運用AI進行自殺分析&預測

國立高雄科技大學
National Kaohsiung University of Science and Technology

班級:四電三甲
製作學生:李福祥、陳家斌、黃千宸
指導老師:戴鴻傑 教授

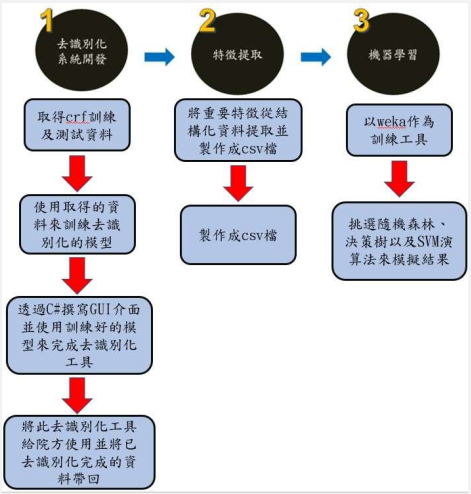
組別:C10

摘要

自殺防治是全世界重要的公共衛生與醫療議題，自殺不僅僅對個人造成傷害，對於家庭也是一個不可抹滅的創傷。導致自殺的成因錯綜複雜，需考量多面向的資訊來達成有效的自殺防治，例如：具有巨大經濟與社會負擔或非致命性自殺企圖的患者皆為自殺的高風險族群。

自殺問題的範圍及嚴重性已引起大量研究關注，隨著近年人工智慧技術的快速發展，**本專題以AI的方式**分析臨床醫生評估報告、患者的自述結果以及電子健康記錄**應用LSTM、CNN深度學習技術**、並結合開發**C#去識別化系統**，保護資料中患者的**隱私**， 將敏感資料去識別化，供院方將相關資料釋出，便於專題研究使用。 取得已去識別化的資料集後，依據資料集中病患是否再次自殺，將群體分為**六個月內再自殺**、**一年內再自殺**、**一年後再自殺**以及**無再自殺**。隨後我們將根據此資料集分別進行**資料前處理**、**特徵工程**以及**特徵選取**並基於**多種不同的機器學習演算法**進行模型的開發並評估其效能，**輔助**醫事人員**評估病患的自殺風險**，落實自殺防治**並降低再自殺人數比例**。

本專題實作過程分為三部份(如圖1），**去識別化系統開發**、**資料前處理**、**機器學習**。



▲圖1

資料前處理

(一)、資料內容

院方釋出去識別化資料後，資料項目多達數百項透過資料前處理刪除非必要欄位、抽取與自殺風險有關的變項。

(二)、資料統計

合作醫院釋出資料中，病患可能多次住院及看診，數量統計結果如下表、圖3。

欄位名稱	筆數
住院主檔adm_d	5130
住院明細adm_p	521405
門診主檔opd_d	102092
門診明細opd_p	821908

	為再自殺	再自殺	<=90天	<=180天	<=365天	>365天
筆數	2007	583	53	124	106	300

▼圖2

(三)、資料整理

1.年齡，我們將患者年齡以10歲為間隔分成十一群，避免年齡特徵過於分散，降低模型準確度。

2.住院及門診，考量到季節會影響人類的心理因素，我們以三個月做為分類，結合醫師診斷碼，用四位元編碼表示90天數內有無包含該類診斷碼。如ICD10（ F32 憂鬱症 ）患者90天內住院2次、90天後門診3次，F32欄編碼為1001。

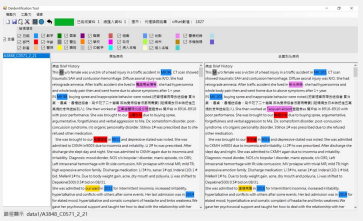
3.國際疾病分類(ICD 10編碼)，原始資料中，ICD10編碼為患者疾病細項，將同為呼吸系統疾病詳細區分為 ICD10（J06.9）急性上呼吸道感染、ICD10（J96.91）呼吸衰竭併缺氧……等，將此細項列為機器學習特徵顯得複雜，因此，我們將細項統計後，將其編碼分為二十二大類，並與院方討論後，僅將精神相關疾病的代碼加以細分，如表，篩選精神相關疾病為重要特徵。

精神相關疾病代碼	疾病名稱
F00	阿茲海默病（失智症）
F01	血管性痴呆
F02	分類於他處的其他疾病引起的痴呆
F03	未特指的痴呆
F04	器質性遺忘綜合症，非由酒精和其他精神活性物質所致
F05	譫妄，非由酒精和其他精神活性物質所致
F06	由於腦損害和機能障礙及軀體疾病引起的其他精神障礙
F07	由於腦部疾病、損害和功能障礙引起的人格和行為障礙
F09	未特指的器質性或徵狀性精神障礙

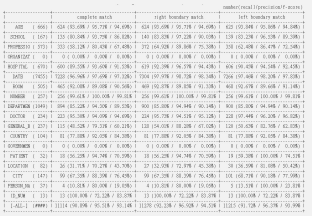
圖3 ▶

去識別化系統開發

專題擬運用合作醫院院區內 2005 至 2020年間所有急診就診患者的相關數據與文字敘述的轉診單做為後續自殺風險評估因子模型的發展資料。然而，此資料包含病患的隱私資訊，在資料可以釋出前，須經過去識別化的處理方可做為研究使用。因此我們應用已發展的去識別化技術結合 C# 程式來開發非結構化文本資料的去識別化系統(如圖4，圖5為去識別化系統模型的跑分），提供合作醫院的大數據中心使用。



▲圖4

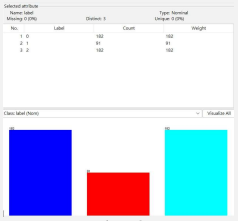


▲圖5

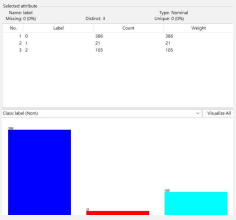
特徵提取、機器學習

1. 利用weka內建的資料處理，使不平衡的數據依照Oversampling及Undersampling兩種方式，使得數據分布可以平衡，讓數據可以提升。由於需要利用虛擬資料來增加原有數據，一開始先以分成8:2的訓練及測試資料，再以訓練資料依照2:1:2的方式來平衡各類的數據，如圖6所示。

2. 再以此訓練資料來利用支持向量機（ Support Vector machine）、隨機森林（ RandomForest ）、決策樹（ Decision Tree ）、貝式分類（Naïve Bayes）四種機器學習來建模。測試資料如圖7所示。



▲圖6



▲圖7

3. 最後測試資料來對分別建好的模型做評分，評分數據及混淆矩陣如下圖所示：

1. 支持向量機(如圖8)

正確率：49.4141%
召回率：0.494
F-score：0.545

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 231  55 100 |   a = 0
   5   9   7 |   b = 1
  50  20  35 |   c = 2
```

▲圖8

2.隨機森林(如圖9)

正確率：59.9609%
召回率：0.6
F-score：0.623

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 250   5 131 |   a = 0
   8   1  12 |   b = 1
  49   0  56 |   c = 2
```

▲圖9

3.決策樹(如圖10)

正確率：53.7109%
召回率：0.537
F-score：0.583

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 231  55 100 |   a = 0
   5   9   7 |   b = 1
  50  20  35 |   c = 2
```

▲圖10

4.貝式分類(如圖11)

正確率：66.7969
召回率：0.668
F-score：0.661

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
 315  13  58 |   a = 0
  11   1   9 |   b = 1
  73   6  26 |   c = 2
```

▲圖11

結論

由於資料取得不易，只能從醫院單方面取得資料，這種情況下，資料分布會顯得十分不平均，進而導致測試分數大大降低，以測試資料分布來看，class b(即為<=90天內在自殺者)為最少類，且在訓練過程中也時常遇到class b預測值為0的情況。以四種機器學習得出的結論，雖以貝式分類平均召回率最高，但若以各類拆分開來看，決策樹判斷class b 為0.429最高分，其次為支持向量機 0.381分。