Diabetes Prediction Using CDC Survey

Biomedical Categorization Team: Donna Chen, Gary Hu, Michelle Lobb, Shayne Plourde
GitHub: https://github.com/chendonna3/diabetes-categorization | **Tableau**

Overview:
Diabetes is a major issue in the world, impacting 8.5% of adults and killing 1.5 million people in 2019 according to the World Health Organization. Diabetes is a chronic disease that affects how the body regulates blood glucose levels. Raised blood glucose levels may lead to serious damage and further complications. The goal of this project is to better understand the relationship between lifestyle factors and diabetes. Further, we aim to predict whether an individual has diabetes based on a survey questionnaire.

CDC Diabetes Health Indicators Dataset:
The CDC annually collects a health-related telephone survey which now includes responses from over 400,000 Americans. This Dataset contains 22 features spanning demographic information, lab test results, and lifestyle survey information, in addition to their diagnosis of diabetes.

Stakeholders: Clinicians, insurance companies, biomedical researchers, patients, health organizations
KPIs: AUC-ROC score for classifying patients with diabetes.
Approach: We used Logistic Regression, K-Nearest Neighbors, Linear Support Vector Machine, Random Forest, and XGBoost

Results:
The non-boosting methods for classification yielded accuracy scores capped at around 0.75, with the random forest model performing the best with 0.756 accuracy, 0.756 AUC-ROC score, and 0.796 recall on the imbalanced binary data set. However, the XGBoost model outperformed the non-boosting methods, with an AUC-ROC score of 0.821. The random forest model also provided relative importance scores for the features; the features with the highest importance scores were general health (0.2153), high blood pressure (0.2038), BMI (0.1279), age (0.1014), and high cholesterol (0.0937).

Model	**AUC-ROC**
Logistic Regression	0.740
K-Nearest Neighbors	0.742
Linear Support Vector Machine	0.751
Random Forest	0.756
XGBoost	**0.821**

Future directions:
We will explore whether factors that influence diabetes risk in the US are similar to those of data from lower-middle-income countries. Another future direction is to conduct a minimal survey to predict the likelihood of diabetes and recommend further treatment or intervention. Subsequently, an app can be developed to track an individual's risk of diabetes based on lifestyle factors.