Heart Disease Prediction and Factors Analysis
CSE 841: Artificial Intelligence - Final Project Writeup

Asadullah Hill Galib

17 December 2020

1 Introduction

This project aims at predicting heart disease effectively with consideration of performance measures and significant factors/attributes analysis. So, it addresses two aspects: how the factors influence heart disease prediction and how well we can predict heart disease. By doing so, heart disease can be analyzed and predicted more effectively.

Several machine learning techniques with various configurations are employed here for prediction and important factors analysis. UCI repository: Cleveland database [1] is used here for evaluation. Finally, this study suggests an analysis of important factors and recommend machine learning techniques to effectively predict heart disease.

2 Formal statement of the computational problem

Formally, this project addresses the following two scientific questions:

1. *How the factors/attributes influence heart disease prediction?* - This question addresses the importance of each factor, analyzes the top important factors in heart disease prediction.

2. *How well can we predict heart disease using machine learning techniques?* - This question addresses heart disease prediction using several machine learning techniques, which models work well, and how well those models perform in predicting heart disease.

3 Related Works

Several prior works address these scientific questions. For instance, Nahar et al. [2, 3] investigate some computational intelligence techniques in the detection of heart disease using six well-known classifiers for the Cleveland dataset. It identifies heart disease risk factors. Medhekar et al. [4] use the Naive Bayes classifier in heart disease prediction. Batii et al. [5, 6] propose a Hybrid Naïve Possibilistic Classifier (HNPC) for heart disease detection from the heterogeneous data.

Kumar et al. [7] predict heart disease using an advanced fuzzy resolution mechanism. Shah et al. [8] extract high impact features using Probabilistic Principal Component Analysis (PPCA) for heart disease prediction. Amin et al. [9] identify significant features for heart disease prediction using data mining techniques. Uyar et al. [10], Fida et al. [11], and Gokulnath et al. [12] employ a genetic algorithm-based technique for prediction. Rani et al. [13] use neural networks for heart disease prediction.

All of the research address the first scientific question. Nahar et al. [2, 3], Shah et al. [8], and Gokulnath et al. [12] also address the second scientific question. All the mentioned works use the Cleveland dataset.

4 Algorithmic Details and Implementation Details

This project is being implemented in Python on the Jupyter Notebook platform. Basically, the following three parts are implemented and analyzed.

4.1 Data Exploration and Preprocessing:

Before data preprocessing, data is being explored. Among 303 instances, 165 instances are of heart attack and 138 instances are not of heart attack. The distribution of individual features of the dataset is plotted using a scatter plot. Figure (1) depicts that distribution. According to Figure (1), the distribution of numeric and categorical values are visualized, and the possibility of outliers can be speculated.

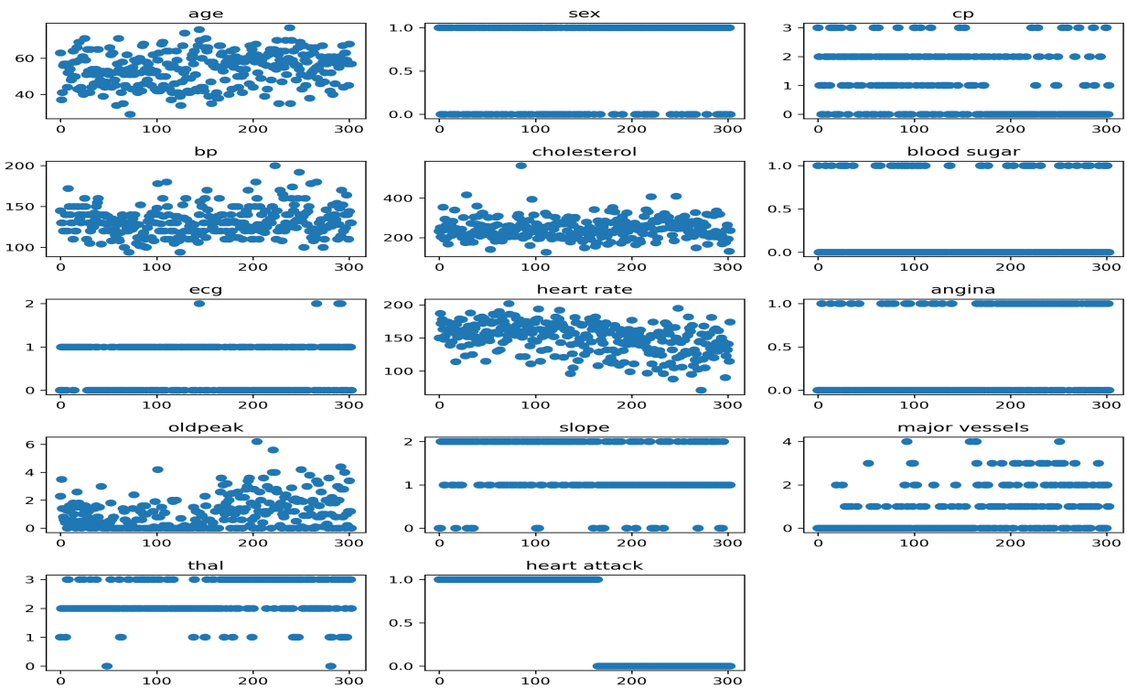

Figure 1: Scatter Plot of All Features

Also, the correlation among features is measured and plotted. According to Figure (2), there is no notable correlation among the features. The highest correlated features are *slope* and *oldpeak*, but their correlation coefficient is not so high: -0.58.

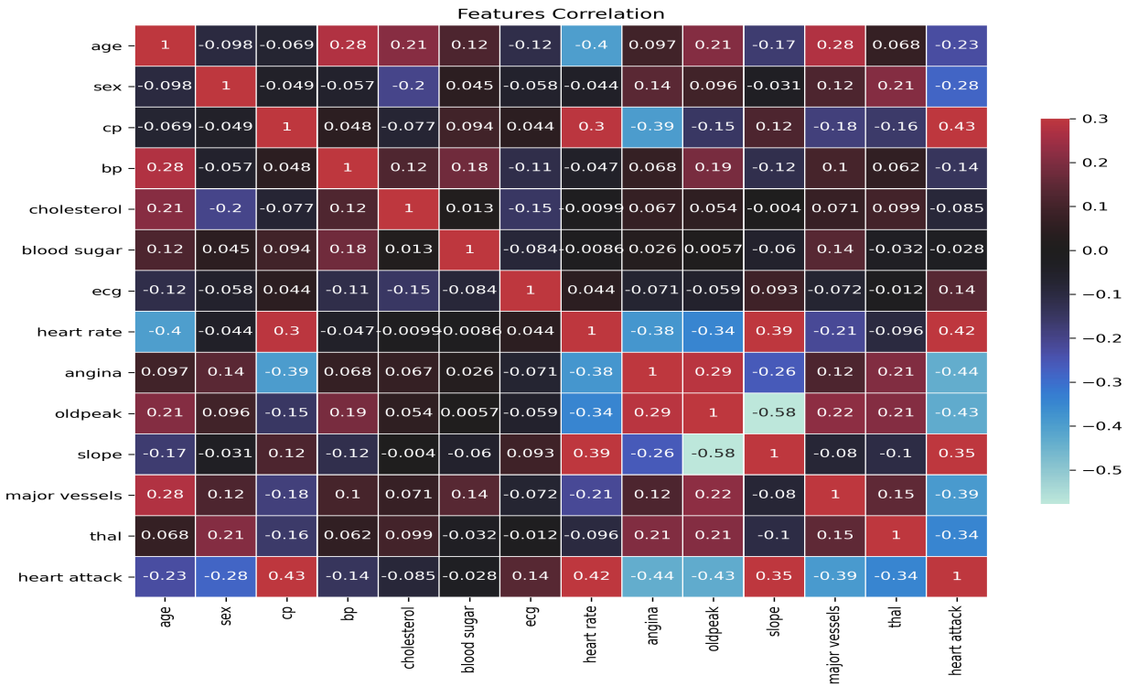

Figure 2: Correlation Among Features

In data preprocessing, data is being normalized first. The values of numeric columns are changed to a standard scale, without distorting the variations between the value ranges.

Afterward, outliers imputation is carried out for numeric features. Data-points outside the two standard deviations are identified as possible outliers as 95% of the data-points lie in two standard deviations. Those possible outliers are imputed with the mean value of the corresponding feature. The following number of outliers have been identified after outliers imputation: (bp: 8, age: 2, cholesterol: 17, heart rate: 15, oldpeak: 8).

Finally, the dataset is split for 10-fold cross-validation. 10-fold cross-validation allows out-of-sample accuracy to be measured more accurately. As each observation is used for both training and test, it triggers more successful use of data.

4.2 Important Factors Identification in Heart Disease

To figure out important factors and their relative importance in heart disease, four machine learning algorithms are being evaluated.

Feature Importance using Chi-Squared Test: SelectKBest scores the features according to the k highest scores. It takes a score function as a parameter [16]. In this study, the chi2 scoring function is employed. This scoring function computes the chi-squared stats between each non-negative feature and class scores accordingly. It tests for which the distribution of the test statistic approaches the χ^2 (Chi-Squared) distribution asymptotically [17].

Feature Importance using Random Forest Classifier: SelectFromModel is a meta transformer that can be used along with any tree-based estimator. It calculates the feature importance of each feature according to fitting the estimator into the data. Based on the feature importance it selects the top N features, where N is predefined [18]. Tree-based estimators are used here as it can classify the significant features by selecting the classification features based on how well they boost the node's purity [19]. Random Forest Classifier is used here as the tree-based estimator.

Feature Importance using Mutual Information Gain: Mutual Information is a non-negative value between two random variables, which measures dependency between variables. It measures the quantity of information gained by analyzing the other random variable involving one random variable. It is equal to zero if there are two independent random variables, and higher values mean higher dependence. [20, 21].

Feature Importance using ROC-AUC: An AU-ROC curve (receiver operating characteristic curve) is a graph representing a classification model output at all classification thresholds. This curve maps two parameters: True Positive Rate and False Positive Rate. The region under the AU-ROC curve is proportional to the probability that a classifier ranks a randomly selected positive instance higher than a randomly selected negative one by using normalized units [22].

All the features are analyzed using the above mentioned feature importance measurement techniques to figure out important features in common.

4.3 Predicting Heart Disease using Machine Learning Techniques

In predicting heart disease using machine learning techniques, 12 machine learning algorithms have been employed with different configurations to get better performance. All the algorithms are very popular and widely used. Also, 10-fold cross-validation is used to evaluate the performance of each model. Five performance metrics are taken into consideration: accuracy, precision, recall, F1-score, ROC-AUC (Receiver Operating Characteristic-Area Under the Curve).

Different configurations of the algorithms used in this study are described below with their corresponding parameters.

- **Support Vector Machine (SVM)**
  - kernel: linear, poly, rbf, sigmoid
  - C: 0.5, 1, 2
- **Random Forest**
  - number of estimators: 100
  - criterion: gini, entropy
  - max depth: 10
- **Extra Trees**
  - number of estimators: 100
  - criterion: gini, entropy
  - max depth: 10
- **Logistic Regression**
  - penalty: l2, l1, none
  - C: 0.5, 1, 2
- **Gradient Boosting**
  - number of estimators: 100
  - loss: deviance, exponential
  - learning rate: 0.1, 0.5

  - criterion: mse
  - max depth: 10

- **AdaBoost**
  - number of estimators: 100
  - learning rate: 0.1, 0.5

- **KNN Classifier**
  - number of neighbors: 5
  - weights: uniform, distance

- **Decision Tree**
  - number of estimators: 100
  - criterion: gini, entropy
  - max depth: 10
  - splitter: best, random

- **Naive Bayes**

- **Linear Discriminant Analysis**
  - number of components: 1, 5
  - solver: svd, lsqr, eigen

- **Principal component analysis (PCA)**
  - number of components: $2\tilde{1}0$

- **Stacking Classifier**
  - estimators:
    * Extra Trees (number of estimators: 100, criterion: gini)
    * Random Forest Classifier (number of estimators: 100, criterion: gini)
    * Support Vector Machine (SVM) (kernel: sigmoid, C:1)
  - final estimator: Extra Trees

Here, the Stacking classifier is used with three top-performing classifiers.

5 Dataset and Experimental Design

This section describes the dataset and experimental design of this study.

5.1 Dataset

The Cleveland dataset (UCI, 1990) is used in this study which is a public dataset and can be found in the University of California Irvine (UCI) Machine Learning Repository [1]. The dataset contains 303 instances of real patient data. There are 14 attributes:

1. **age:** age of the person in years
2. **sex:** gender of the person (1: male, 0: female)
3. **cp:** chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
4. **bp:** resting blood pressure in mm Hg (ranges from 94 to 200)
5. **cholesterol:** serum cholestoral in mg/dl (ranges from 126 to 564)
6. **blood sugar:** fasting blood sugar in mg/dl (categorical values: 0, 1)
7. **ecg:** resting electrocardiograph results (categorical values: 0, 1,2)
8. **heart rate:** maximum heart rate achieved (ranges from 71 to 202)
9. **angina:** exercise induced angina (categorical values: 0, 1)

10. **oldpeak:** ST depression induced by exercise relative to rest (ranges from 1 to 3)
11. **slope:** slope of the peak exercise ST segment (categorical values: 1, 2, 3)
12. **major vessels:** number of major vessels colored by flourosopy (ranges from 0 to 3)
13. **thal** categorical values (0: normal, 1: fixed defect, 2: reversable defect
14. **target class - heart attack:** 0: less chance of heart attack, 1: more chance of heart attack

5.2 Experimental Design

This project is being implemented in Python on the Jupyter Notebook platform. The implementation was carried out on a system with the following specifications.

- Operating System: Windows 10
- RAM: 16.00 GB
- CPU: 3.70GHz AMD Ryzen 5 3400G with Radeon Vega Graphics

6 Results

The experimental results are depicted in this section. The two scientific questions are assessed separately.

6.1 Important Factors in Heart Disease

This subsection addresses the first scientific question using different machine learning algorithms.

Figure 3 depicts the feature importance using the Chi-squared Test. According to the Chi-Squared Test, *angina, cp, major vessels*, and *oldpeak* are the top important factors in predicting heart disease. Conversely, *cholesterol, blood sugar,* and *bp* are the less relevant factors.

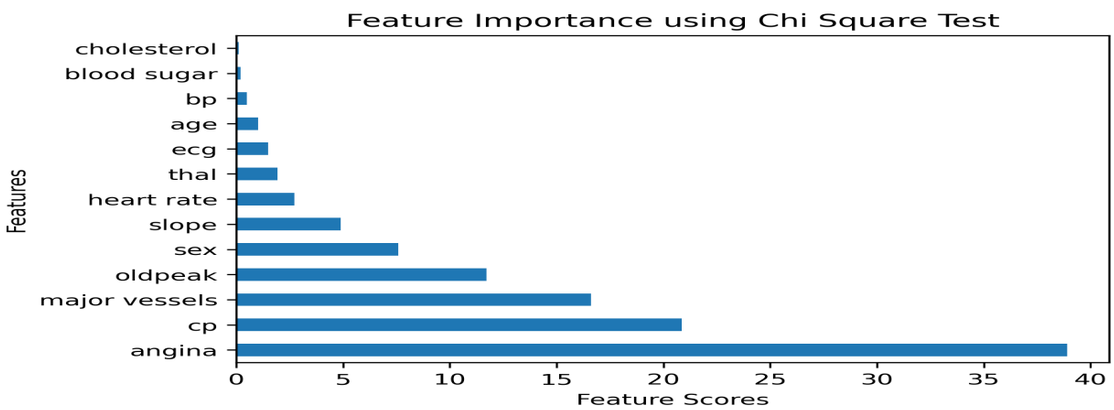

Figure 3: Feature Importance using Chi-Squared Test

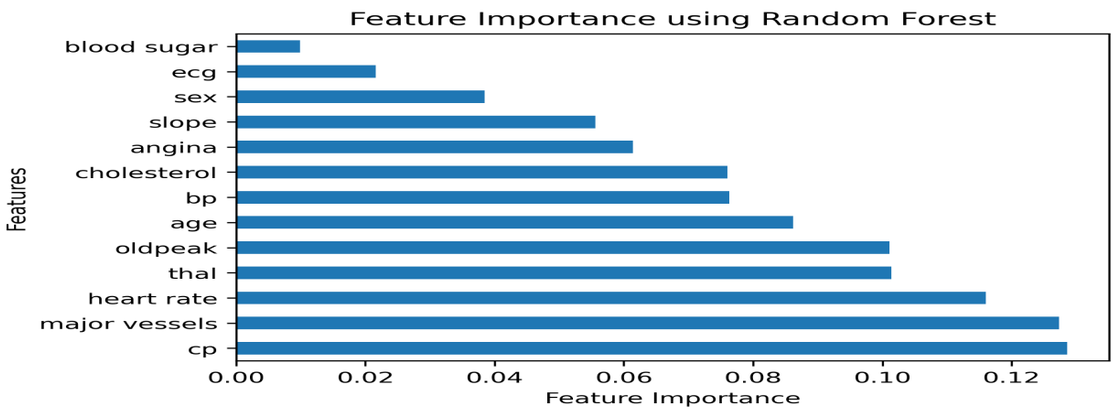

Figure 4: Feature Importance using Random Forest Classifier

Feature importance using Random Forest classifier is depicted in Figure 4. It shows that *cp, major vessels, heart rate, thal*, and *oldpeak* are the top most factors while *blood sugar, ecg*, and *sex* are the less relevant factors.

According to Mutual Information Gain (see Figure 5), *cp, major vessels, slope, oldpeak*, and *thal* are the top important factors in predicting heart disease. Contrarily, *blood sugar, age, bp*, and *ecg* are the less relevant factors.

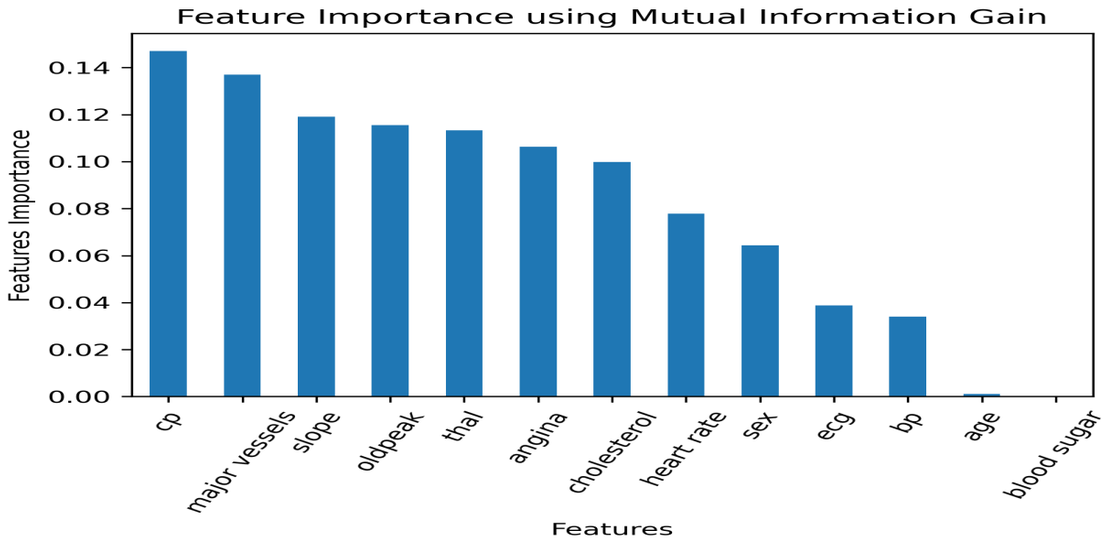

Figure 5: Feature Importance using Mutual Information Gain

Likewise, ROC-AUC scores suggest that (see Figure 6) *slope, cp, thal, oldpeak*, and *major vessels* are the top important factors in predicting heart disease. On the other hand, *blood sugar, bp*, and *sex* are the less relevant factors.

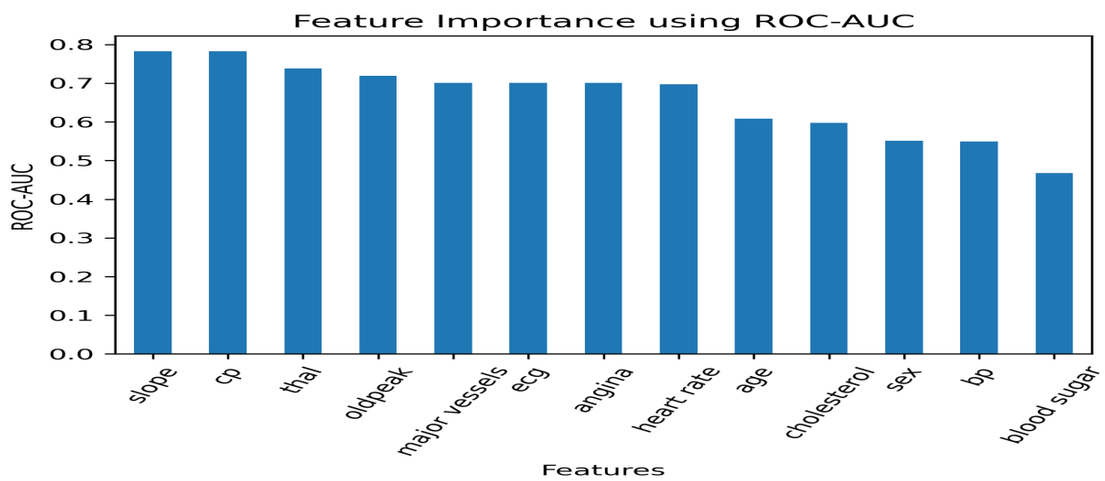

Figure 6: Feature Importance using ROC-AUC

In short, all the algorithms suggest that *cp, thal, oldpeak, slope*, and *major vessels* are top important factors in predicting heart disease whereas *blood sugar, ecg, sex*, and *bp* are less relevant factors with respect to others.

6.2 Heart Disease Prediction

This subsection addresses the second scientific question using five performance metrics: accuracy, precision, recall, F1-score, ROC-AUC.

Figure 7 depicts the accuracy of different machine learning algorithms. According to that, Extra Trees and Stacking classifiers have outperformed all other algorithms in terms of accuracy. Also, Principal Component Analysis (PCA) with 8 components, Random Forest, Support Vector Machine have performed well with respect to others.

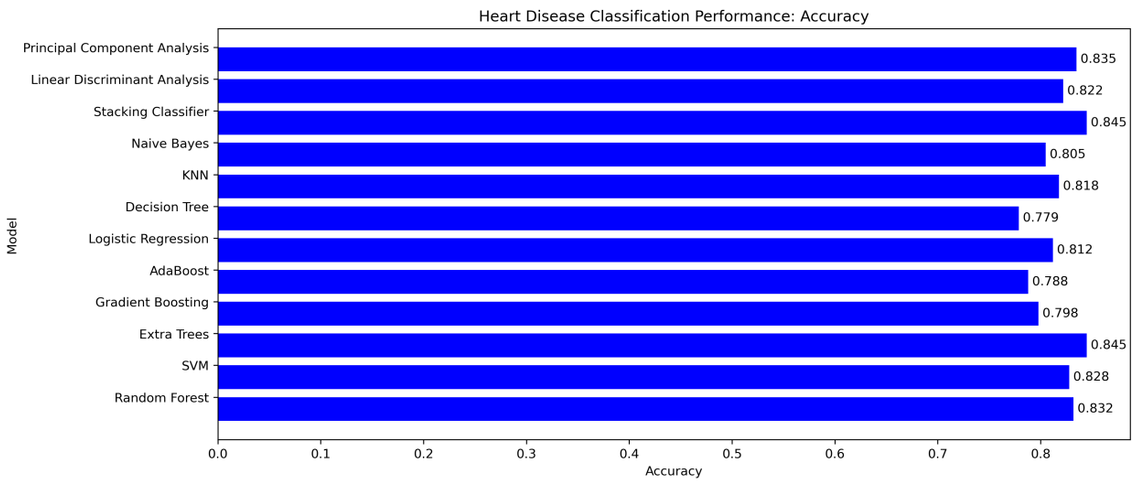

Figure 7: Feature Importance using Chi-Squared Test

According to Figure 8, Linear Discriminant Analysis, and Stacking classifiers have outperformed all other algorithms in terms of precision. Besides, Extra Trees, KNN, Random Forest, and Support Vector Machine have performed well.

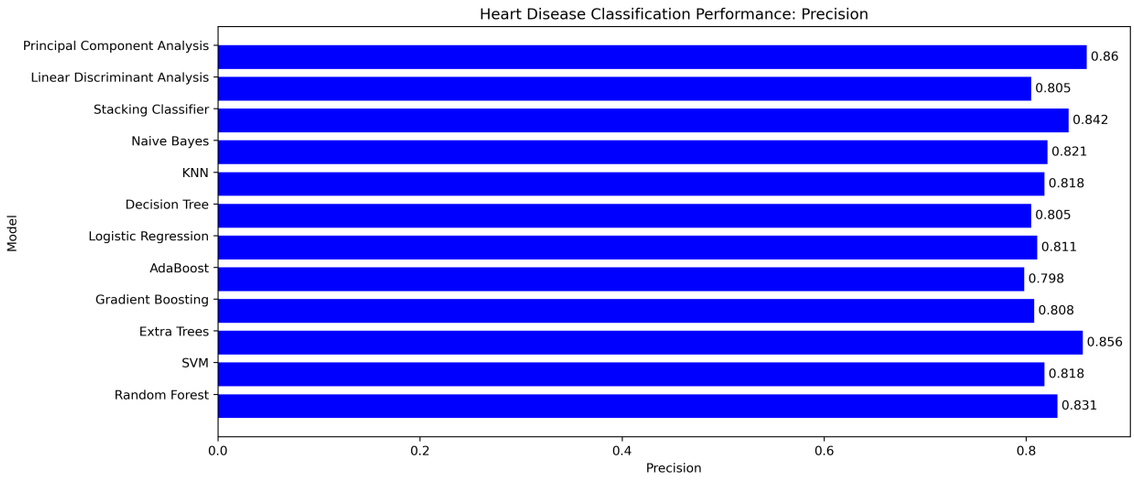

Figure 8: Feature Importance using Random Forest Classifier

Figure 9 depicts the recall of different machine learning algorithms. Here also, Extra Trees and Stacking classifiers have outperformed all other algorithms with regards to recall. Also, Random Forest and Support Vector Machine have shown better recall values.

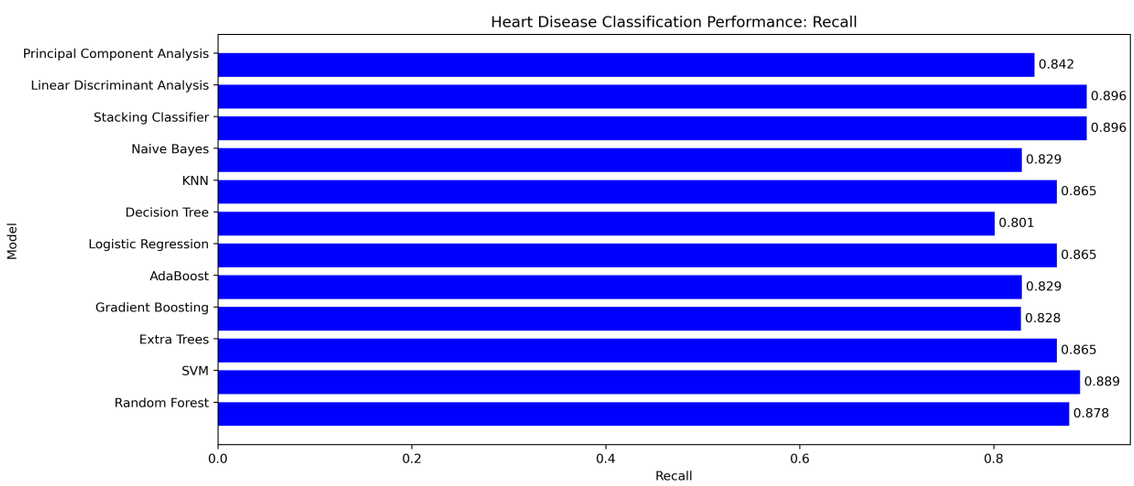

Figure 9: Feature Importance using Mutual Information Gain

According to Figure 10, Extra Trees and Stacking classifiers have the top F1-scores. Besides, Random Forest, Support Vector Machine, and Principal Component Analysis have performed well with respect to others.

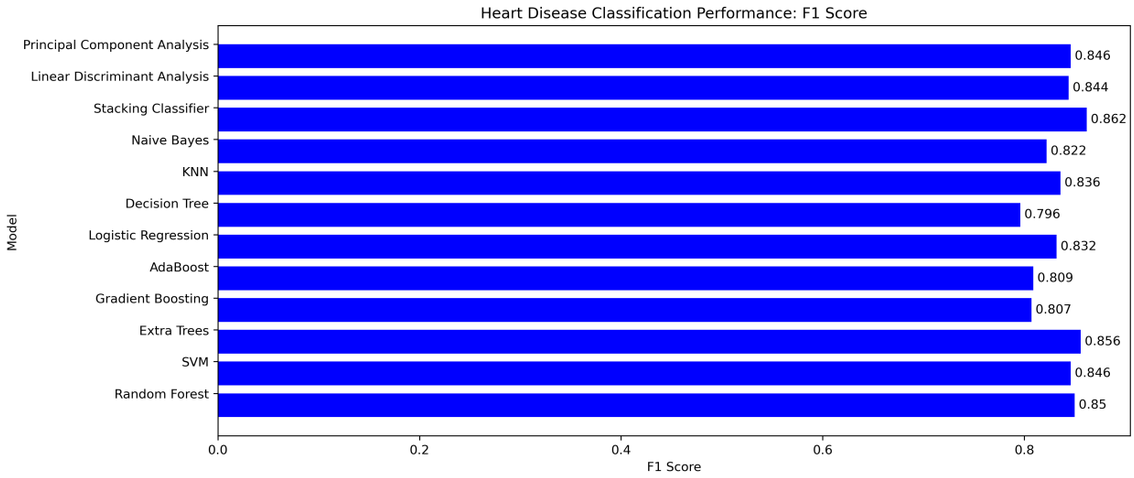

Figure 10: Feature Importance using ROC-AUC

Figure 11 depicts the ROC-AUC scores of different machine learning algorithms. Again, Extra Trees and Stacking classifiers have outperformed all other algorithms in terms of ROC-AUC. Also, Random Forest, Support Vector Machine, Principal Component Analysis, Linear Discriminant Analysis have shown better ROC-AUC scores.

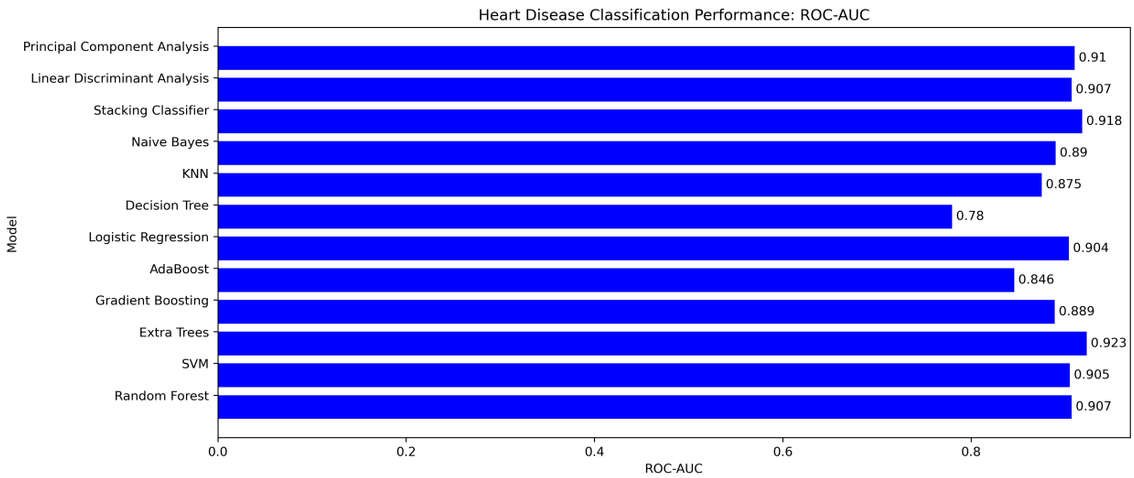

Figure 11: Feature Importance using ROC-AUC

In conclusion, Extra Trees and Stacking classifiers are top machine learning algorithms in predicting heart disease. Apart from that, Random Forest, Support Vector Machine, and Principal Component Analysis also have been performed better than other algorithms. So, using the top-performing classifiers, we can predict heart disease effectively.

7 Conclusions

This study incorporates four machine learning algorithms to analyze important factors and twelve machine learning algorithms to predict heart disease.

Experimental results suggest that *cp, thal, oldpeak, slope*, and *major vessels* are top important factors in predicting heart disease whereas *blood sugar, ecg, sex*, and *bp* are less relevant factors with respect to others. In predicting heart disease, Extra Trees and Stacking classifiers have outperformed others. Also, Random Forest, Support Vector Machine, and PCA have shown sound performance.

With the suggested important factors and recommended classifiers, heart disease can be analyzed and predicted effectively.

8 Future Work

In the future, the genetic algorithm will be employed in factor analysis and prediction. Existing research [10, 11, 12] suggest its applicability in heart disease prediction. Also, the neural network or ensemble learning-based techniques will be incorporated for prediction as those are effective in analyzing discerning features.

References

[1] Dua, D., Graff, C., and Detrano, R. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Nahar, J., Imam, T., Tickle, K. S., Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), 96-104.

[3] Nahar, J., Imam, T., Tickle, K. S., Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

[4] Medhekar, D. S., Bote, M. P., Deshmukh, S. D. (2013). Heart disease prediction system using naive Bayes. Int. J. Enhanced Res. Sci. Technol. Eng, 2(3).

[5] Baati, K., Hamdani, T. M., Alimi, A. M. (2014, August). A modified hybrid naive possibilistic classifier for heart disease detection from heterogeneous medical data. In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR) (pp. 353-358). IEEE.

[6] Baati, K., Hamdani, T. M., Alimi, A. M. (2013, December). Hybrid naive possibilistic classifier for heart disease detection from heterogeneous medical data. In 13th International Conference on Hybrid Intelligent Systems (HIS 2013) (pp. 234-239). IEEE.

[7] Kumar, A. S. (2013). Diagnosis of heart disease using Advanced Fuzzy resolution Mechanism. International Journal of Science and Applied Information Technology, 2(2), 22-30.

[8] Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., Hussain, S. A. (2017). Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications, 482, 796-807.

[9] Amin, M. S., Chiam, Y. K., Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[10] Uyar, K., İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, 588-593.

[11] Fida, B., Nazir, M., Naveed, N., Akram, S. (2011, December). Heart disease classification ensemble optimization using genetic algorithm. In 2011 IEEE 14th International Multitopic Conference (pp. 19-24). Ieee.

[12] Gokulnath, C. B., Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22(6), 14777-14787.

[13] Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. arXiv preprint arXiv:1110.2626.

[14] Tkachenko, R., Izonin, I., Kryvinska, N., Chopyak, V., Lotoshynska, N., Danylyuk, D. (2018). Piecewise-linear Approach for Medical Insurance Costs Prediction using SGTM Neural-Like Structure. In IDDM (pp. 170-179).

[15] Maity, S. K., Panigrahi, A., Mukherjee, A. (2018, August). Analyzing social book reading behavior on goodreads and how it predicts amazon best sellers. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 211-235). Springer, Cham.

[16] sklearn.feature_selection.selectkbest. https://scikit-learn.org/ stable/modules/generated/ sklearn.feature_selection.SelectKBest. html. Online; accessed 12 December 2020.

[17] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 50, no. 302, pp. 157-175, 1900.

[18] 1.13. feature selection. https://scikit-learn.org/stable/modules/ feature_selection.html. Online; accessed 11 December 2020.

[19] 11. feature engineering for machine learning - data science beginners. https://datasciencebeginners.com/ 2018/11/26/ 11-feature-engineering-for-machine-learning/. Online; accessed 13 December 2020.

[20] A. Kraskov, H. St¨ogbauer, and P. Grassberger, "Erratum: estimating mutual information [phys. rev. e 69, 066138 (2004)]," Physical Review E, vol. 83, no. 1, p. 019903, 2011.

[21] B. C. Ross, "Mutual information between discrete and continuous data sets," PloS one, vol. 9, no. 2, p. e87357, 2014.

[22] T. Fawcett, "An introduction to roc analysis," Pattern recognition letters, vol. 27, no. 8, pp. 861-874, 2006.