Early Detection of Diabetes in Women

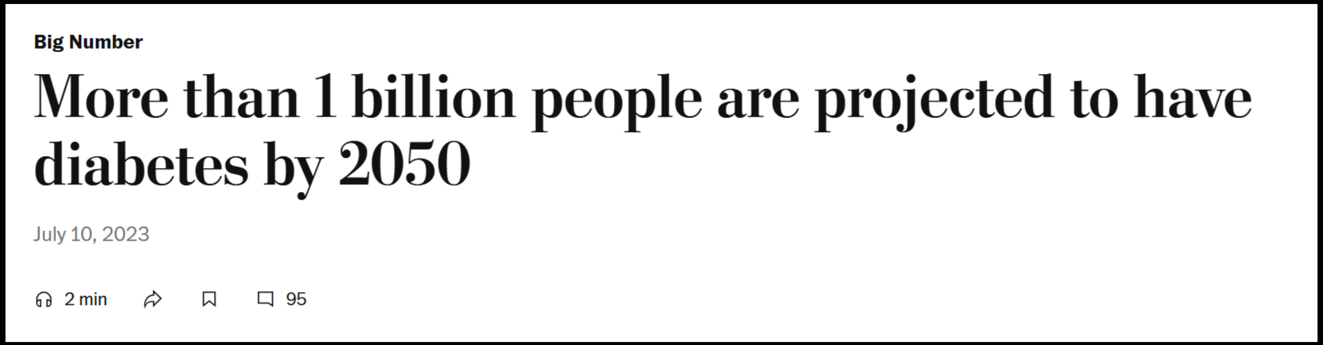
Ananya Soleti, Rishabh Nair, Sahil Gandhi

Background & Motivation

Diabetes is a chronic disease that develops when a person's blood sugar (glucose) level is too high for a prolonged amount of time

We chose this topic specifically because early detection of diabetes is crucial to:

- Prevent complications such as heart disease, kidney failure, and infections
- Reduce the burden on healthcare systems by decreasing the need for costly treatments and prolonged hospitalizations
- Identify high-risk individuals so hospitals can take appropriate actions



Big Number

More than 1 billion people are projected to have diabetes by 2050

July 10, 2023

2 min 95

Washington Post. July 10, 2023

"For people who have diabetes, however, the body may make little or no insulin (Type 1 diabetes) or may not make or use insulin properly (Type 2), leaving too much glucose in the bloodstream. Over time, this can lead to serious health problems, including heart disease, stroke, kidney disease, nerve damage and vision loss."

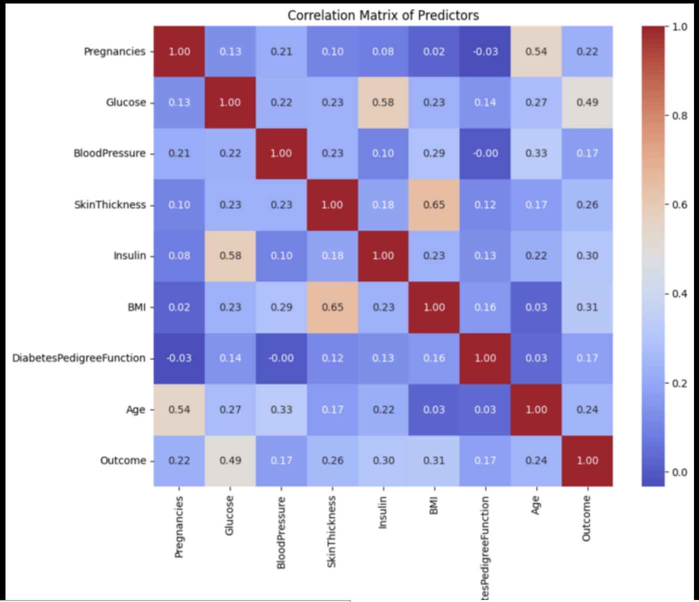
Our Dataset

- 768 observations of female diabetes patients aged 21 years and older
- Eight features
- Target variable: Outcome
  - Indicates whether diabetes is present in a patient

- Continuous:
  - Pregnancies
  - Glucose
  - Blood Pressure
  - Skin Thickness
  - Insulin
  - BMI
  - Diabetes Pedigree Function
  - Age
- Categorical:
  - Outcome

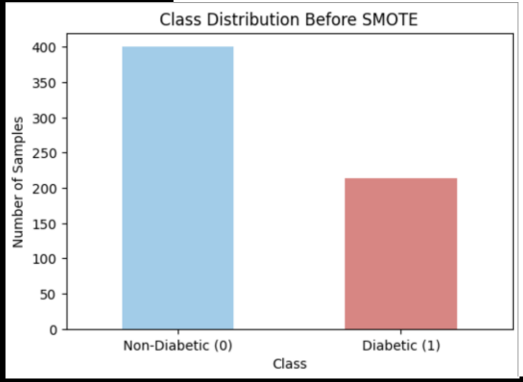
Data Preprocessing

- SMOTE generates synthetic samples by focusing on the nearest neighbors within the minority class. This ensures that the new samples are more representative of the actual data distribution and maintain class-specific feature patterns.
- Adjusting class weights modifies the model’s loss function to give more importance to the minority class. While this helps with imbalance, it doesn’t alter the data distribution itself, leaving the minority class underrepresented during training. Also in this case we have a correlation matrix of predictors as seen on the right

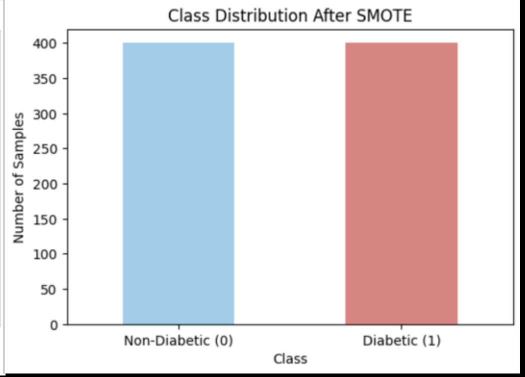


Correlation Matrix of Predictors

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	tesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.21	0.10	0.08	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.22	0.23	0.58	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1.00	0.23	0.10	0.29	-0.00	0.33	0.17
SkinThickness	0.10	0.23	0.23	1.00	0.18	0.65	0.12	0.17	0.26
Insulin	0.08	0.58	0.10	0.18	1.00	0.23	0.13	0.22	0.30
BMI	0.02	0.23	0.29	0.65	0.23	1.00	0.16	0.03	0.31
DiabetesPedigreeFunction	-0.03	0.14	-0.00	0.12	0.13	0.16	1.00	0.03	0.17
Age	0.54	0.27	0.33	0.17	0.22	0.03	0.03	1.00	0.24
Outcome	0.22	0.49	0.17	0.26	0.30	0.31	0.17	0.24	1.00



Class Distribution Before SMOTE

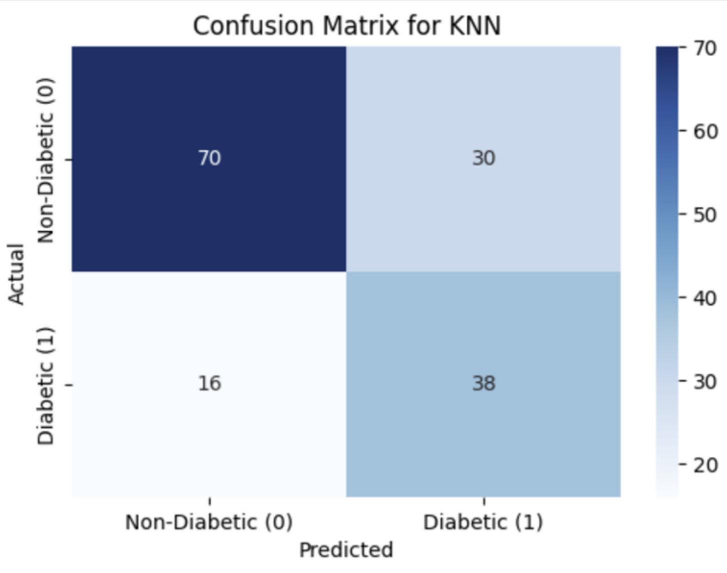


Class Distribution After SMOTE

Method 1 - K-Nearest Neighbors (KNN)

- A model that makes prediction based on nearest data points
- We used as our first method KNN because it is good at
  - o Handling non-linear data
  - o Working with multiple features where relationships aren't obvious (combing BMI, insulin levels, glucose levels)
  - o Working with medium sized datasets such as the one we have

Accuracy Achieved: 0.70

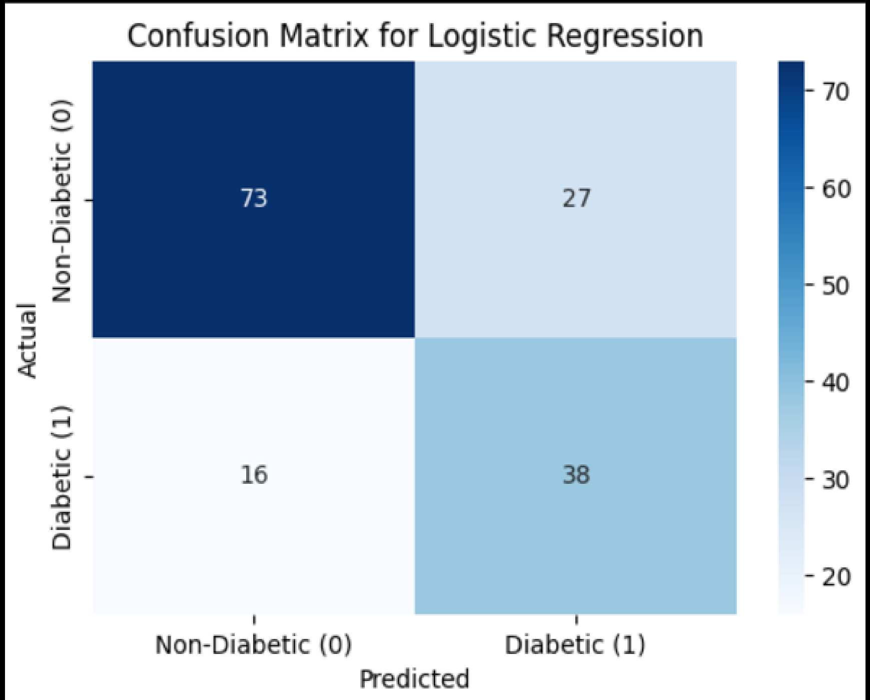


- Identified 70 people as non-diabetic
- incorrectly predicted 30 non-diabetic people as diabetic
- missed 16 actual diabetic cases
- identified 38 people as diabetic

Method 2 – Logistic Regression

- We picked Logistic Regression as our second method because it:
  - Provides clear coefficients for each feature, indicating the direction and magnitude of their influence on the target variable (outcome).
  - Designed for binary classification (diabetic vs non-diabetic)
  - Hyperparameters used:
  - C = 0.1 (Regularization) - Applied to prevent overfitting, 0.1 indicates strong regularization, prevents the model from relying too heavily on any single feature
  - Max_iter = 100 (Maximum Iterations) - finds optimal coefficients by continually minimizing loss function
  - Since we have a small-medium sized dataset, 100 is a good balance efficiency and convergence

ACCURACY ACHIEVED: 0.72



Actual \ Predicted	Non-Diabetic (0)	Diabetic (1)
Non-Diabetic (0)	73	27
Diabetic (1)	16	38

- Identified 73 people as non-diabetic
- incorrectly predicted 27 non-diabetic people as diabetic
- missed 16 actual diabetic cases
- identified 38 people as diabetic

Insights/Observations

- Healthcare Cost Reduction
  - Identify high risk patients through predictive analytics and implement preventative programs that can lower long-term expenses for hospitals
  - Clinics can optimize their resources for high-risk patients
- Custom Insurance Plans
  - Design and promote premium insurance plans for high-risk patients
- Product Development
  - Utilize important features (BMI, glucose) to cater and develop products for diabetes patients

THANK YOU

ANY QUESTIONS?