# Project Proposal: Analyzing Education and Career Success

March 05, 2025

Tyler Leech

CSC240

## 1. Project Choice and Rationale

The chosen dataset for this project is *Education & Career Success* from Kaggle. This dataset aligns with my academic background in computer science and my interest in understanding the relationship between educational background and career success. By analyzing this dataset, I aim to uncover key patterns that may influence career outcomes based on educational attainment, skills, and other relevant factors. The findings could provide insights for students, educators, and policymakers on how education impacts professional success.

## 2. Dataset Description and Research Question

*Education & Career Success* contains information on individuals' education levels, fields of study, career paths, income levels, and other factors that may contribute to career success. The primary research question for this project is: How do various educational factors, such as degree level, field of study, and academic performance, correlate with career success metrics such as salary and job satisfaction?

This dataset has been analyzed by others, allowing for comparative insights with previous research and alternative approaches.

## 3. Team Composition

Tyler Leech; Undergraduate; Computer Science; Solo

## 4. Goal of the Analysis

The goal of this project is to apply data mining techniques to identify trends and relationships between education and career success. I aim to determine which educational factors have the most significant impact on career success; predict salary ranges based on educational and demographic variables; compare my findings with existing research on the topic.

## 5. Planned Technical Approach

The project will follow these steps:

### 5.1 Data Exploration and Visualization

- Perform initial exploratory data analysis to understand the structure and distribution of the data.
- Generate visualizations to highlight key patterns and relationships.

### 5.2 Data Preprocessing

- Handle missing values and clean the dataset.
- Normalize or encode categorical variables.
- Identify and address any outliers that may affect model performance.

### 5.3 Model Selection and Implementation

- The baseline model will be Naive Bayes for classification tasks and Linear Regression for numerical predictions.
- A more sophisticated model such as Random Forest or Support Vector Machines will be used for comparison.
- Feature selection methods such as Principal Component Analysis may be applied to improve model efficiency.

### 5.4 Model Evaluation and Comparison

- Evaluate model performance using appropriate metrics (e.g., accuracy, RMSE, R-squared).
- Compare results with similar studies to validate findings.

## 6. Role of Team Members

As I am working solo, I will be responsible for all the work.