Module Title: Advanced Machine learning and Python

Title: Heart Disease Prediction Using Machine Learning

Lecturer name: Michael Farayola

Student Name: Laeeq Ur Rehman

Student number: 22448606

Submission Date:18/04/2025

Table of Contents

Reflective Report: Heart Disease Prediction using Machine learning

Introduction to Methodology

Machine learning has become an indispensable tool in modern medicine (Naser et al., 2024), particularly in the early detection and prediction of diseases such as heart disease, which remains one of the leading causes of death globally, even in 2025 (Lewis, 2025). With advancements in innovation and the development of predictive tools, machine learning algorithms can utilize structured patient data to assist medical professionals by providing accurate, fast, data-driven predictions. However, as someone's life is on the line, such practices must be completely accurate, effective, and trustworthy; the development methodology must adhere to strict guidelines, be clear, and be evidence based.

This report aims to implement and evaluate a machine learning model that can predict the presence or absence of heart disease using a publicly available dataset. The dataset used was the Cleveland Heart Disease Dataset (Janosi et al.), which was found on the UCI Machine learning Repository, and was also published on Kaggle (Ritwik_B3, 2020). The approach taken in this report was not only shaped by the technical coding requirements of the assignment but also by the practices I observed in published research that applied machine learning across various healthcare datasets. In this section, we outline the research that assisted us in our methodology, followed by a breakdown of how we implemented the code, choice of selection of models, the handling of data, and the optimization of our evaluation process.

Review of Related Works

A peer-reviewed study titled "Prediction of Heart Disease Based on Machine Learning Using the Cleveland Heart Disease Dataset" (Ahmad, A.A. and Polat, 2023) played a crucial role in guiding my approach. This research paper used the same dataset that we used, the Cleveland Heart Disease Dataset from the UCI machine learning Repository. I found that this dataset is used in many cardiology-related machine learning research due to its balanced features, small and manageable data size, and its relevance to real-world problems.

In their methodology, Ahmad, A.A., and Polat, H. (2023) used a variety of classification models, including Random Forests, Logistic Regression, and Support Vector Machines. A key innovation in their study was the application of the Jellyfish optimization for feature selection, which was aimed at enhancing the model's performance and reducing dimensionality. Their Data preprocessing steps included standardizing and handling of missing values. The study encompassed accuracy, precision, recall, F1 score, and ROC-AUC, with an SVM classifier demonstrating the best performance, achieving an accuracy of 98.47% and an AUC of 94.8%. These metrics exceeded many other benchmarks I found in other cardiovascular prediction papers, demonstrating the power of optimized machine learning pipelines when applied to even modestly sized datasets.

This research paper helped me with the structural guidance for my report. We selected the same dataset and followed similar preprocessing steps, including standardization and addressing missing values. By critically examining this research paper, I found various enhancements that would not just improve the robustness of my analysis but also help to differentiate my work. Unlike Ahmad, A.A., and Polat, H. (2023), who focused on improving neural network performance using the Jellyfish Optimization Algorithm and relied solely on a single train-test split, I used a three-way data split (training, validation, and test), to better assess the model's generalizability. I also implemented GridSearchCV for hyperparameter tuning and incorporated a 5-fold cross-validation to ensure a deeper search for optimal model parameters. Additionally, while their study prioritized neural network optimization, I focused on the interpretability and comparative evaluation of traditional classifiers, including Random Forests, Logistic Regression, and Support Vector machines, which, according to other studies, have demonstrated very high performances in heart disease prediction. This focus enabled a deeper exploration of the model's strengths and trade-offs in the context of heart disease prediction.

To better understand the scalability of Machine learning in healthcare, I reviewed the paper "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" by Weng et al. (2017). What set this study apart from ours was the scale and data complexity. We used structured, well-cleaned tabular data with 13 features and 303 samples. In comparison, Weng et al. (2017) worked with noisy, high-dimensional real-world data containing over 378,000 patients and hundreds of clinical variables such as demographic data, lab results, and even medical history. As a result, their model required more complex data preprocessing and dimensionality reduction techniques.

Weng et al. (2017) compared 5 machine learning algorithms: Random Forest, Logistic Regression, Gradient Boosting, Neural Networks, and Naïve Bayes. This was used against the Framingham risk score, which is based on traditional statistical modelling. However, there was some similarity between both studies, which was the use of ensemble learning methods (Random Forests) and the emphasis on interpretability.

Our Methodology and Implementation

With the help of studies that were centred around machine learning in healthcare, our report began by importing the Cleveland Heart Disease dataset into a Google Colab environment using Python's "Pandas" library. First, I checked for missing values in the dataset, which were denoted by "?", and were replaced with NaN and subsequently dropped using dropna(). This step was essential as it ensured data quality, as it eliminated incomplete entries that could introduce bias or outliers that could disrupt model training. Additionally, I checked for duplicate rows to maintain a clean and non-redundant data set.

To align with the binary classification goals seen in research papers (Iacobescu et al., 2024), we transformed the dataset's original multiclass target variable, ranging from 0 to 4, into a binary variable. 0 indicating no heart disease and 1,2,3,4 indicating the severity of the disease. Any value that was greater than 0 was converted to 1 (which would indicate the presence of heart disease), and any value that was 0 remained as is (this indicated heart disease was absent). This binarization step simplified our task and matched the methodology used in other research papers.

After this, all feature variables were then scaled using "StandardScaler", a normalization technique which standardizes features by removing the mean and scaling to unit variance. This step was essential, especially for classification algorithms such as Support Vector Machines and Logistic Regression, which are sensitive to feature magnitudes (Juszczak, Tax and Duin, 2002). A visual inspection of feature distributions before and after scaling using histograms proved the effectiveness of standardization. The raw features were unevenly distributed and varied in value, while the scaled feature was centred around 0 and had a standard deviation of 1. This visual evidence supported the use of “StandardScaler” and ensured compatibility with models sensitive to feature magnitude, such as SVM and Logistic Regression.

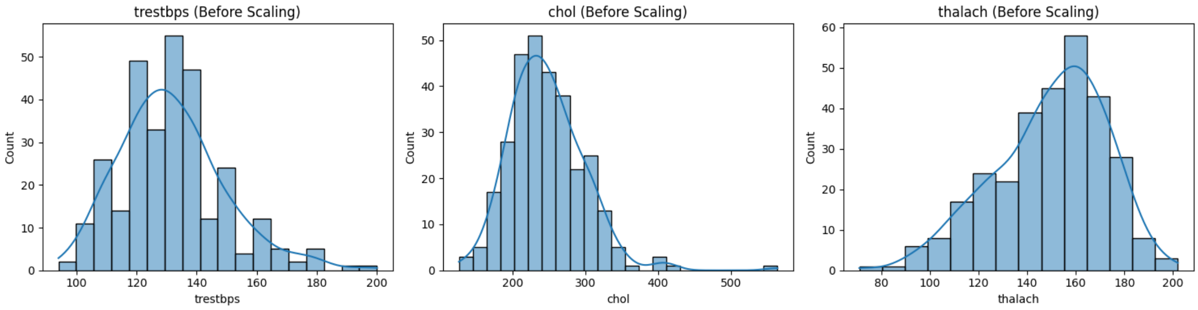


Fig.1- Before Scaling:

Distributions of trestbps, chol , and thalack before standardization. The features are unevenly scaled, which negatively impacts models such as SVM and Logistic Regression that are distance-based.

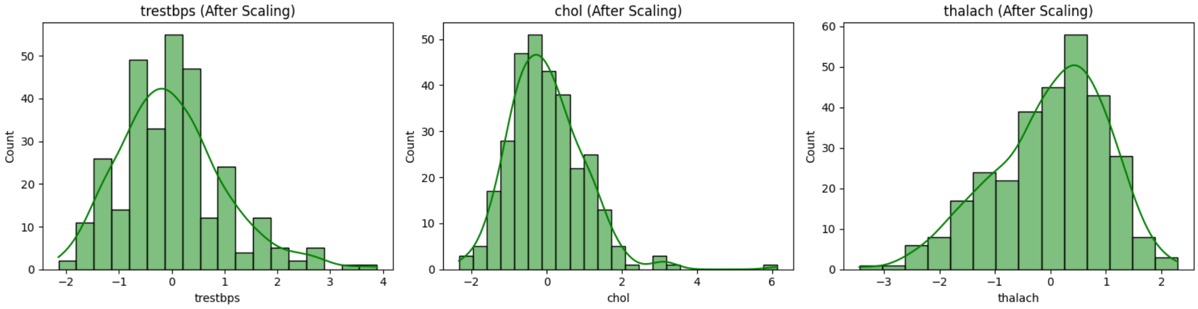


Fig.2- After Scaling:

Distributions of the same feature after applying StandardScaler. The values are now centred around 0 and have unit variance, ensuring all features contribute equally.

The dataset was then split into three partitions using train_test_split():

- 60% Training Set
- 20% Validation Set
- 20% Test Set

This split allowed us to perform hyperparameter tuning using the validation set while preserving an untouched test set for final evaluation. This is a stronger approach than a simple train-test split, and it allows better generalization of model performance.

I chose 3 supervised classification models for implementation:

1. Logistic Regression
2. Support Vector Machine (SVM)

3. Random Forest

Of these, Random Forest was the only model that I subjected to extensive hyperparameter tuning. We used GridSearchCv to evaluate combinations of:

- n_estimators (number of trees)
- max_depth (depth of each tree)
- min_samples_split (minimum number of samples to split)
- max_features (how many features to consider per split)

This tuning was performed using 5-fold cross-validation, which allowed the model to be tested across various data splits, thereby increasing the depth of the data. The best parameters were selected based on cross-validated accuracy scores, and the final model was then applied to the test set.

Evaluation Metrics

Each model's performance was measured using the following metrics:

- Accuracy: The % of correct predictions over the total number of predictions.
- Precision: Proportion of true positives among the total number of positives.
- Recall (Sensitivity): The proportion of true positives among the total number of positives.
- F1-score: A single number that combines precision and recall that tells how good the model is, especially when the data is imbalanced.
- ROC-AUC: The area under the ROC, which measures the model's ability to distinguish between classes.

This was computed by using Scikit-Learn's built-in function. This guided us on our model comparisons and final selection. To visually compare class separation performances, ROC curves were plotted.

Confusion Matrices and Feature Importance

To support our metric-based evaluation, we generated confusion matrices for each model using the function "seaborn.heatmap()". These matrices visualized how each model classified positives and negatives and helped in identifying the type of error (false positives vs false negatives). This is especially relevant in a healthcare context, where different error types have varying clinical consequences.

For Random Forest, we also used feature importance analysis. This was done by using the built-in feature importances_ attribute, which helped us distinguish which clinical features contributed most to the models' decisions. By including this analysis, I aimed to increase the interpretability of our results and draw a relationship between model behaviour and medical knowledge.

Results

After doing model training and validation, each classifier was evaluated using a set of KPI's on the held-out test set. These were: Accuracy, Precision, Recall, F1-score, and ROC-AUC, which together provided a comprehensive view of how well the model performed and generalized on unseen data.

1. Performance Metrics on Test Set

Each model was tested using Scikit-Learn's metric functions. The results showed that SVM was the best performer, performing better than both Logistic Regression and Random Forest across all the major metrics. The scores were reflected by the macro average for weighting across both classes.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Logistic Regression	0.87	0.87	0.87	0.87	0.92
Support Vector Machine	0.90	0.90	0.90	0.90	0.92
Random Forest	0.82	0.83	0.82	0.82	0.93

Table 1: Weighted average of evaluation metrics on the test set for all models

Among these classifiers, SVM demonstrated the highest overall performance with an accuracy of 90% and consistent Precision, Recall, and F1-score of 0.90. Logistic Regression also performed well, with a balanced set of scores across all metrics and an accuracy of 87%. What caught my attention was that the Random Forest underperformed compared to what I expected, especially in recall, which dropped from 0.90 to 0.75 for the positive class (presence of disease) despite achieving high precision for that class, 0.90. This suggests that Random Forest may have misclassified more positive classes (presence of disease) cases compared to SVM and Logistic Regression.

Cross-Validation Accuracy

To ensure the model's reliability, I used 5-fold cross-validation on the training set before the final evaluation. This tests the model across five data splits, helping to identify overfitting and providing a more stable estimate of performance.

Model	**Mean CV accuracy**
Logistic Regression	0.84
SVM	0.86
Random	0.89

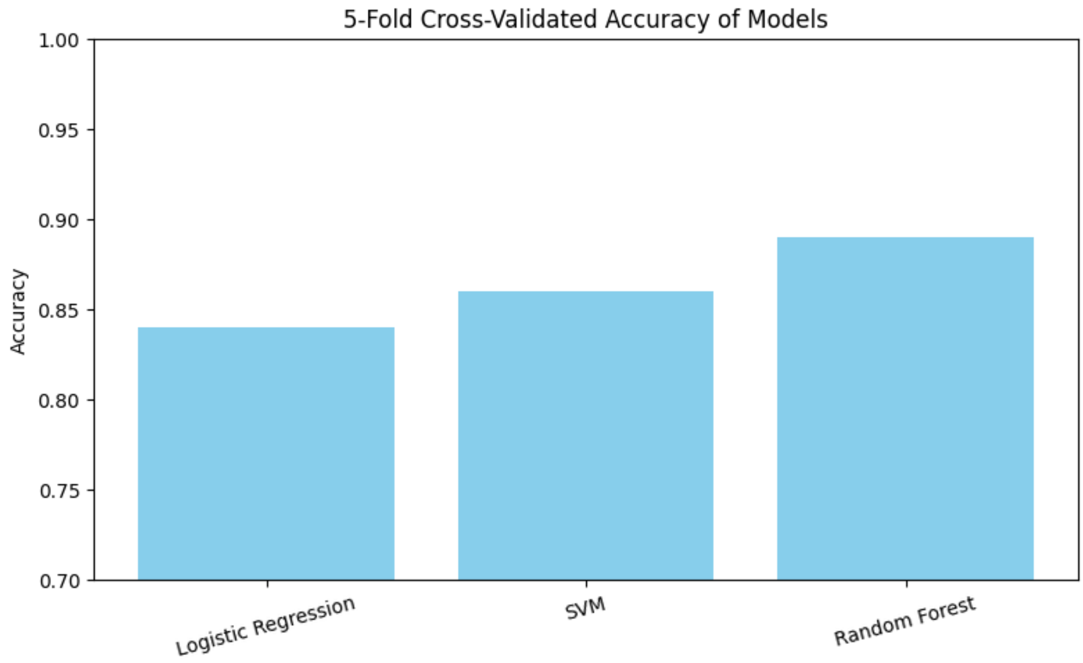


Fig.3

Unexpectedly, Random Forest had the highest cross-validation score, which suggests that while it performed well during training, it may not have performed as well on the test set. This result reminds the importance of using both CV and final test evaluation.

3. ROC Curve Analysis

To visually interpret each model's ability to separate classes across thresholds, ROC curves were plotted for each model, showing true positive rates vs false positive rates at different thresholds. The AUC (Area Under the Curve) summarizes this performance.

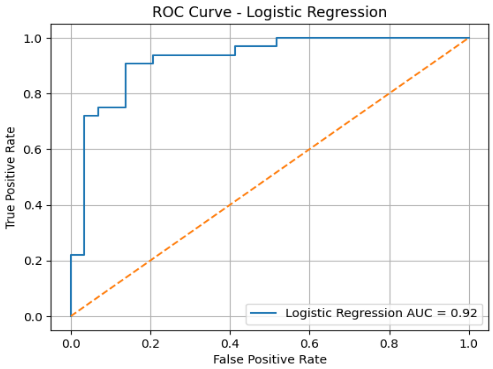


Fig.4

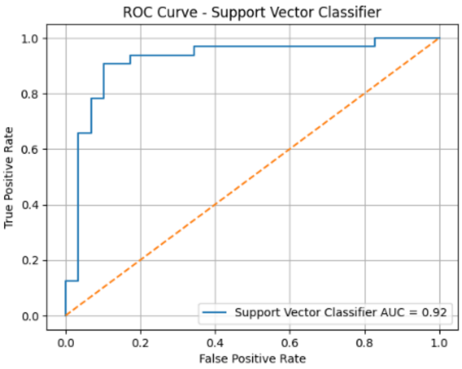


Fig.5

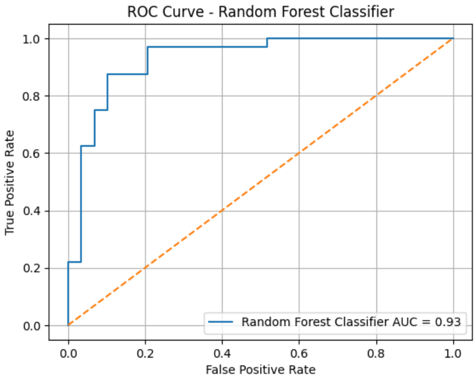


Fig.6

They all performed well, but Random Forest achieved the highest AUC of 0.93, which suggested a superior classification ability, but its lower test accuracy and recall suggest that its predictions may have been less balanced. SVM achieved a high AUC of 0.92 and maintained top scores across all test metrics, reinforcing SVM's status as the best overall performer.

4. Confusion Matrices

To better understand the classification behaviour of each model, I generated confusion matrices. These diagrams helped identify the types of errors each model made.

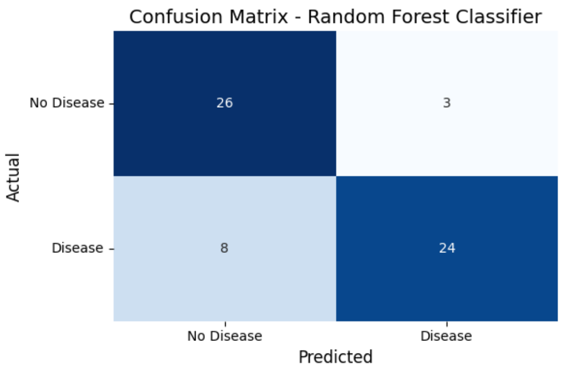
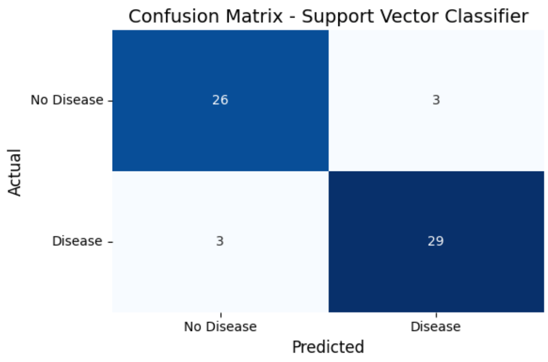


Fig.3

Random forest produced more false negatives (failing to identify if the disease is present), which can be critical in medical contexts.



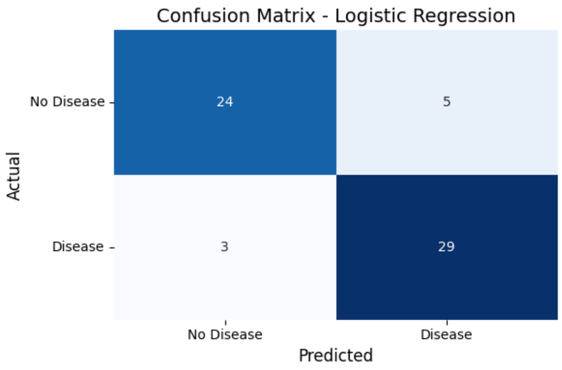


Fig.4
Fig.5

Both SVM and Logistic Regression had a more balanced distribution of predictions, with fewer false positives and negatives.

The confusion Matrices reinforced the superiority if the SVM model, especially in reducing both types of errors.

5. Feature Importance (Random Forest)

Features such as “cp” (chest pain type), “oldpeak” (ST depression), and “thalach” (Maximum heart rate achieved). These features were constantly rated among the top factors for heart disease across many other clinical studies. For example, Iffat Ara Talin et al. (2022 identified these features to be among the most influential in diagnosing heart disease, highlighting their clinical relevance and the model's validity.

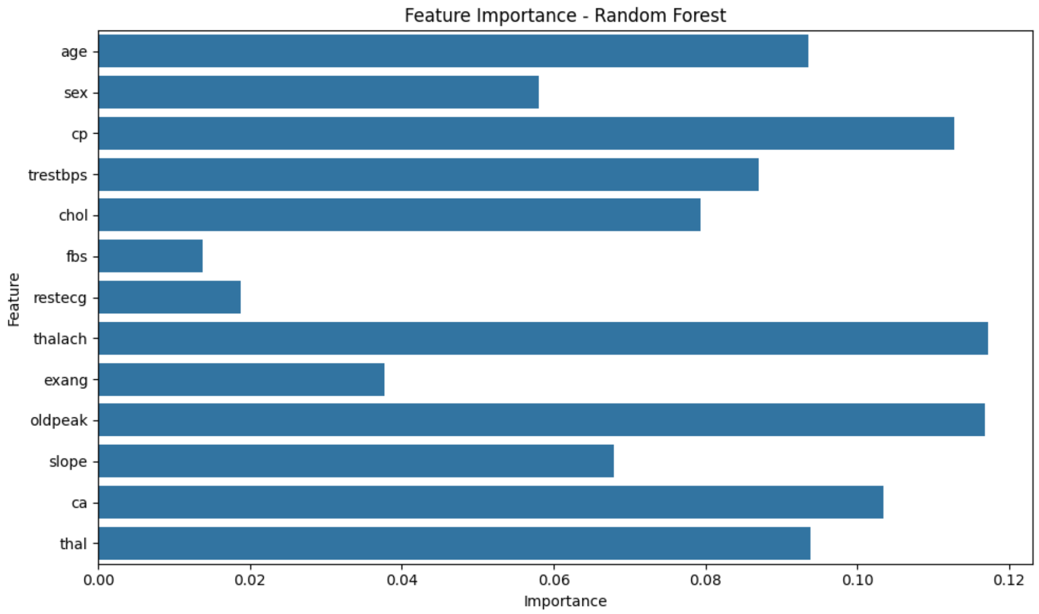


Fig.7

This table not only provides insight into the model's inner workings but also strengthens the interpretability and relevance of the model, which is key for potential real-world medical applications.

6. Summary

These results tell us that SVM was the best-performing model overall, with consistent strength across accuracy, AUC, and class-level precision/recall. This result aligns with many findings in the reviewed literature, where SVMs have demonstrated robust classification performances in medical datasets, especially for heart disease prediction (Sohn, Kim, and Moon, 2007).

Logistic Regression was a close second, it was simple and still effective. This is supported by Dinh et al., (2019) research, which shows Logistic Regression, despite its simplicity, performs effectively in heart disease prediction tasks, and it remains a reliable baseline model in medical datasets.

Finally, Random Forest showed strong cross-validation and an AUC but struggled with generalization to the test set. This same limitation is also highlighted in other studies, where Random Forest models demonstrated high performance on training data but faced overfitting issues and reduced generalizability to unseen data (Lasai Barreñada et al., 2024)

These results demonstrate the importance of using a combination of performance metrics, cross-validation, and ROC analysis when choosing the right model, especially in a critical domain like healthcare.

Discussion

This report set out to explore how machine learning can be used and applied to predict heart disease using structured medical data. While the implementation and evaluation phases provided critical insights into the model's performance, this discussion section allows us to reflect on the implications of our choice, the lessons learned, and the complexity of this work.

Interpreting Model Performance.

Among the 3 models, logistic regression, Support Vector Machine (SVM), and Random Forest. SVM delivered the most consistent and strong results. It achieved the highest test accuracy of 90%, with precision, recall, and F1-score scoring the highest as well. It also produced an impressive AUC of 0.92, indicating good class separation ability. This aligns with a lot of research, which supports the use of SVM in small to medium-sized structured datasets such as (Althnian et al., 2021).

Random forest had the highest cross-validation accuracy of 0.89 but also had the lowest test accuracy of 0.82. This difference suggests potential overfitting; the model may have adapted too well to the training fold and lost generalisability (Janitza and Hornung, 2018). Despite its high AUC of 0.93 and strong feature importance insights make it valuable from an interpretability standpoint. Logistic Regression was simple and reliable, staying consistent across all metrics and had no signs of overfitting. While slightly behind SVM in performance, it remains an excellent baseline model and is highly explainable. In some instances, it was found that Logistical Regression outperformed better than more complex models such as Decision Trees and SVM in breast cancer diagnostics (Arshad, Shahriar, and Anjum, 2023), which reinforces the statement that highlights Logistical Regression's simplicity, interpretability, and computational efficiency.

This multi-model approach showed that no single metric is superior and tells the whole story. For instance, Random Forest AUC did not translate into better real-world performance on the test set, indicating that accuracy, recall, and confusion matrices are all critical in healthcare, where false negatives can impact patient health.

Impact of Preprocessing and Feature Engineering.

The success of this report was heavily influenced by early preprocessing decisions. Binarizing the target variables (from multiclass to binary), handling missing values, and standardizing features were all essential. Scaling significantly improved the performance of both SVM and Logistic Regression, which are extremely sensitive to feature magnitude.

Visual checks before and after scaling confirmed that the data were prepared properly. These steps aren't just needed for technical necessity, they're the foundation for a model to perform fairly and accurately. The report reinforced that a well-preprocessed small dataset can often outperform a poorly managed large dataset.

Evaluation Strategy and the Importance of Cross-Validation

Using 5-fold cross-validation gave us valuable insight into model stability. Random Forest, which excelled in cross-validation, performed very poorly on the test set, showing the importance of evaluating across different folds and not relying solely on a single train-test split.

This strategy also shows that cross-validation scores don’t always guarantee test performance, especially if the validation data is too like the training data. SVM’s consistency across both CV and test performance metrics made it the safest choice overall

Feature Importance and Model Explainability

Random Forest’s built-in feature importance function allowed us to see into the model’s decision-making. This highlighted the most predictive features: chest pain type (cp), St depression (oldpeak), and maximum heart rate (thalach). These are all frequently cited in medical research as the vital indicators of heart disease (Talin et al., 2022). While SVM does not have this level of built-in interpretability. Random Forest helped validate the clinical relevance of the dataset and model behaviour. This highlights the importance of testing on a mix of models for raw performance and other insights.

Limitations and Challenges

There were some limitations in this report. First, the size of the dataset (only 303 records), limited the complexity of models we could train, and made the models at risk to outliers. The likes of Deep learning and ensemble stacking techniques might require larger datasets to avoid overfitting.

The dataset came from a single source (Cleveland Clinic) and may not generalize well to other demographics or clinical settings. We also handled missing data and duplicates, but we didn’t apply any imputation techniques or test different preprocessing strategies that could have improved model performance.

Finally, the model's explainability remained an issue. While Random Forest offers feature importance, neither Logistic Regression nor SVM provides intuitive “decision paths “. Using tools like SHAP or Lime could enhance transparency, which is especially important in clinical applications (Arjunan, 2021).

Future improvements

In the future, many things can be done to build on this work.

- Firstly, test on larger and more diverse datasets (e.g., from multiple hospitals or countries).
- Implement model explainability tools such as SHAP values.
- Explore deep learning models like MLPs or CNNs if the data size is larger.
- Deploy it as a prototype app with synthetic patient data to trial model predictions.
- Expand Feature engineering by combining EHRs, medication history, or real-time vitals from wearables.

Conclusion of Discussion

This report was not just about applying machine learning concepts to train models, it was about making careful decisions at every step. From cleaning the data to its validation, model selection to visualization, the process taught me that success in machine learning is rarely about algorithms alone; it's about understanding the problem, the data, and the trade-offs behind every line of code.

SVM delivered the best overall performance, but each model taught us something unique. In real-world healthcare, it’s not just about accuracy, it's about life or death. Safety, trust, and transparency. This report lays the groundwork for more advanced systems, and using data science does not just to analysing lives but helps and saves them.

Bibliography

Ahmad, A. and Polat, H. (2023). Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. *Diagnostics*, 13(14), pp.2392–2392. doi:https://doi.org/10.3390/diagnostics13142392.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A.B., Alzakari, N., Abou Elwafa, A. and Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences*, [online] 11(2), p.796. doi:https://doi.org/10.3390/app11020796.

Arjunan, G. (2021). Implementing Explainable AI in Healthcare: Techniques for Interpretable Machine Learning Models in Clinical Decision-Making. *International Journal of Scientific Research and Management (IJSRM)*, 9(05), pp.597–603. doi:https://doi.org/10.18535/ijsrm/v9i05.ec03.

Arshad, M.A., Shahriar, S. and Anjum, K. (2023). *The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques -- Case Of Breast Cancer Diagnosis*. [online] arXiv.org. Available at: https://arxiv.org/abs/2306.02449.

Dinh, A., Miertschin, S., Young, A. and Mohanty, S.D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1). doi:https://doi.org/10.1186/s12911-019-0918-5.

Iacobescu, P., Marina, V., Anghel, C. and Anghele, A.-D. (2024). Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities. *Journal of Cardiovascular Development and Disease*, 11(12), p.396. doi:https://doi.org/10.3390/jcdd11120396.

Iffat Ara Talin, Mahmudul Hasan Abid, Md. Al-Masrur Khan, Kee, S.-H. and Abdullah-Al Nahid (2022). Finding the influential clinical traits that impact on the diagnosis of heart disease using statistical and machine-learning techniques. *Scientific reports*, 12(1). doi:https://doi.org/10.1038/s41598-022-24633-4.

Janitza, S. and Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLOS ONE*, 13(8), p.e0201904. doi:https://doi.org/10.1371/journal.pone.0201904.

Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R. (1988). *UCI Machine Learning Repository*. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/45/heart+disease.

Juszczak, P., Tax, D.M.J. and Duin, R.P.W. (2002). *Feature Scaling in Support Vector Data Description*. [online] Available at: https://www.researchgate.net/publication/2535451_Feature_Scaling_in_Support_Vector_Data_Description.

Lasai Barreñada, Dhiman, P., Timmerman, D., Anne-Laure Boulesteix and Calster, B.V. (2024). Understanding overfitting in random forest for probability estimation: a visualization and simulation study. *Diagnostic and Prognostic Research*, 8(1). doi:https://doi.org/10.1186/s41512-024-00177-1.

Lewis, C. (2025). *Heart disease remains leading cause of death as key health risk factors continue to rise*. [online] American Heart Association. Available at: https://newsroom.heart.org/news/heart-disease-remains-leading-cause-of-death-as-key-health-risk-factors-continue-to-rise.

Naser, M.A., Majeed, A.A., Alsabah, M., Al-Shaikhli, T.R. and Kaky, K.M. (2024). A Review of Machine Learning's Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. *Algorithms*, [online] 17(2), p.78. doi:https://doi.org/10.3390/a17020078.

Ritwik_B3 (2020). *Heart Disease Cleveland*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland.

scikit-learn.org. (n.d.). *3. Model selection and evaluation — scikit-learn 0.24.2 documentation*. [online] Available at: https://scikit-learn.org/stable/model_selection.html.

Sohn, S.Y., Kim, H.S. and Moon, T.H. (2007). Predicting the financial performance index of technology fund for SME using structural equation model. *Expert Systems with Applications*, 32(3), pp.890–898. doi:https://doi.org/10.1016/j.eswa.2006.01.036.

UCLA UCLA Electronic Theses and Dissertations Title Heart Disease Prediction Using Machine Learning Algorithms. (n.d.).

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, 12(4), p.e0174944. doi:https://doi.org/10.1371/journal.pone.0174944.

Ahmad, A.A. and Polat, H., 2023. Prediction of heart disease based on machine learning using Jellyfish Optimization Algorithm. *Diagnostics (Basel)*, 13(14), p.2392. doi:10.3390/diagnostics13142392.