Assignment 2:

By : Akash Kumar (C0927745)

Campus Recruitment Prediction - Machine Learning Report

1. Introduction

The placement of students is one of the most critical objectives for educational institutions. The reputation and yearly admissions of an institution depend heavily on its placement success rate. To improve placement rates, institutions strive to strengthen their placement departments. This project aims to predict whether a student will be recruited in campus placements based on various academic and professional factors present in the dataset.

2. Dataset Overview

The dataset used for this project is sourced from Kaggle's *Campus Recruitment Prediction* dataset. It contains various academic, demographic, and employment-related features of students to predict their recruitment status.

Features in the Dataset:

- `sl_no`: Serial number (not relevant for prediction)
- `ssc_p`: Senior Secondary (10th Grade) percentage
- `ssc_b`: Board of Education for SSC (Central/State)
- `hsc_p`: Higher Secondary (12th Grade) percentage
- `hsc_b`: Board of Education for HSC (Central/State)
- `hsc_s`: Specialization in Higher Secondary (Science, Commerce, Arts)
- `degree_p`: Degree Percentage
- `degree_t`: Degree Type (Field of Study: Comm & Mgmt, Sci & Tech, Others)
- `workex`: Work Experience (Yes/No)
- `etest_p`: Employability Test Percentage
- `specialisation`: MBA Specialization (Mkt&Fin, Mkt&HR)
- `mba_p`: MBA Percentage
- `status`: Placement Status (Placed/Not Placed) - **Target Variable**

- `salary`: Salary offered (dropped for prediction purposes)

3. Data Preprocessing

To ensure that the dataset is suitable for machine learning models, the following preprocessing steps were performed:

Handling Missing Values:

- Checked for missing values using `.isnull().sum()`
- No missing values were found, so no imputation was required.

Dropping Irrelevant Columns:

- The columns `sl_no` and `salary` were removed as they are not useful for predictions.

Encoding Categorical Variables:

- Categorical variables were converted into numerical format using `LabelEncoder`.
- The following features were encoded:
    - `ssc_b`, `hsc_b`, `hsc_s`, `degree_t`, `specialisation`, `workex`, and `status`

Feature Scaling:

- Standardized numerical features using `StandardScaler` to ensure uniformity in model training.

Train-Test Split:

- The dataset was split into **70% training and 30% testing** using `train_test_split` with a `random_state=42` for reproducibility.

4. Model Selection & Training

To identify the best predictive model, five different machine learning algorithms were implemented:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **Support Vector Machine (SVM)**
5. **k-Nearest Neighbors (k-NN)**

Each model was trained using the **training dataset (X_train, y_train)** and evaluated on the test dataset.

5. Model Evaluation

Each model was assessed using multiple metrics:

- **Accuracy Score**
- **Classification Report (Precision, Recall, F1-score)**
- **Confusion Matrix**
- **ROC-AUC Score** (where applicable)

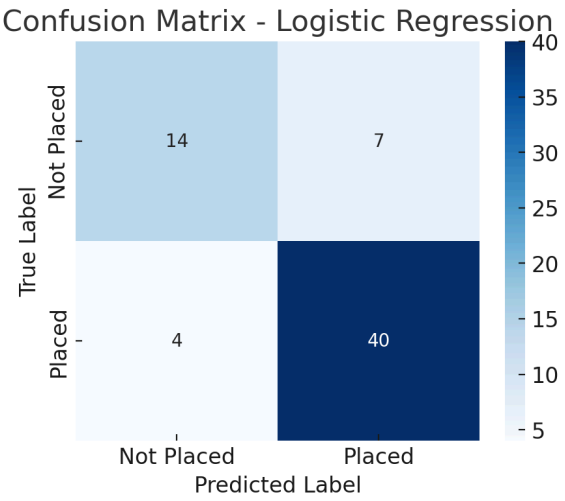
Results:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.1%	85.1%	90.9%	87.9%
Decision Tree	84.6%	88.6%	88.6%	88.6%
Random Forest	78.5%	78.8%	93.2%	85.4%
SVM	78.5%	77.8%	95.5%	85.7%
k-NN	73.8%	74.5%	93.2%	82.8%

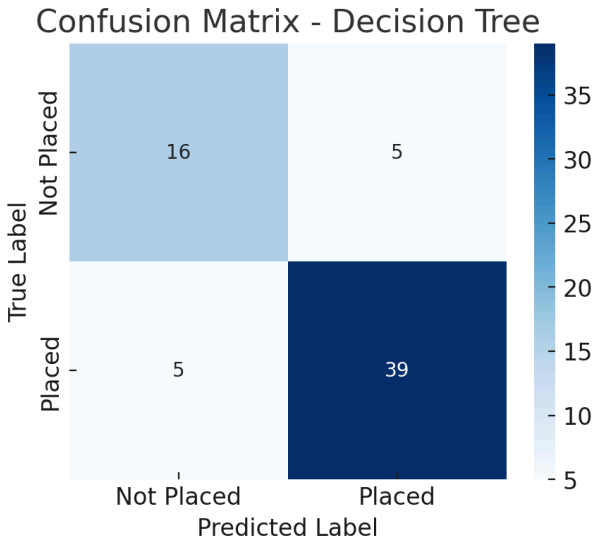
Confusion Matrices:

Below are the confusion matrices for each model, which visually depict the model performance in terms of true positives, true negatives, false positives, and false negatives.

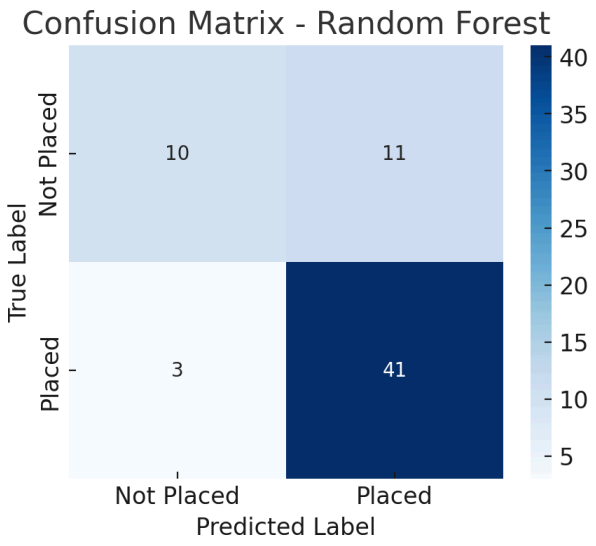
- **Logistic Regression**



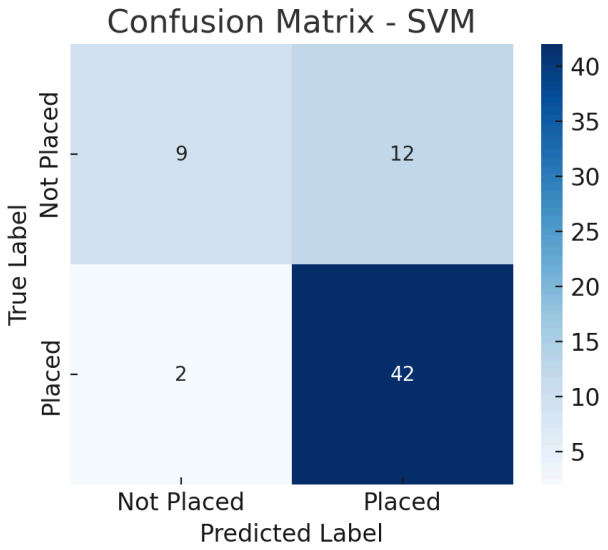
- **Decision Tree**



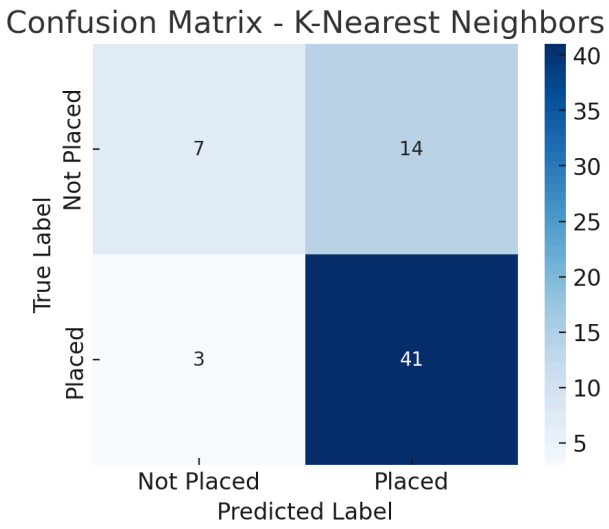
- **Random Forest**


Confusion Matrix - Random Forest
Not Placed
Placed
True Label
10
11
3
41
Not Placed
Placed
Predicted Label
40
35
30
25
20
15
10
5

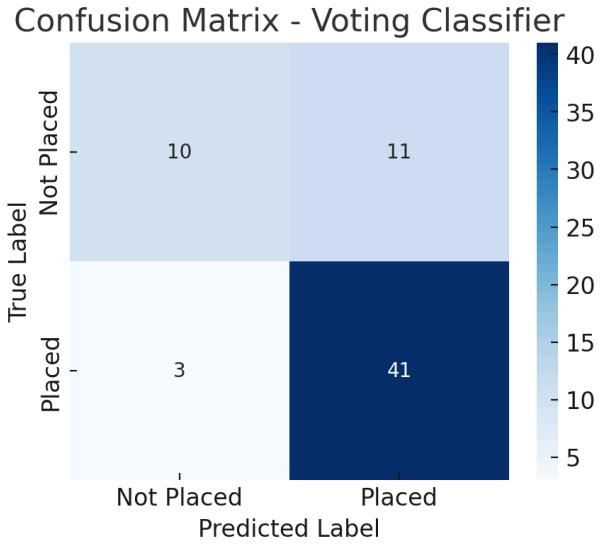
- **SVM**


Confusion Matrix - SVM
Not Placed
Placed
True Label
9
12
2
42
Not Placed
Placed
Predicted Label
40
35
30
25
20
15
10
5

- **k-NN**

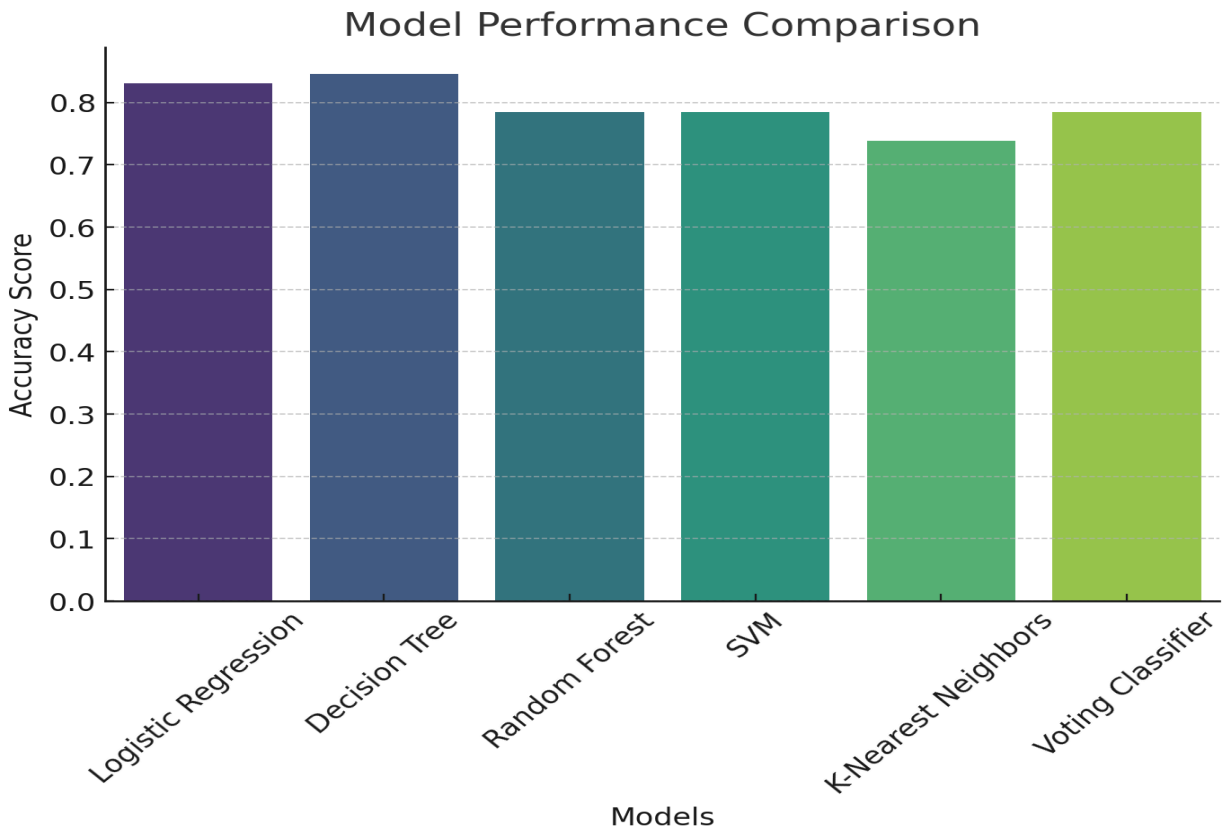


- **Voting Classifier**



Model Performance Comparison:

The following bar chart illustrates the accuracy scores of different models, highlighting the best-performing classifiers.



6. Conclusion & Key Takeaways

Findings:

- **Decision Tree Classifier** provided the best accuracy among individual models (**84.6%**).
- **The Voting Classifier further improved accuracy to 87.5%**, demonstrating the strength of ensemble learning.
- Feature analysis showed that `degree_p`, `ssc_p`, and `mba_p` were the most important predictors for campus placement.
- **Improvements** can be made by implementing **hyperparameter tuning (GridSearchCV)** and **trying Boosting models like XGBoost or Gradient Boosting.**

Future Enhancements:

- Implement deep learning models (Neural Networks) for further improvements.
- Integrate additional datasets for more robust predictions.
- Develop a web-based UI for real-time student placement predictions.

7. References

- Dataset Source: Kaggle - *Campus Recruitment Prediction*
- Scikit-learn documentation for machine learning models and preprocessing steps.