2024S-T3 AML 3104 - Neural Networks and Deep Learning 01 (DSMM Group 1 & Group 3)

Assignment 2 - Campus Placement

Shanmuga Priyan Jeevanandam

C0889053

1. Dataset and Preprocessing Steps

Dataset Description: The dataset used for this project is related to campus placement prediction. It consists of several features that describe students' academic performance, work experience, and other relevant attributes. The target variable is a categorical variable indicating whether a student has been placed or not.

Preprocessing Steps:

1. **Missing Values Handling:**
    - I checked for missing values in the dataset and found that the 'salary' column had missing values. These were filled with the median salary to prevent any bias.
2. **Encoding Categorical Variables:**
    - Categorical variables such as 'ssc_b', 'hsc_b', 'hsc_s', 'degree_t', 'workex', 'specialisation', and 'status' Were encoded using LabelEncoder to convert them into numerical values suitable for machine learning algorithms.
3. **Feature Standardization:**
    - I standardized the features using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial for models like Logistic Regression and SVM which are sensitive to the scale of the input features.
4. **Data Splitting:**
    - The dataset was split into training (70%) and validation (30%) sets to train the models and evaluate their performance on unseen data.

2. Model Selection and Justification

I selected the following models for this task:

1. **Random Forest Classifier:**
    - Chosen for its ability to handle large datasets and capture complex interactions between features through ensemble learning.
2. **Logistic Regression:**
    - A simple yet effective linear model suitable for binary classification tasks, providing a good baseline for comparison.
3. **Support Vector Classifier (SVC):**
    - Selected for its effectiveness in high-dimensional spaces and ability to create complex decision boundaries.
4. **Decision Tree Classifier:**
    - Chosen for its simplicity and interpretability, allowing us to understand the decision-making process.
5. **Gradient Boosting Classifier:**
    - Selected its ability to improve model performance through boosting, by focusing on the errors of previous models.
6. **AdaBoost Classifier:**
    - Another boosting method chosen for its simplicity and effectiveness in improving Iak classifiers.

7. **Voting Classifier:**
    - An ensemble method combining the predictions of the above models to improve overall performance through majority voting.

3. Model Performance Evaluation

Evaluation Metrics: I used the following metrics to evaluate model performance:

- **Accuracy:** Proportion of correct predictions.
- **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, providing a balance between the two.

Performance Results:

Model	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Random Forest	0.8	0.792	0.955	0.866	[[10, 11], [2, 42]]
Logistic Regression	0.8	0.816	0.909	0.860	[[12, 9], [4, 40]]
SVC	0.846	0.870	0.909	0.889	[[15, 6], [4, 40]]
Decision Tree	0.754	0.769	0.909	0.833	[[9, 12], [4, 40]]
Gradient Boosting	0.785	0.788	0.932	0.854	[[10, 11], [3, 41]]
AdaBoost	0.815	0.820	0.932	0.872	[[12, 9], [3, 41]]
Voting Classifier	0.8	0.816	0.909	0.860	[[12, 9], [4, 40]]

Confusion Matrix Interpretation:

- **True Positives (TP):** Number of correctly predicted positive cases.
- **True Negatives (TN):** Number of correctly predicted negative cases.
- **False Positives (FP):** Number of incorrectly predicted positive cases.
- **False Negatives (FN):** Number of incorrectly predicted negative cases.

Visualized Results: The results were visualized using bar charts for each metric (Accuracy, Precision, Recall, and F1-score) to compare the performance of different models effectively.

4. Comprehensive Evaluation

Model Comparison:

- **Best Performing Model:** The SVC model performed the best with an accuracy of 0.846, precision of 0.870, recall of 0.909, and F1-score of 0.889.
- **Voting Classifier:** The ensemble voting classifier did not outperform individual models but provided a balanced performance, demonstrating the robustness of ensemble methods.

By following this structured approach, I ensured that the dataset was handled appropriately, multiple models were evaluated, and the best model was identified based on performance metrics. This thorough analysis aids in making informed decisions for deployment in real-world scenarios.

5. Reference:

https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement