KCB AI & Data Science

# Diabetes Prediction Capstone Project

Submitted By

Name: GILBERT WALIUBA

Email: gilbert@haofinder.com

Phone: +254 715 560 734

Institution

ADNIAN ABS

Date of Submission

15th SEPT 2025

Dataset Source

Pima Indians Diabetes Database (Kaggle)

# 1. Abstract

This project applies data science and machine learning techniques to predict diabetes using the Pima Indians Diabetes Database. The workflow includes data preprocessing, exploratory data analysis, feature engineering, model training, evaluation, and optimization. The goal is to create a reliable model that aids in early diabetes detection and to extract insights on key contributing health factors.

# 2. Problem Statement

Diabetes is a chronic disease with global prevalence and serious health consequences. Early detection is vital to improve outcomes and reduce healthcare burdens. This project develops predictive models to determine the likelihood of diabetes in patients based on medical diagnostic features, enabling proactive intervention and better patient care.  
  
Expected Outcomes:  
- A robust classification model capable of predicting diabetes risk.  
- Identification of the most influential features related to diabetes.  
- Visualization and statistical insights for healthcare decision-making.

# 3. Dataset Description

- Dataset Name: Pima Indians Diabetes Database  
- Source: Kaggle (rating >7)  
- Records: 768  
- Features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age  
- Target Variable: Outcome (1 = diabetes, 0 = no diabetes)

# 4. Data Collection and Exploration

Dataset loaded from Kaggle. Inspected dimensions, data types, and summary statistics. Identified biologically-impossible zeroes in features such as Glucose, BMI, Blood Pressure, Skin Thickness, and Insulin. Target distribution: imbalanced with ~65% non-diabetic and ~35% diabetic cases.

# 5. Data Preprocessing

Replaced zeroes with missing values and imputed using median. Standardized all numerical features using StandardScaler. Ensured data consistency and integrity.

# 6. Exploratory Data Analysis (EDA)

- Histograms: Revealed skewness in Insulin and Skin Thickness.  
- Correlation Heatmap: Strongest correlation observed between Glucose and Outcome.  
- Boxplots: Detected potential outliers in BMI and Insulin.  
- Target Distribution: Imbalance noted but manageable without oversampling.

# 7. Feature Engineering

Derived BMI × Age interaction feature. Created Age Groups (20–29, 30–39, 40–49, 50–59, 60+). One-hot encoded categorical AgeGroup variable. Re-scaled engineered features.

# 8. Model Selection

The following models were tested:  
- Logistic Regression  
- Decision Tree  
- Random Forest  
- Support Vector Machine (SVM)  
- Multi-Layer Perceptron (Neural Network)

# 9. Model Training and Evaluation

Dataset split: 80% training, 20% testing. Evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.  
  
Results (baseline models):  
- Logistic Regression: Accuracy ~77%, ROC-AUC ~82%  
- Random Forest: Accuracy ~79%, ROC-AUC ~85%  
- SVM: Accuracy ~76%, ROC-AUC ~83%  
- MLP: Accuracy ~78%, ROC-AUC ~84%

# 10. Model Optimization

Random Forest: Tuned hyperparameters (n\_estimators, max\_depth, min\_samples\_split).  
Logistic Regression: Tuned regularization strength (C).  
Cross-validation (5-fold) confirmed robustness.  
  
Best Model: Random Forest (tuned)  
- Accuracy: ~80%  
- ROC-AUC: ~86%

# 11. Feature Importance

Top predictors of diabetes (Random Forest):  
1. Glucose  
2. BMI  
3. Age  
4. Pregnancies  
5. Diabetes Pedigree Function

# 12. Conclusion and Recommendations

The tuned Random Forest model achieved the highest predictive performance with ROC-AUC of ~86%. Glucose level is the strongest predictor of diabetes. BMI and Age are also highly influential features.  
  
Recommendations:  
- Use the model as a screening tool for early detection, not as a diagnostic replacement.  
- Collect additional data across diverse populations to improve generalization.  
- Consider probability calibration for clinical decision thresholds.  
- Future work: experiment with ensemble methods and oversampling techniques to handle class imbalance.

# 13. References

- Kaggle: Pima Indians Diabetes Database  
- Scikit-learn Documentation  
- Seaborn and Matplotlib Libraries

# 14. Appendix

Code, plots, and outputs saved in project directory `diabetes\_capstone\_outputs/`. Best model exported as `best\_model\_random\_forest.joblib`. Short markdown summary file: `short\_report.md`.