# 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

Diabetes is a prevalent chronic disease that affects millions worldwide, with its prevalence steadily increasing. Early prediction and diagnosis are vital for managing the disease effectively and preventing complications. Traditional methods for diabetes prediction often rely on clinical criteria such as blood sugar levels, BMI, and family history. However, these approaches can be limited by their reliance on specific thresholds and may not fully account for complex interactions between health metrics. This project seeks to explore the potential of machine learning techniques to enhance the accuracy and reliability of diabetes prediction. Machine learning algorithms, through their ability to uncover patterns in large datasets, may provide more precise predictions by analysing a broader range of health-related features.

## Dataset Description

The dataset used for this project, the “Healthcare Diabetes Dataset” from Kaggle, includes health metrics like age, BMI, blood pressure, glucose levels, and more. The target variable is binary, indicating whether an individual has diabetes (1) or not (0). This dataset provides an opportunity to test the effectiveness of various machine learning models in predicting diabetes.

## Problem definition and objective

The main research question of this project is: How can machine learning models predict the likelihood of diabetes based on various health metrics, and how do these predictions compare to traditional methods in terms of accuracy and reliability? By developing and evaluating machine learning models, this project aims to determine the most effective algorithm for predicting diabetes risk. The key objective is to assess the performance of models such as Logistic Regression and Decision Trees, comparing their effectiveness in handling the complex, nonlinear relationships often present in health data, and to evaluate their ability to identify at-risk individuals for early intervention.

# 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.

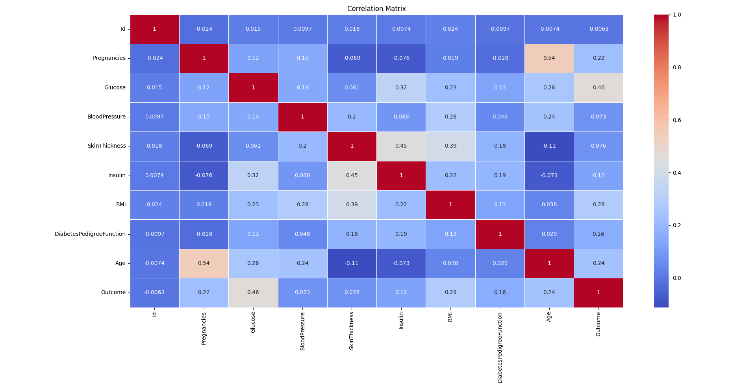


Figure 3: Correlation matrix showcasing the links between features and the target variable

# 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.

*## handling missing values*

*# Replacing 0 values in skewed data with median*

skewed\_columns = ['Insulin', 'DiabetesPedigreeFunction', 'SkinThickness']

for col in skewed\_columns:

    data[col] = data[col].replace(0, data[col].median())

*# Replacing 0 values in normally distributed data with mean*

normal\_columns = ['BloodPressure', 'Glucose', 'BMI']

for col in normal\_columns:

    data[col] = data[col].replace(0, data[col].mean())

*# Plot histograms and boxplots after handling missing values*

plot\_histograms(data, columns\_to\_plot)

plot\_boxplots(data, columns\_to\_plot)

*# Final check - pregnancies and outcome can be 0 but the others are handled*

zero\_values = (data == 0).sum()

**print**("Count of 0 values in each column:")

**print**(zero\_values)

## 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.

*## handling outliers for each feature*

*# Glucose: Cap values at 1st and 99th percentiles (Winsorization)*

data["Glucose"] = data["Glucose"].clip(data["Glucose"].quantile(0.01), data["Glucose"].quantile(0.99))

*# Blood Pressure: Replace 0 values with the median and cap extreme outliers*

data["BloodPressure"] = data["BloodPressure"].replace(0, data["BloodPressure"].median())

data["BloodPressure"] = data["BloodPressure"].clip(data["BloodPressure"].quantile(0.01), data["BloodPressure"].quantile(0.99))

*# Cont…*

*# Skin Thickness: Cap extreme values at the 99th percentile*

data["SkinThickness"] = data["SkinThickness"].clip(data["SkinThickness"].quantile(0.01), data["SkinThickness"].quantile(0.99))

*# Insulin: Use log transformation to reduce skewness*

data["Insulin"] = np.log1p(data["Insulin"])

*# BMI: Winsorization to cap extreme values*

data["BMI"] = data["BMI"].clip(data["BMI"].quantile(0.01), data["BMI"].quantile(0.99))

*# Diabetes Pedigree Function (DPF): Log transform to reduce the impact of large values*

data["DiabetesPedigreeFunction"] = np.log1p(data["DiabetesPedigreeFunction"])

*# Plot histograms and boxplots after handling outliers*

plot\_histograms(data, columns\_to\_plot)

plot\_boxplots(data, columns\_to\_plot)

A group of blue graphs

AI-generated content may be incorrect.

Figure : Histogram post preprocessing, more normally distributed values for all the features

A screenshot of a computer

AI-generated content may be incorrect.

Figure : Boxplot post preprocessing, fewer outliers without compromising data

Post preprocessing the correlation matrix (Figure 6) 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.

Figure : Correlation matrix post preprocessing

# 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 (Figure 7) through coefficient values.

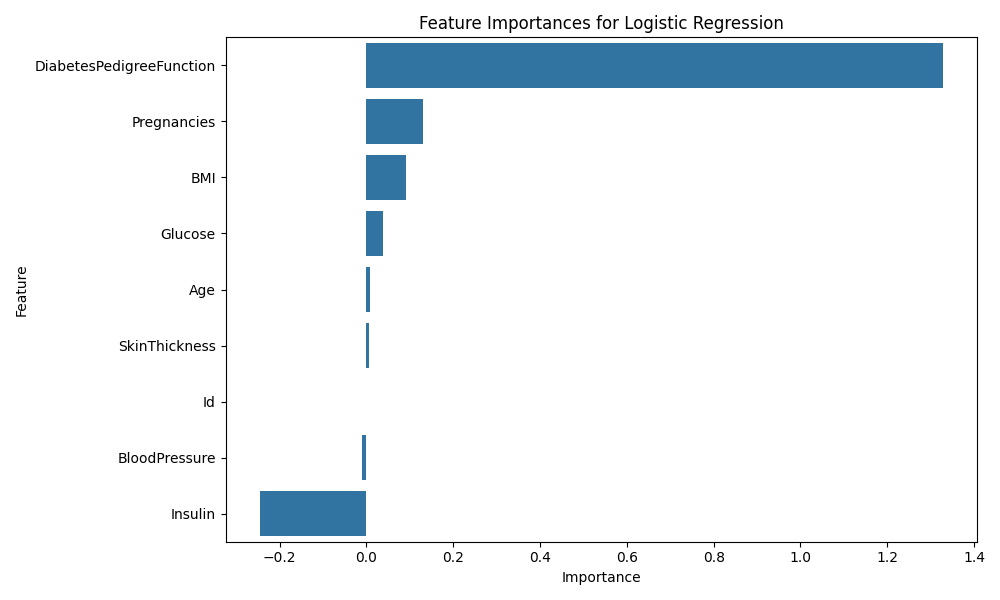


Figure : Feature Importance for Logistic Regression for test diagnosis

## 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.

Figure : Feature Importance for Decision Tree for test diagnosis

## 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.

*# Load the train and test sets*

X\_train = pd.read\_csv('data/X\_train.csv')

X\_test = pd.read\_csv('data/X\_test.csv')

y\_train = pd.read\_csv('data/y\_train.csv')

y\_test = pd.read\_csv('data/y\_test.csv')

*# Convert y\_train and y\_test to 1D arrays*

y\_train = y\_train.values.ravel()

y\_test = y\_test.values.ravel()

*# Function to evaluate model performance*

def **evaluate\_model**(model, X\_train, y\_train, X\_test, y\_test):

    model.fit(X\_train, y\_train)

    y\_pred = model.predict(X\_test)

    accuracy = accuracy\_score(y\_test, y\_pred)

    precision = precision\_score(y\_test, y\_pred)

    recall = recall\_score(y\_test, y\_pred)

    f1 = f1\_score(y\_test, y\_pred)

    roc\_auc = roc\_auc\_score(y\_test, y\_pred)

    return accuracy, precision, recall, f1, roc\_auc

# Results and Evaluation

## Model Accuracy and Performance Metrics

The performance of both Logistic Regression and Decision Tree models were evaluated using key metrics (Figure 9) 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.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Model analysis with key metrics for understanding of model's diagnosis ability

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 (Figure 10) showed a higher false negative rate which indicates that it might have some difficulty in identifying individuals with diabetes. The Decision Tree model (Figure 11) , 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.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Confusion Matrix and Classification Report for Logistic Regression model

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Confusion Matrix and Classification Report for Descision Tree model

## Cross-Validation Results

Cross validation (Figure 12) 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.

A computer screen shot of a number

AI-generated content may be incorrect.

Figure : Cross Validation accuracy scores for Logistic Regression and Decision Tree models

# Discussion

**Key Insights from the Analysis**

The analysis revealed several important 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 patterns in the data, particularly after hyperparameter tuning. The ROC-AUC scores for both models indicated their ability to distinguish between diabetic and non-diabetic individuals, with the Decision Tree model achieving a better balance between precision and recall.

## Critical Analysis and Model Comparison

While both Logistic Regression and Decision Tree models were implemented, their performance differences highlight key aspects of machine learning in medical diagnostics.

Logistic Regression served as a strong baseline model due to its simplicity and interpretability. It assumes a linear relationship between features and the target variable, making it effective in datasets where such relationships hold. However, its performance was slightly weaker in this case, as some features in the dataset exhibited non-linear interactions with diabetes occurrence. The confusion matrix showed that Logistic Regression had a higher false negative rate, meaning it struggled to correctly classify some diabetic individuals. This limitation could have significant implications in real-world applications, where missing a positive diabetes case could delay necessary medical intervention.

On the other hand, the Decision Tree model captured these non-linear relationships more effectively. By iteratively splitting the data based on feature importance, it identified complex decision boundaries that improved classification performance. The Decision Tree’s ability to handle interactions between variables allowed it to reduce false negatives compared to Logistic Regression. However, Decision Trees are prone to overfitting, which was mitigated in this case through hyperparameter tuning, specifically pruning and depth constraints.

## 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

While the comparison was limited to these two models, further improvements could be made by exploring ensemble methods. Techniques such as Random Forest or Gradient Boosting could enhance performance by combining multiple decision trees to reduce overfitting while improving generalization. Although these models were not implemented in this assessment, their potential for improving diabetes prediction should be acknowledged.

Additionally, feature engineering could further refine model performance. Exploring polynomial features, interactions between key health metrics, or feature selection techniques could enhance predictive accuracy. Furthermore, an evaluation of ethical considerations, such as bias in medical datasets, should be included to ensure fairness and reliability in real-world applications.

Overall, this study demonstrates the importance of model selection in predictive healthcare. The results highlight the trade-offs between simplicity, interpretability, and predictive power, emphasizing the need for thoughtful consideration when applying machine learning to medical diagnoses.