A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Yumiko Maria Bejarano Azogue |
| *Student Number* | 2024144 |
| *Module Title* | Machine Learning (10 ETCS) |
| *Assessment Title* | CA 1 – Capstone Project Proposal |
| *Assessment Due Date* | 21st April 2024 |
| *Date of Submission* | 21st April 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Abstract

[NOTE: This section is designated for the abstract. Abstracts are not assigned page numbers and should precede the table of contents. If an abstract is unnecessary for your work, please delete this page.]

Attention: All notes must be removed from the document before submission!!

Contents

[Student Placement Analysis 4](#_Toc164248508)

[1. Introduction 4](#_Toc164248509)

[2. Objectives 4](#_Toc164248510)

[3. Methodology 4](#_Toc164248511)

[Characterization of Data 4](#_Toc164248512)

[Pre-processing 4](#_Toc164248513)

[Techniques for Varying Accuracy Across Training Splits 4](#_Toc164248514)

[Description of Variation in Accuracy Across Three Training Splits using Cross-Validation Techniques 5](#_Toc164248515)

[4. Data Analysis (Characterization of Data) 5](#_Toc164248516)

[4.1 Data Description Data Description 5](#_Toc164248517)

[4.2 Summary Statistics: 6](#_Toc164248518)

[4.3 Exploratory Data Analysis 6](#_Toc164248519)

[5. Pre-processing 7](#_Toc164248520)

[5.1 Handling Missing Values: 7](#_Toc164248521)

[5.2 Encoding Categorical Variables: 7](#_Toc164248522)

[5.3 Normalization/Standardization: 7](#_Toc164248523)

[6.1 Data Splitting: 7](#_Toc164248524)

[6.2 Model Selection: 7](#_Toc164248525)

[Results Obtained 8](#_Toc164248526)

[Correlation Matrix 8](#_Toc164248527)

[Predictive Models 8](#_Toc164248528)

[Conclusion 8](#_Toc164248529)

[References 8](#_Toc164248530)

# Student Placement Analysis

## Introduction

In this project, is about to predict whether students will be placed in jobs based on various factors using machine learning techniques. Job placement is crucial for students and educational institutions as it directly impacts the perceived quality of education. The goal is to develop predictive and classification models that can accurately anticipate student job placement.

## Objectives

The primary objectives of this project are as follows:

* Analyze the demographic distribution of students based on age and gender.
* Investigate the relationship between academic performance (CGPA) and placement status.
* Explore the impact of internship experiences and hostel accommodation on placement outcomes.
* Examine the correlation between different variables in the dataset.
* Develop predictive models, including neural network, decision tree, random forest, logistic regression, and support vector machine, to predict students' placement.

## Methodology

### Characterization of Data

* Data Description: Begin by describing the dataset, including its size, structure, and variables. This involves examining the types of variables (numeric or categorical), their distributions, and any missing or erroneous values.
* Summary Statistics: Calculate descriptive statistics for numerical variables (e.g., mean, median, standard deviation) and frequency tables for categorical variables. This provides an initial understanding of the data's central tendencies and variability.

### Pre-processing

* Handling Missing Values: Address any missing data by imputation or removal, depending on the extent and nature of the missingness.
* Encoding Categorical Variables: Convert categorical variables into a numerical format suitable for analysis, such as one-hot encoding or label encoding.
* Normalization/Standardization: Scale numerical features to a similar range to prevent variables with larger magnitudes from dominating the analysis.

### Techniques for Varying Accuracy Across Training Splits

* Data Splitting: Divide the dataset into training, validation, and testing sets. Typically, an 80-20 or 70-30 split is used for training and testing, respectively.
* Model Selection: Choose appropriate machine learning algorithms for classification tasks, considering factors such as the dataset size, complexity, and interpretability of models.
* Training Models: Train multiple models using the chosen algorithms on the training data. Common algorithms include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.

### Accuracy Variation Across Three Training Splits with Cross-Validation

* Evaluation Metrics: Assess model performance using relevant evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
* Cross-Validation: Employ k-fold cross-validation to validate model performance across different subsets of the data. This technique divides the data into k subsets (folds), trains the model on k-1 folds, and validates it on the remaining fold. Repeat this process k times, rotating the validation fold each time. Calculate the average performance metrics across all folds to obtain a robust estimate of model performance.
* Analysis of Variation in Accuracy: Examine variations in model accuracy across different training splits and cross-validation folds. Identify any patterns or trends in model performance and assess the stability and generalizability of the models.
* Root Cause Analysis: Investigate factors contributing to variations in accuracy, such as data imbalance, feature importance, model complexity, hyperparameter tuning, and inherent randomness in the data splitting process.

## Data Analysis (Characterization of Data)

### 4.1 Data Description Data Description

This section initiates a comprehensive exploration of the dataset, unveiling its size, structure, and constituent variables. The dataset encompasses information on 2966 engineering students, with eight attributes shedding light on students' academic journeys and job placement outcomes.

The dataset's structure is as follows:

* Size: The dataset comprises 2966 entries, each corresponding to an individual student.
* Variables: It includes eight variables, encompassing six numerical and two categorical variables.
* The variables and their descriptions are:
* Age: Represents each student's age.
* Gender: Indicates the gender of each student.
* Stream: Denotes the field of study or specialization pursued by each student.
* Internships: Quantifies the number of internships completed by each student.
* CGPA (Cumulative Grade Point Average): Measures each student's academic performance.
* Hostel: Indicates whether each student resides in hostel accommodations (binary: 1 for yes, 0 for no).
* HistoryOfBacklogs: Indicates whether each student has a history of failed courses (binary: 1 for yes, 0 for no).
* PlacedOrNot (Target Variable): Indicates whether each student has secured employment post-graduation (binary: 1 for yes, 0 for no).

### 4.2 Summary Statistics:

Following the data description, this section proceeds to calculate descriptive statistics for numerical variables and frequency tables for categorical variables. The objective is to gain an initial understanding of the dataset's central tendencies and variability.

The summary statistics for numerical variables are as follows:

* Age: The mean age of the students is approximately 21.49 years, with a standard deviation of 1.32 years. The age distribution ranges from 19 to 30 years, indicating a relatively narrow age range among the student cohort.
* Internships: On average, students completed approximately 0.70 internships, with a standard deviation of 0.74. The distribution of internships completed ranges from 0 to 3, highlighting variations in students' practical industry experiences.
* CGPA: The mean CGPA of students is approximately 7.07, with a standard deviation of 0.97. CGPA scores range from 5.0 to 9.0, reflecting variations in academic performance among students.

For categorical variables, frequency tables were generated to elucidate the distribution of students across different categories, such as gender and stream.

### 4.3 Exploratory Data Analysis

#### Age Distribution:

The age distribution of students in the dataset reveals that the average age is approximately 21.5 years, with most falling between 19 and 30 years of age.

#### Internship Experience:

Internships serve as crucial opportunities for students to gain real-world experience. On average, students have completed approximately 0.7 internships during their academic tenure.

#### Academic Performance (CGPA):

The Cumulative Grade Point Average (CGPA) serves as an indicator of academic performance. On average, students have a CGPA of around 7.1, with scores ranging from 5 to 9.

#### Hostel Accommodation:

Data regarding students' living arrangements indicate that about 27% reside in hostels, while the majority live off-campus.

#### History of Backlogs:

A history of backlogs, representing previously failed courses, offers insights into students' academic challenges. Approximately 19% of students have encountered backlogs during their academic journey.

#### Job Placement Status:

The primary focus revolves around students' job placement status. Out of the total 2966 students, approximately 55% have successfully secured employment post-graduation.

#### Correlation Analysis:

A correlation analysis was conducted to understand the relationships between different variables. Key findings include:

* Age exhibits a slight negative correlation with CGPA (-0.12).
* Internships demonstrate a slight positive correlation with job placement (0.18).
* CGPA shows a moderate positive correlation with job placement (0.59).

#### Model Performance:

Various machine learning models were employed to predict job placement outcomes. The Support Vector Machine (SVM) model emerged as the top performer, achieving an accuracy rate of approximately 88%.

## Pre-processing

### 5.1 Handling Missing Values:

Addressing missing data is a crucial step in ensuring the integrity and accuracy of the analysis. In our dataset, no missing values were detected across any of the variables, as confirmed during the data exploration phase. Therefore, no imputation or removal of missing values is required.

### 5.2 Encoding Categorical Variables:

To facilitate the analysis, categorical variables need to be encoded into a numerical format. Two categorical variables exist in our dataset: "Gender" and "Stream."

* **Gender Encoding:** The "Gender" variable is binary, with two categories: Male and Female. It was encoded using label encoding, where Male was assigned the value 0 and Female was assigned the value 1. This transformation allows the algorithm to interpret gender as numerical data while preserving the ordinal relationship between the categories.
* **Stream Encoding:** The "Stream" variable represents the field of study or specialization pursued by each student. Since there are multiple categories within this variable, one-hot encoding was employed. This technique creates binary columns for each category, where a value of 1 indicates the presence of that category, and 0 indicates its absence. By encoding "Stream" in this manner, we avoid imposing false ordinality or hierarchy among the different fields of study.

### 5.3 Normalization/Standardization:

Normalization and standardization are essential preprocessing steps to ensure that numerical features are on a similar scale, preventing variables with larger magnitudes from disproportionately influencing the analysis.

* **Normalization:** The numerical variables "Age," "Internships," and "CGPA" were normalized to a range between 0 and 1 using min-max scaling. This transformation preserves the relative relationships between data points within each variable while scaling them to a uniform range.
* **Standardization:** Although not explicitly performed in our analysis, standardization (z-score normalization) could be applied if the distribution of a variable significantly deviates from a normal distribution. This technique transforms the data to have a mean of 0 and a standard deviation of 1, making it suitable for algorithms that assume normally distributed data.

## 6. Techniques for Varying Accuracy Across Training Splits:

### 6.1 Data Splitting:

Data splitting involves partitioning the dataset into distinct subsets for training, validation, and testing. This ensures that the model's performance is evaluated on unseen data, providing a more accurate assessment of its generalization capabilities.

In our analysis, we followed a standard data splitting approach:

* **Training Set:** Approximately 80% of the dataset was allocated to the training set. This portion of the data was used to train the machine learning models.
* **Validation Set:** A smaller subset (typically 20% of the dataset) was reserved for the validation set. This set was used to fine-tune model hyperparameters and assess performance during training.
* **Testing Set:** The remaining portion of the dataset, approximately 20% in our case, constituted the testing set. This set was kept completely separate from the training and validation data and was used to evaluate the final model's performance.

By splitting the data in this manner, models are trained on one subset, validated on another, and ultimately evaluated on a completely independent subset, minimizing the risk of overfitting and providing a robust measure of performance..

### 6.2 Model Selection:

Selecting the most appropriate machine learning algorithms is a critical aspect of the analysis. Several factors must be considered when choosing models, including the dataset's size, complexity, and the interpretability of the algorithms.

In the study, a variety of classification algorithms commonly used in predictive modeling tasks were employed:

* **Logistic Regression:** This linear model is well-suited for binary classification tasks and provides interpretable results. It's particularly useful when the relationship between the features and the target variable is linear.
* **Decision Trees:** Decision trees partition the feature space into distinct regions based on feature values, making them intuitive and easy to interpret. However, they can be prone to overfitting, especially with complex datasets.
* **Random Forests:** Random forests mitigate the overfitting tendency of decision trees by aggregating the predictions of multiple trees. They offer improved robustness and performance, making them suitable for a wide range of classification tasks.
* **Support Vector Machines (SVM):** SVMs are powerful classifiers capable of handling complex decision boundaries. They work well in high-dimensional spaces and are effective for both linear and nonlinear classification tasks.
* **Neural Networks:** Neural networks, particularly deep learning architectures, are highly flexible models capable of learning intricate patterns in the data. They excel in capturing nonlinear relationships but may require significant computational resources and data preprocessing.

By leveraging this diverse set of algorithms, insights into the dataset's nuances are gained, identifying the most effective model for predicting job placements among engineering students.

### 6.3 Training Models:

## Results Obtained

### Correlation Matrix

CGPA demonstrates a strong positive correlation with placement status, while hostel accommodation and history of backlogs exhibit weaker correlations.

### Predictive Models

The predictive models yield varying accuracy levels in predicting placement status. Support vector machines achieve the highest accuracy, with an 88.05% accuracy rate on the test dataset.

## Conclusion

In conclusion, this study sheds light on the determinants of students' placement outcomes. Academic performance, as gauged by CGPA, emerges as a pivotal predictor of placement success. Additionally, internship experiences and hostel accommodation may also influence placement outcomes. The predictive models developed herein offer valuable tools for educational institutions and recruiters to gauge students' likelihood of placement, thereby facilitating more informed decision-making.

## References