Machine Learning Classification Algorithms to Predict Diabetes

1. Introduction:

Machine learning is a part of artificial intelligence (AI) that focuses on teaching computers to learn from data and make decisions or predictions. Instead of being programmed with specific instructions for every task, machines can recognize patterns, make choices, and improve their accuracy by learning from examples. This ability to learn from data has made machine learning a key tool in many areas, like diagnosing diseases, predicting financial trends, recognizing images, and understanding human language. With more data and better computing power, machine learning continues to change how we solve problems and make decisions.

Classification Algorithms in Machine Learning

Classification is a supervised machine learning technique where an algorithm learns to categorize input data into predefined classes. This process involves several common algorithms, each with unique characteristics. Logistic Regression is a linear model often used for binary classification, making it simple yet effective for various tasks. This project evaluates 5 such classification algorithms. The K-Nearest Neighbors (KNN) algorithm classifies data points based on the majority class of their nearest neighbors, making it intuitive but sensitive to the dataset's structure. Support Vector Machine (SVM) aims to find the best hyperplane that maximizes the margin between classes, making it robust for high-dimensional spaces. Decision Trees model decisions hierarchically, dividing data based on features, while Random Forests improve on this approach by constructing multiple decision trees to achieve greater accuracy and robustness. Each of these algorithms brings different advantages, making them valuable tools in various contexts.

Dataset Description

The Diabetes dataset, often referred to as the Pima Indians Diabetes Database, contains medical and demographic information about female patients of Pima Indian heritage, aiming to predict the onset of diabetes. This dataset includes 768 observations, each representing an individual patient. The dataset contains eight input features and one output variable, providing insights into health indicators that could potentially influence the likelihood of diabetes.

Input/Independent Variables (Features):

- **Pregnancies**: Number of times the patient has been pregnant.
- **Glucose**: Plasma glucose concentration measured after a two-hour oral glucose tolerance test.
- **BloodPressure**: Diastolic blood pressure (mm Hg).
- **SkinThickness**: Triceps skin fold thickness (mm).
- **Insulin**: Serum insulin (mu U/ml).
- **BMI**: Body mass index, calculated as weight in kg/(height in m)^2.
- **DiabetesPedigreeFunction**: Diabetes pedigree function, representing the likelihood of diabetes based on family history.

- **Age**: Age of the patient in years.

Output/Dependent Variable:

- **Outcome**: A binary variable indicating whether the patient has diabetes (1) or not (0).

2. Different Classification Algorithms

In this section, we'll cover the main classification methods used in this analysis. These methods are widely used to predict categories based on input data.

2.1 Logistic Regression

Logistic Regression is a method that predicts the likelihood of an outcome belonging to one of two groups. It's like a decision-making tool that assesses certain features to calculate the probability of an event happening. For example, it can predict if a patient has diabetes based on their health data. It's straightforward and effective when the differences between the groups are somewhat clear.

2.2 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a method that classifies data points based on their closest neighbors. Imagine you want to know what type of fruit an apple is. If it's surrounded by oranges, KNN will likely classify it as an orange. It compares new data points with existing ones, assigning a category based on the nearest group. This method is simple but can vary in accuracy depending on the chosen number of neighbors.

2.3 Support Vector Machine (SVM)

Support Vector Machine finds the best boundary that separates two groups of data with a "buffer zone" in between. It is like drawing a line between two groups of dots on a paper, ensuring they're as far apart as possible. SVM works well even when the two groups are close to each other and provides a clear separation, making it ideal for complex data where a simple line may not work.

2.4 Decision Tree

A Decision Tree is like a flowchart of choices that lead to a decision. It's a series of "if-then" steps that guide the model to classify data points. For example, a tree might first check if someone's glucose level is high and then ask about their age to make a prediction. Each step narrows down the possible outcomes until a final decision is made. It's easy to understand and visualize, though it can sometimes make too specific decisions if not managed well.

2.5 Random Forest

Random Forest builds multiple decision trees and combines their predictions for a stronger result. Imagine asking a group of doctors for a diagnosis rather than just one—their combined opinion is likely more reliable. Each tree in the forest considers different parts of the data, and together, they create a more accurate and balanced prediction. This method is powerful

because it balances out the individual “opinions” of each tree, making it less likely to overfit or make errors.

3. Results and Analysis:

The results for each classification algorithm used to predict diabetes are as follows, including analysis of both accuracy scores and confusion matrices:

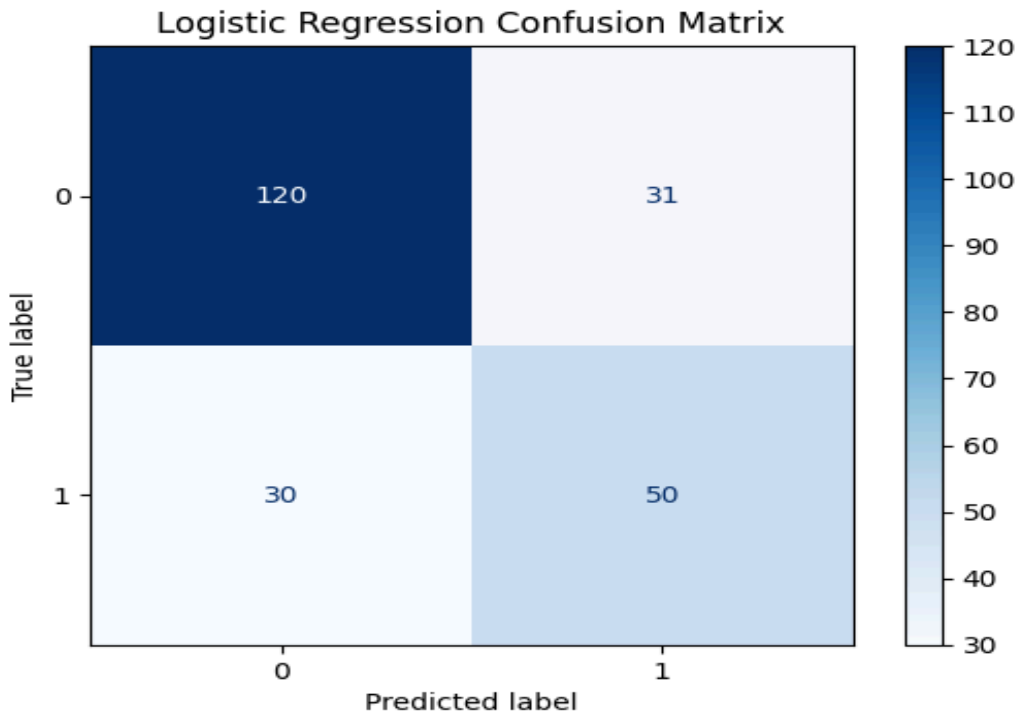
Logistic Regression:

Logistic Regression Confusion Matrix:

[120 31]

[30 50]

Logistic Regression Accuracy Score: 0.7359



Logistic Regression achieved an accuracy of approximately 73.6%. The confusion matrix correctly identified 120 non-diabetic cases and 50 diabetic cases but misclassified 31 non-diabetic cases as diabetic and 30 diabetic cases as non-diabetic. Logistic Regression effectively captured the general patterns, though it struggled slightly with more ambiguous cases, particularly leading to some false positives.

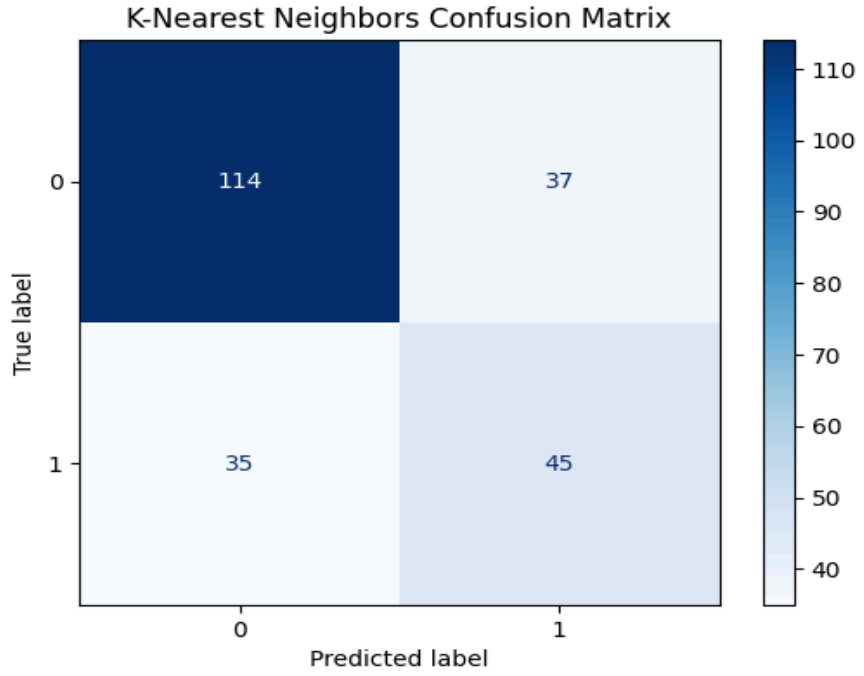
K-Nearest Neighbors (KNN):

K-Nearest Neighbors Confusion Matrix:

[114 37]

[35 45]

K-Nearest Neighbors Accuracy Score: 0.6883



KNN achieved an accuracy of about 68.8%. Its confusion matrix reveals that it correctly classified 114 non-diabetic cases and 45 diabetic cases, while misclassifying 37 non-diabetic cases as diabetic and 35 diabetic cases as non-diabetic. The model's reliance on neighbors made it sensitive to nearby data points, which sometimes led to incorrect classifications, particularly in cases where the diabetic and non-diabetic groups had overlapping characteristics.

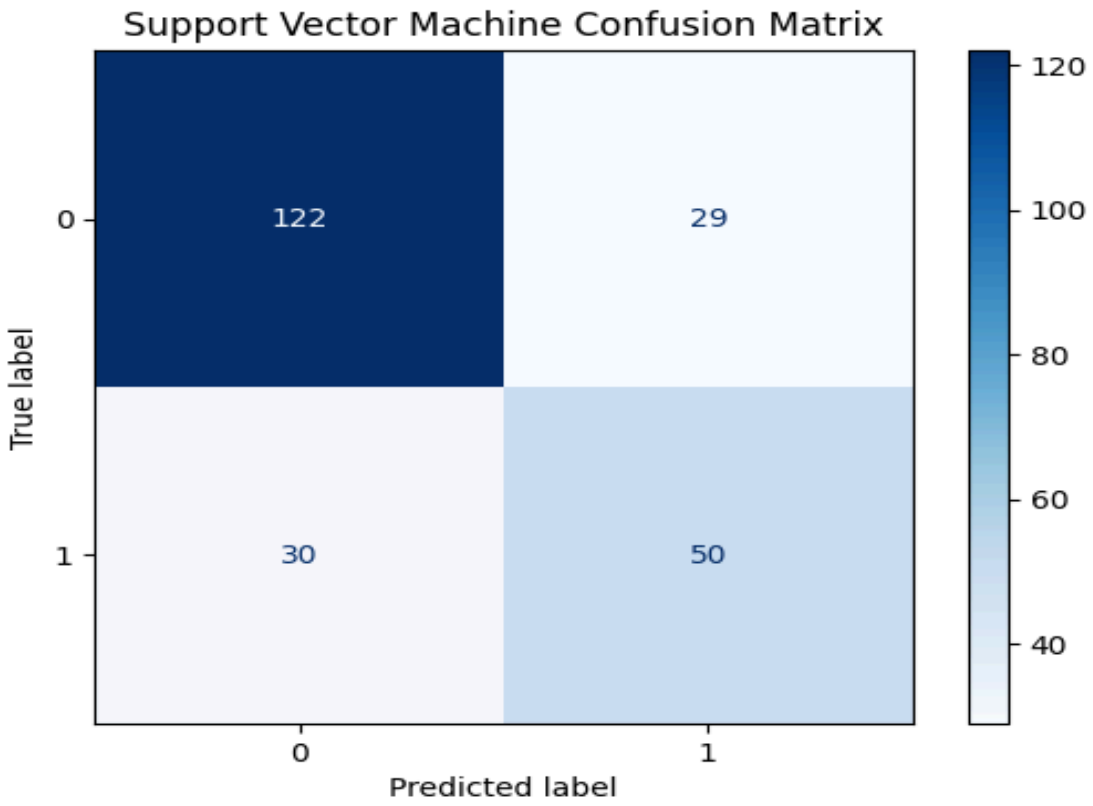
Support Vector Machine (SVM):

Support Vector Machine Confusion Matrix:

[122 29]

[30 50]

Support Vector Machine Accuracy Score: 0.7446



SVM demonstrated an accuracy of 74.5%. The confusion matrix shows that it successfully identified 122 non-diabetic cases and 50 diabetic cases, misclassifying 29 non-diabetic cases and 30 diabetic cases. SVM's boundary-focused approach effectively separated the classes, leading to fewer false positives and false negatives than some other models, making it a solid performer in distinguishing between diabetic and non-diabetic cases.

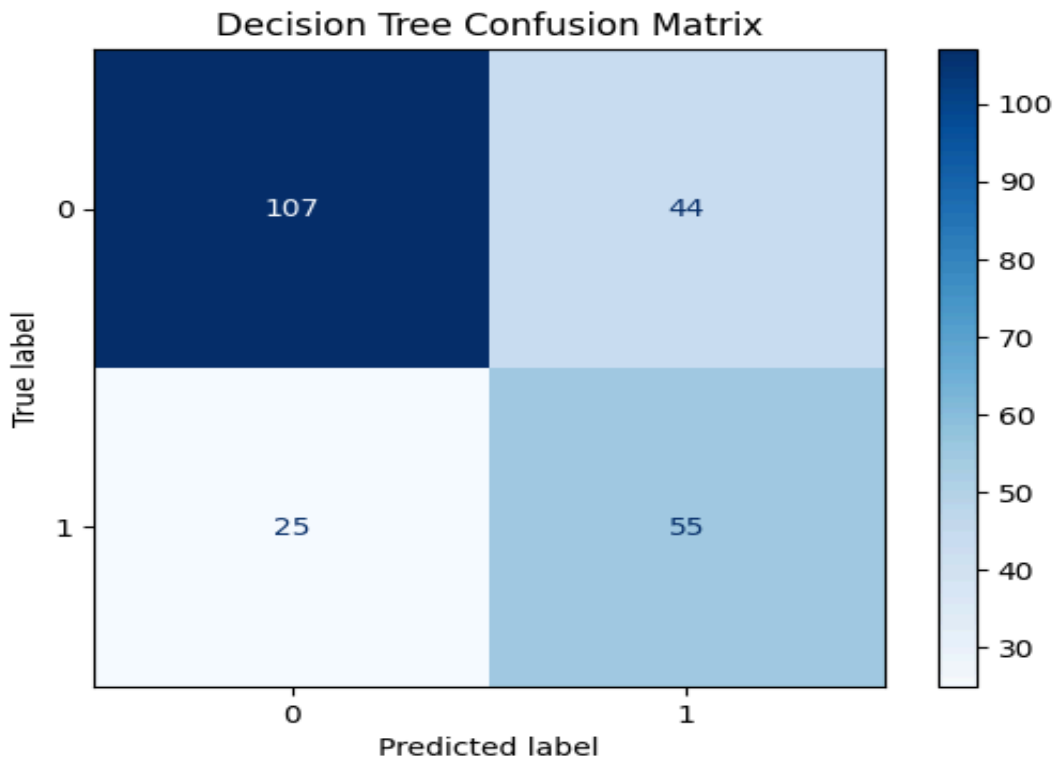
Decision Tree:

Decision Tree Confusion Matrix:

[107 44]

[25 55]

Decision Tree Accuracy Score: 0.7013



The Decision Tree model provided an accuracy of 70.1%. According to the confusion matrix, it correctly classified 107 non-diabetic cases and 55 diabetic cases but misclassified 44 non-diabetic cases and 25 diabetic cases. The model performed well in its straightforward decision-making process but sometimes made overly specific classifications, leading to a relatively higher number of false positives in predicting diabetes.

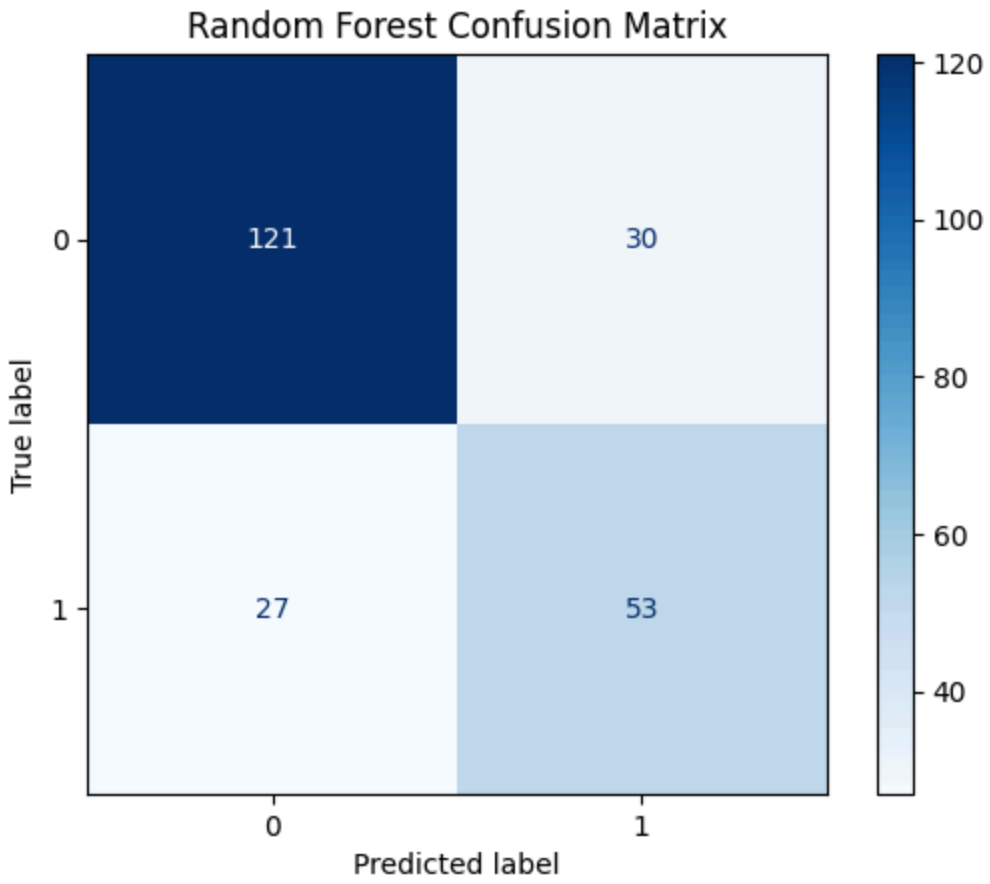
Random Forest:

Random Forest Confusion Matrix:

[121 30]

[27 53]

Random Forest Accuracy Score: 0.7532



Random Forest achieved the highest accuracy at approximately 75.3%. The confusion matrix indicates that it correctly identified 121 non-diabetic cases and 53 diabetic cases, with only 30 non-diabetic cases and 27 diabetic cases misclassified. By aggregating the results from multiple decision trees, Random Forest minimized both false positives and false negatives, providing the most balanced and accurate predictions among the models.

4. Conclusion

This analysis evaluated five popular classification algorithms to predict diabetes in patients using the Pima Indians Diabetes dataset. Each model's performance was reviewed based on accuracy scores and confusion matrix breakdowns. Random Forest demonstrated the highest accuracy and had the fewest misclassifications, making it the most reliable model for predicting diabetes in this dataset. The confusion matrix analysis further illustrated Random Forest's strength in balancing true positives and true negatives, minimizing errors effectively.

5. References:

http://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/discussion?sort=hotness