آزمایشگاه تحلیل پیشرفته کلان داده (ابدال)

طبقه‌بندی – ۲
(ماشین بردار پشتیبان)

استاد: دکتر حسین رحمانی

خرداد ۱۴۰۰

مقدمه

- مهلت ارسال تمرین تا ساعت ۲۳:۵۹ تاریخ ۱۴۰۰/۰۳/۱۰ است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر ۲۵ درصد از نمره تمرین کسر خواهد شد.
- پاسخ به سوالات این تمرین باید در قالب یک گزارش با فرمت PDF ارائه شود.
- به همراه فایل گزارش تمرین، فایل کدهای اجراشده نیز پیوست شود.
- تمامی فایلهای این تمرین (گزارش و کدها) در قالب یک فایل فشرده با نامگذاری زیر ارسال شود.
  StudentNumber_FirstName_LastName_HW5.zip
- فایل تمرین را از طریق سامانه LMS ارسال نمایید.
- رعایت نکات نگارشی در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این تمرین حتما باید از زبان برنامه نویسی پایتون استفاده شود.

۱- معرفی دیتاست

مجموعه داده Insurance Claim مربوط به تراکنش‌های یک شرکت بیمه در کشور آمریکا بوده و از طریق نشانی زیر[1] (و یا فایل پیوست) قابل دسترس می‌باشد. این مجموعه داده شامل ویژگی‌های زیر می‌باشد:

- شامل ۱۰۰۰ تراکنش
- شامل ۳۹ ستون
- بدون مقادیر گمشده[2]
- حجم ۲۶۱ کیلوبایت

این دیتاست شامل ۳۹ ستون بوده و متغیر هدف نیز fraud_reported می‌باشد. به‌طور کلی ستون‌های این مجموعه داده را می‌توان به ۴ دسته تقسیم کرد:

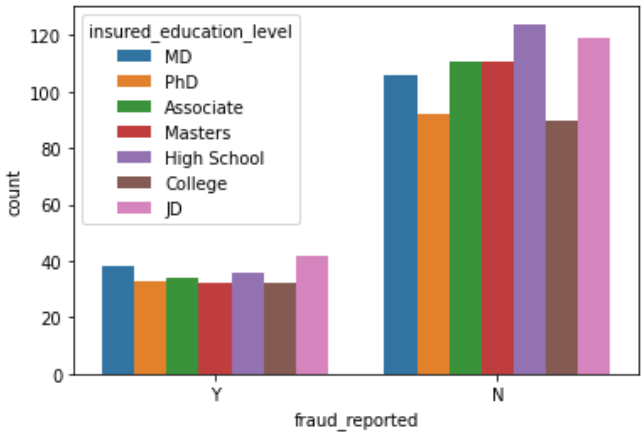
۱-۱- ویژگی‌های بیمه‌نامه

- **policy_Number**: شماره بیمه‌نامه
- **policy_bind_date**: تاریخ عقد بیمه‌نامه
  - ابتدای بازه: ۲۰۰۶/۰۱/۰۱
  - انتهای بازه: ۲۰۱۲/۱۲/۳۱
- **policy_state**: ایالت بیمه‌نامه را نشان می‌دهد و شامل ۳ مقدار می‌باشد.
- **policy_csl** [3]: عددی از پیش تعیین شده‌است که حد پوشش خسارت جانی و مالی در هر حادثه را نشان می‌دهد.
- **policy_deductable**: موارد قابل‌کسر در بیمه‌نامه
- **policy_annual_premium**: حق بیمه سالانه
- **umbrella_limit**: مبالغ اضافی پرداخت‌شده توسط شرکت بیمه

۱-۲- ویژگی‌های فرد بیمه‌شده

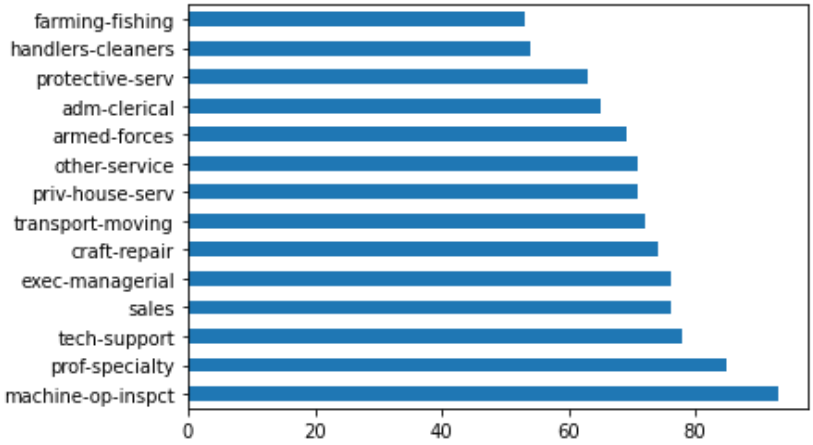
- **month_as_customer**: تعداد ماه‌های عضویت یک مشتری در شرکت بیمه را نشان می‌دهد و مقادیر آن عددی و شامل بازه ۰ تا ۴۷۹ می‌باشد.
- **age**: سن مشتریان
- **insured_zip**: کد پستی فرد بیمه‌شده
- **insured_sex**: جنسیت فرد بیمه‌شده
- **insured_education_level**: سطح تحصیلات فرد بیمه‌شده و مقادیر آن در شکل زیر نمایش داده شده‌است:

[1] https://www.kaggle.com/roshansharma/insurance-claim
[2] Missing Value
[3] Combined single limit (CSL)



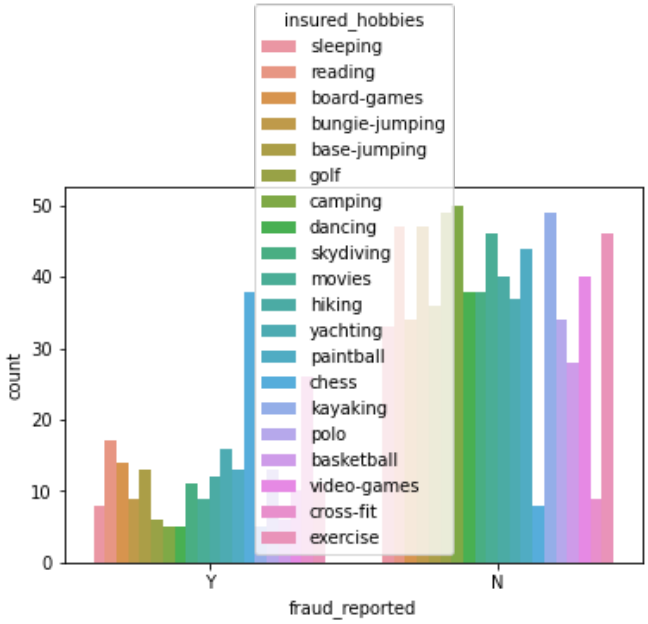
شکل ۱-۱: فراوانی مشتریان در هر کلاس را براساس سطح تحصیلات نشان می‌دهد.

- **insured_occupation**: شغل فرد بیمه‌شده و مقادیر آن در شکل زیر نمایش داده شده‌است:



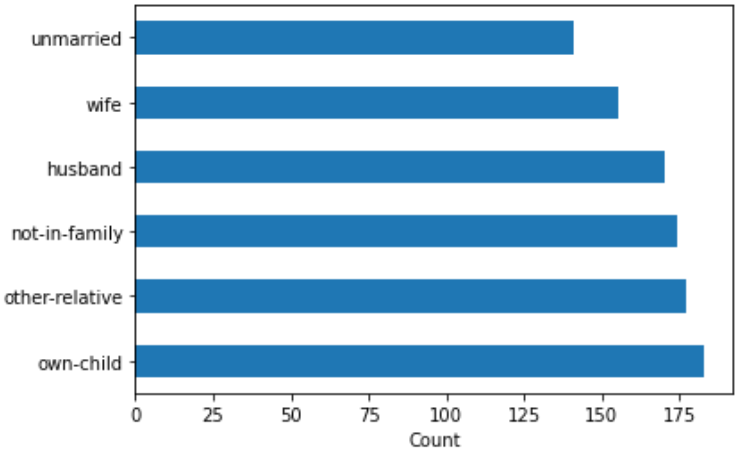
شکل ۲-۱: شغل مشتریان و فراوانی آن‌ها نمایش داده شده‌است.

- **insured_hobbies**: سرگرمی‌های فرد بیمه‌شده و شامل مقادیر ذکر شده در شکل زیر می‌باشد:



شکل ۳-۱: فراوانی مشتریان براساس سرگرمی‌های خود در متغیر هدف نمایش داده شده‌است.

- **insured_relationship**: رابطه با فرد بیمه‌شده را نمایش می‌دهد.



شکل ۴-۱: فراوانی مشتریان برحسب نوع رابطه با فرد بیمه شده، نمایش داده شده‌است.

- **capital-gains**: سود سرمایه و بازه ۰ تا ۱۰۱،۰۰۰ را در بر می‌گیرد.
- **capital-loss**: ضرر سرمایه و بازه ۱۱۱۱۰۰۰- تا ۰ را در بر می‌گیرد.

۳-۱- ویژگی‌های مربوط به حادثه

- **incident_date**: تاریخ حادثه را نشان می‌دهد.
- **incident_type**: نوع حادثه را نشان می‌دهد و شامل مقادیر زیر است:
  - برخورد تنها یک وسیله
  - سرقت وسیله نقلیه
  - برخورد چند وسیله نقلیه
  - وسیله نقلیه پارک‌شده
- **collision_type**: نوع برخورد را نشان می‌دهد و شامل مقادیر زیر است:
  - برخورد جانبی
  - ؟

- برخورد عقب
- برخورد جلو

- **incident_severity**: شدت حادثه را نشان می‌دهد و شامل مقادیر زیر است:
  - خسارت عمده
  - خسارت جزئی
  - خسارت کلی
  - خسارت بی‌اهمیت
- **authorities_contacted**: مراجع مرتبط را نشان می‌دهد و شامل مقادیر زیر است:
  - پلیس
  - آتش‌نشانی
  - آمبولانس
  - سایر
  - هیچ‌کدام (None)
- **incident_state**: ایالت محل وقوع حادثه را نشان می‌دهد.
- **incident_city**: شهر محل وقوع حادثه را نشان می‌دهد.
- **incident_location**: آدرس محل وقع حادثه را نشان می‌دهد و شامل ۱۰۰۰ مقدار یکتا می‌باشد. نمونه‌ای از مقادیر این ویژگی در شکل زیر نمایش داده شده‌است:

```
data['incident_location'].unique()
```

```
array(['9935 4th Drive', '6608 MLK Hwy', '7121 Francis Lane',
       '6956 Maple Drive', '3041 3rd Ave', '8973 Washington St',
       '5846 Weaver Drive', '3525 3rd Hwy', '4872 Rock Ridge',
       '3066 Francis Ave', '1558 1st Ridge', '5971 5th Hwy',
       '6655 5th Drive', '6582 Elm Lane', '6851 3rd Drive',
       '9573 Weaver Ave', '5074 3rd St', '4546 Tree St',
       '3842 Solo Ridge', '8101 3rd Ridge', '5380 Pine St',
       '8957 Weaver Drive', '2526 Embaracadero Ave', '5667 4th Drive',
       '2502 Apache Hwy', '3418 Texas Lane', '2533 Elm St',
       '3790 Andromedia Hwy', '3220 Rock Drive', '2100 Francis Drive',
       '4687 5th Drive', '9038 2nd Lane', '6092 5th Ave',
       '8353 Britain Ridge', '3540 Maple St', '3104 Sky Drive',
       '4981 Weaver St', '6676 Tree Lane', '3930 Embaracadero St',
       '3422 Flute St', '4862 Lincoln Hwy', '5719 2nd Lane',
```

شکل ۵-۱: چند نمونه از مقادیر ستون incident_location نمایش داده شده‌است.

- **incident_hour_of_the_day**: ساعتی از روز، که حادثه رخ داده‌است.
- **number_of_vehicles_involved**: تعداد وسایل درگیر در تصادف را نشان می‌دهد.
- **property_damage**: وجود یا عدم وجود خسارت مالی را نشان می‌دهد و شامل مقادیر Yes، No و ؟ می‌باشد.
- **bodily_injuries**: آسیب‌های بدنی را نشان می‌دهد و شامل مقادیر ۰، ۱ و ۲ می‌باشد.
- **witnesses**: تعداد شاهدان صحنه تصادف را نشان می‌دهد.
- **police_report_available**: در دسترس بودن گزارش پلیس را نشان می‌دهد (با مقادیر Yes، No و ؟).
- **total_claim_amount**: مبلغ کل خسارت
- **injury_claim**: خسارت ناشی از آسیب دیدگی
- **property_claim**: خسارت مالی

۴-۱- ویژگی‌های مربوط به وسیله نقلیه

- **vehicle_claim**: خسارت وسیله نقلیه

- **auto_make**: کارخانه سازنده وسیله نقلیه
- **auto_model**: مدل وسیله نقلیه
- **auto_year**: سال وسیله نقلیه

۲- وارد کردن داده

- داده‌ها را از فایل مورد نظر بخوانید و در dataframe ذخیره کنید
- پیش‌پردازش‌های لازم را روی داده‌ها انجام دهید
- ستون fraud_reported را به عنوان ستونی انتخاب کنید که می‌خواهید پیش‌بینی کنید
- سایر ستون‌ها را به عنوان ویژگی‌هایی انتخاب کنید که برای انجام پیش‌بینی استفاده می‌شود (انتخاب ویژگی‌های مناسب و خلاقیت در پیش‌پردازش آن‌ها به عهده شما است)

۳- ساخت مدل ماشین بردار پشتیبان (SVM)

در ادامه به ساخت دو مدل بر اساس ماشین بردار پشتیبان می‌پردازیم. در هر مرحله اگر چندین حالت وجود دارد فرضیات خود را گزارش کنید و میزان بررسی و گزارش دقیق‌تر شما در نمره‌دهی شما لحاظ می‌شود.

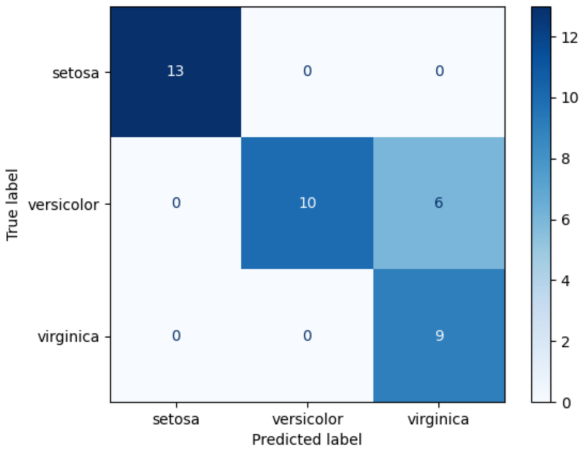
۳-۱-۱- مقیاس‌بندی داده

- داده‌های تست و آموزش را جدا کنید (random_state=42)
- از تابع "scale()" استفاده کنید و داده‌ها آموزش و تست را مقیاس‌بندی کنید

۳-۱-۲- ساخت نمونه اولیه

- یک نمونه اولیه از SVM بسازید
- تفاوت چهار kernel (linear, poly, rbf, sigmoid) و پارامترهای هرکدام را مشخص کرده و به ازای مقادیر مختلف پارامترها تفاوت نتایج را مقایسه کنید

- با استفاده از ماتریس درهم ریختگی (Confusion Matrix) و داده‌های تست عملکرد این ماشین بردار پشتیبان را به ازای ۴ کرنل ارزیابی کنید(رسم ماتریس درهم ریختگی مانند شکل ۶ باشد)



شکل ۶: نمونه از ماتریس درهم ریختگی

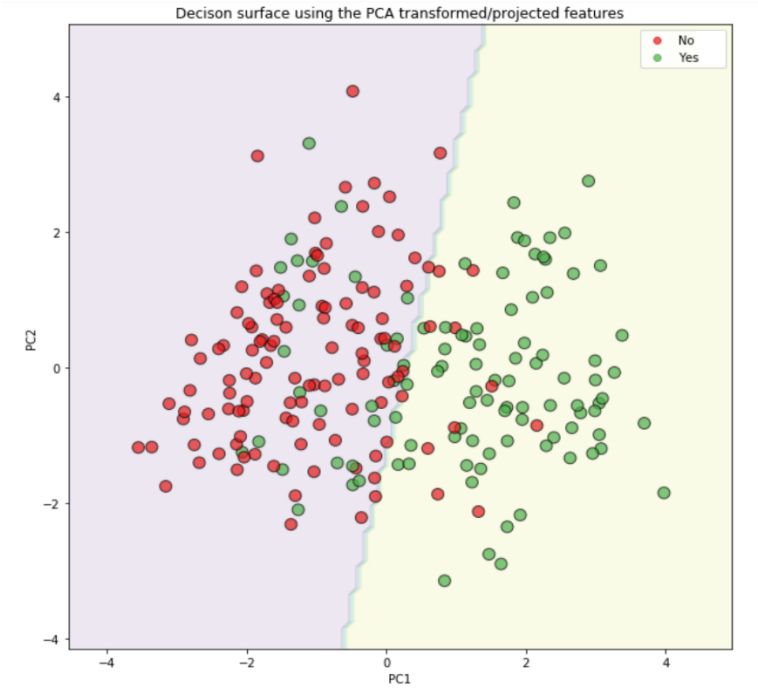
۳-۱-۳- بهینه‌سازی پارامترها

- با استفاده از تابع «GridSearchCV()» بهینه‌ترین مقدار را برای پارامترها بیابید

۳-۱-۴- ساخت، ارزیابی، رسم و تفسیر آخرین نمونه ماشین بردار پشتیبان

- نمونه جدیدی با استفاده از پارامترهای بهینه شده بسازید
- نتایج ارزیابی این مدل را با حالتی که پارامترها بهینه نبودند مقایسه کنید و نمودار ROC را برای هر دو رسم کنید

- **امتیازی**: با استفاده از PCA ویژگی‌ها را به ۲ بعدی کاهش دهید و از آن برای رسم شکلی از طبقه‌بند مانند شکل ۷ استفاده کنید



شکل ۷: نمایش داده و نتیجه طبقه‌بندی