Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms: Insights from the UCI Irvine Public Heart Disease Datasets (Cleveland and Long Beach)

Jonathan Asanjarani

City University of New York Graduate Center

DATA 79000: Capstone Project and Thesis

Advisor: Johanna Devaney

Approval

Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms: Insights from the UCI Irvine Public Heart Disease Datasets (Cleveland and Long Beach)

By

Jonathan Assanjarani

This manuscript has been read and accepted for the Graduate Faculty in Digital Initiatives in satisfaction of the thesis requirement for the degree of Master of Science in Data Analysis and Visualizaton

Approved: January 2025

Professor Johanna Devaney

Director Graduate Center Digital Initiatives Matthew Gold

Abstract

For this capstone project, I investigated how well machine learning models can predict heart disease, while also studying how the patient's gender affects these predictions, as well as determining how well the same model performs across different regions. This project utilizes two clinical datasets from the publicly accessible UCI Machine Learning Repository under the collection "Heart Disease." This repository is composed of data from patients across four distinct locations, capturing varying levels of heart disease severity. These datasets are part of a collection of databases, open to the public, and can be used by anyone looking to conduct empirical analysis.

For this study, I specifically focused on two of the four datasets available: the "processed. cleveland.data," which contains patient records from Cleveland, and the "processed.va.data," which includes data from the Veterans Administration in Long Beach. These datasets provided my foundation for analyzing heart disease risk factors and evaluating predictive models. This project evaluated the predictive accuracy of machine learning algorithms and compared its performance to traditional cardiovascular risk scores, such as the American College of Cardiology's **Atherosclerotic Cardiovascular Disease (ASCVD).** Furthermore, I investigated the expectations of a machine learning models' ability to surpass the predictive accuracy of established ASCVD risk scoring. I also examined whether significant differences arose when evaluating the model's predictive performance for men and women, as well as when applying the same model to datasets from two different cities. This dual focus aims to uncover both gender-specific variations and location-based disparities in model effectiveness.

For my research, I applied the following machine learning algorithms: Random Forest, XGBoost, and an ensemble method combining them. These algorithms were applied to heart disease prediction datasets, using tailored preprocessing and feature engineering techniques.

Three distinct experiments were conducted, each labeled **Transformation 1**, **Transformation 2**, and **Transformation 3**. **Transformation 1** normalized skewed data and applied squared transformations to capture non-linear relationships. **Transformation 2** introduced engineered features such as the combination of ST depression (Oldpeak) with the slope of the ST segment. **Transformation 3** incorporated a custom feature to account for gender differences. Exploratory analysis revealed that chest pain type, ST depression (Oldpeak), and exercise-induced angina were key predictors, while cholesterol and fasting blood sugar contributed little predictive value. The best experiment combined multiple elements from all these transformations and significantly improved the model's performance and predictive accuracy.

The performance of the ASCVD risk score was also evaluated on both datasets to act as a benchmark against the machine learning models. For the Cleveland dataset, the ASCVD score achieved an accuracy of 69.64%, with a precision of 63.58%, recall of 79.14%, and an F1 score of 70.51%. Notably, the model performed better for females than males, with higher recall (88.89% vs. 77.19%) but lower precision (48.88% vs. 68.75%). This variability highlighted challenges in achieving consistent results for both men and women. On the VA Long Beach dataset, the ASCVD score achieved a higher accuracy of 76.82%, precision of 78.95%, and recall of 93.75%, yielding an F1 score of 85.71%. Due to an insufficient number of females in the VA Long Beach dataset, a sex-specific analysis was not feasible.

The Random Forest model utilizing **Transformation 2** demonstrated the best overall predictive accuracy and better predictive accuracy compared to the ASCVD score, achieving 83.33% accuracy for the Cleveland dataset and 82.5% for the Long Beach dataset. However, gender-based disparities emerged in the model's performance. The Random Forest Model performed better overall for the female subgroup in the Cleveland dataset, exhibiting a higher overall precision and recall, meaning that the model was better at making a correct prediction. However, the model exhibited a poor recall of 67% amongst patients with heart disease. This indicated that the model was not good at correctly predicting patients with heart disease. In contrast, the model had much more balanced predictive accuracy when evaluating the male subgroup.

The Transformation 2 experiment utilizing the Random Forest model performed well overall but demonstrated additional biases when evaluated on the Long Beach dataset. For example, in the Cleveland dataset, which had a relatively balanced distribution of heart disease and non-heart disease cases, the model was able to accurately identify both classes. However, in the VA Long Beach dataset, where heart disease cases were more prevalent, the model demonstrated a poor recall for the minority class in the dataset ("patients without heart disease") while showing high precision and recall for the majority class ("heart disease"). This tendency resulted in a recall of 97% for heart disease cases, indicating the model was highly effective at identifying individuals with heart disease. The precision for heart disease was 83%, meaning 17% of those classified as having heart disease were false positives.

For non-heart disease cases, the model's recall was only 40%, highlighting its inability to correctly identify most individuals without heart disease. This trade-off—favoring high recall for the majority class at the expense of accurately identifying minority class cases—illustrates the

challenge of generalizing performance across datasets which include patients from different regions. The class imbalance only further reduces the ability to generalize these results.

Findings suggest that while my machine learning models hold promise for improving heart disease prediction, the gender imbalances and regional variations in these datasets limit the general utility of these findings. Due to the class imbalance, future research that utilizes these datasets should explore multilabel classification for varying disease severities and address sampling biases that arise from an overall low representation of women in these datasets. This will enhance both fairness and the model's application across all populations.

Contents

Tables

Digital manifest

1. Capstone Report (Print and Digital)

- **File Name:** Final_Write_up1.27.25.pdf
- **File Type:** PDF
- **Description:** Full written report detailing research objectives, methodology, results, and discussions.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/Final_Write_up1.27.25.pdf

2. Exploratory Data Analysis (EDA) Notebook

- **File Name:** EDA_4_binary_classification.ipynb
- **File Type:** Google Collab Notebook (.ipynb)
- **Description:** Python notebook detailing data cleaning, univariate, bivariate, and multivariate analyses, including visualization and statistical tests.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/EDA_4_binary_classification.ipynb

3. Machine Learning Model Implementation for Cleveland

- **File Name:** ML_Algo_4_binary_classification.ipynb
- **File Type:** Google Collab Notebook (.ipynb)
- **Description:** Google Collab Notebook containing code for implementing and evaluating machine learning models (Random Forest, XGBoost, and ensemble methods) using the Cleveland dataset.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ML_Algo_4_binary_classification.ipynb

4. Machine Learning Model Implementation for VA Long Beach

- **File Name:** ML_Algo_4_bin_classification_va_longbeach.ipynb
- **File Type:** Google Collab Notebook (.ipynb)
- **Description:** Google Collab Notebook containing code for implementing and evaluating machine learning models (Random Forest, XGBoost, and ensemble methods) using the VA Long Beach dataset.

- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ML_Algo_4_bin_classification_va_longbeach.ipynb

5. Cleveland Processed Dataset

- **File Name:** processed.cleveland.data
- **File Type:** ZIP archive (contains .data files)
- **Description:** Includes cleaned and transformed versions of the Cleveland dataset used in the study.
- **URL:** https://github.com/Jdasanja/masters_thesis/blob/main/processed.cleveland.data

6. VA Long Beach Processed Datasets

- **File Name:** processed.va.data
- **File Type:** ZIP archive (contains .data files)
- **Description:** Includes cleaned and transformed versions of the VA Long Beach dataset used in the study.
- **URL:** https://github.com/Jdasanja/masters_thesis/blob/main/processed.va.data

7. Data Transformation Script Cleveland

- **File Name:** ML_Algo_4_binary_classification.ipynb
- **File Type:** Google Collab Notebook (.ipynb)
- **Description:** Custom Python scripts for data preprocessing and feature engineering, including transformations applied to Cleveland dataset.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ML_Algo_4_binary_classification.ipynb

8. Data Transformation Script VA Long Beach

- **File Name:** ML_Algo_4_bin_classification_va_longbeach.ipynb
- **File Type:** Google Collab Notebook (.ipynb)
- **Description:** Custom Python scripts for data preprocessing and feature engineering, including transformations applied to VA Long Beach dataset.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ML_Algo_4_bin_classification_va_longbeach.ipynb

9. ASCVD Risk Score Implementation Cleveland

- **File Name:** ACSVD_calculation_of_Cleveland.ipynb
- **File Type:** Jupyter Notebook (.ipynb)
- **Description:** Python notebook implementing the ASCVD Risk Calculator for the Cleveland dataset.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ACSVD_calculation_of_Cleveland.ipynb

10. ASCVD Risk Score Implementation VA Long Beach

- **File Name:** ACSDV_Calculation_4_va_longbeach.ipynb
- **File Type:** Jupyter Notebook (.ipynb)
- **Description:** Python notebook implementing the ASCVD Risk Calculator for the Cleveland dataset.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/ACSDV_Calculation_4_va_longbeach.ipynb

11. A Note on Technical Specifications

- **File Name:** A Note on Technical Specifications.pdf
- **File Type:** PDF
- **Description:** PDF that provides an overview of the project's development environment, data sources, processing methods, file formats, version control, and external tools used to ensure reproducibility and transparency.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/A%20Note%20on%20Technical%20Specifications.pdf

12. Data Dictionary

- **File Name:** Data Dictionary.pdf
- **File Type:** PDF
- **Description:** PDF that outlines key variables, transformations, critical functions, and classifiers used in the project, providing detailed descriptions to ensure clarity and reproducibility.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/Data%20Dictionary.pdf

13. Digital References

- **File Name:** Digital References.pdf
- **File Type:** PDF
- **Description:** PDF that provides detailed citations for all software, tools, datasets, and external resources used in the project, ensuring transparency and enabling reproducibility.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/Digital%20References.pdf

14. Data Management Plan

- **File Name:** Data Management Plan Overview.pdf
- **File Type:** PDF
- **Description:** Comprehensive plan outlining data handling, storage, and ethical considerations.
- **URL:** https://github.com/Jdasanja/masters_thesis_final/blob/main/Data%20Management%20Plan%20Overview.pdf

A note on technical specifications

This project used Google Collab as the development environment. Google Collab is a cloud-based Python platform providing access to GPUs for accelerated computation. Python (version 3.8) was used in the Google Collab environment, with additional libraries and frameworks included, such as Scikit-learn, XGBoost, Pandas, NumPy, Matplotlib, and Seaborn, as detailed in the References section. The dataset sources used were gathered from the the UCI Machine Learning Repository from the “Heart Disease” database. Two datasets from this database were used; Cleveland and VA Long Beach datasets. Data cleaning and preprocessing were conducted within Google Colab Notebooks using Python-based libraries, with datasets and code files stored in CSV, Python (.py), and Jupyter Notebook (.ipynb) formats.

Version control was maintained through a GitHub repository that hosted the project’s source code, processed datasets, and supplementary materials. The repository, accessible at [https://github.com/Jdasanja/masters_thesis_final], was updated regularly with a detailed commit history to ensure reproducibility. External tools included the ASCVD Risk Calculator, implemented via an open-source Python package available at [https://github.com/brandones/ascvd/tree/master].

Sex-specific and regional analysis of heart disease prediction using machine learning algorithms: insights from the UCI Irvine public heart disease datasets (Cleveland and Long Beach)

Introduction

The central focus of my capstone project is to explore the effectiveness of machine learning models in predicting heart disease and assess its ability to generalize across different cities and biological sexes. This research highlights the importance of building models that not only achieve high accuracy within a specific dataset or geographic location but are also reliable when applied to both men and women separately, as well as when applied to more than one city in the United States. Traditional cardiovascular risk scores, while reliable, exhibit limitations that impact its performance (Talha, Elkhoudri, and Hilali, 2024). Machine learning models have been shown, in previous research, to have better predictive accuracy than some cardiovascular risk scores (Cho, Kim, Kang, et al., 2021).

Studies have shown that traditional cardiovascular risk models frequently lack validation across diverse populations, leading to miscalculations and reduced sensitivity when applied to groups outside their original development context (Talha, Elkhoudri, and Hilali, 2024). This inability to generalize poses a challenge when providing accurate predictions for underrepresented groups, particularly women. Cardiovascular disease is the leading cause of death globally, responsible for 17.7 million deaths in 2015—a number expected to rise to over 23.6 million annually by 2030. Despite its prevalence, women tend to be undertreated, with their symptoms frequently misdiagnosed or dismissed as non-cardiac issues (Woodward, 2019).

This disparity underscores the urgent need to develop more inclusive and accurate predictive models. Women's risk factors are often underestimated (Abdullah, Beckett, Wilson, et al., 2024). To address these challenges, my project uses machine learning models trained on data from both men and women, with performance tested separately for each gender. By addressing the limitations of traditional risk scores and accounting for differences in predictive accuracy for each gender, my project aims to more wholistically evaluate the accuracy and fairness of heart disease prediction. This is shown in my project when I compared the accuracy of the heart disease machine learning models to the ACSVD risk scores. This evaluation can help contribute to more equitable healthcare outcomes, by acting as an additional point of consideration.

Literature review

Heart disease prediction methods (risk models)

Over the past three decades, numerous risk prediction models have been developed to estimate an individual's likelihood of developing cardiovascular disease (CVD). Among these, the multivariate risk prediction model from the Framingham Study has been particularly influential in estimating future CVD risk (Cui, 2009). Additional models have been created in the United States, such as the Reynolds Risk Score for women, derived from data collected in the

Women's Health Study. Other efforts include a "multi-marker" risk model that incorporates 10 genetic markers, including C-reactive protein and B-type natriuretic peptide (Cui, 2009).

In Europe, separate cardiovascular risk prediction models were developed due to the limited applicability of the Framingham risk scores to the general European population without

recalibration (Cui, 2009). For example, the European Society of Cardiology developed the Systematic Coronary risk evaluation, the SCORE equation, endorsed by the Third Joint

European Task Force on cardiovascular prevention, which has been validated in Spain. In the UK, the Questo Risk Assessment Tool (QRISK) algorithm was created using a population-based clinical research database. Germany contributed both a simple Prospective **Cardiovascular** Münster (PROCAM) score and a more complex neural network model, with the PROCAM score recently updated. Dundee University in Scotland developed the Scottish Intercollegiate Guidelines Network (ASSIGN) risk score, which includes family history of CVD, based on data from the Scottish Heart Health Extended Cohort. Italy introduced the CUORE equation tailored for populations with a low incidence of coronary events (Cui, 2009). These diverse models reflect efforts to address regional differences in CVD risk and provide tailored tools for prevention.

Major limitations of cardiovascular risk scores

A study conducted by Talha, Elkhoudri, and Hilali, in 2024, summarized the best-known limitations of current cardiovascular risk models. Critical analysis revealed numerous limitations that impact performance. Each calculator demonstrates distinct advantages for one population while potentially encountering limitations with another. Some scores lack validation from external cohorts, while others seem to miscalculate risk when applied to populations outside of its origin, limiting its sensitivity, and being unable to explain all cardiac events (Talha, Elkhoudri, and Hilali, 2024).

Numerous cardiovascular risk assessment tools have been developed from large population studies, but only a few have undergone essential external validation. The most

common models, American and European scores, have distinct characteristics and limitations. Understanding these limitations is crucial for improving the effectiveness of these tools (Talha, Elkhoudri, and Hilali, 2024).

Comparisons of established risk prediction models for cardiovascular disease

A study, reviewing 74 previous research articles, aimed to evaluate and compare established cardiovascular risk prediction models to assess their performance. Investigators evaluated the performance of two or more risk prediction models in the same populations. The study extracted information on design, assessed models, and outcomes, examining their performance in terms of discrimination, calibration, and reclassification, while also considering biases favoring newer or author-developed models (Siontis, Tzoulaki, Siontis, et al., 2012).

The review included 74 articles, covering 56 pairwise comparisons of eight models, such as two variants of the Framingham risk score, ASSIGN score, SCORE, PROCAM score, QRISK1 and QRISK2 algorithms, and the Reynolds risk score. Only 10 of the 56 comparisons showed more than a 5% relative difference in predictive performance based on the area under the receiver operating characteristic curve (AUC). This means that most models had similar discriminatory abilities, as most comparisons did not exceed the 5% threshold. This suggests that these risk prediction models perform comparably in distinguishing between individuals at high and low risk of cardiovascular events. The use of other statistical measures like discrimination, calibration, and reclassification was inconsistent (Siontis, Tzoulaki, Siontis, et al., 2012). Outcome selection bias was evident in 32 comparisons, where 78% of the time, the model originally developed using the selected outcome performed better. Additionally, authors tended

to report better AUCs for models they developed, highlighting potential optimism bias (Siontis, Tzoulaki, Siontis, et al., 2012).

The conclusions suggest that while multiple cardiovascular risk prediction models exist, their comparisons would benefit from standardized reporting and consistent statistical evaluation. The reporting/evaluation of these risk models appear to be impacted by outcome selection and optimism biases, emphasizing the need for more rigorous and unbiased comparisons (Siontis, Tzoulaki, Siontis, et al., 2012).

Gender bias in assessing cardiovascular disease

Cardiovascular disease (CVD) rates are higher in males, which has led to it being seen as primarily a men's issue. However, CVD is the leading cause of death and a major cause of disability for women globally. Women are often under-recognized and undertreated for CVD compared to men, and their symptoms can differ, leading to worse outcomes. Female patients treated by male cardiologists fare worse than male patients, while no such difference exists for female cardiologists. Clinical trials often focus on men, despite some drugs having different effects in women. Risk factors like diabetes and smoking increase CVD risk more in women, and factors related to pregnancy and reproductive health add to their vulnerability. Women's health research is often focused on mother and child health and breast cancer, neglecting CVD and other non-communicable diseases. There is a need to broaden the definition of women's health to include the entire lifecycle and emphasize CVD, with sex-specific research analyses becoming the standard (Abdullah, Beckett, Wilson, et al., 2024).

Additionally, a study reviewed the evidence on gender bias in CVD diagnosis, prevention, and treatment. Following Preferred Reporting Items for Systematic Reviews and

Meta-Analyses (PRISMA) guidelines, several databases from 19 studies were searched and analyzed. The findings showed that CVD is less reported in women, who often have milder symptoms or are misdiagnosed with gastrointestinal or anxiety issues. As a result, women's risk factors are often overlooked, especially by male doctors. Women are given fewer diagnostic tests and are less likely to be referred to cardiologists or hospitalized. Even when hospitalized, women receive fewer coronary interventions and are prescribed fewer cardiovascular medications, except for antihypertensive and anti-anginal drugs. Women also tend to perceive themselves at lower risk for CVD than men. This review highlights that women receive fewer diagnostic tests and treatments for CVD, which affects their health outcomes, likely due to a lack of awareness about gender differences in CVD symptoms (Abdullah, Beckett, Wilson, et al., 2024).

Most of this research is based on a gender binary frame, which assumes the existence of only two distinct and opposite genders—male and female—often neglecting the experiences of non-binary and gender-diverse individuals. This binary approach reinforces systemic gaps in understanding the intersection of gender and health outcomes, as it fails to account for how non-binary individuals experience, report, and are treated for CVD. The focus on a binary framework not only limits the inclusivity of cardiovascular research but also perpetuates disparities by oversimplifying the complex interplay of biological sex and gender identity. Expanding research to include non-binary and gender-diverse populations is crucial to developing a more comprehensive understanding of cardiovascular health and ensuring equitable healthcare practices for all individuals.

Pre-existing machine learning methods for predicting cardiovascular disease

Using machine learning methods to predict cardiovascular disease has been an ongoing point of research in the last decade. The following study focused on improving risk prediction using machine learning on healthcare data from 222,998 Korean adults aged 40-79 without prior cardiovascular disease or lipid-lowering therapy. Traditional risk models showed moderate to good performance (C-statistics 0.70–0.80), with the pooled cohort equation (PCE) achieving a C-statistic of 0.738 (Cho, Kim, Kang, et al., 2021). Among various machine learning models tested, the neural network model performed best, with a C-statistic of 0.751, which was higher than PCE. It also showed better agreement between predicted and actual outcomes. Improvements were noted compared to other models like the Framingham risk score, systematic coronary risk evaluation, and QRISK3 (Cho, Kim, Kang, et al., 2021). The study concluded that machine learning algorithms could enhance cardiovascular risk prediction beyond existing models, making them valuable tools for risk assessment and clinical decision-making in healthy Korean adults (Cho, Kim, Kang, et al., 2021).

Additionally, in a study by Stephen F. Weng and colleagues, machine learning was assessed for improving cardiovascular risk prediction using data from 378,256 UK patients. Four algorithms (random forest, logistic regression, gradient boosting, neural networks) were compared to the American College of Cardiology guidelines for evaluating CVD risk. The best-performing algorithm, neural networks, had an AUC of 0.764, improving prediction accuracy by 3.6% over the established method. This approach identified more patients who could benefit from preventive treatment and reduced unnecessary interventions (Weng, Reps, Kai, et al., 2017).

Gender-based approach for diagnosing coronary heart disease

In 2019, Hogo published an article titled “A proposed gender-based approach for diagnosis of coronary artery disease.” In this research article, two separate and individual models were trained and evaluated for each gender to determine whether the patient’s gender affects the structure and performance of a diagnosis model for coronary artery disease. The male diagnosis model achieved an accuracy of 95%, with a sensitivity of 96% and a specificity of 100%, while the female diagnosis model performed slightly better, with an accuracy of 96%, a sensitivity of 97%, and a specificity of 96%. The high-performance results overall highlight the success of the proposed gender-based approach for diagnosing coronary artery disease. The dataset used for this project is from the UCI Machine Learning Repository; specifically, the "Heart Disease Database" and the "Z-Alizadeh Sani Dataset," which comprises records for 270 patients, each with 75 attributes (Hogo, 2020).

Supervised machine learning

My project uses multiple supervised machine learning algorithms on a clinical dataset to predict the presence of cardiovascular disease in patients. Supervised learning is a type of machine learning where the algorithm is trained on labeled data to make predictions or decisions (Mueller and Guido, 2016). It learns to map input data to the correct output. For this project, I conducted a classification task to predict whether a patient has heart disease. Classification is one of the two primary types of supervised learning problems in machine learning (Mueller and Guido, 2016). Specifically, this project focused on binary classification, which involves distinguishing between two distinct classes (Mueller and Guido, 2016). The results of different machine learning models, such as a Random Forest Classifier, and XGB Classifier, were

compared with cross-validation, which was used to ensure the reliability of the performance estimates.

A random forest classifier is a type of ensemble method that combines multiple decision trees to create a more powerful model. Decision trees are a type of diagram/logic that is widely used for classification tasks. Each decision tree is based off a hierarchy of "if/else" questions, leading to a decision tree. In a random forest model, each decision tree is slightly different from the others. The concept behind random forests is that while each individual tree can provide reasonably accurate predictions, it is prone to overfitting specific portions of the data. One decision tree might be overfit in one way, and another might be overfit in another. By averaging their results, one can retain the predictive power of decision trees while reducing overfitting. Random forests incorporate two levels of randomness: first, by selecting a random subset of data points to construct each tree, and second, by randomly selecting a subset of features to evaluate at each split (Mueller and Guido, 2016).

XGBoost, which stands for "Extreme Gradient Boosting", is a highly efficient and scalable library for gradient boosting, specifically designed to optimize the training process of machine learning models. The XGBoost classifier is a gradient boost regression tree, which is another ensemble method that combines multiple decision trees to create a more powerful model. Unlike the random forest method, gradient boosting constructs trees sequentially, with each tree aiming to address the errors made by the preceding one (Mueller and Guido, 2016). The gradient boos trees are typically shallow, with a maximum depth of 1 to 5. These types of models would be considered "weak learners", as they only capture a small portion of the data's complexity. The main idea is to combine many of these models. Each tree can only provide good predictions on

part of the data, and so more and more trees are added iteratively to improve performance (Mueller and Guido, 2016).

Gradient boosting models are generally more sensitive to parameter settings compared to random forest classifiers. This increased sensitivity means that the performance of gradient boosting can vary significantly depending on how parameters are tuned. However, when parameters are properly optimized, gradient boosting has the potential to achieve higher accuracy than random forests (Mueller and Guido, 2016).

Machine learning models were evaluated by calculating the F1-score, recall score, accuracy score, and precision score. The precision score measures how many predicted positive instances were correct. Accuracy measures the overall correctness of the model. The recall score focuses on how well a model evaluated negative prediction. For this dataset, a model is considered to have a negative prediction if a patient is not diagnosed at all with cardiovascular disease. The F1-score combines precision and recall into a single number, balancing the trade-offs. It is the average of precision and recall, with a greater emphasis on the smaller value. The score goes from 0 to 100%, with higher values indicating better performance. It is particularly useful for imbalanced datasets because it considers both false positives and false negatives (Mueller and Guido, 2016).

Materials and Methods

Data

The University of California Irvine (UCI)Machine Learning Repository is a comprehensive resource that provides databases, domain theories, and data generators widely utilized by the machine learning community for evaluating models. For this project, I utilized the database titled “Heart Disease” available in the UCI machine Learning Repository. The “Heart Disease” database from the UCI Machine Learning Repository is a resource that can be used for evaluating models that classify cardiovascular disease. The database includes four distinct datasets, each comprised of anonymized patient records with various social and biological indicators—such as age, exercise-induced angina, and systolic blood pressure—alongside a target column that specifies whether the patient has heart disease and, if so, its severity.

There are four datasets available from four different regions: Cleveland, Hungary, Switzerland, and the VA Long Beach. There are 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Most experiments have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0) (Janosi, Steinbrunn, Pfisterer, et al., 1988). That is what I did for this project. The datasets donated by UC Irvine consists of 76 attributes. However, the processed datasets that I used consisted of a subset of 13 of them. For the purposes of this project, I used the Cleveland dataset, as well as the VA Long Beach Dataset.

The processed Cleveland dataset includes **303 patient records,** each with varying levels of heart disease severity. An additional dataset, consisting of patients from the United States was contributed to the database by the Veterans Administration of Long Beach, California. The VA Long Beach Healthcare System, formerly known as Naval Hospital Long Beach, encompasses a network of Veterans Administration facilities in Long Beach and nearby cities. This data set includes **200 patient records**, most of whom have some degree of heart disease, and features the same 14 columns as the Cleveland dataset.

In both datasets, the presence of heart disease is indicated by values ranging from 1 to 4, corresponding to increasing severity, while a value of 0 represents the absence of the condition. Consistent with prior research on the Cleveland dataset, this study concentrates on distinguishing between the presence (values 1–4) and absence (value 0) of heart disease. The primary focus of my project was on analyzing binary classification.

This dataset provides clinical and demographic information about patients, focusing on attributes associated with heart disease risk. Key variables include age (in years) and sex (0 for female, 1 for male), along with cp (chest pain severity) categorized on a scale of 0 to 4. Diagnostic measurements include restbps (resting blood pressure in mmHg), chol. (serum cholesterol level in mg/dL), and fbs (fasting blood sugar), a binary variable indicating whether fasting blood sugar exceeds 120 mg/dL. Additionally, the dataset captures restecg (resting electrocardiographic results) with categories ranging from normal results to indications of left ventricular hypertrophy, and Thalach, which measures the maximum heart rate achieved.

Other features assess cardiovascular responses to stress, such as exang (exercise-induced angina, binary), oldpeak (ST segment depression on ECG during exercise), and slope, describing the pattern of the ST segment during peak exercise (upsloping, flat, or down sloping). The ca

variable indicates the number of major vessels (0–3) visualized through fluoroscopy, and thal (a thalassemia-related variable indicating heart stress test results) which is also associated with increased heart disease severity. As a side note, this thalassemia related variable seems to become more pronounced as age increases. Each patient's thallium stress test results were categorized as normal, fixed defect, or reversible defect. The target variable, num, ranges from 0 to 4, representing the presence and severity of heart disease. This dataset is comprehensive, combining demographic and clinical variables, making it suitable for exploring the factors influencing heart disease outcomes. To ensure privacy, patient names and social security numbers were removed from the database and replaced with de-identified data.

Exploratory data analysis results

Basic descriptives of the Cleveland training set

The training set for the Cleveland dataset provides insights into the demographic and clinical characteristics of individuals included in the study. The average age of participants is approximately 54.77 years, with ages ranging from 34 to 77 years and an interquartile range (IQR) of 48 to 62 years. The dataset has a higher proportion of males, with the mean value for sex coded as 0.675 (1 for males and 0 for females). For chest pain type (cp), the average value is 3.186, with a range of 1 to 4, reflecting a balanced range of values for classifications of chest pain. Resting blood pressure (trestbps) has a mean of 132.27 mmHg, ranging from 94 to 200 mmHg, with an IQR of 120 to 140 mmHg. The mean cholesterol level (chol) is 249.41 mg/dL, spanning from 126 to 564 mg/dL, and an IQR of 212 to 277 mg/dL.

Fasting blood sugar (fbs) shows a mean of 0.169, indicating a low prevalence of elevated fasting blood sugar (coded as 1 for high and 0 for normal). Maximum heart rate achieved

(thalach) averages 149.84 bpm, with a range of 88 to 195 bpm and an IQR of 136 to 166 bpm. Exercise-induced angina (exang) is less frequent, with a mean of 0.346 (coded as 1 for yes and 0 for no). The mean ST depression induced by exercise (oldpeak) is 1.062, ranging from 0 to 6.2, with an IQR of 0 to 1.8. Finally, the slope of the peak exercise ST segment (slope) has an average value of 1.586, with values ranging from 1 to 3, representing a balanced range of values for slope categories. These descriptive statistics highlight key trends and distributions in the dataset.

Table 1

Basic Descriptives of the Cleveland training data

	count	mean	std	min	25%	50%	75%	max
age	237.000	54.772	9.032	34.000	48.000	56.000	62.000	77.000
sex	237.000	0.675	0.469	0.000	0.000	1.000	1.000	1.000
cp	237.000	3.186	0.956	1.000	3.000	3.000	4.000	4.000
trestbps	237.000	132.270	17.916	94.000	120.000	130.000	140.000	200.000
chol	237.000	249.409	53.162	126.000	212.000	243.000	277.000	564.000
fbs	237.000	0.169	0.375	0.000	0.000	0.000	0.000	1.000
restecg	237.000	0.983	0.996	0.000	0.000	0.000	2.000	2.000
thalach	237.000	149.844	22.398	88.000	136.000	152.000	166.000	195.000
exang	237.000	0.346	0.477	0.000	0.000	0.000	1.000	1.000
oldpeak	237.000	1.062	1.180	0.000	0.000	0.800	1.800	6.200
slope	237.000	1.586	0.616	1.000	1.000	2.000	2.000	3.000

Univariate analysis of the Cleveland training set

The dataset reveals patterns about the participants and their health indicators. Most participants middle-aged, fall between 55 and 65 years old, with males having a majority within the dataset (see Table 1). When it comes to those experiencing chest pain, this had the highest severity (type 4). Both resting blood pressure and cholesterol levels show right-skewed distributions, indicating that while most values are moderate, some individuals have significantly higher levels. Only a small proportion of participants had elevated fasting blood sugar, suggesting that this issue is less common in the dataset. The resting electrocardiogram, the results appear to be normally distributed, showing a balanced spread of values.

The ‘exang’ is a variable that indicates the presence of an exercise-induced angina (1 for presence, 0 for absence), and is inferred from the low mean (see Table 1). The variable reveals that most participants do not experience angina during exercise and suggests that exercise-induced chest pain is less common in this dataset. However, the subset of individuals with a positive value for exang may represent a higher risk for underlying heart issues. Angina during exercise is often a significant indicator leading to coronary artery disease. (see Table 2) (see Table 3).

The distribution of values for the maximum heart rate feature has a strong left skew. Based on this left skew, we

see most participants have a maximum heart rate above 140 beats per minute. The old peak feature, which measures changes in the ST segment of an ECG during exercise, indicates that most participants have values between 0 and 2, reflecting minimal ST depression and better heart health. However, there is a smaller group with higher values (up to 6).

The variable labeled ‘slope’, which describes the shape of the ST segment during peak exercise, shows that most participants have a flat slope (category 2). A smaller number have an upsloping slope (category 1), or a down sloping slope (category 3).

Relationships between distinct variables

A series of scatter points were created to understand the relationship between the distinct variable’s characteristics. The diagonals on the pair plot contain histograms depicting the distribution of specific features. The off-diagonal scatter plots show correlations between pairs of variables. For instance, age and cholesterol show a clear trend where greater cholesterol levels are related directly with older individuals.

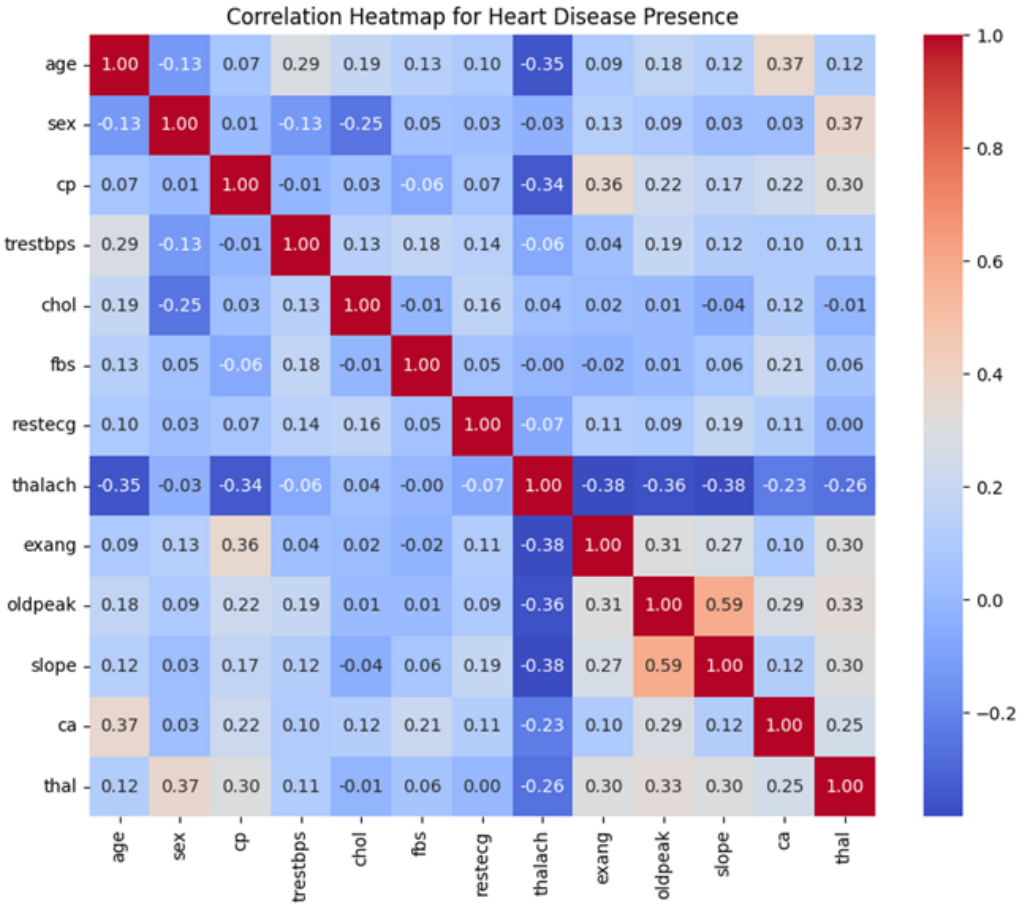
The pair plot shows value distributions and feature relationships in the Cleveland training dataset. Age and maximum heart rate (thalach) exhibit a mild negative correlation, while Oldpeak (ST depression) and Slope display a stronger linear relationship, supporting their combination in feature engineering. Cholesterol (chol) and resting blood pressure (trestbps) have right-skewed distributions with noticeable outliers.

A correlation heatmap was utilized to identify and emphasize strong relationships among variables. Notable findings include a strong positive correlation between the slope of the ST segment and Oldpeak ($r = 0.59$), indicating a close association between these features. Age and maximum heart rate (thalach) show a moderate negative correlation ($r = -0.35$), suggesting that older individuals tend to have lower maximum heart rates. Additionally, chest pain type (cp) exhibits a moderate positive correlation with exercise-induced angina (exang, $r = 0.36$), highlighting a link between specific chest pain types and exercise-related symptoms. Mild correlations were also observed between sex and thalassemia (thal, $r = 0.37$), as well as age and

the number of major vessels (ca, r = 0.37). Conversely, thalach displays weak negative correlations with several variables, including slope (r = -0.38) and Oldpeak (r = -0.36). These insights, highlighted in Figure 1, reveal key variable interactions that can inform the development of predictive models for heart disease.

Figure 1

Correlation amongst individual variables



Distribution of heart disease presence in the training data.

A bar chart was used to illustrate the distribution for binary classification of heart disease, distinguishing between presence (values 1–4) and absence (value 0).

The distribution is slightly imbalanced, with more individuals in the "no heart disease" category compared to the "heart disease" category. In the training set, 128 patients did not have heart disease while there were 109 patients that indicated some level of presence for heart disease. This imbalance, while not extreme, could influence model performance. It is important to note that in a population-representative sample, most people are more likely to be free of heart disease than to have it.

Analyzing individual variable association with heart disease presence.

The analysis highlights different relationships between each feature and the presence of heart disease. A weak but statistically significant positive correlation was observed between age and heart disease, with older individuals modestly more likely to have heart disease (Point-Biserial Correlation Coefficient: 0.1990, p-value: 0.002085). Gender-based analysis revealed some notable patterns, with males more likely to have heart disease and females predominantly without it. This association was statistically significant (p-value: 3.26e-04). Similarly, higher chest pain types showed a strong relationship with heart disease presence. Particularly type 4 (asymptomatic chest pain), was more common among individuals with heart disease (Chi-Square test statistic: 65.29635, p-value: 4.33e-14). (see Table 2) (see Table 3)

Resting blood pressure was significantly higher among individuals with heart disease and exhibited greater variability (Point-Biserial Correlation Coefficient: 0.1632, p-value: 0.01188). Comparatively, cholesterol levels were slightly higher on average among those with heart

disease, however, the relationship was not statistically significant (Point-Biserial Correlation Coefficient: 0.0661, p-value: 0.31068). Fasting blood sugar levels showed no meaningful association with heart disease. Most individuals in both groups had normal levels (Chi-Square test statistic: 0.00129, p-value: 0.9710). The resting electrocardiogram results (restecg) demonstrated a strong association with abnormal results (restecg = 2), and was more common in individuals with heart disease. Normal results (restecg = 0) were prevalent among those without (Chi-Square test statistic: 35.63319, p-value: 2.38e-09). (see Table 2) (see Table 3).

Maximum heart rate (thalach) was inversely correlated with heart disease presence, with lower heart rates more common among individuals with heart disease (negative correlation coefficient, p-value < 0.05). Exercise-induced angina (exang) was noticeably more prevalent in those with heart disease (p-value: 2.38e-09). Oldpeak (ST depression) and slope (the slope of the ST segment during exercise) also exhibited a strong relationship with heart disease and with Oldpeak values (p-value < 0.05). Sharing their relationship with stress test outcomes and heart ischemia markers (see Table 2) (see Table 3).

In summary, the analysis underscores the importance of features such as Oldpeak, slope, chest pain type, and exercise-induced angina as strong indicators of heart disease. While expected trends like the link between age, vascular blockages, and declining cardiovascular efficiency were evident, weaker relationships for variables like cholesterol and fasting blood sugar highlight additional factors which were not associated with heart disease presence.

Table 2

Variable descriptives based on heart disease presence

Variable	HeartDiseasePresence	Count	Mean	Std	Min	25%	50%	75%	Max
Age	0	128	53.117	9.671	34	44.75	52.5	60.25	76
	1	109	56.716	7.825	35	53	58	62	77
Sex	0	128	0.57	0.497	0	0	1	1	1
	1	109	0.798	0.403	0	1	1	1	1
Cp	0	128	2.828	0.965	1	2	3	3.25	4
	1	109	3.606	0.828	1	4	4	4	4
Trestbps	0	128	129.578	16.193	94	120	130	140	180
	1	109	135.431	19.349	100	120	132	145	200
Chol	0	128	246.172	56.375	126	208.75	236	269	564
	1	109	253.211	49.107	164	219	254	284	409
Fbs	0	128	0.164	0.372	0	0	0	0	1
	1	109	0.174	0.381	0	0	0	0	1
Restecg	0	128	0.82	0.984	0	0	0	2	2
	1	109	1.174	0.98	0	0	2	2	2
Thalach	0	128	157.562	19.498	96	147.75	160	172	194
	1	109	140.78	22.263	88	126	144	158	195
Exang	0	128	0.172	0.379	0	0	0	0	1
	1	109	0.55	0.5	0	0	1	1	1
Oldpeak	0	128	0.591	0.726	0	0	0.2	1.2	3
	1	109	1.615	1.36	0	0.5	1.4	2.6	6.2
Slope	0	128	1.383	0.577	1	1	1	2	3
	1	109	1.826	0.575	1	1	2	2	3

Table 3

Correlation Statistic between individual variables and heart disease presence

Variable	Test Type	Correlation Coefficient / Chi2 Statistic	p-value	Degrees of Freedom	Interpretation
age	Point-Biserial	0.199	2.09E-03	-	Significant correlation (p < 0.05)
trestbps	Point-Biserial	0.1632	1.19E-02	-	Significant correlation (p < 0.05)
thalach	Point-Biserial	-0.3742	2.71E-09	-	Significant correlation (p < 0.05)
cholesterol	Point-Biserial	0.0661	3.11E-01	-	No significant correlation (p >= 0.05)
oldpeak	Chi-Square	68.8562	1.60E-03	-	Significant association (p < 0.05)
sex	Chi-Square	12.91475	3.26E-04	1	Significant association (p < 0.05)
cp	Chi-Square	65.29635	4.33E-14	3	Significant association (p < 0.05)
fbs	Chi-Square	0.00129	9.71E-01	1	No significant association (p >= 0.05)
restecg	Chi-Square	7.98928	1.84E-02	2	Significant association (p < 0.05)
exang	Chi-Square	35.63319	2.38E-09	1	Significant association (p < 0.05)
slope	Chi-Square	37.40353	7.55E-09	2	Significant association (p < 0.05)
ca	Chi-Square	46.99451	3.48E-10	3	Significant association (p < 0.05)
thal	Chi-Square	66.31957	3.97E-15	2	Significant association (p < 0.05)

Coding environment

Google Colab, short for "Colaboratory," is a cloud-based platform that allows users to write and execute Python code directly in their web browsers. Requiring no setup, it provides free access to GPUs for accelerated computation and facilitates easy sharing of projects. Google

Colab's computational resources and capabilities are leveraged throughout this project to write and execute code. Allowing a seamless workflow for analysis and modeling.

Coding language

Python is a high-level and versatile programming language and is a central tool for this project. Python facilitates efficient code development for tasks such as data analysis, machine learning, and scientific computing (Python Software Foundation, 2024). Python has a robust standard library and a wide array of third-party packages, providing the necessary tools to handle complex data processing, build predictive models, and perform statistical analyses.

Python libraries

Several Python libraries that are essential for data manipulation, analysis, visualization, and machine learning are used within my research. NumPy serves as a fundamental package for numerical computations, offering support for arrays and matrices alongside a wide range of mathematical functions for handling large datasets. Complementing this is Pandas, which provides data structures similar to Data Frames that streamline the management and analysis of structured data. For visualization, Matplotlib and Seaborn are utilized. Matplotlib facilitates the creation of static visualizations, while Seaborn, built on Matplotlib, enhances statistical graphics through a high-level interface

For machine learning tasks, the notebook relies heavily on Scikit-learn and XGBoost. Scikit-learn is a versatile library that provides tools for preprocessing, model selection, and the implementation of machine learning algorithms, including classification, regression, clustering, and dimensionality reduction. XGBoost, known for its high efficiency and flexibility, is an optimized gradient boosting library designed for supervised learning tasks, particularly large-

scale problems. Together, these libraries create a robust framework for data loading, preprocessing, visualization, and the implementation of advanced machine learning algorithms. By integrating these tools, the notebook demonstrates a comprehensive approach to solving binary classification problems, emphasizing both performance and interpretability.

Splitting into training and testing sets

Before conducting model evaluation, it is essential to split data into training and testing sets. Splitting data into training and testing sets is crucial for building reliable machine learning models. The training set is used to teach the model by allowing it to learn patterns from the data, while the testing set evaluates the model's performance on unseen data. This helps ensure that the model generalizes well and performs accurately on new, real-world data, reducing the risk of overfitting, where the model may perform well on training data but poorly on new data.

To effectively train and evaluate the binary classification model, the dataset was partitioned into training and testing subsets using the train_test_split function from the sklearn.model_selection module. The feature set (X) was created by removing the target column (num_binary)/ (Heart Disease Presence) from the dataset df_binary_col, while the target variable (y) was isolated as the column num_binary. This ensured that only relevant features were included in the model training process while preserving the target variable for prediction.

The data was split such that 80% of the samples were allocated to the training set and 20% to the testing set, specified by the test_size=0.2 parameter. To ensure reproducibility, a fixed random seed (random_state=42) was used, enabling consistent splitting of the dataset across different runs. Additionally, the stratify=y parameter was employed to maintain the proportional distribution of the binary target variable across both subsets. This stratification

ensured that the training and testing sets were representative of the overall class distribution, reducing the risk of bias and enhancing the reliability of the model evaluation. This method of data splitting forms a robust foundation for training machine learning models and assessing their generalization capabilities.

To ensure robustness, multiple train-test splits were conducted for each experiment in order to maintain consistency for model performance. After the general train-test split, the test set (X_test) was further divided into male and female subsets. This additional split aimed to evaluate the model's performance separately for each biological sex. The male and female subsets were used exclusively for testing, ensuring no overlap with the training data to maintain unbiased evaluations. This approach allowed for a detailed assessment of model generalizability for both men and women.

Exploratory data analysis (eda on x_train, Cleveland only)

Exploratory Data Analysis (EDA) was conducted exclusively on the X_train subset of Cleveland’s dataset to understand its structure and guide subsequent transformations and modeling decisions while avoiding data leakage. This analysis included univariate analysis of each variable, how each variable relates to each other, as well as, examining variable distributions and their relationships with the target variable (heart disease presence), which provided insights into the underlying data characteristics. Correlation analysis was performed to detect multicollinearity and redundant features, ensuring that highly correlated variables were appropriately managed in later steps. Additionally, data visualization and transformation were applied to identify and address skewed data that could impact model performance.

To assess the relationship between individual features and heart disease presence, Chi-square tests were performed for categorical variables, such as chest pain type (cp) and sex, to determine significant associations. Continuous variables, including age and resting blood pressure (restbps), were evaluated using point-biserial correlation to assess their linear relationship with the binary target variable. Both tests provided critical insights into feature relevance and informed decisions about feature selection and weighting during the modeling process. The insights gained through EDA also guided transformation strategies, including the application of log transformations to normalize skewed variables like cholesterol (chol) and oldpeak. Notably, EDA was not performed for the VA Long Beach dataset, as this data was reserved exclusively for validation to maintain the integrity of the evaluation process.

Data cleaning: Cleveland

Although there were no NULL values in the dataset, six rows contained the value "?" to indicate missing data. Since this affected only a small number of records, all six rows were removed.

Data cleaning: VA Long Beach

The data cleaning process for the VA Long Beach dataset was significantly more intricate compared to cleaning the Cleveland dataset, reflecting the increased complexity of the data, regarding the level of missing values. This thorough process was essential to address the challenges of missing values, outliers, and inconsistencies, ensuring the dataset's readiness for accurate binary classification modeling.

The first step involved removing columns with excessive missing values (ca and thal), as identified through visualizations. This step reduced noise while retaining the most informative

features. For the remaining columns, a set of imputation strategies were implemented to handle missing values based on their data characteristics. Numerical columns with a normal distribution, such as restbps and thalach, were addressed using mean imputation, where missing values were replaced with the column's mean. In contrast, median imputation was applied to skewed numerical columns, including chol and oldpeak, ensuring that the imputation process did not distort the data's central tendency. For skewed categorical or binary columns (fbs, exang, and slope), mode imputation was employed, replacing missing values with the most frequently occurring category.

Outlier handling posed another challenge. Continuous numerical columns (restbps, chol, thalach, and oldpeak) were processed using the Interquartile Range (IQR) method. This approach identified outliers as values outside 1.5 times the IQR from the first and third quartiles. These outliers were clipped to the calculated bounds, effectively reducing their influence while preserving the overall data distribution.

Following imputation and outlier handling, any additional rows with missing values were removed. The corresponding target labels (y_train and y_test) were realigned to maintain consistency with the cleaned feature datasets. This ensured a seamless alignment of the feature-target pairs, critical for effective model training and evaluation.

To validate the cleaning process, the datasets were examined to confirm the absence of missing values, and descriptive statistics were reviewed to verify that outliers had been successfully addressed. These verification steps confirmed the data sets' integrity, ensuring they were free of anomalies and inconsistencies.

Data transformation and feature engineering

Three distinct data transformation/feature engineering experiments were applied to the Cleveland dataset to evaluate their impact on model performance. These experiments aimed to address issues such as skewness, scaling, and feature encoding to enhance the predictive power of the models. Based on performance evaluation metrics from the Cleveland dataset, the best-performing data transformation and feature engineering strategy was subsequently applied to the VA Long Beach dataset to ensure consistency and generalizability.

Transformation experiments on Cleveland dataset

Comparing logarithmic data transformation.

To extract meaningful results from our machine learning models, it is important to account for outlier values. An outlier indicates a value which is significantly different in value from the rest of the dataset. These values can negatively affect how well a machine learning model can generalize results, as they affect the performance and accuracy of a model (Mueller and Guido, 2016). Outliers can be removed or accounted for using data transformation. Methods, such as logarithmic transformation, can reduce the impact of large outlier values. Alternatively, we can use square root transformation as well which is suitable for positively skewed data (Mueller and Guido, 2016).

For the following experiments, trestbps (resting blood pressure) and chol (cholesterol), and oldpeak were transformed using a logarithmic scale because their distributions are right-skewed.

Side-by-side histograms were created to show the original and log-transformed distributions for resting blood pressure, cholesterol, and oldpeak (ST depression). The first

histogram of the original data revealed a strong right skew for all three features. After log transformation, we observed a clear reduction in skewness. For resting blood pressure and cholesterol, the transformed data shows a significantly more symmetrical shape, indicating that the log transformation effectively normalized these distributions. For oldpeak, while the transformation reduces the skewness, the distribution remains slightly asymmetrical due to the heavy concentration of values near zero. These visual comparisons illustrate how log transformation reshapes the data, making it better suited for modeling and statistical analysis.

Comparing squared data transformation

Thalach (maximum heart rate achieved) has a left-skewed distribution. Therefore, a square transformation was applied to determine if this transformation better normalized the distribution. These visual comparisons highlight the impact of squared transformation on the data's structure, making it clearer how the technique improves the suitability of these features for modeling and statistical analysis

In the first histogram with the original data, there is a slightly longer tail extending toward the lower values, indicating a right skewness. The outlying values seem to be patients with a low maximum heart rate. The histogram that represents the squared-transformed data amplified the range of values, particularly for higher maximum heart rates, while slightly smoothing the irregularities in the original data. This results in a slightly more normal distribution of values, although the overall symmetry of the distribution is preserved.

Transformation 1: logarithmic and square transformations

The first transformation experiment focused on reducing skewness and capturing non-linear relationships within the Cleveland dataset by applying two custom transformations: logarithmic and square transformations. These transformations were implemented using scikit-learn's BaseEstimator and TransformerMixin, enabling integration into preprocessing pipelines. The "Log Transformer" applied a logarithmic transformation (log1p) to specific features, including resting blood pressure (trestbps) and cholesterol (chol), to reduce skewness and normalize their distributions. This adjustment aimed to stabilize variance and improve the compatibility of these features with machine learning algorithms. Similarly, the "Square Transformer" squared the values of maximum heart rate (thalach) to emphasize larger differences and capture potential non-linear relationships. These transformations tailored the preprocessing pipeline to the unique characteristics of the dataset. The integration of these custom transformers into the pipeline ensured seamless preprocessing and highlighted the importance of feature-specific transformations in enhancing data suitability for machine learning models.

Additionally, continuous features such as age, sex, chest pain, fasting blood sugar, resting electrocardiogram results, exercise induced angina, oldpeak, and slope are normalized using `StandardScaler` to ensure they have a mean of zero and a standard deviation of one. Categorical features, including 'number of major vessels (ca)', and 'thalassemia (thal)', are converted into binary format using OneHotEncoder. Any columns not explicitly specified for transformation are dropped, ensuring the preprocessing pipeline is both precise and adaptable to the dataset's needs.

Transformation 2: logarithmic, square, and combined features

The second transformation experiment extended the preprocessing strategy by incorporating logarithmic and square transformations alongside the creation of a new combined feature. As in the first experiment, logarithmic transformations were applied to resting blood pressure (trestbps) and cholesterol (chol) to address skewness and stabilize variance. Similarly, maximum heart rate (thalach) was squared to emphasize its higher values and capture non-linear relationships. Additionally, a combined feature, oldpeak_slope_combined, was created by summing Oldpeak (ST depression induced by exercise) and Slope (the slope of the peak ST segment). These features were identified as having a strong correlation (0.59), suggesting a synergistic relationship that could enhance predictive power. The combined feature aimed to capture the interaction between Oldpeak and Slope, representing their joint contribution to heart disease prediction. After the combined feature was created, the original columns were dropped to streamline the dataset. The transformed data was standardized and processed alongside other features in the pipeline, with the goal of improving model predictive accuracy through enhanced feature engineering.

Transformation 3: logarithmic, square, combined features, and gender-based engineering

The third transformation experiment built upon the second by introducing gender-specific feature engineering to account for potential differences between male and female subgroups. As in the previous experiments, logarithmic transformations were applied to resting blood pressure (restbps) and cholesterol (chol), while maximum heart rate (thalach) was squared to capture non-linear relationships. The combined feature, oldpeak_slope_combined, was also created by summing Oldpeak and Slope to leverage their correlated relationship. In addition to these transformations, gender-based interactions were introduced to explore potential differences in

feature relevance across genders. Interaction terms between gender and key features, such as cholesterol and resting blood pressure, were generated and integrated into the pipeline to evaluate their impact on model performance. By incorporating gender-aware feature engineering, this experiment aimed to enhance the model's ability to predict heart disease while accounting for demographic-specific variations. The transformed data was processed within a standardized pipeline, ensuring compatibility and consistency across features.

Parameter tuning

Three machine learning models—Random Forest Classifier, XGBoost Classifier, and an Ensemble Method combining the predictions from these two classifiers plus a logistic regression model—were used to evaluate the performance of each transformation strategy applied to the Cleveland dataset. For each model, a parameter grid search was conducted to optimize hyperparameters and enhance performance. This process involved systematically testing combinations of parameters, such as the number of estimators, maximum depth, learning rate, and feature split criteria, to identify the optimal configuration for each model. After parameter tuning, the models were trained and tested using multiple train-test splits to ensure robustness and mitigate the impact of variability. Performance metrics, including accuracy, precision, recall, and F1-score were calculated to comprehensively assess the predictive capabilities of each model under different transformations. The best-performing transformation was identified by selecting the strategy that achieved the highest average metrics across the evaluated models, ensuring both accuracy and reliability in subsequent applications.

To determine the most effective strategy for the Cleveland dataset, each experiment was evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provided an assessment of the models' predictive capabilities. To ensure robustness and minimize the impact of variability in train-test splits, multiple splits were applied during the evaluation process.

Among the three transformation experiments, Transformation 2, which incorporated logarithmic transformations for resting blood pressure (trestbps) and cholesterol (chol), along with a squared transformation for maximum heart rate (thalach) and a combined feature (oldpeak_slope_combined), demonstrated the best overall performance. This strategy effectively balanced predictive accuracy with robustness across the evaluation metrics, making it the optimal choice for subsequent application to the VA.

Gender-based evaluation on Cleveland dataset

Transformation 3 aimed to address gender differences identified during the exploratory analysis by introducing gender-specific feature engineering. This involved adding interaction terms between gender and key features, such as cholesterol and resting blood pressure, to capture sex-specific patterns and improve the model's predictive performance for heart disease. The ensemble method, which combined the random forest classifier, XGBoost, and logistic regression, delivered the best results for this experiment.

On the testing data, the model demonstrated strong performance, achieving a mean accuracy of 0.85, an F1 score of 0.85, mean precision of 0.85, and recall of 0.85. Cross-validation revealed a significant drop in performance, with a mean accuracy of 0.77, mean precision of 0.79, mean recall of 0.77, and a mean F1 score of 0.76. These results indicate

potential overfitting, as the model performed well on the testing data but struggled to generalize effectively across validation splits. Furthermore, the cross-validation results were worse than those achieved in previous experiments, Transformation 1 and Transformation 2.

Notably, Transformation 1 and Transformation 2 did not incorporate gender-specific parameter tuning, yet their cross-validation results outperformed Transformation 3. This suggests that while gender-aware parameter tuning improved performance on the test data, it introduced overfitting, reducing the model's generalizability. As a result, the experiment with the best cross-validation performance was selected and further evaluated separately for men and women to ensure a more reliable assessment of its effectiveness.

To evaluate the performance of the best experiment (Transformation 2) and model across male and female subgroups, the Cleveland dataset’s test set was divided into male and female subsets based on the "sex" feature. Models trained using the best experiment (Transformation 2) were assessed separately for each gender. Key performance metrics, including accuracy, precision, recall, and F1-score were calculated for the male and female subsets to understand how the model performed across these demographic groups. This analysis provided valuable insights into potential gender-based discrepancies in model behavior, emphasizing the importance of ensuring fairness and robustness in predictive modeling across demographic subgroups.

Evaluation on VA Long Beach dataset

The best-performing strategy identified from the Cleveland dataset, Transformation 2, was applied to the VA Long Beach dataset to evaluate the regional generalizability of the trained model. This strategy included logarithmic transformations for features with skewed distributions,

such as cholesterol (chol), resting blood pressure (trestbps), and Oldpeak, to stabilize variance and improve compatibility with the model. StandardScaler was then used to standardize continuous features, ensuring that all variables contributed equally during the model training and prediction process. Categorical variables were one-hot encoded, developed similarly as the Cleveland dataset to maintain consistency across the two regions. The goal of this approach was to investigate how a model trained exclusively on one region performs when applied to a different region with potentially varying feature distributions and population characteristics.

While a gender-based performance test was conducted for the Cleveland dataset, it could not be performed for the VA Long Beach dataset due to the limited representation of females in the dataset, with only six female observations. This small sample size precluded meaningful statistical analysis and comparison. Consequently, the gender-based analysis focused exclusively on the Cleveland dataset, where sufficient data for both male and female subgroups was available. By applying the same transformation strategy, the study aimed to make direct comparisons of overall model performance, assessing its robustness and ability to generalize across multiple cities in the U.S.

Application of the ascvd risk calculator to the Cleveland and VA Long Beach datasets

To further enhance the comparison of the predictive models, the ASCVD (Atherosclerotic Cardiovascular Disease) Risk Calculator was applied to both the Cleveland and VA Long Beach datasets. The ASCVD calculator, widely used in clinical settings, estimates the 10-year risk of cardiovascular disease based on key risk factors. The implementation was conducted using the Python package available in the GitHub repository https://github.com/brandones/ascvd/tree/master.

The ASCVD (Atherosclerotic Cardiovascular Disease) Risk Calculator was applied to both the Cleveland and VA Long Beach datasets to estimate the 10-year cardiovascular risk for individual patients. This ASCVD calculated risks using features such as age, sex, total cholesterol, HDL cholesterol, systolic blood pressure (SBP), blood pressure treatment status, diabetes status, and smoking status. Missing or unavailable data were addressed using proxies to ensure compatibility with the ASCVD model. HDL cholesterol was assigned a placeholder value of 50, and ethnicity was uniformly set to non-Black (isBlack = False) due to the absence of explicit data on this characteristic. Hypertension status was derived from SBP values, with readings of 130 or higher classified as hypertensive. Diabetes status was inferred from fasting blood sugar (fbs), with values over 120 converted into a boolean indicator (diabetic = True). Smoking status was approximated using exercise-induced angina (exang), treating the absence of angina as indicative of non-smoking status (False).

The use of proxies, such as placeholder HDL cholesterol values, inferred diabetes and smoking statuses, and uniform assumptions about ethnicity, underscores that the calculated risk scores are not pure ASCVD scores. These proxies, while necessary to accommodate missing or unavailable data, introduce approximations that deviate from the precise data required by the ASCVD model.

Results

Atherosclerotic cardiovascular disease risk calculation on Cleveland dataset

The 2013 ASCVD (Atherosclerotic Cardiovascular Disease) risk score was evaluated on the entire Cleveland dataset without any modifications to the data. The score achieved an accuracy of 69.64%, indicating that approximately 70% of predictions matched actual outcomes. Precision was 63.58%, reflecting the proportion of correctly identified positive cases among all predicted positives, while recall was 79.14%, demonstrating the model's ability to capture actual positive cases. The F1 score, balancing precision and recall, was 70.51%, signifying a moderate trade-off between the two. Additionally, the AUC-ROC score of 70.36% suggests a fair level of discriminatory ability between positive and negative cases. These results indicate that the 2013 ASCVD risk score provides reasonable predictive performance for the Cleveland dataset, with notable strengths in recall but areas for improvement in precision and overall accuracy.

The performance comparison between male and female subgroups reveals notable differences, particularly in the variability of scores for the female group. For females, the model achieved an accuracy of 73.19%, slightly higher than the accuracy of 69.67% observed for males. However, the precision for females was lower at 48.88% compared to 68.75% for males, indicating that the model was less reliable in identifying true positives among predicted positives for the female subgroup. Conversely, the recall for females was significantly higher at 88.89%, compared to 77.19% for males, suggesting that the model was more effective in identifying actual positive cases among females.

The F1 score, which balances precision and recall, highlights the disparity between the subgroups. Females achieved an F1 score of 62.87%, reflecting the impact of lower precision

despite high recall, whereas males had a more balanced F1 score of 72.73%. The AUC-ROC scores further emphasize this difference, with females achieving 78.08% compared to 68.87% for males, indicating better overall discrimination for the female subgroup.

The wide variability in the scores for the female group, particularly the sharp contrast between high recall and low precision, underscores potential challenges in the model's consistency when applied to different demographic subgroups.

Atherosclerotic cardiovascular disease risk calculation on VA Long Beach dataset

The 2013 ASCVD (Atherosclerotic Cardiovascular Disease) risk score was evaluated on the entire VA Long Beach dataset, without any modifications, aside from the imputation of missing data, achieving an accuracy of 76.82%, indicating that over three-quarters of the predictions aligned with the actual outcomes. The precision of 78.95% highlights the model's reliability in identifying true positives among predicted positives, while the recall of 93.75% demonstrates its ability to effectively detect many actual positive cases. The F1 score, balancing precision and recall, was 85.71%, reflecting robust overall performance. However, the AUC-ROC score of 60.98% suggests limited discriminatory power between positive and negative classes, indicating room for improvement in distinguishing cases.

It is important to note that due to the small number of females in the VA Long Beach dataset, a sex-specific analysis was not feasible. This limitation restricts the ability to assess the model's performance across different demographic subgroups and emphasizes the need for more diverse and balanced datasets in future analyses.

Model performance on Cleveland dataset

Results from multiple train-test splits and evaluations indicated that Transformation 2, which applied log transformations to skewed features and introduced a combined feature of Oldpeak and slope, consistently delivered the best performance. Among the models evaluated, the Random Forest Classifier emerged as the top performer (see Table 4).

The Random Forest Classifier model demonstrated strong performance on the test set, achieving an accuracy of 83.33%, an F1 score of 0.8345, a precision of 0.8556, and a recall of 0.8333. The confusion matrix highlights that the model correctly classified 30 true negatives and 20 true positives, with 2 false positives and 8 false negatives. These results indicate a balanced performance, with a slight trade-off between precision and recall for the positive class (see Table 4).

Cross-validation results confirm the model's consistency across folds. The mean accuracy was 88.33%, with a mean precision of 91.79%, a mean recall of 82%, and a mean F1 score of 0.8367. The precision remained high, reflecting the model's ability to minimize false positives, while recall variability suggests opportunities for improvement in detecting true positives (see Table 4).

Overall, the model exhibits strong generalization capabilities, supported by its stable cross-validation performance and robust test set results.

Model performance on VA Long Beach dataset

Transformation 2 was used to evaluate the generalizability of the Random Forest Classifier by applying it to the VA Long Beach dataset (see Table 4).

The new model, without optimized parameters, demonstrated an improvement in terms of its ability to predict instances of the minority class (those with heart disease). The overall accuracy is slightly higher at 82.5%, compared to the 75% accuracy of the optimized model, and overall, has a more balanced performance across both classes (see Table 4).

In this model, the confusion matrix reveals that the classifier correctly predicts 4 out of 10 instances of the minority class (0), resulting in a recall of 40%. Additionally, the precision for the minority class is 80%, with an F1 score of 0.53, indicating a more balanced trade-off between precision and recall for this class. For the majority class (1), the model maintains high precision (83%) and recall (97%), resulting in an F1 score of 0.89 (see Table 4).

The macro-average F1 score of 0.71 and weighted-average F1 score of 0.82 suggest better overall performance compared to the previous model with optimized parameters, which failed to predict any minority class instances. This highlights that, while parameter optimization may increase accuracy for the majority class, it can lead to a complete neglect of the minority class, whereas a more generic configuration balances predictions across both classes more effectively.

Table 4 - Model Results

Experiment #	City	Model	Mean Accuracy	Mean F1 Score	Mean Precision	Mean Recall	Cross-Validation Accuracy	Cross-Validation F1 Score	Cross-Validation Precision	Cross-Validation Recall
ACSVD Risk Score	Cleveland	ACSVD Risk Score	0.69	0.7	0.64	0.79	N/A	N/A	N/A	N/A
ACSVD Risk Score - Male	Cleveland	ACSVD Risk Score	0.68	0.73	0.69	0.77	N/A	N/A	N/A	N/A
ACSVD Risk Score - Female	Cleveland	ACSVD Risk Score	0.73	0.63	0.49	0.88	N/A	N/A	N/A	N/A
ACSVD Risk Score	VA Long Beach	ACSVD Risk Score	0.77	0.86	0.79	0.94	N/A	N/A	N/A	N/A
Transformation 1	Cleveland	Random Forest Classifier - Baseline	0.85	0.85	0.86	0.85	0.8	0.77	0.78	0.77
Transformation 1	Cleveland	Random Forest Classifier - Optimized Parameters	0.83	0.83	0.84	0.83	0.82	0.78	0.78	0.79
Transformation 1	Cleveland	XGB Classifier - Baseline	0.85	0.85	0.85	0.85	0.75	0.72	0.72	0.73
Transformation 1	Cleveland	XGB Classifier - Optimized Parameters	0.8	0.8	0.81	0.8	0.8	0.78	0.76	0.8
Transformation 1	Cleveland	Ensemble Method	0.8	0.8	0.81	0.8	0.78	0.78	0.79	0.78
Transformation 2	Cleveland	Random Forrest Classifier - Baseline	0.83	0.83	0.84	0.83	0.82	0.77	0.81	0.76
Transformation 2	Cleveland	Random Forrest Classifier - Optimized Parameters	0.83	0.83	0.85	0.83	0.88	0.84	0.92	0.82
Transformation 2	Cleveland	XGB Classifier - Baseline	0.87	0.87	0.87	0.87	0.77	0.72	0.75	0.7
Transformation 2	Cleveland	XGB Classifier - Optimized Parameters	0.82	0.81	0.82	0.82	0.8	0.78	0.76	0.8
Transformation 2	Cleveland	Ensemble Method	0.8	0.8	0.81	0.8	0.82	0.81	0.84	0.82
Transformation 3	Cleveland	Random Forrest Classifier - Baseline	0.83	0.83	0.84	0.83	0.78	0.72	0.75	0.72
Transformation 3	Cleveland	Random Forrest Classifier - Optimized Parameters	0.83	0.83	0.84	0.83	0.78	0.74	0.74	0.76
Transformation 3	Cleveland	XGB Classifier - Baseline	0.87	0.87	0.87	0.87	0.78	0.78	0.7	0.73
Transformation 3	Cleveland	XGB Classifier - Optimized Parameters	0.87	0.87	0.87	0.87	0.77	0.72	0.75	0.7
Transformation 3	Cleveland	Ensemble Method	0.86	0.86	0.86	0.86	0.77	0.76	0.79	0.77
Transformation 2 - Male	Cleveland	Random Forrest Classifier - Optimized Parameters	0.78	0.78	0.8	0.78	N/A	N/A	N/A	N/A
Transformation 2 - Female	Cleveland	Random Forrest Classifier - Optimized Parameters	0.94	0.94	0.96	0.94	N/A	N/A	N/A	N/A
Transformation 2	VA Long Beach	Random Forrest Classifier - Optimized Parameters	0.75	0.64	1	0.75	0.75	0.84	0.75	1
Transformation 2	VA Long Beach	Random Forrest Classifier - Baseline	0.83	0.8	0.9	0.82	0.75	0.8	0.76	0.97
Transformation 2	Adjusted Cleveland Dataset	Random Forrest Classifier - Optimized Parameters	0.72	0.72	0.72	0.72	0.78	0.88	0.64	0.69

Discussion

Key findings:

My project leveraged the Cleveland and VA Long Beach datasets, in the "Heart Disease" database, which was donated to the UCI Machine Learning Repository to explore the binary classification of heart disease presence, using the available demographic and clinical features. Through exploratory data analysis (EDA), data cleaning, transformation experiments, and model evaluation, several critical insights emerged. Transformation 2, which included logarithmic transformations for skewed features, a squared transformation for maximum heart rate, and the creation of a combined feature (old peak and slope), was identified as the most effective preprocessing strategy. This approach enhanced feature stability and predictive accuracy on the Cleveland dataset and was subsequently applied to the VA Long Beach dataset to assess regional generalizability.

The Random Forest Classifier for Transformation 2 consistently outperformed other models in terms of prediction accuracy and robustness across multiple train-test splits. On the Cleveland dataset, the Random Forest Classifier achieved an accuracy of 83.33%, with balanced precision (85.56%) and recall (83.33%), supported by consistent cross-validation performance. Applied to the VA Long Beach dataset, the unoptimized Random Forest Classifier demonstrated a performance with an accuracy of 82.5%, precision of 80%, and recall of 40% for the minority class. Unlike its optimized counterpart, this configuration effectively mitigated the neglect of minority class predictions, striking a better balance between the classes.

While the Random Forest Classifier outperformed the ASCVD risk score on average across both datasets, achieving higher accuracy and F1 scores, it did not resolve variability in

female subgroup predictions. On the Cleveland dataset, the ASCVD score achieved an AUC-ROC of 78.08% for females compared to 68.87% for males, and the Random Forest model exhibited similar imbalances. Female recall for the ASCVD was higher (88.89%) but precision was considerably lower (48.88%), resulting in a less balanced F1 score (62.87%) compared to males (72.73%). On the VA Long Beach dataset, the ASCVD score achieved strong recall (93.75%) but limited discriminatory power, with an AUC-ROC of 60.98%. The lack of sufficient female representation in the VA Long Beach dataset precluded any sex-specific analysis, underscoring the importance of diverse and balanced datasets.

Gender-based analysis on the Cleveland dataset revealed clear disparities in machine learning model performance with the Random Forest Classifier. While the model achieved high overall accuracy and precision; the model demonstrated considerably lower recall for female patients with heart disease (0.67). This variability indicates that the model, while improving overall performance compared to the ASCVD score, did not effectively address gender-specific prediction inconsistencies. Features such as Oldpeak and slope emerged as strong indicators of heart disease presence, whereas weaker relationships were observed for cholesterol and fasting blood sugar, suggesting limited predictive value for these variables within these datasets.

In conclusion, the Random Forest Classifier demonstrated superior average performance compared to the ASCVD risk score but fell short of addressing variability in female subgroup predictions. These findings highlight the importance of future work to explore gender-specific adjustments and strategies for achieving equitable performance across demographic groups, while also emphasizing the need for diverse datasets to enhance generalizability.

Regional generalizability

VA Long Beach models with optimized parameters highlight a few possible implications. Parameter optimization of the Random Forest Classifier, while improving the overall accuracy, had significant drawbacks when applied to a dataset with an imbalanced class distribution. The optimized model prioritized the majority class, resulting in a complete inability to predict any instances of the minority class (those with heart disease). This led to a recall of 0% for the minority class, effectively excluding it from the model's predictions. Although overall accuracy increased, this came at the cost of fairness and utility, as the model failed to capture critical instances of the minority class. These findings highlight that, in the case of the Random Forest Classifier, parameter optimization compromised the model's balance, favoring the majority class performance while neglecting the minority class.

The comparison between the Cleveland and VA Long Beach models without optimized parameters further highlights two possible implications: optimized parameters can overgeneralize for the majority class, or they may reduce the model's ability to classify effectively across regions. These potential drawbacks are particularly obscured by the small number of patients in the minority class (those without heart disease) in the VA Long Beach dataset, which makes it challenging to draw definitive conclusions.

Without optimization, the Random Forest Classifier still achieved balanced performance for the minority class on the Cleveland dataset, while still maintaining reasonable performance for the VA Long Beach dataset. However, with optimized parameters, the VA Long Beach model completely failed to predict any instances of the minority class, suggesting that the optimization may have tailored the model too closely to the majority class, resulting in overgeneralization.

These outcomes suggest that the small sample size of the minority class amplifies the difficulty in determining whether the reduced performance is due to overgeneralization for the majority class or a lack of adaptability across regions. This underscores the importance of balancing datasets and carefully evaluating the impact of optimization on both regional performance and minority class predictions.

Sex-specific performance summary

The model assessing the male population achieved an accuracy of 78.05% and an F1-score of 78.31%. For class 0 (no heart disease), it recorded a precision of 68%, recall of 81%, and an F1-score of 74%. For class 1 (heart disease), the model demonstrated a higher precision of 86% but a lower recall of 76%, resulting in an F1-score of 81%. The overall weighted averages for precision, recall, and F1-score were 80%, 78%, and 78%, respectively, reflecting moderately balanced performance on the male population. In contrast, the model applied to the female population exhibited higher overall accuracy (94.74%) and F1-score (94%). The precision for class 0 was higher (94% compared to 67% in the male model), the recall for class 0 was significantly higher as well at 100%, surpassing the male model's recall of 81%. For class 1, the female model achieved perfect precision (100%) but a lower recall of 67%, compared to the male model’s recall of 76%.

These results highlight distinct performance differences between the male and female subpopulations. The model demonstrated better overall accuracy and precision for the female population, but its ability to detect heart disease cases (class 1) was slightly lower in recall compared to the male population. Conversely, the male model exhibited a more balanced trade-off between precision and recall for class 1 but at the cost of a higher false positive rate. These

findings underscore the need for additional tuning to ensure the model performs consistently across gender-specific groups, avoiding potential biases in prediction outcomes.

Data limitations

Due to the limitations of the Cleveland data, with the male test set comprising 41 samples and the female test set comprising only 19 samples, I am unable to perform cross-validation for these subsets. This restriction limits the ability to thoroughly assess model generalizability and robustness across gender-specific groups. As a result, the interpretation of the results must be approached with caution, as the insights drawn may not fully capture the broader performance trends for male and female populations.

The VA Long Beach dataset had significant limitations that must be addressed. One key issue is the severe gender imbalance, with only approximately 6 females included in the dataset. This small number makes it impossible to test the model by gender, as there are insufficient entries to draw any meaningful conclusions for female patients.

Additionally, the dataset suffers from a pronounced class imbalance, with most instances representing individuals with heart disease. This imbalance introduces challenges for the model, as there is limited data available to effectively train the minority class (individuals without heart disease). As a result, there is a strong expectation of underfitting for the minority class, where the model may fail to accurately predict or generalize for these cases.

Another critical issue is the missing data. Columns such as "ca" (99% missing values) and "thal" (83% missing values) have so few valid entries that they needed to be removed from the analysis.

To ensure a fair comparison when evaluating models, I had to create modified Cleveland models for comparison. This was done by removing the "ca" and "thal" columns, aligning the feature set with the limitations of this dataset. This approach allowed for a slightly more balanced evaluation when comparing the models trained on the Cleveland dataset to the Long Beach Models.

Lastly, The ASCVD (Atherosclerotic Cardiovascular Disease) Risk Calculator was applied to both the Cleveland and VA Long Beach datasets to estimate the 10-year cardiovascular risk for individual patients. However, due to missing or unavailable data, proxies were used to ensure compatibility with the ASCVD model. HDL cholesterol was assigned a placeholder value of 50, and ethnicity was uniformly set to non-Black (isBlack = False) because explicit data on this characteristic was not available. Hypertension status was derived from systolic blood pressure (SBP) values, with readings of 130 or higher classified as hypertensive. Diabetes status was inferred from fasting blood sugar (fbs), with values over 120 converted into a boolean indicator (diabetic = True). Smoking status was approximated using exercise-induced angina (exang), with the absence of angina interpreted as non-smoking status (False).

While these proxies enabled the datasets to be used for ASCVD risk estimation, they introduced approximations that deviate from the precise inputs required by the ASCVD model. Consequently, the calculated risk scores are not pure ASCVD scores, but rather adapted estimates. This reliance on proxies adds a layer of uncertainty to the analysis and necessitates cautious interpretation of the results.

Future directions

My research highlighted several areas for improvement and exploration in future research. After evaluating the datasets and outcomes, it became evident that a more effective approach might involve transitioning from binary classification to multilabel classification. Specifically, this approach could predict varying levels of heart disease severity rather than focusing solely on the binary presence or absence of the condition. This shift in focus is motivated by the observation that, aside from the Cleveland dataset, the other cities' datasets predominantly consist of patients with some degree of heart disease. The relative scarcity of individuals without heart disease in these datasets diminishes the utility of binary classification and underscores the potential for a more nuanced multilabel approach.

Additionally, my research revealed a significant gender imbalance in the datasets, with fewer females represented compared to males. This raises critical questions about whether this disparity reflects sampling bias or is indicative of real-world clinical trends. Considering that cardiovascular disease is a leading cause of death among women, it is essential to investigate why females may be underrepresented in these datasets. Future research should aim to address this imbalance, ensuring equitable representation to enhance the generalizability and fairness of predictive models.

Incorporating these changes, future studies could develop more accurate and reproducible models that account for demographic disparities and focus on the varying levels of heart disease severity. This approach would not only provide richer clinical insights but also foster more inclusive and accurate models capable of addressing the diverse needs of populations affected by cardiovascular disease.

Meta-critical reflective appendix

The capstone project is a culmination of the skills and knowledge acquired through coursework in data analysis fundamentals, quantitative research methods, fundamentals of design and visualization, fundamentals of machine learning, as well as coursework detailing how social issues affect our work as analysts. Biases that are reflected in society permeate into data and how it's analyzed.

The capstone project, *"Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms,"* aims to enhance cardiovascular risk assessment by leveraging analytics tools, learned in my studies, to explore machine learning's potential to surpass traditional risk calculators like ASCVD. This project is at the intersection of data science and public health and contributes to digital scholarship by addressing disparities in predictive accuracy across genders and cities in the U.S., offering a nuanced understanding of machine learning's capabilities in healthcare. Through this study, machine learning models were applied to heart disease datasets from Cleveland and Long Beach to evaluate their generalizability and efficacy in predicting cardiovascular risk, particularly for underrepresented groups like women. By emphasizing transparency, equity, and evidence-based decision-making, this capstone extends the field of digital humanities by engaging with ethical concerns and systemic biases in data-driven health solutions.

The project critiques and builds upon existing research in cardiovascular risk prediction by comparing traditional models to machine learning techniques. It addresses the limitations of current methods, such as their failure to account for subgroup-specific variations and incorporates innovative approaches like gender-specific feature engineering. By documenting

and sharing tools, workflows, and findings in a digital repository, the project not only ensures reproducibility but also contributes resources that future scholars and practitioners can utilize and build upon. This study enriches the existing ecosystem by offering actionable insights and models that account for demographic diversity, promoting a more inclusive approach to healthcare analytics while challenging traditional methodologies to evolve in response to contemporary demands.

Early in my education, I learned the fundamentals of quantitative research methods with Jeremy Porter. In this class I covered issues pertaining to the applied research process in the collection, management, and analysis of quantitative data. Foundational tools such as data visualization and correlational analysis were used to tackle a real-world issue. In this case, I began exploring my interest in health data by doing research related to the effects of COVID-19. What were characteristics globally, that led to higher death rates? While I did not continue to directly explore this question, my interest in tackling public health issues became my primary focus.

In the spring of that year, I took the course, “Data Analysis Methods” with Professor Liza Steele, where I was introduced basic statistical techniques necessary for analyzing data. Concepts like descriptive statistics and inferential statistics were explored. Various statistical measures and techniques for analyzing data were applied to real-world data problems. At this point, I began using NYC Open Data repositories, to understand arrest data in New York city. In my thesis project, descriptive statistics were explored to get an overview of the clinical data available to me and inferential statistics were used to understand how the individual variable in the dataset correlated with the presence of heart disease.

In the spring of 2024, I took “Data, Culture, and Society” with Professor James Lowry. In this course, I examined the social, political, and cultural effects of society's growing dependence on large-scale, often real-time, data analysis. I covered the principles of data collection, organization, analysis, and dissemination, while addressing the opportunities, challenges, and implications of using data-driven approaches in social sciences and the humanities. The course also delved into established work within computational social science, digital humanities, and cultural analytics. I explored the history and key concepts behind modern methods for analyzing data patterns, including statistics, data visualization, data mining, and machine learning. While in this class, I began looking into the effectiveness of risk models, which are typically evaluated through summary statistics, as is normal cultural practice. However, using summary statistics for risk models can result in inaccuracies in assessing certain subgroups within the population demographic. Furthermore, the true impact of risk factors for these subgroups, maintains as a pipeline of information not available to the public.

In Fall 2024, I had the opportunity to enroll in “Data Bias - How Big Data Increases Inequality” with Professor Allen Hillery, where I explored what data bias is and how it occurs. Data bias arises when data or information is incomplete or skewed, leading to an inaccurate representation of the population being analyzed or failing to tell the full story. It has the potential to influence individuals, businesses, and even entire societies in profound ways. For my final project in the course, I investigated the implications of data bias within facial recognition technology (FRT), uncovering how it contributes to social and economic inequality. My research highlighted issues such as algorithmic discrimination against underrepresented groups, the lack of transparency and accountability in the development of such systems, and the ethical challenges these technologies pose.

One striking example discussed was the wrongful arrest of Julian-Borchak Williams due to a flawed facial recognition system, which illustrated the real-world consequences of biased data in law enforcement. Through this research, I engaged deeply with the work of scholars like Dr. Brandeis Marshall, who emphasizes the need for ethical frameworks in algorithm design and explored potential solutions such as stricter regulations and more inclusive data practices. This course not only deepened my understanding of data bias but also sharpened my ability to critically analyze the social, cultural, and ethical implications of technology in modern society.

The two courses, “Data, Culture, and Society” and “Data Bias,” significantly influenced the direction and insights of my final paper. In "Data, Culture, and Society," I developed an understanding of how large-scale data analysis is deeply embedded in societal systems and how summary statistics, while useful, can lead to inaccuracies when applied broadly, particularly in subgroup analyses. This provided a foundation for my exploration of heart disease risk models, where I critically examined the limitations of traditional summary statistics in capturing the nuances of diverse populations. The concepts of transparency, fairness, and the need for more equitable data representation became central to my analysis.

Building on this, “Data Bias” with Professor Allen Hillery deepened my awareness of how incomplete or skewed data could perpetuate systemic inequities. This perspective was instrumental in evaluating the gender-specific and regional disparities highlighted in my capstone project. For example, my research into the biases in facial recognition systems informed my approach to identifying and addressing the gender-based disparities in predictive accuracy within heart disease models. Both courses emphasized the critical role of fairness, accountability, and inclusivity in data-driven decision-making, shaping my exploration of how machine learning algorithms can be optimized to minimize biases and maximize effectiveness in

diverse contexts. Together, these courses provided the analytical and ethical framework that underpinned the methodological choices and critical insights in my final paper.

In Summer 2024, my engagement with "Advanced Data Analysis" under Professor Johanna Devaney significantly influenced the technical development of my capstone project. This course provided me with practical skills in machine learning techniques, focusing on the application and interpretation of algorithms using Python and the scikit-learn library. The hands-on approach, particularly with supervised methods like decision trees, support vector machines, and ensemble techniques, directly informed my implementation of machine learning models such as Random Forest and XGBoost in my capstone research.

One of the most impactful lessons from "Advanced Data Analysis" was learning to design features and preprocess data effectively, which I applied extensively in my project. For instance, I implemented tailored transformations, such as normalizing skewed features and engineering new ones (e.g., combining ST depression with the slope of the ST segment), to improve model performance. Additionally, the course emphasized the importance of evaluating models not just for accuracy but also for their ability to generalize across diverse datasets—a concept that influenced my exploration of regional and gender-based disparities in heart disease prediction.

"Advanced Data Analysis" also instilled an understanding of when machine learning techniques are appropriate compared to traditional statistical methods, which helped me critically evaluate and compare the performance of machine learning models against cardiovascular risk scores like ASCVD. This alignment of practical machine learning skills with a broader conceptual framework was invaluable in refining my capstone's methodological rigor and ensuring its results were both interpretable and impactful.

This digital capstone project, “Sex-Specific and Regional Analysis of Heart Disease Prediction Using Machine Learning Algorithms: Insights from the UCI Irvine Public Heart Disease Datasets (Cleveland and Long Beach)”, documents a comprehensive journey of research, development, and evaluation to address critical challenges in heart disease prediction. The process began with identifying the research problem—assessing the performance and biases of machine learning models compared to traditional cardiovascular risk calculators like ASCVD. Using publicly available datasets from the UCI Machine Learning Repository, the project involved meticulous steps of data cleaning, exploratory analysis, feature engineering, and model evaluation. Each stage of the process was carefully recorded to ensure transparency and reproducibility, from addressing missing data and class imbalances to applying advanced data transformations and engineering gender-specific interaction terms.

The product of this project is multifaceted. It includes machine learning models—such as Random Forest and XGBoost—optimized and evaluated on two datasets (Cleveland and VA Long Beach) to assess their predictive accuracy and generalizability. A key feature of the project is its emphasis on subgroup analysis, highlighting the gender-specific and regional disparities in model performance. Additionally, the project integrates tools and methodologies documented in the **Digital Manifest**, which catalogs essential resources, including Python libraries, processed datasets, and detailed Jupyter notebooks hosted in a GitHub repository. The final white paper encapsulates the findings, discussing key metrics like accuracy, recall, and F1-score, as well as the limitations and ethical considerations of the models. By detailing the challenges, values, and technical decisions that shaped this project, the documentation not only enhances the reproducibility of the work but also provides future researchers and practitioners with a critical resource for advancing machine learning applications in healthcare.

Data management plan overview

The data management plan ensures the organized, secure, and ethical handling of all project data. I acquired datasets from the UCI Machine Learning Repository and follow their terms of use. The data was stored securely on a personal computer. I documented all data processing steps, including cleaning, transformation, and analysis, ensuring transparency and reproducibility. The data is already anonymized for individual privacy. Access to the data was restricted to authorized project members only. Upon project completion, I submitted our data and final project documentation to the CUNY Graduate Center Library's digital repository, adhering to their guidelines for online digital deposits. This submission ensures long-term preservation and accessibility of our work. For detailed guidance on data management and submission, I referred to the library's resources available on their website.

Appendix A

Data dictionary

Significant Variables

1. **Age**
    - Type: Integer
    - Description: Patient's age in years.
2. **Sex**
    - Type: Binary (0 for Female, 1 for Male)
    - Description: Biological sex of the patient.
3. **Cp (Chest Pain Type)**
    - Type: Categorical (0–4)
    - Description: Chest pain severity levels, where higher values indicate more severe pain.
4. **Trestbps (Resting Blood Pressure)**
    - Type: Continuous (mmHg)
    - Description: Resting blood pressure in millimeters of mercury. Transformed using logarithmic scaling to reduce skewness.
5. **Chol (Serum Cholesterol)**
    - Type: Continuous (mg/dL)
    - Description: Serum cholesterol level in milligrams per deciliter. Transformed using logarithmic scaling to reduce skewness.
6. **Fbs (Fasting Blood Sugar)**
    - Type: Binary (0 for <120 mg/dL, 1 for ≥120 mg/dL)
    - Description: Indicator of whether fasting blood sugar exceeds 120 mg/dL.
7. **Restecg (Resting ECG Results)**
    - Type: Categorical (0–2)
    - Description: Results of resting electrocardiographic tests (e.g., normal, ST-T wave abnormality, left ventricular hypertrophy).

8. **Thalach (Maximum Heart Rate Achieved)**
    - Type: Continuous (bpm)
    - Description: Maximum heart rate achieved during exercise. Transformed using a squared transformation to emphasize non-linear relationships.
9. **Exang (Exercise-Induced Angina)**
    - Type: Binary (0 for No, 1 for Yes)
    - Description: Presence of exercise-induced angina (chest pain).
10. **Oldpeak**
    - Type: Continuous
    - Description: ST depression induced by exercise relative to rest (ECG measure). Transformed using logarithmic scaling to reduce skewness.
11. **Slope (ST Segment Slope)**
    - Type: Categorical (1 for Upsloping, 2 for Flat, 3 for Downsloping)
    - Description: The slope of the peak exercise ST segment.
12. **Ca (Number of Major Vessels)**
    - Type: Integer (0–3)
    - Description: Number of major vessels (0–3) colored by fluoroscopy. Transformed using one-hot encoding.
13. **Thal (Thallium Stress Test Results)**
    - Type: Categorical (3 for Normal, 6 for Fixed Defect, 7 for Reversible Defect)
    - Description: Results of thallium stress tests. Transformed using one-hot encoding.
14. **Oldpeak_Slope_Combined**
    - Type: Continuous
    - Description: A derived feature combining Oldpeak (ST depression) and Slope (ECG segment pattern during peak exercise).
15. **Gender**-Based Interaction Terms
    - Type: Continuous
    - Description: Interaction features created by multiplying the "Sex" feature with key variables like Chol and Trestbps to account for demographic-specific variations.

Critical Functions

1. **Log Transformer**
    - Purpose: Reduces skewness in variables like Chol, Trestbps, and Oldpeak.
    - Inputs: Skewed numerical features.
    - Outputs: Log-transformed features.
2. **Squared Transformation**
    - Purpose: Captures non-linear relationships in features like Thalach.
    - Inputs: Thalach feature.
    - Outputs: Squared-transformed feature.
3. **Combine Oldpeak and Slope**
    - Purpose: Creates a new feature to enhance model accuracy.
    - Inputs: Oldpeak and Slope features.
    - Outputs: Combined feature reflecting ST segment depression and slope interaction.
4. **Gender-Based Interaction Creation**
    - Purpose: Generates gender-specific interaction terms to capture the influence of demographic variations on key features.
    - Inputs: Sex feature and numerical features such as Chol and Trestbps.
    - Outputs: Interaction features highlighting gender-based relevance.

Classifiers Used

1. **Random Forest Classifier**
    - Purpose: Constructs an ensemble of decision trees for binary classification.
    - Features: Robust against overfitting, useful for datasets with imbalanced classes.
    - Implementation: Optimized using GridSearchCV to select parameters like the number of estimators, maximum depth, and feature importance.
2. **XGBoost Classifier**
    - Purpose: Gradient boosting algorithm designed for efficiency and performance in binary classification tasks.
    - Features: Focuses on minimizing loss functions with parallelized tree construction.

3. **Ensemble Method**
    - Purpose: Combines predictions from Random Forest, XGBoost, and Logistic Regression to improve robustness.
    - Features: Weighted averaging of classifiers to leverage strengths of individual models.
4. **Logistic Regression**
    - Purpose: Serves as a baseline model to compare linear relationships between features and outcomes.
    - Features: Interpretable and effective for datasets with linear separability.

Bibliography

Al Hamid, Abdullah, Beckett, Rachel, Wilson, Megan, et al. “Gender Bias in Diagnosis, Prevention, and Treatment of Cardiovascular Diseases: A Systematic Review.” Cureus, U.S. National Library of Medicine, 15 Feb. 2024, www.ncbi.nlm.nih.gov/pmc/articles/PMC10945154/.

Calomfirescu, Marius Vicea, and Nicoleta Elena Calomfirescu. “Assessing Women’s Cardiovascular Risk.” European Society of Cardiology, 9 Apr. 2012, www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-10/Assessing-women-s-cardiovascular-risk

Cesare, Nina, and Lawrence P O Were. “A Multi-Step Approach to Managing Missing Data in Time and Patient Variant Electronic Health Records.” BMC Research Notes, U.S. National Library of Medicine, 17 Feb. 2022, pubmed.ncbi.nlm.nih.gov/35177096/

Cho, Sang-Yeong, Kim, Sun-Hwa, Kang, Si-Hyuck, et al. “Pre-Existing and Machine Learning-Based Models for Cardiovascular Risk Prediction.” Nature News, Nature Publishing Group, 26 Apr. 2021, www.nature.com/articles/s41598-021-88257-w.

Hogo, Mofreh A. “A Proposed Gender-Based Approach for Diagnosis of the Coronary Artery Disease - Discover Applied Sciences.” *SpringerLink*, Springer International Publishing, 11 May 2020, link.springer.com/article/10.1007/s42452-020-2858-1.

Janosi, Andras, et al. “Heart Disease.” UCI Machine Learning Repository, 1988, archive.ics.uci.edu/dataset/45/heart+disease.

Mueller, Andreas C, and Guido, Sarah. *Introduction to Machine Learning with Python*, 22 Sept. 2016, github.com/dlsucomet/MLResources/blob/master/books/[ML] Introduction to Machine Learning with Python (2017).pdf.

Shaw, Leslee J. “Framingham Risk Score.” Framingham Risk Score - an Overview | ScienceDirect Topics, Interventional Cardiology Clinics, Apr. 2012, www.sciencedirect.com/topics/medicine-and-dentistry/framingham-risk-score

Siontis, George C M, Tzoulaki, Ioanna, Siontis, Konstantinos C, et al. “Comparisons of Established Risk Prediction Models for Cardiovascular Disease: Systematic Review.” The BMJ, British Medical Journal Publishing Group, 24 May 2012, www.bmj.com/content/344/bmj.e3318

Sofogianni, Areti, Stalikas, Nikolaos, Antza, Christina et al. “Cardiovascular Risk Prediction Models and Scores in the Era of Personalized Medicine.” Journal of Personalized Medicine, U.S. National Library of Medicine, 20 July 2022, www.ncbi.nlm.nih.gov/pmc/articles/PMC9317494/

Talha, Ibtissam, Elkhoudri, Noureddine, and Hilali, Abderraouf "Major Limitations of Cardiovascular Risk Scores." Cardiovascular Therapeutics, U.S. National Library of Medicine, 28 Feb. 2024, www.ncbi.nlm.nih.gov/pmc/articles/PMC10917477/

Weng, Stephen F, Reps, Jenna, Kai, Joe, et al. "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?" PloS One, U.S. National Library of Medicine, 4 Apr. 2017, pubmed.ncbi.nlm.nih.gov/28376093/.

Woodward, Mark. "Cardiovascular Disease and the Female Disadvantage." International Journal of Environmental Research and Public Health, U.S. National Library of Medicine, 1 Apr. 2019, pubmed.ncbi.nlm.nih.gov/30939754/.

"Scikit-Learn Machine Learning in Python." *Scikit-Learn*, scikit-learn.org/stable/.

Digital References

Software and Tools Used

1. **Google Colab**
    - Description: Cloud-based Python environment with GPU access for accelerated computation.
    - URL: https://colab.research.google.com
    - Accessed: November 2024
2. **Python**
    - Version: 3.8
    - Description: High-level programming language used for data analysis, modeling, and visualization.
    - URL: https://www.python.org
    - Accessed: November 2024
3. **Scikit-learn**
    - Version: 1.2.0
    - Description: Library for machine learning algorithms, preprocessing, and evaluation.
    - URL: https://scikit-learn.org/stable/
    - Accessed: November 2024
4. **XGBoost**
    - Version: 1.6.0
    - Description: Gradient boosting library optimized for supervised learning tasks.
    - URL: https://xgboost.ai
    - Accessed: November 2024
5. **Pandas**
    - Version: 1.4.3
    - Description: Data manipulation and analysis library for structured data.
    - URL: https://pandas.pydata.org
    - Accessed: November 2024

6. **NumPy**
    - Version: 1.23.0
    - Description: Library for numerical computations and array processing.
    - URL: https://numpy.org
    - Accessed: November 2024
7. **Matplotlib**
    - Version: 3.6.0
    - Description: Visualization library for static and interactive graphics.
    - URL: https://matplotlib.org
    - Accessed: November 2024
8. **Seaborn**
    - Version: 0.12.2
    - Description: Statistical data visualization library built on Matplotlib.
    - URL: https://seaborn.pydata.org
    - Accessed: November 2024
9. **ASCVD Risk Calculator**
    - Version: GitHub Repository
    - Description: Python implementation of the ASCVD Risk Calculator for cardiovascular risk prediction.
    - URL: https://github.com/brandones/ascvd/tree/master
    - Accessed: November 2024

Datasets

1. **Cleveland Heart Disease Dataset**
    - Source: UCI Machine Learning Repository
    - Description: Dataset used for binary classification of heart disease presence.
    - URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
    - Accessed: November 2024
2. **VA Long Beach Heart Disease Dataset**
    - Source: UCI Machine Learning Repository

- Description: Dataset used for regional generalization of machine learning models.
- URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
- Accessed: November 2024

Guidelines and Methodological References

1. **Mueller, Andreas C., & Guido, Sarah**
   - Title: *Introduction to Machine Learning with Python*
   - Publisher: O'Reilly Media
   - Publication Date: 2016
   - URL: https://github.com/dlsucomet/MLResources/blob/master/books/[ML]%20Introduction%20to%20Machine%20Learning%20with%20Python%20(2017).pdf
2. **Software Sustainability Institute**
   - Title: *How to Cite and Describe Software*
   - URL: https://www.software.ac.uk/how-cite-and-describe-software
   - Accessed: November 2024

Additional Resources for Citing Software and Data

1. **Digital Curation Centre**
   - Title: *How to Cite Datasets and Link to Publications*
   - Authors: Ball, A., & Duke, M.
   - Publisher: Digital Curation Centre
   - Publication Date: 2011
   - URL: http://www.dcc.ac.uk/resources/how-guides/cite-datasets
   - Accessed: November 2024
2. **DataCite**
   - Title: *Why Cite Data?*
   - URL: https://www.datacite.org/
   - Accessed: November 2024