Abstract

As of 2022, over 11% of the US population has been diagnosed with diabetes and around 90-95% of those people have type 2 diabetes. The client, Apple, would like to work in junction with the CDC to create a public health campaign to help customers and the general public be more knowledgeable about other potential indicators of diabetes. To accomplish this we need to see if indicators other than blood sugar levels are viable in determining if someone has diabetes. Then from a list of potential factors, we need to see what features were most important in contributing to the likelihood of someone having diabetes. A classification model will be built that would be able to serve both of these purposes.

Design

To see if there were key indicators other than blood sugar levels that could predict the likelihood of someone having diabetes a classification model needed to be built. Data was first obtained by looking online for any datasets that use indicators other than blood sugar levels as a way to predict the likelihood of someone having diabetes. After the data was obtained EDA was conducted on the data to see the distribution of the data and select features. In the classification models that will be built, they need to be able to provide insight into a multitude of possible indicators as well as accurately predict the presence of diabetes with the given data. Various classification models such as Random Forest, Gradient Boosting, XGBoost, Naive Bayes, etc... will be built and the best performing one will be chosen.

Data

Data gathered was a diabetes health indicators dataset obtained from Kaggle (https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset). This dataset has over 70 thousand observations, each row indicating a unique person. It contains various features

such as the status of cholesterol and blood pressure, demographic information, BMI, and lifestyle choices. I hope to be able to determine which key features are key indicators of type 2 diabetes with blood pressure status and physical activity being areas of high interest.

Algorithms

- Exploratory data analysis was done on the raw data to view basic relationships between
- Classification models were built on this data

Tools

- Pandas, Numpy for EDA and data manipulation
- Sklearn for model building and scoring
- XGBoost for creating XGBoost Classifier

Scroll down for communication

Communication