Machine Learning Methodology for Classification of Heart Disease

Koushik Sridhar [*1] **Dylan Yang** [*1] **Rohan Shankar** [*1] **Manas Panchavati** [*1]

Abstract

Heart Diseases (HD) are reported as the leading cause of death and disability worldwide. With the increasing rate of heart disease being a rising concern, we propose a machine learning methodology to classify and predict heart disease incidences among patient pools. In this paper, we perform correlation analyses to understand features, and developed and evaluated the performances of a Logistic Regression Classifier, a Support Vector Machine, a Decision Tree, a Random Forest, a K-Nearest Neighbors Classifier, a Naive Bayes Classifier, and an XGBoost Classifier. Weak to medium strength correlations were found between all variables to HD, with RestingBP and RestingECG having nearly 0 correlation. Best model performance in HD classification was obtained using the XGBoost Classifier. The results are expected to further personalized medicine and display effective machine learning usage in complex conditions.

1. Introduction

Cardiovascular disease (CD) is a type of heart disease that accounts for over 30% of all deaths worldwide. In CD, plaques develop on the walls of arteries which can obstruct blood flow, which could further result in heart attacks and/or strokes. Cardiovascular Disease arises as a result of various risk factors such as physical inactivity, unhealthy diet, and the use of alcohol and tobacco (Tao et al., 2018). By incorporating a healthy lifestyle such as reducing sodium intake, consuming fruits and vegetables, having daily exercise, and reducing alcohol and tobacco use, one can reduce their individual risk of heart disease (Spencer et al., 2020). A potential solution to overcome these problems is to use collections of patient records from various healthcare centers and analyze lifestyle factors and their respective patient outcomes. Early identification of heart disease using a prediction model is beneficial for fatality rate reduction. In addition, detection enables healthcare professionals to begin treatment at a much earlier stage in the disease and for evaluation of treatment response to occur. In addition, such predictive models would reduce the time and money spent on tests.

In this project, we propose the development of an algorithm based on machine learning to analyze patient data, and classify patients as those who are heart disease positive (HD+) and heart disease negative (HD-).

1.1. Research and Analysis Contribution

Machine learning and associated algorithms can be used to diagnose, detect, and forecast disorders in the medical industry. The primary purpose of this work is to analyze data distributions by correlation analysis to make basic analyses and develop and evaluate a series of machine learning classification models. This work creates a Logistic Regression Classifier, a Support Vector Machine, a Decision Tree, a Random Forest, a K-Nearest Neighbors Classifier, a Naive Bayes Classifier, and an XGBoost Classifier. These models are trained on the UCI Heart Disease dataset (Detrano et al., 1989), and each model is evaluated on the bounds of accuracy, precision, sensitivity, and specificity. As a result, future work will be able to create and deliver appropriate treatment due to the early detection nature of these algorithms, potentially avoiding serious effects of the disease by properly administered treatment.

2. Literature Review

Previous work proposed an HD prediction framework based on supervised machine learning algorithms including Support Vector Machines (SVMs), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) models all trained on the Cleveland Dataset from University of California, Irvine (UCI) machine learning repository (Rairikar et al., 2017). The models were developed based on 302 instances and 14 heart disease features in model scope. The research was a comparative analysis of selected techniques mentioned previously, where the NB model performing best with an accuracy of 86.6%. Researchers also employed NB and SVM models, but com-

[*]Equal contribution [1]Department of Computer Science, University of North Carolina - Chapel Hill, Chapel Hill, North Carolina, United States of America. Correspondence to: Koushik Sridhar <sridhark@email.unc.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

bined it with a line of combinational thinking, termed "Hybridization". This work combined several machine learning algorithms into a single model trained on the Cleveland dataset. Preprocessing reduced the Cleveland dataset to 12 features, with classification algorithms applied to the dataset. Specifically, researchers investigated the viability of NB, SVM, KNN, Neural Networks, Random Forests, and Generational Adversarial Networks. Once again, they found that NB and SVM performed better in heart disease prediction with same accuracy of 89.2%. Through these two works, we see that in datasets with 10-15 features involved, a NB or SVM model performs most effectively. However, these models are also prone to overfitting, and not much information was given regarding how overfitting was prevented in the training stages of these models.

Another work that explored possible prediction methods deviated from the NB/SVM models of the past, proposing a HD prediction framework using Multi-Layer Perceptron Neural Networks (MLP) built with back-propagation as the training algorithm (Subhadra & Vikas, 2019). The model was evaluated on the bounds of sensitivity, specificity, precision and accuracy. The data was preprocessed to be built on only 14 features, and resulted in accuracy of 93.39% for a 5-neuron hidden layer. This work further standardizes the feature reduction concept for this heart disease dataset. The work done with MLP inherently reduces concerns about potential overfitting, and displays a high accuracy rate amongst algorithmic approaches.

Alternative work proposed a model that combines descriptive and predictive techniques of data analysis for HDs algorithmically (Shamsollahi et al., 2019). Datasets including 282 CVD instances and 58 features were used. Preprocessing removed missing values and outliers. For the descriptive technique, researchers employed a K-means algorithm as a clustering method for further analysis. Once clusters were determined, various classification algorithms including CHAID, Quest, C5.0, C & RT-DT, and ANN were chosen. Results showed that C & RT-DT performed best with an error of 0.0704 when the complete dataset was used. When performing cluster-specific analysis, C & RT-DT performed better in clusters 1 and 2 with 0.022 and 0.023 errors respectively. Surprisingly, CHAID performed best in cluster 3. However, this work does not explain why the descriptive model developed clusters 1-3 in the manner that it did. Thus, we cannot conclusively determine the basis for why the algorithms performed as well as they did, but further work can expand on potentially using a descriptive and predictive meshed model in an effort to understand the inner workings of predictive models and steer away from "black-box" methodologies.

3. Dataset Analysis

The dataset consists of health and lifestyle factors from patients as part of the Cleveland Heart Disease Database. The dataset has 918 observations (representing individual patients), was cleaned and preprocessed, and was split 80-20 for training and testing accordingly. Each observation contains 12 features, the final feature being the classification of the patient (HD+ or HD-).

3.1. Data Pre-processing

After examining the data, it was apparent that certain features were missing in a select few (less than 1%) of observations. For this missing data, numerical data was filled with the median of that feature, and observations with missing categorical information were dropped. After performing this form of preprocessing, a correlation matrix was developed across all features. The results of this correlation matrix are depicted in Figure 1.

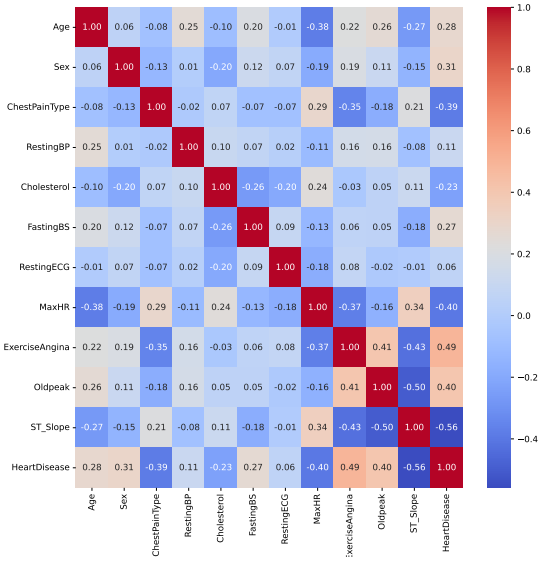

Figure 1. Correlation Matrix depicting relationships between Heart Disease dataset features

The correlations between Heart Disease and each feature were of primary importance in the preprocessing stage of the project. Looking at those values we see no significant correlation between the RestingBP and RestingECG features to Heart Disease. In addition, all other features have weak to medium correlations with Heart Disease outcomes individually. Further analysis is performed to determine the significance of features to heart disease, splitting analysis by categorical and numerical features. Following this analysis, all numerical data was normalized and standardized, and all categorical features were one-hot encoded.

3.2. Feature Engineering

Feature Analysis for Categorical Features: To further understand the relationship and significance of categorical features with Heart Disease predictability, Chi-Square Test was performed as shown in Table 1. RestingECG's low Chi-Square score indicates it is not important for HD prediction and as a result was dropped in model development.

Table 1. Chi-Square Score Results for Categorical Features.

FEATURE	SCORE
CHESTPAINTYPE	160.74
EXERCISEANGINA	133.64
ST-SLOPE	77.49
FASTINGBS	50.30
SEX	18.01
RESTINGECG	1.22

Feature Analysis for Numerical Features: To further understand the relationship and significance of numerical features with Heart Disease predictability, ANOVA test was performed as shown in Table 2. RestingBP's low ANOVA score indicates it is not important for HD prediction, and as a result was dropped in model development.

Table 2. ANOVA Score Results for Categorical Features.

FEATURE	SCORE
OLDPEAK	178.62
MAXHR	174.91
AGE	79.16
CHOLESTEROL	52.46
RESTINGBP	10.73

4. Modeling

This project features the development of a Logistic Regression Classifier, a Support Vector Machine, a Decision Tree, a Random Forest, a K-Nearest Neighbors Classifier, a Naive Bayes Classifier, and an XGBoost Classifier. Most of these models were developed using the facilities of the `sklearn` library, with the XGBoost model being developed with its own `xgboost` library. The models were developed while applying `GridSearchCV` to complete hyperparameter tuning and prevent overfitting. The results of each model will reflect the output due to optimized hyperparamter usage. Confusion Matrices were developed for each model, illustrating the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Each model was evaluated on its accuracy, precision, sensitivity, and specificity. Accuracy is the fraction of predictions the model got right. Precision is a measure of the quality of a positive prediction made by a model. Sensitivity is a measure of a model's ability to designate a patient as positive. Specificity on the other hand is a measure of a model's ability to designate a patient as negative. Each of these metrics are computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
$$Precision = \frac{TP}{TP+FP}$$
$$Sensitivity = \frac{TP}{TP+FN}$$
$$Specificity = \frac{TN}{FP+TN}$$

4.1. Logistic Regression

The first model developed was a Logistic Regression model to serve as a baseline model for HD classification. `GridSearchCV` functionality was completed using the `LogisticRegressionCV` sub-package in `sklearn`. Hyperparameters were autofit due to the sub-package.

4.2. Support Vector Classifier (SVC)

Hyperparameter tuning was performed on possible C values (integers from 0 to 5 inclusive), kernel types (linear, polynomial, rbf, and sigmoid functions), and gamma settings (scale or auto). This resulted in 40 candidate functions being tested over 5 folds, creating 200 possible fits. The most optimal SVC was the one where C=2, kernel=rbf, and gamma=auto.

4.3. Decision Tree Classifier

Hyperparameter tuning was performed on possible maximum depth constraints, minimum sample splits, minimum leaf samples, and maximum features. This resulted in 400 candidate functions being tested over 5 folds, creating 2000 possible fits. The most optimal Decision Tree was one where maximum depth=5, minimum sample split=15, minimum leaf samples = 1, and maximum features=5.

4.4. Random Forest Classifier

Hyperparameter tuning was performed on possible maximum depth constraints, minimum sample splits, minimum leaf samples, and number of estimators. This resulted in 500 candidate functions being tested over 5 folds, creating 2500 possible fits. The most optimal Random Forest was one where maximum depth=30, minimum sample split=5, minimum leaf samples = 10, and number of estimators=100.

4.5. K-Nearest Neighbors Classifer (KNN)

Hyperparameter tuning was performed on possible leaf size, number of neighbors, and power value for computing distance. This resulted in 72 candidate functions being tested over 5 folds, creating 360 possible fits. The most optimal KNN was one where leaf size=1, number of neighbors=3, and power=2 (meaning that the function would compute

Euclidean distance when making calculations).

4.6. Naive Bayes Classifer

Hyperparameter tuning was performed on the smoothing factor. 100 candidate values were tested over 10 folds, creating 1000 possible fits. The most optimal Naive Bayes was one where the smoothing factor equals about 0.000123.

4.7. XGBoost Classifier

Hyperparameter tuning was performed on the number of estimators, max depth, learning rate, column sampling factor, booster type, and minimnum child weight. This resulted in 15390 candidate functions being tested over 10 folds, creating 153900 possible fits. The most optimal XGBoost Classifier was one where the booster= 'gbtree', column sampling factor = 0.7, learning rate= 0.4, max depth= 4, minimum child weight=0.001, and the number of estimators= 14.

5. Results & Discussion

From Table 3, we can see the accuracy, precision, sensitivity, and specificity of the models developed as a result of this project. Across all metrics, we find that the XGBoost model performed the best with 88.6% accuracy, 88.1% precision, 92.3% sensitivity, and 83.8% specificity. On the other hand, the Logistic model featured the worst accuracy, precision, sensitivity, and specificity values across the board. Looking at Figure 2, we can inspect how the XGBoost models placed importance on features. We note that MaxHR and Age were considered as the most important factors while FastingBS and Sex were the least important factors in terms of classification as denoted by their respective F-scores.

Table 3. Classification Metrics for Models on Heart Disease Dataset.

MODEL	ACCURACY	PRECISION	SENSITIVITY	SPECIFICITY
LOGISTIC	82.6	81.6	89.4	73.8
SVC	85.3	82.4	94.2	73.8
DECISIONTREE	84.2	82.1	92.3	73.8
RANDOMFOREST	87.0	84.5	94.2	77.5
KNN	84.2	83.8	89.4	77.5
NAIVEBAYES	83.7	83.0	89.4	76.3
XGBOOST	88.6	88.1	92.3	83.8

5.1. Confounding Variables

The development of such models is dependent on the data that is used to train the models. When performing exploratory data analysis, we found that the distributions of the data overall were typically normally distributed. However, when investigating data distributions under strictly HD+ cases, we saw that the data was not evenly distributed. As a result of this uneven distribution, we believe that the weights

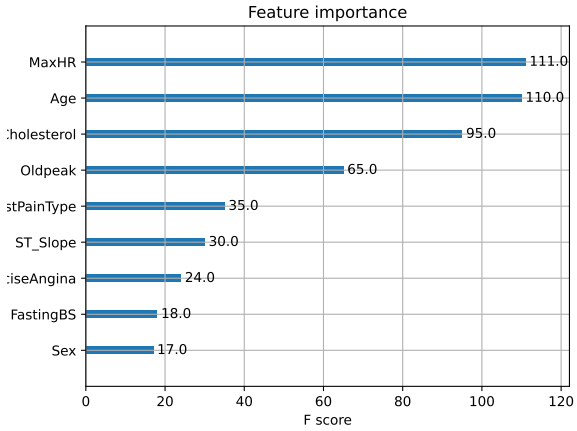

Figure 2. Feature Importance of XGBoost model

associated with each of the features under the models may be affected. For example, in looking at the comparison within sex, we saw that Male patients made up 90% of the HD+ pool which could affect weighting such that male patients are naturally weighted towards being HD+ which is not representative of real life. Similarly, other features within the HD+ or HD- cases may be distributed improperly and thus may sway model decision-making accordingly. Finding more real-world representative data is important for further tuning of these models.

6. Conclusion

This paper provided an analysis of the UCI Heart Disease dataset, inspecting data distributions and correlations in the data as initial forms of analysis. The paper saw the development of seven models: a Logistic Regression Classifier, a Support Vector Machine, a Decision Tree, a Random Forest, a K-Nearest Neighbors Classifier, a Naive Bayes Classifier, and an XGBoost Classifier. Best performances in HD classification were obtained using the XGBoost Classifier across all metrics (accuracy, precision, sensitivity, specificity). Exploratory Analysis displayed issues in data distributions within HD+ which could prove to be a confounding factor, and gives a greater need for more quality heart disease data. The work gives an idea regarding the effectiveness of machine learning-based heart detection methods. Future work may focus on conducting analyses and developing models with more features, and can focus on live integration with healthcare professional tests.

Acknowledgements

The authors would like to thank Jorge Silva for teaching the course on Machine Learning. Through this course, an improved understanding of developing and evaluating machine learning models and algorithms was developed.

References

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., and Lamgunde, A. Heart disease prediction using data mining techniques. In *2017 International conference on intelligent computing and control (I2C2)*, pp. 1–8. IEEE, 2017.

Shamsollahi, M., Badiee, A., and Ghazanfari, M. Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: a case study approach. *Journal of AI and Data Mining*, 7(1):47–58, 2019.

Spencer, R., Thabtah, F., Abdelhamid, N., and Thompson, M. Exploring feature selection and classification methods for predicting heart disease. *Digital health*, 6: 2055207620914777, 2020.

Subhadra, K. and Vikas, B. Neural network based intelligent system for predicting heart disease. *International Journal of Innovative Technology and Exploring Engineering*, 8 (5):484–487, 2019.

Tao, R., Zhang, S., Huang, X., Tao, M., Ma, J., Ma, S., Zhang, C., Zhang, T., Tang, F., Lu, J., et al. Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. *IEEE Transactions on Biomedical Engineering*, 66(6): 1658–1667, 2018.