IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

Ashish Mark Daniel
Department of Computer Science,
SRM Institute of Science and Technology, Chennai

Srijan Srivastava
Department of Computer Science,
SRM Institute of Science and Technology, Chennai

Dr. V. Anbarasu
Department of Computer Science,
SRM Institute of Science and Technology, Chennai

Abstract- In India, the deaths of young adults due to heart related problems has become a pressing issue. The reason behind the high fatality rate of this issue is the unpredictability of heart attacks, and how time is of the essence. We could therefore, reduce the fatality rate here if somehow the person was warned beforehand about the possibility of heart related ailments, and get a check up as a preemptive measure, beforethe situation even gets more serious.

Index Terms- Heart, Disease, Prediction, Machine, learning

I. INTRODUCTION

India has one of the highest burdens of CVD(Cardiovascular disease) worldwide. Coronary heart disease rates in India in urban areas are between 1% to 13.2%. In rural areas, these rates are between 1.6% to 7.4%. The number of deaths caused by CVD has seen a significant increase in the last three decades (from 1990 to 2020), i.e from 2.26 million in 1990 to 4.77 million in 2020. We are attempting to build a system that allowspeople to keep a track of their cardiovascular health regularly, and in an efficient manner by developing a Machine Learning prediction system which warns the user about when they ought to seek a medical check-up regarding the status of their cardiovascular system. This paper includes research about various features affecting cardiovascular health, their relevance,and applying Machine Learning algorithms to the dataset. The dataset is from the UCI Library.

This paper is divided in four sections.
Starting at section I. Introduction, the purpose and objectives are discussed in brief.
Further in Section II. Research Elaborations, The data exploration and understanding of features is discussed.
In section III. Implementation and design, the implementationof the system is elaborated upon.
In section IV. Results, the results and conclusions of all the various implementations are explained.

II. Research Elaborations

There are 13 attributes for every record that are chosen out ofan original of 75, as per relevance to the implementation. The age and sex of the individual are the first two attributes which are self-explanatory. The other attributes are elaborated further:

A. Chest pain type

This attribute has 4 integral values from 1 to 4. Each numberdenotes a type of chest pain.
1 : Typical angina 2 : Atypical angina
3 : Non-anginal pain4 : Asymptomatic

B. Resting blood pressure (mm Hg)

This is the systolic blood pressure (pressure exerted on the wallsof the arteries when the heart beats) measured in millimeters of mercury.

C. Cholesterol (mg dL)

This is a measurement of the amount of cholesterol in the blood which can be obtained with a blood test. It is measured in milligrams per deciliter. A high amount of cholesterol in the blood is a good indicator of possibility of CVD.

D. Fasting blood sugar (mg dL)

Measurement of the amount of sugar in the blood sample afteran overnight fast. This is measured in milligrams per decilitre. However in this case, we use a binary value where 1 indicates avalue greater than 120 mg dL, and 0 indicates a value lowerthan 120 mg dL.

E. Resting ECG

Electrocardiogram reading measures the beats of the heart through electrodes placed on the body. This attribute takes 3 values from 0 to 2.
0 : normal
1 : ST-T Wave abnormality. (ST-T wave is a segment of the graph plotted by the electrocardiogram. Elevations and depressions of the segment point to heart related disease)
2 : left ventricular hypertrophy.

F. Maximum heart rate

The maximum heart rate measurement of the individual. It ismeasured in beats per minute.

G. Exercise induced angina

This refers to chest pain experienced by the person upon physical exertion. This is a binary value where 0 indicates 'no' and 1 indicates 'yes'.

H. ST Depression induced by exercise relative to rest In the result of a cardiac stress test, a depression of 1mm or greater indicates ischemia.

I. Slope of peak exercise ST segment

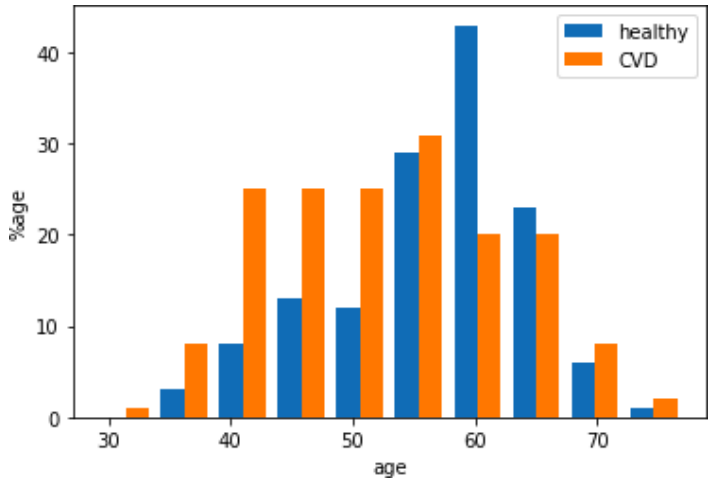
The slope of the ST segment should normally be upwards during exercise. A flat or down sloping ST segment during exercise is abnormal. This attribute takes these 3 conditions as values from 0 to 2.
0 : upsloping
1 : flat
2 : down sloping

J. Number of major vessels colored in fluoroscopy The number of major vessels of the heart colored duringfluoroscopy. This attribute takes values from 0 to 3. The morenumber of colored vessels, the higher the chance of heartdisease. If the vessel is colored, it indicates that blood is not passing through it correctly.

K. Thallium

It is a test in which a radioactive element is injected into the blood. The blood flow is then monitored. This attribute takes values between 0 to 3.
0 : null
1 : fixed defect
2 : normal blood flow3 : reversible defect

III. Data Exploration



The average age is calculated as 54.36, from the 303 available records in the dataset. However, on further analysis, it is noted that 284 of the records are of people having age greater than 40,and only 19 records of people with age less than or equal to 40. We conclude that the dataset is not distributed very well according to age, as data available for younger people is not enough to get a good prediction accuracy. It is analyzed that thesignificant majority of the people in the dataset are aged between 50 to 60 years of age. The ratio of CVD to healthy in the prediction is 45.54% healthy to 54.46% CVD.

Figure 2.1
Age-wise graph of people afflicted with CVD vs notafflicted

The above figure shows a histogram of age distribution where itshows the percentage of people afflicted with CVD in various age groups. We can conclude from this that the ratio of CVD affliction increases with increase in age.

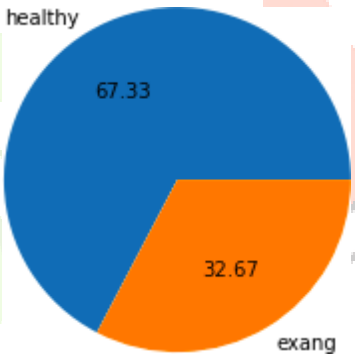


Figure 2.2 - Chest pain affliction

In figure 2.2 we can see the ratio of people who suffer from anginal pain triggered by exercise to the percentage of peoplenot suffering from this. It is concluded that about one third ofthe people are suffering from exercise induced angina.

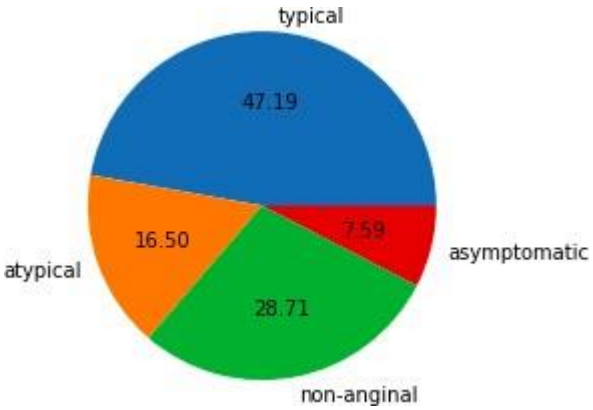


Figure 2.3 - Chest pain type

In figure 2.3, we can see the percentage comparison of people suffering from the 4 different types of chest pain mentioned in the dataset. We can conclude from here that about half of peopleare suffering from typical angina, and only about 7.6% people are asymptomatic.

IV. Implementation and Design

Four models have applied and checked for their accuracy inorder to choose the most accurate one.

A. KNN model

In the KNN algorithm, the prediction is made by calculating theEuclidean distance between two points, where one point contains the values to be predicted and the other points are the k nearest points to by, as per the calculated distance. K is a constant and can be selected while defining the model. By practice, we always choose a small odd number as the value of K as even numbers can cause confusion in the prediction.

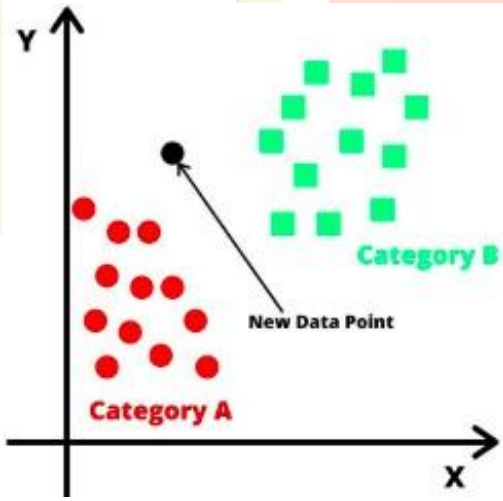


Figure 3.1
Logical visualization of KNN model

B. Decision Tree

The decision tree model is a tree like structure which dividesthe entire dataset repeatedly into two halves at every node, based on some parameter. For example, in our case, one such partition could be age > 55 which would break the dataset into two parts, one with people younger than 55, and the other with people older than 55.

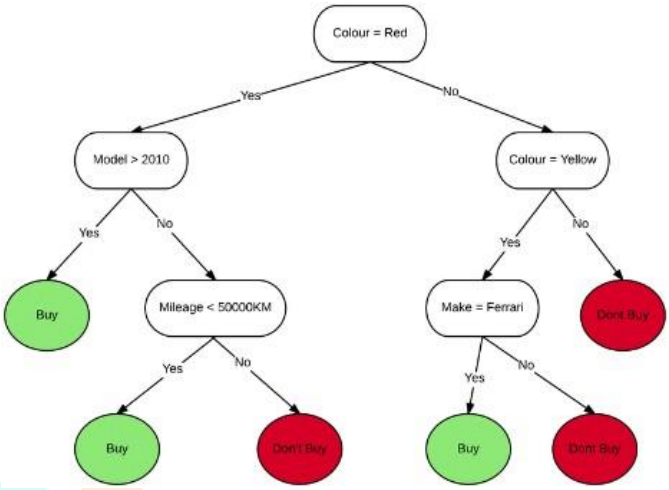


Figure 3.2
Logical visualization of Decision Tree model

C. Adaboost/Adaptive-Boost (ensemble method)

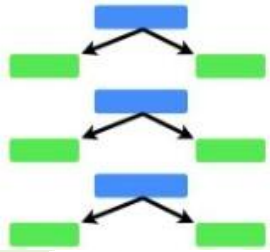


Figure 3.3
Logical visualization of Adaboost ensemble

Adaboost is an ensemble method. This means that it combines multiple models which would individually be weaker classifiersand combines them together in order to produce one strong classifier. In the case of Adaboost, the weak classifiers are decision trees, but instead of having an exponential number of splits like a sole decision tree model does, these have only one split per tree. Due to this reason, these trees are also known as decision stumps.

D. Random Forest (ensemble method)

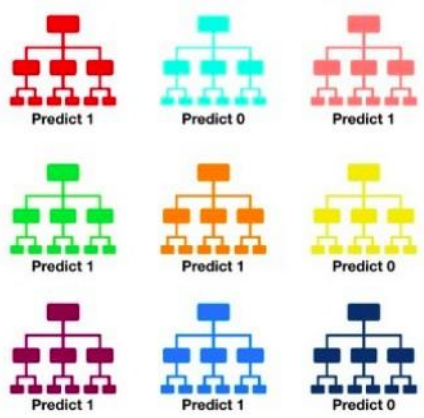


Figure 3.4
Local visualization of Random Forest Ensemble

Adaboost In figure 3.4, as an example there are 9 different decision trees in the forest. Six of them predict 1, and three of them predict 0. In this case, the final prediction would be 1.

Random forest classifier is another ensemble method. This means that it consists of various weak learners that together act as a strong classifier. In the case of Random Forest, the individual classifiers are all decision trees and all are different from each other. We could look at it as a committee of decision trees that take decisions based on the majority vote. The majority vote is a stronger prediction compared to any individual tree.

GridSearchCV

We have implemented hyperparameter tuning on the Random Forest model. The random forest classifier has parameters which can have any values and changes in these values can alterthe accuracy. GridSearchCV is a method to choose the best parameters to improve the accuracy of our model. For this purpose, we have to build a parameter grid. This is basically a python dictionary in which the keys are names of the parameters as strings and the values are python lists of the values which we are going to test for that particular parameter. A combination of all of these values is going to be implemented, and we can directly obtain the list of best values for our parameters. The parameters taken into the grid are: n_estimators, criterion, bootstrap, min_samples_leaf, min_samples_split, and max_features. After running the GridSearchCV successfully, the values obtained are

1. bootstrap: True,
2. criterion: 'gini',
3. max_features: 'sqrt',
4. min_samples_leaf: 1,
5. min_samples_split: 5,
6. n_estimators: 10

IV. Results and conclusion

MODEL	ACCURACY
KNN Model	67.03
Decision Tree Model	72.52
Adaboost Ensemble Model	72.52
Random Forest Ensemble	87.91
Random Forest Ensemble with Grid Search CV	90.10

Table No. 1
Accuracy for each model is tabulated for comparison purpose

A. KNN Model

The accuracy obtained for the KNN model was found to be 67.03%. On attempting to apply scaling of features, the accuracy suffered a decrease so we can conclude that the attributes should not be scaled when using KNN on this dataset.

B. Decision Tree Model

The decision tree model gives us an accuracy of 72.52%. We can conclude that the Decision Tree model is a better fit for the data compared to the KNN model.

C. Adaboost ensemble

The adaboost ensemble method gives us an accuracy of 75.8%. This result makes sense as Adaboost is showing better accuracy than the decision tree as it is using multiple decision stumps for its predictions.

D. Random Forest ensemble

On implementing the Random Forest model, we obtain an accuracy of 87.91%. This accuracy is higher than all the other models which was expected as Random Forest uses multiple decision trees to make a prediction.

E. Random Forest Ensemble using Grid Search CV

Hyperparameter tuning is implemented on the Random Forest model through GridSearchCV. After tuning the parameters, the accuracy is bumped up to 90.10%. This is an increase of 2.19% due to hyperparameter tuning.

Thus we can conclude that the Random Forest ensemble has the highest accuracy.

WEB APPLICATION

A web application is implemented through the Flask micro webframework. We save our machine learning model through pickle. A simple web form is built as an html document and rendered in the python web application file. The model is loaded using pickle into this file and the form values are passedas a python list to the web app. Using python code, we convert this list to a numpy array of the features entered by the patient in the form. This numpy array is passed into the model and it returns a prediction which can then be rendered onto the screenof the patient's computer.

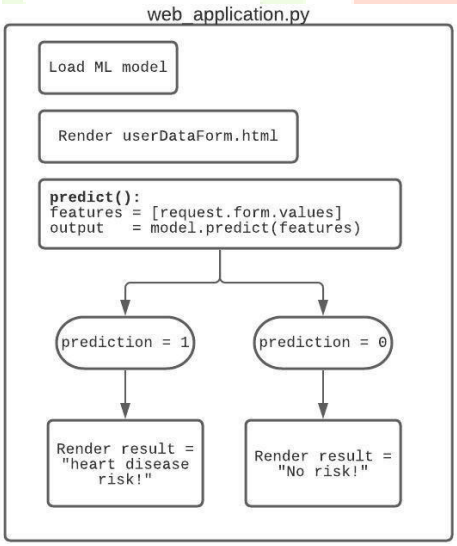


Figure 4.1

Figure 4.1 shows us the logical structure of web applicationand how it obtains the prediction values from the form.

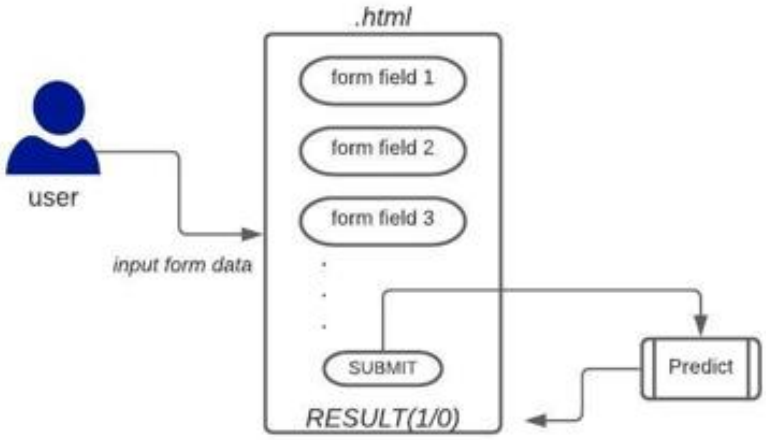


Figure 4.2

Figure 4.2 shows us the html form's structure in which the userenters their data which is then fed to the model.

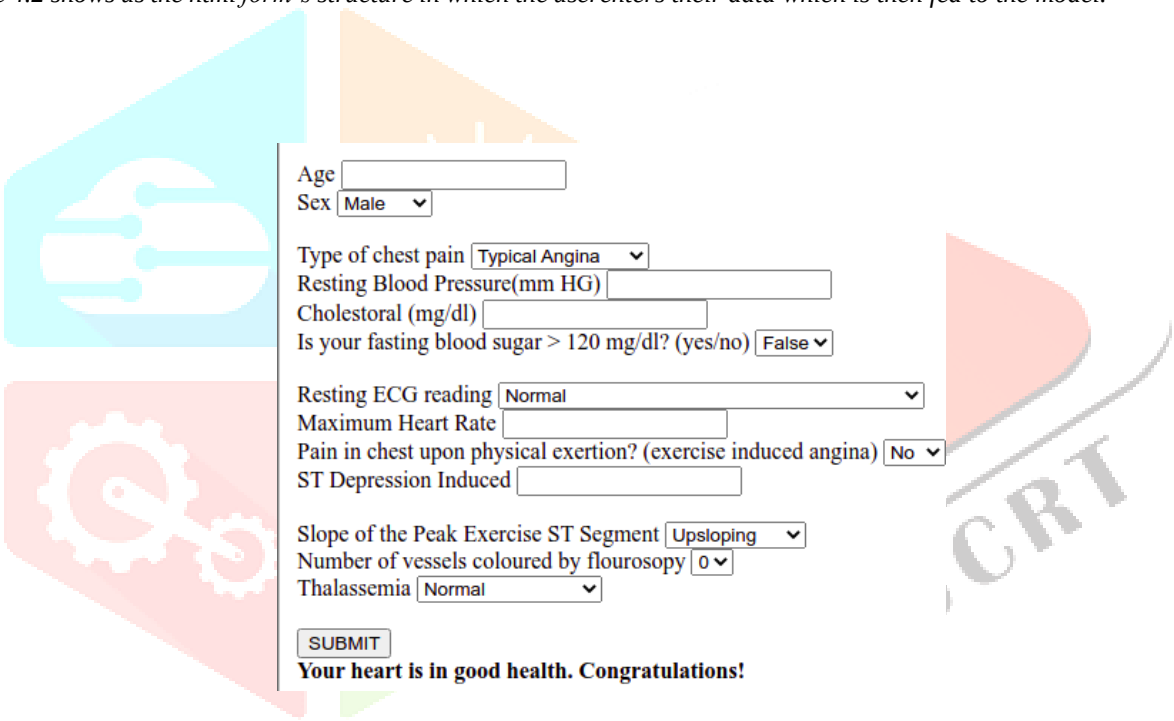


Figure 4.3

Figure 4.3 shows a screenshot of the web application in action with the form values filled, the Result button is pressed and the prediction result is displayed at the bottom.

IV. Future Work

A more representative dataset if available at a later time would give abetter prediction accuracy. More data could be gathered on heart disease as increasing the number of rows in our data could have significant positive effects on the accuracy of each model. The web application could be made into a more commercial grade application which stores user's data and performs analyses with it using a compilation of previous entries

V. REFERENCES

[1] Prediction of Heart Disease Using Machine Learning, 2018, IEEEhttp://ieeexplore.ieee.org/document/8474922

[2] Identification and Classification of Heart Beat by Analyzing ECGSignal using Naive Bayes, 2019, IEEE http://ieeexplore.ieee.org/document/9036455

[3] HeartCare: IoT based heart disease prediction system, 2019, IEEEhttps://ieeexplore.ieee.org/document/9031924

[4] Heart Disease Prediction Using Machine Learning Algorithms,2020, IEEE https://ieeexplore.ieee.org/abstract/document/9122958

[5] Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for FeaturesSelection, 2020, IEEE https://ieeexplore.ieee.org/document/9033683

[6] Prediction of Heart Disease Using Machine Learning Algorithms, 2019, IEEE https://ieeexplore.ieee.org/document/8741465

[7] Scikit Learn - KNN Classifier https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[8] Scikit Learn - Decision Tree Classifier https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[9] Scikit Learn - Adaboost Classifier https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[10] Scikit Learn - Random Forest https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[11] Scikit Learn - GridSearchCV https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[12] Prediction of Sudden Cardiac Death using Classification and Regression Tree Model with Coalesced based ECG and Clinical Data, 2018, IEEE https://ieeexplore.ieee.org/document/8723979

[13] Figure 3.1 from analyticsjobs.in

[14] Figure 3.2 from towardsdatascience.com

[15] Figure 3.3 from statquest.org

[16] Figure 3.4 from towardsdatascience.com

[17] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Arrayprogramming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (https://www.nature.com/articles/s41586-020-2649-2)

[18] Pandas library for dataframe manipulation (https://zenodo.org/record/3715232#.XoqFyC2ZOL8)

(https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pd f)

[19] Matplotlib library for data visualization (https://ieeexplore.ieee.org/document/4160265)

[20] Flask web framework (palletsprojects.com/p/flask/)

[21] Lucidchart for module visualization (lucidchart.com)