Placement Prediction Report

Name: Bo Yang
Student Number: 901042

1. Dataset Description and Preprocessing

• The dataset consists of 215 student records and 15 columns, including both numerical and categorical variables.
• The target variable is status (Placed / Not Placed), a binary classification problem.
• Preprocessing steps included:
  - Dropping the irrelevant sl_no column.
  - Encoding binary columns (gender, workex, status) into 0/1.
  - One-hot encoding for multi-class categorical features: ssc_b, hsc_b, hsc_s, degree_t, and specialisation.
  - Filling 67 missing values in salary with the median salary: 265000.0.
  - Scaling numerical features using StandardScaler.

2. Models Chosen and Rationale

We used three classification models:
1. Logistic Regression: A baseline linear model appropriate for binary classification.
2. Random Forest Classifier: An ensemble tree-based model that handles both non-linearities and feature importance well.
3. Support Vector Machine (SVM): Effective in high-dimensional spaces and with clear decision boundaries.

Each model was selected to represent a different class of learning algorithm:
• Linear (Logistic)
• Tree-based (Random Forest)
• Margin-based (SVM)

3. Model Training & Hyperparameter Tuning

We used GridSearchCV with 5-fold cross-validation for each model to tune hyperparameters:

Model: Logistic Regression
Tuned Parameters: C from [0.01, 0.1, 1, 10]
Best Parameters: C=1

Model: Random Forest

Tuned Parameters: n_estimators, max_depth
Best Parameters: n_estimators=200, max_depth=10

Model: SVM
Tuned Parameters: C, kernel
Best Parameters: C=1, kernel='rbf'

4. Model Evaluation

We used accuracy, precision, recall, and F1-score on the test set:

Model: Logistic Regression
Accuracy: 0.80, Precision: 0.83, Recall: 0.89, F1-score: 0.86

Classification Report: Logistic Regression

	precision	recall	f1-score	support
0	0.72	0.62	0.67	21
1	0.83	0.89	0.86	44
accuracy			0.80	65
macro avg	0.78	0.75	0.76	65
weighted avg	0.80	0.80	0.80	65

- Class 1 (Placed) is predicted well (Precision: 0.83, Recall: 0.89).
- Class 0 (Not Placed) has weaker recall (0.62), meaning many "Not Placed" students were misclassified.
- Balanced performance; a solid linear baseline.

Model: Random Forest
Accuracy: 0.88, Precision: 0.85, Recall: 1, F1-score: 0.92

Classification Report: Tuned Random Forest

	precision	recall	f1-score	support
0	1.00	0.62	0.76	21
1	0.85	1.00	0.92	44
accuracy			0.88	65
macro avg	0.92	0.81	0.84	65
weighted avg	0.90	0.88	0.87	65

- Best performance across all models.
- Class 1: High precision (0.85) and perfect recall (1.00) — it captures all placed students.
- Class 0: Excellent precision (1.00), moderate recall (0.62) — predicts "Not Placed" with certainty but misses some.

- Well balanced and powerful, slightly overconfident on “Not Placed.”

Model: SVM
Overall Metrics:

- **Accuracy:** 0.77
- **Macro Avg F1-score:** 0.68
- **Weighted Avg F1-score:** 0.74

Classification Report: Tuned SVM

	precision	recall	f1-score	support
0	0.80	0.38	0.52	21
1	0.76	0.95	0.85	44
accuracy			0.77	65
macro avg	0.78	0.67	0.68	65
weighted avg	0.78	0.77	0.74	65

- Strong for Class 1 (Placed) with high recall (0.95) and decent F1 (0.85).
- Struggles with Class 0 (Not Placed): recall only 0.38 → many false negatives.
- Imbalanced sensitivity — favors predicting students as “Placed”.

Conclusion

•Random Forest is the most reliable overall.

•Logistic Regression is acceptable and more balanced but weaker than RF.

•SVM may need better tuning or feature engineering — it’s too biased toward predicting placements.

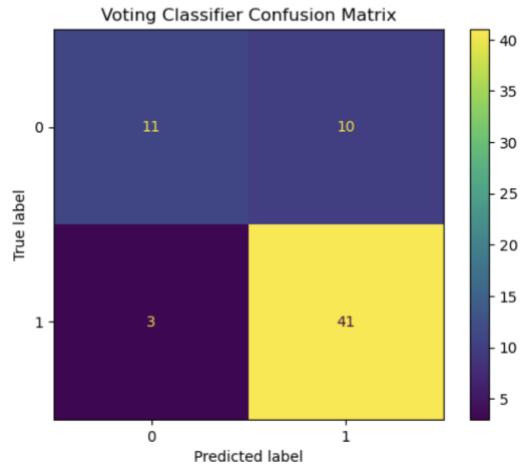
5. Voting Classifier

Overall Metrics:

- **Accuracy:** 0.80
- **Macro Avg F1-score:** 0.80
- **Weighted Avg F1-score:** 0.79

Classification Report: Voting Classifier

	precision	recall	f1-score	support
0	0.79	0.52	0.63	21
1	0.80	0.93	0.86	44
accuracy			0.80	65
macro avg	0.79	0.73	0.75	65
weighted avg	0.80	0.80	0.79	65



A Voting Classifier combining all three models was implemented using hard voting. Interpretation:

- Excellent at detecting “Placed” students (high recall of 0.93), better than Logistic Regression and even Random Forest in that specific area.
- Weak at detecting “Not Placed” students — high number of false positives.
- The ensemble prioritizes recall for Class 1 (likely due to its majority class and consistent strength across base models).6. Report Quality

• The report includes detailed steps, justifications, and metrics.
• All preprocessing and modeling code is well-commented.
• The report is now organized with numbered headings and is grammatically correct.

Conclusion

- The Voting Classifier offers stability and decent overall performance, but:
- It underperforms compared to the Random Forest, especially in precision and F1.

- It behaves similarly to Logistic Regression but sacrifices class 0 recall.
- It significantly outperforms SVM in balance and consistency.
- The advantage of the Voting Classifier is its ability to generalize and balance predictions, but in this case, Random Forest alone is stronger on all metrics.