**Project Proposal (CIS 545)**

1. All Group Members’ Names & Duties:

**Chenkang Zhang (stezhang@seas.upenn.edu)**: Responsible for data preprocessing, data modeling via machine learning, write up of Introduction/Background & Modeling & Potential Next Steps sections of the final report.

**Yinuo Xie (yinuoxie@seas.upenn.edu):** Responsible for data preprocessing, Exploratory Data Analysis of cleaned dataset, write up of Exploratory Data Analysis & Description of Challenges/Obstacles Faced & Potential Next Steps sections of the final report.

**Haoyi Cheng (haoyic@seas.upenn.edu):** Responsible for Exploratory Data Analysis of cleaned dataset, data modeling via machine learning, write up of Exploratory Data Analysis & Modeling & Potential Next Steps sections of the final report.

1. Data Source:

Software Defects Dataset: <https://www.kaggle.com/competitions/playground-series-s3e23/data>

* Estimated Row of Dataset: 101763
* Number of Features (Columns): 23

1. Objective and Value Proposition: Explain what you intend to study with your project. What is the ultimate objective? Why is this project interesting?

The objective of this project is to create a binary classification model capable of predicting the presence of defects in a C program, utilizing various features of the program as indicators. The ultimate aim is to enhance the functionality of Integrated Development Environments (IDEs) by providing users with a tool that can proactively identify potential defects, thereby promoting code accuracy and reducing debugging time. This project holds intrigue due to the potential it embodies for significantly augmenting the code development process, making it a valuable endeavor in the realm of software engineering.

1. Modeling Plan: What types of models are you considering? Classification vs. Regression? What's your target variable?

The primary focus is on employing a binary classification approach, with the intended target variable being a binary outcome (0 or 1) indicating the presence or absence of a defect in the provided program. Specifically, a value of 1 will signify a defect, while a 0 will indicate a defect-free program. Several classification models may be explored to achieve optimal prediction accuracy, including but not limited to logistic regression, decision trees, random forests, XGBoosts. Additionally, the Receiver Operating Characteristic (ROC) curve will be utilized as a key metric in evaluating the performance and the discriminative ability of the model(s), aiding in the selection of the most effective model.

1. Anticipated Obstacles & Challenges: what challenges and obstacles might you anticipate with this project?
   1. **Feature Engineering**: Identifying and crafting the most pertinent features that accurately reflect the presence of defects in a C program.
   2. **Imbalanced Data**: The dataset may have a skewed distribution of defective and non-defective programs.
   3. **Model Complexity**: Balancing model complexity and interpretability might be challenging.
2. Requested TA: No Preference