113 學年度

資料探勘

Term Project Report

第 11 組

313553053 康峻瑋

313554059 林滋隆

313554046 許茗鈞

413551001 黃正鵬

1、 計畫目標的問題 (Target problem)

- 目標問題描述

此計畫的題目為「Predicting Obesity Levels by Linking Personal Information and Lifestyle Factors」，目標為通過分析個人生活方式來預測肥胖程度。從個人生活(包含運動)及飲食習慣的資訊中，透過資料探勘技術找出導致過重或肥胖的原因，並建立機器學習模型來判斷其過重或肥胖的程度。

在世界衛生組織 2024 年 3 月的一份報導[1]中指出，全世界每 8 人中就有 1 人患有肥胖症，從 1990 年以來全球肥胖人口在各年齡層都持續增加，全球有 25 億成年人口有過重或肥胖問題，且超過 3.9 億的兒童及青少年有過重或肥胖問題。這份報導指出，超重或肥胖的成因是來自於熱量攝取(飲食)與熱量消耗(身體活動)之間的不平衡導致。另外，目前診斷過重或肥胖的方式是透過測量個人的身高及體重，並計算身體質量指數(BMI: 體重(kg)/身高^2(m^2))。

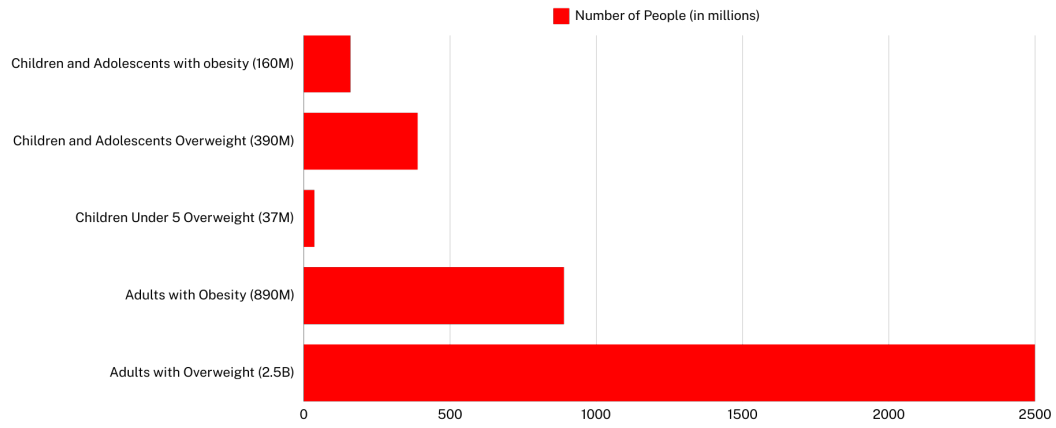


圖: 世界人口中過重與肥胖的統計資訊

若能找出過重與肥胖的成因，尤其針對個人生活飲食與身體活動的關聯性分析，就能及早預防與進行調整，降低罹患肥胖症的機會。健康飲食與運動是眾所周知，可以預防過重或患上肥胖症的生活習慣。但對於其細節與關聯性的認知則普遍沒有清楚的了解，所以我們希望透過資料探勘技術找出個人飲食與活動習慣與肥胖症之間的關係。以及訓練及使用機器學習模型，評估個人目前是屬於哪種過重或肥胖層級。

本計畫目標問題描述如下:

➔ **輸入**: 所使用的資料集包含 16 個特徵欄位，涵蓋個人資訊，如年齡、身高、體重，飲食習慣與活方式等因素，如每日餐數、吸菸狀況等。

➔ **處理過程**

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

- ◆ 肥胖程度預測: 使用分類模型（Random Forest, Gradient Boosting , 並嘗試 Ensemble method）來預測肥胖程度
- ◆ 將生活習慣與肥胖程度聯繫: 尋找某些生活習慣因素（例如: high fast food consumption, low physical activity）經常共同出現，並且是過重或肥胖的案例（使用 ECLAT 或 FP-Growth 方法）

➔ 輸出:

- ◆ 分類模型預測: 過重或肥胖等級。
  - 其類別共有 4 個 Weight Level 、3 個 Obesity Level。
- ◆ 關聯分析結果: 生活方式與 Obesity Level 之間的關聯

- 評估指標
  - 主要指標
    i. Accuracy: 模型預測正確的體重等級或肥胖類型的比例。
       - 在類別不平衡的情況下（某些類別樣本數遠多於其他類別），準確率可能會導致偏差。
    ii. AUROC: 在多類別分類問題中，AUROC 可以擴展為每個類別的曲線，衡量模型在不同肥胖類型或體重等級上的分類能力。
  - 輔助指標
    i. Confusion Matrix: 在肥胖分類中，混淆矩陣可以幫助理解模型在不同體重等級或肥胖類型上的預測是否準確，並看出常出現的錯誤類型。
    ii. Recall: 真實屬於某一肥胖類型的人中，有多少被模型正確預測出來。
    iii. Precision: 被分類為某肥胖類型的人中，有多少人實際屬於該類別。
    iv. F1 Score: 綜合評估模型在正確預測肥胖類型或體重等級時的表現
- 原預估之模型效能與目標

根據過去三年中相關研究主題的論文，其中三篇的成果 (Thamrin, 2021) (Quiroz, 2022) (Cheng, 2021)，我們訂出以下模型效能評估指標與目標。

- Top-1 Accuracy: 90% 以上。
- AUROC: 80% 以上。

表: 相關研究主題的成果

作者	研究目標	結果
Thamrin, Arsyad, et al., 2021	Predicting obesity in adults	Accuracy: 72% AUC: 79%
Cheng et al., 2021	Prediction of the effect of physical activity on obesity	Accuracy: 67% AUC: 64%

Santisteban Quiroz, 2022	Identifying obesity levels based on lifestyle through ML techniques	Accuracy: 97.45% AUC: 99.90%

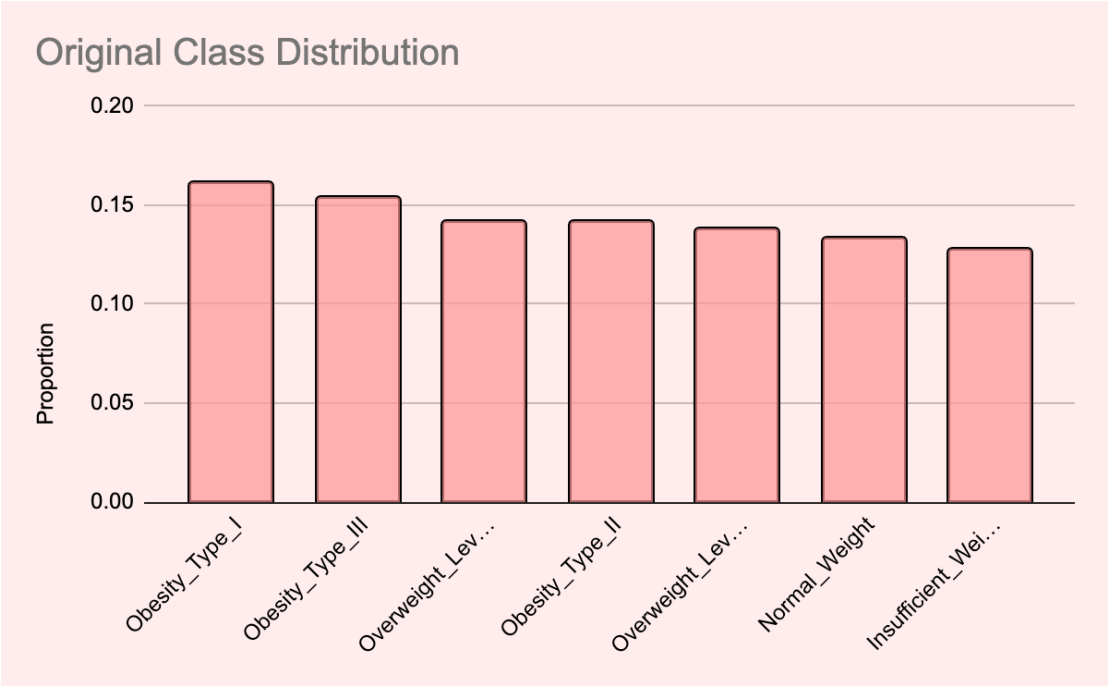
2、 選用的資料集描述 (Descriptions of selected datasets)

- 資料集來源

我們使用的資料集是來自於 Kaggle 上名為「Estimation of obesity levels UCI dataset」的公開資料集 (Palechor & Manotas, 2021)。其原始資料是使用網路問卷調查形式蒐集而來，並在 2019 年 8 月 26 日公開提供於「UC Irvine Machine Learning Repository[2]」平台上，問卷填寫人中包含來自於墨西哥、祕魯以及哥倫比亞三個國家的人，這些資料可以用於分析個人飲食習慣與身體活動等資訊，進而及評估個人肥胖層級。

- 資料集相關描述

此資料集中包含 17 個特徵資訊、2111 筆紀錄，每筆紀錄標有一個類別資訊:「肥胖層級」。這些資訊能作為資料分析與分類任務使用，用來指出個人目前的過重或肥胖層級，此類別資訊的值包含: 過輕、正常、過重 I、過重 II、肥胖 I、肥胖 II 以及肥胖 III 等 7 種類別。 其中 23% 資料是直接來自於網路問卷調查的使用者，其餘 77% 則是透過資料探勘工具 Weka 模擬生成。



2 https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

圖: 目標類別欄位資料分佈狀況

下表說明 17 個特徵資訊 (包含 1 目標類別) 的詳細資訊，初步檢查此資料集並沒有遺漏值的情況，且目標類別欄位也沒有不平衡情形 (參考上圖)。目標欄位有 7 種唯一值，因此屬於多類別的分類任務。根據原始資料與資料集提供來源的說明，目標類別是使用身體質量指數（BMI）對應 WHO 及 Mexican Normativity (LA OBESIDAD & GENERAL, 2010) 所訂出來的肥胖層級，以下條列基於墨西哥成人的 BMI 指數區分層級方式。

- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 40

表: 資料集 17 個特徵資訊

名稱	功能類型	資料型態	資料範圍	說明
Gender	feature	Categorical	Female,Male	-
Age	feature	Continuous	14~61 歲	-
Height	feature	Continuous	1.45~1.98 公尺	-
Weight	feature	Continuous	39~173 公斤	-
family_history_with_overweight	feature	Binary	yes, no	Has a family member suffered or suffers from overweight?
FAVC	feature	Binary	yes, no	Do you eat high caloric food frequently?
FCVC	feature	Integer	1 (never), 2 (sometimes), 3 (always),	Do you usually eat vegetables in your meals?
NCP	feature	Continuous	1 (Between 1 y 2), 3 (Three), 4 (More than three)	How many main meals do you have daily?

CAEC	feature	Integer	1 (no), 2 (Sometimes), 3 (Frequently), 4 (Always)	Do you eat any food between meals?
SMOKE	feature	Binary	yes, no	Do you smoke?
CH2O	feature	Continuous	1 (Less than a liter), 2 (Between 1 and 2 L), 3 (More than 2 L)	How much water do you drink daily?
SCC	feature	Binary	yes, no	Do you monitor the calories you eat daily?
FAF	feature	Continuous	0 (I do not have), 1 (1 or 2 days), 2 (2 or 4 days), 3 (4 or 5 days)	How often do you have physical activity?
TUE	feature	Integer	0 (0–2 hours), 1 (3–5 hours), 2 (More than 5 hours)	How much time do you use technological devices such as cell phone, video games, television, computer and others?
CALC	feature	Integer	1 (no), 2 (Sometimes), 3 (Frequently), 4 (Always)	How often do you drink alcohol?
MTRANS	feature	Categorical	Automobile, Motorbike, Bike, Public_Transportation, Walking	Which transportation do you usually use?
NObeyesdad	target class	Categorical	Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level	Obesity level

			_II, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III	

3、 針對問題設計的分析流程 (Analysis workflow)

- 選用之資料探勘方法以及選用原因

針對此問題設計的分析流程分為五個階段，分別為資料前處理、資料轉換、特徵選取、模型訓練、模型效能評估，以下分別說明各階段所使用的方法及選用原因。

1. 資料前處理 (Data Preprocessing)
    a. 資料型態檢查
        i. 數值型態轉換，如資料型態異常，應該是 Integer 卻被儲存成 Float。對特徵使用分析統計處理時才能得到正確的結果。
    b. 缺失值與異常值的確認與處理
        i. 確認資料是否有缺失值，並透過缺失值填補讓資料能進行模型訓練又不過度影響原始資料分佈。
        ii. 異常值產生的雜訊可能在分類模型中導至錯誤的判斷。
    c. 低變異值欄位確認
        i. 用於判斷特徵是否資訊貢獻低，並從訓練資料集中移除。
    d. 特徵關聯性分析
        i. 如果兩個特徵為高相關的欄位，對於模型訓練能貢獻的資訊會有重複性，因此僅保留一個代表性欄位即可。

2. 資料轉換 (Data Transformation)
    a. 多數模型接是使用數值進行分析與訓練，因此文字類型資料與類別資料需要透過資料轉換處理，才能在模型訓練階段使用。
    b. 由於類別型資料的類別都在 5 類以內，因此全部資料轉換都採用 Label Encoding 處理。另外，資料總筆數約兩千多筆左右，在切分過後的資料集可能發生缺少某一類別的情況，導致在驗證時總欄位數量不一致的情況發生。

3. 特徵選取 (Feature Selection)
    a. 此步驟將依照前處理的分析結果進行特徵挑選，來得到能提供模型最有幫助資訊的特徵。
    b. 由於肥胖層級是由身高體重計算後的 BMI 值進行類別區分，所以將在訓練階段將身高體重欄位排除進行訓練及評估。

4. 模型訓練 (Modeling)
    a. 首先將資料集且分為 80% 訓練資料集和 20% 測試資料集。
    b. 分類模型任務將使用訓練資料集並選擇五個模型進行比較，執行步驟為先進行資料分割成為訓練集與驗證集，再對訓練集做資料平衡處理，由於資料總比數不多，且總共有七個類別，各類的資料會相對稀少，因此採用 SMOTE 的方法來維持資料原有的代表性。
        i. 選用的五個模型如下，所選取的模型都是 Decision Tree-based 的模型。
            1. LGBMClassifier (LightGBM (Ke, 2017))
            2. CatBoostClassifier (CatBoost)
            3. RandomForestClassifier (Random Forest)
            4. XGBClassifier (XGboost)
            5. GradientBoostingClassifier (Gradient Boosting)
        ii. Decision Tree-based 的分類模型在處理資料集較小且屬於多類別分類任務時，具有以下幾個優點與特色:
            1. 決策樹的結構清晰，能夠直接解釋每個決策的過程，便於理解和分析特徵對結果的影響。
            2. 決策樹不依賴於大量數據來訓練，因此在小型資料集上仍然可以有效地運作。
            3. 本身就能支援多類別分類問題，可以在同一棵樹中處理多個目標類別，而不需要特別修改演算法。
            4. 決策樹會自動選擇對分類結果影響較大的特徵，對於含有無關或冗餘特徵的數據表現良好。
            5. 相較於許多複雜模型，決策樹的訓練和推論速度較快。
    c. 關聯分析任務選擇兩個方法進行比較
        i. ECLAT

            透過垂直數據表示法，將每個項目對應到一組出現的交易紀錄索引（TID-list，主要利用集合的交集運算，快速計算頻繁項目集合。)

            - 優點：ECLAT 計算邏輯簡單，特別適合處理稀疏的數據，能在高維度的數據下保持較高的計算效率。
            - 缺點：當數據集規模較大或頻繁項目數量龐大時，TID-list 的操作可能導致需要大量的記憶體。

        ii. FP-Growth

            透過建立 FP-Tree 去避免生成候選項目集合，以遞歸方式壓縮數據結構，利用 conditional pattrern base 去尋找 freqent itemset。

- 優點：FP-Growth 避免了大量的候選集合的產生，能更高效的處理大型數據集，同時 FP-Tree 的壓縮特性使其在記憶體使用上更具優勢。
- 缺點：當數據過於密集時，FP-Tree 的建立和遍歷可能變得複雜。

iii. 比較與結果

根據本專案中資料的特性與目標，FP-Growth 在尋找生活習慣因素與肥胖層級的深層關聯上更為適用，因此在後續分析中更注重於 FP-Growth 結果的應用。

5. 模型效能評估 (Evaluation)
   a. Accuracy: 評估模型預測正確的體重等級或肥胖類型的比例。
   b. AUROC: 衡量模型在不同肥胖類型或體重等級上的分類能力。

- 模型評估方法

我們採用 K-Fold cross validation 來評估模型效果，一方面降低 over fitting 的風險，另一方面能更有效率的使用訓練資料，降低資料偏差議題，也提升找到更具有代表性的資料。尤其在資料量有限的情況下，K-Fold Cross Validation 能夠更高效地利用可用的資料，每個樣本在不同的迴圈中都有機會成為測試集，提升了資料的利用效率。我們將使用 5-Fold 並參考平均性能指標，如 AUROC、F1 score 的平均分數，這類平均分數更能反映模型的穩健性，避免模型過度優化於某個特定的訓練集或測試集分割。因此，使用 K-Fold Cross Validation 是提升模型可靠性與穩健性的重要手段，特別是在資料偏差風險較高或數資料有限的應用場景中。

- 所採用之平台及工具

我們選用 Google Colab 作為研發平台，Google Colab 提供基於雲端的 Jupyter Notebook，可以透過分享連結讓多位使用者即時協作，我們利用其即時共享功能進行團隊合作，確保開發流程的透明與高效。並且透過雲端執行的方式，結合 Python 和 scikit-learn，我們能快速重建一致的實驗環境，降低結果不一致的風險，同時透過 Google Colab 的 Markdown 單元，清楚記錄每個模型的測試與結果，便於後續分析。

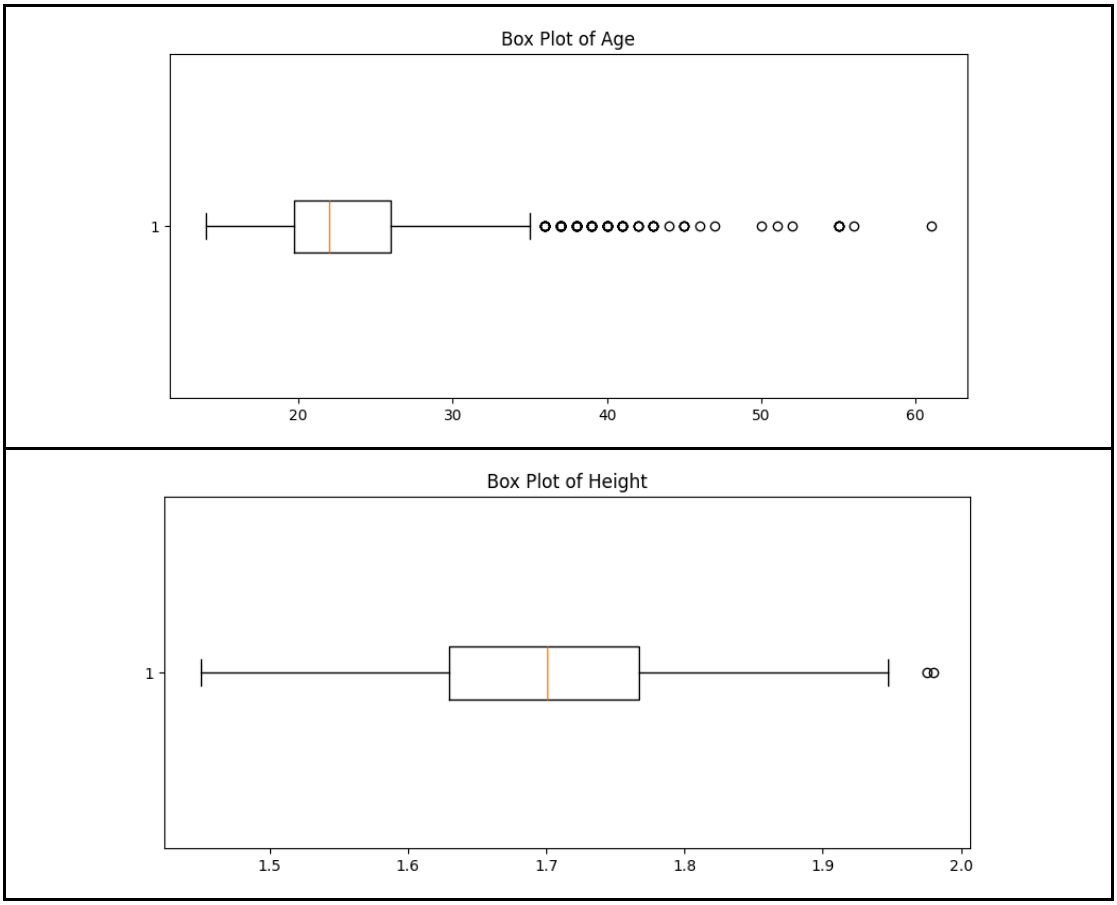
4、 分析結果 (Analysis results)

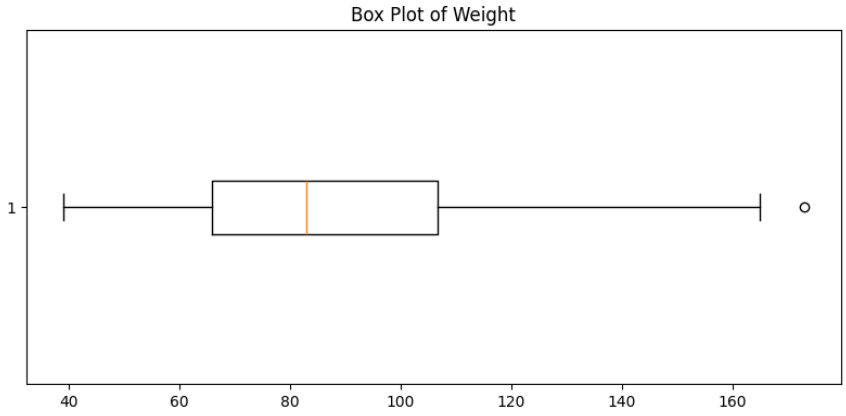
- 實驗結果
  - 資料探索與前處理

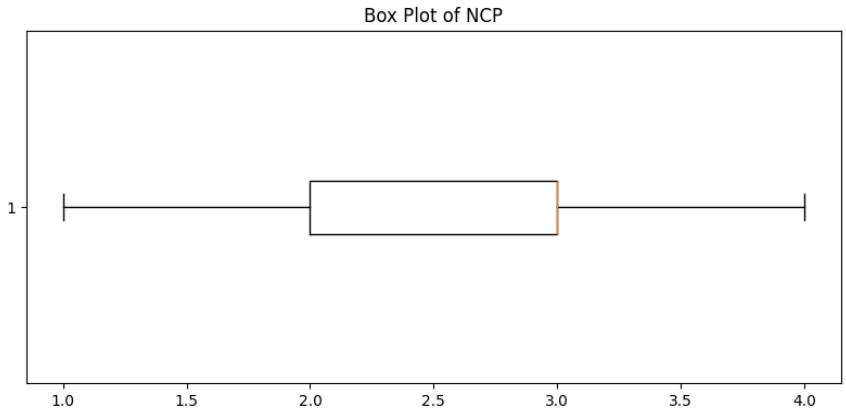
i. 針對類別型欄位進行低變異值欄位確認，下表為各個類別型欄位的箱型圖，能看出異常值(Outliers)以外，還能看出數值集中分佈的範圍落在那些區間，同時得知數值變異範圍是否廣泛。

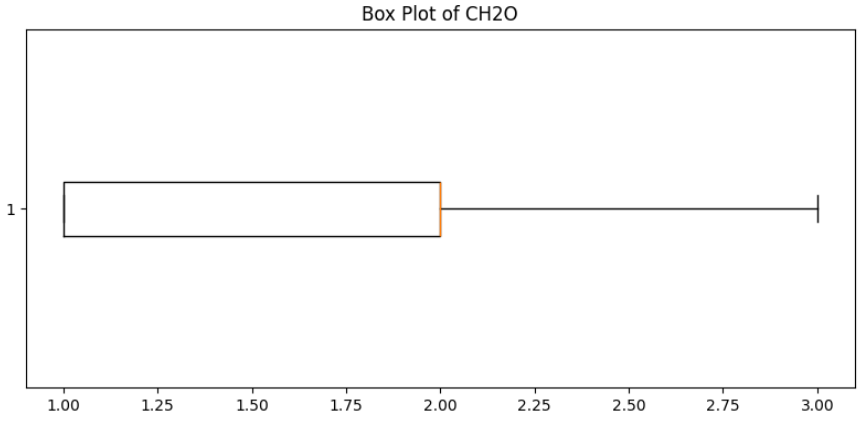
- 年齡 (Age)：主要的數值集中在 20 歲到 30 歲之間。40 歲以上的資料顯示較為分散。
- 身高 (Height)：身高的分佈相對均勻，異常值非常少，集中於約 1.6 到 1.8 公尺之間。
- 體重 (Weight)：體重的分佈集中於 60 到 100 公斤之間。
- 飲食行為變數 (FCVC)、每日餐數 (NCP)、水分攝取 (CH2O)、活動頻率 (FAF)、工作壓力 (TUE)：原始資料已經將文字類別數值化，因此分佈都屬於正常且沒有什麼異常值。

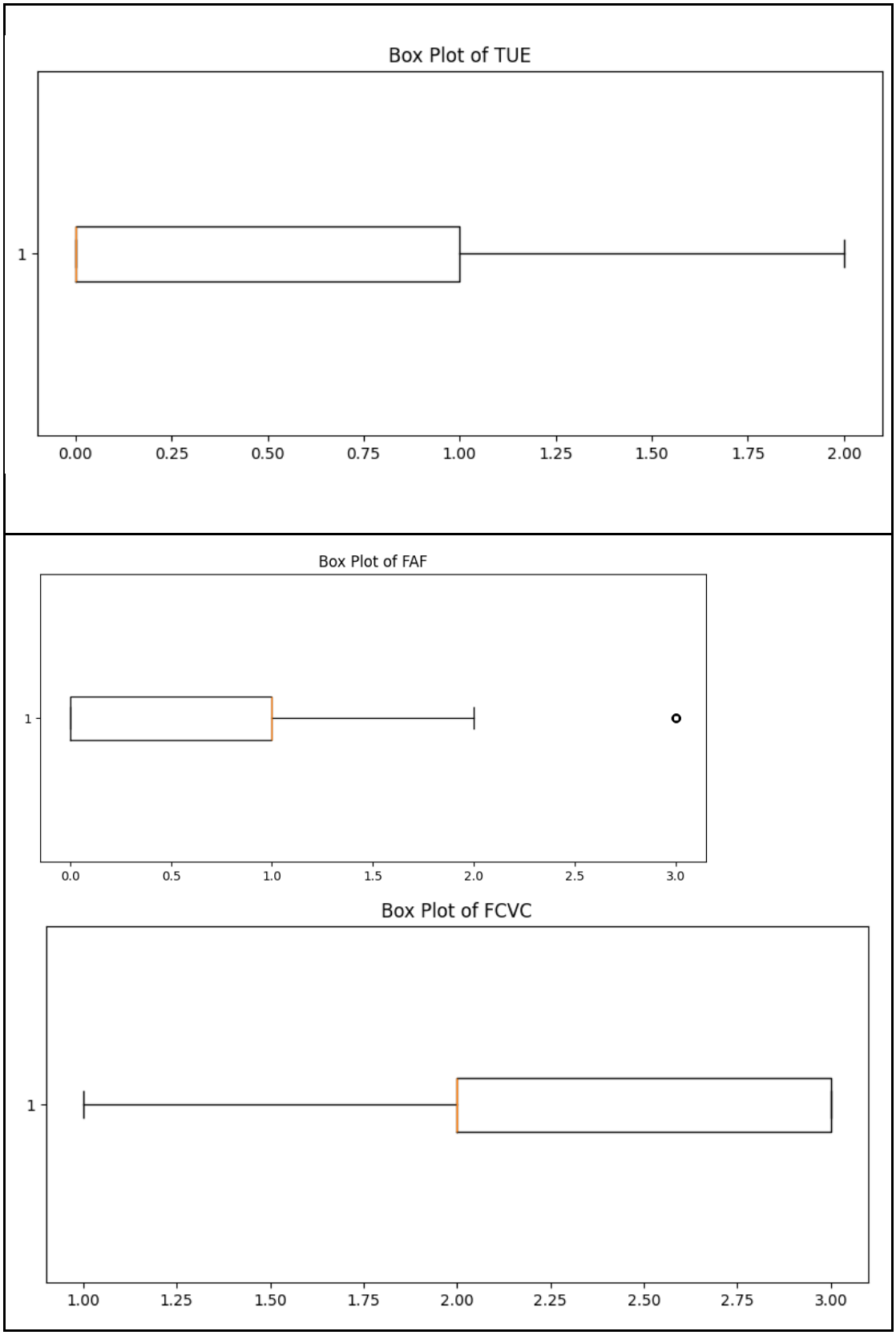
表：類別型欄位箱型圖分析




Box Plot of Weight
1
40
60
80
100
120
140
160


Box Plot of NCP
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Box Plot of CH2O
1
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

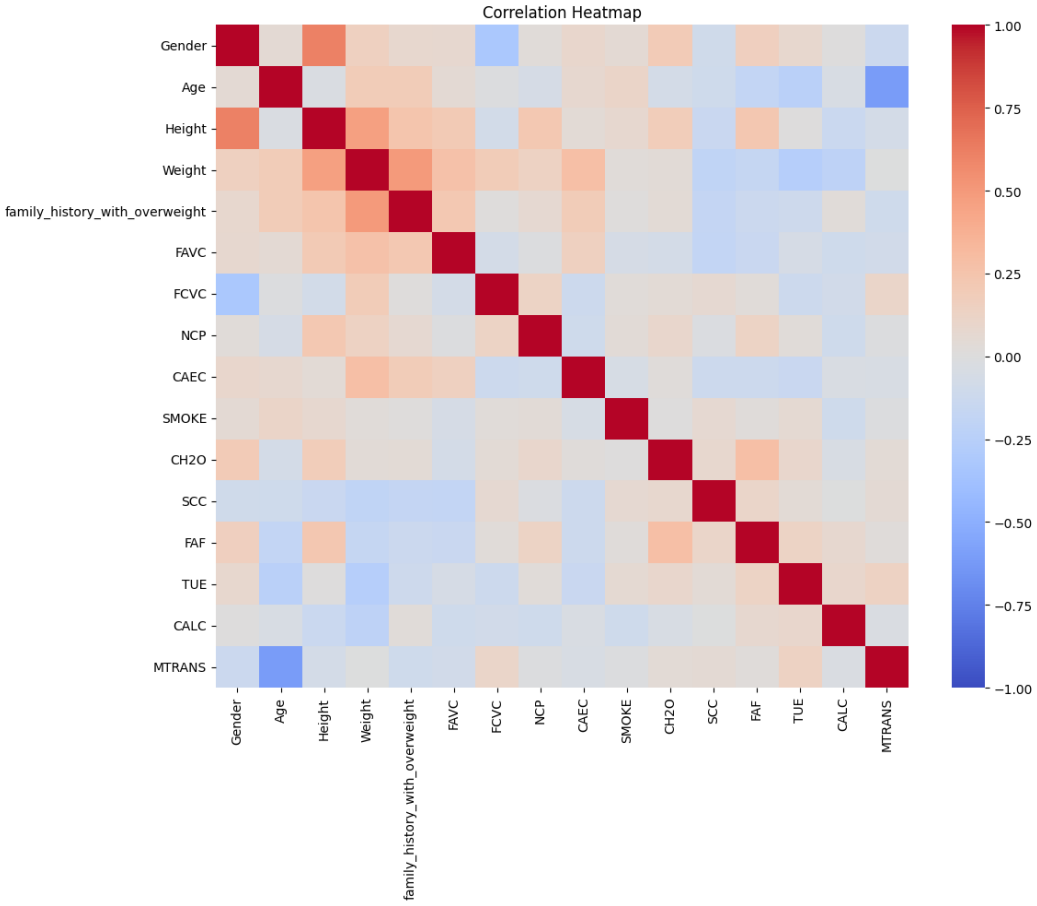


ii. 透過繪製熱力圖（Heatmap）進行特徵相關度檢查與分析。

- 呈現正相關的有：

- 【性別】與【身高】：通常男性身高比女性高
- 【身高】與【體重】：身高高的人通常體重也相對比較高
- 【體重】與【家族肥胖史】：有家族肥胖史通常本身也會有過重或肥胖的情形

- 呈現負相關的有：
  - 【年齡】與【經常搭乘的交通工具】：隨著年齡增長搭乘大眾交通工具次數會高於自行開車或騎腳踏車



圖：所有特徵之間的熱力圖

- 使用分類模型進行肥胖程度預測

i. 使用經過特徵挑選後的訓練資料集進行訓練

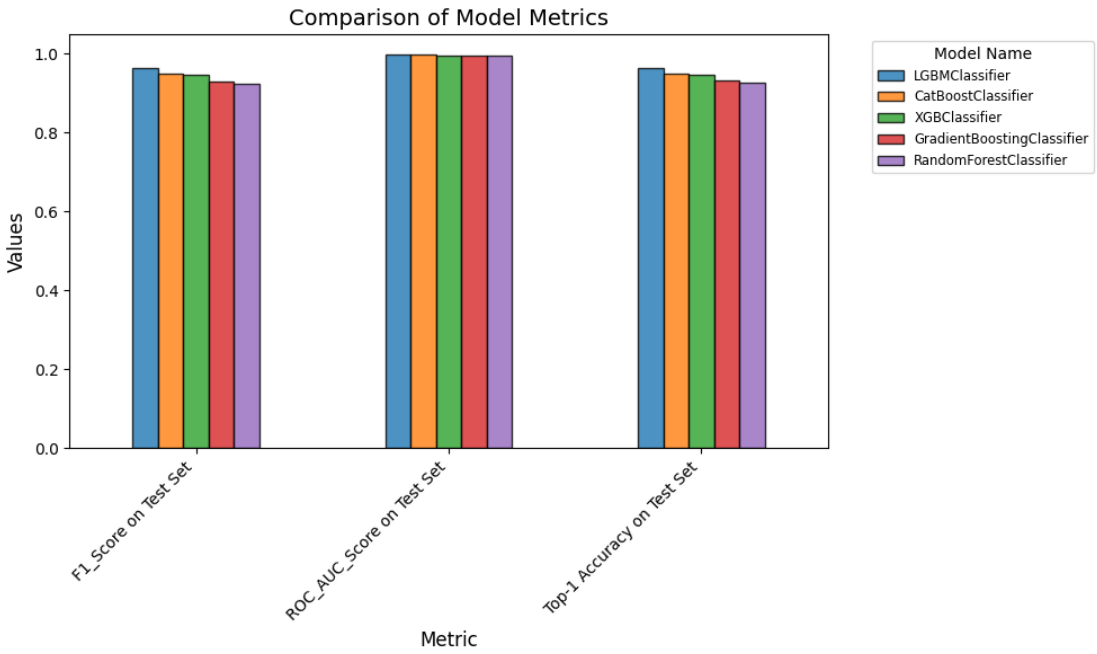
- 從 Top1 Accuracy 及 F1 score 可以看出，模型的準確度大概都在 70%~80% 之間，其中以 LGBMClassifier 的表現最好。
- 從 AUROC 指標可以看出，因為沒有分類資料不平衡問題，訓練出來的模型對於各個類別的預測能力都相當不錯，都達到 90% 以上。

表: 使用經過特徵挑選後的資料集進行訓練

Model	**Top1 Accuracy**	**F1 score**	**AUROC**
LGBMClassifier	0.789941	0.787195	0.957348
CatBoostClassifier	0.775148	0.767360	0.952991
RandomForestClassifier	0.757396	0.749795	0.954587
XGBClassifier	0.751479	0.744141	0.942010
GradientBoostingClassifier	0.748521	0.744003	0.934714

ii. 為提升模型預測能力，第二次使用有包含身高和體中的特徵資料集進行訓練。

- 從五個模型的比較結果來看，整體模型預測能力都提升到 85% 以上。
- 其中 LGBMClassifier 在三個指標 F1 score、AUROC、Accuracy 的表現都優於其他模型
- 因此選定為最終訓練的模型，並使用 5-fold Cross-Validation 評估其模型對於不同資料區塊的表現。



圖：五個模型在各評估指標的表現

iii. 訓練資料集 (來自 80% 的原始資料集) 結果

- 此訓練資料集包含所有特徵，訓練與驗證資料集切分採用 5-fold，在每個 fold 的循環中會對訓練資料 (其他 4 folds) 實行 SMOTE 來維持切分後的分類資料的平衡。
- 從 F1 score 與 Top-1 Accuracy 的分數可以知道，模型準確度大約是 95%。
- 從 AUROC 以及展開的分類報告可以看出，最終訓練模型對於各個類別的表現都是 90% 以上的表現。

表: LGBM 使用 5-fold Cross-Validation 的分數

Metric	Average(mean) score
F1 score	0.9596
Top1 Accuracy	0.9607
AUROC	0.9978

表: LGBM 各分類預測表現狀況

	Precision	Recall	F1 score	Support
Insufficient_Weight	0.92	0.98	0.95	56
Normal_Weight	0.95	0.89	0.92	62
Obesity_Type_I	0.99	0.99	0.99	78
Obesity_Type_II	1.00	0.98	0.99	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	0.93	0.96	0.95	56
Overweight_Level_II	1.00	0.98	0.99	50
accuracy			0.97	423
macro avg	0.97	0.97	0.97	423
weighted avg	0.97	0.97	0.97	423

iv. 測試資料集 (來自 20% 的原始資料集) 結果

- 最終模型在測試資料集各指標結果如下，各指標分數都與訓練資料集的結果相近，表示在所訓練的模型效果在沒看過的資料是有足夠的預測能力。
  - F1-Score: 0.9687

- Top 1 Accuracy: 0.9693
- AUROC: 0.9989

- 從下圖 ROC curve 可以看出各類別的預測能力接近，而線條呈現鋸齒狀，表示在資料集內的樣本數較少才會有此情況。

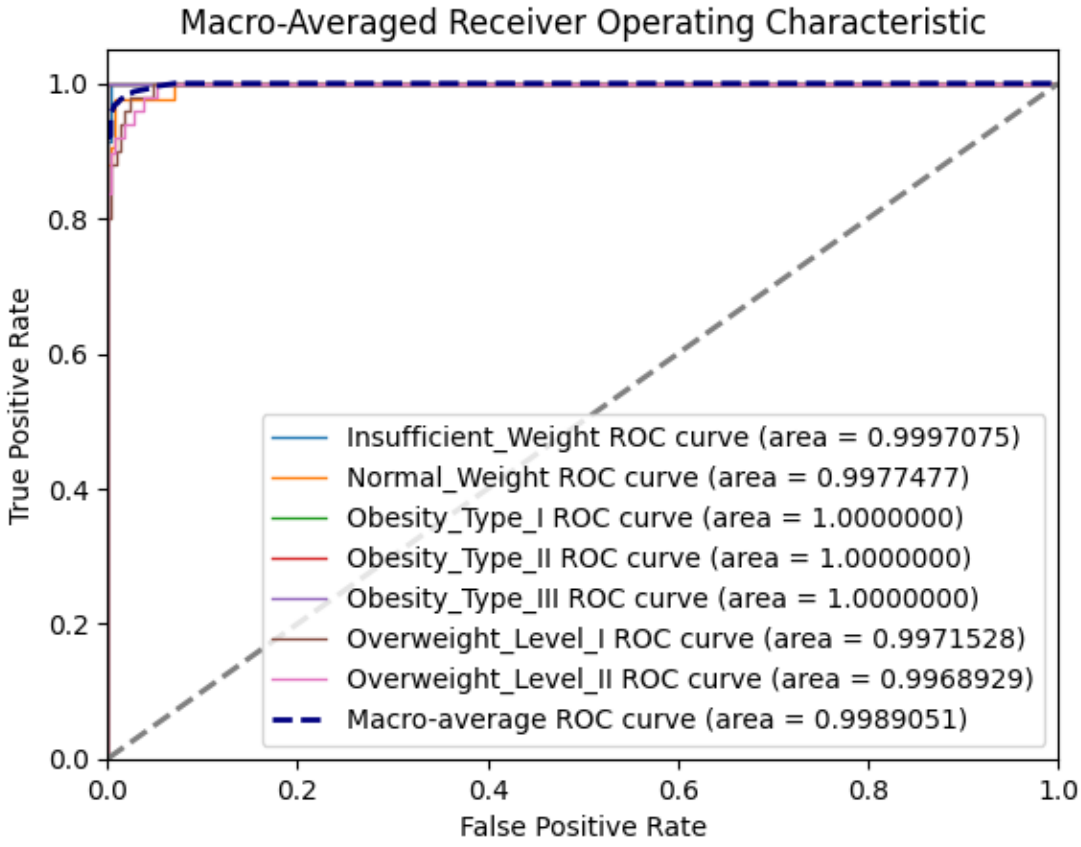


圖: LGBM 在測試資料集的 ROC curve

- 使用關聯分析尋找生活習慣因素與肥胖程度關聯的結果:

i. **不規律的飲食行為與肥胖高度相關:**

有時候攝取高熱量食物 (CAEC_Sometimes) 的男性，並且有家族肥胖史，與第二型肥胖 (Obesity_Type_II) 關聯顯著，LIFT 為 2.35，表示確實具有正相關性。

ii. **有攝取高熱量食物傾向者易肥胖**

頻繁攝取高熱量食物 (FAVC_yes) 的男性，尤其有家族肥胖史者，肥胖風險明顯提升，Confidence 為 24.08%，LIFT 達到 2.19，表示存在正相關性。

iii. **性別差異顯著**

男性族群的肥胖型態，尤其是第二型肥胖，與多重因素的共同作用（如不規律飲食、攝取高熱量食物和家族肥胖史）息息相關，表現出性別差異。

iv. **家族肥胖史是關鍵**

家族中有肥胖史的人，即使僅部分不健康的飲食行為（如有時候進食高熱量食物），仍與肥胖層度有相關，Confidence 為 10.5%，存在一定程度的相關性。

- 針對實驗結果提出看法，並進行細項討論
  - 資料源本身
    i. 樣本數
      - 總樣本數少，在不同特徵廣泛度不足，如: 年齡。
      - 但各類別資料筆數接近，對於各自類別仍具有代表性，足夠進行模型訓練。

    ii. 年齡範圍
      - 多數資料落在 20~30 歲，因此對於其他年齡層的資訊代表性不足。
      - 會產生一定程度的資料偏差，因此分類模型可能對於年齡介於 20~30 之間的預測結果較具有參考性。

    iii. 人種多來自中南美洲
      - 對於不同人種的資訊缺乏，原始特徵沒有標記種族，僅能知道資料來源的國家。
      - 有可能因為種族差異，導致訓練模型僅對於特定人種有較高的預測能力。

    iv. BMI 定義的過重與肥胖層級是基於墨西哥人的數據
      - 如果新的資料是來自於不同地區人種，可能會產生偏差影響模型預測效果。因為種族整體差異是明顯的，同樣的 BMI 區間，在不同種族之間存在不同分級方式。

  - 模型表現
    i. 分類模型任務
      - 使用所有特徵 vs. 不包含身高體重
        - 對於預測肥胖層級的問題，身高與體重在此分類模型中為重要程度較高的特徵。
        - 如果特徵中未包含身高體重資訊，模型表現會下降兩成左右，降至 75% ～ 78% 的預測表現。
      - 訓練資料集 vs. 測試資料集

- 三個主要評估指標都呈現相近的分數，表示模型對於新的尚未看過的資料有良好的表現，同時也對於不同類別的預測能力接近。

- 重要特徵排名
  - 分數為 Gain 值，表示特徵在分裂節點中帶來的平均增益（Information Gain）。
  - 年齡在重要特徵中位於第三僅次於身高與體重。
    - 隨著年齡增長，身體的新陳代謝狀況會影響肥胖層級。
  - 身體活動及飲食的頻率相對於活動及飲食的內容 (例如: 做了什麼、吃了什麼) 對於分類模型的判斷更為重要，以下為對比的範例:
    - 你多久會運動一次? (Gain: 1956) vs. 你每天搭什麼交通工具? (Gain: 875)
    - 你每天吃幾餐主餐? (Gain: 1632) vs. 你常吃高熱量食物嗎？(Gain: 878)

表: LGBM 的 feature importance

特徵欄位	欄位描述	分數(Gain)
Weight	體重	13025
Height	身高	11187
Age	年齡	5422
FAF	你多久會運動一次?	1956
NCP	你每天吃幾餐主餐?	1632
TUE	你每天吃幾餐主餐?	1461
FCVC	你每餐中會吃蔬菜嗎?	1355
CH2O	你每天喝多少水?	1217
CALC	你喝酒的頻率?	1082
CAEC	你喝酒的頻率?	1063
Gender	性別	965
FAVC	你常吃高熱量食物嗎？	878

MTRANS	你每天搭什麼交通工具?	875
family_history_with_overweight	家族肥胖史	651
SCC	你會控制每天攝取的卡路里嗎？	356
SMOKE	抽菸	26

- 綜合探討
  - i. 特徵與肥胖層級的關聯性，跟用於模型進行分類肥胖層級的特徵，存在部分差異。而一般常見認為影響個人肥胖層級的關聯特徵屬於前者（關聯性分析），分類模型主要用於判斷的屬性則屬於個人生理資訊（如：體重、年齡）。
  - ii. 分類模型對於未來預測的可能性: 因為目標類別是基於 BMI 劃分，對於目前的肥胖層級判斷有極高的準確率。但如能用於預測未來此人可能所屬的肥胖層級，例如一個月以後此人可能會從 Overweight_Level_II 變成 Obesity_Type_I 的結果，在實際情況較能有廣泛應用，達到預防的效果。因此，在資料搜集階段，可以考慮填寫問卷問題與提供個人生理資訊存在時間差，這樣訓練的模型可能具有對於未來肥胖層級預測的能力。
  - iii. 需要提高樣本數與廣泛程度，對於不同年齡層與不同地區人種的資料缺少的情況下，對於這類資料的預測表現會變差。
    - 此資料集在 30 歲以上的樣本數不足，因此對於年長年齡層的模型預測將會不足。
    - 蒐集的資料並沒有非中南美洲以外的國家與人種資訊，人種之間的個體差異可能會導致預測結果失準。
    - 肥胖層級欄位的類別定義是基於墨西成人的 BMI 數值進行劃分，同樣會因為人種之間差異導致分類錯誤。

5、 過程中遭遇的挑戰以及總結(Discussion and Conclusion)

- 實作 Project 中遇到的難題，以及對應之解決方法
  - 環境一致性：在初期每個人使用的環境存在差異，如作業系統、程式語言版本、程式版本，因此採用線上協作的方式，一方面不用花時間建立研發環境，另一方面透過共同編輯達到一致性，降低研究時間成本。
  - 不同任務需要的資料前處理不同：由於關聯分析與模型需要對於原始資料進行不同的前處理，我們依照開發階段切分不同的程式檔，前處

理程式的最後會產生處理後的檔案，在模型訓練階段只需載入各自需要的前處理檔案版本。

- 關聯式分析資料範圍過大：年齡、身高、體重這類型資料區間比較廣，因此需要使用區間重新編碼，例如將年齡切分成：青年、壯年、老年等類別，而不是之間使用年齡數字。

- 針對此 Project 總結

本次專案通過個人及生活方式因素預測肥胖程度，我們完成了目標問題的分析、資料處理與模型訓練，並以實驗結果驗證了方法的有效性。此專案目標在於找出與肥胖有關的關鍵因子，並利用資料探勘技術提升對肥胖層級判斷的準確性，希望藉此提供預防策略。

在數據分析過程中，我們針對資料集進行詳細的前處理，例如清理異常值與進行特徵選取，以確保輸入模型的資料品質，讓其具有代表性，在模型訓練階段能有更好的結果。接著透過與多個模型訓練結果比較（如 LightGBM、CatBoost 等模型），其中 LightGBM 的表現最佳，因此選擇此模型作為最佳模型進行訓練與評估。我們在訓練及評估階段使用 K-Fold Cross Validation 的方法評估模型的穩健性及準確性，同時用於避免 overfitting 的發生。實驗結果顯示，有包含身高與體重特徵的訓練模型，能有明顯的效能提升，其準確率達到 95% 以上，在評估指標 AUROC 更是接近 99%。此外，在關聯分析的結果顯示高熱量飲食、不規律飲食與家族肥胖史，在肥胖成因中是相對重要的。

然而，本專案仍存在挑戰，包括資料樣本數不足、資料來源侷限於特定地區，以及 BMI 指標跨種族適用性等問題，為了讓模型能對於更廣泛類型資料仍有相同的表現，未來可考慮增加資料收集範圍，像是不同年齡層、地區、國家、人種的樣本數增加，且針對不同地區的人給予相對應合適的肥胖層級的分類標準。

總結來說，此專案成功完成了起初設立的目標，並發現了對於過重與肥胖的相關研究相關的實用與實作參考，同時我們也認識到資料廣泛性及模型應用性的改進空間，這將作為後續研究的重要方向。

6、 參考文獻 (Reference)

Cheng, X. (2021). Does physical activity predict obesity—A machine learning and statistical method-based analysis. *International Journal of environmental research and public Health*, *18*(8), 3966. PudMed. 10.3390/ijerph18083966

Ke, G. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*(52). ACM. doi/10.5555/3294996.3295074

LA OBESIDAD, S. Y., & GENERAL, C. D. S. (2010, August 4). *Diario Oficial de la Federación*. DOF - Diario Oficial de la Federación. Retrieved December 6, 2024, from http://diariooficial.gob.mx/nota_detalle.php?codigo=5154226&fecha=04/08/2010#gsc.tab=0

Palechor, F. M., & Manotas, A. D. l. H. (2021). *Estimation of obesity levels UCI dataset*. 10.34740/KAGGLE/DSV/2918196

Quiroz, J. P. S. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. *Informatics in Medicine Unlocked*, *29*, 100901. Elsevier. 10.1016/j.imu.2022.100901

Thamrin, S. A. (2021). Predicting obesity in adults using machine learning techniques: an analysis of Indonesian basic health research 2018. *Frontiers in nutrition*, *8*, 669155. PubMed. 10.3389/fnut.2021.669155

7、 組員分工與各自執行細項 (Work distribution chart)

姓名/階段	**Proposal**	**Experiment**	**Final**
康峻瑋	Dataset survey, Slide, Previous work survey	FP-growth, ECLAT, lightGBM	Slide, Video, Report
林滋隆	Dataset survey, Slide, Previous work survey	GradientBoosting, CatBoost	Slide, Video, Report
許茗鈞	Dataset survey, Slide, Previous work survey	ECLAT, XGBoost	Slide, Video, Report

黃正鵬	Dataset survey, Slide, Previous work survey	EDA, Data Preprocessing, RandomForest training, Evaluation	Slide, Video, Report