LAB 10

PCA:

In [1]:

```python
#heart dataset
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.decomposition import PCA
from google.colab import files

# Upload the CSV file
uploaded = files.upload()
df = pd.read_csv(next(iter(uploaded)))
print(df.head())

# 2. Encode categorical columns
label_enc_cols = ['Sex', 'ExerciseAngina']
for col in label_enc_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])

# One-hot encode nominal categorical columns
df = pd.get_dummies(df, columns=['ChestPainType', 'RestingECG', 'ST_Slope'], drop_first=True)

# 3. Split features and target
X = df.drop('HeartDisease', axis=1)
y = df['HeartDisease']

# 4. Scale features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# 5. Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

```python
# 6. Train classifiers and evaluate
models = {
    'Logistic Regression': LogisticRegression(),
    'SVM': SVC(),
    'Random Forest': RandomForestClassifier()
}

print("Without PCA:")
for name, model in models.items():
    model.fit(X_train, y_train)
    preds = model.predict(X_test)
    acc = accuracy_score(y_test, preds)
    print(f"{name} Accuracy: {acc:.4f}")

# 7. Apply PCA
pca = PCA(n_components=5)  # Try fewer components for dimensionality reduction
X_pca = pca.fit_transform(X_scaled)
X_train_pca, X_test_pca, y_train, y_test = train_test_split(X_pca, y, test_size=0.2, random_state=42)

print("\nWith PCA (5 components):")
for name, model in models.items():
    model.fit(X_train_pca, y_train)
    preds = model.predict(X_test_pca)
    acc = accuracy_score(y_test, preds)
    print(f"{name} Accuracy: {acc:.4f}")
```

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving heart.csv to heart.csv
   Age Sex ChestPainType  RestingBP  Cholesterol  FastingBS RestingECG  MaxHR  \
0   40   M           ATA        140          289          0     Normal    172
1   49   F           NAP        160          180          0     Normal    156
2   37   M           ATA        130          283          0         ST     98
3   48   F           ASY        138          214          0     Normal    108
4   54   M           NAP        150          195          0     Normal    122
```

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving heart.csv to heart.csv
   Age Sex ChestPainType  RestingBP  Cholesterol  FastingBS RestingECG  MaxHR  \
0   40   M           ATA        140          289          0     Normal    172
1   49   F           NAP        160          180          0     Normal    156
2   37   M           ATA        130          283          0         ST     98
3   48   F           ASY        138          214          0     Normal    108
4   54   M           NAP        150          195          0     Normal    122

  ExerciseAngina  Oldpeak ST_Slope  HeartDisease
0              N      0.0       Up             0
1              N      1.0     Flat             1
2              N      0.0       Up             0
3              Y      1.5     Flat             1
4              N      0.0       Up             0
Without PCA:
Logistic Regression Accuracy: 0.8533
SVM Accuracy: 0.8750
Random Forest Accuracy: 0.8696

With PCA (5 components):
Logistic Regression Accuracy: 0.8207
SVM Accuracy: 0.8424
Random Forest Accuracy: 0.8533
```

Lab 10

PCA

Q

Feature	Example 1	Example 2	Example 3	Example 4
x_1	4	8	13	7
x_2	11	4	5	14

$$\bar{x_1} = \frac{4+8+13+7}{4} = 8$$

$$\bar{x_2} = \frac{11+4+5+14}{4} = 8.5$$

$$S = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & cov(x_2, x_2) \end{bmatrix}$$

$$S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

~~$0 = (S - \lambda I)$~~

$$0 = |S - \lambda I|$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$0 = \left[\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right]$$

$$= \begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix}$$

$$0 = (14-\lambda)(23-\lambda) - 121$$

$$0 = \lambda^2 - 37\lambda + 201$$

$$\lambda = \frac{1}{2}\left[37 \pm \sqrt{37^2 - 4 \times 201}\right]$$

$$= [30.3849, 6.6151]$$

λ_1 λ_2

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$(S - \lambda I)U = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23.7 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda)v_1 - 11v_2 = 0$$
$$-11v_1 + (23-7)v_2 = 0$$

$$(14-\lambda)v_1 = 11v_2$$
$$\frac{v_1}{11} = \frac{v_2}{14-\lambda} = 1$$

$$v = \begin{bmatrix} 11 \\ 14-\lambda \end{bmatrix}$$

$$\|v\| = 19.7348$$

$$e_1 = \begin{bmatrix} 11 / \|v\| \\ (14-\lambda_1) / \|v\| \end{bmatrix}$$

$$= \begin{bmatrix} 11 / \|v\| \\ (14-\lambda_1) / \|v\| \end{bmatrix}$$

$$= \begin{bmatrix} 11 / 19.7348 \\ (14-30.3849) / 19.7345 \end{bmatrix}$$

$$e_1 = \begin{bmatrix} 0.5374 \\ -0.8303 \end{bmatrix}$$

$$e_1^T \begin{bmatrix} x_{1k} - \bar{x}_1 \\ x_{2k} - \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \end{bmatrix}$$

$$= 0.5574(4-8) - 0.8303(11-8.5)$$
$$= -4.3052$$

Model	Accuracy before PCA	Accuracy after PCA
SVM	0.8369	0.8403
Logistic	0.8478	0.8369
Random forest	0.8864	0.8369