Diabetes Prediction Model

Abstract

This report details the analysis of a diabetes dataset and the development of a predictive machine learning model. The objective was to build a classification model to accurately predict the onset of diabetes in individuals based on various diagnostic measurements. The analysis included data exploration, visualization of key features, and the training and evaluation of two different machine learning algorithms: **Logistic Regression** and **Gaussian Naive Bayes**. The models were evaluated based on their accuracy on a held-out test set, with the Logistic Regression model showing a slightly higher performance.

Methodology

The study utilized the dibeties.csv dataset, which contains 768 entries and 9 columns, including health metrics such as Pregnancies, Glucose, Blood Pressure, BMI, and Age, as well as a target variable, Outcome, indicating a diabetes diagnosis.

Data Preprocessing and Exploration:

- Initial checks confirmed there were no missing values in the dataset. However, some features like Glucose, Blood Pressure, and BMI had a minimum value of 0, which is biologically implausible and likely represents missing data.
- The dataset's class distribution for the Outcome variable was visualized using a count plot, revealing an imbalance with approximately 65.1% of the cases being non-diabetic (Outcome = 0) and 34.9% being diabetic (Outcome = 1).
- The age distribution of the individuals was also analysed, showing ages ranging from 21 to 81.

Model Training:

- The dataset was split into training and testing sets to evaluate the models' performance on unseen data. The split resulted in a training set of 614 entries and a testing set of 154 entries.
- Two machine learning models were trained on the training data:

  Logistic Regression: A linear model suitable for binary classification.

  Gaussian Naive Bayes: A probabilistic classifier based on the assumption that features are conditionally independent.

Model Evaluation:

- The performance of both models was measured using the **accuracy score** on the test set. Accuracy is calculated as the proportion of correctly predicted instances out of the total instances.

Results

The models' performance on the test set is summarized below:

Logistic Regression achieved an accuracy of **74.67%**.

Gaussian Naive Bayes achieved an accuracy of **72.72%**.

The results indicate that both models are effective in predicting diabetes from the given features, with **Logistic Regression** slightly outperforming **Gaussian Naive Bayes** in this specific analysis. The findings suggest that the chosen features are strong predictors of the outcome, and the models are capable of generalizing to new, unseen data. Further improvements could be explored by addressing the data imbalance or missing values in the future.