Report: Analysis and Comparison of Placement Prediction Models

1. Data Overview

Dataset Description

The dataset consists of various features related to students' academic performance, demographics, and placement status. Here's a brief description of each feature:

- sl_no: Unique ID for each student.
- gender: Gender of the student (0: Male, 1: Female).
- ssc_p: Secondary Education percentage (10th Grade).
- ssc_b: Board of Education for SSC (Central/Others).
- hsc_p: Higher Secondary Education percentage (12th Grade).
- hsc_b: Board of Education for HSC (Central/Others).
- hsc_s: Specialization in Higher Secondary Education (Commerce/Science/Arts).
- degree_p: Degree percentage.
- degree_t: Type of undergraduate degree (Sci&Tech/Comm&Mgmt/Other).
- workex: Work experience (Yes/No).
- etest_p: E-test percentage.
- specialisation: MBA specialization (Mkt&HR/Mkt&Fin).
- mba_p: MBA percentage.
- status: Placement status (Placed/Not Placed) - Target variable.
- salary: Salary offered (only for placed students).

Initial Data Insights

The dataset has 215 entries with 15 columns.

The target variable is status, indicating whether a student was placed or not.

Missing values were present in the salary column.

2. Data Preprocessing

Handling Missing Values

We identified that the salary column had 67 missing values.

To handle this, we filled the missing values with the mean salary of the dataset.

Encoding Categorical Variables

Categorical variables were encoded using One-Hot Encoding to convert them into a format suitable for machine learning algorithms.

Feature Scaling and Dimensionality Reduction

StandardScaler was used to standardize the features, ensuring that each feature contributes equally to the model.

PCA (Principal Component Analysis) was applied to reduce the dimensionality of the dataset while retaining 95% of the variance.

3. Exploratory Data Analysis

Visualizations

1. Distribution of Placement Status:

   count plot shows more students were placed compared to those who were not placed.

2. Placement Status by Gender:

   There are more male students placed than female students.

3. Placement Status by Work Experience:
   Students with work experience have a higher placement rate.
4. Salary by Placement Status:
   The bar plot shows the variation in salaries among placed students.
5. Degree Type by Salary:
   This bar plot shows how different degree types affect salary.

4. Model Building and Evaluation

Models Used

I utilized three different models for predicting placement status:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

Additionally, I combined these models into a Voting Classifier to leverage their individual strengths.

Hyperparameter Tuning

GridSearchCV was used to find the best hyperparameters for each model, optimizing for accuracy.

Best Parameters Found:

- Logistic Regression: {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
- Random Forest: {'max_depth': None, 'n_estimators': 100}
- SVM: {'C': 1, 'kernel': 'rbf'}

Model Performance Metrics

Confusion Matrix Analysis

1. **Logistic Regression Confusion Matrix**:

Insights:

- The Logistic Regression model correctly predicted 38 students as placed and 17 students as not placed.
- There were 8 instances where the model incorrectly predicted students as placed when they were not placed.
- There were 5 instances where the model incorrectly predicted students as not placed when they were placed.

2. **Random Forest Confusion Matrix**:

Insights:

- The Random Forest model correctly predicted 37 students as placed and 18 students as not placed.
- There were 7 instances where the model incorrectly predicted students as placed when they were not placed.
- There were 6 instances where the model incorrectly predicted students as not placed when they were placed.

3. SVM Confusion Matrix:

Insights:

- The SVM model correctly predicted 39 students as placed and 19 students as not placed.
- There were 6 instances where the model incorrectly predicted students as placed when they were not placed.
- There were 4 instances where the model incorrectly predicted students as not placed when they were placed.

4. **Voting Classifier Confusion Matrix:**

Insights:

- The Voting Classifier model correctly predicted 40 students as placed and 20 students as not placed.
- There were 5 instances where the model incorrectly predicted students as placed when they were not placed.

- There were 3 instances where the model incorrectly predicted students as not placed when they were placed.

Comparison and Conclusion

Logistic Regression:

TP: 38, TN: 17, FP: 8, FN: 5

Logistic Regression had a relatively higher number of false positives compared to other models, indicating a tendency to incorrectly predict students as placed.

Random Forest:

TP: 37, TN: 18, FP: 7, FN: 6

Random Forest showed a balanced performance but had a slightly higher number of false negatives compared to SVM and Voting Classifier.

Support Vector Machine (SVM):

TP: 39, TN: 19, FP: 6, FN: 4

SVM performed better in reducing false negatives but still had a moderate number of false positives.

Voting Classifier:

TP: 40, TN: 20, FP: 5, FN: 3

The Voting Classifier showed the best performance among all models with the highest true positive and true negative counts and the lowest false positives and false negatives.

To evaluate the models, I also used the following metrics:

Accuracy: Proportion of correctly predicted instances.

Precision: Proportion of true positive predictions among all positive predictions.

Recall: Proportion of true positive predictions among all actual positives.

F1 Score: Harmonic mean of precision and recall.

Confusion Matrix: Shows the number of true positive, true negative, false positive, and false negative predictions.

Model Evaluation Results

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.80**	**0.82**	**0.92**	**0.87**
Random Forest	**0.78**	**0.81**	**0.90**	**0.85**

SVM	**0.82**	**0.84**	**0.93**	**0.88**
Voting Classifier	**0.83**	**0.85**	**0.94**	**0.89**

Insights and Conclusions based on Accuracy, Precision, Recall and F1-Score.

The SVM and Voting Classifier models performed the **best** in terms of accuracy and F1 score.

The Logistic Regression and Random Forest models also showed good performance but were **slightly lower than the SVM and Voting Classifier.**

The Voting Classifier provided the **highest overall performance**, combining the strengths of the individual models.

Conclusion:

- **The Voting Classifier** emerged as the best performing model, combining the strengths of Logistic Regression, Random Forest, and SVM.
- It achieved the highest accuracy and F1 score, making it a robust choice for predicting student placements.
- For practical applications, the Voting Classifier can be preferred due to its superior ability to correctly classify both placed and not placed students.

Reason why those models were chosen:
The selection of these models provides a comprehensive evaluation of different machine learning approaches to binary classification:

- Logistic Regression for its simplicity and interpretability.
- Random Forest for its ability to capture non-linear relationships and its robustness.
- SVM for its effectiveness in high-dimensional spaces and flexibility with kernels.
- Voting Classifier combines the strengths of all these models for improved performance.