Data Science Applications

ደብረ ብርሃን ዩኒቨርሲቲ

DBU

DEBRE BIRHAN UNIVERSITY

TITLE: Diabetes Prediction Data Preparation Project

Name : Yohannis Hailye

ID: 1501575

Submit to : Petros A.

Table Of Content

A. Executive Summary..1
B. Data Quality Assessment...1
C. Methodology and Justification.. 2
D. Results and Discussion...2
E. Challenges and Solutions.. 3

A. Executive Summary

Briefly describe:

- Purpose: To clean, transform, and prepare diabetes data for ML model training.
- Outcome: Final processed dataset ready for predictive modeling with balanced target classes.

Example:

> This project aims to preprocess the Pima Indian Diabetes dataset to ensure data quality, consistency, and readiness for machine learning. The preprocessing pipeline included handling missing values, transforming variables, detecting and treating outliers, reducing dimensionality, and addressing class imbalance. The final dataset provides a balanced, standardized, and feature-engineered version suitable for predictive modeling.

B. Data Quality Assessment

Summarize key findings:

- Missing values identified in Glucose, BloodPressure, SkinThickness, Insulin, and BMI.
- Outliers detected using IQR and Z-Score methods.
- Zero values treated as biologically impossible → replaced with NaN → imputed.
- Dataset now has no missing or invalid entries.

C. Methodology and Justification

Explain why you used each step:

Phase	Description	Method / Justification
Phase 1	Data Understanding	Inspected dataset, data types, distributions
Phase 2	Data Cleaning	Replaced 0s → NaN, imputed missing with mean/median
Phase 3	Transformation	Added Age & BMI categories, scaled numeric features
Phase 4	Reduction	Used PCA for variance visualization
Phase 5	Imbalance Handling	Applied SMOTE+ENN for balanced classes

D. Results and Discussion

Summarize:

- **Imputation** reduced missing rate to 0%.
- **Feature Engineering** added meaningful categorical insights.
- **PCA** showed top components explain ~90% of variance.
- **SMOTE+ENN** achieved perfectly balanced target distribution.

Include visualizations:

- Missing value heatmap
- Before/after outlier boxplots
- PCA variance graph
- Class distribution before/after balancing

E. Challenges and Solutions

Challenge	Solution
Missing & impossible values	Replaced zeros with NaN and applied imputation
High outlier influence	Used IQR method for capping extreme values
Imbalanced target	Used SMOTE+ENN to create class balance
Feature scaling differences	Compared StandardScaler and MinMaxScaler before finalizing