Analyzing the Effectiveness of Various Machine Learning Algorithms in the Classification and Prediction of Cardiovascular Disease

Hasham Zahid

November 27, 2023

Table of Contents

Abstract

Cardiovascular diseases (CVDs) are the leading cause of death globally and are among the most prevalent chronic diseases in the United States, leading to 1 in 5 deaths annually. The term "heart disease" or CVD can refer to several types of heart conditions, with the most common type being coronary artery disease (CAD). Eventually, most of theses heart diseases will lead to a myocardial infarction (heart attack) or a stroke. It is paramount that timely, effective, and precise intervention be taken to ensure that patients are diagnosed and treated properly. In this respect, machine learning (ML) algorithms and approaches can help in the classification and prediction of heart diseases by looking at patient history, chronic diseases, and past behaviours, as sometimes heart disease is not diagnosed until an individual experiences the symptoms of heart failure.

Machine learning algorithms can be used to assess behavioural risk factors for heart disease, the most important of which are: unhealthy diet, physical inactivity, smoking, high alcohol consumption, high cholesterol, hypertension, and obesity, among other things. Identifying those at the highest risk of heart failure and disease can lead to ensuring they receive proper treatment and prevent their premature deaths. The effectiveness of multiple different ML models, including Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests, and XGBoost in the classification and prediction of heart disease was analyzed and compared.

The objective was to assess which model would be the most accurate and appropriate, therefore improving detection methods and preventing negative outcomes. Additionally, the features most useful in predicting heart disease were also outlined for preventative health screening of risk factors. This would give insight into the most associated with and most predicative risk

factors of heart disease and fatality, as well as which subset of risk factors may be used to accurately predict whether an individual has heart disease or not.

The data used in this study was an open dataset from 2019 conducted by the Centre of Disease Control and Prevention's (CDC) annual Behavioral Risk Factor Surveillance System (BRFSS), available from [https://www.cdc.gov/brfss/annual_data/annual_2019.html]. It consisted of 418,268 individuals and had 343 features. The dataset was cleaned, and features were selected according to relevance, including but not limited to, high blood pressure, high cholesterol, body mass index (BMI), smoker, stroke, diabetes, and physical activity and more. The working dataset had 256203 responses and 18 features. Exploratory data analysis was applied to the dataset to understand relationships, correlations, and distributions in Python, using python libraries such as pandas, NumPy, matplotlib, seaborn and scikit. Cross validation was performed on all models and randomized search was used to find the best parameters for the model. Imbalanced target variable was also handled using random under sampling and SMOTE. The dataset was split into train and test sets for each of the following models: Logistic Regression, Random Forests, Support Vector Machines, K-Nearest Neighbors and XGBoost. Each model was evaluated for its accuracy, precision, recall and F1 score, as well as fit time.

Literature Review

Cardiovascular diseases cause approximately one third of all deaths worldwide (Khan et. al, 2020). The most common cause of cardiovascular disease is ischemic heart disease, which clinically manifests as a myocardial infarction or heart attack and can lead to debilitating outcome such as chronic disabilities and even death (Khan et. al, 2020). Over the past few decades, cardiovascular disease has been increasing, mostly due to diet and lifestyle changes, as well as steep rise in population aging (Khan et. al, 2020). There are many factors that play a role in cardiovascular disease, including mental health, physical health, chronic behaviours and risk factors, and sleep quality and length (Khan et. al, 2020). The main risk factors for heart disease are elevated blood pressure, high cholesterol, type 2 diabetes, smoking, alcohol consumption, and obesity (Hopkins et. al, 2010). The increasing financial burden of cardiovascular disease is also a vast problem and is expected to rise to almost one trillion US dollars by the year 2030 (Khan et. al, 2020).

Heart disease and failure detection and diagnosis can be challenging, especially in the early stages of heart failure (Plati et. al, 2021). This is mostly due to heart disease being known as a 'silent killer', where it may be too late to help a patient and the disease does not show until the symptoms of heart disease or worse have already occurred. This where machine learning algorithms are important. Using past research and data, machine learning algorithms can help in predicting and classifying patients with heart disease effectively to prevent illness or death. Prognostic models may help healthcare workers and doctors to select better treatment options, and diagnostic models can be used for screening and recommending the proper tests (Plati et. al, 2021). In this paper, the analysis and comparison of different machine learning algorithms for the prediction of heart disease will be performed and discussed. The goal is not only to predict heart disease, but also to offer analysis on which algorithm and technique would be most valuable to use. Additionally, a general guideline of which risk factors and behaviours are most important will also be detailed.

In the past few years, there have been multiple different studies that have evaluated the accuracy of classification algorithms in the prediction of heart disease. There have also been several studies that have compared and analyzed the effectiveness of these algorithms with each other. However, unlike the current paper, most of these studies used the Cleveland Heart Disease dataset, available from the UCI Irvine Machine Learning Repository. The Cleveland Heart Disease dataset has 303 instances, and 76 attributes, but most studies only used between 13-14 attributes. The target variable is represented from 0 – 4, with increasing severity per level, and 0 representing the absence of disease.

In 2021, Plati and colleagues published a research article a machine learning approaches for chronic heart failure diagnosis. The dataset was a merged data provided from the University College in Dublin, Ireland (410 subjects) and the Department of Cardiology from the University of Ioannina in Greece (77 subjects), totaling to 487 subjects (260 without heart disease, 180 with chronic heart disease, and 47 with acute heart disease) (Plati et. al, 2021). The study used Decision Trees, Random Forests, Rotation Forests, Naïve Bayes, K-Nearest Neighbours, Support Vector Machines, Logistic Model Trees and Bayes Network as the classifiers. 10-fold cross validation was applied for evaluation metrics and accuracy, sensitivity and specificity were recorded (Plati et. al, 2021). This study has several strengths. Firstly, the features and guidelines used are different from the general binary classification of other models, which distinguishes between heart disease or no-heart disease. This study instead provides more support to various healthcare workers in both diagnosis and prognosis of heart disease by using features that can be more closely followed by clinicians (Plati et. al, 2021). This is useful, as the model achieved a 84.12% accuracy when classification was used with only clinical features, which can prove to be of great value if only initial screening tests and questions are being assessed (Plati et. al, 2021). The overall model had an accuracy of 91.23%, specificity at 89.62% and sensitivity at 93.83% (Plati et. al, 2021). This study is unique, as it simulates a clinical approach in diagnosing heart disease using classification algorithms, as well as investigating the impact of different features on accuracy (Plati et. al, 2021).

However, it still suffers from a few weaknesses. Foremost, the dataset, despite being one of the larger ones in the literature, and larger than the previously mentioned Cleveland Heart Disease dataset, is still not a very large and diverse dataset. Therefore, this study does require more validation with a more comprehensive dataset, if this clinical type of approach system is to be used (Plati et. al, 2021). Additionally, the subjects of the datasets were mostly senior citizens, with the mean age being 69 years, and median age being 71 years (Plati et. al, 2021). Although aging does play a role in cardiovascular health, heart disease and failure can happen at any age and the dataset was limited by the diversity of the ages of the participants (Khan et. al, 2020).

TR et. al also conducted a study in 2022 outlining predictive analysis of heart diseases using machine learning approaches (TR et. al, 2022). This study was done using the Cleaveland Heart Disease dataset. Various machine learning algorithms were used, including Random Forests, Regressions, Decision Trees, K-Nearest Neighbors, Naïve Bayes, and Support Vector Machines (TR et. al, 2022). Moreover, k-fold validation was used as cross validation techniques and precision, F1 score, AUC, and recall were used to determine the efficiency of the various classification algorithms (TR et. al, 2022). Overall, the study found Random Forest classification to be the most effective at predicting heart disease and does offer a somewhat comprehensive view on the various models in the algorithm. However, it does suffer from the same problems as Plati and colleagues did, and that is the low number of diversity and size of the dataset.

In terms of comprehensive comparison of machine learning techniques pertaining to heart disease, Pouriyeh et. al wrote a paper to compare the accuracy of different classification techniques and Ensemble Machine Learning Techniques for the prediction of heart disease (Pouriyeh et. al, 2017). The techniques compared were Decision Trees, Naïve Bayes, MultiLayer Perceptron, K-Nearest Neighbors, Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF), and

Support Vector Machines (Pouriyeh et. al, 2017). The dataset used was the Cleaveland Heart Disease dataset, and K-fold cross validation was used on the dataset (Pouriyeh et. al, 2017). The main performance measures were precision, recall, F-score, and ROC (Pouriyeh et. al, 2017). This paper, as opposed to TR et. al, found that Support Vector Machines were the most effective at predicting heart disease (Pouriyeh et. al, 2017). This study ran two different experiments, with the first being applying all the techniques to the whole dataset as baseline. The second experiment applied bagging, boosting and stacking to the different machine learning models and evaluated the response there (Pouriyeh et. al, 2017). The main goal of the research paper was to simply compare the various machine learning techniques to measure effectiveness on a small dataset; the study was purely technical. Thus, it suffers from further answering any questions except the analysis of the machine learning algorithms, such as important clinical risk factors and behavioural risks, or which demographics may be more at risk and what would be the best use case for the specific algorithms in question. It did, however, use techniques that the aforementioned studies did not, such as MultiLayer Perceptron and Single Conjunctive Rule Learner.

In 2020, Yadav et. al also conducted a study outlining applications of machine learning for the detection of heart disease (Yadav et. al, 2020). This study also used the Cleveland Heart Disease dataset, however, unlike the previous studies, this study did not use any Support Vector Machines or Tree-Based Methods, instead opting to use K-Nearest Neighbors and Neural Networks to measure performance instead (Yadav et. al, 2020). Likewise, K-fold cross validation was also used, and performance metrics such as accuracy, recall and precision were measured (Yadav et. al, 2020). This study is one of two studies being analyzed, the other being Pouriyeh et. al, that used neural networks to measure detection of heart disease. However, this study did not compare Tree-based methods, Support Vector Machines or Naïve Bayes in its analysis, thus falling short of being a comprehensive overview in terms of analysis of various algorithms.

Finally, a study by Lakshmanarao and colleagues in 2019 went over various machine learning techniques for heart disease predictions (Lakshmanarao et. al, 2019). Unlike the other studies, this study used a larger dataset acquired from Kaggle, which had 4239 instances and 15 features (Lakshmanarao et. al, 2019). This dataset was unbalanced so researchers used three different oversampling methods to deal with this problem: Random Over-Sampling, Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling Approach (Lakshmanarao et. al, 2019). This study was useful, as it analyzed how the different sampling approaches affected model performance, with different oversampling techniques favoring different models (Lakshmanarao et. al, 2019).

The purpose of this research paper is to analyze the effectiveness of different machine learning models in the prediction and detection of heart disease, as well as determine which general features would be most important in a survey from a healthcare expert in the determining of heart disease. The aforementioned studies are also attempting to do the same, with some of them focusing on different aspects or different techniques overall. For example, Plati et. al is the only study to focus on both the clinical aspect (surveys) and the technical aspect (machine learning models) with the research. Most of the studies mentioned only focus on the technical aspects.

Additionally, most of the studies were using the Cleveland Heart Disease dataset, which is a small dataset, and despite other studies using slightly larger datasets, they still lacked diversity. Therefore, we believe that this paper would be beneficial and worth doing due to focusing on a multitude of different aspects, such as mental health, specific demographics, physical health, diet, and clinical risks. Moreover, this paper also has a very large dataset consisting of over 250000 instances and 18 features. Although previous research is closely related to this paper, it is believed that this research would still be beneficial and of great value in the diagnostic and screening of heart disease and may further help validate some of the previous findings with a larger dataset.

Data Description and Initial Analysis

The Behavioral Risk Factor Surveillance System (BRFSS) is a project conducted by the Centre for Disease Control and Prevention (CDC) in all states in the United States. This dataset is done via survey over the phone, and collects data on health risk behaviours, chronic diseases and conditions, access to healthcare and much more. The raw dataset was cleaned and only 18 attributes that would be relevant to heart disease prediction and classification were selected, such as specific demographics, health indicators, risk factors, and general health and wellness.

According to the CDC, important risk factors for heart disease are high blood pressure, high cholesterol, diabetes, smoking, obesity, unhealthy diet, and physical inactivity. From the 343 variables in the dataset, the health indicators in this dataset were chosen from this information. The attributes are split into different categories, such as demographics, clinical risk factors, and general health and wellness. The features of the dataset are the target variable, heart disease, age, sex, income, education, high cholesterol, high blood pressure, BMI (obesity), alcohol consumption, smoking, vegetables in diet, fruits in diet, diabetes, physical activity, stroke, general health, mental health and physical health.

The target variable in this analysis is 'HeartDisease' and is a yes or no attribute declaring if the individual has ever had coronary heart disease (CHD) or a myocardial infarction (MI) before. This variable has a heavy imbalance, almost 90% of the responses are no (Figure 1). Therefore, this problem will need to be handled before models can be created.

The correlation between the variables (except the target variable) was also explored in the initial analysis. Some things of note were age being correlated with high cholesterol and blood pressure (Figure 2). This makes sense, as these things are likely to be higher in those that are older, and lower in those that are younger. Additionally, general health, physical health, and mental health were all positively correlated to some degree (Figure 2). Diabetes was also correlated with BMI, blood pressure and cholesterol (Figure 2).

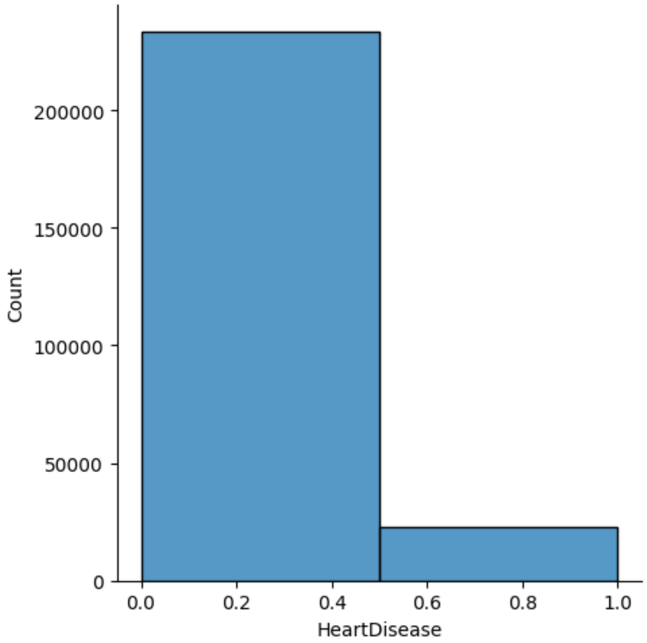


Figure 1: Heart Disease variable for the dataset. Note the imbalance of the classes.

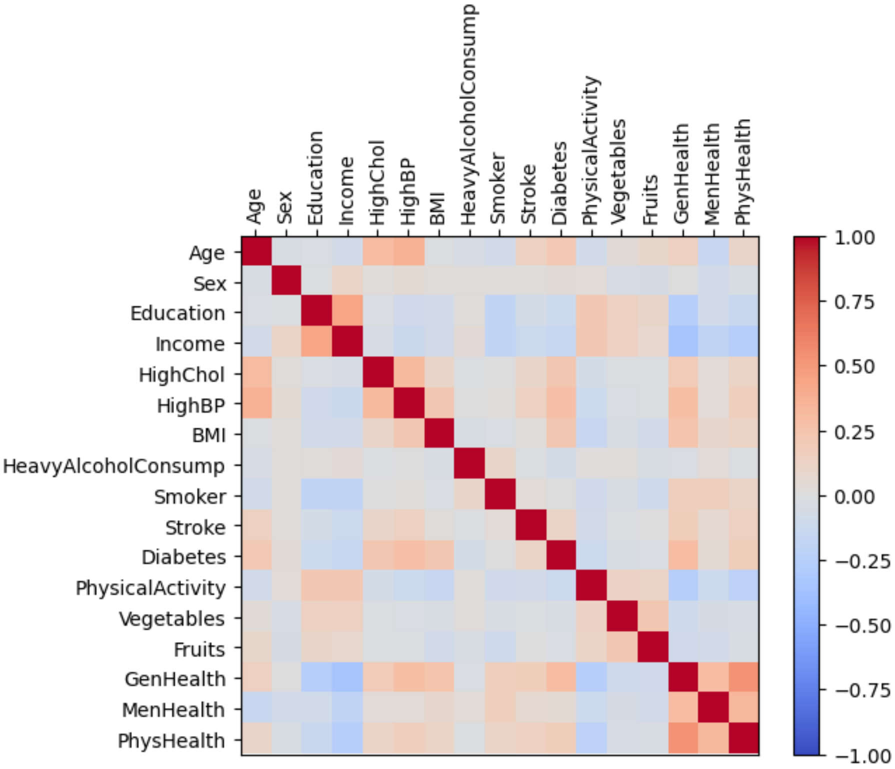


Figure 2: Correlation between different features, excluding the target variable.

Overall Methodology and Data Approach

The general goal of this project is to not only create a machine learning algorithm that is effective at predicting and classifying heart disease, but to compare the effectiveness of various machine learning algorithms with each other to determine which one would be the most accurate. Additionally, a secondary objective to see which features and survey questions are more likely to be important in the prediction of heart disease compared to others was also determined (e.g., what features to look out for in patients).

Firstly, in the planning phase, research will be conducted in choosing the most relevant features of the dataset. The dataset will be cleaned, missing values handled, and outliers detected. The dataset will be made as consistent as possible. Initial analysis will be performed to determine the distributions of the attributes, especially the target variable, and the correlation of the attributes will be analysed. A literature review will also be conducted at this stage to determine which techniques to use and how similar studies have handled certain problems (e.g., imbalanced data).

Secondly, a more in-depth exploratory data analysis will be conducted on the dataset, and key takeaways from the data will be outlined. The data analysis will focus on the relationship between the target variable and other variables, as well as the relationship between non-target variables. Afterwards, the data will be balanced using SMOTE or RandomUndersampling and will be split into a train-test split. Features will be chosen, and a 5-fold cross validated randomized search will be performed on all models to find the best parameter. Randomized search was chosen over grid search to save computation time and complexity. 10-fold cross validation will then be used on the best parameters specifically to determine the stability of the model and the model will then be used on the test set to determine performance and fit time.

Finally, the models will be trained and evaluated using accuracy, F1-score, precision and recall as the evaluation metrics. The techniques used will be Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Random Forests, and XGBoost. In this phase, features that are most important in determining the correct output will also be analyzed using the Random Forests and XGBoost models, so they may act as general guideline in survey questions.

Exploratory Data Analysis

The percentage of people suffering from heart disease is far less than those who are not, thus the dataset ends up with a very imbalanced target variable (Figure 1). Additionally, despite being the minority class in the dataset, the proportion of males suffering from heart attack is higher than females (Figure 3). This may indicate that men are more at risk of heart disease than women (Figure 3).

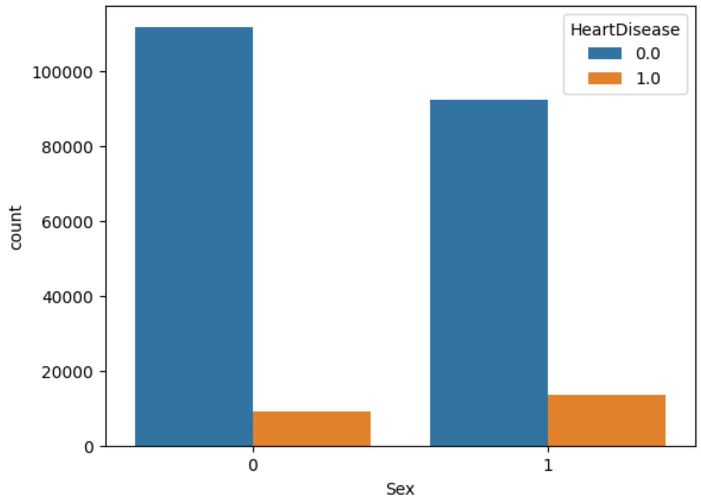


Figure 3: Relationship between sex and heart disease. Males are 1, females are 0. There are less incidents of CVD for females despite majority.

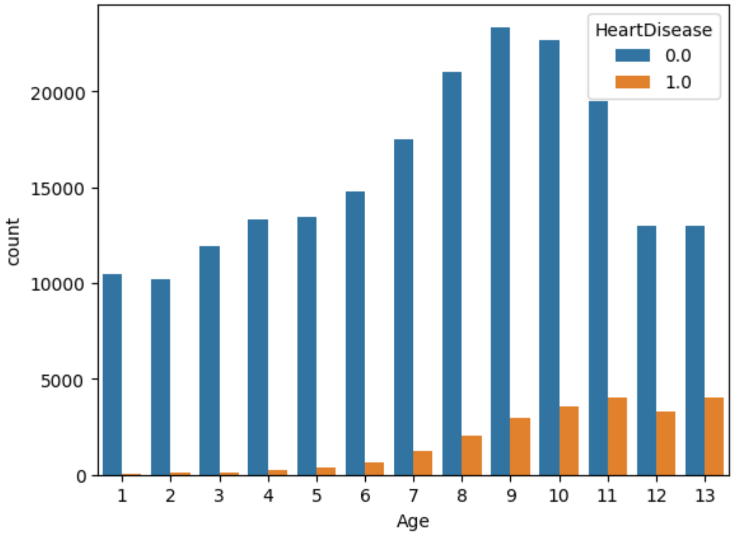


Figure 4: Age is separated into 13 categories in ascending order. Age is highly correlated with heart disease.

Furthermore, the analysis found that as age increases, so does the risk of heart disease, mostly in a linear fashion (Figure 4). High cholesterol and high blood pressure also seem to play a role in higher chance of heart disease (Figure 5, 6). They are both correlated highly with age as well, suggesting that as age increases so does the risk of developing high cholesterol and high blood pressure, which again may lead to increased risk CVD (Figure 2). Not a very significant link with BMI and heart disease found.

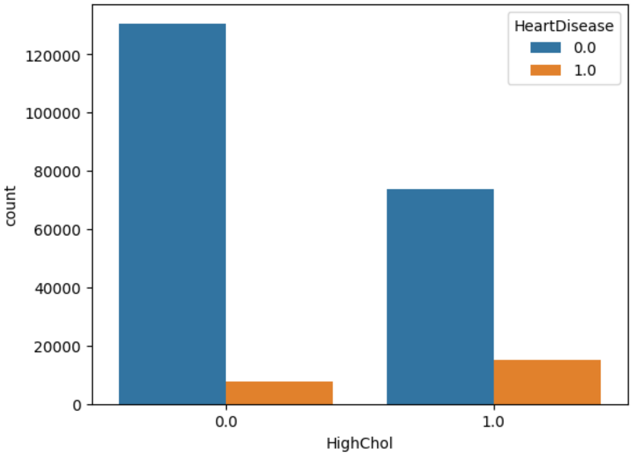


Figure 5: Frequency of high cholesterol in the target variable class.

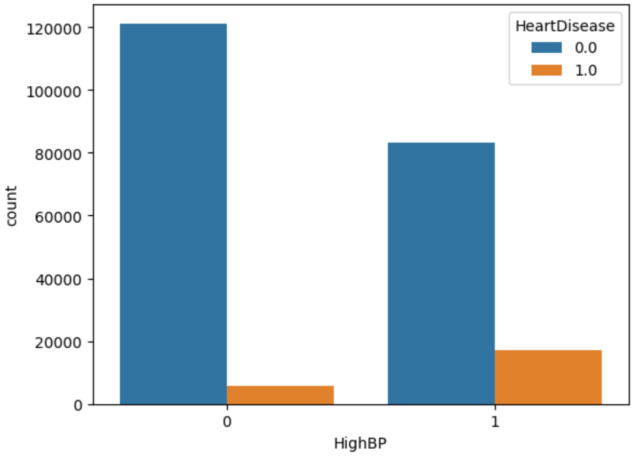


Figure 6: Frequency of high blood pressure in the target variable.

Interestingly, despite being a small percentage of the entire dataset, people with diabetes made up a significant portion of all positive cases (Figure 7). Stroke and heart disease seem to go hand in hand, as previous history of having a stroke does seem to increase likelihood of having CVD or MI. Moreover, income does not seem to play a huge factor in heart disease and conversely, has a negative relationship with heart disease (Figure 2). This may be due to the fact that most of the individuals within the dataset were in the highest bracket of income listed. Most of the individuals in the dataset made $75-80k/year (Figure 8). The increase from class to class in the income variable was mostly consistent, as was the risk of heart disease.

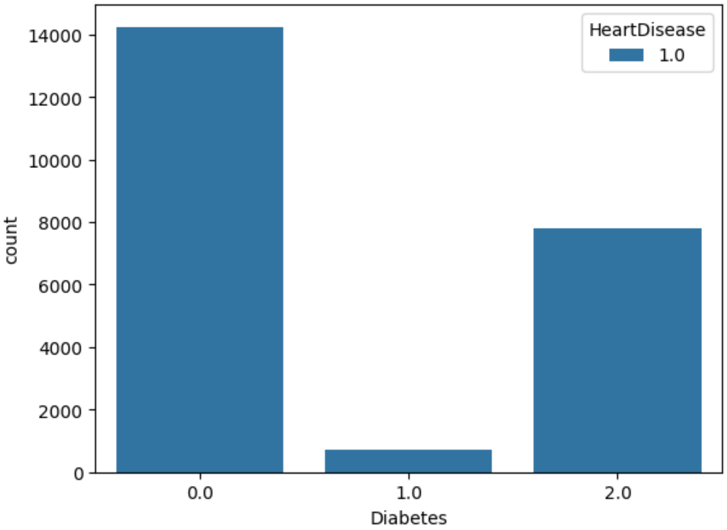


Figure 7: The distribution of the diabetes variable classes for those who have heart disease. 0, 1, 2 are for no diabetes, pre-diabetes, and diabetes respectively. As shown, despite diabetes patients being a low percentage of individuals in the dataset, over half of heart disease cases have diabetes.

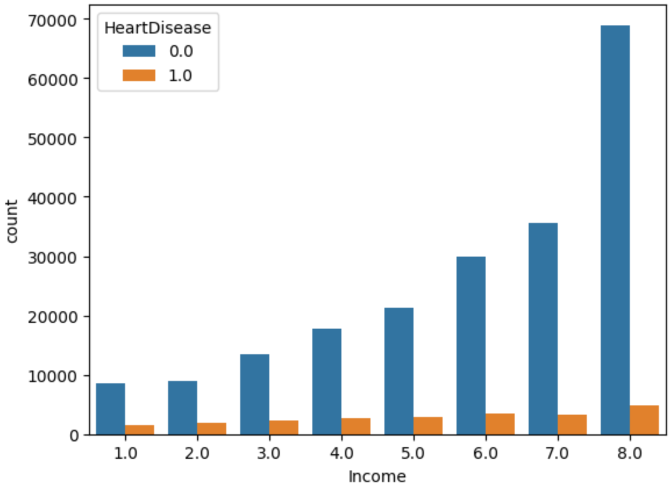


Figure 8: Distribution of income in relation to heart disease. Higher number is equal to higher income.

Mental health may play a small factor in heart disease (most likely due to increased stress). Individuals who described having poor mental health 30 days of the month did illicit more heart disease cases generally. However, overall, it seems to have very little to do with increasing the risk of heart disease. Individuals who described their general health to be poorer on average, generally had a higher proportion of positive cases of heart disease; ergo, worse general health may contribute to higher risk of heart disease.

There does not seem to be a significant correlation between eating fruits and vegetables daily and heart disease in this dataset. Additionally, neither heavy alcohol consumption nor smoking seem to have a significant correlation with heart disease in the current dataset. General Health and Physical Health seem to be correlated, suggesting that good physical health may contribute to a good general health. Moreover, good general health seems to decrease the risk of heart disease. It can be inferred from the data that good physical health, which subsequently leads to good general health, is quite important in reducing the risk CVD or MI.

Imbalanced Classes and Feature Engineering

The target variable, ‘HeartDisease’, was imbalanced with a 9:1 ratio as illustrated by Figure 1. To solve this problem, Synthetic Minority Oversampling Technique (SMOTE) and

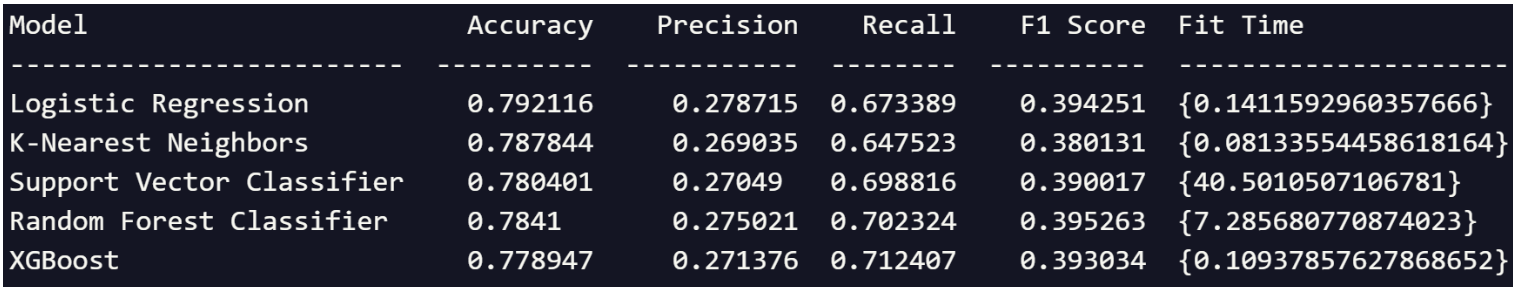
RandomUndersampling were used to balance the classes to an acceptable ratio. This was performed on the training data only, and the test set data was left imbalanced to mimic real world scenarios. After testing and consideration, only RandomUndersampling was used to under sample the majority class, instead of SMOTE to oversample, or SMOTE and RandomUndersampling together. After RandomUndersampling the target variable had a 2:3 split which was adequate. This left the training dataset with 49,756 samples and the test set with 22705 samples.

In terms of feature selection, the following features were removed from the final models due to either being redundant because they are highly correlated with another feature, or because they did not have a high correlation with the target variable. The features removed were education, physical health, fruits, vegetables, mental health, heavy alcohol consumption, smoking, BMI, and physical activity. The final data had 8 features (not including the target variable), which consisted of age, sex, high blood pressure, high cholesterol, stroke, diabetes, income, and general health.

Results

Each of the machine learning algorithms was evaluated for its accuracy, precision, recall, F1 score and time to train the model (fit time). Logistic regression had the highest accuracy at 0.792, as well as an F1- score of 0.394 (Table 1). The fit time for logistic regression was also quick at 0.141s. K-Nearest neighbors had an accuracy of 0.787, and the poorest F1-score from all the models (Table 1). It also had the lowest precision and recall; however it was the quickest model with a fit time of 0.081s (Table 1). Support Vector Classifier had an accuracy of 0.78, and an F1-score of 0.39. However, it was much more computationally expensive than all the other models and had a fit time far longer than the rest at 40.5 seconds (Table 1). Random Forest classifier had an accuracy of 0.784 and the highest F1-score at 0.395. However, it also had a relatively long fit time at 7.26s (Table 1). Finally, XGBoost had an accuracy of 0.779, the lowest of all the models, and the F1-score was 0.393. However, it had the highest recall of all the models, as well as being the second quickest algorithm with a fit time of 0.109s (Table 1).

Table 1: Demonstrates the evaluation metrics and fit times for all five techniques. Recall and accuracy are more important than precision in this case. Fit time is also important to save computational resources.



```
Model                        Accuracy    Precision    Recall    F1 Score  Fit Time
-------------------------  ----------  -----------  --------  ----------  ---------------------
Logistic Regression          0.792116     0.278715  0.673389    0.394251  {0.1411592960357666}
K-Nearest Neighbors          0.787844     0.269035  0.647523    0.380131  {0.08133554458618164}
Support Vector Classifier    0.780401     0.27049   0.698816    0.390017  {40.5010507106781}
Random Forest Classifier     0.7841       0.275021  0.702324    0.395263  {7.285680770874023}
XGBoost                      0.778947     0.271376  0.712407    0.393034  {0.10937857627868652}
```

The importance of features for random forest classifier were (in descending order): age, general health, high blood pressure, high cholesterol, stroke, sex, diabetes, and income (Figure 9). Conversely, with XGBoost, the importance of features (in descending order) was high blood pressure, age, high cholesterol, general health, stroke, sex, diabetes, and income (Figure 10). High blood pressure was by far the most important feature in the dataset, having a value of 0.386 in feature importance, compared to 0.153 for age, which was the second most important feature (Figure 10). Income was almost a non-factor in XGBoost classifier and was the lowest rated of feature for both models.

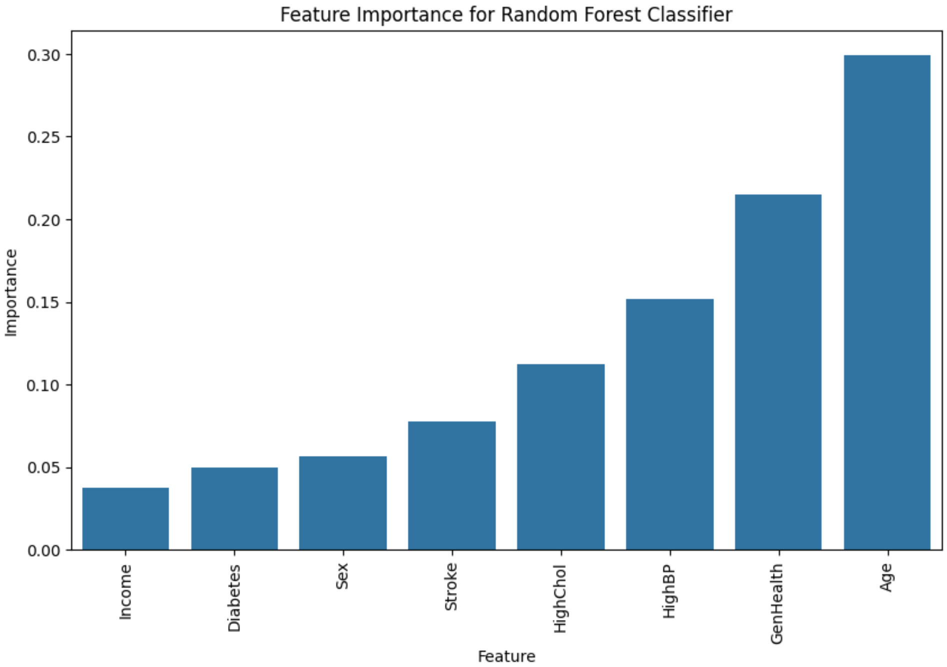


Figure 9: Feature importance for random forest classifier. Age being the most important feature and income being the least important.

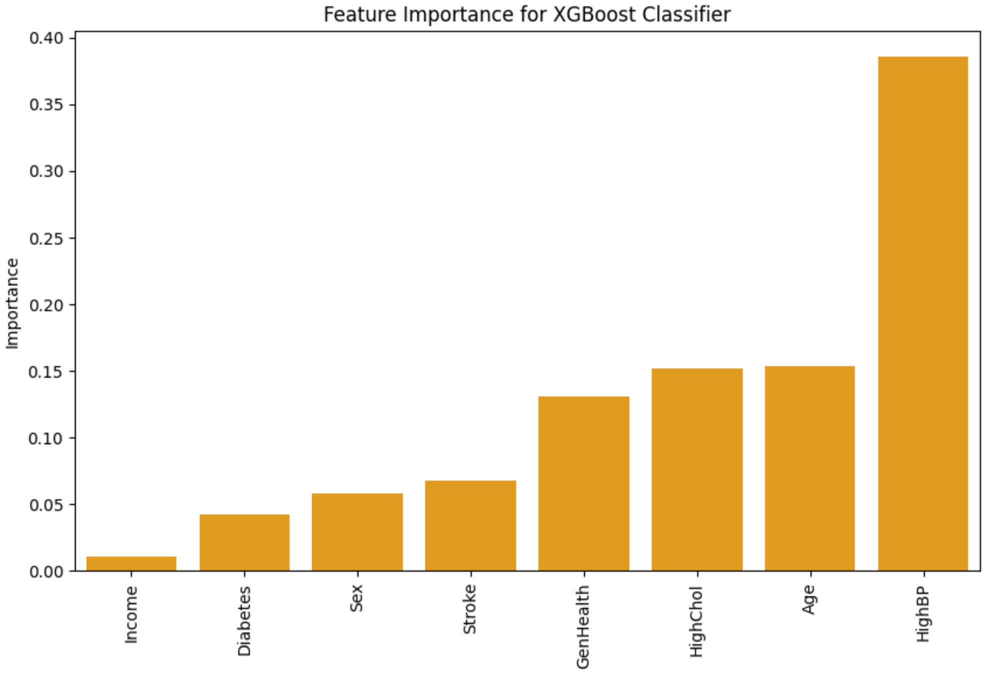


Figure 10: Feature importance for XGBoost. Note the value of high blood pressure, it is by a wide margin the most important feature. Income is the least important by a wide margin as well.

Discussion

Almost all the models were very similar in performance metrics, excluding fit times which differed between the various algorithms. The highest accuracy was logistic regression, with the lowest being XGBoost. Both logistic regression and XGBoost however had almost identical F1-scores, with logistic regression being slightly better. However, the recall for logistic regression was lower than XGBoost, and since we are dealing with healthcare data where the recall is an important metric, the XGBoost algorithm is preferred compared to logistic regression as catching false negatives is vital. The fit times were also different, with XGBoost performing quicker than logistic regression.

In terms of fit time, SVM was the worst performing model with extremely long fit times. This is most likely due to SVM being computationally complex and the dataset being large with multiple features, even after under sampling. The quickest model was K-Nearest neighbors, but it also had the lowest F1- score of all the models as well as the lowest precision and recall It was also the only F1-score less than 0.39. Random forest classifier was the second longest fit time, but not nearly as long as SVM. It also happened to have the best F1-score out of all the models, and average accuracy at 0.784.

The accuracy of the models is on par with most of the previous studies conducted on this type of problem, around the high 70s. However, the F1-score is quite low, most likely due to the test-set being imbalanced and the algorithms being overfitted. This problem may be solved with more data but that would also be quite demanding computationally. Since the data and problem in question is sensitive health data, recall is prioritized over precision, however, ideally both being in an acceptable range would be the best-case scenario. In our case, the recall was somewhat acceptable for all the models, ranging from 0.65 to 0.71. However, precision was severely lacking with all models being in the low 20s.

In terms of feature importance, both random forests and XGBoost were quite similar, with the top four and bottom four features being the same across both models. High blood pressure, general health, age, and high cholesterol seem to be the most important factors in classification of heart disease. Diabetes, Stroke and Sex also play a role to some degree but not as much as the former features. Income seems to have least affect on determining and classifying heart disease and this was true for both random forest classifier importance and XGBoost importance.

In a clinical setting, with some form of health expert or researcher surveying patients or individuals, the most important features to focus on the in the dataset would be age, followed closely by high blood pressure and high cholesterol. General health should also have a significant impact in screening, as it affects multiple areas of life such as physical activity and diet. Sex, previous history of stroke and diabetes should also be important factors to consider.

Overall, considering the accuracy, the F1-score, the trade off between precision/recall (recall is slightly more important for this specific dataset), and the fit times, the preferred model would be XGBoost. It was the second quickest model to fit at 0.11s and had a high F1-score at 0.393. The recall was also the highest of all the models at 0.71. The accuracy, despite being the lowest was still good at 77.9%. However, it should be noted that if model fit times are not a problem and the infrastructure to run complex algorithms is available, then the preferred model would have to be random forest classifier, which had the overall best metrics with high recall, f1-score, and accuracy.

Contributions, Limitations and Project Continuity

The objective of this research paper was to analyze the effectiveness of different machine learning models in the prediction and detection of heart disease, as well as determine which general features would be most important in the determining of heart disease, so healthcare experts may weight those features more heavily that others. This would give insight into the technical and clinical aspects of using machine learning tools to prevent heart disease or classify those at high risk for it. In terms of this paper, most of the studies conducted in the past were using the Cleveland Heart Disease dataset, which is a small dataset, and despite other studies using slightly larger datasets, they still lacked diversity. This dataset also has the benefit of being focused on past behaviours and chronic illnesses, something that other datasets lack. Due to these reasons, it is believed that this research paper is worthwhile in contributing to the overall knowledge in this area of study.

However, there are still some limitations of the study. For example, a strength of the dataset is the fact that it looks at the history of an individual and their past chronic illnesses to determine heart disease. This, however, is also a weakness, as the dataset does not consider other more recent factors such as ECG's, recent bouts of arrhythmia or other conditions of the nature. Another limitation of the study was computational complexity, which lead to algorithms not being utilized to the fullest. In this case, randomized search was used instead of a grid search, despite the grid search being more comprehensive in determining the best parameters for a model. Finally, as comprehensive as this analysis was, it was only performed using machine learning techniques and it still does not cover any deep learning techniques or artificial neural networks, such as some of the studies that were performed in the past. Therefore, it is unknown how well those powerful algorithms would perform for this data. Lastly, in terms of sampling the data, only SMOTE and RandomUndersampling were used, and whether other sampling techniques may have performed better is unknown, which is also a limitation of the study.

Looking forward to the future, this research could further be developed by using more resources to perform a better parameter search. Additionally, deep learning techniques can be employed to discover any other possible solutions to the problem. The study may also benefit from more data being utilized for the training and testing. Finally, some other sampling techniques such as AdaSyn may also be used to see if the algorithms perform better.

Conclusion

In the technical aspect, the preferred model for this specific dataset and problem would have to XGBoost, due to its high recall, acceptable accuracy, and low fit time. However, if the fit time is not a large problem and the resources are available, then random forest classifier would be preferred as it does perform slightly better than XGBoost overall, suffering only from being computationally expensive. In the clinical aspect of this study, the most important features and concerns that healthcare experts should be looking out for in patients regarding heart disease are: age of the patient (the older the more at risk), whether they have high blood pressure or high cholesterol, and whether their general health is good or not. Sex and previous history of stroke and diabetes may also be important factors to consider.

References

Hopkins J, Agarwal G, Dolovich L. Quality indicators for the prevention of cardiovascular disease in primary care. Can Fam Physician. 2010 Jul;56(7)

Khan MA, Hashim MJ, Mustafa H, Baniyas MY, Al Suwaidi SKBM, AlKatheeri R, Alblooshi FMK, Almatrooshi MEAH, Alzaabi MEH, Al Darmaki RS, Lootah SNAH. Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. Cureus. 2020 Jul 23;12(7).

Lakshmanarao, A & Swathi, Y & Pullela, Sundareswar. (2020). Machine Learning Techniques For Heart Disease Prediction. International Journal of Scientific & Technology Research. 8. 374.

Plati DK, Tripoliti EE, Bechlioulis A, Rammos A, Dimou I, Lakkas L, Watson C, McDonald K, Ledwidge M, Pharithi R, et al. A Machine Learning Approach for Chronic Heart Failure Diagnosis. *Diagnostics*. 2021; 11(10):1863.

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. 2017 IEEE Symposium on Computers and Communications (ISCC).

TR, R. ., Lilhore, U. K., M, P. ., Simaiya, S. ., Kaur, . A. ., & Hamdi, M. . (2022). PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES. Malaysian Journal of Computer Science, 132–148.

Yadav, S. S., Jadhav, S. M., Nagrale, S., & Patil, N. (2020). Application of Machine Learning for the Detection of Heart Disease. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).