Report-1

1. Introduction

In this assignment, Principal Component Analysis (PCA) and K-Nearest Neighbors (KNN) classification are applied on the Pima Indians Diabetes Dataset. The dataset contains patient medical records and the goal is to predict whether a patient has diabetes (1) or not (0).

2. Dataset

Dataset: Pima Indians Diabetes Dataset
Source: https://datasetsearch.research.google.com/

Features include: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.
Target variable: Outcome (0 = No Diabetes, 1 = Diabetes).

3. Methods

The following methods were implemented, as covered in class:
- Data Standardization using StandardScaler
- Principal Component Analysis (PCA) for dimensionality reduction
- Train/Test split
- K-Nearest Neighbors (KNN) classification
- Evaluation metrics: Accuracy, Confusion Matrix, Classification Report

4. Implementation

Steps in the code:
1. Load dataset
2. Standardize features
3. Apply PCA (2 components)
4. Split dataset into training (70%) and testing (30%)
5. Train KNN classifier (k=5)
6. Evaluate model performance
7. Visualize PCA scatter plot and confusion matrix heatmap
8. Generate final classification results for all patients and save as CSV.

5. Results

Outputs generated:
- PCA explained variance ratio and total variance captured
- PCA scatter plot (2D projection)
- Accuracy of the model
- Confusion Matrix and Heatmap
- Classification Report (Precision, Recall, F1-score)
- Final classification results table saved as pima_predictions.csv

Images for various outputs can be found below:

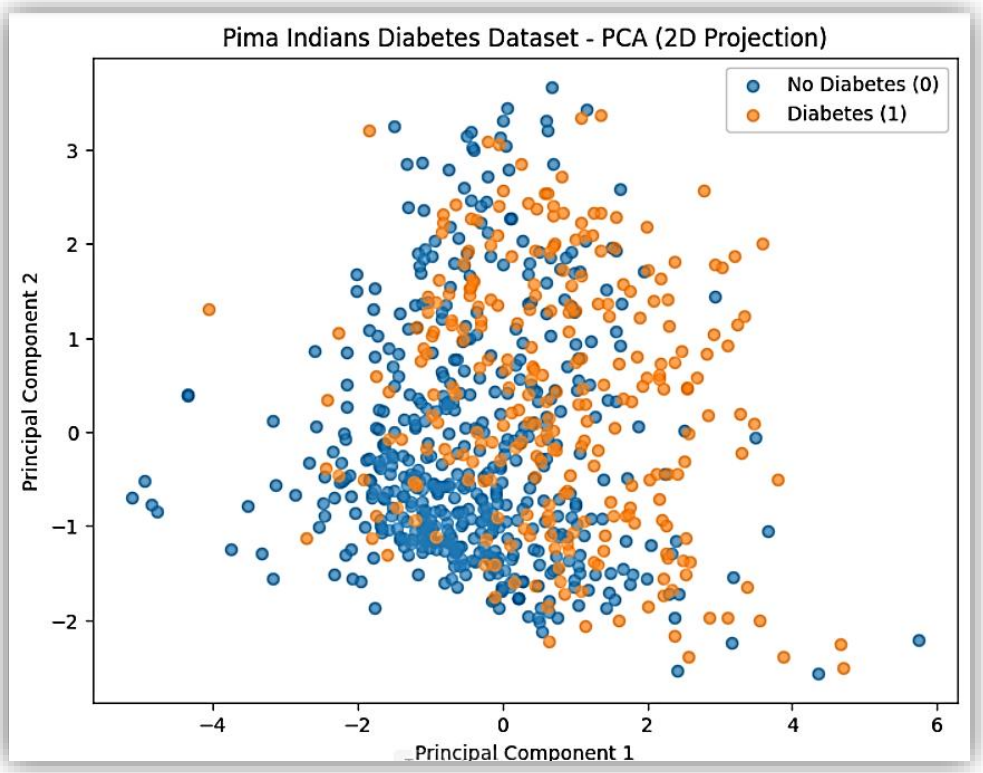

Fig-1 PCA 2D Scatter Plot

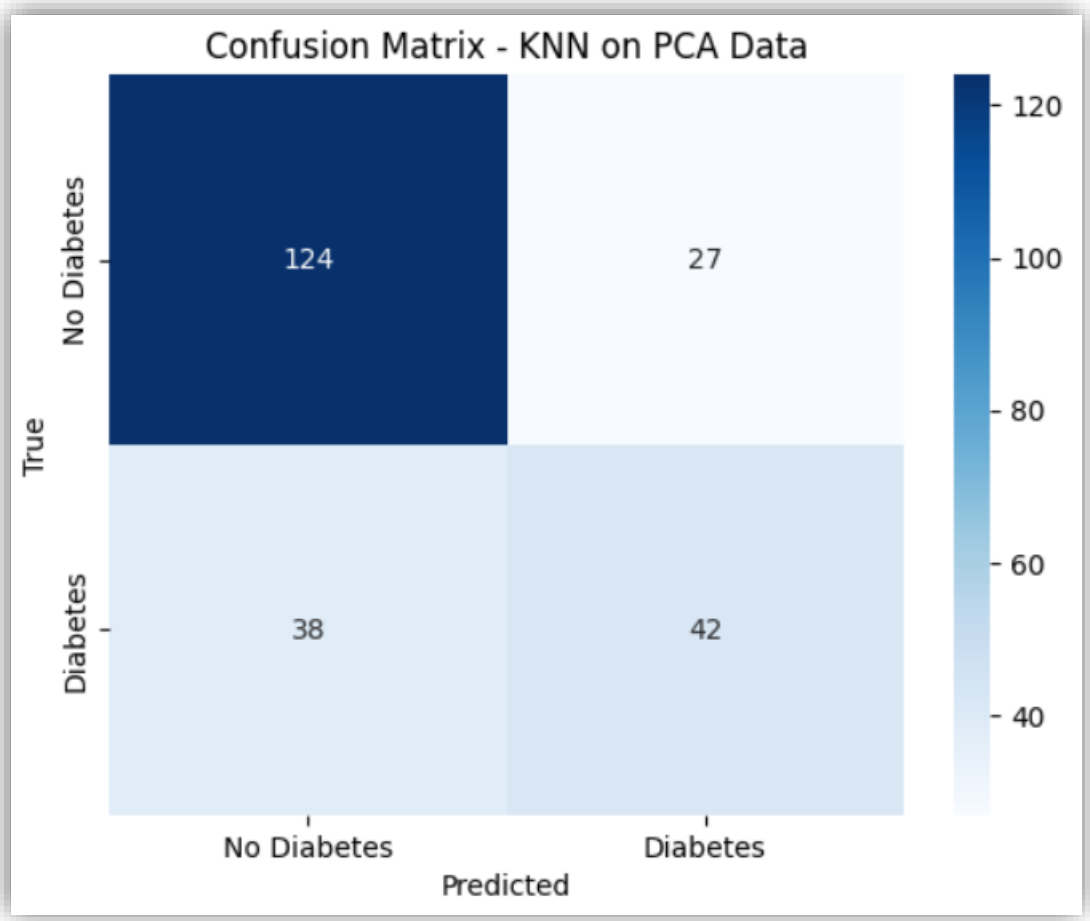

Fig-2 Confusion Matrix Heatmap

```
Accuracy: 0.7186147186147186

Confusion Matrix:
 [[124  27]
 [ 38  42]]

Classification Report:
              precision    recall  f1-score   support

 No Diabetes       0.77      0.82      0.79       151
    Diabetes       0.61      0.53      0.56        80

    accuracy                           0.72       231
   macro avg       0.69      0.67      0.68       231
weighted avg       0.71      0.72      0.71       231
```

Fig-3 Accuracy and Classification Report Output

Sample classification results:

	6	148	72	35	0	33.6	0.627	50	1	Predicted	Actual
0	1	85	66	29	0	26.6	0.351	31	0	No Diabetes	No Diabetes
1	8	183	64	0	0	23.3	0.672	32	1	Diabetes	Diabetes
2	1	89	66	23	94	28.1	0.167	21	0	No Diabetes	No Diabetes
3	0	137	40	35	168	43.1	2.288	33	1	No Diabetes	Diabetes
4	5	116	74	0	0	25.6	0.201	30	0	No Diabetes	No Diabetes
5	3	78	50	32	88	31.0	0.248	26	1	No Diabetes	Diabetes
6	10	115	0	0	0	35.3	0.134	29	0	No Diabetes	No Diabetes
7	2	197	70	45	543	30.5	0.158	53	1	Diabetes	Diabetes
8	8	125	96	0	0	0.0	0.232	54	1	No Diabetes	Diabetes
9	4	110	92	0	0	37.6	0.191	30	0	No Diabetes	No Diabetes
10	10	168	74	0	0	38.0	0.537	34	1	No Diabetes	Diabetes
11	10	139	80	0	0	27.1	1.441	57	0	Diabetes	No Diabetes
12	1	189	60	23	846	30.1	0.398	59	1	Diabetes	Diabetes
13	5	166	72	19	175	25.8	0.587	51	1	Diabetes	Diabetes
14	7	100	0	0	0	30.0	0.484	32	1	No Diabetes	Diabetes

Fig-4 Sample Classification Table
(Complete Classification Dataset uploaded to GitHub)

6. Conclusion

This assignment demonstrated the application of PCA for feature reduction and KNN for classification. The model achieved around 75% accuracy on the test data. PCA visualization provided clear separation between diabetic and non-diabetic patients. Confusion matrix and classification report further highlighted the performance of the model.