CA5314.Machine Learning Laboratory **Exercise Sheet 2**
Instructor: Dr. M. Deivamani Multilinear Regression

Aim:

The goal of this assignment is to:

- Preprocess the dataset and prepare it for machine learning using **numpy** and **pandas**.
- Implement Multilinear regression from scratch using **numpy** and **pandas**.
- Evaluate the performance of the models and interpret the results using metrics such as mean squared error (MSE) and R-squared (coefficient of determination) on the test set.

Preprocessing / Data Loading

The few datasets require some preprocessing before it can be used for machine learning. Specifically, you will need to:

- Load the dataset into a panda DataFrame.
- Handle any missing values in the dataset.
- Remove any outliers from the dataset.
- Normalize (Regularize) the data.
- Split the dataset into a training set and a test set.

Multilinear Regression

Once you have pre-processed the dataset, you will implement linear regression from scratch using **numpy** and **pandas**. Specifically, you will:

Problem 1. Basketball (Dataset 4)

The following data (X1, X2, X3, X4, X5) are for each player.

- X1 = height in feet
- X2 = weight in pounds
- X3 = percent of successful field goals (out of 100 attempted)
- X4 = percent of successful free throws (out of 100 attempted)
- X5 = average points scored per game

Problem 2. Crime (Dataset 5)

The data (X1, X2, X3, X4, X5, X6, X7) are for each city.

- X1 = total overall reported crime rate per 1 million residents
- X2 = reported violent crime rate per 100,000 residents
- X3 = annual police funding in $/resident
- X4 = % of people 25 years+ with 4 yrs. of high school
- X5 = % of 16 to 19 year-olds not in highschool and not highschool graduates.
- X6 = % of 18 to 24 year-olds in college
- X7 = % of people 25 years+ with at least 4 years of college

Note: You should not use external libraries such as scikit-learn for the implementation of Multilinear regression