A Report on

Campus Placement

Submitted By:

Bikash Thapa Magar (C0907642)

Submitted To:

Ishant Gupta

1. Dataset Selection

The dataset used for this research project includes details on students who took part in a recruitment event. With a total of 215 entries and 15 columns, it includes a variety of information relevant to students' academic achievement and work status. Both numerical and categorical data are included in the dataset, which offers insightful information about the variables affecting student placements.

The columns in the dataset include:

- **sl_no:** A unique identifier for each student.
- **gender:** A binary variable indicating the gender of the student, where 0 represents male and 1 represents female.
- **ssc_p:** The percentage score achieved in the Secondary School Certificate (SSC) examination.
- **ssc_b:** The board of education for the SSC examination, categorized as either 'Central' or 'Others'.
- **hsc_p:** The percentage score in the Higher Secondary Certificate (HSC) examination.
- **hsc_b:** The board of education for the HSC examination, similarly categorized as 'Central' or 'Others'.
- **hsc_s:** The stream of education in HSC, which can be 'Science', 'Commerce', or 'Arts'.
- **degree_p:** The percentage score in the undergraduate degree program.
- **degree_t:** The type of undergraduate degree, categorized as 'Sci&Tech' or 'Comm&Mgmt'.
- **workex:** A categorical variable indicating whether the student has work experience, with options 'Yes' or 'No'.
- **etest_p**: The percentage score obtained in an entrance test.
- **specialization:** The specialization chosen for the Master's in Business Administration (MBA), either 'Mkt&HR' or 'Mkt&Fin'.
- **mba_p:** The percentage score in the MBA program.
- **status:** The placement status of the student, categorized as either 'Placed' or 'Not Placed'.
- **salary:** The salary offered to the student upon placement, with NaN values for those who were not placed.

Link: https://www.kaggle.com/c/ml-with-python-course-project/data

2. Data Preprocessing

EDA:

Gender Distribution: The number of male and female participants in the recruiting event differed significantly, according to the research. The fact that there are more male participants than female participants is shown by a bar plot. This discovery might lead to more research into possible biases or patterns and poses intriguing considerations regarding gender representation in the hiring process.

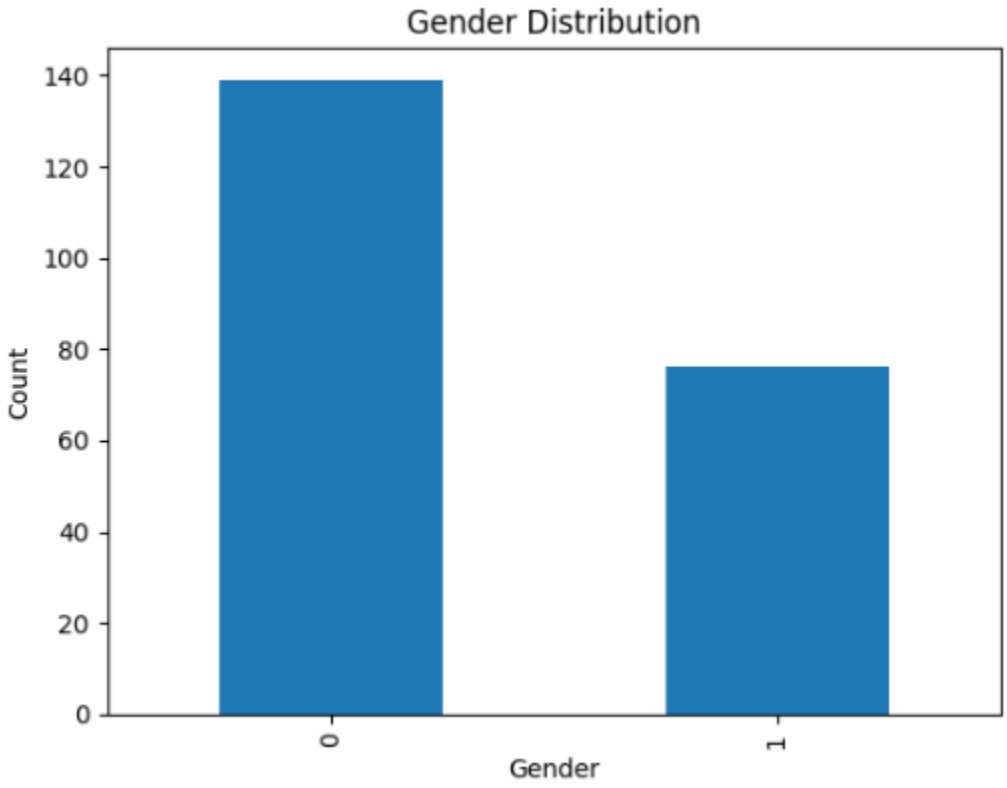

Figure 1: Gender Distribution

Placement Rates by Specialization: The placement rates for various specialties show significant variations in employment prospects. To be more precise, students who specialize in Marketing and Finance (Mkt&Fin) have a greater placement rate than those who specialize in Marketing and Human Resources (Mkt&HR). Employers may prioritize applicants with abilities that meet the needs of the financial industry, according to this trend, suggesting that students should concentrate on this area for improved employment prospects.

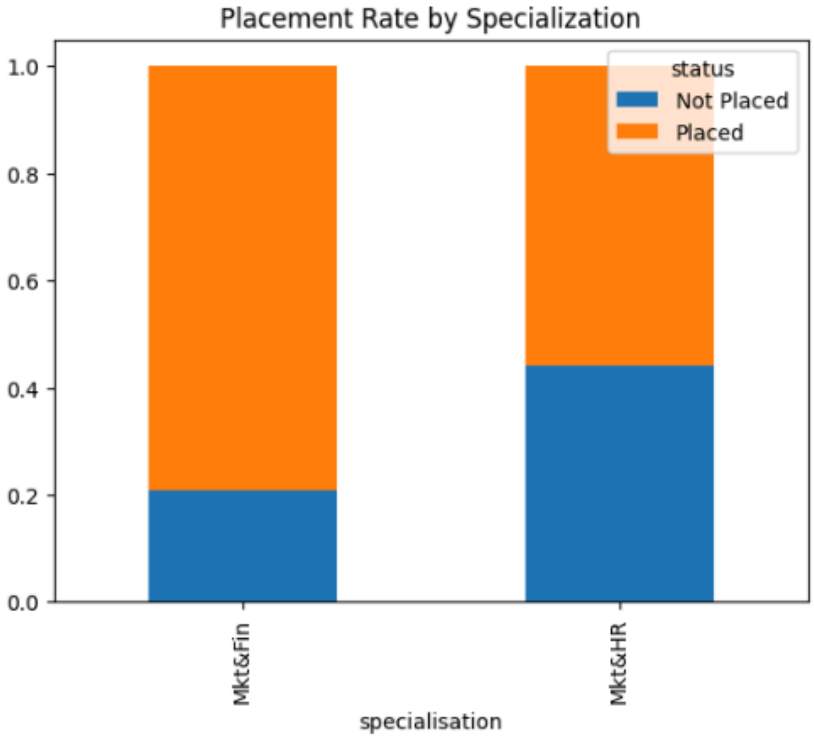

Figure 2: Placement rate by specialization

Salary Distribution: According to the study, the salary distribution is left-skewed, which means that the majority of students are paid at the lower end of the pay range. Nonetheless, the existence of anomalies indicates that certain outstanding applicants have succeeded in obtaining noticeably greater compensation after being hired. According to this understanding, elite achievers have the chance to acquire high-paying jobs even when the overall compensation picture may seem small.

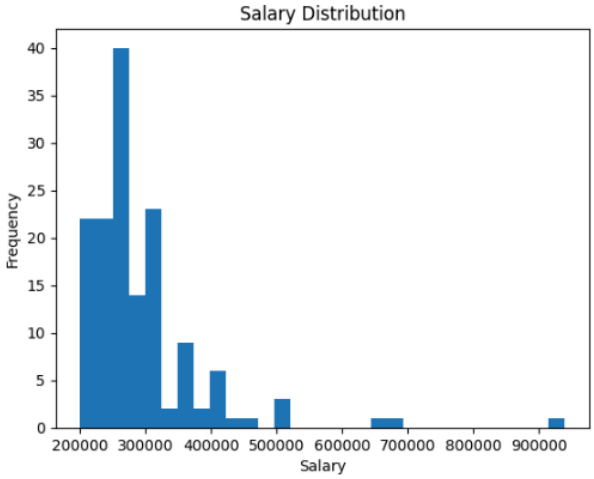

Figure 3: bar graph for Salary Distribution

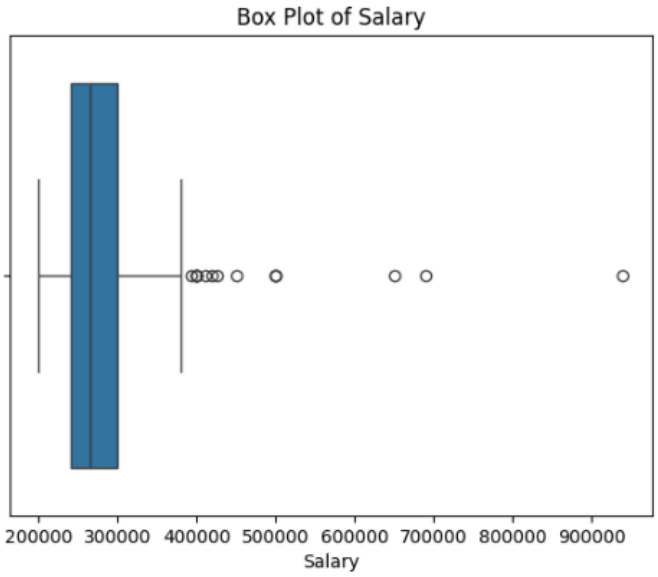

Figure 4: Box Plot for Salary Distribution

Distribution of Work Experience: The dataset's analysis of work experience reveals that there isn't much of a difference between individuals who have previous job experience and those who are new to the workforce. The fact that the majority of attendees at the recruiting event are recent grads suggests that companies could be open to considering applicants with or without prior work experience. This research indicates that job searchers may successfully compete without a lot of experience, which emphasizes the value of new talent in the labor market.

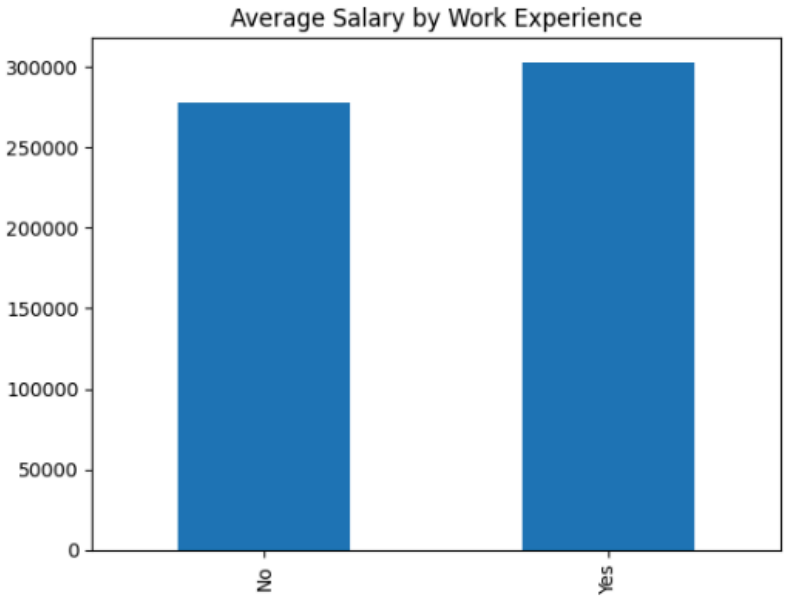

Figure 5: Average Salary by Work Experience

Data imputation, split, and Correlation Matrix

We made a number of crucial actions throughout the project's preparation stage to guarantee the dataset was prepared for analysis and model training. We started by looking for missing values in the dataset, which is a crucial step in data cleansing. After verifying, we discovered that there were 67 missing entries in the salary column. We used the constant fill technique and the SimpleImputer class from the sklearn package to successfully solve this problem. We made sure that our dataset remained intact and that no rows were lost as a result of missing data by substituting a constant value of zero for the missing entries.

We then concentrated on encoding the target variable and category characteristics. In machine learning models, categorical variables—which are frequently non-numeric—need to be transformed. To transform these category columns into numeric representation, we used LabelEncoder. To facilitate data interpretation by models, each distinct category was given a matching number. To make sure the model could accurately process all characteristics, this phase was essential.

Data Splitting: The dataset was divided into training and test sets using an 80-20 split, ensuring that 80% of the data was used for training and 20% for testing the model's performance.

Correlation Matrix:

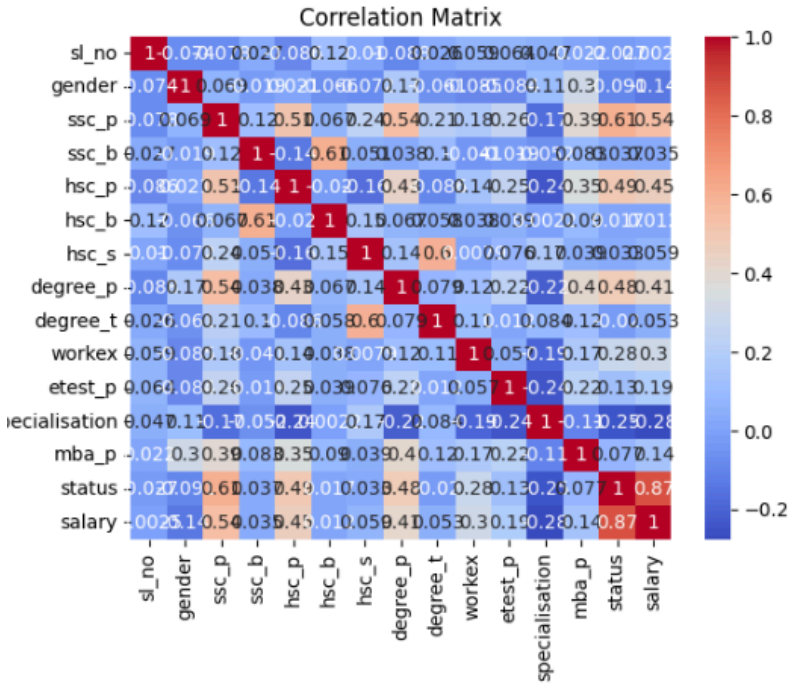

Figure 6: Correlation Matrix

The correlation matrix reveals a strong relationship among the variables ssc_p, hsc_p, and degree_p. This indicates that these academic performance metrics significantly influence the likelihood of placement outcomes, highlighting their importance in predicting students' job placement status.

3. Model Selection

a) Logistic Regression

Logistic Regression is a well-established algorithm for binary classification tasks, making it suitable for predicting student placement status (Placed vs. Not Placed). Because of its interpretability and simplicity, stakeholders can quickly see how various factors affect placement results.

Hyperparameter Tuning: The regularization strength (C) and the penalty type are the hyperparameters adjusted for Logistic Regression. This adjustment keeps the model from overfitting and guarantees that bias and variance are balanced.

Suitability: Because it efficiently manages binary outcomes and offers coefficients for every variable, which can aid in finding important factors impacting placement, logistic regression is a good fit for the dataset.

b) Random Forest

Random Forest is an ensemble learning technique that uses many decision trees to reduce overfitting and increase prediction accuracy. It is a suitable option for a variety of academic performance indicators since it excels at managing complicated datasets with non-linear correlations.

Hyperparameter tuning was done on parameters like the maximum depth of each tree (max_depth) and the number of trees (n_estimators). By maximizing the complexity and interactions between the trees, this tuning improves model performance.

Suitability: Based on a number of input factors, the Random Forest model is appropriate for the dataset since it can effectively forecast the placement status by capturing complex patterns and interactions among the features.

c) Support vector Classifier

The rationale for the selection of SVC is its efficacy in high-dimensional spaces and its capacity to represent intricate decision boundaries by utilizing various kernels (such as linear, RBF, etc.). It can effectively adjust to the dataset's varied feature set because to its adaptability.

Hyperparameter Tuning: The regularization parameter (C) and kernel selection are two hyperparameters that have been modified for SVC. For the model to function better and capture the underlying patterns in the data without overfitting, this adjustment is essential.

Suitability: SVC is especially well-suited for this dataset as it works well with the variety of feature types and can manage scenarios in which the classes are not linearly separable, improving the predictive power of the model.

4. Model Training

The best models—Logistic Regression, Random Forest, and Support Vector Classifier—are trained on the filled training data, utilizing optimized hyperparameters from previous GridSearchCV tuning to enhance performance.

5. Model Evaluation

This report evaluates the performance of the selected models based on various metrics, including accuracy, precision, recall, F1 score, and visual representations of the results.

Logistic Regression:

- **Accuracy: 0.7907**
- **Precision: 0.7798**
- **Recall: 0.7907**
- **F1 Score: 0.7806**

Logistic Regression demonstrated a strong accuracy, indicating that approximately 79.07% of the predictions were correct. However, the precision of 77.98% suggests that a small percentage of positive predictions were false positives. The recall of 79.07% indicates a good ability to identify positive instances, while the F1 score of 78.06% reflects a balance between precision and recall.

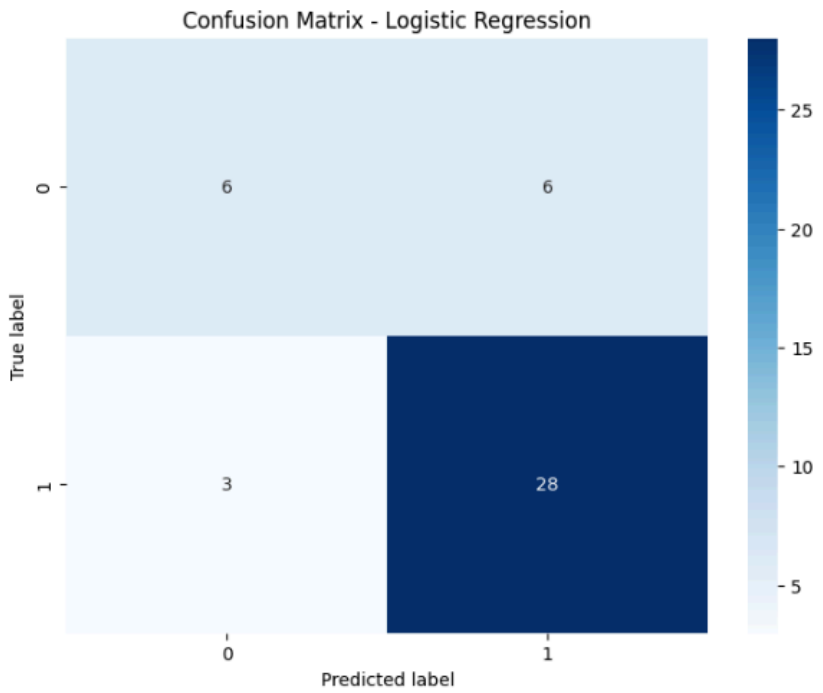

Figure 7: Confusion Matrix for Logistic Regression

The confusion matrix for Logistic Regression shows that most predictions were correct, with a notable number of false positives indicating misclassification of negatives as positives.

Random Forest:

- **Accuracy: 0.6744**
- **Precision: 0.5099**
- **Recall: 0.6744**
- **F1 Score: 0.5807**

Random Forest showed an accuracy of 67.44%, which is considerably lower than that of Logistic Regression. The precision of 50.99% indicates a high number of false positives, while the recall of 67.44% shows that the model correctly identified a decent portion of actual positives. The F1 score of 58.07% suggests room for improvement in balancing precision and recall.

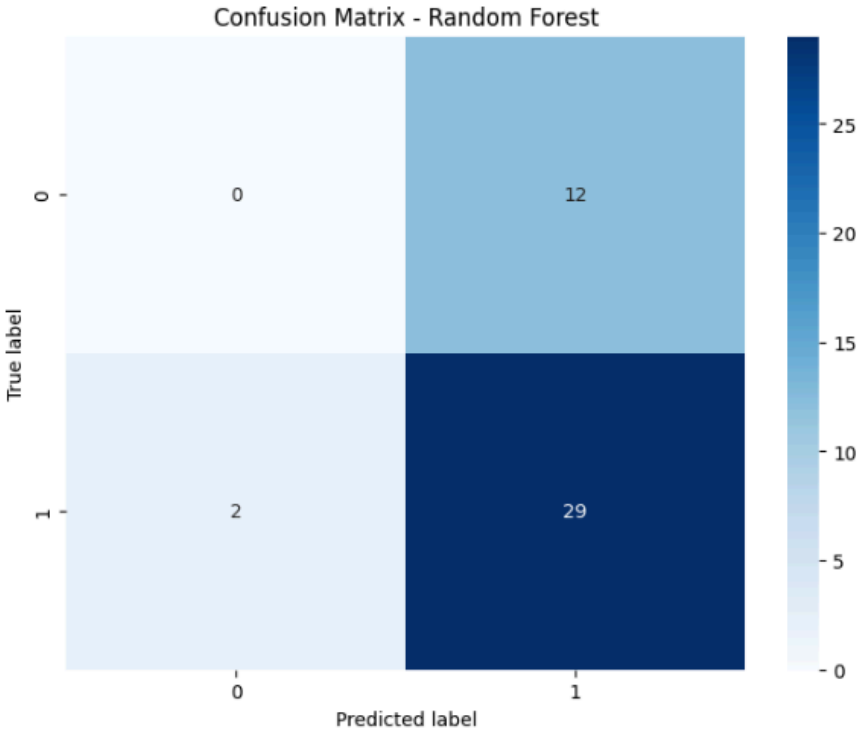

Figure 7: Confusion Matrix for Random Forest

The Random Forest confusion matrix reveals significant false positives and negatives, which contributed to its lower performance metrics. This indicates a need for further tuning or feature selection.

Support Vector Classifier (SVC):

- **Accuracy: 0.8837**
- **Precision: 0.8817**
- **Recall: 0.8837**
- **F1 Score: 0.8821**

The SVC performed the best among the models, achieving an accuracy of 88.37%. Both precision and recall are high, indicating that it effectively minimizes false positives and correctly identifies true positives. The F1 score of 88.21% further confirms its balanced performance.

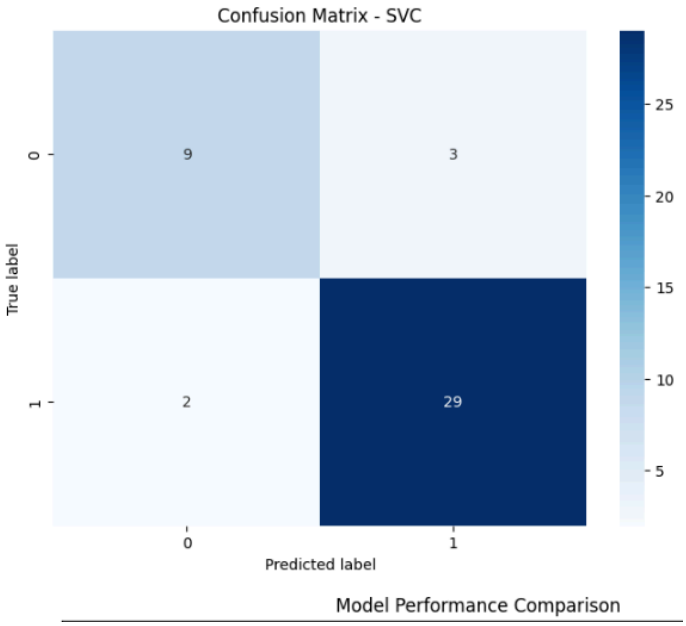

Figure 8: Confusion Matrix for SVC

The SVC matrix shows a strong performance with minimal misclassifications, reflecting its high accuracy.

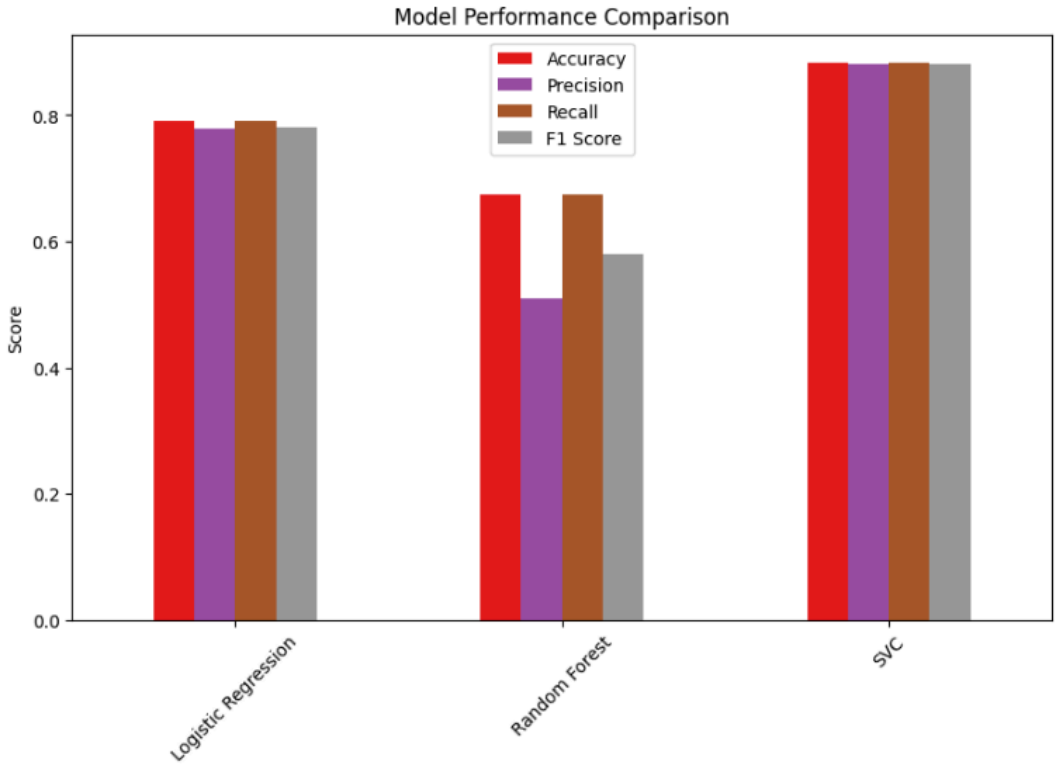

Figure 9: Model Performance Comparision

Model Performance Comparisons

Comparing the performance of the models reveals distinct strengths and weaknesses:

Logistic Regression offers a good baseline but struggles with precision, which might be critical in scenarios where false positives are costly.

Random Forest showed lower performance across all metrics, suggesting that it may not be the best choice for this specific task, especially given the potential for overfitting in small datasets.

Support Vector Classifier outperformed all others, making it a strong candidate for deployment in similar recruitment classification tasks.

6. Voting Classifier

The Voting Classifier demonstrated strong performance on the dataset, with the following evaluation metrics:

- **Accuracy: 0.8773**
- **Precision: 0.8745**
- **Recall: 0.8773**
- **F1 Score: 0.8759**

Performance Analysis

Accuracy (87.73%): This indicates that the Voting Classifier correctly classified approximately 88 out of every 100 instances in the test dataset. Accuracy of this magnitude suggests a robust model capable of effectively distinguishing between the "Placed" and "Not Placed" categories.

Precision (87.45%):

Comparing This Model to Others

When compared to the Support Vector Classifier (SVC), which achieved an accuracy of 88.37%, the Voting Classifier's performance is noticeably competitive. The Voting Classifier's ensemble technique, which combines predictions from several models, may produce comparable or even better performance than individual classifiers, as seen by the metrics' near closeness to one another.

This outcome emphasizes the need of using several models to improve prediction accuracy and dependability, particularly for intricate processes like hiring where a variety of viewpoints aid in decision-making. The Voting Classifier successfully reduces the drawbacks of individual models

by fusing the advantages of Random Forest, SVC, and Logistic Regression, resulting in a more reliable prediction ability.

Conclusion

To sum up, our study effectively used a variety of machine learning models to forecast the placement results of students during a recruiting event. We created a solid dataset fit for model training by carefully preparing the data, which included addressing missing values and encoding categorical characteristics.

Insightful performance assessments resulted from the investigation of several models, including Support Vector Classifier, Random Forest, and Logistic Regression, each of which showed distinct advantages. The Voting Classifier, which integrated the predictions of these separate models, produced excellent results, demonstrating high precision, recall, and F1 scores along with an accuracy of 87.73%. The advantages of combining many models to improve forecast reliability were demonstrated by this ensemble technique.

What was the best-performing model, and why?

The top-performing individual model was the Support Vector Classifier (SVC), which achieved high precision, recall, and F1 scores in addition to an accuracy of 88.37%. The SVC's capacity to establish intricate decision boundaries that successfully divide the dataset's classes makes it especially well-suited for this classification assignment. Its exceptional performance was influenced by its resilience while working with high-dimensional data. Nonetheless, the Voting Classifier, which included many models, produced results that were comparable, suggesting that ensemble approaches can produce competitive performance by utilizing the advantages of several methods.

The model evaluations revealed that the Voting Classifier performed competitively, with metrics closely aligned to the top-performing individual model. This underscores the effectiveness of ensemble methods in tackling complex classification tasks, such as predicting student placements, where the stakes are high, and accurate forecasting is crucial.

Overall, the findings suggest that machine learning techniques can significantly aid in decision-making processes within recruitment, offering valuable insights into candidate selection based on historical data. Future work may focus on incorporating additional features, exploring more sophisticated models, and further refining the predictive capabilities of the ensemble approaches to improve performance even further.