

Phase 5 Capstone

Diabetes Classification

Heath Rittler

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Business Problem

6.7%

Diabetes Prevalence

Description: According to the International Diabetes Federation, in 2021, an estimated **537 million** people worldwide had **diabetes**, and this number is projected to rise to 642 million by 2040.

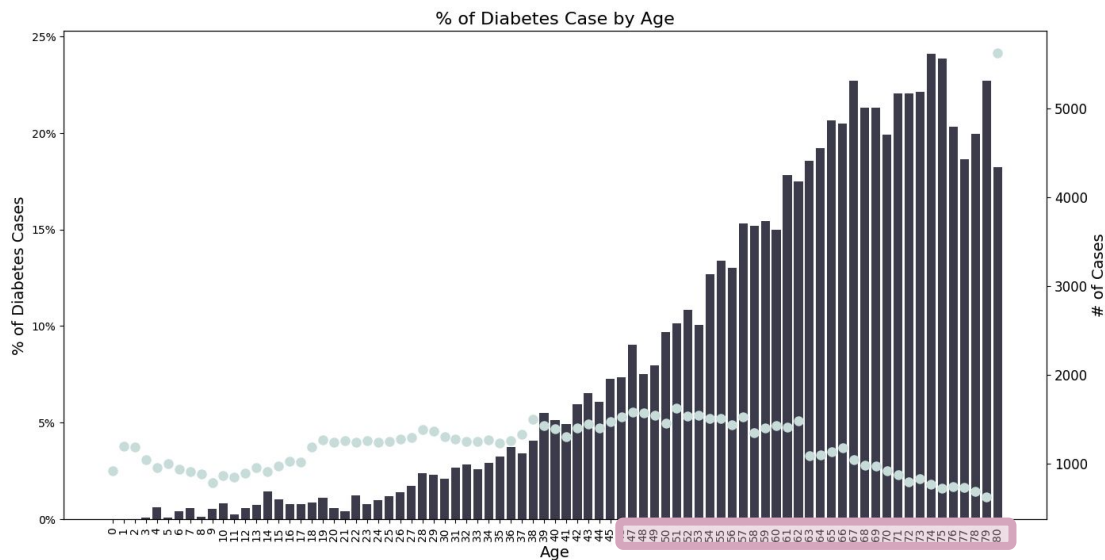
Goal: Create a classification system to help **identify individuals with Diabetes** within a given population.

Data

100k individuals

8 independent variables + our
diabetes flag

Roughly **8.5%** with Diabetes



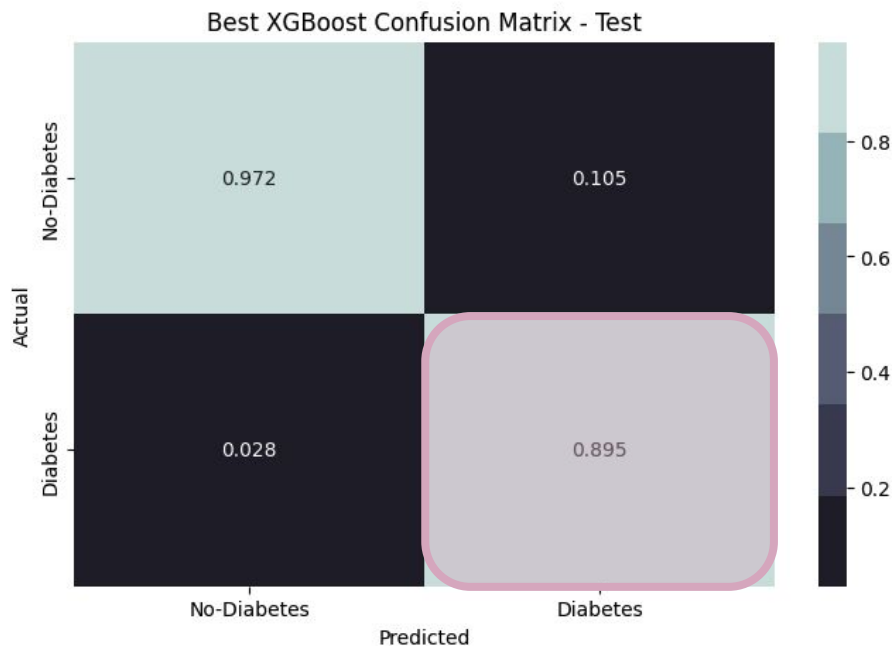
<https://www.kaggle.com/datasets/iammustafatz/diabetes-pre-diction-dataset>

Approach & Goals

Classification model to determine Diabetes diagnoses

Evaluate multiple methodologies, and pick the best one (Logistic Regression, Random Forest, Decision Tree, **XGBoost**, Neural network)

Correct predictions **89%** of the time



Final Model

Minimize false positives -
Precision

Decision Tree & Random
Forest had better precision
metrics but **sacrificed recall**

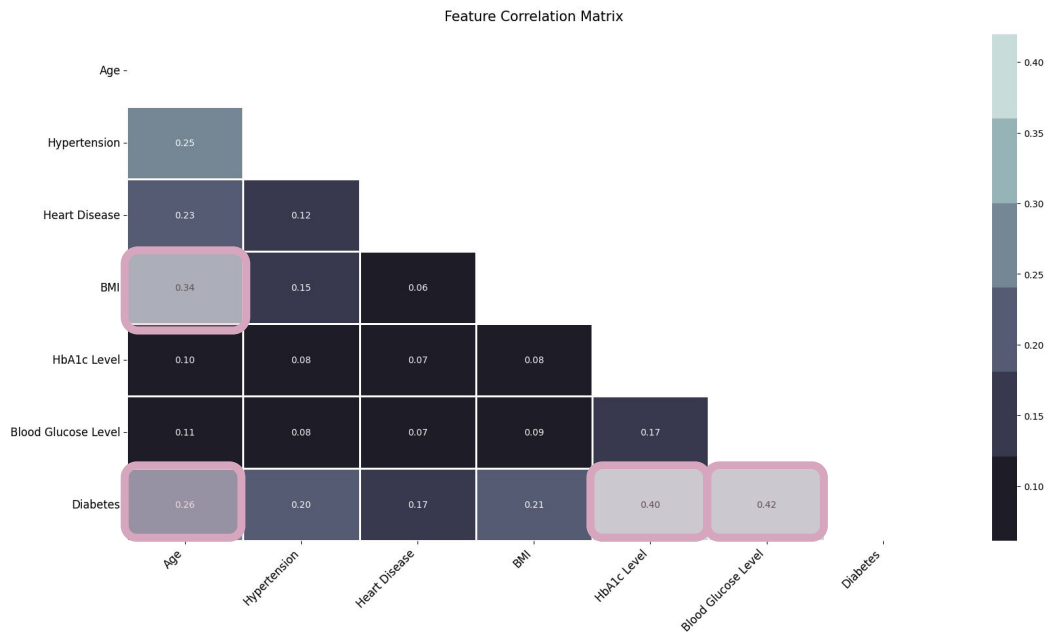
Best XGBoost chosen for
consistency & overall
performance

	Model	Train Precision Score	Test Precision Score	Train F1 Score	Test F1 Score
3	Best Decision Tree	1	1	0.801026	0.803713
5	Best Random Forest	1	1	0.801026	0.803713
7	Best XGBoost	0.884868	0.89547	0.784971	0.788142
6	Baseline XGBoost	0.93223	0.883058	0.836596	0.78499
4	Baseline Random Forest	0.994138	0.753302	0.994374	0.740895
2	Baseline Decision Tree	0.996498	0.699173	0.994441	0.715686
1	Best Logreg	0.430175	0.442249	0.558478	0.568822
0	Baseline Logreg	0.419432	0.432226	0.551109	0.561701

Insights

HbA1c and blood glucose highly **correlated** with Diabetes

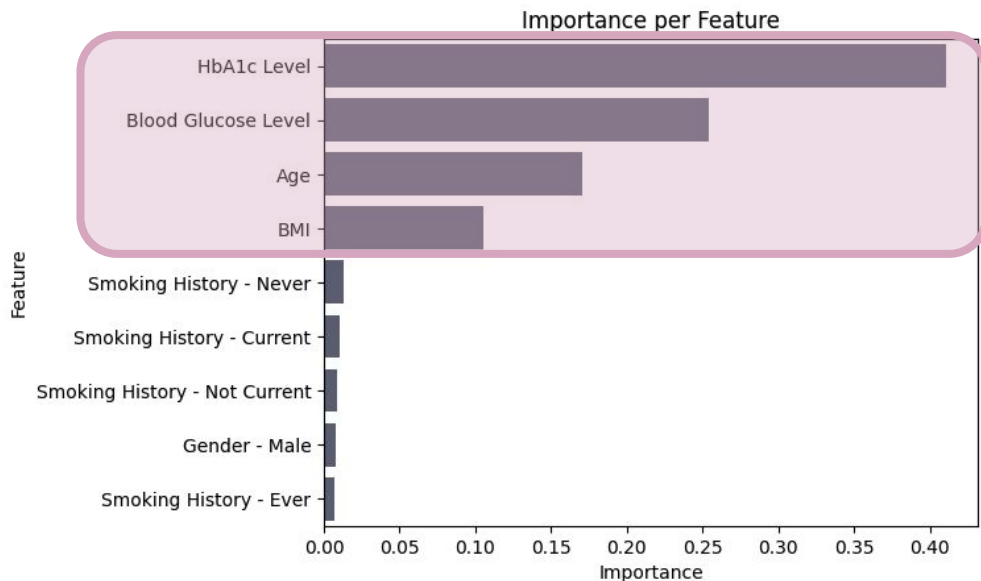
Age, BMI, and Smoking also risk factors for Diabetics



Feature Importance

Lifestyle factors for
intervention - HbA1c,
Blood Glucose, BMI,
Smoking

Similar impact to
correlation matrix



Recommendations

Run the algorithm on new data.

Evolve the dataset.

Understand time, and impact.

Evaluate interventions with A/B testing.

Load data into centralized repository.

Thank you!

Email: hrittler@gmail.com

Github: [@heathlikethecandybar](https://github.com/heathlikethecandybar)

LinkedIn: [linkedin.com/in/heathrittler](https://www.linkedin.com/in/heathrittler)