Machine Learning-Based Early Prediction of Diabetes Using the PIMADataset

Edward Tsatsu Akorlie (ID: 22424530)
Daniel Kpakpo Adotey (ID: 22424924)
Kwame Ofori-Gyau (ID: 22424324)
Francis Aboraa Sarbeng (ID: 22424635)
Caleb Abakah Mensah (ID: 22424188)

UNIVERSITY OF GHANA

Dataset Description

Source & Background
Sourced from Kaggle. The dataset comprises medical records of Pima Indian women, a population known for high diabetes prevalence.

768

Patients

8

Columns

Features

- Pregnancies
- Glucose (Plasma)
- Blood Pressure (mm Hg)
- Skin Thickness (mm)
- Insulin (mu U/ml)
- BMI (Weight/Height²)
- Diabetes Pedigree Func.
- Age (Years)

Population Demographics
•Female patients only
•Pima Indian heritage
•Aged 21 years and above

TARGET VARIABLE: OUTCOME

0 = No Diabetes
1 = Diabetes

Data Sufficiency Met
Dataset exceeds requirements: ≥ 500 rows (768) and ≥ 3 features (8).

UNIVERSITY OF GHANA

Key Research Questions

Results

- The dataset shows a moderate class imbalance, with non-diabetic cases forming the majority.
•Non-diabetic (Outcome = 0): 500 patients (65.1%)
•Diabetic (Outcome = 1): 268 patients (34.9%)
This indicates that roughly **1 in 3 patients** in the dataset has diabetes.

- There is a clear and substantial difference in average glucose levels between the two groups:
•Average Glucose (Non-diabetic): 109.98
•Average Glucose (Diabetic): 141.26

-BMI and glucose are positively related overall, but the relationship differs by outcome. Non-diabetic patients generally maintain lower glucose levels even as BMI increases, indicating that high BMI alone does not strongly raise glucose. In contrast, diabetic patients show consistently higher glucose levels across most BMI ranges, with the combination of high BMI and elevated glucose occurring much more frequently.

01

Class Distribution

Is there imbalance between diabetic vs non-diabetic patients?

02

Glucose Impact

How do glucose levels differ between diabetic and non-diabetic patients

03

Feature Correlations

Which clinical features show strongest correlation with diabetes?

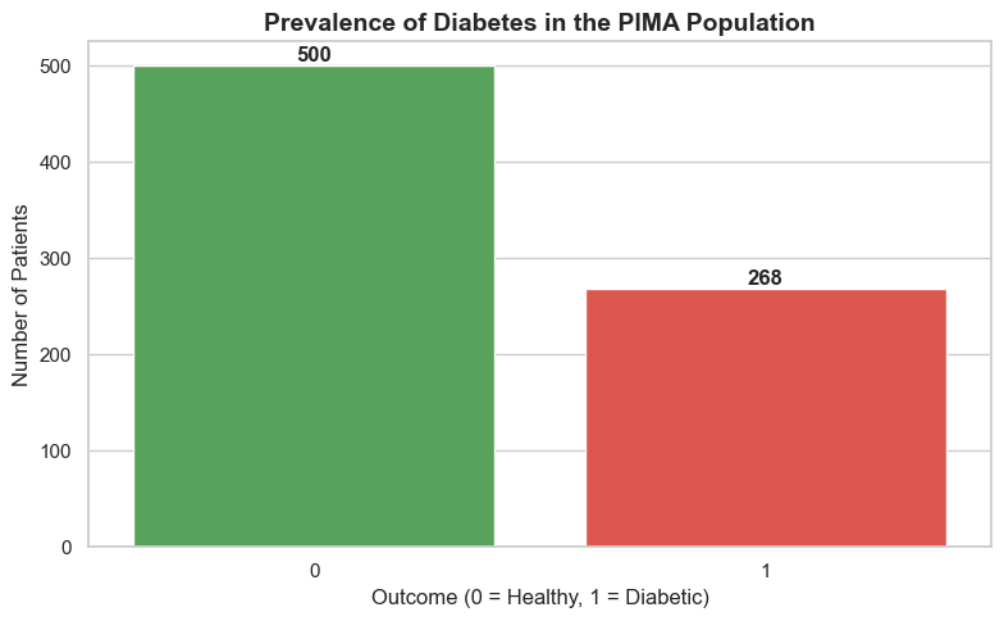
04

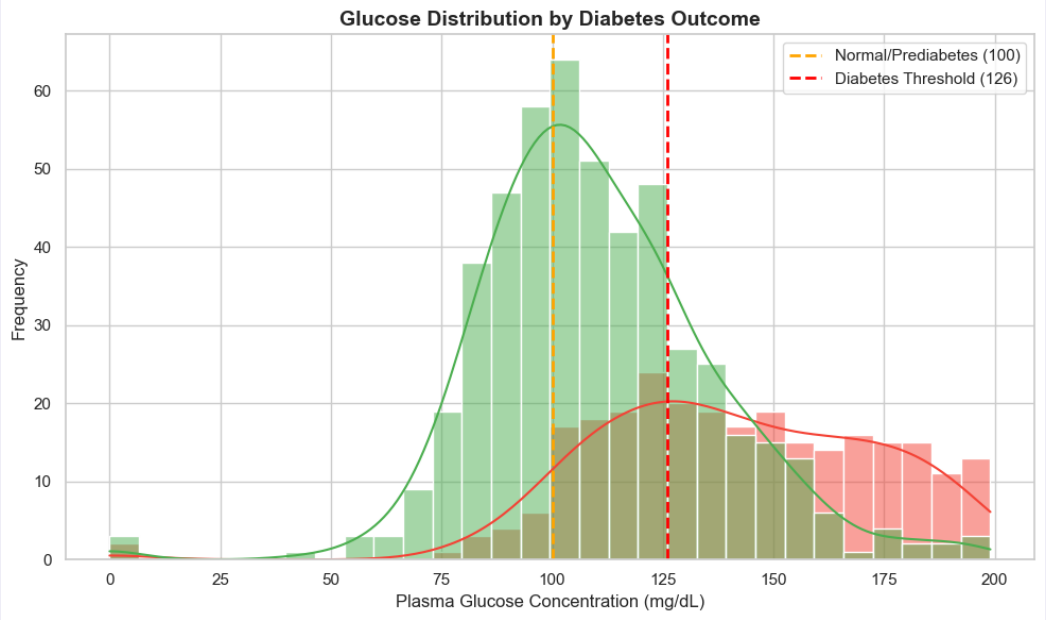
Risk Factors

Does age correlate with diabetes risk?

UNIVERSITY OF GHANA

Outcome Distribution & Glucose Analysis



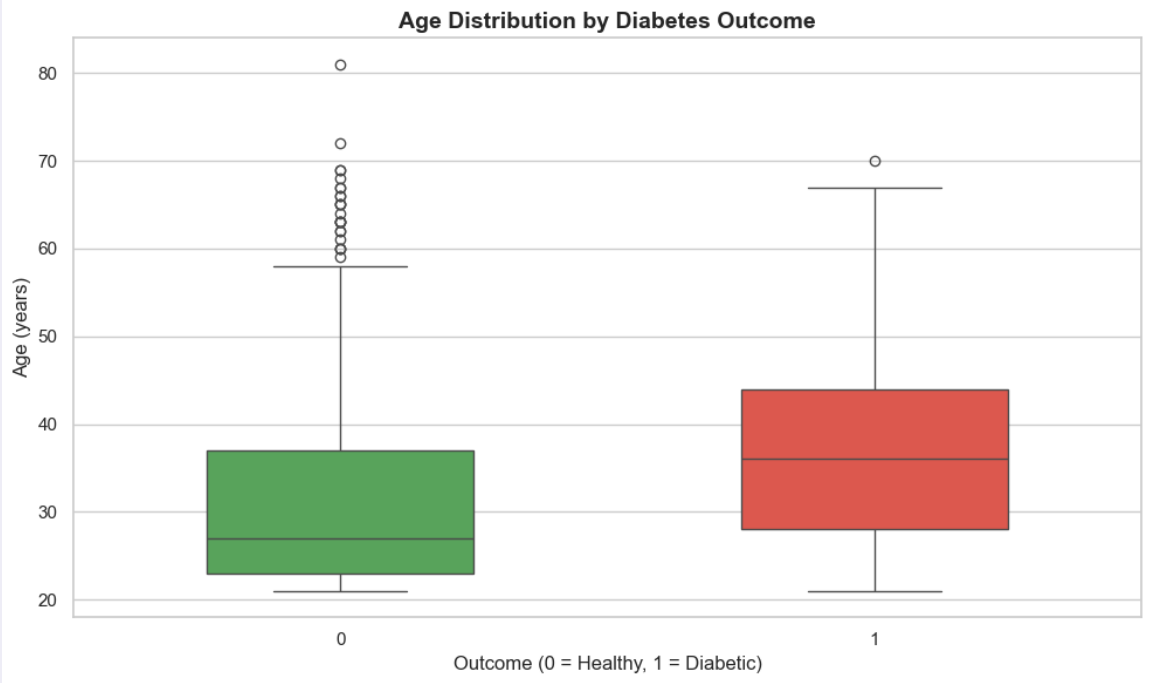


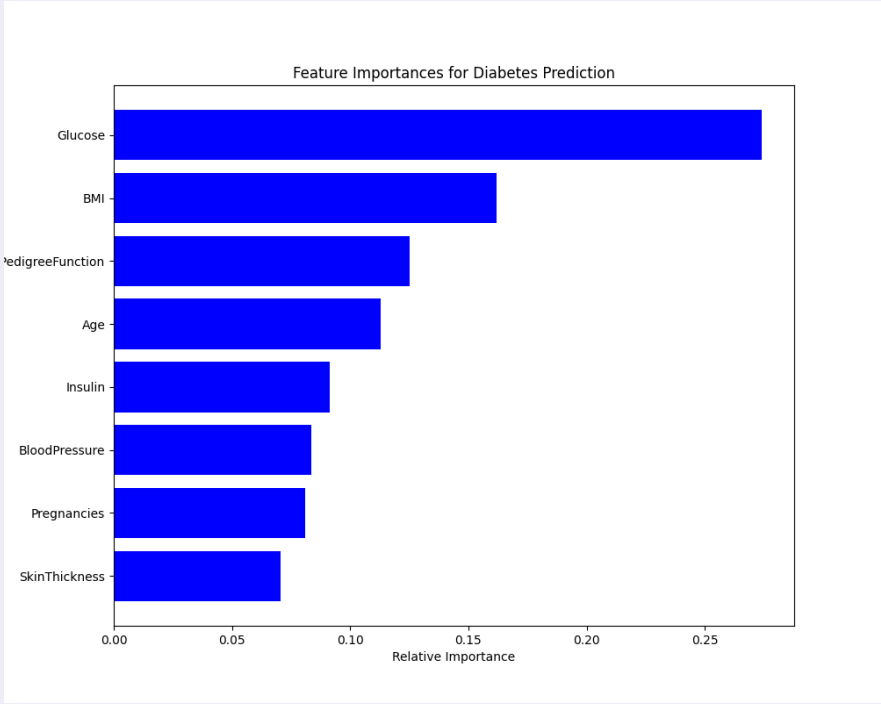
Class Imbalance: 500 non-diabetic vs 268 diabetic patients. Accuracy alone may mislead F1-score and recall are critical.

Clear Separation: Diabetic patients show significantly higher median glucose (~140 mg/dL vs ~100 mg/dL). Strong predictive power.

UNIVERSITY OF GHANA

Feature Correlations & Importance





Diabetic patients are typically older (median 37 years) than non-diabetic patients (median 27 years). While risk increases with age, diabetes also occurs in younger individuals, making age a contributing but not deterministic risk factor.

Glucose most important feature.

UNIVERSITY OF GHANA

Machine Learning Models & Evaluation

Models Experimented

Logistic Regression
Baseline probabilistic classifier

Decision Tree & Random Forest
Tree-based ensemble methods

Support Vector Machine (SVM)
Optimal hyperplane separation

K-Nearest Neighbors (KNN)
Distance-based classification

Key Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.779221	0.717391	0.611111	0.660000	0.817870
Logistic Regression	0.707792	0.600000	0.500000	0.545455	0.812963
SVM	0.740260	0.652174	0.555556	0.600000	0.796389
KNN	0.753247	0.660000	0.611111	0.634615	0.788611
Decision Tree	0.681818	0.553191	0.481481	0.514851	0.635741

UNIVERSITY OF GHANA

Random Forest Classifier

77%

Accuracy

81%

ROC-AUC

72%

Precision

66%

F1-Score

Highest stability across cross-validation folds

Handles non-linear feature interactions (Age & BMI)

Robust to outliers in insulin and skin thickness data

UNIVERSITY OF GHANA

Conclusion & Impact

Key Findings

- Glucose & BMI identified as strongest predictors
- Random Forest outperformed with highest F1-score
- Effective for screening at-risk populations over age 21

Clinical Benefits

- Rapid risk stratification at point of care
- Data-driven support for referrals & testing
- Early identification of pre-diabetic cases

Limitations

- Demographic Bias
- Restricted to Pima Indian females; generalization to other ethnicities requires validation.

1

Feature Enrichment

Incorporate HbA1c and lifestyle data

2

External Validation

Test on diverse global datasets

3

Model Explainability

Integrate SHAP values for insights

4

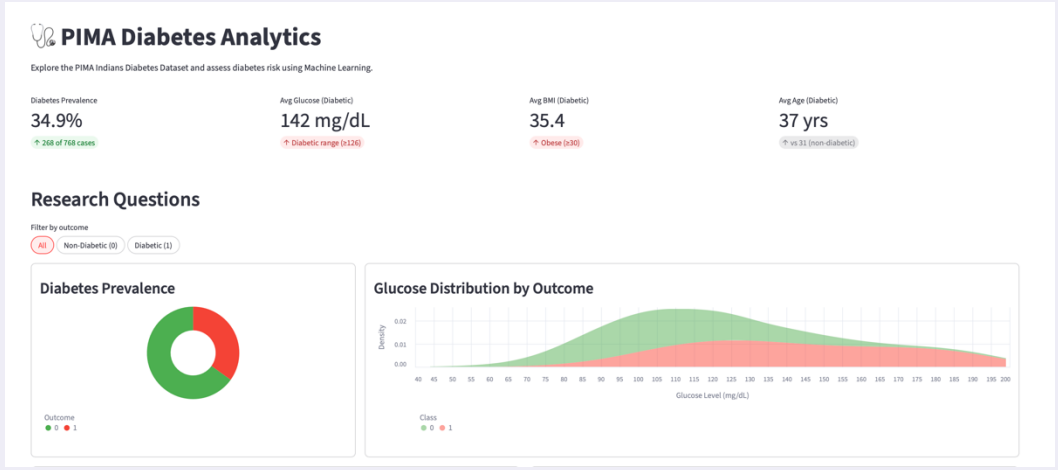
App Integration

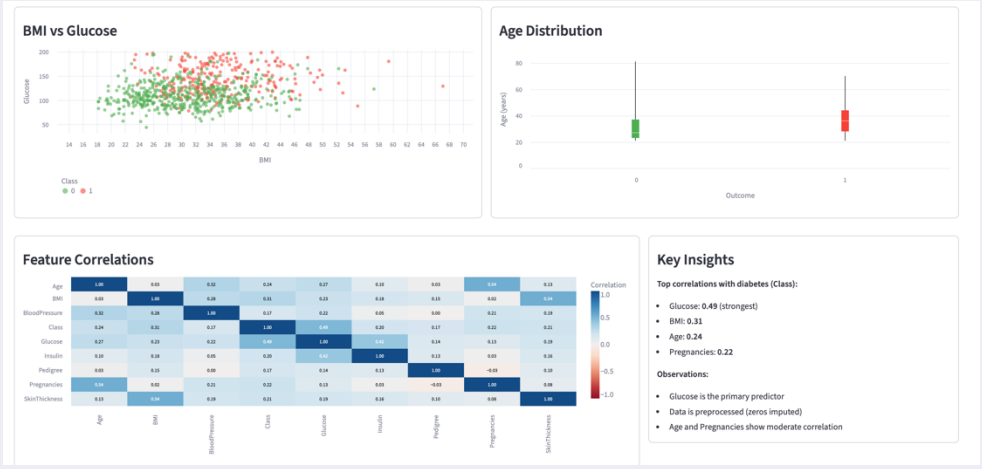
Deploy as mobile clinical tool

UNIVERSITY OF GHANA

Appendix

Streamlit





Graphs