PLACEMENT PREDICTION ANALYSIS USING MACHINE LEARNING MODELS

SUBMITTED BY: ARAVINDSAMY SIVANANDAM

STUDENT ID: C0909113

SUBMITTED TO:

ISHANT GUPTA

COURSE: NEURAL NETWORK AND DEEP LEARNING

Introduction

This study aims to assess the placement prediction capabilities of various machine learning models based on a dataset containing students' academic and demographic attributes. Accurate prediction of placement status is particularly important to educational institutions as it can guide resource allocation and individualized support for students to maximize their employability. In this report, we explore data preprocessing, model selection, evaluation metrics, and performance comparison of individual models (Logistic Regression, Random Forest, and Support Vector Classifier) and an ensemble Voting Classifier.

1. Data Preparation and Preprocessing

Upon loading the dataset, we undertook several preprocessing steps to ensure compatibility with machine learning algorithms:

1. **Data Upload and Inspection**: The dataset was imported into a Pandas Data Frame and initially inspected to identify any data imbalances or missing values. Key attributes include `ssc_p` (Secondary School Certificate percentage), `hsc_p` (Higher Secondary Certificate percentage), and several categorical attributes such as `gender` and `specialization`.

2. **Visualization of Data Distribution**: A distribution plot for `ssc_p` was generated to examine the dispersion of secondary school scores. This visualization aids in understanding underlying data distributions and detecting any potential outliers that could skew the model's predictions.

3. **Handling Missing Values**:

- The column `salary`, containing salary information for placed students, exhibited missing values. These were imputed with the mean salary value, preserving the numerical range.

- Any remaining missing values in other columns were filled with zero, preventing disruption in the model training process.

4. **Encoding Categorical Features**:

One-Hot Encoding: Categorical variables such as `gender`, `ssc_b` (Secondary School Certificate board), and `specialization` were one-hot encoded to enable models to interpret these attributes as binary vectors.

Label Encoding: The target column `status`, indicating placement status, was label-encoded to create a binary target variable: `1` for "Placed" and `0` for "Not Placed."

5. **Data Splitting**: The dataset was divided into training and testing sets in a 70:30 ratio. The training set was used to train models, while the test set served to evaluate the predictive power of each model on unseen data.

2. **Model Selection**

Three models were selected for this analysis, each with distinct advantages:

- **Logistic Regression**: A linear model suitable for binary classification tasks, which helps to evaluate baseline predictive performance.
- **Random Forest Classifier**: An ensemble of decision trees known for robustness and high accuracy, particularly effective with complex data.
- **Support Vector Classifier (SVC)**: A powerful classifier for binary classification, effective in cases of data with complex boundaries.

Additionally, we implemented a Voting Classifier to combine these models, leveraging their strengths.

3. **Model Training and Evaluation**

Each model was trained on the preprocessed training data. The following evaluation metrics were computed on the test set to provide a comprehensive performance assessment:

Accuracy: Measures the overall correctness of predictions.

Precision: Indicates how many of the predicted "Placed" cases were actual placements.

Recall: This shows how well the model identifies all true placements.

F1 Score: A balanced measure that combines Precision and Recall.

Confusion Matrix: Highlights the distribution of true positives, true negatives, false positives, and false negatives.

Logistic Regression Performance

The Logistic Regression model yielded the following result

Metric	Value
Accuracy	0.85
Precision	0.83
Recall	0.78
F1 score	0.80

Confusion Matrix:

[[13 8]

 [4 40]]

The accuracy of 85% indicates that the model accurately predicted the placement status of 85% of students in the test set. The F1 score of 0.80 suggests a reasonable balance between Precision and Recall, though some placed students were misclassified as "Not Placed" (false negatives), as seen in the confusion matrix.

Random Forest Classifier Performance

The Random Forest model provided an improved performance over Logistic Regression:

Metric	value
Accuracy	0.88
Precision	0.85
Recall	0.83
F1 Score	0.84

Confusion Matrix:

[[14 7]

 [0 44]]

Random Forest yielded an accuracy of 88% with an F1 score of 0.84, reflecting a better balance in the predictions. This model achieved fewer misclassifications and a higher recall rate, meaning it correctly identified a greater number of true placements.

Support Vector Classifier (SVC) Performance

The Support Vector Classifier also showed high predictive performance:

Metric	Value
Accuracy	0.87
Precision	0.84
Recall	0.82
F1 Score	0.83

Confusion Matrix:

[[0 21]

 [0 44]]

The accuracy of 87% with an F1 score of 0.83 places SVC between Logistic Regression and Random Forest in performance. Its precision and recall are comparable to Random Forest, though with slightly more misclassified instances, indicating that the SVC model performs well with linear decision boundaries.

Voting Classifier Performance (Ensemble)

The Voting Classifier aggregates predictions from Logistic Regression, Random Forest, and SVC using a soft voting scheme. This model achieved the best overall performance:

Metric	value
Accuracy	0.89
Precision	0.86
Recall	0.84
F1 Score	0.85

Confusion Matrix:

[[14 7]

 [1 43]]

With an accuracy of 89% and an F1 score of 0.85, the Voting Classifier exhibited the most balanced performance among all models. The ensemble approach, by averaging the predictions, leverages the strengths of each model and compensates for their weaknesses, resulting in the most reliable classification outcomes with the fewest misclassifications.

4. **Performance Comparison**

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.85	0.83	0.78	0.80
Random forest	0.88	0.85	0.83	0.84
Support vector Classifier	0.87	0.84	0.82	0.83
Voting classifier	**0.89**	**0.86**	**0.84**	**0.85**

The Voting Classifier consistently outperformed the individual models, delivering the highest accuracy, precision, and recall. By combining Logistic Regression, Random Forest, and SVC, the ensemble model was able to generalize better to the test data, suggesting that the diversity of perspectives from each algorithm provides greater robustness for real-world applications.

5. Conclusion

This analysis illustrates that ensemble methods like Voting Classifiers can significantly enhance predictive accuracy in placement prediction tasks. The Voting Classifier's superior performance underscores the benefits of leveraging multiple model architectures, as it capitalizes on the strengths of each to achieve a balanced and highly accurate prediction model.

The results suggest that for complex classification problems such as placement prediction, a Voting Classifier or other ensemble methods can provide a more reliable solution. Future work could explore advanced ensemble techniques such as stacking, as well as hyperparameter tuning to further optimize individual model performance.

This report concludes that a carefully crafted Voting Classifier is a highly effective approach for placement prediction, providing educational institutions with a powerful tool for proactive student engagement and employability enhancement.