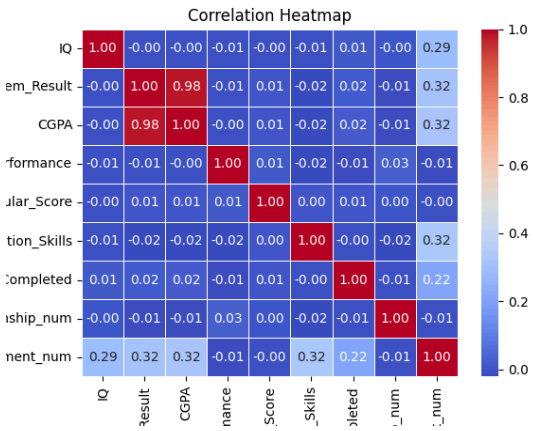
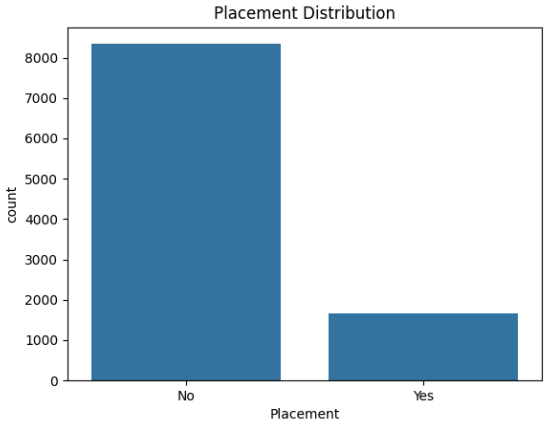
Graduate Employability Analysis Report

Dataset: Grad Employment - Kaggle
Number of Records: 10,000
Target Variable: Placement (Yes/No)
Author: Gavin Cardeno

1. Executive Summary

This analysis investigates the factors influencing graduate employability using a dataset of 10,000 students, each with academic, behavioral, and experiential attributes. The primary goal is to understand which features most strongly predict employment (placement) outcomes and to evaluate the performance of predictive models such as Logistic Regression and Random Forest.





Key Findings:

- The dataset shows a class imbalance, with approximately **8,500 students unplaced** and **1,500 students placed**.
- **IQ, communication skills, CGPA, and projects completed** are key differentiators between placed and unplaced students.
- **Random Forest** outperformed other models, achieving near-perfect accuracy and highlighting **Communication Skills** and **IQ** as the most important predictors.

2. Dataset Overview

Feature	Type	Description
College_ID	Categorical	Unique student identifier
IQ	Numeric	Student IQ score
Prev_Sem_Result	Numeric	Previous semester grade
CGPA	Numeric	Cumulative grade point average
Academic_Performance	Numeric	Academic evaluation score
Internship_Experience	Categorical	Internship completion (Yes/No)
Extra_Curricular_Score	Numeric	Extracurricular activities score
Communication_Skills	Numeric	Communication skill rating
Projects_Completed	Numeric	Number of projects completed
Placement	Categorical	Placement outcome (Yes/No)

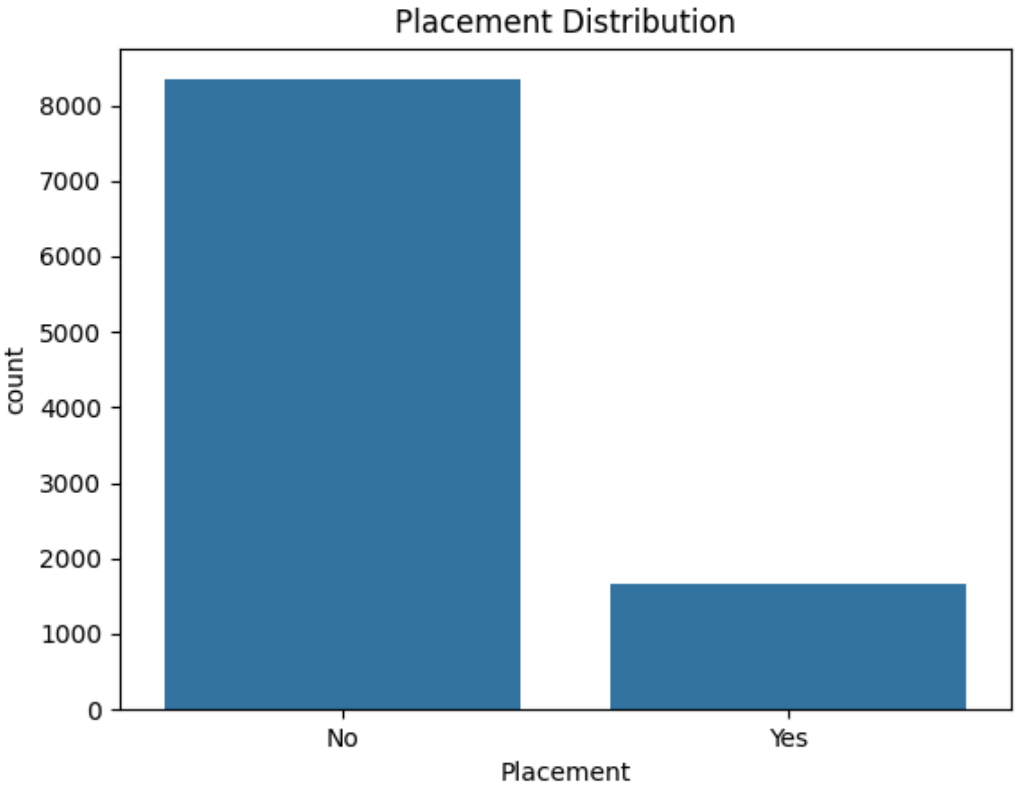
Data Quality Check

- **No missing values** detected (`df.isnull().sum()` returned all zeros).
- Categorical variables were encoded prior to modeling (Categorical to Binary)
- Target variable (`Placement`) is **imbalanced**, with 85% labeled "No" and 15% labeled "Yes".

---

3. Exploratory Data Analysis (EDA)

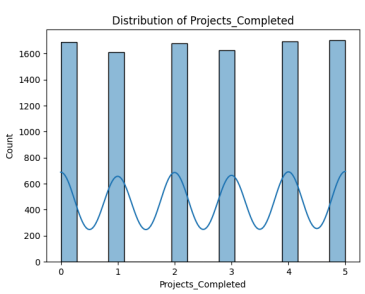
3.1 Distribution

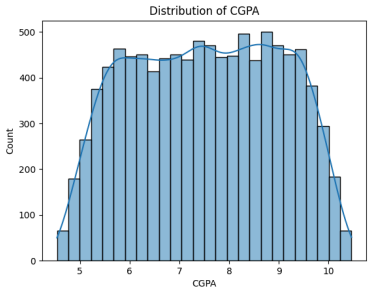


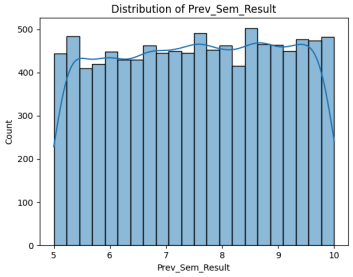
The dataset is heavily skewed toward **non-placement**, with only about **15% of students placed**. This imbalance may impact classification performance, necessitating model evaluation with precision and recall in addition to accuracy.

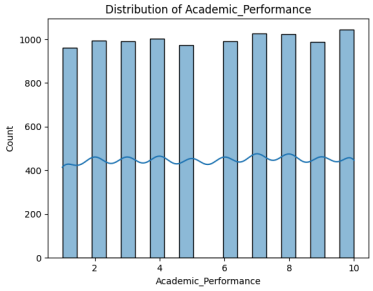
3.2 Numerical Feature Distributions

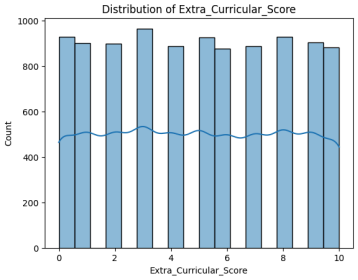
Histograms across numeric variables (e.g., IQ, CGPA, Academic_Performance, Communication_Skills, Projects_Completed) appear **normally distributed and evenly spread**, indicating consistent data quality and no major outliers.

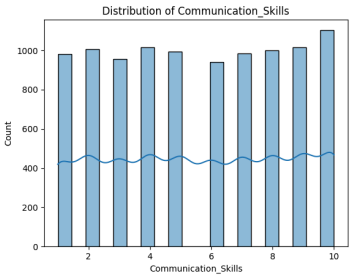


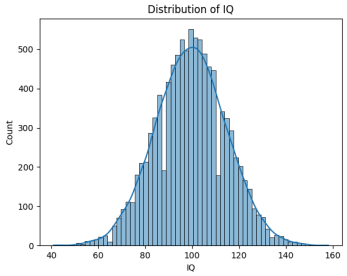












3.3 Feature Relationships with Placement

Boxplots were used to compare numeric features across placement outcomes.
Key observations include:

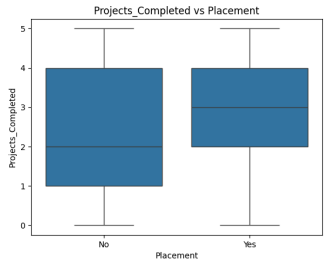
- **IQ:** Students with higher IQ scores were more likely to be placed.

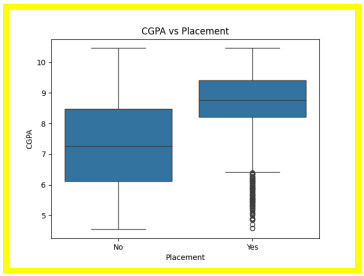
- **Projects Completed:** Placed students tended to complete slightly more projects on average.

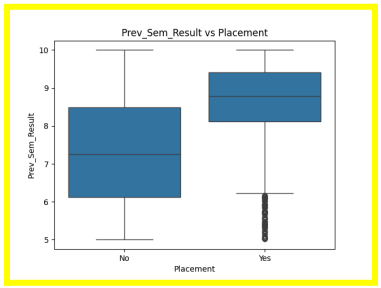
- **Previous Semester Result:** Higher academic results were associated with placement success.
- **Communication Skills:** Significantly higher for placed students, indicating strong soft skills influence employability.

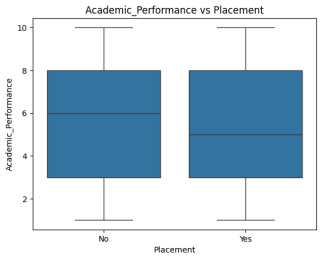
- **CGPA:** Slightly higher averages for placed students, but not as strong a differentiator as IQ or communication.

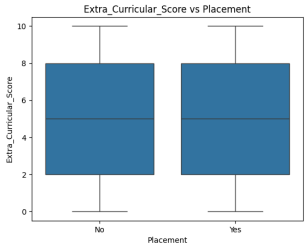
- Other variables appeared evenly distributed across placement outcomes.

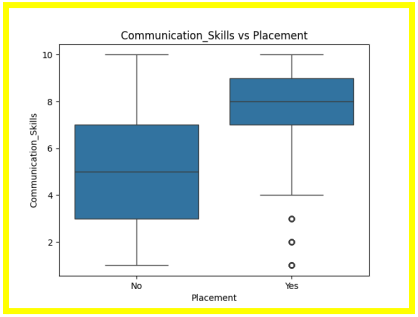


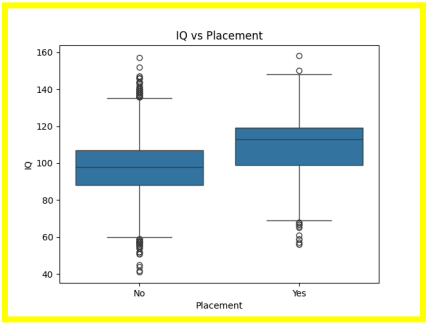












4.0 Data Preparation for Modeling

Before model training, the dataset was divided into training and testing subsets using an 80/20 split. Stratified sampling was applied to maintain proportional representation of the imbalanced placement classes (≈ 8500 “No”, 1500 “Yes”).

To improve model stability and comparability, all numeric features were standardized using `StandardScaler`.

Code Snippet:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(
```

```
    X, y, test_size=0.2, random_state=42, stratify=y
)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

---

5. Predictive Modeling

5.1 Logistic Regression - **Accuracy:** 90.35%

- The model served as a strong baseline, performing well at identifying unplaced students but less effective at detecting placed ones.

Metric	Not Placed	Placed
Precision	0.92	0.78
Recall	0.97	0.59
F1-score	0.94	0.67

Interpretation: Logistic Regression performs adequately for the majority class but underperforms for placed students due to class imbalance.

5.2 K-Nearest Neighbors (KNN) – Accuracy: 94.8%

KNN achieved high overall accuracy and balanced performance across both classes. Data was standardized, ensuring all predictors contributed equally to the distance calculations. However, KNN is computationally heavier on larger datasets and may not generalize well beyond the training data.

Metric	Not Placed	Placed
Precision	0.95	0.91
Recall	0.98	0.77
F1-Score	0.97	0.83

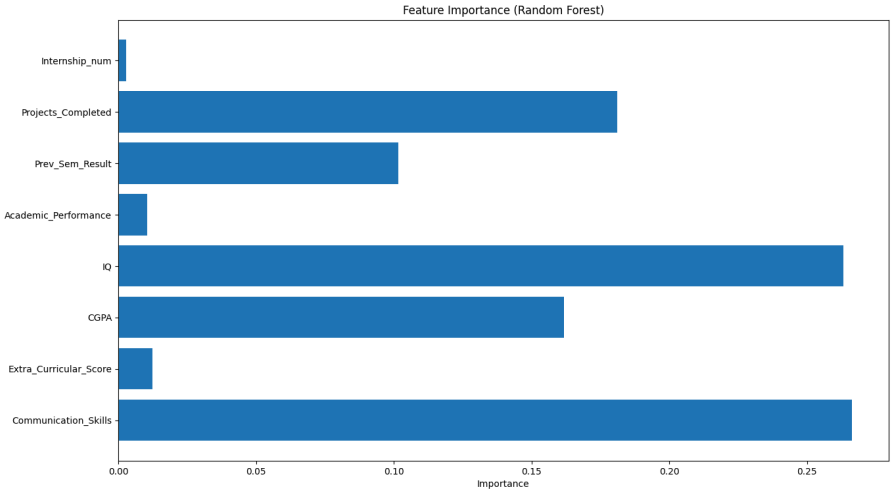
5.2 Random Forest – **Accuracy:** 99.95%

Random Forest delivered highly accurate predictions, outperforming Logistic Regression.

Metric	Not Placed	Placed
Precision	1.00	1.00
Recall	1.00	1.00
F1-score	1.00	1.00

Feature Importance (Top Predictors):

1. **Communication Skills and IQ** – Highest importance
2. **CGPA** – Moderate predictor
3. **Projects Completed** – Slightly above 0.15 in importance



---

6. Conclusions

Strong Predictors:

- **Communication Skills** and **IQ** were the most critical features for predicting placement.
- **CGPA** and **Projects Completed** also contributed meaningfully.

Model Performance:

- **Random Forest** demonstrated superior performance and generalization.
- **Logistic Regression** offered interpretability but struggled with minority class detection.

Actionable Insights:

- Encourage **communication and soft skills development** through workshops and practical training.
- Promote **internship participation** and **project completion** to enhance employability.
- Focus on **academic performance support programs** to maintain high CGPA and semester results.

---

7. Recommendations / Next Steps

- Implement **data balancing techniques** (e.g., SMOTE, class weighting) to improve prediction fairness.
- Incorporate **additional variables** such as certifications, interview performance, and domain specialization.
- Develop a **predictive dashboard** for academic institutions to identify and support at-risk students.
- Explore **ensemble and explainable AI methods** (e.g., SHAP, XGBoost) for more robust and interpretable modeling.

---

8. Appendices / Visuals

- Correlation Heatmap
- Placement Distribution Bar Chart
- Placement by Internship Experience
- Numeric Feature Histograms
- Boxplots (Feature vs. Placement)
- Random Forest Feature Importance Chart