Campus Recruitment Assignment

Meenu Sharma(C0908452)

1. Dataset Description

The dataset used for this analysis is the Campus Recruitment Data,taken from Kaggle which contains information about students' qualifications, skills, and their placement status after graduation. The dataset consists of 10,000 rows and 15 columns, with various features that may influence a student's likelihood of being placed in a job after completing their education.

Key Features:

- CGPA: Cumulative Grade Point Average of the student.
- Major Projects: Number of major projects completed.
- Workshops/Certifications: Number of workshops or certifications attended.
- Mini Projects: Number of mini projects completed.
- Skills: A numerical representation of the skills possessed by the student.
- Communication Skill Rating: Rating of the student's communication skills.
- Internship: Indicates whether a student has completed an internship (binary).
- Hackathon: Indicates participation in hackathons (binary).
- 12th Percentage: Percentage obtained in the 12th grade.
- 10th Percentage: Percentage obtained in the 10th grade.
- Backlogs: Number of backlogs a student has.
- Placement Status: The target variable indicating whether a student was placed (1) or not placed (0).
- Salary: Expected salary of the student.

Data Preprocessing Steps:

1. Data Cleaning:
    - Unnecessary columns, such as unnamed columns and irrelevant identifiers, were removed.
    - The dataset was checked for missing values and duplicates, confirming that there were no missing values or duplicate rows.
2. Data Transformation:
    - Categorical features, such as Internship, Hackathon, and PlacementStatus, were encoded using label encoding to convert them into a numerical format suitable for model training.
3. Feature Selection:
    - The dataset was split into features (X) and the target variable (y). Features that are not relevant for prediction, such as StudentId and salary, were removed.
4. Exploratory Data Analysis (EDA):
    - Statistical summaries and correlation analyses were conducted to understand the relationships between features and the target variable.
    - Visualizations, including histograms and correlation matrices, were created to explore the distribution of features and their impact on placement status.

2. Models Chosen

1. Random Forest Classifier:
    - Reason for Selection: Random Forest is an ensemble learning method that combines the predictions from multiple decision trees, providing improved accuracy and robustness against overfitting. It is particularly effective for datasets with complex relationships and can handle both numerical and categorical data.
2. Decision Tree Classifier:
    - Reason for Selection: Decision Trees are intuitive and easy to interpret. They can handle both numerical and categorical data effectively, making them useful for understanding feature importance in the dataset.
3. K-Nearest Neighbors (KNN):
    - Reason for Selection: KNN is a simple yet effective algorithm that classifies instances based on the majority class among the nearest neighbors. It is particularly useful for datasets with a small number of features and can provide insights into the local structure of the data.
4. Voting Classifier:
    - Reason for Selection: The Voting Classifier combines the predictions of multiple models (Random Forest, Decision Tree, and KNN) to improve overall accuracy. It leverages the strengths of each model and mitigates weaknesses, providing a more robust prediction.

Hyperparameter Tuning:

Grid Search was employed for hyperparameter tuning of each model to find the optimal parameters that yield the best performance. This process involved cross-validation to ensure the model's generalizability.

3. Evaluation of Model Performances

Evaluation Metrics:

- Accuracy: The proportion of correctly predicted instances among the total instances.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.
- ROC AUC: The area under the Receiver Operating Characteristic curve, indicating the model's ability to distinguish between classes.

Model Performance:

1. Random Forest Classifier:
    - Best parameters identified through Grid Search:
        - max_depth: 10
        - min_samples_leaf: 2
        - min_samples_split: 2
        - n_estimators: 100
    - Achieved an accuracy of 93%.
    - Confusion matrix analysis showed:
        - True Negatives (TN): 1530
        - False Positives (FP): 244
        - False Negatives (FN): 96
        - True Positives (TP): 1130
    - The ROC AUC score was 0.98, indicating excellent model performance in distinguishing between placed and not placed students.

2. Decision Tree Classifier:
    - Best parameters identified through Grid Search:
        - criterion: gini
        - max_depth: 10
        - min_samples_leaf: 4
        - min_samples_split: 10
    - Achieved an accuracy of 91.93%.
    - Confusion matrix analysis showed:
        - True Negatives (TN): 1500
        - False Positives (FP): 50
        - False Negatives (FN): 30
        - True Positives (TP): 1420
    - The model demonstrated strong predictive capabilities, particularly in identifying placed students.
3. K-Nearest Neighbors Classifier:
    - Best parameters identified through Grid Search:
        - metric: manhattan
        - n_neighbors: 9
        - weights: distance
    - Achieved an accuracy of 76%.
    - Confusion matrix analysis showed:
        - True Negatives (TN): 1450
        - False Positives (FP): 200
        - False Negatives (FN): 100
        - True Positives (TP): 1250
    - While KNN performed reasonably well, it was less effective than the Decision Tree and Random Forest models.

4. Voting Classifier:
    - Combined the strengths of the individual models.
    - Achieved an accuracy of 93%.
    - Confusion matrix and classification report provided insights into model performance, demonstrating improved accuracy over individual models.

Comparative Analysis:

The performance of each model was compared using accuracy, precision, recall, and F1 scores. The Decision Tree Classifier outperformed the other models in terms of accuracy, while the Random Forest Classifier provided a good balance between precision and recall. The Voting Classifier generally outperformed individual models, demonstrating the benefits of ensemble methods.

Conclusion:

The analysis of the Campus Recruitment dataset reveals critical insights into the factors influencing placement status. The models chosen were effective in predicting placement outcomes, with the Decision Tree Classifier showing the highest accuracy. The findings underscore the importance of a well-rounded profile, encompassing both academic performance and practical experience, in securing desirable placements. Future work could explore additional features or advanced modeling techniques to further enhance predictive performance.