Campus Recruitment Prediction

Introduction

The placement of students is a critical objective for educational institutions. The reputation and yearly admissions of an institution are heavily influenced by the success of its placement efforts. Therefore, institutions strive to strengthen their placement departments to enhance overall performance. This project aims to predict whether a student will be recruited in campus placements based on various factors provided in the dataset.

Dataset Description

The dataset contains information on students, including their academic performance, personal details, and placement status. The key columns include:

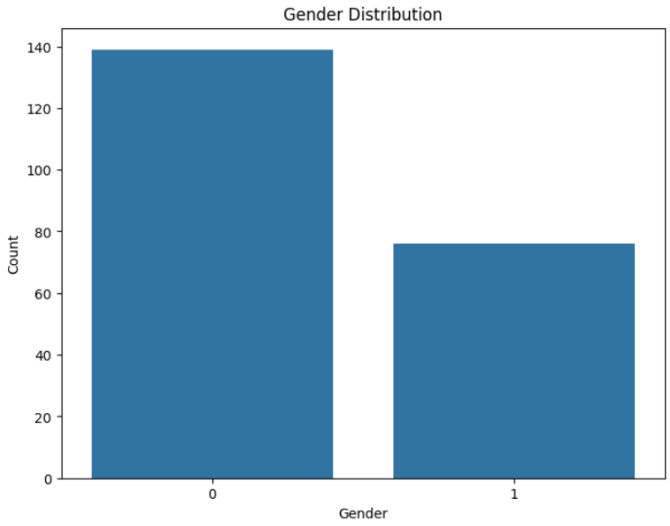
- **gender**: Gender of the student (Male/Female)
- **ssc_p**: Secondary Education percentage (10th Grade)
- **ssc_b**: Board of Education (Central/Other)
- **hsc_p**: Higher Secondary Education percentage (12th Grade)
- **hsc_b**: Board of Education (Central/Other)
- **hsc_s**: Specialization in Higher Secondary Education
- **degree_p**: Degree Percentage
- **degree_t**: Type of Degree (Science/Commerce/Arts)
- **workex**: Work Experience (Yes/No)
- **etest_p**: E-test percentage
- **specialisation**: Post-graduation (MBA) Specialization
- **mba_p**: MBA percentage
- **status**: Placement Status (Placed/Not Placed)

Data Preprocessing

1. Exploratory Data Analysis (EDA)

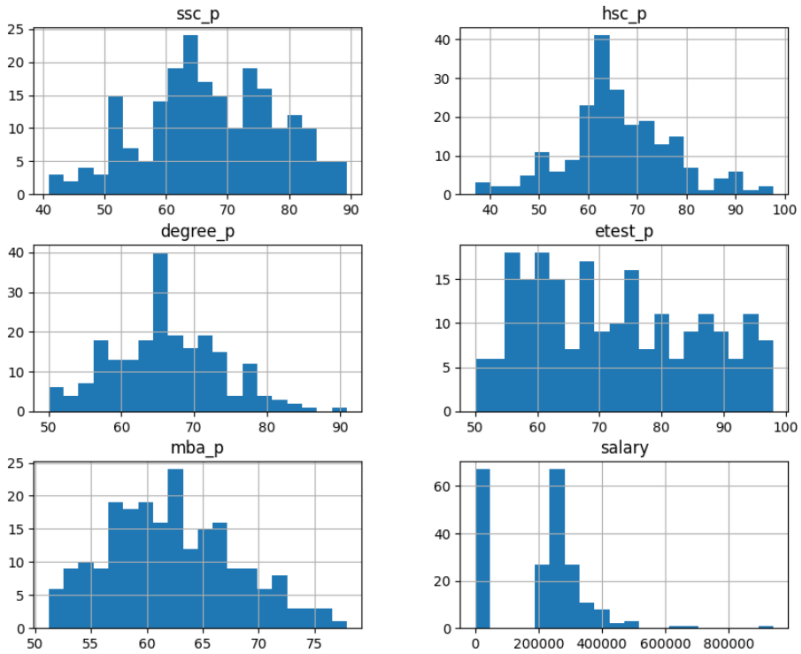
EDA helps in understanding the dataset better. The initial steps included generating statistical summaries and visualizing the distributions of numerical features. Key observations from EDA included:

- **Gender Distribution**: The dataset had a balanced gender distribution.



The bar chart visualization shows that there are more male (0) participants than female (1).

- **Academic Performance**: Histograms showed the spread of percentages across various educational stages.

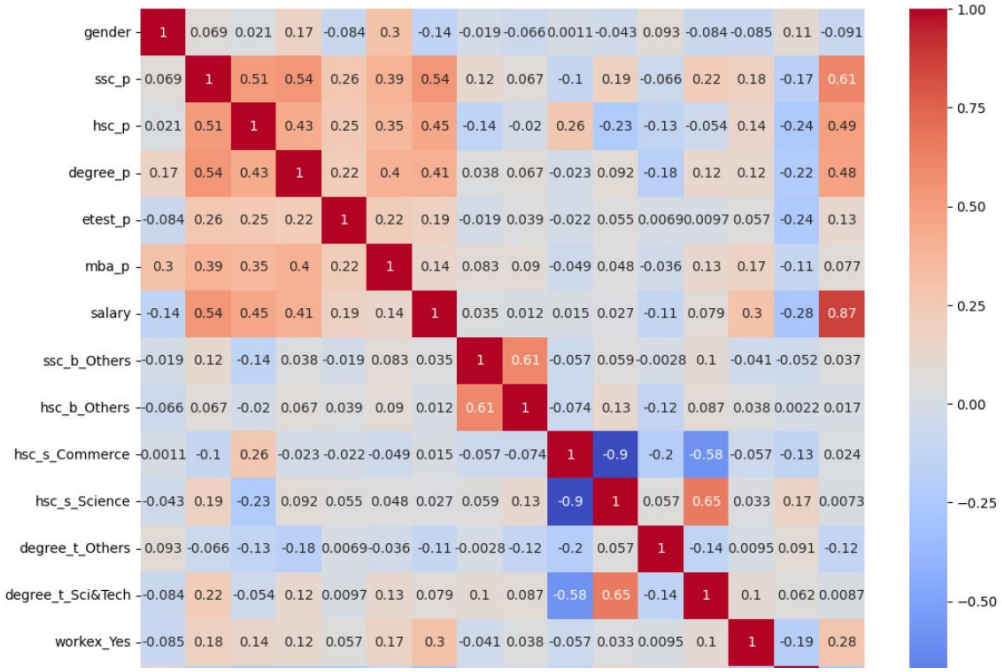


From the above visualization, we can see histograms displaying the distribution of secondary, higher secondary, degree, e-test, and MBA percentages along with salary distributions highlighting key trends and peaks in student performance and placement outcomes.

- **ssc_p**: The **ssc_p** histogram depicts that most students' secondary education percentages lie between 50%-90% with a peak around 60-70%.

- **hsc_p**: Similar to ssc_p histogram, the **hsc_p** histogram shows the distribution of higher secondary education percentages where majority of students scored between 50% and 80% with a noticeable peak around 60-70%.
- **degree_p**: The histogram of **degree_p** depicts the same trend as in **hsc_p**.
- **etest_p**: The histogram of e-test percentage shows a more widespread distribution compared to previous ones. Scores range from 50% to 100% with multiple peaks indicating varying levels of performance.
- **mba_p**: The MBA percentage distribution is similar to the degree percentage distribution, with scores primarily between 50% and 75%. There is a peak around 60-65%.
- **salary**: This histogram displays the distribution of salaries among placed students. There is a large number of students with a salary of 0, indicating that these students were not placed. For those who were placed, salaries are mostly concentrated between 200,000 and 400,000, with a few outliers earning higher amounts.

- Correlation: A heatmap revealed the correlation between different numerical features, indicating potential multicollinearity.



- ssc_p (Secondary Education percentage) and degree_p (Degree percentage) have a notable positive correlation (0.54).
- hsc_s_Commerce and hsc_s_Science have a very strong negative correlation (-0.9), reflecting that students who chose Commerce did not choose Science and vice versa.

2. Handling Missing Values

The dataset contained missing values in the salary column that were handled appropriately to ensure data quality. We checked the corresponding column named **status** and they were set to '**Not Placed**' which implied that their salary would be zero. So instead of dropping rows with null values, we imputed 0 in place of the null value.

Additionally, we dropped the sl_no column considering it was of no use for the model building.

3. Encoding Categorical Features

Categorical features ('ssc_b', 'hsc_b', 'hsc_s', 'degree_t', 'workex', 'specialisation', 'status') were encoded using one-hot encoding to convert them into numerical format suitable for machine learning models.

4. Data Splitting

The dataset was split into training and testing sets in a 70:30 ratios to evaluate model performance on unseen data.

Model Selection

Three different models were chosen for this classification task:

- **Logistic Regression**: A simple and interpretable linear model.
- **Random Forest Classifier**: An ensemble model that provides robustness and handles non-linear relationships.
- **Support Vector Classifier (SVC)**: A powerful model for classification tasks, particularly with a clear margin of separation.

These models were selected based on their suitability for the dataset and the classification task at hand.

Model Training

Each model was trained on the training data. The training process was documented to ensure reproducibility and clarity.

Model Evaluation

The models were evaluated using metrics such as accuracy, precision, recall, and F1-score. Confusion matrices were also generated to provide deeper insights into the model performance.

Voting Classifier

A voting classifier was implemented to combine the predictions from multiple models, aiming to improve overall performance.

Conclusion

The project successfully predicted campus placements using various machine learning models. Logistic Regression, Random Forest, and SVM models were trained and evaluated. The Voting Classifier provided an ensemble approach, combining the strengths of individual

models. The final evaluation metrics and confusion matrices highlighted the models' performance, with the Voting Classifier showing promising results.