1. **Introduction:**

Diabetes is a metabolic disorder characterized by high blood sugar and insulin resistance. Globally, almost half a Billion people have Diabetes - 90% of which are adult onset (type two). While hereditary (type one) and gestational (type 3) diabetes also exist, the majority of negative health consequences come from type two diabetes. The symptoms include frequent urination, increased appetite, and increased level of thirst, and if left untreated or undertreated, a diabetic could experience severe heart disease, nerve damage, kidney failure, and many other catastrophic health complications. A diabetic will typically be prescribed medication that lowers blood sugar, and uses insulin for acute glucose spikes. Insulin is a hormone produced in the pancreas which is responsible for converting glucose to energy, and when a diabetic’s body does not respond in a typical way to insulin, exrra-corporeal insulin must be introduced. While some people are genetically predisposed to diabetes, a healthy diet and regular exercise can mitigate the chances of developing type two diabetes. Most health care practitioners monitor patients’ health during yearly physicals to ensure that the incipient warning signs of diabetes are not present, and often prescribe medications to pre-diabetics. In many cases, however, the onset of diabetes is too fast or far along for these interventions to be medically significant.

1. **Objective:**

In our project, we will attempt to correlate warning signs of diabetes with its development, and prescribe a quantitative threshold with which to initiate medical interventions. Additionally, we will create regimens for these medical interventions and suggest quantitative ways that patients can lower their risk to below a threshold that we will determine.

We will use the dataset in order to make these quantitative suggestions. Overall, we plan to reduce the rate of diabetes in adults with these monitoring and corrective procedures.

1. **Dataset:**

We obtained several datasets that may be used for predicting the development of diabetes. The first is a dataset from Kaggle, originally acquired from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains eight diagnostic measurements that will serve as predictor variables and one binary column that indicates whether a particular patient developed diabetes. There are 768 patients total – all of which are females at least 21 years-old from Pima Indian heritage. A table of variables included in the data can be seen in Appendix A. While the data is relatively clean, there are several features with missing values that need imputation. Additionally, there is a slight class imbalance in the data that will make classification more difficult for machine learning algorithms.

Link to data: <https://www.kaggle.com/datasets/whenamancodes/predict-diabities>

The second dataset, also acquired from Kaggle, contains 253,680 survey responses to The Behavioral Risk Factor Surveillance System (BRFSS) – a telephone survey conducted annually by the CDC. The survey aims to collect information from Americans on chronic health conditions, health-related risk behaviors, and preventative services. A table with information on variables can be found in Appendix B. Although this dataset does not have any missing values, it has a more extreme class imbalance problem.

Link to data: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download>

In the end, there are several strengths and weaknesses of both datasets. The first dataset contains more quantitative measurements of an individual’s health, but it is much smaller and not generalizable to the population at large. The second dataset is much larger and more generalizable, but it contains many variables that are subjective and self-reported such as an individual’s perception of their overall health. The following link provides an initial exploratory data analysis of both datasets:

<https://drive.google.com/file/d/1EHriHygQ3UIb8YGD4c2p-pzmUhjiOIgR/view?usp=sharing>

1. **Approach/Deliverables:** 
   1. **Predictive Algorithms: Logistic Regression, CART, XGBoost**

We will use the above mentioned data and build an algorithm which predicts the probability of a person developing diabetes. Using the diagnostic measurements listed in the data, we will implement three machine learning algorithms: Logistic Regression, CART (Decision Tree), and XGBoost (Gradient Boosted Decision Tree). After building each model, we will identify the variables which are most influencing the outcome. Although XGBoost is likely the most accurate model, it is also the least interpretable. As a result, the final model will need to find a balance between accuracy and interpretability.

* 1. **Prescriptive Algorithm: Minimizing Lifestyle Changes in Diabetes Prevention**

Each patient in the dataset can make certain lifestyle changes to lower their chance of developing diabetes as predicted by the machine learning algorithm. For example, a patient could start eating veggies, increase their physical activity level, or lower their cholesterol level. Some of these treatments are much less costly to the individual’s lifestyle than others. As a result, the goal of the prescriptive portion of the project is to create a tailored diabetes prevention plan that minimizes each patient’s lifestyle cost while reducing their probability of developing diabetes to a safe threshold.

We can create this optimization formula by assigning unique costs to each preventative treatment. Then, we will optimize over the combinations of treatments that will reduce the probability of diabetes to a safe threshold. The final optimization model will give a unique set of interventions for each patient that are not only preventive but also the least costly to implement. This type of tailored healthcare would be extremely valuable in preventing diabetes in the real world.

1. **Challenges:** 
   1. **Missing details in the data:** Some of the diagnostic measurements/features have missing values in the data, hence we will be imputing the missing values with the most accurate inference from the rest of the available observations.
   2. **Class imbalance in the data**: Many machine learning algorithms are designed to maximize overall accuracy by default and do not handle imbalance problems well. Potential solutions to this problem include balanced metrics such as AUC and F1 Score as well as resampling techniques such as minority class oversampling, majority class undersampling, and SMOTE.
   3. **Interpretability of algorithm:** The more interpretable the models are, the easier it is for someone to comprehend and trust the model. Models such as deep neural networks and gradient boosted trees are often referred to as black-box models because they are extremely difficult to interpret. Although some techniques such as Shapley plots help us understand deeper models, we would prefer using decision trees and linear regression models that are easy to understand.

**Appendix A - National Institute of Diabetes and Digestive and Kidney Diseases Dataset Description**

|  | **Data Field** | **Data Type** | **Description** |
| --- | --- | --- | --- |
| 1 | Pregnancies | Numeric | Number of pregnancies |
| 2 | Glucose | Numeric | Glucose level in blood |
| 3 | BloodPressure | Numeric | Blood pressure measurement |
| 4 | SkinThickness | Numeric | Thickness of skin |
| 5 | Insulin | Numeric | Insulin level in blood |
| 6 | BMI | Numeric | Body Mass Index |
| 7 | DiabetesPedigreeFunction | Numeric | Likelihood of diabetes based on family history |
| 8 | Age | Numeric | Age |
| 9 | Outcome | Binary | 1 = diabetes, 0 = no diabetes |

**Appendix B - The Behavioral Risk Factor Surveillance System Dataset Description**

|  | **Data Field** | **Data Type** | **Description** (no = 0, yes = 1) |
| --- | --- | --- | --- |
| 1 | Diabetes\_012 | Categorical | 0 = no diabetes, 1 = prediabetes, 2 = diabetes |
| 2 | HighBP | Binary | High blood pressure |
| 3 | HighChol | Binary | High cholesterol |
| 4 | CholCheck | Binary | Cholesterol check in the last 5 years |
| 5 | BMI | Numeric | Body Mass Index |
| 6 | Smoker | Binary | Smoked at least 100 cigarettes in your entire life |
| 7 | Stroke | Binary | Have had a stroke |
| 8 | HeartDiseaseorAttack | Binary | Coronary heart disease (CHD) or myocardial infarction (MI) |
| 9 | PhysActivity | Binary | Physical activity in past 30 days |
| 10 | Fruits | Binary | Consume fruit 1 or more times per day |
| 11 | Veggies | Binary | Consume veggies 1 or more times per day |
| 12 | HyvAlcoholConsump | Binary | Men (women) having more than 14 (7) drinks per week |
| 13 | AnyHealthcare | Binary | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. |
| 14 | NoDocbcCost | Binary | Time in past 12 months when needed to see a doctor but could not because of cost |
| 15 | GenHlth | Categorical | In general, your health is \_\_ on a scale of 1-5 |
| 16 | MentHlth | Numeric | For how many days in the last 30 \_\_\_ |
| 17 | PhysHlth | Numeric | For how many days during the last 30 \_\_\_ |
| 18 | DiffWalk | Binary | Serious difficulty walking or climbing stairs |
| 19 | Sex | Binary | 0 = female, 1 = male |
| 20 | Age | Categorical | 13-level age category; 1 = 18-24, etc. |
| 21 | Education | Categorical | Education level scale 1-6 |
| 22 | Income | Categorical | Income scale 1-8 |