## Project 2 – Milestone 1

**Topic –**

This project with aim to analyze factors that influence whether a student is placed in a job after college and build a predictive model for student placement outcomes.

**Business Problem –**

This project/ analysis will be aimed at Colleges and Universities. They continuously aim to improve graduate employability, but sometimes it can be unclear which academic or demographic factors most influence job placement success.

With this analysis I seek to answer the following questions: 'What students characteristics are most predictive of being placed?' and ' Can we build a model to predict a student's placement status based on their academic and personal data?

**Datasets –**

This data set comes from the College Student Placement data set accessed through Kaggle. It consists of the records of 10,000 college students with various academic, technical, and personal development metrics. The ones this project will focus on the most include Academic Performance, Demographics, and Placement, Status. While these are the main factors that will be analyzed. Other factors, such as practical experience, will be taken into consideration to ensure the results are as accurate as possible without overcomplicating the analysis.

**Methods –**

I will apply the following data analysis and modeling techniques in order to do this:

*Exploratory Data Analysis (EDA)* – This will include summary statistics to summarize key variables and visualization to understand the distributions and relationships.

*Data Processing* – This will deal with any missing values, encoding categorical variables, and future engineering. This may include creating new features from the existing data to capture patterns not explicitly present in this raw data set.

*Predictive Modeling* – I will use predictive modeling techniques such as Logistic Regression, Decision Trees, and Random Forest Models to classify the placement outcomes of students in the data set. Logistic regression will provide a baseline model that provides interpretable coefficients and probable outputs. The decision tree classifier will help visualize decision paths and capture nonlinear relationships. Random Forest will help

to improve performance and reduce overfitting. An XGBoost model may be included if necessary to help improve accuracy.

*Model Evaluation* – These models will be assessed using Confusion Matrix Metrics, ROC-AUC Scores, and Cross-Validation to evaluate overall performance and accuracy. The tree based models will help assess which features are most influential in determining placement outcomes.

## Ethical Considerations –

There are a few ethical considerations that should be made while conducting this analysis. One being to make sure there is as little bias in the data set as possible. We would like to ensure the data does not reflects historical or social biases. With this it is important the model is evaluated for fairness among groups while also providing accurate information. It is also important to make sure that the identity of the students remains private, if this is not taken into consideration there is a risk if personal attributes are misused.

## Challenges/ Issues –

There are also a few challenges to consider when going into this. With this data, there may be class imbalances that skew model results. We may also see limitations to the results because of the dataset size. While there are 10,000 students included in the dataset, millions of students graduate college each year. This data set size may have an impact on results. Missing or inaccurate values may also introduce some issues, however there are steps planned to limit these being used for the analysis.

## References –

National Center for Education Statistics. (2023). *Fast facts: Employment outcomes of college graduates*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/fastfacts/display.asp?id=40

Sumedh. (2022). *College student placement* [Data set]. Kaggle. https://www.kaggle.com/datasets/sumedh1507/college-student-placement?resource=download