Diabetes Analysis Project - Data Dictionary

This document provides information about all datasets and variables used in the diabetes analysis project.

Raw Dataset

The raw dataset (`diabetes_raw.csv`) contains information about female patients of Pima Indian heritage, with various health measurements and a binary outcome indicating diabetes diagnosis.

Variable	Description	Type	Range/Units	Potential Issues
Pregnancies	Number of times pregnant	Integer	0-17	None
Glucose	Plasma glucose concentration at 2 hours in an OGTT	Integer	0-199 mg/dL	Contains zero values
BloodPressure	Diastolic blood pressure	Integer	0-122 mm Hg	Contains zero values
SkinThickness	Triceps skin fold thickness	Integer	0-99 mm	Contains zero values
Insulin	2-Hour serum insulin	Integer	0-846 mu U/ml	Contains zero values
BMI	Body mass index	Float	0-67.1 kg/m^2	Contains zero values

DiabetesPedigreeFunction	Diabetes pedigree function (genetic influence)	Float	0.078-2.42	None
Age	Age in years	Integer	21-81	None
Outcome	Class variable (0: No diabetes, 1: Diabetes)	Binary	0 or 1	None

Cleaned Dataset

The cleaned dataset (`cleaned_data.csv`) is a processed version of the raw data with missing values imputed and other cleaning steps applied.

- Zero values in physiologically impossible columns (Glucose, BloodPressure, SkinThickness, Insulin, BMI) were replaced with NaN
- Missing values were imputed using median values stratified by outcome class
- Duplicates were removed
- Outliers were identified but retained in the dataset

Engineered Dataset

The engineered dataset (`features_engineered.csv`) includes the original variables plus derived features.

Added Feature	Description	Formula
Diabetes_Risk_Index	Combined risk based on BMI and glucose	BMI * Glucose / 100
Insulin_Sensitivity	Ratio of insulin to glucose	Insulin / Glucose
Age_BMI_Factor	Age-adjusted BMI metric	Age * BMI / 100

Pregnancies_Age_Ratio	Pregnancy frequency relative to age	Pregnancies / Age
Genetic_Physical_Risk	Combined genetic and physical risk	DiabetesPedigreeFunction * BMI
Glucose_BP_Ratio	Ratio of glucose to blood pressure	Glucose / BloodPressure

Derived Categorical Variables

The following categorical variables were created during analysis:

Categorical Variable	Description	Categories
AgeGroup	Age grouped into ranges	'20-30', '30-40', '40-50', '50-60', '60+'
BMI_Category	BMI grouped by standard categories	'Underweight', 'Normal', 'Overweight', 'Obese'
Glucose_Category	Glucose levels by medical categories	'Low', 'Normal', 'Prediabetes', 'Diabetes'

Scaled Dataset

The scaled dataset (`features_engineered_scaled.csv`) contains normalized versions of the features to improve model performance:

- Standard scaling (zero mean, unit variance)
- Min-max scaling (0-1 range)
- Robust scaling (based on quantiles, less sensitive to outliers)

Feature Importance

Based on model analysis, the following features were identified as most important for predicting diabetes:

1. Glucose
2. BMI
3. Diabetes_Risk_Index (engineered feature)
4. Age
5. DiabetesPedigreeFunction

Model Outputs

The project includes several trained models:

Model File	Description
random_forest.pkl	Random Forest base model
logistic_regression.pkl	Logistic Regression base model
gradient_boosting.pkl	Gradient Boosting base model
random_forest_optimized.pkl	Hyperparameter-tuned Random Forest
logistic_regression_optimized.pkl	Hyperparameter-tuned Logistic Regression
gradient_boosting_optimized.pkl	Hyperparameter-tuned Gradient Boosting
ensemble_model.pkl	Voting ensemble of best-performing optimized models