به نام خدا

گزارش تمرین های سری دوم درس داده کاوی

استادان گرامی:

جناب آقای دکتر فرهانی و جناب آقای دکتر خردپیشه

دستیار آموزشی : جناب آقای شریفی

گردآورنده: سلاله شیخیان

الف: آیا داده پرت در دیتاست وجود دارد ؟ در صورت وجود آنها را حذف کنید

داده‌های پرت باعث می‌شود ارتباط بین دو متغیر ضعیف شود یا از بین برود، اگرچه ممکن است در واقعیت یا بر اساس مبانی نظری ارتباط بین دو متغیر وجود داشته باشد اما نتایج به علت ورود داده‌های پرت ممکن است مخدوش شود و ارتباط بین متغیرها معنادار نشود. پس داده های پرت در تشخیص روابط داده ها اختلال ایجاد می کند.

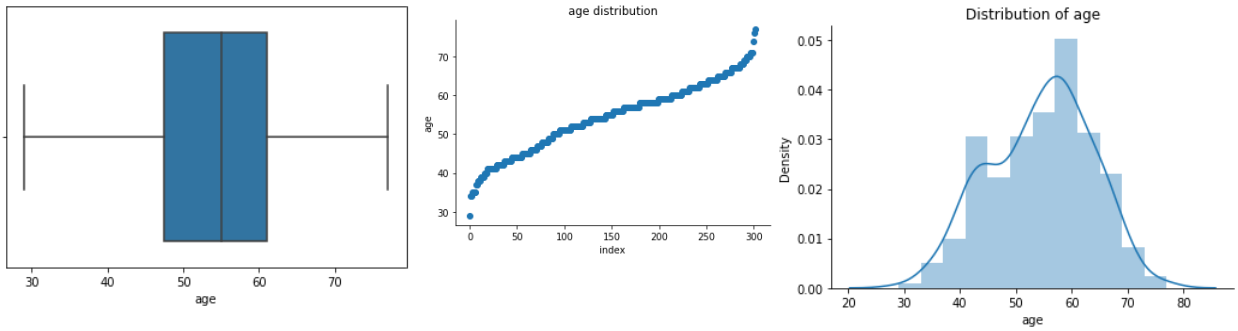
برای این که ببینم آیا داده پرت وجود دارد یا نه چندین راه وجود دارد:

هر ستون را به صورت تکی بررسی کنم و اگر نمودار جعبه ای آن، داده های پرت را نشان داد، دامنه آن ستون را محدود نمایم تا داده های پرت حذف شوند.
راه حل دوم این است که برای تشخیص داده های پرت،دو یا چند ستون را با هم در نظر بگیریم.
راه حل بهتر این است که از الگوریتم جنگل ایزوله استفاده نماییم.

الگوریتم جنگل ایزوله، یکی از مدل‌های مبتنی بر درختاست که براساس تقسیم و تفکیک مشاهدات به نقاط با فراوانی کم و متفاوت از بقیه عمل می‌کند.
به این ترتیب نقطه‌ای تصادفی در بین کوچکترین و بزرگترین مقدار انتخاب شده و همسایگی حول آن سنجیده می‌شود. اگر تعداد همسایه یک نقطه نسبت به بقیه نقاط، کم باشد، آن نقطه به عنوان مقدار مشکوک و ناهنجار شناسایی می‌شود.

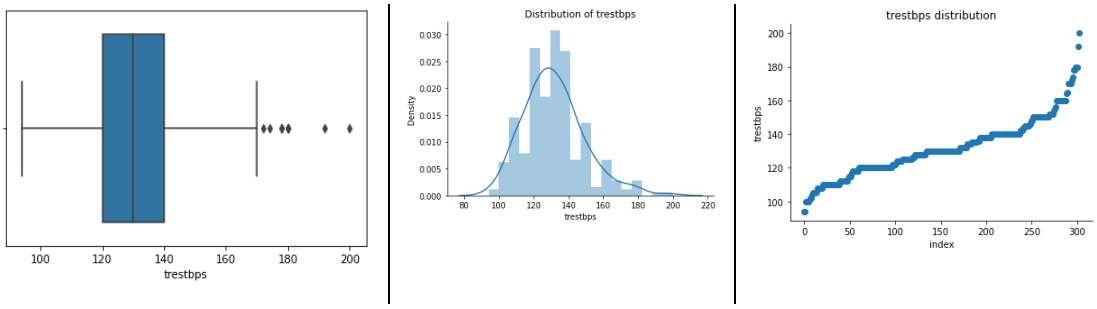
استفاده از یک متغیر برای شناسایی مشاهدات ناهنجار برای داده‌های چند متغیره مناسب به نظر نمی‌رسد. ممکن است با استفاده از یک متغیر، بعضی از مشاهدات ناهنجار بوده در حالیکه هنگام استفاده از متغیر دیگر، چنین مشاهداتی کاملا معقول به نظر برسند. بنابراین بهتر است در مواجه با مجموعه داده‌های چند بعدی از تکنیک‌های چند متغیره استفاده شود.
اما در اینجا ما از روش تک متغیره استفاده می نماییم. و بر اساس نمودار جعبه ای داده های پرت را حذف می کنیم.

نمودار سن:



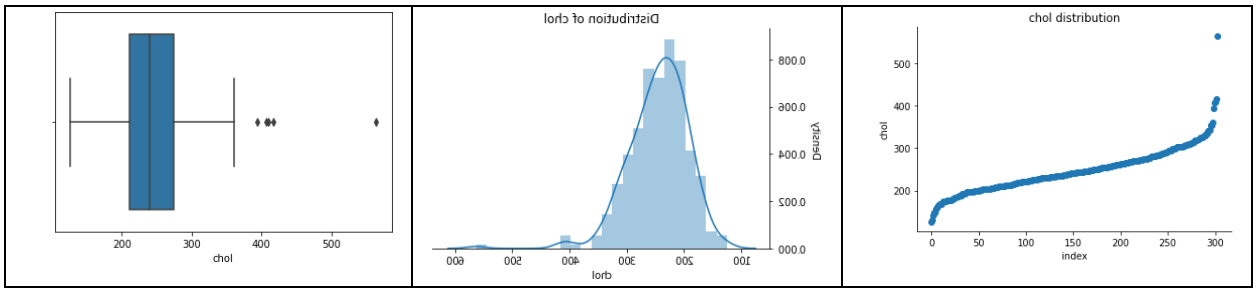
این سه نمودار نشان می دهد که این ویژگی در داده ها داده پرتی را نشان نمیدهد.

نمودار trestbps

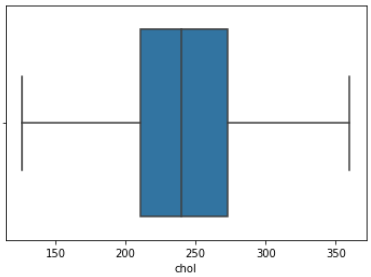


برای این ویژگی هر سه نمودار داده های پر را نمایش میدهد و این داده هایی که دارای این ویژگی بیشتر از ۱۷۰ هستند را حذف میکنیم.

نمودار chol

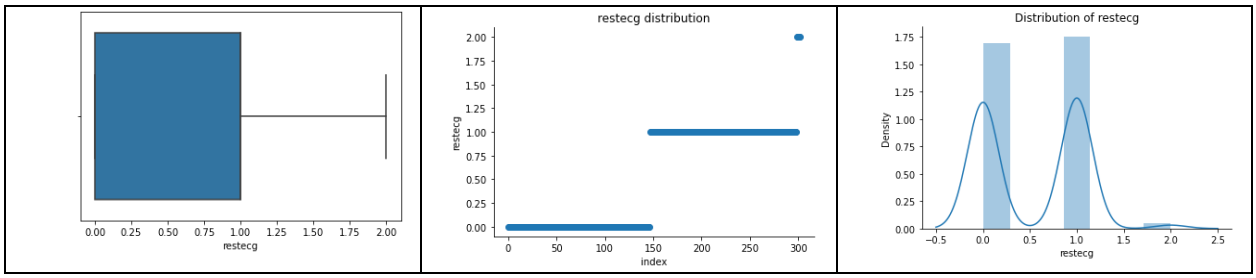


این ویژگی هم دارای داده پرت هست که پس از حذف آن ها نمودار به صورت زیر خواهد بود:



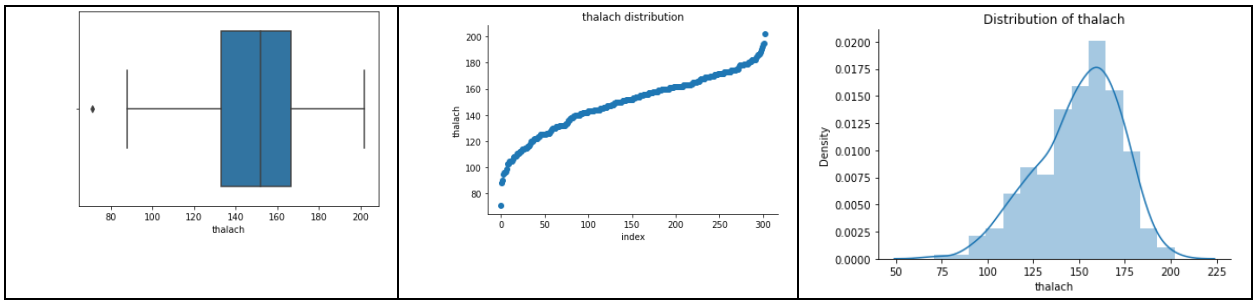
نمودار fbs

این ویژگی با این که دارای دو دسته بزرگ و دسته کوچکی است اما طبق نمودار ها این دسته کوچک داده پرت به حساب نمی آید:

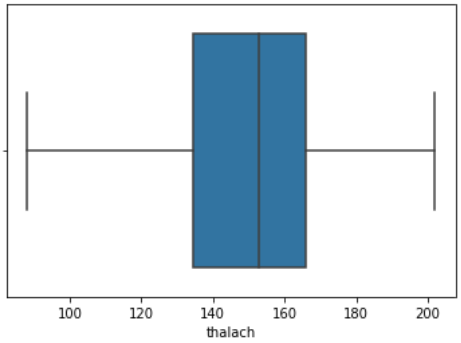


نمودار thalach

این ویژگی هم دارای اطلاعات پرت هست که آن ها را حذف مینماییم:



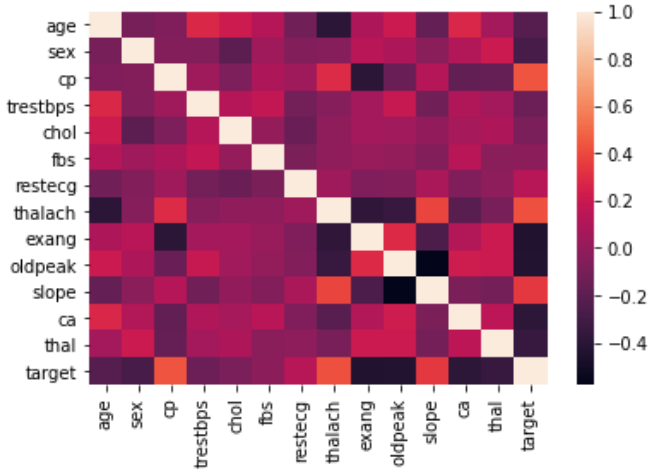
داده های اصلاح شده این صورت هستند:



برای سایر داده ها هم همانطور که در کد مشاهده مینمایید داده های پرت را حذف نمودیم و برای خلاصه نمودن فایل گزارش از درج آن ها در این فایل خودداری نمودیم.

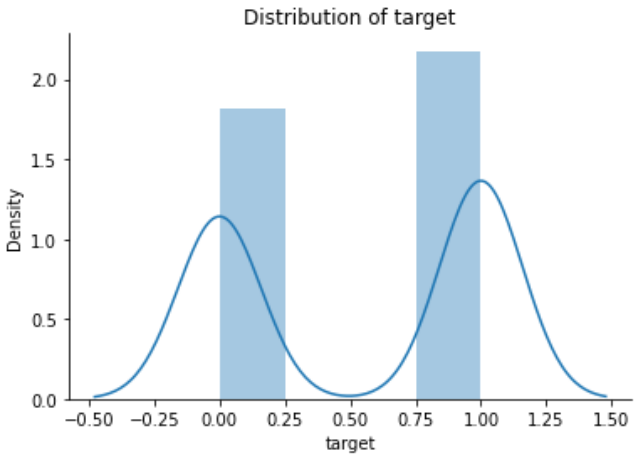
چند متغیره:

برای این که نقاط پرت را بهتر بدست آوریم، میتوایم به جای درنظر گرفتن یک متغیر *دو* متغیر که با هم همبستگی دارند را درنظر بگیریم. سپس داده هایی که بر اساس همبستگی این تو متغیر پرت هستند را به عنوان داده پرت در نظر بگیریم. پس ابتدا لازم است که ضرایب همبستگی متغیر ها را بدست آوریم



ب- بررسی کنید آیا تعداد نمونه ها در هر کلاس متوازن است ؟ (به صورت مختصر توضیح دهید اگر داده ها متوازن نباشد چه مشکلاتی ممکن است پیش بیاید وچه راه حل هایی برای آن وجود دارد

داده های دو کلاس متوازن هستند. اگر داده های یک دسته کمتر از ۵درصد باشد یعنی داده های کلاس ها متوازن نیستند و این مسئله باعث می شود که الگوریتم های یادگیری جهتدار شوند. برای مثال در مسئله یافتن تقلب و یا دزدی برق که تعداد داده های متقلب بسیار کمتر از تعداد بدون تقلب هست. بسیاری از الگوریتم ها این داده های کم را کم اهمیت در نظر گرفته و جهتدار تصمیم گیری میکنند.



در چنین وضعیتی، مدل پیشگویانه ای که با به کارگیری الگوریتم های یادگیری ماشین ایجاد شده است، جهت دار و یک‌طرفه شده و دقت آن بسیار پایین خواهد بود.

این اتفاق بدین خاطر می‌افتد که الگوریتم‌های یادگیری ماشین معمولا طوری طراحی شده‌اند که با کاهش خطا، دقت مدل را افزایش دهند. بنابراین، این الگوریتم ها توزیع/نسبت یک کلاس نسبت به کل کلاس ها، یا توازن کلاس ها را در محاسبات خود به حساب نمی‌آورند.

رویکردهای متنوعی برای حل مشکل داده‌های نامتوازن وجود دارد که تکنیک‌های نمونه‌برداری مختلفی را به‌کار می‌گیرند.

حل مشکل کلاس های نامتوازن در الگوریتم های پیش بینی

رویکردهای مختلفی برای مواجهه با داده های نامتوازن وجود دارند که در زیر فهرستی از آن‌ها آورده شده است:

الف) رویکرد در سطح داده: تکنیک‌هایResampling

Random Under Sampling

Random Over Sampling

Cluster-Based Over Sampling

Informed Over Sampling: Synthetic Minority Over Sampling Technique

Modified synthetic minority oversampling technique (MSMOTE)

ب) تکنیک های الگوریتمی تجمعی(Algorithmic Ensemble Techniques)

Bagging Based

Boosting-Based

Adaptive Boosting- Ada Boost

Gradient Tree Boosting

XG Boost

۲- نمونه های موجود در دیتاست را با نسبت ۸۰ به ۲۰ به دو بخش داده های آموزشی و داده های تست تقسیم بندی کنید . برای این کار میتوانید از پکیج sklearn ۱ بهره ببرید .
این تقسیم بندی در کد انجام شده است.

۳- قضیه بیز را در حداقل یک پاراگراف بیان کنید . سپس دسته بند های Gaussian Naive Bayes، Multinomial Naive Bayes، Bernoulli Naive Bayes را با یکدیگر مقایسه کنید و بیان کنید هر کدام از این دسته بندها بیشتر در کجا کاربرد دارند.

قضیه بیز از روشی برای دسته‌بندی پدیده‌ها بر پایه احتمال استفاده می‌کند و احتمال رخ احتمال رخداد پیشامد A به شرط B برابر است با احتمال رخداد پیشامد B به شرط A ضرب در احتمال رخداد پیشامد A تقسیم بر احتمال رخداد پیشامدB

$$P(C|B)=\frac{P(B|A)P(A)}{P(B)}$$

دسته بندی به این صورت انجام می شود که : احتمال این که یک نمونه در هر کدام از دسته ها باشد را بدست ی آوریم و نمونه را در دسته ای قرار میدهیم که مقدار احتمال آن بیشتر است. در قضیه بیز فرمول بالا به صورت زیر تعریف میشود:

$$P(C|X)=\frac{P(x|c)P(c)}{P(x)}$$

که در آن c دسته مورد نظر و x مقادیر نمونه است.

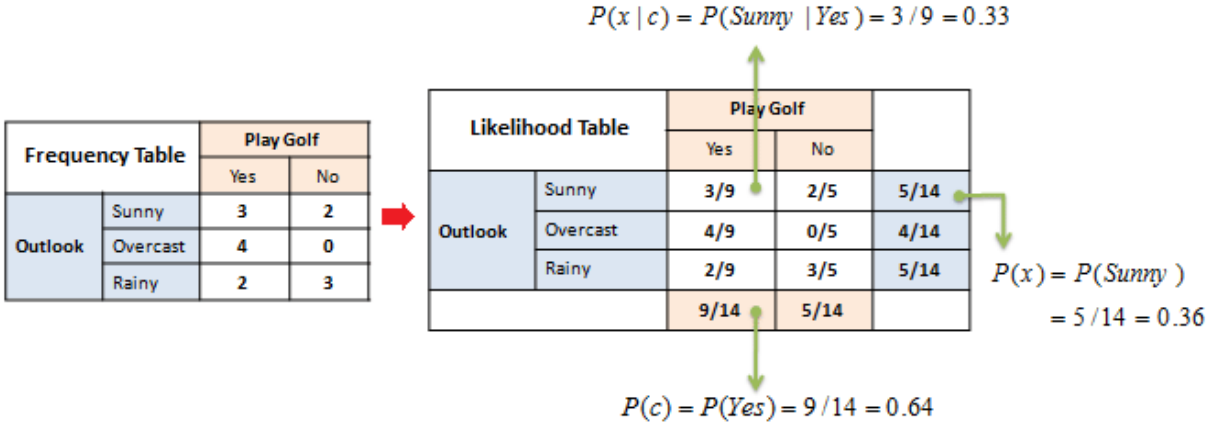
- $P(c \mid x)$ احتمال پیش بینی کننده (ویژگی) است.
- $P(c)$احتمال قبلی کلاس است.
- $P(x \mid c)$این احتمال است که احتمال کلاس پیش بینی کننده داده شده است.
- $P(x)$احتمال قبلی پیش بینی کننده است.

در مثال زیر بهتر میتوانیم هرکدام از این احتمالات را با مثال بدست آوریم:

$p(x|c)$ احتمال این که نمونه

$p(c)$ از تقسیم تعداد برچسب های آن دسته به نسبت کل داده ها بدست می آید

$p(x)$ احتمال قبلی که احتمال وقوع مقدار مورد نظر برای نمونه را نشان می دهد



قضیه بیز بسته به داده هایی که داریم میتواند به صورت های مختلفی به کار برده شود. مثلا برای داده های پیوسته و دارای توزیع نرمال از الگوریتم دسته بند بیز گاوسی استفاده می شود. در ادامه سه دسته بند را معرفی و با هم مقایسه مینماییم.

- gaussian naive bayse

برای داده های پیوسته و با توزیع نرمال مناسب است و احتمال آن با فرمول زیر محاسبه می شود:

$$p(x = v|c_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

- **Multinomial naive Bayes**

برای زمانی که feature vectors ها دارای خاصیت احتمال چندجمله ای هستند مناسب است.

- **Bernoulli naive Bayes**

در این دسته بند ویژگی ها مستقل هستند.

۴-با در نظر گرفتن فیچر ها chol ، trestbps، thalachو لیبل target یک دسته بند ۲Bayes Naive Gaussian را از پایه پیاده سازی کنید برای این کار شما نیاز است که در دیتاست آموزشی خود اعضای مختلف قاعده بیز را محاسبه کنید .

برای این کار ابتدا در کلاس gaussClf داده ها را به داده های تست و آموزش تقسیم می کنیم.

سپس در تابع دوم که برای تخمین p(c) است به این ترتیب عمل میکنیم: خط اول برای هر کلاس تعداد را شمرده ذخیره میکند. خط دوم دو دسته صفر و یک هستند که برای هرکدام عدد بدست آمده را تقسیم بر کل داده ها میکند تااحتمال آن کلاس را بدست بیاورد.

حال برای محاسبه تابع چگالی احتمال و واریانس ر تابع calculate_mean_variance به این صورت عمل میکنیم:

برای بدست آوردن تابع چگالی احتمال باید در هرکلاس برای هرکدام از ویژگی ها تابع چگالی و احتمال را بدست بیاوریم پس از آنجا که ۳ ویژگی و ۲ کلاس داریم باید ۶ جفت واریانس و انحراف معیار بدست آوریم. برای این کار از کتابخانه پانداز استفاده کرده و برای هر ستون میانگین و انحراف معیار را محاسبه میکنیم.

محاسبه تابع چگالی گوسی: با توجه به فرمول تابع چگالی احتمال گوسی که فقط به واریانس و میانگین نیاز دارد، برای هرکدام از متغیر ها تابع چگالی را بر اساس اطلاعات بدست آمده از تابعی که بالا تعریف کرده ایم محاسبه میکنیم

پیشبینی: حالا که همه چیز را سر جای خود داریم ، وقت آن است که کلاس های خود را پیش بینی کنیم. برای این کار تابع predict را تعریف میکنیم. آنچه در این تابع انجام می شود ، تکرار آن از طریق مجموعه آزمون است و برای هر نمونه احتمال هر کلاس را با استفاده از قضیه بیز محاسبه می کند. تنها تفاوت در اینجا این است که ما از احتمالات log استفاده می کنیم.

بررسی دقت الگوریتم: هنگامی که پیش بینی ها را بدست آوردیم ، می توانیم آنها را با مقدار کلاس موجود در مجموعه داده آزمایش مقایسه کنیم ، بنابراین می توانیم نسبت درست ها را به تعداد کل پیش بینی ها محاسبه کنیم.

۶ . با استفاده از پکیج sklearn و GaussianNB یک مدل بسازید و بر روی داده های آموزشی ، ترین کنید سپس بر روی داده های تست همانند سوال قبل سه معیار را گزارش دهید .

این بخش از کدر را در فایل جداگانه و در محیط ژوپیتر پیاده سازی نمودم. این بخش در فایل GaussianNB قرار دارد.