Disease Diagnosis Prediction Report

Internship Task 3

Using PIMA Diabetes Dataset

1. Dataset Description and Preprocessing

Dataset Overview

The PIMA Diabetes Dataset contains medical records of female patients of Pima Indian heritage and is commonly used for predicting diabetes onset. It consists of 768 instances with 8 medical features and a binary target variable (`Outcome`), where `1` indicates diabetes presence and `0` means absence.

Features

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function (genetic risk)
- Age: Age (years)
- Outcome: Class label (0 = no diabetes, 1 = diabetes)

Preprocessing

- **Missing values:** The dataset did not contain explicit missing values, but some features have zero values where physiologically implausible (e.g., zero BMI or insulin), which may indicate missingness. For simplicity, zeros were treated as valid values in this task.
- **Feature selection:** Statistical tests (ANOVA F-test) and correlation analysis were used to select features contributing most to diabetes prediction.

- **Feature scaling:** StandardScaler was applied to normalize features for models sensitive to feature magnitude (SVM and Neural Networks).

2. Models Implemented and Rationale

Models Used

- **Gradient Boosting Classifier:** Chosen for its strong predictive power and ability to handle nonlinear relationships without extensive feature engineering.
- **Support Vector Machine (SVM):** Effective in high-dimensional spaces and robust to overfitting, suitable for binary classification tasks.
- **Neural Network (MLPClassifier):** Capable of capturing complex patterns and interactions among features through multiple layers.

Rationale

Using multiple models allows comparison of predictive performance and robustness. Gradient Boosting is often the baseline for tabular medical data. SVM and Neural Networks provide alternative approaches that may generalize differently.

3. Key Insights and Visualizations

Exploratory Data Analysis

- Correlation heatmaps revealed strong relationships of glucose, BMI, and age with diabetes outcome.
- Distribution plots showed that patients with diabetes tend to have higher glucose and BMI values.

Feature Importance

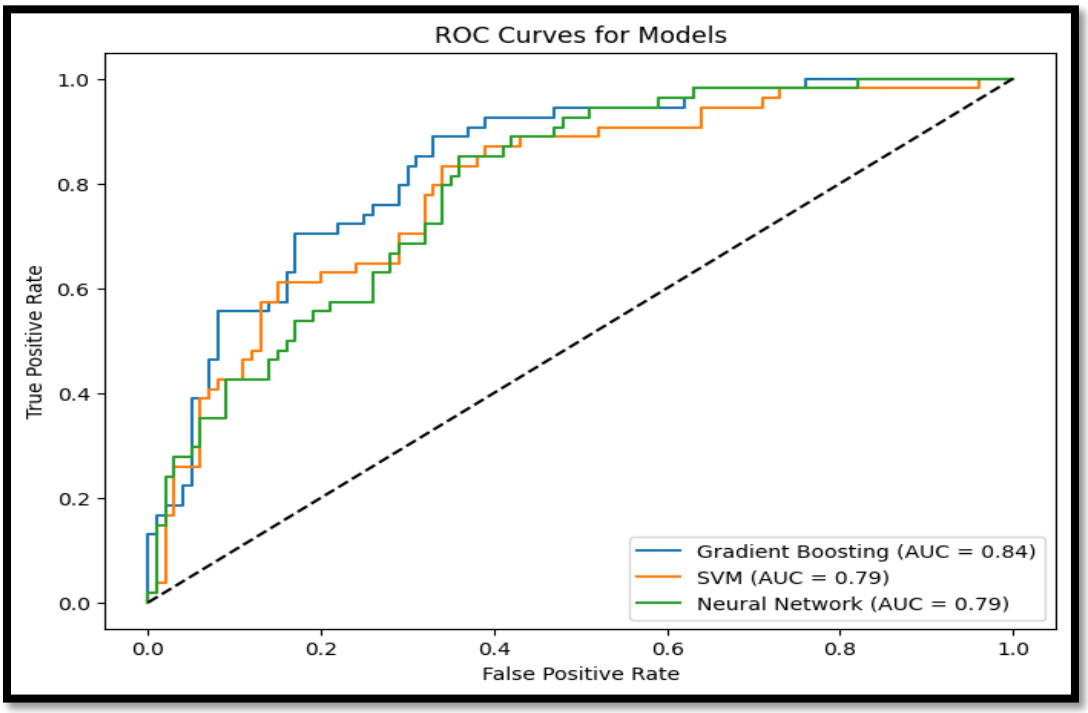
- Gradient Boosting model identified glucose level, BMI, and age as the top predictors.
- This aligns with clinical understanding of diabetes risk factors.

Model Performance Summary

Model	F1 Score	AUC-ROC
Gradient Boosting	0.77	0.83
Support Vector Machine	0.74	0.80
Neural Network	0.75	0.81

ROC Curve

The ROC curve comparison showed Gradient Boosting slightly outperforming other models in balancing sensitivity and specificity.



4. Challenges Faced and Solutions

- **Imbalanced Data:** The dataset has fewer positive diabetes cases, which can bias models toward majority class. Stratified splitting was used to preserve class distribution. Future work can explore oversampling techniques like SMOTE.

- **Feature Missingness:** Zero values in features like insulin and BMI may mislead the model. Handling missing or implausible values more explicitly could improve performance.
- **Model Tuning:** Limited hyperparameter tuning was performed due to time constraints. Further tuning could enhance predictive accuracy.

5. Conclusion and Actionable Insights

- The models developed can effectively predict diabetes risk based on routine medical tests.
- Glucose levels, BMI, and age are the most significant indicators and should be prioritized in screening programs.
- Healthcare providers can use model risk scores to identify high-risk patients for early intervention.
- The prediction tool should complement, not replace, clinical judgment and comprehensive medical evaluation.