**Loan Classifier Machine Learning Project**

Author/s: Abdelrahman Emad 2021/07436  
Mario Magdy  
Youssef Sherif

Abstract:

This report outlines the process and findings of a machine learning project aimed at classifying loan applications. The project utilized several machine learning models to predict loan approval status based on a variety of applicant features. The data for this project was provided by our Teacher Assistant Sarah Nabil. Key steps included data preprocessing, model training, and evaluation using various performance metrics

Introduction:

### Problem Statement:

### The primary objective of this project is to build a classifier that can accurately predict whether a loan application will be approved or not based on the applicant's information.

### Objectives:

* To preprocess the provided loan application data to make it suitable for modeling.
* To train and evaluate various machine learning models.
* To compare the performance of different models and select the best one.

## Methodology:

### Data Collection:

The dataset was provided by our Teacher Assistant Sarah Nabil. It contains information on loan applications including features like applicant's income, loan amount, credit history, and more.

### Data Preprocessing:

* **Handling Missing Values:**
  + Dropped columns with more than 65% null values.
  + Filled numeric columns with the mean if null percentage was more than or equal to 7%, otherwise used KNN Imputer.
  + Filled non-numeric columns with the mode.
* **Encoding:**
  + Encoded non-numeric columns using appropriate encoding techniques (e.g., one-hot encoding).
* **Feature Selection:**
  + Identified and retained columns with a correlation greater than 0.1 and less than 1, and dropped the rest.
  + PCA was tested but only had bad results, even after changing the number of features retained to various ranges

### Model Selection:

The following machine learning models were selected for evaluation:

* XGB Classifier
* Gradient Boosting
* Decision Trees
* Random Forest
* K Neighbors
* Logistic Regression
* MLP Classifier

### Evaluation Metrics:

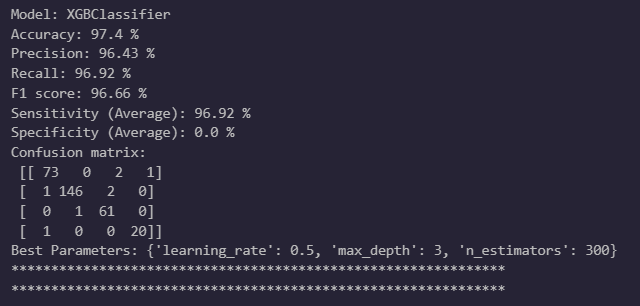
The models were evaluated using the following metrics:

* Accuracy
* Precision
* Recall
* F1 Score
* Sensitivity
* Specificity
* Confusion Matrix

Results:

Model Performance:

Note: All performance metrics were evaluated using K-Fold cross-validation and optimized through GridSearchCV

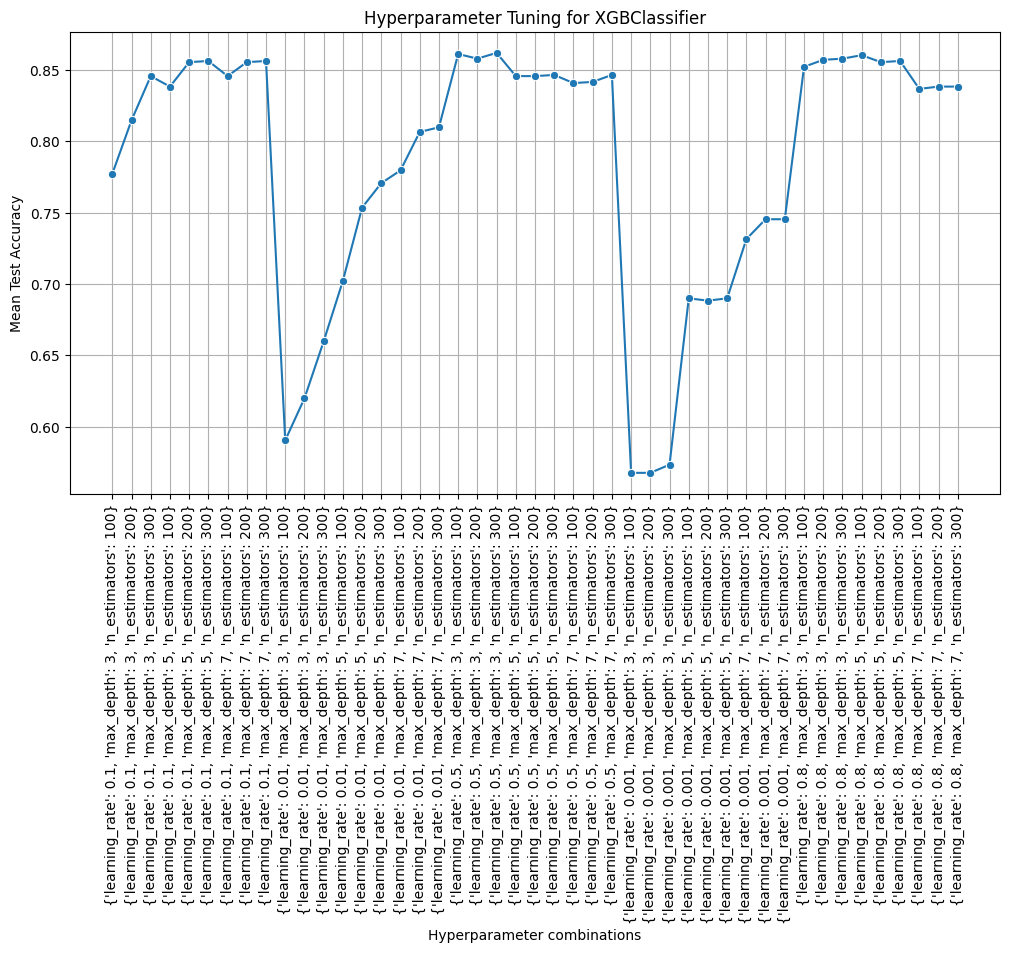
XGB Classifier Performance Metrics:

Accuracy: 97.4%

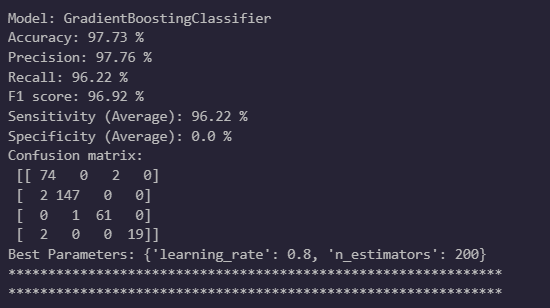
Precision: 96.43%

Recall: 96.92%

F1 Score: 96.66%

Sensitivity (True Positive Rate): 96.92%

Gradient Boosting Classifier Performance Metrics:



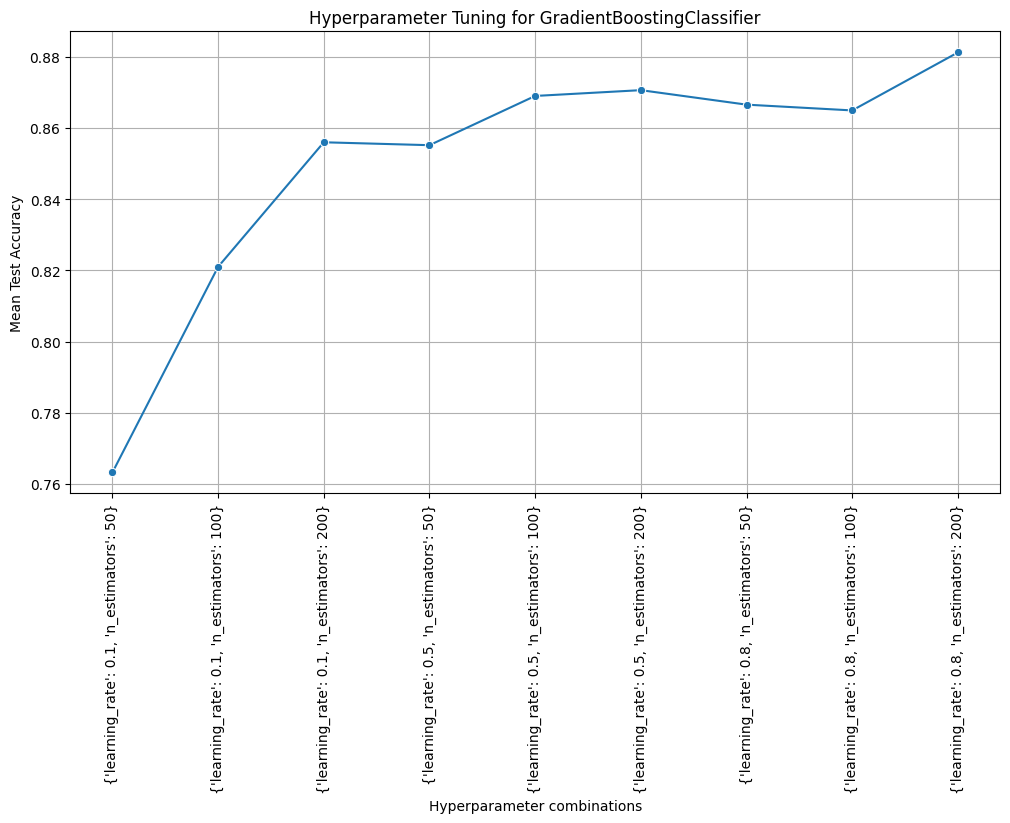
Accuracy: 97.73%

Precision: 97.76%

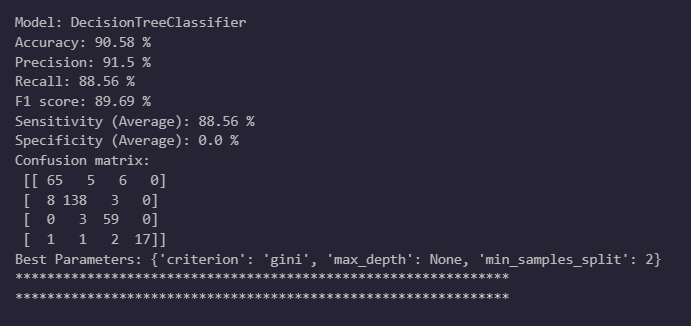
Recall: 96.22%

F1 Score: 96.92%

Sensitivity (True Positive Rate): 96.22%



Decision Tree Classifier Performance Metrics:



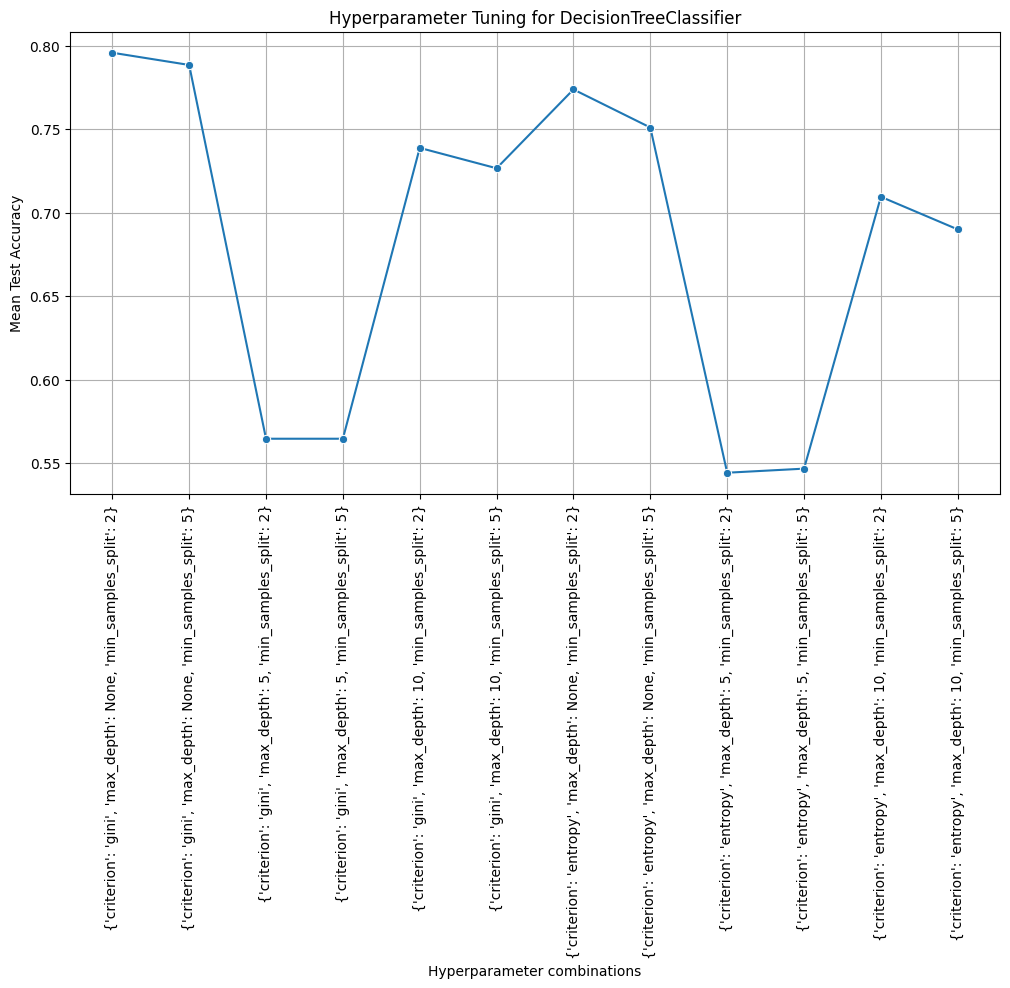
Accuracy: 90.58%

Precision: 91.5%

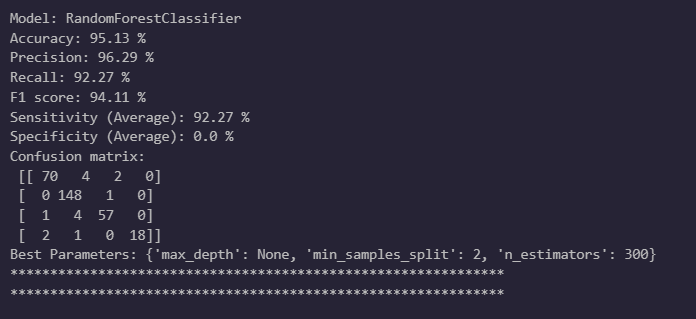
Recall: 88.56%

F1 Score: 89.69%

Sensitivity (True Positive Rate): 88.56%



Random Forest Classifier Performance Metrics:



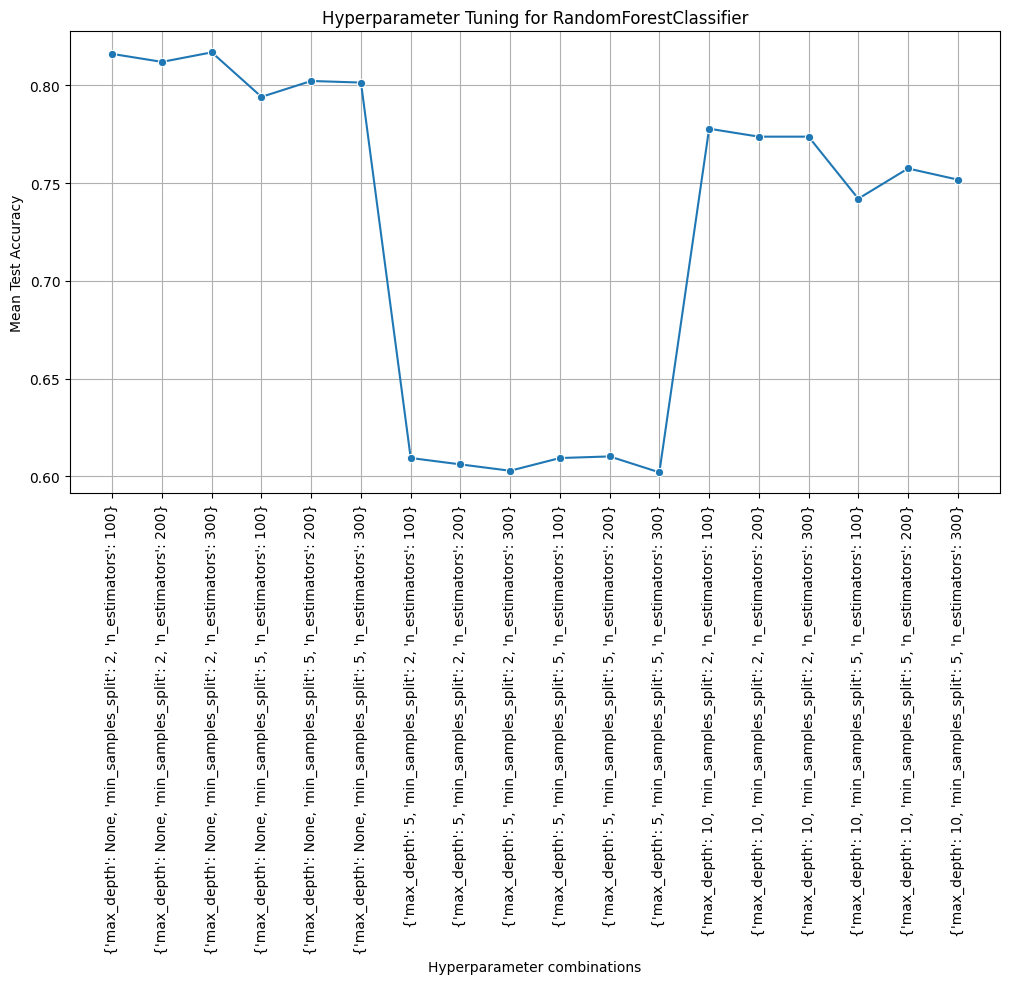
Accuracy: 95.13%

Precision: 96.29%

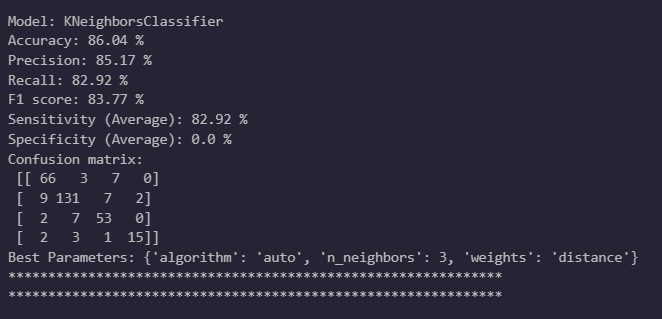
Recall: 92.27%

F1 Score: 94.11%

Sensitivity (True Positive Rate): 92.27%



K Neighbors Classifier Performance Metrics:



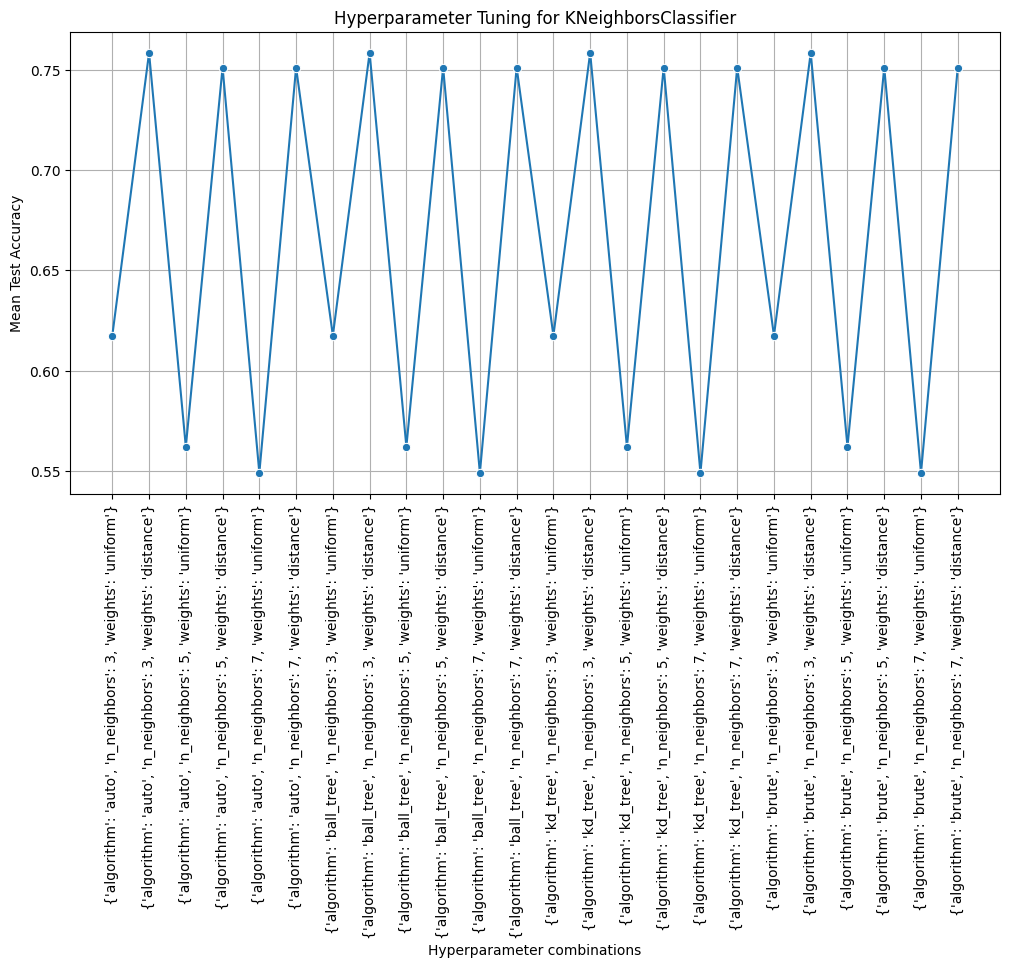
Accuracy: 86.04%

Precision: 85.17%

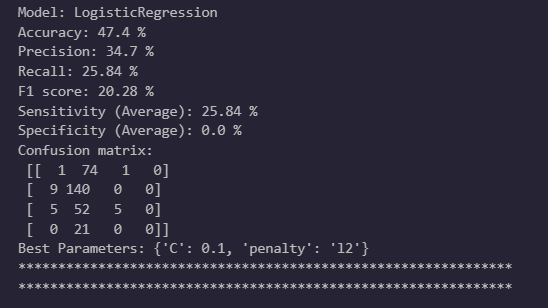
Recall: 82.92%

F1 Score: 83.77%

Sensitivity (True Positive Rate): 82.92%



K Neighbors Classifier Performance Metrics:



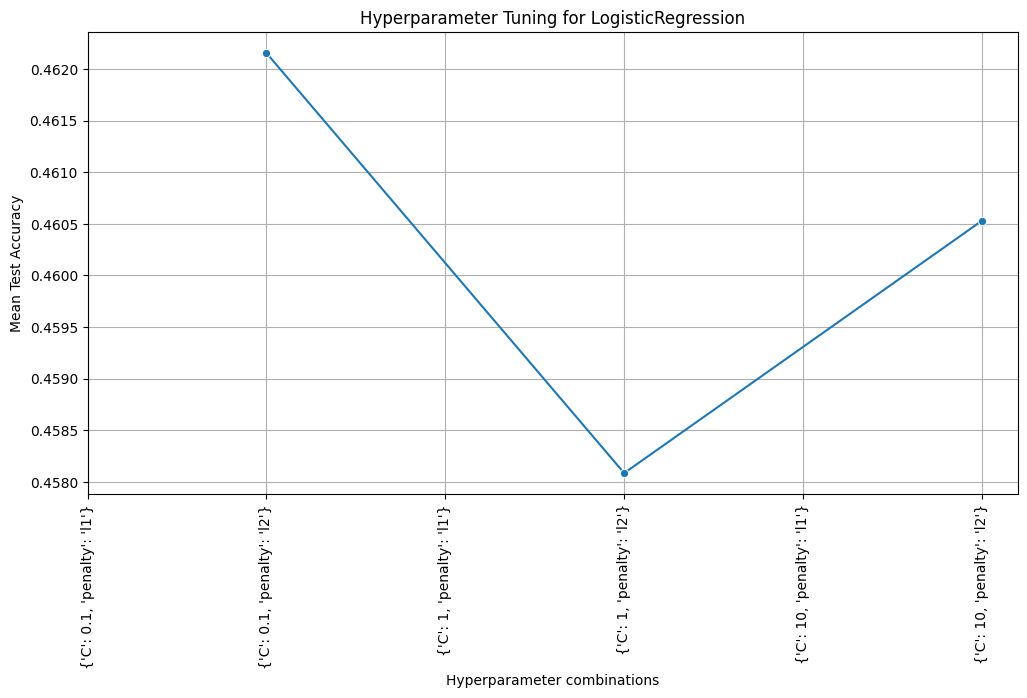
Accuracy: 47.4%

Precision: 34.7%

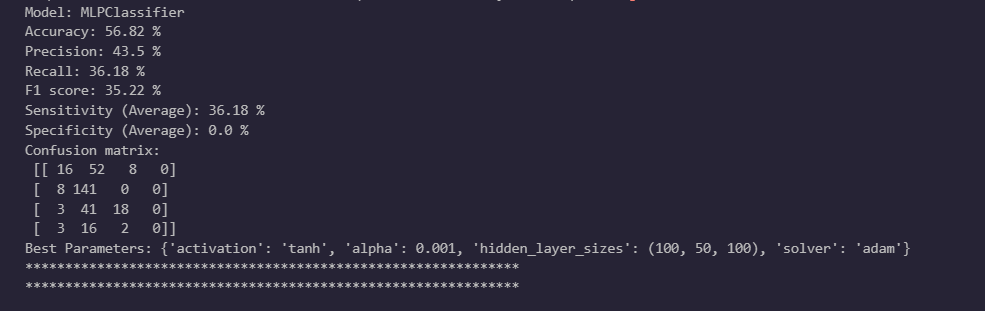
Recall: 25.84%

F1 Score: 20.28%

Sensitivity (True Positive Rate): 25.84%



MLP Classifier Performance Metrics:



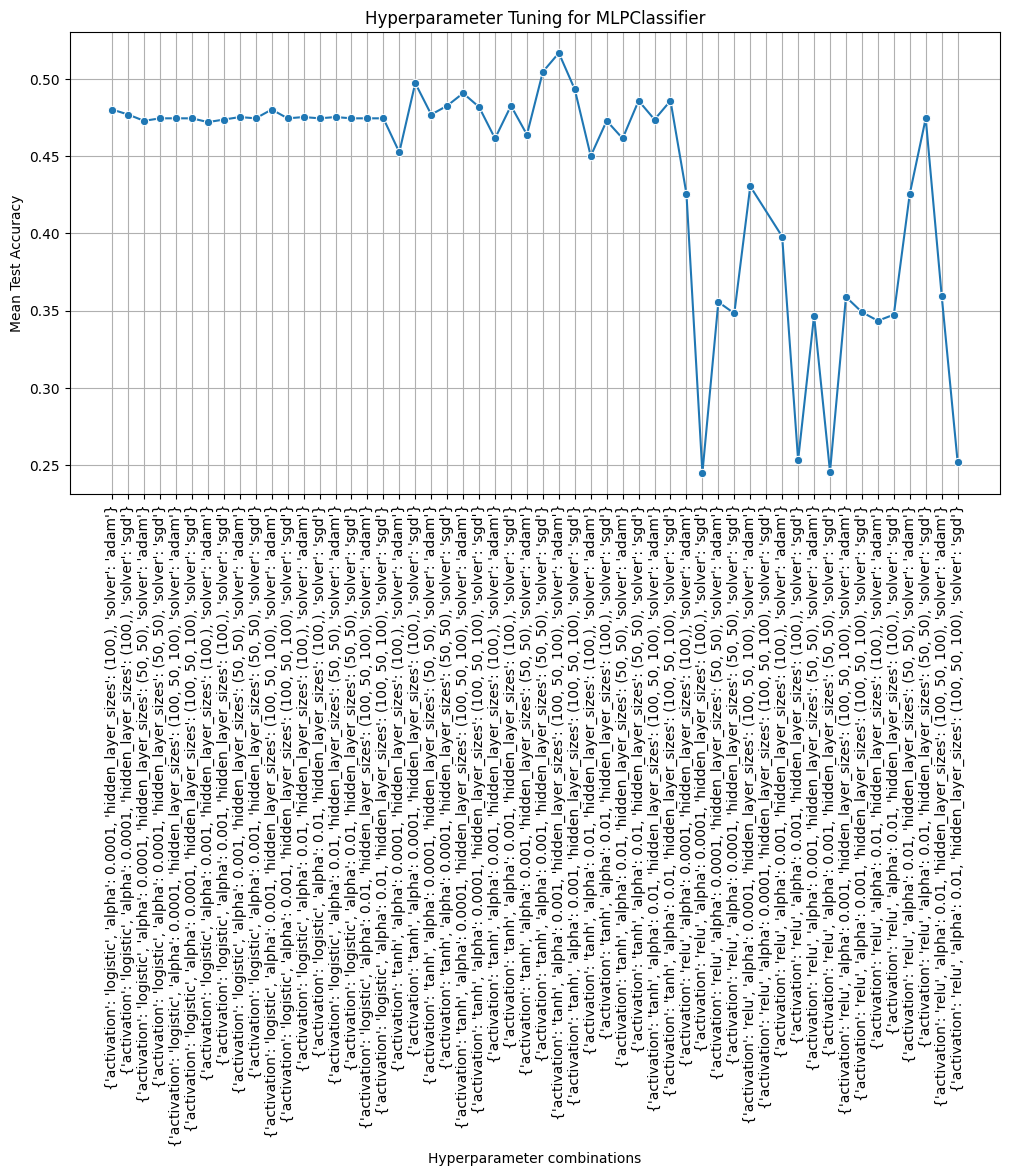
Accuracy: 56.82%

Precision: 43.5%

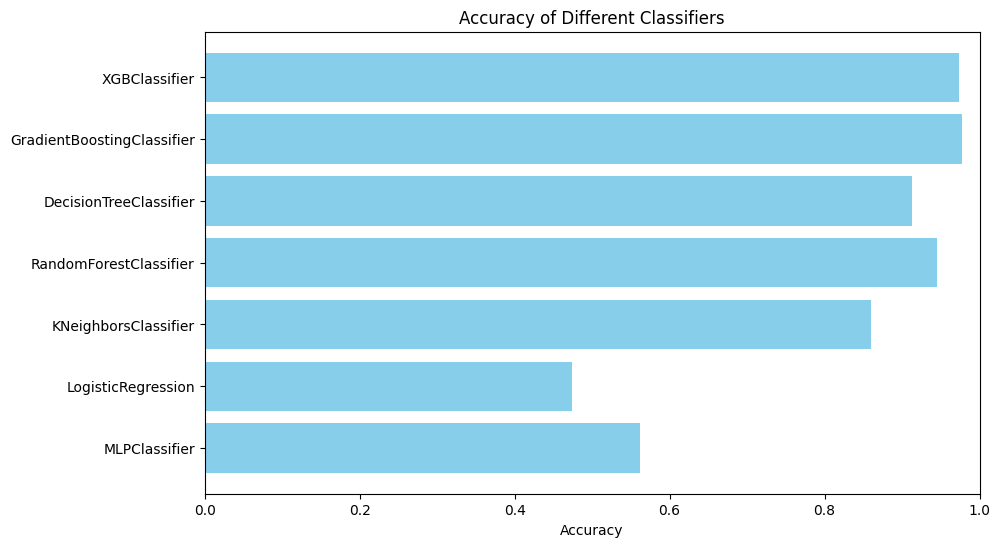
Recall: 36.18%

F1 Score: 35.22%

Sensitivity (True Positive Rate): 36.18%



Accuracy of All Classifiers:



Conclusion:

In summary, Gradient Boosting stood out as the top performer, achieving an impressive mean accuracy of 97.73%. With specific data combinations, it even reached 98.4% accuracy. The XGB Classifier followed closely with 97.4% accuracy. However, the KNN model trailed behind with only 47.4% accuracy.