گزارش پروژه طبقه بندی داده ها

نام : علیرضا یوسفی سعید

طرح مسئله

در این پژوهش داده های مربوط به تصادفات ایالت کایفرنیا را بررسی می‌کنیم. عامل های متعددی در این تصادفات نقش دارند ولی عامل اصلی مد نظر ما مصرف مشروبات الکلی بوده و بر همین اساس از کل داده تصادفات قسمت مد نظر را جدا نموده‌ایم. حال مسئله اصلی اینجا مطرح می‌شود که چه عواملی بر شدت این تصادفات اثرگذارند؟ آیا می‌توان مدلی قابل اعتماد و کارا در این زمینه توسعه داد که با دقت خوبی بتواند این تصادفات را بر اساس شدت آن‌ها دسته‌بندی کند؟

در این پژوهش سعی بر این داریم به این سوال‌ها پاسخ دهیم.

تحلیل کاوشگرانه داده ها

در مرحله اول سه ویژگی cnty_rte ، CASENO ، CCC را به دو دلیل حذف می‌نماییم:

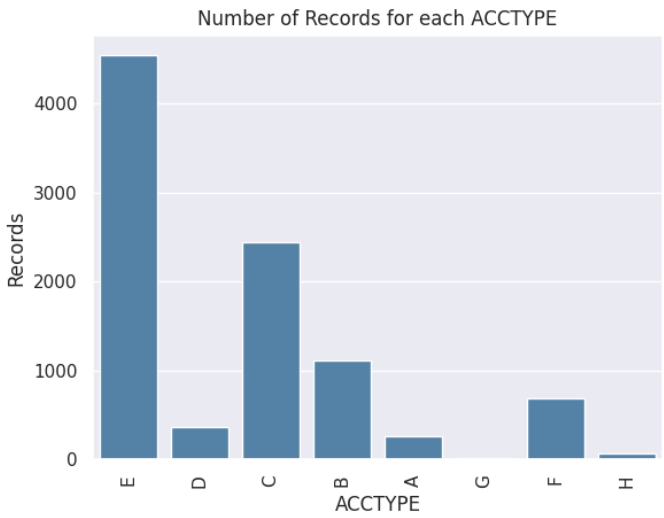
- بی اهمیت بودن نسبت به خروجی هدف(SEVERITY)
- افزایش کارایی مدل

پس از صدا کردن متد ()Describe متوجه می‌شویم که انحراف معیار ویژگی ACCYR صفر بوده و مستقیما هیچ تاثیری بر روی خروجی هدف نمی‌گذارد و صرفا از کارایی مدل کم می‌کند. پس تصمیم به حذف این ویژگی می‌گیریم.

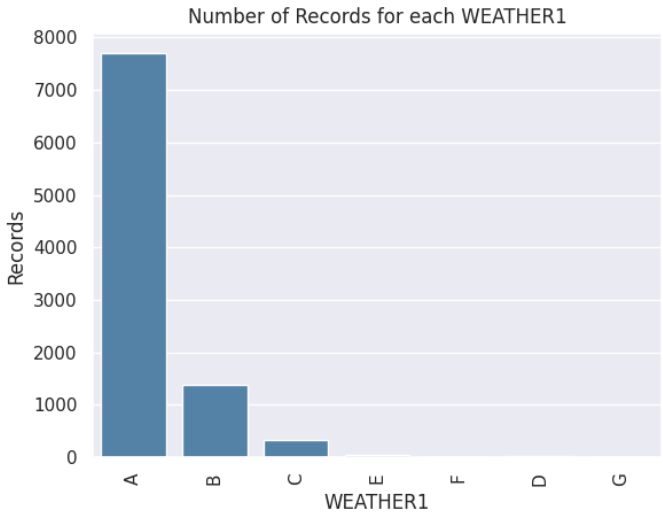
ویژگی ACC_DATE به علت ماهیت وجودی خود که از تاریخ می باشد مهم و قابل استفاده در این تحقیق نبوده و صرفا با پیچیده تر کردن مدل از کارایی آن کم می‌کند. پس تصمیم به حذف این ویژگی نیز می‌گیریم.

نمایش نموداری ویژگی‌های از جنس طبقه بندی(Categorical)

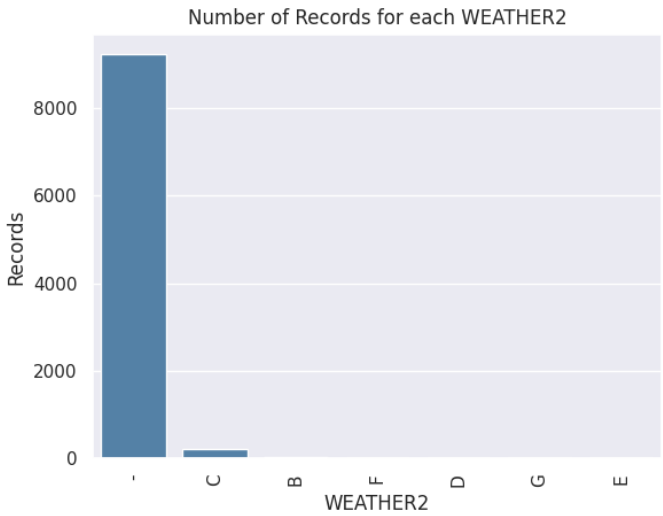
ویژگی ACCTYPE :



ویژگی WEATHER1 :

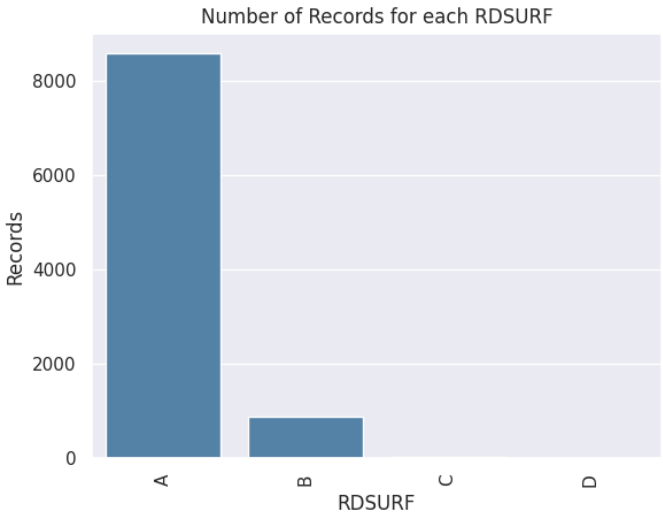


ویژگی WEATHER2 :

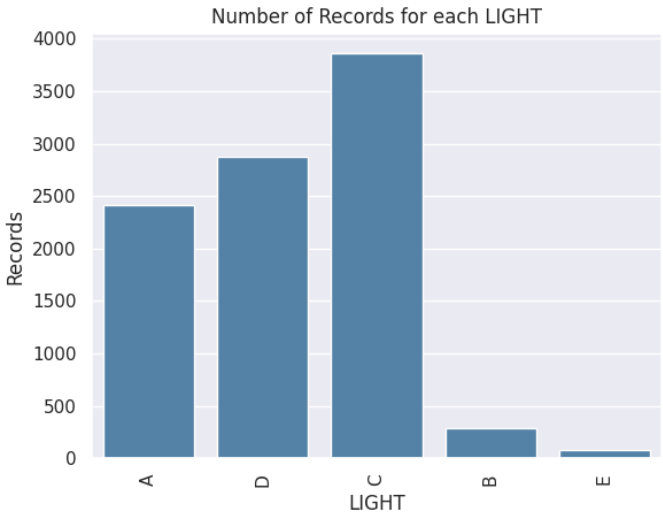


همانطور که مشاهده می کنیم مقادیر زیادی از این ویژگی نامعتبر و یا گزارش نشده می‌باشد. به منظور رفع این مشکل تصمیم به حذف ویژگی می‌گیریم چون جایگزین کردن مقادیر نامعتبر با مقادیری همچون میانگین به علت حجم زیاد آن ها، از لحاظ منطقی مدل را دچار مشکل می‌کند. از طرفی حذف کردن همه‌ی نمونه‌های نامعتبر نیز غیر منطقی می‌باشد.

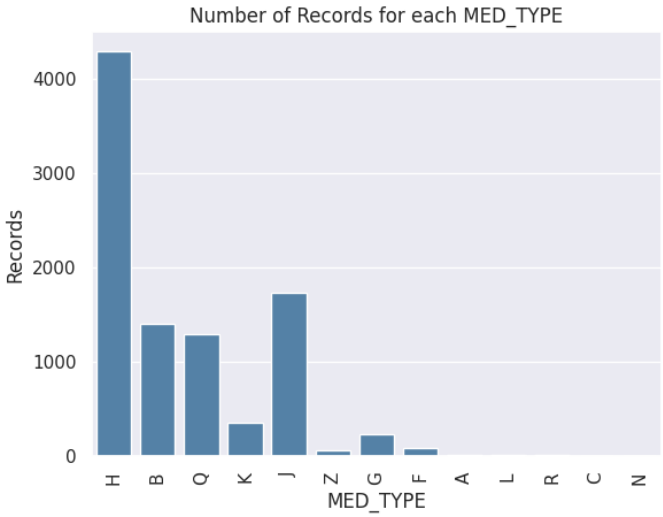
ویژگی RDSURF :



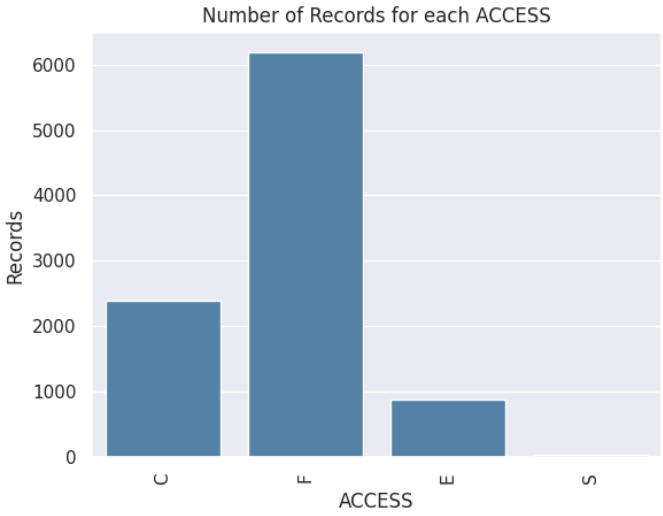
ویژگی LIGHT :



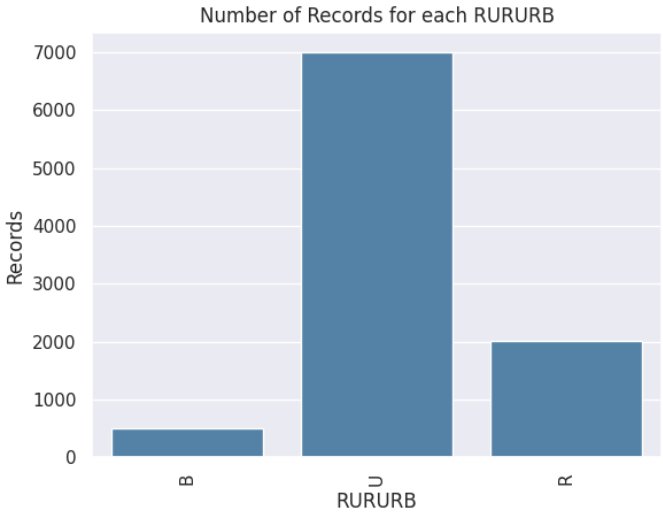
ویژگی MED_TYPE :



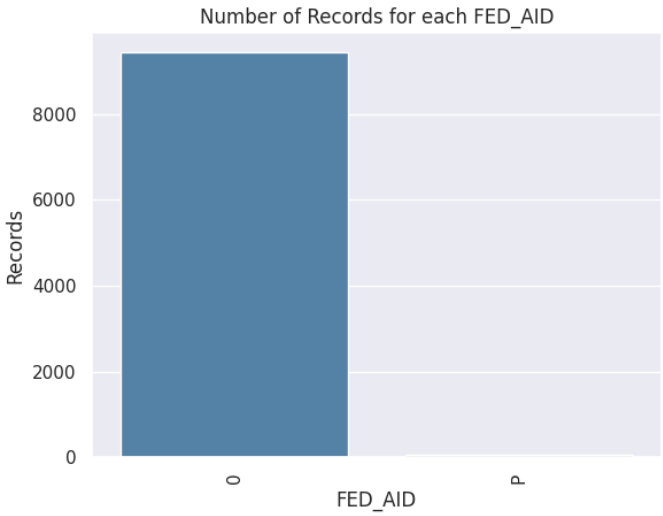
ویژگی ACCESS :



ویژگی RURURB :

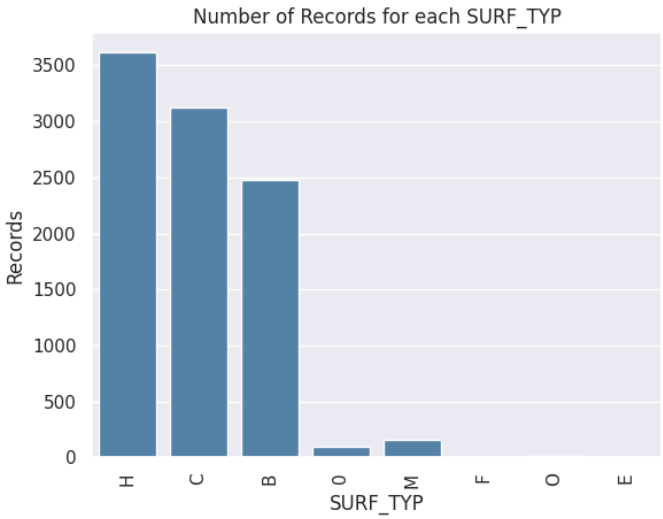


ویژگی FED_AID :

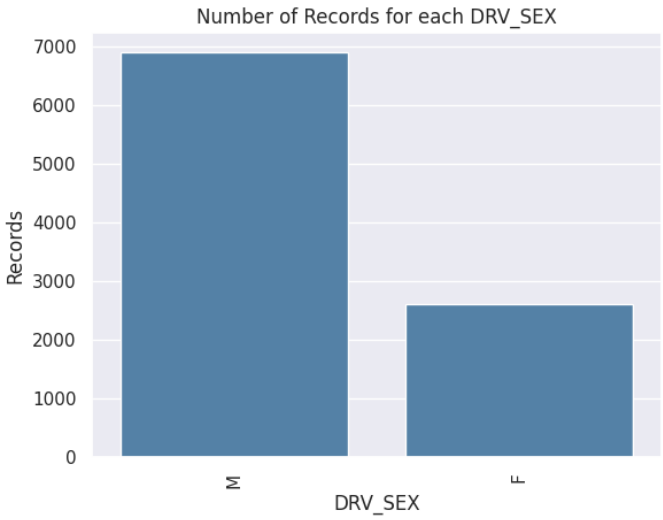


این ویژگی تفاوت زیاد دو طبقه را نشان می دهد که باید به‌صورت جداگانه و عددی تحلیل شده و در مورد آن تصمیم گرفته شود.

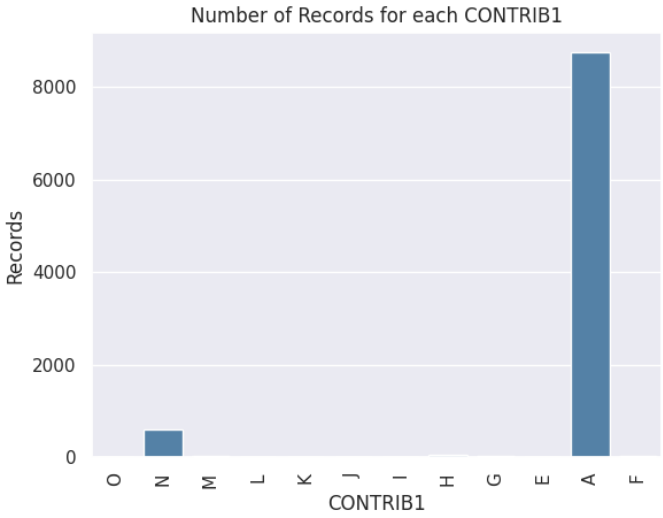
ویژگی SURF_TYP :



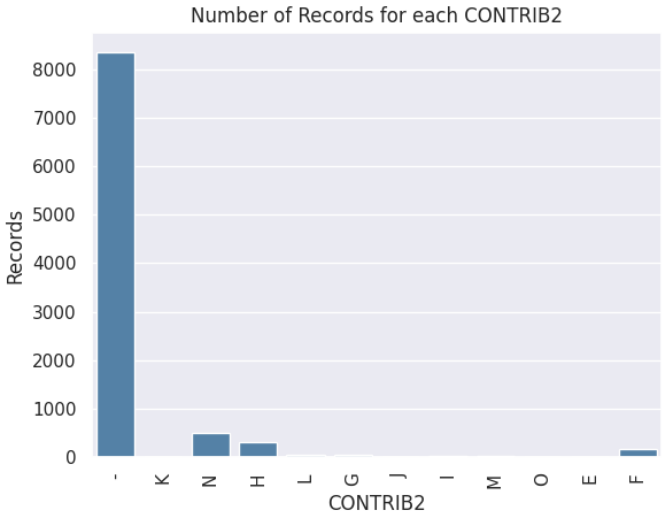
ویژگی DRV_SEX :



ویژگی CONTRIB1 :

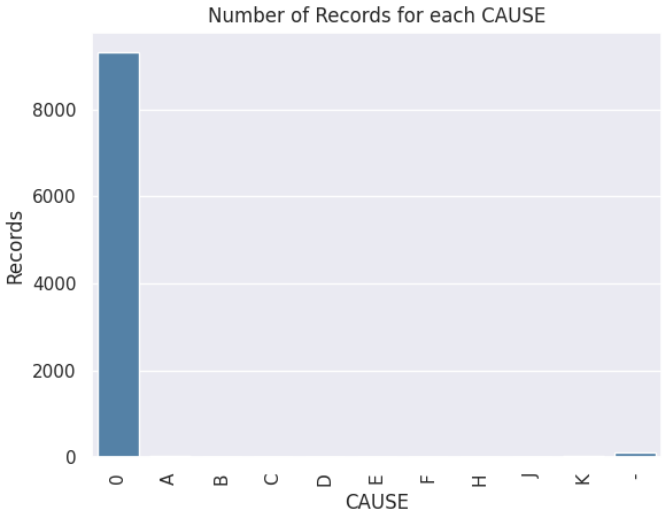


ویژگی CONTRIB2 :



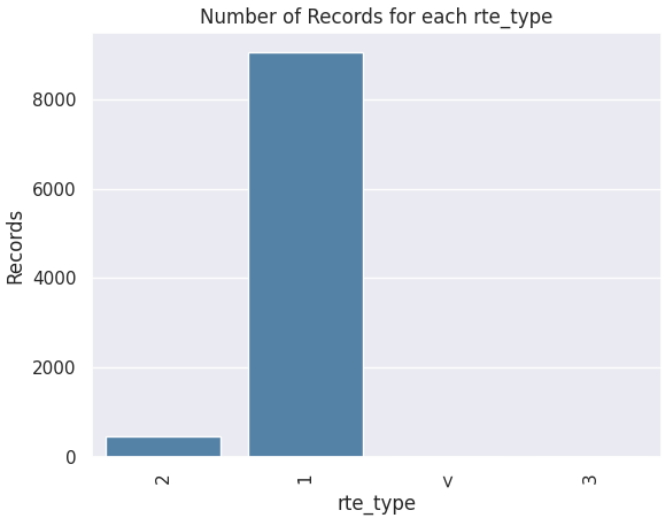
با توجه به نامعتبر بودن مقادیر زیادی از نمونه‌ها در این ویژگی، همانند ویژگی WEATHER2 با این ویژگی رفتار می‌کنیم.

ویژگی CAUSE :



با توجه به نامعتبر بودن تعداد زیادی از مقادیر تصمیم به حذف این ویژگی نیز می‌گیریم.(همانند ویژگی‌های قبلی از این قبیل)

ویژگی rte_type :



پس از بررسی عددی تصمیم به عدم حذف ویژگی FED_AID گرفتیم.

تغییر متغیرهای از جنس طبقه بندی به عددی

با استفاده از LabelEncoder، متغیرهای طبقه بندی را به متغیر های عددی تبدیل می‌کنیم تا مدل دچار مشکل نشود. خروجی لیبل گذاری متغیرهای طبقه بندی به این شکل می‌باشد:

```
ACCTYPE : [4 3 2 1 0 6 5 7]
WEATHER1 : [0 1 2 4 5 3 6]
RDSURF : [0 1 2 3]
LIGHT : [0 3 2 1 4]
MED_TYPE : [ 5  1 10  7  6 12  4  3  0  8 11  2  9]
ACCESS : [0 2 1 3]
RURURB : [0 2 1]
FED_AID : [0 1]
SURF_TYP : [5 2 1 0 6 4 7 3]
DRV_SEX : [1 0]
CONTRIB1 : [11 10  9  8  7  6  5  4  3  1  0  2]
rte_type : [1 0 3 2]
```

پس از اتمام لیبل گذاری، ویژگی های عددی را بررسی می‌کنیم. در این مرحله اطلاعات خاصی به ما افزوده نشده و تغییر خاصی به کل داده‌ها اعمال نمی‌کنیم.

در انتها سایز کلی داده‌ها به 9502 نمونه با در نظرگیری 33 ویژگی تغییر می‌یابد.

تقسیم داده‌ها به دو بخش train و test

داده‌ها را به صورت رندوم و با نسبت 0.33 به دو بخش train و test تقسیم می‌کنیم.(6366 عدد داده برای train و 3136 عدد داده برای test اختصاص می‌یابد).

مدل Random Forest با پارامترهای پیش‌فرض

در اینجا مدل به صورت پیش‌فرض با 10 عدد درخت تصمیم ساخته شد. پس از یادگیری دقت مدل با کمک داده های تست سنجیده شد که خروجی زیر بدست آمد:

```
Model accuracy score with 10 decision-trees : 0.8555
```

مدل Random Forest با پارامتر n_estimatores=100

در این مرحله با افزایش تعداد درخت های تصمیم سعی بر تحلیل و بررسی این تغییر بر دقت مدل داریم. پس از یادگیری و انجام تست خروجی زیر بدست می‌آید:

```
Model accuracy score with 100 decision-trees : 0.8555
```

با مشاهده خروجی بدست‌آمده در می‌یابیم که این تغییر پارامتر تاثیری در دقت مدل ایجاد نکرده است.

یافتن ویژگی‌های مهم با استفاده از RF

تا به اینجا ما از همه‌ی ویژگی‌ها در مدل استفاده کردیم. حال قصد این را داریم که ابتدا ویژگی‌های مهم را شناسایی کنیم، سپس مدل را با این ویژگی‌ها آموزش داده و تاثیر این تغییر را در دقت مدل بسنجیم.

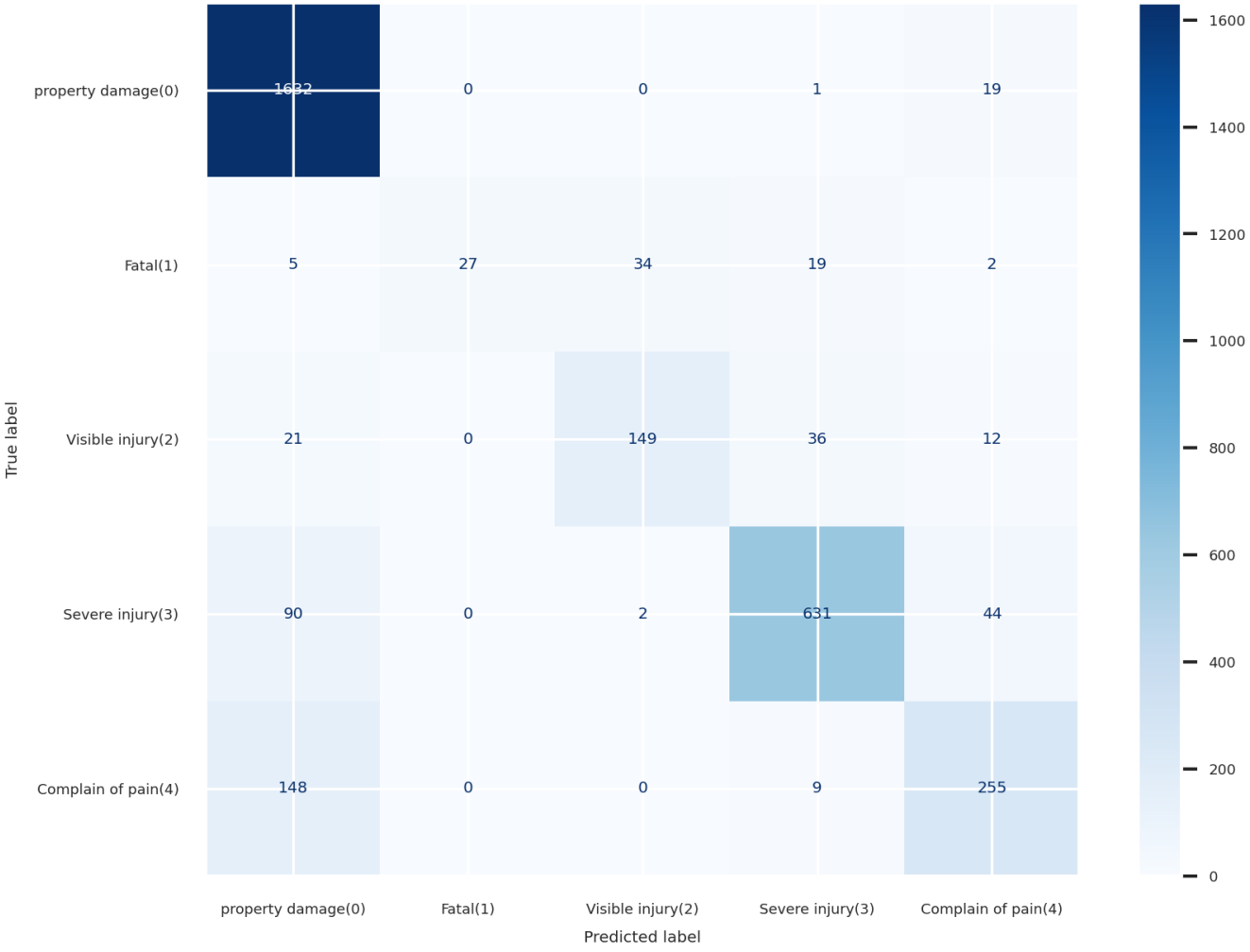
به این منظور ابتدا یک مدل RF با پارامتر n_estimatores=100 ساخته و داده ها را به آن فیت می‌کنیم. سپس از متغیر feature_importance استفاده می‌کنیم تا امتیاز اهمیت هر ویژگی را بهتر درک کنیم. امتیاز همه‌ی ویژگی‌ها در جدول صفحه بعد نشان داده شده است. همانطور که مشاهده می‌کنیم ویژگی DRV_INJ مهم ترین و ویژگی FED_AID کم اهمیت ترین ویژگی‌ها هستند.

بررسی و تحلیل دقت مدل با حذف ویژگی‌های کم اهمیت

با حذف کم اهمیت ترین ویژگی(FED_AID)دقت مدل به 0.8562 افزایش پیدا می‌کند. در مرحله بعدی پس از حذف دو ویژگی دیگر(rte_type, ACCESS) علاوه بر کم اهمیت ترین ویژگی، دقت مدل به 0.8591 افزایش می‌یابد. در مرحله بعدی پس از حذف ویژگی کم اهمیت بعدی(RDSURF) دقت مدل به 0.8581 کاهش می‌یابد. در این مرحله متوقف می‌شویم و دیگر ادامه نمی‌دهیم.

Feature	Importance_Score
DRV_INJ	0.482324808
HOUR	0.055814654
drv_age	0.048027146
milepost	0.046669745
ACCTYPE	0.033330472
AADT	0.025382315
SEG_LNG	0.023725032
ENDMP	0.023378103
numvehs	0.022666155
county	0.021061033
POP_GRP	0.019703772
LIGHT	0.017131306
BEGMP	0.015488553
MEDWID	0.015199052
LANEWID	0.013397369
SURF_WID	0.012912317
LSHLDWID	0.011728032
PAV_WDL	0.011471445
NO_LANES	0.010496074
DRV_SEX	0.010434987
RSHLDWID	0.01019592
WEATHER1	0.009950458
PAV_WIDR	0.009730029
DESG_SPD	0.009262103
MED_TYPE	0.008557874
SURF_TYP	0.007208106
CONTRIB1	0.0059696
RURURB	0.004989304
RDSURF	0.00481315
ACCESS	0.004182871
rte_type	0.004179589
FED_AID	0.000618624

ماتریس درهم ریختگی



با توجه به ماتریس درهم ریختگی مدل عملکرد خوب و قابل قبولی از خود در مواجهه با قسمت تست از خود نشان داده است.

گزارش دسته‌بندی

	precision	recall	f1-score	support
0	0.86	0.99	0.92	1652
1	1.00	0.31	0.47	87
2	0.81	0.68	0.74	218
3	0.91	0.82	0.86	767
4	0.77	0.62	0.69	412
accuracy			0.86	3136
macro avg	0.87	0.68	0.74	3136
weighted avg	0.86	0.86	0.85	3136

نتیجه گیری

در این پزوهش سعی بر این بود که مدلی قابل اعتماد و با دقت بالا به منظور دسته بندی شدت تصادفات که عامل اول آن ها الکل بوده طراحی و توسعه شود. پس از استخراج ویژگی های مهم، کالیبره کردن پارامترهای مدل خروجی 0.86 بدست آمد که خروجی نسبتا قابل قبولی است. در خلال تحلیل و بررسی ویژگی های تاثیرگذار بر شدت تصادفات نتایج جالبی اعم از بی اهمیت بودن برخی ویژگی های دولتی، ساختار جاده ای و ... بدست آمد و متوجه شدیم مهم ترین عامل تاثیرگذار، جراحت وارده به راننده اتوموبیل بوده که خیلی دور از ذهن نمی‌باشد. با توجه به ماتریس درهم ریختگی از نقاط ضعف مدل تشخیص ضعیف آن در مورد شدت هایی از جنس منجربه مرگ می‌باشد که با توجه به اهمیت و حساس بودن، این موضوع خیلی قابل قبول نمی‌باشد.