



پروژه‌ی اول
درس تحلیل رگرسیون

گامی به سوی سیاست‌گذاری هوشمند جاده‌ای

اعضای گروه:

ارشیا دهقان (۴۰۱۱۰۰۳۸۲)

ابوالفضل مسلمی (۴۰۱۱۰۰۵۰۶)

علی محمدزاده شبستری (۴۰۱۱۰۶۴۸۲)

استاد:

جناب آقای علیرضا کدیور

فروردین ۱۴۰۴

فهرست مطالب

۱. مقدمه	۳
۲. هدف پروژه	۳
۳. آشنایی با داده‌ها	۳
۳.۱. ساختار و دقت داده‌ها	۳
۳.۲. تحلیل اکتشافی داده‌ها	۴
۴. تحلیل‌ها و تفسیرهای کلیدی	۶
۴.۱. مدل‌های منتخب و چارچوب تحلیل	۶
۴.۲. تفسیر تخلفات رانندگی	۷
۴.۲.۱. تحلیل و بررسی	۷
۴.۲.۲. توصیه‌های سیاستی پیشرفته	۷
۴.۳. تفسیر استان‌ها	۸
۴.۳.۱. تحلیل و بررسی	۸
۴.۳.۲. توصیه‌های سیاستی پیشرفته	۸
۴.۴. تفسیر تعداد خودروهای عبوری	۹
۴.۴.۱. تحلیل و بررسی	۹
۴.۴.۲. توصیه‌های سیاستی پیشرفته	۹
۴.۵. تفسیر اثر سال‌ها	۱۰
۴.۵.۱. تحلیل و بررسی	۱۰
۴.۵.۲. توصیه‌های سیاستی پیشرفته	۱۰
۴.۶. تفسیر میانگین سرعت	۱۱
۴.۶.۱. تحلیل و بررسی	۱۱
۴.۶.۲. توصیه‌های سیاستی پیشرفته	۱۱
۴.۷. تفسیر اثر آب و هوا	۱۲
۴.۷.۱. تحلیل و بررسی	۱۲
۴.۷.۲. توصیه‌های سیاستی پیشرفته	۱۲
۴.۸. تحلیل و تفسیر نتایج	۱۳
۵. پیوست علمی و فنی	۱۳
۵.۱. تعریف مسئله و اهداف پروژه	۱۳
۵.۲. جمع‌آوری و آماده‌سازی داده‌ها	۱۴
۵.۳. پیش‌پردازش داده‌ها	۱۴
۵.۴. مدیریت هم‌خطی (Multicollinearity Management):	۱۵
۵.۵. انتخاب، تیونینگ و ارزیابی مدل‌ها	۱۶
۵.۶. رتبه‌بندی و انتخاب مدل‌های نهایی برای تفسیر	۱۷
۶. جمع‌بندی	۲۰
۷. محدودیت‌ها و فرصت‌های آینده	۲۰

۱. مقدمه

تصادفات جاده‌ای سهم قابل توجهی در آمار مرگ و میر سالانه دارند و پیامدهای متعددی در حوزه‌های مختلف به جا می‌گذارند. در کشورهای در حال توسعه، سهم بالای این حوادث در بیماری‌ها و هزینه‌های اقتصادی-اجتماعی، لزوم سیاست‌گذاری مبتنی بر داده را برجسته می‌سازد. در عین حال، کمبود تحلیل‌های جامع که داده‌های چندمنبعی را با رویکرد پیش‌بینی ترکیب کنند، یکی از خلأهای پژوهشی مهم در حوزه ایمنی ترافیک است. این پروژه در همین راستا تلاش می‌کند تا با استفاده از داده‌های تردد، تخلف و متوفیات، مسیر تحلیلی برای شناسایی و پیش‌بینی الگوهای پرخطر ترسیم کند.

۲. هدف پروژه

هدف اصلی این پروژه، توسعه و ارزیابی یک چارچوب مدل‌سازی آماری برای پیش‌بینی تعداد متوفیات ناشی از تصادفات رانندگی، در سطح استان و به تفکیک سال است.

با تکیه بر داده‌های چندمنبعی شامل تردد وسایل نقلیه، تخلفات، اطلاعات هواشناسی و آمار متوفیات، این پروژه تلاش می‌کند تا علاوه بر پیش‌بینی دقیق ریسک، عوامل کلیدی تأثیرگذار را نیز شناسایی کند. به‌طور خاص، تمرکز بر استفاده از مدل‌های رگرسیون با قابلیت تفسیر بالا، امکان استخراج بینش‌های قابل اعتماد را برای سیاست‌گذاری در حوزه ایمنی جاده‌ای فراهم می‌کند.

این پروژه در نهایت، گامی است در جهت تصمیم‌سازی داده‌محور برای مدیریت هوشمند ریسک در سامانه حمل‌ونقل جاده‌ای کشور.

۳. آشنایی با داده‌ها

۳.۱. ساختار و دقت داده‌ها

جدول ۱- منابع داده‌های مورد استفاده در ویژگی‌های آن‌ها

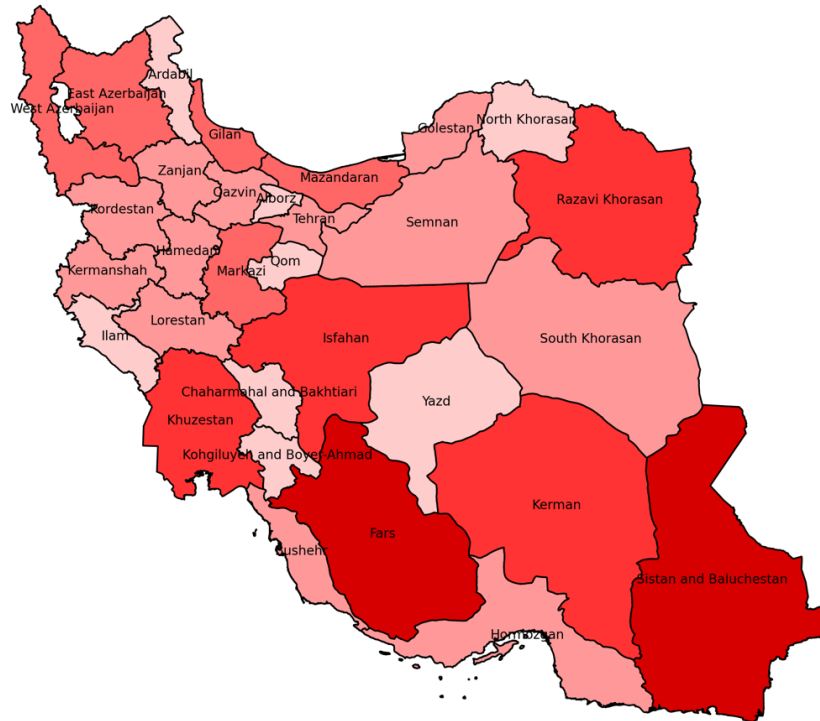
داده	منبع	ویژگی‌ها	دقت اولیه	روش جمع
ترافیکی	سامانه ۱۴۱	تردد خودروها، تخلفات، سرعت	جاده‌ای - روزانه	جمع‌زدن و میانگین‌گیری
هواشناسی	World Bank، Time & Date	دما و بارش سالانه	استانی - سالانه	نیاز به جمع ندارد
متوفیات	سازمان پزشکی قانونی	تعداد متوفیات حوادث رانندگی	استانی - سالانه	نیاز به جمع ندارد

به دلیل محدودیت در دقت زمانی و مکانی داده‌های هدف (فقط در سطح استانی-سالانه)، داده‌های ورودی نیز با رویکردهای مشخص (جمع‌زدن برای داده‌های شمارشی، میانگین‌گیری برای داده‌های پیوسته) به همین سطح جمع شده‌اند.

۲.۳. تحلیل اکتشافی داده‌ها

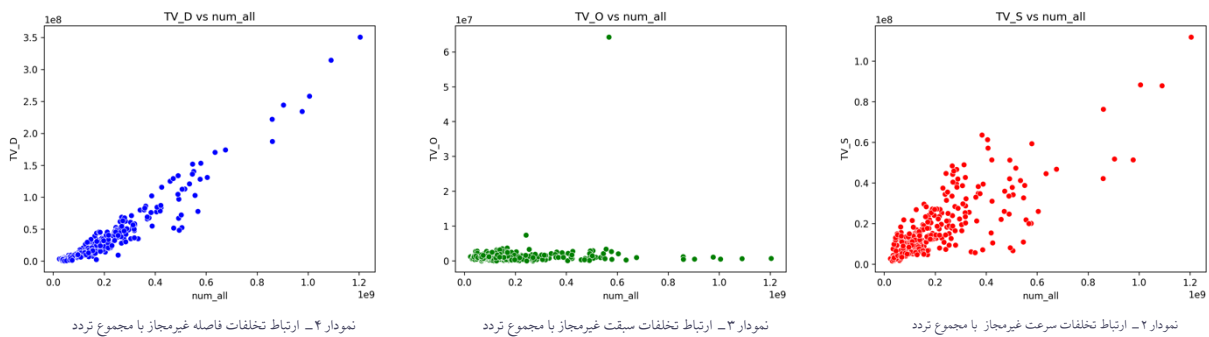
تحلیل اکتشافی داده‌ها (EDA) به درک اولیه از ساختار داده و آماده‌سازی مناسب برای مدل‌سازی کمک می‌کند.

ابتدا برای مقایسه‌ی متوفیات حوادث رانندگی در استان‌های مختلف نمودار زیر را می‌سازیم.



نمودار ۱ - پراکندگی شدت تعداد متوفیات حوادث رانندگی در استان‌های ایران

یک مسئله‌ی مهم دیگر، نحوه‌ی ارتباط سه نوع تخلف سرعت، سبقت و فاصله‌ی غیرمجاز با تعداد خودروهای عبوری است. به عبارت دیگر آیا با بیشتر شدن تعداد خودروها، تعداد تخلفات نیز بیشتر می‌شود؟

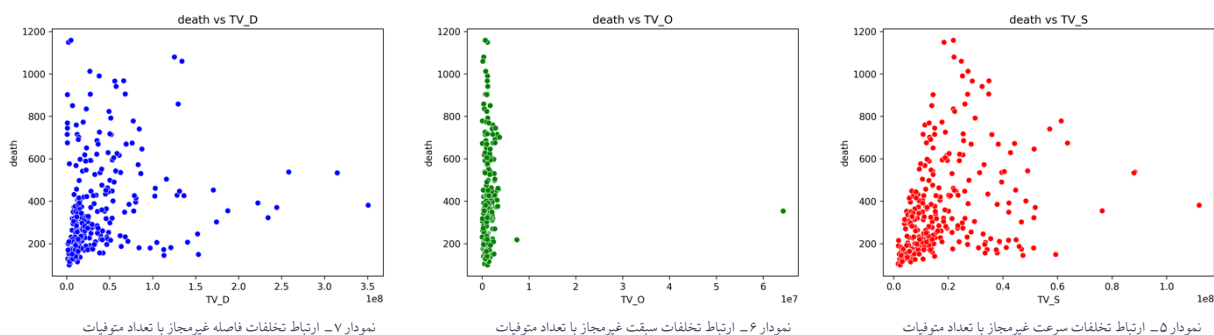


نمودار ۴ - ارتباط تخلفات فاصله غیرمجاز با مجموع تردد

نمودار ۳ - ارتباط تخلفات سبقت غیرمجاز با مجموع تردد

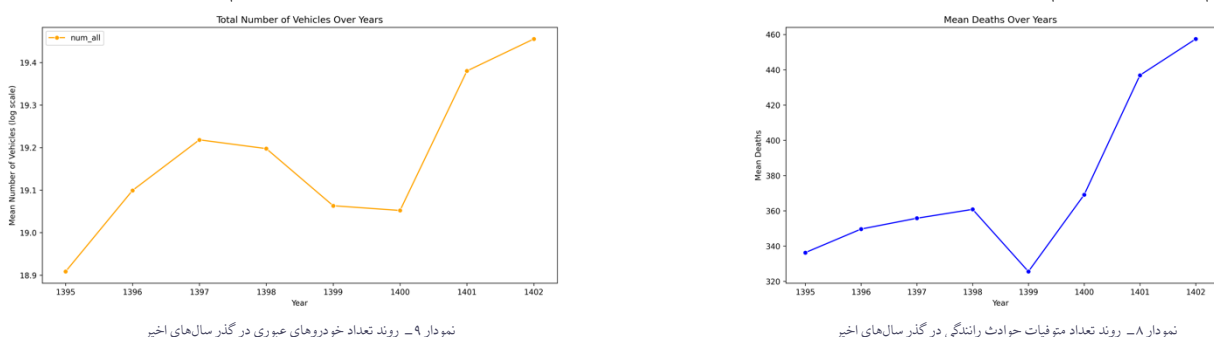
نمودار ۲ - ارتباط تخلفات سرعت غیرمجاز با مجموع تردد

در سه نمودار فوق ارتباط خطی (همبستگی) قوی و مثبتی بین آن‌ها وجود دارد. پس به طور کلی با افزایش تراکم و تعداد خودروهای عبوری، تخلفات رانندگی بیشتر می‌شود. در ادامه، بررسی می‌کنیم که آیا همبستگی و ارتباط خطی به همان قدرت بین انواع تخلفات و تعداد متوفیات وجود دارد یا خیر.

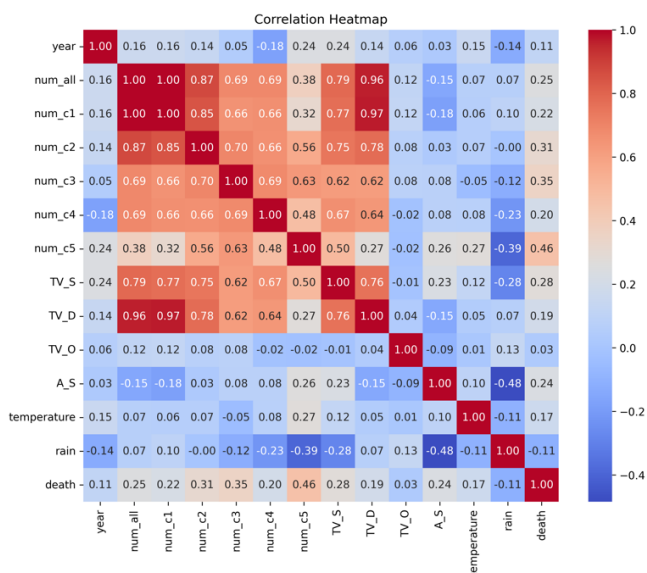


مطابق نمودارهای فوق، گاهی با تعداد تخلف تقریباً یکسان، تعداد متوفیات بسیار متفاوت هستند. پس باید سراغ مدل‌های رگرسیونی برویم که ترکیبی از چند ویژگی و عامل را کنار یکدیگر در نظر بگیرد.

همچنین می‌توانیم روند تعداد متوفیات و تعداد خودروهای عبوری در طول سال‌ها را مصور کنیم.



همانطور که انتظار می‌رفت، در اوایل دوره‌ی کرونا (سال ۱۳۹۹) هر دو متغیر کاهش شدید داشته‌اند. برای پاسخ به پرسش‌های احتمالی دیگر در خصوص ارتباط متغیرها با یکدیگر، از ماتریس همبستگی استفاده می‌کنیم.



نمودار ۸- ماتریس همبستگی متغیرها (Correlation Heatmap)

علاوه بر پرسش‌هایی از جنس ارتباط بین متغیرها یا روند تغییرات آن‌ها در گذر زمان، نمایش توزیع آن‌ها (مثل نقشه‌ی استان‌ها) می‌تواند مفید باشد. در ادامه به تحلیل رگرسیون می‌پردازیم.

۴. تحلیل‌ها و تفسیرهای کلیدی

۴.۱. مدل‌های منتخب و چارچوب تحلیل

پس از ارزیابی دقیق و چندمرحله‌ای ۱۳ مدل مختلف، دو مدل **ElasticNet** و **Lasso** به‌عنوان گزینه‌هایی با بیشترین پایداری، دقت، و قابلیت تفسیر انتخاب شدند. عملکرد مطلوب این دو مدل از نظر دقت، تعادل بین خطا و پیچیدگی (bias-variance)، و ثبات ضرایب، پایه استنتاج‌های تحلیلی این مطالعه را تقویت می‌کند.

در این گزارش، تفسیر ضرایب، مقادیر p-value، و بازه‌های اطمینان ۹۵٪ برای متغیرهای کلیدی ارائه شده و برای هر دسته از ویژگی‌ها تحلیل عددی همراه با پیشنهادهای سیاستی ارائه گردیده است.

در اینجا مقایسه‌ی عملکرد همه‌ی مدل‌ها را مشاهده می‌کنید. جزئیات کامل فرآیند ارزیابی و نمودارهای مقایسه‌ای دو مدل برتر در پیوست فنی آمده است.

جدول ۲- اطلاعات آماری تارگت در داده‌های آموزشی و تست (برای مقایسه با خطاهای جدول ۲ و نمودارهای ۱ الی ۴)

Train Y Mean	Train Y STD	Test Y Mean	Test Y STD
378.37	231.76	351.22	200.54

جدول ۳- مقایسه عملکرد مدل‌های رگرسیون بر اساس معیارهای مختلف ارزیابی شامل RMSE و R^2 در داده‌های آموزش، تست، اعتبارسنجی و بوت‌استرپ

Model	Train RMSE	Test RMSE	Crossval RMSE	Train R2	Test R2	Crossval R2	R2 (Bootstrap 95% CI)	RMSE (Bootstrap 95% CI)
CatBoost Regressor	60.1	66.37	114.25 ± 35.40	0.93	0.89	0.76 ± 0.09	0.883, 95% CI: [0.827, 0.918]	66.26, 95% CI: [51.46, 80.55]
Decision Tree Regressor								
Gradient Boosting Regressor	103.77	105.01	160.77 ± 36.04	0.8	0.72	0.50 ± 0.12	0.714, 95% CI: [0.599, 0.821]	103.96, 95% CI: [73.73, 136.68]
LGBM Regressor	70.52	101.33	175.18 ± 31.55	0.91	0.74	0.40 ± 0.10	0.720, 95% CI: [0.498, 0.829]	101.17, 95% CI: [78.34, 122.63]
Random Forest Regressor	60.68	101.72	158.09 ± 41.12	0.93	0.74	0.52 ± 0.16	0.734, 95% CI: [0.614, 0.835]	100.30, 95% CI: [69.13, 131.25]
XGBoost Regressor	98.03	102.62	163.99 ± 41.29	0.82	0.73	0.49 ± 0.16	0.773, 95% CI: [0.648, 0.857]	91.96, 95% CI: [69.71, 113.22]
K-Nearest Neighbor	61.85	81.23	112.66 ± 14.72	0.93	0.83	0.72 ± 0.10	0.813, 95% CI: [0.428, 0.974]	75.27, 95% CI: [33.68, 131.13]
ElasticNet Regressor	49.19	48.99	71.98 ± 13.50	0.95	0.94	0.89 ± 0.09	0.94, 95% CI: [0.89, 0.97]	48.7, 95% CI: [38.9, 58.3]
Lasso Regressor	49.19	48.99	71.98 ± 13.50	0.95	0.94	0.89 ± 0.05	0.94, 95% CI: [0.89, 0.97]	48.7, 95% CI: [38.9, 58.3]
Linear Regressor	49.19	49.07	71.98 ± 13.49	0.95	0.94	0.89 ± 0.05	0.94, 95% CI: [0.89, 0.97]	48.82, 95% CI: [38.99, 58.35]
Ridge Regressor	49.11	50.05	73.20 ± 12.95	0.95	0.94	0.88 ± 0.05	0.932, 95% CI: [0.882, 0.964]	49.79, 95% CI: [40.13, 59.50]
Gaussian Process Regressor	39.99	47.65	73.45 ± 14.82	0.97	0.94	0.88 ± 0.06	0.937, 95% CI: [0.880, 0.970]	47.42, 95% CI: [37.68, 58.12]
Support Vector Regressor								

جدول ۴- مقادیر R^2 حاصل از Nested Cross-Validation در تکرارهای مختلف با seedهای تصادفی متفاوت

Model	R2 (Nested CV, Seed 42)	R2 (Nested CV, Seed 52)	R2 (Nested CV, Seed 62)	R2 (Nested CV, Seed 72)	R2 (Nested CV, Seed 82)
CatBoost Regressor	0.77 ± 0.04	0.80 ± 0.05	0.76 ± 0.07	0.81 ± 0.04	0.81 ± 0.02
Decision Tree Regressor	0.28 ± 0.20	0.09 ± 0.44	0.17 ± 0.25	-0.09 ± 0.19	-0.06 ± 0.21
Gradient Boosting Regressor	0.48 ± 0.08	0.47 ± 0.07	0.43 ± 0.08	0.53 ± 0.09	0.48 ± 0.08
LGBM Regressor	0.41 ± 0.15	0.45 ± 0.04	0.41 ± 0.07	0.46 ± 0.05	0.50 ± 0.07
Random Forest Regressor	0.50 ± 0.09	0.52 ± 0.13	0.53 ± 0.08	0.59 ± 0.04	0.56 ± 0.08
XGBoost Regressor	0.54 ± 0.11	0.53 ± 0.10	0.52 ± 0.11	0.60 ± 0.09	0.57 ± 0.07
K-Nearest Neighbor	0.79 ± 0.07	0.78 ± 0.03	0.73 ± 0.03	0.79 ± 0.06	0.72 ± 0.08
ElasticNet Regressor	0.90 ± 0.04	0.90 ± 0.02	0.89 ± 0.08	0.90 ± 0.02	0.90 ± 0.03
Lasso Regressor	0.90 ± 0.04	0.89 ± 0.02	0.90 ± 0.03	0.91 ± 0.03	0.91 ± 0.03
Linear Regressor	0.90 ± 0.04				
Ridge Regressor	0.90 ± 0.04	0.89 ± 0.02	0.90 ± 0.03	0.91 ± 0.03	0.91 ± 0.03
Gaussian Process Regressor	0.90 ± 0.04	0.91 ± 0.01	0.90 ± 0.03	0.92 ± 0.03	0.93 ± 0.03
Support Vector Regressor	0.44 ± 0.12	0.48 ± 0.08	0.44 ± 0.12	0.49 ± 0.13	0.43 ± 0.06

۲.۴. تفسیر تخلفات رانندگی

جدول ۵- مقایسه ضرایب و سطوح معناداری تخلفات رانندگی در مدل‌های Lasso و ElasticNet

	Feature	Lasso Coef (p)	ElasticNet Coef (p)
0	TV_S	+29.05 (0.018)	+31.80 (0.013)
1	TV_O	-4.93 (0.417)	-4.31 (0.483)
2	TV_D	+3.73 (0.839)	+16.02 (0.521)

۱.۲.۴. تحلیل و بررسی

- **تخلف سرعت (TV_S):** این متغیر در هر دو مدل رابطه‌ای مثبت و معنادار با تعداد متوفیات دارد. ضریب مثبت و بازه اطمینان ۹۵٪ که به‌طور کامل در ناحیه مثبت قرار دارد، بیانگر اثر قابل اطمینان این عامل در سطح استانی-سالانه است. مقدار ضریب (+۳۰) نشان می‌دهد که افزایش یک انحراف معیار در تخلفات سرعت با افزایش حدود ۳۰ مورد مرگ و میر در همان سال و استان همراه است.
- **تخلف سبقت (TV_O) و فاصله (TV_D):** ضرایب این دو متغیر در هر دو مدل از نظر آماری معنادار نبودند. این امر ممکن است ناشی از جذب اثر اصلی توسط متغیر تخلف سرعت، کیفیت پایین‌تر ثبت داده‌ها، پراکندگی مکانی اثرات در سطح جاده‌ای، یا قدرت آماری محدود ناشی از حجم نمونه ($n=244$) باشد.

۲.۲.۴. توصیه‌های سیاستی پیشرفته

«تمرکز راهبردی بر مدیریت سرعت»

یافته‌ها بر اولویت کنترل سرعت غیرمجاز تأکید دارند. پیشنهادهای کلیدی عبارتند از:

- **اجرای هوشمند:** به‌جای افزایش کلی دوربین‌ها، از تحلیل داده‌ها برای شناسایی پویای نقاط و زمان‌های پرریسک سرعت (Dynamic Speeding Hotspots) استفاده شود. ترکیب دوربین‌های ثابت در نقاط مزمن با دوربین‌های سیار (پلیس، پهپاد) توصیه می‌شود.
- **جریمه‌های متغیر:** طراحی سیستم جریمه مبتنی بر شدت تخلف، سابقه راننده، و شرایط محیطی (Dynamic Fines). اجرای این طرح نیازمند زیرساخت فناوری اطلاعات و اصلاحات قانونی است.
- **فناوری‌های درون‌خودرویی و V2X:** حمایت از استفاده از سامانه‌های هشدار سرعت هوشمند (ISA) در خودروهای جدید و توسعه ارتباط خودرو با زیرساخت (V2X) برای هشدارهای محلی شده و آنی در مناطق پرخطر.

۳.۴. تفسیر استان‌ها

جدول ۶- مقایسه ضرایب و سطوح معناداری استان‌ها (پنج‌تای برتر) در مدل‌های Lasso و ElasticNet

	Province	Lasso Coef (p)	ElasticNet Coef (p)
0	Fars	+635.25 (5.17e-28)	+643.42 (2.35e-27)
1	Sistan and Baluchestan	+474.84 (6.17e-14)	+473.97 (7.59e-14)
2	Kerman	+401.81 (1.83e-13)	+399.36 (2.99e-13)
3	Razavi Khorasan	+261.08 (7.56e-08)	+263.01 (6.82e-08)
4	Isfahan	+360.84 (1.29e-07)	+357.53 (1.83e-07)

۳.۴.۱. تحلیل و بررسی

ضرایب متغیرهای استانی در هر دو مدل معنادار و نشان‌دهنده تفاوت‌های قابل‌توجه در مرگ‌ومیر، حتی پس از کنترل سایر عوامل اند. استان‌هایی مانند فارس، کرمان، و سیستان و بلوچستان ضرایب مثبت بالایی دارند.

- نقش عوامل غیرقابل مشاهده در مدل فعلی: این تفاوت‌ها احتمالاً ناشی از عوامل زمینه‌ای و ساختاری اند که در مدل اندازه‌گیری نشده‌اند، از جمله:

- زیرساخت جاده‌ای: کیفیت طراحی و نگهداری، علائم، روشنایی، حفاظ‌ها و ...
- فرهنگ رانندگی و شرایط اجتماعی-اقتصادی: الگوهای رفتاری، سطح آموزش، وضعیت اقتصادی و امدادسانی.
- نظارت و اجرای قانون: تفاوت در شدت و کیفیت اعمال قوانین توسط پلیس.
- شرایط محیطی و جغرافیایی: توپوگرافی، تراکم جمعیت، و الگوهای سفر.
- وضعیت ناوگان خودرویی: میانگین عمر و ایمنی خودروهای مورد استفاده.

۳.۴.۲. توصیه‌های سیاستی پیشرفته

«مداخلات هدف‌مند و اولویت‌بندی منطقه‌ای»

ضرایب مثبت بالا در برخی استان‌ها، ضرورت تمرکز منابع و اقدامات متناسب با شرایط محلی را نشان می‌دهد. پیشنهادهای کلیدی:

- ممیزی ایمنی: انجام ممیزی‌های دقیق برای شناسایی و اصلاح نقاط پرخطر در استان‌های پرریسک.
- تقویت نظارت: افزایش ظرفیت پلیس و بهره‌گیری از فناوری‌های نظارتی در این استان‌ها.
- آموزش منطقه‌ای: طراحی کمپین‌های آموزشی متناسب با ویژگی‌های فرهنگی-رفتاری هر منطقه.
- سیاست‌گذاری مبتنی بر ریسک: بررسی امکان لحاظ ریسک استانی در سیاست‌های کلان (مانند تعرفه بیمه یا تخصیص بودجه) با توجه به ملاحظات عدالت اجتماعی.

۴.۴. تفسیر تعداد خودروهای عبوری

جدول ۷- مقایسه ضرایب و سطوح معناداری تعداد خودروهای عبوری در مدل‌های Lasso و ElasticNet

	Feature	Lasso Coef (p)	ElasticNet Coef (p)
0	num_c2	-50.16 (0.0039)	-43.60 (0.0256)
1	num_c3	-18.16 (0.221)	-16.35 (0.277)
2	num_c4	-2.67 (0.845)	-0.86 (0.951)
3	num_c5	+18.17 (0.353)	+17.52 (0.371)
4	num_c1	—	-33.86 (0.465)

۴.۴.۱. تحلیل و بررسی

- اثر محافظتی کلاس ۲ وسایل نقلیه: تعداد بالاتر خودروهای کلاس ۲ (کامیونت، مینی‌بوس) با کاهش معنادار مرگ‌ومیر در هر دو مدل همراه بوده است. این رابطه منفی، حتی پس از کنترل سایر عوامل، نشان‌دهنده اثری غیرمنتظره و قابل تأمل است. فرضیه‌های احتمالی شامل:
 - **تعدیل سرعت جریان ترافیک:** کاهش میانگین سرعت به واسطه حضور بیشتر وسایل نقلیه کندتر (Pacing Effect).
 - **نمایندگی متغیرهای پنهان:** همبستگی با نوع خاصی از جاده‌ها یا شرایط اقتصادی
 - **رفتار رانندگان حرفه‌ای:** احتمال رعایت بیشتر مقررات توسط رانندگان این وسایل.
- **سایر کلاس‌ها:** تردد اتوبوس‌ها (num_{c3}) و کامیون‌های سنگین (num_{c4} ، num_{c5}) اثر معناداری نشان نمی‌دهد. ضریب مثبت ولی غیرمعنادار num_{c5} ممکن است به ریسک بالقوه اشاره داشته باشد، اما برای نتیجه‌گیری نیاز به داده‌های دقیق‌تر وجود دارد.

۴.۴.۲. توصیه‌های سیاستی پیشرفته

«انجام پژوهش‌های متمرکز بیشتر در این زمینه»

یافته معنادار درباره ارتباط منفی تردد خودروهای کلاس ۲ با تعداد متوفیات، فرصت مهمی برای گشودن مسیرهای تحقیقاتی جدید فراهم می‌کند. برای درک بهتر این پدیده و ارزیابی ظرفیت‌های آن در سیاست‌گذاری، مطالعات تکمیلی زیر پیشنهاد می‌شود:

- **تحلیل تفکیکی تصادفات:** بررسی اینکه چه نوع و شدت تصادفاتی در حضور یا غیاب این خودروها رخ می‌دهد، با تمرکز بر سهم احتمالی آن‌ها در کاهش تصادفات شدید.
- **کنترل متغیرهای پنهان:** شناسایی و تحلیل متغیرهای زیرساختی یا اقتصادی-اجتماعی که ممکن است با تردد این نوع خودروها همبستگی داشته و در مدل کنترل نشده باشند.

۵.۴. تفسیر اثر سال‌ها

جدول ۸- مقایسه ضرایب و سطوح معناداری سال‌ها در مدل‌های Lasso و ElasticNet

	Year	Lasso Coef (p)	ElasticNet Coef (p)
0	1402	+107.97 (4.60e-08)	+113.96 (1.19e-07)
1	1401	+101.47 (1.36e-11)	+105.61 (6.30e-11)
2	1400	+38.69 (1.89e-03)	+38.45 (2.06e-03)
3	1399	-14.19 (2.51e-01)	-16.28 (2.00e-01)
4	1398	+35.75 (2.75e-03)	+35.95 (2.65e-03)
5	1397	+31.55 (1.83e-02)	+31.26 (1.96e-02)
6	1396	+29.19 (3.84e-02)	+26.87 (6.31e-02)
7	1395	-8.19 (6.28e-01)	-13.88 (4.56e-01)

۵.۴.۱. تحلیل و بررسی

ضرایب متغیرهای سال‌ها نشان‌دهنده روند زمانی معنادار در تعداد متوفیات است، حتی پس از کنترل عوامل ترافیکی، تخلفات و استانی.

- **روند افزایشی پس از کرونا:** سال‌های ۱۴۰۱ و ۱۴۰۲ دارای ضرایب مثبت، بزرگ و بسیار معنادار هستند که نشان‌دهنده افزایش قابل توجه ریسک پایه در این دو سال است. سال‌های ۱۳۹۶ تا ۱۴۰۰ نیز ضرایب مثبت و عمدتاً معنادار دارند که نشان می‌دهد روند افزایشی ریسک احتمالاً قبل از ۱۴۰۱ شروع شده، اما در دو سال اخیر تشدید شده است. روند افزایشی تلفات ممکن است تحت تأثیر مجموعه‌ای از عوامل بیرونی باشد که در مدل‌های فعلی لحاظ نشده‌اند. از جمله این عوامل می‌توان به تغییرات پس از دوران کرونا اشاره کرد، مانند افزایش سفرها، دگرگونی در الگوهای رفت‌وآمد و تغییر رفتارهای روانی و اجتماعی رانندگان. همچنین شرایط اقتصادی نامطلوب می‌تواند به کاهش کیفیت نگهداری خودروها و افزایش رفتارهای پرریسک بین رانندگان منجر شود. فرسودگی ناوگان خودروبی و افزایش میانگین عمر وسایل نقلیه نیز از عوامل اثرگذار تلقی می‌شوند. در نهایت، فشار مضاعف بر زیرساخت‌های جاده‌ای و احتمال فرسودگی آن‌ها در برخی مسیرهای پرترافیک نیز می‌تواند نقش مهمی در تشدید ریسک ایفا کند.

۵.۴.۲. توصیه‌های سیاستی پیشرفته

«تحلیل عمیق علل روند افزایشی»

- **مطالعات چندوجهی آسیب‌شناسانه:** بررسی نقش عوامل اقتصادی، اجتماعی، روانی، وضعیت ناوگان و تغییرات در اجرای قوانین با استفاده از داده‌های مکمل.
- **تحلیل تفکیکی روند:** بررسی اینکه آیا این روند افزایشی در تمام استان‌ها یکسان بوده یا در برخی شدیدتر است؟ آیا نوع تصادفات یا گروه‌های قربانیان خاصی در این روند نقش بیشتری داشته‌اند؟

۶.۴. تفسیر میانگین سرعت

جدول ۹- مقایسه ضرایب و سطوح معناداری میانگین سرعت در مدل‌های Lasso و ElasticNet

	Feature	Lasso Coef (p)	ElasticNet Coef (p)
0	A_S	-16.33 (0.225)	-19.77 (0.166)

۶.۴.۱. تحلیل و بررسی

ضریب میانگین سرعت در هر دو مدل منفی است اما به سطح معناداری آماری نمی‌رسد ($p > 0.15$). علامت منفی در نگاه اول ممکن است غیرمنتظره باشد (انتظار می‌رود سرعت بالاتر با خطر بیشتر همراه باشد). با این حال، در حضور متغیر تعداد تخلفات سرعت (TV_S)، این یافته می‌تواند به این معنا باشد:

- نقش پراکندگی سرعت: میانگین سرعت به تنهایی شاخص مناسبی برای ارزیابی ریسک نیست؛ بلکه پراکندگی سرعت و وجود سرعت‌های بسیار بالا (که شاخص TV_S آن را بهتر نمایان می‌کند) نقش مهم‌تری در افزایش خطر ایفا می‌کنند. جاده‌هایی با میانگین سرعت بالا اما یکنواخت ممکن است ایمن‌تر از مسیرهایی با سرعت ناپایدار و تخلفات مکرر باشند.
- اثر نوع جاده: میانگین سرعت بالاتر ممکن است نشان‌دهنده جاده‌های با کیفیت بهتر (بزرگراه‌ها) باشد که علی‌رغم سرعت بالاتر، به دلیل طراحی ایمن‌تر، ممکن است نرخ تصادف کمتری (به ازای هر واحد تردد) نسبت به جاده‌های با میانگین سرعت پایین‌تر اما ناامن‌تر داشته باشند.
- اثر تجمیع: میانگین‌گیری سرعت در سطح استان و سال، بسیاری از جزئیات حیاتی مربوط به الگوهای سرعت در مقاطع و زمان‌های خاص را از بین می‌برد.

۶.۴.۲. توصیه‌های سیاستی پیشرفته

«فراتر از میانگین؛ تمرکز بر توزیع و ناهنجاری سرعت»

سیاست‌گذاری نباید صرفاً بر کاهش میانگین سرعت تمرکز کند.

- پایش و تحلیل واریانس سرعت: به جای میانگین سرعت، شاخص‌هایی مانند واریانس سرعت یا درصد خودروهای با سرعت بسیار بالا/پایین (Speed Harmonization Metrics) در مقاطع جاده‌ای پایش و تحلیل شوند. این امر نیازمند داده‌های دقیق‌تر از حسگرهای ترافیکی است.
- مدیریت فعال ترافیک: (ATM) پیاده‌سازی سیستم‌های مدیریت فعال ترافیک که از سرعت‌های مجاز متغیر (Variable Speed Limits – VSL) بر اساس شرایط ترافیکی، آب‌وهوایی و حوادث لحظه‌ای استفاده می‌کنند، می‌تواند به کاهش نوسانات خطرناک سرعت و یکنواخت‌سازی جریان ترافیک کمک کند.

۷.۴. تفسیر اثر آب و هوا

جدول ۱۰ - مقایسه ضرایب و سطوح معناداری آب و هوا در مدل‌های Lasso و ElasticNet

	Feature	Lasso Coef (p)	ElasticNet Coef (p)
0	rain	+23.45 (0.500)	+24.17 (0.488)
1	temperature	+5.74 (0.892)	+3.27 (0.939)

۷.۴.۱. تحلیل و بررسی

هیچ‌یک از متغیرهای آب و هوایی (میانگین بارش و دمای سالانه) در مدل اثری معنادار بر تعداد متوفیات نشان ندادند. این عدم معناداری به احتمال زیاد ناشی از وضوح پایین داده‌هاست، نه بی‌اهمیتی واقعی شرایط جوی. پدیده‌های آب و هوایی پرریسک مانند بارش شدید، یخبندان یا مه معمولاً موضعی و کوتاه‌مدت هستند، و تجمیع آن‌ها در مقیاس سالانه و استانی موجب از بین رفتن اثرات دقیق آن‌ها می‌شود. ضریب مثبت (غیرمعنادار) بارش می‌تواند بازتابی از افزایش بالقوه روزهای نامساعد رانندگی در سال‌های پربارش باشد، اما در این سطح از تجمیع، شناسایی دقیق چنین اثراتی ممکن نیست.

۷.۴.۲. توصیه‌های سیاستی پیشرفته

«تحلیل با استفاده از داده‌های آب و هوا با وضوح بالا»

این یافته بیش از هر چیز، ضرورت نیاز به داده‌های آب و هوایی مرتبط با جاده (RWIS) و با وضوح زمانی و مکانی بالا (ساعتی، در سطح مقاطع جاده‌ای کلیدی) را برجسته می‌کند. چرا که می‌خواهیم حدس قوی‌مان مبنی بر اثر آب و هوا را بررسی کنیم.

- توسعه زیرساخت RWIS: سرمایه‌گذاری در نصب و توسعه شبکه ایستگاه‌های هواشناسی جاده‌ای مجهز به سنسورهای بارش، دما، رطوبت، سرعت و جهت باد، وضعیت سطح جاده (خشک، تر، یخ‌زده)، و دید افقی (مه).
- ادغام داده‌های هواشناسی در مدل‌های ریسک: استفاده از داده‌های RWIS ساعتی به عنوان ورودی برای مدل‌های پیش‌بینی ریسک تصادف کوتاه‌مدت در سطح مقاطع جاده‌ای.
- سیستم‌های هشدار و توصیه‌گر پویا: توسعه سیستم‌های اطلاع‌رسانی به رانندگان از طریق تابلوهای پیام متغیر (VMS)، اپلیکیشن‌های مسیریابی، رادیو (در مورد شرایط خطرناک آب و هوایی در مسیر پیش رو و ارائه توصیه‌های مشخص (کاهش سرعت، استفاده از مسیر جایگزین، تعویق سفر).

۴.۸. تحلیل و تفسیر نتایج

تحلیل ما نشان داد که **تخلف سرعت (TV_S)** مهم‌ترین عامل قابل مداخله در افزایش تلفات رانندگی است. این یافته نه تنها با تحقیقات جهانی در حوزه ایمنی ترافیک همسو است، بلکه بر ارزش داده‌محور بودن سیاست‌های کنترل سرعت در ایران تأکید می‌کند. برخلاف تصور رایج، میانگین سرعت به‌تنهایی معیار مناسبی برای ارزیابی خطر نبود؛ این نکته بیانگر آن است که سیاست‌گذاری مؤثر باید بر ناهنجاری‌های رفتاری (مانند سرعت‌های بیش از حد و تکرار تخلف) تمرکز کند، نه صرفاً بر کاهش میانگین.

یافته غافل‌گیرکننده‌ای که در مورد **کلاس ۲ خودروها (کامیونت و مینی‌بوس)** مشاهده شد، احتمالاً نشانگر یک اثر ثانویه تعدیل سرعت یا رفتار حرفه‌ای رانندگان آن‌هاست. این موضوع نیاز به مطالعات تکمیلی دارد ولی می‌تواند بر طراحی کمپین‌های آموزشی یا اصلاح قوانین محدودیت تردد خودروهای سنگین اثر بگذارد. همچنین، تفاوت شدید ضرایب بین استان‌ها نشان می‌دهد که ایمنی راه مسئله‌ای تک‌متغیره یا ملی نیست، بلکه تابعی از زیرساخت، فرهنگ، شرایط جغرافیایی، و کیفیت اجراست. این نکته به وضوح تأکید می‌کند که **بهبتر است سیاست‌های ایمنی جاده‌ای به‌صورت منطقه‌ای و بومی‌شده طراحی شوند.**

۵. پیوست علمی و فنی

۵.۱. تعریف مسئله و اهداف پروژه

- **هدف اصلی:** تحلیل و پیش‌بینی تعداد متوفیات و مصدومین رانندگی بین‌شهری استان‌ها به عنوان نماینده‌ای از خطر رانندگی.
- **هدف ایده‌آل اولیه:** تفسیر و پیش‌بینی در سطح جاده‌ای-ساعتی.
- **اولین چالش داده:** محدودیت دقت داده تارگت (متوفیات/مصدومین) که در سطح استانی-سالانه موجود است.
- **تصمیم کلیدی:** کاهش دقت مدل از جاده‌ای-ساعتی به استانی-سالانه به دلیل محدودیت داده تارگت. این تصمیم، اگرچه دقت مطلوب اولیه را کاهش می‌دهد، اما تنها راه عملی برای استفاده از داده‌های موجود و انجام مدل‌سازی رگرسیونی است که با اهداف آموزشی و پژوهشی پروژه همخوانی دارد [۱].
- **دومین چالش داده:** داده موجود برای تعداد مصدومین به طور مجزا برای درون‌شهری موجود نبود، پس به اجبار تارگت را از جمع تعداد متوفیات و مصدومین به فقط تعداد متوفیات تغییر دادیم.
- **هدف نهایی:** تفسیر و پیش‌بینی تعداد متوفیات رانندگی بین‌شهری استان‌ها به عنوان نماینده‌ای از خطر رانندگی.

۵.۲. جمع‌آوری و آماده‌سازی داده‌ها

(۱) منابع داده:

- تردد (فیچر): دقت جاده‌ای-روزانه (141.ir)
- هواشناسی (فیچر): استانی-سالانه (climateknowledgeportal.worldbank.org, timeanddate.com)
- متوفیات/مصدومین (تارگت): دقت استانی-سالانه (lmo.ir)

(۲) تجميع داده‌ها (Aggregation):

- چالش: نیاز به تبدیل دقت داده‌های تردد (جاده‌ای-روزانه) به دقت تارگت (استانی-سالانه).
- رویکرد اتخاذ شده [۲]:
 - جمع‌کردن: برای تعداد تردد کلاس‌های مختلف خودرو و تعداد تخلفات.
 - میانگین‌گیری: برای سرعت متوسط.

(۳) تعریف فیچرها و تارگت نهایی:

- فیچرها:

تعداد خودروهای کلاس ۱ تا ۵ = $num_{c1}, num_{c2}, num_{c3}, num_{c4}, num_{c5}$

سرعت متوسط = A_S , تعداد تخلفات سبقت = TV_O , تعداد تخلفات فاصله = TV_D , تعداد تخلفات سرعت = TV_S

سال = $year$, نام استان = $province$, میانگین بارش = $rain$, میانگین دما = $temperature$

- تارگت:

$death$ = تعداد متوفیات

۵.۳. پیش‌پردازش داده‌ها

❖ حذف داده‌های ناقص [۳]:

- کل داده‌های سال ۱۴۰۳ (به دلیل نبودن داده متوفیات دوازده ماهه)
- آذربایجان غربی ۱۴۰۲، زنجان ۱۴۰۲، یزد ۱۴۰۲ و تهران ۱۴۰۰ به دلیل ناقص بودن داده تردد

تعداد دیتاپوینت (بدون ۴ مورد ناقص):

$$۴ - \text{هشت سال} \times ۳۱ \text{ استان} = ۲۴۴ \text{ دیتاپوینت.}$$

❖ تقسیم داده:

- تقسیم داده‌ها به مجموعه آموزش (Train) و آزمون (Test) با نسبت ۲۰ / ۸۰
- استفاده از $RandomState = ۴۲$ [۴].

❖ انکودینگ ویژگی‌های رسته‌ای (۸ سال و ۳۱ استان)

▪ **CatBoost, LGBM**:

استفاده از انکودر داخلی و بهینه‌ی داخلی مدل [۵].

▪ **XGBoost, RandomForest, GradientBoosting**:

استفاده از OrdinalEncoder برای استان و `pd.get_dummies()` برای سال [۶].

▪ **DecisionTree**:

استفاده از `pd.get_dummies()` برای استان و سال [۷].

▪ **Lasso, Ridge, ElasticNet, LinearRegression, KNN, GPR, SVR**:

استفاده از `pd.get_dummies()` برای استان و سال [۸].

❖ **Scaling**:

• روش: استفاده از StandardScaler روی فیچرهای عددی [۹].

• زمان اعمال: بعد از تقسیم داده [۱۰].

• مدل‌های اعمال شده: Lasso, Ridge, ElasticNet, LinearRegression, KNN, GPR, SVR [۱۱].

۵.۴. مدیریت هم‌خطی (Multicollinearity Management):

(۱) علت بررسی: با توجه به هدف اصلی پروژه یعنی تفسیر، بررسی و مدیریت هم‌خطی، برای مدل‌های Lasso, LinearRegression, KNN, GPR, SVR، اهمیت بالایی دارد تا ضرایب مدل قابل اعتماد باشند [۱۲].

(۲) روش تشخیص: محاسبه‌ی فاکتور تورم واریانس (VIF) برای فیچرهای عددی [۱۳].

(۳) فیچرهای با VIF بیش از ۱۰: num_{c1} (۴۴,۵) و TV_D (۲۳,۲) [۱۴].

(۴) تصمیم: طبق مقایسه نتیجه تست حذف هر دو فیچر (که در پایین آمده است) و اهداف پروژه تصمیم بر حذف num_{c1} شد [۱۵].

جدول ۱۲ – نتیجه‌ی حذف فیچر TV_D

feature	VIF
const	416.853983
num_c2	5.031534
num_c3	3.389973
num_c4	2.694589
num_c5	3.097035
TV_S	4.160234
TV_D	5.558066
TV_0	1.090483
A_S	1.721482
temperature	1.261639
rain	1.864629

جدول ۱۱ – نتیجه‌ی حذف فیچر num_{c1}

feature	VIF
const	483.950242
num_c1	10.657368
num_c2	7.345182
num_c3	3.439798
num_c4	2.730589
num_c5	3.373377
TV_S	4.567717
TV_0	1.073141
A_S	2.063794
temperature	1.296223
rain	1.866320

۵.۵. انتخاب، تیونینگ و ارزیابی مدل‌ها

(۱) استراتژی جامع انتخاب مدل:

- شروع با طیف وسیعی از مدل‌ها: انتخاب ۱۳ مدل اولیه که الگوریتم‌های مختلفی (خطی، درختی، مبتنی بر فاصله، مبتنی بر کرنل ...) را پوشش می‌دهند.
- مرحله اول فیلترینگ (مبتنی بر پایداری و عملکرد پایه):
 - روش: استفاده از Repeated Nested Cross-Validation روی ۵ random seed مختلف و 5-Fold خارجی و 3-Fold داخلی [۱۶].
 - تیونینگ هایپرپارامتر: استفاده از GridSearchCV در حلقه داخلی Nested CV برای یافتن بهترین پارامترها برای هر fold خارجی [۱۷].
 - معیار فیلترینگ: $(mean_{R^2} - std_{R^2}) \geq 0.40$. این معیار تضمین می‌کند که مدل نه تنها عملکرد خوبی دارد بلکه پایدار است و حتی در بدترین حالت $(mean - std)$ نیز حداقل قدرت توضیح‌دهندگی قابل قبولی ($R^2 = 0.4$) دارد [۱۸].
- در این مرحله مدل‌هایی که حذف شدند مدل‌هایی بودند که این معیار را حتی در یک seed هم رعایت نکردند [۱۹].
- مرحله دوم فیلترینگ/تیونینگ (مبتنی بر تعادل Bias-Variance و Overfitting):
 - روش: در مرحله اول بررسی دقیقی روی خروجی تابع full_report برای مدل‌هایی که از فیلتر اول عبور کردند انجام شد. این خروجی شامل Learning Curves (R^2 , RMSE)، مقایسه R^2 و RMSE بین Train/Test/CV، و فاصله بین منحنی‌های یادگیری در انتها (Rightmost Gap) است [۲۰].

در مرحله دوم بطور دستی هایپرپارامترهای هر مدل در راستای همدل کردن تعادل Bias-Variance و Overfitting تنظیم شدند و سپس مدل تیون شده دوباره با روش فیلترینگ اولیه بررسی شدند و مدل‌هایی که این معیار را حتی در یک seed هم رعایت نکردند حذف شدند [۲۱].

مدل‌های حذف شده بعد از دو مرحله فیلترینگ: SVR, DecisionTree, LGBM

(۲) نحوه ارزیابی نهایی بر روی ۱۰ مدل منتخب [۲۲]:

- ❖ **full_report function:** Raw Performance, Overfitting, Underfitting, Stability, Learning Behavior
- ❖ **Bootstrap 95% CI:** Stability of Test Estimate, Uncertainty Quantification, Robustness to Sampling Noise
- ❖ **Repeated Nested CV:** Less Biased Generalization Estimate, Overall Stability, Hyperparameter Robustness

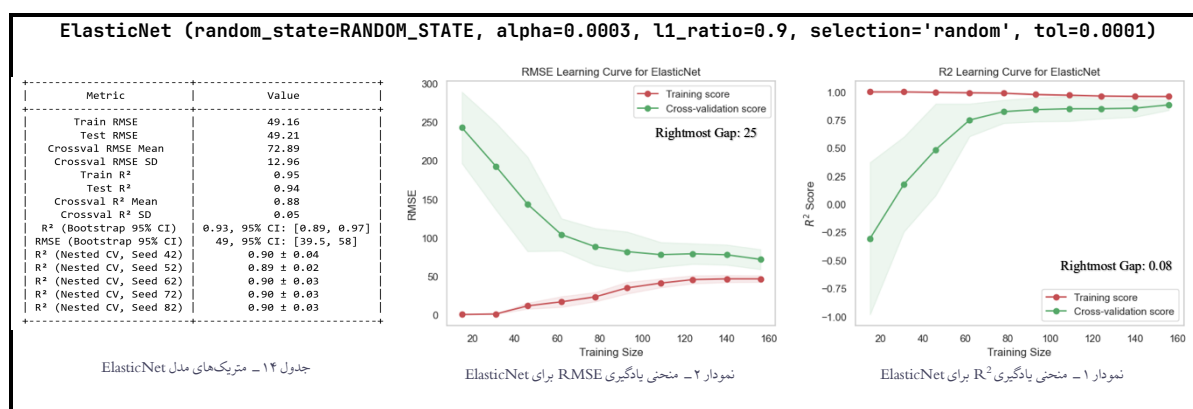
۵. ۶. رتبه‌بندی و انتخاب مدل‌های نهایی برای تفسیر

بر اساس خروجی سه متد گفته شده برای ارزیابی رتبه‌بندی مدل‌ها به شکل زیر شد (چپ ترین مدل بهترین و راست ترین مدل بدترین است):

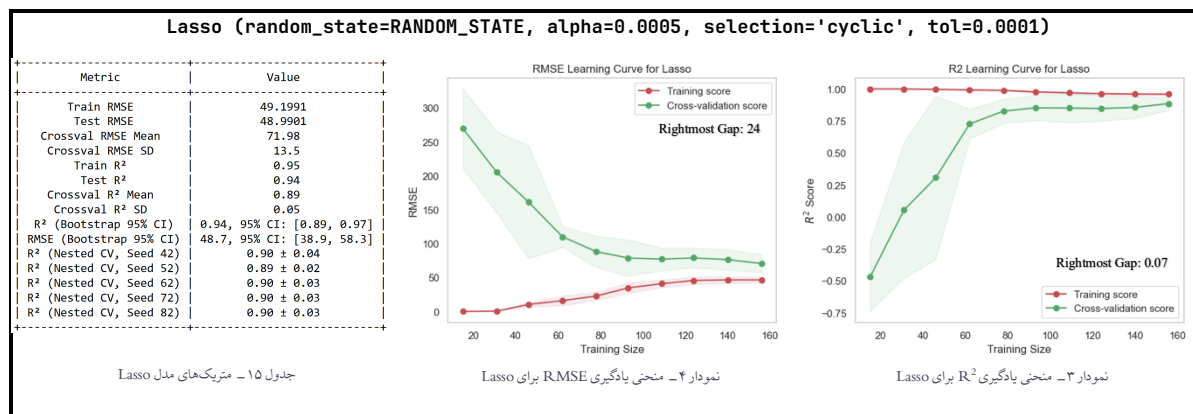
جدول ۱۳ – رتبه‌بندی مدل‌ها در گروه اول و دوم

رتبه‌بندی گروه اول (۷ مدل، بدون حذف فیچر num_{c1})	رتبه‌بندی گروه دوم (۳ مدل، با حذف فیچر num_{c1})
۱) ElasticNet	۱) Lasso
۲) Ridge	۲) LinearRegression
۳) CatBoostRegressor	۳) GaussianProcessRegressor
۴) XGBRegressor	
۵) RandomForestRegressor	
۶) KNN	
۷) GradientBoosting	

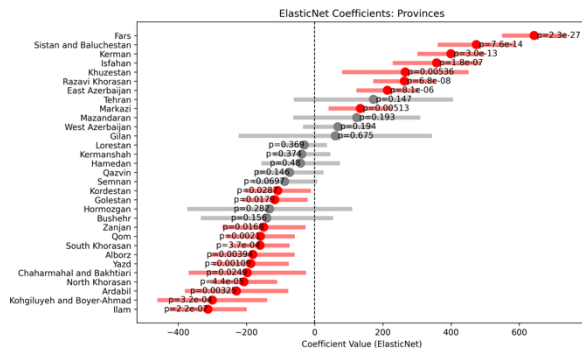
❖ منتخب گروه اول: ElasticNet



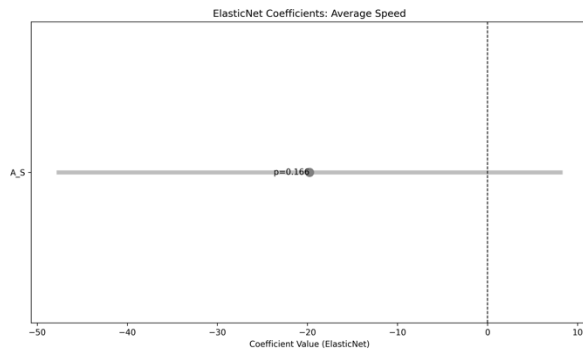
❖ منتخب گروه دوم: Lasso



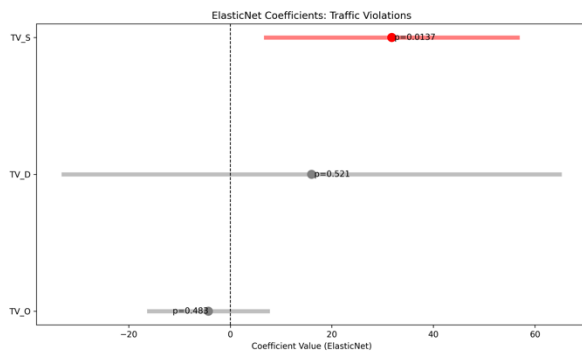
❖ نمودارهای ضرایب مدل ElasticNet



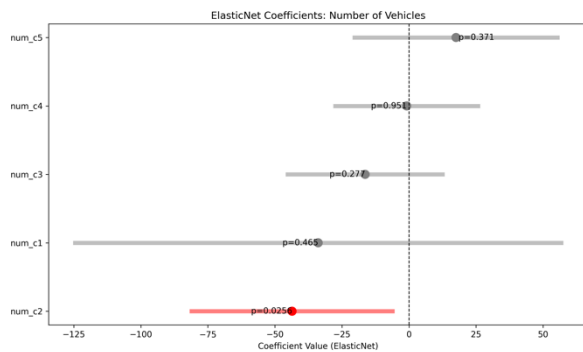
نمودار ۶- ضرایب مدل Elastic برای استان‌ها



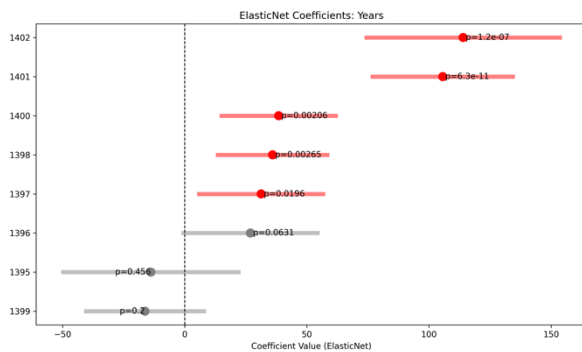
نمودار ۵- ضرایب مدل Elastic برای میانگین سرعت



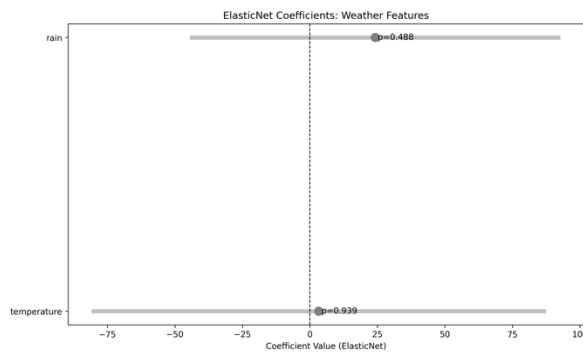
نمودار ۸- ضرایب مدل Elastic برای تخلفات رانندگی



نمودار ۷- ضرایب مدل Elastic برای تعداد خودروهای عبوری



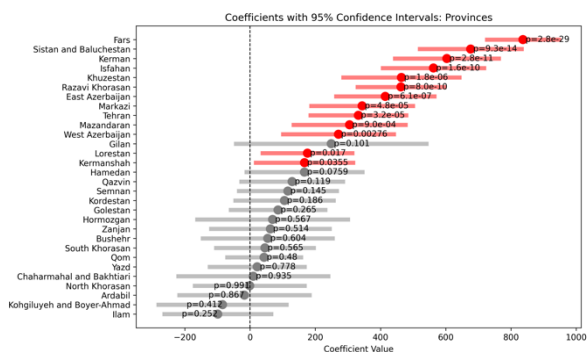
نمودار ۱۰- ضرایب مدل Elastic برای سال‌ها



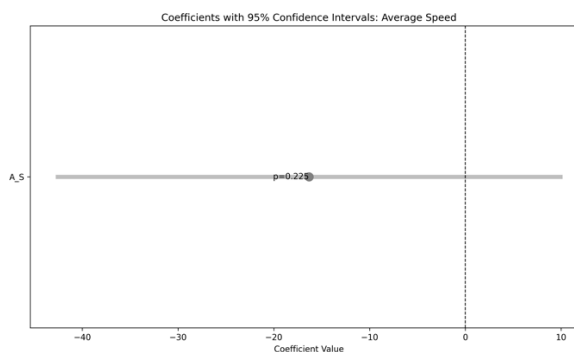
نمودار ۹- ضرایب مدل Elastic برای آب‌وهوا (دما و بارش)

تحلیل ضرایب مدل ElasticNet نشان می‌دهد که **تخلف سرعت (TV_S)** با اختلاف، مهم‌ترین عامل قابل مداخله در افزایش تلفات است و رابطه‌ای مثبت و معنادار دارد. ($\beta \approx +29$ و $p < 0.05$) در مقابل، سایر تخلفات معنادار نیستند و احتمالاً تأثیرشان توسط تخلف سرعت جذب شده است. ضرایب سال‌های اخیر (۱۴۰۱ و ۱۴۰۲) نیز مثبت، بزرگ و معنادارند، که بیانگر افزایش ریسک پایه در این سال‌هاست. از نظر جغرافیایی، استان‌هایی مثل فارس، کرمان و سیستان و بلوچستان ضرایب بسیار بالایی دارند که نشان از ضرورت مداخلات منطقه‌ای هدفمند دارد.

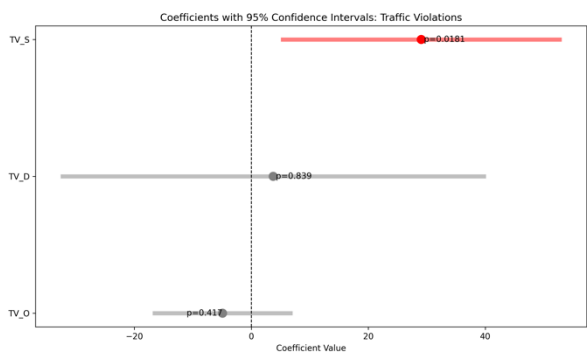
❖ نمودارهای ضرایب مدل Lasso



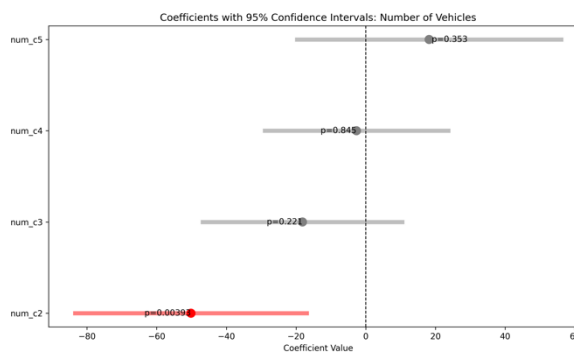
نمودار ۱۲ - ضرایب مدل Lasso برای استانها



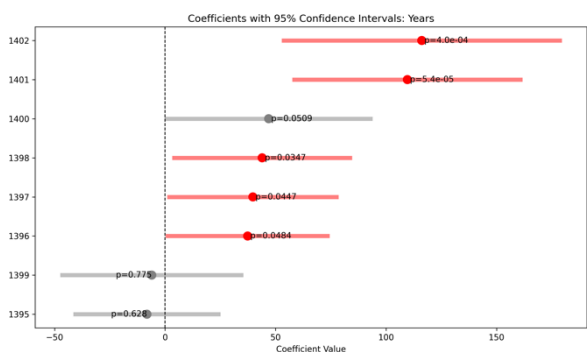
نمودار ۱۱ - ضرایب مدل Lasso برای میانگین سرعت



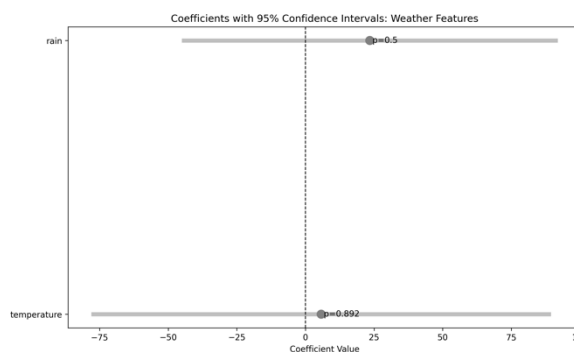
نمودار ۱۴ - ضرایب مدل Lasso برای تخلفات رانندگی



نمودار ۱۳ - ضرایب مدل Lasso برای تعداد خودروهای عبوری



نمودار ۱۶ - ضرایب مدل Lasso برای سالها



نمودار ۱۵ - ضرایب مدل Lasso برای آب و هوا (دما و بارش)

تحلیل نمودارهای ضرایب مدل Lasso تقریباً مشابه ElasticNet است.

۶. جمع‌بندی

در این پروژه، با تحلیل داده‌های تردد و تصادف در سطح استانی و سالانه، مدلی برای ارزیابی خطر جاده‌ها توسعه داده شد. نتایج نشان می‌دهند که حتی با وجود دقت محدود داده‌ها، می‌توان الگوهای معناداری از ریسک جاده‌ای استخراج کرد. با این حال، پتانسیل واقعی این مدل، در صورت دسترسی به داده‌های دقیق‌تر (مثلاً در سطح جاده‌ای و زمانی)، به مراتب بیشتر خواهد بود. مدل پیشنهادی می‌تواند در شرایط فعلی نیز به عنوان ابزاری مؤثر برای سیاست‌گذاری، تخصیص منابع، و ارتقاء ایمنی راه‌ها مورد استفاده قرار گیرد.

۷. محدودیت‌ها و فرصت‌های آینده

با توجه به محدودیت‌های موجود در سطح تجمیع داده‌ها و نداشتن دسترسی به برخی متغیرهای کلیدی، توسعه آتی پژوهش در چند محور اصلی پیشنهاد می‌شود:

- توسعه مدل به دقت جاده‌ای-ساعتی با کمک داده‌های دقیق‌تر پزشکی قانونی
- استفاده از داده‌های RWIS و تصاویر هوایی برای لحاظ کردن اثر هندسه و وضعیت جاده
- ترکیب داده‌های سنسور خودروها (VX, telematics) برای مدل‌سازی دقیق‌تر رفتار رانندگی
- طراحی مدل‌های دو مرحله‌ای: مرحله اول برای پیش‌بینی فیچرهای آینده، مرحله دوم برای پیش‌بینی خطر