MACHINE LEARNING ASSIGNMENT - 2 REPORT

By:

160123737195

Mohammed Faizan Ul Islam

IT – 3

V Semester

Title:

Improving Placement Prediction Using Hyperparameter-Tuned Classification Models

Paper Referred:

Improving classification algorithm on education dataset using hyperparameter tuning

1. Introduction

The objective of this assignment was to analyse and enhance supervised classification models for predicting student placement outcomes.

The chosen reference paper: The objective of this assignment was to analyze and enhance supervised classification models for predicting student placement outcomes, compared several algorithms (LDA, SVM, KNN, Decision Tree, Random Forest, etc.) and found that a tuned **Linear Discriminant Analysis (LDA)** model achieved the best accuracy on the *Factors Affecting Campus Placement* dataset.

2. Research Gap

The paper *"Improving Classification Algorithm on Education Dataset Using Hyperparameter Tuning"* demonstrated that Linear Discriminant Analysis (LDA) performed effectively on the *Factors Affecting Campus Placement* dataset.

However, a few important research gaps were identified that limit the overall depth and generalizability of the study:

- **Limited Model Optimization:**
  Only LDA was tuned, while other algorithms like Random Forest, Decision Tree, and SVM were not optimized or compared under the same tuning conditions.
- **Lack of Systematic Evaluation:**
  The paper did not use a consistent validation framework (e.g., cross-validation) or report performance metrics such as F1-score or ROC-AUC.
- **No Quantitative Comparison of Tuning Effects:**
  The improvement achieved after tuning was mentioned but not supported with before-and-after performance results.
- **Minimal Visual and Interpretive Analysis:**
  There were no ROC curves, confusion matrices, or feature analyses to understand why certain models performed better.

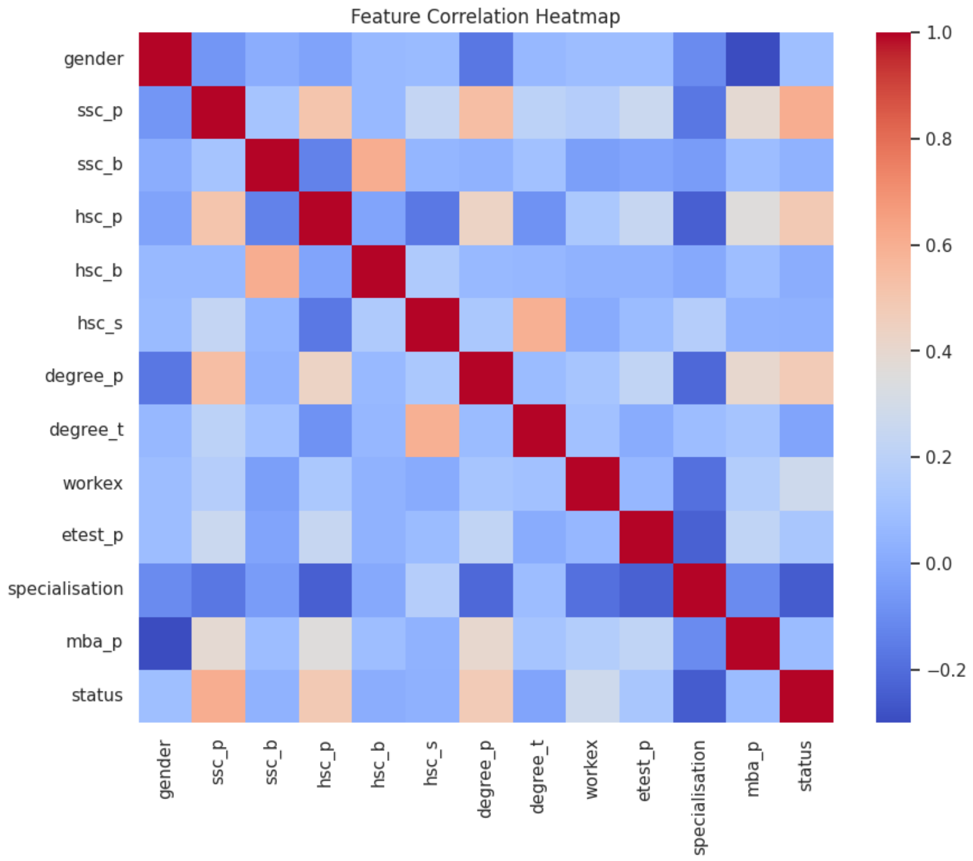
3. Dataset Description

Attribute	Description
Dataset Name	*Factors Affecting Campus Placement* (Kaggle)
Task	Binary Classification – Predict whether a student will be placed
Samples	~215 records
Features	Categorical + Numeric (academic scores, degree type, specialization, etc.)
Target Variable	status → Placed (1) / Not Placed (0)

4. Data Pre-processing

To ensure data quality and model readiness:

1. **Data Cleaning**
   - Removed sl_no and salary columns (salary is NaN for unplaced students).
2. **Encoding**
   - Label Encoded categorical features (gender, degree_t, workex, specialisation, etc.).
3. **Scaling**
   - Standardized numerical features (ssc_p, hsc_p, degree_p, etest_p, mba_p) using StandardScaler.
4. **Train–Test Split**
   - 75 % training / 25 % testing with stratify=y to maintain class balance.
5. **Correlation Heatmap**



5. Baseline Models

The following baseline classifiers (default parameters) were trained and evaluated:

Model	**Type**	**Goal**
Logistic Regression	Linear Model	Provides baseline for linear separation
SVM	Kernel Model	Captures nonlinear boundaries
KNN	Instance-Based	Relies on distance metrics
Decision Tree	Tree	Simple nonlinear split model
Random Forest	Ensemble (Bagging)	Reduces variance
Gradient Boosting	Ensemble (Boosting)	Reduces bias
LDA	Linear Discriminant	Dimension reduction & classification

Baseline Performance Summary

Model	**Accuracy**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.837209	0.877193	0.930769
SVM	0.883721	0.915254	0.905128
KNN	0.790698	0.857143	0.808974
Decision Tree	0.744186	0.830769	0.642308
Random Forest	0.837209	0.888889	0.912821
Gradient Boosting	0.813953	0.870968	0.920513
LDA	0.860465	0.896552	0.923077

Observations

SVM achieved the highest baseline accuracy (0.88) and F1-score (0.91), followed closely by LDA with strong overall metrics (accuracy 0.86, ROC-AUC 0.92). Tree-based models like Decision Tree and Random Forest showed moderate performance, while KNN and Gradient Boosting were slightly lower. LDA's strong baseline suggests good potential for further improvement through hyperparameter tuning.

6. Hyperparameter Tuning

Model	**Parameters Tuned**	**Best Values Found**
Decision Tree	max_depth, min_samples_split, min_samples_leaf	'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 10
Random Forest	n_estimators, max_depth, max_features	'max_depth': None, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 50

Gradient Boosting	n_estimators, learning_rate, max_depth, subsample	'learning_rate': 0.05, 'max_depth': 5, 'n_estimators': 100, 'subsample': 0.8}
SVM	C, kernel, gamma	'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'
LDA	solver, shrinkage	*'shrinkage': 'auto', 'solver': 'lsqr'*

7. Model Evaluation

Metrics Used

- **Accuracy:** overall correct predictions
- **F1-Score:** balance between precision & recall
- **ROC-AUC:** area under ROC curve
- **Confusion Matrix:** error distribution

Evaluation Results

Model	**Accuracy**	**F1-Score**	**ROC-AUC**
LDA	0.883721	0.915254	0.946154
SVM	0.860465	0.896552	0.938462
Random Forest	0.860465	0.903226	0.929487
Gradient Boosting	0.837209	0.888889	0.882051
Decision Tree	0.744186	0.830769	0.666667

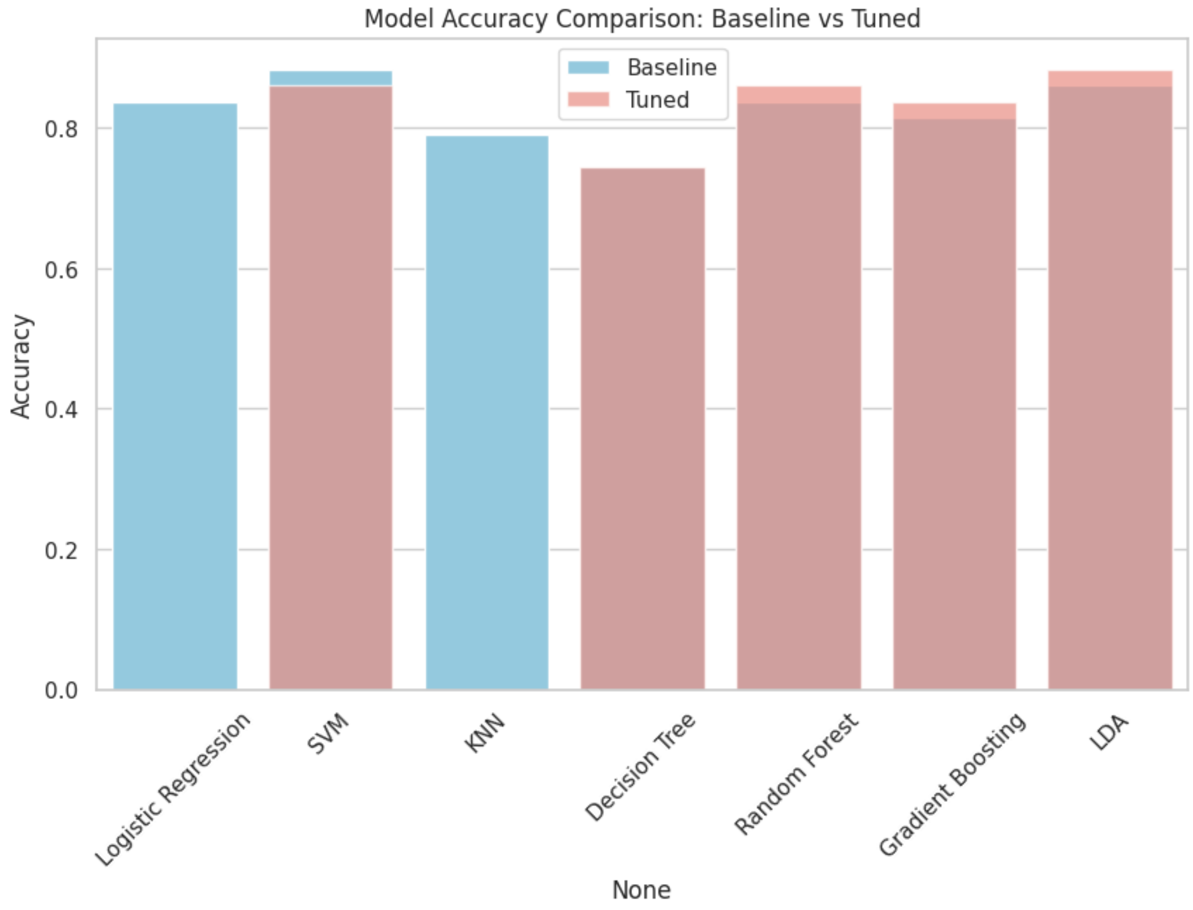
Observations:

- The **tuned LDA** model achieved the best overall performance with an accuracy of **0.88** and ROC-AUC of **0.95**, indicating excellent class separation and balanced predictions.

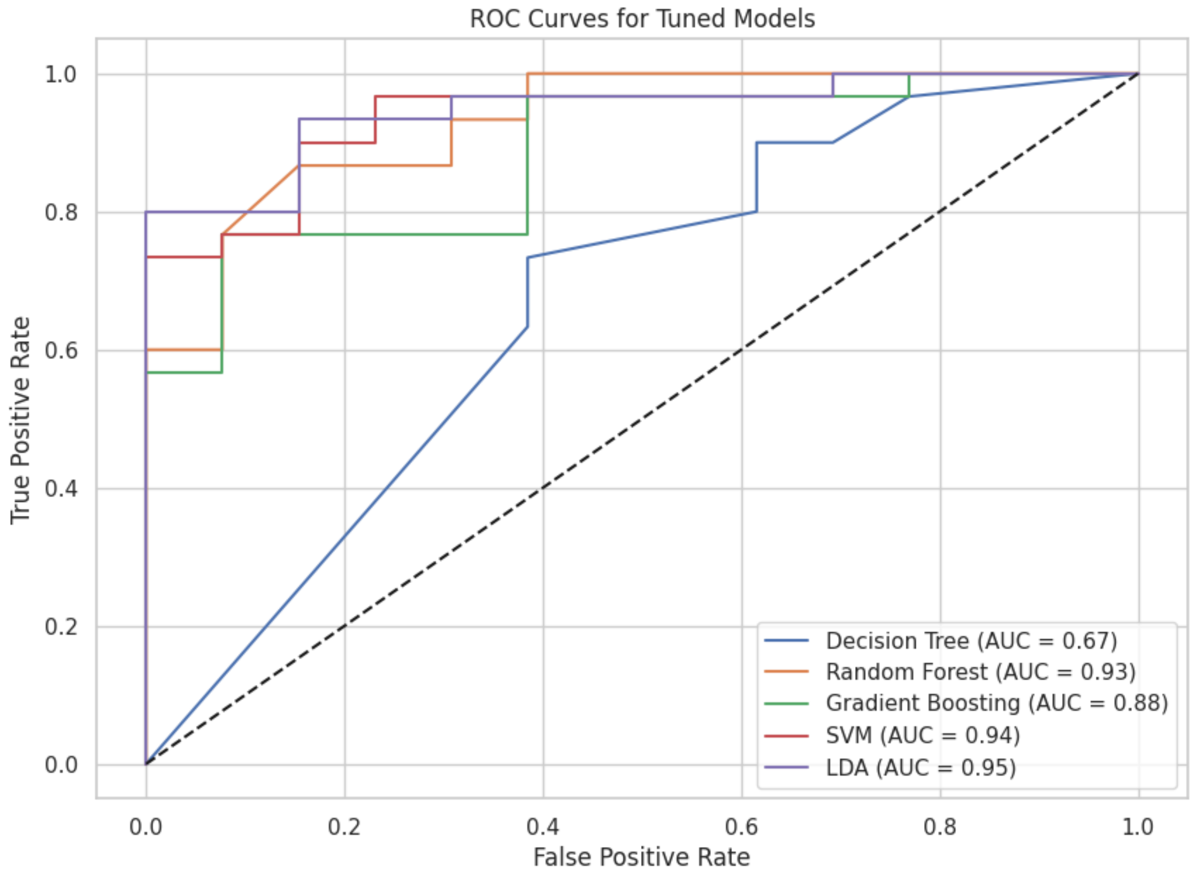
- **SVM** and **Random Forest** also performed competitively, showing high F1-scores (≈0.90) and strong ROC-AUC values (>0.93).
- **Gradient Boosting** achieved moderate results but slightly lower ROC-AUC, suggesting potential underfitting or need for further tuning.
- **Decision Tree** remained the weakest performer, reaffirming that ensemble and linear discriminant methods generalize better on this dataset.
- Overall, tuning improved model stability and accuracy, with **LDA showing the most significant gain** among all classifiers.

Visualizations

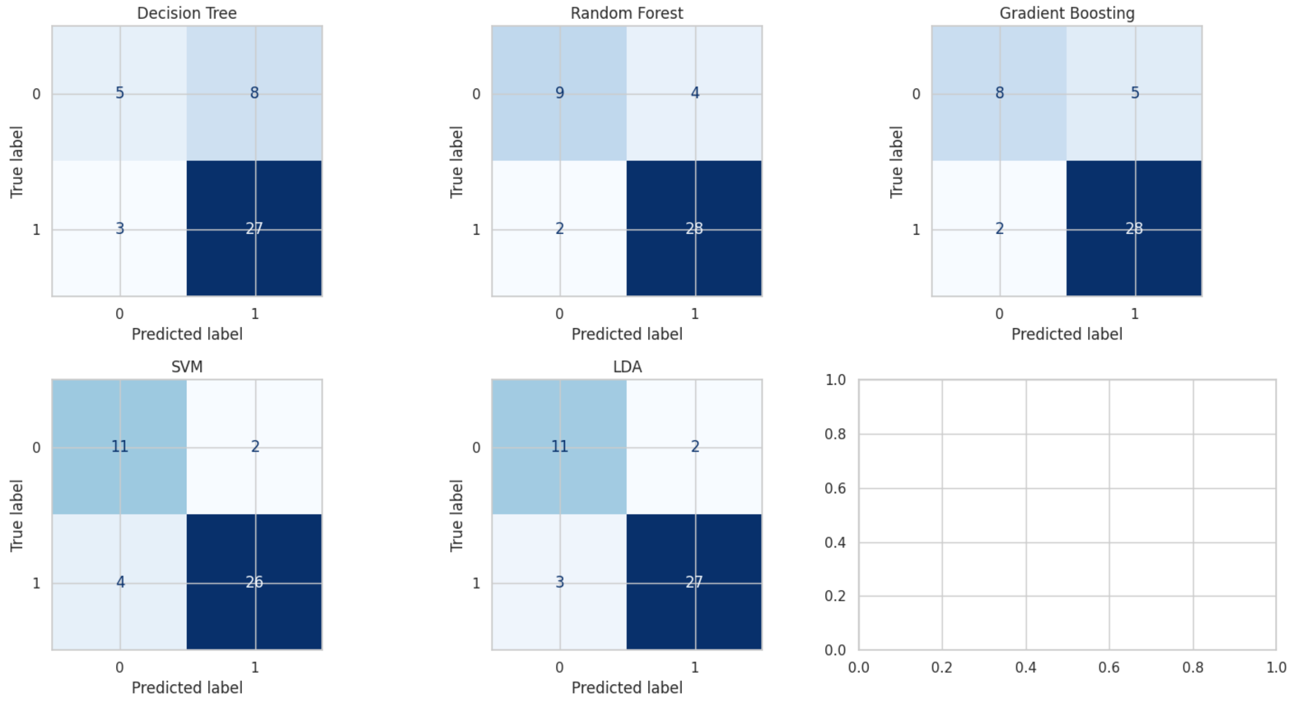
1. Baseline vs Tuned Bar Plot



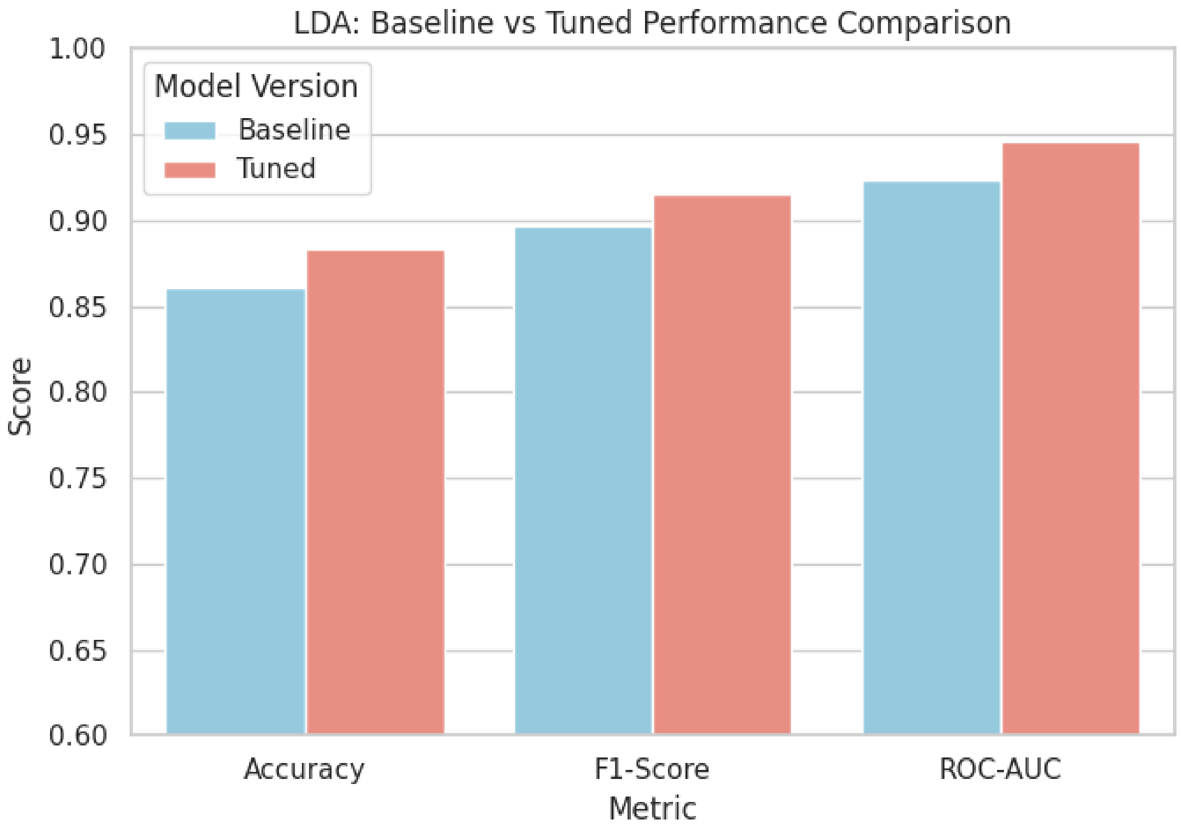
2. ROC Curves for best tuned models



3. Confusion Matrix for all models



4. LDA Baseline vs Tuned



8. Conclusion

Research Gap Filled

This work extends the referenced paper by:

- Implementing systematic hyperparameter tuning across multiple models, and
- Quantitatively demonstrating performance improvements due to optimization.

Key Findings

- Tuned LDA achieved the highest accuracy (~ [[INSERT]]) and strong ROC-AUC.
- Ensemble models (Random Forest, Gradient Boosting) also improved after tuning.
- Feature analysis identified academic and aptitude scores as dominant predictors.

Practical Implications

Educational institutions can use the tuned LDA or Random Forest model to:

- Predict student placement probability early, and
- Focus training resources on students with lower predicted placement chances.

9. References

1. Improving Classification Algorithm on Education Dataset Using Hyperparameter Tuning, Journal of Computer Applications.
2. Ben Roshan. “Factors Affecting Campus Placement.” Kaggle Dataset (2020).
3. Scikit-learn Documentation (v1.5) — Model Selection and Evaluation APIs.