

**SCHOOL OF COMPUTER TECHNOLOGY**

AASD 4001 Mathematical Concepts for Machine Learning

**Term Project:**

1. Obtain a large classification database. The database needs to have at least 2 classes, more than 5 features and over 200 samples. Each group has to use a unique dataset.
2. Provide a description of the dataset used including explanation of various features.
3. Pre-process and clean the dataset as appropriate.
4. Use following approaches for classification of the dataset:
   1. Logistic Regression
   2. Decision Tree
   3. Random Forest
   4. SGD
   5. SVM
5. Use GridSearchCV to tune the parameter of each of the above models. Can you obtain better results in this step for any of the models? Discuss your observations.
6. Randomly (or based on certain hypothesis) remove some features and re-evaluate the models. Document your observations with respect to models performances.
7. Present your work in-person including approach and findings during the class on January 28th, 2025. Each group will have a maximum of 15 minutes to present their project. It is advised that your PowerPoint files to be no longer than 15 slides.
8. Prepare a written technical report of no longer than 15 pages to discuss the problem statement, various steps conducted, summary of findings and conclusions. Submit the report and the notebook file (with proper headings, explanatory comments and code sections) by the midnight of February 2nd, 2025.

**Dataset:**

* **Customer Satisfaction Response to AI**
  + [**https://www.kaggle.com/datasets/pooriamst/online-shopping/data**](https://www.kaggle.com/datasets/pooriamst/online-shopping/data)
    - **ONLINE SHOPPING – LogisticRegression**
    - [**https://www.kaggle.com/code/hainescity/online-shopping-logisticregression**](https://www.kaggle.com/code/hainescity/online-shopping-logisticregression)

**AASD 4001 機器學習數學概念**

**學期專題：**

1. **獲取大型分類數據庫**：  
   數據庫必須至少包含兩個類別，超過五個特徵，以及至少200個樣本。每組必須使用不同的數據集。
2. **提供數據集描述**：  
   包括對各個特徵的說明。
3. **數據預處理與清理**：  
   根據需要進行適當的數據預處理與清理。
4. **使用以下方法對數據集進行分類**：
   * 邏輯回歸 (Logistic Regression)
   * 決策樹 (Decision Tree)
   * 隨機森林 (Random Forest)
   * 隨機梯度下降 (SGD)
   * 支持向量機 (SVM)
5. **使用 GridSearchCV 調參**：  
   調整上述模型的參數。討論此步驟是否能為某些模型取得更好的結果，並記錄觀察結果。
6. **隨機移除（或基於某些假設移除）部分特徵並重新評估模型**：  
   記錄觀察結果，尤其是對模型表現的影響。
7. **現場展示成果**：  
   在2025年1月28日課堂內進行現場展示，每組最多15分鐘。建議簡報頁數不超過15頁。
8. **撰寫技術報告**：  
   報告篇幅不超過15頁，內容應包括問題陳述、實施步驟總結、發現與結論。報告及包含標題、註釋和代碼段的Notebook檔案，需於2025年2月2日午夜前提交。

是的，許多平台提供公開的數據集，您可以選擇符合學期專題要求（至少2類別、5個特徵、200樣本）的數據集進行分析。以下是一些常用的公開數據集資源：

**1. Kaggle**

* **網站**：<https://www.kaggle.com/datasets>
* Kaggle 提供大量公開數據集，涵蓋從分類到回歸等多種應用場景。
* **範例數據集**：
  + [Iris Dataset](https://www.kaggle.com/uciml/iris)：經典的分類數據集，有4個特徵和3個類別。
  + [Titanic Dataset](https://www.kaggle.com/c/titanic)：有2個類別（存活與否），適合分類任務。
  + [MNIST](https://www.kaggle.com/c/digit-recognizer)：手寫數字分類，有10個類別和大量特徵。

**2. UCI Machine Learning Repository**

* **網站**：<https://archive.ics.uci.edu/ml/index.php>
* UCI 是機器學習領域最著名的數據集存儲庫之一。
* **範例數據集**：
  + [Wine Dataset](https://archive.ics.uci.edu/ml/datasets/wine)：針對葡萄酒的分類問題，有13個特徵和3個類別。
  + [Bank Marketing](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)：基於客戶數據預測是否接受銀行產品。
  + [Adult Dataset](https://archive.ics.uci.edu/ml/datasets/adult)：收入分類數據（年收入是否超過50K）。

**3. OpenML**

* **網站**：<https://www.openml.org/>
* OpenML 提供多領域的數據集，支援直接導入至 Python。
* **範例數據集**：
  + openml.datasets.get\_dataset(31)：經典的 Credit Scoring 數據集。
  + openml.datasets.get\_dataset(61)：Sonar數據集，適用於二元分類。

**4. Google Dataset Search**

* **網站**：<https://datasetsearch.research.google.com/>
* Google 提供強大的數據集搜索工具，可以找到來自學術機構或開放平台的數據。

**5. AWS Public Datasets**

* **網站**：<https://registry.opendata.aws/>
* 提供科學與社會科學領域的大型數據集，例如天氣、交通或醫學領域的數據。

**6. Data.gov**

* **網站**：<https://www.data.gov/>
* 美國政府數據平台，涵蓋健康、教育、能源等多個領域的公開數據。

**7. その他國內數據集來源（特別適用於中文資料分析）**

* **台灣開放政府數據平台**：<https://data.gov.tw/>
* **中國國家統計局數據**：<http://data.stats.gov.cn/>