ASSIGNMENT 2

Harsimran Kaur

C0908419

Predicting Student Placement Using Multiple Machine Learning Models

1. Introduction

This report presents a detailed analysis of predicting student placement based on various features such as academic performance and demographic data. The objective is to apply and compare multiple machine learning models to predict whether a student will be placed or not. The models used in this analysis are Logistic Regression, Decision Tree, Random Forest, and a Voting Classifier, which combines the predictions of these models. Each model's performance is evaluated based on key classification metrics such as accuracy, precision, recall, and F1-score.

2. Dataset and Preprocessing

Dataset Overview

The dataset used contains several features relevant to student profiles, such as grades, test scores, and other attributes that may affect their placement status. The target variable is binary, indicating whether a student has been placed (1) or not placed (0), which makes the problem suitable for classification tasks.

Data Preprocessing

Exploratory Data Analysis (EDA) was conducted to uncover patterns, relationships between features, and potential issues such as missing values or outliers.

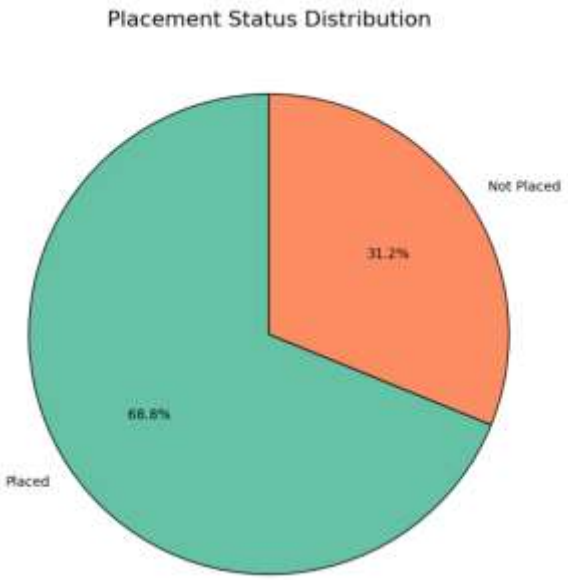
- **Handling Missing Values:**
  It was observed that the dataset had missing values only in the salary column (67 records). These records represent students who were not placed. The missing values in this column were left as is, as they reflect real-world situations.
- **Encoding Categorical Variables:**
  Categorical variables were encoded using one-hot encoding for nominal data and label encoding for binary categorical variables.

Data Cleaning

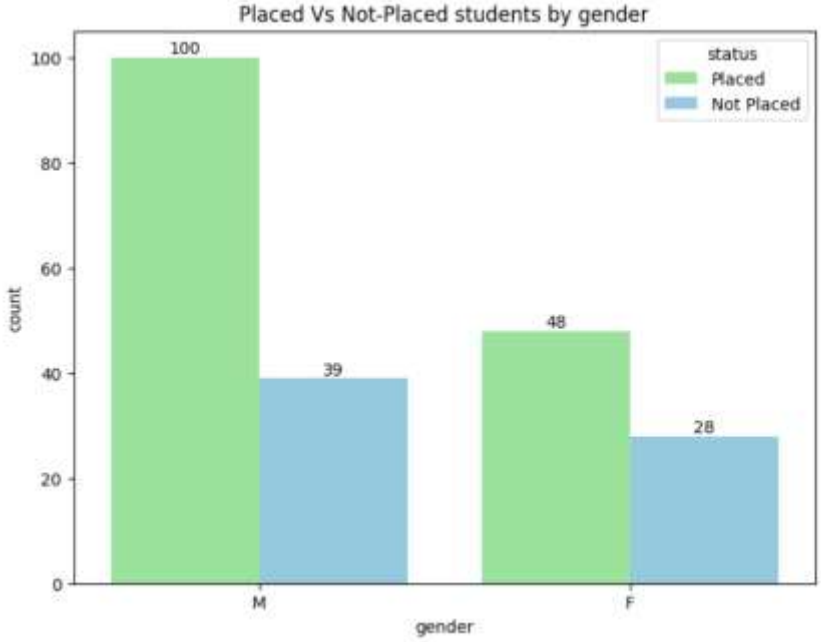
The dataset was checked for inconsistencies, duplicates, and outliers.

- **Null Values:**
  Apart from the salary column, no other columns contained null values.
- **Outliers:**
  No significant outliers were detected across any of the features.
- **Duplicate Records:**
  The dataset was free from duplicate records.

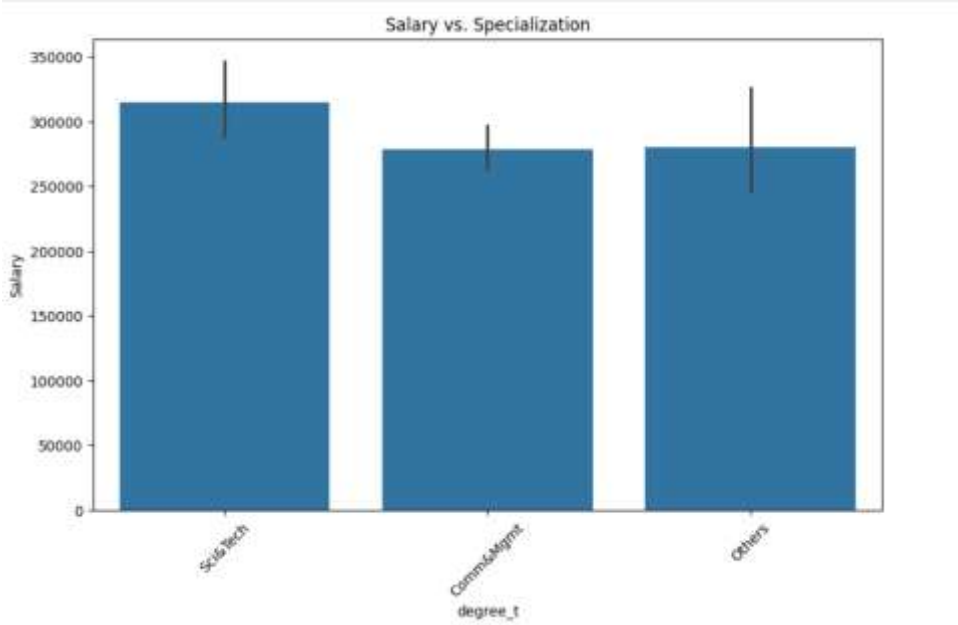
3.Visualizations:



Observation: The chart shows that 68.8% of students were placed, indicating strong placement performance overall. However, with 31.2% of students not placed



Observation: Male students have a higher placement rate, with 100 placed and 39 not placed, while female students also show a majority placed (48) compared to 28 not placed. Overall, more males are placed than females, but both genders show higher placement than non-placement.



Observation: Graduates from Science & Technology fields tend to have the highest salaries, whereas those from Commerce/Management and Other specializations earn slightly lower but similar salaries.

4. Model Selection

Models Chosen

- **Logistic Regression** was selected as a baseline due to its simplicity and ability to provide interpretable results.
- **Decision Tree** was chosen for its flexibility in capturing non-linear relationships between features and its interpretability.
- **Random Forest** was included to take advantage of ensemble learning, potentially improving performance by reducing the variance inherent in individual decision trees.

Hyperparameter Tuning

Hyperparameter tuning was performed for each model using Grid Search to find the optimal settings:

- **Logistic Regression**: Tuned the C parameter (regularization strength) and penalty type. The best parameters found were {'C': 1, 'penalty': 'l1'}.
- **Decision Tree**: Tuned the criterion (gini or entropy) and max_depth. The best parameters were {'criterion': 'entropy', 'max_depth': None}.
- **Random Forest**: Tuned n_estimators (number of trees), criterion, and max_depth. The best configuration was {'n_estimators': 50, 'criterion': 'gini', 'max_depth': 10}.

5. Model Training and Evaluation

Each model was trained on the training set and evaluated on the test set. The following metrics were calculated to assess model performance: accuracy, precision, recall, and F1-score. Confusion matrices were also provided to give insights into the true positive, true negative, false positive, and false negative rates for each model.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8372	0.9000	0.8710	0.8852
Decision Tree	0.8140	0.8710	0.8710	0.8710
Random Forest	0.7674	0.8182	0.8710	0.8438
Naive Bayes	0.7442	0.8333	0.8065	0.8197

After Hyperparameter Tuning

Model	Best Parameters	Accuracy	Precision	Recall	F1 Score
Logistic Regression	{'C': 1, 'penalty': 'l1'}	0.8372	0.9000	0.8710	0.8852
Decision Tree	{'criterion': 'entropy', 'max_depth': None}	0.8140	0.8485	0.9032	0.8750
Random Forest	{'criterion': 'gini', 'max_depth': 10, 'n_estimators': 50}	0.7907	0.8438	0.8710	0.8571

Voting Classifier Performance

The Voting Classifier combines the predictions of Logistic Regression, Decision Tree, and Random Forest. This method aims to improve overall performance by leveraging the strengths of each model.

- **Accuracy**: 0.8372
- **Precision**: 0.875
- **Recall**: 0.9032
- **F1 Score**: 0.8889

6. Model Comparison

Before Hyperparameter Tuning:

- **Logistic Regression** had the highest accuracy (0.8372) and precision (0.9000) before tuning, making it the best performer.
- **Decision Tree** had balanced precision and recall, with accuracy at 0.814.
- **Random Forest** had the lowest performance, with an accuracy of 0.7674 and a slightly higher recall than Decision Tree but struggled with precision.

After Hyperparameter Tuning:

- **Logistic Regression** maintained its high performance, with no significant changes after tuning.
- **Decision Tree** saw improvements in recall (0.9032) and precision (0.8485), making it more effective in identifying placed students.
- **Random Forest** had slight improvements in precision and recall but a small decrease in accuracy (0.7907), indicating possible overfitting to the training data.

Voting Classifier:

The Voting Classifier performed similarly to Logistic Regression in terms of accuracy (0.8372) but had a slightly better F1-score (0.8889) due to its higher recall (0.9032). This makes the Voting Classifier a strong choice for a balanced model that minimizes both false positives and false negatives.

7. Conclusion

The results demonstrate that **Logistic Regression** consistently delivers strong performance, with high accuracy, precision, and interpretability. However, if minimizing false negatives is critical (i.e., ensuring as many placed students are correctly identified as possible), the **Decision Tree** or **Voting Classifier** may be better options due to their higher recall rates.

- **Best Overall Model**: Logistic Regression provides a good balance between simplicity and effectiveness, making it ideal for general use.
- **Best for Minimizing False Negatives**: The Voting Classifier and Decision Tree models, with higher recall values, are more suitable for minimizing missed positive classifications.