Project Title: Comparative Analysis of Classification Models on the Pima Indians Diabetes Dataset

Introduction

This project aims to apply various supervised learning classification models to the Pima Indians Diabetes dataset to predict whether individuals have diabetes based on diagnostic measurements. The objective is to evaluate which model provides the best balance of precision, recall, and accuracy, thus offering a reliable tool for early diabetes detection.

Data Description

The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and age. The dataset comprises 768 instances, with the Outcome variable being binary (0 or 1), where 1 indicates the presence of diabetes.

Data Preprocessing

The dataset underwent several preprocessing steps:

Handling Missing Values: Fields like Glucose and Blood Pressure had zero values, which were replaced with NaN and then imputed with the mean of their respective columns.

Normalization: Features were scaled to have a uniform range, ensuring no single feature dominated others in the model training process.

Splitting Data: The data was divided into a training set (70%) and a testing set (30%) to evaluate the model performance.

Models Evaluated

Three classification models were evaluated:

Random Forest Classifier

Support Vector Machine (SVM)

Gradient Boosting Classifier

Results and Model Comparison

Performance Metrics:

Precision: Measures the accuracy of positive predictions. Formally, it is the ratio of true positives to the sum of true and false positives.

Recall: Measures the ability of a model to find all relevant instances. It is the ratio of true positives to the sum of true positives and false negatives.

F1-Score: Harmonic mean of precision and recall. It provides a balance between precision and recall in one number.

Accuracy: Overall, how often the model correctly predicts the outcome.

Confusion Matrix:

Shows the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

Results Summary:

Random Forest Classifier:

Precision: 0.81 (Class 0), 0.64 (Class 1)

Recall: 0.81 (Class 0), 0.64 (Class 1)

F1-Score: 0.81 (Class 0), 0.64 (Class 1)

Accuracy: 75%

Support Vector Machine:

Precision: 0.79 (Class 0), 0.63 (Class 1)

Recall: 0.81 (Class 0), 0.59 (Class 1)

F1-Score: 0.80 (Class 0), 0.61 (Class 1)

Accuracy: 74%

Gradient Boosting Classifier:

Precision: 0.83 (Class 0), 0.64 (Class 1)

Recall: 0.77 (Class 0), 0.74 (Class 1)

F1-Score: 0.80 (Class 0), 0.69 (Class 1)

Accuracy: 74%

Analysis

The Gradient Boosting Classifier shows a slightly better balance between precision and recall, particularly for Class 1 (diabetic patients), which is critical for medical diagnostics where missing a positive case (diabetes present) can have serious consequences. Though its overall accuracy ties with the SVM, its superior recall for diabetic cases (Class 1) makes it potentially more valuable in practical applications.

Conclusion and Recommendations

The Gradient Boosting Classifier is recommended as the best model due to its higher recall for diabetic cases and overall balance in precision and recall across classes. Future work could explore additional feature engineering, more sophisticated ensemble techniques, or deep learning approaches to further improve the predictability and reliability of the model.