**AI-Powered Diabetes Prediction System - Data Loading and Preprocessing**

**Introduction**

This document outlines the process of loading and preprocessing the dataset for the AI-powered diabetes prediction system. The dataset contains medical features such as glucose levels, blood pressure, BMI, etc., along with information about whether the individual has diabetes or not. The purpose of this document is to describe the steps taken to prepare the data for machine learning model training and analysis.

**Data Loading**

The dataset was loaded into a pandas DataFrame using the pd.read\_csv('file.csv') function. The file.csv represents the path to the CSV file containing the dataset.

**Data Exploration**

After loading the dataset, an initial exploration was performed to understand the data's characteristics. This included:

Checking the first few rows of the dataset using data.head().

Examining the data types of each column using data.dtypes.

Identifying missing values using data.isnull().sum().

Obtaining summary statistics using data.describe().

**Data Preprocessing**

The following preprocessing steps were carried out:

**1.Handling Missing Values:**

Rows with missing target labels were removed, as these records are not useful for training the model.

Missing values in the feature columns were imputed using the median value of the respective columns.

**2.Data Normalization:**

Numerical features were standardized to have a mean of 0 and a standard deviation of 1. This was done to ensure that the model performs well regardless of the scale of the features.

**3.Feature Encoding:**

Categorical features were encoded using one-hot encoding, which converts categorical data into a format that can be fed into machine learning algorithms.

**4.Feature Selection:**

A correlation matrix was computed to identify features that are highly correlated with the target variable (diabetes status). Features with a low correlation were dropped to simplify the model and reduce overfitting.

**Data Analysis**

After preprocessing, various analyses were performed to gain insights into the data, including:

Visualizing the distribution of the target variable to check for class imbalance.

Plotting histograms and box plots of the feature variables to understand their distributions and identify outliers.

Analyzing the relationships between features and the target variable using scatter plots and pair plots.

**Conclusion**

The dataset has been successfully loaded and preprocessed for the diabetes prediction system. The data exploration and analysis have provided valuable insights into the characteristics of the data, which will guide the model selection and training process. The next step in the project is to train machine learning models using the preprocessed data and evaluate their performance in predicting diabetes risk.