FINAL PROJECT: PREDICTING THE PRESENCE OF HEART DISEASE IN PATIENTS WITH MACHINE LEARNING

BY

ANUOLUWAPO ALEEM

ABSTRACT

Heart disease has been proven to be the leading cause of death for both men and women, killing about 697,000 people in the US in 2020, which makes it a major concern to be dealt with. Identifying heart disease in patients could be quite challenging due to several contributory risk factors, which requires some high level of techniques. Machine learning proves to be effective in predicting heart disease in patients given the contributory risk factors. In this project, we aim to use seven different machine learning classifiers such as K-Nearest Neighbors, SVM, Decision Tree, Random Forest, Logistic Regression, Neural Network, and AdaBoost classifiers to predict whether a person is suffering from heart disease or not, using the Cleveland Heart Disease dataset from the UCI Repository, and also to determine the best performing classifier. Predictive analysis is carried out on the dataset using the seven different classifiers. The performance of each classifier is also estimated and compared to determine the best performing classifier on the dataset. Linear support vector machine and Elastic net logistic regression have the best overall performance with an overall performance of 89%.

INTRODUCTION

Heart diseases carry a high mortality rate if not detected and treated in its infancy. Early detection of the disease helps to reduce the death rate caused by this killer disease, and this could be quite difficult due to its numerous risk factors. However, this constraint can be overcome by machine learning. Machine learning can be used to diagnose, detect, and forecast many disorders in the medical industry. Early identification of heart disease using a machine learning prediction algorithm can be recommended generally for fatality rate reduction, and decision-making is improved for further treatment and prevention. The primary purpose of this study is to help clinicians determine the best performing machine learning algorithm able to predict the presence of heart disease in people at an early stage. To achieve this, we will use seven different machine learning classifiers such as K-Nearest Neighbors, SVM, Decision Tree, Random Forest, Logistic Regression, Neural Network, and AdaBoost classifiers to train the Cleveland Heart Disease dataset, and select the classifier that produces the best performance.

METHODS

In this project, the classification goal is to predict the presence of heart disease using seven machine learning classifiers on the Cleveland Heart Disease dataset, and select the best classifier in terms of their performance on the dataset. To achieve this, we will take the following steps.

- Data Preprocessing & Cleaning: This is the part where we load and clean the downloaded Cleveland Heart Disease dataset gotten from UCI Repository in python. The dataset is described below.
    - Cleveland Heart Disease dataset: consists of 14 variables measured on 303 patients who are grouped into five levels of heart disease. The “outcome of interest” or “label” measures the presence or absence of heart disease in the patient. It is integer valued from 0 to 4, where 0 indicates the absence of heart disease while the rest represents 4 different levels of heart disease, which will be regrouped to 1 indicating the presence of the disease

We noticed that the dataset contains some missing value denoted by “?” and didn’t contain column names, so at the point of loading the data into python we assigned a name to each column using the data description given from the dataset and changed “?” in each row to nan value. There are 5 continuous and 8 discrete features, which do not contain the correct datatype. So, we changed the datatype of the discrete and integer features from float to categories and integer respectively. The label which is valued from 0 to 4 is recoded into 2 dummy outcomes, 0s and 1s. $Y_i = \begin{cases} 1 & Presence\ of\ heart\ disease \\ 0 & Absence\ of\ heart\ disease \end{cases}$. We dropped six missing values from the dataset since they are few. Now, we can begin our analysis where the largest part of the work was done, taking about a week to complete.

- Exploratory Data Analysis: We try to explore the dataset to understand how the outcome of interest (disease diagnosis of patient) is distributed and related to each feature. Also, we try to identify the groups of people who are at higher risk of having the disease. This analysis will be reported in the result section.

- Predictive Analysis: Before starting to train the data, we defined our hyperparameters for each classifier that will be used. We used different sets of hyperparameters for each method and selected the ones that gave the highest performance. We carried out the analysis following this summarized process: first, we converted the discrete features to integer to make it possible to train and test the data, then we split the datasets into 70% training data and 30% test data, after which we use the training data to train seven machine learning algorithms which are:
  - K Nearest Neighbors (KNN): This classifier looks for the classes of K nearest neighbors of a given data point and based on the majority class, it assigns a class to this data point. We used 1, 3, 5, 15, 25, and 50 number of neighbors with 2 types of weight: ‘uniform’ & ‘distance’ used in prediction. We selected the combination with the highest performance.
  - Support Vector Machine (SVM): For this classifier, we used three kernels: linear, poly, and rbf with regularization parameters: 0.025, 0.05, 0.5, 1 and selected the 1 with the best performance.
  - Decision Tree: It creates a decision tree based on which it assigns the value of a target variable. Here, we used six different values of maximum number of features: 5, 10, 15, 20, 25, 30 to train the classifier.
  - Random Forest: fits a number of decision tree classifiers on dataset. We can vary the number of trees that will be used to predict the value of the target variable, and here we used 10, 100, 200, 500, and 1000 trees with 5, 10, 15, & 20 values of maximum tree depth and number of features ranging from 1 to 5
  - Logistic Regression: We will use 3 types of logistic regression: Lasso, Ridge and Elastic Net which uses l1, l2 and combination of l1 and l2 regularization respectively. We will then select the 1 with the best performance.
  - Neural Network: We will use 50 & 100 number of hidden layers with 2 different activation functions: relu & tanh, and 3 different values of initial learning rate: 0.001, 0.01, 0.1. We will also use 0.1 & 1 alpha which is the strength of regularization term. The best performing combination of hyperparameters will be selected.
  - AdaBoost Classifier: We applied 3 different maximum number of estimators: $\{20,50,100\}$ and 3 learning rates: $\{0.01, 0.1, 1\}$ and select the best performing combination.

The performance of each classifier will then be evaluated on the test data and will be compared with one another to determine the classifier that achieved the best performance. The following five metrics were used to evaluate and compare the performance of the classifiers.

Recall (R)	Precision (P)	F1	Accuracy (A)	Average Score
$\frac{TP}{(TP + FN)}$	$\frac{TP}{(TP + FP)}$	$\frac{2 * P * R}{(P + R)}$	$\frac{TP + TN}{Total\ Predictions}$	$\frac{R + P + F1 + A}{4}$

Figure 1. *Metric used to evaluate & compare classifiers Performance where TP = # of True Positives, TN = # of True Negatives, FP = # of False Positives & FN= # of False Negative.*

RESULTS

The exploratory data analysis result is given in the figures below.

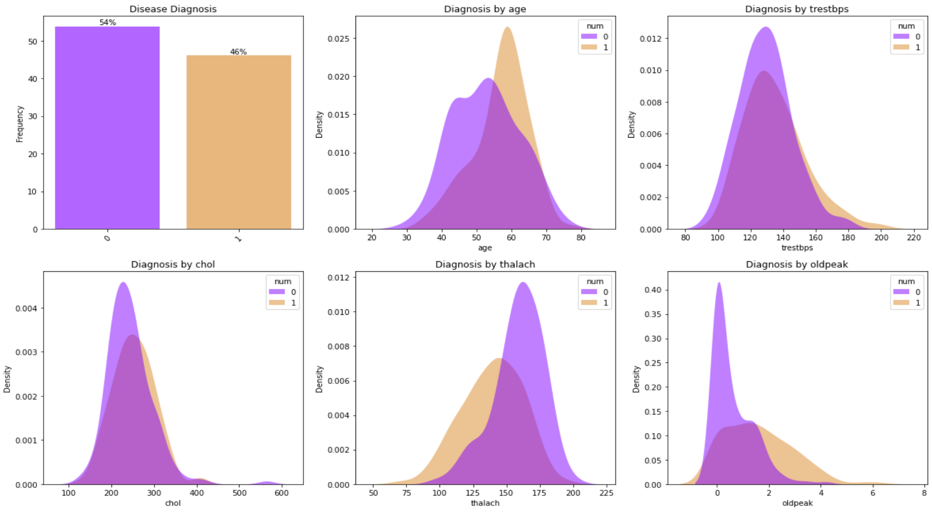


Figure 2. *Distribution of target variable (heart disease diagnosis) by the numeric features*

Inference: We can see that there is an almost equal distribution of patients with no disease and those that have it. We observe that older people from the age range of 51 to 65 have higher risk of having the disease since the number of patients with the disease is high at that age range with a mean of 56.75. We also noticed that most patients with higher resting blood pressure above 120-140 have a higher risk of having the disease. People having high Serum cholesterol and those with low maximum heart rate are at higher risk of having the disease.

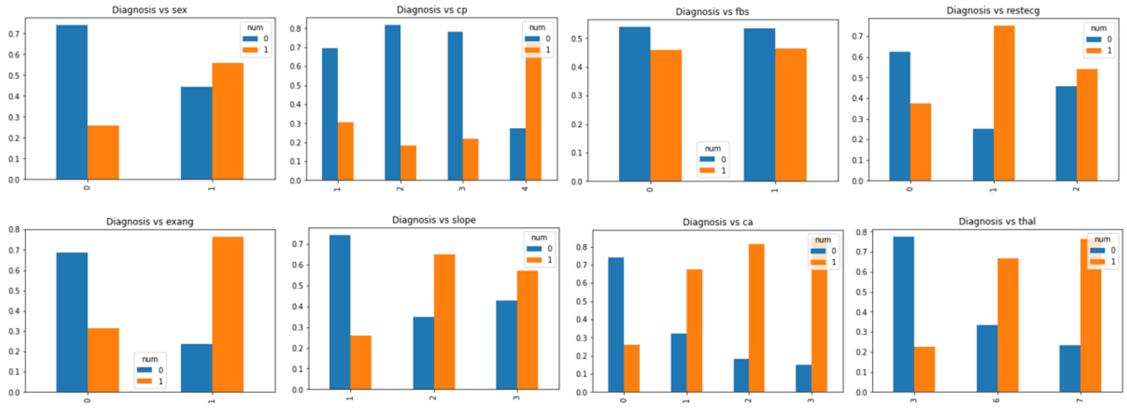
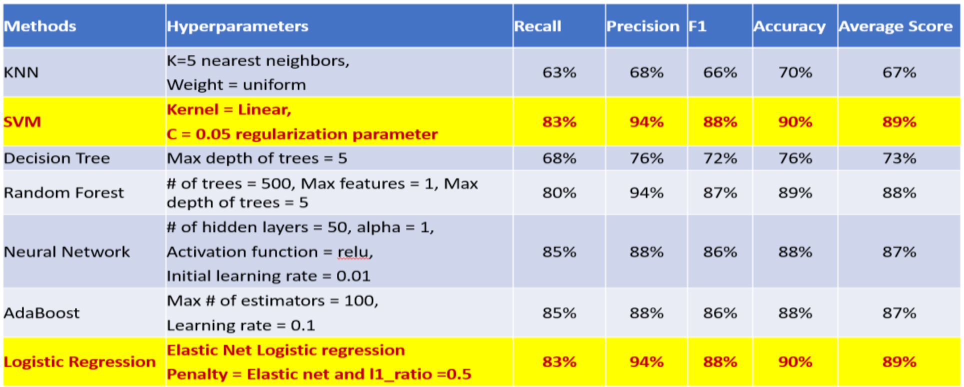


Figure 3. *Distribution of target variable (heart disease diagnosis) by the Discrete features*

Inference: This analysis shows that males are at higher risk of having the disease than females, also those that have no chest pain are at higher risk than other types of chest pain. Those that have fasting blood sugar are at almost the same risk as those that do not have. We also observed that, people at higher risk of heart disease are those having exercise induced Angina, those having ECG of wave abnormality (category 1), those who falls in the category of flat slope of peak exercise(category 2), and those having reversible defect thallium heart scan.

For more information about the individual distribution of each feature kindly refer to the Appendix file. The performance of all the machine learning classifiers used in training the dataset are shown in the table below:



Methods	Hyperparameters	Recall	Precision	F1	Accuracy	Average Score
KNN	K=5 nearest neighbors, Weight = uniform	63%	68%	66%	70%	67%
SVM	**Kernel = Linear,** **C = 0.05 regularization parameter**	**83%**	**94%**	**88%**	**90%**	**89%**
Decision Tree	Max depth of trees = 5	68%	76%	72%	76%	73%
Random Forest	# of trees = 500, Max features = 1, Max depth of trees = 5	80%	94%	87%	89%	88%
Neural Network	# of hidden layers = 50, alpha = 1, Activation function = relu, Initial learning rate = 0.01	85%	88%	86%	88%	87%
AdaBoost	Max # of estimators = 100, Learning rate = 0.1	85%	88%	86%	88%	87%
Logistic Regression	**Elastic Net Logistic regression** **Penalty = Elastic net and l1_ratio =0.5**	**83%**	**94%**	**88%**	**90%**	**89%**

Figure 4. *Performance of all machine learning classifiers using Hyperparameters that yield the best result*

From the performance table given above, Linear support vector machine classifier and Elastic net logistic regression have the best overall performance in terms of accuracy, precision, f1, and average score compared to all the other methods used.

CONCLUSION

In other to help clinicians determine the best performing machine learning algorithm able to predict the presence of heart disease in people at an early stage, we carried out our analysis using python where we loaded and preprocessed the Cleveland Heart Disease dataset so as to prepare the data for EDA and predictive analysis. A huge load of work was done in the EDA and predictive analysis part of this project taking about a week to complete. We carried out EDA on the preprocessed data and we discovered some groups of people that are at higher risk of having heart disease: older people with the age range of 51-65 years. People with resting blood pressure above 120-140, high Serum cholesterol, and low maximum heart rate are at higher risk. Males are at higher risk of having the disease than females. Also, those having exercise induced Angina, ECG of wave abnormality (category 1), those who falls in the category of flat slope of peak exercise(category 2), and those having reversible defect thallium heart scan (category 3) are at higher risk. We split our data set into 70% training and 30% test data, and train different classifiers on the training data using different set of hyperparameters (Report can be found in the Appendix file) and we evaluated the performance of each classifier on the test data selecting the combination of hyperparameter with the best result. Linear support vector machine classifier and Elastic net logistic regression turned out to give the best overall performance and will be highly recommended for use.

REFERENCES

- Dataset: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
- Related Work: http://csjournals.com/IJCSC/PDF7-1/18.%20Tejpal.pdf