APR Assignment 1

Heart Disease Classification Using Machine Learning: Logistic Regression and Support Vector Machines

2201AI02
Akash Sinha

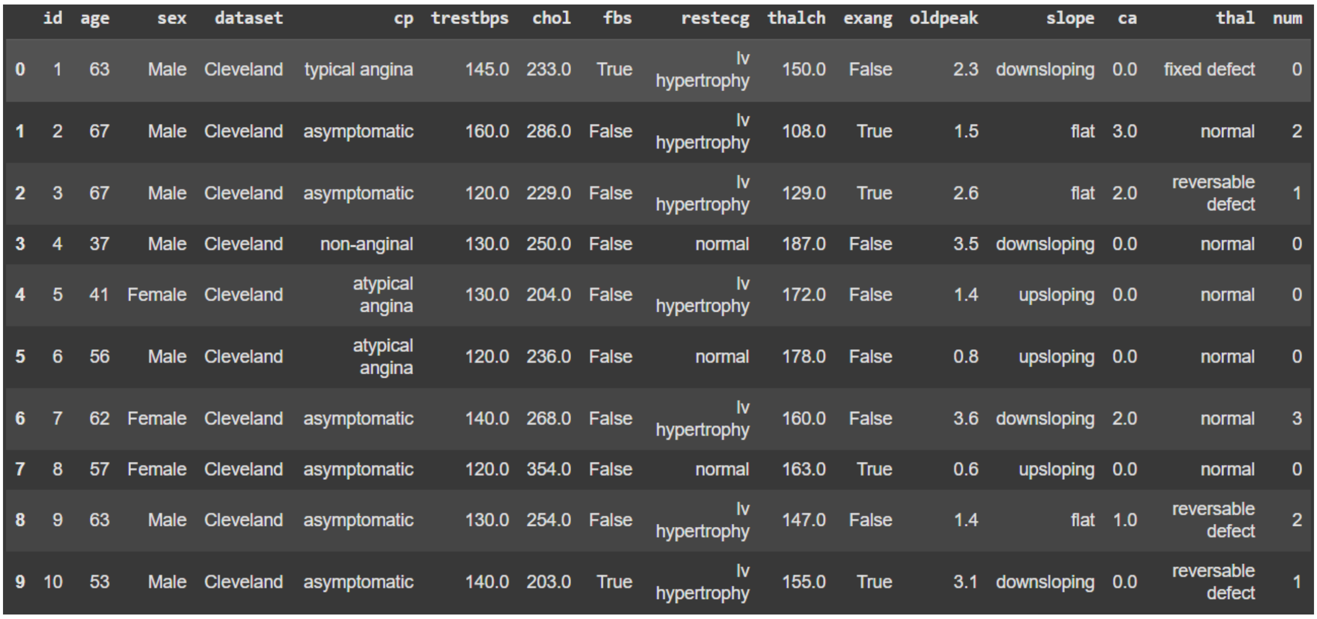
September 19,2025

Introduction

Heart disease is a major medical condition that affects the heart’s structure and function, leading to serious health complications if not diagnosed and managed early. Accurate prediction and diagnosis are crucial for effective treatment and patient care. Machine learning approaches like Logistic Regression and Support Vector Machines (SVM) are widely used to automate and enhance the accuracy of heart disease classification. These models analyse clinical and diagnostic data to efficiently differentiate between individuals with and without heart disease.

Dataset Information

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has heart disease or not.

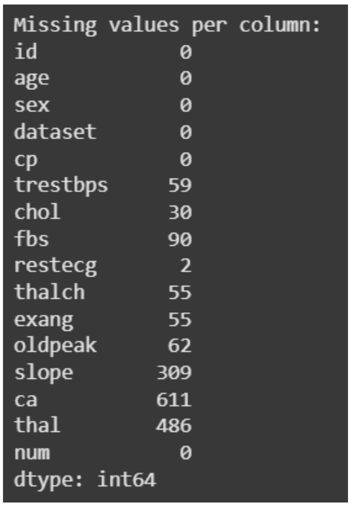


	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0
5	6	56	Male	Cleveland	atypical angina	120.0	236.0	False	normal	178.0	False	0.8	upsloping	0.0	normal	0
6	7	62	Female	Cleveland	asymptomatic	140.0	268.0	False	lv hypertrophy	160.0	False	3.6	downsloping	2.0	normal	3
7	8	57	Female	Cleveland	asymptomatic	120.0	354.0	False	normal	163.0	True	0.6	upsloping	0.0	normal	0
8	9	63	Male	Cleveland	asymptomatic	130.0	254.0	False	lv hypertrophy	147.0	False	1.4	flat	1.0	reversable defect	2
9	10	53	Male	Cleveland	asymptomatic	140.0	203.0	True	lv hypertrophy	155.0	True	3.1	downsloping	0.0	reversable defect	1

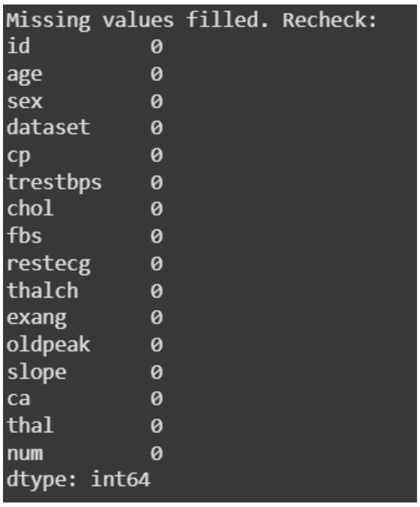
Data Imputation

Real-world clinical datasets often contain missing values, inconsistent formats, and categorical variables. Preprocessing is a crucial step to ensure the dataset is suitable for machine learning algorithms. Common preprocessing steps include:

- **Handling Missing Values:** Numerical missing values can be imputed with the mean or median, while categorical missing values are typically filled using the mode.
- **Binary Conversion:** In the UCL Heart Disease dataset, the target variable may have multiple classes indicating the presence and severity of heart disease. For classification purposes, it is often converted into a binary variable: 0 for no disease and 1 for disease.



```
Missing values per column:
id            0
age           0
sex           0
dataset       0
cp            0
trestbps     59
chol         30
fbs          90
restecg       2
thalch       55
exang        55
oldpeak      62
slope       309
ca          611
thal        486
num           0
dtype: int64
```



```
Missing values filled. Recheck:
id          0
age         0
sex         0
dataset     0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalch      0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
num         0
dtype: int64
```

Data Splitting, Encoding, and Normalization

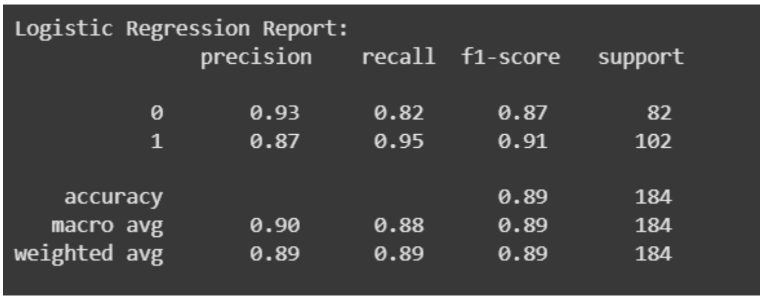
- **Train-Test Split:** Dataset divided into 80% training and 20% testing sets; stratified to preserve class distribution.
- **Feature Types:** Numerical features (e.g., age, cholesterol) and categorical features (e.g., gender, chest pain type) are identified.
- **Normalization:** Numerical features are standardized (mean=0, std=1) to ensure all features contribute equally and improve model performance.
- **Categorical Encoding:** One-hot encoding converts categorical features into binary vectors for machine learning algorithms.
- **Preprocessing Pipeline:** Column Transformer applies normalization and encoding in a single step for consistent and error-free preprocessing

Logistic Regression: Theory Overview

Logistic Regression is a statistical model used for binary classification. It estimates the probability that a given input belongs to a particular class by applying the logistic function to a linear combination of input features. This model is interpretable and effective for problems where the relationship between features and the target is approximately linear. It predicts class membership by choosing a decision threshold on the output probability.

Logistic Regression: Results

The logistic regression model was trained on the prepared dataset and evaluated on the test set. Its performance metrics are summarized below:



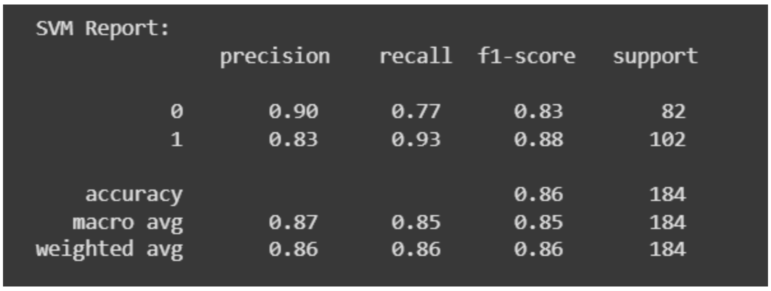
```
Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.93      0.82      0.87        82
           1       0.87      0.95      0.91       102

    accuracy                           0.89       184
   macro avg       0.90      0.88      0.89       184
weighted avg       0.89      0.89      0.89       184
```

Support Vector Machines (SVM): Theory Overview

Support Vector Machine (SVM) is a supervised learning algorithm designed to classify data by finding the optimal hyperplane that maximally separates different classes. It focuses on the margin, defined as the distance from the hyperplane to the nearest data points. The use of kernel functions allows SVM to handle non-linear class boundaries by transforming the data into higher-dimensional feature spaces.

SVM: Results

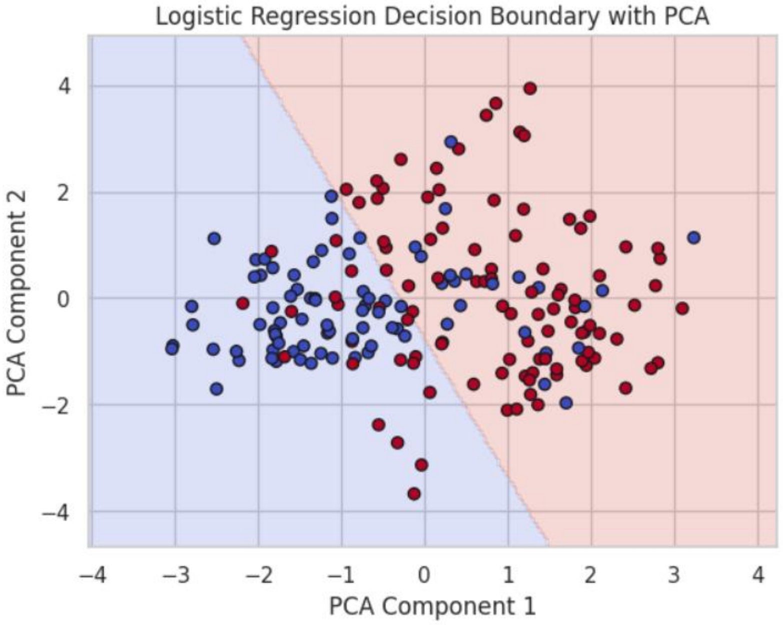
An SVM with a radial basis function (RBF) kernel was trained and the classification results for the test set are summarized below:

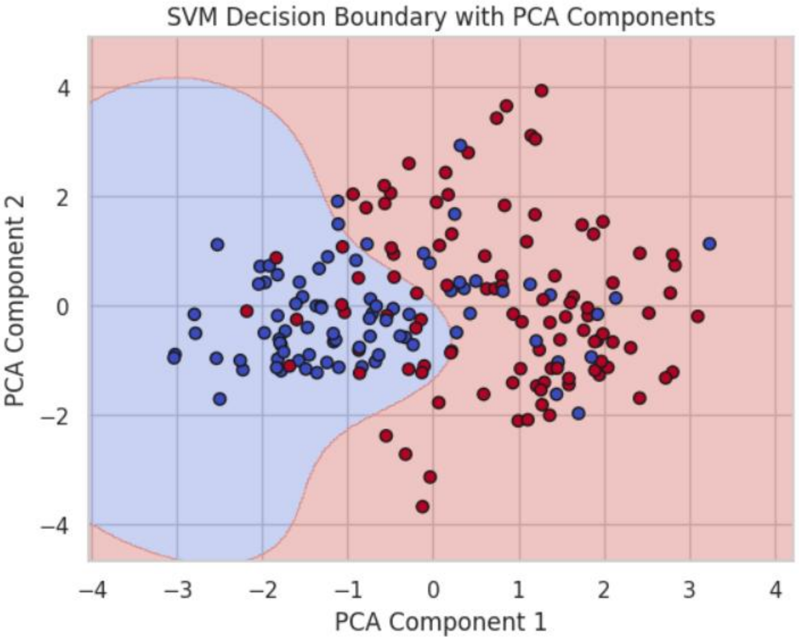


```
SVM Report:
              precision    recall  f1-score   support

           0       0.90      0.77      0.83        82
           1       0.83      0.93      0.88       102

    accuracy                           0.86       184
   macro avg       0.87      0.85      0.85       184
weighted avg       0.86      0.86      0.86       184
```

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a high-dimensional dataset into a smaller set of new variables called principal components. These components capture the maximum variance present in the original data while being uncorrelated with each other. PCA helps to simplify complex datasets, reduce noise, and improve computational efficiency, particularly useful for visualization in two or three dimensions. By projecting data onto the principal components, PCA retains the most important information and allows models to work on simplified inputs.

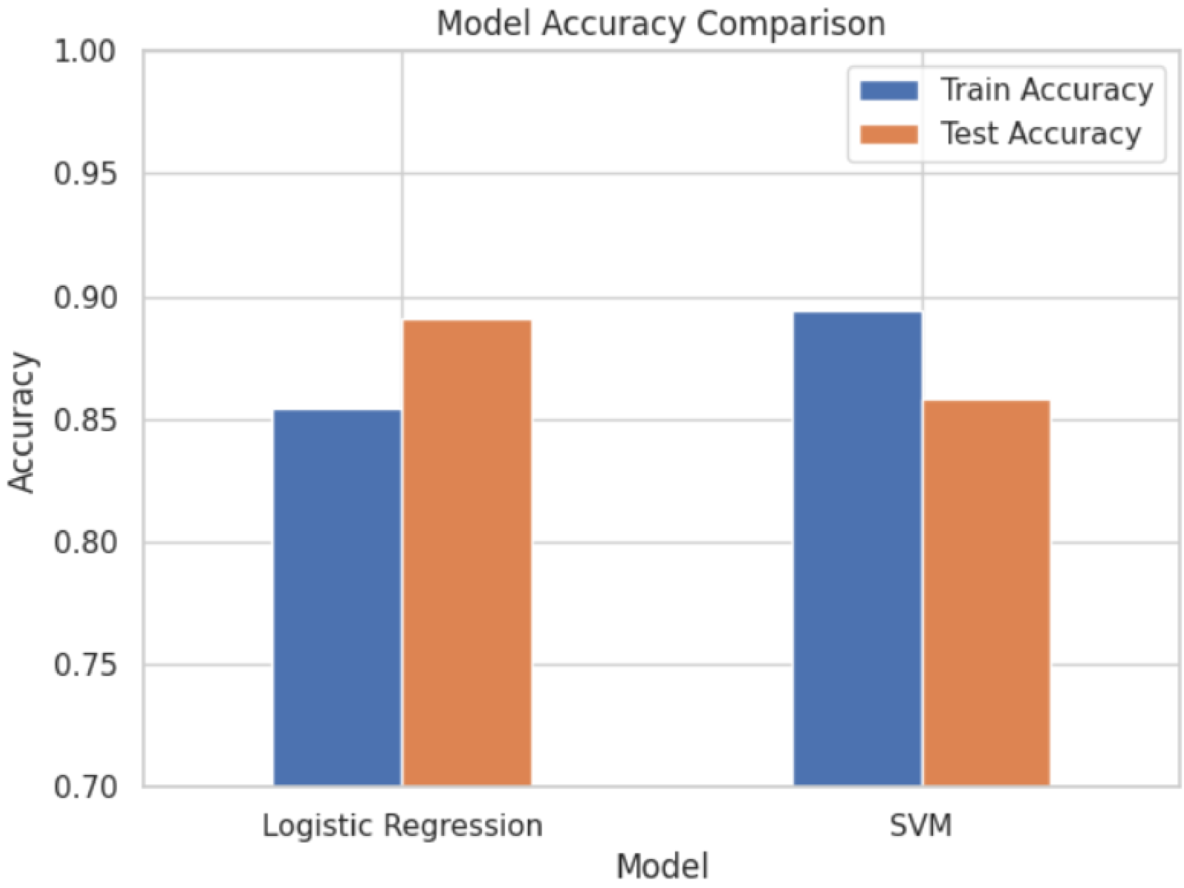


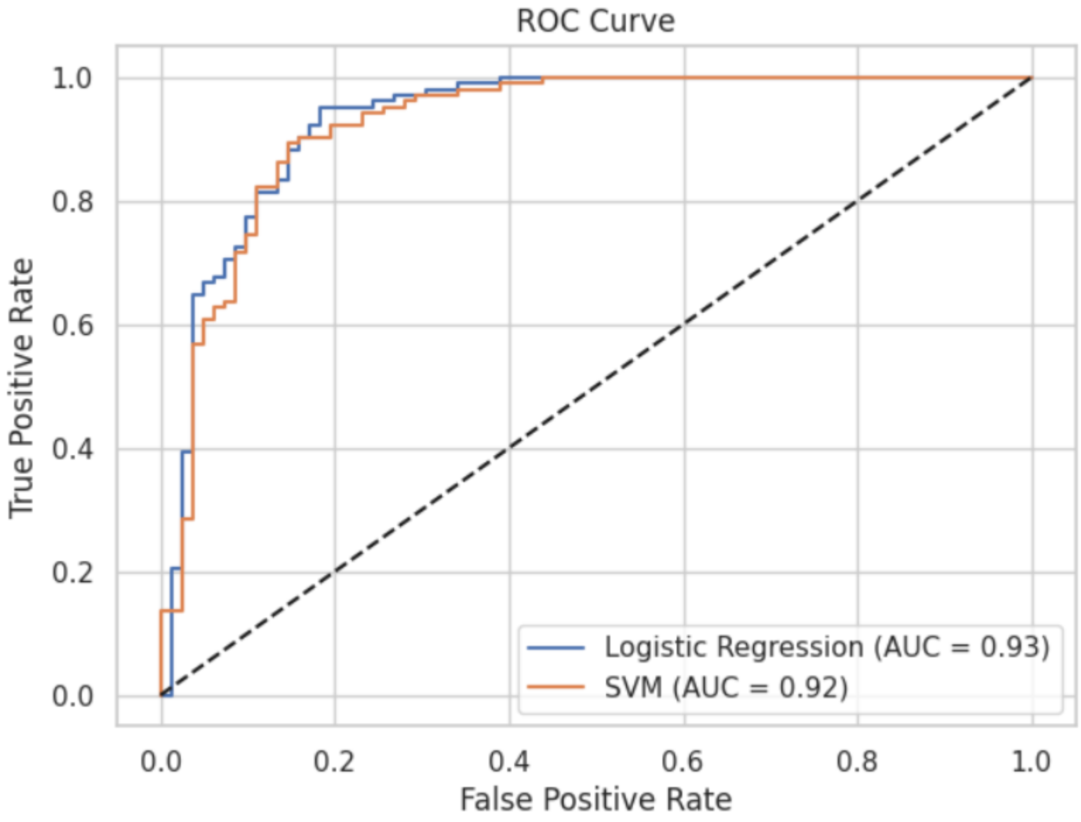


Comparing Logistic Regression and SVM for Classification

Both Logistic Regression and Support Vector Machines (SVM) are widely used for classifying heart disease, but they differ in methodology and strengths. Logistic Regression estimates the probability of a patient having heart disease using a logistic function and works effectively when the data is approximately linearly separable. Its coefficients are interpretable, allowing insight into the influence of each clinical feature. SVM, in contrast, identifies the optimal decision boundary that maximizes the margin between patients with and without heart disease, and can handle non-linear patterns using kernel functions.

In this study, both models achieved high performance. Logistic Regression showed slightly higher test accuracy of 0.89 and an AUC of 0.93, while SVM achieved a slightly lower test accuracy of 0.86 and an AUC of 0.92. Logistic Regression's probabilistic predictions make it useful for risk estimation, whereas SVM provides robustness to outliers and can capture complex relationships in the data. The choice between these models depends on whether interpretability or capturing complex patterns is prioritized for heart disease prediction.





Conclusion

Both Logistic Regression and SVM effectively classified heart disease using clinical data, demonstrating the promise of machine learning in supporting early and accurate diagnosis. Logistic Regression provided interpretability, while SVM showed slightly higher robustness on this dataset. Applying PCA aided in understanding patterns and visualizing decision boundaries. Future work could explore advanced feature engineering, ensemble models, and explainable AI techniques to further enhance predictive performance and clinical utility.