Capstone Project: Heart Disease Prediction using Machine Learning

By

Rezwanur Rahman (Rez)

Background: Heart disease is very common among different ages. People with certain blood sugar level, cholesterol level etc., may exhibit symptoms for heart disease. It is important to find a smart way to speculate possibility of heart disease given certain physiological status. UCI ML database has a large, collected data for heart disease will be used to develop a ML model. Historically great effort has been made to detect prime arbitrator of heart disease given various parameters [1]. This work highlights on few analytical metrics without proposing a ML prediction methodology(s). There has been a necessity to do a systematic analysis on “primary” causes of heart disease to trigger alarm for a person. Overall analytics and prediction can be compiled as a streamlined process to work on any probable future datasets. As a part of the work author will be creating a ML pipeline to predict possibility of heart disease using multiple candidate ML methods. This will be done to ensure reliability of overall prediction.

Problem Statement: This data has several parameters that encompasses a relationship between person’s general health status and potential heart risk. A ML model can be used to extract important parameter (aka features) while analyzing the data. Eventually ML model can be used to predict heart risk.

Dataset source and Input: In this project UCI ML dataset on heart disease has been used [2]. The dataset has a wide range of input features collected from patients at a Cleveland V.A. hospital. The raw data from UCI website needed rigorous pre-processing. Since the scope of the project is to develop ML scheme with preliminary data analytics, we have collected the dataset from Kaggle website [4] as a csv file (filename: 'heart\_disease\_dataset\_UCI.csv').

Dataset description: The csv file has 14 columns. 13 of them are features and 14th column is the Boolean target variable defining chance of heart risk or not. Detail variable description can be seen as:

* **age**: The person's age in years
* **sex**: The person's sex (1 = male, 0 = female)
* **cp**: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic
* **trestbps**: The person's resting blood pressure (mm Hg on admission to the hospital)
* **chol**: The person's cholesterol measurement in mg/dl
* **fbs**: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
* **restecg**: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
* **thalach**: The person's maximum heart rate achieved
* **exang**: Exercise induced angina (1 = yes; 0 = no)
* **oldpeak**: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
* **slope**: 0 = upsloping, 1 = flat, 2 = downsloping
* **ca**: The number of major vessels colored by flourosopy (0-3)
* **thal**: A blood disorder called thalassemia: 0: NULL (dropped from the dataset previously), 1: fixed defect (no blood flow in some part of the heart), 2: normal blood flow, 3: reversible defect (a blood flow is observed but it is not normal)
* **target**: Heart disease (0 = no, 1 = yes)

Numerical data is stored in df\_heart\_disease dataframe:

Table

Description automatically generated

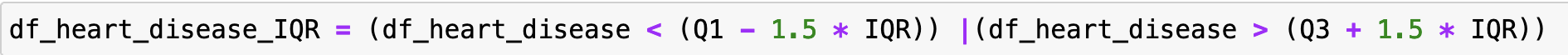
Table: 1

1. Data pre-processing: As part of pre-processing following steps will be done to cleanup the data:

1. Find out missing values by looking for NULL values: At this step our goal is to see if there is any missing value i.e. NaN existing in the dataset. If any missing value found that can be replaced or removed by whole row pertaining to that. Key steps:
   1. Execute df\_heart\_disease.isna().sum()
   2. Visualize the table using package “missingno”
2. Find out duplicate attributes: Checked for any duplicate columns:

Text

Description automatically generated with medium confidence

1. Look for outliers in the data: In this step we look for any outliers in the data. If the outlier resides outside a prescribed range, those are discarded. Any data outside 5%-95% quantile region will be removed.
   1. Use Pandas IQR function: 
   2. Visualize distribution of accepted outliers as usable data (Figure: 1):

Calendar

Description automatically generated with medium confidence

Figure: 1

1. Rename features if necessary: For the purpose of analysis and understanding numerical features with categorical nature (e.g. thal, chol, sex, etc) were converted to categorical form. This was used in exploratory data analysis:

Table

Description automatically generated with low confidence

Table: 2

2. Data Analytics and Visualization: In this section exploratory data analysis (EDA) has been performed using several visualization approaches. Goal is to develop a clear understanding on the relation between heart risk and different features, how different features are distributed, quality of correlation between features etc.

***Feature data spatial distribution***: In this case we can visualize how features are distributed. Any feature with unbalanced spatial distribution may trigger quality of data and eventually affect ML results (Figure: 2). Few outliers are observed in the data distribution.

Graphical user interface, diagram, application, Word

Description automatically generated Chart

Description automatically generated

Figure: 2 Figure: 3

In a similar manner histogram analysis is done to understand data distribution over certain range. Most of the features seem to have skewness over certain range (Figure: 3).

***Feature analysis***: In this section we can internal variations in each feature and it’s relationship with targe i.e. chance of heart risk. First thing comes to mind is relationship between patient age, gender and chance of heart risk. As shown in Figure: 4(a) age has an interesting relationship with hear risk. People in their late 50’s-60’s have more risk of heart disease. Out of men and women, men have higher heart risk. To supplement that observation we can see male have higher cholesterol than women and also mode defective heart than women (Figure: 4(b, d)). This is probably dietary habit and lifestyle of men in the timeframe of 1980-1990. As we move away from age and gender, we can see from Figures: 4(c, e, f, g) higher cholesterol level and angina has stronger effect on heat disease.

Graphical user interface, application

Description automatically generated

Figure: 4

Further digging into the data gives us some understanding between heart functionality and heart disease. Figures: 5(a, b) tells us that flat slope in peak exercise has relationship with heart risk. This is not really a trait, more of a behavioral patten of heart. On a same note, detected vessels de to fluoroscopy has some relationship with heart risk. Patients without heart risk has no vessel detected as a very frequent case. Figures: 5(c, d) confirms us that max heart rate, high cholesterol and blood pressure so have internal dependency in triggering heart risk. These figures are created for patients with heart risk only.

Chart, scatter chart

Description automatically generated

Figure: 5

***Feature correlation***: It is important to understand how different features are correlated. If they are highly correlated, further feature extraction or reduction would be necessary. As shown in Figures: 6(a, b), features don’t have great correlation in general. Maximum positive value for correlation is with the range of 0.4. Hence it is sufficient to move ahead with these feature vectors for further analysis.

Chart, treemap chart

Description automatically generated

Figure: 6

***Feature importance***: As a next part of EDA we are getting into extracting important features from the feature vector. Initially SKlearn’s univariate feature selection function “SelectKBest” was used to sort important features based on their scores. To supplement this result both random forest and XGBoost ML techniques were used to extract important features (Figure: 7). Finally first 80% important features were considered which are common in these three cases.

Chart

Description automatically generated

Figure: 7

From Figure: 7 looks like 6-7 features are important enough to drive the classification model performance. However, there are not same in both these cases. Further principle component analysis was done to get an idea on how many features are actually significant. Figure: 8(a) tells us maximum 4 features are very important. So considering 80% of the features should suffice. Extracted features to be considered are shown in Figure: 8(b). Given this subset of feature we can visualize a simple decision tree to see how they help detecting heart risk.

3. Prediction using ML: In this section we have developed multiple ML models for binary classification problem. Initial efforts were given to make a short list of candidates as ML models. This was followed by hyper-parameter tuning relevant to specific ML models. Our workstream can be summarized as:

Diagram

Description automatically generated

Figure: 8

***ML Model development and tuning strategy***:

1. Data split: Data has been split by 75% training data and 25% testing data.
2. Only best features detected from previous step were considered.
3. Develop ML models: In this case we develop ML binary classification models with important hyper parameters. Base models always need tuning to achieve better result. In this case *Random Search based hyper-parameter tuning was performed with 5-fold cross validation*. This was done to ensure an computationally efficient yet effective model was obtained. Model parameter summary is:
   1. XGBoost model with hyper-parameter range

*params = {*

*'objective':['binary:logistic'],*

*'learning\_rate': [0.001, 0.005, 0.01, 0.1,0.3,0.5,0.7,1],*

*'max\_depth': [1, 2, 3, 4, 5, 6, 7],*

*'min\_child\_weight': [1e-5, 1e-3, 1e-2],*

*'subsample': [0.01, 0.1, 0.3,0.5,0.7,1],*

*'colsample\_bytree': [0.7,1],*

*'n\_estimators': [100, 200, 300, 400, 500, 1000]}*

* 1. Random Forest model with hyper-parameter range

*params = { 'bootstrap': [True, False],*

*'max\_depth': range(1,10, 1),*

*'max\_features': ['auto', 'sqrt'],*

*'min\_samples\_leaf': [1, 2, 4],*

*'min\_samples\_split': [2, 5, 10],*

*'n\_estimators': [100, 200, 300, 400, 500]}*

* 1. Logistic regression (LGR) model with hyper-parameter range

*params = {"max\_iter": range(100,500,2),*

*"solver" : ['newton-cg', 'lbfgs', 'liblinear'], "C": [0.5, 0.1, 1.0]}*

* 1. Light GBM model with hyper-parameter range

*params = {'num\_leaves':range(10,100, 10), 'min\_child\_samples':range(5,25,5),'max\_depth': range(5, 15, 1),*

*'learning\_rate':[0.05,0.1,0.2],'reg\_alpha': [0,0.01,0.03]}*

* 1. Linear Discriminant Analysis (LDA) model with hyper-parameter range (ideally don’t have any since closed form model)

*params = {"solver" : ["svd"],*

*"tol" : [0.0001,0.0002,0.0003]}*

1. After “best\_estimator” was obtained it was used to predict on test dataset. Multiple metrics were developed to quantify model’s performance.
2. Given all the trained model an ensemble ML model was developed based on Sklearn’s voting classifier.
3. Model performance metrics are:
   1. Confusion metrics
   2. F1 score
   3. ROC\_AUC score
   4. Precision score
   5. Recall score
   6. Accuracy score
   7. Cross validation score

In this project we treated the problem as: Classification Problem (instead of Regression Problem) since many of our feature vectors and target variable are discrete in nature. It is unlikely to have 0.5 assigned to heart risk/no-risk condition. Decision trees are very good at classification problems. We also included well know approaches like LGR and LDA since we are dealing with “Binary Classification” problem.

***ML Model performance***:

*Training performance*: It is important to evaluate model training performance with cross-validation samples. This will ensure us to conclude that models are not under trained. As shown in Figure: 9, as expected training scores tend to become steady with increasing CV samples. This is an indication of models are being stable and well trained on the data. LDA and LGR models seem to be behaving batter while training.

Graphical user interface, chart, line chart

Description automatically generated

Figure: 9

*Model accuracy metrics*: Based on Figure: 10 we can have a clear understanding on the model accuracy. Figures: 10 (1-e) shows LGR and LDA models have slightly better accuracy compared to XGBoost, Random Forest and Light GBM models. LDA and LGR has higher True positive and almost same true negative (In this case: No Risk is considered positive and risk is negative). In this case False Positive is more crucial since we don’t want to label a patient with heart risk as no risk condition. Given that we can see LGR and LDA are slightly better by having lower false positives. ROC\_AUC scores also depict how the models are with True vs. False positive rates. As expected LGR and LDA have steepest increase in True Positive rates given lower range of False Positive Rate. Based on these matrices we can see LGR and LDA models have better prospect in this dataset.

*Model performance evaluation*: In order to build a deeper understanding on the model accuracy we calculated some of the key metrices. Goal is to conclude which model can be better

Diagram

Description automatically generated

Figure: 10

performing for predicting heart risk. As measuring scales we used f1, precision, recall, accuracy, roc\_auc, and cross-validation scores. In this dataset we are more interested in precision, f1 and accuracy scores to start with. Since the data has almost balanced distribution of “No Risk” and “Risk” conditions, we can use accuracy as a good metric to get overall picture.

Table

Description automatically generated

Table: 3

As shown in Table: 3, LGR and LDA models have higher accuracy. In addition, f1 and precision scores are higher compared to other models. Recall score seems to be higher for LGR and LDA but RF classifier exhibits similar accuracy. Cross-validation tells us that models are reasonably well fitted since all of them are with in similar range of values.

*Benchmark model*:

Perform exploratory data analysis to develop an understanding on relationship between heart disease and a given attribute. For example, a histogram in Fig. 1 depicts a good idea on relationship between heart disease and age.

Chart, bar chart, histogram

Description automatically generated

Fig. 1

* 1. Develop clustering models to use unsupervised learning in finding a pattern of self-similarity in the data. For example, can we say “Young females have tendency to have low heart disease”?
  2. Develop a regression model(s) to conclude a relationship between most significant features and risk of heart disease.

Benchmark Model: Data can be split into train and test parts. Model can be benchmarked based on its performance with test dataset. Model performance can be compared with literature data [1, 3].

Evaluation Metrics: Performance of the ML model can be determined in terms of ROC curve and confusion matrix.

Project Design: For this project ML package Scikit-learn will be used for both supervised and unsupervised ML methods. Sklearn is a rich an highly applied python package. As ML methodologies for regression we will use models like: RandomForest Regression, Logistic regression, XGboost regression etc. Idea is to use effective regression method capable of capturing both linear and non-linear relationships in the data. To avoid overfitting we will be using k-fold cross validation method on training set. K-nearest neighbor based clustering will be used to find any self similar pattern in the data. Model will be deployed on GitHub repo.

Reference:

1. Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP Conference Series: Materials Science and Engineering. Vol. 1022. No. 1. IOP Publishing, 2021.
2. <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
3. Albahr, Abdulaziz, et al. "Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer." Computational Intelligence and Neuroscience 2021 (2021).
4. https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci