2024

# Machine Learning in Healthcare: Heart Disease

MASTER OF SCIENCE IN INFORMATION MANAGEMENT (MSIM)
UNIVERSITY OF WASHINGTON
GROUP 3: MARIE CLAIRE O'CONNELL, TIMOTHY MBUTHIA, NEBIAT ABRAHA

IMT 574 A: DATA SCIENCE II: MACHINE LEARNING SPRING 2024

# Contents

Executive Summary....2
Practical Explanation....3
Importance of Heart Disease in Society....4
Importance of Heart Disease in Technology....4
Importance of Heart Disease in Business....5
Theoretical Explanation....5
Support Vector Machine (SVM)....7
Random Forest....8
Data Processing and Computation....8
Introduction....8
Data Exploration and Understanding....10
Data Preprocessing....12
Model Selection and Training....15
Model Evaluation....19
Conclusion....23
References....24

# Executive Summary

Heart Disease is one of the leading causes of death and remains has a significant impact to economies globally. This poses a significant challenge to healthcare systems and society since treatment is costly and unsuccessful at times. One challenge is not being able to catch it beforehand, preventing its exacerbation. The purpose of this machine learning project is to build and evaluate a predictive model that can identify patients presenting with or at considerable risk for heart disease. With the use of classification algorithms like Support Vector Machine (SVM), this project supports early diagnosis and intervention, which can improve patient outcomes and reduce healthcare costs.

The data used to train and test the SVM model was sourced from the Cleveland Clinic Foundation database, pulled from the UC Irvine Machine Learning Repository. The data originally had 76 features, however only a subset of 14 was available for use. The metrics used to evaluate model performance were accuracy, precision, and recall.

A few models were testing Logistic Regression, Random Forest, and SVM for accuracy to start. After running a few different randomization searches and reviewing the visualizations, the focus was turned to improving the SVM model. The SVM Radial Basis Function (RBF) model shows strong performance with an accuracy of 0.89 and high F1 scores for both training and testing datasets. These metrics suggest that the model is well-tuned and capable of making accurate predictions.

With the insights gained from predictive machine learning models, healthcare professionals could potentially detect illnesses in patients earlier and implement

customized proactive or preventative treatment plans based on risk analysis. Using LIME (Local Interpretable Model-Agnostic Explanations), medical providers can use a model to determine which features are aiding a patient's heart conditions and provide better care.

The complete project report will explore the theoretical foundations and practical implementations of the SVM model, including data processing techniques, the significance of each feature, and the broad impacts of the findings. This comprehensive analysis will provide a thorough understanding of how predictive machine learning models can influence early diagnosis and personalized treatment in healthcare.

## Practical Explanation

Globally, heart disease is one of the main causes of hospitalization and death. It significantly accounts for approximately 17.9 million deaths annually (WHO, 2019). As per the American Heart Association (2020), coronary heart disease was the leading cause (42.6%) of deaths attributable to cardiovascular disease in the US in 2017. This project aims to prove predictive modeling emerges as a ray of hope as we grapple with the complexity of this worldwide epidemic, providing game-changing answers in the fields of business, technology, and society. Machine learning has been touted as an essential development that can improve clinical decision-making, optimize resource allocation, and, most significantly, help people with cardiac diseases to live long and comfortable lives. (Abbas,2024)

This report will cover the importance of heart disease predictive modeling in society, technology, and business areas, highlighting its significance for public health, innovation, and economic growth.

Predictive models can enhance patient outcomes and lower healthcare costs by assisting in the early detection and avoidance of cardiac disease. These algorithms can identify those individuals who are at considerable risk and facilitate fast action through the examination of many risk elements and patient data. By taking proactive steps, the impact of heart disease on society and medical care can be minimized.

## Importance of Heart Disease in Society

Heart disease can cause mental anguish, financial strain, and reduced standards of life for people, families, and communities. (McHorney et al., 2021). Through early detection, prompt intervention, and individualized treatment regimens, healthcare professionals can lessen these burdens by utilizing predictive models, ultimately improving quality of life and lowering society's healthcare expenditures.

## Importance of Heart Disease in Technology

Developments in data mining, machine learning, and artificial intelligence approaches have made it easier to construct predictive models for cardiac diseases by leveraging electronic health records. This can be transformed to identify patterns in heart disease, enabling more accurate predictions and personalized treatment plans (Habehh et al., 2021). Healthcare providers and insurance businesses can deliver more targeted interventions and proactive healthcare strategies, thereby enhancing patient outcomes and advancing the frontier of personalized medicine (Chornyy , 2019)

Currently, a portion of the population is accustomed to wearing health devices. The wearables and mobile health applications monitor patients' vital signs and provide real-time feedback, encouraging healthy behaviors and reducing the risk of heart disease. (Moshawrabet al., 2023). The devices also offer insight to medical professionals to keep track of a patient's recovery and deliver patient-customized interventions and proactive healthcare strategies, thereby improving outcomes and lowering recovery periods.

### Importance of Heart Disease in Business

As highlighted earlier, heart disease is a major cause of death that carries significant healthcare costs globally, with estimated annual costs of over $1 trillion in the United States alone (American Heart Association, 2020). Predictive models would be crucial to facilitate early intervention, lower hospitalization rates, and improve resource allocation. Predictive analytics provide a viable option for companies in the healthcare industry to cut expenses overall, simplify healthcare delivery, and allocate resources optimally.

These models might accelerate the availability of efficient medicines and therapies for heart disease by aiding drug development and clinical trial processes. Innovation and economic growth can be stimulated by the creation of novel heart disease therapies and technology (Bohr et al., 2020).

## Theoretical Explanation

This data science project centers around medical diagnosis, specifically, the diagnosis of heart disease. Heart disease is a term often used to describe types of

conditions that affect the heart and vessels. The data utilized in this analysis was sourced from the Cleveland Clinic Foundation database and was pulled from the UC Irvine Machine Learning Repository. The data originally had 76 features, but only a subset of 14 were available for use. Machine learning is useful in medical diagnostics because it is an efficient way to analyze large multi-dimensional datasets to classify data or provide predictive analysis. This means that health care professionals could potentially detect illnesses in patients earlier and implement customized treatment plans that are proactive, or preventative based on risk analysis. Machine learning algorithms can dramatically impact the approach to health care for the better because they can ensure accurate and reliable models that can be used for diagnosis.

There are several machine learning algorithms that can evaluate medical data; however, classification is particularly valuable. By analyzing a patient's medical history and current health status, classification algorithms can identify patterns, assess risk levels, and categorize patients into subgroups. Customized treatment plans can then be implemented by subgroups, and the symptoms experienced. Often in medical diagnostics, the variable of interest is whether a patient has a disease or not, for example, whether a patient has heart disease or not. For this data science project, there were initially two classifiers used to study heart disease, Support Vector Machine (SVM) and Random Forest. However, after some preliminary analysis, SVM provided promising results and was selected as the main classifier for this project.

## Support Vector Machine (SVM)

Support Vector Machine, SVM, is a supervised learning method that can be used for classification and regression. SVM is used as a binary classification algorithm for this data science project because it is focused on medical diagnostics with a target variable that has two possible classes. When applied to this heart disease dataset, SVM works to find the optimal hyperplane, the plane with the largest margin, that best separates the data into the two classes of heart disease or no heart disease. Once SVM finds the optimal hyperplane and maximizes the margins, a decision boundary is created that categorizes patient data with more accuracy, prevents overfitting, and is well suited to handle unseen data.

Medical datasets, like the one used in this project, often have a high number of features because of the amount of patient data and clinical measurements recorded. SVM is a popular machine learning algorithm used in healthcare because of its ability to handle high-dimensional datasets. SVM has several hyperparameters that control the behavior of this algorithm. For example, the Kernel type hyperparameter allows SVM to model non-linear and complex relationships between variables which is often seen in high dimensional medical datasets. For this project, other hyperparameters like C, Kernel, Gamma, Degree and Coef0 were tuned to find the optimal hyperplane, thus optimizing the model's performance.

It is imperative that healthcare professionals understand the logic behind machine learning predictions, and with SVM the decision boundaries, support vectors and margins all provide insight into how the output is calculated. The combination of

high accuracy, flexibility and interpretability makes SVM a valuable tool in improving diagnostic outcomes and supporting clinical decision-making.

### Random Forest

Though it is not the main tool for analysis of this heart disease dataset, Random Forest is another powerful classifier used in medical diagnostics because of its ensemble nature. Ensemble learning is a machine learning technique that aggregates the predictions of multiple models to produce an accurate and reliable final model. Random Forest is a machine learning method that falls under the category of ensemble learning because it builds many decision tree models and then aggregates and outputs the classes or predictions of multiple trees. It uses a technique called Bootstrap Aggregating or bagging which trains each tree in the Random Forest model on a random subset of data and splits on a random subset of features. This means each tree is trained on different data, creating a diverse set of models that can potentially uncover hidden patterns in the dataset. Once each tree has classified the data, the classification predictions are aggregated, and the final output is chosen. Like SVM, Random Forest is a powerful machine learning technique used in medical diagnostics because it is reliable, accurate, interpretable, can handle unseen data and works well with numerous features.

## Data Processing and Computation

### Introduction

#### Objective

The goal of the machine learning project was to build a model to determine the probability of angiographic coronary artery disease. Concentrating on simply distinguishing presence (value 1,2,3,4) from absence (value 0). (Andras Janosi, n.d.)

## Scope

This paper aims to explore and understand a dataset from the Cleveland Clinic Foundation to develop a predictive model for diagnosing heart disease. The scope includes:

Data Exploration and Understanding: Summarize the dataset, including instances, features, and preprocessing steps. Provide descriptive statistics and a data dictionary for insights into feature distributions and relevance.

Data Preprocessing: Ensure data integrity by handling errors, duplicates, and missing values. Select relevant features using statistical tests and domain knowledge. Engineer new features to enhance predictive power.

Model Development: Standardize the dataset for model training. Implement pipelines to streamline preprocessing and modeling steps.

Model Selection and Training: Compare initial models (Logistic Regression, Random Forest, SVM) to identify the best approach. Optimize SVM hyperparameters through randomized search for improved performance.

Model Evaluation: Use confusion matrices and classification reports to assess model performance. Apply LIME for interpretability and understanding of feature importance in the SVM model.

Conclusion: Summarize findings, highlighting the SVM model's performance with an RBF kernel.

Emphasize the importance of data preprocessing, feature selection, and model optimization in developing predictive models. This research provides a comprehensive methodology for developing and evaluating a heart disease diagnostic model, demonstrating best practices in machine learning.

## Data Exploration and Understanding

### Data Overview:

Number of Instances:

Cleveland: There were 303, but after dropping the rows with NA to run the Random For est model, there were 297.

| Name | Role | type | Demographic | Description | Units | Missing Values |
|---|---|---|---|---|---|---|
| **age** | Feature | Integer | Age | age in years | years | no |
| **sex** | Feature | Categorical | Sex | 1 = male; 0 = female | | no |
| **cp** | Feature | Categorical | | Chest Pain Type: Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic | | no |
| **trestbps** | Feature | Integer | | Resting blood pressure (on admission to the hospital) | mm Hg | no |
| **chol** | Feature | Integer | | serum cholesterol | mg/dl | no |
| **fbs** | Feature | Categorical | | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) | | no |
| **restecg** | Feature | Categorical | | resting electrocardiographic results, Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria | | no |
| **thalach** | Feature | Integer | | maximum heart rate achieved | | no |
| **exang** | Feature | Categorical | | exercise induced angina (1 = yes; 0 = no) | | no |

| | | | | | | |
|---|---|---|---|---|---|---|
| **oldpeak** | Feature | Integer | | ST depression induced by exercise relative to rest | | no |
| **Slope** | Feature | Integer | | the slope of the peak exercise ST segment: Value 1: upsloping, Value 2: flat, Value 3: down sloping | | no |
| **ca** | Feature | Integer | | number of major vessels (0-3) colored by fluoroscopy | | no |
| **thal** | Feature | Integer | | 3 = normal; 6 = fixed defect; 7 = reversable defect | | no |
| **num** | Target | Integer | | diagnosis of heart disease | | no |

## Initial Inspection

Below is a table with the description of each of the features and target data.

| Info Type | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | num |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 | 297 |
| **mean** | 54.5 | 0.68 | 3.16 | 131.69 | 247 | 0.14 | 0.9966 | 149.6 | 0.33 | 1.0556 | 1.6 | 0.68 | 4.73 | 0.95 |
| **std** | 9.05 | 0.47 | 0.96 | 17.763 | 52 | 0.35 | 0.9949 | 22.942 | 0.47 | 1.1661 | 0.62 | 0.94 | 1.94 | 1.23 |
| **min** | 29 | 0 | 1 | 94 | 126 | 0 | 0 | 71 | 0 | 0 | 1 | 0 | 3 | 0 |
| **25%** | 48 | 0 | 3 | 120 | 211 | 0 | 0 | 133 | 0 | 0 | 1 | 0 | 3 | 0 |
| **50%** | 56 | 1 | 3 | 130 | 243 | 0 | 1 | 153 | 0 | 0.8 | 2 | 0 | 3 | 0 |
| **75%** | 61 | 1 | 4 | 140 | 276 | 0 | 2 | 166 | 1 | 1.6 | 2 | 1 | 7 | 2 |
| **max** | 77 | 1 | 4 | 200 | 564 | 1 | 2 | 202 | 1 | 6.2 | 3 | 3 | 7 | 4 |

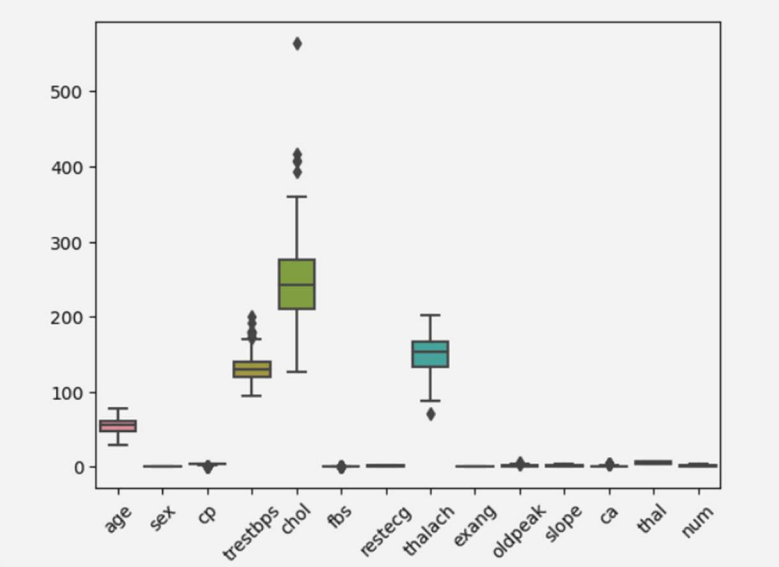
## Data Dictionary

This directory contains one database concerning heart disease diagnosis. All attributes are numeric. The data was collected from the Cleveland Clinic Foundation (Cleveland.data). While the databases have 76 raw attributes, only 14 are used. (Andras Janosi, n.d.)

# Data Preprocessing

## Data Cleaning

Any identified data errors were fixed or corrected to ensure the accuracy and consistency of the dataset. This involved reviewing for duplicates and missing values and resolving inconsistencies with the data. All the data was in the range required for each feature; there were a few rows that had missing data, which were dropped.  There was one outliner that was dropped.



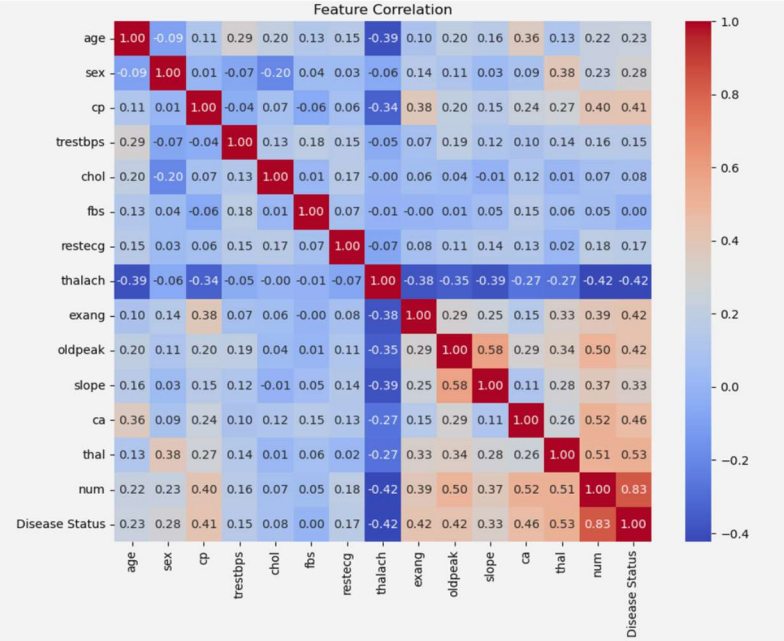

## Feature Selection

Even though the data was already reduced, wanting to ensure the feature selection would work for the model, the data was split into test and train, imported chi2 from sklearn.feature_selection to calculate chi-square statistics and p-values for the training dataset. A DataFrame was created to display the attributes, chi-square statistics, and p-values. See chart below.

| Attribute | Chi-square statistic | P-value |
|---|---|---|
| thalach | 150.4258881 | 1.6499E-31 |
| ca | 64.75266458 | 2.90119E-13 |
| oldpeak | 58.19140002 | 6.95669E-12 |
| thal | 51.52051143 | 1.73761E-10 |
| exang | 35.13569888 | 4.35654E-07 |
| chol | 32.37762546 | 1.60153E-06 |
| age | 26.29676258 | 2.757E-05 |
| trestbps | 23.51807868 | 9.97544E-05 |
| cp | 12.59592766 | 0.013428601 |
| restecg | 11.06536181 | 0.025839176 |
| slope | 6.826918915 | 0.145322391 |
| fbs | 4.901542292 | 0.297549876 |
| sex | 4.122326057 | 0.38970388 |

A higher chi-square statistic indicates a stronger relationship between the attribute and the target variable. A lower p-value, typically below a significance level, e.g., 0.05, indicates higher statistical significance, suggesting stronger evidence against the null hypothesis (no relationship between the attribute and the target variable). After running the above code a few different time, with different random states and test_size when creating the training data, age ,cp, trestbps, chol, restecg, thalach, exang, oldpeak, ca, thal have a P-value of <0.05. Slope, fbs and Sex all have a P-value >0.05.

However, domain knowledge is utilized to identify and select features. Healthcare domain knowledge of coronary artery disease (CAD) states that Slope, Sex, and FBS are important features.




A heat map was created to see if significant corrections existed between any of the selected features. With SVM, if there was high correction between two features, we could drop one without affecting the model and reducing the computational time. Since there were no real corrections between the features, no features were dropped.

## Feature Engineering

The Disease Status feature was created by defining a function heart_disease to convert any number not equal to zero to heart disease if num = 0, then no_disease

Before fitting the SVM model, the dataset is preprocessed using StandardScaler() to standardize the features. This ensures that each feature has a mean of 0 and a standard deviation of 1. Normalizing the features.

A pipeline was also used. Using a pipeline in machine learning workflows offers several advantages, especially when preprocessing steps like feature scaling are

involved, particularly in conjunction with preprocessing steps like StandardScaler() for SVM.

For example, a pipeline provides a convenient way to encapsulate multiple preprocessing steps and the model into a single object. This simplifies the code and makes it easier to understand and maintain. When performing cross-validation or hyperparameter tuning, it's essential to ensure that preprocessing (such as scaling) is applied separately to the training and testing datasets. Pipelines handle this automatically, ensuring preprocessing is applied consistently to both datasets within each cross-validation fold. Using pipelines helps streamline the machine learning workflow, ensures consistency in preprocessing, reduces the risk of errors, and can improve computational efficiency. These benefits make pipelines a common practice in machine learning projects, especially when preprocessing steps like feature scaling are involved.

### Data splitting

When splitting the data, the test_size was set at 30% and  random_state starting at 42.

## Model Selection and Training

The initial models considered were Logistic Regression, Random Forest, and Support Vector Machine (SVM). The accuracy scores obtained were as follows: Logistic Regression achieved 86.67%, Random Forest achieved 84.4%, and both a basic SVM and SVM using a pipeline with StandardScaler achieved 70%.

Despite logistic regression yielding the highest accuracy, the decision was made to further understand SVM for the sake of gaining valuable experience, and it's possibilities.

## Hyperparameter Tuning

A small, randomized search was conducted to determine the best C value and kernel type for the SVM model to ensure we addressed all possible issues. A dictionary was created with the given hypo parameters.

C = uniform (loc=0.0000000001, scale=10) for Uniform distribution

kernel = ['linear', 'poly', 'rbf', 'sigmoid']

gamma= ['scale', 'auto'] Gamma options for 'rbf', 'poly', 'sigmoid'

degree= [2, 3, 4, 5] Degree options for 'poly'

coef0= uniform (loc=0, scale=10) Uniform distribution from 0 to 10 for 'poly' and 'sigmoid'

class_weight= [None, 'balanced'] Class weight options

shrinking= [True, False] Shrinking options

Best parameters found: {'C': 6.235636967959723, 'class_weight': None, 'coef0': 4.3758

72112626925, 'degree': 4, 'gamma': 'scale', 'kernel': 'linear', 'shrinking': True}

These parameters had a cross-validation accuracy of 0.81 ± 0.07

We then did a randomization search just for rbf using the dictionary below.

Kernel= ['rbf']

C=uniform (loc=0.0000000001, scale=10) Uniform distribution from 0.00000001 to 10

gamma= ['scale', 'auto'] Gamma options for 'rbf', 'poly', 'sigmoid'

class_weight= [‘balanced’]

shrinking= [True, False] Shrinking options

probability= [True]

Best parameters found: {‘4.375872112726925, ‘class_weight’: ‘balanced’, ‘gamma’: ‘sca le’, ‘kernel’: ‘rbf’, ‘probability’: True, ‘shrinking’: True}

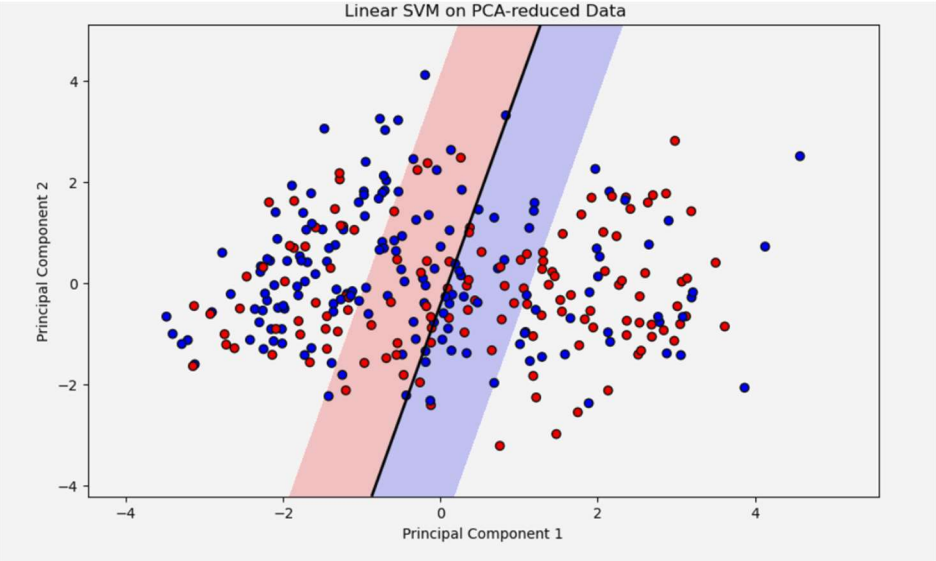
Resulting in a Cross-validated accuracy: 0.69 ± 0.05

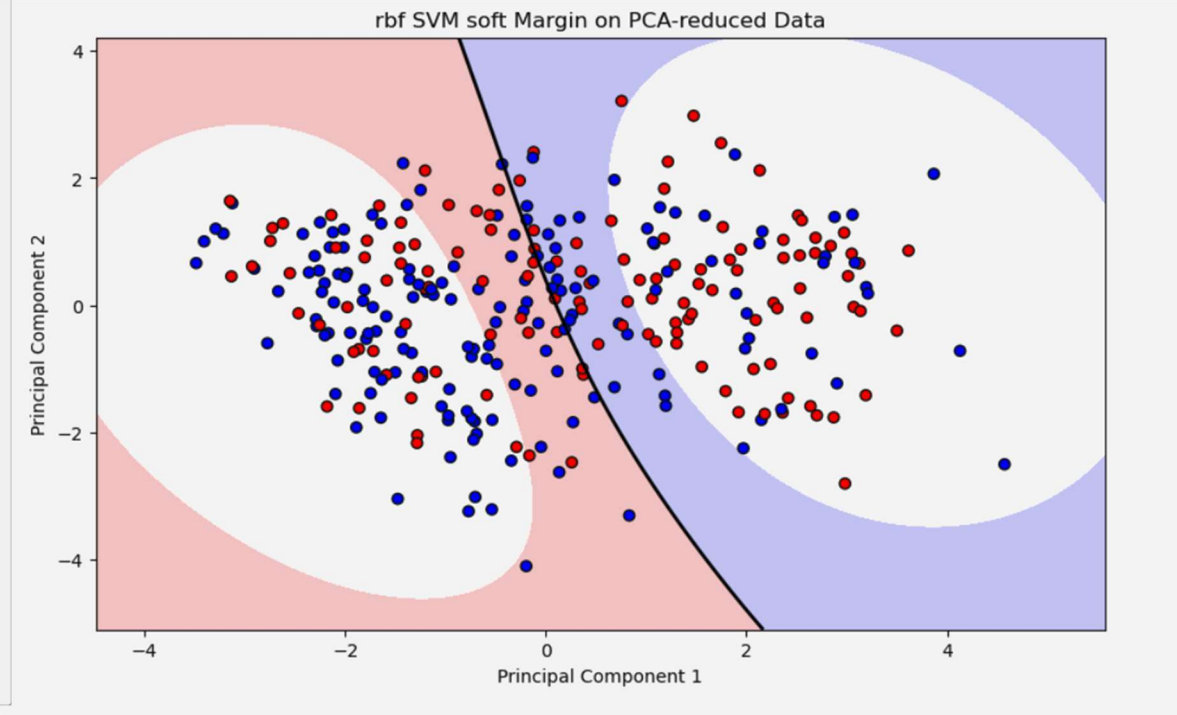
## SVM PCA (Principal Component Analysis)- Reduced Graphs

We utilized Principal Component Analysis (PCA) to visualize the data for the linear SVM and the Radial Basis Function (RBF) SVM models. Our analysis revealed that the linear model failed to capture the data’s complexity adequately. Additionally, the visualization demonstrated significant overlap among the data points, indicating that the data was not linearly separable. A nonlinear decision boundary would be necessary to address the data’s complexity. This involved using a kernel trick (radial basis function kernels) to transform the feature space into a higher-dimensional space where linear separation becomes possible.

Given these findings, we opted for the RBF SVM model, which provided the best performance. The RBF kernel allows for nonlinear decision boundaries, making it better suited to handle the complex and overlapping nature of the data. Through experimentation and evaluation, we determined that the RBF SVM model yielded the

highest accuracy and effectively captured the underlying patterns in the dataset.



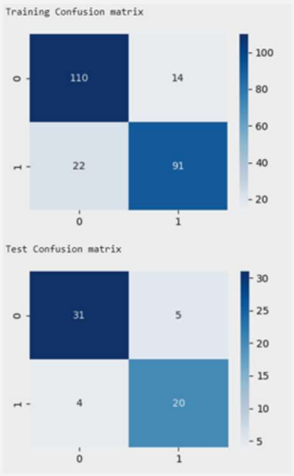

This version clarifies the use of PCA for visualization, specifies the types of SVM models compared (Linear and RBF), and emphasizes the rationale for choosing the RBF SVM model based on the characteristics of the data and the model's performance.

# Model Evaluation

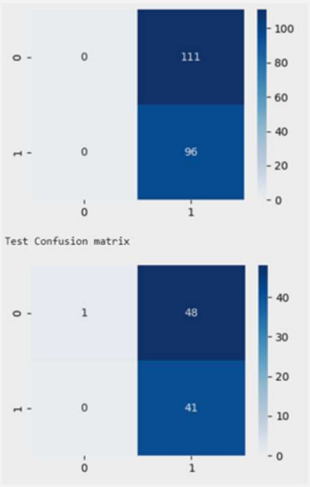
## Performance Metrics

### Confusion Matrixs

**svm.SVC linear**




**svm.SVC Radial Basis Function Hard Margins**




**svm.SVC Radial Basis Function Soft Margins**


### Classification Reports

**svm.SVC linear**

Training Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.89 | 0.86 | 124 |
| 1 | 0.87 | 0.81 | 0.83 | 113 |
| accuracy | | | 0.85 | 237 |
| macro avg | 0.85 | 0.85 | 0.85 | 237 |
| weighted avg | 0.85 | 0.85 | 0.85 | 237 |

Test Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.86 | 0.87 | 36 |
| 1 | 0.80 | 0.83 | 0.82 | 24 |
| accuracy | | | 0.85 | 60 |
| macro avg | 0.84 | 0.85 | 0.84 | 60 |
| weighted avg | 0.85 | 0.85 | 0.85 | 60 |

**svm.SVC Radial Basis Function Hard Margins**

Training Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 111 |
| 1 | 0.46 | 1.00 | 0.63 | 96 |
| accuracy | | | 0.46 | 207 |
| macro avg | 0.23 | 0.50 | 0.32 | 207 |
| weighted avg | 0.22 | 0.46 | 0.29 | 207 |

Test Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.02 | 0.04 | 49 |
| 1 | 0.46 | 1.00 | 0.63 | 41 |
| accuracy | | | 0.47 | 90 |
| macro avg | 0.73 | 0.51 | 0.34 | 90 |
| weighted avg | 0.75 | 0.47 | 0.31 | 90 |

**svm.SVC Radial Basis Function Soft Margins**

Training Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.94 | 0.91 | 124 |
| 1 | 0.92 | 0.87 | 0.89 | 113 |
| accuracy | | | 0.90 | 237 |
| macro avg | 0.91 | 0.90 | 0.90 | 237 |
| weighted avg | 0.90 | 0.90 | 0.90 | 237 |

Test Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.92 | 0.90 | 36 |
| 1 | 0.87 | 0.83 | 0.85 | 24 |
| accuracy | | | 0.88 | 60 |
| macro avg | 0.88 | 0.88 | 0.88 | 60 |
| weighted avg | 0.88 | 0.88 | 0.88 | 60 |

For label ‘1’ (heart disease):

Recall (True Positive Rate) = True Positives / (True Positives + False Negatives)

It represents the percentage of actual heart disease cases that the model correctly predicts as positive.

A high recall indicates that the model is good at identifying true positive cases (minimizing false negatives).

For label ‘0’ (no heart disease):

Recall = True Negatives / (True Negatives + False Positives)

It represents the percentage of actual non-heart-disease cases that the model correctly predicts as negative.

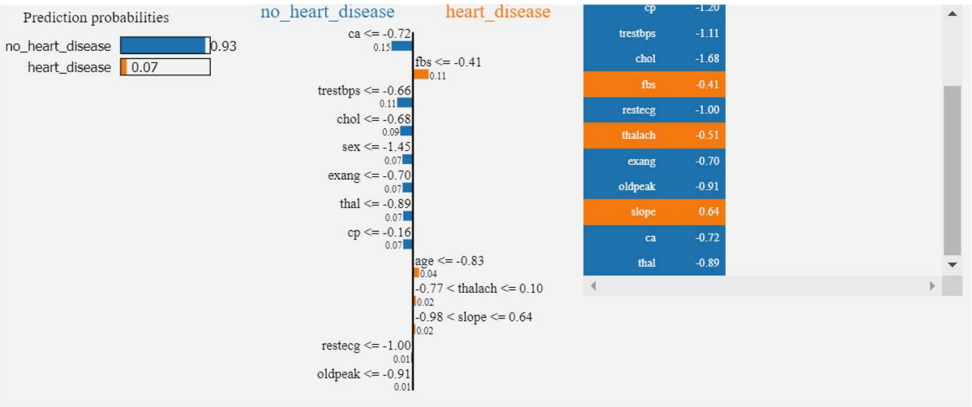
A high recall for label ‘0’ means the model can identify true negative cases (minimizing false positives).

Local Interpretable Model-Agnostic Explanations (LIME)

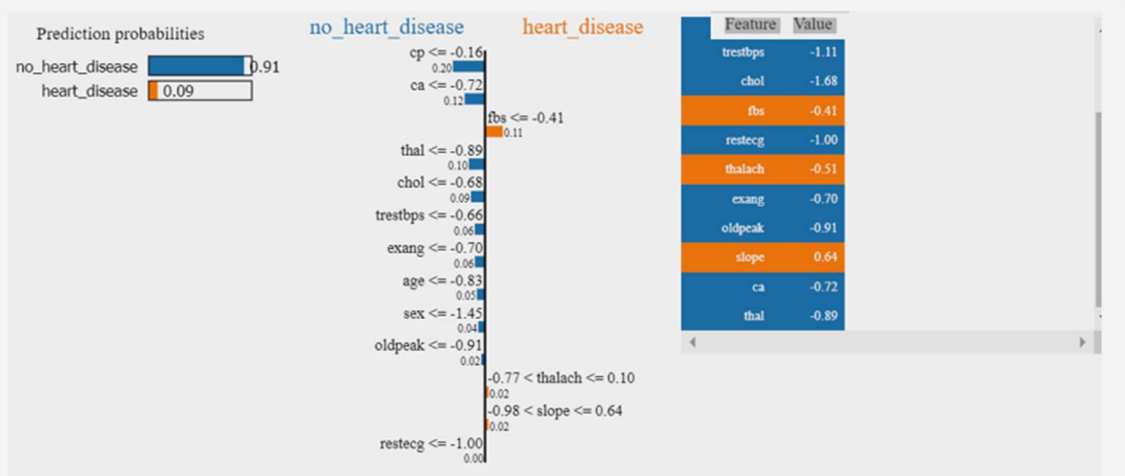
The SVM decision boundary, especially with non-linear kernels, can be challenging to interpret, posing a disadvantage in fields valuing model transparency. Options like LIME and SHAP aid in model review, yet LIME may not accurately depict complex model behavior, particularly in regions with high non-linearities or discontinuities.

**Patient One**

**svm.SVC Linear LIME summary**



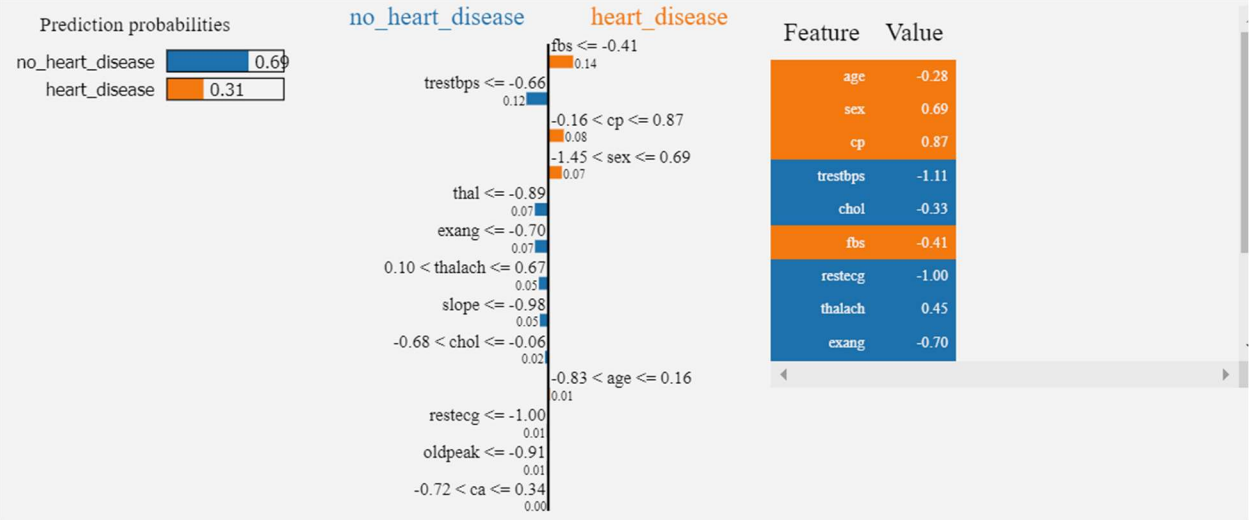
**svm.SVC Radial Basis Function Soft Margins LIME summary**



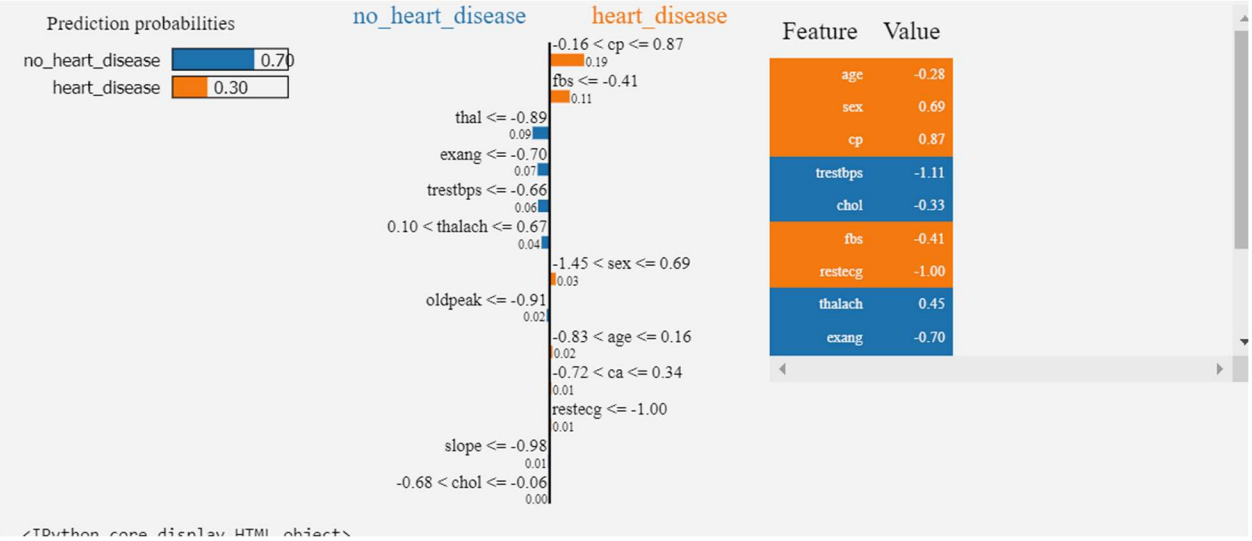

Above is the LIME summary for one patient using the linear and RBF models. displays two bar graphs side by side, illustrating how various features influence heart disease predictions. The left graph represents "no_heart_disease," while the right represents "heart_disease." Each bar, representing a feature such as 'cp' or 'thalach,' indicates its impact on prediction probabilities, with shades of blue contributing to "no_heart_disease" and shades of orange contributing to "heart_disease." For example, a higher 'thalach' (maximum heart rate achieved) increases the probability of "no_heart_disease." Notably, 'slope' is the only positive feature that is an indicator of heart disease. According to this LIME summary for the model, for both models. The patient has a 9% chance of having heart disease if your 'slope' value is between -0.98 and 0.64 in the RBF model and a 7% in the linear model. This is logical because the RBF model is better on true negative.

**Patient two**

**svm.SVC Linear LIME summary**




**svm.SVC Radial Basis Function Soft Margins LIME summary**




For this patient, they have a 30% chance of having heart disease, where the again positive orange features are the main contributors.

## Conclusion

The SVM with linear and Radial Basis Function (RBF) kernel using soft margins achieved the highest accuracy, 85% for the linear, 90% for rbf on the training dataset, respectively, and 85% and 88% accuracy on the test set. Based on accuracy alone, these models perform equally well.

The superior performance of the RBF kernel confirms its suitability for our dataset when using SVM. We started with an initial accuracy of 70% and improved the model to achieve an 89% accuracy.

# References


1. American Heart Association. (2020). Heart disease and stroke statistics—2020 update. Retrieved from https://www.ahajournals.org/doi/10.1161/CIR.0000000000000757
2. World Health Organization. (2019). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
3. McHorney, C. A., Mansukhani, S. G., Anatchkova, M., Taylor, N., Wirtz, H. S., Abbasi, S., Battle, L., Desai, N. R., & Globe, G. (2021). The impact of heart failure on patients and caregivers: A qualitative study. PloS one, 16(3), e0248240. https://doi.org/10.1371/journal.pone.0248240
4. Habehh, H., & Gohel, S. (2021). Machine Learning in Healthcare. Current genomics, 22(4), 291–300. https://doi.org/10.2174/1389202922666210705124359
5. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review. Sensors (Basel, Switzerland), 23(2), 828. https://doi.org/10.3390/s23020828
6. Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare, 25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2
7. Chornyy , R. (2019, December 19). Predictive analytics for better decision-making in Healthcare. Predictive Analytics for Better Decision-Making in Healthcare. https://binariks.com/blog/predictive-analytics-in-healthcare-use-cases-benefits/
8. Abbas, Ghulam H. MD. AI-based predictive modeling: applications in cardiology. International Journal of Surgery: Global Health 7(2):e0419, March 2024. | DOI: 10.1097/GH9.0000000000000419
9. Aurélien Géron (September 2019) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition
10. https://archive.ics.uci.edu/dataset/45/heart+disease is the heart dataset repository
11. lime package — lime 0.1 documentation. (n.d.). https://lime-ml.readthedocs.io
12. Welcome to the SHAP documentation — SHAP latest documentation. (n.d.). https://shap.readthedocs.io
13. Extracting, transforming and selecting features - Spark 2.1.0 Documentation. (n.d.). https://spark.apache.org/docs/2.1.0/ml-features.html
14. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X. https://archive.ics.uci.edu/dataset/45/heart+disease