Cardiovascular Disease Risk Classification by Using Kernel SVM with Dimensionality Reduction Techniques

Abstract

Cardiovascular disease is one of the leading causes of mortality worldwide, making early risk detection an important healthcare challenge. This project aims to classify the risk factor of cardiovascular disease and find out which data mining method has significant performance.

A cardiovascular dataset consisting of approximately 70,000 anonymized records with 11 features and one binary target variable was obtained from Kaggle. The dataset includes attributes such as age, gender, blood pressure, cholesterol level, glucose level, and lifestyle indicators.

Data preprocessing and correlation analysis were performed before model training. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were applied for dimensionality reduction and feature transformation. Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel was employed as the primary classification model.

The experimental results show that Correlation analysis has shown that no single feature shows a strong linear relationship with cardiovascular disease ($0< r < 0.22$). The raw SVM model achieved the highest classification accuracy of approximately 72%. The performance of PCA-SVM and LDA-SVM is not ideal, and the possible reason will be discussed in this report later.

Based on these analysis results, we can conclude that the relationships between the features and cardiovascular disease status are not well described by linear dependencies. Instead, the data exhibits complex multi-dimensional and non-linear interactions among features, which explains that the non-linear SVM classifier has superior performance compared with another linear dimensionality reduction methods.

Table of contents

Table of Contents

1. Introduction

1.1 Background

Cardiovascular disease is one of the leading causes of mortality worldwide, making early risk detection an important healthcare challenge. This project used data mining techniques taught in COMP S460F and COMP S381F – specifically dimensionality reduction (PCA/LDA), Cross-Validation (for evaluation) and kernel-based classification (SVM) – to predict the presence of cardiovascular disease risk from patient features such as age, blood pressure, cholesterol, glucose levels, and lifestyle factors.

1.2 Dataset overview

This project used a cardiovascular dataset from Kaggle which is reported to be collected from medical examinations. However, there is not any detailed documentation or verifiable evidence regarding the data collection process; medical institutions' approval is provided. As a result, this dataset is treated as an anonymized research dataset for educational and experimental machine learning purposes, rather than confirmed real-world clinical patient records.

This dataset consists of approximately 70,000 anonymized records with 11 features, and one binary target variable was obtained. The attributes include age, gender, blood pressure, cholesterol

level, glucose level, and lifestyle indicators. Also, this dataset has complete data and there is not any missing value.

2. Problem Description

Binary Classification

This study is a binary classification problem, where the object is to predict whether a subject is at risk of cardiovascular disease based on multiple clinical and lifestyle features. Given a dataset consisting of N subjects, each subject is represented by an 11-dimensional feature vector

$$\boldsymbol{x_i} = [x_{i1}, x_{i2}, \ldots, x_{i11}] \in R^{11}$$

The formula contains clinical and lifestyle attributes such as age, blood pressure, cholesterol level, glucose level, and physical activity. The corresponding target label is denoted as

$$y_i \in \{0,1\}$$

Which $\boldsymbol{y_i}$ = 1 indicates the presence of cardiovascular disease; $\boldsymbol{y_i}$ = 0 indicates a healthy case.

$$f: R^{11} \rightarrow \{0,1\}$$

It maps the feature vector to the cardiovascular disease risk status. The optimal classifier is learned by minimizing the classification error over the training samples and is evaluated using classification accuracy under a five-fold cross-validation framework.

3. Methodology

3.1 Data Preprocessing and Standardization

The dataset consists of 11 numerical and categorical features. This dataset does not consist of any missing value. Therefore, no missing value imputation is required in the preprocessing. For the model training, all features are standardized using z-score normalization to ensure that each feature has zero mean and unit variance.

3.2 Training set- test split

The dataset is randomly divided into a training set and a testing set with a 7:3 split ratio, where 70% of the data will be used for modelling training; 30% is reserved for performance evaluation.

3.3 Correlation Analysis

Pearson correlation analysis is the linear relationships among the input features and between each feature and the target variable. A larger absolute value of r represent a stronger linear association between two variables. A correlation heatmap will be generated for visualization.

3.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) used to reduce the dimensionality of a dataset. We used it to project the original 11-dimensional feature into a 2-dimensional for visualization It' s transformed features are used for inputting SVM classifier for performance comparison.

3.5 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is employed as a supervised dimensionality reduction method to maximize class separability. Since our project problem is the binary classification, the LDA projection reduces the feature space into 1-dimension. It' s transformed features are also used for inputting to the SVM classifier for performance comparison with PCA-based approaches.

3.6 Support Vector Machine (SVM) Classification

Support Vector Machine(SVM) with a Radial Basis Function (RBF) kernel is used as the classification model. RBF kernel enables the classifier to find a non-linear decision boundary in the high-dimensional (11) feature space. SVM model aims to find the optimal hyperplane separation different classes in data. Three classification method are constructed and compared:

1. Raw SVM using all 11 standardized features
2. PCA with SVM (PCA + SVM)
3. LDA with SVM (LDA + SVM)

3.7 Cross-Validation Strategy (From COMP381F)

To ensure an unbiased performance evaluation, we used the 5-fold cross-validation to randomly divide into five equal subsets. In each fold, 4-subsets are used for training and the remaining subset is used for testing. This process is repeated 5 times. The classification performance is reported as the mean accuracy of 5-folds.

3.8 Performance Evaluation Metric

The classification performance is evaluated by using accuracy. It is the ratio of correctly classified sample to the total number of samples.

4. Analysis

4.1 Correlation Analysis

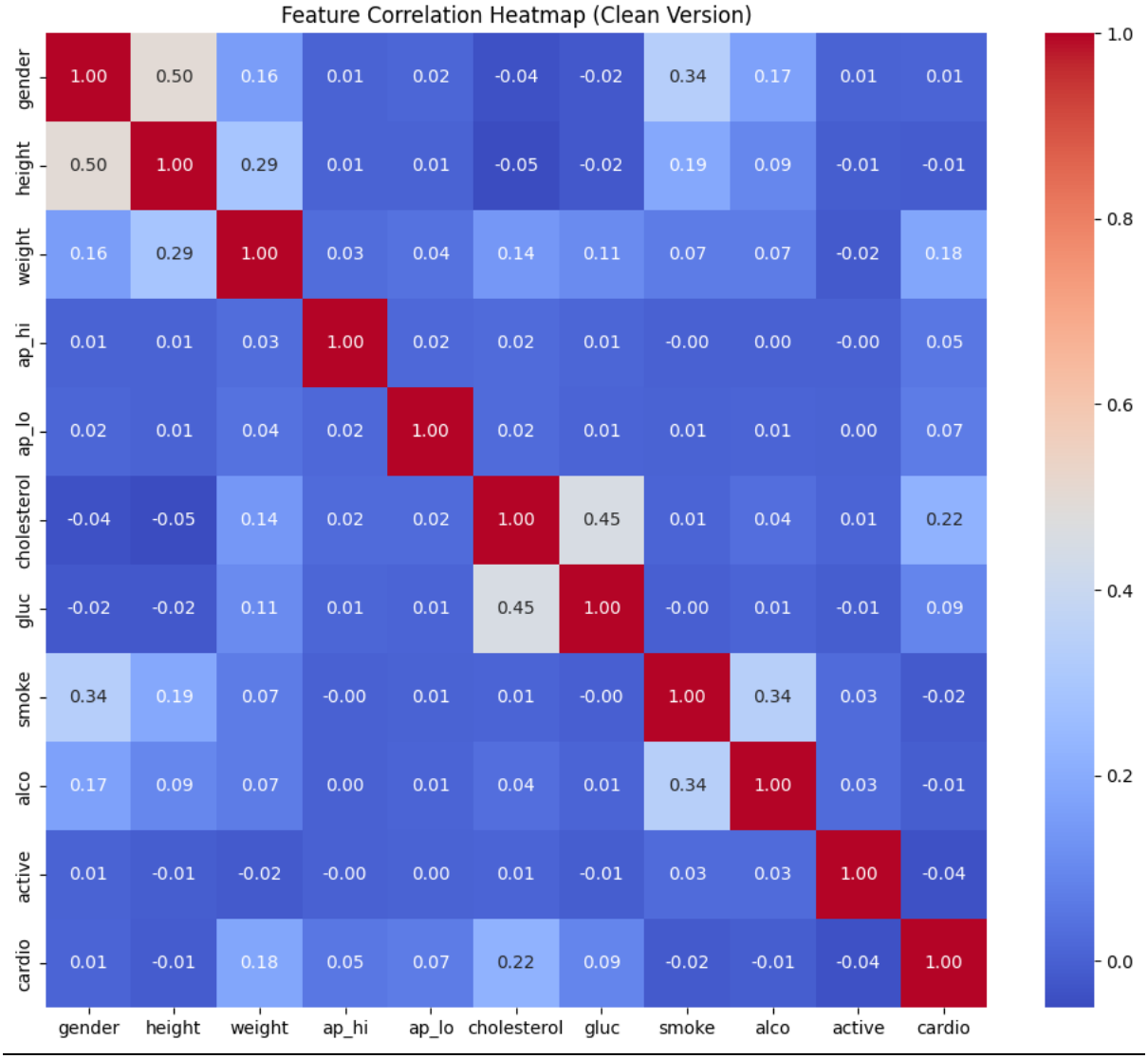

Figure 1: Correlation Heatmap of data

Figure 1 has illustrated the Pearson correlation heatmap among the input features and the target variable. The results show that most features exhibit weak linear correlations with cardiovascular disease. The strongest correlation with the target variable is observed for cholesterol ($r \approx 0.22$), followed by weight ($r \approx 0.18$). There is none of a single feature demonstrates a strong direct linear relationship with the disease.

From the input features, gender and height ($r \approx 0.50$) show the strongest inter-feature correlation, while cholesterol and glucose ($r \approx 0.45$) also exhibit moderate positive

association which means cardiovascular disease risk cannot be reliably explained by any single linear factor. This observation implies that the predictive structure of the dataset is likely non-linear and involves multi-dimensional feature interactions.

4.2 PCA Projection Analysis

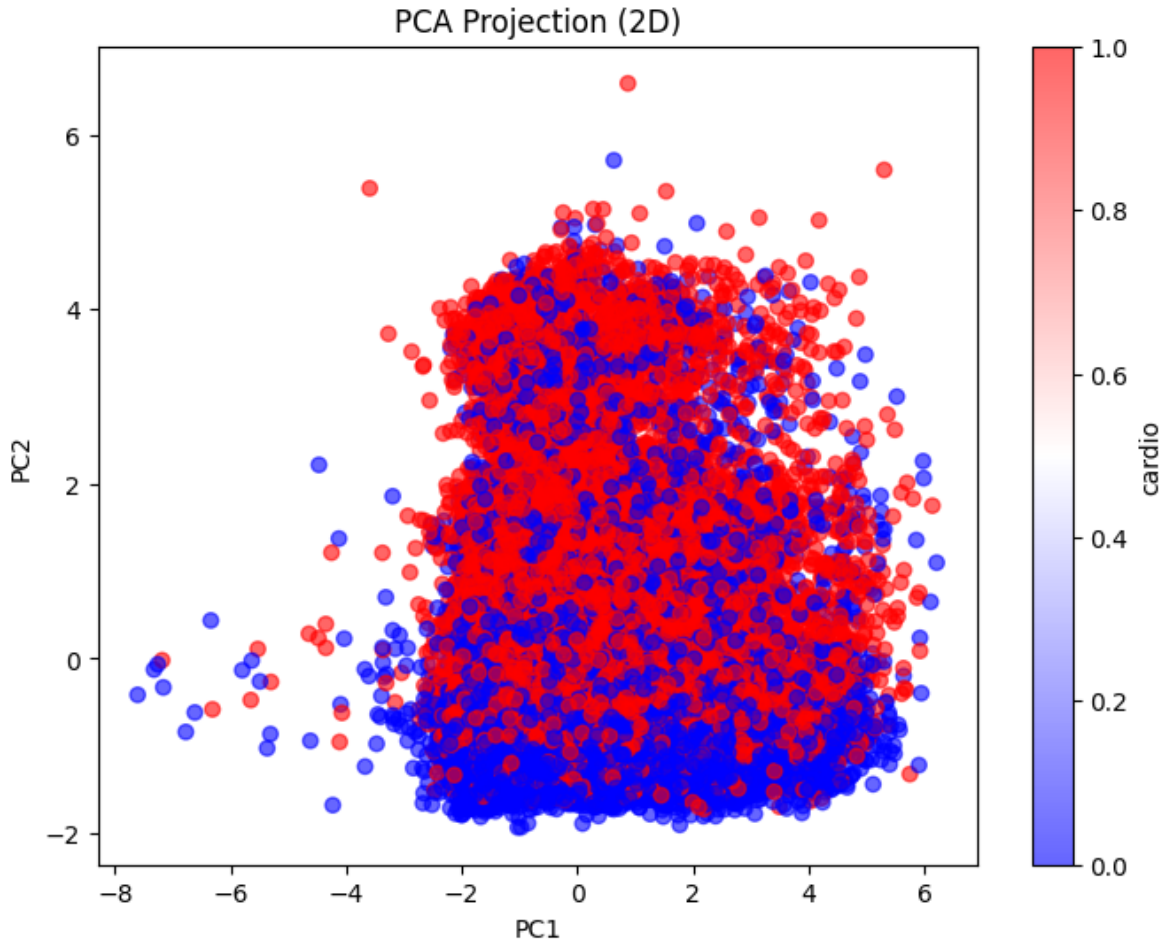

Figure 2: PCA scatter plot

Figure 2 shows the two-dimensional PCA projection of the dataset. While PCA preserves the directions of maximum variance, the projected data points from the healthy and cardiovascular disease classes exhibit overlap. No clear linear decision boundary can be formed in the reduced space. This indicates that variance-maximizing directions do not necessarily correspond to directions of maximum class separability.

4.3 LDA Projection Analysis

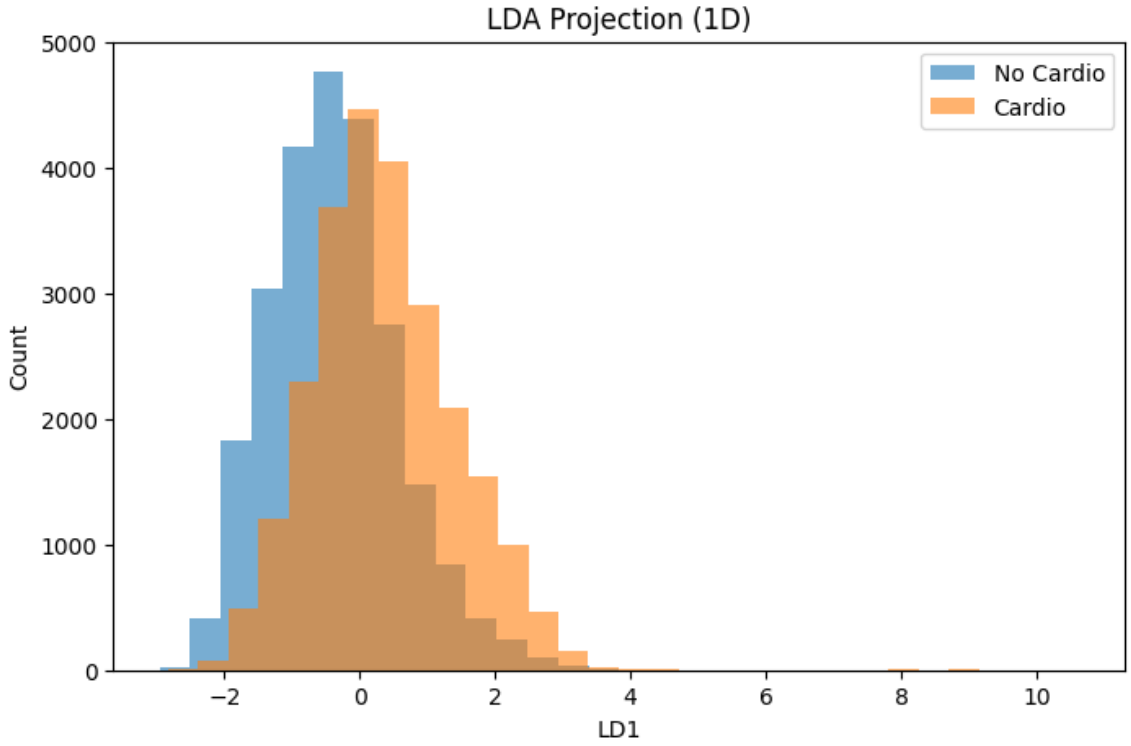

Figure 3: 1D LDA histogram

Figure 3 demonstrates the 1-dimensional LDA projection of the dataset. LDA aims to maximize the separation between class means relative to within-class variance. However, although a slight shift between the two class distributions is observed, there is still a large degree of overlap.

This overlapping distribution indicates that a single linear discriminant axis is insufficient to fully separate the cardiovascular disease and healthy classes. It also explains why purely linear discriminant-based classification methods cannot achieve high prediction accuracy on this dataset.

4.4 SVM Performance Comparison Analysis

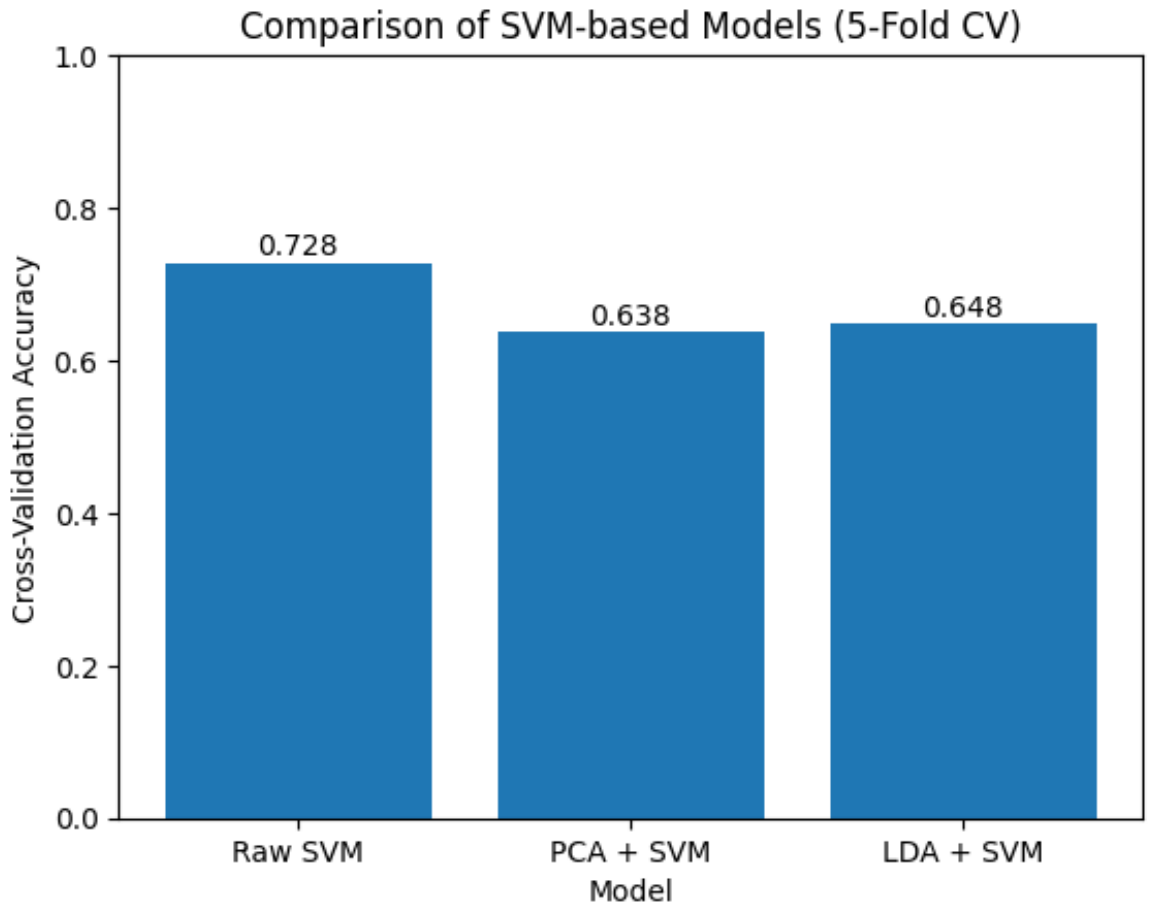

Figure 4: Raw SVM vs PCA+SVM vs LDA+SVM Bar Chart

Figure 4 compares the classification accuracy of different models under five-fold cross-validation by a bar chart. The Raw SVM model achieves the highest accuracy of 72.8%, significantly outperforming the PCA+SVM (63.8%) and LDA+SVM (64.8%). These results confirm that the predictive structure of the dataset is highly non-linear and multi-dimensional.

5. Conclusion

Based on these analysis results, we can conclude that the relationships between the features and cardiovascular disease status are not well described by linear dependencies. Instead, the data exhibits complex multi-dimensional and non-linear interactions among features, which explains that the non-linear SVM classifier has superior performance compared with another linear dimensionality reduction methods.

For future work, model performance can potentially be improved through another data mining tool like FLD or applying SMOTE for imbalanced data and using XGBoost to enhance classification performance and minority-class detection. Based on our search and study, XGBoost is a robust ensemble classifier suitable for medical prediction, while SMOTE is an effective oversampling method to address class imbalance and enhance minority-class detection and they may improve our model performance effectively.

6. Appendix

Data Preprocessing Code (split dataset & Standardization)

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)
```

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Correlation Heat Map

```
import matplotlib.pyplot as plt
import seaborn as sns

heatmap_features = df[['gender','height','weight','ap_hi','ap_lo',

'cholesterol','gluc','smoke','alco','active','cardio']]

plt.figure(figsize=(12,10))
sns.heatmap(heatmap_features.corr(), cmap="coolwarm", annot=True,
fmt=".2f")
plt.title("Feature Correlation Heatmap (Clean Version)")
plt.show()
```

PCA Implementation

```
from sklearn.decomposition import PCA
```

```
pca = PCA(n_components=2)
X_train_pca = pca.fit_transform(X_train_scaled)
X_test_pca = pca.transform(X_test_scaled)
```

```
plt.figure(figsize=(8,6))
plt.scatter(X_train_pca[:,0], X_train_pca[:,1], c=y_train,
cmap='bwr', alpha=0.6)
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("PCA Projection (2D)")
plt.colorbar(label="cardio")
plt.show()
```

LDA Implementation

```
from sklearn.discriminant_analysis import
LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis(n_components=1)
X_train_lda = lda.fit_transform(X_train_scaled, y_train)
X_test_lda = lda.transform(X_test_scaled)
```

```
plt.figure(figsize=(8,5))
plt.hist(X_train_lda[y_train==0], bins=30, alpha=0.6, label="No
Cardio")
plt.hist(X_train_lda[y_train==1], bins=30, alpha=0.6,
label="Cardio")
plt.xlabel("LD1")
plt.ylabel("Count")
plt.title("LDA Projection (1D)")
plt.legend()
plt.show()
```

SVM Training and Evaluation (5-fold Cross Validation)

```
cv = KFold(n_splits=5, shuffle=True, random_state=42)

svm_pipeline = Pipeline([
    ("scaler", StandardScaler()),
```

```
    ("svm", SVC(kernel="rbf", gamma="scale"))
])

scores = cross_val_score(svm_pipeline, X, y, cv=cv,
scoring="accuracy")
print("Raw SVM CV Mean:", scores.mean())

pca_svm_pipeline = Pipeline([
    ("scaler", StandardScaler()),
    ("pca", PCA(n_components=2)),
    ("svm", SVC(kernel="rbf", gamma="scale"))
])

scores_pca = cross_val_score(pca_svm_pipeline, X, y, cv=cv,
scoring="accuracy")
print("PCA + SVM CV Mean:", scores_pca.mean())

lda_svm_pipeline = Pipeline([
    ("scaler", StandardScaler()),
    ("lda", LinearDiscriminantAnalysis(n_components=1)),
    ("svm", SVC(kernel="rbf", gamma="scale"))
])

scores_lda = cross_val_score(lda_svm_pipeline, X, y, cv=cv,
scoring="accuracy")
print("LDA + SVM CV Mean:", scores_lda.mean())
cv_results
```

6. Reference:

Ulianova, S. (2020). Cardiovascular Disease Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset