CAMPUS PLACEMENT PREDICTION REPORT

1.Introduction

Campus placements are a critical metric for educational institutions, often influencing reputation and enrollment. With machine learning, we can predict student placement likelihood based on a range of factors, offering valuable insights for students aiming to boost employability and for institutions optimizing placement strategies.

2.Objective and Approach

This project aims to predict campus placement outcomes using a structured machine learning pipeline:

1. **Data Exploration and Preprocessing**: Cleaning the dataset to ensure accuracy.
2. **Feature Engineering**: Creating new attributes to uncover additional data patterns.
3. **Model Selection and Training**: Testing various models and tuning hyperparameters.
4. **Model Evaluation and Comparison**: Using metrics like accuracy, F1 score, and ROC AUC.
5. **Ensemble Model**: Employing a Voting Classifier to combine top-performing models for enhanced prediction accuracy.

The dataset, sourced from Kaggle's "Campus Recruitment Prediction" project, comprises 215 records with 15 attributes including personal details, academic scores, and placement status.

3.Data Preprocessing

Data preprocessing involved handling missing values, visualization, feature engineering, encoding, and scaling:

- **Handling Missing Values**: Missing salary values were imputed with the mean salary.
- **Data Visualization**: Distribution plots highlighted trends, such as higher placement rates for students with higher academic scores.
- **Feature Engineering**: New features included:
    - ssc_hsc_ratio: Ratio of secondary to higher secondary scores.
    - ssc_degree_ratio: Ratio of secondary to degree scores.
    - total_academic_score: Aggregate score from secondary, higher secondary, and degree percentages.
- **Encoding Categorical Variables**: Label encoding was used to convert categorical data to numerical format.

- **Data Splitting and Scaling**: The data was split into training and test sets (70-30), followed by feature scaling for model compatibility.

4.Model Selection and Training

Several algorithms were used to capture varied predictive approaches:

1. **Logistic Regression**: Effective in binary classification, tuned with the regularization parameter C.
2. **Decision Tree Classifier**: Captures non-linear patterns; max_depth values were varied.
3. **Random Forest Classifier**: Combines decision trees for enhanced accuracy, with n_estimators and max_depth tested.
4. **Support Vector Machine (SVM)**: Efficient in high-dimensional spaces; linear and RBF kernels tested.
5. **k-Nearest Neighbors (k-NN)**: A simple, distance-based classifier.
6. **Gradient Boosting Classifier**: A boosting technique for improved performance, with tuning of n_estimators and learning_rate.

5.Hyperparameter Tuning

Each model was tuned for optimal performance. For instance, the Gradient Boosting model tested different n_estimators and learning_rate values, balancing model complexity with accuracy.

6.Model Evaluation

Performance metrics used included:

- **Accuracy**: Measures the proportion of correct predictions.
- **Precision & Recall**: Precision indicates the avoidance of false positives, and recall reflects the capture of true positives.
- **F1 Score**: Balances precision and recall.
- **ROC AUC**: Assesses the model's ability to differentiate between classes.

Precision-Recall and ROC Curves provided visual insight into model performance on imbalanced data. The Gradient Boosting model demonstrated high recall and an AUC close to 1, showing strong distinction between placed and non-placed students.

Ensemble Model: Voting Classifier

A weighted Voting Classifier was implemented, using top models based on F1 scores. With soft voting, this ensemble achieved:

- **Accuracy**: 0.9846
- **Precision**: 0.9778
- **Recall**: 1.0
- **F1 Score**: 0.9888
- **ROC AUC**: 1.0

The Voting Classifier outperformed individual models, demonstrating the benefit of ensemble methods in complex classification tasks.

7.Conclusion

The Voting Classifier provided the most accurate placement predictions. Its robust accuracy, precision, and recall make it a reliable model for campus placement predictions. Future work could refine this model with additional data for improved generalization, supporting institutions and students in maximizing placement success.
