CS 4710: Final Project Group Proposal

Team Members: Asim Koirala (ak3vmg), Mujtaba Khan (hhv6bx), Carlos Revilla (cfr5spw)

**Project Idea:** Utilizing AI to Predict Diabetes in the Pima Indian Population using the Pima Indians Diabetes Database from kaggle.com.

**Motivation + Social Good:**
The Pima Indian population has been observed to have a high prevalence of type 2 diabetes, making it a significant public health concern. By using the Pima Indians Diabetes database, our project aims to develop an AI model that can accurately predict the onset of diabetes within this community based on various diagnostic measurements from the dataset. The goal is to create models that can correctly identify patterns in the Pima Indians database which can be extrapolated to form a prediction for the early diagnosis of type 2 diabetes, which would significantly impact patient care.

**Methods:**
Our project plans to use ML techniques, particularly focusing on supervised learning algorithms such as Logistic Regression and the Naive Bayes Classifier. These ML methods will be applied to predict the 'Outcome' variable based on the medical predictor variables provided in the dataset. By experimenting with different algorithms, we want to allow for comparison in order to easily catch any inconsistencies and give us the opportunity to fine tune our models for a more accurate prediction.

**Data and Library:**
The Pima Indians Database includes several medical predictor variables and one target variable, Outcome.

| **Variable Name** | **Data type** | **Description** |
|---|---|---|
| Pregnancies | Integer | Number of times pregnant |
| Glucose | Integer | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| BloodPressure | Integer | Diastolic blood pressure (mm Hg) |
| SkinThickness | Integer | Triceps skin fold thickness (mm) |

| | | |
|---|---|---|
| Insulin | Integer | 2-Hour serum insulin (mu U/ml) |
| BMI | Float | Body mass index (weight in kg/(height in m)^2) |
| DiabetesPedigreeFunction | Float | Diabetes pedigree function |
| Age | Integer | Age (years) |
| Outcome (Target) | Integer | Class variable (0 or 1) |

For our analysis, we plan to utilize Python along with python libraries such as Pandas for data manipulation, Scikit-learn for machine learning models, and Matplotlib for data visualization.

**Experiments:**
Our experimental setup will be focused on employing Logistic Regression and Naive Bayes algorithms. The experiments are designed to not only assess the performance of these models but also to understand the data and its implications for diabetes prediction. The process is as follows:

*Exploratory Data Analysis (EDA):*

- Initially, we will utilize Data Cleaning and Preprocessing techniques to handle any missing values/anomalies in our data.
- We can then use statistical methods such as Normal distribution analysis to understand the distributions of various data columns from our csv file. This phase will include visualization of the data distributions, correlation analysis of our variables, and identification of potential predictors for diabetes outcomes.
- Implementing standardization - Since Logistic Regression and Naive Bayes can be sensitive to the scale of input features, we will standardize the dataset to have a mean of 0 and a standard deviation of 1 for each predictor variable.

*Model Implementation and Analysis:*

- Based on insights from EDA, we may implement new techniques or adjust existing models to improve the overall model performance. This could include normalization of skewed variables.
- Model Training will train both Logistic Regression and Naive Bayes by splitting the dataset with possibly ⅔ for training set and ⅓ for testing set.
- Feature Importance: For the Logistic Regression model, we will examine the coefficients to interpret the importance and influence of each feature on the prediction outcome.