Predicting Diabetes to Facilitate Early Intervention

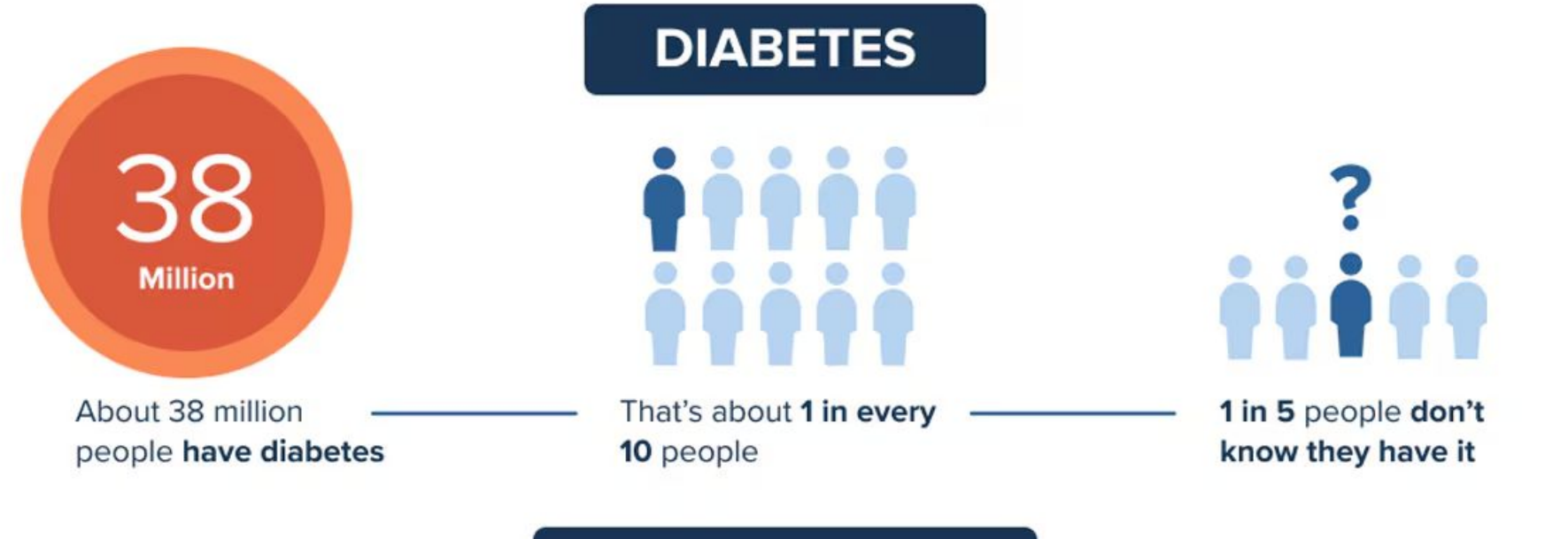
Sarah Berkin

Data Science Capstone Project, Sept 2024 Cohort

The Problem

If left untreated, diabetes leads to severe health complications like nerve damage, cardiovascular diseases, & vision problems

Early detection is crucial for effective management and improved quality of life



Understanding the Problem

Factors that may cause Diabetes

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- Body mass index (BMI)
- Family History (Diabetes Pedigree Function)
- Age

Data Information for this Analysis

Data Source

The dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases

Sample Population

768 women above the age of 21 were sampled in May 1990 near Phoenix, AZ, USA

Variables

Target variable:
Presence of Diabetes as a binary Yes/No

All other factors are independent variables to assess

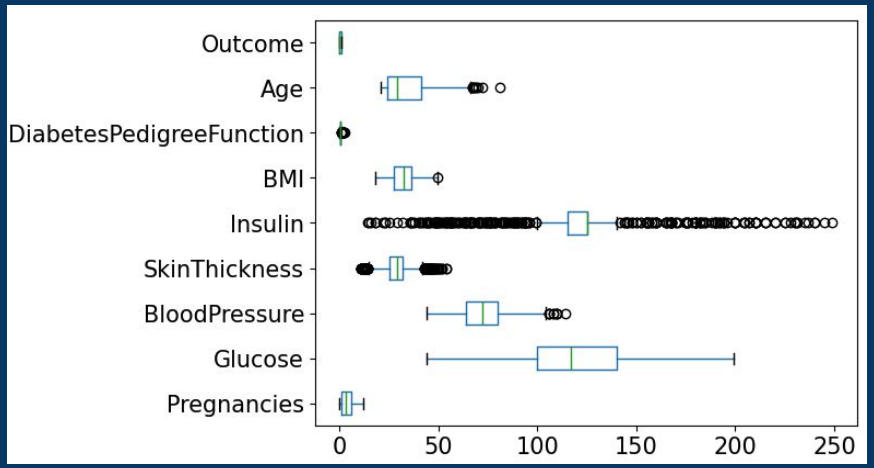
Data Cleaning

Completeness of Data Entry

Filled in data entry gaps with averages (mean or median, as appropriate)

Reduced Extreme Outliers

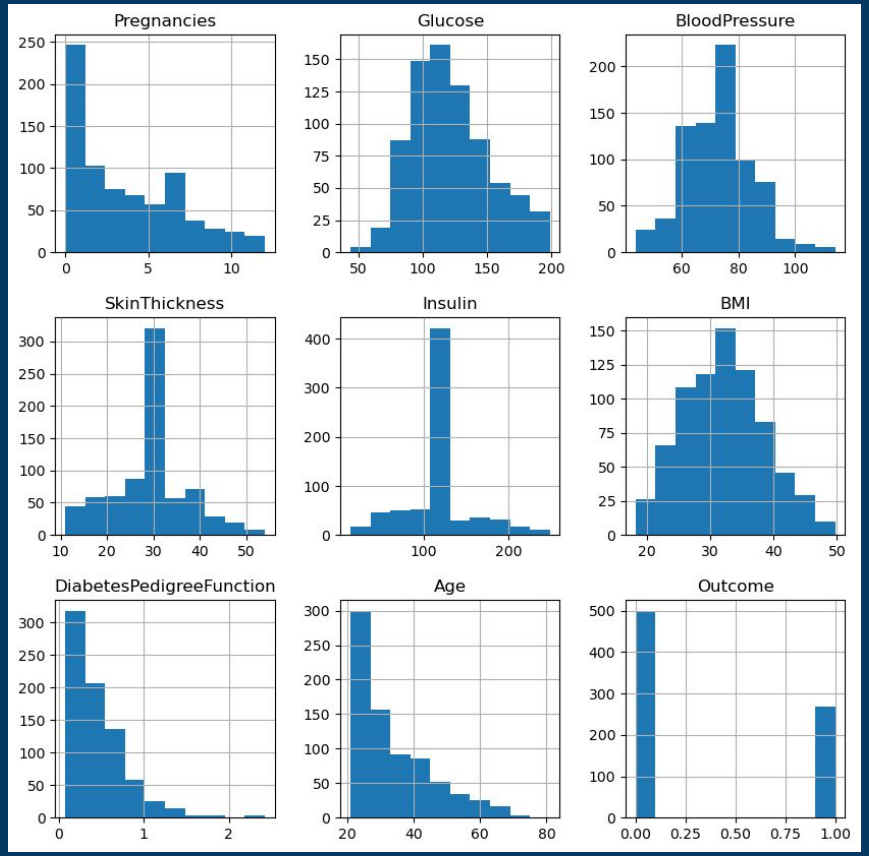
Removed the most extreme outliers that skewed the impact seen from each variable



Exploratory Data Analysis

Visualizing the Distribution of the Variables

Histograms provide a bird's eye view of the values in each variable



Exploratory Data Analysis

Visualizing Correlations between the Variables

Heatmaps provide a way to see how deeply correlated (or not) each variable is with another

Exploratory Data Analysis

Which Variables are most correlated to the Outcome?

Pearson Correlation Coefficient tests showed that 1) Glucose, 2) Insulin, and 3) BMI are the variables most connected to whether a person gets diabetes.

Glucose P-value

```
corr_glucose = pearsonr(list_gluc, list_outcome)
print(corr_glucose)
```

```
PearsonRResult(statistic=0.23416325004247598, pvalue=0.002116090405336514)
```

Insulin P-value

```
corr_insulin = pearsonr(list_insu, list_outcome)
print(corr_insulin)
```

```
PearsonRResult(statistic=0.2042761957857276, pvalue=0.007539590419075425)
```

BMI P-value

```
corr_bmi = pearsonr(list_bmi, list_outcome)
print(corr_bmi)
```

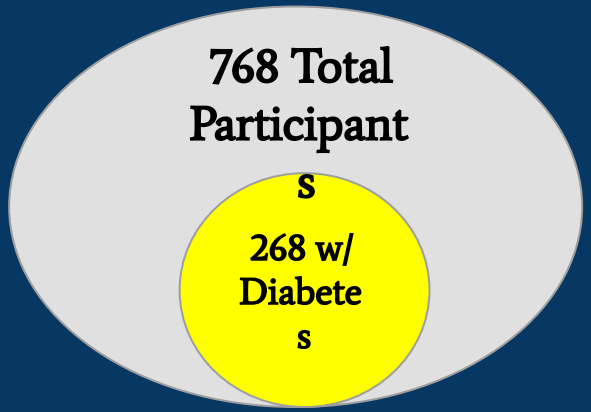
```
PearsonRResult(statistic=0.1484872188601701, pvalue=0.05330039210472585)
```

Modeling Steps

- Data Pre-Processing
- Cross-Validation for Hyper-Parameter Tuning
- Classifier Training

Data Pre-Processing

Balancing the Data Set



The data set was balanced by resampling with dummy variables.

Data Splitting

Train Test Split method

80% Train / 20% Test

K-Fold Cross-Validation

10 Folds

Classifier Training: Cross-Validation (CV)

Logistic Regression	Support Vector Machine (SVM)	Gradient Boosting Classifier
Mean CV 84.5%	Mean CV 83.4%	Mean CV 90.1%
Standard Deviation +/- 6.4%	Standard Deviation +/- 7.3%	Standard Deviation +/- 4.6%

Classifier Training: Modeling Results

Logistic Regression	Support Vector Machine (SVM)	Gradient Boosting Classifier
Accuracy: 75.8%	Accuracy: 74.8%	Accuracy: 82.1%
ROC-AUC CV Mean: 84.5% Std Dev: +/- 3.2%	ROC-AUC CV Mean: 83.4% Std Dev: +/- 3.7%	ROC-AUC CV Mean: 90.1% Std Dev: +/- 2.2%

Winning Model

Gradient Boosting Classifier

With K-Fold Cross Validation

Highest scores across the board

Mean CV: 90.1% / Std Dev +/- 4.6%

Accuracy: 82.1% /

ROC-AUC CV: Mean 90.1% / Std Dev +/- 2.2%

Recommended Business Use Cases

Targeted Marketing Campaigns	Health-related products & services
Product Development & Innovation	Low-sugar food products, fitness tracking tools, or preventive health services
Enhanced Customer Segmentation	Risk assessment processes for health insurance companies

Future Work

Incorporate Additional Data	Include variables like family medical history, physical activity levels, and dietary habits
Address Class Imbalance	SMOTE (Synthetic Minority Over-sampling Technique) to balance dataset & improve performance on minority
Feature Selection	Identify and retain the most impactful variables

The Team

Project by:

Sarah Berkin

Data Science Trainee

Special thanks to:

Jaleed Khan, PhD

Data Science Mentor
Sr. Researcher
Honorary Research Fellow
Lecturer