Can a machine save a heartache? Predicting heart disease with Machine Learning models.

Executive Summary

Purpose: This report summarises the process of developing machine learning models in predicting heart disease, utilizing existing heart disease databases from UCI Machine Learning Repository[1][2] with the aim of predicting heart disease in new and unseen data.

Background: An exploratory data analysis (EDA) was performed on all datasets which consist of patient anthropometric data and cardiac health-related data. The target variable is the presence of heart disease.

As part of data preprocessing, we applied Random Forest Imputation method and set a binary classification target. We trained three models: Logistic Regression Classification (LC), Random Forest Ensemble Classification (RFE), and XGBoost Classification (XGB). Performance metrics were analysed while hyperparameters tuning informed required changes for the improvement of models.

Key Findings: Our models are comparable with the existing literature that had trained these model types on identical datasets. We compared the performance metrics and concluded that the Random Forest Ensemble model[3] is the most generalisable with 86% in accuracy and precision of 91% when tested against a combined dataset.

Future Directions and Recommendations*:* It would be interesting to explore the minimum features required to predict while maintaining performance. Selecting key features would not only reduce the data dimensionality but potentially improve performance. Investigating the impact of different imputation methods on the performance of these models may be valuable.

Conclusion: Despite being comparable to current literature, these models were trained on datasets from the 1980s'. Heart disease predictive models remain clinically relevant to serve as cardiac risk predictors, even with the advancement in cardiac diagnostics and risk assessments. However, we posit our models require further training and tuning with unseen datasets to remain contemporary and dependable.

[1] UCI Machine Learning Repository visited on the 27th of September 2024

[2] Heart Disease - UCI Machine Learning Repository visited on the 27th of September 2024.

[3] Trained on DS1 Statlog dataset.

Contents

Introduction

Cardiovascular disease is still one of the leading causes of death. According to the World Health Organization (WHO), several factors contribute to increased heart disease risks such as raised blood pressure, blood glucose, lipids, and other factors such as obesity[4]. There has been literature on utilising machine learning (ML) models to create predictive modelling to help inform clinicians and patients to manage cardiovascular health (Detrano et al 1989). A brief search on Google Scholar indicated that this area has been well explored and researched by academia[5].

Why Predict?

The gold standard for detecting heart disease is coronary angiography, which uses contrast dye to visualise the cardiac vessels. However, this is an invasive procedure. Although various non-invasive techniques (such as cardiac Magnetic Resonance Imaging) are now available, it is costly both to the individual and the healthcare system.

Predicting heart disease risks through ML models could be useful for providing insightful triage, stratifying risks and allocating the often-limited healthcare resources appropriately. Additionally, identifying elevated risk groups could also advocate for early preventative measures, resulting in better long-term health outcomes for the individual and reducing the healthcare burden. Therefore, it is pertinent that health predictive models should be robust and dependable for the given population groups.

With that in mind, this report presents the process of exploring the heart disease-related datasets to build ML models to predict the presence of heart disease, followed by its key findings, learning points, and recommendations.

[4] Cardiovascular diseases (who.int)

[5] A search on Google Scholar with the key phrase ‘heart disease prediction using machine learning’ returned 888 thousand results.

Exploratory Data Analysis (EDA)

The datasets originated from the open-sourced UCI Machine Learning Repository. While the datasets consist of two parts (DS1 and DS2), they contain patient anthropometric data such as sex and age and cardiovascular metrics. The DS2 datasets are location specific: Cleveland (USA), Hungary, Switzerland, and Long Beach (California, USA) while DS1 is a complete subset of DS2 Cleveland.

All datasets have thirteen prognostic attributes followed by one target variable (see appendix for list of attributes). DS1 Cleveland and DS2 Hungarian have binary targets (i.e. presence/absence of heart disease), while the remaining have a range from 0 to 4 to classify heart disease severity (see Table 1).

Table 1 Summary of Datasets

DS1	- Subset of DS2 Cleveland - 270 entries. No missing values. - Binary target classification (0=absence, 1= present)
DS2	
Cleveland	- 303 entries, minor missing values - Multiclass target (0=absence, 1-4=present + severity)
Hungarian	- 294 entries, significant missing values in some features. - Binary target classification (0=absence, 1= present)
Switzerland	- 123 entries, significant missing values in some features. - Multiclass target (0=absence, 1-4=present + severity)
Long Beach	- 200 entries, significant missing values in some features. - Multiclass target (0=absence, 1-4=present + severity)

DS1 EDA

The continuous attributes, such as age, blood pressure and cholesterol level are normally distributed (see Figure 1 for example). The QQ plot at Figure 2 showed that max heart rate at exercise mostly follow the theoretical normal distribution, but the observed showed a slight deviation at the higher rates. From a PCA plot, while we were able to visualize a degree of separation, we also found features that contributed the most variance (see table 2).

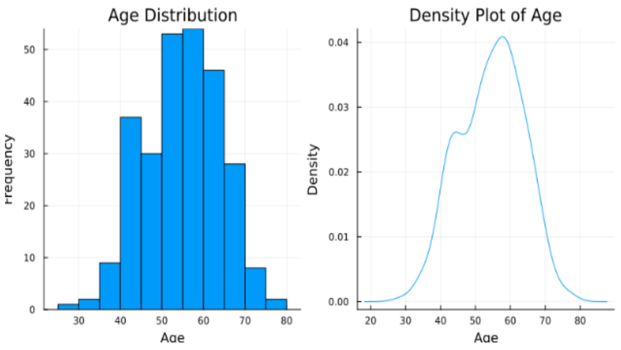

Figure 1 DS1 Age Distribution

Table 2 First Two Principal Components and its attributes

PC1	Old Peak> Max heart rate> Slope
PC2	Sex> Cholesterol> Age

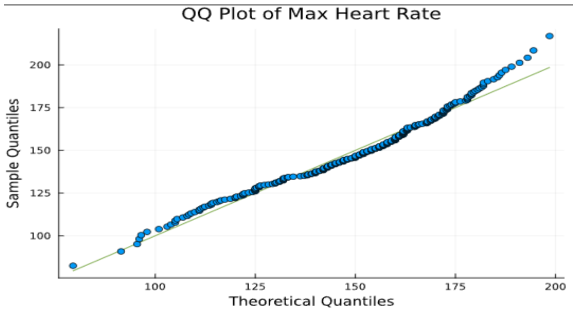

Figure 2 Max Hr DS1

Correlations

Correlation Matrix

In DS1, attributes such as sex, age, thal, angina history, old peak, slope, and vessels have positive correlation with the target variable. In other words, these attributes positively contribute to the presence of heart disease. However, maximum heart rate is negatively correlated (see Figure 3).

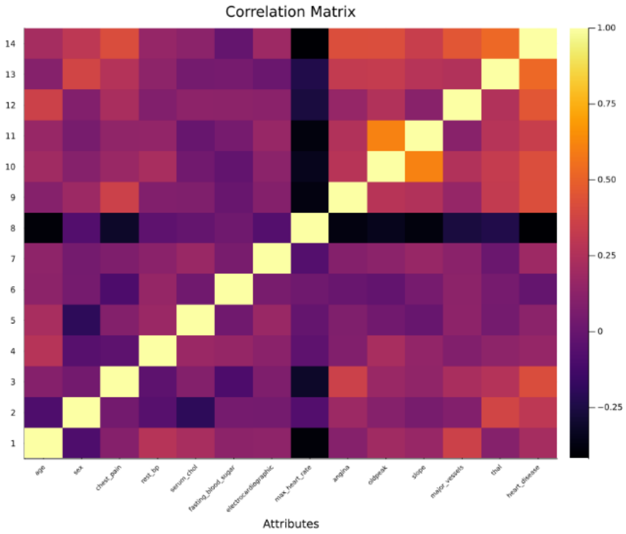

Figure 1 Correlation Matrix of DS1 heatmap

Biserial Corelation[6]

This showed that continuous attributes such as old peak, blood pressure, and age correlates positively with heart disease while maximum heart rate is the opposite (see figure 4). This corroborates with the correlation matrix.

[6] Biserial correlation is used to measure correlation between a binary attribute with continuous attributes.

Chi-square Test

Interestingly, the Chi-square test indicated that all categorical features except blood sugar are contributing risk factors (see Figure 5). This supports the notion that these categories have association with the target variable.

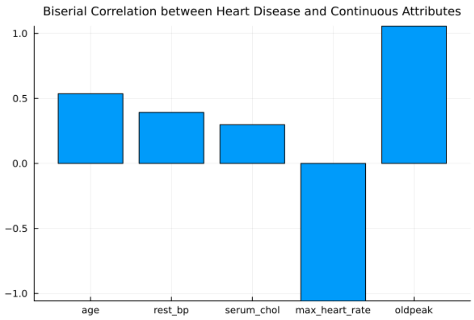

Figure 4 DS1 Biserial Correlation

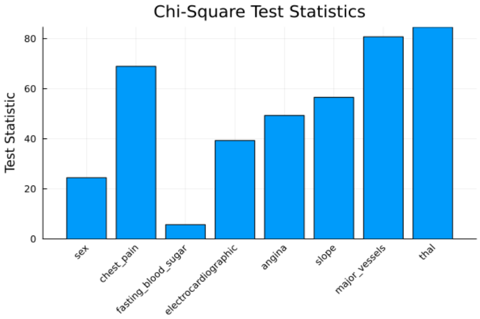

Figure 5 DS1 Chi-Square Test

Comparisons between DS1 and DS2 Datasets[7]

Overall, most of the continuous attributes are normally distributed in all datasets, with a few exceptions. Some anthropometric data were significantly different (see Figure 5). For cardiac metrics, one example would be the maximum heart rate between DS1 and Hungarian, where they are statistically different (see Figure 6). Other examples that are different are ECG readings between DS1 and DS2 Long Beach (see Figure 7). The significant differences between datasets indicated that the population groups were different and should be taken into consideration for model training.

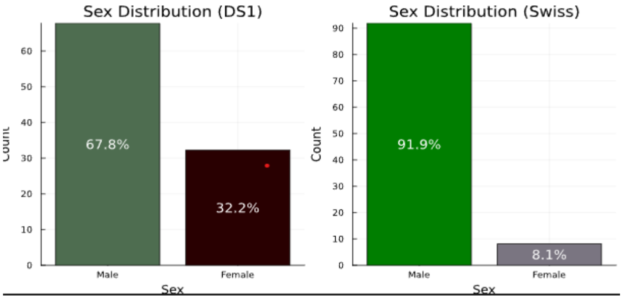

Figure 5 DS1 Versus DS2 Switzerland

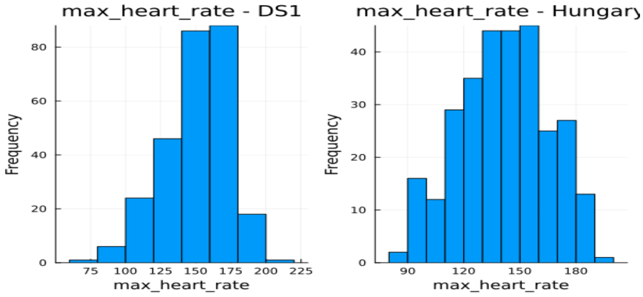

Figure 6 Max HR DS1 Vs Hungary

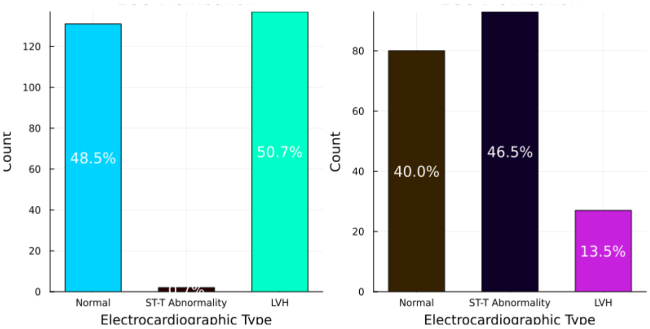

Figure 7 ECG DS1 Vs Long Beach

[7] Other comparisons are on the notebooks.

Data Imputation

We used a Random Forest (RF) approach because it can manage categorical and continuous data while able to capture non-linear interactions between variables despite being aware that RF can be harder to interpret and has higher computational cost.

Based our assessment, there were significant missing values in DS2 datasets such as the DS2 Hungarian as shown (see figure 8)[8]. Using the same example, we found that after imputation, the 'thal' attribute values were significantly different from its original. This also corroborates with the high % missing values.

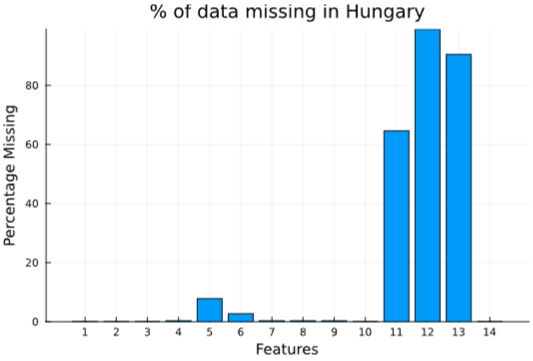

Figure 8 Percentage of Missing Values in Hungary

Although it could be 'Missing at Random' (MAR)[9], however we would argue that apart from the Cleveland datasets, the missing values from Hungary, Switzerland and Long Beach may be due to collection protocol differences.

It is also worth noting that we reclassify datasets with multi-class target classification into a binary classification to simplify our model selection (i.e. class 0= 0, class 1-4=1).

[8] For in-depth details, refer to notebooks on EDA.

[9] MAR means that the probability of data being missing depends only on observed data not the missing data itself.

Model Selection

Given the dataset characteristics, we investigated a binary classification predictor using three models: **Logistic Regression, Random Forest Ensemble Classification and XGBoost Classification**. We chose these for a variety of reasons while understand the drawbacks of each model (see Figure 9).

From each model type, we have trained a subset of five models (3 model types x 5 datasets). In total, we have trained fifteen models and applied their hyperparameters tunings to optimise their performance. We trained each model on partitioned of training and testing sets.

Logistic Regression

- Easy to implement, suitable for binary classification.
- Assumes linear distribution of data.
- May struggle with complex and large datasets.

Random Forest Ensemble

- Excels at handling large, non-linear datasets.
- Less likely to overfit but also could be slower due to multiple trees approach.

XGBoost

- Uses gradient boosting to improve iteratively.
- Can be more complicated to tune and may require more computation power.

Figure 9 Rationale of Model Selection

Performance Metrics and Hyperparameter Tuning

We applied K-fold cross-validation test for all model hyperparameter tuning and measured the log loss (cross-entropy) as a performance metric for all our models because it penalises the training model when it is confidently incorrect (i.e., false positive and false negative). It also provides a probabilistic output which indicates how 'confident' with the classification outcome. Additionally, it manages imbalance datasets especially when there are non-linear relationships. To compare our metrics across different models, we calculated accuracy, precision, AUC-ROC, log loss, and confusion matrix as part of our assessment.

Findings and Insights

Limitations:

Before we describe our findings, we must acknowledge limitations of this experiment. Firstly, we did not standardise the datasets given the mixture of categorical and continuous attributes in the datasets. However, reflecting on this decision, we are aware with methods such as one-hot encoding could be useful.

With the awareness of the attributes with high missing values percentage, we chose to train models with all thirteen attributes. The results may differ with explicit feature selection. While we did not investigate, it would be interesting to assess on un-imputed data, particularly with XGB as it can tolerate missing data and more robust to handle against real-world data.

Although binary classification can differentiate between those with or without heart disease, it does not indicate severity, which could further inform priority of care in real-world clinical settings.

Highlights:

From the EDA, we noted maximum heart rate has negative correlation against presence of heart disease. This means that the higher achievable heart rate during exercise, the better the cardiac health of the person. Conversely, if the person has more narrowed vessels and history of angina during exercise, they are positively correlated with heart disease. Further exploration of these attributes could lead to more intentional feature selection.

Like Paladino et al 2023, we evaluated our models with a combined dataset approach (DS1 and DS2). Our best performing model is the **Random Forest Ensemble model**, trained on DS1 Cleveland dataset with accuracy of 85.97% and precision of 91.27% (see Figure 10). It outperformed the baseline[10] results from the UCI website. We believe our results were comparable with the same study by Paladino et al who explored the use of automated machine learning (AutoML). The authors reported accuracies from 54% to 83% when assessed against the Cleveland, Hungarian and a

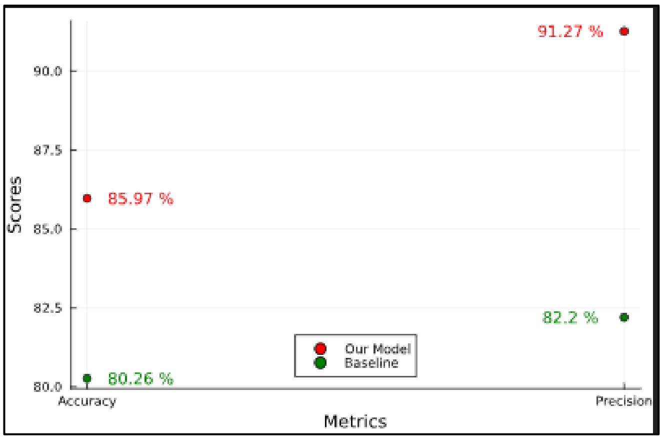

Figure 10 Best Performing: RFE DS2 Cleveland Model

[10] The UCI baseline RFE performance measures accuracy and precision

combined of the two datasets. Our RFE DS1 model however has a range accuracy from 69% to 97.41%.

As for the Logistic Regression model, and not surprisingly, the DS2 Cleveland trained model performed the best on the combined dataset. XGBoost model performed similarly as the RFE although less precise (87.52%).

On one hand, we did not consider certain models to be generalisable such as the DS2 Switzerland models, regardless of model type. They were only about 52% accurate with less than 1.25% precision. On the other hand, reflecting on the EDA on the Switzerland dataset, we noted that the dataset only has very few negative target values of '0' and we reclassified 1-4 as the new positive ('1') value. Despite training with k-folds and shuffle, it is likely that the models trained insufficiently and therefore not applicable to other unseen datasets. When applying our best RFE model to evaluate the DS2 Switzerland dataset, we found that the results were less than satisfactory.

We found training RFE models were computationally extensive when compared to XGB and LC models. It is due to its characteristics of multiple decision trees at random while LC does not have iterative tree-building, which makes it significantly easy to train. XGB on the other hand builds tree sequentially while optimizing for speed and performance.

Learnings and Recommendations

Team Learnings

From this experiment, we gained significant insights and experience in developing ML models. In EDA and data imputation, it was quite iterative as we learn about the datasets especially different methods of imputations. To achieve the tasks, we had new learnings, such as dashboard creation and developing the XGBoost model. We also understood the importance of consistency in areas such as performance metrics.

There were also reflective learnings along the process particularly with the decisions made. These are the key lessons for us to reconsider for the future development:

Standardisation of data and protocols

We learnt quite early that as a team we need to establish our goals and directions for the task at hand. However, we had to establish consistency standards of our protocols to avoid repetition of work. Reflectively, we would apply early standardisation of our data and execution of tasks.

Intentional feature engineering and selection

Imputation was an important part of the data preprocessing. It was challenging to establish what was essential to developing ML models. However, removing features such as those with high missing values may be beneficial.

Literature

Recent literature suggests that these datasets are still been actively used by researchers in machine learning and data mining (Rani et al 2024, Tougui et al 2020, Paladino et al 2023). Particularly in the review paper by Rani et 2024, it described extensive publications of various methods of machine learning models on these datasets. However, given that the dataset was from the 1980s, we opined that this may not be generalisable to today's population. Interestingly, there are existing tools in the public domain to estimate risk of heart disease[11][12].

In an ensemble of only two members, it was important there were delegation of tasks based on strengths and weaknesses. Using the Agile development approach, we believed that regular communication, use of a sharing platform[13] and willing to learn and adapt were the keys to success. We concluded this experiment with the recommendation of further refinement of our models.

Summary

We explored the characteristics of the heart disease datasets and processed them into useful sets. We developed three potential classification ML models from various datasets and tuned their hyperparameters to achieve the optimum performance metrics. We are pleased to report that the Random Forest Ensemble model performed the best in this controlled environment. While our current work and findings are not yet exhaustive, there is much to investigate particularly testing with other unseen incomplete datasets. We believe a contemporary predictive model could be achieved if our models are retrained with more recent datasets followed by finetuning of hyperparameters.

As a final note, it also worth noting that predictive modelling in healthcare carries its own ethical dilemmas and risks such as potential misdiagnosis and trustworthiness of its outputs. For any predictive ML models to be applied in clinical settings, it is quintessential that academic scrutiny and rigorous benchmarking are applied as part of developing such an instrument.

[11] My Heart Check. Heart Foundation

[12] CVD Risk Assessment (nzssd.org.nz)

[13] GitHub was used as our platform.

References:

Statlog (Heart) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57303.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Detrano, R.C., Jánosi, A., Steinbrunn, W., Pfisterer, M.E., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology, 64 5*, 304-10.

Rani, P., Kumar, R., Jain, A. *et al*. An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction. *Arch Computat Methods Eng* **31**, 3331–3349 (2024). https://doi.org/10.1007/s11831-024-10075-w

Paladino LM, Hughes A, Perera A, Topsakal O, Akinci TC. Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction. *AI*. 2023; 4(4):1036-1058. https://doi.org/10.3390/ai4040053

Tougui, I., Jilbab, A. & El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. *Health Technol.* **10**, 1137–1144 (2020). https://doi.org/10.1007/s12553-020-00438-1

Appendix

Dataset Attributes:

Variable Name	Role	Type	Description
age	Feature	Continuous	Age
sex	Feature	Binary	Sex
chest-pain	Feature	Categorical	Chest pain type
rest-bp	Feature	Continuous	Resting blood pressure
serum-chol	Feature	Continuous	Serum cholesterol (mg/dl)
fasting-blood-sugar	Feature	Binary	Fasting blood sugar > 120 mg/dl
electrocardiographic	Feature	Categorical	Resting electrocardiographic results
max-heart-rate	Feature	Continuous	Maximum heart rate achieved
angina	Feature	Binary	Exercise induced angina
oldpeak	Feature	Continuous	ST depression induced by exercise relative to rest
slope	Feature	Integer	The slope of the peak exercise ST segment
major-vessels	Feature	Continuous	Number of major vessels (0-3) coloured by fluoroscopy
thal	Feature	Categorical	Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
heart-disease	Target	Integer	Diagnosis of heart disease

Contributions:

Exploratory Data Analysis	
- General Statistics and PCA	- John
- Correlations/Chi-Square	- Leonard
Data Cleaning and Preprocessing	- John
Data Imputation	- Leonard and John
Model Building/Hyperparameter Tuning	
- Logistic Classifier	- John
- Random Forest Ensemble Classifier	- Leonard
- XGBoost Classifier	- Leonard
Dashboard	- John
Report	- Leonard