Team 2 Project Proposal

The Diabetes Prediction dataset is a comprehensive collection of medical and demographic information from patients, labeled with their diabetes status (positive or negative). The dataset includes key health indicators such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c levels, and blood glucose levels. With this data, machine learning models can be developed to predict the likelihood of diabetes in patients, leveraging both medical history and demographic attributes to enhance early detection and preventive care efforts.

SMART Questions:

- What are the primary risk factors (e.g., blood glucose level, BMI, age, hypertension) for diabetes in this population, and how accurately can they predict the likelihood of a diabetes diagnosis?
- Are there any notable differences in diabetes prevalence based on gender, age, or smoking history? For example, does smoking history combined with high blood glucose level increase the risk?
- How does the distribution of BMI and HbA1c levels differ between those with and without diabetes?
- How sensitive are the models to changes in certain variables (e.g., slight increases in blood glucose or BMI)? Can we identify an actionable threshold for intervention?
- Can unsupervised learning techniques (K-means clustering) identify subgroups within the dataset that are at higher or lower risk for diabetes? What are the defining characteristics of these clusters?

Modeling Methods:

- **Logistic Regression:** This is a simple and interpretable model for binary classification that can help identify which features have the strongest association with diabetes. It's a good baseline model for understanding the dataset.
- **Decision Trees:** Decision trees are effective for handling both numerical and categorical data. They provide easy-to-interpret rules and can help identify important feature splits that indicate higher diabetes risk.
- **Random Forests**: A method that builds multiple decision trees to improve prediction accuracy. Random forests are robust to overfitting and can highlight the most important predictors across many decision paths, making them suitable for datasets with mixed data types.
- **Gradient Boosting Machines (ex: XGBoost)**: This method combines multiple weak learners to create a powerful predictive model. It often achieves high accuracy in classification tasks and can capture complex interactions among variables.

Github Link: https://github.com/tgormley/DATS6103_FinalProject_Team2

Dataset Source: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset