Roshan Joseph

C0904407

Placement Prediction Model Performance Report

Introduction

This report details the performance of various machine learning models trained on a placement dataset. The primary goal is to predict the placement status of students based on various attributes, which can be critical for educational institutions and students alike.

Data Preprocessing

The dataset underwent several preprocessing steps to ensure the models could learn effectively from the data:

1. **Label Encoding**: Categorical features such as 'gender', 'work experience', 'specialization', and 'status' were converted into numerical representations using Label Encoding. This transformation allows the models to interpret categorical data appropriately.
2. **One-Hot Encoding**: Multi-categorical features like 'ssc_b', 'hsc_b', 'hsc_s', and 'degree_t' were transformed using One-Hot Encoding. This method creates binary columns for each category, preventing the model from assuming any ordinal relationship between categories.
3. **Missing Value Imputation**: Missing 'salary' values were filled using the median salary corresponding to each specialization. This approach helps maintain the integrity of the data without introducing bias.
4. **Feature Scaling**: Numerical features were standardized using **StandardScaler**, which normalizes the data to have a mean of 0 and a standard deviation of 1. This step is crucial for models sensitive to the scale of input features, such as K-Nearest Neighbors and SVM.

Model Training and Evaluation

Multiple classification models were trained and evaluated using the following metrics: accuracy, precision, recall, and F1-score. A confusion matrix was also generated for each model to visualize performance.

Models Evaluated

The following models were evaluated:

- K-Nearest Neighbors (KNN)
- Logistic Regression
- Gradient Boosting
- Decision Tree

- Naive Bayes
- Random Forest
- Support Vector Machine (SVM)
- XGBoost
- Voting Classifier (combination of multiple models)

Model Performance Summary

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**R-Squared**
K-Nearest Neighbors	0.790698	0.805556	0.935484	0.865672	0.780
Logistic Regression	0.883721	0.906250	0.935484	0.920635	0.850
Gradient Boosting	0.813953	0.828571	0.935484	0.878788	0.900
Decision Tree	0.837209	0.900000	0.870968	0.885246	0.920
Naive Bayes	0.744186	0.812500	0.838710	0.825397	0.750
Random Forest	0.790698	0.789474	0.967742	0.869565	0.880
Support Vector Machine	0.790698	0.805556	0.935484	0.865672	0.800
XGBoost	0.837209	0.852941	0.935484	0.892308	0.820

Best Performing Model

The best performing model based on the F1-score is **Logistic Regression**, achieving an F1-score of **0.9206**. This model demonstrates a strong balance between precision and recall, making it particularly effective for the placement prediction task.

Model Performance Comparison (Visualization)

A plot illustrating the performance of different models across the metrics of accuracy, precision, recall, and F1-score was created. This visual representation aids in quickly assessing the strengths and weaknesses of each model.

Voting Classifier Results

A **Voting Classifier**, which combines the predictions of all the individual models, was also tested. Its performance is summarized below:

- **Accuracy**: 0.8372
- **Precision**: 0.8529
- **Recall**: 0.9355
- **F1-score**: 0.8923

The Voting Classifier demonstrates promising performance, indicating that combining multiple models can enhance overall prediction accuracy.

Conclusion

In conclusion, **Logistic Regression** shows the best performance in predicting student placement based on the F1-score, indicating its effectiveness in balancing precision and recall. The **Voting Classifier** also shows competitive performance, suggesting that ensemble methods can be beneficial in improving prediction outcomes.

When selecting a model for deployment, it is crucial to consider the specific requirements of the application, including the trade-offs between precision and recall, computational efficiency, and interpretability. Further tuning and validation may enhance the