**運用監督式學習建立道路交通事故之預測模型-以桃園市為例**

**Use supervised learning to analyze and forecast factors that cause road traffic accidents-Take Taoyuan City as an example**

**邱守燦、黃榛涵、包家豪、柯孟岑**

**Shou-Can Qiu, Chen-Han Huang, Chia-Hao Pao, Meng-Tsen Ke**

國立雲林科技大學 工業工程與管理所

National Yunlin University of Science and Technology

Department of industrial engineering and management

[m11021007@gemail.yuntech.edu.tw](mailto:m11021007@gemail.yuntech.edu.tw)

[m11021045@gemail.yuntech.edu.tw](mailto:m11021045@gemail.yuntech.edu.tw)

[m11021052@gemail.yuntech.edu.tw](mailto:m11021052@gemail.yuntech.edu.tw)

[m10921055@gemail.yuntech.edu.tw](mailto:m10921055@gemail.yuntech.edu.tw)

**運用監督式學習建立道路交通事故之預測模型-以桃園市為例**

**Use supervised learning to analyze and forecast factors that cause road traffic accidents-Take Taoyuan City as an example**

**摘要**

道路交通事故直接造成生命或財產損失，同時衍生相當大的社會成本與代價。本研究採取政府資料開放平台所提供桃園市於民國106年1月1日至109年12月31日的交通事故紀錄資料，以273,084件交通事故作為研究件數。基於發生交通事故的當下，大多以警察單位的交通事故分析初判表為主的形式，並無公開標準及考量當時車禍之天氣、光線等道路狀況因素；因此，本研究透過Scikit-Learn開源軟體，分別使用決策樹、隨機森林、支援向量機(SVM)與極限梯度提升(XGBoost)，建立監督式學習演算法的模型，並依照個別模型的準確率與績效來探討不同模型對於交通事故的分析成效，並找出最適合的預測結果，再進一步分析找出交通事故的主要因素。使最終結果能協助相關單位作為道路規劃的參考依據，並預先做好應對措施，減輕交通事故所造成的後果。

關鍵詞：交通事故、決策樹、隨機森林、支援向量機、XGBoost

**Abstract**

**​​**Road traffic accidents cause loss of life or property directly, and considerable social costs and prices at the same time. This study adopts the traffic accident record data of Taoyuan City from January 1,2017 to December 31, 2020 provided by the government information open platform, and takes 273,084 traffic accidents as the number of studies. Based on the moment of the traffic accidents, most of them are mainly in the form of the preliminary judgment of the traffic accident analysis of the police. There are no public standards and consideration of the weather, light and other road conditions of the traffic accident. Therefore, the research uses the Scikit-Learn open source software, and using Decision Tree, Random Forest, Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost) respectively to establish models of supervised learning algorithms, and explore the analysis of traffic accidents by different models according to the accuracy and performance of individual models and find out the most suitable forecast results. And then further analyze and find out the main factors of traffic accidents. So that the final result can assist relevant units as a reference for road planning, and prepare countermeasures in advance to reduce the consequences of traffic accidents.

Keywords: Traffic Accident, Decision Tree, Random Forest, Support Vector Machine(SVM), eXtreme Gradient Boosting( XGBoost)

**​​**

**1、緒論**

**1.1研究背景**

交通事故不僅會造成人員的傷亡及財產損失，還會帶來交通壅堵，甚者交通中斷，事故發生的原因，會受到駕駛人行為、路面狀況及氣候等各種因素影響。根據「道路交通事故處理辦法」第十條，交通事故發生當警察單位抵達現場後，會先進行現場勘查、蒐證等動作，然後依肇事因素索引表中的項目去判定肇事主因，接著製作道路交通事故分析初判表，但目前警政單位所公布之肇因索引表僅列舉駕駛人、行人、機件及交通設施等九大類別，雖有紀錄肇事現場天候、光線等相關資訊，但在肇因考量方面，卻未將其納入考量，且此僅為警察單位片面的初步研判意見，而未更客觀忠實的呈現結果。

根據警政署統計之交通事故件數概況108年交通事故件數為341,972件，109年為362,393件，年增長率6%，109年桃園市交通事故件數為49,311件，為全台交通事故率六都之冠，因此本研究以桃園市作為主要的研究對象，並針對其公路系統中省道、縣道、鄉道及市區道路，不包含非屬汽車專用道及國道之其它道路。

**1.2研究動機與目的**

為了針對現有交通事故肇事要因分析時，仍就以警察單位的交通事故分析初判表為主的形式，本研究利用桃園市政府警察局所提供106年度至109年度的交通事故資料表，運用監督式學習的決策樹、隨機森林、支援向量機、XGBoost四種演算法建立預測模型，找出交通事故的關鍵肇事因素，並比較不同模型的準確率，同時探討不同模型對於交通事故資料的分析成效，進而找出最適合分析交通數據的模型，以供未來研究者做為參考，本研究結果可協助道路安全分析師及交通單位管理人員進行比較及確定問題，以做為未來政府道路規劃的參考；期望能提前針對發生交通事故的原因加強防範，做好相對應的措施及時預防或減輕事故發生造成的嚴重後果。

1. **文獻探討**

**2.1 Scikit-Learn**

Scikit-Learn又稱為Sklearn，以Python程式語言與其他延伸套件(如:Pandas、Numpy)當作基礎[17]，為實際操作機器學習的開源軟體，許多機器學習相關的演算法由Scikit-Learn實現。本論文使用到的研究方法為監督式學習的演算法，包括:隨機森林(Random Forest)、決策樹(Decision Tree)、支援向量機(Support Vector Machine,SVM)、XGBoost，這些演算法為主要的工具進行分析與研究。

**2.2決策樹(Decision Tree，DT)**

決策樹也稱為DT是一種用於分類和預測的建模方法，從大量資料提取多變相的關鍵因素，並找尋因素間關聯性，再依決策規則向下分支節點[3]，從每個節點向下延伸去找尋可能的答案，而每個葉節點代表對所考慮問題的解決方案的預測，以此來模擬我們選擇的過程，並且最終依照選擇的分支導引到不同的結果[5]。Moral-Garcia學者將決策樹運用於分析西班牙市區新手駕駛在不同交通號誌及環境下其發生交通事故的事故原因，結果顯示當事故類型為與行人發生碰撞，其肇事原因是超速行駛所導致[19]；美國馬里蘭州公路安全辦公室收集2015年至2019年的車輛碰撞數據，藉由建構能夠預測新道路碰撞程度的分類，來降低碰撞的嚴重程度或降低事故的發生頻率。[7]； 針對加州1973年至2014年間的交通事故資料進行，透過決策樹在有限資源下有效地採取對策，分析結果顯示以下五個為重要類型：碰撞類型、違規類別、在交叉路口、洲際公路[18]；以上學者皆運用決策樹建構模型並預測，以解決交通事故相關問題[2]。

**2.3隨機森林(Random Forest，RF)**

隨機森林(Random Forest)也稱為RT，用於分類跟迴歸的分類方法，模型原理是將眾多的決策樹透過隨機分配的方式，合併為所有決策的集合。隨機森林的樣本間互相獨立，訓練方式為裝袋法(bagging)，將三分之二的資料分袋，稱為袋內樣本，用於訓練決策樹，剩餘樣本被稱為袋外樣本，視為驗證集，用於交叉驗證訓練集的準確率。隨機森林可用於解決不平衡的交通違規資料，進而預測裝設自動監測系統的交叉路口其交通違規率[9]；更用於針對交通事故駕駛所面臨的傷害程度進行特徵分析，分別從駕駛員、車輛、環境、其他，例如：車齡、貸款條件等四個面向中，找出其內含項目裡的重要因素，作為特徵進行後續分析[23]；此方法更用於針對鐵路平交道的安全進行模擬預測，來分配安全機構的相關資源[22]；[Yunxiang Liu](https://ieeexplore-ieee-org.libdb.yuntech.edu.tw:3001/author/37086284647)等學者利用隨機森林針對道路交通堵塞建立模型進行預測；透過汽車間的追撞事故，根據駕駛員、車輛、環境等相關因素預測領先車輛的反應，利用隨機森林結果顯示，對前方車輛駕駛員反應影響最大的因素[16]。

**2.4支援向量機(Support Vector Machine , SVM)**

支援向量機(SVM)主要解決小量樣本、線性函數、高維度資料等類型的資料[4]。支援向量機可分為線性支援向量機與非線性支援向量機，本研究的支援向量機為線性支援向量機，線性支援向量機的核心觀念為:各個分類之間的差距需達到最大值，在進行訓練的資料中，若有一個超平面可將資料分為兩類，將超平面做區分與分類，並要求符合線性支援向量機的定義。支援向量機曾應用在交通號誌的辨識與交通事故研判的研究當中[2,15,21]，Fu等學者利用其針對擁擠交通環境進行分層車輛辨識系統[8]，以上方法皆利用支援向量機進行號誌或道路影像偵測，完成分類並建構辨識系統。汽車碰撞程度模型的建構分析，是交通安全的重要研究議題，透過SVM預測可過預測，Wang學者等人則將其運用於預測上海中環高速公路的碰撞情向，結果顯示穩定的駕駛速度以及減少突然加速的駕駛行為有助於減少碰撞[20]。

**2.5極限梯度提升(eXtreme Gradient Boosting , XGBoost)**

XGboost 為 eXtreme Gradient Boosting的簡寫，被稱為極限梯度提升決策樹，利用集成學習(ensemble learning)找到最佳解的一種方法，集成學習是由多個較弱的學習器結合而成的學習法，集成學習可分為Bagging和Boosting[1]。經由XGBoost可以有效地建構決策樹，透過並行的方式解決分類和回歸問題[11]，XGBoost在面對回歸問題時，會一直增加新的回歸樹模型，接著再透過新生成的樹對應之前的每一棵樹比較殘差，將結果作為預測值[12]。XGBoost可針對交通壅塞模型預測進行預測[14]，XGBoost也應用於短期交通流量的預測當中，相較於支援向量機(SVM)的預測結果，XGBoost的預測精確度與運算時間皆相對於SVM更有效率[6]。XGBoost更被應用於檢測及分析芝加哥的高速公路交通事故數據，並發現離市區越近越容易造成事故發生，同時天氣及車道等也會影響事故，結果更顯示XGBoost的預測系統準率高並具有穩定度[13]。在洛杉磯交通事故的死亡率與肇事特徵之間的關係中，發現XGboost相較SVM和隨機森林等其他模型預測的更準確[1]。

**3、研究方法**

本研究根據政府資料開放平台中桃園市政府警察局所提供的，民國106/1/1-109/12/31的交通事故資料表進行分析，4年度資料筆數共計273,084筆[2]。將分別運用決策樹(Decision Tree)、隨機森林(Random Forest)、支援向量機(Support Vector Machine)、XGBoost建立預測模型。

**3.1資料清理**

在資料分析前將原始資料進行以下的前處理過程，以提升後續資料分析時，資料的可用性及一致性。

**3.1.1資料編碼及消除**

本研究使用labelEncoder將資料進行編碼，編碼方式主要沿用警政署的事故欄位。並將資料中日期、時間、年度、傷亡人數、國籍、性別、年齡、速限及牌照種類等與研判肇事因素無關的項目，進行消除。人員的傷亡狀況，例如人員當場死亡、24小時以內或超過24小時死亡的資料，因其同樣不與肇因相關，進行消除。

本研究目的為協助相關管理單位作為未來道路規劃的考量，是以在分析階段不將道路類別中的國道納入考量，因國道的道路規劃為交通部高速公路局負責，並非各縣市主管機關負責，且國道工程的改善及規劃計畫，期程至少5年，非短期的規劃計畫，且本組僅針對行駛於道路上的汽車作為研究對象，所以消除當事者類別中的其他項目。

**3.1.2資料整合**

將原始資料經過編碼、清理、確定研究目標的特徵值等前置處理，依循研究目的整理出主要的交通事故要因包括：天候、光線、當事者、道路型態、事故位置、路面狀況、交通號誌、車道環境、事故型態、肇因研判，共計10項的資料欄位整合為一份資料集，之後為模型的建立與訓練做準備。

**3.1.3資料轉換**

完成資料整合的資料集後，使用Z-Score標準化進行資料數據轉換，並將各資料索引的值域縮小，其目的在於減少資料分析產生的偏誤；經過資料轉換後的資料集，將在後續的模型建立與訓練中使用到。

**4、研究結果**

完成資料清理等資料處理後，在訓練和測試模型時，使用273,084筆資料中的 80% 進行訓練，其餘 20% 用作測試和交叉驗證並且評估模型的性能。本節將討論模型預測的準確率及其分析各模型的績效。

**4.1特徵選取**

透過隨機森林方法針對已經前處理完畢的資料進行驗證(validation)，以此分析交通事故因素，並選取出關鍵特徵因素，作為後續模型建立的特徵，圖1為隨機森林特徵重要分析圖，評估指標y軸對應車道劃分設施，可得知對模型訓練與預測而言，重要的特徵分別為1光線、2當事者區分類別，也就是發生車禍的主要類別例如:大客車、小貨車及小客車等；4為道路型態所對應的項目有交岔路路及平交道等、10與11為車道劃分分向設施大類別以及子類別、12車道環境、6事故位置包含交岔路口與路段及交流道、15事故類型及型態，代表本此交通事故為車與車之間發生的碰撞還是汽機車本身、17與19分別對應肇因研判大類別以及子類別；透過所獲得的特徵分析可得知不只駕駛人行為是肇事原因，環境及道路規劃同為影響肇事的主要特徵。

其中編號11的車道劃分設施分向設施，所代表的是車道設有中央分向島、雙向禁止超車線、單向禁止超車線及行車分向線，此一類別為最主要的特徵，可知車道規劃中分向規劃是影響肇事因素的關鍵。

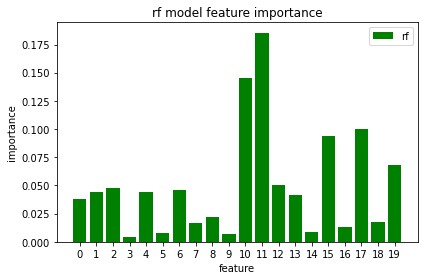


圖1、隨機森林特徵重要分析圖

**4.2模型建立**

本研究將預測模型進行準確率的估計及預測績效的衡量，最大深度MD(Max-Depth)為數據訓練的最大極限層數，並根據最大深度MD(Max-Depth)當作指標，評估不同深度的準確率與預測績效，當超過指定層數的樹將會被剪枝處理不進行訓練，防止分類器產生過度擬合(overfitting)的現象。

**4.3模型預測之準確率**

本研究所使用的四種預測模型之準確率如表1所示，可以發現在各個深度的預測模型中，皆是以XGBoost預測模型的準確率為最高。

表1、各預測模型之準確率

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DT** | **RF** | **XGBoost** | **SVM** |
| **MD=5** | **66.72%** | **64.61%** | **68.31%** | **18.92%** |
| **MD=10** | **69.36%** | **68.95%** | **70.78%** | **23.18%** |
| **MD=15** | **71.30%** | **70.07%** | **72.01%** | **27.55%** |
| **MD=20** | **72.93%** | **73.51%** | **75.41%** | **31.26%** |
| **MD=25** | **73.42%** | **73.52%** | **77.18%** | **33.59%** |
| **MD=30** | **73.54%** | **73.57%** | **80.05%** | **34.74%** |

**4.4 預測績效指標(MAPE、RMSE、MAE)**

**4.4.1 MAPE(平均絕對百分比誤差)**

MAPE(平均絕對百分比誤差)，為預測模型的績效指標，預測模型在預估目標未來值所產生的評估上的誤差，即是實際值和預測值之間差異的平均值，同時以實際值的百分比來表示，當MAPE越小越接近零，代表預測模型的差異越小，並且越接近完美，透過表2可知，整體而言以XGBoost建立的模型為最佳。

表2、各監督學習方法所建立之預測模型的MAPE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DT** | **RF** | **XGBoost** | **SVM** |
| **MD=5** | **15.12%** | **16.33%** | **15.01%** | **41.64%** |
| **MD=10** | **14.78%** | **15.53%** | **14.67%** | **40.98%** |
| **MD=15** | **14.66%** | **14.29%** | **14.16%** | **39.87%** |
| **MD=20** | **14.32%** | **14.21%** | **13.53%** | **36.62%** |
| **MD=25** | **14.11%** | **13.94%** | **13.18%** | **35.49%** |
| **MD=30** | **13.97%** | **13.85%** | **12.92%** | **33.75%** |

**4.4.2 RMSE(均方根誤差)**

RMSE(均方根誤差)，為預測模型的績效指標，用於衡量預測值與實際值之間的差異，當RMSE值越小代表誤差越小，表示預測的模型更佳的精準。透過表3可知，RandomForest在RMSE中比起其他模型擁有更準確的預測。

表3、各監督學習方法所建立之預測模型的RMSE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DT** | **RF** | **XGBoost** | **SVM** |
| **MD=5** | **0.6759** | **0.6029** | **0.6540** | **105.1752** |
| **MD=10** | **0.6566** | **0.6021** | **0.6471** | **103.2536** |
| **MD=15** | **0.6564** | **0.5997** | **0.6394** | **100.6379** |
| **MD=20** | **0.6543** | **0.5840** | **0.6381** | **97.9901** |
| **MD=25** | **0.6536** | **0.5774** | **0.6335** | **96.1848** |
| **MD=30** | **0.6524** | **0.5718** | **0.6319** | **95.2164** |

**4.4.3 MAE(平均絕對誤差)**

MAE(平均絕對誤差)，是一種用於預測模型的損失函數，MAE是目標值和預測值的絕對值總和。與平均誤差相比，平均絕對誤差中由於離差被絕對值化，不會出現正負相抵消的情況，因此，MAE能更好反應預測值誤差的實際狀況。當MAE值越小，說明預測有更好的精準度。透過表4可知，XGBoost在MAE中比起其他模型擁有更精準的預測。

表4、各監督學習方法所建立之預測模型的MAE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DT** | **RF** | **XGBoost** | **SVM** |
| **MD=5** | **0.4274** | **0.4294** | **0.4130** | **6.9353** |
| **MD=10** | **0.4071** | **0.4142** | **0.4018** | **6.3677** |
| **MD=15** | **0.4039** | **0.3938** | **0.3903** | **5.8589** |
| **MD=20** | **0.4018** | **0.3743** | **0.3656** | **5.4924** |
| **MD=25** | **0.4003** | **0.3678** | **0.3624** | **5.3102** |
| **MD=30** | **0.3998** | **0.3662** | **0.3591** | **5.2348** |

**4.5 研究結果**

基於警政署收集的桃園市政府交通事故碰撞數據進行模型訓練與測試，分別輸入至決策樹、隨機森林、XGBoost和SVM四種建模方法，並預測交通事故發生的原因，觀察四種模型的準確率與預測績效，整體而言以XGBoost的訓練結果為最佳的模型，當資料為多屬性資料且數量龐大時，運用此方法模型進行訓練可以克服過度擬合的情形，並得到理想的預測結果，因此可優先選擇XGBoost來進行，當未來研究交通相關數據時，也可用此模型當作參考。

透過本研究使用監督式機器學習方法的特徵選取及肇事因素分析，可得知環境及道路規劃也是重要的考量因素，環境及道路規劃所指的是燈光、車道劃分、車道環境等，其中道路規劃中的車道劃分分向設施為主要特徵，分向設施例如中央分隔島。本研究透過隨機森林分類選取特徵，得出的10項特徵值進行模型的建立並進行預測，如表5所示，由模型的預測結果得知當駕駛人變換車道與行駛方向不當時，事故位置發生在交叉路口內的次數比起一般車道高出許多，且同樣肇事原因而發生於交叉路口附近的車禍數明顯比一般道路來的多，同時發現左轉彎未依規定不當為機車待轉區最主要的肇事因素，在易肇事路段之一的快車道來看，分別可以發現未依規定讓車、未保持行車安全距離、未保持行車安全間隔皆為快車道的主要肇因；基於此模型的預測結果，可以找出肇事主因和需要重點改善的易肇事區域與環境。

表5、預測交通事故次數

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 交叉路口內 | 交叉路口附近 | 機車待轉區 | 快車道 | 坡道 |
| 變換車道或方向不當 | 14886 | 15618 | 8 | 275 | 24 |
| 未依規定讓車 | 2595 | 2109 | 211 | 41563 | 5156 |
| 左轉彎未依規定 | 3643 | 15462 | 16180 | 120 | 82 |
| 未保持行車安全距離 | 310 | 327 | 67 | 36937 | 6193 |
| 未保持行車安全間隔 | 328 | 380 | 71 | 33255 | 6219 |

**5、研究結論**

本研究使用桃園市政府公開的4年交通事故資料集，運用監督式學習的演算法中的決策樹、隨機森林、支援向量機及XGBoost建立預測交通事故的模型，透過模型訓練交叉驗證進行分析，可得知使用XGBoost演算法在預測交通事故資料上為最佳建立模型的方法。透過警政署交通事故檔所建立的道安資訊事故查詢網站，對應109年桃園市政府不分區肇事熱點資訊，可得知在桃園市龍潭區的台3縣、民豐一街、中興路九龍355巷、中興路之交叉路口中，該路口109年總肇事件數為58件，肇事原因為左轉彎未依規定的車禍發生了9件，為該路口的交通事故排行第一名，未保持行車安全距離以及未注意前車狀況也各發生了8件，同為該路口排行第二高的交通事故，以上三個肇事原因肇事件數共為25件，佔該路口109年將近半數的肇事數量，而與本研究所預測的，在交叉路口中以左轉未依規定、未保持安全距離、未注意前車狀況這三者的肇事數量最多，此實際狀況與本研究預測結果相符。

**參考文獻**

[1] 李敏聞, "透過資料融合以梯度提升決策樹方法預測高速公路旅行時間之研究," 碩士, 運輸與物流管理學系, 國立陽明交通大學, 2021. [Online]. Available: <https://hdl.handle.net/11296/84zvve>

[2] 施建合, "車載攝影系統之交通標誌辨識技術開發," 碩士, 電機工程研究所, 國立中正大學, 嘉義縣, 2016. [Online]. Available: <https://hdl.handle.net/11296/24x2k4>

[3] 張俊偉, "決策樹與支撐向量機方法於電力系統暫態穩定度評估," 碩士, 電機工程學系, 國立臺灣海洋大學, 基隆市, 2018. [Online]. Available: <https://hdl.handle.net/11296/v59gn7>

[4] 黃崇豪, "使用支持向量機於二變量製程之非隨機同步管制圖樣式辨識," 碩士, 工業工程與管理系, 國立雲林科技大學, 雲林縣, 2015. [Online]. Available: <https://hdl.handle.net/11296/rfs8zb>

[5] 歐陽萱晏, "使用決策樹建立口罩購買意圖分類分析," 碩士, 經營管理學系, 亞洲大學, 台中市, 2021. [Online]. Available: <https://hdl.handle.net/11296/u4633h>

[6] X. Dong, T. Lei, S. Jin, and Z. Hou, "Short-Term Traffic Flow Prediction Based on XGBoost," in *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, 25-27 May 2018 2018, pp. 854-859, doi: 10.1109/DDCLS.2018.8516114.

[7] S. Elyassami, Y. Hamid, and T. Habuza, "Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees," in *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, 5-12 June 2021 2020, pp. 520-525, doi: 10.1109/CiSt49399.2021.9357298.

[8] H. Fu, H. Ma, Y. Liu, and D. Lu, "A vehicle classification system based on hierarchical multi-SVMs in crowded traffic scenes," *Neurocomputing,* vol. 211, pp. 182-190, 2016/10/26/ 2016, doi: <https://doi.org/10.1016/j.neucom.2015.12.134>.

[9] Y. Li, M. Li, J. Yuan, J. Lu, and M. Abdel-Aty, "Analysis and prediction of intersection traffic violations using automated enforcement system data," *Accident Analysis & Prevention,* vol. 162, p. 106422, 2021/11/01/ 2021, doi: <https://doi.org/10.1016/j.aap.2021.106422>.

[10] J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan, and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," *IEEE Access,* vol. 7, pp. 148059-148072, 2019, doi: 10.1109/ACCESS.2019.2946401.

[11] H. Nguyen, X.-N. Bui, H.-B. Bui, and D. T. Cuong, "Developing an XGBoost model to predict blast-induced peak particle velocity in an open-pit mine: a case study," *Acta Geophysica,* vol. 67, no. 2, pp. 477-490, 2019/04/01 2019, doi: 10.1007/s11600-019-00268-4.

[12] S. Pan, Z. Zheng, Z. Guo, and H. Luo, "An optimized XGBoost method for predicting reservoir porosity using petrophysical logs," *Journal of Petroleum Science and Engineering,* vol. 208, p. 109520, 2022/01/01/ 2022, doi: <https://doi.org/10.1016/j.petrol.2021.109520>.

[13] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis & Prevention,* vol. 136, p. 105405, 2020/03/01/ 2020, doi: <https://doi.org/10.1016/j.aap.2019.105405>.

[14] D. Ran, H. Jiaxin, and H. Yuzhe, "Application of a Combined Model based on K-means++ and XGBoost in Traffic Congestion Prediction," in *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 13-14 June 2020 2020, pp. 413-418, doi: 10.1109/ICSGEA51094.2020.00095.

[15] J. L. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Physica a-Statistical Mechanics and Its Applications,* vol. 517, pp. 29-35, Mar 2019, doi: 10.1016/j.physa.2018.10.060.

[16] Y. Xu, S. Bao, and A. K. Pradhan, "Modeling drivers’ reaction when being tailgated: A Random Forests Method," *Journal of Safety Research,* vol. 78, pp. 28-35, 2021, doi: 10.1016/j.jsr.2021.05.004.

[17] J. G. Hao and T. K. Ho, "Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language," (in English), *J. Educ. Behav. Stat.,* Software Review vol. 44, no. 3, pp. 348-361, Jun 2019, doi: 10.3102/1076998619832248.

[18] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," (in English), *Accid. Anal. Prev.,* Article vol. 75, pp. 1-15, Feb 2015, doi: 10.1016/j.aap.2014.11.005.

[19] S. Moral-Garcia, J. G. Castellano, C. J. Mantas, A. Montella, and J. Abellan, "Decision Tree Ensemble Method for Analyzing Traffic Accidents of Novice Drivers in Urban Areas," (in English), *Entropy,* Article vol. 21, no. 4, p. 15, Apr 2019, Art no. 360, doi: 10.3390/e21040360.

[20] J. H. Wang, T. Y. Luo, and T. Fu, "Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach," (in English), *Accid. Anal. Prev.,* Article vol. 133, p. 12, Dec 2019, Art no. 105320, doi: 10.1016/j.aap.2019.105320.

[21] Z. S. Wang, J. G. Wang, S. Yang, and J. Q. Gao, "Traffic identification and traffic analysis based on support vectormachine," (in English), *Concurr. Comput.-Pract. Exp.,* Article vol. 32, no. 2, p. 8, Jan 2020, Art no. e5292, doi: 10.1002/cpe.5292.

[22] X. Y. Zhou, P. Lu, Z. J. Zheng, D. Tolliver, and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," (in English), *Reliab. Eng. Syst. Saf.,* Article vol. 200, p. 9, Aug 2020, Art no. 106931, doi: 10.1016/j.ress.2020.106931.

[23] G.Pillajo-Quijia, B. Arenas-Ramirez, C. Gonzalez-Fernandez, and F. Aparicio-

Izquierdo, "Influential Factors on Injury Severity for Drivers of Light Trucks and. Vans with Machine Learning Methods," (in English), *SUSTAINABILITY,* vol. 12, no. 4, FEB 2 2020, doi: 10.3390/su12041324.