1. Explain your implementation which get the best performance in detail.
    a. KNN 算法是基於距離進行的異常檢測，適合用於特徵空間中正常點和異常點有明顯分隔的情況
    b. KNN 是無監督學習算法，不需要標記異常點的先備知識，而是直接從數據中學習異常的分布
    c. KNN 的異常分數直觀易懂，是通過樣本與其最近的距離來量化異常程度

2. Explain the rationale for using auc score instead of F1 score for binary classification in this homework.
    a. 處理不平衡的數據：因為 AUC 對類別不平衡的狀況更穩定，F1 score 是 precision 和 recall 的調和平均數，而在類別不平衡的情況下，F1 score 可能會偏向多數類，導致對少數類的檢測效果不佳。AUC 則考慮了所有閾值下的分類性能，可以更好的反應模型在不平衡數據中的表現
    b. 全面評估模型性能：透過 ROC 曲線反映模型在各個閾值下的 TPR 和 FPR，相較之下，F1 score 只考慮一個固定閾值下的 precision 和 recall，無法全面反映模型的整體性能，F1 評分則依賴選定的分類閾值，不同的閾值會得到不同的 F1 評分。因此，AUC 可以避免因閾值選擇帶來的效能評估偏差，提供更穩定且可靠的效能指標
    c. 閾值獨立性：AUC 與特定的值無關，提供了對模型性能的全局視角，F1 評分則依賴選定的分類閾值，不同的閾值會得到不同的 F1 評分。因此，AUC 可以避免因閾值選擇帶來的效能評估偏差，提供更穩定且可靠的效能指標
    d. 更好的區分能力：AUC 反映了模型區分正負類別樣本的能力。如果模型在 AUC 上表現良好，表示它能有效地將正類樣本與負類樣本區分開來。這在異常檢測任務中非常重要，因為異常通常是少數且難以檢測的樣本

3. Discuss the difference between semi-supervised learning and unsupervised learning.
    a. semi-supervised learning
        - 數據標註：半監督學習使用的訓練數據既包含有標籤的數據，也包含大量沒有標籤的數據
        - 目標：透過利用少量的標註數據和大量的未標註數據，提高模型的泛化能力和分類性能
        - 常見演算法： 半監督支援向量機（S3VM）、圖半監督學習（Graph-based Semi-supervised Learning）、自我訓練（Self-training）、共訓練（Co-training）等

- 提高分類效能： 半監督學習的主要目標是透過結合少量標註資料和大量未標註資料來提高分類器的效能。它假設未標註資料能夠提供關於資料分佈的信息，從而輔助分類器的訓練
- 應用場景： 文字分類、影像分類、醫學影像分析等。在這些場景中，標註資料昂貴或難以取得，而未標註資料相對容易取得
- 應用於標註資料稀缺的分類任務： 半監督學習適用於那些標註資料難以取得但未標註資料相對豐富的任務。例如，在醫學影像分類中，醫師標註的影像有限，但可以獲得大量未標註的影像
- 結合標註與未標註資料： 半監督學習利用少量的標註資料來引導模型學習，同時利用大量的未標註資料來增強模型的泛化能力
- 文字分類： 使用少量已標註的文件和大量未標註的文檔，透過半監督學習演算法（如自訓練或共訓練）訓練文字分類器，從而提高分類器的準確性
- 影像分類： 使用少量標註的影像和大量未標註的影像，透過圖半監督學習演算法（如圖卷積神經網絡，GCN）訓練影像分類模型，以更好地識別影像中的對象

b. unsupervised learning
- 資料標註：無監督學習使用的訓練資料沒有任何標籤，即所有資料都是未標註的
- 目標：透過分析資料的內在結構和模式，對資料進行聚類、降維或異常檢測等任務
- 常見演算法：K-means、層次聚類、DBSCAN、PCA（主成分分析）、t-SNE（t 分佈隨機鄰域嵌入）等
- 資料模式發現：無監督學習的主要目標是發現資料的內在結構或模式。例如，透過聚類找到資料中的自然分組，透過降維方法找到資料的主要特徵或模式
- 應用場景：客戶細分、影像壓縮、異常偵測、資料視覺化等
- 應用於資料探索和模式發現：無監督學習通常用於初步的資料探索、模式發現和資料理解。例如，市場區隔、客戶分組和異常檢測等
- 無需人工標註：由於不需要標註數據，無監督學習適用於數據量大且無法人工標註的情況