**Project Type 1: Classification of different objects based on their images.**
**https://github.com/Horea94/Fruit-Images-Dataset**
Source: Horea Muresan, Mihai Oltean, Fruit recognition from images using deep learning, Acta Univ. Sapientiae, Informatica Vol. 10, Issue 1, pp. 26-42, 2018.

The dataset contains images of fruits. Consider different species of apples from the fruit database and frame it as a classification problem.

**Step 1:** Load the data and extract the features.  Many of the image-processing techniques would be relevant here, you may include details of the procedure that you have used to extract the features.

**Step 2:** Construct a tabular form of the data with all the extracted features and the response variable, that is the name of the folder.

**Step 3:** Consider different classification techniques such as Logistic regression, Discriminant analysis, and Random Forest Classification Trees.
For all the methods, construct a classification table and calculate the accuracy and misclassification rate to compare the performance of the results. Remember, for comparing different methods you should perform the accuracy of the procedure for the testing set only.

**Project Type 2: Classification of handwriting data**
**Repository1:**
**https://didadataset.github.io/DIDA/**

Source: Huseyin Kusetogullari, Amir Yavariabdi, Johan Hall, Niklas Lavesson, "DIGITNET: A Deep Handwritten Digit Detection and Recognition Method Using a New Historical Handwritten Digit Dataset", Big Data Research, 2020, DOI: 10.1016/j.bdr.2020.100182.

**Repository2: https://www.kaggle.com/datasets/olafkrastovski/handwritten-digits-0-9/code**

**Source: https://www.kaggle.com/datasets/olafkrastovski/handwritten-digits-0-9/download?datasetVersionNumber=2**

The dataset contains different handwritten digits of numbers. Consider it as a classification problem.

**Step 1:** Load the data and extract the features.  Many of the image-processing techniques would be relevant here, you may include details of the procedure that you have used to extract the features.

**Step 2:** Construct a tabular form of the data with all the extracted features and the response variable, that is the name of the folder.

**Step 3:** Consider different classification techniques such as Logistic regression, Discriminant analysis, and Random Forest Classification Trees etc.
For all the methods, construct a classification table and calculate the accuracy and misclassification rate to compare the performance of the results. Remember, for comparing different methods you should evaluate their performance on a testing set only.

**Project Type 3: Modelling a continuous Response.**
**Repository1:**

The market historical data set of real estate valuation is collected from Sindian Dist., New Taipei City. The primary objective of the analysis is to develop a prediction model for the housing price.

**Source:**
https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set

**Reference:** The UCI page mentions the following as the original source of the data set:
Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271

Cite: Real estate valuation data set. (2018). UCI Machine Learning Repository. https://doi.org/10.24432/C5J30W.

The inputs are as follows: X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.) X2=the house age (unit: year) X3=the distance to the nearest MRT station (unit: meter) X4=the number of convenience stores in the living circle on foot (integer) X5=the geographic coordinate, latitude. (unit: degree) X6=the geographic coordinate, longitude. (unit: degree) The output is as follow Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

**Step 1:** Load the data and extract the features. Perform some exploratory analysis.

**Step 2:** Consider different techniques including multiple Regression, Regression trees, Random-Forest, etc. to develop prediction models.

**Step 3:** Compare the performance of all the methods applied. Remember, for comparing different methods you should evaluate the efficiency of the procedure based on a testing set.

**Project Type 4: Modelling a continuous Response.**
**Repository1:** https://users.stat.ufl.edu/~winner/data/airline_costs.txt

Source: J.W. Proctor and J.S. Duncan (1954). "A Regression Analysis
of Airline Costs," Journal of Air Law and Commerce, Vol.21, #3, pp.282-292.

Description: Regression relating Operating Costs per revenue ton-mile to 7 factors: length of flight, speed of plane, daily flight time per aircraft, population served, ton-mile load factor, available tons per aircraft mile, and firms net assets. Regression based on natural logarithms of all factors, except load factor. Load factor and available tons (capacity) for Northeast Airlines was imputed from summary calculations.

Variables/columns
Airline 1-20
Length of flight (miles) 22-28
Speed of Plane (miles per hour) 30-36
Daily Flight Time per plane (hours) 38-44
Population served (1000s) 46-52
Total Operating Cost (cents per revenue ton-mile) 54-60
Revenue Tons per Aircraft mile 62-68
Ton-Mile load factor (proportion) 70-76
Available Capacity (Tons per mile) 78-84
Total Assets ($100,000s) 86-92
Investments and Special Funds ($100,000s) 94-100
Adjusted Assets ($100,000s) 102-108

**Step 1:** Load the data and extract the features. Perform some exploratory analysis.

**Step 2:** Consider different techniques including multiple Regression, Regression trees, Random-Forest, etc. to develop prediction models.

**Step 3:** Compare the performance of all the methods applied. Remember, for comparing different methods you should evaluate the efficiency of the procedure based on a testing set.

**Project Type 5: Modelling a continuous Response.**
**Repository1: https://archive.ics.uci.edu/dataset/9/auto+mpg**
This data set has been sourced from the Machine Learning Repository of University of California, Irvine Auto MPG Data Set (UC Irvine).

Source: Quinlan,R.. (1993). Auto MPG. UCI Machine Learning Repository. https://doi.org/10.24432/C5859H.

Description: This intermediate level data set has 398 rows and 9 columns and provides mileage, horsepower, model year, and other technical specifications for cars. This data set is recommended for learning and practicing your skills in exploratory data analysis, data visualization, and regression modelling techniques.

**Step 1:** Load the data and extract the features. Perform some exploratory analysis.

**Step 2:** Consider different techniques including multiple Regression, Regression trees, Random-Forest, etc. to develop prediction models.

**Step 3:** Compare the performance of all the methods applied. Remember, for comparing different methods you should evaluate the efficiency of the procedure based on a testing set.