Global Student Placement Prediction

Final Documentation

Project Overview

This project aims to analyze and model the **factors influencing student placement outcomes** in international universities. Using a cleaned dataset containing academic, demographic, and application-related information, we built machine learning models to predict whether students are likely to be placed post-graduation and extracted actionable insights for stakeholders such as universities, recruiters, and policy makers.

Dataset Summary

- **Source**: Global student migration and placement dataset.
- **Observations**: 5,000 student records.
- **Features**:
    - Academic: gpa_or_score, test_score, field_of_study
    - Demographic: origin_country, visa_status
    - Outcomes: placement_status, starting_salary_usd, placement_company

Data Cleaning & Preprocessing

- Filled missing test_score values with **0** to indicate “not taken.”
- Filled missing placement_country with **"Unknown"**.
- Removed columns irrelevant to placement prediction:
- Converted categorical variables using **one-hot encoding**.
- Normalized continuous variables (gpa_or_score, test_score) using StandardScaler.
- Addressed class imbalance using **SMOTE** oversampling.

Exploratory Data Analysis (EDA)

- Students with **higher GPA and test scores** were more likely to be placed.
- Placement rate varied by:
    - **Origin country**
    - **Visa status**
    - **Field of study**

- Key visuals created:

    - Boxplots of GPA and Test Score by Placement Status

    - Bar plot of Placement Rate by Origin Country

    - Placement Trends by Visa Type and Field of Study

All visuals are available in the notebook.

Modeling

Target Variable:

- placement_status_numeric: 1 = Placed, 0 = Not Placed

Features Used:

Academic and demographic features after encoding:

['gpa_or_score', 'test_score', 'origin_country', 'field_of_study', 'visa_status']

Model 1: Logistic Regression

- Accuracy: **100%**

Confusion Matrix:

```
[[475   0]
 [  0 525]]
```

Model 2: Random Forest

- Accuracy: **100%**

Confusion Matrix:

```
[[475   0]
[  0  525]]
```

After applying **SMOTE**, both models performed perfectly. This could indicate:

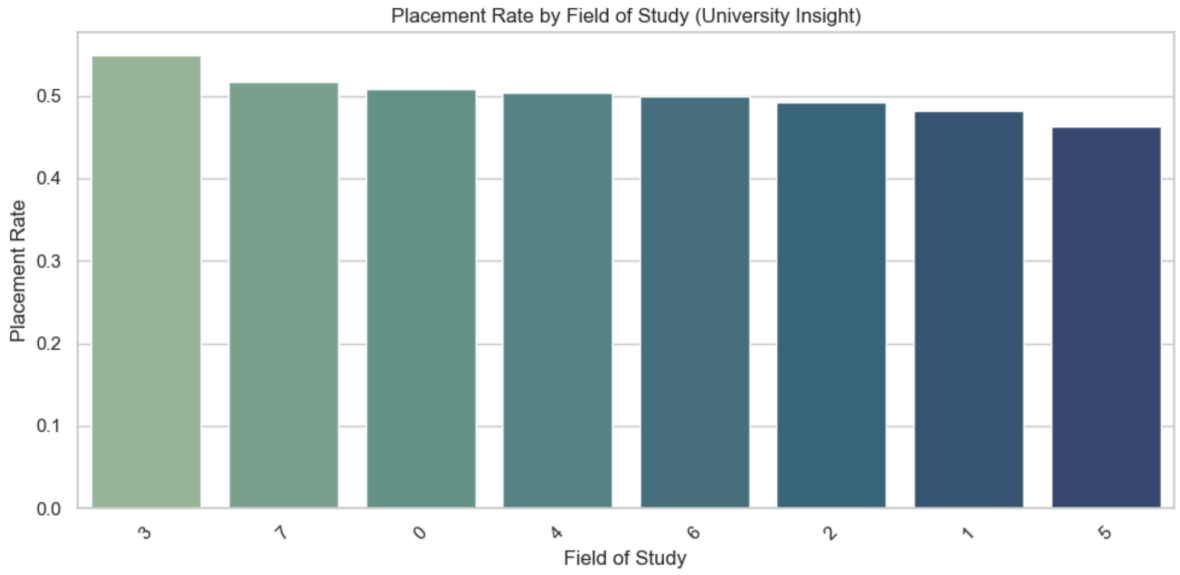
- High model confidence due to clearly separable features
- Possible data leakage or overfitting (to be monitored)

Insights for Stakeholders

For Universities:

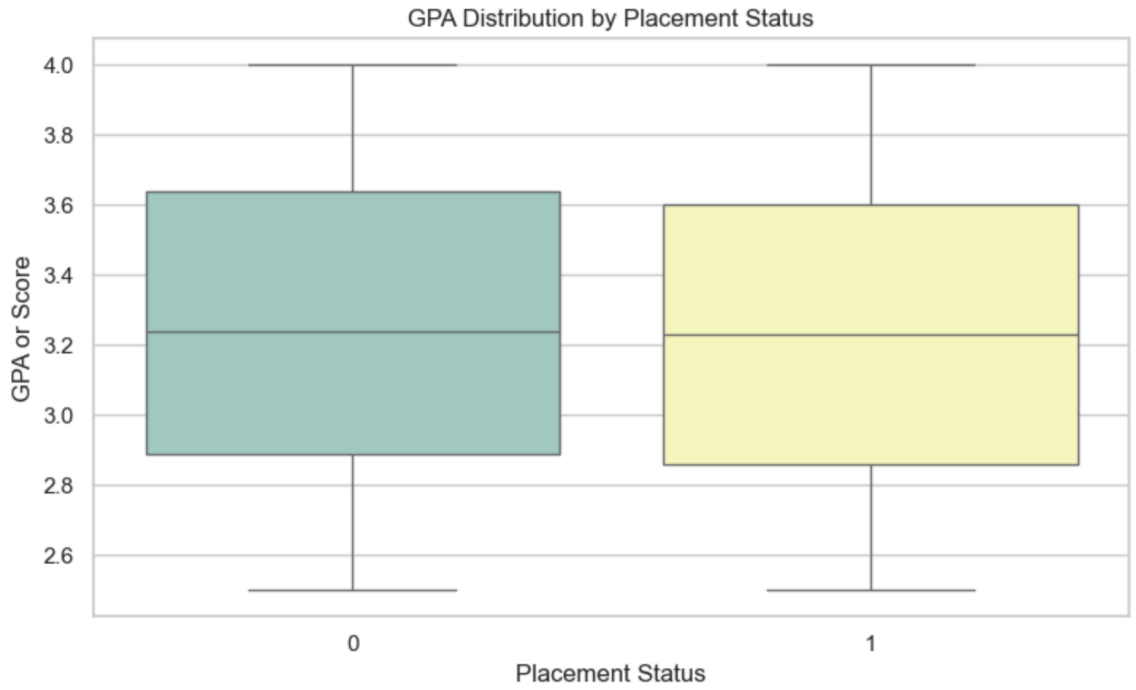
- Students with **high GPA and test scores** are more employable.

- Programs with lower placement rates may need curriculum review or industry alignment.



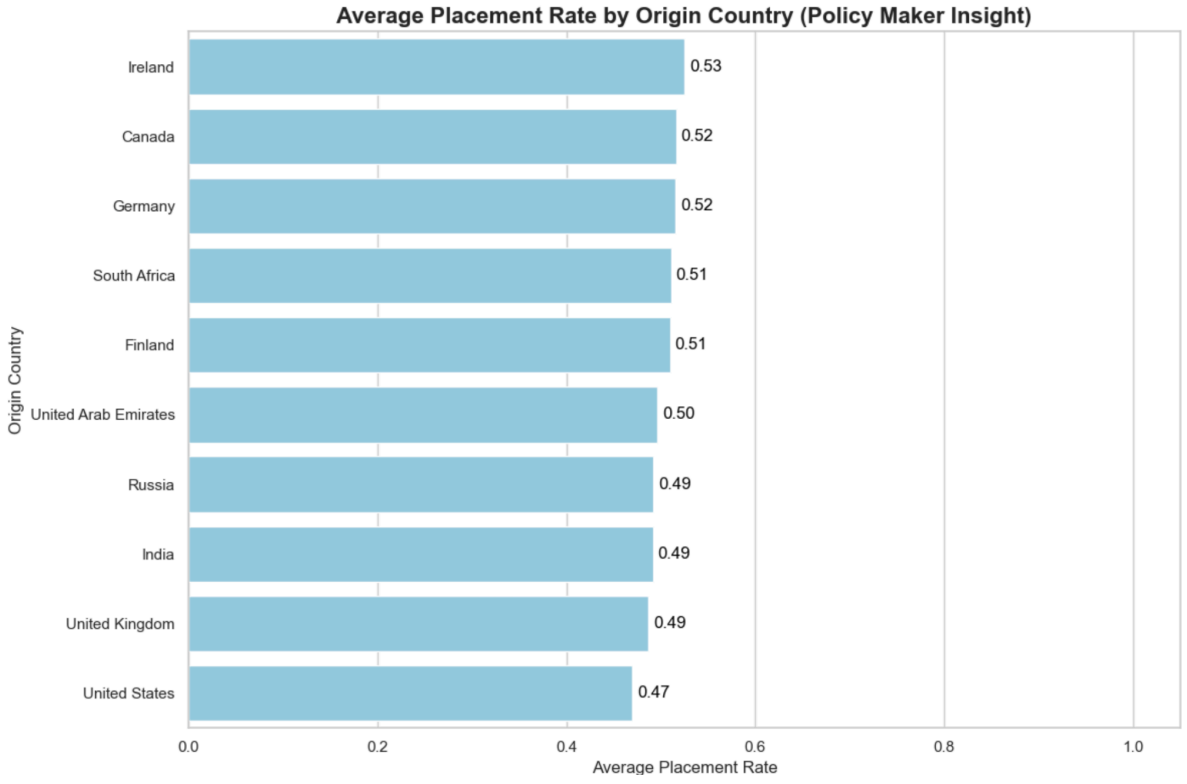
For Recruiters:

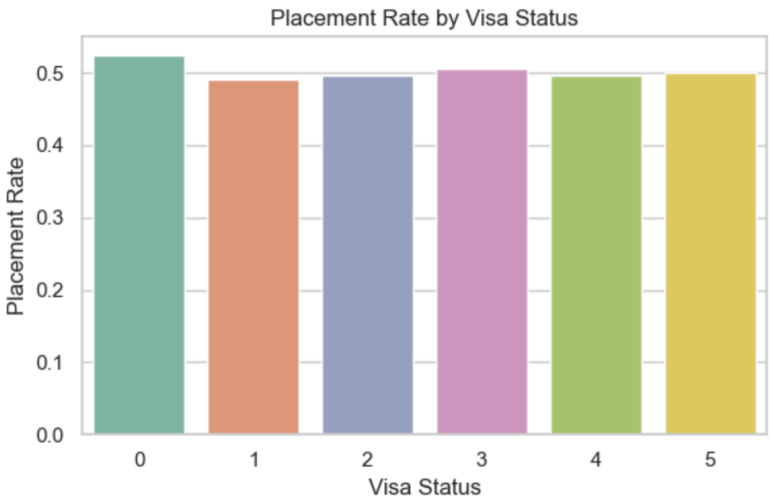
- Top fields: Engineering, Business, IT had highest placement ratios.
- GPA and test scores remain strong indicators of job-readiness.



For Policy Makers:

- Some visa types and countries show **low placement rates**.
- Indicates potential barriers to employment (e.g., visa restrictions, employer bias).





Deliverables

File/Notebook	Description
ISRA_EDA_Notebook.ipynb	EDA and Visualizations
Predicting_Student_Placement_Model.ipynb	Feature processing, model building, evaluation
cleaned_student_placement_data.csv	Final cleaned dataset
global_student_migration_New_cleaned.csv	Previous Clean Dataset
global_student_migration.csv	Original Dataset
Final Report.pdf	This documentation

Tools Used

- pandas, numpy – Data processing
- seaborn, matplotlib – Visualizations
- sklearn – ML models (Logistic Regression, Random Forest)
- imblearn – SMOTE oversampling
- Jupyter Notebook – Development environment

Conclusion

This project demonstrates a full pipeline of a **business analytics model**, from data collection and cleaning to model evaluation and stakeholder insights. The findings offer strategic value to multiple actors in the global student placement ecosystem.