**Documentation**

**Problem statement:** To predict the likelihood of diabetes in patients based on health attributes using machine learning techniques.

**Approach:**

**Step-1:** collection of data

* The required dataset was collected from **Kaggle**, containing patient health records such as glucose level, BMI, insulin, blood pressure, and age for diabetes prediction.

**Step-2:** preprocessing

* preprocessing steps including column renaming, handling missing values, removing outliers via Z-score, and descriptive statistical analysis. Exploratory Data Analysis (EDA) is conducted through distribution plots, boxplots, and correlation heatmaps to understand feature significance**.**

**Step-3:** model development and evaluation

* Multiple classification algorithms were tested with hyperparameter tuning to optimize performance. Each model was evaluated using accuracy score, confusion matrix, classification report, and cross-validation score to ensure reliability and robustness in diabetes prediction.

**Step-4:** model selection

* Among the tested classification algorithms, Random Forest delivered the best performance across evaluation metrics. Therefore, it was selected as the final predictive model for diabetes detection.

**Problems encountered during model building:**

**-** I attempted to categorize Blood Pressure and BMI, but all models performed poorly, so I decided not to proceed with this.

- The dataset is very small, so I handled the zeros present in the SkinThickness and Insulin columns by replacing them with the mean values.