**Problem Statement**

**Objective:**

The objective of this assignment is to develop a predictive model to assess the **living conditions** of a **hostel** based on environmental factors such as **temperature**, **humidity**, **CO2 levels**, and **light intensity**. This project aims to identify whether the living conditions in the hostel are **harmful** or **not harmful**, which will allow hostel management to take appropriate actions for improving the environment.

By applying advanced data analysis and machine learning techniques, the goal is to predict whether the living conditions are harmful based on the environmental factors that impact the comfort and health of the residents.

**Dataset: Hostel Environmental Conditions**

**Dataset link:** [**Hostel Dataset**](https://drive.google.com/file/d/1SCzR_fBvBTvGsxMU1YlUqcp2ih8SBZ9c/view?usp=sharing)

The dataset gives an insight on the quality of air in some University Halls of Residence to make inform decision regarding safety and healthy living of the students living therein. The data received by the microcontroller was stored in the microSD card in a text format and later retrieved. The data was then converted from the text file to a csv file for further analysis.

The dataset contains data on the environmental conditions in a hostel, collected over a period. The features include temperature, humidity, CO2 levels, and light intensity. The dataset has been classified as **harmful(1)** or **not harmful(0)** based on predefined thresholds, which will allow students to apply classification techniques to predict the conditions.

**Assignment Tasks**

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 Karthik M. (karthik.m@wilp.bits-pilani.ac.in)***