Heart Disease analysis and Prediction

ARCHITECTURE DOCUMENT

Bhushan Raut, Sana T. Tadvi.

INTRODUCTION

The major killer cause of human death is Heart Disease (HD). Many people die due to this disease. Lots of researchers have been discovering new technologies to prognosticate the disease early before it's too late for helping healthcare as well as people. These processes are still under research phase.

For predicting Heart Disease, a lot of research scholars contributes their effort in this work using various techniques and algorithms such as Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Linear Regression, etc.

In order to give some effort on this work, we are going to develop a Web-based Heart Disease Prediction System by applying SVM ML algorithms.

The limitation of this project is to have only predicted the presence of heart disease. In future work, we can enhance the project by appending more detail prediction of Heart Disease at patient and incorporate with smart wear devices that integrate to Hospital Emergency System.

ML project is done by the following steps:

- Defines a problem statement.
- Classifying the problem into ML problems.
- Selecting suitable ML algorithms based on their type of problems.
- Collecting and cleaning the data.
- Training a Model from data.
- Test the Model from test data
- Evaluate a model from their accuracy

This work is closely related to the supervised problem of ML. This system will do by Python programming language which is handled entire development of this system using its ML's libraries. At the initial phase, a Heart Disease Prediction Model will build by one of both mentioned algorithms through the comparison of them. This process is done through the ML Process in order to build a model. After that, the model will deploy into web application by implanting Python Flask server-side. The proposed system used by Doctors that can access the system in order to decide whether the patient having Heart Disease or not. This system provides the level of Heart Disease presence such as no Heart Disease, having Heart Disease, and most likely having Heart Disease. This system has one admin user that manage and control the overall system and data of doctors and patients reports.

Problem context

Health is real wealth in the pandemic time we all realized the brute effects of covid-19 on all irrespective of any status. You are required to analyze this health and medical data for better future preparation.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The “goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Attribute Information:> 1. age> 2. sex> 3. chest pain type (4 values)> 4. resting blood pressure> 5. serum cholesterol in mg/dl> 6. fasting blood sugar > 120 mg/dl> 7. resting electrocardiographic results (values 0,1,2)> 8. maximum heart rate achieved> 9. exercise induced angina> 10. oldpeak = ST depression induced by exercise relative to rest> 11. the slope of the peak exercise ST segment> 12. number of major vessels (0-3) colored by flourosopy> 13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

Find key metrics and factors and show the meaningful relationships between attributes.

Target users

The proposed system will cover a particular field of hospital and healthcare center and the target users of this system are also specific like heart-related doctors who have directly used this system in hospitals and healthcare centers.

Aim and objectives

- Developing a Web-based HDPS using ML method for doctors as well as for the hospitals.
- To study about HDs prediction using various algorithms and identify the important attributes of HD
- To study and research on NB and DT algorithms for comparing accuracy.
- To evaluate and identify the best out of two algorithms
- To build a HD Prediction Model by utilizing UCI Datasets and deploy the Model into a web app.

Limitation of project

A web-based HDPS application is reserved for the specific task of HD due to researched knowledge of the domain and technical. Considering the time to accomplish this project, it will conduct the following tasks only:

- HDPS can only find the presence of HD of patients. In future, we will enhance the model for predicting specific type of HD.
- HD prediction model can be trained only 303 data of HD patient due to difficult of collecting Nepalese heart patient data but in future, we will collect large data and train model with their high accuracy.
- HDPS can be run with the internet and can be open in the only browser. In the future, we will develop this system as an offline based application.
- HDPS can only reserve heat disease prediction. In the future, we will integrate a healthcare system.

LITERATURE REVIEW

Heart Disease is defined a range of conditions that affect your heart. It is describing any disorder of the heart. The umbrella of Heart Disease consists of different type of Heart Disease such as blood vessel diseases (coronary artery disease, and arrhythmias) and heart defects when you're born with congenital heart defects, among others. The term "Heart Disease" is always used interchangeably with the term "Cardiovascular Disease (CVD)". CVD generally refers to conditions that involve blocked or narrowed blood vessels that can lead to a heart attack, stroke or chest pain (angina).

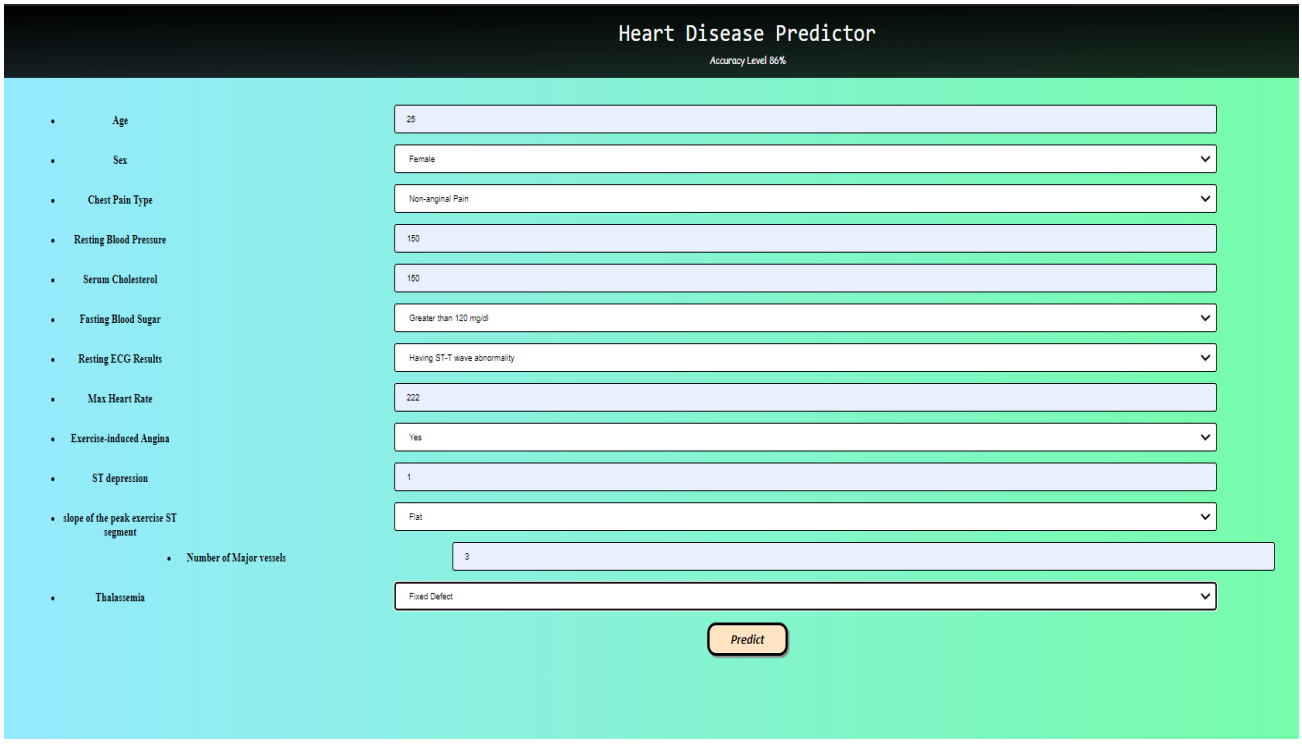
SYSTEM ARCHITECTURE

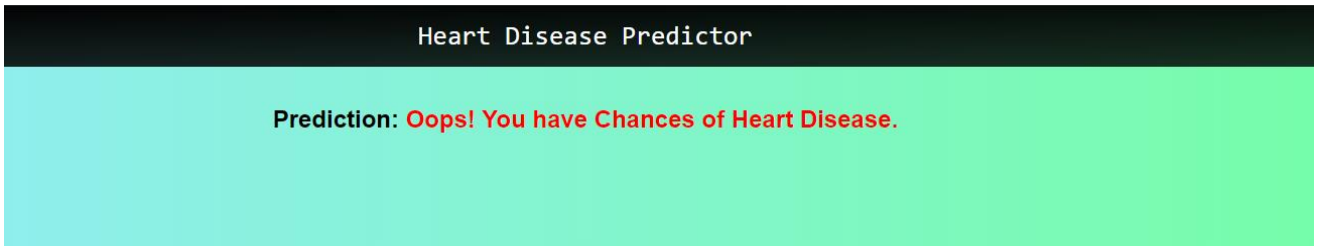
The system architecture (Kautish et al, 2016, 2018, 2019) is like a blueprint of any object. It is a conceptual model to integrate between business logic and physical system in an organized way. It demonstrates the structure, view, behavior, features, and functionalities of the system. It is the way of portraying the desired system in visualizing a way to well understand for people. The System architecture is the foundational orchestrate of a system that incorporated in its elements, their relationships of elements, and the science of its design (MITRE, 2019). HDPS is a web-based application that runs on the browser. This system is embodied in a web application. The web application architecture of the Heart Disease Prediction System is to define the communication between applications, and database on the web. It also helps to know about other third-party application required like python's packages. It represents the represent the relationship between them and visualizes how they work together simultaneously.

Abstract Architecture

Abstract Architecture of Heart Disease Prediction System has demonstrated the comprehensive structure of the system that easily understands all components and their relationship. The way to represent the system as a structural way can beneficial to understand the system for technical and non-technical users. The abstract architecture consists of a system, database, and interface design to visualize the better way of viewing. It describes the overall design of the system to easily understand the system elements, feature and functionalities, and behavior of the system.

Screenshot for Predict Heart Disease Deployment

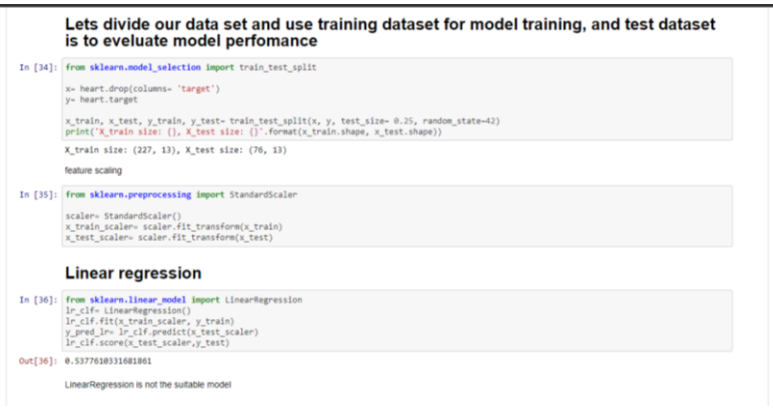




Description

The above page shows the Predict Heart Disease page that appears only access by doctors. This page demonstrates the form of prediction Heart Disease that contains details about the patient's data. It has input validation to restrict the garbage input that displays error message within this form. Must of input data are dropdown list due to restricting the garbage input. When the doctor successfully predicts the Heart Disease, it will automatically open the success page that displays the result of the patient. Furthermore, this function also stores the patient's report into the database table.

Sample Codes of Heart Disease Prediction Model Development



Lets divide our data set and use training dataset for model training, and test dataset is to eveluate model perfomance

```
In [34]: from sklearn.model_selection import train_test_split

         x= heart.drop(columns= 'target')
         y= heart.target

         x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=42)
         print('X_train size: {}, X_test size: {}'.format(x_train.shape, x_test.shape))
```

X_train size: (227, 13), X_test size: (76, 13)

feature scaling

```
In [35]: from sklearn.preprocessing import StandardScaler

         scaler= StandardScaler()
         x_train_scaler= scaler.fit_transform(x_train)
         x_test_scaler= scaler.fit_transform(x_test)
```

Linear regression

```
In [36]: from sklearn.linear_model import LinearRegression
         lr_clf= LinearRegression()
         lr_clf.fit(x_train_scaler, y_train)
         y_pred_lr= lr_clf.predict(x_test_scaler)
         lr_clf.score(x_test_scaler,y_test)
Out[36]: 0.5377610331681861
```

LinearRegression is not the suitable model

SVC

```
In [49]: from sklearn.svm import SVC

         svc_clf= LogisticRegression()
         svc_clf.fit(x_train_scaler, y_train)
         y_pred_svc= svc_clf.predict(x_test_scaler)
         svc_clf.score(x_test_scaler,y_test)
Out[49]: 0.868421052631579

In [50]: print('Classification Report\n', classification_report(y_test, y_pred_svc))
         print('Accuracy: {}%\n'.format(round((accuracy_score(y_test, y_pred_svc)*100),2)))
```

Classification Report

	precision	recall	f1-score	support
0	0.90	0.80	0.85	35
1	0.84	0.93	0.88	41
accuracy			0.87	76
macro avg	0.87	0.86	0.87	76
weighted avg	0.87	0.87	0.87	76

Accuracy: 86.84%

```
In [51]: cm = confusion_matrix(y_test, y_pred_svc)
         cm
Out[51]: array([[28,  7],
                [ 3, 38]], dtype=int64)

In [52]: import pickle

         filename = 'Healthcare_Analysis_on_Heart_Disease.pkl'
         pickle.dump(rf_clf, open(filename, 'wb'))

In [ ]:
```

Support Vector Machine: SVM is a technique for ramification of both linear and non-linear data. It applies a non-linear mapping method so that it can transform the training data into a higher dimension. A hyperplane is a kind of line which separates the input variable space in SVM. The hyperplane can separate the points in the input variable space containing their class that is either 0 or 1. In two-dimensions, one can visualize this as a line and it is assumed that each input points can be completely separated by this line. The distance between the hyperplane and adjacent data

coordinates is called margin. The line which has the largest margin can distinguish between the two classes is known as the optimal hyperplane. These points are called support vectors, as they define or support the hyperplane. In practice, there is an optimization algorithm which is used to calculate the values for the parameters that maximize the margin.

Logistic Regression: Logistic regression is a technique of machine learning which is taken from the field of statistics. This method can be used for binary classification where values are distinguished with two classes. Logistic regression is similar to linear regression where the goal is to calculate the values of the coefficients within every input variable. Unlike linear regression, here the prediction of the output is constructed using a non-linear function which is called a logistic function. The logistic function transforms any value within the range of 0 to 1. The predictions made by logistic regression are used as the probability of a data instance concerning to either class 0 or class 1.

CONCLUSION

Heart Disease is a killer disease of death in the world. According to WHO and other statistical facts, HD is the most dangerous disease that is the cause of death of a human. In 2017, the latest fact data of the World Health Organization (WHO) published that Nepal has reached 18.72% or 30,559 deaths from Coronary HD. The rate of age fixed death is 158.35 out of 100,000 population and world rank is #41. (World Life Expectancy, 2019). After identifying the problem statement, we strived to find a way of solving the problem through ML. After that, we conducted requirement analysis and planning of the system. We made the line of research boundary to complete this project by planning the system requirement and planning. System requirement and planning consists of aim, objective, deliverables, and target audience of the system. We conducted the literature review by studying various journal papers and articles to understand the way of solving the problem as well as collecting important attributes of HD that lead role of the HD in patient. We found a various way to resolve the problem but dilemma to select the best algorithm. Eventually, we chose two algorithms (DT and NB) among them which have high accuracy in HD datasets. The dataset of HD is retrieved from UCI repertory which has important 13 attributes among 76 attributes that play the role of HD. it has 303 records that a small dataset. It may difficult to achieve a high accuracy to predict an HD. In technical research, we found a Python programming language that is suitable for the HDPS by comparing two other languages such as R and Java. Python is only one language to handle the both function such as building and deploying a Prediction Model and developing a web application.

For building an Heart Disease Predicting Model, we chose Jupyter notebook for interactive development environment. We selected some essential Python libraries that supports to build an HDP Model such as NumPy, Pandas, Matplotlib, and Sk-learn. For developing web application, we select Python Flask Server-side scripting language to deploy the HDP Model and develop a web application. Furthermore, we selected Heroku, and Google Chrome for web browser by critical evaluating.

It is clear that SVM showed higher accuracy rates of more than 86 percent which make them considerable models for biomedical applications of disease detection and prediction. Moreover, to determine the best performing model for the considered databases in this study, other analysis criteria are considered. The accuracy, sensitivity ,specificity, precision, and F-score of SVM is higher than both Random Forest and Simple Logistic models.

Eventually, we are able to achieve all the proposed objectives of this project. We developed a web application of HDPS that has all features and functionalities what we planned. The research of this project is successfully met in the system. This system resolves the real-environment problem. It is successfully predicted the presence of HD in patient. It is also store and manage the prediction report of heart patient by doctor account. Admin user can handle create doctor account, mange doctor account and view the report of the patient. The overall system can solve the problem statement of the project and make a novel tool to predict an HD at hospital.