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

Table of Contents

[Introduction 2](#_Toc193116486)

[Overview of the Dataset 2](#_Toc193116487)

[Problem definition and objective 2](#_Toc193116488)

[Exploratory Data Analysis 2](#_Toc193116489)

[Data Summary and Initial Observations 2](#_Toc193116490)

[Visualization Techniques (e.g., Histograms, Scatter Plots) 2](#_Toc193116491)

[Correlation Analysis 2](#_Toc193116492)

[Data Preprocessing 2](#_Toc193116493)

[Handling Missing Data 2](#_Toc193116494)

[Encoding Categorical Variables 2](#_Toc193116495)

[Outlier Detection and Removal 2](#_Toc193116496)

[Model Selection and Implementation 2](#_Toc193116497)

[Logistic Regression 2](#_Toc193116498)

[Decision Trees 2](#_Toc193116499)

[Results and Evaluation 3](#_Toc193116500)

[Model Accuracy and Performance Metrics 3](#_Toc193116501)

[Confusion Matrix and Classification Report 3](#_Toc193116502)

[Cross-Validation Results 3](#_Toc193116503)

[Comparison with Other Models 3](#_Toc193116504)

[Key Insights from the Analysis 3](#_Toc193116505)

[Challenges Encountered and Solutions 3](#_Toc193116506)

[Implications of the Results 3](#_Toc193116507)

[Summary of Findings 3](#_Toc193116508)

[Future Work and Improvements 3](#_Toc193116509)

[References 3](#_Toc193116510)

# 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.

## Problem definition and objective

# Exploratory Data Analysis

## Data Summary and Initial Observations

## Visualization Techniques (e.g., Histograms, Scatter Plots)

## Correlation Analysis

# Data Preprocessing

## Handling Missing Data

## Encoding Categorical Variables

## Outlier Detection and Removal

# Model Selection and Implementation

## Logistic Regression

## Decision Trees

# Results and Evaluation

## Model Accuracy and Performance Metrics

## Confusion Matrix and Classification Report

## Cross-Validation Results

## Comparison with Other Models

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.