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

November 25th, 2024

Abstract

This capstone project investigates the application of machine learning models for predicting heart disease, emphasizing their sex-specific and regional generalizability. Using the Cleveland and VA Long Beach datasets from the UCI Machine Learning Repository, this study evaluates the predictive accuracy of machine learning algorithms and compares them to traditional cardiovascular risk scores, such as the ASCVD. The research explores whether machine learning models can outperform conventional methods while addressing demographic and biological variations to ensure equitable healthcare outcomes.

Machine learning algorithms, including Random Forest, XGBoost classifiers, and an ensemble method combining these models, were applied to datasets preprocessed with tailored transformations. These included three distinct transformations: the first transformation included logarithmic normalization of skewed features and squared transformations to emphasize non-linear relationships The second transformation involved the creation of engineered features, such as combining ST depression (Oldpeak) with the slope of the ST segment. Lastly, the third transformation attempted to utilize a custom feature to account for gender differences. Exploratory data analysis highlighted the importance of features such as chest pain type, ST depression (Oldpeak), and exercise-induced angina, while demonstrating limited predictive value for cholesterol and fasting blood sugar. The best transformation strategy combined these methods to enhance predictive accuracy.

The Random Forest model from Transformation 2 achieved superior performance compared to the ASCVD score, with accuracy rates of 83.33% for Cleveland and 82.5% for Long Beach. However, disparities were observed in gender-specific evaluations: while female subgroup models demonstrated higher recall, they suffered from reduced precision, indicating

inconsistencies in predictive performance across sexes. Regional analysis revealed that optimized models sometimes overgeneralized to the majority class, underscoring challenges in maintaining balanced performance across different datasets.

Findings suggest that while machine learning models hold promise for improving heart disease prediction, gender imbalances and regional variations in datasets limit their generalizability. Future research using this dataset should explore multilabel classification for varying disease severities and address sampling biases to enhance fairness and applicability across diverse populations.

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)

The central focus of my capstone project is to explore the effectiveness of machine learning models in predicting heart disease and assess their ability to generalize across different regions and biological sexes. Additionally, the project investigates whether machine learning models can identify cardiovascular disease (CVD) as accurately as, or better than, traditional risk score assessments. This research highlights the importance of building models that not only achieve high accuracy within a specific dataset or geographic location but also remain robust and reliable when applied to diverse populations, accounting for demographic, geographic, and biological variations. Ensuring the generalizability of these models is critical for their application in real-world scenarios.

Traditional cardiovascular risk scores, while widely used, often face significant limitations. Studies have shown that these 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 to 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 are often 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, especially by male

physicians, leading to delayed or insufficient treatment. To address these challenges, this project leverages machine learning to create models trained on data from both men and women, while testing performance separately for each gender. Additionally, separate models will be trained and tested exclusively on women and men to evaluate their specific predictive capacities. By addressing the limitations of traditional risk scores and accounting for biological sex differences, this project aims to enhance the accuracy and fairness of heart disease prediction, contributing to more equitable healthcare outcomes worldwide.

Literature Review

Heart Disease Prediction Methods

Several cardiovascular risk assessment models, each with distinct methodologies and applications, can be used to predict the likelihood of heart disease. Understanding the strengths and limitations of these models is crucial for effective risk stratification and prevention strategies.

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 SCORE equation, endorsed by the Third Joint European Task Force on cardiovascular prevention, has been validated in Spain. In the UK, the QRISK algorithm was created using a population-based clinical research database. Germany contributed both a simple PROCAM score and a more complex neural network model, with the PROCAM score recently updated. Scotland developed the 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

Cardiovascular disease is the leading cause of death globally, responsible for 17.7 million deaths in 2015, or 31% of all deaths. By 2030, this number is expected to rise to over 23.6 million annually (Talha, Elkhoudri, and Hilali, 2024).

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 relative prognostic 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). 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 literature appears affected by outcome selection and optimism biases, emphasizing the need for more rigorous and unbiased comparisons (Siontis, Tzoulaki, Siontis, et al., 2012).

Female Disadvantages in assessing cardiovascular disease

Cardiovascular disease (CVD) rates are higher in men, 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 standard (Abdullah, Beckett, Wilson, et al., 2024).

Additionally, a study reviewed the evidence on gender bias in CVD diagnosis, prevention, and treatment. Following 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).

Pre-existing Machine Learning Methods for Predicting Cardiovascular Risk

Predicting cardiovascular disease risk is essential for prevention. This 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, significantly 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. 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 are evaluated to assess whether the patient’s gender affects the structure and performance on a diagnosis model for coronary artery disease. The accuracy of the male diagnosis model was 95%, with a sensitivity of 96% and specificity of 100%. The accuracy of the female diagnosis model was 96%, with a sensitivity of 97% and specificity of 96%. The high-performance results prove the success of the proposed gender-based approach for the diagnosis of coronary artery disease. The dataset used for this project is also the UCI machine learning repository and named as “heart disease dataset and Z-Alizadeh Sani dataset.” This dataset is comprised of 270 patients’ records, each with 75 attributes (Hogo, 2020).

Supervised Machine Learning

This project will use a supervised machine learning algorithm 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. This machine learning technique is optimal for classification tasks. The results of different machine learning models, such as a Random Forest Classifier, and XGB Classifier, will be compared, with cross-validation used to ensure the reliability of the performance estimates. Machine learning models will be 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 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

Dataset

The University of California at Irvine has a public dataset repository donated in 1988 available in the UC Irvine Machine Learning Repository. There are four databases available from 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 we will be doing here. The datasets donated by UC Irvine consists of 76 attributes. However, the processed Cleveland dataset consists of a subset of 13 of them. For the purposes of this project, we will be using the Cleveland database, as well as the VA Long Beach Dataset.

For this analysis, on the processed Cleveland dataset, which includes **303 patient records** with varying levels of heart disease severity. An additional dataset from the United States was contributed 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 the dataset, the presence of heart disease is indicated by values ranging from 1 to 4, corresponding to increasing severity, while a value of 0 denotes 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 the project will be on analyzing binary classification.

This dataset provides detailed clinical and diagnostic 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 trestbps (resting blood pressure in mmHg), chol (serum cholesterol level in mg/dL), and fbs, 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 downsloping). The ca variable indicates the number of major vessels (0–3) visualized via fluoroscopy, while thal reflects thallium stress test results (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, clinical, and diagnostic variables, making it suitable for predictive modeling and exploring the factors influencing heart disease outcomes.

To ensure privacy, patient names and social security numbers were removed from the database and replaced with dummy identifiers.

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. It requires no setup, provides free access to GPUs for accelerated computation, and facilitates easy sharing of projects. With its user-friendly interface, Colab is widely used by students, data scientists, and researchers to develop and run machine learning models and data analyses. For this project, I will utilize Google Colab to write and execute code, leveraging its computational resources and collaborative features to streamline the analysis and modeling processes.

Coding Language

Python, a high-level and versatile programming language, will be a central tool for this project. Renowned for its simplicity and extensive capabilities, Python facilitates efficient code development for tasks such as data analysis, machine learning, and scientific computing (Python Software Foundation, 2024). With its robust standard library and a wide array of third-party packages, Python provides the necessary tools to handle complex data processing, build predictive models, and perform statistical analyses. Leveraging Python's open-source platform and its active community support, I will utilize it to execute the analytical and modeling components of this project effectively.

Python Libraries

This project employs several Python libraries that are essential for data manipulation, analysis, visualization, and machine learning. These libraries form the backbone of the notebook's capability to handle binary classification tasks efficiently. 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 like 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, enabling a deeper understanding of data relationships and patterns.

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

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 transformation, capturing variability in 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 across different biological groups.

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, 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, outlier detection methods were applied to identify and address anomalies that could skew 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 (trestbps), 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, in regard to 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 and robust 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 tailored imputation strategies was implemented to handle missing values based on their data characteristics. Numerical columns with a normal distribution, such as trestbps 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 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 critical challenge. Continuous numerical columns (trestbps, 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 rows with residual 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 datasets' integrity, ensuring they were free of anomalies and inconsistencies.

Data Transformation

Three distinct transformation strategies were applied to the Cleveland dataset to evaluate their impact on model performance. These transformations 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 transformation strategy was subsequently applied to the VA Long Beach dataset to ensure consistency and generalizability.

Transformation Experiments on Cleveland Dataset

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. 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 (trestbps) 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.

Model Evaluation

Three machine learning models—Random Forest Classifier, XGBoost Classifier, and an Ensemble Method combining the predictions from these two classifiers plus a linear 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

optimization, 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.

Best Transformation Selection for Cleveland

To determine the most effective transformation 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 transformation strategy effectively balanced predictive accuracy with robustness across the evaluation metrics, making it the optimal choice for subsequent application to the VA.

Gender-Based Testing on Cleveland Dataset

To evaluate the performance of the selected transformation 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 transformation (Transformation 2) were

assessed separately for each gender. Key performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, were calculated for the male and female subsets to understand how the model performed across these demographic groups. Additionally, statistical tests, such as t-tests, were conducted to determine whether observed performance differences between the male and female subgroups were statistically significant. 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.

Application to VA Long Beach Dataset

The best-performing transformation 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 using the same schema developed for 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 geographic regions.

Application of the ASCVD Risk Calculator to the Cleveland and VA Long Beach Datasets

To further enhance the analysis for predictive modeling, 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 analysis required the preparation of essential 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 inputs required by the ASCVD model.

Results

Exploratory Data Analysis Results

Basic Descriptives of the 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 the various 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). Resting electrocardiographic results (restecg) have an average of 0.983, with values ranging from 0 to 2. 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 (likely 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 different slope categories. These descriptive statistics highlight key trends and distributions in the dataset, informing subsequent analysis and model development.

Univariate analysis of the training set

The dataset reveals several key patterns about the participants and their heart health indicators. Most participants are middle-aged, falling between 55 and 65 years old, with males making up roughly two-thirds of the dataset. When it comes to those who experience chest pain, the majority experience 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 have elevated fasting blood sugar, suggesting that this issue is less common in the dataset. Resting electrocardiogram results appear to be normally distributed, showing a balanced spread of values.

The exang variable, which indicates the presence of exercise-induced angina (1 for presence, 0 for absence), reveals that most participants do not experience angina during exercise. This 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 those at higher risk for underlying heart issues, as angina during exercise is often a significant indicator of coronary artery disease.

The maximum heart rate distribution has a strong left skew, meaning 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, a smaller group with higher values (up to 6) may be at greater risk for heart issues.

The slope feature, which describes the shape of the ST segment during peak exercise, shows that most participants have a flat slope (category 2), often associated with underlying heart problems. A smaller number have an upsloping slope (category 1), typically linked to better heart health, or a down sloping slope (category 3), which is more often connected to severe heart disease. Together, these features provide valuable insights into the varying levels of heart health risk within the 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 experiment, 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.

Relationships between distinct variables

A series of scatter matrices were created to understand the relationship between distinct variables. The diagonals of the pair plot contain histograms depicting the distribution of specific features. Off-diagonal scatter plots show correlations between pairs of variables. For example, age and cholesterol show a clear trend in which greater cholesterol levels are related with older individuals.

The pair plot provides an overview of feature distributions and relationships within the Cleveland dataset, revealing key patterns and correlations. Continuous features, age and maximum heart rate (thalach), exhibit a mild negative correlation, while Oldpeak (ST depression) and Slope show a stronger linear relationship, supporting their combination in feature engineering. Most features, such as cholesterol (chol) and resting blood pressure (trestbps), display right-skewed distributions with visible outliers. Categorical variables, including sex and chest pain type (cp), are clearly separated into distinct groups.

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 several significant relationships between clinical features 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 notable disparities, with males more likely to have heart disease and females predominantly without it. This association was statistically significant (p-value: 3.26e-04). Similarly, chest pain types showed a strong relationship with heart disease presence, particularly type 4 (asymptomatic chest pain), which was more common among individuals with heart disease (Chi-Square test statistic: 65.29635, p-value: 4.33e-14).

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). While cholesterol levels were slightly higher on average among those with heart disease, 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, as most individuals in both groups had normal levels (Chi-Square test statistic: 0.00129, p-value: 0.9710). Resting electrocardiogram results (restecg) demonstrated a strong association, with abnormal results (restecg = 2) being more common in individuals with heart disease, while

normal results (restecg = 0) were prevalent among those without (Chi-Square test statistic: 35.63319, p-value: 2.38e-09).

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) also emerged as a significant indicator, being 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) exhibited a strong relationship with heart disease, with higher Oldpeak values and certain slope categories (e.g., slope = 2) more frequently observed among those with heart disease (p-value < 0.05). These features were closely correlated (correlation coefficient: 0.59), reflecting their shared relationship with stress test outcomes and heart ischemia markers.

The correlation heatmap further illustrated key patterns, including the positive relationship of the number of major vessels (ca) with age and thal, and the negative correlation of age and maximum heart rate. Notably, weak or negligible correlations were observed for cholesterol and fasting blood sugar, suggesting limited relevance within the dataset's structure.

In summary, the analysis emphasizes 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 association of age with vascular blockages and declining cardiovascular efficiency, were evident, weaker relationships for variables like cholesterol and fasting blood sugar highlight the need for a focused approach when analyzing predictors of heart disease.

Model Performance on Cleveland Dataset

Results from multiple train-test splits 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 and the Ensemble Method emerged as the top performers.

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.

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.

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.

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 new model, without optimized parameters, demonstrated an improvement in terms of its ability to predict instances of the minority class (those with heart disease). While the overall accuracy is slightly higher at 82.5%, compared to the 75% accuracy of the optimized model, the primary improvement lies in its balanced performance across both classes.

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.

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.

Atherosclerotic Cardiovascular Disease Risk Calculation on Cleveland Dataset

The 2013 ASCVD (Atherosclerotic Cardiovascular Disease) risk score was evaluated on the Cleveland dataset, yielding key performance metrics. 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 most 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. This variability suggests that further optimization or subgroup-specific adjustments may be necessary to ensure more balanced and equitable performance across genders.

Atherosclerotic Cardiovascular Disease Risk Calculation on VA Long Beach Dataset

The 2013 ASCVD (Atherosclerotic Cardiovascular Disease) risk score was evaluated on the VA Long Beach dataset, 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 the majority of 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.

Discussion

Key Findings:

This study leveraged the Cleveland and VA Long Beach datasets to explore the binary classification of heart disease presence, with a focus on demographic and clinical predictors. 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 (oldpeak 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 balanced 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 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 meaningful sex-specific analysis, underscoring the importance of diverse and balanced datasets.

Gender-based analysis on the Cleveland dataset revealed clear disparities in model performance, with the Random Forest Classifier achieving higher recall but lower precision for females. 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, slope, chest pain type, and exercise-induced angina 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 regional generalizability.

Regional Generalizability

The comparison between the Cleveland and VA Long Beach models without optimized parameters highlights two possible implications: optimized parameters may lead to overgeneralization for the majority class in the VA Long Beach dataset (presence of heart

disease), or they may reduce the model's ability to generalize 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 achieved balanced performance for the minority class on the Cleveland dataset, with a recall of 61% and an F1 score of 0.69, 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 will need to be removed from the analysis.

To ensure a fair comparison when evaluating models, I had to modify the Cleveland models 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 of whether the success of these models in the Cleveland dataset translated effectively to data from different regions.

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

This study 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, this study 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 are 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 targeted and interpretable 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.

Data Management Plan Overview

Our data management plan ensures the organized, secure, and ethical handling of all project data. We will acquire datasets from the UCI Machine Learning Repository and follow their terms of use. The data will be stored securely on a personal computer. We will document 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 will be restricted to authorized project members only. Upon project completion, we will submit 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 will ensure long-term preservation and accessibility of our work. For detailed guidance on data management and submission, we will refer to the library's resources available on their website.

Bibliography

1. 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/.
    a. In-Text Citation: (Abdullah, Beckett, Wilson, et al., 2024)
2. 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
    a. In-text citations: (Calomfirescu and Calomfirescu, 2012)
3. 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/
    a. In-text citations: (Cesare and Lawrence, 2022)

4. 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.
    a. In-text Citation: (Cho, Kim, Kang, et al., 2021)
5. 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.
    a. In-text Citation: (Hogo, 2020)
6. Janosi, Andras, et al. “Heart Disease.” UCI Machine Learning Repository, 1988, archive.ics.uci.edu/dataset/45/heart+disease.
    a. In-Text Citation: (Janosi, Steinbrunn, Pfisterer, et al., 1988)
7. 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.
    a. In-Text Citation: (Mueller and Guido, 2016)
8. 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
    a. In-Text Citations: (Shaw, 2012)
9. 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

a. In-text Citation: (Siontis, Tzoulaki, Siontis, et al., 2012)

10. 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/

    a. In-Text Citations: (Sofogianni, Stalikas, and Antza, 2022)

11. 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/

    a. In-Text Citation: (Talha, Elkhoudri, and Hilali, 2024)

12. 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/.

    a. In-text Citation: (Weng, Reps, Kai, et al., 2017)

13. 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/.

    a. (Woodward, 2019)

14. “Scikit-Learn Machine Learning in Python.” *Scikit-Learn*, scikit-learn.org/stable/.