# تحلیل داده های مربوط به تصادفات رانندگی

نام کارآموز:
فاطمه طاهر

اسفند ماه ۱۴۰۲

# چکیده

پیش بینی شدت تصادفات و شدت جراحات وارده بر انسان‌ها در صورتی که با روش‌های دارای مبانی علمی انجام گیرد می تواند به عنوان ابزاری مهم در مدیریت سلامت انسان‌ها و ایمنی راه‌ها و مهندسی حمل و نقل محسوب شود. در این پروژه از یادگیری ماشین برای شناسایی و پیش بینی و شدت تصادفات در راههای بین شهری ارائه شده است. در این پروژه تحلیل اکتشافی داده ها و نتایج امار توصیفی بر روی مجموعه داده مورد بررسی قرار گرفته است و به منظور پیش بینی شدت صدمات وارده بر افراد بعد از تصادف از الگوریتم های مختلفی همچون رگرسیون خطی، ماشین بردار پشتیبان، کا نزدیک ترین همسایه، درخت تصمیم، جنگل تصادفی، تقویت گرادیان و تقویت سازگار استفاده شده است.

**واژه‌های کلیدی:** تحلیل داده، رگرسیون، شناسایی الگو، تصادف.

## فهرست مطالب

# فهرست اشکال

# فهرست جداول

## مقدمه

سهم زیادی از تصادفات در جهان مربوط به کشورهای با درآمد متوسط و پایین است. در این میان، آمارهای مجروحین وفوتی های تصـادفات ایران روندی صـعودی به خود گرفته اسـت که نشـان دهنده لزوم توجه و تمرکز هرچه بیشتر بر تحلیل تصادفات ترافیکی و یافتن علل موثر بر شدت تصادفات برای ارتقاء ایمنی راه ها و کاهش پیامدهای ناشـی از آن می‌باشـد. در این پرروژه به بررسـی عوامل موثر بر شـدت تصـادفات با توجه به عوامل و فاکورهای مختلف با الگوریتم های مختلف یادگیری ماشـین پرداخته شـده اسـت. بدین منظور از مجموعه داده آمار تصادفات جاده ها طی میانه سال ۲۰۲۲ استفاده شده است. پس از فرآیند پاکسازی داده ها، مدل ها در محیط برنامه نویسی ژوپیتر نوت بوک توسعه داده شدند.

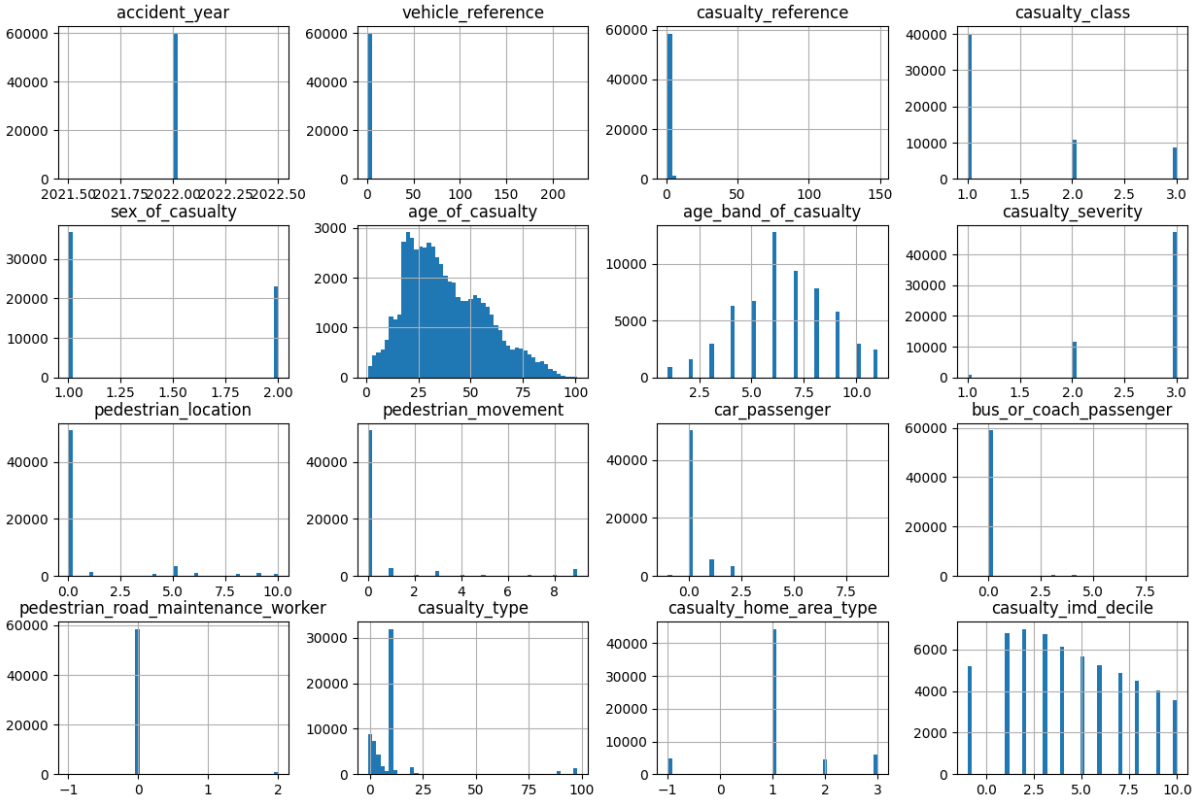
# مروری بر منابع

شدت تصادفات و تلفات جاده ای نگرانی بسیار بزرگی در جهان به ویژه در کشورهای توسعه نیافته است. درک عوامل اولیه و کمک کننده است. یاسین و همکاران در پژوهش خود مهم ترین عوامل کمک کننده برای تشخیص شدت تصادفات جاده ای را شناسایی کردند .آنها در این مطالعه با استفاده از یک رویکرد ترکیبی الگوریتم های K-means و جنگل تصادفی را برای پیش بینی شدت تلفات تصادفات جاده ای استفاده کردند. این مدل دقت ۹۹.۸۶٪ را به دست اورد و تجربه راننده، شرایط جاده، سن راننده و تعداد کارکرد خودرو را به عنوان ویژگی های موثر برای سطوح مختلف شدت شناسایی کردند. در مطالعه دیگر مارسیو و همکاران به این نتیجه رسیدند که عملکرد یک مدل پیش بینی شدت تلفات رانندگی عمدتا به کیفیت داده های آن و پیکربندی تقسیم داده های مناسب بستگی دارد. از سوی دیگر شاکیاپ و همکاران دریافتند که استفاده از شبکه های عصبی، ماشین بردار پشتیبان و الگوریتم های درخت تصمیم برای پیش بینی شدت تصادفات رانندگی عملکرد خوبی دارند. این مطالعات پتانسیل یادگیری ماشین را در پیش بینی تلفات حوادث و اهمیت کیفیت داده ها و انتخاب ویژگی ها در ساخت مدل های موثر برجسته می کند.

## تحلیل و بررسی مجموعه داده:

برای تحلیل و شناسایی عوامل موثر بر شدت صدمات وارده بر سرنشینان خودرو در تصادفات داده‌ها و ویژگی‌های مختلفی مورد نیاز است. مجموعه داده مورد استفاده در این تسک، مجموعه داده آمار تصادفات جاده ها طی میانه سال ۲۰۲۲ است. این مجموعه داده اطلاعات دقیقی در مورد تصادفات جاده ای گزارش شده طی چندین سال ارائه می دهد. مجموعه داده شامل ویژگی های مختلف مربوط به وضعیت تصادف، خودرو و شدت صدمات و تلفات است که شامل عواملی مانند جزئیات عابر پیاده، انواع تلفات، مشارکت کارکنان تعمیر و نگهداری جاده، و دهک شاخص محرومیت چندگلنه[1] (IMD) برای مناطق مسکونی قربانیان است. در ادامه به بررسی و بصری سازی برخی ستون‌های این مجموعه داده می‌پردازیم. در شکل زیر توزیع متغیرهای مجموعه داده به تصویر کشیده شده اند.

[1] Index of Multiple Deprivation



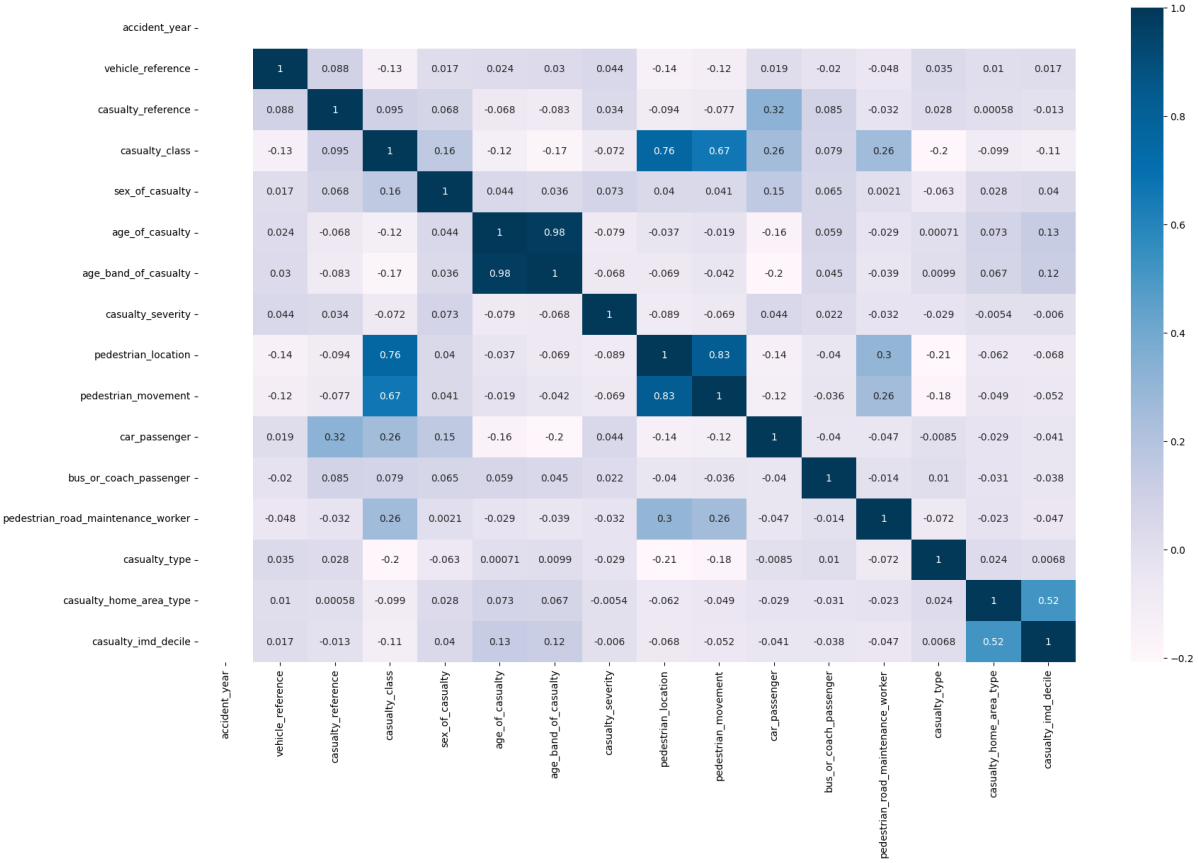

شکل ۱-۰ توزیع متغیر ها

## تحلیل اکتشافی داده‌ها و بصری سازی آن‌ها :

این مجموعه داده دارای مقادیر گمشده نمی باشد اما جنسیت و سن افراد آسیب دیده دارای مقادیر غیر نرمال هستند. به عنوان مثال جنسیت افراد دارای دو مقدار غیرعادی 1- و ۹ و سن آن‌ها دارای ۱۳۵۰ مقدار منفی می باشد. پس به همین دلیل این مقادیر از مجموعه داده حذف شده اند.

## ضریب همبستگی بین متغیر های مجموعه داده:

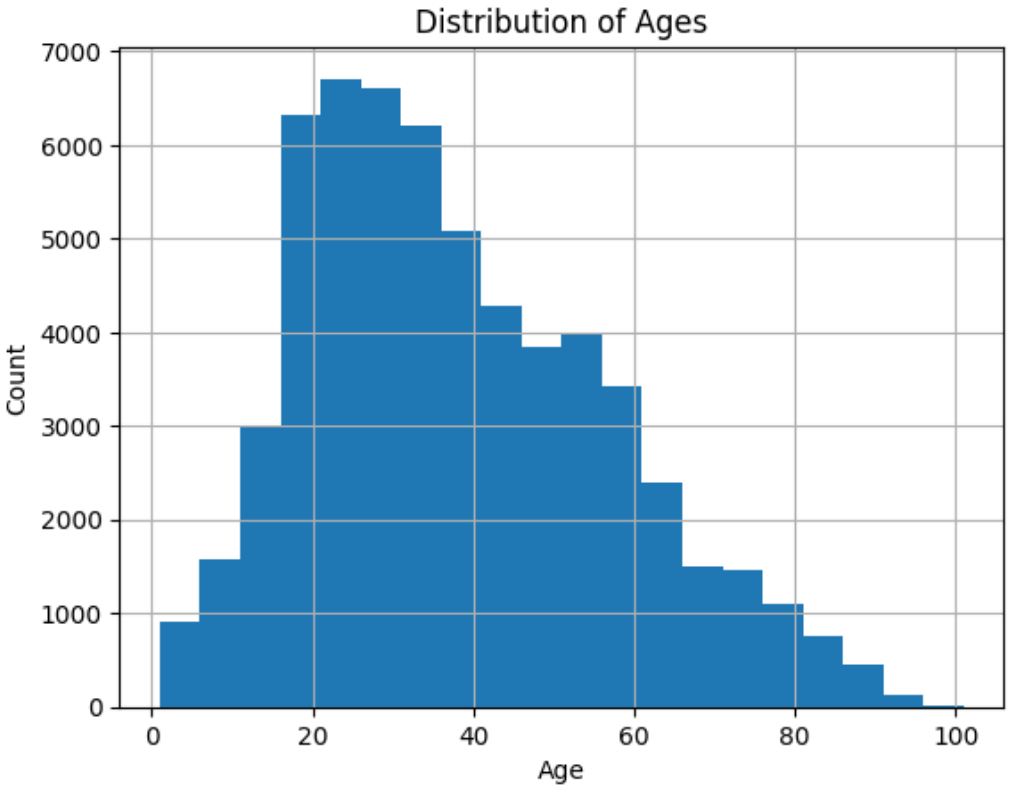
در تحلیل‌های چند متغیره آماری، روش‌های گوناگون محاسباتی برای اندازه‌گیری وابستگی یا ارتباط بین دو متغیر تصادفی وجود دارند. منظور از ضریب همبستگی بین دو متغیر، قابلیت پیش‌بینی مقدار یکی از آن متغیرها بر اساس دیگری است. به عنوان مثال، عرضه و تقاضا دو پدیده وابسته به یکدیگر هستند.



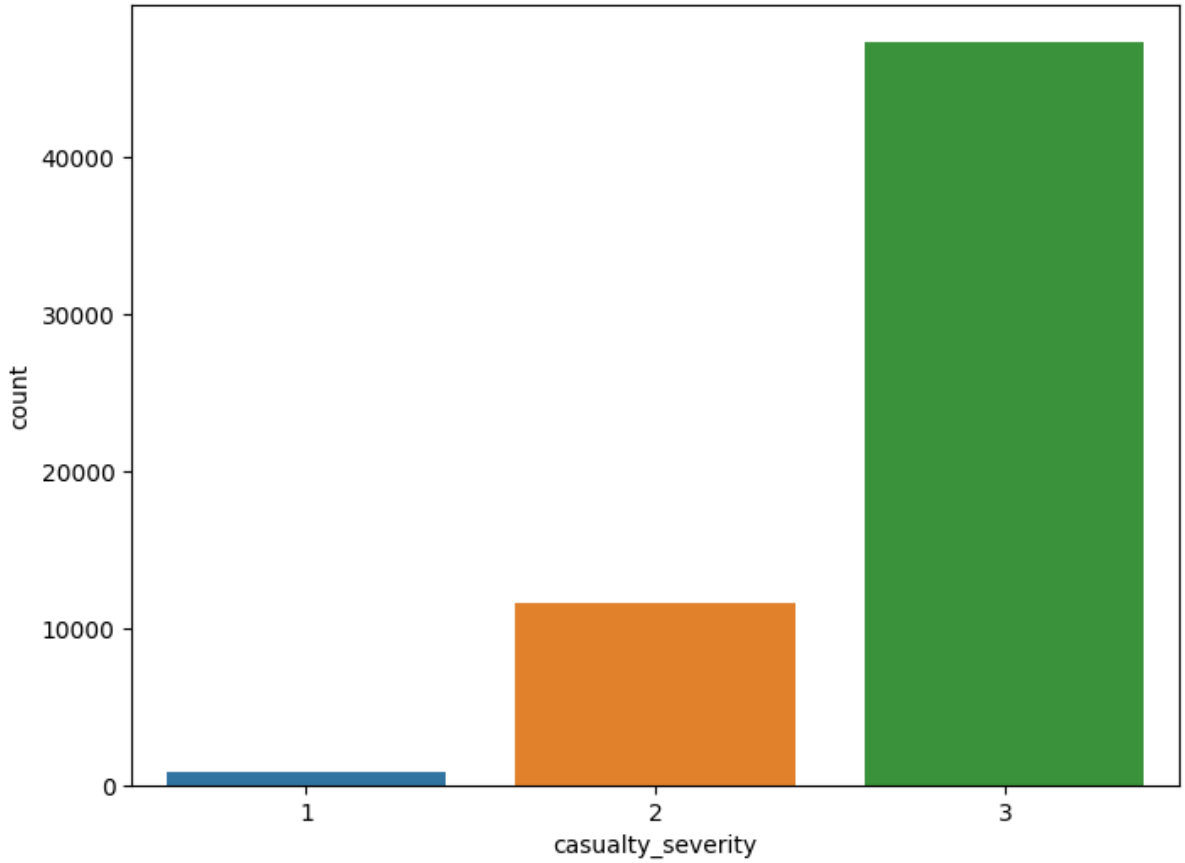

شکل ۲-۰ ضریب همبستگی بین متغیر های مجموعه داده

## ۱. توزیع سن افراد:

توزیع سن افراد آسیب دیده در نمودار فراوانی زیر نمایش داده شده است.



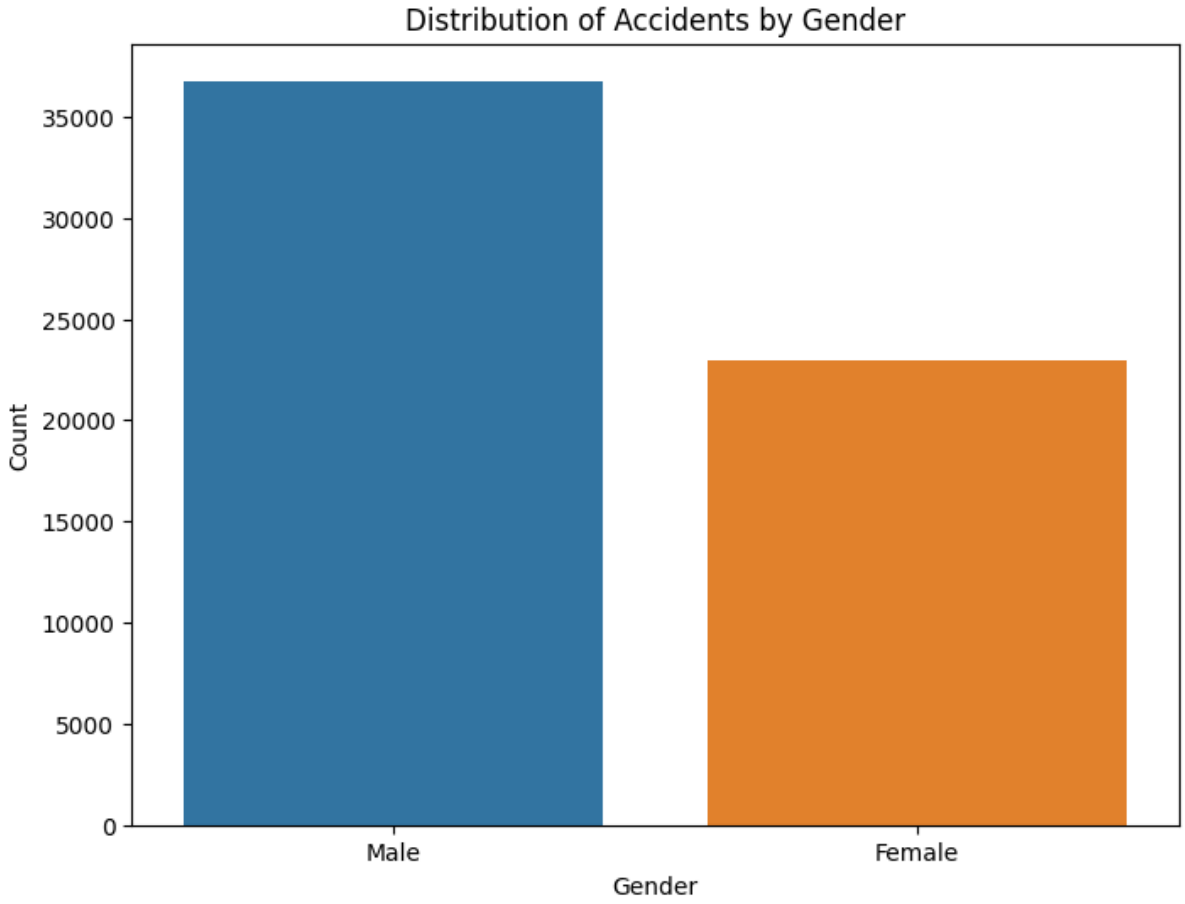

شکل ۳-۰ نمودار جعبه ای متغیر سن




شکل ٤-٠ نمودار شمارشی شدت صدمات و تلفات

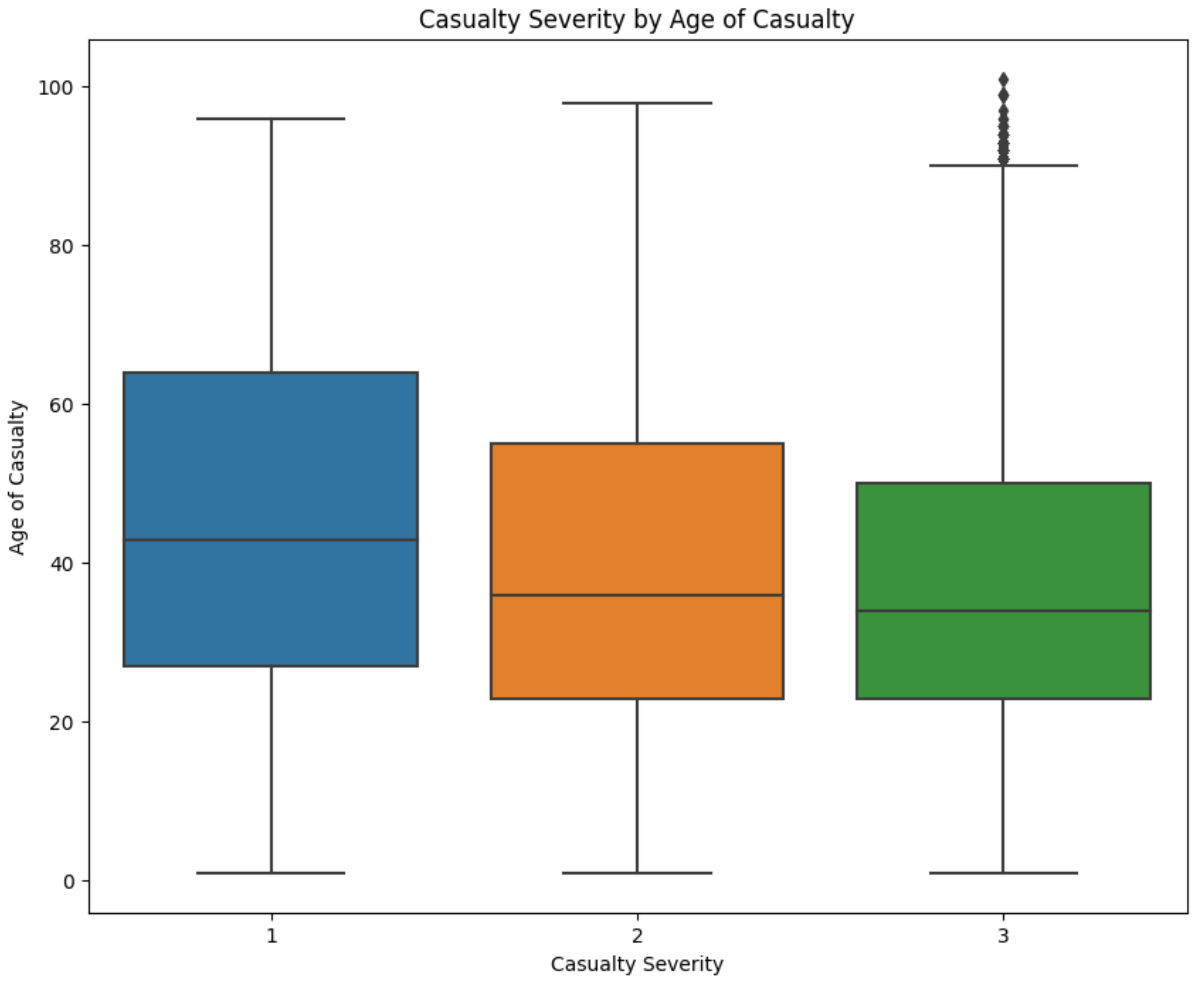
## ۲. جنسیت افراد (Gender):

برای نمایش بهتر جنســـیت، اعداد متناظر آن‌ها را با متغیر های کیفی "مونث" و "مذکر" جایگزین می‌کنیم. شکل زیر جنسیت کاربران را نمایش می‌دهد. همانطور گه مشخص است اکثر تلفات مردان بوده اند.



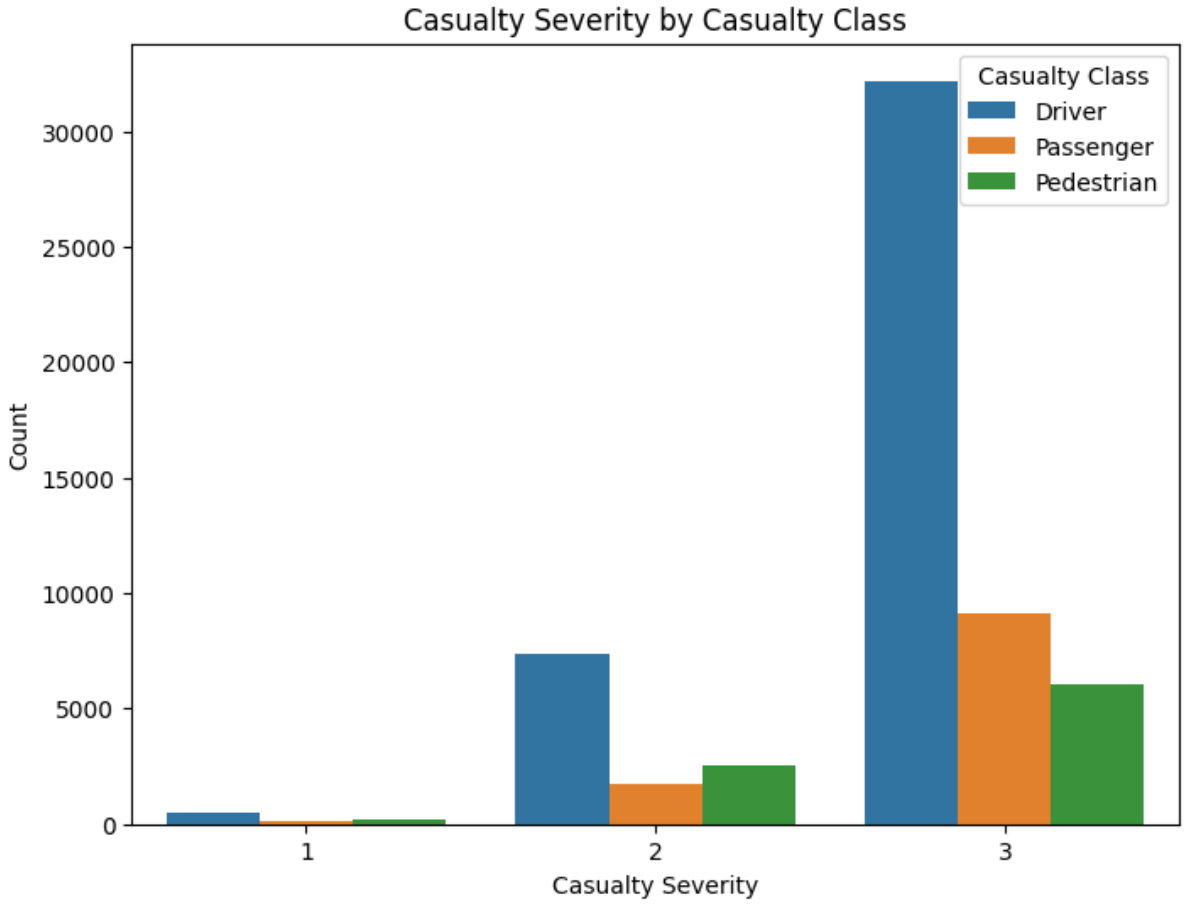

شکل ۵-۰ توزیع جنسیت افراد

در شکل زیر نمودار جعبه شدت تلفات را بر اساس سن افراد نشان داده شده است. همانطور که در شکل مشخص است سن افراد که در گروه اول جای گرفته اند حدودا ۴۴ بوده است و حدود ۱۰ داده پرت در سن افراد گروه ۳ وجود دارد.




شکل ۶-۰ نمودار جعبه ای شدت صدمات بر اساس سن

در شکل زیر شدت تلفات بر اساس کلاس تلفات نشان داده شده است که رانندگان در هر سه گروه بیشترین تلفات را داشته اند.




شکل ۷-۰ نمودار شدت تلفات بر اساس کلاس تلفات

## پیش پردازش داده ها:

به منظور پیش پردازش داده ها ۸۰ درصـــد داده ها به داده های آموزش و ۲۰ درصـــد داده ها به آزمایش اختصاص داده شده اند و به منظور مقیاس بندی و نرمال سازی ویژگی ها داده ها به میانگین صفر و واریانس واحد نرمال شده اند.

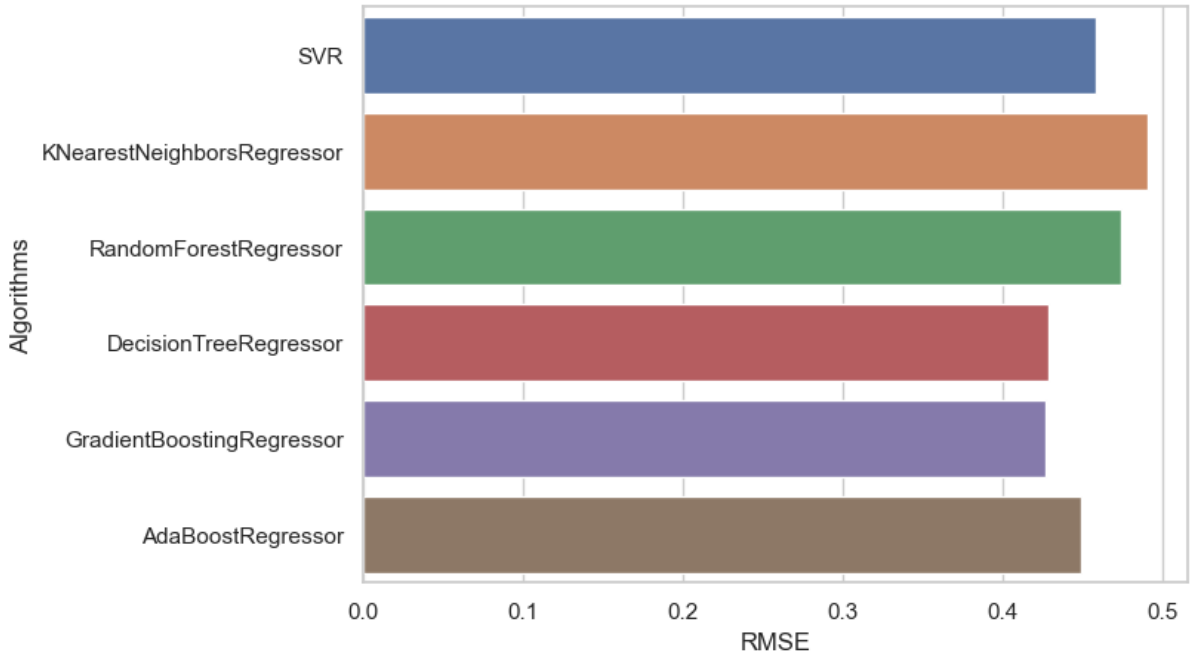
## آموزش و ارزیابی مدل‌ها:

با تجزیه و تحلیل این داده‌ها، می‌توان الگوهایی را شناسایی کرد که به ما کمک می‌کند تا بفهمیم چه عواملی شدت تلفات و صدمات را در تصادفات جادگی بیشتر می‌کنند. همچنین می‌توان این داده‌ها را برای ساخت مدل‌های پیش‌بینی استفاده کرد که می‌توانند با دقت قابل قبولی پیش‌بینی کنند که شدت صدمات احتمالی چقدر خواهند بود. پس از پاکســازی و نرمال ســازی داده ها اقدام به آموزش مدل های یادگیری ماشــین و بدســت اوردن معیار های ارزیابی مدل ها می کنیم. الگوریتم‌های مورد اســتفاده جهت شــناســایی الگوها و عوامل موثر در تصادفات عبارت اند از الگوریتم های مختلفی همچون رگرسیون خطی، ماشین بردار پشتیبان، کا نزدیک ترین همســایه، درخت تصــمیم، جنگل تصــادفی، تقویت گرادیان و تقویت ســازگار که معیارهای ارزیابی میانگین مربعات خطا[1]، جذر میانگین مربعات خطا[2] و ضــریب تعیین[3] ($R^2$) برای هر کدام از آنها محاسبه شده اند.

[1] Mean Squared Error

[2] Root Mean Squared Error

[3] Root Squared




شکل ۸-۰ مقدار جذر میانگین مربعات خطا برای الگوریتم های مختلف

همانطور که در شکل بالا مشاهده می شود مقدار جذر میانگین مربعات خطا برای الگوریتم تقویت گرادیان از دیگر الگوریتم ها کمتر اســت که این امر نشــان دهنده عملکرد بهتر و قدرت پیش بینی بالاتر این مدل می باشد. مقادیر میانگین مربعات خطا، جذر میانگین مربعات خطا و ضریب تعیین برای الگوریتم های مختلف در جدول زیر نمایش داده شده اند.

جدول ۱۰- معیار های ارزیابی برای الگوریتم های مختلف

| Algorithms | MSE | RMSE | R2 |
|---|---|---|---|
| LinearRegression | 0.191245 | 0.437315 | 0.026329 |
| SVR | 0.210265 | 0.458546 | -0.070507 |
| KNearestNeighborsRegressor | 0.241324 | 0.491247 | -0.228634 |
| RandomForestRegressor | 0.225414 | 0.474778 | -0.147634 |
| DecisionTreeRegressor | 0.322791 | 0.568147 | -0.643406 |
| GradientBoostingRegressor | 0.181990 | 0.426602 | 0.073448 |
| AdaBoostRegressor | 0.230170 | 0.479760 | -0.171848 |

## نتیجه گیری و پیشنهادات:

به منظور ساخت مدلی برای پیش بینی شدت صدمات و تلفات حادثه رانندگی در جاده ها مدل هایی از جمله رگرسیون خطی، ماشین بردار پشتیبان، کا نزدیک ترین همسایه، درخت تصمیم، جنگل تصادفی، تقویت گرادیان و تقویت سازگار به کار گرفته شدند و الگوریتم تقویت گرادیان بهترین عملکرد را میان الگوریتم های انتخاب شده داشت.

در حالت کلی عملکرد یک مدل پیش بینی شدت تلفات رانندگی نیازمند بدست اوردن ویژگی هایی با ضریب همبستگی بالا می باشند و عملکرد بهتر مدل رگرسیون عمدتا به کیفیت داده های ان و پیکربندی داده های مناسب بستگی دارد. به همین دلیل استفاده از شبکه های عصبی به دلیل اینکه می توانند الگوهای پیچیده و زیادی را بین داده ها کشف کنند پیشنهاد میشود . از سوی دیگر ماشین بردار پشتیبان و الگوریتم های درخت تصمیم نیز برای پیش بینی شدت تصادفات رانندگی عملکرد بهتری نسبت به دیگر الگوریتم های یادگیری ماشین دارند.

# پایان