# Data Mining and Decision Tree Analysis for Diabetes Prediction: An Exploration of Preprocessing, EDA, and Classification Models

Student Number: 100611584 Module: Data Mining and Foundations of AI

# Introduction

## Project Overview

This project aims to predict whether a person had diabetes using machine learning algorithms. Diabetes is a chronic disease that affects millions globally and early prediction is crucial for effective management. Predictive models can help healthcare professionals identify high-risk individuals and take preventive actions

## Dataset Description

The dataset used for this project is the “Healthcare Diabetes Dataset” available on Kaggle. It contains data from a variety of health-related features, including age, BMI, blood pressure, glucose levels and more. The target variable is binary which indicates whether an individual had diabetes (1) or not (0).

## Problem definition and objective

The main goal and problem definition is to predict the likelihood of diabetes based on a patient’s health metrics. By building a predictive model we can achieve this and evaluate different machine learning algorithms to select the best-performing model for accurate predictions.

# Exploratory Data Analysis

## Data Summary and Initial Observations

Upon inspecting the dataset, it was apparent that the data only contains numerical features. A preliminary examination identified missing values, outliers and inconsistencies. It was clear that some features contained zero values where they were not logically possible and caused a skewedness in the data, such as BMI and blood pressure which required data imputation.

## Visualization Techniques

Visualisations were used to understand the distribution of key variables. Histograms provided insight into the distribution of key variables (Figure 1) like BMI, glucose, and insulin levels. Boxplots were used to identify and analyse outliers (Figure 2). A correlation matrix was generated to assess relationships between features and diabetes occurrence.

A group of blue graphs

AI-generated content may be incorrect.

Figure 1: Histogram before Preprocessing showing distribution of key variables

A screenshot of a graph

AI-generated content may be incorrect.

Figure 2: Boxplots before preprocessing to showcase outliers to be handled

## Correlation Analysis

The heatmap of the correlation matrix (Figure 3) revealed that Glucose had the strongest correlation with diabetes, followed by BMI and Age, aligning with established medical research.

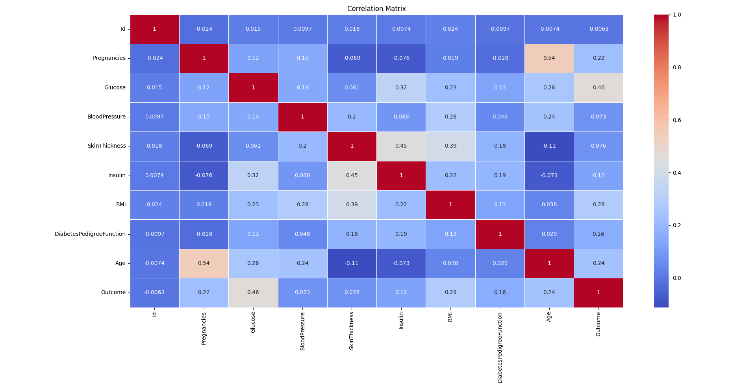


Figure 3: Correlation matrix showcasing the links between features and the target variable

# Data Preprocessing

## Handling Missing Data

Median imputation was applied to skewed variables such as Insulin and DiabetesPedigreeFunction. Mean imputation was used for normally distributed variables like Blood Pressure, Glucose, and BMI.

\*\* Code snippet handling missing values \*\*

A group of blue and white graphs

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## Outlier Detection and Removal

Winsorization was employed to cap extreme values in Blood Pressure, BMI, and Skin Thickness. Log transformation was applied to Insulin and DiabetesPedigreeFunction to reduce skewness.

\*\* Code snippet handling outlier values/removal \*\*

A group of blue graphs

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Post preprocessing the correlation matrix was again analysed to check for any differences in the links between features and the target variable.

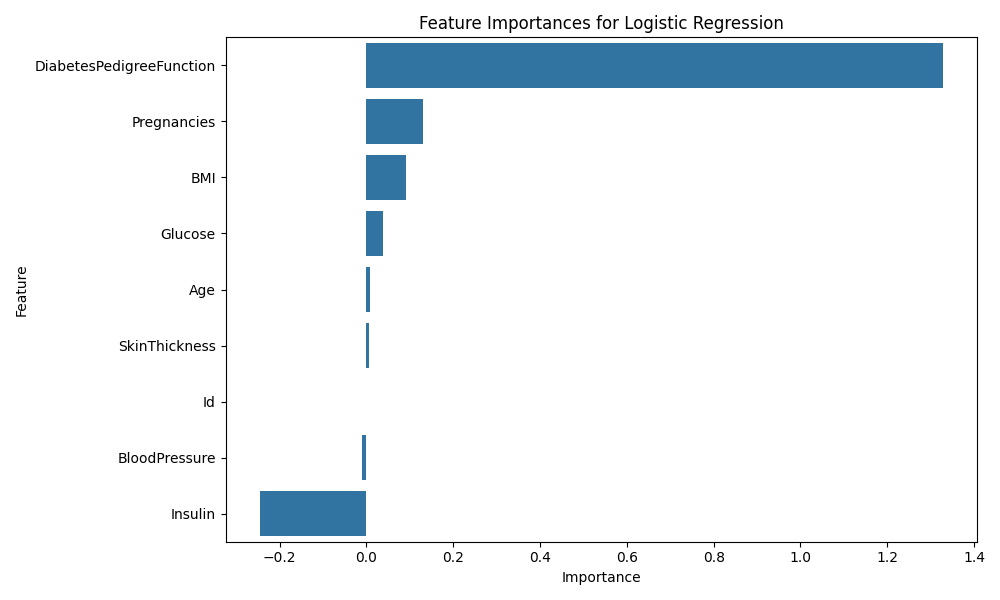
A blue and red squares with white text

AI-generated content may be incorrect.

# Model Selection and Implementation

## Logistic Regression

Logistic regression was selected as a baseline model due to its simplicity and interpretability. It’s a statistical method that models the probability of an outcome based on input features. Since the dataset involves binary classification, Logistic Regression provides a straightforward way to estimate the likelihood of diabetes. Despite its linear nature, it offers valuable insights by determining feature importance through coefficient values.



\*\* Example results \*\*

## Decision Trees

Decision Trees were implemented to capture non-linear relationships between features and diabetes occurrence. The model splits the dataset into hierarchical decision rules to make it easier to interpret. While able to handle both categorical and numerical data without feature scaling, decision trees are prone to overfitting which was mitigated in this implementation by applying pruning techniques and setting depth constraints.

A graph with blue bars

AI-generated content may be incorrect.

\*\* Example Results \*\*

## Model Training and Evaluation

The dataset was split into 80% training and 20% testing, with feature scaling to standardise variables. Hyperparameter tuning was also applied to optimise the performance of the models.

\*\* Code snippet to add for model training \*\*

# Results and Evaluation

## Model Accuracy and Performance Metrics

The performance of both Logistic Regression and Decision Tree models were evaluated using key metrics such as accuracy, precision, recall, F1-score and ROC-AUC. These provide a comprehensive understanding of the model’s ability to correctly classify instances of diabetes.

Logistic Regression achieved competitive results, showcasing its effectiveness in capturing linear relationships between feature sand the target variable. Meanwhile, the Decision Tree model, specifically after hyperparameter tuning, showcased improved performance by capturing non-linear patterns in the data.

The ROC-AUC scores for both models indicated their ability to distinguish between positive and negative cases effectively.

## Confusion Matrix and Classification Report

To further analyse the models’ performance, confusion matrices and classification reports were generated. These provided further insights into the number of true positives, true negatives, false positives and false negatives which helped identify areas in which the models might struggle.

For example, the Logistic Regression model showed a higher false negative rate which indicates that it might have some difficulty in identifying individuals with diabetes. The Decision Tree model, meanwhile, after tuning, reduced the negative rate. This made it more reliable for this classification task. The classification reports highlighted the precision and recall trade-offs, with the Decision Tree model achieving a better balance.

## Cross-Validation Results

Cross validation was performed using 5-fold cross validation to ensure robustness of the models and reduce the risk of overfitting. The mean cross-validation accuracy for Logistic Regression was consistent across folds, indicating stable performance. Similarly, the Decision Tree model showed improved cross-validation accuracy after hyperparameter tuning, confirming its ability to generalise well to unseen data. These results validate the reliability of the models and provide confidence in their predictive capabilities.

# Discussion

## Key Insights from the Analysis

The analysis revealed several insights. Glucose levels, BMI and age were identified as the most significant predictors of diabetes, aligning with established medical research. Logistic regression provided a strong baseline model, effectively capturing linear relationships between features and the target variable. However, the Decision Tree model demonstrated superior performance by capturing non-linear patters in the data, particularly after hyperparameter tuning. The ROC-AUX scores for both models indicated at their strong ability to distinguish between diabetic and non-diabetic individuals, with Decision Tree model achieving a better balance between precision and recall.

## Challenges Encountered and Solutions

Several challenges were identified while working on the data and models. Missing values in key features, such as BMI and blood pressure posed a significant issue as well as outliers in features like insulin and skin thickness. The prior was handled with imputation techniques, with median imputation and applied to skewed variables and mean imputation for normally distributed values. While outliers were mitigated with the use of Winsorisation and log transformations to reduce

## Implications of the Results

The results of the analysis have important implications for diabetes prediction as the models developed in this project can help assist healthcare professionals in identifying high-risk individuals and enable earlier intervention and management to better the lives of individuals and reduce the impact of diabetes on their lifestyles. The insights gained could also guide future research and data collection efforts, focussing on the most impactful predictors of diabetes. Furthermore, the success of hyperparameter tuning highlights the importance of model optimisation in achieving reliable and accurate predictions.

# Conclusion

## Summary of Findings

This project successfully developed and evaluated machine learning models for predicting diabetes based on health metrics. Logistic Regression served as a robust baseline model, while Decision Tree model demonstrates superior performance by capturing non-linear relationships. Ky features such as glucose levels, BMI and age were identified as the most significant predictors of diabetes. The models themselves were validated using cross-validation, ensuring robustness and generalisability to unseen data.

## Future Work and Improvements

Future work can have more focus on various areas to further enhance both the models themselves and their applicability, on diabetes and other medical conditions. Additional data for collection, particularly for underrepresented populations, could be used to improve the model’s generalisation. The use of advanced machine learning algorithms, such as Random Forest, Gradient Boosting or Neural Networks, could be explored to achieve higher predictive accuracy. Feature engineering, such as creating interaction terms or binning continuous variables, could also be explored to help improve model performance. Finally, deploying modes in real-world healthcare settings and integrating them into decision-support systems could maximize their impact on diabetes prevention and management.