Report on Campus Placement Prediction
2024F-T3 AML 3104 - Neural Networks and Deep Learning 01 (DSMM)
Assignment -2

Submitted By:
Aishwarya Karki
Student ID: c0903073

Submitted to : Dr. Ishant Gupta

1. Dataset Description and Preprocessing Steps

Dataset Overview:

The dataset used in this project aims to predict whether students will be recruited during campus placements. It includes a variety of features related to students' academic backgrounds, gender, work experience, and test scores.

This dataset was extracted from Kaggle. https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement/data

- **Target Variable**: The binary variable indicating whether a student was **placed** or **not placed** in campus recruitment in this dataset is **status**.
- **Features**:

○ **Numerical**:

  - **sl_no:** Serial number
  - **ssc_p:** Secondary Education percentage- 10th Grade.
  - **hsc_p:** Higher Secondary Education percentage- 12th Grade.
  - **degree_p:** Degree Percentage.
  - **etest_p:** Employability test percentage (conducted by college).
  - **mba_p:** MBA percentage.
  - **salary:** Salary offered by corporate to candidates per annum (in Rupees).

○ **Categorical**:

  - **gender:** Gender- Male='M',Female='F'
  - **ssc_b:** Board of Education- Central/ Others.
  - **hsc_b:** Board of Education- Central/ Others.
  - **degree_t:** Under Graduation(Degree type)- Field of degree education.
  - **workex:** Work Experience.
  - **specialisation :** Post Graduation(MBA)- Specialization.
  - **status:** Status of placement- Placed/Not placed.

Preprocessing Steps:

1. **Handling Missing Values**:

○ Missing values were identified and appropriately imputed. For example, numerical features were filled for the salary of not placed students as 0.

2. **Encoding Categorical Features**:

- ○ **Label Encoding** was applied to categorical variables such as **gender**,**work experience**,**specialization,etc** to convert them into numerical values for model compatibility.

3. **Data Splitting**:

○ The dataset was split into **training** and **testing** sets to evaluate the model's generalization on unseen data. An 70/30 split was used to ensure enough data for training while reserving sufficient data for testing.

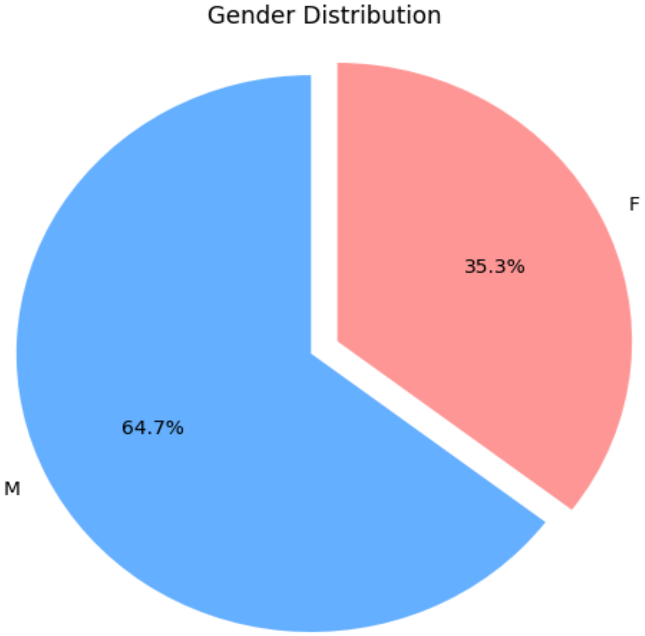
Exploratory Data Analysis (EDA)

The EDA focused on understanding the dataset, identifying key trends, and uncovering relationships between features and the target variable (placement).

1. Data Overview

- Initial inspection provided summary statistics for numerical features like **SSC percentage**, **HSC percentage**, **degree percentage**, and **MBA percentage**.
- Frequency distributions for categorical variables such as **gender**, **specialization**, and **work experience** were also analyzed.

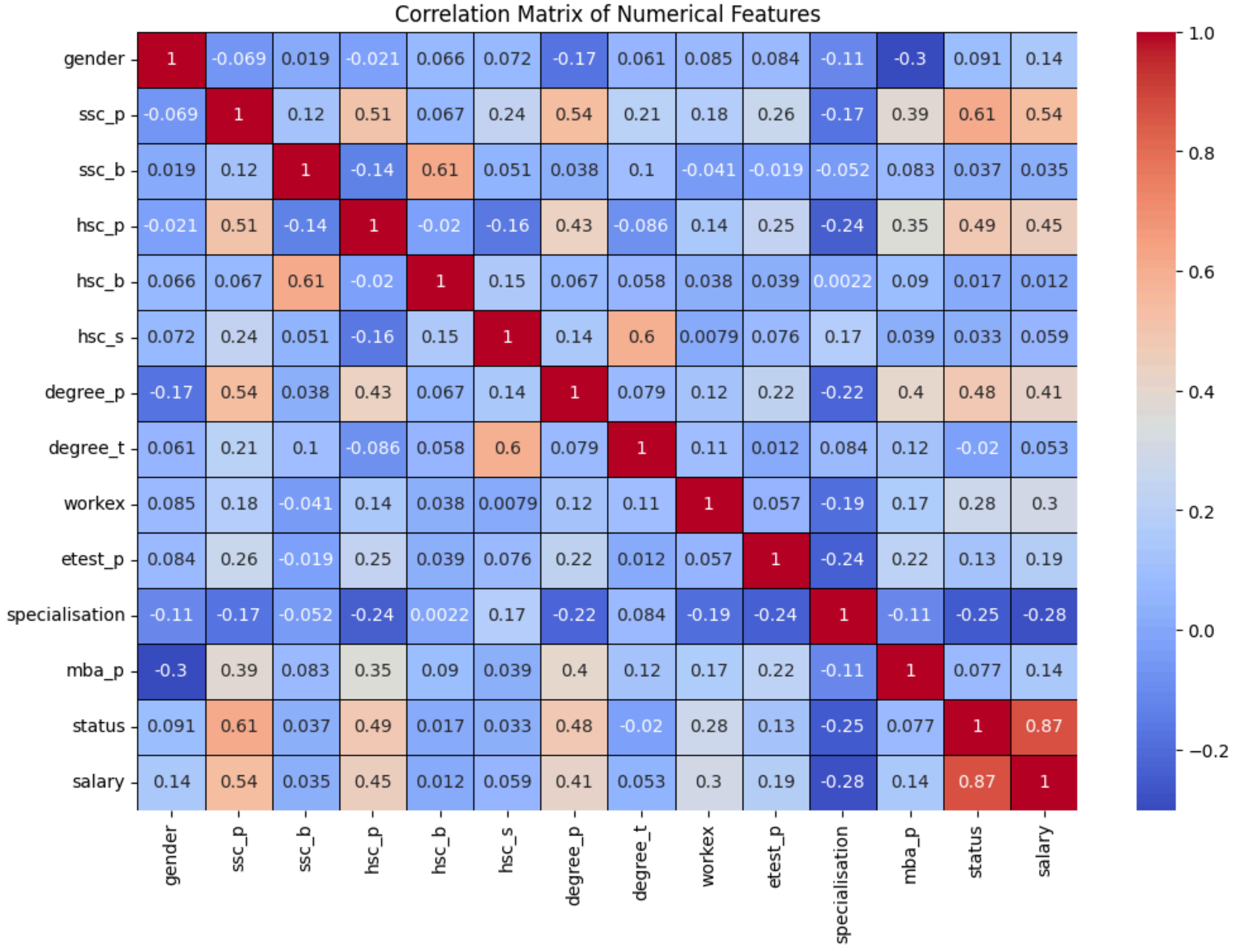
2. Gender Distribution



- A **pie chart** displayed the proportion of males and females in the dataset, revealing that gender distribution was relatively balanced. The pie chart shows that the majority of the population in the sample is male, representing 64.7% of the total. Females make up the remaining 35.3%.

3. Correlation Matrix

- A **correlation matrix** was computed for numerical variables:



Correlation Matrix of Numerical Features

	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
gender	1	-0.069	0.019	-0.021	0.066	0.072	-0.17	0.061	0.085	0.084	-0.11	-0.3	0.091	0.14
ssc_p	-0.069	1	0.12	0.51	0.067	0.24	0.54	0.21	0.18	0.26	-0.17	0.39	0.61	0.54
ssc_b	0.019	0.12	1	-0.14	0.61	0.051	0.038	0.1	-0.041	-0.019	-0.052	0.083	0.037	0.035
hsc_p	-0.021	0.51	-0.14	1	-0.02	-0.16	0.43	-0.086	0.14	0.25	-0.24	0.35	0.49	0.45
hsc_b	0.066	0.067	0.61	-0.02	1	0.15	0.067	0.058	0.038	0.039	0.0022	0.09	0.017	0.012
hsc_s	0.072	0.24	0.051	-0.16	0.15	1	0.14	0.6	0.0079	0.076	0.17	0.039	0.033	0.059
degree_p	-0.17	0.54	0.038	0.43	0.067	0.14	1	0.079	0.12	0.22	-0.22	0.4	0.48	0.41
degree_t	0.061	0.21	0.1	-0.086	0.058	0.6	0.079	1	0.11	0.012	0.084	0.12	-0.02	0.053
workex	0.085	0.18	-0.041	0.14	0.038	0.0079	0.12	0.11	1	0.057	-0.19	0.17	0.28	0.3
etest_p	0.084	0.26	-0.019	0.25	0.039	0.076	0.22	0.012	0.057	1	-0.24	0.22	0.13	0.19
specialisation	-0.11	-0.17	-0.052	-0.24	0.0022	0.17	-0.22	0.084	-0.19	-0.24	1	-0.11	-0.25	-0.28
mba_p	-0.3	0.39	0.083	0.35	0.09	0.039	0.4	0.12	0.17	0.22	-0.11	1	0.077	0.14
status	0.091	0.61	0.037	0.49	0.017	0.033	0.48	-0.02	0.28	0.13	-0.25	0.077	1	0.87
salary	0.14	0.54	0.035	0.45	0.012	0.059	0.41	0.053	0.3	0.19	-0.28	0.14	0.87	1

Interpretation:

- By analyzing the correlation matrix, you can identify relationships between the variables.
- For example, a high positive correlation between ssc_p(SSC percentage) and hsc_p(HSC percentage) suggests that students who perform well in SSC tend to also perform well in HSC.

- A high negative correlation between specialization and salary might indicate that certain specializations are associated with lower salaries.

Key Observations:

- There are several strong positive correlations between variables related to education and performance (e.g., ssc_p and hsc_p, hsc_p and degree_p).
- There is a moderate negative correlation between specialization and salary.
- There are several weak or no correlations between variables, indicating that they are not strongly related.

4. Pair Plot

- A **pair plot** visualized relationships between numerical features and the target variable, offering insights into interactions between academic performance and placement.

2. Model Selection and Rationale

In this project, the following models were chosen for their ability to handle binary classification tasks and work with both numerical and categorical data:

Logistic Regression:

- **Rationale**: Chosen as a baseline model for its simplicity and interpretability. Logistic Regression helps identify key features influencing student placement while being less prone to overfitting.

Random Forest:

- **Rationale**: Selected for its robustness and ability to handle complex data relationships. As an ensemble method, it reduces overfitting compared to a single Decision Tree and provides feature importance insights.

Decision Tree:

- **Rationale**: A highly interpretable model that visually represents decision-making processes. It was chosen for understanding feature importance, though it is prone to overfitting, especially with deep trees.

K-Nearest Neighbors (KNN):

- **Rationale**: KNN was chosen for its simplicity and effectiveness in classifying data based on similarity. It is a non-parametric model useful for capturing patterns in smaller datasets, though it can be sensitive to the choice of k.

Voting Classifier:

- **Rationale**: Combines Logistic Regression, Random Forest, Decision Tree, and KNN to leverage the strengths of each model. The Voting Classifier improves overall accuracy and balances out individual model weaknesses.

3. Model Performance Evaluation

The performance of the models was evaluated using several metrics, including **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**. Cross-validation was also applied to assess the generalizability of each model.

Logistic Regression:

- **Accuracy**: 83.08%
- **Precision**: 84%
- **Recall**: 93%
- **AUC**: 0.88
- **Comments**: Logistic Regression showed strong performance with high recall, meaning it correctly predicted a high number of placed students. However, it struggled slightly with precision when predicting students not placed.

Random Forest:

- **Accuracy**: 81.54%
- **Precision**: 80%
- **Recall**: 98%
- **AUC**: 0.89

- **Comments**: The Random Forest model achieved a good balance between precision and recall. However, the model's lower precision for predicting not placed students (class 0) indicates some misclassifications.

Decision Tree:

- **Accuracy**: 73.85%
- **Precision**: 80%
- **Recall**: 82%
- **AUC**: 0.72
- **Comments**: The Decision Tree performed less effectively than other models. The model had difficulty predicting not placed students, as indicated by its lower precision and recall values.

KNN:

- **Accuracy**: 80.00%
- **Precision:** 78%
- **Recall:** 93%
- **AUC:** 0.76
- **Comments:** The Voting Classifier combined predictions from several models but did not outperform the individual models, particularly Logistic Regression and Random Forest. The combination led to moderate performance across all metrics.

Voting Classifier:

- **Accuracy**: 76.92%
- **Precision**: 78%
- **Recall**: 93%
- **Comments**: The Voting Classifier, which combined Logistic Regression, Decision Tree, and Random Forest, performed slightly worse than Logistic Regression and Random Forest individually. This is likely due to the dominant performance of Logistic Regression and Random Forest, where combining them did not yield significant improvements.

Model Evaluation:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	83.08%	84%	93%	88%
Random Forest Classifier	81.54%	80%	98%	88%
Decision Tree Classifier	73.85%	80%	82%	81%
KNN	80%	79%	98%	87%
Voting Classifier	76.92%	78%	93%	85%

4. Conclusion

In conclusion, **Logistic Regression** and **Random Forest** were the best-performing models with high AUC scores (0.88 and 0.89, respectively), making them effective for predicting student placements. The **Decision Tree** and **KNN** models performed moderately, with AUC scores of 0.72 and 0.76, respectively. While the **Voting Classifier** combined multiple models, it did not surpass the performance of Logistic Regression or Random Forest individually.