Diabetes Data Analysis and Model Development Report

Introduction

This report presents the results of a data analysis and model development process using a diabetes dataset.

The goal is to predict whether a patient has diabetes based on various diagnostic measurements.

The dataset, provided by the National Institute of Diabetes and Digestive and Kidney Diseases, includes features such as pregnancies, glucose levels, blood pressure, BMI, and age.

Exploratory Data Analysis (EDA)

The initial step in this project involved exploratory data analysis (EDA) to understand the data structure, detect outliers, and explore relationships between variables.

Key steps in this phase included:

- Distribution plots of all numerical features to assess their spread.
- A correlation matrix to understand the relationships between features.
- A count plot to visualize the distribution of diabetic and non-diabetic patients.
- A pair plot to observe relationships between features and their interactions with the target variable (diabetes outcome).

Model Development

The next phase focused on model selection and development for predicting diabetes.

The following steps were performed:

- **Model Selection:** Several classification models were considered, including Logistic Regression, Decision Trees, and Random Forest.
- **Model Training:** The dataset was split into training and test sets (70% training, 30% test) to ensure the model could generalize well.
- **Model Evaluation:** Precision, recall, and F1 scores were calculated to assess model

performance.

- **Hyperparameter Tuning:** Grid search was used to optimize hyperparameters such as regularization strength (C) and solver type. The best-performing parameters were identified as C=0.1 and solver='lbfgs'.

Regression Task

Although the primary goal of this project was classification, a regression model was also developed to predict BMI values based on other features.

A linear regression model was trained and evaluated on the dataset using metrics such as R-squared and Mean Squared Error (MSE).

Conclusion

In this project, a complete data analysis and model development process was conducted.

EDA helped reveal important insights about the data, and a classification model was developed to predict diabetes with optimized performance using hyperparameter tuning.

Additionally, a regression task was carried out to predict BMI values. This project demonstrates the typical steps in a data science workflow, from initial data exploration to model evaluation and optimization.