UNIVERSITY OF GHANA

COLLEGE OF BASIC AND APPLIED SCIENCES

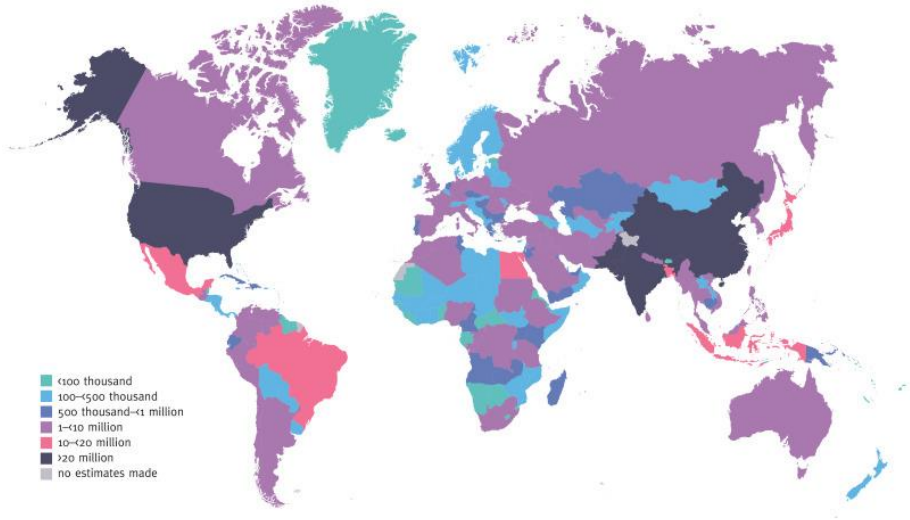
DSCD 611 Final Project Report
Machine Learning-Based Early Prediction of Diabetes Using the PIMA Dataset

Group B15
Group Leader: 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)

This project develops a supervised machine learning pipeline for binary classification to predict Type 2 Diabetes onset using a clinical dataset from female Pima Indian patients. The analysis identifies key metabolic markers, such as plasma glucose concentration and body mass index, as reliable early-warning indicators for the disease.

Type 2 Diabetes is a chronic condition characterized by insulin resistance, primarily driven by genetics, lifestyle factors, and obesity. Medical literature from organizations like the National Institute of Diabetes and Digestive and Kidney Diseases highlights that Pima Indians experience one of the highest global prevalence rates. Traditional screening methods rely on resource-intensive glucose tolerance tests, which prove challenging in rural settings. Machine learning addresses this gap by uncovering patterns in routine, non-specialized health metrics to augment clinical decision-making.(Looker et al., 2023)

The topic holds profound relevance amid a global diabetes crisis affecting over 422 million people, a leading cause of blindness, kidney failure, heart attacks, and stroke. Early detection through this project can prevent irreversible organ damage and alleviate economic pressures on healthcare systems. In low-resource environments, the resulting automated tool enables efficient triage of high-risk patients for targeted testing. Scientifically, it illustrates how predictive modeling transforms raw clinical data into actionable health insights with real-world impact.

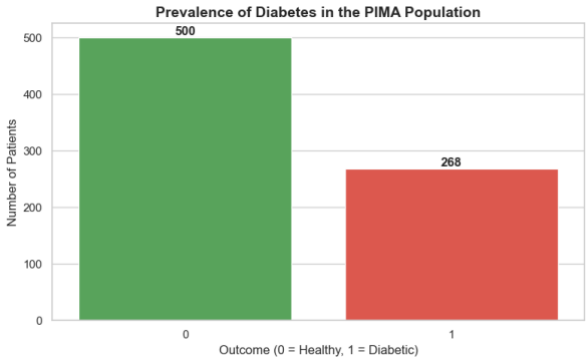


The study utilized the PIMA Indians Diabetes Dataset, comprising 768 samples across nine medical features and a binary target variable for Outcome (0 for no diabetes, 1 for diabetes). These features encompass the number of pregnancies, plasma glucose concentration from a 2-hour oral glucose tolerance test, diastolic blood pressure in mm Hg, triceps skin fold thickness in mm, 2-hour serum insulin in mu U/ml, body mass index calculated as weight in kg divided by height in m squared, a diabetes pedigree function scoring family history likelihood, and age in years. Feature selection drew from clinical theory on metabolic syndrome markers like glucose, BMI, and insulin, alongside demographic risks such as pregnancy history and age, confirmed by statistical evidence from exploratory data analysis.

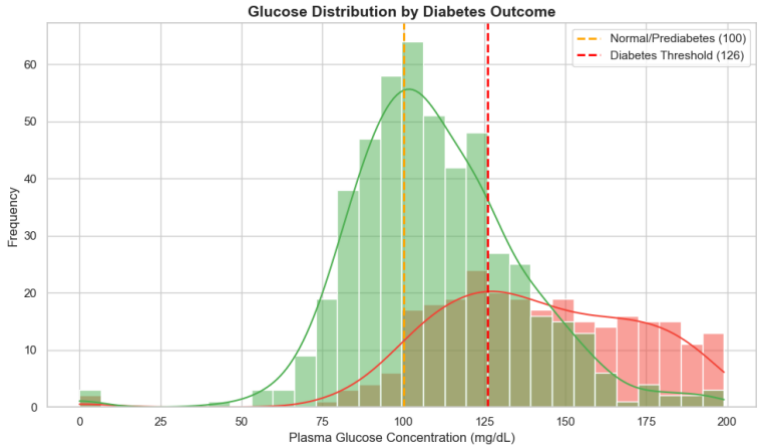
To guide the work, the team formulated four analytical research questions: the prevalence of diabetes outcomes in this demographic, the variation in plasma glucose levels between diabetic and non-diabetic groups, the interplay between BMI and glucose across outcomes, and does age correlate with diabetic risk.

The dataset reveals a moderate class imbalance in diabetes outcomes among Pima Indian females, with non-diabetic cases (Outcome = 0) comprising the majority at 500 patients or 65.1% of the total 768 samples, while diabetic cases (Outcome = 1) account for 268 patients or 34.9%. This distribution indicates that roughly 1 in 3 patients in the cohort has

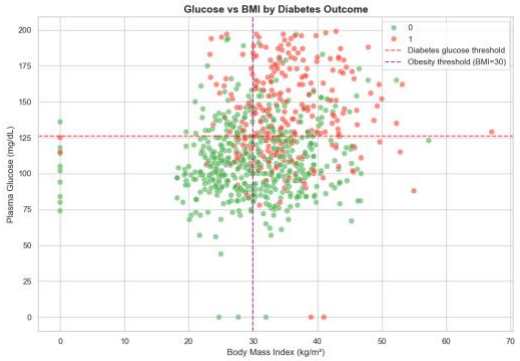
diabetes, underscoring the elevated prevalence in this high-risk population compared to global averages and highlighting the need for targeted predictive tools to address the minority diabetic class effectively.



Analysis uncovers a clear and substantial difference in average plasma glucose levels between diabetic and non-diabetic groups, with non-diabetic patients averaging 109.98 mg/dL and diabetic patients averaging 141.26 mg/dL a gap of over 30 mg/dL that aligns with clinical thresholds for impaired glucose tolerance. This stark contrast positions glucose as a pivotal early-warning marker, as elevated levels consistently signal insulin resistance in the diabetic cohort.

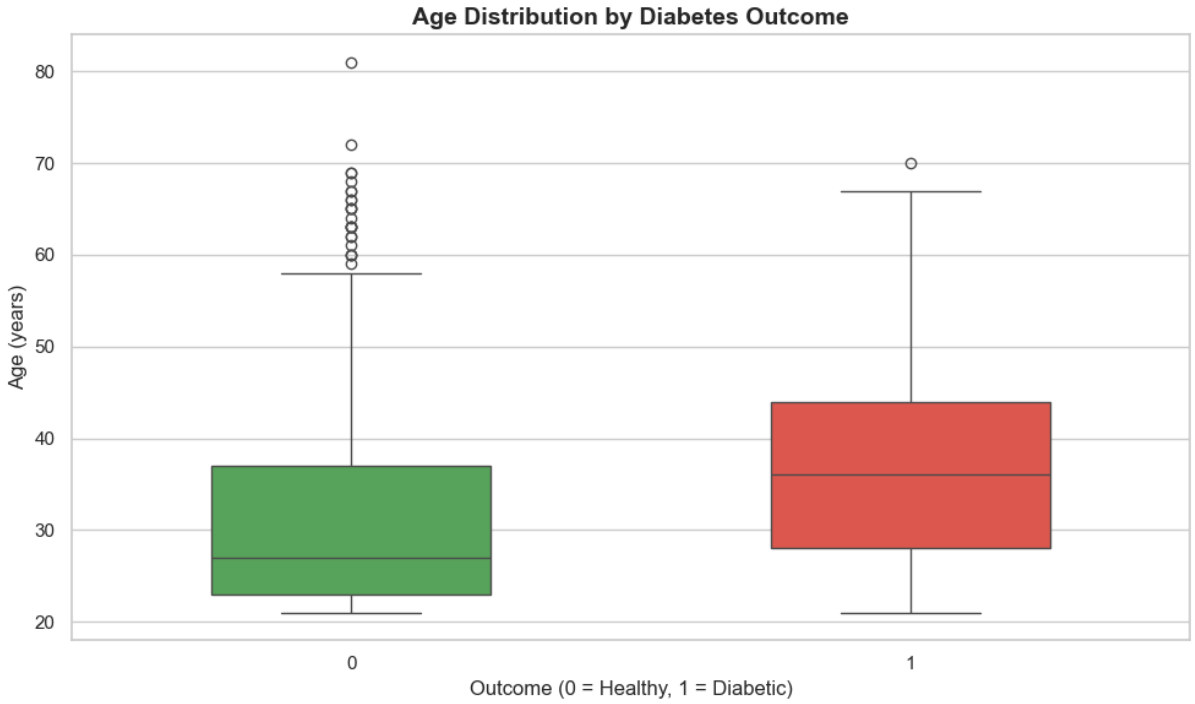


BMI and glucose exhibit a positive relationship overall, yet this interaction varies markedly by outcome: non-diabetic patients sustain lower glucose levels even as BMI rises, suggesting high BMI alone does not strongly elevate glucose in healthy individuals, whereas diabetic patients display consistently higher glucose across most BMI ranges, with the dangerous synergy of high BMI and elevated glucose appearing far more prevalent in this group and amplifying disease risk.



The analysis indicates that age does correlate with diabetes risk, as diabetic patients have a higher median age (37 years) compared to non-diabetic patients (27 years), reflecting

a notable 10-year difference. The upward shift of the diabetic age distribution and its wider interquartile range suggest that diabetes becomes more prevalent and variable with increasing age. However, the presence of younger individuals in both groups shows that age alone does not determine diabetes onset. Rather, age acts as a non-modifiable risk factor that increases susceptibility, especially when combined with other factors such as glucose levels and BMI, rather than serving as a standalone predictor.

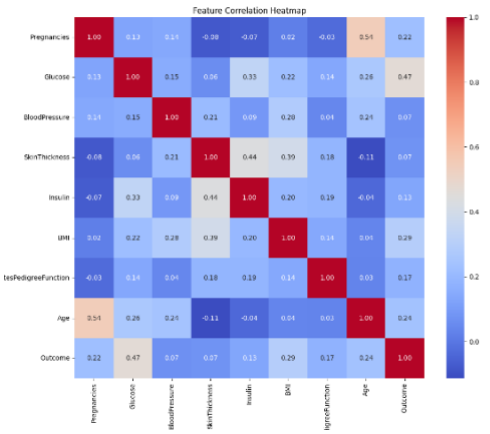


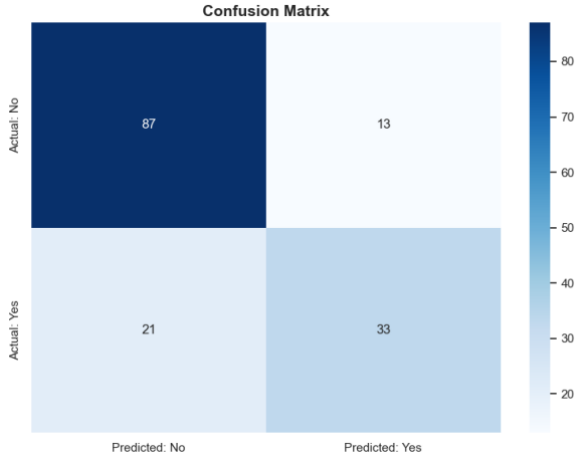
Model performance centered on accuracy, precision, recall, F1-score, and ROC-AUC, with Random Forest achieving the highest scores: 0.7792 accuracy, 0.7174 precision, 0.6111 recall, 0.6600 F1-score, and 0.8179 ROC-AUC.

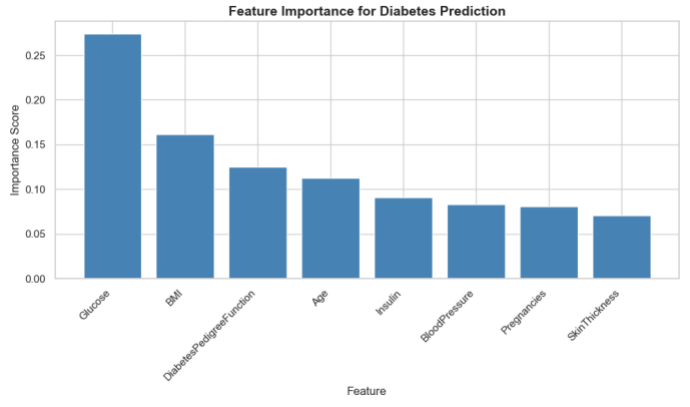
KNN followed at 0.7532 accuracy, SVM at 0.7403, Logistic Regression at 0.7078, and Decision Tree at 0.6818. Glucose and BMI emerged as the dominant predictors. These findings enable early interventions that could save thousands of lives through preventative care in high-risk populations.

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

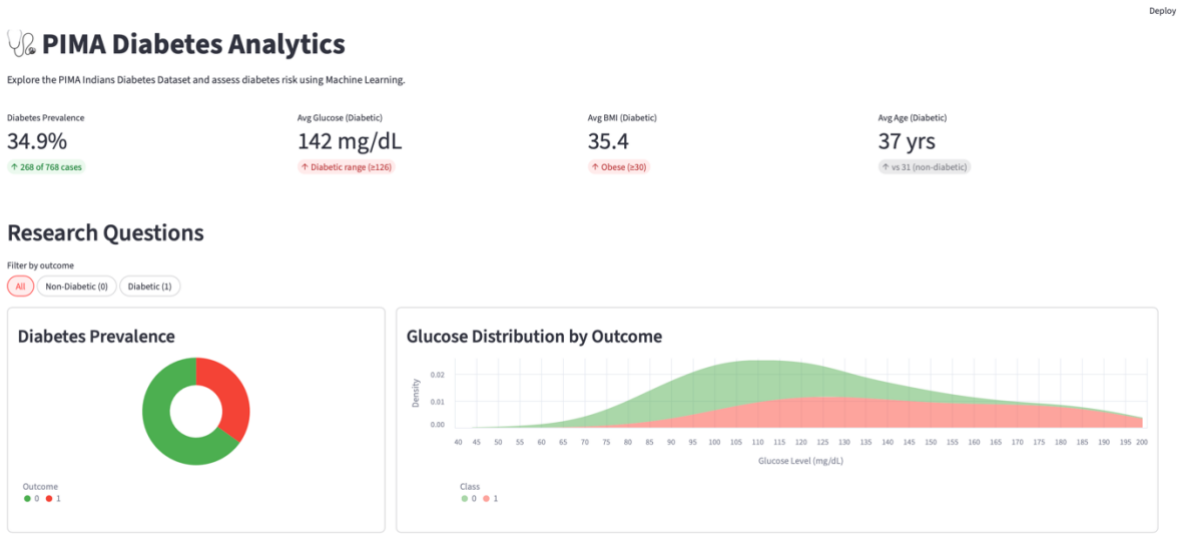
Appendix







Streamlit Dashboard



Reference

Looker, H. C., Chang, D. C., Baier, L. J., Hanson, R. L., & Nelson, R. G. (2023). Diagnostic criteria and etiopathogenesis of type 2 diabetes and its complications: Lessons from the Pima Indians. *La Presse Médicale*, *52*(1), 104176. https://doi.org/10.1016/J.LPM.2023.104176

Pima Indians Diabetes Database. (n.d.). Retrieved February 1, 2026, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database