**Problem Statement**

Software defect prediction is the process of identifying parts of the software, such as modules or instances that are likely to contain defects or bugs using predictive models. Software defect prediction is a crucial step in software engineering domain. The objective is to develop a mathematical model to classify the bugs using defect datasets.

Dataset: defect\_dataset-ant-1.3.csv (uploaded as separate file)

Metadata file: metadata\_defectdataset.txt (uploaded as separate file)

1. **Import Libraries/Dataset** 
   1. Download the dataset.
   2. Import the required libraries.
2. **Data Visualization and Exploration [1M]** 
   1. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   2. Provide appropriate data visualizations to get an insight about the dataset.
   3. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**
3. **Data Pre-processing and cleaning [2M]** 
   1. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell.
   2. Apply appropriate feature engineering techniques. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.
4. **Model Building [5M]**
   1. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1M]
      1. Train = 80 % Test = 20%
      2. Also, try to split the dataset with different ratios of your choice.
   2. Build model using Logistic model and decision tree [4 M]
      1. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.
5. **Performance Evaluation [2M]**
   1. Compare the performance of the Logistic Regression and Decision Tree models using appropriate evaluation metrics.
   2. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**

***For clarifications, contact Madhusmita D. (madhusmita.das@wilp.bits-pilani.ac.in)***