**Milestone project**

In the course milestone project, we are going to practice how to implement supervised learning, dimensionality reduction and clustering models in Python. There are two datasets considered for this project including:

1. ***Breast Cancer Dataset***
   1. *Available through sklearn (use load\_breast\_cancer as part of sklearn)*
   2. ***Link to the original dataset****:* [*http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)*](http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))
   3. ***Name****: Wisconsin breast cancer dataset*
   4. ***Summary****: Identifying if there is a malignant tumor or not using features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.*
   5. ***number of features****: 30 (real, positive)*
   6. ***Number of data points (instances)****: 569*
   7. ***dataset accessibility****: Dataset is available as part of sklearn package.*
2. ***Facebook metrics Dataset***
   1. *Download the file form the link and read it in python using pandas*
   2. ***Link to the dataset****:* [*https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand*](https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand)
   3. ***Name****: Facebook Live Sellers in Thailand dataset*
   4. ***Summary****: Posts of a different nature (video, photos, statuses, and links) on Facebook pages of 10 Thai fashion and cosmetics retail sellers.*
   5. ***Number of features****: 12 (we will work with 9 columns as features)*
   6. ***Number of data points (instances)****: 7051 (we will work with the first 1000 data points (rows))*
3. ***Parkinson Dataset***
4. *Parkinsons is a disease that can cause a nervous system disorder and affects the movement. Parkinson dataset contains biomedical measurements, 195 records of people with 23 different attributes. This data is useful in differentiating healthy people and people with Parkinsons disease.*
5. *Download the fil form the link and read it in python using pandas*
6. *Link to dataset*[*https://archive.ics.uci.edu/ml/datasets/parkinsons*](https://archive.ics.uci.edu/ml/datasets/parkinsons)
7. ***Name****: Parkinsons Data Set*
8. *Project Idea: You can build a model that can separate healthy people from people having Parkinson’s disease.*
9. ***The Boston Housing Dataset:***
10. *This is a popular dataset used in pattern recognition. It contains information about the different houses in Boston based on crime rate, tax, number of rooms, etc. It has 506 rows and 14 different variables in columns. You can use this dataset to predict house prices.*
11. ***Name****: Facebook Live Sellers in Thailand dataset*
12. ***Summary****: This is a popular dataset used in pattern recognition. It contains information about the different houses in Boston based on crime rate, tax, number of rooms, etc.*
13. ***Number of features****: 14 features*
14. ***Number of data points (instances)****: 506*
15. *Data Link:* [*Boston dataset*](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)

***Stay***

1. *Data Science Project Idea: Predict the housing prices of a new house using linear regression. Linear regression is used to predict values of unknown input when the data has some linear relationship between input and output variables.*

***5)* Credit Card Fraud Detection Dataset *:***

1. The dataset contains transactions made by credit cards, they are labeled as fraudulent or genuine. This is important for companies that have transaction systems to build a model for detecting fraudulent activities.
2. **Data Link:** [**Credit card fraud detection datase**](https://www.kaggle.com/mlg-ulb/creditcardfraud)
3. Data Science Project Idea: Implement different algorithms like decision trees, logistic regression, and artificial neural networks to see which gives better accuracy. Compare the results of each algorithm and understand the behavior of models.
4. Data: 492 frauds out of 284,807 transactions.

**Step 1:**

You need to implement two important tasks in this step:

1. Read the data files in each dataset and prepare them as dataframes.
2. Implement three supervised machine learning methods on each dataset including
   1. Regression
   2. Random forest
   3. k-NN

**Step 2:**

You need to practice two important tasks in this step:

1. Implement the following dimensionality reduction methods on each one of these datasets
   1. PCA
   2. t-SNE
   3. UMAP
2. Implement PCA and then implement three supervised machine learning methods on each dataset including
   1. Regression
   2. Random forest
   3. k-NN

**Step 3:**

You need to practice clustering in this step:

1. Read the data files in each dataset and prepare them as data frames.
2. Implement different clustering on each dataset including

**Note**. You can also practice proper visualization and working with different performance measures in each step.

**Code submission**: You need to work on all these steps and submit your code before the last session of the course.

**Final presentation**: Half of the last session is dedicated to your presentations for the milestone project. Prepare a 5 minutes presentation focusing on the main results of each step and what you learned from them. Your presentation should be 5 slides maximum. You need to also submit your presentation after the last session.