Techniques for training classifiers with class-imbalanced data

1. Group Member
   - 109550040 高鈺鴻
   - 109550008 王禹博

2. Video Link

   https://youtu.be/q1WQGBU3nvo

3. Overview

   機器學習在處理分類問題時，許多資料集會有類別不平衡的問題。然而，許多演算法較適合或被設計應用在類別分布較均勻的資料，當面對不平衡資料時，可能會嚴重影響演算法的表現。另一方面，由於演算法的表現常常被用準確度來評估其好壞，因此，在面對不平衡的資料時，只要分類器傾向於把全部的輸入判別為多數類(majority class)，就能夠達到相當高甚至接近100%的整體準確度，然而，這種高準確度並不實際，因為在此類問題上，少數類的準確度通常更為重要的。

   以醫學圖像判別的應用舉例，在藉由分類一張醫學圖像來檢測病症的二分類問題時，如果分類器將大部分的圖像判別為陰性(多數類)，雖然可以取得相當高的整體準確度，但同時也把多數陽性(少數類)的案例判別為陰性，如此造成的後果可能不符合我們的期望。

   因此，如何解決不平衡學習 (Imbalanced Learning) 就顯得相當重要，其中不同的方法包括採樣(Sampling)、代價敏感(cost-sensitive)學習等，在這個 Final Project中，我們專注於處理不平衡學習問題上的採樣方法，通過採樣可以修改訓練資料集使各類別的資料數量趨於平衡，而採樣方法主要可分為兩大類別：過採樣(Oversampling)及欠採樣(Undersampling)。

   在這個 Project 中，我們選擇了多個與採樣方法有關的論文，將這些方法利用 scikit-learn 實作並選擇了一個極度不平衡(579:1)的資料集進行實驗，通過實驗結果來觀察機器學習演算法在使用不平衡資料和採樣過後的平衡資料訓練後的效果表現差異與提升程度，以及觀察各個不同的採樣方法的特徵及不同處。

4. Selected Papers

在這個 Final Project 中，我們參考多個論文提出的採樣方法，其中包含過採樣方法和欠採樣方法，分別為 NearMiss[1]、SMOTE[2]、Borderline-SMOTE[3]、ADASYN[4]。

- NearMiss:

  欠採樣方法中，一種簡單而常見的方法為隨積欠採樣(Random UnderSamping)，藉由隨機移除不平衡資料中的多數類來達到資料的平衡。然而，隨機移除多數類的資料可能會導致多數類中訊息的損失而對效果造成影響。因此，論文中提出的NearMiss是一種以K-近鄰演算法(KNN，K Nearest Neighbors)為基礎的欠採樣方法，此方法的概念是通過啟發式(heuristic)的規則從多數類別的樣本中選擇具有代表性的樣本進行訓練，可以減緩隨機欠取樣中的重要資訊容易丟失的問題，(該方法)在論文中提出三種版本，分別為NearMiss-1、NearMiss-2和NearMiss-3。

  - NearMiss-1: 分別計算每個多數類別樣本與每個少數類樣本之間的距離，之後選擇最近的 k 個少數類別樣本並計算這 k 個少數類別樣本與該多數類別樣本之間的平均距離，最後進行比較，保留平均距離最小的多數類別樣本。
  - NearMiss-2: 分別計算每個多數類別樣本與每個少數類樣本之間的距離，之後選擇最遠的 k 個少數類別樣本並計算這 k 個少數類別樣本與該多數類別樣本之間的平均距離，最後進行比較，保留平均距離最小的多數類別樣本。
  - NearMiss-3: 分別計算每個少數類別樣本與每個多數類別樣本之間的距離，保留距離最近的k個多數類別樣本，以保證每個少數類別樣本都被多數類別樣本包圍。

  通過使用這些NearMiss方法，我們可以更有針對性地選擇欠採樣後的樣本，以保留重要資訊並改善模型的預測能力。

- SMOTE(Synthetic Minority Over-sampling Technique):

  在過採樣方法中，最簡單常見的方法是隨機過採樣(Random Oversamping)，會從少數類中選擇樣本進行重複取樣。然而，隨機過採樣的缺點在於只通過重複抽樣少數類樣本來平衡類別數量，並不會增加樣本的多樣性，可能會使模型對於少數類有過擬合(overfitting)的問題。而SMOTE是基於隨機過採樣所改進的方法，會藉由人工生成新的少數類資料來避免此問題。

SMOTE的具體流程是先對於每個少數類的樣本找到k個最近的且同樣屬於少數類的鄰居，當要產生新樣本時，在k個鄰居中隨機選擇一個樣本，然後計算該少數類樣本與被選中樣本在特徵空間中的距離(Ex: Euclidian distance)，最後通過先將計算出的距離先乘以一個介於0和1之間的隨機數，再和該少數類樣本相加以生成新的合成樣本。

- Borderline-SMOTE:

  Borderline-SMOTE是基於SMOTE的改進版本，由於SMOTE演算法在生成合成樣本時會同等地考慮少數類中的每一個樣本，沒有額外考慮少數類中那些較靠近多數類樣本邊界的樣本(邊界樣本)。因此，Borderline-SMOTE中改良的點在於會先找出少數類中那些較靠近類別邊界的樣本，並使用這些樣本來進行新樣本的合成，可以更好地解決類別間的重疊問題，能夠使模型更好地分辨類別間的邊界。

- ADASYN(Adaptive Synthetic Sampling):

  ADASYN同樣是SMOTE的改進版本，且和Borderline-SMOTE同樣是希望對於邊界樣本的考慮比重能夠變多。不同的點在於，Borderline-SMOTE會找出一個少數類中較靠近邊界的樣本集合，但只考慮該集合內的樣本做為合成新樣本的基準點，而ADASYN的方式則是會針對少數類中的每一個樣本算出一個值，如果該樣本較靠近多數類的話該值也會越大，並根據這些值形成一個機率分布，接著依照此機率分布來合成新樣本，藉此讓值越大越靠近多數類的那些少數類樣本在合成新樣本階段時更容易被選為基準點，也能達到多考慮了邊界樣本的效果。

5. Expirements

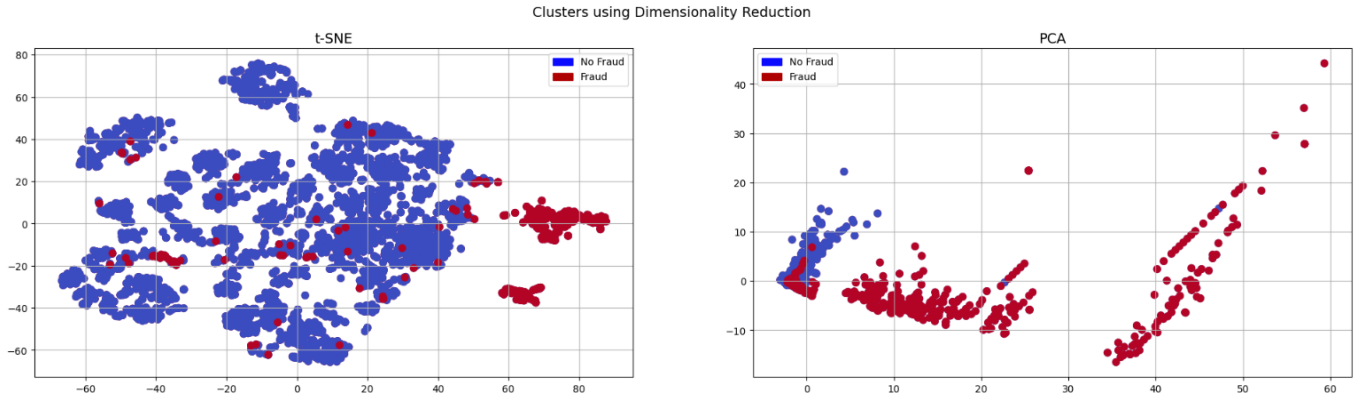
在實驗中，我們使用了來自 Kaggle 上的公開資料集 Credit Card Fraud Detection(*https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud*)，這個資料集被廣泛應用於評估不平衡學習方法的有效性。此資料集蒐集了歐洲信用卡持有人在2013年9月兩天內的交易紀錄，總共有284807筆交易，其中只有492筆試詐騙交易，比例約579:1，是全部交易中的極少數。這個資料集提供了每筆交易中的31個不同的資訊，包含時間、交易金額、是否為詐騙，以及28個因為敏感性問題經由主成分分析(PCA)變換所獲得的資訊。

決策樹 (Decision Tree) 是一個貪婪 (Greedy) 的演算法，透過遞迴的方式由上至下從根結點開始選擇一個特徵將資料分割成兩個子節點，而因為在這種數據極為不平衡的訓練資料中，因為有一種分類是佔非常少數，所以這種少數類的葉節點 (leaf node) 會非常的少，甚至在檢查純度 (purity) 的時候，少數類可能會被直接忽略掉。因為考量了上述的各種特徵，我們決定使用決策樹來當作我們評估各種解決方法的一個標準。

在進行實驗前，我們先把原資料集用分層的方式依照 8:2 的比例切分為訓練集和測試集，可以讓測試集保有原本不平衡資料的情形，藉此確保可以分析驗證訓練出來的模型實際在分類不平衡資料時的表現。在訓練上，我們分別用原始的訓練集以及分別經過隨機欠採樣、隨機過採樣、NearMiss、SMOTE、Borderline SMOTE 和 ADASYN 處理過後的資料來訓練模型並使用交叉驗證(由於直接使用過採樣方法會造成資料過多、訓練過久，所以我們在使用上述屬於過採樣的方法前先將多數類隨機採樣並降低樣本的數量為少數類數量的十倍，再對少數類進行過採樣)，再分析比較決策樹在使用原始訓練資料和不同採樣方法在測試集表現的提升程度，並特別關注模型在分類少數類上的召回率(Recall)表現，以下是各個方法的呈現與結果。

- Original Training Set

  在原始的訓練集中，保有了不平衡資料的分布情形，所以少數類只佔了全部訓練資料的 0.18%，圖一為使用t-SNE和PCA的視覺化呈現。由於多數類占比過多，所以預期的表現結果是決策樹可能會將過多的詐騙交易紀錄判別為非詐騙紀錄，圖二為訓練出的決策樹在測試集的表現，可以看出，雖然準確率幾乎達到100%，但在少數類的召回率只有69%，也就是有超過三成的詐騙紀錄沒有被分類出來，可能在實際上不是期待的結果。



圖(一): 原始資料集的分布情形，藍色點為正常交易，紅色點為詐騙交易

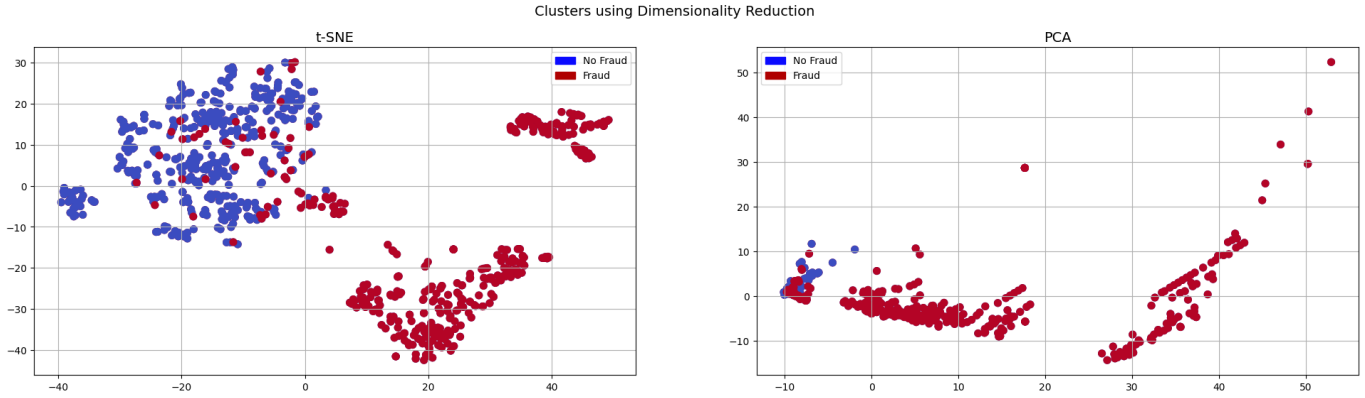
Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56872
1	0.87	0.69	0.77	90
accuracy			1.00	56962
macro avg	0.94	0.84	0.88	56962
weighted avg	1.00	1.00	1.00	56962

圖(二): 使用原始資料經由決策樹所產生的分類報告

- Random Undersamping

  隨機欠採樣會隨機保留多數類中的樣本，並使不同類別中的樣本數趨於平均，圖三為經過隨機欠採樣後的訓練集視覺化呈現，圖四為實驗結果，可以看出由於不平衡資料的情形得到緩解，所以決策樹在少數類的召回率表現有所提升，能夠有超過八成的詐騙紀錄被分類出來。



圖(三): 經隨積欠採樣後的分布情形

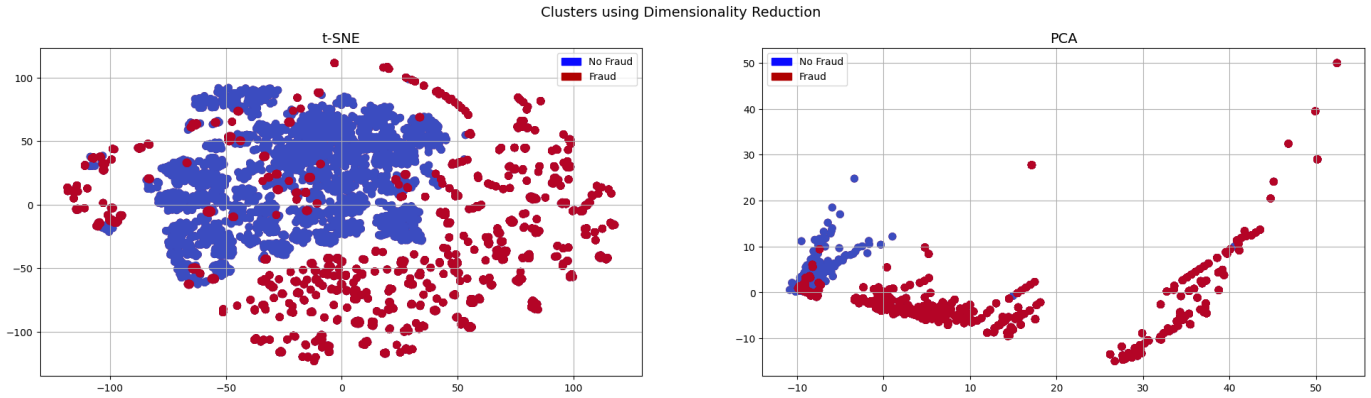
Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	56872
1	0.05	0.82	0.10	90
accuracy			0.98	56962
macro avg	0.53	0.90	0.54	56962
weighted avg	1.00	0.98	0.99	56962

圖(四): 使用隨積欠採樣資料訓練後決策樹所產生的分類報告

- Random Oversamping

  隨機過採樣和隨積欠採樣的策略相反，會隨機複製少數類的樣本，圖五為經隨機過採樣後的訓練集視覺化呈現，圖六為實驗結果，可以看出效果和使用隨機欠採樣相像，也能在少數類達到超過八成的召回率。



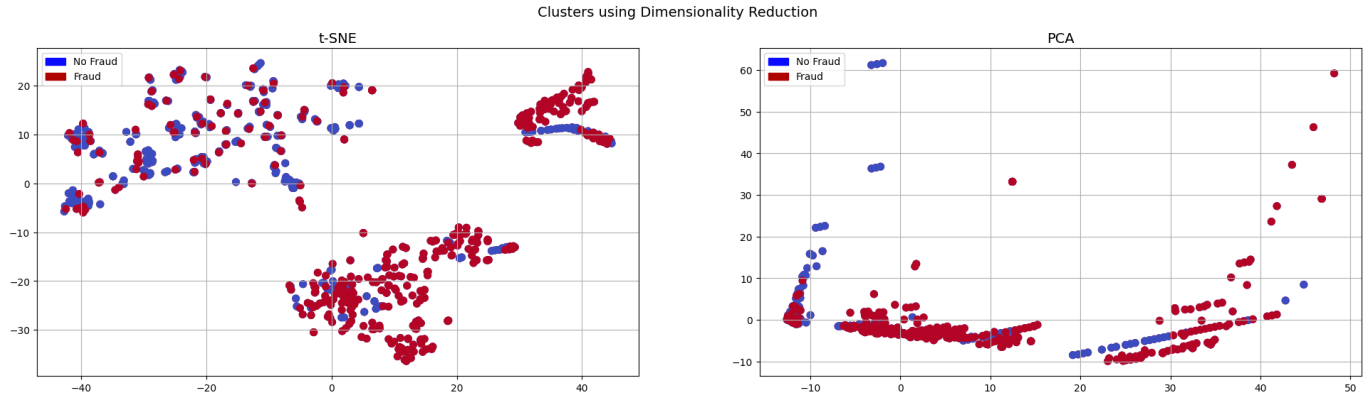
圖(五): 經隨積過採樣後的分布情形

Testing Classification Report:				
	precision	recall	f1-score	support
0	1.00	0.98	0.99	56872
1	0.05	0.83	0.09	90
accuracy			0.97	56962
macro avg	0.52	0.90	0.54	56962
weighted avg	1.00	0.97	0.99	56962

圖(六): 使用隨積過採樣資料訓練後決策樹所產生的分類報告

- NearMiss

  NearMiss欠採樣方法有三種版本，經我們實驗過後版本三是表現最好的，雖然NearMiss-1和NearMiss-2在少數類的召回率很高能超過九成，但會把過多甚至超過50%的正常交易紀錄分類為詐騙紀錄，圖七為經NearMiss-3欠採樣後的訓練集視覺化呈現，圖八為實驗結果，從結果可看出，雖然在少數類的召回率上相比使用原始訓練集有所提升，但由於NearMiss-3的策略特性是傾向讓每個少數類樣本周圍都有多數類樣本留下，可能使得少數類召回率的提升幅度有限。



圖(七): 經NearMiss-3欠採樣後的分布情形

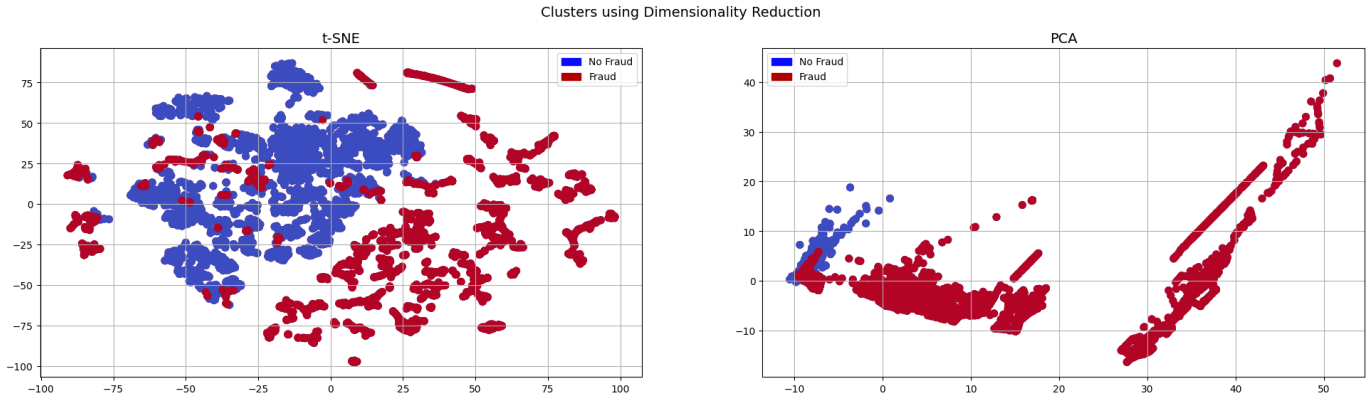
Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	56872
1	0.09	0.77	0.15	90
accuracy			0.99	56962
macro avg	0.54	0.88	0.57	56962
weighted avg	1.00	0.99	0.99	56962

圖(八): 使用NearMiss-3欠採樣資料訓練後決策樹所產生的分類報告

- SMOTE

SMOTE會藉由在少數類別中合成新樣本來達到平衡不同類別的樣本數， 圖九為經SMOTE過採樣後的訓練集視覺化呈現，圖十為實驗結果，從結果可看出SMOTE用在訓練資料上效果相當不錯，能夠使決策樹在測試資料上中的少數類達到將近九成的召回率。



圖(九): 經SMOTE過採樣後的分布情形

Testing Classification Report:

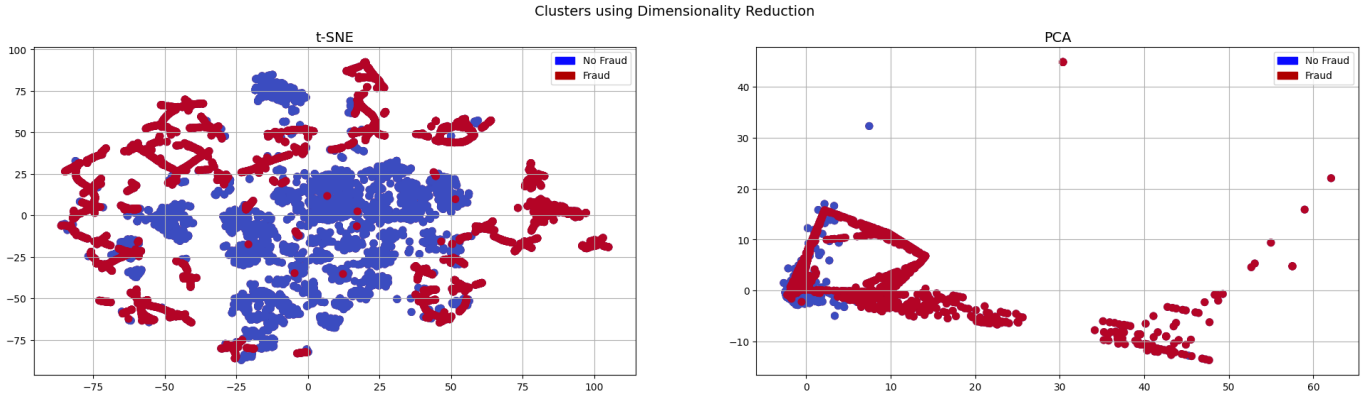
	precision	recall	f1-score	support
0	1.00	0.97	0.98	56872
1	0.04	0.88	0.08	90
accuracy			0.97	56962
macro avg	0.52	0.92	0.53	56962
weighted avg	1.00	0.97	0.98	56962

圖(十): 使用SMOTE過採樣資料訓練後決策樹所產生的分類報告

- Borderline-SMOTE

Borderline-SMOTE相比SMOTE會針對於少數類中的邊界樣本來進行新樣本的合成，從圖十一中的視覺化呈現可看出差異，而圖十二為實驗結果，從結果可看出雖然召回率超過了85%，但並沒有比SMOTE好，原因可能

在於此資料集的類別重疊問題不嚴重，SMOTE的合成樣本可能有更均勻的優點。



圖(十一): 經Borderline-SMOTE過採樣後的分布情形

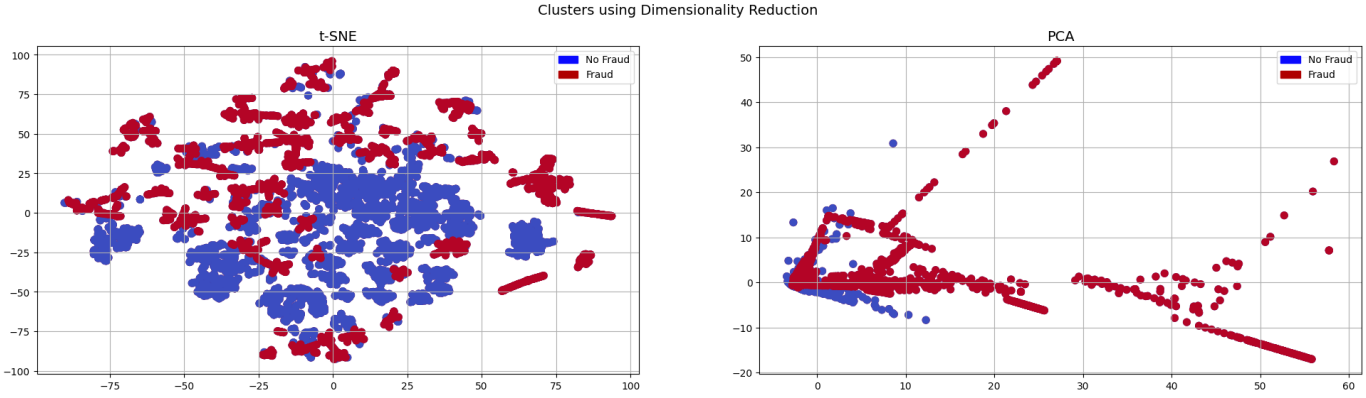
Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	0.95	0.98	56872
1	0.03	0.86	0.05	90
accuracy			0.95	56962
macro avg	0.51	0.90	0.51	56962
weighted avg	1.00	0.95	0.97	56962

圖(十二): 使用Borderline-SMOTE過採樣資料訓練後決策樹所產生的分類報告

- ADASYN

ADASYN過採樣同樣針對少數類中的變邊界樣本進行合成，越靠近多數類的樣本更容易被選擇來合成新樣本，圖十二為經ADASYN過採樣後的訓練集視覺化呈現，而圖十三為實驗結果，從結果可看出ADASYN在少數類召回率的表現略差於SMOTE，原因可能在於ADASYN的策略在此資料集上有噪聲放大(noise amplication)的問題。



圖(十一): 經ADASYN過採樣後的分布情形

Testing Classification Report:

	precision	recall	f1-score	support
0	1.00	0.92	0.96	56872
1	0.02	0.84	0.03	90
accuracy			0.92	56962
macro avg	0.51	0.88	0.50	56962
weighted avg	1.00	0.92	0.96	56962

圖(十二): 使用ADASYN過採樣資料訓練後決策樹所產生的分類報告

6. Discussion

不平衡學習的議題來自於分類問題中不同類別樣本數的不均衡，導致訓練出來的分類器可能有偏向於多數類別的情況，而這個情形常見於機器學習在各種現實生活中的應用場景，比如說罕見疾病偵測、垃圾郵件檢測、欺詐檢測等等，為了應對這個問題，很多不同類別的方法被提出，而在這個 Project 中我們主要參考的文獻屬於使用採樣方法來解決不平衡資料的問題，其中包含過採樣方法和欠採樣方法，我們認為不同的演算法可能會適合用在不同的應用或資料集上，以我們實驗中選擇資料集來說，SMOTE在結果的表現上是最好的，能透過合成新的少數類樣本來填補樣本空間，使各類別樣本數均衡後能夠明顯地提高對少數類別的學習能力。

7. Reference

[1] Mani, Inderjeet, and I. Zhang. "kNN approach to unbalanced data distributions: a case study involving information extraction." *Proceedings of workshop on learning from imbalanced datasets*. Vol. 126. ICML, 2003.

[2] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

[3] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1*. Springer Berlin Heidelberg, 2005.

[4] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008.