**DS110 / ML PROJECT / HEART DISEASE PREDICTION**

**ABDUL RAFAY / MINSEO KIM**

**Introduction / Objective :**

In this project our objective was to successfully predict whether someone has heart disease based on their attribute using two different machine learning methods. The dataset we found at University of California Irvine’s machine learning repository, **“Heart Disease UCI”**, provided us with valuable data good fit for our project. We chose this project because heart disease is a major public health problem, and its successful and essential treatment is an early diagnosis. One of our team members had research on electrochemical immunosensor for detection of heart failure cases and its real-world application, and the interest continues in this  project.

**Description about datasets from UCI ML Repository is as follows :**

The **Heart Disease UCI** comprises of multiple datasets with data collected from multiple locations. They all were medical centers with direct access to significant number of patients with their attributes. The dataset contains **76 attributes** in total, but we only used some of the **14** that is deemed to be related to heart disease. Most of the other published studies using the UCI dataset only used **Cleveland database**, with 14 of those attributes. They mostly focused simply on attempting to distinguish presence from absence of heart disease.

Unlike most other studies which only used Cleveland database, we used all of them. This was to increase our sample size to potentially increase our accuracy of our predictions. So, the final dataset we used included information of **920 patients**, with **14 relevant features**. From which we used our own selection of features as needed to accurately predict the presence of heart disease in patients.

The UCI dataset can be found at : <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

Same dataset in CSV format was accessed at : <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?resource=download>

**The data was collected from the following locations by :**

1. Cleveland Clinic Foundation (cleveland.data) by Robert Detrano, M.D., Ph.D.

2. Hungarian Institute of Cardiology, Budapest (hungarian.data) by Andras Janosi, M.D.

3. V.A. Medical Center, Long Beach, CA (long-beach-va.data) by Robert Detrano, M.D., Ph.D.

4. University Hospital, Zurich, Switzerland (switzerland.data) by William Steinbrunn, M.D.

5. University Hospital, Basel, Switzerland (switzerland.data) by Matthias Pfisterer, M.D.

**Columns (Attributes) :**

**id** (Unique id for each patient); **age** (Age of the patient in years); **dataset** (place of study); **sex** (Male/Female); **cp** chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic]); **trestbps** resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital)); **chol** (serum cholesterol in mg/dl); **fbs** (if fasting blood sugar > 120 mg/dl); **restecg** (resting electrocardiographic results) -- Values: [normal, stt abnormality, lv hypertrophy]; **thalch**: maximum heart rate achieved; **exang**: exercise-induced angina (True/ False); **oldpeak**: ST depression induced by exercise relative to rest; **slope**: the slope of the peak exercise ST segment; **ca**: number of major vessels (0-3) colored by fluoroscopy; **thal**: [normal; fixed defect; reversible defect]; **num**: the predicted attribute;

**Our Coding Progression :**

**Load Data – Preprocessing – Normalization(SVC) – DecisionTreeClassifier – SVC – Evaluation**

**Preprocessing Data :**

Data had to be preprocessed before we could perform machine learning techniques on it. It had some columns with missing portion, which we filled up with the median value of the column. Column **‘ca’** had significant missing portion, so it was exempted from filling and was just dropped. Which was then followed by further dropping of columns, which were thought to be insignificant. This included additional drop of columns : **‘fbs’**, **‘exang’**. Then **‘num’** column was renamed to **‘heart\_disease’**.

A picture containing text

Description automatically generated

**Initial data (horizontal lines showing missing portions) :**

Chart, diagram

Description automatically generated

**After preprocessing :**

Chart

Description automatically generated

**Label Encoder :**

Some of the columns in preprocessed data included categorical data which had to be encoded for it to be processing by the classifiers. We used label encoder to convert the following columns : **‘sex’, ‘cp’, ‘restecg’, ‘slope’, ‘dataset’, ‘thal’, ‘heart\_disease’**. The **‘heart\_disease’** column was then further converted to binary system instead of values of 0 to 4.

**Table

Description automatically generated**

**Train Set & Test Set :**

Train set and test set were split with test size being **20%** and random state being **42**.

**Normalization (SVC) :**

The data then further had to be normalized for SVC method. We used **StandardScaler** to transform **X\_train** and **X\_test** and labeled them as **X\_train\_scaled** and **X\_test\_scaled**.

**ML Methods :**

**Decision Tree Classifier** was accessed from **scikit-learn** library’s **DecisionTreeClassifier** class; Maximum depth of the tree was varied to achieve the best possible performance outcome. Maximum depth of 25 and maximum leaf nodes of 4 yielded highest accuracy of 83%.

**Accuracy: 0.8369565217391305**

**Classification Report:**

**Precision recall f1\_score support**

**0 0.83 0.76 0.79 75**

**1 0.84 0.89 0.87 109**

**Accuracy 0.84 184**

**macro avg 0.83 0.82 0.83 184**

**macro avg 0.84 0.84 0.84 184**

**Decision Tree Visuals :**

**Diagram

Description automatically generated**

**Decision Tree Confusion Matrix :**

**Chart, treemap chart

Description automatically generated**

**Decision Tree Binary Classification Performance :**

**Chart, line chart

Description automatically generated**

**Vector Machine Classifier** was accessed using **scikit-learn** library but this time, accessing an **SVC** class; This time we varied the regularization parameter and kernel function to achieve best possible performance. Kernel was set to ‘linear’, C equal to 1 and gamma being ‘auto’. The best performance we got was accuracy of 67%.

**Accuracy: 0.6793478260869565**

**Classification Report:**

**Precision recall f1\_score support**

**0 0.70 0.37 0.49 75**

**1 0.67 0.89 0.77 109**

**Accuracy 0.84 184**

**macro avg 0.69 0.63 0.63 184**

**macro avg 0.68 0.68 0.65 184**

**SVC Scatter Confusion Matrix :**

Chart, treemap chart

Description automatically generated

**Overall Data Correlation Visual :**

A picture containing chart

Description automatically generated

**Conclusions :**

Upon assessing our results, we can reasonably conclude that machine learning models can be trained to predict the presence of heart disease with acceptable accuracy. However, in our project, one method turned out to have better performance: decision tree. So, we also learned that the method one chooses to predict these types of data, could vary the performance.

On the side note, it is also important to consider the ethical implications of using machine learning models in healthcare, particularly regarding patient privacy and potential biases in the data used for training. Therefore, it is crucial to establish clear guidelines and regulations for the development and deployment of these models.

Also, we believe further research should be done for better understanding of the complex interplay of factors that contribute to heart disease and to develop more accurate and personalized risk prediction models, as this can vary by a lot based on the patient. Such models in which one can guess the future health risks based on their health condition would be a promising avenue for improving healthcare outcomes.