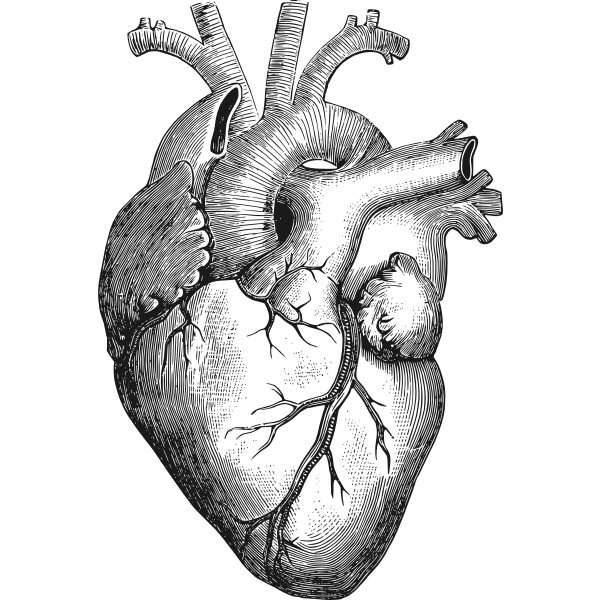
Heart Disease Project



Philopatir Bebway

[Literature Review 2](#_75rf4vta81ax)

[Multi Vessel Coronary Artery Disease Presenting as a False Negative Myocardial Perfusion Imaging and True Positive Exercise Tolerance Test: A Case of Balanced Ischemia 2](#_x2rgpx8mb194)

[Setup 3](#_lquiyrwpy6ke)

[Introduction 3](#_okpnclfca93l)

[Problem Statement 3](#_x957cgbtktrm)

[Approach 4](#_kn5uvgo00ajj)

[Proposed model and any benchmark models for comparison 4](#_xcz6s9w9zm6q)

[Implications for research question 4](#_nkpmfogqixgg)

[Model Experimentation 4](#_fhg5vy9jb3u)

[Linear Regression Classifier 4](#_qtxqnjuxcc8f)

[Support Vector Classifier 5](#_j8npvpz2lrzt)

[Tree Based Classifiers 6](#_sydsbli1k2t3)

[Decision Trees 6](#_p8oumtkfh20w)

[Random Forest Algorithm 6](#_it0wip21vgxq)

[Extra Random Trees 7](#_mrdz9j1d977i)

[Neural Network Classifier 7](#_rk6pyyaz61pn)

[Non-Linear Polynomial Regression 7](#_hq34j73cqxlp)

[Neural Network Regression 8](#_cfpwt3w9c7dr)

[Results 10](#_chou9188p6co)

[Conclusion 12](#_4576q1q7hd8w)

[Appendix 13](#_fxkpcdidec29)

[References 14](#_aaslcnc1pz4s)

# 

# Literature Review

Feature Selection: Analysing the impact of feature selection on the accuracy of heart disease prediction

This study tackles the issues of feature multicollinearity and dimensionality by testing multiple feature selection methods for Machine Learning Models. The researchers conducted a feature correlation analysis using correlation heatmaps displaying their corresponding coefficients to determine feature significance with respect to other features and the target variable. The analysis explored the inter-dependencies between variables and importance in feature elimination based on multiple high collinearity. (Pathan et al., 2022)

Multivariate Analysis of Risk Factors for Coronary Heart Disease:

The study samples men from Sweden, who did not exhibit signs of CHD. The researchers analysed nine probable risk factors and found that six of the nine probable factors increased the risk of CHD significantly. The study identifies multiple significant relationships between features significantly correlated with heart disease hence demonstrating the importance of multivariate analysis in identifying the complex interplay between different risk factors for CHD. The findings of the study have important implications to make sure our model looks at multiple risk factors simultaneously (Preis et al., 2009).

Model Selection: Implementation of Machine Learning Model to Predict Heart Failure Disease

Using the UCI repository datasets, this study attempts to test several Classification models on Heart disease diagnosis including Logistic Regression, SVM, Decision Tree and RandomForest. The research exhibits good overall performance across all models ranging between 86% and 96% accuracy scores. Given the similarities in datasets used in this study, a replication of models used in this study could potentially yield similar significant results. (Alotaibi, 2019)

Neural Networks: Performance evaluation of different machine learning techniques for prediction of heart disease

This study explores several classification models including Artificial Neural networks on the prediction of Heart disease. The study demonstrates ANN as a proficient method alongside Logistic regression in minimising misclassification and the presence of Heart disease, moreover maximising the potential of deep learning for boosting accuracy in classifying the presence of Heart disease. (Dwivedi, 2016)

# Multi Vessel Coronary Artery Disease Presenting as a False Negative Myocardial Perfusion Imaging and True Positive Exercise Tolerance Test: A Case of Balanced Ischemia

This study emphasises the role of non-invasive techniques in medical practices in diagnosis before further action on suggested medical practices. The study concludes placing further significance on not ignoring false negatives whereby there is a presence of a positive case as the ramifications of no actions could lead to a high probability of severe or critical heart disease. (Baqi et al., 2020)

# Setup

## Introduction

According to the World Health Organization (WHO), the two most fatal diseases around the world are heart related, making the jobs of healthcare specialists critical and the accuracy of precautionary testing extremely important. The Healthcare industry sits on a pool of untapped potential for predicting and preventing such fatal diseases through their collection of data recorded in clinical notes and Electronic Health Records (EHRs). This potential can be utilised through Machine Learning (ML) techniques used for diagnosis and risk assessment, including the identification of high risk for medical emergencies such as relapse or transition into another disease state.

During this project, we shall aim to study datasets related to Heart Disease and underlying attributes that would possibly indicate the development of the disease while also testing ML algorithms to predict the prevalence of the disease. The results of the ML models will be analysed using the following performance metrics: For Regression, Mean Squared Error (MSE) and Mean Absolute Error (MAE) ; For Classification, Accuracy Score, F1-score and Confusion Matrix.

## Problem Statement

“We shall analyse the various risk factors for Heart Disease and apply various statistical and computational models to predict the probability of Heart Disease in individuals.”

Research Question

Given a set of parameters for a patient, which variables are the most indicative of potential heart disease and how can we best predict its presence?

# Approach

## Proposed model and any benchmark models for comparison

For this experiment, various different models were trained. The initial stage focused on implementing Classification models while the secondary stage relied on Regression based prediction modelling and Neural Networks.

We initially focused on models with lower flexibility but faster computation time, which allowed us to experiment with as many models as possible. To further reinforce our findings, we implemented a neural network classifier to see if any more accuracy could be gained.

## Implications for research question

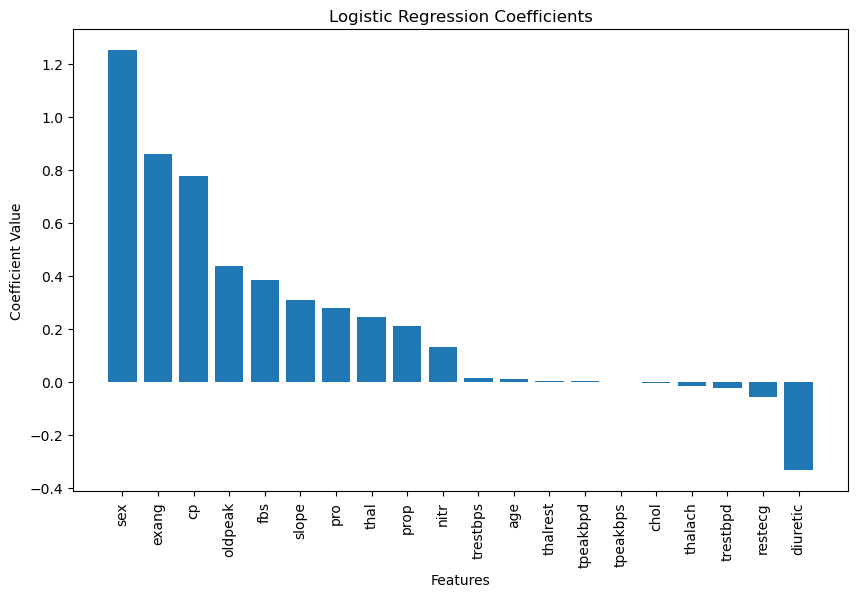
Chronic diseases like heart disease is not a state but a process which means a patient does not just transition from not having a disease to having a disease once a threshold is met (Liebman, M, 2019). Risks can actually progress over time and even reverse with lifestyle changes. Therefore, given the application of this experiment in a medical diagnosis domain, we moved to implement regression models so a continuum of heart disease risk can be provided for each patient. These regression model predictions would in turn complement our existing classification predictions and reinforce our suggestions to the respective medical domain/system.

## Model Experimentation

#### Linear Regression Classifier

The easiest model to implement, our implementation of logistic regression set the tone for our classification analysis. While it lacks in the number of hyper-parameters to tune and work with, it is much easier to work with and gives us a generic idea of the work going forward.

For the hyper-parameters, we identified the best solver for our dataset as "newton-cholesky", and balanced the class weights for the target variable.



Based on the logistic regression model coefficients, the features with the largest positive change in magnitude against the dependant variable include sex, exang, cp, oldpeak, while the most negative features are diuretic, restecq and trestbpd. Among the positive features, sex has the highest coefficients, followed by exang.

#### Support Vector Classifier

The second model we implemented is a popular choice simply due to the way the model works to visually partition a dataset. It does have a large number of hyper-parameters, however, in using this model we found value in changing only the kernel and gamma.

As such, at the cost of longer processing and computing times, the best hyper-parameters tuned were :

|  |  |  |
| --- | --- | --- |
| **C** | **Gamma** | **Kernel** |
| 1.0 | 0.001 | “linear” |

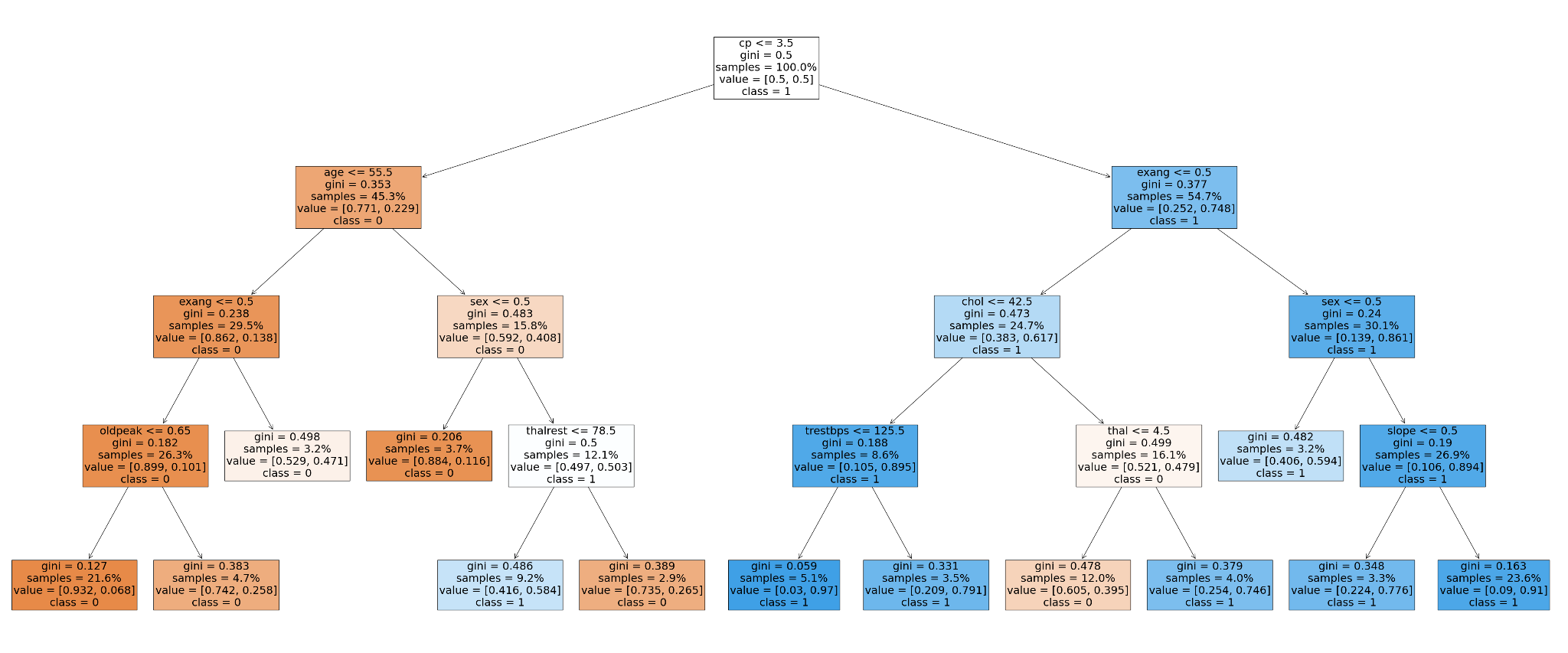
#### Tree Based Classifiers

##### Decision Trees

The first of the tree based classifiers, decision trees offer the most customizability, while suffering the most from overfitting and inaccurate training scores.

Of all the possible hyper-parameters for the model, the ones that yielded the best results against investment in time and avoiding overfitting were :

|  |  |
| --- | --- |
| **max\_depth** | **min\_samples\_leaf** |
| 4 | 20 |



The tree diagram shows the optimal rules for determining the target variable. For example the first node splits the data by patients that have a chest pain greater or equal to 3.5. The group with the high cp rating mostly consists of patients that have heart disease whilst the group with the low cp rating mostly consists of patients that don’t have the disease.

##### 

##### Random Forest Algorithm

Offering a much greater system of elimination and searching, this ensemble learning algorithm runs multiple decision trees in parallel and selects the one which offers the best split conditions and accuracies. However, the criteria for randomly selecting cut-offs and splits comes at the expense of longer computation times. Given our dataset size, we did not mind.

The most productive values for hyper-parameters were :

|  |  |  |
| --- | --- | --- |
| **n\_estimators** | **min\_samples\_split** | **max\_depth** |
| 50 | 100 | 8 |

##### 

##### Extra Random Trees

An extension of the Random Forest Algorithm, Extra Random Trees further the advantages and disadvantages of randomly selecting values and moving forward with the best splits and trees. We gain much more computation complexity, but also drastically increase our computation time.

The hyper-parameters searched for while tuning were :

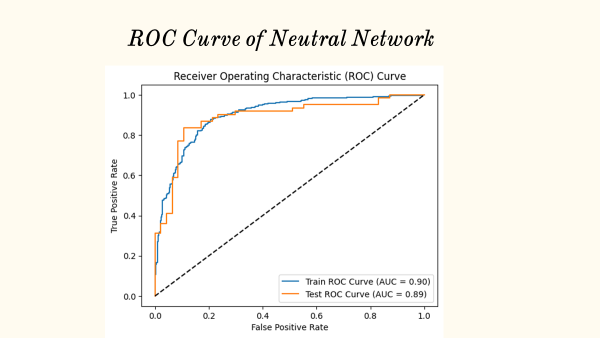
|  |  |  |  |
| --- | --- | --- | --- |
| **n\_estimators** | **min\_samples\_split** | **max\_depth** | **max\_leaf\_nodes** |
| 227 | 210 | 14 | 25 |

#### Neural Network Classifier

Using our findings from the previous classification models, we decided to implement a neural network to see if we could further improve on our results.

The parameters used for training were :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **hidden\_layer\_sizes** | **batch\_size** | **activation** | **solver** | **Learning\_ rate\_init** | **Learning\_ rate** | **max\_iters** |
| 512 | 32 | “relu” | “lbfgs” | 0.0025 | “adaptive” | 200 |

**A picture containing text, screenshot, rectangle, diagram

Description automatically generated**

**A picture containing text, screenshot, line, diagram

Description automatically generated**

The above ROC curve (receiver operating characteristics) represents performance of a binary classification model. The graph plots the true positive rate against the false positive rate. The train ROC curve (AUC=0.9) suggests that the model performs well in differentiating between positive and negative classes. The test AUC=0.89 suggests that the testing AUC is slightly lower than the training AUC. However, a higher level of AUC suggests that the model performs well on this classification task.

Overall, both the training and testing ROC curve indicate that the model has a relatively strong ability to classify instances correctly.

Based on the feature importance values, it seems that the most important features vary between layers. The two layers share some common patterns. The top features are thalach, chol, which suggest that they are highly influential in these layers. The least important features are prop, restecg, diuretic and pro, which indicates that they do not play important roles in the classification model.

In general, the both layers show similar patterns in terms of feature importance, with chol, thalach, tpeakbps

#### Non-Linear Polynomial Regression

Since we were unable to obtain a model that afforded us an accuracy rate of 90% or more, we decided to implement a regression model to predict values based on the range of the target variable. This prediction system allowed us to avoid misclassification, as well as worked in conjunction with our classification system to reinforce the results and suggestions.

We trained a non-linear regression with a number of polynomial features of 2, which afforded us an extremely low mean square error and mean absolute error rate, which coincided with the expected implementation.

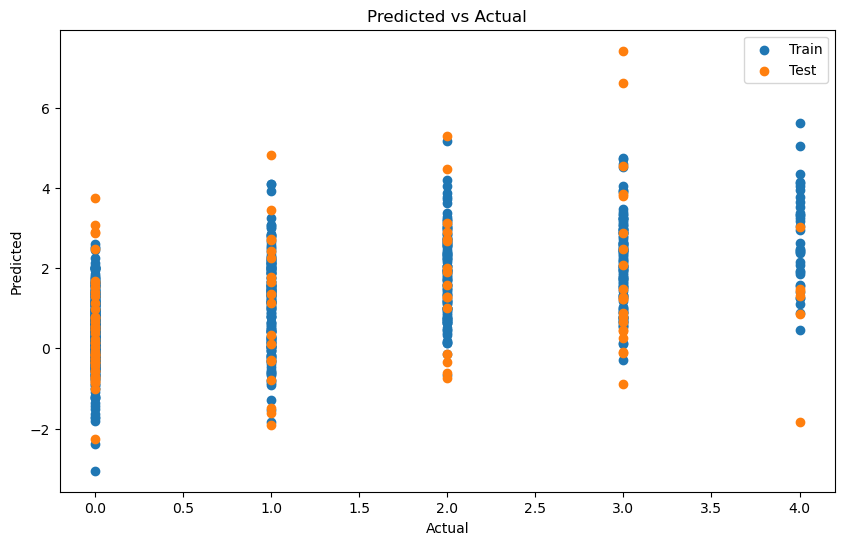
#### 

#### 

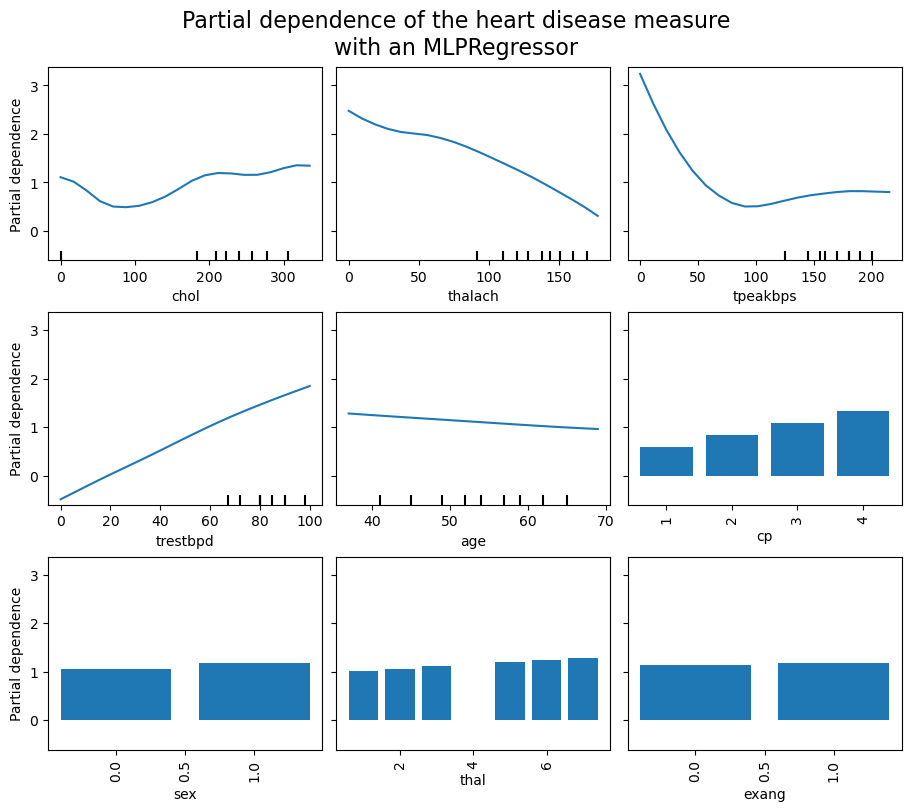
#### Neural Network Regression

As the final stage of our modelling, we implemented a Neural Network based regression model to try and further improve on our non-linear model. The result of this attempt was the following hyper-parameters :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **hidden\_layer\_sizes** | **batch\_size** | **activation** | **solver** | **Learning\_ rate\_init** | **Learning\_ rate** | **max\_iters** |
| 128 | 32 | “relu” | “adam” | 0.001 | “constant” | 5 |



When plotting the predicted against the actual, the values do show pretty good correlation for the 0 - 2 level of heart disease severity, the predicted values do concentrate around the actual. For categories 3-4 the results were less conclusive. Potentially for severe heart disease there might be other health issues or metrics that might not have been outside the dataset.



A partial dependence plot (PDP) was created to assist the interpretability of the neural network model. It shows the marginal effect that the features have on the prediction produced by the model. For chol (Cholesterol) it seems to be a significant factor to indicate heart disease in patients as it goes up, although there is an increase around zero as well but this might be due to some data issues.

In tpeaksbps (peak exercise blood pressure) the PDP is a hockey stick shape. This might suggest very low blood pressure during exercise is a predictor of heart disease but so is very high blood pressure. Whilst in cp, the higher the chest pain rating, the higher the prediction of a patient having the disease

# Results

Summary of Results

The table below summarises the performance of our Model testing section. By placing the overall performance of the training and testing split side by side, it is quite evident that our models perform well in avoiding the common issue of overfitting, especially with a relatively small dataset. Despite this, the only model that reduces in accuracy score while predicting on the test split is the randomforest. This can be attributed towards multiple factors such as small dataset as randomforest benefits from having larger data to build more decision trees. Further investigation demonstrates that inappropriate hyperparameter tuning may lead to overfitting such as increasing the max depth. More testing is required to validate these theories of overfitting.

As mentioned in our literature review, we anticipated SVM to outperform the other models, which stands at a 6% increase in accuracy between the training and testing predictions. It was noted that SVM performs extremely well on datasets that have a balanced class, which is the case with our set. With our focus on minimising false negatives, while maintaining a high accuracy score, the SVM alongside the Neural Network classifier perform well, recording the lowest false negative classifications. A result that was anticipated based on prior research, and verified by consistent scores across 10 folds of cross validation performed for final verification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name.** | **Training** | | **Testing** | |
| **Confusion Matrix.** | **Accuracy and F1 Scores.** | **Confusion Matrix.** | **Accuracy and F1 Scores.** |
| Logistic Regression | [[289, 68],  [ 77, 357]] | Accuracy Score : 81.66%  F1 Score : 81.66% | [[ 41, 6],  [ 9, 52]] | Accuracy Score : 86.11 %  F1 Score : 86.15 % |
| Support Vector Classifier | [[286, 71],  [ 72, 362]] | Accuracy Score : 81.92%  F1 Score : 81.92% | [[ 41, 6],  [ 7, 54]] | Accuracy Score : 87.96 %  F1 Score : 87.98 % |
| Decision Tree Classifier | [[289, 68],  [ 91, 343]] | Accuracy Score : 79.89%  F1 Score : 79.93% | [[ 41, 6],  [ 13, 48]] | Accuracy Score : 82.41 %  F1 Score : 82.48 % |
| Random Forest Algorithm | [[281, 76],  [ 58, 376]] | Accuracy Score : 83.05 %  F1 Score : 83.01 % | [[ 41, 6],  [ 13, 48]] | Accuracy Score : 82.41 %  F1 Score : 82.48 % |
| Extra Random Trees | [[281, 76],  [ 84, 350]] | Accuracy Score : 79.77%  F1 Score : 79.79% | [[ 40, 7],  [ 12, 49]] | Accuracy Score : 82.41 %  F1 Score : 82.48 % |
| Neural Network Classifier | [[282, 75],  [ 60, 374]] | Accuracy Score : 82.93 %  F1 Score : 82.89 % | [[ 37, 10],  [ 8, 53]] | Accuracy Score : 83.33 %  F1 Score : 83.29 % |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name.** | **Training** | | **Testing** | |
| **Mean Squared Error.** | **Mean Absolute Error.** | **Mean Squared Error.** | **Mean Absolute Error.** |
| Non Linear Polynomial Regression |  |  | 3.389318630858085 | 1.4331509426788047 |
| Neural Network Regression | 1.5014565400532573 | 0.9631860837070207 | 1.703759707085431 | 0.9846019936949404 |

# 

# Evaluation

# Limitations of Research

1. **Limited Sample size**

* A small dataset exhibits issues of overfitting, underfitting and misclassification due to the relatively low and redundant number of information to learn from.

1. **Data quality**

* The increased number of missing values require actions such as pruning which can result in alternative noise in the reliability of the study

1. **Feature selection**

* With the presence of over 70 features, the method of feature selection is crucial and inappropriate methods can lead to suboptimal performance.

# Obstacles Overcome

1. **Dataset Impurity and Visualisation :**

* Researching a number of fields with poor descriptions and information.
* Optimising dataset splitting due to size and shape constraints.
* Visualising a dataset with missing and corrupted fields, while ensuring accuracy.

1. **Feature elimination and correlation :**

* Determining the features we needed versus those that impacted accuracy.
* Eliminating seemingly balanced features before and after merging the datasets into one.

# Future Improvements

1. **Dataset generalisation and expansion :**

* Interaction and feedback with professionals in the medical fields.
* Increasing the number of entries.
* Decreasing the number of extraneous variables.
* Establishing guidelines for measurement.

1. **Applying a combination of more complex and interconnected models, allowing better performance :**

* Researching and training better models, specifically neural networks.
* Collecting better information for metrics of performance and deployment.

# Conclusion

First and foremost, with respect to the business and the outcomes in that domain, we have researched, trained and tuned a model that is more than satisfactory in terms of performance and repeatability. We hope that it will assist medical professionals in better serving their patients, as well as improve efficiency and treatment procedures.

We also have a stronger grasp on the various machine learning models we studied, along with their unique strengths and weaknesses. Additionally, nuanced features such as multicollinearity, optimised dataset splitting and hyper-parameter tuning are now much easier to detect, implement and use effectively.

Various obstacles did present themselves in the course of our work, however, they also brought opportunity to grow and advance our knowledge. From a poorly explained dataset that required weeks of research to corrupted fields and values, we persevered to ensure representative testing and training, and to visualise the data and processes effectively.

Going forward, we have learned where the shortcomings in our project lie and wish to focus on them in future tasks and reviews. We plan to work with professionals in the medical domain to better understand the features and their importance in a practical setting, in turn working towards a more generalised and concise dataset, increasing the number of entries we have for training and testing even further.

We also hope to leverage the advances made in the domain of machine learning with newer models and better performance metrics, to train better and more precise networks that will ensure better accuracy and even better recommendations.

# Appendix

## Data Dictionary

| Variables | Description |
| --- | --- |
| age | age in years |
| sex | sex (1 = male; 0 = female) |
| cp | chest pain type  -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| trestbps | resting blood pressure (in mm Hg on admission to the hospital) |
| chol | serum cholesterol in mg/dl |
| cigs | cigarettes per day |
| years | number of years as a smoker |
| fbs | fasting blood sugar > 120 mg/dl (1 = true; 0 = false) |
| dm | 1 = history of diabetes; 0 = no such history |
| famhist | Family history of coronary artery disease (1 = yes; 0 = no) |
| restecg | Resting electrocardiographic results  -- Value 0: normal  -- Value 1: having ST-T wave abnormality  -- Value 2: showing probable left ventricular hypertrophy by Estes' criteria |
| prop | Beta blocker used during exercise ECG: 1 = yes; 0 = no |
| nitr | Nitrates used during exercise ECG: 1 = yes; 0 = no |
| pro | Calcium channel blocker used during exercise ECG: 1 = yes; 0 = no |
| diuretic | Diuretic used used during exercise ECG: 1 = yes; 0 = no |
| thalach | Maximum heart rate achieved |
| thalrest | Resting heart rate |
| tpeakbps | Peak exercise blood pressure (first of 2 parts) |
| tpeakbpd | Peak exercise blood pressure (second of 2 parts) |
| trestbpd | Resting blood pressure |
| exang | Exercise induced angina (1 = yes; 0 = no) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | The slope of the peak exercise ST segment  -- Value 1: upsloping, Value 2: flat , Value 3: downsloping |
| ca | Number of major vessels (0-3) coloured by fluoroscopy |
| thal | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| num | Diagnosis of Heart Disease (angiographic disease status)  -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing  (in any major vessel: attributes 59 through 68 are vessels) |

# 

# References

1. Preis, S. R., Pencina, M. J., Hwang, S.-J., D'Agostino, R. B., Savage, P. J., Levy, D., & Fox, C. S. (2009). Trends in cardiovascular disease risk factors in individuals with and without diabetes mellitus in the Framingham Heart Study. Circulation, 120(3), 212–220. https://doi.org/10.1161/circulationaha.108.846519
2. Baeßler, B., Götz, M., Antoniades, C., Heidenreich, J. F., Leiner, T., & Beer, M. (2023). Artificial Intelligence in coronary computed tomography angiography: Demands and solutions from a clinical perspective. Frontiers in Cardiovascular Medicine, 10. https://doi.org/10.3389/fcvm.2023.1120361
3. Luo, C., Zhu, Y., Zhu, Z., Li, R., Chen, G., & Wang, Z. (2022). A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. Journal of Translational Medicine, 20(1). https://doi.org/10.1186/s12967-022-03340-8
4. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K., Lee, S., &amp; Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310. <https://doi.org/10.1016/0002-9149(89)90524-9>
5. Sidey-Gibbons, J. A., &amp; Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. BMC Medical Research Methodology, 19(1). <https://doi.org/10.1186/s12874-019-0681-4>
6. Sisodia, D., &amp; Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585. [https://doi.org/10.1016/j.procs.2018.05.12](https://doi.org/10.1016/j.procs.2018.05.122)2
7. WILHELMSEN, L. A. R. S., WEDEL, H. A. N. S., &amp; TIBBLIN GÖSTA. (1973). Multivariate analysis of risk factors for coronary heart disease. Circulation, 48(5), 950–958. <https://doi.org/10.1161/01.cir.48.5.950>
8. Australian Institute of Health and Welfare. (n.d.). Heart, stroke and vascular disease: Australian facts, coronary heart disease. AIHW, Australian Government. Retrieved March 25, 2023, from https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/sum mary-of-coronary-heart-disease-and-stroke/coronary-heart-disease
9. Dwivedi, A. K. (2016). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, *29*(10), 685–693. https://doi.org/10.1007/s00521-016-2604-1
10. Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, 100060. https://doi.org/10.1016/j.health.2022.100060
11. Alotaibi, F.S. (2019) ‘Implementation of machine learning model to predict heart failure disease’, *International Journal of Advanced Computer Science and Applications*, 10(6). doi:10.14569/ijacsa.2019.0100637.
12. Baqi, A., Ahmed, I. and Nagher, B. (2020) *Multi vessel coronary artery disease presenting as a false negative myocardial perfusion imaging and true positive exercise tolerance test: A case of balanced ischemia*, *Cureus*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717085/ (Accessed: 23 May 2023).
13. Liebman, M. (2019). Medicine: It may be precise but is it accurate? *INFORMS News*. doi:https://doi.org/10.1287/orms.2019.05.04.