**Problem Statement**

**Diamonds Cut Type Classification**

The objective of this assignment is to **develop a predictive model to accurately determine the cut quality of diamonds**, leveraging a dataset containing nearly 54,000 diamonds with attributes such as price, carat weight, color, clarity, and physical dimensions. This project aims to apply advanced data analysis and visualization techniques to uncover patterns and relationships among these attributes, ultimately enabling the prediction of diamond cut quality. The insights gained from this analysis will help improve the accuracy and efficiency of diamond appraisal processes in the industry, providing a robust tool for gemologists and jewelers to assess cut quality based on measurable characteristics.

Dataset: [Diamonds Dataset](https://docs.google.com/spreadsheets/d/1QIQtfO9YooUC05GjblLv8lYtaVm6DXqdYT6sV7arS6Q/edit?usp=sharing)

Metadata file: [Diamonds Metadata](https://docs.google.com/spreadsheets/d/1rCjOYpAQemstdBmqtZ0kVCbMlWlDlwMVgYBPMODXROw/edit?usp=sharing)

1. **Import Libraries/Dataset** 
   1. Download the dataset.
   2. Import the required libraries.
2. **Data Visualization and Exploration [1M]** 
   1. Print at least 5 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
   2. Print the description and shape of the dataset.
   3. Provide appropriate visualization 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.**
4. **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.
5. **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 regression and decision tree [4 M]
      1. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.
6. **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.**

**Instructions for Assignment Evaluation**

1. Organise your code in separate sections for each task. Add comments to make the code readable.
2. Deep Learning Models are strictly not allowed. You are encouraged to learn classical Machine learning techniques and experience their behaviour.
3. Notebooks without output shall not be considered for evaluation.

***For clarifications, contact Murtuza (murtuza.dahodwala@wilp.bits-pilani.ac.in)***