it allows health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and lower costs.

Related Work (/Problem survey/Literature review):

Heart diseases are the leading cause of death globally. An estimated 17.9 million people died from cardiovascular diseases in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. It is important to detect heart disease as early as possible so that management with counselling and medicines can begin.

Changes in lifestyle, medicine, and sometimes surgeries can all help to manage heart disease. The symptoms of heart disease can be reduced and the heart's function can be improved with the right treatment. People can be protected from heart disease with the help of predicted outcomes, and the cost of surgical treatments will be reduced.

The primary goal of this study is to create a heart prediction system. A historical heart data set can be used to discover and extract hidden disease knowledge. The goal of the heart disease prediction system is to use data mining techniques on medical data sets to aid in the prediction of heart diseases. Data mining has enormous potential in the healthcare industry because it allows health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and lower costs.

Data set description:

A dataset is a collection of individual items of related data that can be accessed individually or in combination and managed as a whole. Records are organized in some sort of data structure.

The table is the whole list of data for our project. This has patient information like age, sex, cp, trestbps, cholesterol level, restcg, thalach, exang, oldpeak, slope, ca, thal, target as each column name. The values of cp, sex, fbs, restecg, exang, target lies between 0 and 1.

S. No	Attributes	Values	Description
1	age	29-77	age in years
2	sex	0-male, 1-female	gender
3	cp	0 - asymptomatic; 1-typical angina; 2-atypical angina 3-non-anginal pain;	chest pain type
4	trestbps	Numeric value(140mm/Hg)	resting blood pressure in mm/Hg
5	chol	Numeric value(289mg/dl)	serum cholesterol in mg/dl
6	fbs	1-true, 0-false	fasting blood pressure>120mg/dl
7	restecg	0-normal, 1-having ST-T, 2-hypertrophy	resting electrocardiographic results
8	thalach	140,173	maximum heart rate achieved
9	exang	1-yes, 0-no	exercise induced angina
10	oldpeak	Numeric value	ST depression induced by exercise relative to rest
11	slope	1-upsloping, 2-flat, 3-downsloping	the slope of the peak exercise ST segment
12	ca	0-3 vessels	number of major vessels coloured by fluoroscopy
13	thal	0: < 50% diameter narrowing 1: > 50% diameter narrowing	thalassemia
14	target	1 = no, 0= yes	diagnosis of heart disease (angiographic disease status)

The dataset we used here in our project is the categorization of different health reports of the patient. We took the dataset from the Kaggle website. The dataset shown in our project has 1025 rows × 14 columns. This table is called a data frame.

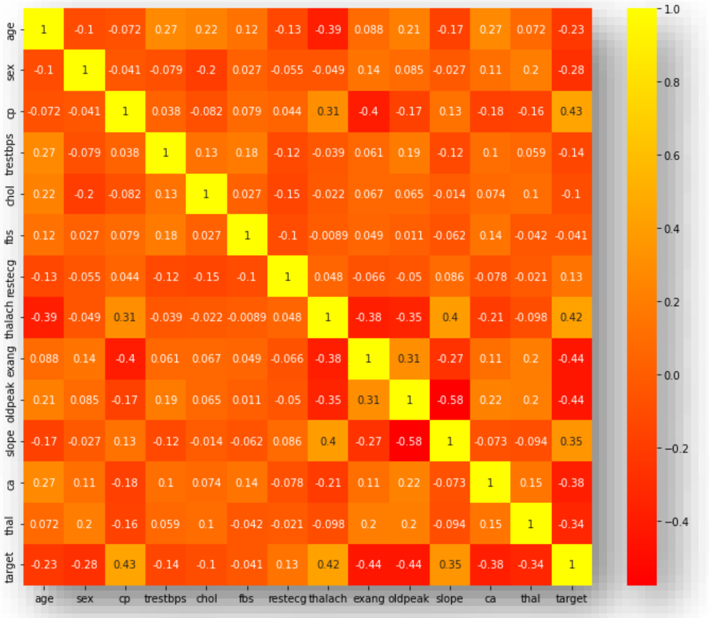
One occurrence in a record is called a data point. The label class in this case is a genre hint, a special attribute that is expected based on all input attributes. An identifier is a unique property used to find or identify a particular record. Identifiers are often used as a lookup key to group different records together. These do not provide useful information for creating data science models and should be avoided during the actual modelling phase.

Data Pre-processing:

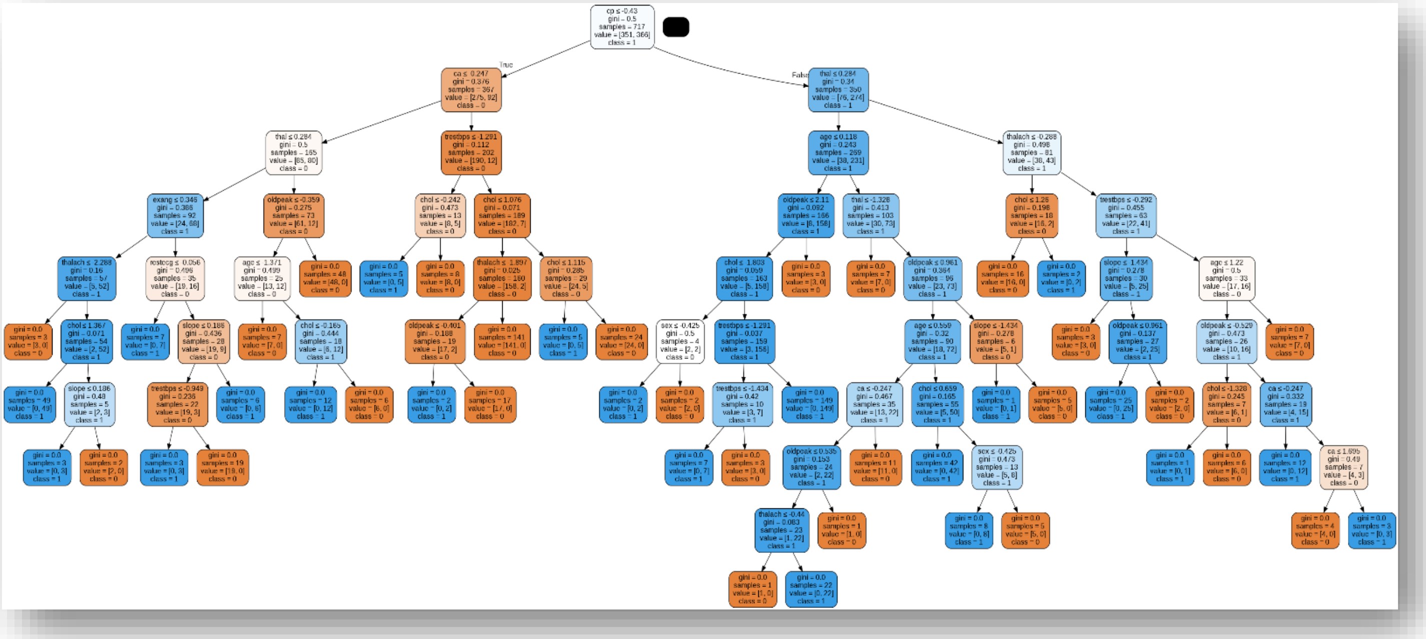
It is the first step in the data mining process. It converts the data into a format that can be processed more easily and effectively. We pre-process the data because it improves data quality by removing inconsistencies and incomplete data from the database. It improves the quality of the data, which in turn improves the quality of the mining results. The data is pre processed by checking if they have any null value.

Results:

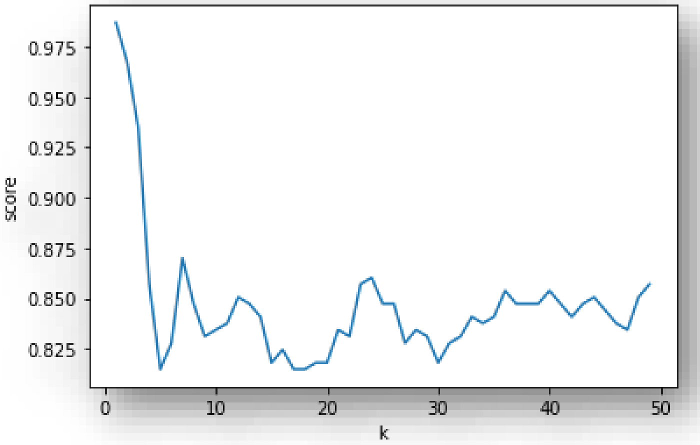
Correlation matrix for the heart data set:



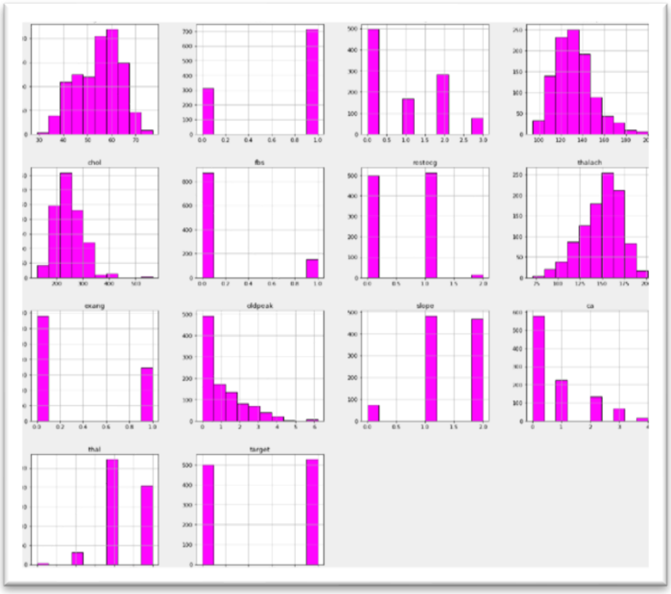
Decision tree for the heart data set:



K-neighbours score for 50 data points:



Histogram for every attribute:



CONCLUSION:

In this project, we mentioned how data mining techniques can be used to uncover hidden patterns in order to make better decisions in the healthcare field. We concentrated on data mining classification methods utilized in data discovery. For data categorization and knowledge extraction, many data mining classification algorithms offer advantages and disadvantages. Furthermore, decision trees can be examined in greater depth in order to develop a useful algorithm for healthcare organizations. In this project, we implemented different algorithms such as Logistic regression, Support vector classifier (SVC), K neighbour, Naive Bayes, and Decision tree. Logistic regression improves the performance, pre-processing of corpus like Cleaning, finding the missing values are done. SVC machine learning technique is a supervised learning model with associated learning algorithms that analyse data for classification and regression analysis. It is mainly used for small data. K nearest neighbours is a non-parametric method as there is not a particular finding of parameters to a particular functional form. It does not make any type of assumptions about the features and output of the dataset. A decision tree is used to retrieve the details associated with each patient. Based on the accurate result prediction, the performance of the system is analysed. The Naive Bayes Classifier is a simple and effective Classification algorithm that aids in the development of fast machine learning models capable of making quick predictions. It predicts the class of the test data set is simple and quick. It also excels at multi-class prediction. When the assumption of independence is met, a Naive Bayes classifier outperforms other models such as logistic regression and requires less training data. When compared to numerical variables, it performs well with categorical input variables(s). From these, we can choose the KNN algorithm because, by comparison, of all the algorithms accuracy the highest accuracy is the KNN algorithm. So, we consider the KNN is the best prediction algorithm. Comparing the remaining algorithms by involving features, the KNN algorithm is much faster than other algorithms. As a result, the KNN algorithm does not require training before making predictions, new data can be added without affecting the algorithm's accuracy. To put it another way, there is no training period for it. It saves the training dataset and only learns from it when making real-time predictions. As a result, the KNN algorithm is much faster than other algorithms that require training, such as SVM, Linear Regression, etc.

References:

1. **https://www.ijrte.org/wp-content/uploads/papers/v8i2S3/B11630782S319.pdf**

2.**https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download**