Pima Indians Diabetes Dataset Proposal

Dataset Name: Pima Indians Diabetes Dataset
Source: Kaggle (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)
Original Provider: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)
Dataset Size: 768 observations, 8 features, 1 target variable

Objective

The goal is to predict whether a patient has diabetes based on diagnostic health measurements. The target variable is Outcome, where 1 = diabetic and 0 = non-diabetic. Several people in my family have type 1 diabetes, so I find this dataset particularly interesting. Seeing what variables directly influence diabetes, and what might not could be very helpful. This dataset has 8 features, so EDA will be easier, but at 768 data entries, it gives me plenty of meat to train different models on. This dataset also has no missing values which makes training data a lot easier.

Dataset Variables

Variable	**Type**	**Description**
Pregnancies	Integer	Number of times pregnant
Glucose	Integer	Plasma glucose concentration
BloodPressure	Integer	Diastolic blood pressure (mm Hg)
SkinThickness	Integer	Triceps skinfold thickness (mm)
Insulin	Integer	2-Hour serum insulin (mu U/ml)
BMI	Float	Body Mass Index (kg/m^2)
DiabetesPedigreeFunction	Float	Diabetes risk score from family history
Age	Integer	Age in years
Outcome	Binary	1 if diabetic, 0 if not