運輸計劃季刊
第五十一卷 第四期
民國一一一年十二月
頁 275～頁 301

Transportation Planning Journal
Vo1. 51 No. 4
December 2022
PP. 275～301

數據不平衡下以機器學習方法預測交通事故嚴重性之分析

MACHINE LEARNING METHODS FOR TRAFFIC ACCIDENT SEVERITY PREDICTION UNDER IMBALANCED DATA

胡大瀛 Ta-Yin Hu [1]
李岳洪 Yueh-Hung Li [2]

(110 年 4 月 26 日收稿，110 年 11 月 2 日第一次修正，110 年 11 月 9 日接受)

摘 要

降低事故的嚴重程度是近年來全世界努力的方向，全球已經發展出許多被動式安全系統來減緩事故嚴重程度，如安全帶、安全氣囊、煞車輔助系統等等，建立預測事故嚴重性的模型也是許多學者研究的目標，近年來機器學習以及深度學習的方法取代統計方法，可以達到較高的準確度以及運算效率，然而進行模型訓練時需要大量的數據，但肇事資料庫中存在著數據不平衡的問題，因此如何處理這種狀況將是一項重要的課題。

本研究將交通事故嚴重性分為死亡、受傷、未受傷三個等級，為多元分類問題，並收集臺南市的公開資料庫且利用過採樣以及欠採樣兩種資料預處理的方法，對於不平衡的數據進行重新採樣，分別使用 SMOTE 和 Cluster Centroid 這兩種演算法去進行；在模型訓練的部分，採用基於集成學習 (Ensemble Learning) 的兩種分類模型，本文使用 Random Forest 和 Catboost 這兩種演算法來進行兩種集成的學習，研究結果顯示，在欠採樣及過採樣的資料中，兩種模型分別都有 97.69%以及 86.84%以上的準確度，此結果未來可以應用於自駕車上或是給予相關單位作為制定決策時的一些證據。

1. 成功大學交通管理學系教授 (聯絡地址：臺南市東區大學路 1 號成功大學交管系館 62504，電話：06-2757575#53224，E-mail：tyhu@mail.ncku.edu.tw)。
2. 成功大學交通管理學系碩士。

關鍵詞： 事故嚴重性、不平衡數據、機器學習、集成算法

ABSTRACT

Reducing traffic accident severity is an effective approach to improve road safety. To decrease traffic severity, there are many passive safety systems like safety belts, airbags, brake assist systems and so on. In recent years, building models to predict traffic accident severity is also the subject that many researchers focus on. There are a lot of machine learning and deep learning approaches instead of statistical methods. They can get higher accuracy and faster calculate speed. It needs large datasets to train the model, but there is usually an imbalanced data problem in the datasets. Therefore, it must preprocess these sets.

This study divides the traffic accident severity into three levels: death, injury, and non-injury. It is a multi-class classification problem. We collect data from Tainan open datasets and utilize over-sampling and under-sampling methods to resample the imbalanced data. To implement the resample process, we apply SMOTE and Cluster Centroid algorithms separately. We apply two classification models based on the ensemble learning to train the model. This study uses Random Forest and Catboost to execute the two ensemble learning methods. The research results denote that these two models have more than 97.69% and 86.84% accuracy separately in the under-sampling and over-sampling datasets. This result can apply in autonomous vehicles in the future or provide related apartments some suggestions for making the decision.

Key Words： *Traffic accident severity, imbalanced data, machine learning, ensemble method*

一、前　言

全世界每年都有超過 120 萬人死於道路上 [1]，交通事故所造成的損失往往是很難去衡量的，包含人損、財損等等，且額外造成的社會外部成本更是難以估計，每當有車禍發生，便會影響一般的車流，導致多餘的繞道時間以及交通壅塞情形發生，且根據世界衛生組織的統計，交通事故為全球第八死因，如圖 1。

因此，預防道路交通事故已經成為許多國家關鍵的交通政策，在臺灣，2019 年的時候就發生了 30 幾萬起的交通事故 (A1+A2)，造成了將近 3000 人死亡，因為私有運具的普及，導致市中心都會區地帶 (CBD, Central Business District) 面臨上下班的尖峰，都會產生經常性的壅塞 (recurrent congestion)，交通的曝光量增加，就更容易有事故的發生。

臺灣人使用機車的習慣非常之普遍，旅次長度大多較短，而許多機車駕駛人的觀念並

不完善，導致其事故頻傳，雖然各相關機關都有為此進行各種的努力以及措施，但私有運具的成長越來越快，還是沒有辦法有效的降低事故發生。

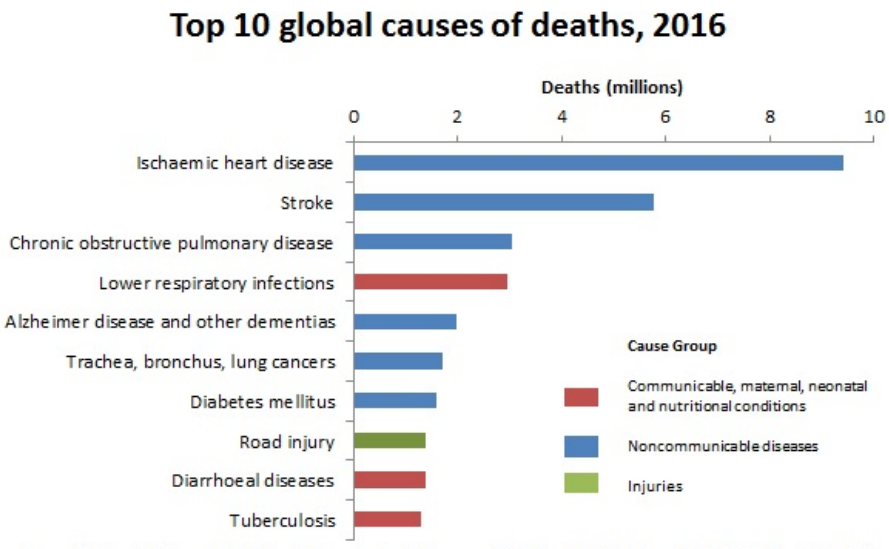


圖 1　全球死因統計圖 [2]

本研究將針對交通事故嚴重性的的問題提出分析架構，期望能預測事故的嚴重程度且找出影響事故嚴重性的關鍵因素，並根據臺南市的交通事故資料庫進行探討以及建立模型，主要考慮到兩個重要的因素：機器學習以及資料預處理，所使用的機器學習方法為集成學習，是屬於監督式學習的一種，其主要的核心為建立很多個小分類器最後合併為一個強分類器；在資料預處理的部分會採用過採樣以及欠採樣的方式進行資料的平衡。

在進行預測模型的建立時，利用實際的肇事資料庫進行訓練，時常會有資料不平衡的狀態發生，因為在現實中，A1、A2、A3 類別的事故比例其實差距很大 (A1 是指造成人員當場或二十四小時內死亡之交通事故、A2 是指造成人員受傷或超過二十四小時死亡之交通事故、A3 是僅有財物損失之交通事故)，尤其是 A1 數量遠遠少於其他兩種，以臺南市 108 年的肇事資料為例，A1 的事故比例約為 0.25%，A2 約為 64.9%，A3 約為 34.8%，存在著嚴重數據不平衡的問題，如果不對於數據進行處理，模型則沒辦法學習到比例較小的樣本特性，意即無法有效的識別 A1 事故，會將所有的事故歸類為 A2 或 A3，模型的準確度雖然會很高，但對於 A1 事故沒有分辨能力。

由於科技的發展，AI 的技術越來越成熟，機器學習發展出許多的分支，如圖 2。

機器學習的內容包含迴歸、強化學習、深度學習、神經網路、決策樹、集成學習、集群等等許多分支，強化學習、深度學習等這些需要經由多層的神經網路稱為非監督式學習，也就是事先建立好各個神經層，再讓模型自己去學習資料中的特徵，不必人工去標記

資料，進行特徵工程的部分；決策樹、集成學習、集群等等這些屬於監督式學習，必須要人工去對於資料進行標籤，並選取適當的特徵及變數給模型做訓練。

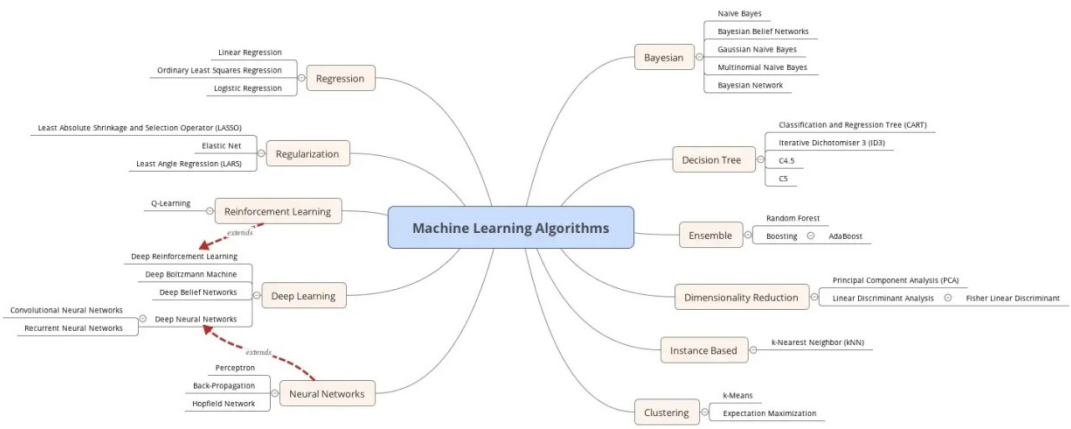


圖 2　機器學習分支圖 [3]

因臺南市 A1 事故死亡人數自 105 年起已連續三年高居全臺第一名，事故發生頻繁，本文將針對此進行相關研究，擷取 108 年臺南市政府公開的事故資料，並進行相關的整理及分析，本文的研究目的主要是藉由利用真實的肇事資料庫，並處理不平衡的數據、進行重新採樣，並建立預測事故嚴重性的機器學習模型，從中得知哪些因素是預測模型中的關鍵因素；期望能夠建立一套流程，能夠對數據事先處理，接著建立分類模型，預測事故嚴重程度，結果可以提供給交通部門單位作為交通安全宣導相關策略制定時的參考依據。

在第二節時會回顧交通安全、機器學習、事故嚴重性相關等文獻，參考他人對於此方面所做的一些研究，在第三節時提出研究流程，包含資料預處理、模型建立、評估標準等，第四節以臺南市 108 年的肇事資料庫為例，進行實證分析，在第五節會對於研究結果進行討論及分析，最後於第六節總結並給予相關建議。

二、文獻回顧

在第二節當中，會分為三個部分來回顧相關的研究，2.1 節會回顧有關交通事故的研究，2.2 節中會回顧機器學習與深度學習的相關應用，2.3 節中會探討對於分類事故嚴重性的相關文獻。

2.1 交通事故的相關研究

Zheng et al. [4] 探討因為交通事故所引起的非重現性壅塞，此種壅塞常會讓駕駛覺得很困擾，因為它常會導致意料之外的旅行時間延遲，進而導致錯過重要的會議或約會。非重

現性壅塞的主要原因包括惡劣的天氣條件、自然災害和交通事故。儘管已有大量研究調查不利的天氣條件和自然災害如何影響城市道路中的壅塞，但很少有研究來探討由交通事故所引起的非重現性壅塞。當城市道路發生交通事故時，塞車的情形將會蔓延並影響到鄰近的道路。這篇研究調整了 Dijkstra 演算法，以識別事故附近受影響的交通。研究結果顯示：(1) 交通壅塞程度主要與交通事故的類型、所涉及車輛的類型以及發生的時間有關；(2) 對於三種類型的交通事故，即車輛之間的擦撞、碰撞到固定物體 (自撞) 和追撞事故，與前兩種類型的壅塞程度是差不多的，而因為第三類事故所造成的壅塞會更加的嚴重；(3) 由於公車或是卡車事故所造成的壅塞程度會比小客車之間碰撞所造成的壅塞程度高約 6%；(4) 就發生時間而言，與尖峰小時相關的壅塞程度會比與非尖峰時段相關的壅塞程度高。

Naqvi et al. [5] 則是考慮油價對於交通事故的影響，在過去的十年內，在已發展國家中，由於汽車、道路設計、醫療技術以及駕駛教育和培育方面的改善，這些國家的道路交通事故有所減少。不過，最近有其他的研究顯示燃油價格的變化會透過其他中介因素對道路交通事故產生重大影響，舉例來說，透過減少汽車行駛和更有效率的駕駛 (例如在高速道路上降低速度) 來減少交通曝光量。而這篇文章主要是在研究英國國內透過旅行行為的變化來量化燃油價格對道路交通事故頻率的影響。使用一階自回歸 AR (1) 的模型以及季節性自回歸綜合移動平均模型 (SARIMA) 研究了道路交通事故的影響。結果發現，燃油價格每上漲 1%，致命道路交通事故的數量就會減少 0.4%。

在印尼 [6]，交通事故也是政府長久以來頭痛的問題，雖然許多學者一致認為交通事故是三種不同類型因素的結果，包括人為因素、車輛因素和外部因素 (包括道路狀況)，但根據全球的統計，人為因素的影響最大。不過，鑑於印尼的交通事故不斷增加，作者試圖找出非人為因素的問題，其中機車涉及的事故佔大多數，對於機車的快速增長，突顯出印尼許多城市的交通運輸系統供需不平衡，此外還有缺乏足夠道路基礎設施等其他因素，為了要降低機車的使用並減少交通事故，必須要提供充足且負擔得起的大眾運輸系統。印尼的雅加達正在進行一些公共交通方式的開發項目，而其他大城市的目標也差不多，若是有涉及私人部門的各種籌資計劃，可能還會更加速公共交通的提供並改善總體交通安全。最後建議三個可能的改進方案：(1) 強力推動及發展公共交通 (2) 改善汽機車道路比例 (3) 採取相關交通管理措施。

會造成事故的原因大多數都是人為因素造成的，其中疲勞駕駛是交通事故的主要原因，但許多人仍不太清楚其潛在危害，所以疲勞駕駛又被稱為「沉默的殺手」。因此，有必要對於疲勞駕駛和其潛在的危險因素進行研究。G. Zhang et al. [7] 分析了中國廣東省 2006 - 2010 年的交通事故數據。研究資料來自中國公安部交通事故資料庫，並且使用羅吉斯回歸模型評估駕駛員特徵、車輛類型、道路狀況和環境因素對與疲勞相關的交通事故的發生、嚴重程度的影響。結果顯示，男性司機在午夜至凌晨行駛的卡車以及早晨的尖峰時間被確定為與疲勞相關的碰撞危險因素，但不一定會造成嚴重的人員傷亡，在沒有路燈的情況下夜間駕駛會導致疲勞相關的事故和嚴重的人員傷亡。在另一方面，雖然經驗不足的

駕駛員、不安全的車輛狀態、濕滑的道路以及週末行駛等因素對與疲勞相關的撞車沒有明顯影響，但是與這些因素相關的事故很可能會有嚴重的人員傷亡。

而在教育的方面，也有學者進行相關的研究，Nakai and Usui [8] 討論不同交通運輸方式之間 (例如，汽車與行人之間，汽車與自行車之間或汽車與摩托車之間) 的道路交通事故時常發生。在這種兩種不同運具可能會發生碰撞的情況下，重要的是要考慮另一方的觀點以及角度，這就有關到正確判斷道路上其他人如何看待周圍環境以及下一步打算做什麼的能力。在這篇文章中，作者進行兩種類型的研究，在第一個研究中，分析了 65 歲至 74 歲之間涉及老年人的事故，這些事故與行人或騎自行車的人相撞，並考慮到事故現場的事故類別和道路類型。結果顯示，是否擁有駕照與騎自行車的事故無關，但對於與行人發生的事故，沒有駕照的老年人可能會涉及更多的事故。在第二個研究中，作者審查了 875 名要考駕照的考生，並根據他們的駕照持有狀態 (摩托車駕照、輕型摩托車駕照或沒有任何駕照) 對他們進行了分類，並比較他們左轉彎的能力以及駕駛習慣。結果顯示，擁有摩托車或輕型摩托車駕照的人傾向於更安全地左轉。總結來說，對於擁有不同運輸方式的經驗可能會減少發生事故的風險。這些研究結果可用於普及交通教育相關規劃，鼓勵各種運輸方式的用戶考慮其他用路人的觀點。

2.2 機器學習與深度學習於交通上之應用

近年來，AI 發展迅速，各領域的研究越來越多採用人工智慧的方式來進行，而在機器學習當中，又可分為無監督學習和監督式學習，其中有無監督是以是否需要人工去標籤以及處理資料來做區分，無監督學習大多是為了要處理特徵眾多且難以人工方式整理的資料，像是圖像、音樂、影片、自然語言處理等等，由於沒辦法各個去標籤每張圖片的特徵值，需要預先建立好神經網路或是相關模型，使機器自己可以能夠去學習資料裡面的特徵，並模仿人類的神經網路系統，藉由大量的數據來進行訓練。

監督式學習則是為了進行數據的分類或是回歸，需要人工事先設定資料的標籤值，讓機器去學習在各種變數的影響之下，其結果為何，舉例來說，像是要分析消費者的消費習慣，就必須先將消費者根據其社經條件或是消費狀況分為各個類別，並針對每種不同的分群進行不同的行銷策略，來提升整體的銷售業績，其中學習如何分類的這個步驟就可以交給機器來學習，只要給予足夠大量的資料，就可以訓練出不錯的預測模型，準確判斷消費者屬於哪一群。

辨識事故的集中區域，或稱為交通事故黑點，有助於提供交通管理部門一些有利的證據，Fan et al. [9] 基於蘇州工業園區交通事故數據，對交通事故黑點涉及的多元影響因素進行分析，利用支援向量機 (SVM) 訓練模型和學習事故黑點的判別。同時，針對交通事故數據的快速增長，提出了一種基於深度神經網絡的黑點識別算法，建立了相關數據類別訊息的深度神經網絡，以驗證模型識別事故黑點的能力。此外，建立了動態自適應機器學習架構，改善了黑點識別的準確性。最後結合人、車輛、道路和環境等 10 個特徵的模型，在

機器學習中使用支持向量機對交通黑點進行預測和分析，獲得了 63%的精確率和 61%的召回率。

而分辨駕駛行為也是許多學者研究的方向，若是能準確的判斷駕駛目前是否分心或是疲勞等，即可及時提醒駕駛，避免事故的發生，Osman et al.[10] 建立了一個雙層模型，利用速度、縱向加速度、橫向加速度、踏板位置以及偏移率等五項變數，去分辨駕駛是否進行別的行為以及所進行分心的行為是屬於哪種，並使用九種分類樹的方法去評估各分類樹的表現好壞，其中隨機森林 (Random Forest) 可以達到最高的準確性。隨著車路聯網的技術逐漸發展，可以將此輔助識別模型加入車載系統中，用以檢測危險的駕駛行為，並提醒駕駛員注意駕駛以及前方路況。

在高速公路的事件偵測上，也可以應用機器學習的方法來進行，Parsa et al. [11] 使用極限梯度提升 (XGBoost)，透過交通、網路、人口統計、土地使用和天氣特徵組成的即時數據來檢測事故的發生，收集 2016 年 12 月至 2017 年 12 月在芝加哥高速公路的數據，其中包括 244 起交通事故和 6073 起非事故樣本。此外，採用 SHAP Value 來解釋結果並分析各個特徵的重要性。結果顯示，XGBoost 能夠以 99%，79%和 0.16%的準確度、檢測率和誤報率，尤其是事故發生前 5 分鐘和事故發生後 5 分鐘之間的速度差異，對事故的發生有相對較大的影響。

而在深度學習的部分，Z. Zhang et al. [12] 調查了美國兩個大都市地區一年中超過 300 萬條的推文內容，並運用深度學習從社交媒體數據中檢測交通事故。結果顯示，作者所建立的模型可以捕獲與事故相關的推文中固有的關係以及規則，並進一步提高交通事故檢測的準確性。並研究了兩種深度學習方法：深度信念網絡 (DBN) 和長短期記憶網路 (LSTM)，使用大約 44 個單獨的特徵和 17 個配對的特徵，DBN 可以獲得 85%的準確性而且其分類結果優於支持向量機 (SVM)。最後，為了驗證這項研究，作者將與事故相關的推文與高速公路上的交通事故記錄和來自 15,000 個環路探測器的本地道路上的交通數據進行了比較。結果發現，事故記錄中幾乎有 66%與事故相關的推文可以找到，其中 80%以上與附近的異常交通數據相關。並透過比較，提出了使用 Twitter 檢測交通事故的幾個重要問題，包括位置和時間偏差以及有影響力的用戶和標籤的特徵，給予未來要繼續研究這個領域的學者一些建議。

在交通量預測的部分，Yao and Ye [13] 利用長短期記憶網路 (Long Short Term Memory, LSTM) 取代傳統的預測方法，來推估未來的交通流量，同時對於車輛的車牌、車身顏色等屬性特徵進行辨識，去判斷車輛的組成，研究結果顯示基於 LSTM 的預測模型可以非常接近實際的狀況，誤差小，且預測能力高，以平均絕對百分比誤差 (MAPE) 來衡量模型的好壞，最終模型的誤差都小於 15%，能夠達到良好的預測效果。

2.3 事故嚴重性的分類

交通事故嚴重性的分類模型一直是交通安全研究的重要課題，涉及一些統計的理論，

以及數據挖掘或是機器學習的技術去深入了解影響或與碰撞的嚴重程度相關，預測未來事故的嚴重程度以及結果的嚴重級別。Iranitalab and Khattak [14] 比較了四種分類方法，分別為多項式羅吉特 (Multinomial Logit)、最近鄰分類 (Nearest Neighbor Classification)、支援向量機 (Support Vector Machines)、隨機森林 (Random Forests)，使用 2012 年至 2014 年的數據進行訓練，利用 2015 年的數據來做測試，並使用 K 均值模式 (K-means Clustering) 以及潛在類別模式 (Latent Class Clustering) 進行數據的聚類，結果顯示，最近鄰有最好的分類準確度，再來則是支援向量機和隨機森林，多項式羅吉特則為最差的預測模型，最後定義事故所造成的成本，去觀察每種分類方法在預估事故成本時的預測能力。

Kwon et al. [15] 收集從 1973 年以來加州高速公路巡邏隊所記錄的交通事故報告，每份事故報告都包含大約 100 個欄位，總共約 130 幾萬筆的事故資料。其中，作者調查了 2004 年至 2010 年之間與車禍最相關的 25 個領域。使用樸素貝葉斯分類器和決策樹分類器這兩種分類方法，去探討交通事故間的風險因素，將分類器的性能進行相互比較，發現這些高風險因素彼此高度依賴。

Chen et al. [16] 針對追尾事故 (Rear-end Collision) 進行事故嚴重性的探討，儘管多項 Logit 模型和貝葉斯網路方法都分別用於建立模型和分析，但是它們各自都有自己的應用程序限制和局限性。在這項研究中，開發了一種混合方法，將多項式 Logit 模型與貝葉斯網路方法互相結合，基於 2010 年至 2011 年在新墨西哥州所收集的碰撞數據，對追尾碰撞中的駕駛員傷害嚴重性進行分析，探索用於調查和識別導致駕駛傷害嚴重程度的重要因素，這些嚴重程度可分為三類：無受傷、受傷和死亡。然後利用確定的重要因素來建立貝葉斯網路，包括駕駛員行為、人口特徵、車輛因素、幾何和環境特徵等。研究結果顯示，提出的混合方法表現相當良好，卡車參與、惡劣的照明條件、風大的天氣條件、所涉車輛數量等在內的因素可能會增加追尾事故中駕駛員受傷的嚴重性。

不過在統計資料當中，事故的資料時常是不平衡的，因為通常死亡的事故會遠小於受傷的事故，但這並不代表死亡的事故就不重要，反而是我們要想辦法去處理的部分，Jeong et al. [17] 收集了從 2016-2017 年的 297113 起車禍，與任何其他交通事故的資料庫類似，並沒辦法平衡的代表不同的事故嚴重性等級。為了解決不平衡數據的問題，使用了幾種技術，包括欠採樣和過採樣，這邊所使用的採樣方式都是以隨機複製或是隨機刪除的方式來進行，再來使用五種分類學習模型 (即邏輯回歸，決策樹，神經網路，梯度提升模型和樸素貝葉斯分類器)，對事故嚴重程度進行分類，其中隨機森林的分類模型加上過採樣的資料前處理有最好的分類表現。

AlKheder et al. [18] 使用了三種機器學習的方法，分別為決策樹、貝葉斯網路、支援向量機，來分析交通事故嚴重程度的潛在風險因素，將事故嚴重程度分為四個等級：輕度傷害、中度傷害、嚴重傷害、致命傷害，蒐集 2008 年至 2013 年總共 5740 筆交通事故的資料並進行模型的建立，比較特別的是這篇研究有針對車內乘客的位置不同而設定不同的虛擬變數，結果顯示，行人是道路中受傷嚴重程度最高的族群，男性、前排乘客、年長的駕

駛這幾種特徵也都會更容易受到嚴重傷害或是死亡，使用安全帶可以有效的降低事故的傷害程度，而在模型準確度的部分，決策樹可以達到 66.06%，貝葉斯網路可以達到 66.18%。

2.4 小節

根據上述文獻，各國都對於交通安全這項議題非常重視，舉凡事故熱點的偵測、交通事故的預測、分辨駕駛行為等等都是熱門的研究議題，期望能夠降低交通事故的嚴重程度，減少事故所造成的額外成本，也有許多應用機器學習或是深度學習的方法進行分析和研究的，僅少數有考慮到數據不平衡的問題，但若是以不平衡的數據去訓練模型，其實對於模型來說並無法學習到較少數資料的特徵。

本研究將針對此問題進行資料的預處理，使用兩種不同的採樣方法，對原本的數據進行平衡，再利用兩種機器學習的模型進行事故嚴重性的分類，分別採用隨機森林以及 Catboost 這兩種集成學習 (Ensemble Learning) 的方法，後續會在第三節中詳細介紹。

三、研究方法

3.1 研究流程

本研究的研究架構流程如圖 3 所示，可分為五個部分：資料收集、資料合併和標籤、數據前處理、模型訓練、驗證分析。以下分別說明每個部分的細節：

1. 資料收集：

本研究收集臺南市的交通事故資料庫 108 年資料，由於資料庫中記錄肇事資料的部分區分為兩種，一種是以事件為單位作紀錄，另一種則以當事人為單位作紀錄。

2. 資料合併和標籤：

將兩種資料先以案件編號進行合併，並將資料標籤，給予不同的編號，例如事故地點以城鄉作為區分，給予 1 或 2 的虛擬變數。

3. 數據前處理：

利用過採樣以及欠採樣方法同時處理原始的不平衡數據，使得資料得以平衡，並產生出兩種經過不同採樣的資料。

4. 模型訓練：

進行模型訓練時，切割訓練以及測試集，本研究採用 75%作為訓練、25%作為測試集，測試集為完全不參與訓練及驗證的資料，而在訓練集中，採用十折交叉驗證，將訓練集分為 10 等份並每次抽取不一樣的子集進行驗證，進行參數的校估，在測試前建立一個在訓練集上表現最好的模型。

5. 驗證分析：

最後再利用測試集，用以檢驗模型對於新的樣本之預測能力為何，輸出混淆矩陣，觀察模型的分類結果，接著計算出模型的各項分數，在兩種採樣方式下，分析兩種機器學習的模型表現，分析其特徵顯著程度，觀察各項變數，判斷重要的風險因子。

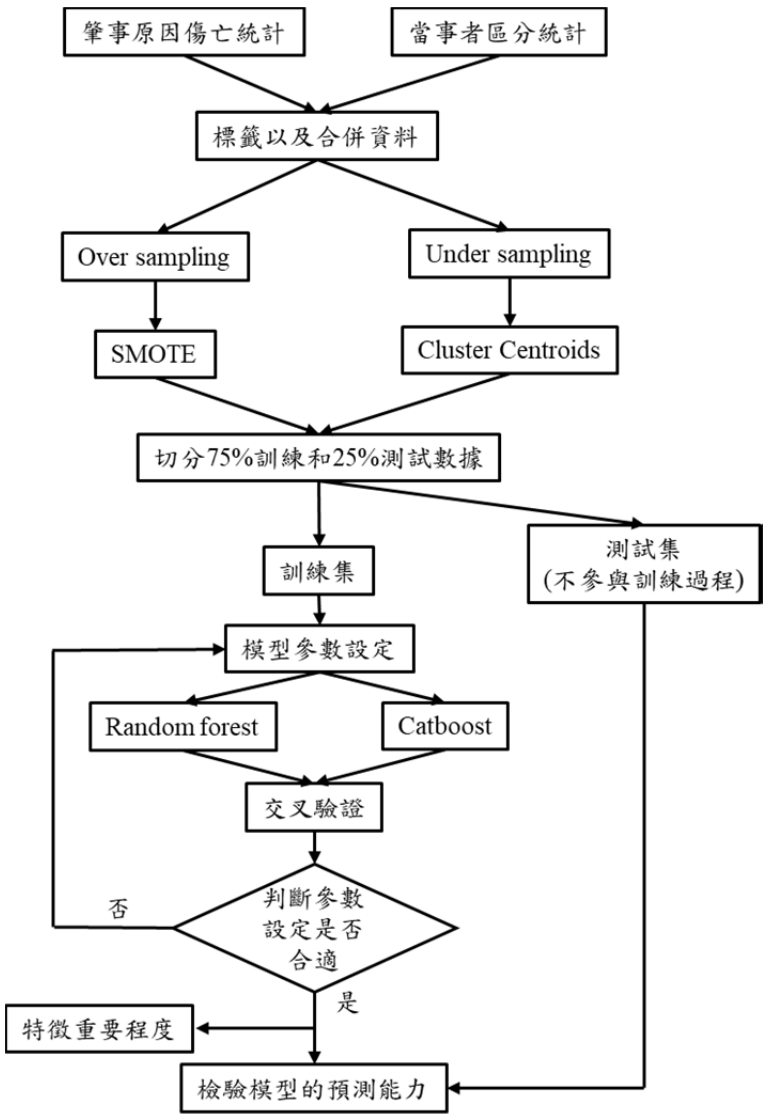
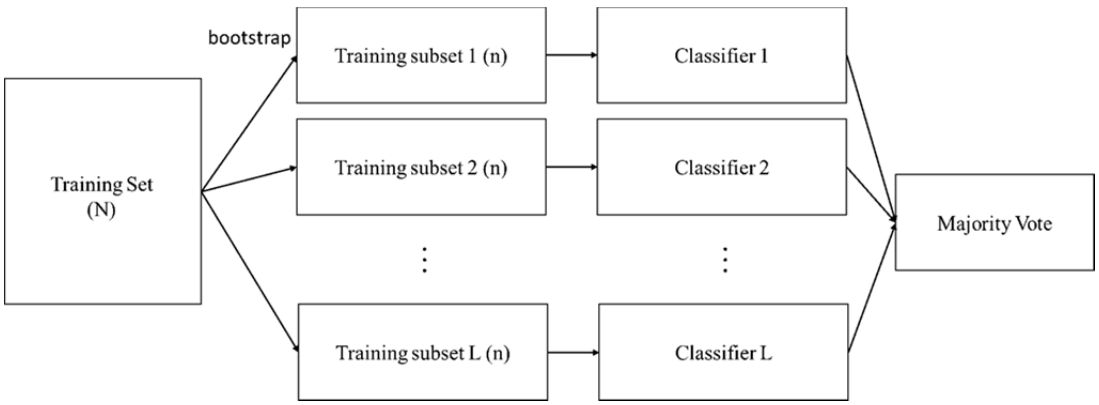


圖 3　研究流程圖

3.2 集成學習 (Ensemble Learning)

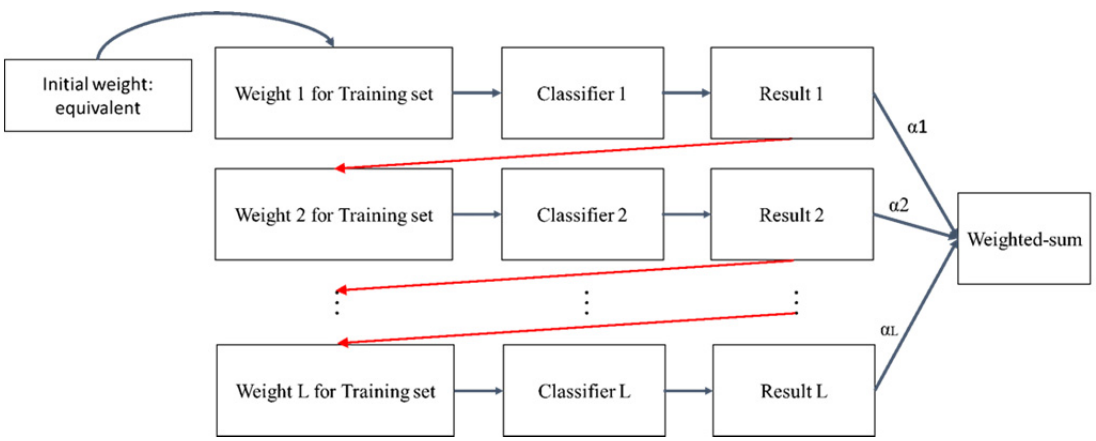
集成學習 (Ensemble Learning) 的基本概念為建立多個小的模型，並整合所有小模型

的結果，因為傳統的建模技術都只著重於調整單一模型的參數，那如果訓練很多個小模型，最終合併起來他們的結果，模型的表現以及預測能力都會提升，這就是集成算法的由來，而常見的訓練方法有 Bagging 和 Boosting 兩種，如圖 4 及圖 5 所示。



圖片來源：Tommy (2018)

圖 4　Bagging 示意圖 [19]



圖片來源：Tommy (2018)

圖 5　Boosting 示意圖 [19]

Bagging 的概念主要是從訓練資料中隨機抽取樣本訓練多個分類器，每個分類器的權重一致最後用投票方式 (Majority vote) 得到最終結果，而這種抽樣的方法在統計上稱為 bootstrap，其主要核心在於從樣本中抽樣。Bagging 的優點在於原始訓練樣本中有不好的資料 (噪聲)，透過 Bagging 抽樣就有機會不讓有噪聲資料被訓練到，所以可以降低模型的不穩定性。常見的應用舉例來說隨機森林 (Random Forest) 就是決策樹加上 Bagging 的集成演算法之應用。

Boosting 算法是將很多個弱的分類器 (weak classifier) 進行合成變成一個強分類器 (Strong classifier)，和 Bagging 不同的是分類器之間是有關聯性的，是透過將舊分類器的錯誤資料權重提高，然後再訓練新的分類器，這樣新的分類器就會學習到錯誤分類資料 (misclassified data) 的特性，進而提升分類結果。因為舊的分類器在訓練有些資料落在混淆區間 (confusion area)，如果再用全部的資料下去訓練，錯的資料永遠都會是錯的，因此我們需要針對錯誤的資料去學習 (將錯誤的資料權重加大)，那這樣新訓練出來的分類器才能針對這些錯誤判讀的資料得到好的結果 [19]。常見的 Boosting 演算法有 Adaboosting、XGboosting 等等，都是在 Kaggle 競賽中時常使用的方法。

3.3 Random Forest

隨機森林 (Random Forest) 是由 Breiman [20] 所提出的一種方法，是決策樹的延伸，利用建立許多個決策樹，集合成一個森林，而每一棵樹之間是沒有關聯的，當要進行預測時，會輸入要預測的特徵，並藉由所有的樹進行判斷並投票，最後根據投票，選出最後的分類結果，每棵樹之間的權重相等，都有決定最後結果的權利，而整個森林的建立則是以隨機的方式建立，故稱之為隨機森林。

以這種方式所建立的決策樹其實單棵是很弱的，並沒有太好的分類效果，不過若是組合全部的樹，那整體的準確度會高非常多，因為每棵樹都是會專注強化某部分的分類，並不會學習到所有的特徵，集合所有樹則表示可以用各種不同的角度去判斷以及看待，最後再經由大家的投票，來做為最後的決策，可以避免模型因為專注於某個部分而過度擬和。

以圖 6 說明隨機森林的預測過程，整個模型是很多棵樹所組成，而每棵決策樹在判斷新的樣本時都會有不同的結果，最後則是透過整合所有的樹，得到最後的分類結果，好處是可以處理高維度的資料，因為會有專精某些特徵的決策樹來判斷，而且對於不平衡的資料，可以平衡誤差，整個森林的訓練過程也非常快速，已經廣泛被運用在各領域的研究上。

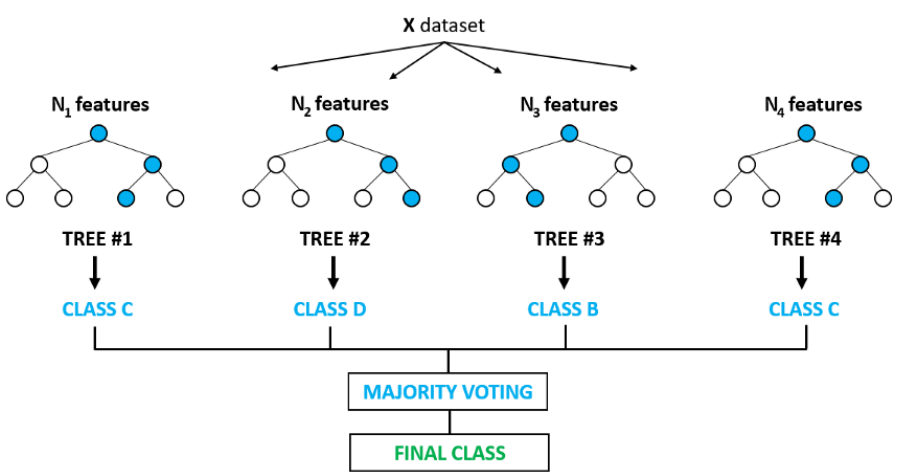


圖 6 隨機森林示意圖 [21]

3.4 Catboost

CatBoost 是俄羅斯的搜尋引擎公司 Yandex 在 2017 年所開發的開源機器學習庫，其名稱是來自於 Category (Categorical Features，類別型特徵) BoostGradient Boosting，梯度提升)，是基於梯度提升決策樹的機器學習框架，可以很好地處理類別變數的分類問題，會自己組合特徵之間的關係，形成新的標籤。

參考 Catboost 官網 (https://catboost.ai/) 的圖來做說明：

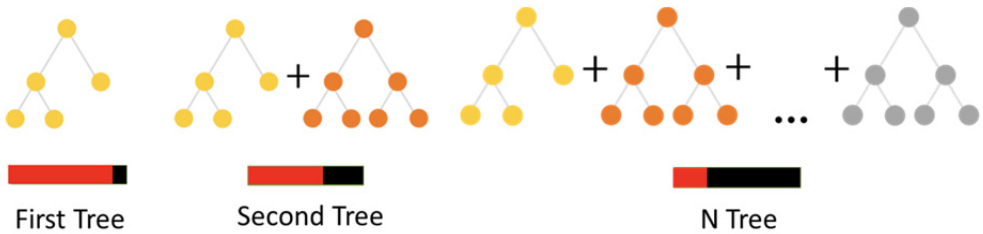


圖 7 Catboost 示意圖 [22]

圖 7 下方的紅色長條表示誤差的比例，黑色表示正確預測的部分，而這項演算法也是透過建立多個分類樹來進行預測，不過與隨機森林不同，每個樹之間是有關連的，圖 7 的第二棵樹是學習第一棵樹的錯誤，並進行修正，利用梯度提升的概念不斷減少誤差，直到第 N 棵樹的時候會有最小的誤差，Catboost 會根據誤差的大小給予每棵樹不同的權重，而最後則是根據每棵樹的分類結果乘上其權重，得出最後的分類結果。

梯度提升是一種功能強大的機器學習算法，已廣泛應用於多種類型的分類及回歸問題，例如詐騙檢測、推薦項目、預測、信用卡盜刷等等，而且效果也很好，與需要從大量數據中學習的深度學習模型不同，它還可以以相對較少的數據產生非常好的結果，其相關更深入的數學模式以及細節可以參考 Dorogush et al. [23]、Prokhorenkova et al. [24] 這兩篇文章。

其優點包含運算速度快，而且對於參數的調整的要求不高，不須花太多時間在模型的調教，並降低過度擬合的情況，可以與 python 的介面整合，上手容易，也支援自定義損失函式，自由調整模型收斂方向，若是需要大量運算，也可使用 GPU 加速。

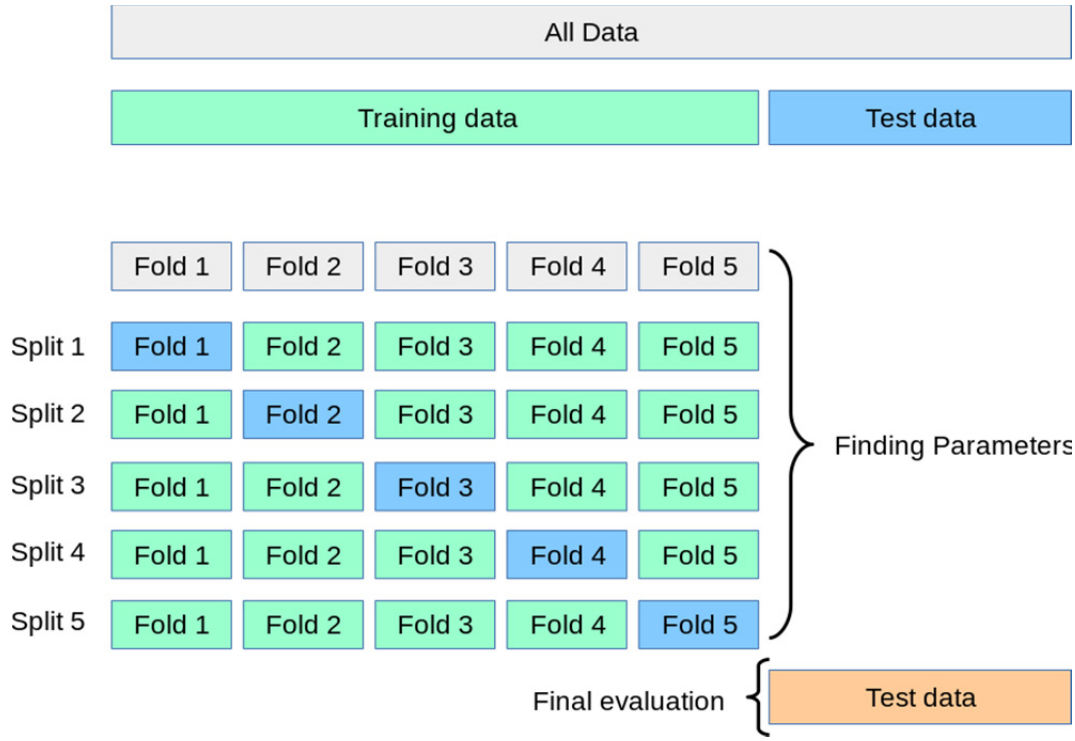
因本研究幾乎所有變數皆為類別變數，因此很適合使用此 boosting 的演算法來進行分類，故採用這項新的機器學習方法。

3.5 交叉驗證

交叉驗證是為了減少模型誤差以及過度擬合的一種方法，若是直接將訓練的資料放入模型讓他學習，那有可能會導致模型會過度專注於學習現有的資料上，使得放入新資料

時，沒有辦法有效的去預測新的數據集，因此交叉驗證則是利用各種方法來處理訓練集，並切分為訓練以及驗證，再經過多次的交叉比對之後得出最後平均分數，如此一來即可得到較為客觀的模型結果，也可利用此結果來評估模型的參數設定是否適合。

一般而言，交叉驗證的方法可分為以下幾種：Holdout Cross Validation、Leave-one-out Cross Validation、K-fold Cross Validation 等等，最常使用的為 K-fold CV，也是本研究所使用的交叉驗證方法，中文譯為 K 折交叉驗證，舉例來說，假設將訓練集切為 5 等份，則使用其中 4 等份來做訓練，並使用其中的 1 份來做驗證，分別進行 5 次確定每等份的資料皆有被訓練或是驗證過後，取其結果的平均，如圖 8 所示。



圖片來源：Pedregosa et al. (2011)

圖 8　交叉驗證示意圖 [25]

本研究先將資料切分為 75%訓練集以及 25%測試集，其中訓練集是以 10 折做為切割數量的設定，意即將資料分為 10 份，每次分別取 9 份來訓練以及 1 份作為驗證的資料，來進行模型的內部驗證，並根據訓練結果調整各模型的參數設定，以求達到最佳的狀態後，再放入測試集中的資料 (完全不參與訓練以及驗證的資料)，測試此模型是否有預測能力。

3.6 模型評估標準

混淆矩陣 (confusion matrix) 又可稱為模糊矩陣，是一種用矩陣呈現的可視化工具，用於監督 (Supervised) 學習的模型評估上，無監督學習通常用匹配矩陣 (matching matrix) 來衡量。

混淆矩陣的名字來自它容易表示出多個類別是否有混淆，可以透過混淆矩陣的一些指標衡量算法的精度。混淆矩陣可以用來總結分類模型預測結果的情形分析表，以矩陣形式將數據集中的記錄按照真實的類別與分類模型預測的類別。而其中矩陣的行表示真實值，矩陣的列表示預測值，在對角線上的格子則為準確判斷的樣本數，以下為二分類的混淆矩陣：

TP (True Positive)：實際是正例，預測為正例。

FN (False Negative)：實際是正例，預測為負例。

FP (False Positive)：實際是負例，預測為正例。

TN (True Negative)：實際是負例，預測為負例。

以表格呈現混淆矩陣如表 1。

表 1　混淆矩陣表

	預測為正	預測為負
原始為正	TP	FN
原始為負	FP	TN

利用矩陣計算指標，可從不同角度觀察分類器，主要有以下幾項指標：

(1) 正確率 (Accuracy)：(TP+TN)/(TP+TN+FN+FP)，用於反映整體預測是否與實際結果一致，即不管是哪種類別，預測結果與實際結果一致的比例。

(2) 精確率 (Precision)：TP/(FP+TP)，在預測的所有結果中，有多少是正確預測的比例。

(3) 召回率 (Recall)：TP/(TP+FN)，在實際為正的情況下，有多少是正確預測的比例。

(4) F-measure：又稱為 F-Score，是準確率和召回率加權調和平均，用於綜合反映整體的指標。計算公式為：(2 * Precision * Recall)/(Precision + Recall)。

四、實證分析

4.1 資料來源

本研究將使用臺南市政府 [26] 的公開資料庫 (https://data.tainan.gov.tw/) 做為實證分析的資料來源，其資料庫是為方便各界更易於取得臺南市政府所提供之開放資料，而臺南

市政府智慧發展中心期望透過建立一個共通的平台，將臺南市的開放資料 (open data) 以通用開放資料格式，集中存放於該平台，以提供較佳的品質與存取方式，此平台的價值不只在於公告各項應公開的施政資訊，也可視為一個內容服務平台，使得相關需要資料的業者或是學者能夠更加有效率地去取得，後續只要專注於研究或是開發軟體即可。

若發生交通事故且有民眾報案，警察將會抵達現場並進行筆錄，而在進行筆錄的過程中會分為兩個表來做填寫，分別為原因傷亡統計表以及當事者區分統計表，一個是以事件為主軸做紀錄，一個則用當事人作為主軸填寫，以下分別介紹兩表紀錄內容：

4.1.1 臺南市道路交通事故原因傷亡統計

此表為紀錄該事件所發生之各項資料，分別為案件編號、發生日期、發生時間、經度、緯度、案件類別、地址類型、地點、死亡人數、受傷人數、天候、速限、道路型態、事故位置、號誌種類、事故類型及型態、肇因研判因素，皆有詳實介紹，本研究在這些資料當中所使用之變數有發生日期、發生時間、地點、天候、速限、事故位置、號誌種類、肇因研判因素，108 年的統計資料中，此部分共有 16000 筆。

4.1.2 臺南市道路交通事故當事者區分統計

第二張表則針對每位事故當事人作紀錄，其資料分別有案件編號、當事者順位、國籍、當事者屬性別代碼、當事者屬性別名稱、當事者事故發生時年齡、當事者區分類別大類別名稱-車種、當事者區分類別子類別名稱-車種、牌照種類名稱、受傷程度名稱、主要傷處名稱、保護裝備名稱、肇因研判子類別名稱等，本研究所使用此統計資料中的內容包含當事者性別、年齡、大類別名稱-車種、受傷程度名稱等，來衡量依據個人之社經變數對於事故嚴重程度的影響，以 108 年度統計資料來看，一共有 89865 筆資料。

4.2 變數定義

統整 3.1 節所介紹之兩種資料，利用 python 裡面對於數據處理的函式庫 pandas 來做合併的步驟，將事故原因傷亡統計表之資料根據案件編號與當事者資料進行合併，並刪除有遺漏或是記載錯誤之樣本，而原因傷亡統計表是以案件作為單位，當事者區分統計是以當事者作為單位，因此傷亡統計表的數量將會少於當事者統計表，故必須利用案件編號來兩種資料進行比對以及整理，最後處理完一共剩下 88708 筆樣本。

且因肇事原因之種類過多，警察在紀錄時會根據當時的狀況選擇不同的肇因，雖然紀錄詳細，但是時常會有幾種類別的原因只會被紀錄到很少筆的狀況發生，可能會導致訓練機器模型的時候，沒辦法很有效地進行訓練以及建模，因此對於 50 幾種的肇事因素，並根據內政部警政署以及參考 Al-Ghamdi [27]，將原本的肇因統整為 10 種大類別，分別為超速、闖紅燈、跟車過緊、車道錯誤、未禮讓、其他駕駛相關因素、機件故障、行人或乘客過失、交通管制 (設施缺陷)、非上所述之其他因素等十大種類，並統整所有變數如表 2。

表 2 變數總表

變數名稱	模型中設定之名稱	判斷依據	標籤
月份	month	根據當月作為標籤	1~12
時間	time	以小時作為單位	0~24
天氣	weather	晴	1
		陰	2
		雨	3
速限	speed	根據當時的速限	0~100
事故位置	road type	交叉路口	1
		路段	2
		其他	3
號誌類型	sign	無號誌	1
		行車管制號誌	2
		行車管制號誌(附設行人專用號誌)	3
		閃光號誌	4
發生事故鄉鎮名稱	location	人口數大於兩萬五(都市)	1
		人口數小於兩萬五(鄉村)	2
肇事原因	reason	超速	1
		闖紅燈	2
		跟車過近	3
		車道錯誤	4
		未禮讓行人或車輛	5
		其他駕駛相關因素	6
		機件故障	7
		行人、乘客過失	8
		交通管制(設施)缺陷	9
		非上所述之其他因素	10
性別	gender	男生	1
		女生	2
年齡	age	根據當時表上所紀載之年齡	1~117
車種	vehicle	小客車及小貨車	1
		機車	2
		行人	3
		其他(大客車、大貨車、連結車、曳引車、特種車)	4
事故種類	injury	死亡	1
		受傷	2
		未受傷	3

4.3 不平衡數據的處理

由於死亡和受傷事故的比例懸殊，在 8 萬多筆樣本當中，只有 230 筆為死亡，57556 筆為受傷，30922 筆為未受傷，如圖 9 所示。

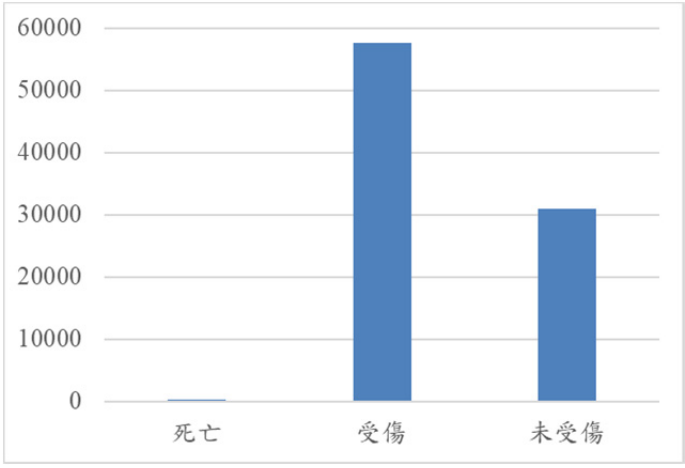


圖 9　肇事比例分布圖

由長條圖可以看出三種類型的嚴重程度數量差距非常大，死亡的數量遠低於其他兩種，若直接將未處理的樣本放入模型中進行訓練，出來的結果很可能會是可以很好的分辨受傷跟未受傷的事故，但沒辦法分辨死亡的事故，因為機器學習的目標都是讓損失函數最小化，所以若沒有進行處理，機器一定會往誤差最小的方向進行收斂，因此我們有必要對於數據做一些處理，使得模型可以平衡的學習到三種類別特徵。

4.3.1 欠採樣 (Under Sampling)

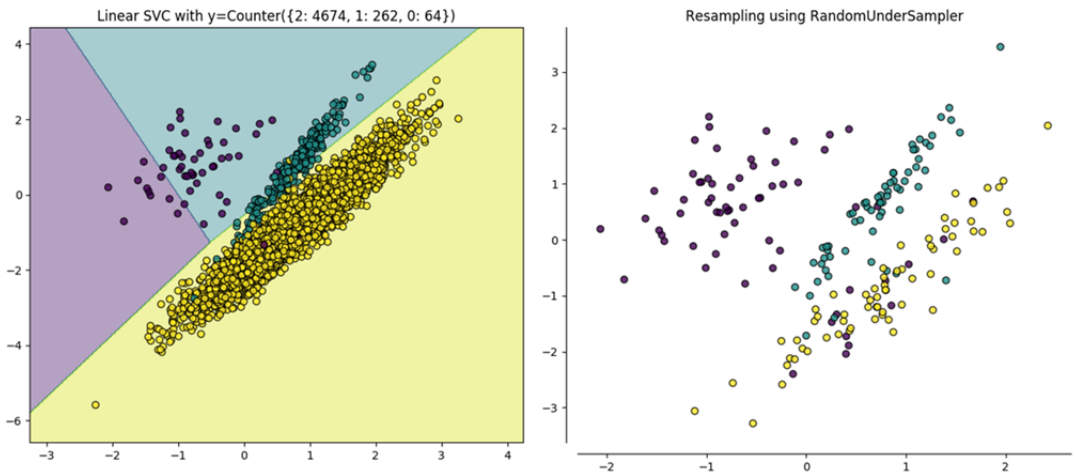
在數據欠採樣 (Under Sampling) 的部分，採用 Leisch [28] 所提出的 Cluster Centroid (集群分析中心法)，在多數欠採樣的方法中，大多數的主要目標都是將資料中不重要的樣本刪除。在 Cluster Centroid 中，使用特徵空間幾何上的聚類概念將樣本劃分為重要和不重要的對象。

聚類 (Cluster) 是一種無監督的學習方法，透過在特徵空間中屬於多數類的數據點上獲得所有特徵的平均特徵向量，可以找到集群的中心。找到多數類的集群中心後，距離中心最遠的樣本則會被認為是最不重要的樣本，越接近特徵空間中心的樣本則是越重要的樣本，此方法是根據不平衡的比例，刪除不重要的樣本，優點是可以在不影響資料樣態的情況下刪除大多數不重要的樣本，使得兩邊可以平衡。

舉個例子來說明，假設一個資料庫中有 50 個陽性樣本，2000 個陰性樣本，則在這種情況下使用 Cluster Centroid 的採樣技術，可以形成 50 個群集，並利用這 50 集群來產生替

換 2000 個陰性樣本，以這 50 個群集的中心為剩下的樣本，刪除其他相對不重要的樣本，使得陰性陽性的數據可以平衡。

下面以圖 10 來視覺化資料，可以更清楚的理解其核心運作模式。



圖片來源：Lemaître et al. (2017)

圖 10　欠採樣示意圖 [29]

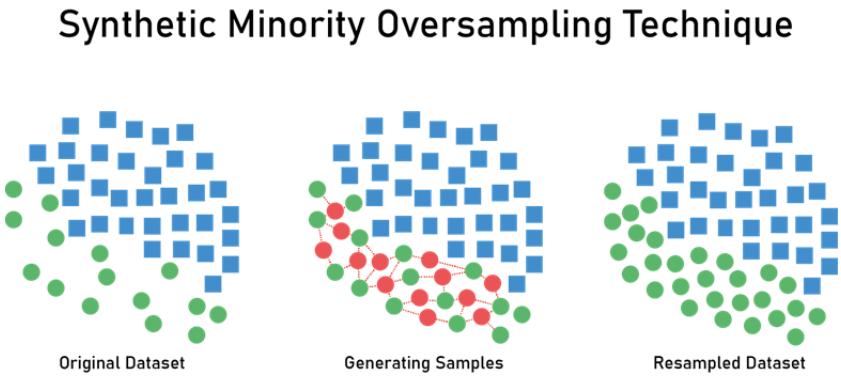
圖 10 左邊為尚未處理過的資料，可以發現其黃色的點非常多，遠遠大於紫色以及綠色的資料，而右邊是處理過的資料，黃色的點雖然被刪除很多，但大多還是能看出其資料原有的樣態，就是 Cluster Centroid 的基本概念，本研究將使用 imblearn 函式庫 [29] 裡面的套件來進行欠採樣的資料處理。

4.3.2 過採樣 (Over Sampling)

在數據過採樣 (Over Sampling) 的部分，將使用合成少數類過採樣技術 (SMOTE, Synthetic Minority Oversampling Technique)，是 Chawla et al. [30] 所提出基於隨機過採樣 (Random Over Sampling) 演算法的一種改進方法，由於隨機過取樣採取簡單複製樣本的策略來增加少數類樣本，這樣容易產生模型過擬合 (Over Fitting) 的問題，即使得模型學習到的資訊過於解釋目前的樣本，並沒辦法有太好的預測能力，以迴歸曲線中的 R2 來做說明，若是我們建構出一條 R2=0.99 的迴歸線，有可能只能解釋目前的數據，如果放入新的資料，此模型可能就沒辦法順利地預測出我們要的數值，就是所謂的過度擬合，所以本研究不採用隨機過採樣的方法，避免這種狀況發生。

而 SMOTE 演算法的基本思想是對少數類樣本進行分析並根據少數類樣本人工合成新樣本新增到資料庫中，簡單來說，就是選取相鄰的點，然後再點跟點中間做內插，內插得到的點，就當作新的樣本，用這樣來增加樣本數，本研究將使用 imblearn 函式庫 [29] 裡面 Over Sampling 的套件—SMOTE，來對數據進行不平衡的處理。

再來以圖 11 來做說明。



圖片來源：Charfaoui (2019)

圖 11　SMOTE 示意圖 [31]

根據圖 11 所示，最左邊為原始的資料樣態，而 SMOTE 為藉由計算少量樣本間的歐式距離 (又稱歐幾里得距離，計算樣本間的距離，為畢氏定理的一種推廣)，並在少量樣本之間合成新的人工樣本，來使得資料可以平衡，中間以及右邊的圖為進行合成樣本的狀況，可以利用這種方法來產生新的資料，讓模型更能有效的去學習各種類型的特徵。

五、結果與分析

5.1　混淆矩陣

將測試集的資料放入校估好的模型當中，並觀察其預測能力為何，預測的結果利用混淆矩陣來視覺化，原始合併後的資料共有 88708 筆，其中有 230 筆死亡、57556 筆受傷、30922 筆無受傷，藉由過採樣處理後，總筆數有 172668 筆，三種受傷程度皆為 57556 筆，切分 75%訓練以及 25%測試，最後總共會使用 129501 筆訓練、43167 筆測試，最後會得到 43167 筆預測結果，如圖 12 所示。

從矩陣可以很明顯看出，兩個模型對於過採樣的資料都有很不錯的分類效果，皆能有效地分辨出各事故的受傷嚴重程度，惟在分類未受傷時表現較差，在兩個模型中都有兩千多筆資料會辨別為受傷，但整體表現上皆可接受，而對於事故中最嚴重的受傷嚴重程度-死亡，都可以預測得很準確。

再來是欠採樣的部分，利用 Cluster Centroid 的欠採樣處理後，將多數的樣本刪除，總共剩下 690 筆資料，其中死亡、受傷、未受傷皆為 230 筆，並切割為 75%訓練集以及測試集，最以 517 筆作為訓練以及交叉驗證，173 筆為測試資料，其預測結果如圖 13。

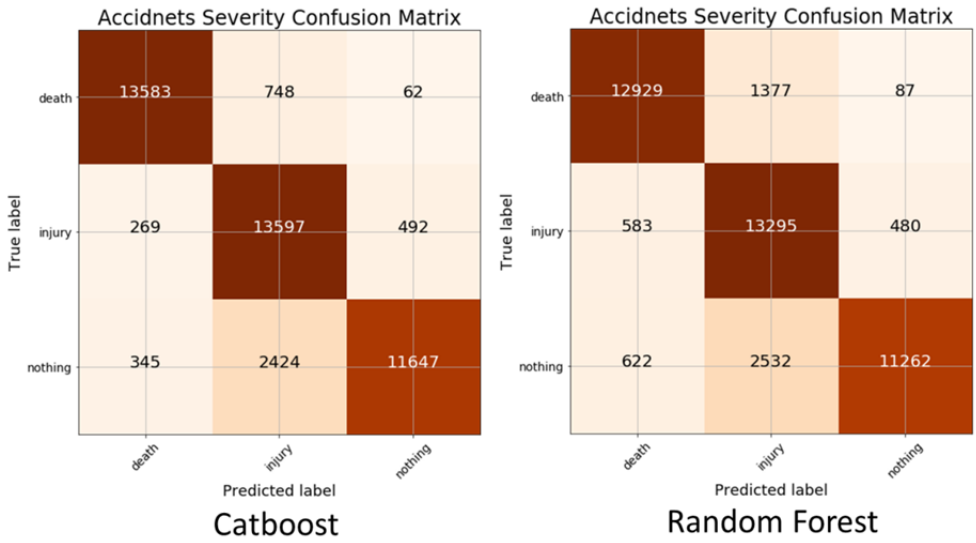


圖 12　過採樣之混淆矩陣

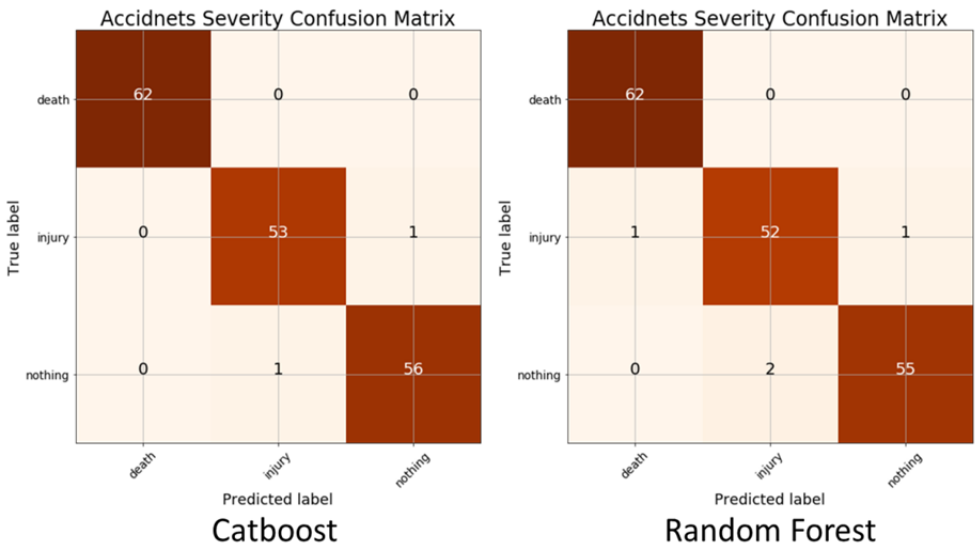


圖 13　欠採樣之混淆矩陣

模糊矩陣中顯示，兩種不同的集成模型在欠採樣的資料預處理下，皆有不錯的分類效果，對於死亡的種類，分類非常準確，幾乎可以全部預測正確，表示藉由訓練的過程中，模型能學習到這三種類別的各項特徵，並判斷其事故的嚴重性。

以各種評分標準，去衡量模型的表現，如表 3。

表 3　各項分數總表

		Accuracy	Precision	Recall	F-measure
Over Sampling	Catboost	89.95%	90.74%	89.95%	89.97%
	Random Forest	86.84%	87.99%	86.85%	86.9%
Under Sampling	Catboost	98.84%	98.8%	98.84%	98.84%
	Random Forest	97.69%	97.64%	97.6%	97.6%

因本研究為多元分類問題，因此在計算多元分類分數時，其計算的方式為 3.6 節之延伸，大略可分為巨觀(Macro)、微觀(Micro)這兩種計算方式，其中 Accuracy 的計算為微觀的計算，亦即對於每個格子不分類別進行統計與計算，以對角線 (預測正確的) 為分子，以所有格子的數值加總為分母，來計算整體的準確度，與第三節所介紹之二元分類的計算相同。

其餘三個分數則採用巨觀的計算模式，亦即將各類別的分類準確度加總起來取其平均，以過採樣中 Catboost 的 Precision 來做說明，Precision 的意涵為在所有預測值中，正確預測的比例，以死亡作為舉例，13583/(13583+269+34) 可得 0.95675，為死亡類別的精準度，同樣的方式類推受傷以及未受傷的計算，分別可得 0.81084 和 0.95459，最後將此三種分數加總並取其平均，則為最後表 3 所呈現的 90.74%，是基於類別的一個平均水準。

以此方法類推計算每個嚴重程度的 Precision、Recall、F-measure，最後取一個平均值，可以去衡量整體的表現，但容易受到小樣本的影響，而導致分數的變化，但因本研究已經對於每個類別進行不平衡數據的處理，因此可採用這種計算方式，來去評估模型在預測各類別時，平均的預測能力。

5.2 模型特徵顯著程度

兩種模型不管是在過採樣或是欠採樣的部分皆能準確分類，在參數設定以及模型建立後，可以產生特徵顯著程度 (Feature importance)，去觀察在模型中，哪些變數對於模型來說是相對重要的，但要特別注意在這邊的顯著程度是指變數對於模型的影響程度，並不是代表變數對於結果的相對關係，能較容易了解哪些特徵是模型比較關注的，圖 14 為過採樣的資料下，兩個模型的特徵顯著程度。

藉由特徵顯著程度，並依照其顯著高低進行排序，可以很明確的知道哪些變數對於模型是比較重要的，哪些則是模型比較不考慮的，根據圖 14，車種對於模型的影響程度最大，也就是說模型大多是透過車種的不同來判斷事故的嚴重性，Catboost 其次是用年齡來去判斷事故的嚴重程度，而 Random Forest 排序第二名的特徵為發生事故鄉鎮名稱，也就是會透過都市與鄉村來判別事故的傷亡程度。而年齡和性別同樣都是位於兩個模型的前四名重

要特徵當中，表示除了車種的不同外，模型還會根據性別和年齡去判斷事故的嚴重性，但此特徵重要程度僅能代表變數在模型當中的重要程度，並無法去解釋何種車種或是年齡發生事故會較易受傷或死亡。

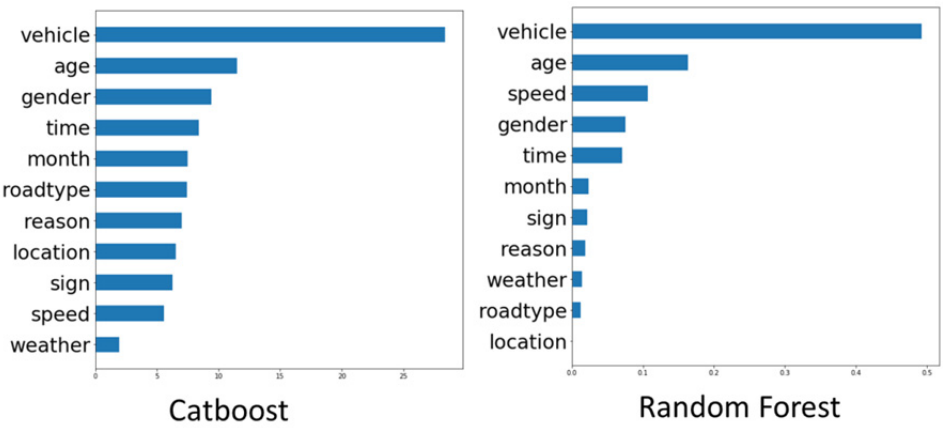


圖 14　過採樣之模型特徵顯著程度

再來是欠採樣的特徵顯著程度，如圖 15。

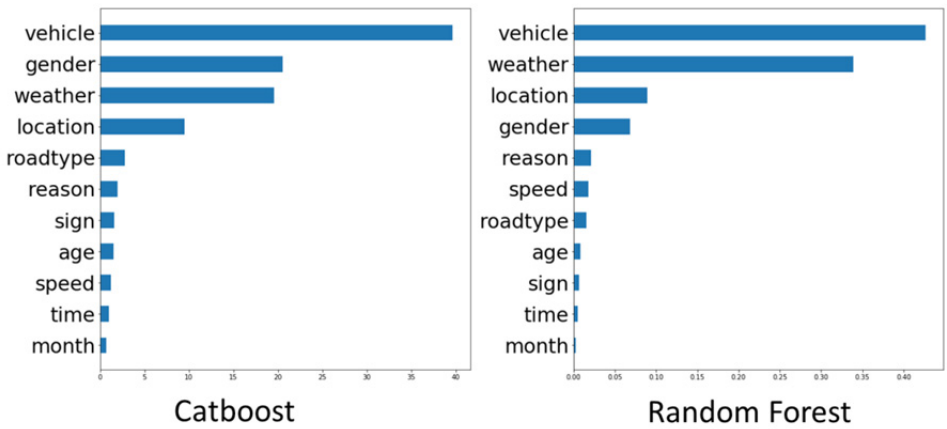


圖 15　欠採樣之模型特徵顯著程度

由圖 15 欠採樣之特徵顯著程度，並對照過採樣之顯著程度，可以發現不管是哪種資料處理方法，對於模型來說，車種的類別都是最重要的特徵，藉由判斷車種的不同來預測事故的嚴重程度，是很合理的，因為駕駛汽車，是以金屬包覆著人，而駕駛機車時，人是裸露於車外的，因此模型在預測的過程中，將駕駛的車輛種類選為最重要的特徵。

再來統整四個模型前五大顯著特徵，如表 4。

表 4　模型顯著特徵統計表

資料處理	分類模型	1	2	3	4	5
Over Sampling	Catboost	車種	年齡	性別	發生時間	發生月份
	Random Forest	車種	年齡	車速	性別	發生時間
Under Sampling	Catboost	車種	性別	天氣	地點	道路種類
	Random Forest	車種	天氣	地點	性別	肇事原因

表 4 中的數字表示特徵重要排序，1 為最重要的特徵，2 為次重要的特徵，依此類推。可以看出對於所有的模型來說，不論分類為何種方法，車種都是最重要依據。在過採樣的模型當中，年齡、性別、發生時間皆是重要的特徵；在欠採樣中，性別、天氣、地點則為比較重要的特徵。在不同模型中，其重要的特徵若是相似，表示這一套預測事故嚴重性的流程是可靠的，並不會因為模型不同而有差異，不過由於採樣的方式不同，重要的特徵會有些許的差異。

天氣特徵，在過採樣中相對是較不顯著的特徵，在欠採樣當中卻是非常顯著的特徵。可能的原因為欠採樣雖然能將多數的樣本降低，並保持原有的資料樣態，但由於死亡事故與受傷事故的數量差距非常之大，在刪除的過程中，可能會導致刪除過多的資料，造成模型在學習時，只能藉由少數的樣本去判斷事故嚴重性，因此其顯著的特徵才會與過採樣有所不同。為了病免可能問題，未來可以採用混和採樣，結合兩種不同的採樣方式來訓練模型，可能會有更好的預測表現。

本研究之模型結果不論是過採樣或是欠採樣，車種的類別都是模型最關注的變數，也就是模型在分類嚴重性時，絕大部份都是依據車種的不同來區別，再來會依據採樣方式的不同而有異。

六、結論與建議

交通事故所造成的傷亡勢必會越來越嚴重，不過由於科技的日新月異，自駕車以及車聯網的技術已經逐漸成熟，將來實際上路後，也許能夠對於交通安全的問題有所突破，若是能夠提出相關策略或是演算法，來改進自駕車的對於事故的判斷以及預防，將會大大提升用路人的安全，本研究則對於預測事故的嚴重程度進行探討。

利用臺南市的公開資料庫做為資料來源，並考慮到實際的資料中數據不平衡的問題，因在現實中，死亡的事故一定會比受傷或是沒受傷來的少很多，倘若直接將這些原始數據拿去訓練模型，模型一定會朝著最少誤差的方式進行收斂，最後並沒辦法有效的分類出死亡事故，本研究根據過採樣以及欠採樣這兩種採樣的概念，並分別採用 SMOTE 以及 Cluster

Centroid 兩種數據預處理的方法，對於原始的數據進行重新採樣，使每種傷亡程度的數據得以平衡，再放入模型中訓練才能使其真正學習到各種嚴重程度的特徵。

本研究所使用的模型為基於兩種集成算法 (Ensemble) 的機器學習模型，分別為 Random Forest 和 Catboost，前者為 Bagging 的延伸，後者則為 Boosting 的延伸，但同樣都是以建立很多個弱分類器來合成一個強分類器，使得預測能力可以比傳統的統計模型來的好很多，其分類結果以及整個流程未來可以結合自駕車或導航系統建立相關資料庫，幫助駕駛員或是電腦在高風險的情況下提早採取迴避的措施，減少事故的嚴重程度。

而 Catboost 演算法在過採樣以及欠採樣中，其模型表現分數都比 Random Forest 來的好，在過採樣中各項分數都超過隨機森林大約 3%，在欠採樣中大約超過 1%左右，但這細小的分數差異，其應用於實務上卻有非常大的影響，因實際上死亡事故的比例佔整體事故的百分比大多都低於 1%，要能夠精準預測，可能還必須更進一步的改善模型。

根據模型特徵顯著程度去解釋各個變數，可藉由此來研擬相關改善措施。本研究之模型結果顯示車種的類別都是模型最關注的變數。在過採樣的模型底下，年齡是第二重要的影響因子，在欠採樣的模型下，天氣是次重要的影響因子。

而本研究的額外貢獻為將來可以利用預測交通事故的嚴重程度，來給予保險公司一些相關的證據來確定客戶的保費並進行一些分析，包含客戶可能所發生的車禍成本，或是提供醫院對於交通事故的急診處理相關資訊，預測事故之嚴重性並安排合適人力進行搶救。

本研究採用兩種不同的採樣方法以及機器學習的模型，對於交通事故嚴重性的預測得到 85%以上精準的結果，未來可以考慮的發展包括：

(1) 採用不同的採樣方法對於資料進行預處理，像是結合過採樣以及欠採樣的混合採樣，可以同時對於多數類別樣本進行欠採樣，對於少數類別樣本進行過採樣。

(2) 應用其他的分類模型，像是利用 Stacking 的概念結合多重的分類模型並建立雙層的預測模型，或是採用深度學習的神經網路去進行模型的訓練。

(3) 結合動態的特徵去進行預測，可以利用車上的行車紀錄器，並進行影像辨識，取得前車距離、轉向、偏移程度等等資料，更能符合實際的交通狀況。

參考文獻

1. World Health Organization, *Global status report on road safety 2018: Summary*, 2018.
2. WHO, “he top 10 causes of death”, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2018.
3. Idrissi, Y. S., “An introduction to Machine Learning”, https://medium.com/ datadriveninvestor/ an-introduction-to-machine-learning-60decab24a2a, 2018.
4. Zheng, Z., Wang, Z., Zhu, L., & Jiang, H., “Determinants of The Congestion Caused by a Traffic Accident in Urban Road Networks”, *Accident Analysis and Prevention*, Vol.136,

105327, 2020.

5. Naqvi, N. K., Quddus, M. A., & Enoch, M. P., “Do Higher Fuel Prices Help Reduce Road Traffic Accidents?”, *Accident Analysis and Prevention*, Vol.135, 105353, 2020.
6. Jones, P., “The Evolution of Urban Mobility: The Interplay of Academic and Policy Perspectives”, *IATSS Research*, Vol.38, Iss.1, 2014, pp. 7-13.
7. Zhang, G., Yau, K. K., Zhang, X., & Li, Y., “Traffic Accidents Involving Fatigue Driving and Their Extent of Casualties.” , *Accident Analysis and Prevention*, Vol.87, 2016, pp. 34-42.
8. Nakai, H., & Usui, S., “How Do User Experiences with Different Transport Modes Affect the Risk of Traffic Accidents? From the Viewpoint of Licence Possession Status”, *Accident Analysis and Prevention*, 99(Pt A), 2017, pp. 242-248.
9. Fan, Z., Liu, C., Cai, D., & Yue, S., “Research on Black Spot Identification of Safety in Urban Traffic Accidents Based on Machine Learning Method”, *Safety Science*, Vol.118, 2019, pp. 607-616.
10. Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S., “A Hierarchical Machine Learning Classification Approach for Secondary Task Identification from Observed Driving Behavior Data”, *Accident Analysis and Prevention*, Vol.123, 2019, pp. 274-281.
11. Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K., “Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis”, *Accident Analysis & Prevention*, Vol.136, 15405, 2020.
12. Zhang, Z., He, Q., Gao, J., & Ni, M., “A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data”, *Transportation Research Part C: Emerging Technologies*, Vol.86, 2018, pp. 580-596.
13. Yao, J., & Ye, Y., “The Effect of Image Recognition Traffic Prediction Method Under Deep Learning and Naive Bayes Algorithm on Freeway Traffic Safety”, *Image and Vision Computing*, Vol.103, 2020.
14. Iranitalab, A., & Khattak, A., “Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction”, *Accident Analysis and Prevention*, Vol.108, 2017, pp. 27-36.
15. Kwon, O. H., Rhee, W., & Yoon, Y., “Application of Classification Algorithms for Analysis of Road Safety Risk Factor Dependencies”, *Accident Analysis and Prevention*, Vol.75, 2015, pp.1-15.
16. Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., & Guan, H., “A Multinomial Logit Model-Bayesian Network Hybrid Approach for Driver Injury Severity Analyses in Rear-End Crashes.”, *Accident Analysis and Prevention*, Vol.80, 2015, pp.76-88.
17. Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N., “Classification of Motor Vehicle Crash Injury Severity: A Hybrid Approach for Imbalanced Data”, *Accident Analysis and Prevention* , Vol.120, 2018, pp.250-261.
18. AlKheder, S., AlRukaibi, F., & Aiash, A., “Risk Analysis of Traffic Accidents' Severities: An Application of Three Data Mining Models”, *ISA Transactions*, Vol.106, 2020, pp.213-220.

19. Tommy, H., “Ensemble learning:Bagging, Boosting, and AdaBoost”, https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ensemble-learning%E4%B9%8Bbagging-boosting%E5%92%8Cadaboost-af031229ebc3, 2018.
20. Breiman, L., “Random forests”, *Machine Learning*, Vol.45, No.1, 2001, pp. 5-32.
21. Abilash, R., “Applying Random Forest (Classification) — Machine Learning Algorithm From Scratch With Real Datasets”, https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57, 2018.
22. Catboost., “CatBoost Enables Fast Gradient Boosting on Decision Trees Using GPUs”, https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus, 2018.
23. Dorogush, A. V., Ershov, V., & Gulin, A., “CatBoost: Gradient Boosting with Categorical Features Support”, arXiv preprint arXiv:1810.11363., 2018.
24. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A., “CatBoost: Unbiased Boosting with Categorical Features”, NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems,pp.6639–6649.
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D. , Brucher, M., Perrot, M., and Duchesnay, É., “ Scikit-learn: Machine learning in Python”, *The Journal of Machine Learning Research*, Vol.12, No.85, 2011, pp. 2825-2830.
26. 臺南市政府，「Data.Tainan 臺南市政府資料開放平台」，https://data.tainan.gov.tw/，民國 110 年。
27. Al-Ghamdi, A. S., “Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity”, *Accident Analysis & Prevention*, Vol.34, Iss.6, 2002, pp. 729-741.
28. Leisch, F., “A Toolbox For -Centroids Cluster Analysis.”, *Computational Statistics & Data Analysis*, Vol.51, No.2, 2006, pp. 526-544.
29. Lemaître, G., Nogueira, F., & Aridas, C. K., “Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”, *Journal of Machine Learning Research*, Vol.18, Iss.1, 2017, pp.559-563.
30. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., “SMOTE: Synthetic Minority Over-Sampling Technique”, *Journal of Artificial Intelligence Research*, Vol.16, 2002, pp. 321-357.
31. Charfaoui, Y., “Resampling to Properly Handle Imbalanced Datasets in Machine Learning.”, https://heartbeat.fritz.ai/resampling-to-properly-handle-imbalanced-datasets-in-machine-learning-64d82c16ceaa, 2019.