國立臺灣大學工學院工程科學及海洋工程學系

碩士論文

Department of Engineering Science and Ocean Engineering

College of Engineering

National Taiwan University

Master Thesis

以特徵重要度與間距保留法改良 k 匿名演算法以提高匿名資料之機器學習成效

Improving k-Anonymization Algorithm to Enhance Machine Learning Performance of Anonymous Data through Feature Importance and Margin Preservation

李丞彥

Cheng-Yen Lee

指導教授: 張瑞益 博士

Advisor: Ray-I Chang, Ph.D.

中華民國 112 年 7 月

July, 2023

doi:10.6342/NTU202303509

摘要

機器學習應用蓬勃發展，為各領域創造難以估量之商業價值，其中高品質資料是為訓練出優良模型之關鍵；然而個人資料之使用具有隱私侵害風險，因此各地區組織皆持續完善法規以規範資料控制者之收集、存取行為，如我國行政院於 2023 年修訂之《個人資料保護法》與歐盟執委會 (European Commission，EC) 於 2018 年訂定之《General Data Protection Regulation》等，此些法規對於遭存取之個人資料須經去識別化處理之限制，將降低其中包含之資訊，是為機器學習應用之一大挑戰。

個人資料之使用授權須透明且嚴謹，因此經常先釐清應用需求，方才尋求途徑收集資料並取得資料主體之授權，其中各國為促進節能減碳之公共利益所推動之智慧電表布建即為一例；而為提升資料共享效率進而挖掘資料潛在價值，先前研究已提出有限誤差資料隱私保護 (Bounded-Error Data Privacy Protection, BEDPP) 架構，以實現具隱私保護之物聯網資料共享服務，惟該架構尚缺乏明確指標量化去識別化程度，是為本研究欲探討之應用情境之一。

k 匿名是為一常見之去識別化手段，適於導入 BEDPP 架構，然而過往之研究皆致力於最佳化整體性誤差指標，並未將機器學習目標或特徵重要度於應用情境之高異質性納入考量，因此本研究以使用匿名資料訓練分類預測模型為例，於 k 匿名演算法中依據特徵重要度 (Feature Importance，FI) 作為權重分派誤差，亦或是先基於目標特徵對資料進行分群以保留類別間距 (Margin Preserving，MP)，再將各資料群匿名化後合併，以提高匿名資料於機器學習之成效。

實驗結果顯示，相較於透過原 k 匿名演算法匿名化資料，再經本研究採用之機器學習模型進行分類預測，考量特徵重要度 k 匿名化資料平均改良機器學習成

doi:10.6342/NTU202303509

效之幅度可達 10.7%，而先對資料進行類別間距保留分群再匿名化各群後合併之方法平均改良機器學習成效之幅度可達 17.40%。

關鍵字：去識別化、k 匿名、資料隱私、特徵重要度、機器學習

doi:10.6342/NTU202303509

Abstract

Machine learning applications have been proliferating and creating incalculable business value across various fields. High-quality data plays a crucial role in training SOTA models. However, the use of personal data poses risks to privacy infringement. Therefore, organizations worldwide continually improve regulations to govern the collection and access behaviors of data controllers. Examples of such regulations include Taiwan's Personal Data Protection Act established by the Executive Yuan in 2015 and the European Commission's General Data Protection Regulation introduced in 2018. These regulations impose restrictions on the handling of accessed personal data, requiring de-identification processes that reduce the information contained therein, presenting a significant challenge for machine learning applications.

The authorization of personal data applications should be transparent and rigorous. Therefore, it is common to clarify the application requirements first and then seek ways to collect data with the consent of data subjects. An example of this is the deployment of smart meters driven by the public interest in energy conservation and carbon reduction in various countries. To enhance data sharing efficiency and unlock the potential value of data, previous studies have proposed the Bounded-Error Data Privacy Protection (BEDPP) framework for privacy-preserving IoT data sharing services. However, this framework

doi:10.6342/NTU202303509

lacks a clear indicator for quantifying the level of de-identification, which is one of the application context explored in this research.

k-anonymity is a common de-identification method. However, previous research has mainly focused on optimizing holistic error metrics without considering the high heterogeneity of machine learning objectives or feature importance in the application context. Therefore, our research used training a classification model on anonymous data as an example. In our k-anonymity, errors are assigned based on feature importance as weights or cluster the data first based on the target features to present margin preserving and then anonymize each cluster. These approaches aim to improve the performance of the machine learning model trained by anonymous data.

Compare to anonymizing data using the original k-anonymity algorithm and subsequently performing classification prediction, the experimental results demonstrate that anonymizing data considering feature importance leads to average model performance improvement up to 10.7% and presenting margin preserving before anonymizing data shows a average model performance improvement up to 17.40%.

Keywords: De-identification, k-Anonymity, Data Privacy, Feature Importance, Machine Learning

doi:10.6342/NTU202303509

目錄

doi:10.6342/NTU202303509

doi:10.6342/NTU202303509

doi:10.6342/NTU202303509

圖目錄

doi:10.6342/NTU202303509

表目錄

doi:10.6342/NTU202303509

第一章　緒論

1.1　研究背景

資料分析之應用情境與日俱增，企業、政府乃至個人對於資料之重視程度亦日益提升，然而當資料涉及個人特徵時，資料主體則有被識別進而導致隱私遭侵害之風險；為維護隱私權益，各地區組織皆陸續頒布相關法規以規範個人資料自收集、儲存至授權、處理及使用等行為，如我國行政院於 2023 年修訂之《個人資料保護法》[1] 與歐盟執委會 (European Commission，EC) 於 2018 年訂定之《General Data Protection Regulation》(GDPR) [2] 等，然而出於各地區之風俗民情、施政方針差異，法規內容與執行進度不盡相同。

由於個人資料之使用具隱私侵害風險，授權之使用形式須明確且嚴密，因此經常先釐清應用需求，方才尋求途徑收集資料並取得資料主體之授權，於公共利益之促進上即有不少案例，如我國衛生福利部為推動醫藥衛生相關領域研究而建置之全民健康保險研究資料庫 [3] 與歐盟為實現節能減碳而倡導之我的能源資料 (My Energy Data，MED) [4] 服務等；根據研究 [5] 所述，MED 服務始於電力管理需求，歐盟成員國預期藉由能源用戶之需求面進行需量反應，以提升能源使用效率，進而實現節能減碳目標，然而電力使用習慣是為個人隱私之一環，故該資料之收集、共享須遵守 GDPR 之規範，其建置方式如圖 1.1 所示。

doi:10.6342/NTU202303509

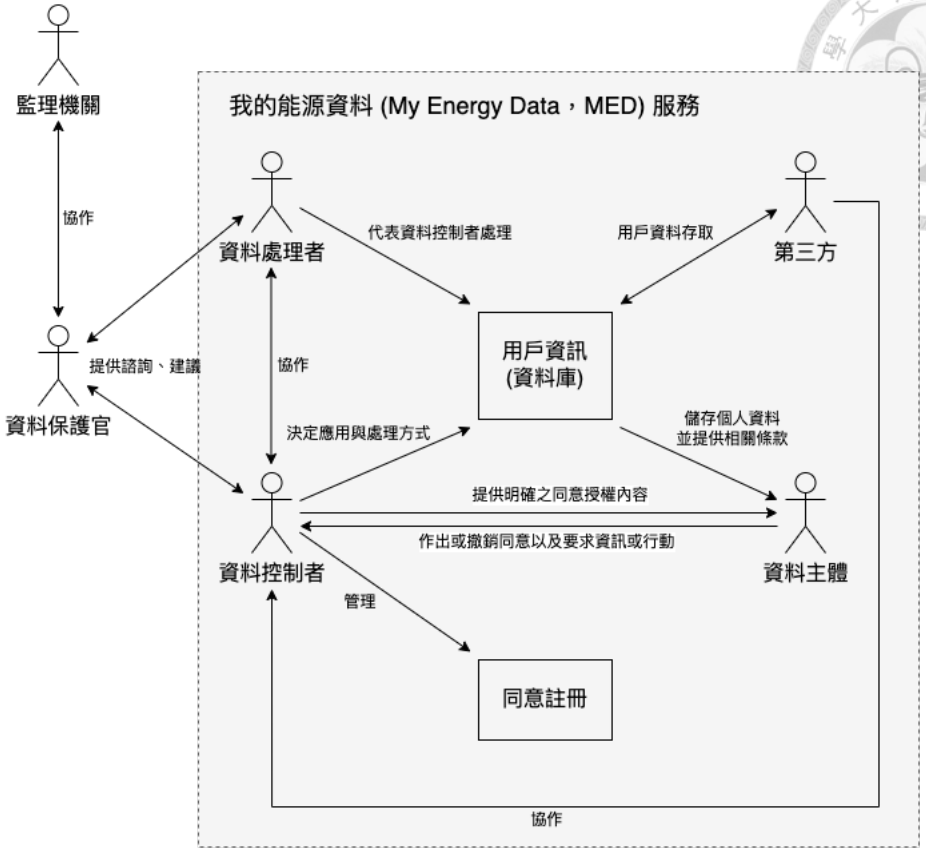

Figure 1.1: MED 服務於 GDPR 規範下之建置方式

考量資料共享之流程繁瑣，且中心化形式之資料儲存、管理於信任、安全和移植性上有所缺乏，先前研究曾提出有限誤差資料隱私保護 (Bounded-Error Data Privacy Protection, BEDPP) 系統架構 [6]，以去中心化形式實現完整之物聯網資料收受與共享流程，其於智能合約交易介面所提供之誤差級別參數設計，使得資料控制者能供應不同隱私保護層級之資料集；BEDPP 可劃分為有限誤差物聯網系統 (Bounded-Error IoT，BEIoT) 與基於區塊鏈之有限誤差資料市集 (Bounded-Error Data Market on Blockchain，BEDMoB)，其中 BEIoT 涵蓋物聯網資料自終端裝置收集至資料庫伺服器儲存之流程，其與本研究較無相關，故在此並不詳述，而 BEDMoB 則涵蓋資料共享之完整流程，以下將進一步介紹。

BEDMoB 是為一去中心化資料共享平台，資料控制者能於以太坊 [7] 區塊鏈公開資料集資訊，而第三方能以智能合約 [8] 查詢資料集資訊與發起交易；若

doi:10.6342/NTU202303509

交易成立，則資料控制者得以委派資料處理者進行資料處理，透過先前研究提出之分層有限誤差運行長度編碼壓縮演算法 (Layered Bounded-Error Run Length Encoding, LBE-RLE) 與非對稱加密技術實現安全且具多層級隱私保護，並藉由 IPFS [9] 轉移資料集，完整之流程如圖 1.2 所示。

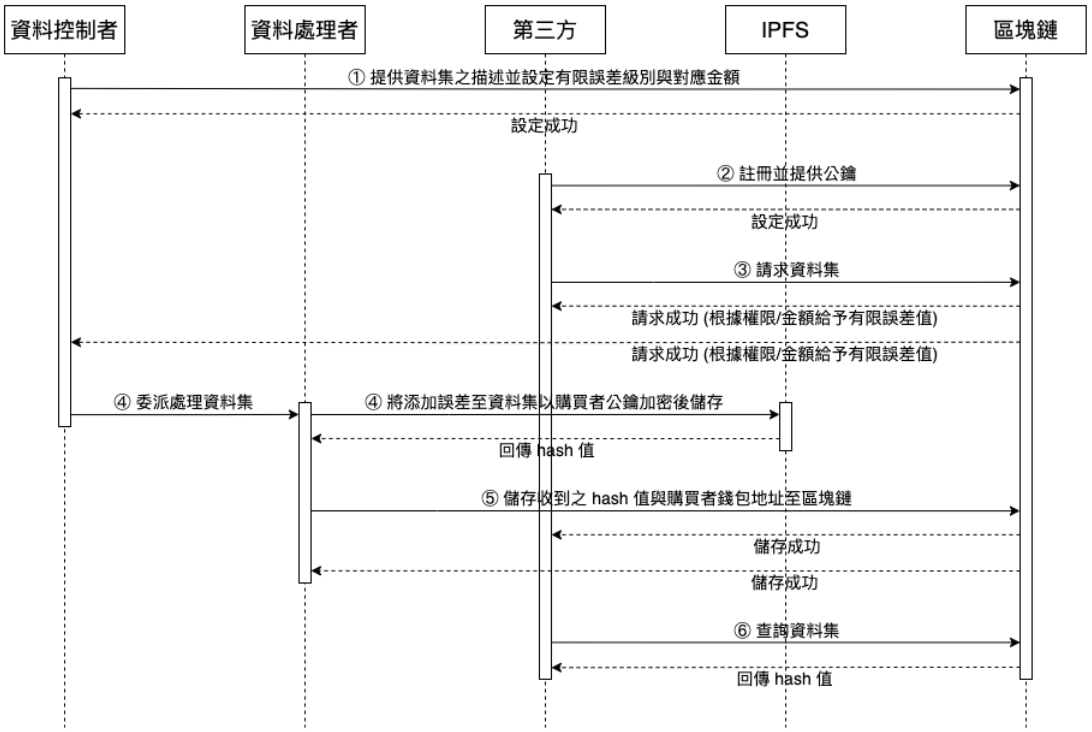

Figure 1.2: BEDPP 中 BEDMoB 之資料交易流程

1.2 研究動機

於 GDPR 之規範下，個人資料之收集、共享流程將有三個關鍵角色—資料主體、資料控制者及資料處理者，資料主體是為能透過資料所含之識別碼 (如: 身分證字號、姓名、網路 IP 等) 直接識別亦或是亞識別碼 (如: 性別、職業、戶籍地等) 間接識別之自然人；而資料控制者為決定個人資料處理目的和方式之自然人、法人、公權力機關或其代表人，該角色會決定所蒐集之個人資料範疇、利用方式以及共享對象，其完整之決策權亦使其須承擔個人資料處理之最終責任；資料處理

doi:10.6342/NTU202303509

者則為代表資料控制者處理個人資料之自然人、法人、公權力機關、其他機構或其代表人，該角色與資料控制者間須簽訂契約並遵照資料控制者之指示行動，對於個人資料之處理方式無決策權，僅能對儲存與處理之技術架構進行選擇。

BEDPP 作為具隱私保護之資料共享解決方案，亦須遵守 GDPR 之規範，方能應用於生產環境，儘管 GDPR 尚無明訂之技術安全標準，僅要求資料處理者應採用「適當」之技術確保資訊安全，然而原架構之隱私保護措施尚無法提供具公信力之指標以量化資料之去識別化程度，是為可改良之處；而 k 匿名 [10] 即為一有效之去識別化技術，於該方法之處理下，每筆資料皆將無法從至少 k 筆資料中遭識別，其參數 k 是為明確之去識別化指標，適於完善 BEDPP 架構。

欲訓練出優良之機器學習模型，高品質資料是為不可或缺之要素，然而面對個人資料時，卻必須藉由添加誤差將資料去識別化以保護資料主體之隱私，兩目標於資料精度上之對立使得以去識別化資料訓練出精準之機器學習模型成為一大挑戰；盤點過往研究提出之去識別化演算法皆致力於最佳化整體性誤差指標，並未將機器學習目標或特徵重要度之高異質性納入考量，可能導致去識別化資料無法進一步適配機器學習應用情境，是為本研究欲探討之議題。

1.3 研究目的

本研究以匿名資料之分類預測為例，根據應用情境之特徵重要度與機器學習目標，於 k 匿名流程導入特徵重要度概念或間距保留法以提高匿名資料之機器學習成效；其中特徵重要度是為不同特徵對於同一機器學習目標之相對重要程度，本研究於 k 匿名演算法中依據特徵重要度作為權重分派誤差，使較重要特徵之誤差得以降低，以減少誤差對機器學習成效之影響；而間距保留法是為基於目標特

doi:10.6342/NTU202303509

徵對資料進行分群以保留類別間距之方法，針對機器學習分類目標，本研究則將欲匿名之資料集先以間距保留法進行分群，再將各資料群匿名化後合併，以降低資料匿名化對機器學習分類模型所造成之類別混淆。

doi:10.6342/NTU202303509

第二章　文獻探討

2.1　特徵選擇演算法

特徵選擇於資料前處理的流程中常見且重要，其旨在從原始特徵集合中選擇最優的特徵子集，以減少後續分析之運算成本並提高效能和結果之可解釋性。文獻 [11] 探討過濾法 (Filter Method)、包裝法 (Wrapper Method)、嵌入法 (Embedded Method) 及混合法 (Hybrid Method) 等常見之特徵選擇方法，並整理此些方法之基本原理、優缺點以及在特殊情況下的選型問題；其中包裝法是為一種依據機器學習目標選擇模型以衡量特徵子集優劣之方法，儘管該方法於計算資源方面需求較高，然而普遍能獲得最佳解；包裝法須要採取搜尋策略以枚舉所衡量之特徵子集，而反向消除即為常見策略之一，該策略藉由從原始特徵集合移除不同特徵以產生特徵子集，因此當包裝法採用反向消除搜尋策略以進行特徵選擇時，即可觀察移除特定特徵所產生之特徵子集對於模型成效之影響，藉此評估是否需將該特徵真正移除。

當以採用反向消除搜尋策略之包裝法時，會藉由逐一移除特徵以評估該特徵對於機器學習任務之影響，其移除特徵之方法與匿名化過程將特徵進行徹底泛化對資料集所產生之影響極為相似，是為啟發本研究發展特徵重要度衡量方法之關鍵。

2.2　k 匿名演算法

k 匿名演算法藉由減少資料所含之資訊以避免資料主體遭識別，經其處理之資料集中，每筆資料皆無法從至少 k 筆資料中遭識別；其實現能分為兩步

doi:10.6342/NTU202303509

驟，首先，原始資料集將會劃分為多個稱為等價類之子資料集，並且每個等價類至少須包含 k 筆資料，接著，針對屬於同一等價類之資料，其被視為準識別碼 (Quasi-Identifier，QI) 之特徵都將經過處理而一致化；其中 QI 是為有助於識別資料主體，但單獨存在不足以確定識別之資訊，而等價類是為資料集中經過 k 匿名演算法處理後，被歸屬於同一集合之資料，此些資料通常具有相似之特徵，並且在匿名化完成後將具有相同之 QI，於過往之研究中經常使用標準化確定性懲罰 (Normalized Certainty Penalty，NCP)[12] 作為衡量等價類資訊遺失程度之指標。

於眾多一致化方法中，泛化 (Generalization) 是相當常見之技巧，泛化之策略又可分為全域和局部兩種，全域策略會將同一特徵之相同值進行相同之泛化處理，而局部方法則允許同一特徵之相同值進行不同之泛化處理 (如: 資料集中存在兩筆身高皆為 170 的資料，其中一筆資料之身高可能被泛化為 165 - 170，另一筆資料之身高可能被泛化為 170 - 175)；泛化處理存在多種實作方式，其中值泛化階層樹 (Value Generalization Hierarchy，VGH) [13] 是為一常見之實作，VGH 是一種樹狀結構，每個節點代表特徵值之集合或範圍，並且每個節點都是父節點之子集合或子範圍，當特徵值之一致化結果趨向根節點稱為泛化 (Generalization)，反之趨向葉節點則稱為特化 (Specialization)；為減少演算法之間的差異性，本研究選擇採用 VGH 對等價類進行泛化，本小節將介紹本研究所使用之 k 匿名演算法。

2.2.1 k-NN Clustering-Based 匿名法

k-NN Clustering-Based 匿名法 (k-Nearest Neighbors Clustering-Based Anonymization，CBA) [14] 存在多種實作版本，然而其之間相異度並不高；於本研究採用之實作中，透過多輪迭代以產生等價類，每一輪迭代將隨機選擇一筆資料作為基準，以 NCP 作為距離公式搜尋與該資料最近之 k - 1 筆資料，而此 k 筆資料將被

doi:10.6342/NTU202303509

合併為一等價類並從原始資料集移除；迭代過程將會持續進行，直到所有資料皆被移除，亦或是剩餘之資料數量少於 k，若為後者，則將其逐一分配至與其距離最近之等價類。

2.2.2 Top-Down Greedy 匿名法

Top-Down Greedy 匿名法 (Top-Down Greedy Anonymization，TDGA) [12] 是為一種結合貪婪與啟發概念之 k 匿名演算法，其以遞迴之形式持續切割資料集再進行局部泛化以形成等價類；其每一層遞迴旨在貪婪地降低最終結果之 NCP，因此以 NCP 作為距離公式尋找欲切割子資料集中距離最遠之兩筆資料，並以此兩筆資料作為中心對其它資料進行最小距離分群，此遞迴過程將持續至切割得少於 k 筆資料之子資料集 (假設大小為 s) 為止；而為使該子資料集能滿足 k 匿名之限制，將會於所有大小至少為 2k - s 之等價類搜尋大小為 k - s 之最近子資料集，比較與該子資料集合併及與最近等價類合併將會產生之 NCP，並選擇較小者進行合併；而為有效率地尋得距離最遠之兩筆資料，該演算法採用一種啟發式方法取代暴力搜尋法，首先，隨機選擇一筆資料 u 作為起點，搜尋距離最遠之資料 v，接著，以 v 取代 u 再次進行搜尋，反覆該流程數次，最終得到之資料 u 和 v 將近似為最佳結果，且時間複雜度將能自 $O(n^2)$ 降低至 $O(n)$。

2.2.3 Mondrian 匿名法

Mondrian 匿名法 (Mondrian Anonymization，MA) [15] 是遞迴地切割資料集以產生等價類之 k 匿名演算法；於每一層遞迴，其將計算欲切割之子資料集中各特徵之標準化值範圍，並選擇最大者於中位數進行切割，該遞迴流程將持續至所有特徵皆無法切割出兩個大小大於 k 之子資料集為止。

doi:10.6342/NTU202303509

2.2.4 Optimal Lattice 匿名法

Optimal Lattice 匿名法 (Optimal Lattice Anonymization，OLA) [16] 是一種最佳化 k 匿名演算法，其以 VGH 作為核心，通過組合所有特徵之 VGH 階層產生節點，並依據 VGH 之父子節點關係建立路徑，以形成包含所有泛化可能性且具階層性之網格；根據此網格，該演算法將於其中尋找最佳節點，首先，基於網格之階層性採用二元搜索策略篩選出符合 k 匿名限制之節點集合，接著，各版本依據自定義指標，於符合 k 匿名限制之節點中選擇最佳節點。

2.3 機器學習模型

機器學習模型是為一種數學模型或演算法，其能自既有資料集中提取有用資訊以調整模型參數或建構模型輸出，進而達成不同任務；依據機器學習之目標，模型將能劃分為回歸、分類、聚類、生成等，本小節將介紹本研究所使用之機器學習分類模型。

2.3.1 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) 是為一基於相似度之機器學習模型，其採用距離作為相似度之依據，以輸入資料最接近之 k 筆資料進行類別統計，並以最高數量之類別作為預測結果；由於該模型之預測結果深受訓練資料集之類別比例影響，當該資料集偏離母體之分佈時，將難以得到良好之預測成效。

doi:10.6342/NTU202303509

2.3.2 Support Vector Machine

Support Vector Machine (SVM) 是為一劃分類別邊界之機器學習模型，其極大地保留類別間距之特性，當面對少量訓練資料亦能有良好之預測成效；此外，其還能藉由採用不同核函數以適應不同特性之資料集，如線性核函數、多項式核函數及高斯核函數等，而本研究所選擇之線性核函數具有快速收斂和避免過度擬合之優勢，惟該模型對懲罰係數敏感程度較高，須審慎調整。

2.3.3 Random Forest

Random Forest (RF) [17] 是為一決策樹之集成模型，其以不同子資料集訓練出多個決策樹，再藉由投票之形式結合此些決策樹之預測結果；儘管該模型改善單個決策樹容易過度擬合之缺陷，但須謹慎訂定決策樹之數量以於模型之準確性與計算效率上取得平衡。

2.3.4 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGB) [18] 是為一梯度提升模型，與 RF 相似，其結合多個決策樹進行預測以達到高穩健性，然而，相異之處在於該模型採用提升之技巧迭代地訓練出愈趨精準之決策樹，同時亦藉由正則化方法控制其複雜度以有效地避免過度擬合之發生。

doi:10.6342/NTU202303509

2.4 k 匿名演算法於機器學習之可用性探討

探討 k 匿名演算法可用性之研究稀缺，僅有文獻 [19] 基於多個資料集、k 匿名演算法 (CBA、TDGA、MA、OLA) 及機器學習模型 (k-NN、SVM、RF、XGB) 觀察不同變因組合對於機器學習成效之影響，其中「重要特徵之泛化程度與機器學習成效具高度相關性」是該文獻之重要結論，亦為啟發本研究以特徵重要度改良 k 匿名演算法之關鍵。

doi:10.6342/NTU202303509

第三章　研究方法設計

本研究透過考量特徵重要度或結合間距保留法改良 k 匿名流程以提高匿名資料之機器學習成效，參考文獻 [19] 之實驗流程，將實驗情境設置為匿名資料之分類預測，並採用 CBA、TDGA、MA、OLA 等 k 匿名演算法及 k-NN、RF、XGB 等機器學習模型進行實驗；針對機器學習模型之超參數，k-NN 之最近鄰居數量設為 10，RF 之決策樹數量設為 300，XGB 之決策樹數量設為 100，而為避免 k-NN 與 k 匿名演算法之參數混淆，將以 10-NN 代稱 k-NN，本章節將分別就兩面向之改良方法進行介紹。

3.1　考量特徵重要度之 k 匿名演算法

本研究提出一創新之演算法改良思維，能使得既有之 k 匿名演算法能依據特徵重要度分派誤差，因此統稱為考量特徵重要度之 k 匿名演算法；改良之方式採兩步驟進行，首先，將藉由反向消除策略列舉特徵子集，並根據機器學習成效影響衡量所缺乏特徵之重要度，接著，將衡量得之特徵重要度導入原 k 匿名演算法調節特徵泛化或特化程度之步驟。

為衡量改良幅度，本研究以 k 匿名演算法作為實驗變因，以資料集、k 匿名之參數 k、機器學習模型作為控制變因，比較多種控制變因組合下所得之機器學習成效，實驗流程如圖 3.1 所示。

doi:10.6342/NTU202303509

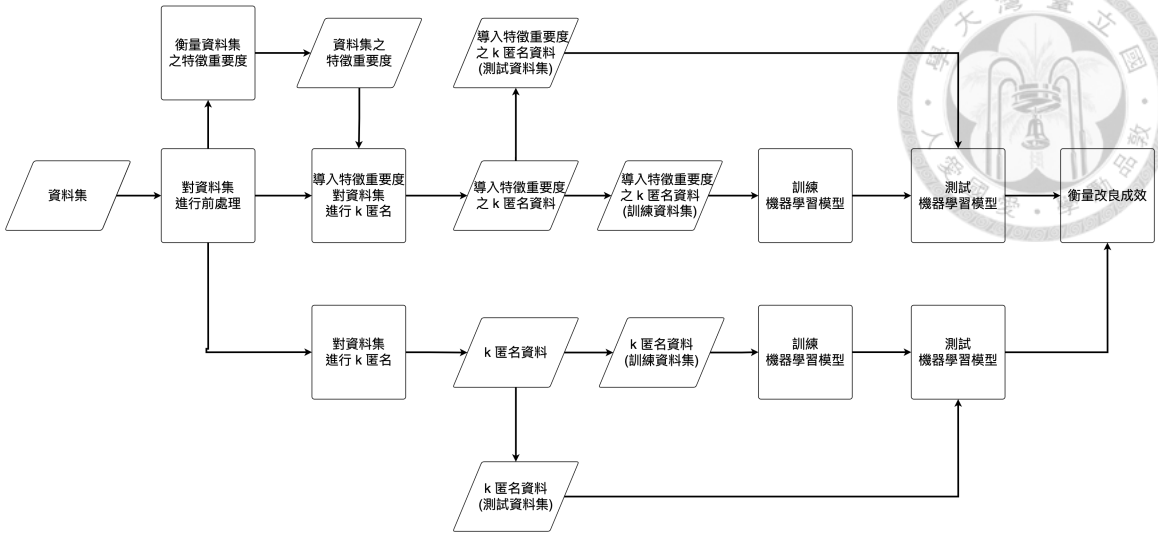

Figure 3.1: 考量特徵重要度之 k 匿名演算法改良成效衡量

針對特徵重要度衡量和導入特徵重要度至 CBA、TDGA、MA、OLA 所得之 CBAFI (CBA with Feature Importance)、TDGAFI (TDGA with Feature Importance)、MAFI (MA with Feature Importance)、OLAFI (OLA with Feature Importance) 實作將於本節依序作介紹。

3.1.1 特徵重要度衡量

本研究受採用反向消除法之包裝法概念啟發，發展出一套特徵重要度衡量方法，其基於泛化特徵與移除特徵之相似性，以將特定特徵移除資料集後對機器學習成效之影響作為該特徵重要度之根據；針對用於實驗之資料集，將會先使用完整資料集進行機器學習，得到一成效指標 P，而後輪流將欲衡量之特徵自資料集移除，並再次進行機器學習，所得到之成效指標 P' 與 P 之差即代表該特徵之重要度，衡量流程如圖 3.2 所示。

doi:10.6342/NTU202303509

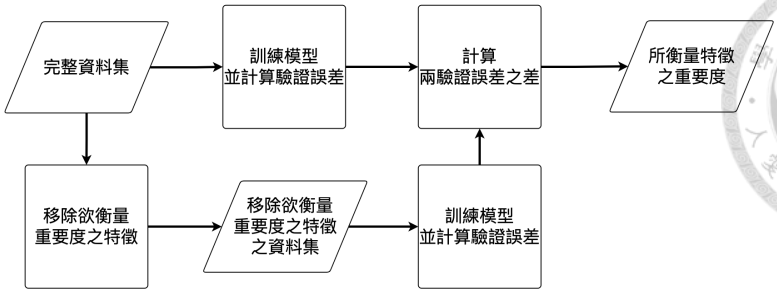

Figure 3.2: 特徵重要度衡量方法

3.1.2 CBAFI

CBA 以 NCP 作為距離公式，迭代地尋找最接近之 k 筆資料合併為等價類；其中 NCP 公式可視為計算所有特徵誤差之總和，將能影響各特徵於等價類之誤差大小，進而決策其泛化程度，因此本研究將該公式修改為特徵誤差依據重要度之加權總和，以導入特徵重要度，所得之 CBAFI 之運行流程如下、虛擬碼如演算法 1 所示。

1. 於資料集中隨機選擇 1 筆資料。
2. 以重要度加權各特徵誤差之總和計算距離，搜尋資料集中與該資料最近之 k - 1 筆資料。
3. 將 k 筆資料合併為一等價類並自資料集移除。
4. 若資料集仍有 k 筆以上之資料回到步驟 1 執行，若否則往後執行。
5. 將剩餘資料分配至距離 (計算方式同步驟 2) 最近之等價類。

doi:10.6342/NTU202303509

3.1.3 TDGAFI

TDGA 是以貪婪之策略最小化 NCP，其以 NCP 作為距離公式，選擇最遠之兩筆資料為群中心進行最小距離分群，遞迴地切割資料集，以產生等價類；其中 NCP 公式之功能等同其於 CBA 之作用，因此本研究採取相同之技巧修改 NCP 公式，以導入特徵重要度，所得之 TDGAFI 之運行流程如下、虛擬碼如演算法 2 所示。

1. 將整個資料集視為一資料群。
2. 切割資料群:
   (a) 以重要度加權各特徵誤差之總和計算距離，搜尋資料集中距離最遠之 2 筆資料。
   (b) 以該 2 筆資料作為中心對其他資料進行最小距離 (計算方式同步驟 2-(a)) 分群。
3. 針對所得之子資料群遞迴執行步驟 2，直到欲切割之資料群所含資料不足 k 筆 (假設大小為 s)。
4. 於所有大小至少為 2k - s 之等價類搜尋與該子資料群距離 (計算方式同步驟 2-(a)) 最近之 k - s 筆資料，比較與該些資料合併及與最近等價類合併將會產生之 NCP，並選擇較小者合併為一等價類。

3.1.4 MAFI

MA 是以遞迴切割資料集之形式產生等價類，其選擇特徵進行切割之步驟，亦是減少所選擇之特徵於等價類之誤差，進而特化該特徵，因此本研究將該演算

doi:10.6342/NTU202303509

法選擇特徵依據之標準化值範圍以重要度進行加權，從而導入特徵重要度，所得之 MAFI 之運行流程如下、虛擬碼如演算法 3 所示。

1. 計算資料集中各特徵之值範圍。
2. 將整個資料集視為一資料群。
3. 切割資料群:
   (a) 計算資料群各特徵之標準化值範圍並以重要度進行加權。
   (b) 選擇得到結果最大者，於該特徵之中位數進行切割。
4. 針對所得之子資料群遞迴執行步驟 3，直到欲切割之資料群所含資料不足 2k 筆。
5. 所得之資料群即為一等價類。

3.1.5 OLAFI

OLA 是於所有特徵之 VGH 階層組合篩選出篩選出符合 k 匿名限制之候選方案，再以所有特徵之 VGH 階層和作為指標，選擇最低者作為最佳方案；其中特徵之 VGH 階層即表示其泛化程度，因此本研究將方案選擇指標修改為特徵之 VGH 階層依據重要度之加權總和，以導入特徵重要度，所得之 OLAFI 之運行流程如下、虛擬碼如演算法 4 所示。

1. 將所有特徵之 VGH 階層進行組合產生節點，並依據 VGH 之父子節點關係建立路徑，以形成包含所有泛化可能性且具階層性之網格。
2. 以二分搜索策略篩選出符合 k 匿名限制之節點集合。

doi:10.6342/NTU202303509

3. 針對集合中之節點，以重要度加權各特徵 VGH 階層之總和計算分數。

4. 選擇分數最低者，採用該節點之 VGH 階層組合匿名化資料集。

Algorithm 1 CBAFI

```
1:  function NCP(a, b)
2:      ncp ← 0
3:      for feature of a do
4:          featureDeviation ← the feature deviation produced by a and b are anonymized
    together
5:          ncp ← ncp + (featureDeviation × featureImportanceWeight)
6:      end for
7:      return ncp
8:  end function
9:
10: function CBAFI(dataset, k)
11:     equivalenceClasses ← ∅
12:     while dataset.size() ≥ k do
13:         record ← random one in the dataset
14:         knn ← the k - 1 nearest neighbors in the dataset (use NCP(record, other-
    Record))
15:         equivalenceClass ← record, knn
16:         equivalenceClasses ← equivalenceClasses ∪ {equivalenceClass}
17:     end while
18:     for record in dataset do
19:         nearestEquivalenceClass ← the nearest equivalence class in equivalence-
    Classes (use NCP(record, equivalenceClass))
20:         nearestEquivalenceClass ← nearestEquivalenceClass ∪ record
21:     end for
22:     anonymizedDataset ← {anonymized records in equivalenceClasses}
23:     return anonymizedDataset
24: end function
```


doi:10.6342/NTU202303509

Algorithm 2 TDGAFI

```
1: function NCP(a, b)
2:     ncp ← 0
3:     for feature of a do
4:         featureDeviation ← the feature deviation produced by a and b are anonymized together
5:         ncp ← ncp + (featureDeviation × featureImportanceWeight)
6:     end for
7:     return ncp
8: end function
9:
10: function TDGAFI(dataset, k)
11:     equivalenceClasses ← ∅
12:     partitions ← the partition with all the records in dataset
13:     while partitions ≠ ∅ do
14:         partition ← partitions.pop()
15:         if partition cannot be splitted then
16:             equivalenceClasses ← equivalenceClasses ∪ {partition}
17:             continue
18:         end if
19:         center1, center2 ← the two records produced max NCP when being anonymized together in partition (use NCP(record1, record2))
20:         partition1 ← {center1}
21:         partition2 ← {center2}
22:         distribute the other records in the partition to partition1 or partition2 which produces the smaller NCP (use NCP(record, center1 or center2))
23:         if partition1.size() or partition2.size() < k then
24:             validPartition, invalidPartition ← partition1, partition2 (or partition2, partition1)
25:             ncp1 ← NCP produced by merged validPartition and invalidPartition
26:             ncp2 ← NCP produced by merged invalidPartition and the closest required records in the same equivalenceClass
27:             if ncp1 > ncp2 then
28:                 partitions ← partitions ∪ {validPartition}
29:                 augmentedPartition ← merge invalidPartition and the closest required records in the same equivalenceClass
30:                 partitions ← partitions ∪ {augmentedPartition}
31:             else
32:                 partitions ← partitions ∪ {validPartition ∪ invalidPartition}
33:             end if
34:         else
35:             partitions ← partitions ∪ {partition1, partition2}
36:         end if
37:     end while
38:     anonymizedDataset ← {anonymized records in equivalenceClasses}
39:     return anonymizedDataset
40: end function
```


doi:10.6342/NTU202303509

Algorithm 3 MAFI

```
1: function chooseFeature(partition)
2:     chosenfeature ← -1
3:     maxNormalizedWidth ← MIN_INT
4:     for feature of partition do
5:         featureNormalizedWidth ← the normalized width of feature
6:         weightedFeatureNormalizedWidth ← featureNormalizedWidth × featureImportanceWeight
7:         if weightedFeatureNormalizedWidth > maxNormalizedWidth then
8:             chosenFeature ← feature
9:             maxNormalizedWidth ← weightedFeatureNormalizedWidth
10:        end if
11:    end for
12:    return chosenfeature
13: end function
14:
15: function MAFI(dataset, k)
16:    equivalenceClasses ← ∅
17:    partitions ← {the partition with all records in dataset}
18:    while partitions ≠ ∅ do
19:        partition ← partitions.pop()
20:        if partition cannot be splitted then
21:            equivalenceClasses ← equivalenceClasses ∪ {partition}
22:            continue
23:        end if
24:        feature ← chooseFeature(partition)
25:        partition1, partition2 ← split partition on the median of feature
26:        partitions ← partitions ∪ {partition1, partition2}
27:    end while
28:    anonymizedDataset ← {anonymized records in equivalenceClasses}
29:    return anonymizedDataset
30: end function
```


doi:10.6342/NTU202303509

Algorithm 4 OLAFI

```
function GETWEIGHTEDGENERALIZATIONLEVEL(node)
    level ← 0
    for featureGeneralizationLevel of node do
        level ← level + (featureGeneralizationLevel × featureImportanceWeight)
    end for
    return level
end function

function OLAFI(dataset, k)
    lattice ← all possible generalization steps
    validNodes ← all nodes in the lattice meet the k-anonymity constraint
    optimizedNode ← node in validNodes with min step (use getWeightedGeneral-
izationLevel(node))
    anonymizedDataset ← anonymized dataset produced by optimizedNode
    return anonymizedDataset
end function
```

3.2 結合間距保留法之 k 匿名流程

針對機器學習之分類任務，本研究創新地提出間距保留法以降低資料匿名化對機器學習分類模型所造成之類別混淆；間距保留法旨在基於目標特徵對資料進行最大化間距分群，本研究藉由 SVM 之特性，先採用該模型對資料集進行分類預測，再基於所得之預測類別將資料分群，以形成具有最大化類別間距之資料群集；由於每一資料群各代表一類別，並且資料群集之間距得益於 SVM 之特性極大地被保留，因此分別將各資料群匿名化後再合併所產生之匿名資料，將能降低機器學習分類模型混淆不同類別之可能性，匿名流程如圖 3.3 所示。

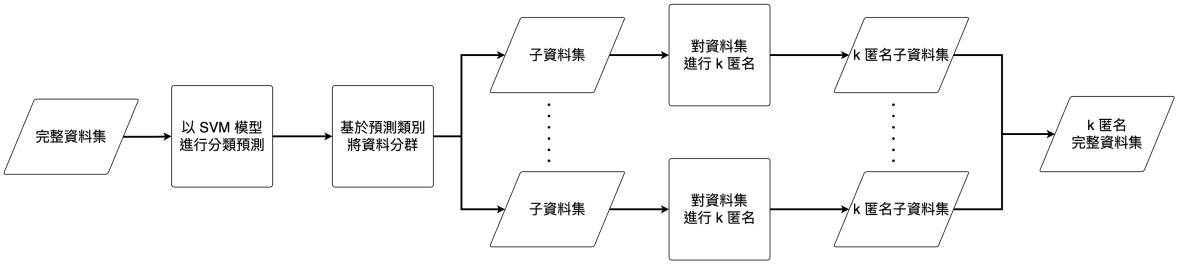

Figure 3.3: 結合間距保留法之 k 匿名流程

doi:10.6342/NTU202303509

同樣地，為衡量改良幅度，本研究以 k 匿名演算法作為實驗變因，以資料集、k 匿名之參數 k、機器學習模型作為控制變因，比較多種控制變因組合下所得之機器學習成效，實驗流程如圖 3.4 所示。

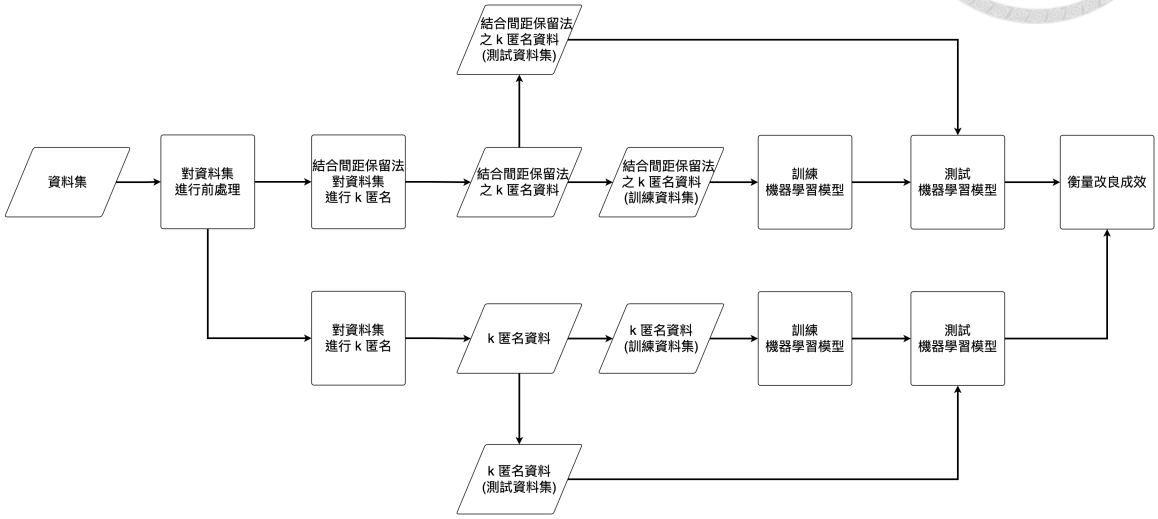

Figure 3.4: 結合間距保留法之 k 匿名流程改良成效衡量

針對 CBA、TDGA、MA、OLA 與間距保留法結合所得之匿名流程後續將以 CBAMP (CBA with Margin Preserving)、TDGAMP (TDGA with Margin Preserving)、MAMP (MA with Margin Preserving)、OLAMP (OLA with Margin Preserving) 代稱。

doi:10.6342/NTU202303509

第四章　實驗結果與討論

本研究透過考量特徵重要度或結合間距保留法改良 k 匿名流程以提高匿名資料之機器學習成效，於實驗中共採用 4 種資料集、4 種 k 匿名演算法及 3 種機器學習演算法，並將 k 匿名之參數 k 設定為 2 至 100，從而於兩種改良實驗之各 4752 組實驗數據，觀察機器學習成效指標 f1score 之變化，本章節將依序介紹本研究所採用之資料集與自兩種改良實驗所得之數據總結之洞察。

4.1　資料集

- Adult 資料集 (ADULT) [20]: 美國人口普查資料集，包含性別、年齡、種族、婚姻狀態、教育程度、國籍、職業別、居住地及薪資級別等特徵，本研究以薪資級別作為敏感資訊、其他特徵作為準識別碼並進行薪資級別預測。

- California Housing Prices 資料集 (CAHOUSING) [21]: 美國加州之房價資料集，包含社區屋齡中位數、社區房價中位數、社區屋主收入中位數、經緯度及距海遠近等特徵，本研究以距海遠近作為敏感資訊、其他特徵作為準識別碼並進行距海遠近預測。

- Contraceptive Method Choice 資料集 (CMC) [22]: 印度避孕措施普查資料集，包含伴侶年齡、伴侶教育程度、孩童數量及避孕措施使用頻率等特徵，本研究以避孕措施使用頻率作為敏感資訊、其他特徵作為準識別碼並進行避孕措施使用頻率預測。

- Mammographic Mass 資料集 (MGM) [23]: 乳腺腫瘤影像資料集，包含患者年齡、系統序數、腫瘤形狀、腫瘤邊界、腫瘤密度及嚴重程度等特徵，本研究

doi:10.6342/NTU202303509

以嚴重程度作為敏感資訊、其他特徵作為準識別碼並進行嚴重程度預測。

4.2 導入特徵重要度改良 k 匿名演算法以提高匿名資料之機器學習成效

為評估改良幅度，本研究觀察多種控制變因組合下有無考量特徵重要度進行匿名化之資料所訓練出之機器學習模型其成效差異，發現當採用特定 k 匿名演算法或資料集改良表現較佳，本小節將複用式 4.1 呈現不同控制變因組合之改良幅度並探討所歸納出之洞察。

$$D = \{ADULT, CAHOUSING, CMC, MGM\}$$

$$KANON = \{CBA, TDGA, MA, OLA\}$$

$$MODEL = \{10 - NN, RF, XGB\}$$

$$\text{機器學習成效} = Perf(d, kanon, k, model)$$

$\cdot\, d$ 為資料集 $\cdot\, kanon$ 為 k 匿名演算法 $\cdot\, k$ 為 k 匿名之參數 $\cdot\, model$ 為機器學習模型

$$\begin{aligned} \text{改良率} &= Improv(d, kanon\ with\ FI, k, model) \\ &= \frac{Perf(d,kanon\ with\ FI,k,model) - Perf(d,kanon,k,model)}{Perf(d,kanon,k,model)} \\ \text{成功改良} &= SI(d, kanon\ with\ FI, k, model) \\ &= \begin{cases} 1, & if\ Improv(d, kanon\ with\ FI, k, model) > 0 \\ 0, & otherwise \end{cases} \end{aligned} \tag{4.1}$$

doi:10.6342/NTU202303509

4.2.1 適於導入特徵重要度之 k 匿名演算法特性

本研究根據式 4.2 計算採用 4 種不同 k 匿名演算法進行實驗之改良表現，數據如表 4.1 所示，其中 CBAFI 與 TDGAFI 演算法相較其他兩者具有較高之成功改良比例，改良表現穩定，其原因可能為各 k 匿名演算法調節特徵泛化或特化程度之實作差異。

CBAFI 與 TDGAFI 演算法是為將特徵重要度導入至 CBA 與 TDGA 演算法之距離公式所產生，該二演算法皆基於大量之資料間距決策等價類之組成資料，其精細至單筆資料顆粒度之作法能使得特徵重要度之作用能極大程度地展現，進而得到較好之改良表現。

MAFI 演算法是為將特徵重要度導入至 MA 演算法之特化特徵選擇步驟所產生，該演算法藉由持續基於所選擇之特徵進行切割以產生等價類，儘管導入特徵重要度能提高重要度較高之特徵被選擇之可能，進而降低該特徵之誤差，然而每一次切割卻僅能降低單一特徵之誤差，使得誤差難以確切地按照特徵重要度進行分配。

OLAFI 演算法是為將特徵重要度導入至 OLA 演算法選擇最佳方案之決策方式所產生，該演算法視 VGH 階層為誤差，因此選擇候選方案中各特徵之 VGH 階層以特徵重要度進行加權總和之最小者作為最終方案，然而不同 VGH 之相同階層所產生之誤差並未統一，導致誤差分配將同時依賴特徵重要度與 VGH 之設計，從而難以產生穩定之改良。

doi:10.6342/NTU202303509

Table 4.1: 採用 4 種不同 k 匿名演算法進行實驗之改良表現

	平均改良率	成功改良比例
CBAFI	6.00%	88.05%
TDGAFI	5.72%	87.63%
MAFI	3.04%	60.86%
OLAFI	10.70%	61.36%

$$\begin{aligned}&\text{k 匿名演算法 } kanon\ with\ FI \text{ 之平均改良率}\\&= \frac{\sum_{d\in D}\sum_{k=2}^{100}\sum_{model\in MODEL} Improv(d, kanon\ with\ FI, k, model)}{|D|\times(100-1)\times|MODEL|}\\&\text{k 匿名演算法 } kanon\ with\ FI \text{ 之成功改良比例}\\&= \frac{\sum_{d\in D}\sum_{k=2}^{100}\sum_{model\in MODEL} SI(d, kanon\ with\ FI, k, model)}{|D|\times(100-1)\times|MODEL|}\end{aligned} \tag{4.2}$$

4.2.2 適於以考量特徵重要度匿名化之資料集特性

本研究根據式 4.3 計算採用 4 種不同資料集進行實驗之改良表現，數據如表 4.2 所示，其中採用 ADULT、CAHOUSING 資料集之實驗具有較佳之成功改良比例，該現象可能源自於不同資料集之特徵間相關性差異，由於本研究提出之特徵重要度衡量方法是針對單一特徵進行，特徵間之高度相關性將難以呈現於衡量結果，進而達到更佳之改良表現，根據圖 4.1 呈現之 4 種資料集之特徵相關性，較於其他兩者，可見 CMC、MGM 資料集具有較高之特徵相關性。

doi:10.6342/NTU202303509

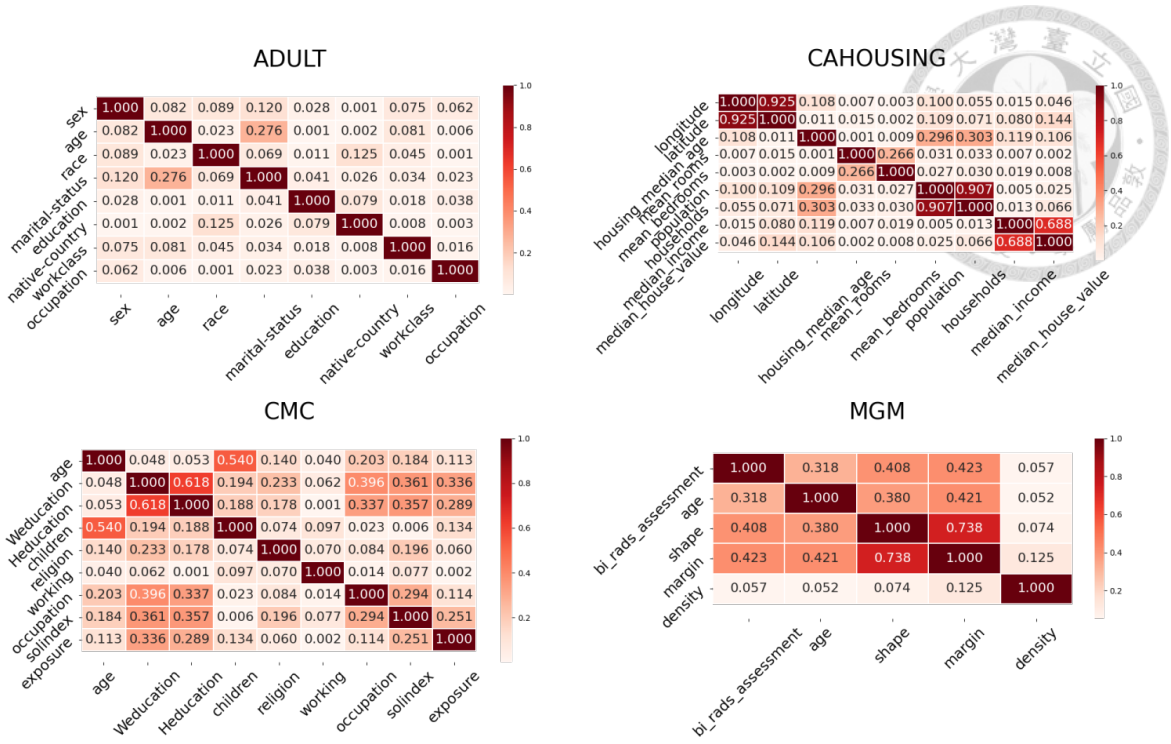

Figure 4.1: 4 種資料集之特徵相關性

Table 4.2: 採用 4 種不同資料集進行實驗之改良表現

	平均改良率	成功改良比例
ADULT	4.88%	85.77%
CAHOUSING	9.30%	88.72%
CMC	3.37%	71.89%
MGM	7.91%	51.52%

於資料集 d 之平均改良率

$$= \frac{\sum_{kanon \in KANON} \sum_{k=2}^{100} \sum_{model \in MODEL} Improv(d, kanon\ with\ FI, k, model)}{|KANON| \times (100-1) \times |MODEL|} \tag{4.3}$$

於資料集 d 之成功改良比例

$$= \frac{\sum_{kanon \in KANON} \sum_{k=2}^{100} \sum_{model \in MODEL} SI(d, kanon\ with\ FI, k, model)}{|KANON| \times (100-1) \times |MODEL|}$$

doi:10.6342/NTU202303509

4.2.3 適於導入特徵重要度之匿名資料分類預測解決方案

本研究根據式 4.4 計算採用不同 k 匿名演算法與機器學習模型組合進行實驗之改良表現，數據如圖 4.2 所示，於本研究之 12 種變因組合中，CBAFI 演算法結合 XGB 模型之解決方案具有最佳之改良表現，其原因可能為 CBAFI 能使特徵重要度精細地作用於每一資料之每一特徵上，並且 XGB 較不擅於學習之低相關性資料，恰為考量特徵重要度之 k 匿名演算法改良表現較佳之類型；此外，該解決方案亦為所有變因組合中機器學習成效最佳者，適於作為匿名資料分類預測之通用性解決方案。

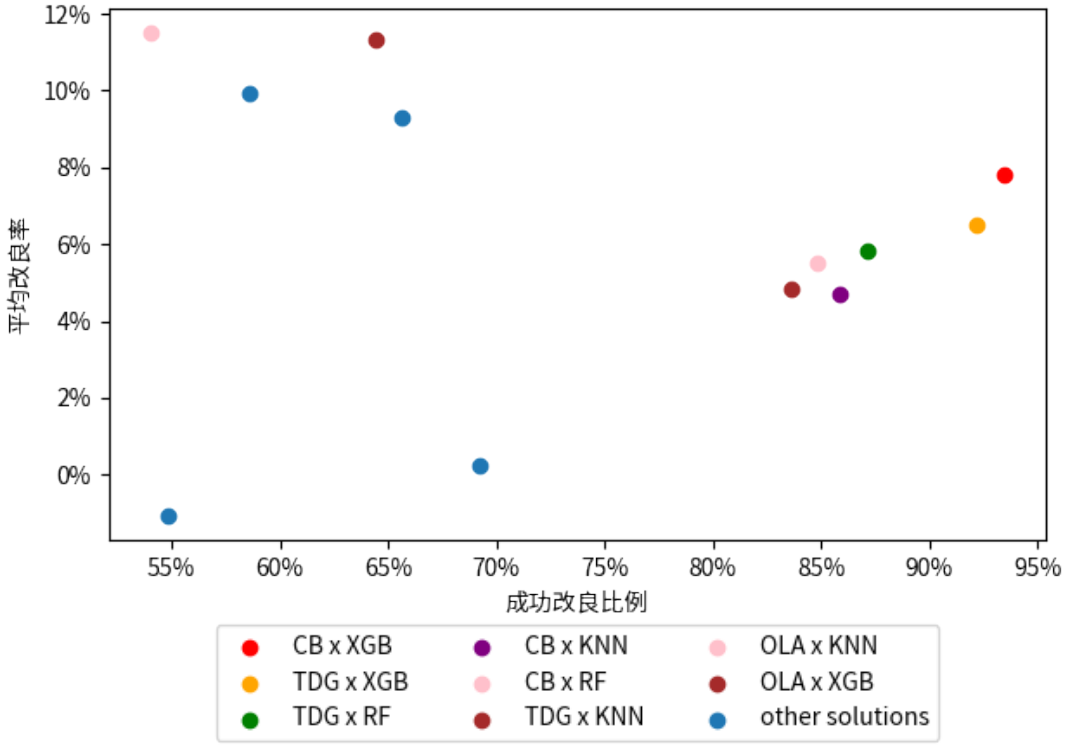

Figure 4.2: 12 種匿名資料分類預測解決方案之改良表現

doi:10.6342/NTU202303509

k 匿名演算法 $kanon\ with\ FI$ 結合機器學習模型 $model$ 解決方案之平均改良率

$$= \frac{\sum_{d \in D} \sum_{k=2}^{100} Improv(d, kanon\ with\ FI, k, model)}{|D| \times (100-1)}$$

k 匿名演算法 $kanon\ with\ FI$ 結合機器學習模型 $model$ 解決方案之成功改良比例

$$= \frac{\sum_{d \in D} \sum_{k=2}^{100} SI(d, kanon\ with\ FI, k, model)}{|D| \times (100-1)} \tag{4.4}$$

4.3 結合間距保留法與 k 匿名演算法以提高匿名資料之機器學習成效

為評估改良幅度，本研究觀察多種控制變因組合下有無結合間距保留法進行匿名化之資料所訓練出之機器學習模型其成效差異，發現當採用特定 k 匿名演算法、機器學習模型或參數 k 設定改良表現較佳，本小節將複用式 4.5 呈現不同控制變因組合之改良幅度並探討所歸納出之洞察。

doi:10.6342/NTU202303509

$$D = \{ADULT, CAHOUSING, CMC, MGM\}$$

$$KANON = \{CBA, TDGA, MA, OLA\}$$

$$MODEL = \{10 - NN, RF, XGB\}$$

$$\text{機器學習成效} = Perf(d, kanon, k, model)$$

· d 為資料集 · $kanon$ 為 k 匿名流程 · k 為 k 匿名之參數 · $model$ 為機器學習模型

$$\begin{aligned} \text{改良率} &= Improv(d, kanon\ with\ MP, k, model) \\ &= \frac{Perf(d,kanon\ with\ MP,k,model) - Perf(d,kanon,k,model)}{Perf(d,kanon,k,model)} \end{aligned}$$

$$\begin{aligned} \text{成功改良} &= SI(d, kanon\ with\ MP, k, model) \\ &= \begin{cases} 1, & if\ Improv(d, kanon\ with\ MP, k, model) > 0 \\ 0, & otherwise \end{cases} \end{aligned} \tag{4.5}$$

4.3.1 適於結合間距保留法之 k 匿名演算法特性

本研究根據式 4.6 計算採用 4 種不同 k 匿名演算法進行實驗之改良表現，數據如表 4.3 所示，其中採用 OLA 結合間距保留法進行之實驗相較於其他三者具有最佳之改良表現，原因可能為 OLA 構建泛化方案之形式使得所有資料之同一特徵皆須泛化至同一 VGH 階層，若針對一特徵存在特定 VGH 節點對應資料不足 k 筆之情境，則僅能藉由提升該特徵於所有資料之 VGH 階層以滿足 k 匿名限制；然而結合間距保留法進行匿名化時，不僅能降低前述情境所影響之資料筆數，亦能降低一等價類存在多種類別資料之可能性。

doi:10.6342/NTU202303509

Table 4.3: 採用 4 種不同 k 匿名演算法進行實驗之改良表現

	平均改良率	成功改良比例
CBAMP	7.24%	93.18%
TDGAMP	6.18%	86.53%
MAMP	11.92%	85.61%
OLAMP	17.40%	86.62%

$$\begin{aligned}&\text{k 匿名流程 } kanon\ with\ MP \text{ 之平均改良率}\\&= \frac{\sum_{d\in D}\sum_{k=2}^{100}\sum_{model\in MODEL} Improv(d, kanon\ with\ MP, k, model)}{|D|\times(100-1))\times|MODEL|}\\&\text{k 匿名流程 } kanon\ with\ MP \text{ 之成功改良比例}\\&= \frac{\sum_{d\in D}\sum_{k=2}^{100}\sum_{model\in MODEL} SI(d, kanon\ with\ MP, k, model)}{|D|\times(100-1)\times|MODEL|}\end{aligned} \tag{4.6}$$

4.3.2 適於結合間距保留法匿名化資料之機器學習模型特性

本研究根據式 4.7 計算採用 3 種不同機器學習模型進行實驗之改良表現，數據如表 4.4 所示，其中採用 10-NN 模型進行之實驗相較於其他兩者具有最佳之改良表現，原因可能為 10-NN 模型基於鄰近資料之類別統計產生預測結果之特性；當未結合間距保留法匿名化資料時，多種類別資料歸屬於同一等價類之可能性高，此些資料將會泛化為相同結果，進而使得相對距離相同之資料具有不同類別，造成 10-NN 模型混淆；然而當結合間距保留法匿名化資料時，匿名化作用之資料群集是為經過間距保留法分群所得，其所含之資料類別較純粹，從而避免未結合間距保留法匿名化資料易產生之情境。

doi:10.6342/NTU202303509

Table 4.4: 採用 4 種不同機器學習模型進行實驗之改良表現

	平均改良率	成功改良比例
10-NN	15.23%	83.84%
RF	8.59%	90.09%
XGB	8.24%	90.03%

$$\text{於機器學習模型 } model \text{ 之平均改良率}$$
$$= \frac{\sum_{d \in D} \sum_{kanon \in KANON} \sum_{k=2}^{100} Improv(d, kanon\ with\ MP, k, model)}{|D| \times |KANON| \times (100-1)} \tag{4.7}$$
$$\text{於機器學習模型 } model \text{ 之成功改良比例}$$
$$= \frac{\sum_{d \in D} \sum_{kanon \in KANON} \sum_{k=2}^{100} SI(d, kanon\ with\ MP, k, model)}{|D| \times |KANON| \times (100-1)}$$

4.3.3 適於結合間距保留法匿名化資料之參數 k 設定

本研究根據式 4.8 計算採用 99 種不同 k 匿名參數設定進行實驗之改良表現，數據如圖 4.3 所示，參數 k 與平均改良率呈正比關係，其原因可能為參數 k 之成長將會增加等價類之大小；當未結合間距保留法匿名化資料時，等價類之大小愈大，則所包含資料之類別將愈混雜，易造成模型混淆進而降低機器學習成效；然而當結合間距保留法匿名化資料時，匿名化作用之資料群集是為經過間距保留法分群所得，其所含之資料類別較純粹，從而能達到抑制等價類所含之資料類別愈趨混雜之狀況。

doi:10.6342/NTU202303509

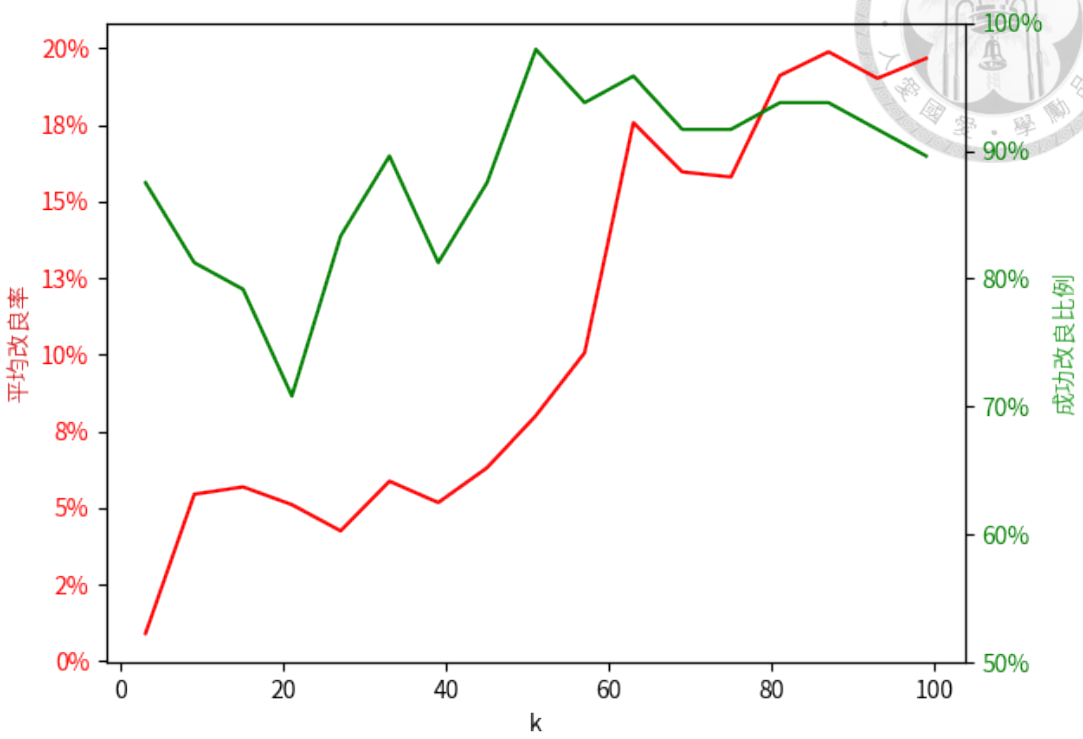

Figure 4.3: 採用 99 種不同 k 匿名參數設定進行實驗之改良指標

於 k 匿名參數 k 之平均改良率

$$= \frac{\sum_{d \in D} \sum_{kanon \in KANON} \sum_{model \in MODEL} Improv(d, kanon\ with\ MP, k, model)}{|D| \times |KANON| \times |MODEL|} \tag{4.8}$$

於 k 匿名參數 k 之成功改良比例

$$= \frac{\sum_{d \in D} \sum_{kanon \in KANON} \sum_{model \in MODEL} SI(d, kanon\ with\ MP, k, model)}{|D| \times |KANON| \times |MODEL|}$$

doi:10.6342/NTU202303509

第五章　結果與未來展望

5.1　結論

本研究考量應用情境之特徵重要度與機器學習目標提出之 k 匿名流程改良方式—導入特徵重要度與結合間距保留法，經實驗數據所證，實能提高匿名資料之機器學習成效，相較於採用原 k 匿名演算法匿名化資料，再經本研究之機器學習模型進行分類預測，採用考量特徵重要度之 k 匿名演算法平均改良機器學習成效之幅度可達 10.7%，而結合間距保留法匿名化資料平均改良機器學習成效之幅度可達 17.40%。

其中於本研究進行之導入特徵重要度改良 k 匿名演算法以提高匿名資料之機器學習成效實驗中，共歸納出以下 3 項洞察。

- 當 k 匿名演算法決策特徵泛化程度之步驟愈精細，則特徵重要度所能產生之影響愈大，因此提高匿名資料之機器學習成效之表現得以愈穩定。
- 針對特徵相關性低之資料集，採用本研究提出之特徵重要度衡量方法時，所得之特徵重要度資訊較完整，因此將其導入至 k 匿名演算法以提高匿名資料之機器學習成效之成功比例較高。
- 於本研究所探討之匿名資料分類預測解決方案中，CBAFI 搭配 XGB 是為改良表現最佳之解決方案，並且亦為所有組合中機器學習成效最佳者，適於作為匿名資料分類預測之通用性解決方案。

而於本研究進行之結合間距保留法與 k 匿名演算法以提高匿名資料之機器學習成效中，共歸納出以下 3 項洞察。

doi:10.6342/NTU202303509

- 當 k 匿名演算法之設計統一相同特徵於所有資料之 VGH 階層時，結合間距保留法匿名化資料將能縮小該設計作用之範圍，因此能有效降低受分佈資料較少之值範圍或值集合影響而泛化之資料筆數，從而較大幅度地提高匿名資料之機器學習成效。

- 當機器學習模型依據訓練資料直接產生預測結果時，結合間距保留法匿名化資料將能使得等價類所含之資料類別更加純粹，從而避免不同類別之資料具有相同匿名化結果以造成模型混淆，以較穩定地提高匿名資料之機器學習成效。

- 參數 k 與等價類之大小成正比，結合間距保留法匿名化資料時，將能使 k 匿名演算法分別於所含資料類別較純粹之資料群集構建等價類，進而降低 k 值增大所損失之機器學習成效。

此外，承研究背景所述，個人資料授權明確且嚴密之使用形式使得資料處理者應得以依據授權資訊客製化地調整去識別化方法，而本研究提出之導入特徵重要度與結合間距保留法即為二有效之策略；以先前研究提出之 BEDPP 架構為例，導入本研究成果至其中，不僅能提供明確之去識別化指標，亦能添增根據應用情境之特徵重要度與機器學習目標提高機器學習成效之可能性，擴充 BEDPP 所得之資料交易流程如圖 5.1 所示。

doi:10.6342/NTU202303509

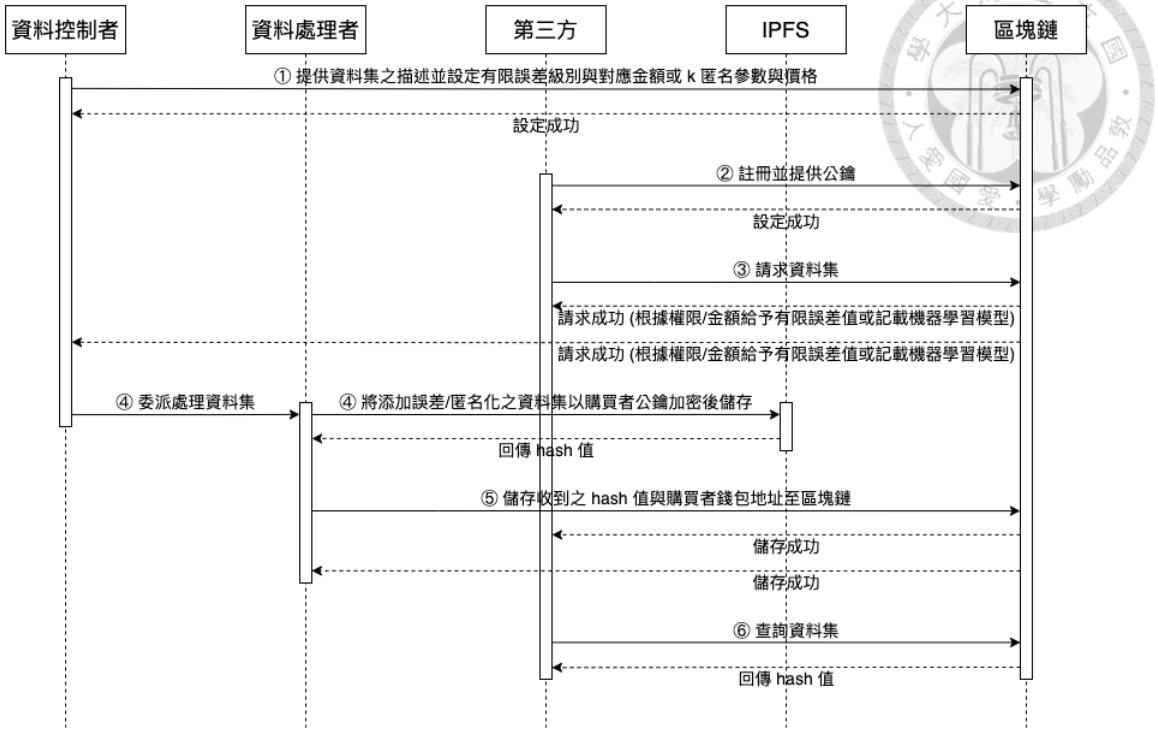

Figure 5.1: 導入本研究成果擴充之 BEDPP 中 BEDMoB 之資料交易流程

5.2 未來展望

本研究創新地提出兩種提高匿名資料之機器學習成效之方法，並採用 4 種資料集、4 種 k 匿名演算法、3 種機器學習模型進行實驗，儘管所得之實驗數據證明該二方法能有效提升匿名資料之機器學習成效，然而生產環境應用之資料集、k 匿名演算法、機器學習模型之異質性高，未來若能採用更多種類之資料集、k 匿名演算法、機器學習模型進行實驗，則將有機會歸納出更多洞察。

於導入特徵重要度改良 k 匿名演算法以提高匿名資料之機器學習成效實驗中，本研究提出一特徵重要度衡量方法，該方法著重於衡量單一特徵之重要度，因此較適用於特徵相關性低之資料集，針對特徵相關性高之資料集設計特徵重要度衡量方法是未來能嘗試之研究。

doi:10.6342/NTU202303509

本研究所提出之兩種提高匿名資料之機器學習成效之方法源自考量應用情境之特徵重要度與機器學習目標異質性而發展，然而應用情境包含多樣因素與流程，不僅存在許多因素本研究尚未涉及，並且當前之實驗流程設計較似於機器學習模型部署於雲端之運作流程，針對機器學習模型部署於終端之情境亦未納入其中，因此考量不同之因素與流程組合研究相對應之匿名流程改良方法是為未來能持續發展之研究方向。

doi:10.6342/NTU202303509

參考文獻

[1] 中華民國行政院. 個人資料保護法, 2023. Available: https://law.moj.gov.tw/LawClass/LawAll.aspx?PCode=I0050021.

[2] European Commission. General data protection regulation, 2018. Available: https://gdpr-info.eu/.

[3] 中華民國衛生福利部. 全民健康保險研究資料庫, 2000. Available: https://nhird.nhri.edu.tw/.

[4] European Commission. Smart grids and meters. European Commission Energy. Available: https://energy.ec.europa.eu/topics/markets-and-consumers/smart-grids-and-meters_en.

[5] 林瑞珠, 朱丹丹. 推動智慧電表布建所需之資訊安全與隱私保護規範. 臺灣能源期刊, 5(4):315–330, 2018.

[6] 曾郁凱. 有限誤差與區塊鏈在物聯網資料安全保護之應用. Master's thesis, 國立臺灣大學, 2021.

[7] V. Buterin. Ethereum, 2014. Available: https://ethereum.org/en/.

[8] N. Szabo. Smart contract, 1994. Available: https://en.wikipedia.org/wiki/Smart_contract.

[9] J. Benet. Interplanetary file system, 2014. Available: https://en.wikipedia.org/wiki/InterPlanetary_File_System.

[10] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.

doi:10.6342/NTU202303509

[11] A. Jović, K. Brkić, N. Bogunović. A review of feature selection methods with applications. International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), (38):1200–1205, 2015.

[12] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. W.-C. Fu. Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06), page 785, New York, NY, USA, 2006.

[13] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571–588, 2002.

[14] J.-L. Lin, M.-C. Wei. An efficient clustering method for k-anonymization. In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society (PAIS '08), pages 46–50, New York, NY, USA, 2008.

[15] K. LeFevre, D. J. DeWitt, R. Ramakrishnan. Mondrian multidimensional k-anonymity. In 22nd International Conference on Data Engineering (ICDE'06), pages 25–25, Atlanta, GA, USA, 2006.

[16] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley. A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association, 16:670–682, 2009.

[17] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001.

[18] T. Chen, C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings

doi:10.6342/NTU202303509

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pages 785–794, San Francisco, CA, USA, 2016.

[19] D. Slijepčević, M. Henzl, L. D. Klausner, T. Dam, P. Kieseberg, M. Zeppelzauer. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. Computers Security, 111:102488, 2021.

[20] B. Becker, R. Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[21] C. Nugent. California housing prices. Kaggle, 2017. Available: https://www.kaggle.com/datasets/camnugent/california-housing-prices.

[22] T.-S. Lim. Contraceptive method choice. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C59W2D.

[23] M. Elter. Mammographic mass. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C53K6Z.

doi:10.6342/NTU202303509