# 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 like BMI, glucose, and insulin levels. Boxplots were used to identify and analyse outliers. A correlation matrix was generated to assess relationships between features and diabetes occurrence.

\*\* Visualisations pre-pre-processing \*\*

## Correlation Analysis

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

# 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 \*\*

\*\* Visualization after handling \*\*

## 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 \*\*

\*\* Visualization after handling \*\*

Following these preprocessing steps, histograms and boxplots were used to verify data consistency.

# 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.

\*\* 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.

## Cross-Validation Results

# Discussion

## Key Insights from the Analysis

## Challenges Encountered and Solutions

## Implications of the Results

# Conclusion

## Summary of Findings

## Future Work and Improvements

# References

There are no sources in the current document.