Predicting Heart Disease
Using Machine Learning Classification Methods

Laura Carbaugh [*1] **Simon Tao** [*] **Atty Ehui** [*]

1. Introduction Laura Carbaugh Atty Ehui

Heart disease is one of the leading causes of death worldwide, making the early and accurate prediction of cardiac risk a critical public health goal to prevent fatalities. Using a data set from the UC Irvine Machine Learning Repository, we plan to explore factors that are involved in predicting whether an individual has heart disease or not. The UCI data set combines patient records from four different sources: Cleveland, Hungary, Switzerland, and VA Long Beach. Although this database contains 76 variables, previous research focuses on a subset of 14 variables that capture the most relevant information. By building and evaluating numerous models on this data, we aim to improve our understanding of various cardiovascular risk factors and illustrate how early intervention can provide better patient outcomes. Our study offers a chance to compare different modeling strategies in a real-world health context, highlighting how careful algorithm selection contributes to meaningful insights.

2. Variable Description Simon Tao

For this project, we are using a dataset from the UC Irvine Machine Learning Repository. The data has 303 observations across 14 different variables regarding the presence or absence of heart disease in a patient. Below is a description of each of the variables that will be used.

age Patient age in years.

sex Biological sex of the patient (1 = male, 0 = female).

cp Chest pain type: 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic.

trestbps Resting blood pressure (mm Hg).

chol Serum cholesterol (mg/dL).

fbs Indicator of fasting blood sugar > 120 mg/dL (1 = true, 0 = false).

restecg Resting electrocardiographic results: 0 = normal; 1 = ST–T wave abnormality; 2 = left ventricular hypertrophy.

thalach Maximum heart rate achieved during exercise.

exang Exercise-induced angina (1 = yes, 0 = no).

oldpeak ST depression induced by exercise relative to rest.

slope Slope of the peak exercise ST segment: 1 = upsloping; 2 = flat; 3 = downsloping.

ca Number of major coronary vessels (0–3) colored by fluoroscopy.

thal Thalassemia status: 3 = normal; 6 = fixed defect; 7 = reversible defect.

num Target variable: diagnosis of heart disease, originally coded 0–4 to reflect severity (often binarized to 0 vs. 1 for classification).

2.1. Key Variables Simon Tao

The variables include demographic factors, laboratory measurements, and clinical assessments which allows for a robust modeling of heart disease risk. While all of these variables are important for modeling, “cp,” “trestbps,” and “ca” are powerful predictors of heart disease. Other continuous variables such as “age,” “chol,” and “thal” are risk factors that can influence long-term disease development. “Sex” has also shown predictive value because men have a higher incidence of heart disease at earlier ages. Exercise-related measurements also play a key role. “Oldpeak,” which measures ST depression induced by exercise relative to rest, and “slope,” describing the shape of the ST segment during peak exercise, have strong diagnostic relevance.

3. Data Cleaning Laura Carbaugh

In order to build a reliable classification model, the data must be cleaned so that it is ready for testing and analysis. This data set does include missing values, particularly in variables such as “ca” which is the number of major vessels colored by fluoroscopy, and “thal,” which indicates thalassemia status. Missing values must either be handled through imputation methods, such as replacing them with the median for numerical features or the mode for categorical ones, or by removal. Because this dataset contains

303 observations, removal must be handled carefully as this can affect model performance if the sample becomes much smaller.

This database contains both numeric and categorical variables, which require different processing techniques. All numeric variables must be converted to integers to ensure that the model interprets these values correctly. Continuous variables such as "trestbps" (resting blood pressure) and "chol" (serum cholesterol) may need normalization to ensure that features measured on different scales do not dominate the algorithm. Categorical variables such as "cp" (chest pain type) and "restecg" (resting electrocardiographic results) must be encoded either using one-hot or ordinal encoding so that it is presented in a readable format for the model. The target variable "num" records the severity of heart disease on a scale from 0 to 4, but this variable may need to be transformed into a binary outcome by grouping all nonzero values as "disease present." After all necessary changes, the dataset must be inspected to confirm that no null values remain and that every feature is in a readable format for classification.

4. Preparation for Analysis

Atty Ehui

After cleaning and refining the data, the next step is to prepare it for model building and evaluation. The processed data will be randomly divided into training and testing subsets to allow for unbiased assessment of the model's capabilities. Any transformations done to the data will be fit on the training set and then applied to the test set in order to prevent data leakage that could harm the model. These measures ensure that the final classification models are trained on consistent, well-formatted data.

5. Exploratory Data Analysis

Simon Tao Laura Carbaugh

An exploratory analysis was conducted to better understand the structure and relationships of variables within the dataset before model building. Summary statistics and visualizations were used to examine distributions, detect potential outliers, and identify correlations between key variables. Exploring the relationship between these variables helps us determine which model techniques will be effective.

Preliminary exploration of the merged heart disease dataset revealed several important trends. The summary statistics show that the average patient age is approximately 54 years, and most participants have resting blood pressure values around 132 mmHg and serum cholesterol near 200 mg/dL. The standard deviations suggest moderate variability across most numeric features, with cholesterol and maximum heart rate (`thalach`) showing the greatest spread.

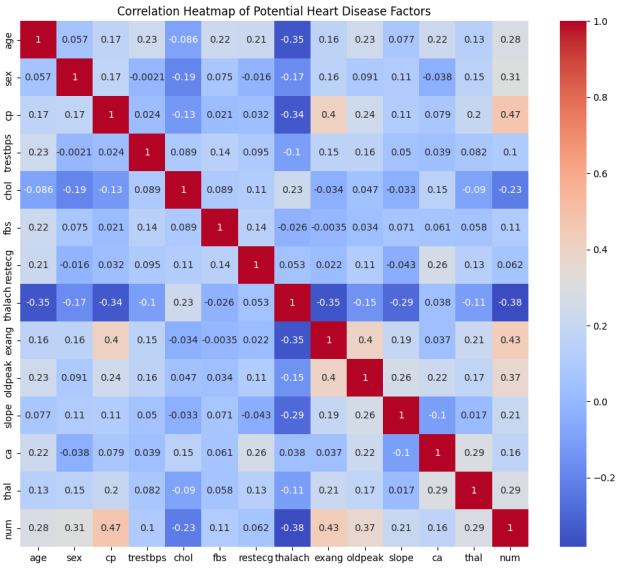

Figure 1. Correlation Heatmap of Continuous Predictors and Target Variable `num`.

The correlation heatmap highlights notable relationships between variables. Chest pain type (`cp`) and maximum heart rate (`thalach`) both display moderate positive correlations with the presence of heart disease, while exercise-induced angina (`exang`) and ST depression (`oldpeak`) are negatively correlated with heart health. These relationships suggest that exercise response variables and pain characteristics may serve as meaningful predictors in the classification models.

Density plots of numeric variables classified by disease status show that patients with heart disease tend to have slightly lower maximum heart rates and higher `oldpeak` values, indicating greater ST depression after exercise. These visualizations and summary tables help us identify which variables could be key factors in determining the presence of heart disease.

6. Methods

Simon Tao

The goal of this project is to build a model that will be able to classify whether or not an individual in the dataset has heart disease. This process will involve building and comparing multiple different models and evaluating their performance. The prediction of heart disease presence is based on the predictors in the dataset that were listed in section 2. In order to find a strong classification model for this data, we are going to explore a variety of machine learning models, such as different types of regression, k-nearest neighbors, and decision trees. These models are

Table 1. Summary statistics for continuous variables

Variable	Count	Mean	SD	Min	P25	P50	P75	Max
age	920	53.51	9.42	28.00	47.00	54.00	60.00	77.00
trestbps	920	132.00	18.45	0.00	120.00	130.00	140.00	200.00
chol	920	199.91	109.04	0.00	177.75	223.00	267.00	603.00
thalach	920	137.69	25.15	60.00	120.00	140.00	156.00	202.00
oldpeak	920	0.85	1.06	-2.60	0.00	0.50	1.50	6.20
ca	920	0.23	0.63	0.00	0.00	0.00	0.00	3.00

Table 2. Counts for categorical variables

Variable	Category	Count
sex	1	726
	0	194
fbs	0	782
	1	138
exang	0	583
	1	337
num	1	509
	0	411
cp	4	496
	3	204
	2	174
	1	46
restecg	0	553
	2	188
	1	179
slope	2	654
	1	203
	3	63
thal	3	682
	7	192
	6	46

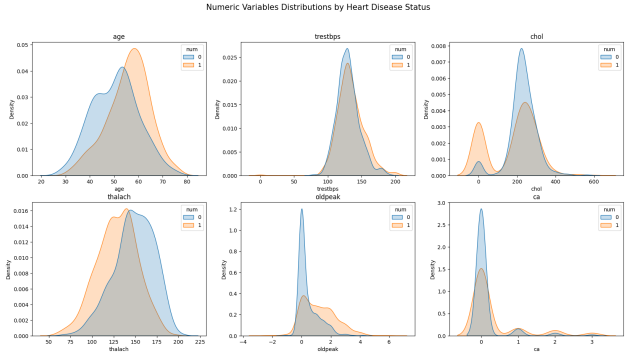

Figure 2. Distributions of Key Numeric Variables Classified By Heart Disease Status

appropriate because of their interpretability and predictive accuracy. The modeling process will involve integrating and cleaning the data, exploratory data analysis, model training, and validation and performance comparison of the different models.

6.1. Models Used and Justification Laura Carbaugh

There are many different methods that we could use to predict the presence of heart disease based on clinical and demographic factors. Logistic regression is a potential option due to its simplicity and interpretability. This method is appropriate because it estimates the probability that an observation belongs to the "disease" or "no disease" class, making it ideal for binary classification problems. The coefficients from the model are easily understandable, which is important in a healthcare context where outcomes must be transparent and explainable. Additionally, logistic regression allows for the identification of statistically significant predictors, making it useful for understanding which variables have the strongest association with heart disease risk. Another option that we would like to explore is k-nearest neighbors (kNN), in which observations are classified based on their nearest neighbors in the feature space. This specific method does not assume linearity, which makes it flexible for finding complex or irregular patterns within the data. KNN is also advantageous because it does not require a predefined model structure or assumptions about the data distribution. However, it can be sensitive to the choice of k and to differences in variable scales, so appropriate feature scaling and model tuning will be important steps during implementation.

The third technique we plan to explore is decision trees, which will allow us to examine non-linear relationships between the predictors and target variables. Decision trees split the data based on certain attributes, forming a series of decision rules that can be visualized in a hierarchical manner. This approach is particularly effective for identifying key variables that contribute the most to predicting heart disease, and it can handle both numeric and categorical data without requiring feature scaling. Decision trees also provide a level of interpretability that is valuable in medical analysis, as the

decision paths can be clearly traced and explained.
Finally, we will also evaluate using Random Forests as a more advanced model to help mitigate the overfitting that is often observed in single decision trees. By exploring these models, we aim to evaluate the trade-offs between interpretability, efficiency, and accuracy. Comparing their results will allow us to determine which method performs best on our dataset while still offering meaningful insights into the underlying factors that contribute to heart disease.

6.2. Model Training Procedure

Atty Ehui

Before beginning model training, the four datasets from Cleveland, Hungary, Switzerland, and VA Long Beach will be combined into a single dataset and cleaned for analysis. The combined dataset will then be divided into a training set and a testing set. We will use an 80/20 split so that the majority of the data is used to train the models while a smaller portion is used for evaluating performance on unseen data. This separation helps assess how well the models generalize beyond the data they were trained on. Additionally, to prevent data leakage, any transformations applied to the training data will be applied to the testing data as well.

During model training each of the described methods will be fitted on the training data. For logistic regression, regularization may be considered to prevent overfitting, which is useful given the potential multicollinearity among similar variables. For k-nearest neighbors, the optimal value of k will be determined using cross-validation to balance bias and variance. Decision trees will be trained with various splitting criteria to identify key patterns without becoming too complex. In addition, a Random Forest classifier will be explored as an ensemble-based extension of decision trees, allowing the model to capture nonlinear relationships more effectively while reducing overfitting through bootstrap aggregation.

6.3. Model Validation Plan

Laura Carbaugh

When assessing the validity and performance of the models, we plan to evaluate accuracy, the confusion matrix, and the AUC value. The accuracy score will measure the overall proportion of correct classifications across both classes. The confusion matrix, which includes sensitivity and specificity, will provide insight into how effectively each model distinguishes between individuals with and without heart disease. The Area Under the Curve (AUC) will assess the trade-off between sensitivity and specificity across different classification thresholds, offering a more complete view of discriminative ability. Together, these metrics will provide a well-rounded understanding of model performance and reliability. For the Random Forest model in particular, these evaluation measures will highlight how ensemble-based methods often improve sensitivity and overall accuracy by reducing the variance associated with individual decision trees.

6.4. Model Implementation

Atty Ehui

After cleaning and merging the data, the models will be implemented in Python using libraries such as scikit-learn, NumPy, and pandas. Each model will follow the same preprocessing steps to ensure a fair comparison across methods. For logistic regression, different regularization strengths and penalty types will be tested to prevent overfitting and identify the most influential predictors. For kNN, we will experiment with multiple k values using cross-validation while ensuring that all numerical features are properly scaled. For decision trees, model depth and splitting criteria will be tuned to balance predictive accuracy with interpretability, helping prevent overly complex trees that fail to generalize well. Random Forests will also be implemented by testing different numbers of decision trees and evaluating their combined performance, as ensemble methods often provide substantial improvements in stability and overall accuracy. All models will be evaluated using consistent metrics and cross-validation procedures to enable meaningful and transparent comparison.

7. Results

7.1. Overview

Atty Ehui

Following the exploratory data analysis, four supervised learning models were developed to predict the presence of heart disease: Logistic Regression, k-Nearest Neighbors, a Decision Tree classifier, and a Random Forest ensemble model. In order to select hyperparameters that generalize well to unseen data, each model underwent 5-fold cross-validation during the training phase. After tuning, all models were evaluated on the held-out 20% test set using accuracy, sensitivity, specificity, AUC, and confusion matrix interpretation. This section summarizes the final performance of each model and discusses the insights gained from their comparative behavior.

7.2. Hyperparameter Tuning

Simon Tao

Each model required selecting hyperparameters that appropriately balanced model flexibility, interpretability, and generalization performance. For Logistic Regression, eight combinations of penalty type (L1 or L2 regularization) and regularization strength ($C = 0.01, 0.1, 1, 10$) were evaluated. The L2-regularized model with $C = 0.01$ achieved the highest cross-validated AUC of 0.888, reflecting strong discrimination during training and was therefore selected as the final model.

For the kNN classifier, values of k ranging from 2 to 10 were tested. Validation accuracy steadily increased as k grew, and $k = 10$ produced the highest cross-validated accuracy of 0.814. This value was subsequently used in the final tuned model.

The Decision Tree model was tuned by exploring multiple maximum tree depths, including {None, 2, 4, 6, 8, 10}. A maximum depth of 4 yielded the best cross-validated AUC (0.827), indicating that moderate pruning helped reduce overfitting while preserving the model's ability to capture nonlinear relationships in the data.

Finally, a Random Forest model with 200 estimators was trained as an advanced model. Ensemble methods such as Random Forests reduce the variance of a single tree and typically achieve stronger and more stable predictive performance, making them a natural extension to the initial set of models.

7.3. Logistic Regression

Simon Tao

Logistic Regression exhibited consistently strong performance across evaluation metrics. After tuning, the model achieved a test accuracy of 0.788, indicating that it correctly classified nearly 79% of individuals. The confusion matrix revealed 85 true positives and 60 true negatives, with 19 false negatives and 20 false positives. The model's sensitivity of 0.817 suggests that it correctly identified more than 81% of heart disease cases, an important characteristic for medical prediction tasks in which missed diagnoses can have severe consequences. Its specificity of 0.750 shows that the model successfully identified three-fourths of individuals without heart disease. Logistic Regression also achieved an AUC of 0.872, one of the highest among the models, demonstrating strong ranking ability across various classification thresholds. Overall, Logistic Regression proved to be a well-balanced and reliable approach with strong discriminative performance and stability.

7.4. k-Nearest Neighbors

Simon Tao

The tuned kNN model, using $k = 10$ neighbors, achieved the highest accuracy among the baseline models, with a test accuracy of 0.793. This indicates that the model correctly classified nearly 80% of all cases. The confusion matrix showed 87 true positives and 59 true negatives, with 17 false negatives and 21 false positives. The sensitivity of 0.837 was slightly higher than that of Logistic Regression, suggesting that kNN captured subtle patterns in the data that improved its ability to detect positive cases. Specificity was 0.738, indicating a reasonable but slightly weaker performance in identifying negative cases. The model achieved an AUC of 0.849, demonstrating moderate ranking ability, although not as strong as the linear model. Overall, kNN performed competitively, offering strong discrete classification results, but its lower AUC indicates less stability in risk ranking across thresholds.

7.5. Decision Tree

Laura Carbaugh

The tuned Decision Tree model improved upon the untuned version but remained the weakest performer among the individual models. With a maximum depth of 4, the model achieved a test accuracy of 0.777, correctly identifying a substantial proportion of cases but still trailing the other models. The confusion matrix showed 89 true positives and 54 true negatives, along with 15 false negatives and 26 false positives. Its sensitivity of 0.856 indicates strong ability to detect heart disease, but this came at the expense of specificity, which dropped to 0.675. The AUC score of 0.841, while improved from the untuned tree, remained lower than those of Logistic Regression and the Random Forest. These results reflect the inherent limitations of single-tree models: although they can capture nonlinear relationships, they are prone to overfitting and may struggle to generalize as effectively as more robust methods.

7.6. Random Forest

Laura Carbaugh

The Random Forest model demonstrated the strongest overall performance across nearly all evaluation metrics. On the test set, it achieved an accuracy of 0.815, the highest among all models. The confusion matrix showed 90 true positives and 60 true negatives, with 14 false negatives and 20 false positives, leading to a sensitivity of 0.865 and a specificity of 0.750. This combination of high sensitivity and solid specificity highlights the model's ability to accurately detect heart disease cases while maintaining a reasonable rate of correct negative classifications. The Random Forest achieved an AUC of 0.870, comparable to Logistic Regression and reflective of strong discriminative capability across thresholds. These results illustrate the advantages of ensemble learning, particularly in capturing nonlinear patterns and reducing variance, making Random Forest a highly effective model for this prediction task.

7.7. Summary of Results

Atty Ehui

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.788	0.817	0.750	0.872
kNN (k=10)	0.793	0.837	0.738	0.849
Decision Tree (depth=4)	0.777	0.856	0.675	0.841
Random Forest	0.815	0.865	0.750	0.870

Table 3. Performance metrics for all tuned models.

Table 3 presents a comparison of the predictive performance of the four machine learning models after hyperparameter tuning. Logistic Regression and Random Forest achieved the strongest overall discriminative performance, with AUC values of 0.872 and 0.870, respectively. Random Forest

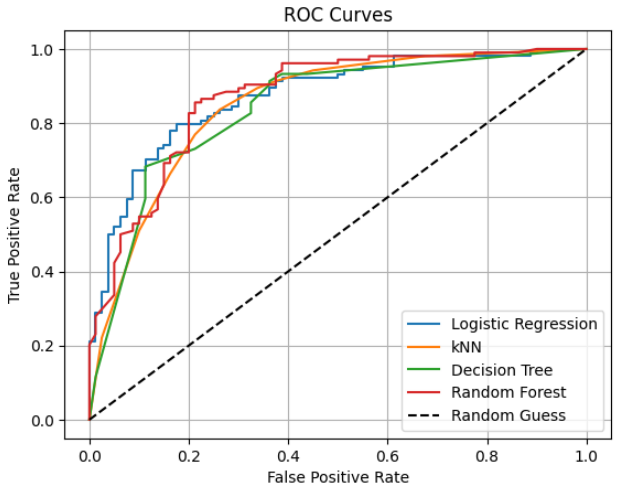

Figure 3. ROC Curves for Logistic Regression, kNN, Decision Tree, and Random Forest Models against a random guess baseline `num`.

produced the highest accuracy of 0.815 and sensitivity of 0.865, reflecting its ability to capture nonlinear interactions while maintaining low bias. The kNN model performed competitively, achieving the second-highest accuracy of 0.793, although its lower AUC indicates weaker ranking ability compared to the top-performing models. The Decision Tree, while improved after tuning, remained the least stable model due to its tendency to overfit. However, it achieved strong sensitivity of 0.856, showing that it correctly identified many positive heart disease cases. Overall, the table highlights that ensemble methods and regularized linear models provide the most balanced and reliable results for this dataset.

7.8. ROC Curve Comparison Simon Tao

The ROC curves provide an additional perspective on model performance by illustrating the full range of sensitivity–specificity trade-offs. Logistic Regression and Random Forest produced the strongest curves, both maintaining high true positive rates across a wide range of false positive rates. Their similar AUC values indicate comparable and consistently strong ranking performance. The kNN curve performed moderately well, showing that the model is effective at certain thresholds but less stable across the full range. The Decision Tree curve lay closest to the diagonal, reflecting its more limited ability to rank patients by predicted risk. Together, the ROC analysis reinforces that Logistic Regression and Random Forest offer the most robust and reliable discriminative performance among the models evaluated.

Beyond performance comparisons, the models provide insight into the factors most predictive of heart disease. From both exploratory analysis and model behavior, variables such as maximum heart rate achieved (thalach), ST-depression during exercise (oldpeak), and the number of major vessels visualized by fluoroscopy (ca) consistently emerged as the most influential predictors. Lower thalach values and higher oldpeak values were strongly associated with heart disease, consistent with reduced exercise capacity and ischemic response patterns. Similarly, higher ca values indicate structural arterial abnormalities, which align with known cardiovascular pathology. The consistency of these findings across models and visual analyses suggests that the predictive signals captured here reflect genuine clinical risk factors rather than artifacts of the modeling process.

7.9. Key Insights Into Heart Disease Prediction Atty Ehui

Across both the exploratory analysis and the final model evaluations, several predictors consistently emerged as important indicators of heart disease. Variables such as maximum heart rate achieved (thalach), ST-depression induced by exercise (oldpeak), and the number of major vessels visualized by fluoroscopy (ca) showed clear separation between individuals with and without heart disease and repeatedly influenced model behavior. These findings align with established clinical understanding: reduced exercise capacity, ischemic changes on an ECG, and observable coronary blockages are well-known markers of cardiovascular dysfunction. The strong agreement between clinical intuition and the model-derived insights enhances confidence in the validity and interpretability of the predictive patterns identified in this study.

8. Conclusion Simon Tao Laura Carbaugh

This project set out to build predictive models for identifying heart disease using a combined dataset from Cleveland, Hungary, Switzerland, and VA Long Beach. After completing data cleaning, exploratory analysis, and model development, several clear patterns emerged regarding both model performance and the clinical factors most associated with heart disease. Using hyperparameter tuning, 5-fold cross-validation, and multiple evaluation metrics allowed for a more rigorous comparison across models and strengthened the reliability of the findings.

Across all models, regularized Logistic Regression and Random Forests demonstrated the strongest and most stable performance. Random Forest achieved the highest accuracy and sensitivity by capturing nonlinear interactions and reducing overfitting, while Logistic Regression maintained excellent discriminative power and offered the most interpretable coefficients among the baseline models. k-Nearest Neighbors performed competitively, particularly in sensitivity, though its AUC suggested weaker ranking ability. The Decision Tree model, even when tuned, remained the least

consistent due to its susceptibility to overfitting. These comparisons illustrate that while simple models provide useful baselines, ensemble methods and regularized approaches tend to be more effective for complex medical prediction tasks.

In addition to evaluating algorithms, the analysis highlighted several important predictors of heart disease. Maximum heart rate achieved (thalach), exercise-induced ST depression (oldpeak), and the number of major vessels visualized (ca) consistently emerged as influential features across exploratory graphs and model behavior. These findings align well with established clinical understanding and enhance confidence in the validity of the predictive patterns observed.

While the project produced strong and interpretable results, there are limitations to acknowledge. The merged dataset originates from multiple hospitals and time periods, which may introduce variation in measurement practices or patient populations. The limited set of available clinical variables also restricts the complexity of relationships that models can detect. Additionally, although cross-validation and a held-out test set were used to evaluate performance, real-world deployment would require further considerations such as calibration, long-term model drift, fairness across demographic groups, and interpretability for clinical decision-making.

Another limitation arises from the structure of the original dataset. Each of the four contributing hospitals initially contained 303 observations, but due to missing values and differences in available variables, the final merged dataset used in this project contained 920 complete records rather than the full 1,212. While merging the datasets increased sample diversity and improved model stability, it also introduces potential imbalance across sites and highlights the fact that some information was lost through necessary cleaning and preprocessing. This constraint reflects the challenges of working with clinical data and suggests that future work incorporating more complete or modern datasets could further strengthen predictive performance.

Overall, this project demonstrates that carefully applied supervised learning methods, supported by statistical rigor and clinical insight, can be effective tools for predicting heart disease. The work provides a solid foundation for future extensions, which could include incorporating additional clinical features, testing more advanced ensemble or deep learning models, or evaluating the models within a broader clinical workflow.

References

Simon Tao

[1] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository: Heart Disease Data Set*. University of California, Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.