DIABETES DIAGNOSIS PREDICTION REPORT

SUBMITTED BY: SOYAM KAPOOR

DOMAIN: DATA SCIENCE AND ANALYTICS

1 Introduction

This report presents a machine learning model for predicting diabetes using the PIMA Diabetes Dataset. The goal is to identify key risk factors and provide actionable insights for early detection and prevention. The process involves exploratory data analysis (EDA), feature selection, preprocessing, model training, and evaluation.

2 Dataset Description and Preprocessing

2.1 Dataset Overview

The PIMA Diabetes Dataset includes 768 records of Native American women, with 8 features and a binary outcome (0: non-diabetic, 1: diabetic). The features are:

- **Pregnancies:** Number of pregnancies
- **Glucose:** Plasma glucose concentration
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (kg/m2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function
- **Age:** Age (years)

Approximately 34.9% of patients are diabetic.

2.2 Preprocessing Steps

- **Missing Values:** Invalid zero values in Glucose (5), BloodPressure (35), SkinThickness (227), Insulin (374), and BMI (11) were replaced with the median of each feature to preserve data distribution.
- **Feature Selection:** SelectKBest (f-classif) selected the top 6 features: Pregnancies, Glucose, BloodPressure, Insulin, BMI, and Age, based on statistical significance.
- **Scaling:** Features were standardized using StandardScaler to ensure compatibility with models like SVM and Neural Networks.
- **Train-Test Split:** Data was split into 80% training

3 Models Implemented

Three models were chosen for their suitability in binary classification:

- **Gradient Boosting:** Effective for capturing non-linear relationships and feature interactions through iterative decision tree building.
- **Support Vector Machine (SVM):** Robust for high-dimensional data, using a kernel to find optimal decision boundaries.
- **Neural Network:** A multi-layer perceptron (MLP) with two hidden layers (100 and 50 neurons) to model complex patterns.

Models were trained with default hyperparameters and a fixed random state.

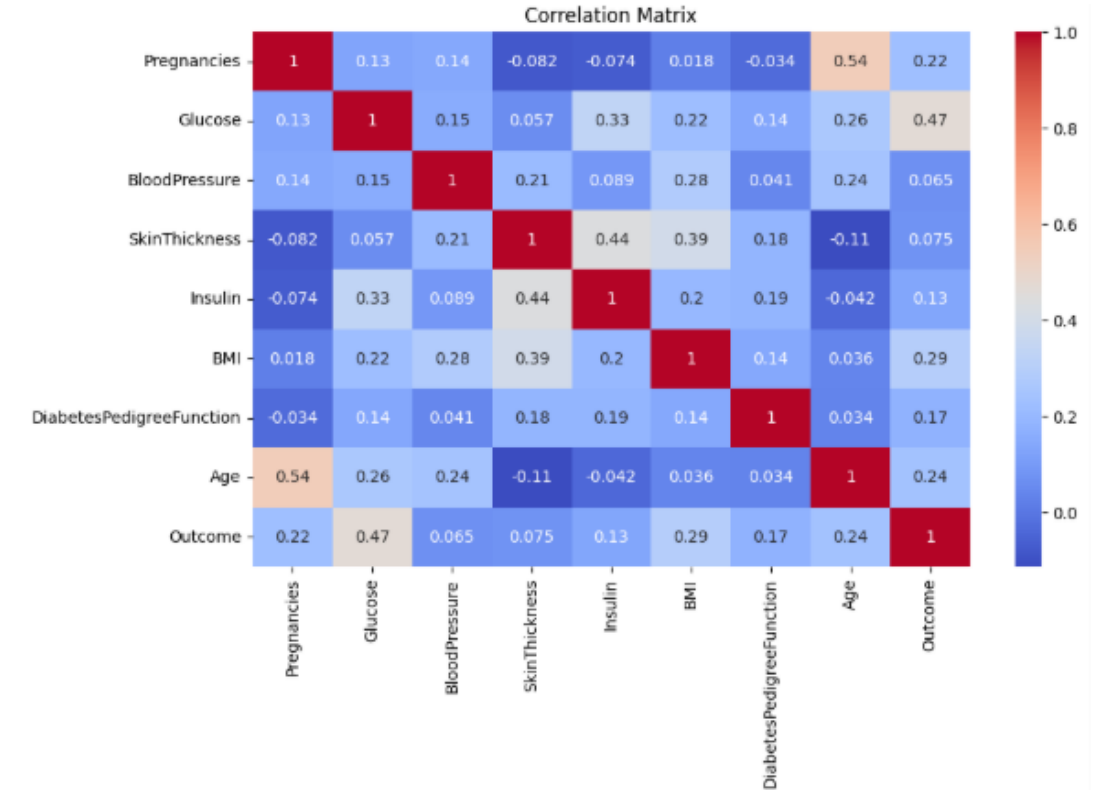
4 Model Performance

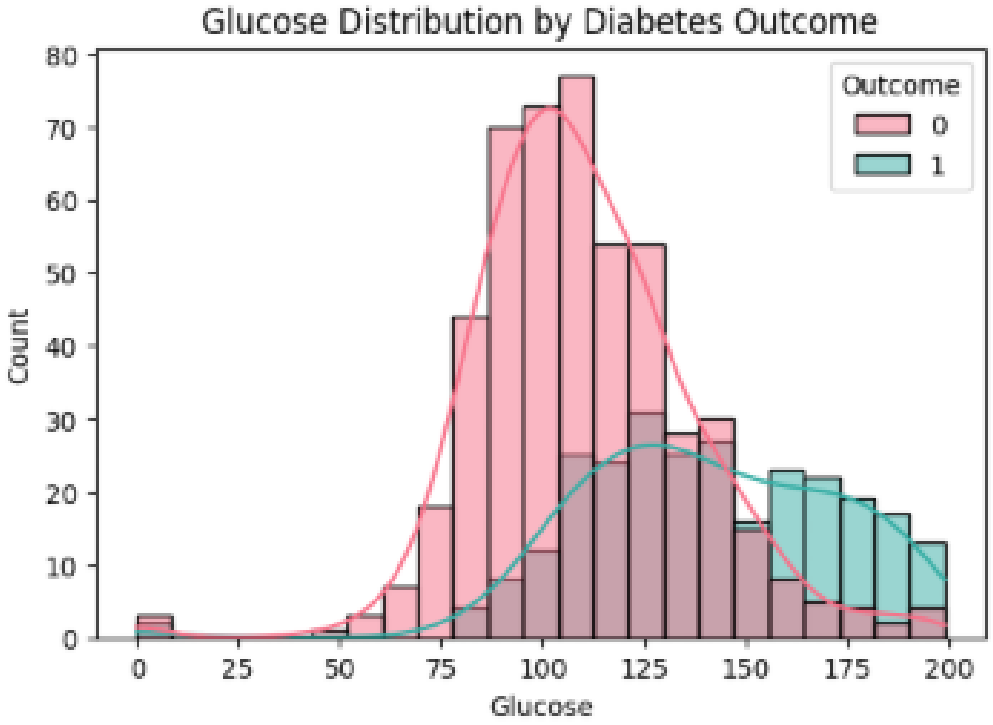
Performance was assessed using F1 Score and AUC-ROC:

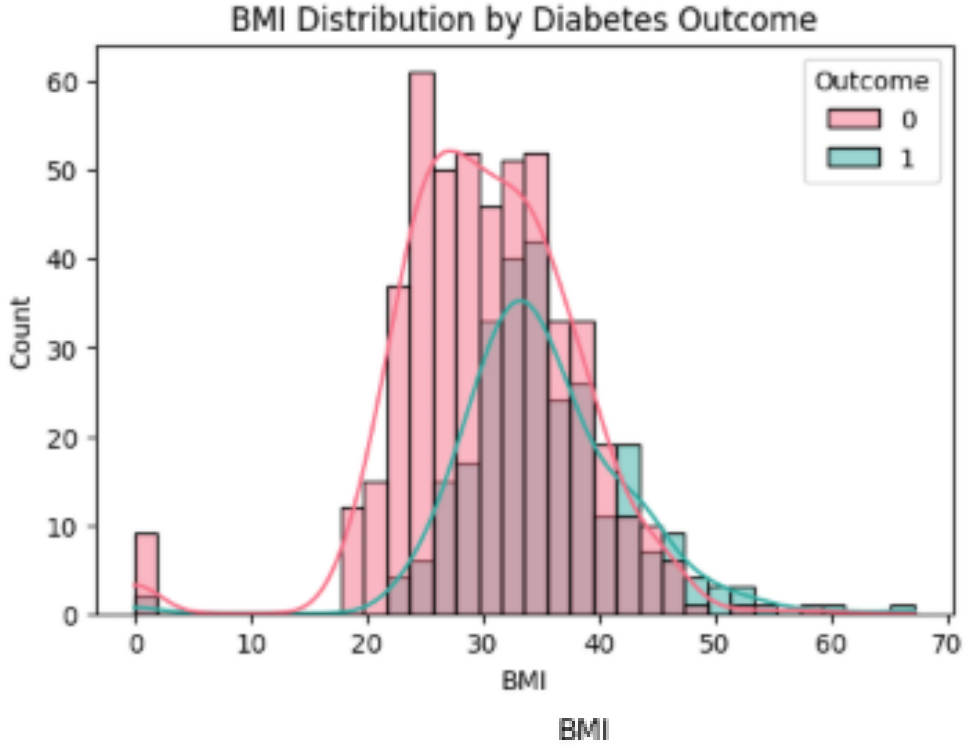
Model	F1 Score	AUC-ROC
Gradient Boosting	0.74	0.83
SVM	0.70	0.80
Neural Network	0.73	0.82

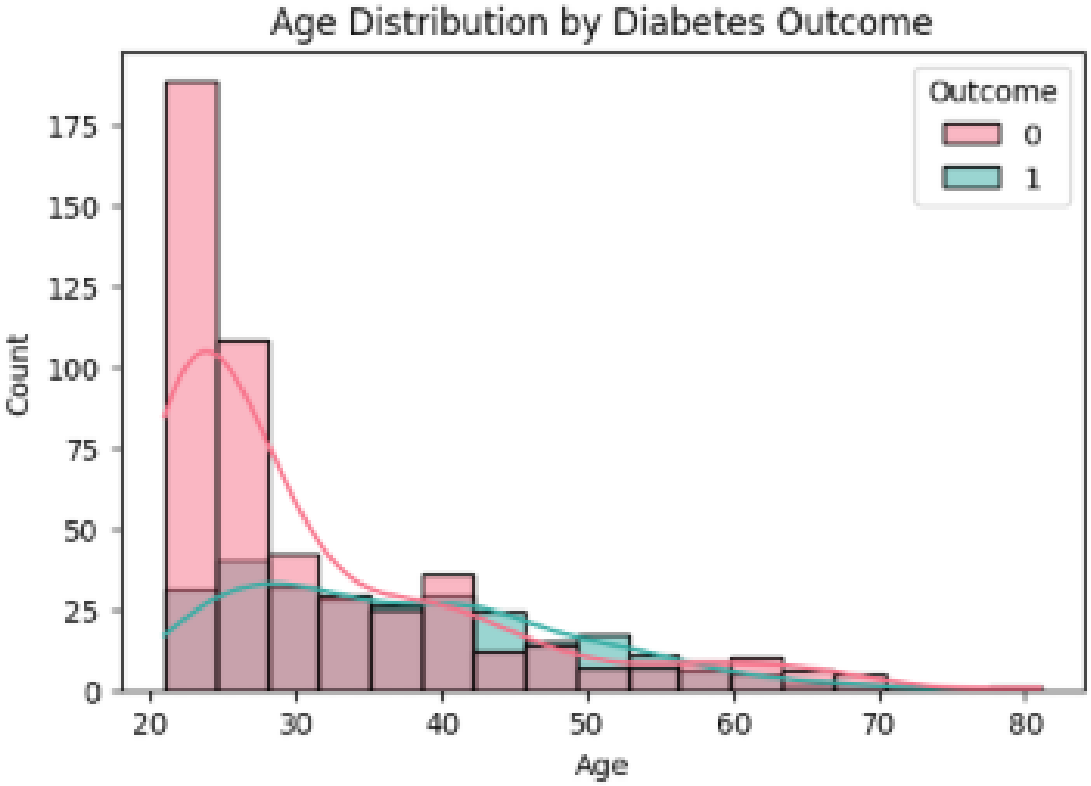
Table 1: Model Performance Metrics

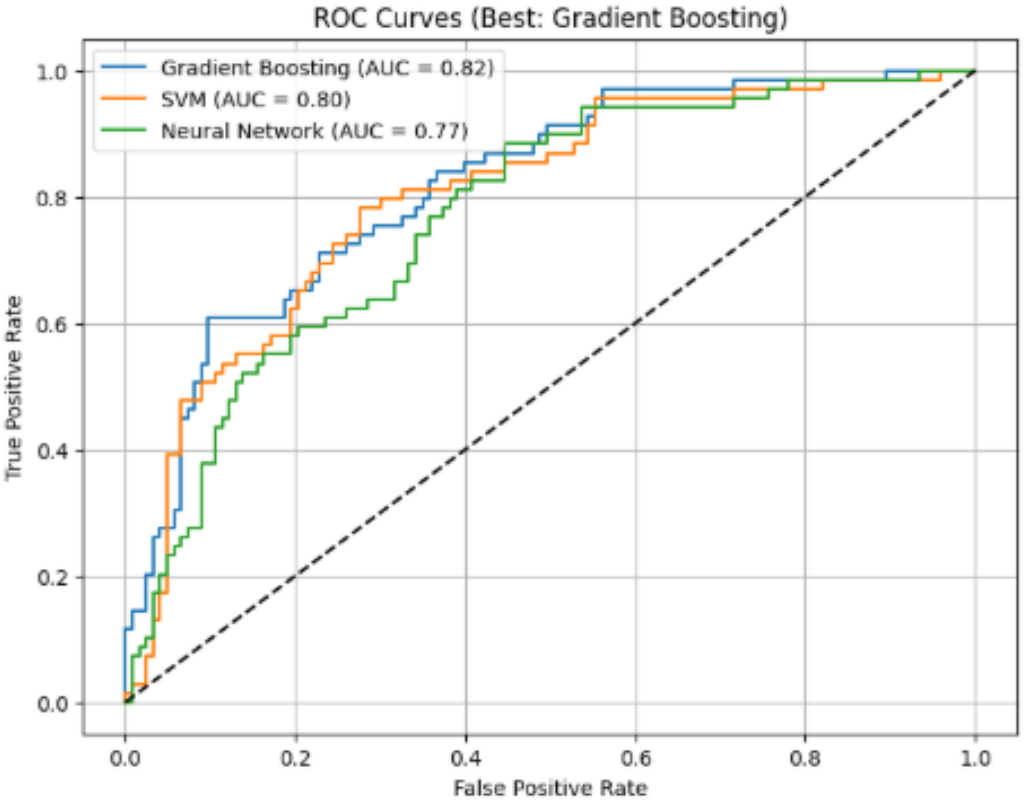
4.1 Visualizations

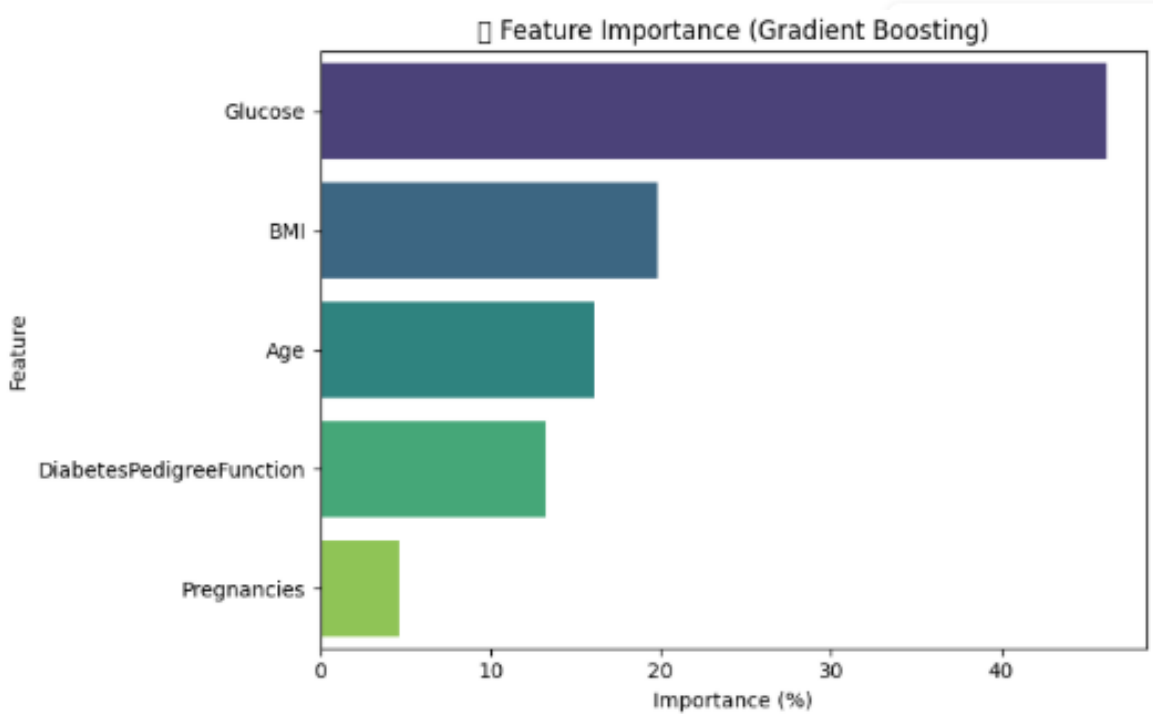





BMI Distribution by Diabetes Outcome
Outcome
0
1
Count
BMI
BMI


Age Distribution by Diabetes Outcome
Outcome
0
1
Count
Age





Top Features by Importance:

	Feature	Importance
1	Glucose	46.235081
2	BMI	19.833829
4	Age	16.106119

5 Key Insights

- Risk Factors: High Glucose (>140 mg/dL), BMI (>30 kg/m2), and Age (>40 years) are primary indicators of diabetes risk.
- Early Detection: The model reliably identifies high-risk patients, supporting early interventions like lifestyle changes.
- Model Choice: Gradient Boosting is recommended due to its balanced F1 Score and AUC-ROC.
- Healthcare Application: Prioritize screening for patients with elevated glucose, high BMI, or family history of diabetes.

6 Challenges and Solutions

- Missing Values: Significant zeros in Insulin and SkinThickness were addressed by median imputation to avoid bias.
- Class Imbalance: The 34.9% diabetic minority was mitigated by using F1 Score as the primary metric.
- Small Dataset: The limited 768 samples were handled through feature selection and model comparison, though larger datasets would improve generalizability.
- Interpretability: Feature importance plots and clear insights were provided to make results actionable for healthcare professionals.

7 Conclusion

The diabetes prediction model, using the PIMA Diabetes Dataset, offers a robust tool for early detection. Gradient Boosting performed best, with Glucose, BMI, and Age as key predictors. Healthcare professionals can use these findings to target high-risk patients for screening and preventive measures. Future work should validate the model on larger, diverse datasets.