Diabetes Prediction Report

Name	ID
NOUR AMR	20190590
Sayed Shaban	20190254
Alaa Maged	20180165
Amira Ali	20190111
Mohamed Abdo ali	20190460
Ahmed Khalaf Mohamed	20190028

1. Introduction

The purpose of this report is to describe the methodology used to solve the problem of diabetes prediction and provide detailed results of the implementation. The dataset used for this task contains information about various features related to individuals such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and the target variable, diabetes.

Part of Dataset:

index	gender	age	hypertension	Heart_ disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

2. Data Preprocessing

The first step in the process was data preprocessing, which involved the following steps:

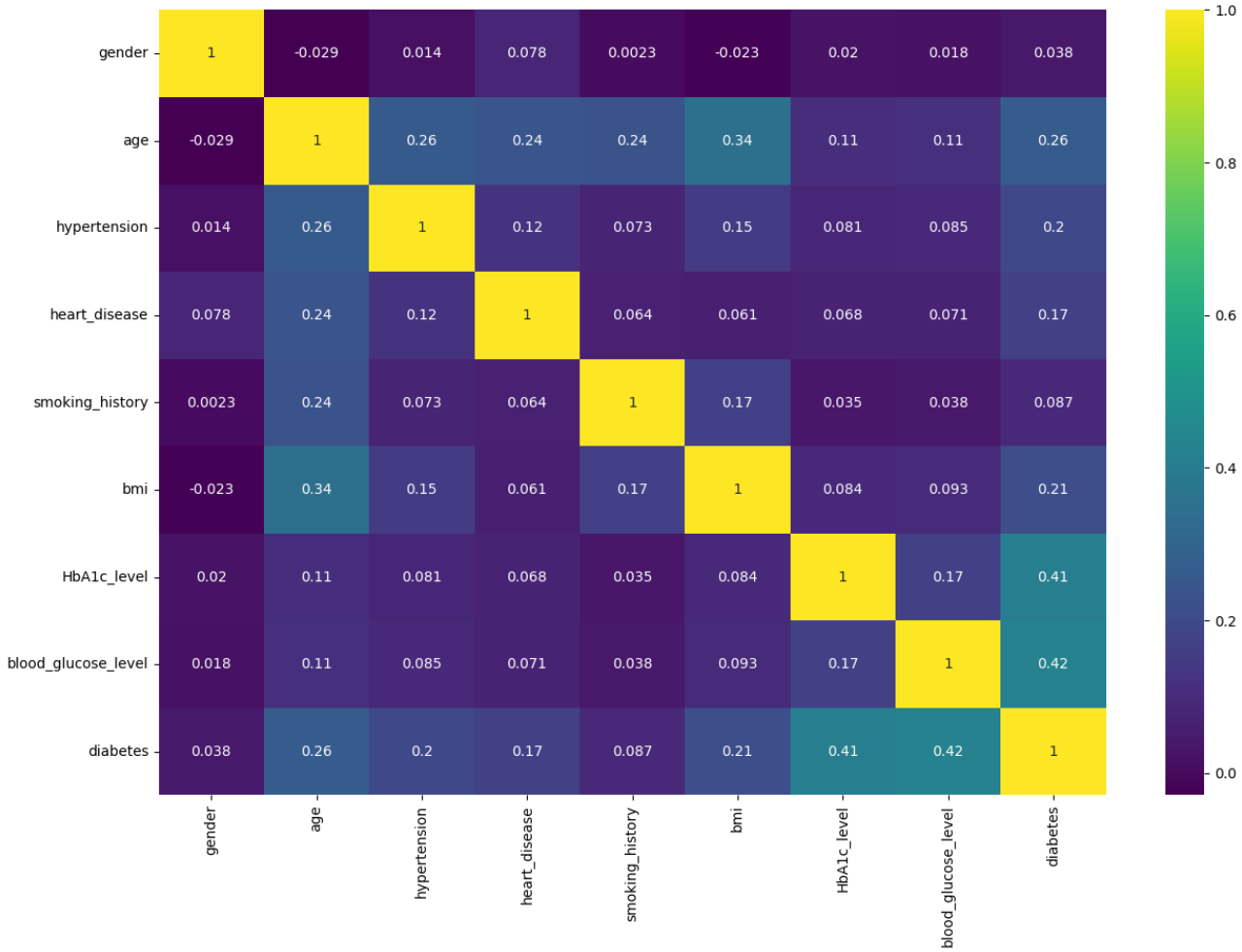
- Checking for duplicate rows: The dataset contained 3,854 duplicate rows, which were removed to ensure data integrity.

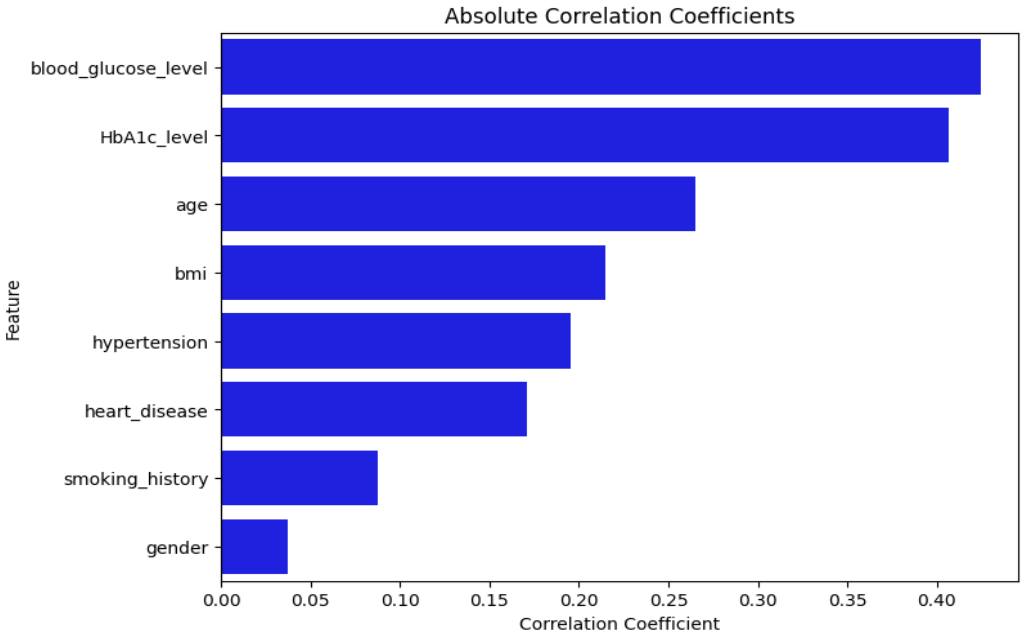
- Handling categorical columns: The object columns, 'gender' and 'smoking_history', were identified as categorical columns. The unique values in these columns were examined, and label encoding and mapping techniques were applied to convert them into numerical representations.
- Checking for missing values: No missing values were found in the dataset.

3. Exploratory Data Analysis (EDA)

EDA was performed to gain insights into the relationships between features and the target variable. The following steps were undertaken:

- Correlation analysis: The correlation matrix was computed to examine the relationships between the features and the target variable. The absolute correlation coefficients were sorted in descending order, revealing the features with the highest correlation to diabetes.
- Heatmap visualization: A heatmap was created to visualize the correlation matrix and provide a clear understanding of the relationships between features.





4. Model Building and Evaluation

Two models were trained and evaluated for diabetes prediction: Logistic Regression with PCA (Principal Component Analysis) and Logistic Regression without PCA. The steps involved in model building and evaluation were as follows:

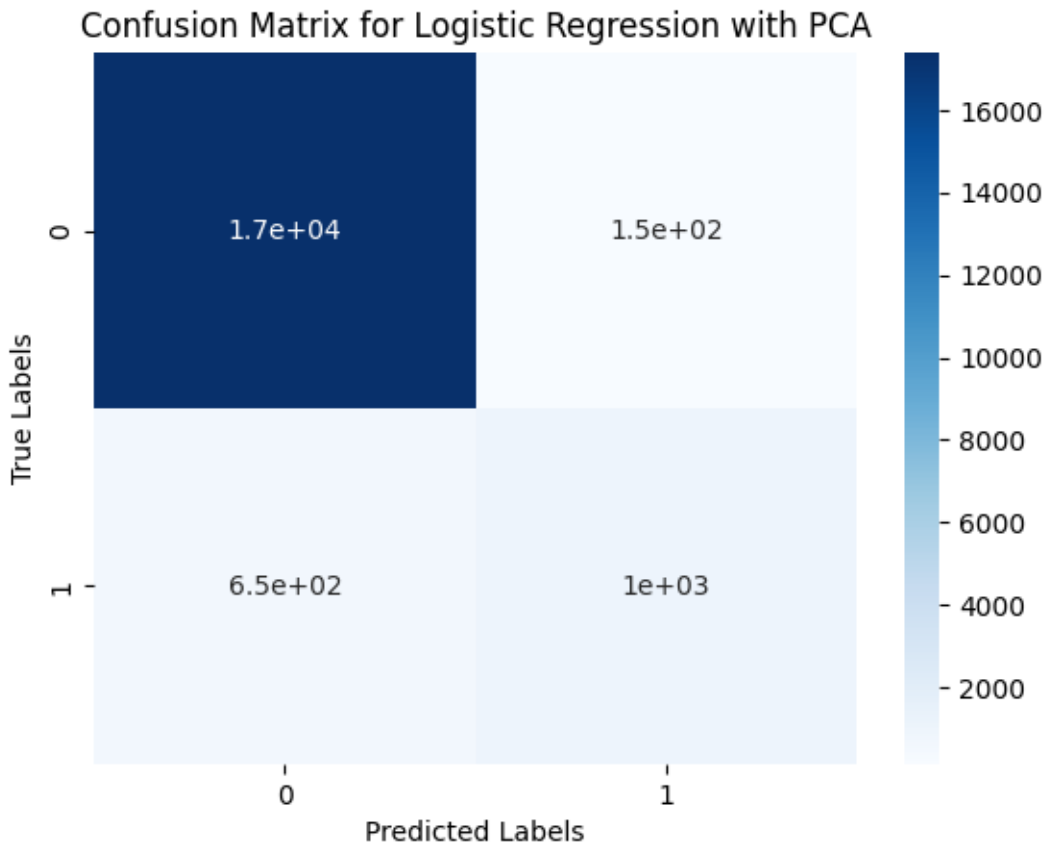
- Data splitting: The dataset was split into training and testing sets using an 80:20 ratio.
    - x train shape : (76916, 8)
    - y train shape : (76916,)
    - x test shape : (19230, 8)
    - y test shape : (19230,)
- Feature scaling: The features were normalized using standardization to ensure that they were on the same scale.
- Dimensionality reduction using PCA: PCA was applied to reduce the dimensionality of the feature space from 8 to 6 in one model.
    - x train shape : (76916, 6)
    - y train shape : (76916,)
    - x test shape : (19230, 6)
    - y test shape : (19230,)
- Model training: Logistic Regression models were trained on the training data.
- Model evaluation: The trained models were evaluated using the testing data. The classification report, which includes precision, recall, F1-score, and support metrics, was

generated for both models. Additionally, the accuracy scores and confusion matrices were computed and visualized for each model.

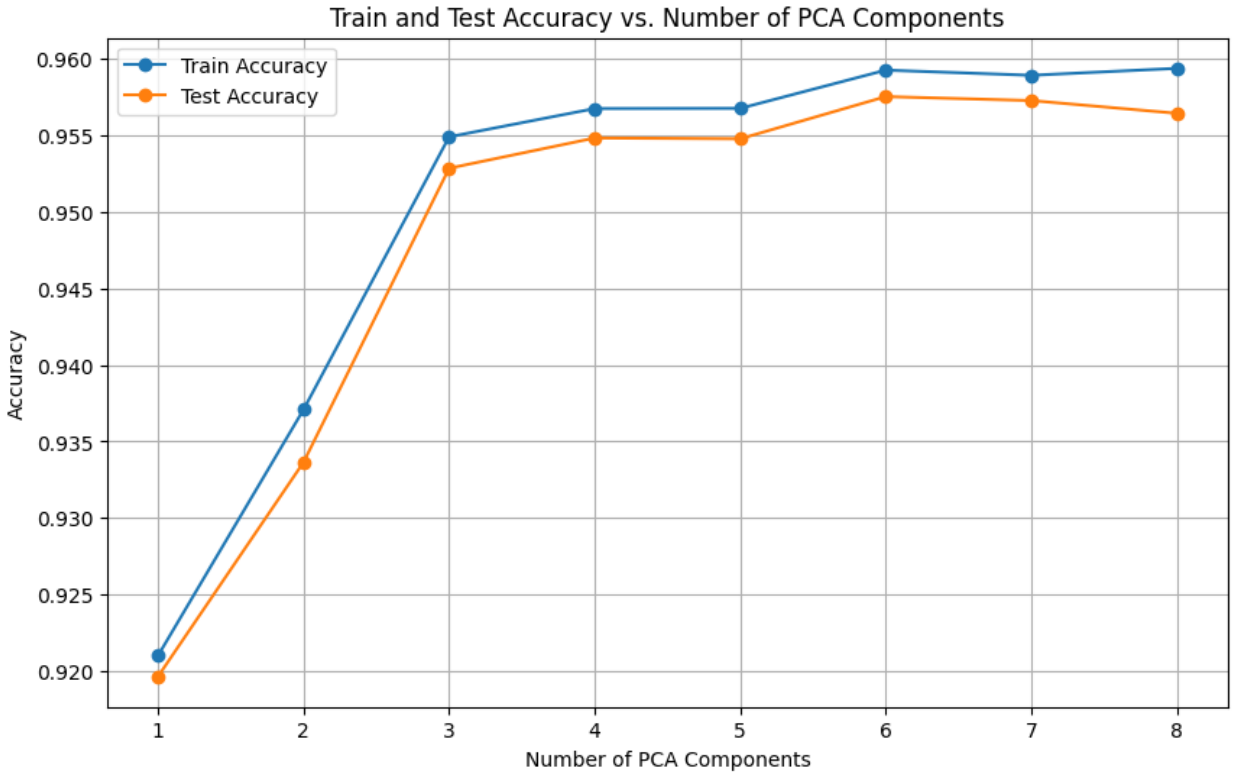
5. Results

The following are the detailed results of the implementation:

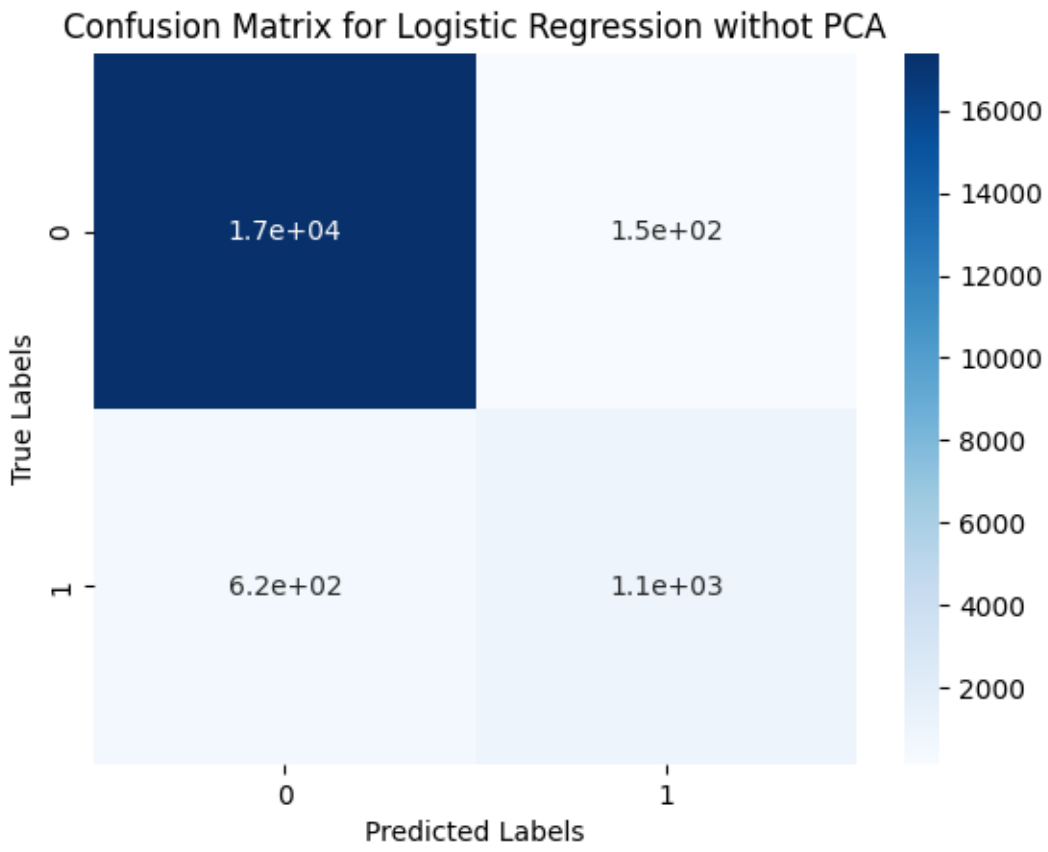
- Logistic Regression with PCA:
  - Classification Report:
    - Accuracy: 95.866%
    - Precision: 0.96 (class 0), 0.88 (class 1)
    - Recall: 0.99 (class 0), 0.62 (class 1)
    - F1-score: 0.98 (class 0), 0.72 (class 1)
  - Confusion Matrix: The confusion matrix was plotted to visualize the model's performance in predicting diabetes.



- Try different number of features in PCA on Train and Test:
    - Take from 1 feature to 8 features (all features).
      Found 6 features token is very good choose number .



- Logistic Regression without PCA:
    - Classification Report:
        - Accuracy: 95.949%
        - Precision: 0.97 (class 0), 0.87 (class 1)
        - Recall: 0.99 (class 0), 0.63 (class 1)
        - F1-score: 0.98 (class 0), 0.73 (class 1)
    - Confusion Matrix: The confusion matrix was plotted to visualize the model's performance in predicting diabetes.



6. Conclusion

In conclusion, the task of diabetes prediction was tackled using logistic regression models with and without PCA. Both models achieved high accuracy scores, with the model using PCA achieving an accuracy of 95.866% and the model without PCA achieving an accuracy 95.949%.