**Bonertz, Brian**

**Project 2**

**Week 5 Project Proposal:**

**Title: Diabetes Predictive Analysis**

**Introduction**

Diabetes is a chronic condition that affects millions of people globally. Understanding lifestyle conditions, personal habits and family history can better prepare individuals for what lies ahead. This project aims to analyze genetic predispositions, lifestyle habits, and key health metrics that influence the onset of diabetes.

**Problem Statement**

Using data and predictive analysis, can individuals be identified as high risk for Type 2 diabetes or prediabetes through evaluation of genetic predisposition and analysis of lifestyle factors?

**Objectives**

1. To analyze the dataset to identify key factors affecting the diagnosis of prediabetes and Type 2 diabetes.
2. To develop prediction models for identifying individuals at high risk of prediabetes or Type 2 diabetes.
3. Evaluate behavioral insights that contribute to diagnosis of prediabetes and Type 2 diabetes.
4. To examine demographic trends that correlate with individuals with prediabetes and Type 2 diabetes.

**Data Description**

The dataset named “Diabetes in Youth vs Adult in India” was sourced from Kaggle website and consists of over 100,000 rows of data and 22 features containing both categorical and numerical data types. Data features are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | ID | Unique identifier per individual | Numerical (INT) |
| 2 | Age | Age of the individual sampled | Numerical (INT) |
| 3 | Gender | Male or Female designation of individual | Categorical |
| 4 | Region | Region located in India | Categorical |
| 5 | Family\_Income | Annual income of individual sampled | Categorical |
| 6 | Family\_History\_Diabetes | Determination of family history of diabetes | Categorical |
| 7 | Parent\_Diabetes\_Type | Diabetes type of the individual sampled | Categorical |
| 8 | Genetic\_Risk\_Score | Genetic risk factor of the individual | Numerical (INT) |
| 9 | BMI | Body mass index | Numerical (INT) |
| 10 | Physical\_Activity\_Level | Activity assessment of individual sampled | Categorical |
| 11 | Dietary\_Habits | Dietary intake rating of individual | Categorical |
| 12 | Fast\_Food\_Intake | Fast food rating ranked from 1 - 10 | Numerical (INT) |
| 13 | Smoking | Smoking designation of the individual | Categorical |
| 14 | Alcohol\_Consumption | Alcohol consumption (Yes / No) | Categorical |
| 15 | Fasting\_Blood\_Sugar | Blood sugar measurement | Numerical (INT) |
| 16 | HbA1c | Measure of average blood glucose levels | Numerical (INT) |
| 17 | Cholesterol\_Level | Cholesterol measurement of individual | Numerical (INT) |
| 18 | Prediabetes | Indication of prediabetes (Yes / No) | Categorical |
| 19 | Diabetes\_Type | Diagnosis of Diabetes Type by individual | Categorical |
| 20 | Sleep\_Hours | Average hours of sleep measured | Numerical (INT) |
| 21 | Stress\_Level | Stress level assessment | Numerical (INT) |
| 22 | Screen\_Time | Hours of social media screen time | Numerical (INT) |

**Methodology**

* Data Preprocessing: Handle missing values, encode categorical variables, and normalize numerical values.
* Exploratory Data Analysis: Identify patterns, correlations, and key features influencing diabetes.
* Model Selection: Because of their predictive abilities, I will evaluate linear regression and random forest to find the most suitable model. If their performance is not remarkable, other machine-learning algorithms will be investigated.
* Model Training and Testing: Data will be split into training and testing sets to train the model and evaluate its performance.
* Validation: Cross-validation techniques to ensure the model’s generalizability.

**Expected Outcomes**

* Insights into the key factors influencing prediabetes and Type 2 diabetes.
* Insights into variables that correlate or contribute to diabetes, enabling better planning and lifestyle habits to reduce risk.

**References**

Kaggle Dataset Website: [Diabetes in Youth Vs Adult in India](https://www.kaggle.com/datasets/ankushpanday1/diabetes-in-youth-vs-adult-in-india/data)

GitHub Dataset:

Note: Other datasets may be used in the analysis to help shape the project and support objectives.