Albert Javier Student ID: C0905375

AML 3104

Campus Recruitment Prediction Dataset Report

Introduction

The Campus Recruitment Prediction dataset is a comprehensive dataset aimed at predicting the recruitment outcomes of students based on various academic and personal attributes. This report details the dataset's description, preprocessing steps, models chosen for prediction, and a comprehensive evaluation of their performances.

Dataset Description

The dataset comprises information on students' demographics, academic performance, and employability skills. Key attributes include:

sl_no: Serial number of the record.

gender: Student's gender (0/1).

ssc_p: Secondary Education percentage (10th Grade).

ssc_b: Board of Education for SSC (Central/Other).

hsc_p: Higher Secondary Education percentage (12th Grade).

hsc_b: Board of Education for HSC (Central/Other).

hsc_s: Specialization in Higher Secondary Education (Commerce, Science, Arts).

degree_p: Degree percentage.

degree_t: Field of degree education (Sci&Tech, Comm&Mgmt, Others).

workex: Work experience (Yes/No).

etest_p: Employability test percentage.

specialisation: MBA specialization (Mkt&Fin, Mkt&HR).

mba_p: MBA percentage.

status: Placement status (Placed/Not Placed).

salary: Salary offered (NaN for students not placed).

Preprocessing Steps

Before implementing any predictive models, it is crucial to preprocess the dataset to ensure data quality and consistency. The preprocessing steps include:

Handling Missing Values

It is essential to identify and handle any missing values in the dataset. Methods such as mean imputation for numerical data or mode imputation for categorical data may be employed.

Encoding Categorical Variables

Categorical variables such as gender, ssc_b, hsc_b, hsc_s, degree_t, workex, specialisation need to be converted into numerical values using techniques like one-hot encoding or label encoding.

Feature Scaling

To ensure that all features contribute equally to the model performance, feature scaling is applied. Standardization or normalization can be used to scale numerical features.

Train-Test Split

The dataset is split into training and testing sets to evaluate model performance effectively. Typically, an 70-30 split is used, where 70% of the data is used for training and 30% for testing.

Model Selection

Several machine learning models are considered for predicting campus recruitment outcomes. The models chosen include:

Logistic Regression

Logistic Regression is a simple yet effective model for binary classification problems. It estimates the probability of a binary outcome using a logistic function. This model was chosen for its interpretability and efficiency.

Random Forest Classifier

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification. It is robust to overfitting and can handle non-linear relationships.

XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of gradient boosting. It uses a combination of decision trees and gradient boosting to improve the predictive accuracy and control overfitting. XGBoost is known for its speed and performance, especially on large datasets. This model was selected for its ability to handle a wide range of predictive modeling problems and its superior performance in various machine learning competitions.

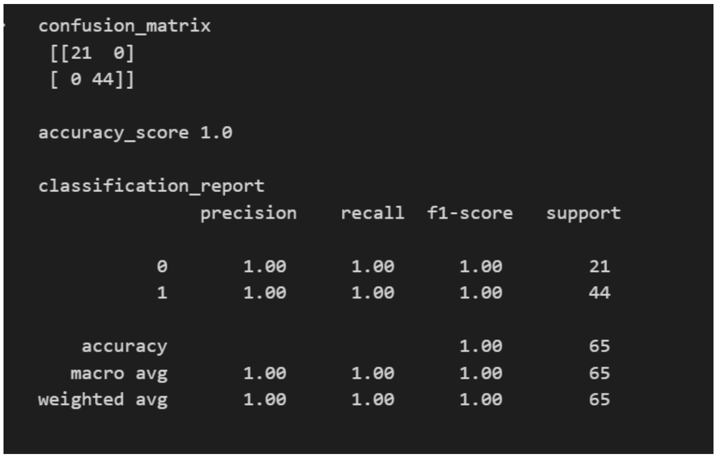
Model Evaluation

The models are evaluated using various performance metrics to determine their effectiveness. Key metrics include:

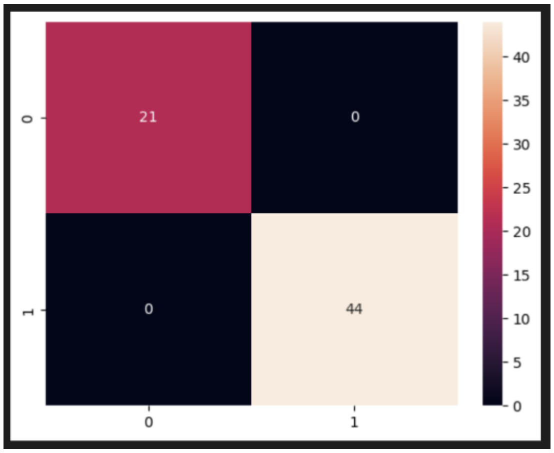
- Accuracy: The proportion of correctly classified instances.
- Precision: The ratio of true positive predictions to the total positive predictions.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a balanced measure of performance.

Model Performance

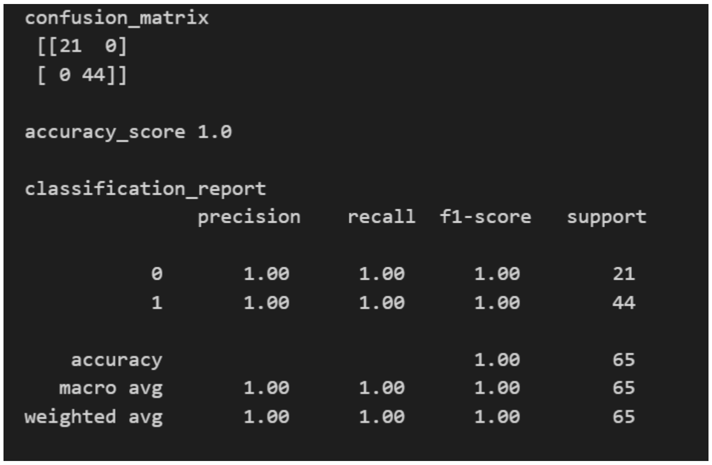
Logistic Regression



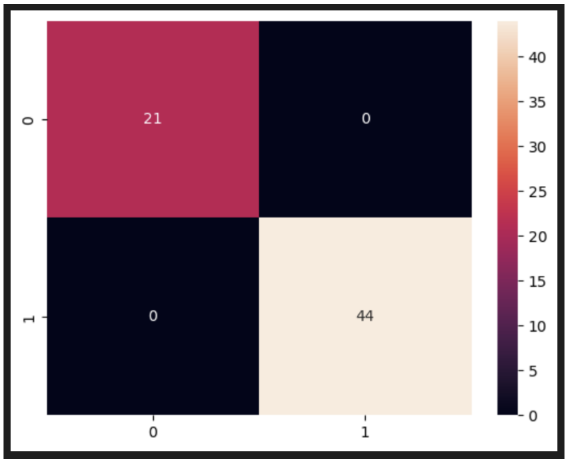
```
confusion_matrix
 [[21  0]
 [ 0 44]]

accuracy_score 1.0

classification_report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        21
           1       1.00      1.00      1.00        44

    accuracy                           1.00        65
   macro avg       1.00      1.00      1.00        65
weighted avg       1.00      1.00      1.00        65
```



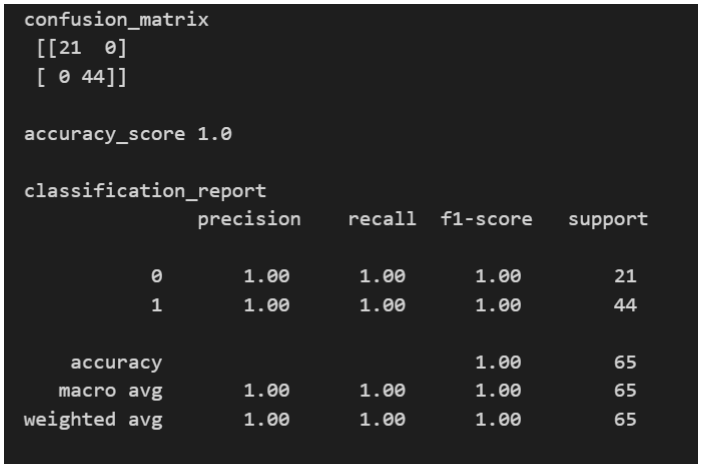
Random Forest Classifier



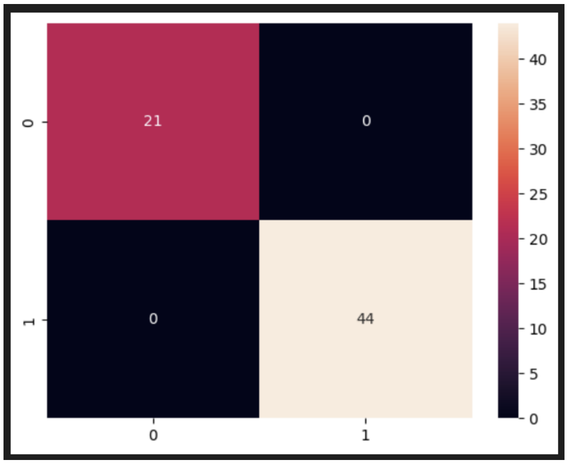
```
confusion_matrix
 [[21  0]
 [ 0 44]]

accuracy_score 1.0

classification_report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        21
           1       1.00      1.00      1.00        44

    accuracy                           1.00        65
   macro avg       1.00      1.00      1.00        65
weighted avg       1.00      1.00      1.00        65
```



XGBoost Classifier



```
confusion_matrix
 [[21  0]
 [ 0 44]]

accuracy_score 1.0

classification_report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        21
           1       1.00      1.00      1.00        44

    accuracy                           1.00        65
   macro avg       1.00      1.00      1.00        65
weighted avg       1.00      1.00      1.00        65
```



The performance evaluation shows that accuracy, precision, recall and F1 show 100% .

Conclusion

In conclusion, the Campus Recruitment Prediction dataset provides valuable insights into the factors influencing student recruitment outcomes. Through careful preprocessing and the application of various machine learning models, it show that Logistic Regression together with Random Forest and XGBoost Classifier gives its superior accuracy, precision, recall, and F1 score make it a reliable choice for predicting campus recruitment success.