Diabetes Prediction

Stark Health Clinic

A Supervised Machine Learning Project by Ololade Folashade

Problem Statement

Diabetes is a chronic illness that has plagued the world with high global impact.

Stark health is currently struggling with early detection of diabetes, leading to high costs and low patient outcomes.

The objective of this project is there to build a predictive model to patients at risk of diabetes using historical health records.

Finally, the goal for the project is to empower stark health to enable early interventions and optimally allocate health care resources for improved patient outcome.

Overview of Dataset

Source: Publicly available anonymized healthcare dataset

Key Variable: Diabetes (1 = diabetic, 0 = non-diabetic)

Key Features:

- gender
- age
- hypertension
- heart_disease
- smoking_history
- bmi
- HbA1c_level
- blood_glucose_level

Approach

Tools Used: Python, Jupyter Notebook, Pandas, Seaborn, Scikit-Learn

Process Flow:

1. Data Cleaning & Handling Missing Values
2. Exploratory Data Analysis (EDA)
3. Feature Engineering (encoding + creation)
4. Model Selection & Training
5. Evaluation & Interpretation

Exploratory Data Analysis - 1

Univariate, Bivariate and Multivariate Analysis

Univariate Analysis:

Age, HbA1c, Blood Glucose Level, BMI: Skewed distributions.

Binary Fields (Hypertension, Heart Disease): Mostly zeros — imbalance noted.

Exploratory Data Analysis - 2

Bivariate Analysis:

Box Plots: Patients with diabetes have visibly higher HbA1c and blood glucose.

Crosstabs: Higher diabetes rates among those with hypertension or heart disease.

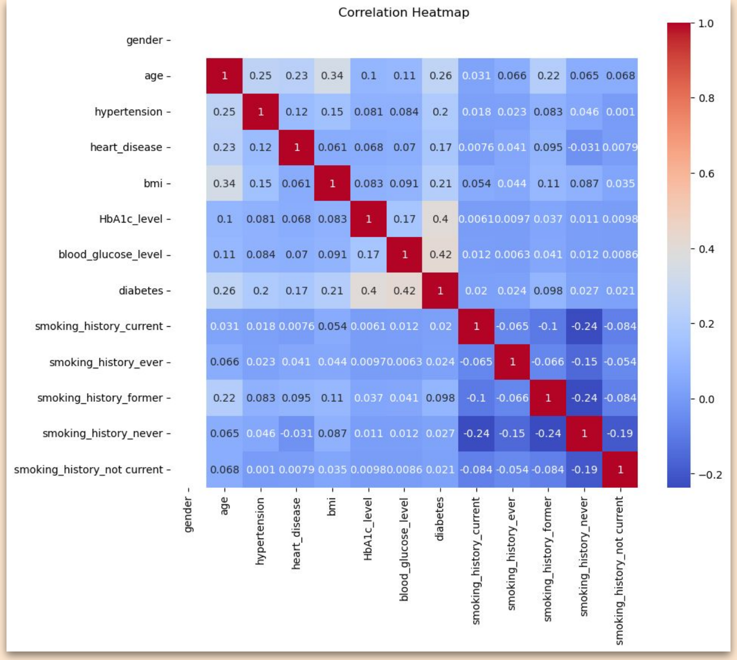
Exploratory Data Analysis - 3

Multivariate Analysis:

OLS Regression: Key predictors include blood glucose, HbA1c, age, and BMI.

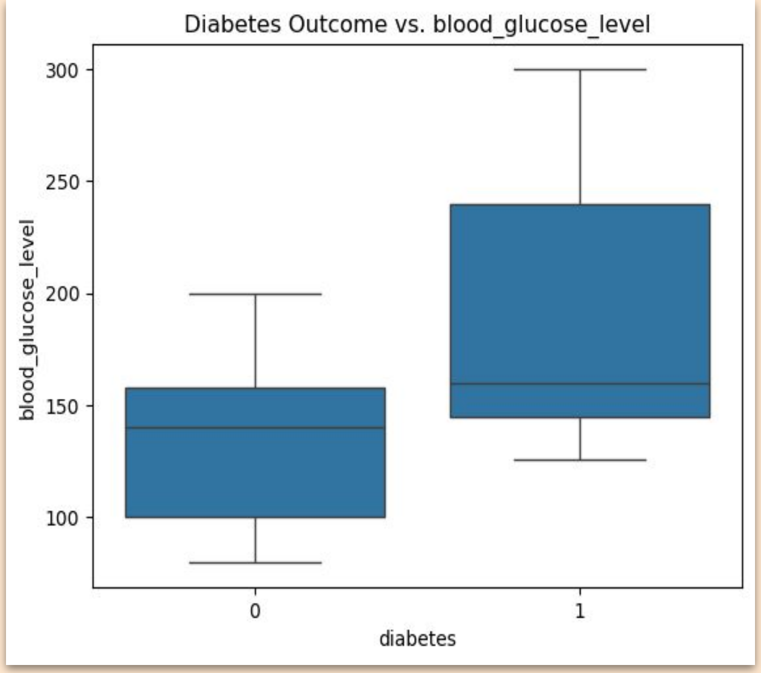
Multicollinearity managed by selecting variables with $p < 0.05$.

Visualizations - Correlations



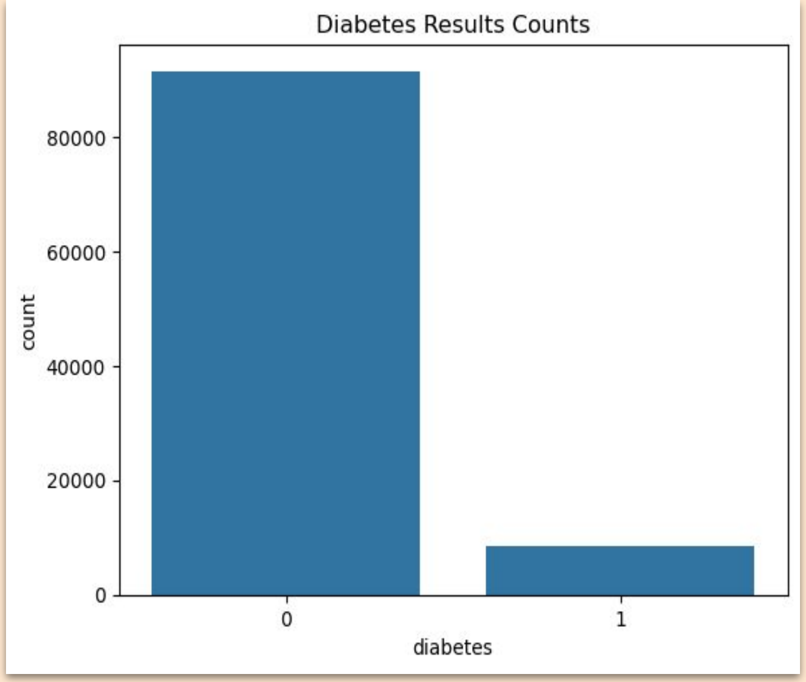
Correlation Heatmap – Shows strongest predictors. HbA1c and blood glucose level strongly correlate with diabetes.

Visualizations - Diabetes vs. Glucose



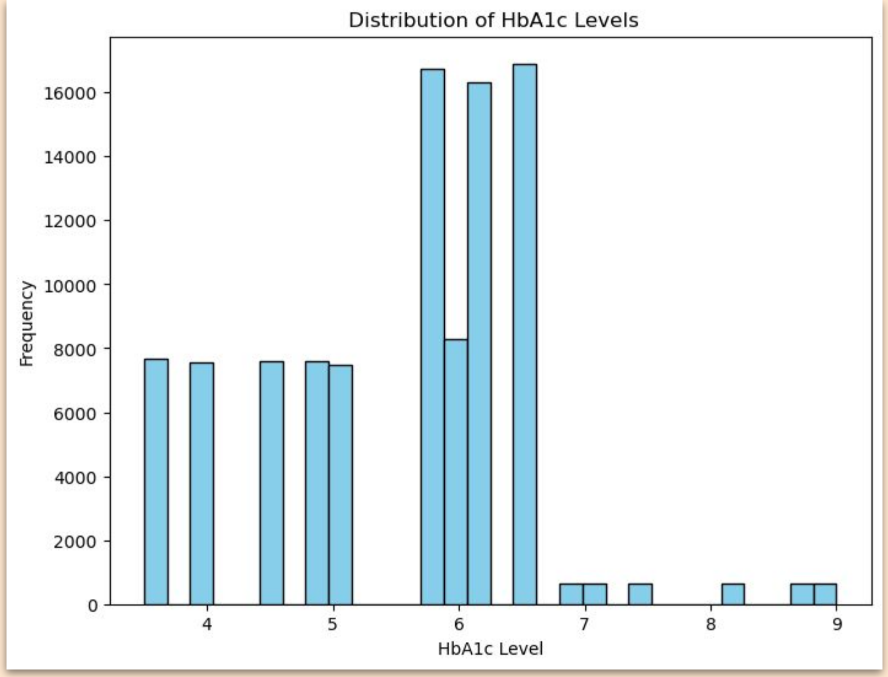
Boxplot of Blood Glucose vs Diabetes – Clear separation. Diabetic patients show higher glucose concentration.

Visualizations - Diabetes Outcome



Majority of patients in dataset are non-diabetic, highlighting data imbalance.

Visualizations



Higher HbA1c levels correlate with increased diabetes risk.

Model Building

Models Trained: Logistic Regression, Decision Tree, Random Forest

Train/Test Split: 80/20

Encoding: Label encoding for categorical features

Modeling Approach:

- Fit each model on training set
- Used accuracy, precision, recall, and F1-score to evaluate

Outcome

Model	Accuracy	Precision	Recall	F1-Score
Logistic Reg.	95%	96%	99%	98%
Decision Tree	97.21%	97%	100%	99%
Random Forest ✅	97.22%	97%%	100%	99%

Best Model: Random Forest

Strongest Predictors: HbA1c, Age, Hypertension

Business Insight: These factors should trigger proactive screening protocols.

Conclusion

- **Outcome:** Successfully built a diabetes prediction model with good performance & **Identified** key risk indicators using EDA and model explainability.

- **Limitations:** Dataset size was limited & **May** need more clinical features for better accuracy.

- **Business Value:** Supports early screening and patient segmentation & **Reduces** long-term treatment costs.

Next Steps

Short-Term:

Expand feature set (include cholesterol, glucose, family history)

Test model on Stark Health's real patient data

Long-Term:

Deploy as web-based prediction tool in clinics

Educate staff on interpreting model predictions

Continuously retrain model with live data