Cardiovascular disease prediction

TEAM: Sai Abhilash Bojja
Jaisuprabhath Reddy Pichili
Prudhvi Chowdary

Abstract :

Heart disease is one of the most fatal problems in the whole world, which cannot be seen with a naked eye and comes instantly when its limitations are reached. Therefore, it needs accurate diagnosis at an accurate time. Health care industry produces a huge amount of data every day related to patients and diseases. However, this data is not used efficiently by the researchers and practitioners. Today the healthcare industry is rich in data however poor in knowledge. There are various data mining and machine learning techniques and tools available to extract effective knowledge from databases and to use this knowledge for more accurate diagnosis and decision making. Increasing research on heart disease predicting systems, it becomes significant to summarize the completely incomplete research on it. The main objective of this research paper is to summarize the recent research with comparative results that has been done on heart disease prediction and also make analytical conclusions. From the study, it is observed Naive Bayes with Genetic algorithm; Decision Trees and Artificial Neural Networks techniques improve the accuracy of the heart disease prediction system in different scenarios. In this paper commonly used data mining and machine learning techniques and their complexities are summarized

PROBLEM SURVEY**:**

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If the function of the heart fails, the brain and various other organs will stop working, and within a few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases. Heart diseases have emerged as one of the most prominent causes of death all around the world. According to the World Health Organisation, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India has lost up to $237 billion, from 2005-2015, due to heart related or cardiovascular diseases. Thus, feasible and accurate prediction of heart related diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful.

Dataset Description:

A Data set is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable.

Data sets describe values for each variable for unknown quantities such as height, weight, temperature, volume, etc of an object or values of random numbers. The values in this set are known as a datum. The data set consists of data of one or more members corresponding to each row. The Dataset is taken from the Internet which is available in Kaggle.

age - age in years

sex - (1 = male; 0 = female)

cp - chest pain type

trestbps - resting blood pressure (in mm Hg on admission to the hospital)

chol - serum cholestoral in mg/dl

fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg - resting electrocardiographic results

thalach - maximum heart rate achieved

exang - exercise induced angina (1 = yes; 0 = no)

oldpeak - ST depression induced by exercise relative to rest

slope - the slope of the peak exercise ST segment

ca - number of major vessels (0-3) colored by flourosopy

thal - 3 = normal; 6 = fixed defect; 7 = reversable defect

target - have disease or not (1=yes, 0=no)

Pre-processing:

Pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends, and is likely to contain many errors.

Pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

```
df = pd.read_csv("heart.csv")
df.head()
sns.countplot(x="target", data=df, palette="bwr")
plt.show()

countNoDisease = len(df[df.target == 0])
countHaveDisease = len(df[df.target == 1])
print("Percentage of Patients Haven't Heart Disease: {:.2f}%".format((countNoDisease / (len(df.target))*100)))
```

```
print("Percentage of Patients Have Heart Disease: {:.2f}%".format((countHaveDisease / (len(df.target))*100)))

sns.countplot(x='sex', data=df, palette="mako_r")
plt.xlabel("Sex (0 = female, 1= male)")
plt.show()

countFemale = len(df[df.sex == 0])
countMale = len(df[df.sex == 1])
print("Percentage of Female Patients: {:.2f}%".format((countFemale / (len(df.sex))*100)))
print("Percentage of Male Patients: {:.2f}%".format((countMale / (len(df.sex))*100)))

df.groupby('target').mean()


pd.crosstab(df.age,df.target).plot(kind="bar",figsize=(20,6))
plt.title('Heart Disease Frequency for Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('heartDiseaseAndAges.png')
plt.show()
```

Flow Chart:

Goes this way

Start -> collecting heart disease from the dataset -> Extract Significant Variables->Data Preprocessing -> Splitting ->

->Training Data->Training (using various data mining technologies)

-> Testing Data

From above two we classify data -> and then we Compute Accuracy

=>Evaluation and analysis Result