Campus Placement Prediction: Leveraging Machine Learning for Student Career Success



**Student Name:** Sai Nikhil Sadhula

**Batch:** AI Elite 16



**Project Name:** Campus Placement Prediction

**Domain:** Human Resources

**Type of ML:** Supervised Learning

**Type of Problem:** Classification

**Project Methodology:** CRISP-ML (Q)



**Phase I: Business and Data Understanding**

**a) Business Understanding:  
  
Business Objective**: The objective of this project is to predict whether a student will be placed or not based on various factors such as their academic performance, gender, work experience, etc.

**Constraints:**

* Limited dataset size (215 entries).
* Imbalanced dataset (more placed candidates than not placed).
* Incomplete salary data (missing values in the 'salary' column)

**Success Criteria:**

1. *ML success criteria: Achieve an average accuracy of 85%.*
2. *Business Success criteria: Increase the efficiency of campus placement processes.*
3. *Economy Success criteria: N/A*

**b) Data Understanding:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Feature Name** | **Data Type** |
| 1 | Gender | Categorical |
| 2 | SSC Percentage | Continuous |
| 3 | SSC Board | Categorical |
| 4 | HSC Percentage | Continuous |
| 5 | HSC Board | Categorical |
| 6 | HSC Specialization | Categorical |
| 7 | Degree Percentage | Continuous |
| 8 | Degree Type | Categorical |
| 9 | Work Experience | Categorical |
| 10 | Employability Test Percentage | Continuous |
| 11 | MBA Specialization | Categorical |
| 12 | MBA Percentage | Continuous |
| 13 | Status | Categorical |
|  |  |  |

**Phase 2:  Data Preparation  
  
a) Exploratory Data Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type** | **Feature Names** | **Observation** |
| 1 | Missing Values | NA | 0 |
| 2 | Duplicates | NA | 0 |
| 3 | Outliers | hsc\_p, degree | 8,1 |
| 4 | Distributions | ssc\_p, hsc\_p, degree\_p, etest\_p, mba\_p | skewed |
| 5 | Noisy data | - | - |
|  |  |  |  |

**b)  Data Cleaning/wrangling:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Type of Cleaning** | **Technique** | **Feature Name** | **Reason** |
| 1 | Missing value | Imputing with mean | Employability Test Percentage, MBA Percentage | No outliers |
| 2 | Encoding | One hot | Categorical variables | Nominal |
| 3 | Scaling | Standard Scaling | Numerical features | Robustness |

**Phase 3: Model Building:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Type of Problem** | **Approach** | **Algorithm Name** |
| 1 | Classification | Probability based | Multinomial Naive Bayes |
|  |  |  |  |
|  |  |  |  |

**Phase 4: Model Evaluation:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Metric Name** | **Metric Score** |
| 1 | Accuracy | 86 |
| 2 | Precision | 86 |
| 3 | Recall | 97 |
| 4 | F1-score | 91 |
|  |  |  |

**Phase 5: Model Deployment:  
  
Deployment Platform:**    Stream lit **Link/URL:**provide the link

A Comprehensive Analysis

Data Preparation

a) Exploratory Data Analysis (EDA):

* Missing Values: No missing values were observed.
* Duplicates: No duplicate entries were found.
* Outliers: Outliers were present in some features.
* Distributions: Some features showed skewed distributions.
* Noisy Data: No noisy data was observed.

b) Data Cleaning/Wrangling:

* Missing Value Imputation: Missing values in 'Employability Test Percentage' and 'MBA Percentage' were imputed with the mean since there were no outliers.
* Encoding: Categorical variables were encoded using Ordinal Encoder.
* Scaling: Numerical features were scaled using Standard Scaling for better performance with KNN algorithm.

Model Building

a) Approach:

* Classification: The problem was treated as a classification task since the objective was to predict the placement status of students.

b) Algorithm Selection:

* Multinomial Naive Bayes (NB): Chosen as the algorithm for its simplicity, efficiency, and suitability for categorical features.

Model Evaluation

a) Evaluation Metrics:

* Accuracy Score: Multinomial NB achieved an accuracy score of 86% on the test data.

b) Hyperparameters:

* Default Parameters: The model was trained using default hyperparameters.

Conclusion

* The dataset underwent meticulous cleaning to address missing values, duplicates, and outliers. Imputation and scaling techniques were effectively employed, ensuring data readiness for modelling.
* Multinomial Naive Bayes was chosen as the preferred classification algorithm for its simplicity and efficiency. Despite assumptions of feature independence, it proved effective in predicting placement status, given the dataset's categorical features.
* The Multinomial NB model demonstrated commendable performance, achieving an accuracy score of 86%, meeting the set success criteria. The confusion matrix provided valuable insights into the model's predictive capabilities, affirming its effectiveness in predicting placement status.
* Future efforts should focus on exploring additional feature engineering techniques, hyperparameter tuning, and ensemble methods to further optimize model accuracy and reliability. Additionally, collecting more data to address imbalanced datasets would enhance model robustness.