Problem Statement

This study's goal is to use a set of diagnostic parameters to predict whether a patient has diabetes. The dataset that was supplied, The aim is to create a classification model that can reliably predict the Outcome variable, where 1 denotes a positive diabetes diagnosis and 0 denotes no diabetes. diabetes2.csv contains a variety of health indicators for multiple patients.

Abstract

This article describes a machine learning project that uses a collection of diagnostic metrics to predict diabetes in patients. The main objective is to create a predictive model that uses a collection of characteristics to categorise patients as either diabetes or non-diabetic. A logistic regression model, an appropriate approach for binary classification tasks, is used in this research. To determine how well the model identifies diabetes cases, the data is pre-processed and its performance is assessed using important metrics like accuracy, recall, and F1 score in addition to a confusion matrix.

Methodology

The methodology followed a standard machine learning workflow:

1. **Data Loading**: The diabetes2.csv dataset is loaded into a Pandas Data Frame. The dataset includes features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, with 'Outcome' as the target variable.
2. **Data Splitting**: The dataset is split into training and testing sets to prepare for model training and evaluation. The training set is used to train the model, while the test set is used to evaluate its performance on unseen data.
3. **Feature Scaling**: The features in the dataset are scaled using Standard Scaler. This is an important preprocessing step for logistic regression to ensure that all features contribute equally to the model, as they are on different scales.

4. **Model Training**: A logistic regression model is initialized and trained on the scaled training data. This model learns the relationship between the features and the target variable to make predictions.
5. **Model Evaluation**: The trained model's performance is evaluated on the test set using several metrics.

Results and Output

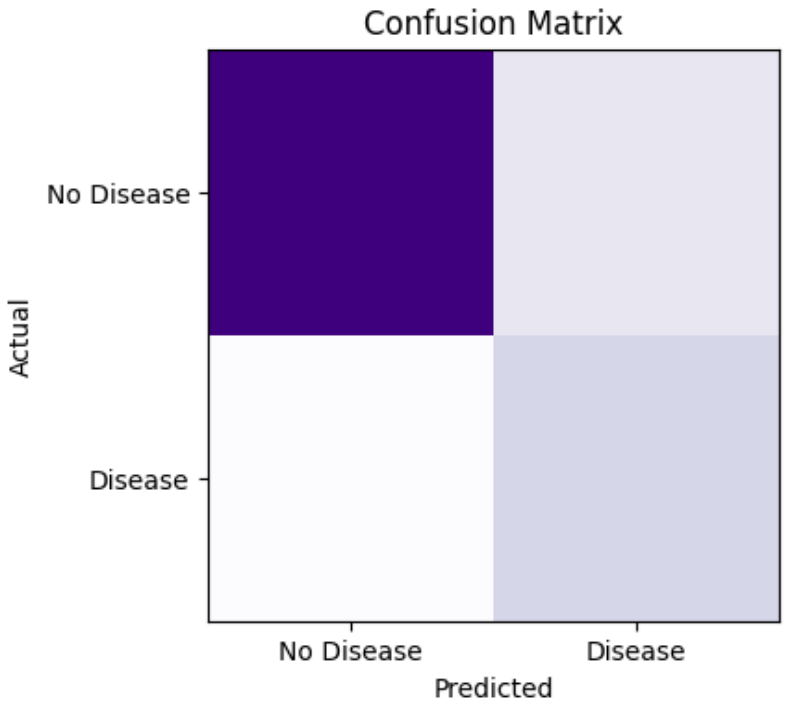
The performance of the logistic regression model on the test data is evaluated using accuracy, recall, F1 score, and a confusion matrix. The results are as follows:

- **Accuracy**: 0.7532
- **Recall**: 0.7111
- **F1 Score**: 0.6274
- **Confusion Matrix**: The confusion matrix provides a detailed breakdown of the model's predictions.
- **Classification report**

The confusion matrix shows the following:

- **True Negatives (TN)**: 84 (Correctly predicted non-diabetic)
- **False Positives (FP)**: 25 (Incorrectly predicted as diabetic)
- **False Negatives (FN)**: 13 (Incorrectly predicted as non-diabetic)
- **True Positives (TP)**: 32 (Correctly predicted as diabetic)

```
Accuracy:  0.7532467532467533
Recall:  0.7111111111111111
f1 score:  0.6274509803921569
Confusion Matrix:  [[84 25]
 [13 32]]
```



Classification report

	precision	recall	f1-score	support
0	0.87	0.77	0.82	109
1	0.56	0.71	0.63	45
accuracy			0.75	154
macro avg	0.71	0.74	0.72	154
weighted avg	0.78	0.75	0.76	154

Conclusion

Based on the given dataset, the logistic regression model performed admirably in predicting diabetes. The model accurately anticipated the result for three-quarters of the test instances, with an accuracy of roughly 75.3%. The model was able to accurately identify a sizable percentage of real diabetes cases, according to the recall score of 71.1%. The model's precision and recall are balanced by its F1 score of 62.7%. Although there is room for improvement by investigating alternative models or feature engineering approaches, the results indicate that the model is a practical tool for diabetes prediction.