A PROJECT REPORT

on

“Heart Disease Detection”

By

ANKIT ANSHUMAN MOHAPATRA - 2006006

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
May 2020

ABSTRACT

Over 17.9 million people die annually due to Cardiovascular Diseases(CVDs), representing 32% of all global deaths. Of these deaths, 85% are directly caused by heart attack and stroke. Earlier Machine learning algorithms were used to detect whether people are suffering from cardiovascular diseases by considering certain attributes like chest pain, cholesterol level, age of the person & resting blood pressure levels. Previous researchers were focused on identifying the significant features to heart disease prediction, however, less importance was given to contributing features that would help in identifying, visualizing or predicting the outcome to these features.

This paper emphasizes on the development of an artificial intelligence-based heart disease diagnosis system using deep learning & machine learning algorithms. This study measures the strength of the contributing features such as FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope that were previously been neglected by research papers, aimed at predicting heart disease based on the scores of the contributing features.Here in this research I predicted the scores of different algorithms contributing to heart disease prediction using the 300+ sample & 14 attribute, The University of California Irvine Heart Disease Dataset. The diagnostic application achieves highest confidence score/accuracy of 88.52% in predicting heart disease. This research has managed to provide a leap through in predicting heart disease using machine learning & artificial intelligence.

Keywords:
Heart Disease Prediction, Cardiovascular Diseases (CVDs), Machine Learning Algorithms (KNN, SVM, Random Forest & Decision Tree), Deep Learning (Artificial Neural Network), The University of California Irvine Dataset.

Section - 1 Introduction

Cardiovascular Diseases(CVDs) are often referred to as heart disease. These diseases often refer to clogged or narrowed arteries that lead to stroke, chest pain or angina, and heart disease. Other types of heart disease, such as those that affect the heartbeat, valves, or muscles, are other types of heart disease. Machine learning, on the other hand, is important in determining whether a person has heart disease. But if these are anticipated, it will be easier for doctors to access important information for treating and diagnosing patients. Heart disease is often a false symptom of coronary artery disease. Also another known CVD, which is a disease of the blood vessels.

Here I have used Python, which is a content-oriented programming language that is well-designed and can be developed quickly. According to my analysis it is considered one of the best programming language and has many applications in medicine. It is also considered a popular and well-received language with applications driven by AI-based software development and many other web applications. As mentioned the python framework is easy to use for building desktop or web-based applications. According to the use of python programming in medicine, especially in cardiac diagnosis, doctors and hospitals can better benefit patients by using scalable and dynamic. However, the coding packages and libraries used in this project are Pandas, Matplotlib, IPython, Numpy, Python, Seaborn, etc.

Data mining provides many techniques to discover hidden patterns or s imilarities in data. Therefore, in this paper, a machine learning algorithm is proposed, which will be used to a validate heart disease prediction system on two open-access disease datasets.

Section - 2
Basic Concepts/ Literature Review

This project involves detecting heart disease using Python via ML & AI. The data set include 14 attributes. Project, matplotlib, Numpy, Pandas, alert etc. It uses several other libraries such as Correlation matrices, histograms, support vector classifiers, K-neighbor classifiers, random forest classifiers, and decision tree classifiers are used to evaluate the results of the data set using the Python programming language. In addition, Python is considered an open language that supports the development of new solutions for healthcare and provides better outcomes for patients, thereby improving health.

One of the best-known machine learning algorithm tasks is data classification. In this context, machine learning is often important for extracting information from business data and transferring it to larger datasets. Most machine learning methods rely on these features that directly or indirectly affect the complexity of the model to explain the behavior of the algorithm.

The model will also be deployed in this project. After the data is loaded and saved in the variable used to copy the data. Finally, the dataset will be exported to file and processed. However, when the results are examined, it is seen that the K-neighbor classifier algorithm scores 88.5%, while the support vector, decision tree, and random forest classifiers score 83%, 81%, and 85%.

Section - 3

Problem Statement / Requirement Specification

Over 17.9 million people die annually, due to CVDs. So it is the need to build an efficient Machine Learning model, having the maximum accuracy to predict whether a person is diagnosed with heart disease or not! So in this section this SRS document gives the model planning, purpose, analysis & design that we have built.

3.1 **Model Planning**

The formulation of the model begins with gathering data from past research papers & creating a improved, idealistic & reliable model. Then we fulfilled the crucial requirement by selecting The University of California Irvine Heart Disease Data set & further optimized it.

3.2 **Project Analysis**

After the requirements are collected then it is loaded for Exploratory Data Analysis, under which we further normalized / standardized the data. Then using various models & techniques we optimized the models to get best accuracy.

3.3 **System Design**

We Further Organized the data to perform Visualizations depicting the relationship between the attributes. Visualization gives a visual summary of the data set and customized plots such as Scatter plot, Missing number Matrix & Bar graphs are to embed into the applications, Here library’s like Seaborn & Matplotlib is primarily used for statistical calculations on plots.

Section - 4 Data-Set (UCI)

Here we are operating on The University of California Irvine data set, which is easy and effective to use, the UCI dataset uses 14 factors such as gender, age, cp (type of chest pain), cholesterol level, fasting blood sugar level, exercise causes angina, maximum heart rate (talaq), Done. maximum heart rate (thal) and major arteries. These biological parameters often represent heart health, and using them for data-driven prediction will lead to good models. There are about 300 models.

Here are it’s 14 attributes used:
1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

Section - 4 Implementation

In this section, I presented the implementation done by me during the project development.

4.1 Methodology

This section explains proposed process and shows all the materials, methods and tools used to create the entire process. Creating an intelligent and user-friendly heart disease predictor requires efficient tools to train large datasets and compare different machine learning algorithms. After choosing the most robust KNN algorithm which gives the maximum efficiency of 88.5, Which i will deploy in a web application using streamlit via the spyder IDE.

4.2 Flow Diagram

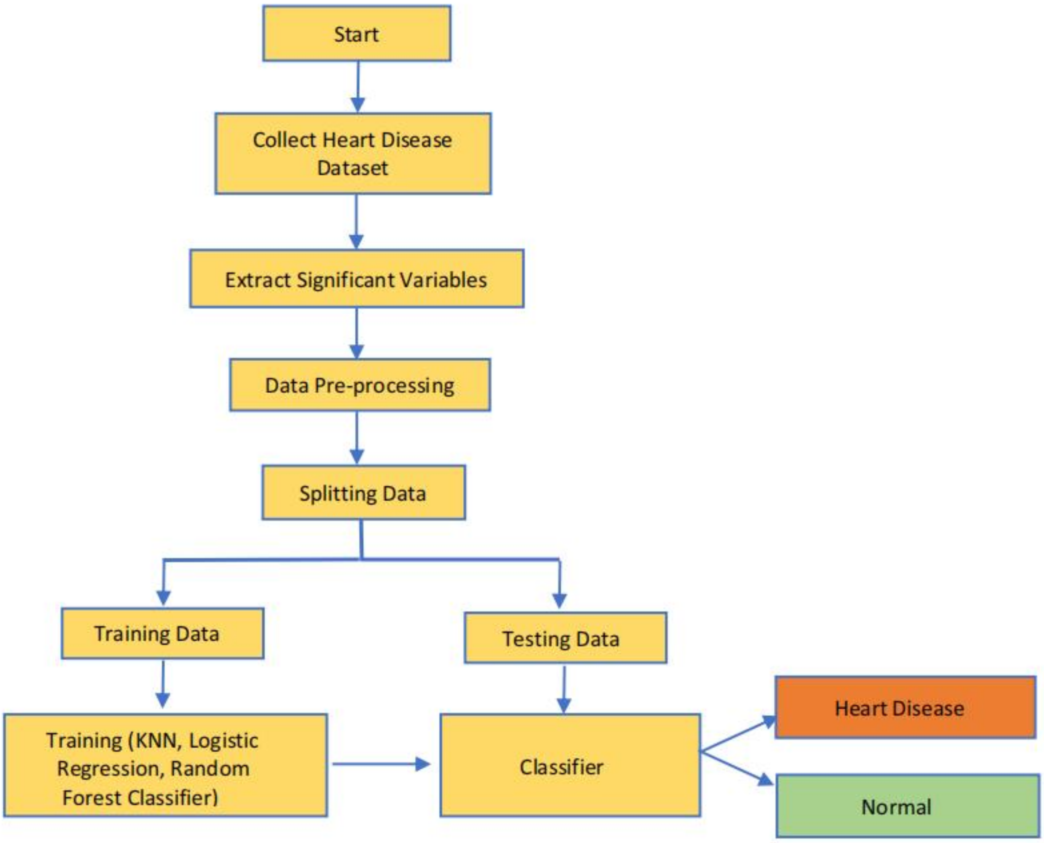
Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Splitting Data
Training Data
Testing Data
Training (KNN, Logistic Regression, Random Forest Classifier)
Classifier
Heart Disease
Normal

Figure 1. Proposed Model

4.3 Proposed Algorithms

K-Neighbors Classifier The k-Neighbors classifier classifies vectors according to most of their neighbors. The classification of this classification is very important. The distance between this vector and its neighbors is denoted by d and the weight is given as 1/d.

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Decision tree classifiers create trees from observational data. This file has branches attached to the branches, which is its main purpose for making leaves. The correct value for the decision tree model is calculated by entering the XTrain parameter and the appropriate image, YTrain. The score of the decision tree is then found and the parameters XTest and YTest are passed around the system score() function, which looks for the score of the decision tree.

Random forest distribution is considered as a learning method for solving learning problems such as distribution and retrieval. This random forest distribution for the heart attack algorithm works by generating multiple decision trees. Among other things, this method uses a technique called "bootstrap aggregation".

$$ni_j = w_j C_j - w_{\mathrm{left}(j)} C_{\mathrm{left}(j)} - w_{\mathrm{right}(j)} C_{\mathrm{right}(j)}$$

Support Vector Machine or SVM is a technique used to classify linear and nonlinear data. It uses a non-linear mapping method so the can transform the training data into longer ones. The hyperplane is a kind of line that separates the variable input space in SVM. The hyperplane can separate points of different ideas that have classes in space, for example 0 or 1

Artificial Neural Networks or ANN also known as Multilayer Perceptrons, are said to be inspiring and have the ability to model poor performance errors. Artificial neural networks are one of the important tools used in machine learning. As the "Neural" name suggests, these are brain-driven systems designed to replicate the way humans learn.

Section - 5 Result Analysis

5.1 Result discussion

After evaluating the models, we found KNN to be the best performer. So we summarize that that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Finally, we deployed it using streamlit using spyder IDE. It it turns out that we are successfully able to predict, whether the person is diagonosed with Heart disease or not.

5.2 Tabular analysis

Model_IDs	***Algorithms***	***Accuracy***
1	KNN	88.52
2	SVM	86.88
3	Random Forest	85.24
4	ANN	83.60
5	Decision Tree	81.96

5.3 Graphs

We Further Organized the data to perform Visualizations depicting the relationship between the attributes. Visualization gives a visual summary of the data set and customized plots such as Scatter plot, Missing number Matrix & Bar graphs are to embed into the applications, Here library's like Seaborn & Matplotlib is primarily used for statistical calculations on plots or graphs.

Section - 6
Conclusion and Future Scope

6.1 Conclusion

Heart disease is one of the most devastating and fatal chronic diseases that rapidly increase in both economically developed and undeveloped countries and causes death. This damage can be reduced considerably if the patient is diagnosed in the early stages and proper treatment is provided to her. In this paper, I developed an intelligent predictive system based on contemporary machine learning algorithms for the prediction and diagnosis of heart disease.

This is a heart disease prediction project based on machine learning & Deep Learning field. While in making this project different data sets are collected.These data sets are again filtered and then suitable machine learning models are selected based on different accuracy. KNN works best (88.5% accuracy). So we used KNN.At last this model is deployed and get ready for final use. This research has managed to provide a leap through in predicting heart disease using machine learning & artificial intelligence.

6.2 Future Scope

Current research data provides the best insights into different machine learning-based heart disease diagnostic methods. This work may be modified in the future by adding more features to cardiac data and making it more interactive for users. It can also be done as a mobile application, which reduces computation time and complexity. We will update the system by connecting to the hospital database.

Section - 7 References

[1] Loku L., Fetaji B., Krstev A., Fetaji M., Zdravev Z. Using python programming for assessing and solving health management issues South East Eur. J. Sustain. Dev., 4 (1) (2020)

[2] .Overview of machine learning in healthcare Machine Learning Applications using Python, A Press, Berkeley, CA (2019), pp. 1-11

[3] Guleria P., Sood M. Intelligent learning analytics in healthcare sector using machine learning Machine Learning with Health Care Perspective, Springer, Cham (2020), pp. 39-55

[4] Ali, F., Hasan, B., Ahmad, H., Hoodbhoy, Z., Bhuriwala, Z., Hanif, M., et al.(2021). Protocol: Detection of subclinical rheumatic heart disease in children using a deep learning algorithm on digital stethoscope: A study protocol. BMJ Open 11:e044070. doi: 10.1136/bmjopen-2020-044070

[5] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, A. (2017) Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology, 10, 2137-2159.

[6] The IEEE website. [Online]. Available: http://www.ieee.org/

[7] BelItSoft Python in healthcare BelItSoft (2017) Available at-

https://belitsoft.com/custom-application-development-services/healthcare-software-development/python-healthcare , [Accessed on 5th March, 2021]

[8] Panesar A. Machine Learning and AI for Healthcare (1–73) Apress, Coventry, UK (2019)

https://iedu.us/wpcontent/uploads/edd/2020/01/Arjun_Panesar_Machine_Learning_and_AI_for_Health-iedu.us_.pdf , [Accessed on 5th March, 2021]