**ABSTRACT**

Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. 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. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, logistic regression and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

*Keywords: Machine Learning, Logistic regression, Cross-Validation, Backward Elimination, REFCV, Cardiovascular Diseases.*

**TABLE OF CONTENTS**

**CHAPTERS PAGE NO**

**1. INTRODUCTION**

1.1 INTRODUCTION TO PROJECT

1.2 STATEMENT OF THE PROBLEM

1.3 BRIEF DESCRIPTION OF THE PROJECT

1.4 SOFTWARE AND HARDWARE SPECIFICATION

**2. LITERATURE SURVEY  
3. SYSTEM ANALYSIS**

3.1 EXISTING SYSTEM

3.2 LIMITATIONS OF EXISTING SYSTEM

3.3 PROPOSED SYSTEM

3.4 ADVANTAGES OF PROPOSED SYSTEM

3.5 FEASIBILITY STUDY

- TECHNICAL FEASIBILITY

- ECONOMICAL FEASIBILITY

- OPERATIONAL FEASIBILITY

4. **SYSTEM DESIGN AND DEVELOPMEN**

4.4 DATAFLOW DIAGRAM

4.2 DATASETS

4.3 ATTRIBUTS

4.11 MODULE DESCRIPTION

4.5 METHODOLOGIES

4.6 TESTING OR EVALUATION METRIX

4.7 DISCUSSION ON RESULTS

**5. CODING**

5.1 PSEUDO CODE

**6. SOFTWARE TESTING**

**7. CONCLUSION**

**8. FUTURE ENCHANCEMENT**

**BIBILOGRAPHY**

**1. INTROUDUCTION**

**1.1 INTROUDUCTION TO THE REPORT**

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. 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 reduce the complications. 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.

**STATEMENT OF THE PROBLEM**

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 overall 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.

**BRIEF DESCRIPTION OF THE PROJECT**

Among all fatal disease, heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don’t give enough rest for themselves and eat what they get and they even don’t bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease

**SOFTWARE AND HARDWARE SPECIFICATION**

|  |  |
| --- | --- |
| OS: windows 8 and above  IDE: jupyter, kaagle  Language: python  Browser: chrome, Edge  RAM: 4GB  ROM: 20GB |  |
|  |  |
|  |  |
|  |  |
|  |  |

**2. LITERATURE SURVEY**

Machine learning is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

**3. SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

Todays world are almost upgraded to new technology, but still for hear disease prediction test some of them are using manual testing , so it needed large amt of time and money. Even we can’t get accurate results from this tests.

**3.2 LIMITATIONS OF EXISTING SYSTEM**

1. require large time

2. large amount

3. less accuracy

**3.3 PROPOSED SYSTEM**

In this system we are implementing effective heart attack prediction system using Naïve Bayes algorithm. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is Naïve Bayes**.** After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

**3.4 ADVANTAGES OF PROPOSED SYSTEM**

1. less time

2. less money

3. more accuracy

4. easy identify disease

**3.5 FEASIBILITY STUDY**

The feasibility study proposes one or more feasible conceptual solutions to the problem set of the project. The conceptual solutions give an idea of what the new system will Look like. They indicate what inputs are needed by the system and what outputs will be produced. Three things to be done to established feasibility. First, it must be checked that the project is technically feasible. Second, operational feasibility must be established. For this, it is necessary to consult the system users to see if the proposed solution satisfies user objectives and can be fitted in to current system operation. Third, economic feasibility must be checked. The study must determine whether the project’s goal can be achieved within the resource limits allocated to it. It must also determine whether it is worthwhile to proceed with the project at all or whether the benefits obtained from the new system are not worth the cost, in which case the project will be terminated.

Feasibility study is necessary to determine whether the proposed system is feasible considering the technical, operational and economical factors. By having detailed feasibility study one can have a clear view of the proposed system with respect to its benefits and draw backs.

For a successful feasibility study of system feasibility, the existing systems and proposed system are studied carefully.

**System Feasibility**

The feasibility study is carried out to determine whether the proposed system can be developed with the available resources.

* Technical Feasibility
* Economical Feasibility
* Behavioral Feasibility
* Motivational Feasibility
* Schedule Feasibility
* Operational Feasibility

**Technical Feasibility**

Technical feasibility is the study of resource availability that may affect the ability to achieve an acceptable system. Technical feasibility is the most difficult area to ensure at initial stages. Since the objectives functions and performance cannot be predicted to its fullest, everything seems possible provided proper assumptions are made. It is essential that the process of technical feasibility. The consideration that is normally associated with technical feasibility included resource availability at the organization where the project is to be developed and implemented.

**Economical Feasibility**

An evaluation of development cast weighted against the ultimate income or benefit derived from the developed system. Economical economic justification includes a broad range of concerns that include cost-benefit analysis. Cost benefit delineates costs for project development and weighs them against tangible and intangible benefits of a system. Regarding the cost and benefits, the project, which is to man-hours with compared to man that are required to record data about activity task report manually and also in terms of money benefits by the selling of this system as a product. Thus this project work is economically feasible for the development in any company.

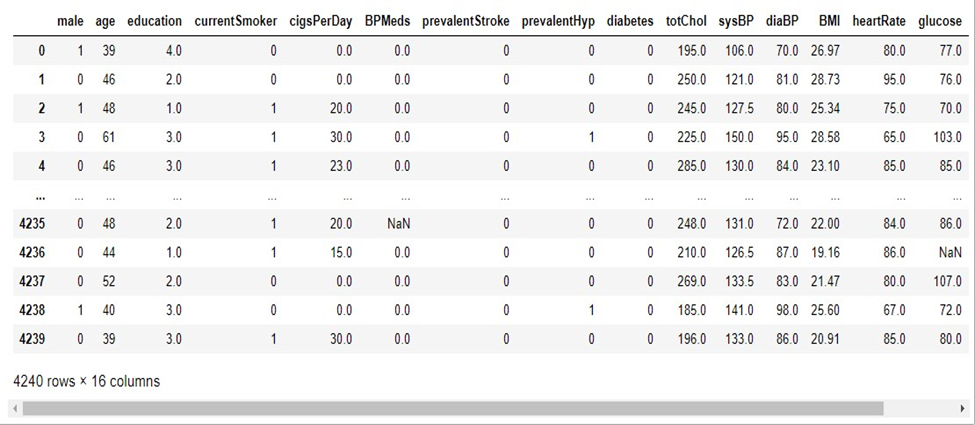
4. **SYSTEM DESIGN AND DEVELOPMEN**

**4.1 DATAFLOW DIAGRAM**



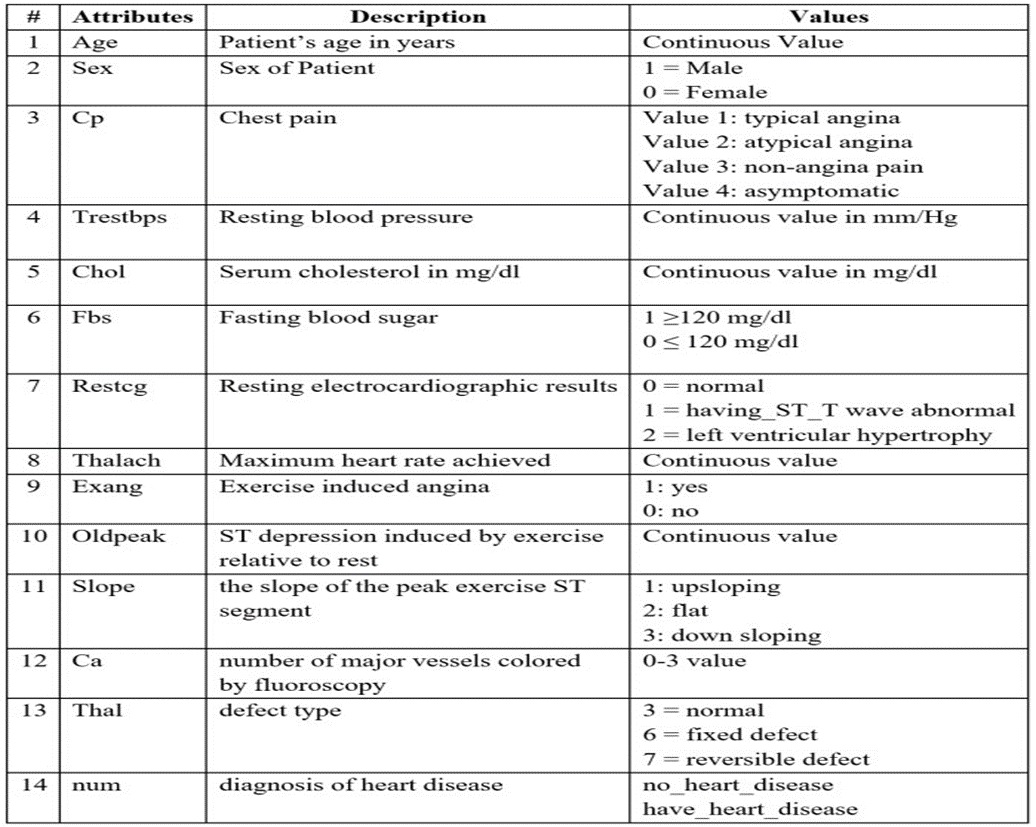
**4.2 DATASETS**

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

****

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further with this dataset pre-processing and experiments are then carried out.

**4.3 ATTRIBUTES**

****

**4.4 MODULES DESCRIPTION**

**DIMENSIONALITY REDUCTION**:

Dimensionality Reduction helps to picking a mathematical representation like that one can relate the maximum of, but not all, the variance inside the given data, so including only major significant data. The data identified for a problem , may consist of many attributes. But not all the dimensions may equally influence the output. A most number of dimensions, may affect the computational complexity and then lead to overfitting then it leads to poor result. It is a very major step considered while building any model. It can achieved by, a)feature extraction and b) feature selection.

**FEATURE EXTRACTION:**

Here , a new set of features are derived from the base features set. It involves in a transformation of the dimensions. This transformation is not reversible as less or few or may many , usable data is erased in the process , in [3] and [4] PCA is used for feature extraction. PCA mainly used for linear transformation algorithm. The feature space , it finds the way that many variances and also finds an way that are mutually orthogonal. It is famous for reconstructing the data.

**FEATURE SELECTION**:

Here, subset of base features sets are selected. CFS(correlation based feature selection) helps to subset evaluation with best first search method to reduce dimensionality and select the most significant feature using chi square method.

**4.5 METHODOLOGIES**



**B) NAÏVE BAIS**

It is a supervised algorithm and simple classification algorithm using bayes theorem . it is a mathematical concept and helps obtain a probability. Its neither related to each other nor corelation to one a other. It provide a path to calculating posterior probability.

**C) Support Vector Machine**

It can be used for classification as well as regression also, but most of the time used in classification problem. SVM basically a differen classes in a hyperplane in multidimensional space. The goal is to split the dataset into 2 different classes to find a max marginal hyperplane.

**D) K-Nearest Neighbour**

It is a method of supervised learning and classifies the things dependant on near neighbour. It using the Euclidean distance to find the distance between 2 object and then data are clustered based on similarity. Its simple to carry out because its not going to making assumption or create model. It is one of the versatile and used for regression ,classification and even search, but the minus point is noise and irrelevant features affect its accuracy.

**E) Decision Tree**

It works on categorical and numerical data. It is used to create a tree structure. Its going to splits the data into 2 or more analogous sets based on some important indicators. First its create the entropy to each attributes then calculate, and data are divided, predictors having maximum entropy and information gain: results are easy to read and interpret. It has higher accuracy rate then compare to other model, because it analyse the data like in tree structure. The disadvantage is data are over classified at a time.

**F) Random Forest**

Here, many trees are create a forest. class with most votes build a models forest and which one having a more number of trees is the accuracy. It is slow to obtain prediction.

**4.6 TESTING OR EVALUATION METRIX**

**Confusion Matrix**

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made.

|  |  |
| --- | --- |
| TP=3569 | FP=27 |
| FN=599 | TN=45 |

*Table 1: Confusion Matrix Obtained after training the data (feature selection by backward elimination)*

|  |  |
| --- | --- |
| TP=3582 | FP=14 |
| FN=600 | TN=44 |

*Table 2: Confusion Matrix Obtained after training the data (feature selection by RFECV method)*

**Accuracy** 𝑇𝑃+𝑇𝑁

The accuracy is calculated as: Accuracy =

𝑇𝑃+𝑇𝑁+𝐹𝑃+𝐹𝑁

Where,

* True Positive (TP) =Observation is positive, and is predicted to be positive.
* False Negative (FN) = Observation is positive, but is predicted negative.
* True Negative (TN) = Observation is negative, and is predicted to be negative. • False Positive (FP) =Observation is negative, but is predicted positive

The obtained accuracy during training the data after feature selection using backward elimination was 86 % and during testing was 83%.

The obtained accuracy during training the data after feature selection using REFCV method was 86 % and during testing was 85 %.

**Recall**

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN). Recall is calculated as:

Recall = 𝑇𝑃

𝑇𝑃+𝐹𝑁

The obtained recall during training the data after feature selection using backward elimination was and during testing was 0.99.

The obtained recall during training the data after feature selection using REFCV method was 1.00 and during testing was 0.99.

**Precision**

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP). Precision is calculated as: Precision = 𝑇𝑃

𝑇𝑃+𝐹𝑃

The obtained precision during training the data after feature selection using backward elimination was 0.86 and during testing was 0.84.

The obtained precision during training the data after feature selection using REFCV method and during testing was 0.86

**4.7 DISCUSSION ON RESULTS**

When performing various methods of feature selection, testing it was found that backward elimination gave us the best results among others. The various methods tried were Backward Elimination with and without KFold, Recursive Feature Elimination with Cross Validation. The accuracy that was seen in them ranged around 85% with 85.5% being maximum. Though both methods gave similar accuracy but it was seen that in Backward Elimination we found that the number of misclassifications of True Negative was more and it was observed that the accuracy had more variance compared to RFEV. The precision of Backward Elimination and RFEV are 84% and 86% respectively. And the recalls are 0.99 and 1 respectively. The precision and recall also shows that the number of misclassifications is less in RFECV than in Backward Elimination.

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| Logistic regression | 75.41% |
| SVM | 77.05% |
| Naïve baise | 73.77% |
| Random forest | 86.89% |
| Decision tree | 77.05% |
| KNN | 75.83% |
|  |  |

**5.CODING**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

%matplotlib inline

sns.set\_style("whitegrid")

plt.style.use("fivethirtyeight")

df = pd.read\_csv("/kaggle/input/heart-disease-uci/heart.csv")

df.head()

df.info()

pd.set\_option("display.float", "{:.2f}".format)

df.describe()

df.target.value\_counts().plot(kind="bar", color=["salmon", "lightblue"])

Out[7]:

df.isna().sum()

categorical\_val = []

continous\_val = []

for column in df.columns:

print('==============================')

print(f"{column} : {df[column].unique()}")

if len(df[column].unique()) <= 10:

categorical\_val.append(column)

else:

continous\_val.append(column)

plt.figure(figsize=(15, 15))

for i, column in enumerate(categorical\_val, 1):

plt.subplot(3, 3, i)

df[df["target"] == 0][column].hist(bins=35, color='blue', label='Have Heart Disease = NO', alpha=0.6)

df[df["target"] == 1][column].hist(bins=35, color='red', label='Have Heart Disease = YES', alpha=0.6)

plt.legend()

plt.xlabel(column)

*# Create another figure*

plt.figure(figsize=(10, 8))

*# Scatter with postivie examples*

plt.scatter(df.age[df.target==1],

df.thalach[df.target==1],

c="salmon")

*# Scatter with negative examples*

plt.scatter(df.age[df.target==0],

df.thalach[df.target==0],

c="lightblue")

*# Add some helpful info*

plt.title("Heart Disease in function of Age and Max Heart Rate")

plt.xlabel("Age")

plt.ylabel("Max Heart Rate")

plt.legend(["Disease", "No Disease"]);

*# Let's make our correlation matrix a little prettier*

corr\_matrix = df.corr()

fig, ax = plt.subplots(figsize=(15, 15))

ax = sns.heatmap(corr\_matrix,

annot=True,

linewidths=0.5,

fmt=".2f",

cmap="YlGnBu");

bottom, top = ax.get\_ylim()

ax.set\_ylim(bottom + 0.5, top - 0.5)

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

def print\_score(clf, X\_train, y\_train, X\_test, y\_test, train=True):

if train:

pred = clf.predict(X\_train)

clf\_report = pd.DataFrame(classification\_report(y\_train, pred, output\_dict=True))

print("Train Result:**\n**================================================")

print(f"Accuracy Score: {accuracy\_score(y\_train, pred) \* 100:.2f}%")

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print(f"CLASSIFICATION REPORT:**\n{clf\_report}**")

print("\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_")

print(f"Confusion Matrix: **\n** {confusion\_matrix(y\_train, pred)}**\n**")

elif train==False:

pred = clf.predict(X\_test)

clf\_report = pd.DataFrame(classification\_report(y\_test, pred, output\_dict=True))

print("Test Result:**\n**================================================")

print(f"Accuracy Score: {accuracy\_score(y\_test, pred) \* 100:.2f}%")

5.1 SNAPSHOTS

Chart, waterfall chart

Description automatically generated

**A screenshot of a computer

Description automatically generated with medium confidence**

**Chart, scatter chart

Description automatically generated**

**A screen shot of a computer

Description automatically generated with low confidence**

**6. CONCLUSION**

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project resolved the feature selection i.e. backward elimination and RFECV behind the models and successfully predict the heart disease, with 85% accuracy. The model used was Logistic Regression. Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced models.

**7. FUTURE ENHANCEMENT**

The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

**8. REFERENCES**

[1]A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

[2]M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3]K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c. [Accessed 2 March 2020].