|  |
| --- |
| TECHNICAL REPORT TEMPLATE |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| SPRING 22 |  |



Contents

[Effective Heart Disease Prediction 2](#_Toc102482781)

[Abstract 4](#_Toc102482782)

[Methodology 6](#_Toc102482783)

[Results Section 9](#_Toc102482784)

[Discussion 10](#_Toc102482787)

[Conclusion 11](#_Toc102482788)

[Contributions/References 11](#_Toc102482789)

|  |
| --- |
| Effective Heart Disease Prediction |

|  |  |  |
| --- | --- | --- |
| The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases. We created a flask web app to showcase our results using AWS EC2 instance. An end user can enter various building parameters on the html page and when submitted it displays the predicted output.  The Github repository for the project is given below :  <https://github.com/abhi5movva/DSCI6007-FinalProject> | | |
| **Team Members:**  **Abhinaya Movva**  **Harsha Kolla**  **Venkatesh Inturi** | **Questions?**  Contact: amovv1@unh.newhaven.edu |  |
|  | | |

Motivation

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (HDPS) is developed using Naive Bayes and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The HDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g. Relationships between medical factors related to heart disease and patterns, to be established. We have employed the multi layer perceptron neural network with back propagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

## 

## Abstract

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the data set. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

The above features are used in estimating the output of a given meter reading.

▪ AWS EMR – Cluster is created, and modelling is done with the given dataset in jupyter notebook

▪ Jupyter – a web interface to execute python code

▪ AWS EC2 – to create an instance to connect to a virtual system

▪ Flask – it is used to create a simple web interface for the built model.

Cover Page

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (HDPS) is developed using Naive Bayes and Decision Tree algorithms for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The HDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g. Relationships between medical factors related to heart disease and patterns, to be established. We have employed the multi layer perceptron neural network with back propagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

## 

## Methodology

The data set is very small and quite complex. None of the model alone can read the pattern. Although some of the models does a good work in finding the underlying pattern.

There remains some scope on top of the best model to increase the score.

In this work, we have created a special type of Ensemble of linear and Tree models

Heart disease data is pre-processed after collection of various records. The data set contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the data set and the remaining 297 patient records are used in pre-processing. The multi class variable and binary classification are introduced for the attributes of the given data set. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set TABLE 1. UCI data set attributes detailed information. TABLE 2. UCI data set range and datatype. to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of 0 indicating the absence of heart disease

# 3 age: age in years

# 4 sex: sex (1 = male; 0 = female)

# 9 cp: chest pain type

# -- Value 1: typical angina

# -- Value 2: atypical angina

# -- Value 3: non-anginal pain

# -- Value 4: asymptomatic

# 10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)

# 12 chol: serum cholesterol in mg/dl

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

# 19 restecg: resting electrocardiography results

# -- Value 0: normal

# -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

# -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

# 32 thalach: maximum heart rate achieved

# 38 exang: exercise induced angina (1 = yes; 0 = no)

# 40 old peak = ST depression induced by exercise relative to rest

# 41 slope: the slope of the peak exercise ST segment

# -- Value 1: Up sloping

# -- Value 2: flat

# -- Value 3: down sloping

# 44 ca: number of major vessels (0-3) colored by fluoroscopy

# 51 thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

# 58 num: diagnosis of heart disease (angiographic disease status)

# -- Value 0: < 50% diameter narrowing

# -- Value 1: > 50% diameter narrowing

Step-1: Loading the Data

* Data was loaded as a dataframe using pandas library

Step-2: Data Exploration

* Various information of the dataset was explored such as shape of the dataframe and description of various parameters of the dataframe using shape and describe() methods. And also value\_counts of various features were also found.
* Pair plots of every feature in the dataset were plotted using seaborn pairplot function.

Step-3: Mapping categorical variables to Strings

* Various categorical variables are converted into their respective strings and created dummy values for these strings and finally concatenated into the dataset as individual columns, better displaying the significance of each label.

Step-4: Splitting the dataset into training and testing datasets

* The dataset was divided into 70% training dataset and 30% testing dataset.

Step-5: Scaling the features

* StandardScalar() preprocessing technique was used to standardize the features in the dataset for better performance of the model.

Step-6: Model Building

Step-7: Making predictions

* After fitting the model with training data, it was used to predict on our testing data.

We get Logistic Regression as the best performing model compared to other models. We pickle this machine learning model and store it in the cluster for the future predictions.

An html page is created and uploaded in the bucket so the user can easily interface and gets the output.

## 

## Results Section

Website link - <http://3.87.112.216/heartPrediction>

For various input patient records are entered in html page, the results are displayed back onto html page. One such input patient records result screen shot is as follows:

## Graphical user interface Description automatically generated with medium confidence

## Table Description automatically generated

## Discussion

The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

## 

## Conclusion

Identifying the processing of raw healthcare data of heart information will help in the long-term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible.

## 

## Contributions/References

GitHub Project repository – <https://github.com/abhi5movva/DSCI6007-FinalProject>

Presentation Link - <https://www.youtube.com/watch?v=Ax7A_-bVlRo>