

Heart Disease Prediction using PCA and SVM

1. Introduction

Heart disease is one of the leading causes of death globally. Early prediction of heart disease can help in effective treatment and prevention. In this project, we use Machine Learning techniques to predict heart disease using the UCI Heart Disease dataset (<https://www.kaggle.com/datasets/nagavedareddy/heartdiseasedata>)

The primary objectives are:

- To preprocess the dataset and handle categorical variables.
- To apply Principal Component Analysis (PCA) for dimensionality reduction.
- To train and optimize an SVM classifier.
- To evaluate the performance using standard metrics.

2. Dataset Description

The dataset used is the Heart Disease Dataset from the UCI repository. It contains 303 samples with 14 attributes.

Key Features:

- age: Age of the patient (years)
- sex: Gender (1 = male, 0 = female)
- cp: Chest pain type (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)
- trestbps: Resting blood pressure (mm Hg)
- chol: Serum cholesterol (mg/dl)
- fbs: Fasting blood sugar (>120 mg/dl, 1 = true, 0 = false)
- restecg: Resting electrocardiographic results (0,1,2)
- thalach: Maximum heart rate achieved
- exang: Exercise-induced angina (1 = yes, 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: Slope of the peak exercise ST segment (0,1,2)
- ca: Number of major vessels colored by fluoroscopy (0-3)
- thal: Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
- target: Presence of heart disease (1 = disease, 0 = no disease)

```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
      'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')

Target value count:
target
1      165
0      138
Name: count, dtype: int64
```

3. Methodology

3.1 Data Preprocessing

- Categorical features (sex, cp, fbs, restecg, exang, slope, thal) were one-hot encoded.
- Features scaled using StandardScaler.

3.2 Principal Component Analysis (PCA)

- PCA applied to reduce dimensionality.
- Variance explained per component calculated.
- Optimal number of components chosen to retain ~95% variance.

3.3 Model Selection – Support Vector Machine (SVM)

- Used GridSearchCV to find best hyperparameters.
- Kernels tested: linear, RBF.
- Parameters tuned: C, gamma.

4. Results

4.1 PCA Variance Plot

A scree plot showed that 14 components explain approximately 95% of the variance.

4.2 Model Evaluation

The best model was SVM with RBF kernel.

- Best Parameters: {C=1, gamma=0.01, kernel='rbf'}
- Accuracy: 82.33%

```
➡ Fitting 5 folds for each of 12 candidates, totalling 60 fits
Best Parameters: {'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}
```

Classification Report

	Precision	Recall	F1-Score
No Disease	0.86	0.68	0.76
Disease	0.77	0.91	0.83

Confusion Matrix

	Predicted No Disease	Predicted Disease
Actual No Disease	19	9
Actual Disease	3	30

5. Conclusion

This project demonstrated that SVM with PCA is effective for heart disease prediction. PCA reduced data dimensionality while preserving information, and SVM provided high accuracy.

Future Scope:

- Try ensemble models (Random Forest, XGBoost).
- Apply deep learning for larger datasets.
- Deploy as a web-based prediction tool.

6. Plots and Results

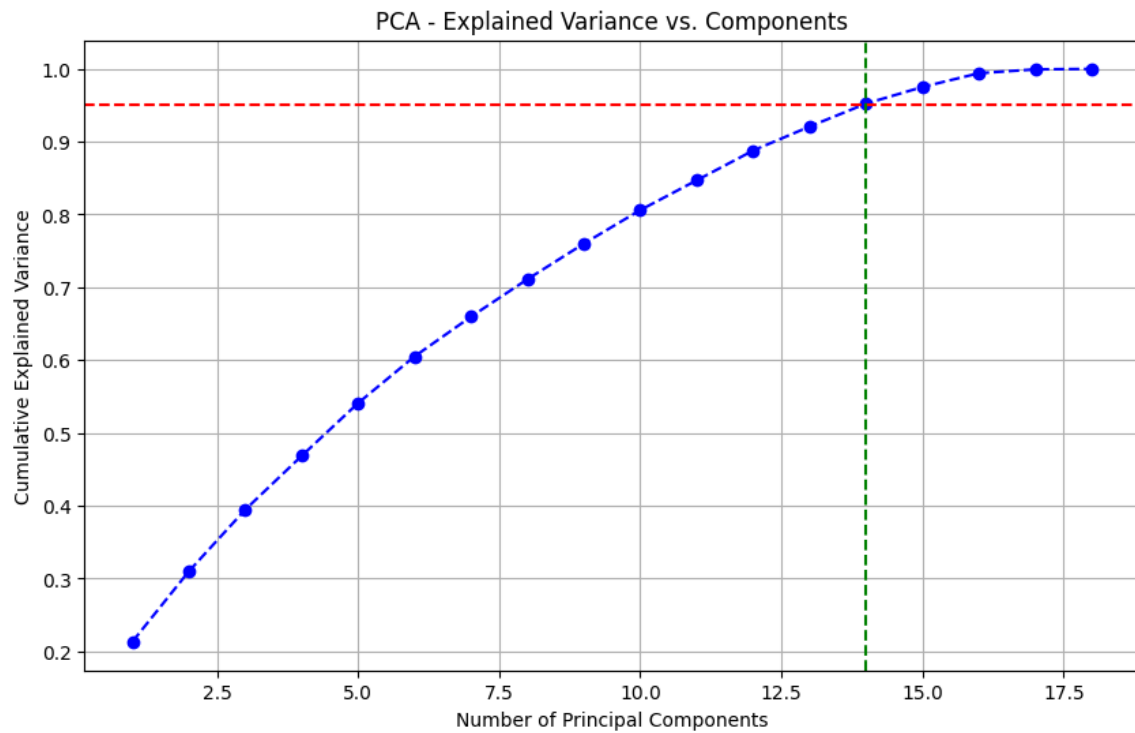


Figure 1: Scree Plot for PCA

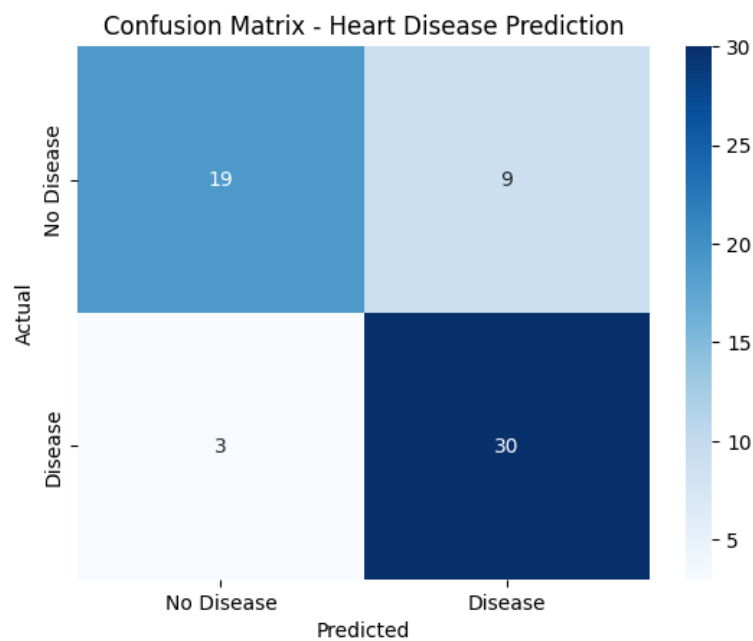



Figure 2: Confusion matrix based on test data



Accuracy: 80.33%

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.68	0.76	28
1	0.77	0.91	0.83	33
accuracy			0.80	61
macro avg	0.82	0.79	0.80	61
weighted avg	0.81	0.80	0.80	61

Figure 3: Evaluation metrics on test data