March 6, 2025
Berkin, Sarah

Capstone Project Report – Diabetes Prediction Modeling

1. Problem Statement

Diabetes is a chronic condition characterized by elevated blood glucose levels due to the body's inability to produce or effectively use insulin. If left untreated, it can lead to severe health complications, including nerve damage, cardiovascular diseases, and vision problems. Early detection is crucial for effective management and improved quality of life. The objective of this project is to develop a predictive model that can accurately identify individuals at risk of diabetes, thereby facilitating early intervention.

2. Data Description

The dataset utilized originates from the National Institute of Diabetes and Digestive and Kidney Diseases. It comprises various medical predictor variables and one target variable: the presence or absence of diabetes. The predictor variables include:

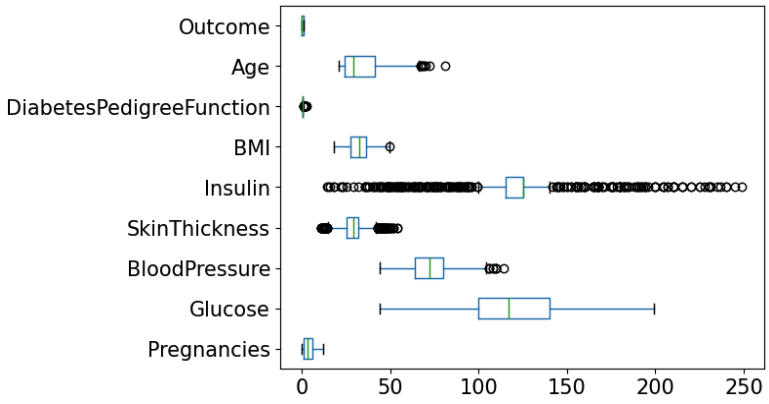
- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skinfold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction**: Diabetes pedigree function
- **Age**: Age in years

3. Methodology

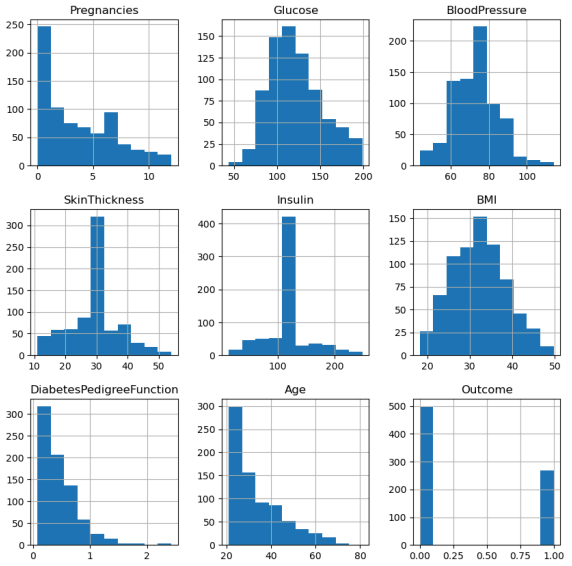
The analysis followed these key steps:

- **Data Cleaning**: Addressed missing or zero values in critical columns like Glucose, BloodPressure, SkinThickness, Insulin, and BMI by replacing them with the median values.
    - **Problem 1:** Inconsistencies in completeness of data entry.
    - **Solution 1:** Null values were filled in with the mean or median, as appropriate for the various independent variables.
    - **Problem 2:** Outliers were present throughout the independent variable values.

- **Solution 2:** Outliers were visualized using boxplots, then severe outliers were trimmed from the data.



- **Exploratory Data Analysis (EDA)**: Conducted univariate and bivariate analyses to understand the distribution of variables and their relationships with the target variable.
  - Histograms were created to visualize the distribution of the variables after outliers were trimmed.
  - A heatmap was generated to visualize correlations between the variables.



  - The 3 independent variables which had the lowest p-value from the Pearson Correlation Coefficient test were:
    - 1) Glucose: RR Stat= 0.51, P-value= 1.77
    - 2) Insulin: RR Stat= 0.26, P-value= 1.4
    - 3) BMI: RR Stat= 0.33, P-value= 1.59

- **Pre-Processing**:
    - **Problem:** Out of the 768 people in the original sample, only 268 of them have diabetes. If the expressions of the dependent variable are not balanced in the dataset, it would be tough to get any model to accurately predict the outcome because the baseline would already be leaning towards a negative outcome.
    - **Solution:** I created dummy variables in order to balance the dataset.
- **Modeling**: I used scikit-learn for training my recommendation system. Based on the fact that my dependent variable, whether or not a person has diabetes, is a binary Yes/No, I determined that the 3 algorithms which made the most sense are 1) Logistic Regression, 2) Support Vector Machine (SVM), & 3) Gradient Boosting Classifier.
    - I first ran each algorithm by performing a train test split on the data to create training and testing clusters of the data.
    - I next ran the same algorithms with K Fold cross validation to create clusters instead of train test split.

4. Findings

- When using Train, Test, Split– SVM has the best Accuracy score & Gradient Boosting Classifier comes in second.
- When using Train, Test, Split– Gradient Boosting Classifier has the best ROC-AUC scores training scores, while SVM has the better ROC-AUC scores test scores.
- When using Train Test Split, there does not appear to be any scenario that Logistic Regression is the best choice.
- When using K Fold Cross-Validation– Gradient Boosting Classifier has the best Accuracy & Logistic Regression comes in second.
- When using K Fold Cross-Validation– Gradient Boosting Classifier has the best ROC-AUC scores, while Logistic Regression comes in second.
- **Winning Model: Gradient Boosting Classifier with K Fold Cross-Validation**
    - For the sake of erring on the more conservative side, I will choose the K Fold CV. K Fold CV helps guard against overfitting and makes sure the entire dataset is used for training and testing while validating unseen data.
    - Gradient Boosting Classifier model performed the best in both Accuracy & ROC-AUC CV.

5. Recommended Business Use Cases

The predictive model developed in this project can support various business objectives:

1. **Targeted Marketing Campaigns:**
   Businesses can use the model to identify individuals at higher risk of diabetes and create

tailored health and wellness campaigns. For example, companies offering health-related products or services could promote specific offerings, such as nutritious meal plans, fitness programs, or medical devices, to this audience.

2. **Product Development & Innovation:**
   Companies in the healthcare and wellness industries could leverage the insights to develop new products that meet the needs of at-risk populations. This could include creating low-sugar food products, fitness tracking tools, or preventive health services, thereby opening new revenue streams.

3. **Enhanced Customer Segmentation:**
   Insurance companies could integrate the model into their risk assessment processes to better segment customers. This could lead to more accurate pricing of health insurance policies and the development of personalized health management programs that reduce overall claims costs.

6. Future Work

To enhance the predictive accuracy and applicability of the model, consider the following:

- **Incorporate Additional Data**: Integrate more comprehensive datasets that include variables like family medical history, physical activity levels, and dietary habits.

- **Address Class Imbalance**: Apply techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset and improve model performance on minority classes.

- **Feature Selection**: Utilize advanced feature selection methods to identify and retain the most impactful variables, potentially enhancing model efficiency and accuracy.