Python資料分析與機器學習應用
Data Analysis and Machine Learning with Python

I. Course Information

Instructor: Cheng-Yuan Ho (Tommy Ho)
Office: 2C, Teaching and Researching Hall
Email: tommyho@ntu.edu.tw
Office hours: Please send an email to make an appointment first, and indicate your class. DO NOT send NTU Cool's internal letter
Curriculum Number: IM1013
Curriculum Identity Number: 705E14200
Class: 03
Required/Elective: Elective
Time: Tuesday 2, 3, 4 (9:10~12:10)

II. Course Syllabus

Course Description:

This course will be your guide to learning how to use the power of Python to **analyze data** and use powerful **machine learning** algorithms, and then create beautiful **visualizations** for the analysis results and predictions.

This course is designed for beginners with some programming experience, the guys who already know some **Python** and are ready to dive deeper into using those Python skills for data analysis and Machine Learning, or experienced developers looking to make the jump to **Data Science (blue)**.

I want to help guide students to learning a set of skills to make them extremely hirable in today's workplace environment. Therefore, I'll teach you how to program with Python, how to create amazing data visualizations, and how to use Machine Learning with Python! Enroll in the course and become a data scientist!

Course Objective:

1. Use Python for Data Analysis and Machine Learning
2. Understand how to Use Tools to Analyze Data
3. Understand how to Use Existing Machine Learning Modules/Packages

4. Learn Related Modules and Tools in Python, like NumPy, Pandas, Matplotlib, and SciKit-Learn

Course Prerequisites:

Take Python programming or related course before, or understand Python

Designated reading: No

References:

1. Introduction to Machine Learning with Python: A Guide for Data Scientists
2. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython
3. Machine Learning with Python Cookbook：Practical Solutions from Preprocessing to Deep Learning
4. Python Cookbook
5. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts,Tools, and Techniques to Build Intelligent Systems

Progress:

Week	Date	Topic (The actual teaching content will be adjusted according to the class status)
Week 1	**2/21**	**Course Overview, What is Data Analysis, What is Machine Learning**
Week 2	**2/28**	**228 holidays**
Week 3	**3/07**	**Steps and Flows in Data Analysis, Introduction to Pandas Functions and Module in Pandas**
Week 4	**3/14**	**Machine Learning: Supervised vs Unsupervised vs Reinforcement Learning Grouping due and Homework 1 assigned**

Week 5	**3/21**	**Algorithms and Modules in Supervised Machine Learning I**
Week 6	**3/28**	**Algorithms and Modules in Supervised Machine Learning II** **Homework 1 due**
Week 7	**4/04**	**Ching Ming holidays**
Week 8	**4/11**	**Midterm Project Presentation**
Week 9	**4/18**	**Algorithms and Modules in Unsupervised Machine Learning** **Homework 2 assigned**
Week 10	**4/25**	**Data Modeling, Tuning, and Explanation (by Manual)**
Week 11	**5/02**	**Data Modeling, Tuning, and Explanation (by Automatic and Functions)**
Week 12	**5/09**	**Implementation and Discussion** **Homework 2 due**
Week 13	**5/16**	**Case Study: Energy Saving and Product Defect Detection**
Week 14	**5/23**	**Case Study: Prognostics and Health Management (PHM)**
Week 15	**5/30**	**Lecture/Talk (topic: TBD)**
Week 16	**6/06**	**Final Project Presentation**

Homework & Project:

Item	Description
Homework 1	**Practice how to analyze data with Python**
Homework 2	**Practice how to predict data with Python Machine Learning modules**
Midterm Project	**Design what you want to show in the final project, including but not limited to the following items:** **1. Target: what problem in which field** **2. Dataset: where it comes from, or collected by web crawler** **3. Packages and tools** **4. System structure and modules** **5. Analysis or prediction or both** **6. Expected result** **7. Schedule** **8. Team work assignments** **9. Pre-completed items, such as data collection** **Note: DO NOT copy problem and solution from the Internet, but can use the existing dataset**
Final Project	**Complete "ALL" your design in the midterm project** **Upload the Python source code and report to NTU Cool** **Demo in the class**

Score Rules:

15%	**Attendance** The first time the random check was conducted, half of the scores were deducted if the checked student was not here. The second time the random check was conducted, if the checked student was not here again, he/she did not meet the requirements and was directly considered as a failure of this course. (Not in the same day) **(If you have applied for leave through the system, it will not be counted against you. However, please remember to send an email to Tommy and TAs as a backup.)**

20%	**Homework 1 & 2 (late submission: 1 week: 20% off; 2 weeks: 40% off; over 2 weeks: 0)**
25%	**Midterm project (including intra- and inter-group evaluations)**
40%	**Final project (including intra- and inter-group evaluations)**
Extra Points	**1. Participate in program language-related competitions and provide supporting materials, you will earn extra 2 points to the course grade. The maximum times is up to 3** **2. Tommy is discussing a project with iKala about the use of iKala's KOL Rader dataset. If succeed, you or your group can create another project with this dataset (i.e., different from the final project) to earn the extra 5 points. This project should make creative, analytical results or predictive models. only once**

TAs:

1. 吳琦艾**(WU, CHI-AI)** r10725023@ntu.edu.tw
2. 吳晨瑋 **(WU, CHEN-WEI)** r10725051@ntu.edu.tw

其他:

1. During class, if there is something you don't understand, please ask questions immediately or after class, I will try my best to let you understand it.

2. If I ask you or your group to think, talk, discuss, share and express, please try to participate in the activity. Moreover, I will find ways to let you participate in it.

3. About online, e-learning, live streaming, and recorded video:
There are no online classes, e-learning, and live streaming, but I will record videos during the class and upload them to Youtube (unlisted) at night.

4. About extending number of student for adding course:
You must come to the class in the first week. Otherwise, you cannot be added into the course.
The max number of all students is 120, the capacity of B01, Building 1, College of Management. (Right now, the number of students is 79.)
If the total number is larger than 120, the priority for adding course is as follows.

a. fourth grade and above
b. others (random selection)

5. Registered Code and Auditor Application Form:
https://forms.gle/NfQE6wdg5LQtv4qf8
(Open: 12:20 pm, Feb. 21; Close: 11:59 pm, Feb. 24)
6. Dispatch Registered Code: March 3 morning via TAs

Video lists:
1. 2023 111-2 Data Analysis and Machine Learning with Python Tuesday Morning
https://www.youtube.com/playlist?list=PLCGdg-V2_I4iKR0j5lQi9AxFe6ow6ZCu2
2. 2022 111-1 Python Programming Foundation Course
https://www.youtube.com/playlist?list=PLCGdg-V2_I4jrLMFPBeBt8dnZ8jMSp_26

What is Data Analysis? (source: wikipedia)

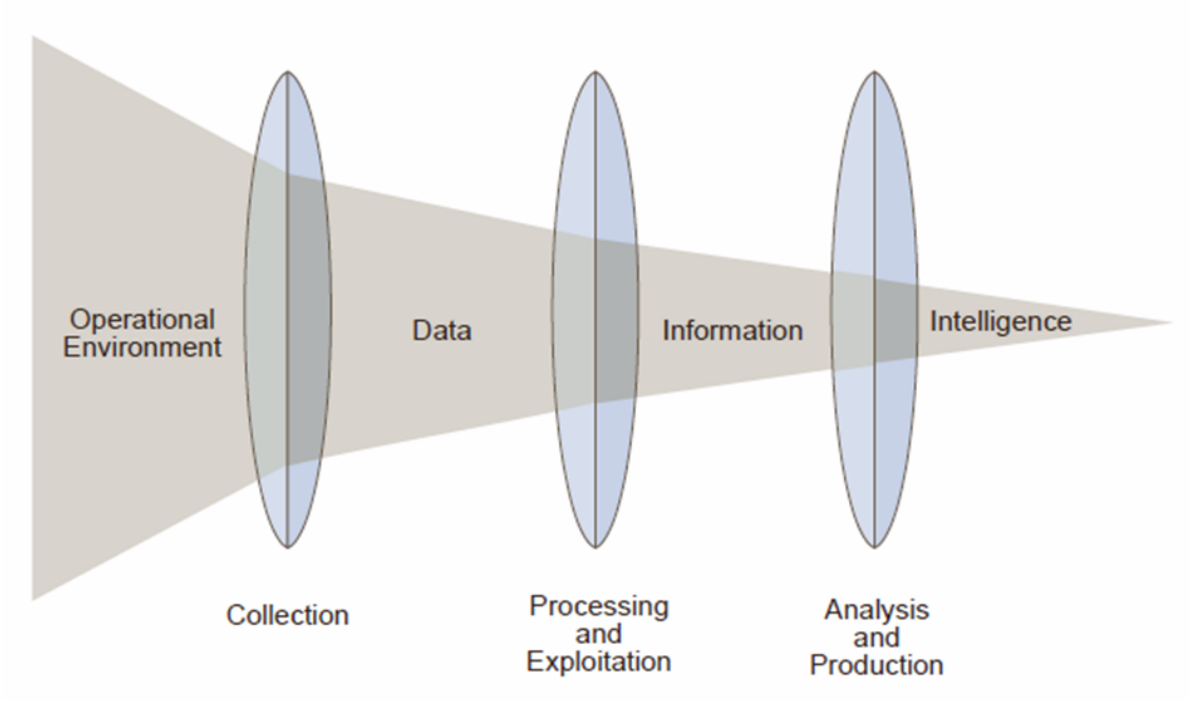
Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.

There are several phases that can be distinguished, described below.

- Data requirements
- Data collection
- Data processing
- Data cleaning
- Exploratory data analysis
- Modeling and algorithms
- Data and information visualization

Relationship of Data, Information and Intelligence



Source: Joint Intelligence / Joint Publication 2-0 (Joint Chiefs of Staff)

Practice Problem: What is Machine Learning?

Please try to follow the steps (questions) to answer the problem, “what is machine learning?”

1. What data do you need in order to present “Machine Learning?”
2. Where do you collect data?
3. Do you need to clean or process collected data?
4. Is there any conflict among the collected data? If yes, what will you do?
5. Can you answer the problem? Please show it with data visualization if possible.

Upload your document to the following Google Drive.
Please set File name as Student ID-Name, like B12345678-Tommy
https://drive.google.com/drive/folders/1dLG7UMWedWluT0-MYZopl3MXygFsJziD?usp=sharing