



Comparative Evaluation of Machine Learning Algorithms for Predicting Victim Road-User Type and Road Type in Fatal Vulnerable Road User Crashes in Tehran, Iran

Fatemeh Salarvandian^{1*} | Zeinab Ebadi² | Rozhin Shahintab³

1. Corresponding Author, Assistant Professor of Geography and Urban Planning, Department of Tourism Management, Faculty of Accounting and Management, Allameh Tabataba'i University, Tehran, Iran. Email: fsalarvandian@atu.ac.ir

2. PhD of Transportation Planning, Faculty of Civil Engineering, Science and Research Branch, Islamic Azad university, Tehran, Iran. Email: zeinabebadi60@gmail.com

3. M.Sc. of Architecture Engineering, Faculty of Architecture and Urban Planning, Qazvin Branch, Islamic Azad university, Qazvin, Iran. Email: rozhinparsa.1981@gmail.com

ARTICLE INFO

Article type:
Research Paper

Article History:

Received 26 June 2026

Revised 27 July 2026

Accepted 22 September 2026

Published Online 01 January 2027

Keywords:

Vulnerable road users,
Fatal crashes,
Tehran,
Machine learning,
Clustering.

ABSTRACT

Traffic crashes are a major public health and urban safety challenge in Tehran, where vulnerable road users account for a substantial share of annual fatalities. Using fatal crash records from 2017 to 2025, this study applied descriptive statistics, the chi-square test, and Cramér's V to assess associations among categorical variables; eight supervised machine learning algorithms to predict deceased road-user type and road type; and K-Means clustering to identify latent patterns. Annual crash counts increased markedly. Pedestrians and motorcyclists constituted a large share of vulnerable road user fatalities, and highways were the most common crash location. Based on consistency across cross-validation folds, XGBoost performed best in predicting deceased road-user type (accuracy = 0.8091). For road type, severe class imbalance was addressed by comparing 14 imbalance-handling methods according to recall for the minority class, secondary streets (7.3% of records). Only an adjusted random forest substantially improved secondary-street recall (from 0 to 0.3333) without a marked loss in overall accuracy (0.4704). Clustering identified four profiles: Cluster 0, characterized by a notable pedestrian presence; Cluster 1, by a higher share of motorcyclists and highway crashes; Cluster 2, by pedestrian predominance and the highest proportion of hit-and-run incidents; and Cluster 3, the largest, by motorcyclist predominance. These differences underscore the need for interventions tailored to each group to improve traffic safety management in Tehran.

Cite this article: Salarvandian, F.; Ebadi, Z. & Shahintab, R. (2027). Comparative Evaluation of Machine Learning Algorithms for Predicting Victim Road-User Type and Road Type in Fatal Vulnerable Road User Crashes in Tehran, Iran. *Urban Development Policy Making*, 3 (4), 709-732. DOI: <http://doi.org/10.22034/judpm.2026.595340.1146>



© Authors retain the copyright and full publishing rights.
DOI: <http://doi.org/10.22034/judpm.2026.595340.1146>

Introduction

Road traffic injuries are a leading cause of death worldwide, disproportionately burdening low- and middle-income countries and vulnerable road users (VRUs) such as pedestrians and motorcyclists. Tehran, Iran's capital, records a substantial share of such fatal crashes daily. While earlier studies of Iranian traffic mortality relied mainly on classical parametric models, machine learning has been increasingly shown better capture complex interactions in crash data. This study analyzes spatial, temporal, and causal patterns of fatal crashes among Tehran's vulnerable road users, evaluates

supervised machine learning algorithms predicting victim type and road type, identifies key predictors, and applies unsupervised clustering to uncover hidden patterns.

Materials and methods

This quantitative, retrospective study analyzed 5,347 officially registered fatal crash records from Tehran spanning 1396-1404 (2017-2025), covering 19 individual, temporal, road/environmental, and vehicle-related variables. After removing records with missing values, 4,803 observations remained and were split into training (3,842) and test (961) sets. Chi-square tests and Cramer's V assessed categorical associations, and Pearson correlation examined numerical features. Eight supervised algorithms (logistic regression, decision tree, random forest, k-nearest neighbors, naive Bayes, gradient boosting, histogram-based gradient boosting, and XGBoost) predicted victim type and road type, evaluated via a combination of consistency across cross-validation folds such as CV Balanced Accuracy, CV F1 Macro, Test Balanced Accuracy, Test F1 Macro accuracy. K-means clustering with principal component analysis detected latent structures in the data.

Results

Pedestrians (42.8%) and motorcyclists (40.6%) together accounted for over 83% of fatalities, whereas car occupants represented only 16.6%. Highways were the predominant crash location (58.8%), followed by main streets (37.5%) and side streets (3.7%). Annual fatalities, stable between 2021 and 2023, rose sharply in 2025 to 786 deaths (+26.4%). Pedestrian deaths were concentrated in evening/nighttime hours, motorcyclist deaths peaked during rush hours; The mean age of the pedestrian population was significantly higher than that of motorcyclists, and men represented 86/8% of fatalities depending on victim type. Failure to observe the road ahead was the leading crash cause, followed by loss of vehicle control and sudden lane changes. Chi-square analysis confirmed significant associations between victim type and road type (Cramer's V = 0.128) and between victim type and time period (V = 0.180), while road type and season showed no significant association (V = 0.012). XGBoost performed best in predicting the type of deceased road user based on a combination of consistency across cross-validation folds (with an accuracy of 0.8091). For road type, given the severe class imbalance, 14 imbalance-handling methods were evaluated on their ability to improve recall for the minority class, Secondary streets (7.3% of the data). Only the adjusted random forest significantly increased secondary -street recall (from 0 to 0.3333) without a substantial loss of overall accuracy (0.4704). This improvement remains modest in absolute terms, so the road-type results should be interpreted with caution. The clustering identified four distinct profiles: Cluster 0, characterized by a notable presence of pedestrians; Cluster 1, by a higher share of motorcyclists and highway crashes; Cluster 2, by pedestrian predominance and the highest proportion of hit-and-run incidents; and Cluster 3 (the largest cluster), by motorcyclist predominance. These differences underscore the need for interventions tailored to the characteristics of each group in managing traffic safety in Tehran.

Discussion and conclusion

The prediction of road type proved considerably more difficult than the prediction of the deceased road-user type, mainly because of the severe class imbalance in the data and results should be interpreted with caution. Overall, combining descriptive statistics, association tests, supervised prediction, and clustering provided a complementary view of fatal crash patterns in Tehran, and the clusters showed that the risk profiles of elderly pedestrians and young motorcyclists differ in meaningful ways. This study has several limitations. The data were retrospective and secondary in nature, incomplete records were excluded, and the findings may not generalize beyond Tehran. The results support prioritizing highway safety measures, given the high share of fatal crashes occurring on highways, and designing interventions tailored to the profile of each cluster, such as pedestrian protection in pedestrian-dominated clusters, measures targeting motorcyclists in motorcyclist-dominated clusters, and stronger post-crash response and enforcement for the cluster with the highest proportion of hit-and-run incidents. Future research should validate these findings using data from other cities and incorporate additional contextual variables.



ارزیابی تطبیقی الگوریتم‌های یادگیری ماشین در پیش‌بینی نوع متوفی و نوع راه در تصادفات فوتی کاربران آسیب‌پذیر شهر تهران

فاطمه سالاروندیان^{۱*} | زینب عبادی^۲ | روژین شاهین طبع^۳

۱. نویسنده مسئول، استادیار جغرافیا و برنامه‌ریزی شهری، گروه مدیریت جهانگردی، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: fsalarvandian@atu.ac.ir
۲. دانش‌آموخته دکتری برنامه‌ریزی حمل‌ونقل، دانشکده مهندسی عمران، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران. رایانامه: zeinabebadi60@gmail.com
۳. کارشناسی ارشد مهندسی معماری، دانشکده مهندسی معماری و شهرسازی، دانشگاه آزاد اسلامی واحد قزوین، قزوین، ایران. رایانامه: saba.amouzade@ut.ac.ir

چکیده

اطلاعات مقاله

نوع مقاله:

پژوهشی

تاریخ‌های مقاله:

تاریخ دریافت: ۱۴۰۵/۰۴/۰۵

تاریخ بازنگری: ۱۴۰۵/۰۵/۰۵

تاریخ پذیرش: ۱۴۰۵/۰۶/۳۱

تاریخ انتشار: ۱۴۰۵/۱۰/۱۱

کلیدواژه:

کاربران آسیب‌پذیر راه،

تصادفات فوتی،

تهران،

یادگیری ماشین،

خوشه‌بندی.

سوانح ترافیکی یکی از چالش‌های مهم سلامت عمومی و ایمنی شهرها از جمله تهران است که سهم عمده‌ای از تلفات فوتی سالانه آن به کاربران آسیب‌پذیر راه مربوط است. پژوهش حاضر بر پایه تصادفات فوتی ثبت‌شده طی سال‌های ۱۳۹۶ تا ۱۴۰۴ از آمار توصیفی، آزمون کای‌دو و ضریب وی کرامر برای بررسی روابط بین متغیرهای طبقه‌ای، از هشت الگوریتم یادگیری ماشین نظارت‌شده به منظور پیش‌بینی نوع متوفی و نوع راه و از خوشه‌بندی K-Means برای شناسایی الگوهای پنهان استفاده کرده است. یافته‌ها بیانگر افزایش قابل ملاحظه روند سالانه تصادفات، بالا بودن سهم عابران پیاده و راکبان موتورسیکلت به عنوان کاربران آسیب‌پذیر راه و سهم بالای بزرگراه‌ها به عنوان اصلی‌ترین محل وقوع تصادفات فوتی است. بر اساس ترکیب معیارهای پایداری، الگوریتم XGBoost با دقت ۰/۸۰۹۱ بهترین عملکرد را در پیش‌بینی نوع متوفی دارد. در خصوص نوع راه به دلیل عدم تعادل شدید در توزیع داده، ۱۴ روش مقابله با عدم توازن بررسی و از نظر توانایی در فراخوانی کلاس خیابان فرعی (۳/۷ درصد داده‌ها) مقایسه شدند؛ تنها روشی که توانست بدون کاهش شدید دقت مدل (دقت ۰/۴۷۰۴)، فراخوانی کلاس خیابان فرعی را به طور معنادار افزایش دهد (از ۰ به ۰/۳۳۳۳)، جنگل تصادفی اصلاحی بود. نتایج خوشه‌بندی نیز نشان داد خوشه صفر با حضور قابل توجه عابران پیاده، خوشه ۱ با سهم بالاتر موتورسیکلت و متغیر بزرگراه، خوشه ۲ با غلبه حضور عابر پیاده و بالاترین سهم فرار از صحنه و خوشه ۳ با غلبه حضور موتورسیکلت (بزرگ‌ترین خوشه) قابل تشخیص هستند. درخور یادآوری است که این تفاوت‌ها، ضرورت استفاده از راهکارهای متناسب با ویژگی‌های هر گروه را در مدیریت ایمنی ترافیک تهران نشان می‌دهد. این پژوهش با تلفیق هم‌زمان روش‌های آماری، پیش‌بینانه و اکتشافی، شواهدی کاربردی برای اولویت‌بندی اقدامات ایمنی در بزرگراه‌ها و برای گروه‌های سنی پرخطر فراهم می‌آورد.

استناد: سالاروندیان، فاطمه؛ عبادی، زینب و شاهین طبع، روژین (۱۴۰۵). ارزیابی تطبیقی الگوریتم‌های یادگیری ماشین در پیش‌بینی نوع متوفی و نوع راه در تصادفات فوتی کاربران آسیب‌پذیر شهر تهران. *سیاستگذاری پیشرفت شهری*، ۳ (۴) ۷۰۹-۷۳۲.

DOI: <http://doi.org/10.22034/judpm.2026.595340.1146>

© نویسندگان

DOI: <http://doi.org/10.22034/judpm.2026.595340.1146>



۱. مقدمه

تصادفات جاده‌ای یکی از چالش‌های مهم بهداشت عمومی و ایمنی ترافیکی در سراسر جهان به شمار می‌رود و ایران نیز از این قاعده مستثنا نیست. بر اساس گزارش سازمان جهانی بهداشت، بیش از نیمی از مرگ‌های ناشی از تصادفات جاده‌ای در میان کاربران آسیب‌پذیر راه، از جمله عابران پیاده، دوچرخه‌سواران و موتورسواران رخ می‌دهد؛ افزون بر این، بیش از ۹۰ درصد مرگ‌های جاده‌ای جهان در کشورهای کم و متوسط درآمد ثبت می‌شود [۱ و ۲]. تهران به عنوان پایتخت و بزرگ‌ترین شهر ایران، با داشتن شبکه گسترده بزرگراه‌ها، خیابان‌های اصلی و فرعی، روزانه صحنه تعداد قابل توجهی از تصادفات منجر به فوت است. این تصادفات نه تنها خسارت‌های جانی و مالی سنگینی به همراه دارند، بلکه تأثیرات عمیق روانی و اجتماعی بر خانواده‌ها و جامعه به جا می‌گذارند. بررسی الگوهای این حوادث و شناسایی عوامل مؤثر بر وقوع آن‌ها می‌تواند کمک شایانی به سیاستگذاران و مدیران شهری در اتخاذ تصمیمات بهینه کند. بر اساس گزارش اداره کل مهندسی و ایمنی ترافیک تهران در سال ۱۴۰۲، ۵۸ درصد تصادفات فوتی مربوط به کاربران آسیب‌پذیر بوده که این رقم در سال ۱۴۰۴ به ۶۴ درصد رسیده است [۳ و ۴]. کاربران آسیب‌پذیر راه شامل عابران پیاده، دوچرخه‌سواران و موتورسواران هستند و در برخی چارچوب‌های سیاستی، گروه‌هایی مانند کارگران راه^۱، کاربران وسایل کمکی حرکتی^۲، کودکان و سالمندان نیز در این دسته قرار می‌گیرند. هدف کلی این پژوهش، تحلیل الگوی تصادفات کاربران آسیب‌پذیر راه در شهر تهران (از منظر مکانی، زمانی و علی) و کاربرد الگوریتم‌های یادگیری ماشین در پیش‌بینی و طبقه‌بندی تصادفات است. مطالعات نشان می‌دهند هزینه‌ها و تبعات اقتصادی ناشی از سوانح ترافیکی می‌تواند درصد قابل توجهی از تولید ناخالص داخلی را تشکیل دهد و فشار زیادی بر سیستم‌های درمانی و بیمه‌ای وارد سازد. بنابراین، تولید شواهد دقیق درباره نقاط پرخطر، علل زمینه‌ای و راهکارهای مؤثر به تصمیم‌گیرندگان شهری، پلیس راهنمایی و رانندگی و دستگاه‌های مسئول مدیریت شهری امکان می‌دهد تا منابع محدود را هدفمند و مبتنی بر شواهد هزینه اثربخش تخصیص دهند. به خلاف اغلب مطالعات پیشین که به آزمون یک یا دو الگوریتم بسنده کرده‌اند، این پژوهش هشت الگوریتم یادگیری ماشین نظارت‌شده (از مدل‌های خطی ساده تا مدل‌های جمعی درختی پیشرفته) را به طور هم‌زمان و تحت شاخص‌های عملکردی یکسان مقایسه می‌کند تا امکان انتخاب مدل بر پایه شواهد تجربی، نه فقط ترجیح روش‌شناختی، فراهم شود. همچنین، پژوهش به خلاف رویکرد رایج در ادبیات مرور شده، دو مسئله طبقه‌بندی چندکلاسه مستقل و مکمل پیش‌بینی نوع متوفی و پیش‌بینی نوع راه را به طور جداگانه برای کاربران آسیب‌پذیر راه مدل‌سازی می‌کند. این پژوهش با تلفیق هم‌زمان روش‌های پیش‌بینانه نظارت‌شده و روش اکتشافی خوشه‌بندی میانگین کا، امکان کشف الگوهای پنهان چندبعدی را فراهم می‌آورد که به‌تنهایی از مدل‌های پیش‌بینی نظارت‌شده قابل استخراج نیست. بنابراین، پژوهش با تمرکز بر داده‌های بلندمدت و بافت‌محور شهر تهران به دنبال پاسخ به پرسش‌های ذیل است:

- الگوها و روابط آماری حاکم بر تصادفات فوتی شهر تهران از نظر ویژگی‌های فردی، زمانی، محیطی و عوامل مؤثر بر وقوع حادثه چگونه است؟
- الگوریتم‌های مختلف یادگیری ماشین تا چه اندازه در پیش‌بینی نوع متوفی و نوع راه در تصادفات فوتی عملکرد مناسبی دارند و کدام الگوریتم از دقت بیشتری برخوردار است؟
- مهم‌ترین متغیرهای مؤثر در پیش‌بینی نوع متوفی و نوع راه کدام‌اند؟
- آیا با استفاده از روش‌های یادگیری ماشین بدون نظارت می‌توان الگوها و خوشه‌های متمایزی از تصادفات فوتی شهر تهران شناسایی کرد؟

این مقاله در پنج بخش اصلی سازمان‌دهی شده است؛ بخش اول ضرورت، اهداف و سؤالات پژوهش را تبیین می‌کند؛ بخش دوم مروری بر مطالعات داخلی و خارجی در حوزه ایمنی کاربران آسیب‌پذیر راه با تأکید بر یادگیری ماشین ارائه می‌دهد؛ بخش

1. Road Workers

2. People With Disabilities or Reduced Mobility

سوم چارچوب روش‌شناختی پژوهش شامل منابع داده، فرایند تحلیل آماری و فلوچارت مراحل پژوهش را شرح می‌دهد؛ بخش چهارم نتایج تحلیل‌های آماری تصادفات، الگوها و روابط آماری حاکم بر تصادفات فوتی با استفاده از الگوریتم‌های مختلف یادگیری ماشین ارائه می‌دهد؛ در نهایت بخش پنجم تفسیر یافته‌ها در پرتو ادبیات موجود، محدودیت‌های پژوهش و پیشنهاد‌های سیاست‌گذاری برای کاهش تصادفات شهری با نگاه ویژه به ارتقای ایمنی کاربران آسیب‌پذیر راه تهران را مطرح می‌سازد.

ادبیات بین‌المللی به‌روشنی نشان می‌دهد مرگ‌ومیر ترافیکی یک مسئله عمده سلامت عمومی است که بار آن به طور نامتوازن بر دوش کشورهای کم و متوسط درآمد و کاربران آسیب‌پذیر راه قرار می‌گیرد [۵]. در این کشورها سهم عابران از کل مرگ‌های ترافیکی به طور معناداری بالاتر است، امری که لزوم سیاست‌گذاری متناسب با بافت محلی را برجسته می‌سازد [۶] و مطالعات متمرکز بر کاربران آسیب‌پذیر نیز همواره عابران و موتورسواران را در معرض بیشترین ریسک شدت و مرگ نشان داده‌اند [۷]. در ایران، تحلیل عوامل خطر مرگ ناشی از سوانح نشان داده ریسک مرگ عابران نسبت به رانندگان و سرنشینان حدود ۲/۰۷ برابر است؛ [۸] این یافته می‌تواند پشتوانه مستقیمی برای تمرکز پژوهش حاضر بر کاربران آسیب‌پذیر فراهم کند. در سطح ملی، روند مرگ‌ومیر ترافیکی ایران طی ۲۰۰۴ تا ۲۰۱۱ کاهشی بوده، [۹] اما شواهد جدیدتر بیانگر بازگشت روند صعودی پس از ۲۰۱۷ است [۱۰]، که ضرورت بازتحلیل داده‌های اخیر را تقویت می‌کند. عوامل مرتبط با مرگ در ایران محدود به رفتار راننده نیستند و هم‌زمان به محیط راه، نوع کاربر و شرایط زمانی وابسته‌اند. در مطالعه شش استان پرجمعیت، حضور عابر، تصادف شبانه، نبود کنترل تقاطع و حضور بزرگراه با افزایش شانس مرگ همراه بوده‌اند [۱۱]. بررسی‌های دیگر نیز خیابان‌های شهری، نقص راه، تجاوز از سرعت مجاز و سبقت خطرناک را با مرگ مرتبط دانسته‌اند. هرچند تصویری جامع از عوامل مرگ در ایران هنوز در دسترس نیست [۸]، در تهران، تحلیل فضایی - زمانی نشان داده بزرگراه‌ها و آزادراه‌ها با نرخ بالاتر تصادفات شدید و نقاط داغ مرگبار همراه‌اند و ساختار شهری، کاربری زمین و نوع راه الگوی بروز و شدت تصادفات را تعیین می‌کنند [۱۲]. از منظر روش‌شناختی، مطالعات پیشین عمدتاً بر رگرسیون لجستیک^۱، پروبیت ترتیبی و مدل‌های پارامتریک تکیه داشتند [۷ و ۱۱]، اما این مدل‌ها به فروض ازپیش‌تعیین‌شده وابسته‌اند و در برهم‌کنش‌های پیچیده داده‌های تصادف ممکن است برآورد سوگیرانه بدهند [۱۳]؛ به همین دلیل، یادگیری ماشین به رویکرد غالب‌تر در پیش‌بینی شدت تصادف تبدیل شده است [۱۴]. در مقایسه‌های چندمدلی، مدل‌های جمعی درختی (جنگل تصادفی^۲، ایکس‌جی‌بوست^۳، تقویت گرادیانت) معمولاً از مدل‌های ساده‌تر بهتر عمل کرده‌اند؛ هرچند برترین مدل بسته به داده و مسئله تغییر می‌کند [۱۳ و ۱۵]. برای کاربران آسیب‌پذیر به طور خاص، جنگل تصادفی عملکرد مقاوم‌تری نشان داده [۱۶] و ایکس‌جی‌بوست در مطالعاتی دیگر متغیرهایی همچون نوع برخورد، شرایط جوی و روشنایی را عوامل حساس معرفی کرده است [۱۳]. بررسی پیشینه نشان‌دهنده وجود الگوی پایداری از مهم‌ترین پیش‌بین‌ها مانند نوع راه، سرعت، روشنایی، نوع وسیله، سن، جنسیت و نوع برخورد است [۵]. به طور مثال، وو و یوهان (۲۰۲۱)^۴ زمان روز را مهم‌ترین پیش‌بین کاربران آسیب‌پذیر معرفی کرده‌اند [۱۷]. با این حال، عدم توازن کلاس‌ها در هر طبقه‌بندی، به دلیل نادر بودن نسبی رخداد‌های مرگبار، از چالش‌های مهم مطالعات در این حوزه است [۱۸]. مطالعات موجود همچنین از خوشه‌بندی به‌خصوص ترکیب میانگین کا^۵ با آزمون کای‌دو برای کشف الگوهای پنهان مکانی، رفتاری و جمعیتی [۱۲]، بررسی ارتباط خوشه‌ها با ویژگی‌های جمعیتی [۱۹] و شناسایی الگوهای پرخطر در میان جوانان و موتورسواران [۲۰] استفاده کرده‌اند. در داده‌های ایران، شواهد مستقیمی برای سودمندی تحلیل روابط طبقه‌ای وجود دارد؛ طاهریان و همکاران (۲۰۲۲) مقدار وی کرامر^۶ بین نوع قربانی و نوع راه را معادل ۰/۲۷ و بین سازوکار حادثه و نوع راه ۰/۲۸ گزارش کرده‌اند [۱۰]. در ادبیات داخلی نیز

1. Logistic Regression
 2. Random Forest
 3. XGBoost
 4. Wu, Yuan
 5. K-Means
 6. Cramér's V

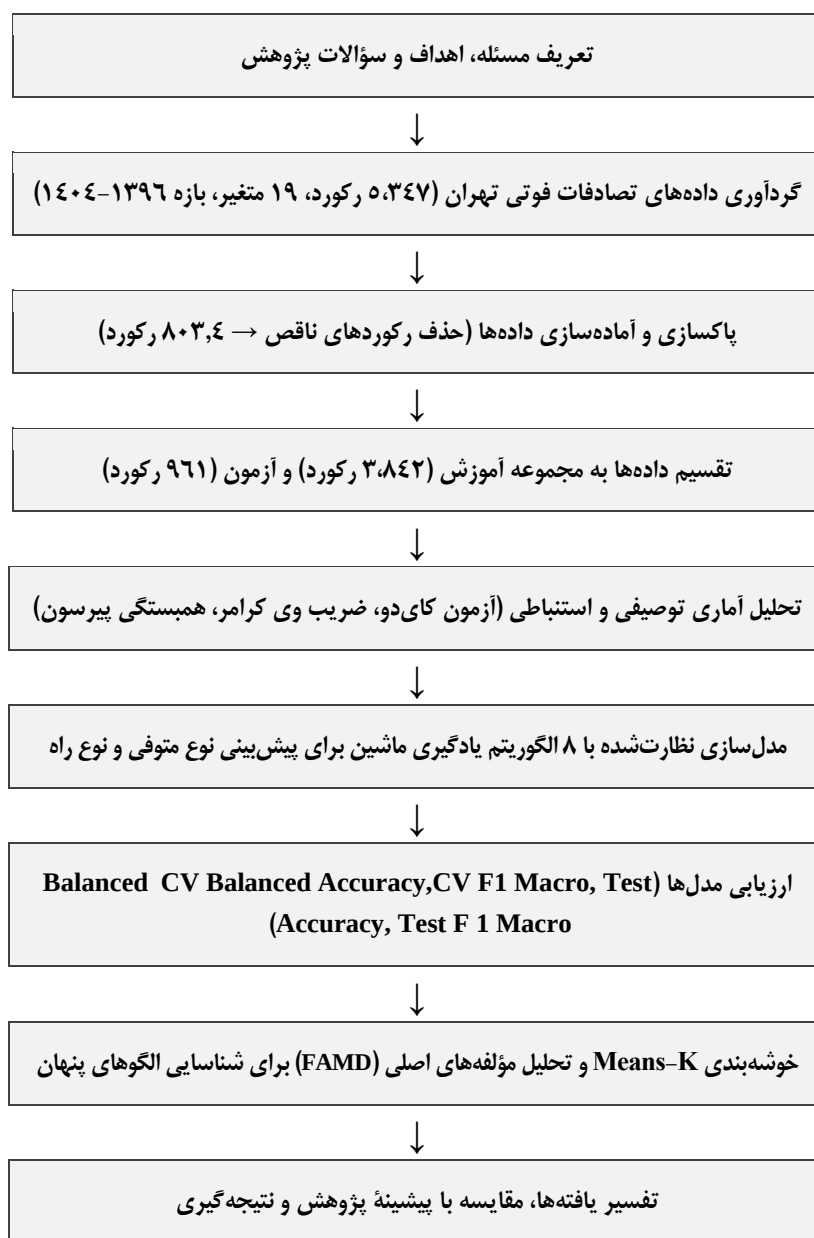
طی سال‌های اخیر استفاده از یادگیری ماشین در تحلیل و پیش‌بینی شدت تصادفات مورد توجه قرار گرفته است؛ برای نمونه، خواجه سلیمی و همکاران (۲۰۱۹) در مطالعه‌ای بر داده‌های تصادفات جاده‌های بین شهری ایران، عملکرد الگوریتم ماشین بردار پشتیبان را در پیش‌بینی شدت تصادفات در دو سطح جرحی و فوتی بررسی کردند و نشان دادند حضور وسیله نقلیه سنگین و خودروی سواری، در کنار خستگی و خواب‌آلودگی، حضور عابر پیاده و انحراف به چپ، از عوامل مهم در تبیین شدت تصادف هستند؛ همچنین در این مطالعه، مدل درخت تصمیم^۱ نسبت به ماشین بردار پشتیبان عملکرد بهتری نشان داد [۲۱]. در مطالعه‌ای دیگر، محمدپور و محمدزاده مقدم (۲۰۲۶) با تمرکز بر تصادفات جاده‌ای ایران، نوعی رویکرد یکپارچه پیش‌پردازش داده مبتنی بر الگوریتم‌های یادگیری ماشین ارائه کردند و نشان دادند توجه به داده‌های گمشده، عدم توازن کلاس‌ها و انتخاب متغیرهای مؤثر، نقش تعیین‌کننده‌ای در بهبود پیش‌بینی شدت تصادف دارد؛ به گونه‌ای که اصلاح عدم توازن، دقت پیش‌بینی تصادفات فوتی را به طور معناداری افزایش داد [۲۲]. همچنین، توکلی کاشانی و همکاران (۲۰۲۴) در مطالعه استان زنجان، با به‌کارگیری ماشین بردار پشتیبان و درخت تصمیم برای پیش‌بینی شدت تصادفات جاده‌ای، نشان دادند نوع تصادف، نوع وسیله نقلیه و کیلومتر وقوع تصادف از مهم‌ترین متغیرهای پیش‌بین هستند و درخت تصمیم از نظر قدرت پیش‌بینی عملکرد بهتری داشته است [۲۳]. به طور کلی، بسیاری از مطالعات انجام‌شده تنها یک یا چند الگوریتم محدود را آزموده‌اند و مقایسه هم‌زمان مجموعه‌ای از چند مدل کمتر دیده می‌شود [۲۴ و ۲۵]. اگرچه مطالعات متعددی بر متغیرهای راه، محیط و وسیله نقلیه تأکید کرده‌اند، هنوز در بسیاری از پژوهش‌ها مدل‌سازی اختصاصی برای زیرگروه‌های خاص و از جمله کاربران آسیب‌پذیر راه به اندازه کافی توسعه نیافته است [۱۴]. در حال حاضر، بخش قابل توجهی از مطالعات بر داده‌های محدود، بازه‌های کوتاه یا مجموعه داده‌های غیر قابل تعمیم استوار بوده‌اند و خود مرورهای جدید نیز بر مشکل انتقال‌پذیری مدل‌ها بین مناطق مختلف تأکید کرده‌اند. بنابراین، نوآوری پژوهش حاضر در پر کردن خلأهای روش‌شناختی نهفته است و می‌تواند شکاف مطالعاتی را در سطح شهری و بافت‌محور تا حدی پوشش دهد.

۲. مواد و روش‌ها

پژوهش حاضر در چارچوب پارادایم اثبات‌گرایی، با رویکرد کمی و از نوع مطالعات مشاهده‌ای - گذشته‌نگر مبتنی بر داده‌های ثانویه انجام شد. جامعه آماری شامل تمامی تصادفات منجر به فوت ثبت‌شده در شهر تهران طی بازه ۱۳۹۶ تا ۱۴۰۴ بود؛ پایگاه داده اولیه ۵،۳۴۷ رکورد و ۱۹ متغیر را در چهار گروه فردی، زمانی، محیطی/راه و مربوط به حادثه/وسیله نقلیه در بر می‌گرفت. پس از بررسی کامل بودن و سازگاری داده‌ها، رکوردهای ناقص از مرحله مدل‌سازی حذف شدند و ۴،۸۰۳ مشاهده برای تحلیل باقی ماند که به دو مجموعه آموزش (۳،۸۴۲ رکورد) و آزمون (۹۶۱ رکورد) تقسیم شد. تحلیل داده در چند گام انجام گرفت: ابتدا آمار توصیفی برای شناخت الگوهای کلی، سپس آزمون کای‌دو و ضریب وی کرامر برای سنجش رابطه و شدت ارتباط میان متغیرهای طبقه‌ای و ضریب همبستگی پیرسون به‌منظور بررسی روابط بین متغیرهای عددی، همگی در سطح اطمینان ۹۵ درصد استخراج شد. برای پیش‌بینی، دو مسئله طبقه‌بندی چندکلاسه مستقل (نوع متوفی و نوع راه) تعریف و هشت الگوریتم یادگیری ماشین نظارت‌شده شامل رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، نزدیک‌ترین همسایه^۲، بیز ساده^۳، گرادیان بوستینگ^۴، هیستوگرادیان بوستینگ^۵ و ایکس‌جی‌بوست مقایسه شدند؛ برای انتخاب مدل برتر Balanced CV Accuracy به عنوان معیار اصلی، Macro 1CV F به عنوان معیار ثانویه، Test Balanced Accuracy به عنوان معیار سوم و Macro 1Test F به عنوان معیار نهایی تعیین شدند. برای مقابله با عدم تعادل شدید در توزیع داده‌های نوع راه ۱۴ روش

1. Decision Tree
2. KNN
3. Naive Bayes
4. Gradient Boosting
5. HistGradientBoosting

کلاسیک و پیشرفته مانند Ensemble, Hybrid, Oversampling, Class Weighting تخصصی، و در نهایت sensitive-Cost و Threshold Tuning بررسی و مقایسه شده‌اند. درخور یادآوری است که در این پژوهش کلیه مدل‌ها طبقه‌بندی را انجام می‌دهند و برای پیش‌بینی وقوع تصادفات آینده طراحی نشده‌اند. در نهایت، برای کشف ساختارهای پنهان بدون تعریف متغیر هدف، از خوشه‌بندی میانگین کا با تعیین کای بهینه (از طریق روش آرنج^۱ و مقدار سیلوئت^۲) و تحلیل مؤلفه‌های اصلی برای کاهش ابعاد و نمایش تصویری استفاده شد. کلیه مراحل در محیط برنامه‌نویسی پایتون و در یک بستر محاسباتی یکپارچه اجرا شد تا تکرارپذیری فرایند پژوهش تضمین شود. شکل ۱ مراحل این روند را از جمع‌آوری داده تا ارزیابی نهایی به صورت فلوچارت نشان می‌دهد.



شکل ۱. فرایند روش‌شناسی پژوهش

۳. یافته‌ها

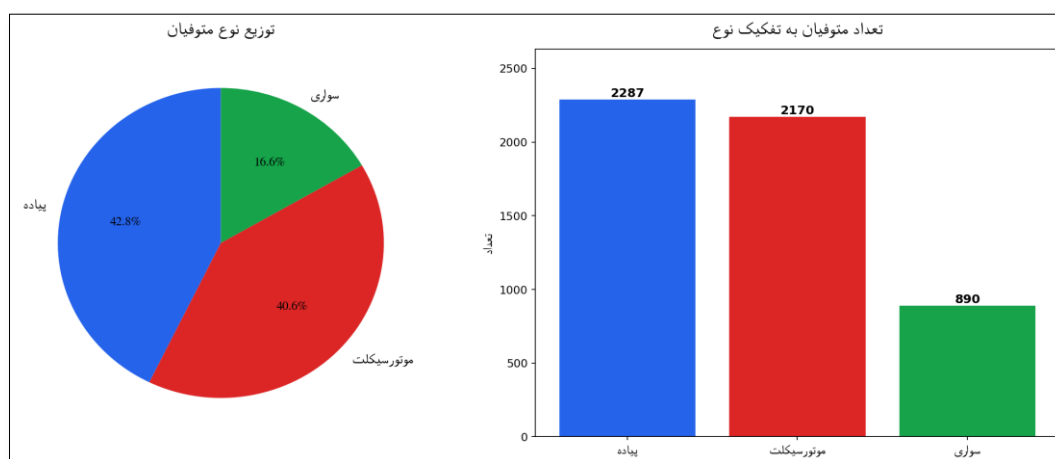
۳-۱. یافته‌های توصیفی

۳-۱-۱. توزیع متوفیان بر اساس کاربران راه

همان‌طور که در جدول ۱ نشان داده شده است، داده‌های این پژوهش شامل ۵,۳۴۷ رکورد از تصادفات فوتی شهر تهران در بازه زمانی ۱۳۹۶ تا ۱۴۰۴ و مشتمل بر ۱۹ ویژگی است. میانگین سن متوفیان ۴۳/۶ سال بوده و ۷۸/۲ درصد از قربانیان را مردان تشکیل می‌دهند که نشان‌دهنده غلبه مرگ‌ومیر در میان مردان است. از نظر نوع کاربر، عابران پیاده با ۲,۲۸۷ مورد (۴۲/۸ درصد) و موتورسیکلت‌سواران با ۲,۱۷۰ مورد (۴۰/۶ درصد) بیشترین سهم را در میان متوفیان دارند و پس از آن‌ها، سرنشینان وسایل نقلیه سواری با ۸۹۰ مورد (۱۶/۶ درصد) قرار می‌گیرند. این توزیع نشان می‌دهد بار اصلی تلفات فوتی در تهران بر دوش کاربران آسیب‌پذیر راه قرار دارد و لازم است تحلیل الگوهای مکانی، زمانی و رفتاری با تمرکز ویژه بر این گروه‌ها انجام شود.

جدول ۱. ویژگی‌های توصیفی داده‌ها

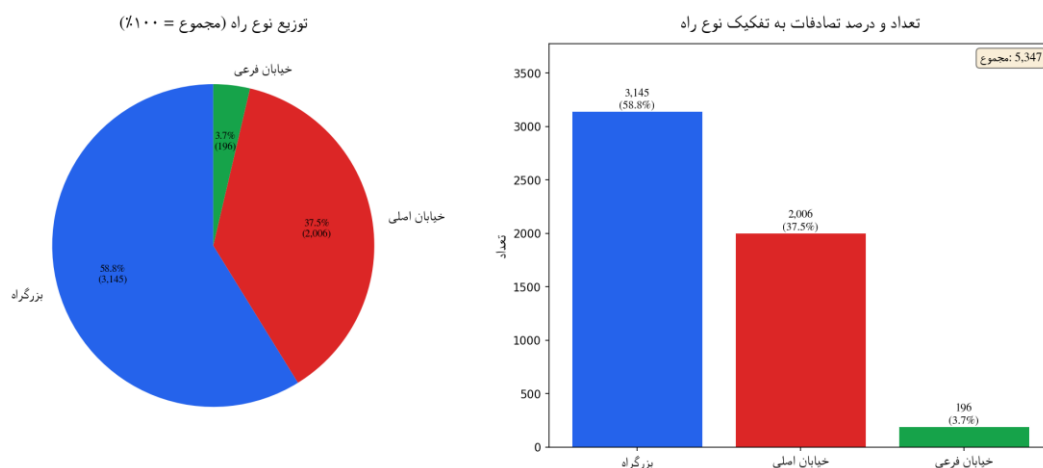
شاخص	مقدار
تعداد کل رکوردها	۵,۳۴۷ (پیش از حذف مقادیر گمشده)
بازه زمانی	۱۳۹۶ تا ۱۴۰۴
تعداد ویژگی‌ها	۱۹
میانگین سن متوفیان	۴۳/۶ سال
درصد مردان	۷۸/۲٪
تعداد پیاده	۲,۲۸۷
تعداد موتورسیکلت	۲,۱۷۰
تعداد سواری	۸۹۰
درصد مردان	۷۸/۲٪



شکل ۲. توزیع متوفیان بر اساس کاربران راه

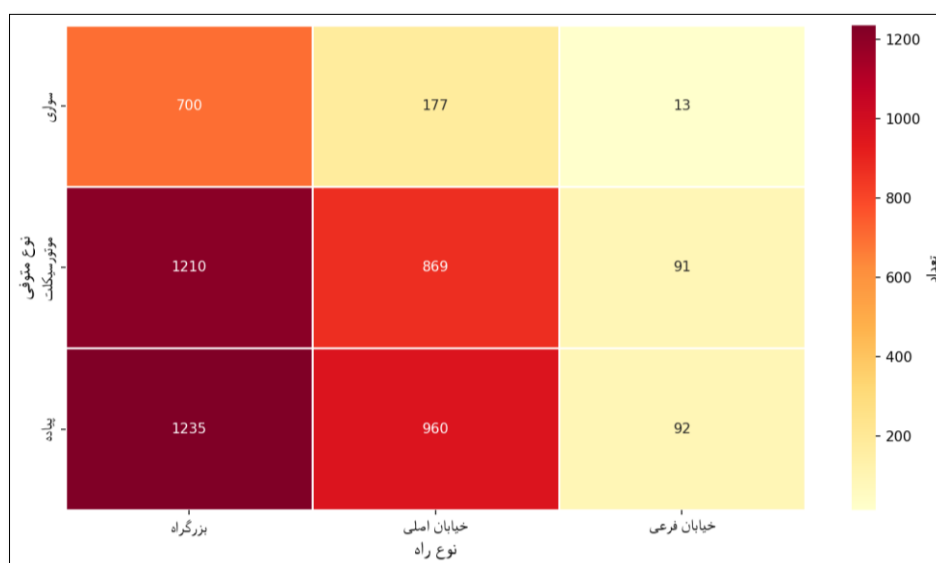
۳-۱-۲. توزیع نوع راه

تحلیل توزیع مکانی تصادفات (شکل ۳) نشان می‌دهد بزرگراه‌ها با ۵۸/۸ درصد اصلی‌ترین محل وقوع تصادفات فوتی هستند. خیابان‌های اصلی با ۳۷/۵ درصد و خیابان‌های فرعی با تنها ۳/۷ درصد در رتبه‌های بعدی قرار دارند. غلبه چشمگیر بزرگراه‌ها در آمار فوتی‌ها بیانگر سرعت بالای ترافیک و شدت بالای تصادفات در این نوع معابر است. همچنین، تمرکز بالای تصادفات در خیابان‌های اصلی نسبت به فرعی نشان‌دهنده حجم بالای ترافیک و تعامل بیشتر وسایل نقلیه و عابران در این معابر است.



شکل ۳. توزیع نوع راه در تصادفات فوتی

جدول متقاطع بین نوع متوفی و نوع راه (شکل ۴)، نیز نشان می‌دهد بیشترین تعداد فوتی مربوط به عابران پیاده در بزرگراه‌ها و سپس در خیابان‌های اصلی است؛ در حالی که راکبان موتورسیکلت عمدتاً در بزرگراه‌ها دچار حادثه شده‌اند. این الگو نشان‌دهنده آن است که بزرگراه‌ها با وجود ایمنی‌سازی‌های بیشتر به دلیل سرعت بالای وسایل نقلیه، خطرناک‌ترین نوع معبر برای تمام گروه‌های کاربری هستند. نکته قابل توجه آن است که سهم عابران پیاده در خیابان‌های اصلی نیز قابل توجه است که نشان‌دهنده نقص در زیرساخت‌های عابرگذار و اقدامات کنترل سرعت در این معابر است.



شکل ۴. جدول متقاطع نوع متوفی و نوع راه

۳-۱-۳. روند سالانه فوتی‌ها

همان‌طور که جدول ۲ نشان می‌دهد، تعداد خام متوفیان با نوسانی نامنظم در نرخ تغییرات سالانه (کاهش ۹/۷۱ درصدی در ۱۴۰۰، افزایش تدریجی در ۱۴۰۱ و ۱۴۰۲، کاهش جزئی در ۱۴۰۳ و جهش ۲۶/۳۷ درصدی در ۱۴۰۴) از ۵۳۰ نفر در سال ۱۴۰۰ به ۷۸۶ نفر در سال ۱۴۰۴ رسیده است. با این حال، تفسیر این ارقام خام به‌تنهایی می‌تواند گمراه‌کننده باشد؛ زیرا بخشی از این افزایش می‌تواند فقط بازتاب رشد حجم سفرهای درون‌شهری باشد که در همین بازه از ۱۸/۸ به ۲۱ میلیون سفر افزایش

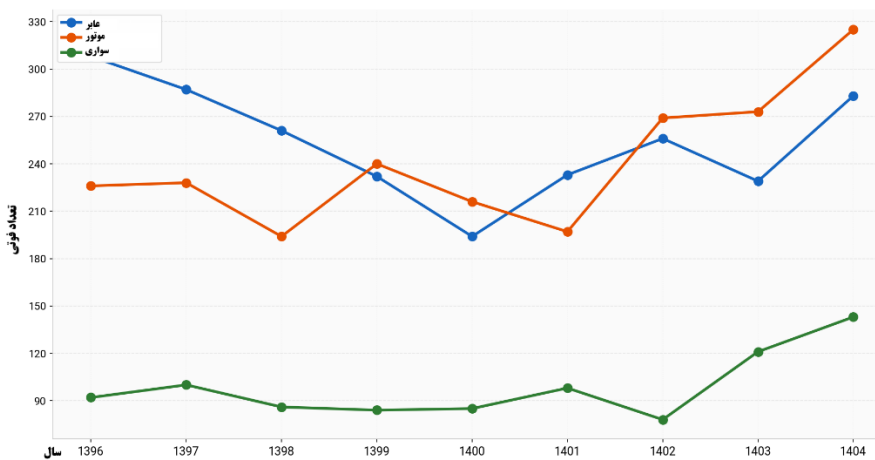
یافته است. برای رفع این محدودیت، نرخ فوتی به ازای هر میلیون سفر نیز محاسبه شد؛ این نرخ نیز نمایانگر روند صعودی (از ۰/۰۷۱ در سال ۱۴۰۰ به ۰/۰۸۱ در سال ۱۴۰۳) بوده که نشان می‌دهد افزایش تلفات فقط ناشی از رشد حجم سفرها نیست و ریسک فوتی به ازای هر سفر نیز به طور مستقل افزایش یافته است. اهمیت این یافته از این نظر است که نشان می‌دهد بهبود ایمنی ترافیکی فقط با کنترل حجم سفر (مثلاً از طریق توسعه حمل‌ونقل عمومی) کافی نبوده و نیازمند راهکارهای مستقیم مانند بهبود زیرساخت و اجرای قوانین در ریسک هر سفر است. درخور یادآوری است که به دلیل عدم انتشار رسمی آمار حجم سفر سال ۱۴۰۴ در زمان نگارش این پژوهش، محاسبه نرخ تعدیل‌شده این سال ممکن نبوده و افزایش ۲۶/۳۷ درصدی این سال باید با احتیاط و در انتظار داده‌های تکمیلی تفسیر شود. در ادامه بررسی روند زمانی فوتی‌ها نشان‌دهنده پیشی گرفتن تلفات موتورسیکلت از گروه فوتی عابران پیاده است (شکل ۵).

جدول ۲. تعداد متوفیان و نرخ تغییرات سالانه

سال	تعداد متوفیان*	نرخ تغییرات %	تعداد سفرهای درون‌شهری (میلیون)	نرخ فوتی به ازای هر میلیون سفر
۱۴۰۰	۵۳۰	-۹.۷۱	۱۸/۸	۰/۰۷۷
۱۴۰۱	۵۴۰	۱.۸۹	۱۹/۸	۰/۰۷۴
۱۴۰۲	۶۲۴	۱۵.۵۶	۲۰/۳	۰/۰۸۴
۱۴۰۳	۶۲۲	-۰.۳۲	۲۱	۰/۰۸۱
۱۴۰۴	۷۸۶	۲۶.۳۷	اعلام نشده	—

منبع: محاسبات نگارندگان بر اساس تعداد سفرهای شهری ارائه‌شده در آمارنامه‌های شهر تهران (۱۴۰۰-۱۴۰۳)

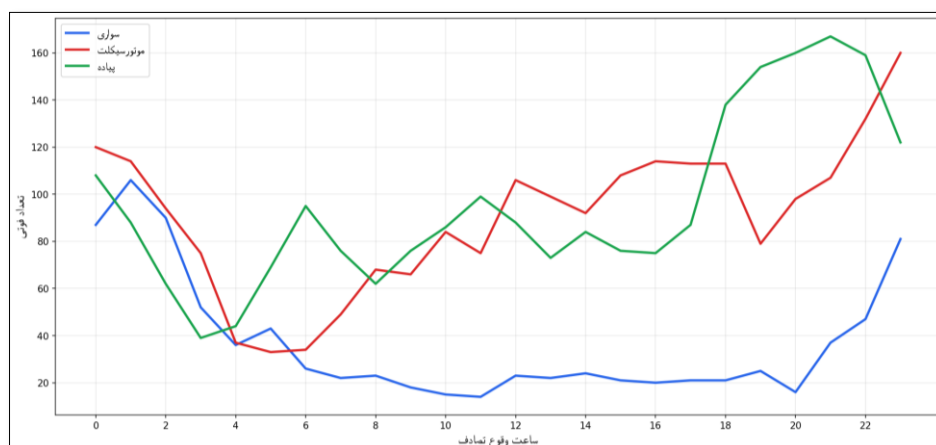
* بر اساس گزارش اداره کل مهندسی و ایمنی معاونت حمل‌ونقل تهران (۱۴۰۵)



شکل ۵. روند سالانه فوتی‌ها به تفکیک نوع متوفی

۳-۱-۴. توزیع ساعتی

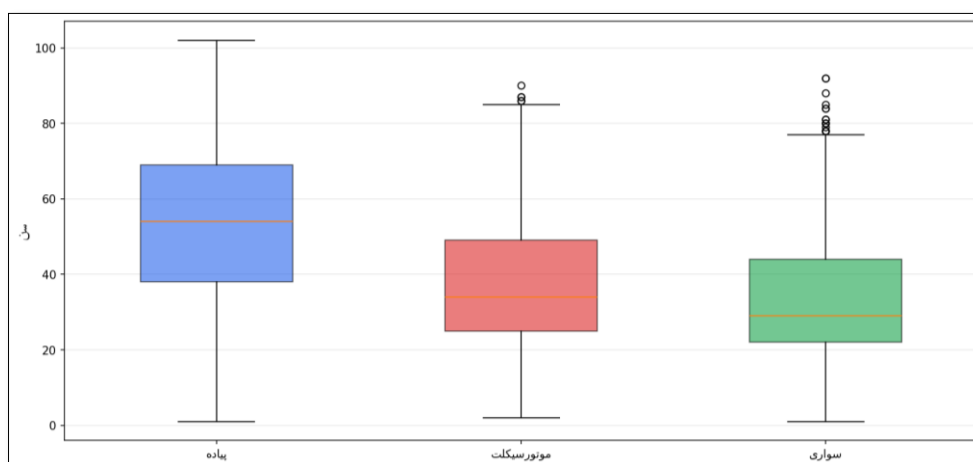
تحلیل توزیع ساعتی تصادفات فوتی نشان می‌دهد الگوی زمانی وقوع حوادث بین سه گروه متوفی تفاوت‌های معناداری دارد. عابران پیاده بیشتر در ساعات عصر و شب (بین ساعت‌های ۱۸ تا ۲۲) دچار حادثه می‌شوند که این موضوع می‌تواند به کاهش دید رانندگان و همچنین، افزایش خستگی در ساعات پایانی روز مرتبط باشد (شکل ۶). همچنین، تمرکز تصادفات راکبان موتورسیکلت با الگوی مشابهی بیشتر در ساعات اوج ترافیک رخ داده است (ساعت ۱۹ تا ۲۴). این یافته‌ها تأیید می‌کند که اقداماتی مانند بهبود روشنایی معابر، نصب چراغ‌های خطرناک و پایش سرعت در ساعات خاص می‌تواند تأثیر چشمگیری در کاهش فوتی‌ها داشته باشد.



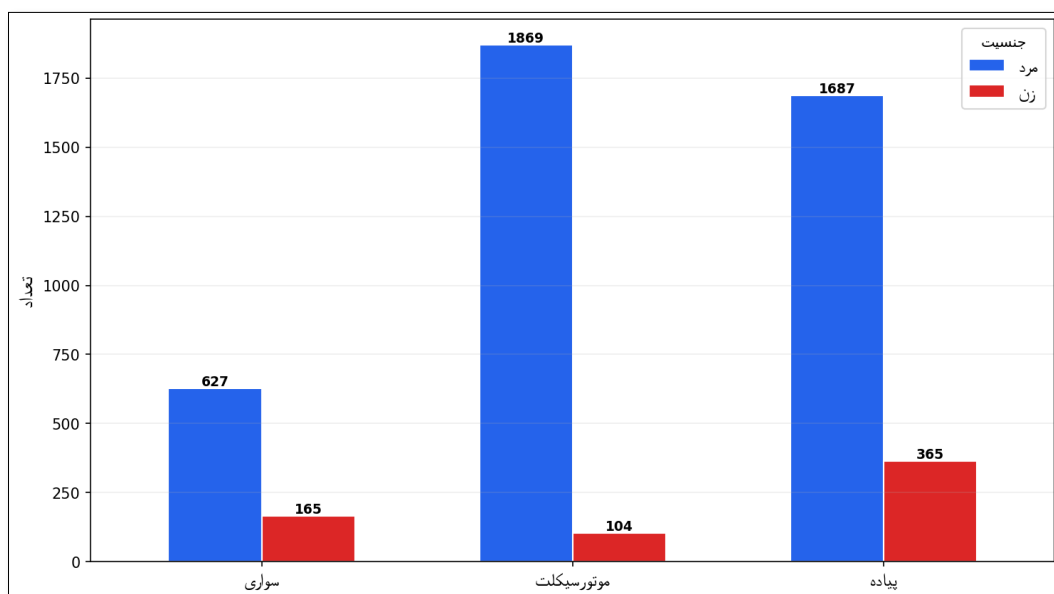
شکل ۶. توزیع ساعتی فوتی‌ها به تفکیک نوع متوفی

۳-۱-۵. توزیع سنی و جنسیتی

میانگین و میانه سن متوفیان ناشی از تصادفات رانندگی در تهران به ترتیب $۴۳/۶$ و ۴۰ سال بود که بیانگر تمرکز مرگ‌ومیر در سنین جوانی و میانسالی، در کنار سهم قابل توجه سالمندان، است. اگرچه افراد ۶۰ سال و بالاتر با $۲۴/۶$ درصد بیشترین سهم را داشتند، $۷۵/۴$ درصد از کل فوت‌شدگان کمتر از ۶۰ سال سن داشتند و گروه‌های ۲۹.۲۰ و ۳۹.۳۰ سال به ترتیب $۲۱/۱۳$ درصد و $۱۷/۴$ درصد از مرگ‌ها را تشکیل دادند. این الگو با گزارش سازمان جهانی بهداشت مبنی بر آنکه حدود دوسوم مرگ‌های ناشی از سوانح ترافیکی در افراد ۱۸ تا ۵۹ سال و در سنین فعالیت اقتصادی رخ می‌دهد همخوانی دارد [۳۶]. با این حال، در مطالعه دیگری میانه سن تصادفات فوتی در سطح ایران ۳۴ سال گزارش شده که این تفاوت را می‌توان به سهم بیشتر سالمندان و ویژگی‌های تصادفات در محیط شهری تهران، از جمله نقش عابران پیاده، مربوط دانست [۳۰ و ۳۹]. بررسی توزیع سنی به تفکیک نوع متوفی نیز نشان می‌دهد عابران پیاده میانگین سنی بالاتری نسبت به راکبان موتورسیکلت دارند. این یافته منطقی به نظر می‌رسد، چرا که بخش قابل توجهی از عابران فوتی را افراد مسن تشکیل می‌دهند. در مقابل، راکبان موتورسیکلت عمدتاً جوانانی هستند که ریسک‌پذیری بالاتر و تجربه رانندگی کمتری دارند (شکل ۷). از نظر جنسیتی، مردان با حدود $۸۶/۸$ درصد بیشترین سهم از تصادفات فوتی را دارند که می‌تواند نشان‌دهنده حضور بیشتر مردان در فضای ترافیک به‌ویژه به عنوان راننده و راکب موتورسیکلت است (شکل ۸).



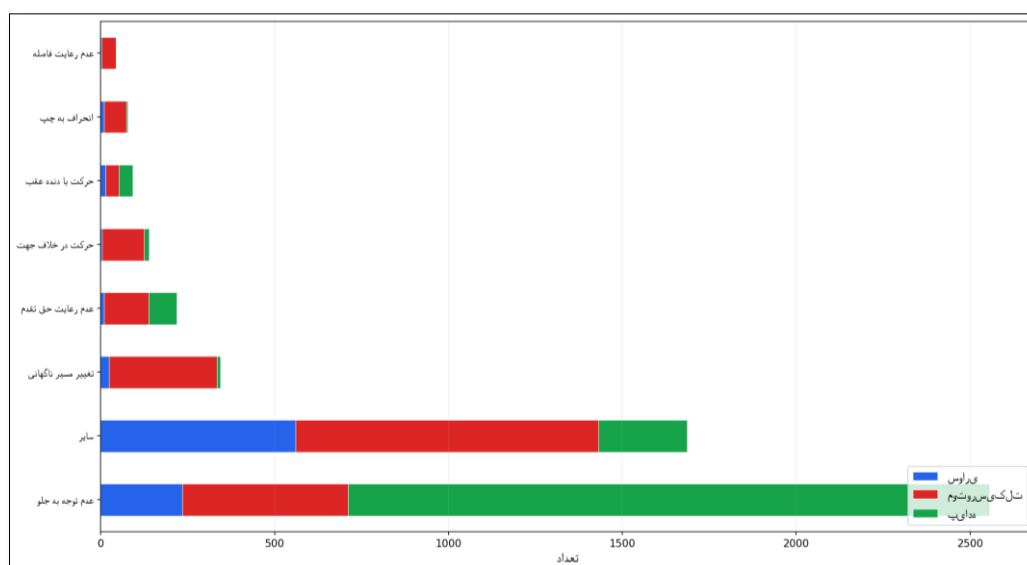
شکل ۷. توزیع سنی متوفیان (نمودار جعبه‌ای)



شکل ۸. توزیع جنسیتی متوفیان به تفکیک جنسیت

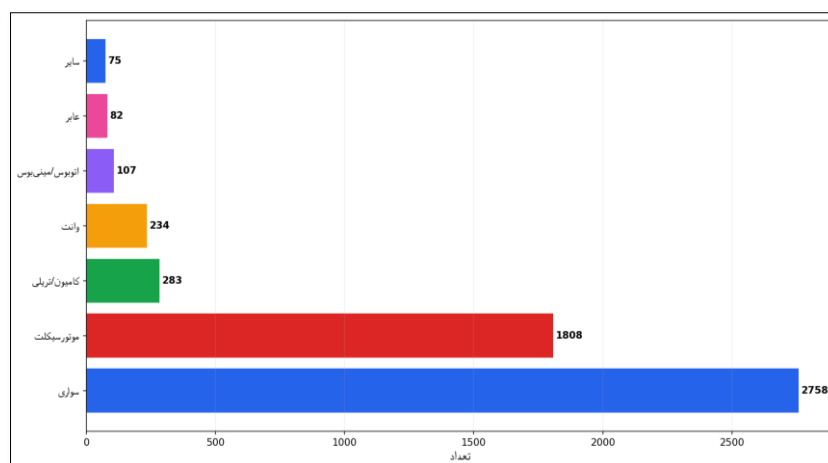
۳-۱-۶. علل تصادف و نوع مقصر

بررسی علل تصادفات نشان می‌دهد بی‌توجهی به جلو با اختلاف زیاد، اصلی‌ترین علت تصادفات فوتی است؛ عدم توانایی در کنترل وسیله نقلیه در رتبه دوم و تغییر مسیر ناگهانی در رتبه سوم قرار دارند (شکل ۹). این سه علت به‌تنهایی مسئول بیش از ۸۰ درصد کل تصادفات فوتی هستند. عدم رعایت حق تقدم و حرکت در خلاف جهت نیز از علل مهم دیگر به شمار می‌روند. نکته درخور توجه تفاوت علل منجر به فوت بین سه گروه کاربران راه است؛ به طور مثال، بی‌توجهی به جلو بیشترین تأثیر را بر فوتی عابران و موتورسیکلت‌سواران داشته، در حالی که عدم کنترل وسیله نقلیه نقش بیشتری در فوتی سرنشینان سواری ایفا می‌کند.



شکل ۹. علت تصادف به تفکیک نوع متوفی

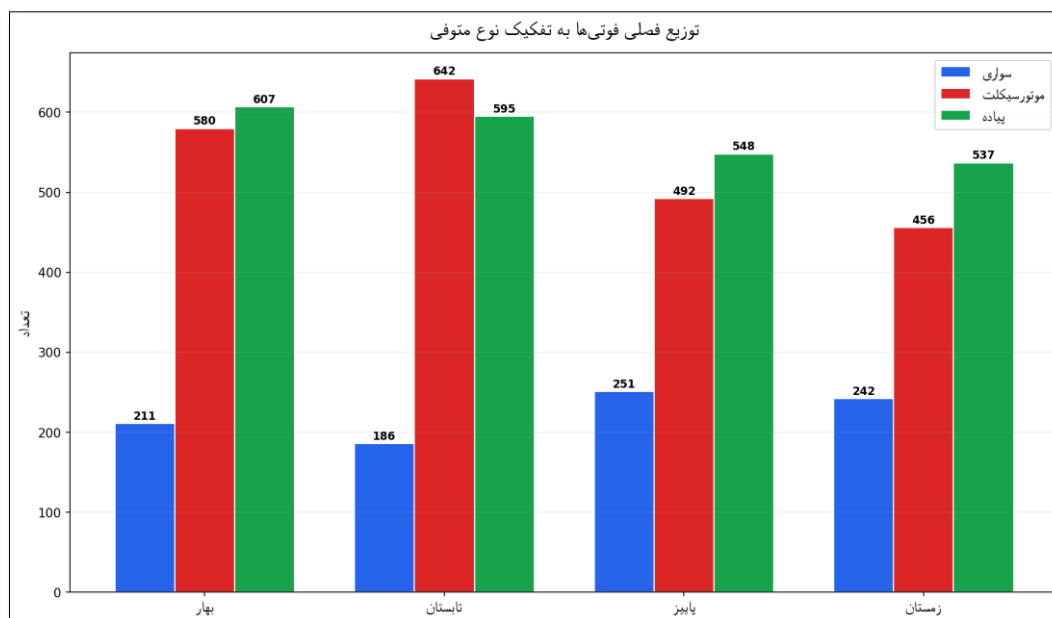
توزیع نوع مقصر (شکل ۱۰) نیز نشان می‌دهد سواری و موتورسیکلت بیشترین سهم در وقوع تصادفات فوتی را داشته‌اند.



شکل ۱۰. توزیع نوع وسیله مقصر

۳-۱-۷. توزیع فصلی

تحلیل فصلی فوتی‌ها نشان می‌دهد تفاوت معناداری بین فصل‌های مختلف وجود دارد (شکل ۱۱). تابستان و بهار معمولاً با افزایش فوتی‌ها همراه هستند که می‌تواند ناشی از افزایش حجم ترافیک، سفرهای تابستانی و همچنین، افزایش فعالیت موتورسیکلت‌ها در فصل‌های گرم سال باشد؛ در حالی که پاییز و زمستان با کاهش نسبی فوتی‌ها همراه بوده که احتمالاً به دلیل کاهش فعالیت‌های بیرونی و افزایش احتیاط رانندگان در شرایط جوی نامساعد باشد. این الگوهای فصلی می‌تواند در برنامه‌ریزی تخصیص منابع پلیس و اورژانس و همچنین، اجرای طرح‌های ایمنی فصلی مورد استفاده قرار گیرد.



شکل ۱۱. توزیع فصلی فوتی‌ها به تفکیک نوع متوفی

۳-۲. یافته‌های تحلیلی

۳-۲-۱. تحلیل آماری (آزمون کای دو)^۱

برای بررسی وجود رابطه آماری معنادار بین متغیرهای طبقه‌ای، از آزمون کای دو و ضریب وی کرامر^۲ به عنوان معیار شدت ارتباط استفاده و نتایج در جدول ۳ ارائه شده است. تمام آزمون‌ها در سطح معناداری ۰/۰۵ انجام شده‌اند؛ قوی‌ترین ارتباط آماری بین نوع متوفی و بازه زمانی روز مشاهده شده است ($V=0/1800$) که نشان می‌دهد الگوی زمانی وقوع تصادفات به طور معناداری بین عابران پیاده، راکبان موتورسیکلت و سرنشینان سواری متفاوت است. ارتباط معناداری نیز بین نوع متوفی و نوع راه وجود دارد ($V=0/1282$) که تأیید می‌کند مکان وقوع تصادف با نوع فرد متوفی مرتبط است. در مقابل، ارتباط بین نوع راه و فصل بسیار ضعیف ($V=0/0124$) و از نظر آماری معنادار نیست؛ به این معنا که توزیع تصادفات در انواع راه طی فصل‌های سال تقریباً ثابت بوده است.

جدول ۳. نتایج آزمون کای دو و ضریب وی کرامر

متغیر ۱	متغیر ۲	کای دو	p-value	V کرامر	معنادار
نوع متوفی	نوع راه	۱۷۵۸۰	۵۰۹۶e-۳۷	۰۰۱۲۸۲	بله
نوع متوفی	بازه زمانی	۳۴۶۰۴۷	۸۰۸۴e-۷۲	۰۰۱۸۰۰	بله
نوع متوفی	فصل	۳۹۶۰	۵۰۴۶e-۰۷	۰۰۰۶۰۹	بله
نوع راه	فصل	۱۰۶۶	۹۰۴۸e-۰۱	۰۰۰۱۲۴	خیر
نوع راه	بازه زمانی	۷۷۰۶۶	۱۰۰۹e-۱۴	۰۰۰۸۵۲	بله

۳-۲-۲. ماتریس همبستگی

ماتریس همبستگی پیرسون بین ویژگی‌های عددی محاسبه و در شکل ۱۲ نمایش داده شده است. بررسی این ماتریس نشان می‌دهد همبستگی‌های قوی بین ویژگی‌ها وجود ندارد که می‌تواند نشان‌دهنده نبود چندخطی شدید بین متغیرهای مستقل و دقت مدل‌سازی پژوهش باشد؛ بیشترین همبستگی مثبت بین متغیرهای «روی گذرگاه عابرپیاده» و «نوع متوفی» مشاهده شده که منطقی به نظر می‌رسد؛ زیرا وضعیت عبور از گذرگاه عمدتاً برای عابران پیاده ثبت می‌شود. همبستگی بین الکل و فرار از صحنه نیز قابل توجه است که نشان‌دهنده آن است که رانندگان مست بیشتر تمایل به فرار از صحنه حادثه دارند.

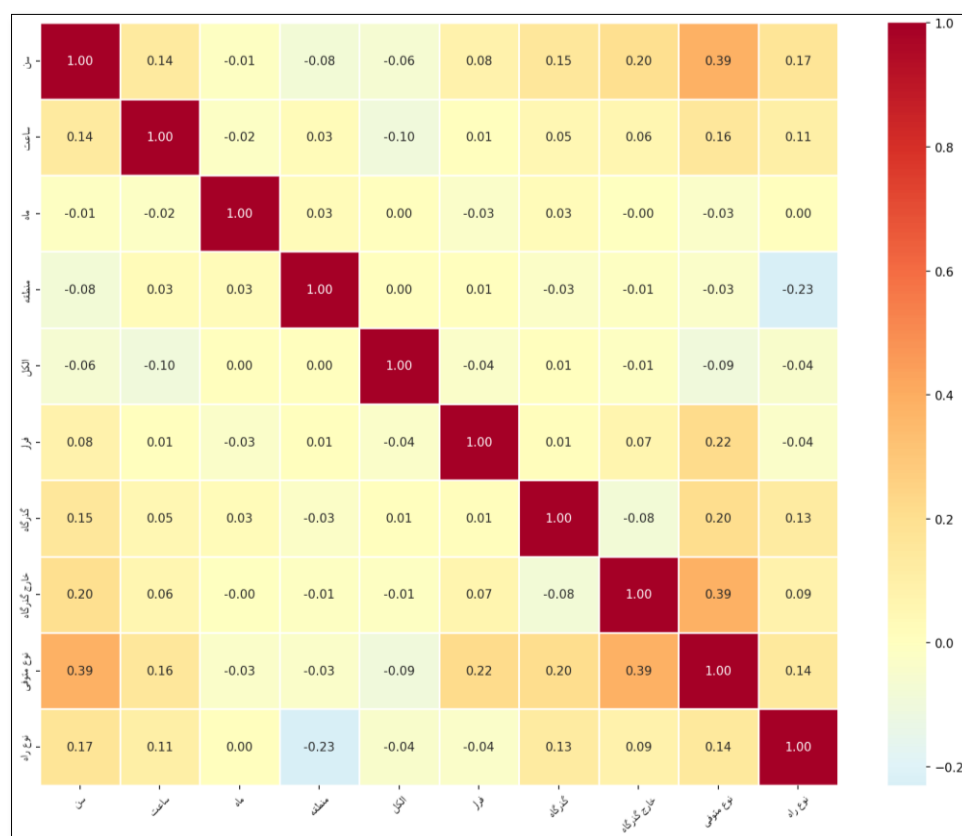
۳-۲-۳. مدل‌های طبقه‌بندی

طبقه‌بندی نوع متوفی

مجموعه داده شامل ۴۰۸۰۳ رکورد (با حذف مقادیر گم‌شده جنسیت و سن) بود که از این تعداد ۳۸۴۲ مورد برای آموزش و ۹۶۱ مورد برای آزمایش استفاده شد. جدول ۴ عملکرد مدل‌ها را نشان می‌دهد.

1. Chi-Square Test

2. Cramer's V



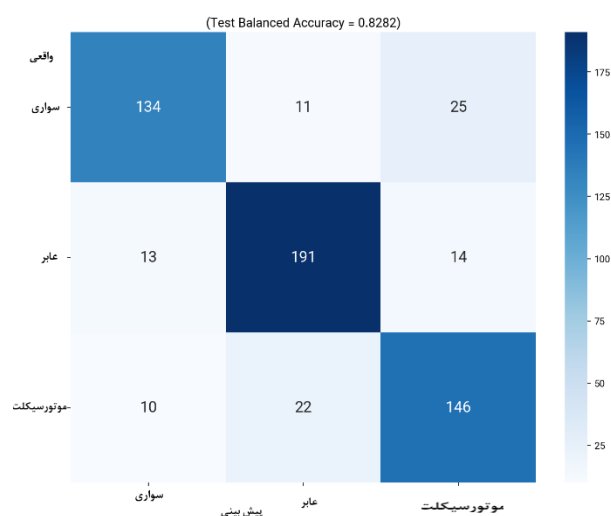
شکل ۱۲. ماتریس همبستگی ویژگی‌های عددی

جدول ۴. عملکرد مدل‌ها

رتبه	مدل	Acc.Bal CV	m1F CV	Acc Test	Test Acc.Bal	m1F Test
۱ ★	XGBoost	۰٫۰۱۷±۰٫۸۰۹۱	۰٫۰۱۶±۰٫۸۱۰۶	۰٫۸۳۲۲	۰٫۸۲۸۲	۰٫۸۲۹۴
۲	Gradient Boosting	۰٫۰۲۱±۰٫۷۹۸۸	۰٫۰۲۱±۰٫۸۰۰۵	۰٫۸۴۱۰	۰٫۸۳۴۴	۰٫۸۳۷۶
۳	stGradientBoostingHi	۰٫۰۲۳±۰٫۷۸۹۵	۰٫۰۲۲±۰٫۷۹۰۸	۰٫۸۲۳۳	۰٫۸۲۰۲	۰٫۸۲۰۵
۴	Random Forest	۰٫۰۲۲±۰٫۷۸۷۴	۰٫۰۲۱±۰٫۷۹۱۱	۰٫۸۱۹۸	۰٫۸۱۰۸	۰٫۸۱۵۲
۵	Logistic Regression	۰٫۰۲۲±۰٫۷۸۱۸	۰٫۰۲۲±۰٫۷۸۲۶	۰٫۸۳۲۲	۰٫۸۲۶۷	۰٫۸۲۹۰
۶	Decision Tree	۰٫۰۱۷±۰٫۷۲۸۶	۰٫۰۱۷±۰٫۷۲۸۵	۰٫۷۶۵۰	۰٫۷۶۳۱	۰٫۷۶۲۲
۷	KNN	۰٫۰۲۹±۰٫۶۹۳۲	۰٫۰۲۷±۰٫۶۹۴۸	۰٫۷۳۸۵	۰٫۷۳۵۷	۰٫۷۳۸۰
۸	Naive Bayes	۰٫۰۵۴±۰٫۵۰۳۳	۰٫۰۶۳±۰٫۴۵۱۴	۰٫۵۲۸۳	۰٫۵۴۱۷	۰٫۵۱۰۰

برای انتخاب مدل برتر CV Balanced Accuracy به عنوان معیار اصلی، CV F 1CV Macro به عنوان معیار ثانویه، Test Balanced Accuracy به عنوان معیار سوم و CV F 1Test Macro به عنوان معیار نهایی تعیین شده‌اند. این معیار به گونه‌ای طراحی شده که هم توانایی تعمیم (CV) و هم مقاومت به عدم تعادل (Balanced) را در نظر بگیرد. نتایج نشان می‌دهد تفاوت عملکرد بین سه مدل برتر (XGBoost، Gradient Boosting و Random Forest) بسیار کم (در حد ۱ - ۲ درصد) بوده و در عمل، هر یک از این مدل‌ها می‌توانند به عنوان مدل برتر استفاده شوند. با توجه به اینکه مدل XGBoost، بالاترین CV Balanced Accuracy در بین همه ۸ مدل را دارد (0.8091 ± 0.0169) به عنوان مدل برتر معرفی می‌شود. همان‌طور که در جدول ۴ نشان داده شده است XGBoost در طول ۵ معیار مختلف Validation-Cross، بهترین عملکرد متوسط را در

curacyBalanced Ac داشت. Balanced Accuracy به طور خاص برای داده‌هایی همچون داده‌های این پژوهش که با عدم تعادل روبه‌رو هستند معیار مناسب‌تری نسبت به Accuracy است. علاوه بر این، XGBoost در Macro1CV F نیز با مقدار 0.8106 ± 0.0162 بالاترین عملکرد را دارد. تحلیل ماتریس درهم‌ریختگی (شکل ۱۳) نیز نشان می‌دهد اگرچه مدل XGBoost در تشخیص هر سه کلاس عملکرد خوبی دارد؛ اما در تشخیص عابران پیاده بهترین عملکرد را داشته و بیشترین خطای مدل بین موتورسیکلت و سواری است. مهم‌ترین دلیل این تفاوت می‌تواند در همپوشانی ویژگی‌های (مانند سن جوان، وقوع در بزرگراه، ساعت شب) دو مدل سواری و موتورسیکلت و تمایز ویژگی‌های عابران (مانند سن بالاتر، وقوع در گذرگاه یا خارج از گذرگاه، نوع ضربه) باشد. جدول ۵ نیز جزئیات عملکرد مدل ایکس‌جی‌بوست را برای هر کلاس نشان می‌دهد.



شکل ۱۳. ماتریس درهم‌ریختگی مدل برتر ایکس‌جی‌بوست (نوع متوفی)

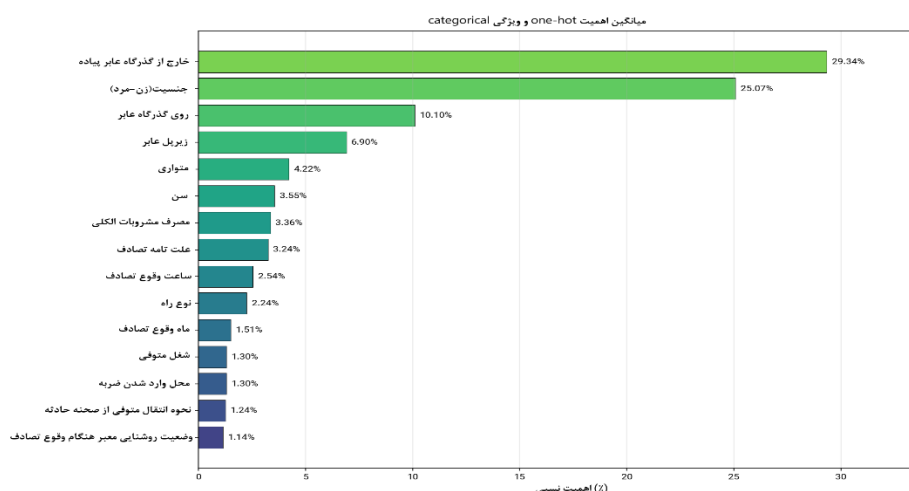
جدول ۵. عملکرد مدل ایکس‌جی‌بوست

طبقه	دقت	فراخوانی	مقدار F1
سواری	۰.۸۵۳۵	۰.۷۸۸۲	۰.۸۱۹۶
عابر	۰.۸۵۳۷	۰.۸۷۶۱	۰.۸۶۴۳
موتورسیکلت	۰.۷۸۹۲	۰.۸۲۰۲	۰.۸۰۴۴

یکی دیگر از خروجی‌های مهم در مدل یادگیری ماشین، اهمیت ویژگی‌ها است. این معیار نشان می‌دهد هر ویژگی چقدر در تصمیم‌گیری مدل مؤثر بوده است. همان‌طور که شکل ۱۴ نشان می‌دهد مهم‌ترین پیش‌بینی‌کننده‌های نوع متوفی، ویژگی‌هایی هستند که مستقیم به عابران پیاده مربوط می‌شوند (خارج از گذرگاه، روی گذرگاه، زیر پل عابر). این یافته منطقی است، زیرا این ویژگی‌ها به طور ذاتی منحصراً برای عابران صدق می‌کنند و بنابراین قوی‌ترین نشانه‌های عابر بودن متوفی هستند. ویژگی‌های دیگر مانند جنسیت و متواری شدن مقصر به طور غیرمستقیم با نوع متوفی مرتبط هستند. شایان یادآوری است که ویژگی‌هایی مانند «سن»، «ساعت» و «ماه» نیز در رتبه‌های بعدی قرار دارند و هر یک به میزان کمتری در تعیین نوع متوفی نقش دارند.

طبقه‌بندی نوع راه

برای پیش‌بینی نوع راه نیز از ۱۹ ویژگی (با اضافه شدن نوع متوفی به عنوان ویژگی ورودی و حذف متغیر نوع راه جهت جلوگیری از نشت داده) استفاده شد (جدول ۶).



شکل ۱۴. اهمیت ویژگی‌ها در پیش‌بینی نوع متوفی در مدل برتر ایکس‌جی‌بوست

جدول ۶. ویژگی‌های ورودی دو مسئله

ردیف	ویژگی	برای پیش‌بینی نوع متوفی	برای پیش‌بینی نوع راه
۱	جنسیت	✓	✓
۲	سن (عددی)	✓	✓
۳	تحصیلات	✓	✓
۴	وضعیت روشنایی	✓	✓
۵	ساعت وقوع	✓	✓
۶	ماه وقوع	✓	✓
۷	نحوه انتقال متوفی	✓	✓
۸	محل وارد شدن ضربه	✓	✓
۹	محل فوت	✓	✓
۱۰	شغل متوفی	✓	✓
۱۱	جهت مسیر حرکت	✓	✓
۱۲	علت تأیید تصادف	✓	✓
۱۳	نوع راه	✓ (ویژگی)	X (هدف - حذف شد)
۱۴	سازمان مالک وسیله	✓	✓
۱۵	مصرف مشروبات الکلی	✓	✓
۱۶	متواری	✓	✓
۱۷	روی گذرگاه	✓	✓
۱۸	زیر پل عابر	✓	✓
۱۹	خارج از گذرگاه	✓	✓
۲۰ (مورد ۱۹)	نوع متوفی (اضافه شده)	X (هدف)	✓ (ویژگی اضافه شد)

بررسی توزیع کلاس‌ها در وظیفه پیش‌بینی نوع راه بیانگر عدم تعادل شدید در توزیع داده‌ها (بزرگراه ۵۸/۸ درصد، خیابان اصلی ۳۷/۵ درصد، خیابان فرعی تنها ۳/۷ درصد) است، به طوری که نسبت کلاس اکثریت به کلاس اقلیت حدود ۱۶ به ۱ است. در نتیجه، مدل‌های سنتی بدون در نظر گرفتن این عدم تعادل، به طور سیستماتیک به نفع کلاس اکثریت سوگیری می‌کنند و کلاس اقلیت (خیابان فرعی) را نادیده می‌گیرند. برای مقابله با عدم تعادل کلاس‌ها ۱۴ روش کلاسیک و پیشرفته بررسی و مقایسه شده‌اند. این دسته‌بندی بر اساس مکانیسم اصلی هر روش برای مقابله با عدم تعادل انجام شده است. شش دسته اصلی عبارت‌اند از: Baseline (بدون مقابله با عدم تعادل)، Class Weighting (وزن‌دهی به کلاس‌ها)، Oversampling (افزایش نمونه‌های اقلیت)، Hybrid (ترکیب undersampling و oversampling)، Ensemble تخصصی (الگوریتم‌های مخصوص داده‌های نامتعادل)، و در

نهایت sensitive-Cost و Tuning Threshold (تنظیم هزینه خطا یا آستانه تصمیم‌گیری). لازم به ذکر است که در همه روش‌ها از همان پایه restRandom Fo به‌جز مواردی که ماهیت روش متفاوت است (مانند XGBoost یا RUSBoost که AdaBoost است) استفاده شد. جدول ۷ نتایج همه ۱۴ روش را در ۸ معیار مختلف نشان می‌دهد. این جدول به گونه‌ای مرتب شده که بهترین روش‌ها (بر اساس دقت متعادل) در بالا قرار دارند. ستون آخر نشان‌دهنده توانایی مدل در شناسایی طبقه اقلیت (خیابان فرعی) است. بر اساس این جدول تنها روشی که توانست فراخوانی کلاس خیابان فرعی را به طور معنادار افزایش دهد (از ۰ به ۰/۳۳۳۳)، جنگل تصادفی اصلاحی بود. در ادامه برای اطمینان بیشتر روش‌های ensemble تخصصی (leEasyEnsemb، Balanced Bagging، RUSBoost) که پیشرفته تر هستند مورد آزمایش قرار گرفتند. علاوه بر این، XGBoost با weight_sample و روش Tuning Threshold نیز برای تکمیل مقایسه اجرا شدند (درمجموع، ۱۴ روش). خروجی‌ها بیانگر آن بود که EasyEnsemble و BalanceCascade هم به طور کامل مشکل عدم تعادل را حل نکردند. EasyEnsemble به همان عملکرد جنگل تصادفی اصلاحی رسید (فراخوانی خیابان فرعی = ۰/۳۳۳۳) و Balanced Bagging حتی ضعیف‌تر عمل کرد (فراخوانی خیابان فرعی = ۰/۰۹۵۲). اما روش (Threshold Tuning) توانست فراخوانی اقلیت را به ۰/۵۲۳۸ برساند که بیش از ۵۰ درصد است، اما این موضوع باعث کاهش شدید دقت مدل (به ۰/۴۱) شد. در نتیجه، می‌توان گفت که روش‌های رفع عدم تعادل تا حدی مشکل را بهبود بخشیدند، با این حال، حتی Balanced RF و EasyEnsemble هم نتوانستند به فراخوانی طبقه اقلیت بالای ۰/۳۳۳۳ دست یابند. این یعنی یک‌سوم از نمونه‌های خیابان فرعی به‌درستی شناسایی می‌شوند، که هنوز هم فاصله زیادی با عملکرد ایده‌آل دارد. این نتایج نشان‌دهنده آن است که در داده‌های با عدم تعادل به‌شدت شدید، نمی‌توان هم‌زمان دقت و فراخوانی بالا برای کلاس اقلیت داشت. انتخاب بین این دو به کاربرد عملی بستگی دارد. اگر هدف شناسایی هر خیابان فرعی حتی به بهای هشدارهای کاذب زیاد باشد، Threshold Tuning مناسب است. اما اگر هدف دقت کلی مدل است، باید پذیرفت که کلاس اقلیت نادیده گرفته می‌شود. بنابراین، پیشنهاد می‌کند که در پژوهش‌های آتی به جای تلاش برای بهبود مدل، روی جمع‌آوری ویژگی‌های متمایزکننده بیشتر برای خیابان فرعی (مانند عرض خیابان، تعداد باندها، تراکم ساختمانی) تمرکز شود.

جدول ۷. مقایسه نتایج استراتژی‌های رفع عدم تعادل در مسئله نوع راه

ردیف	روش	دقت	دقت متعادل	میانگین وزنی 1F بر اساس تعداد هر کلاس	میانگین ساده 1F هر کلاس	فراخوانی بزرگراه	فراخوانی اصلی	فراخوانی فرعی
۱	Balanced Random Forest	۰,۵۴۹۵	۰,۴۷۰۴	۰,۵۷۸۱	۰,۴۲۶۱	۰,۶۴۷	۰,۴۳۱	۰,۳۳۳
۲	-RF + Borderline SMOTE	۰,۶۵۱۹	۰,۴۳۰۱	۰,۶۲۹۳	۰,۴۲۳۹	۰,۸۲۸	۰,۴۶۲	۰,۰۰۰
۳	XGBoost + Class Weight	۰,۶۲۰۱	۰,۴۲۹۳	۰,۶۲۰۹	۰,۴۳۱۰	۰,۷۴۷	۰,۴۹۳	۰,۰۴۸
۴	Baseline RF (بدون مقابله)	۰,۶۴۸۴	۰,۴۲۷۲	۰,۶۲۴۱	۰,۴۱۹۷	۰,۸۲۸	۰,۴۵۳	۰,۰۰۰
۵	RF + ADASYN	۰,۶۴۳۱	۰,۴۲۵۳	۰,۶۲۲۲	۰,۴۱۹۲	۰,۸۰۹	۰,۴۶۷	۰,۰۰۰
۶	RF + SMOTE	۰,۶۳۴۳	۰,۴۱۹۳	۰,۶۱۳۹	۰,۴۱۳۴	۰,۸۰۰	۰,۴۵۸	۰,۰۰۰
۷	TomekRF + SMOTE	۰,۶۳۴۳	۰,۴۱۹۳	۰,۶۱۳۶	۰,۴۱۲۶	۰,۸۰۰	۰,۴۵۸	۰,۰۰۰
۸	RF + Threshold Tuning (۰,۰۵۰=t)	۰,۴۰۶۴	۰,۴۱۸۶	۰,۴۵۴۰	۰,۳۱۹۱	۰,۵۷۲	۰,۱۶۰	۰,۵۲۴
۹	RF + Class Weight	۰,۶۳۲۵	۰,۴۱۵۱	۰,۶۰۶۸	۰,۴۰۷۳	۰,۸۱۹	۰,۴۲۷	۰,۰۰۰
۱۰	SMOTE-RF + SVM	۰,۶۲۷۲	۰,۴۱۵۱	۰,۶۰۶۸	۰,۴۰۸۳	۰,۷۸۷	۰,۴۵۸	۰,۰۰۰
۱۱	EasyEnsemble	۰,۴۶۱۱	۰,۴۰۴۷	۰,۵۰۲۹	۰,۳۵۸۱	۰,۵۸۸	۰,۲۹۳	۰,۳۳۳
۱۲	Balanced Bagging	۰,۵۵۱۲	۰,۳۹۹۱	۰,۵۷۳۴	۰,۳۹۹۳	۰,۶۵۳	۰,۴۴۹	۰,۰۹۵
۱۳	RF + SMOTEENN	۰,۴۸۵۹	۰,۳۹۳۸	۰,۴۹۶۲	۰,۳۶۹۵	۰,۴۰۳	۰,۶۳۶	۰,۱۴۳
۱۴	RUSBoost	۰,۵۸۸۳	۰,۳۹۲۶	۰,۵۷۷۴	۰,۳۸۸۰	۰,۷۱۶	۰,۴۶۲	۰,۰۰۰

مدل خوشه‌بندی میانگین کا

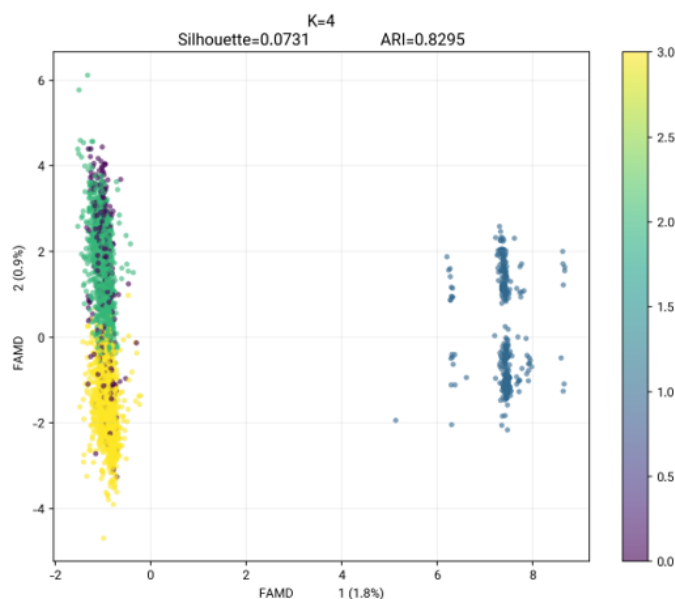
برای کشف الگوهای پنهان در داده‌ها نیز از الگوریتم خوشه‌بندی میانگین کا استفاده شد. ابتدا برای مقابله با وزن‌دهی نامتناسب ناشی از انکدینگ متغیرهای دسته‌ای، سه گروه متغیر بر اساس ماهیت آماری‌شان به طور متفاوت پیش‌پردازش شدند. به این ترتیب که متغیرهای پیوسته (سن، ساعت) با StandardScaler به میانگین صفر و واریانس یک استاندارد شدند؛ متغیرهای ترتیبی (گروه سنی، تحصیلات، بازه زمانی، فصل) با نرمال‌سازی به بازه [۱,۰] کدگذاری شدند تا ترتیب سطوح حفظ شود؛ و متغیرهای باینری و One-Hot (جنسیت، مصرف الکل، منطقه) در مقیاس طبیعی ۱/۰ خود بدون تبدیل باقی ماندند. بررسی میانگین و واریانس هر گروه پس از این پیش‌پردازش (به ترتیب ۱/۰۰۰، ۰/۰۹۶ و ۰/۰۵۷) نشان می‌دهد متغیرهای باینری وزن نسبی کمتری نسبت به متغیرهای پیوسته در محاسبه فاصله دارند. این موضوع یک محدودیت شناخته‌شده در خوشه‌بندی داده‌های ترکیبی است و پیشنهاد می‌شود در پژوهش‌های آتی به کمک فاصله Gower رفع شود. برای ارزیابی پایداری خوشه‌بندی نیز، Means-K با ۱۰ seed تصادفی متفاوت (۴۲ تا ۵۱) و با دو معیار آرنج و مقدار سیلوئت برای هر کا اجرا شد. سپس میزان توافق بین نتایج به‌دست‌آمده از seedهای مختلف با معیار ARI (Adjusted Rand Index) مقایسه شد. ARI معیاری در بازه [۱, -۱] است که ۱ نشان از توافق کامل، ۰ نشان از توافق تصادفی و مقادیر منفی نشان از توافق کمتر از تصادفی داشته و یک خوشه‌بندی پایدار باید ARI بالای ۰/۸۰ داشته باشد. این روش به ما نشان می‌دهد آیا نتایج خوشه‌بندی به انتخاب seed اولیه حساس هستند یا خیر؛ نتایج محاسبه در جدول ۸ خلاصه شده است. همان‌طور که مشاهده می‌شود، دو معیار به نتایج متفاوتی منتهی می‌شوند که این خود نشان‌دهنده پیچیدگی ذاتی داده‌ها است. بر اساس روش Elbow، نقطه آرنج در $K=4$ شناسایی شد. بالاترین Silhouette Score در $K=9$ به دست آمد. اما همان‌طور که مشاهده می‌شود، تفاوت‌های Silhouette بین کاهای مختلف بسیار کم است (بین ۰/۰۵ تا ۰/۰۸) که این نشان می‌دهد هیچ کایی در این داده‌ها به طور واضح برتری ندارد. این یافته نشان می‌دهد در انتخاب کای نهایی باید علاوه بر معیارهای کمی، پایداری و قابلیت تفسیر معنایی خوشه‌ها را نیز در نظر گرفت. با توجه به بالاترین پایداری در معیار آرنج ($ARI = 0/83$) و دومین مقدار پایداری در معیار سیلوئت ($0/731 = Silhouette$)، حجم نسبتاً معقول هر خوشه نسبت به سایر مقادیر کا (۱۱/۳ درصد تا ۴۴/۳ درصد) و تفکیک معنایی واضح بین خوشه‌ها (عابران مسن، عابران جوان‌تر، موتورسیکلت و ترکیب)، $K=4$ به عنوان بهترین خوشه‌بندی انتخاب شد.

جدول ۸. مقایسه دو معیار برای مقادیر مختلف کا

K	Inertia	Silhouette	پایداری (ARI)	انحراف معیار
۲	۶۷۳,۴۶۳	۰,۰۶۳۷	۰,۰۰۰±۱,۰۰۰	۰,۰۰۰
۳	۶۶۷,۸۵۱	۰,۰۶۸۳	۰,۳۶۵±۰,۷۰۵۱	۰,۳۶۵۱
۴	۶۵۰,۴۶۹	۰,۰۷۳۱	۰,۱۳۱±۰,۸۲۹۵	۰,۱۳۰۸
۵	۶۴۶,۲۹۸	۰,۰۶۶۱	۰,۲۴۷±۰,۶۶۵۲	۰,۲۴۶۸
۶	۶۳۵,۱۷۳	۰,۰۷۴۱	۰,۱۲۲±۰,۶۳۷۳	۰,۱۲۲۴
۷	۶۲۶,۸۸۲	۰,۰۵۹۲	۰,۱۳۹±۰,۶۴۱۶	۰,۱۳۹۴
۸	۶۱۸,۱۸۸	۰,۰۷۹۷	۰,۱۱۴±۰,۶۱۱۱	۰,۱۱۳۹
۹	۶۱۴,۲۸۹	۰,۰۸۰۲	۰,۱۷۷±۰,۶۴۶۵	۰,۱۷۷۴
۱۰	۶۱۳,۶۶۱	۰,۰۵۷۳	۰,۱۰۴±۰,۴۹۱۹	۰,۱۰۳۸

به طور کلی در این پژوهش مدل‌های یادگیری ماشین با تنظیمات پیش‌فرض به منظور مقایسه منصفانه اجرا شدند و بهینه‌سازی ابرپارامترها انجام نشده است؛ بنابراین دقت گزارش‌شده نمایانگر سقف عملکرد قابل‌دستیابی هر الگوریتم نیست. درخور یادآوری است که رکوردهای ناقص (۱۰/۲ درصد از داده‌های اولیه) بدون تحلیل الگوی گمشدگی حذف شدند که می‌تواند نوعی محدودیت در پژوهش حاضر محسوب شود؛ لذا، پیشنهاد می‌شود که پژوهش‌های آتی با بررسی این الگو و به‌کارگیری

روش‌های جایگذاری داده‌های گمشده به جای حذف مستقیم، از دست‌رفتن اطلاعات را کاهش دهند. در ادامه برای نمایش تأیید تفکیک‌پذیری خوشه‌ها نیز، از $FAMD^1$ استفاده شد که مخصوص داده‌های ترکیبی اسمی و ترتیبی طراحی شده است؛ نتایج نشان‌دهنده مقدار واریانس کم دو مؤلفه است (۲/۷۵ درصد)؛ که به نظر می‌رسد به دلیل بالا بودن متغیرهای ترتیبی با تعداد دسته‌های متفاوت باشد (شکل ۱۵). علاوه بر این، این یافته بر عدم قطعیت نمودار ۲ بعدی به روش $FAMD$ برای تأیید تفکیک‌پذیری خوشه‌ها و ضرورت در نظر گرفتن پایداری (ARI) و معیارهای کمی (Silhouette) تأکید دارد.



شکل ۱۵. نمایش خوشه‌ها با $FAMD$

نتایج تحلیل خوشه‌بندی نشان داد تصادفات موجود در مجموعه داده مورد مطالعه، از نظر ترکیب کاربران راه، ویژگی‌های مرتبط با نوع معبر و برخی شاخص‌های رفتاری و جمعیتی، از الگوی یکسانی پیروی نمی‌کنند. در راه‌حل چهارخوشه‌ای، ۴۷۹۸ مورد تصادف در چهار گروه طبقه‌بندی شدند که اندازه آن‌ها به ترتیب ۵۴۲، ۵۴۵، ۱۵۸۴ و ۲۱۲۷ مورد بود. این توزیع نشان می‌دهد دو خوشه کوچک‌تر هر یک حدود ۱۱ درصد از مشاهدات را تشکیل می‌دهند، در حالی که دو خوشه بزرگ‌تر در مجموع بیش از سه‌چهارم موارد را شامل می‌شوند. با وجود این تفاوت در اندازه خوشه‌ها، تفسیر اهمیت ایمنی هر گروه نباید فقط بر اساس تعداد تصادفات انجام شود. تعداد رخدادها، ترکیب کاربران راه، ویژگی‌های مکانی، شدت پیامدها و امکان اجرای راهکارها، همگی در تعیین اولویت‌های مدیریت ایمنی نقش دارند.

در خوشه صفر سهم حضور عابران پیاده ۵۹ درصد، میانگین سنی ۴۶/۹ سال و بزرگراه‌ها ۴۷/۲ درصد گزارش شده که ضرورت تمرکز راهکارها بر شناسایی نقاطی که در آن‌ها حضور عابر پیاده با تداخل حرکتی وسایل نقلیه همراه است را بیش از پیش آشکار می‌کند. در این چارچوب، ارزیابی کیفیت و محل گذرگاه‌های عابر، دیدپذیری، شرایط توقف وسایل نقلیه در نزدیکی گذرگاه و نحوه کنترل حرکت در تقاطع‌ها و بهبود عبور از بزرگراه در ساعات پرتردد روز می‌تواند در قالب ممیزی ایمنی مورد استفاده قرار گیرد.

خوشه ۱، ترکیب ناهمگنی از عابر و موتورسیکلت را نشان می‌دهد و می‌تواند بازتاب تصادفاتی باشد که به طور مشترک بین چند نوع کاربر راه رخ می‌دهند. تفسیر دقیق‌تر این خوشه نیازمند تحلیل زیرگروه یا افزودن ویژگی‌های تکمیلی (مانند نوع دقیق تقاطع) در پژوهش‌های آتی است.

خوشه ۲، یکی از دو خوشه بزرگ مجموعه است که با توجه به اندازه این خوشه (۳۳ درصد کل داده‌ها) و سهم بالای عابر پیاده (۹۲/۹ درصد)، می‌توان آن را به عنوان یکی از گروه‌های مهم برای در برنامه‌ریزی ایمنی کاربران آسیب‌پذیر راه در تهران در نظر گرفت. همچنین، بالا بودن میانگین سنی متوفیان در این خوشه نسبت به سایر خوشه‌ها می‌تواند بیانگر آسیب‌پذیری بیشتر افراد مسن در تصادفات بزرگراهی باشد. در این خوشه ویژگی فرار راننده از صحنه نگران‌کننده و نیازمند تشدید برخورد با رانندگان متواری است.

خوشه ۳ با ۴۴/۳ درصد از کل نمونه‌ها بزرگ‌ترین خوشه مجموعه است و ویژگی‌های متمایزکننده ۷۶ درصد موتورسیکلت (بالا‌ترین مقدار در بین خوشه‌ها)، میانگین سن ۳۵/۸ سال (پایین‌ترین در بین خوشه‌ها) می‌تواند راکبان موتورسیکلت جوان را نمایندگی کند. از راهکارهای متناسب برای این خوشه می‌توان به آموزش اجباری موتورسیکلت‌سواران و الزام به استفاده از کلاه ایمنی اشاره کرد. جدول ۹ جزئیات مربوط به خوشه‌ها را نشان می‌دهد.

جدول ۹. پروفایل خوشه‌ها

خوشه	تعداد	سن	ساعت	عابر	موتور	سواری	بزرگراه	الکل	فرار از صحنه
خوشه ۰	۵۴۲ (۱۱,۳٪)	۲۱,۷±۴۶,۹	۱۳,۲	۵۹,۰٪	۱۴,۶٪	۲۶,۴٪	۴۷,۲٪	۲,۴٪	۹,۴٪
خوشه ۱	۵۴۵ (۱۱,۴٪)	۱,۲±۴۰,۰	۱۳,۰	۴۳,۵٪	۳۸,۵٪	۱۸,۰٪	۵۷,۱٪	۱,۸٪	۱۲,۷٪
خوشه ۲	۱,۵۸۴ (۳۳,۰٪)	۲۰,۵±۵۲,۸	۱۳,۵	۹۲,۹٪	۳,۸٪	۳,۳٪	۵۶,۶٪	۱,۰٪	۲۳,۵٪
خوشه ۳	۲,۱۲۷ (۴۴,۳٪)	۱۵,۱±۳۵,۸	۱۲,۳	۲,۰٪	۷۶,۰٪	۲۱,۹٪	۵۷,۷٪	۲,۳٪	۶,۳٪

۴. بحث و نتیجه‌گیری

بر اساس تحلیل ۵,۳۴۷ رکورد تصادف فوتی در بازه ۱۳۹۶ - ۱۴۰۴، کاربران آسیب‌پذیر راه بخش بزرگی از قربانیان تصادفات فوتی شهر تهران را تشکیل می‌دهند. عابران پیاده با ۴۲/۸ درصد و راکبان موتورسیکلت با ۴۰/۶ درصد روی هم بیش از ۸۳ درصد متوفیان را در بر گرفته‌اند. این یافته با تعریف کاربران آسیب‌پذیر راه ارائه‌شده از سوی سازمان بهداشت جهانی و انجمن جهانی راه و با شواهد بین‌المللی مبنی بر تحمیل نامتناسب بار مرگ‌ومیر ترافیکی بر عابران و موتورسواران کاملاً همسو است [۲، ۵ و ۷]. در سطح ملی نیز این یافته سهم بالای عابران را تقویت می‌کند و با گزارش ریسک نسبی حدود ۲/۰۷ برابری مرگ عابران نسبت به رانندگان و سرنشینان همخوانی دارد [۸]. هرچند سهم بسیار بالای موتورسواران در داده‌های حاضر (۴۰/۶ درصد) نسبت به آنچه معمولاً در مطالعات سطح ملی گزارش می‌شود، نشان‌دهنده ترکیب حمل‌ونقلی خاص کلان‌شهر تهران و نیازمند اختصاص دادن وزن بیشتر به گروه در برنامه‌ریزی‌های ایمنی ترافیک است. غلبه بزرگراه‌ها به عنوان محل وقوع ۵۸/۸ درصد تصادفات فوتی، با یافته‌های پیشین درباره ارتباط بزرگراه‌ها و آزادراه‌ها با نرخ بالاتر تصادفات شدید و نقاط داغ مرگبار در تهران مطابقت دارد [۱۲]. همچنین، با مطالعه شش استانی که حضور بزرگراه را از عوامل افزایش‌دهنده شانس مرگ معرفی کرده بود، همسوست [۱۱]. این تکرار در دو بازه زمانی و دو سطح جغرافیایی متفاوت (استانی و شهر تهران) نشان می‌دهد خطر بالای بزرگراه‌ها فقط یک ویژگی موردی نبوده و الگوی پایدار در داده‌های ایران است. در مقابل، سهم قابل توجه عابران در بزرگراه و خیابان‌های اصلی با یافته‌های پیشین درباره نقش خیابان‌های شهری در تصادفات مرگبار ایران هم‌راستا بوده [۸] و نشان می‌دهد نقص زیرساخت گذر عابر، محدود به بزرگراه‌ها نیست. روند سالانه فوتی‌ها نیز تصویری متفاوت از دو خط سیر پیشین ارائه می‌دهد و کاهش مرگ‌ومیر ترافیکی گزارش‌شده برای سال‌های ۲۰۰۴ تا ۲۰۱۱ [۹] در داده‌های حاضر دیده نمی‌شود؛ برعکس، از ۱۴۰۰ تا ۱۴۰۴ تعداد متوفیان با نرخ رشد ۲۶/۳۷ درصدی در سال آخر به بالاترین سطح خود (۷۸۶ نفر) رسیده است. این الگو

با شواهد جدیدتر مبنی بر بازگشت روند صعودی مرگ‌ومیر ترافیکی ایران پس از ۲۰۱۷ همخوانی دارد [۱۰] و در واقع تداوم و حتی تشدید همان روند صعودی را در سطح شهر تهران و در بازه‌ای متأخرتر (تا ۱۴۰۴) را نشان می‌دهد. تمرکز فوتی عابران در ساعات‌های عصر و شب و فوتی موتورسواران در ساعات‌های اوج ترافیک، با یافته‌ای که ساعت وقوع را مهم‌ترین پیش‌بین در تصادفات کاربران آسیب‌پذیر معرفی کرده بود، همسو است [۱۷]. همچنین، سهم قابل توجه سالمندان بالای ۶۰ سال و جوانان بین ۲۰ تا ۲۹ سال با الگوی شناخته‌شده در مطالعاتی که سن، جنسیت و سرعت مجاز را با شدت جراحت مرتبط دانسته‌اند، سازگار است [۱۵]. بی‌توجهی به جلو به عنوان اصلی‌ترین علت تصادفات فوتی نیز با تأکید مطالعات داخلی بر نقش رفتار راننده و عوامل محیطی و رفتاری هم‌زمان در مرگ‌ومیر ترافیکی ایران قابل تبیین است [۸ و ۱۱]. در بخش مدل‌سازی پیش‌بینانه، برتری مدل‌های مبتنی بر درخت و تجمیعی بر مدل‌های خطی و ساده، با شواهد گسترده‌ای که مدل‌های جمعی درختی را در پیش‌بینی شدت و نوع تصادف برتر از مدل‌های پارامتریک کلاسیک معرفی کرده‌اند، مطابقت دارد [۱۳ و ۱۵]. این یافته با نتیجه مطالعاتی که ده مدل یادگیری ماشین را بر داده‌های تصادف اتیوپی آزمود و جنگل تصادفی را برترین مدل معرفی کرد نیز هم‌راستا است. با این حال، در پژوهش حاضر ایکس‌جی‌بوست اندکی بر سایر مدل‌ها در پیش‌بینی نوع متوفی پیشی گرفت، در حالی که برای پیش‌بینی نوع راه، جنگل تصادفی اصلاحی بهترین عملکرد را داشت؛ این تفاوت جزئی بین دو مسئله طبقه‌بندی، با این نکته از ادبیات که بهترین مدل بسته به ساختار داده و مسئله تغییر می‌کند، سازگار است و نشان می‌دهد نمی‌توان یک الگوریتم واحد را برای همه مسائل پیش‌بینی تصادف توصیه کرد [۱۳]. با توجه به عدم توازن شدید در داده‌های مربوط به نوع راه، ۱۴ روش کلاسیک و پیشرفته مقابله به کار گرفته شد و تنها روشی که توانست بدون کاهش شدید دقت مدل (دقت ۰/۴۷۰۴)، فراخوانی کلاس خیابان فرعی را به طور معنادار افزایش دهد (از ۰ به ۰/۳۳۳۳)، جنگل تصادفی اصلاحی بود. این مشکل با آنچه در ادبیات تحقیق درباره تأثیر عدم توازن کلاس‌ها بر عملکرد مدل‌های پیش‌بینی شدت تصادف تأکید شده است همخوانی دارد [۸] با این تفاوت که در پژوهش حاضر این مسئله برای اولین بار در قالب پیش‌بینی نوع راه (و نه فقط شدت تصادف) بررسی شده است. اهمیت نسبی متغیرها نیز الگوی شناخته‌شده‌ای را بازتولید کرد؛ معیارهای خارج از گذرگاه، روی گذرگاه، زیر پل عابر (به عنوان ویژگی‌هایی که مستقیم به عابران پیاده مربوط می‌شوند) در کنار جنسیت، متواری شدن مقصر، سن، مصرف مشروبات الکلی، علت و ساعت وقوع تصادف و نوع راه به عنوان مهم‌ترین پیش‌بین‌ها ظاهر شدند که با مجموعه پایدار متغیرهای پیش‌بین معرفی شده در مطالعات پیشین (نوع راه، سرعت، روشنایی، نوع وسیله، سن، جنسیت و نوع برخورد) هم‌راستاست [۱۶]؛ و با یافته‌ای که نوع برخورد و نوع وسیله را از عوامل حساس در پیش‌بینی شدت تصادف دانسته، مطابقت دارد [۱۳]. خوشه‌بندی میانگین کا نیز چهار الگوی متمایز را آشکار کرد که مقایسه آن‌ها نشان می‌دهد تحلیل خوشه‌ای می‌تواند در کنار روش‌های متعارف تحلیل تصادفات، به شناسایی گروه‌هایی با ویژگی‌های متفاوت کمک کند. خوشه‌های ۰ و ۲، با توجه به سهم بالای حضور عابران پیاده، زمینه‌ای برای بررسی اختصاصی ایمنی کاربران آسیب‌پذیر راه فراهم می‌کنند. خوشه‌های ۱ و ۳ نیز با توجه به سهم بالاتر موتورسیکلت، امکان تمرکز بر الگوهای مرتبط با این گروه از کاربران را فراهم می‌سازند. در این میان، خوشه ۲ به دلیل سهم بالاتر فرار از صحنه، نیازمند بررسی تکمیلی این ویژگی در ارتباط با مکان، شدت و نوع تصادف است. بر اساس مجموع یافته‌های توصیفی و تحلیلی می‌توان گفت که الگوهای فردی، زمانی و محیطی حاکم بر تصادفات فوتی تهران به روشنی سهم بالای عابران و موتورسواران، تمرکز مکانی در بزرگراه‌ها و خیابان‌های اصلی، و روند صعودی اخیر را نشان می‌دهند. توجه ویژه به دو گروه عابران پیاده معمولاً مسن در بزرگراه‌ها و موتورسواران نسبتاً جوان در بزرگراه‌ها و معابر اصلی در برنامه‌ریزی‌های ایمنی شهر تهران توصیه می‌شود.

منابع

- Godthelp H, Ksentini A. Specific road safety issues in low- and middle income countries (LMICs): an overview and some illustrative examples. *Traffic Saf Res.* 2024;8:e000068. doi:10.55329/sdtu9515
- World Health Organization. Road traffic injuries \[Internet\]. Geneva: World Health Organization; \[cited September2026\]. Available from: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
- Tehran Municipality, Deputy of Transportation and Traffic. Selected statistics on the identification and treatment of traffic accident black spots in Tehran in 2024 \[Persian\]. General Directorate of Traffic Engineering and Safety, Deputy of Transportation and Traffic, Tehran Municipality; 2025\.
- Tehran Municipality, Deputy of Transportation and Traffic. Fatal traffic crash analysis report, 2025 \[Persian\]. General Directorate of Traffic Engineering and Safety, Deputy of Transportation and Traffic, Tehran Municipality; 2026\.
- Ahmed SK, Mohammed MG, Abdulqadir SO, El-Kader RA, El-Shall N, Chandran D, et al. Road traffic accidental injuries and deaths: a neglected global health issue. *Health Sci Rep.* 2023;6. doi:10.1002/hsr2.1240
- Naci H, Chisholm D, Baker T. Distribution of road traffic deaths by road user group: a global comparison. *Inj Prev.* 2009;15:55-9. doi:10.1136/ip.2008.018721
- Olszewski P, Szagała P, Rabczenko D, Zielińska A. Investigating safety of vulnerable road users in selected EU countries. *J Safety Res.* 2019;68:49-57. doi:10.1016/j.jsr.2018.12.001
- Yousefifard M, Toloui A, Ahmadzadeh K, Gubari MI, Neishaboori AM, Amraei F, et al. Risk factors for road traffic injury-related mortality in Iran; a systematic review and meta-analysis. *Arch Acad Emerg Med.* 2021;9. doi:10.22037/aaem.v9i1.1329
- Bahadorimonfared A, Soori H, Mehrabi Y, Delpisheh A, Esmaili A, Salehi M, et al. Trends of fatal road traffic injuries in Iran (2004–2011). *PLoS One.* 2013;8. doi:10.1371/journal.pone.0065198
- Taherian MR, Maleki F, Talebi M, Varnosfaderani MR, Mirtorabi S, Forouzesh M, et al. Rising trend in traffic accident mortality in Iran after a decade of decline (2006–2022): time to raise the alarm. *BMC Public Health.* 2025;25. doi:10.1186/s12889-025-22905-y
- Jahanjoo F, Sadeghi-Bazargani H, Asghari-Jafarabadi M. Modeling road traffic fatalities in Iran's six most populous provinces, 2015–2016. *BMC Public Health.* 2022;22. doi:10.1186/s12889-022-14678-5
- Aghasi N. Spatiotemporal analysis of urban traffic accidents: a case study of Tehran city, *Journal of Geographic Information System.* 2018;10; 05\ DOI: 10.4236/jgis.2018.105032
- Jamal A, Zahid M, Rahman MT, Al-Ahmadi HM, Almoshaogeh MI, Farooq D, et al. Injury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study. *Int J Inj Contr Saf Promot.* 2021;28:408-27. doi:10.1080/17457300.2021.1928233
- Hamdan N, Sipos T. Advancements in machine learning for traffic accident severity prediction: a comprehensive review. *Period Polytech Transp Eng.* 2025\ doi:10.3311/pptr.40369
- Zahid M, Habib MF, Ijaz M, Ameer I, Ullah I, Ahmed T, et al. Factors affecting injury severity in motorcycle crashes: different age groups analysis using Catboost and SHAP techniques. *Traffic Inj Prev.* 2024;25:472-81. doi:10.1080/15389588.2023.2297168
- Komol MMR, Hasan MM, Elhenawy M, Yasmin S, Masoud M, Rakotonirainy A. Crash severity analysis of vulnerable road users using machine learning. *PLoS One.* 2021;16. doi:10.1371/journal.pone.0255828
- Wu S, Yuan Q, Yan Z, Xu Q. Analyzing accident injury severity via an extreme gradient boosting (XGBoost) model. *J Adv Transp.* 2021; Article ID: 771640\ doi:10.1155/2021/3771640
- Santos D, Saias J, Quaresma P, Nogueira V. Machine learning approaches to traffic accident analysis and hotspot prediction. *Computers.* 2021;10:157. doi:10.3390/computers10120157
- Shafabakhsh G, Famili A, Bahadori M. GIS-based spatial analysis of urban traffic accidents: case study in Mashhad, Iran. *J Traffic Transp Eng.* 2017;4:290-9. doi:10.1016/j.jtte.2017.05.005

20. Nafiiyah N, Wardhani R, Huda MN, Setyan AP, Prakasa E. K-means clustering for data-driven traffic accident in East Java. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika). 2025\, doi:10.29100/jipi.v10i4.9236
21. Khwajeh Salimi M, Khabiri MM, Falahnejad MS. Prediction and analysis of factors associated with road traffic accidents using the support vector machine algorithm \[Persian\]. Journal of Civil Engineering and Environmental Systems. 2019\, doi:10.22034/ceej.2019.9649
22. Mohammadpour S, Mohammadzadeh Moghadam A. Analysis of road traffic accident severity in Iran using an integrated data preprocessing approach based on machine learning algorithms \[Persian\]. 2026\, Available from: <https://isnac.ir/XGFB-KDZCH>
23. Tavakoli Kashani A., Medghalchi A., Mohammadi A., Jazvanaqi M. Applying Machine Learning Methods to Predict Crash Severity at Rural Roads- Case Study of Zanjan Province. Road. 2024;32(119):107-118 (In Persian). doi: 10.22034/road.2023.397089.2169
24. Ardakani SP, Liang X, Mengistu KT, So RS, Wei X, He B, et al. Road car accident prediction using a machine-learning-enabled data analysis. Sustainability. 2023;15.
25. Yassin S, Pooja P. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Appl Sci. 2020;2. doi:10.1007/s42452-020-3125-1