AI BASED DIABETES PREDICTION SYSTEM

NAME: M RUDHRESHKUMAR REGISTER NO: 410121243045 DOCUMENT SUBMISSION PHASE 3 ---

When building a project for an AI-based diabetic prediction system, loading and preprocessing the dataset are crucial steps. Here's how you might approach it:

DATA COLLECTION:

Data collection : obtain a dataset that contains relevant information for predicting diabetes. This dataset could include features like age, BMI, glucose levels, family history, and more.

DATA LOADING:

Data Loading: Use a programming language like Python and libraries like pandas to load the dataset from a ﬁle (e.g., CSV) or from a database.

DATA EXPLORATION:

Data Exploration: Explore the dataset to understand its structure and the distribution of features. This might include checking for missing values and visualizing data to identify patterns.

DATA PREPROCESSING:

Handling Missing Data: Deal with any missing values in the dataset. You can choose to impute missing values or remove rows with missing data.

Feature Scaling/Normalization: Depending on the algorithms you plan to use, you might need to scale or normalize features to ensure they are on the same scale.

Feature Selection: Identify which features are most relevant for predicting diabetes and consider removing less important ones.

Data Splitting: Split the dataset into training, validation, and test sets. This allows you to train and evaluate your prediction model eﬀectively.

Data Visualization: Create visualizations to understand the relationships between diﬀerent features and how they relate to diabetes outcomes. This can help you gain insights into the data.

DATA AUGMENTATION:

Data Augmentation (if applicable): In some cases, you might augment the data with additional features or generate synthetic samples to improve the performance of the model.

HANDLING CLASS IMBALANCE:

Handling Class Imbalance (if applicable): If the dataset has an imbalance between diabetic and non-diabetic cases, you may need to address this issue by oversampling, undersampling, or using other techniques.

ENCODING CATEGORICQL VARIABLE:

Encoding Categorical Variables (if applicable): If your dataset contains categorical features, you may need to encode them into a numerical format suitable for machine learning models.