دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

عنوان: تکلیف دوم درس مبانی داده کاوی

نام و نام خانوادگی: افروز شیخ الاسلامی

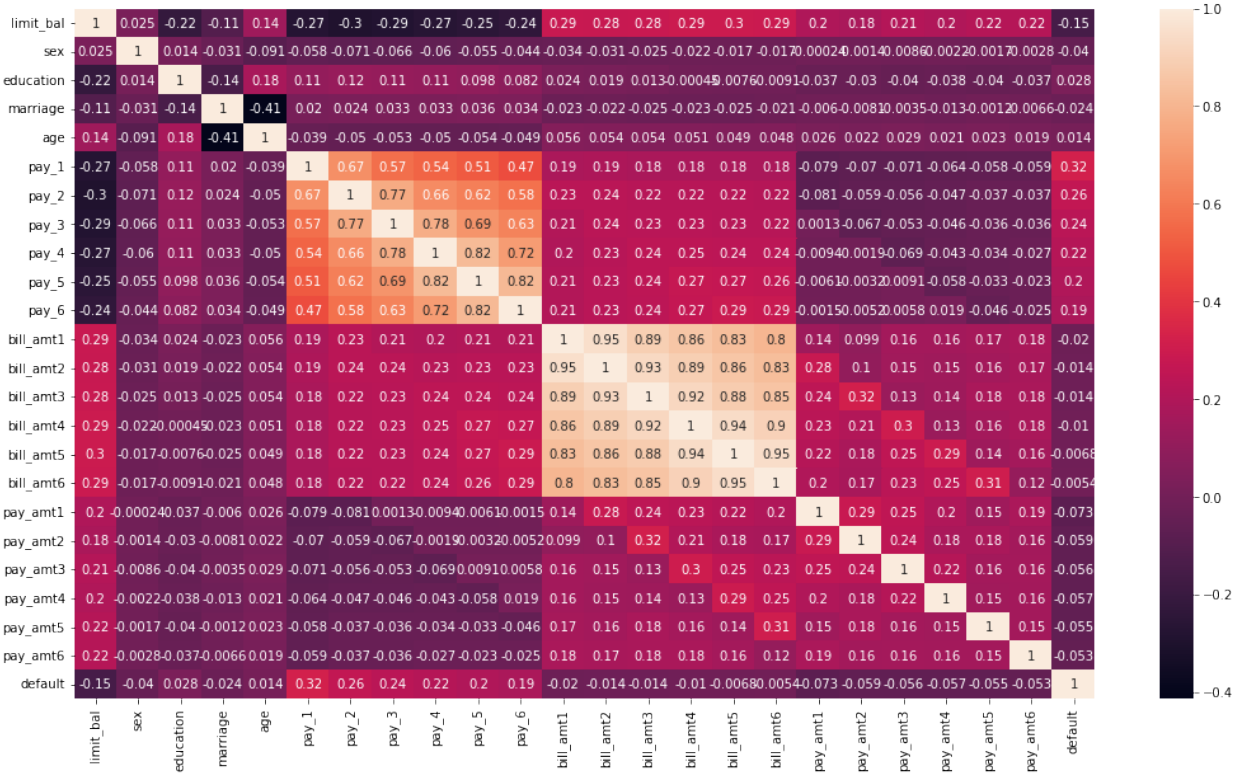
شماره دانشجویی: ۹۷۲۹۳۹۳

نیم‌سال تحصیلی: زمستان ۱۴۰۰

مدرّس: دکتر ناصر قدیری

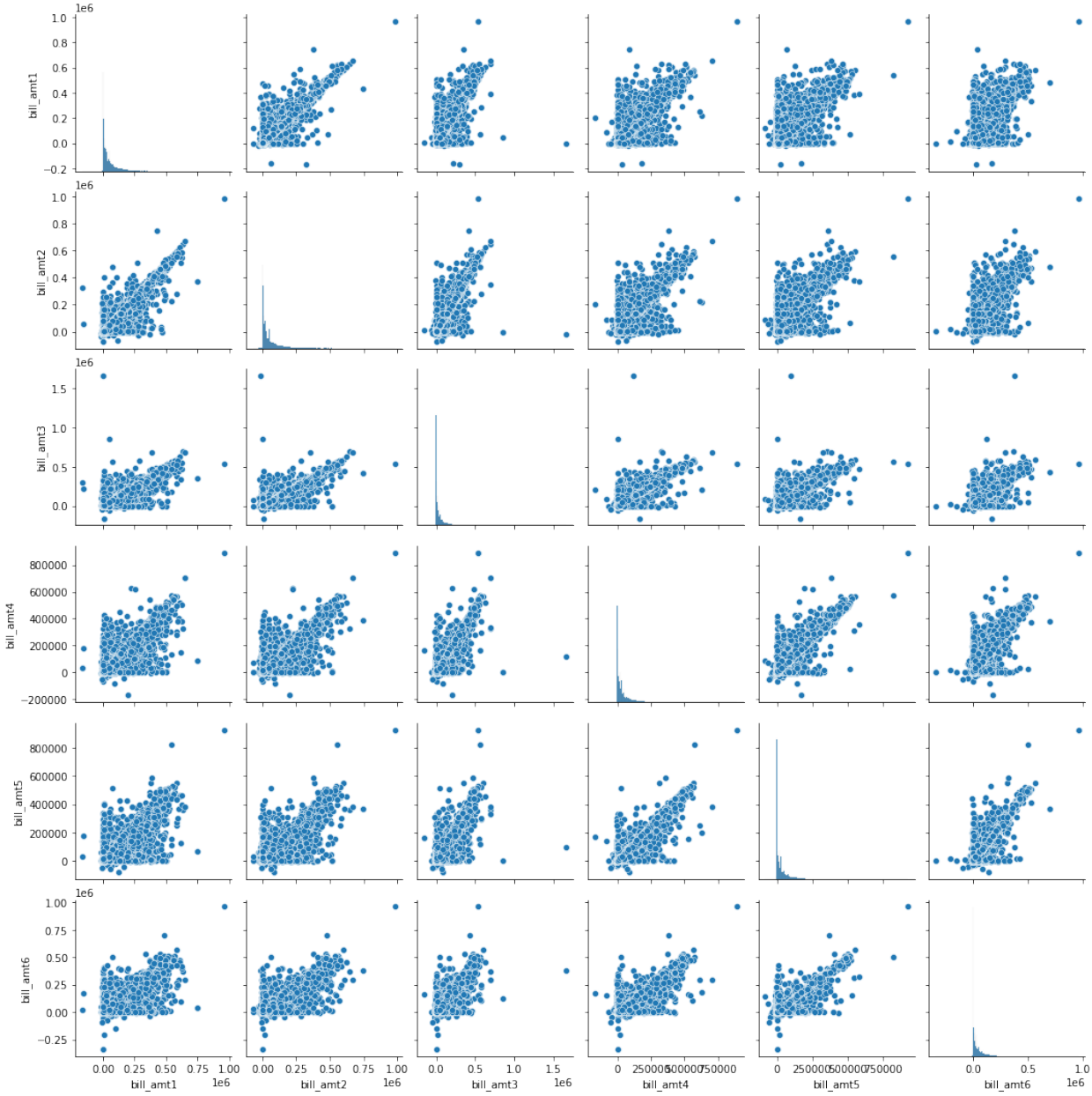
۱ سوال اول

۱.۱ بخش اول سوال اول



ویژگی های pay-1 تا pay-6 نشان دهنده پرداخت های پیشین مشتریان از ماه آپریل تا سپتامبر می باشد.همانطور که در confusion matrix دیده می شود، این ۶ ویژگی بسیار به یکدیگر نزدیک هستند. همچنین ویژگی های bill-amt1 تا bill-amt6 که نشان دهنده مقدار صورت حساب پرداختی مشتریان در ماه های مختلف است نیز، با یکدیگر correlation بالایی دارند. طبیعی است که این اتفاق بیوفتند. چرا که افراد در ماه های مختلف مقدار نسبتا مشخصی برای قبض خود می پردازند. همچنین در این نمودار، ویژگی های bill-amt و pay وابستگی کمی دارند. و بین دیگر ویژگی ها با یکدیگر، رابطه خاصی دیده نمی شود.

۲.۱ بخش دوم سوال اول



به طور تقریبی اکثر این ۶ ویژگی با یکدیگر وابستگی مثبت دارند ولی در بعضی از آنها که نمودار کمتر به شکل یک خط راست است، نمودار نکته خاصی را نشان نمی دهد.

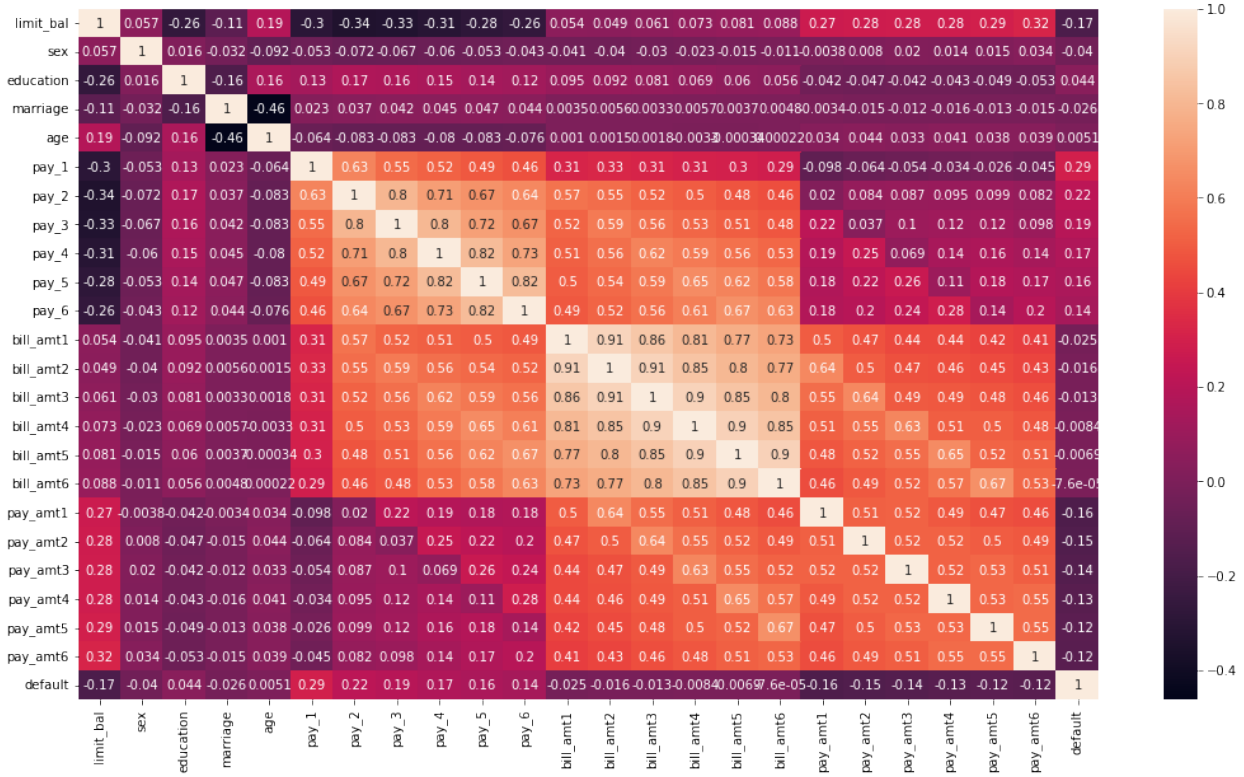
۳.۱ بخش سوم سوال اول

از بین تمامی ویژگی ها، ویژگی های pay-1 تا pay-6 بیشترین همبستگی را با ویژگی هدف دارند. نمودار ها در فایل HW2-Q1 موجود است.

۴.۱ بخش چهارم سوال اول

باز هم مانند قسمت قبل، از بین تمامی ویژگی ها، ویژگی های pay-1 تا pay-6 بیشترین همبستگی را با ویژگی هدف دارند. و نتیجه در حالت کلی با قسمت قبل تفاوتی نداشت، اما می توان گفت که با یکدیگر تفاوت جزئی داشتند. نمودار ها در فایل HW2-Q1 موجود است.

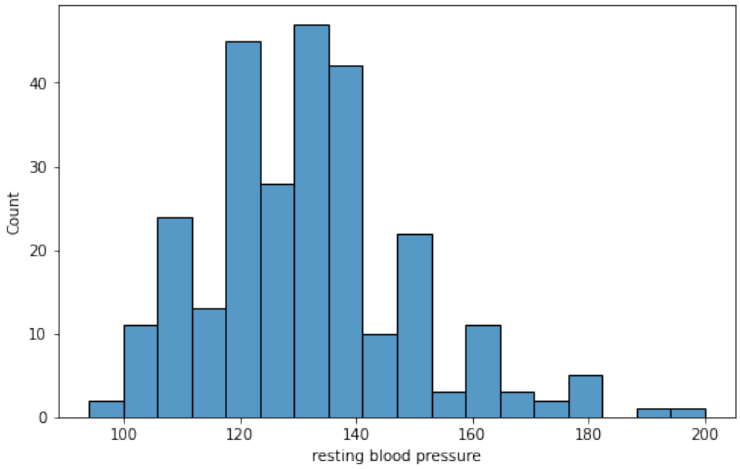
۵.۱ بخش پنجم سوال اول



به طور کلی، نسبت به حالت a تمامی ویژگی های pay-amt هم به یکدیگر و هم به ویژگی های bill-amt همبستگی بیشتری پیدا کرده اند. همچنین ویژگی های bill-amt و تمامی ویژگی های pay نیز، نسبت به حالت قبل بیشتر به یکدیگر همبسته شده اند.

۲ سوال دوم

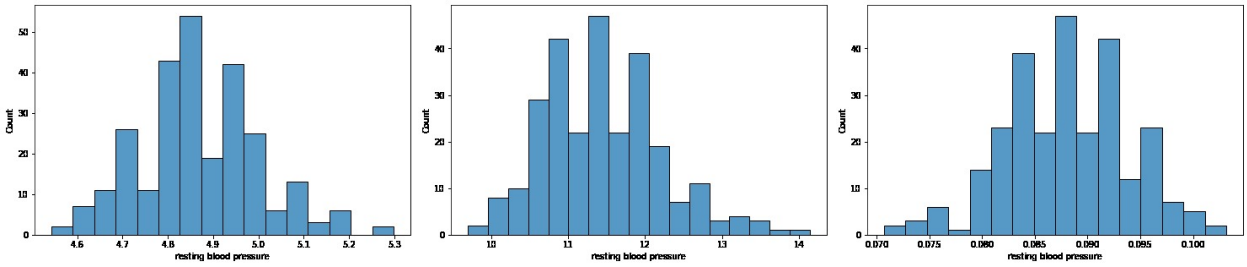
۱.۲ بخش اول سوال دوم



همانطور که مشخص است، کجی راست داریم. $Skewness = 3 * (mean - median)/standard - deviation = 0.225$

۲.۲ بخش چهارم سوال دوم

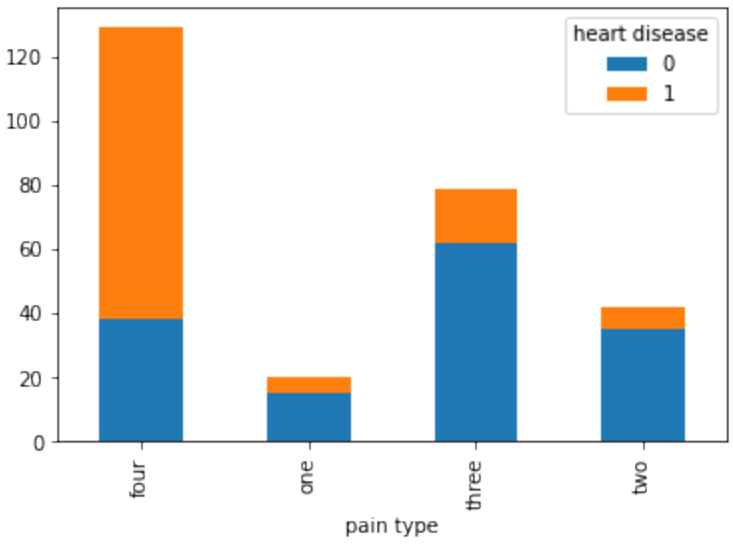
مقایسه کجی سه روش ارائه شده:



شکل ها به ترتیب از چپ به راست مربوط به روش های لگاریتم طبیعی، جذر و معکوس جذر می باشد. مقادیر کجی آنها به ترتیب، 0.0315 و 0.129 و 0.0675 می باشد. بنابراین بهترین روش حل مشکل skewness ، لگاریتم طبیعی بوده است.زیرا مقدار skewness آن از همه روش ها کمتر شده است.

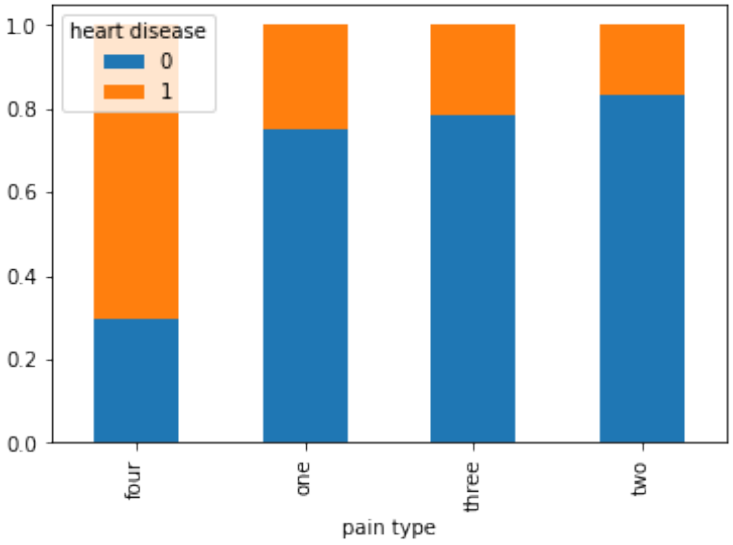
۳ سوال سوم

۱.۳ بخش اول سوال سوم



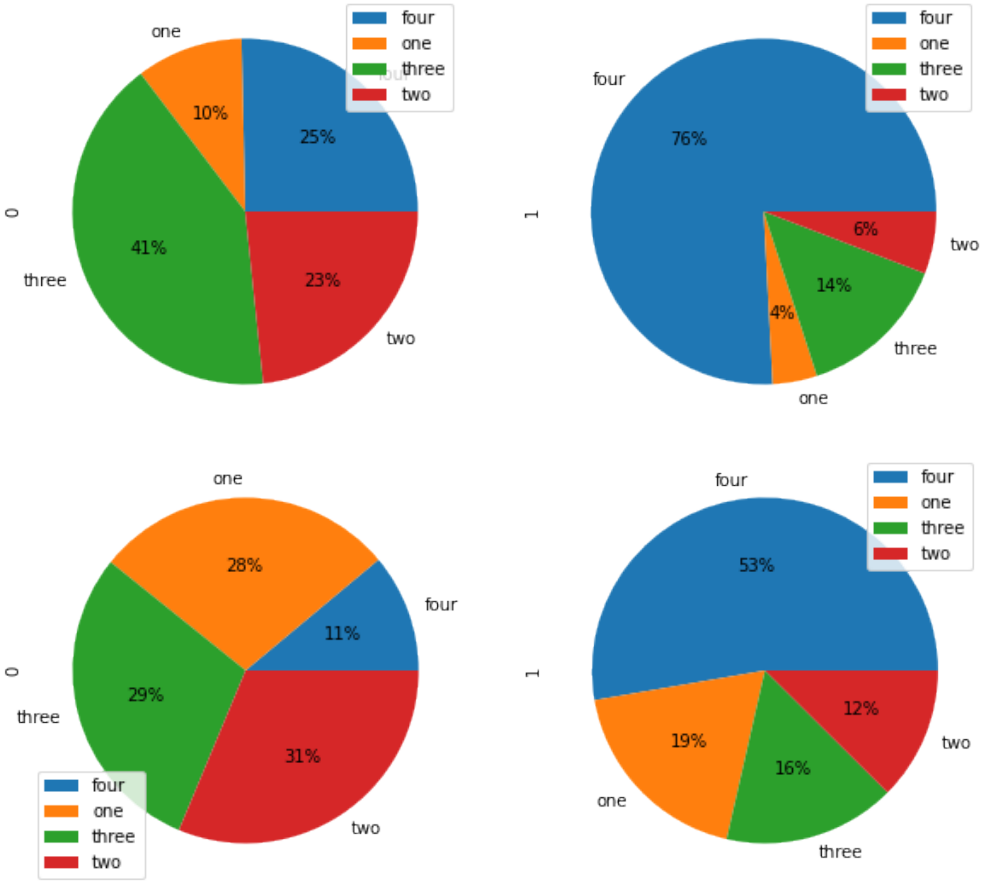
به طور کلی این نمودار نشان می دهد که کسانی که نوع درد آنها یا pain type آنها از نوع چهارم بوده است، بیشتر در معرض بیماری های قلبی بوده اند. همچنین افرادی که درد نوع سوم را داشته اند، ، بیشتر به بیماری های قلبی دچار نشده اند ترتیب تعداد افرادی که به بیماری های قلبی دچار شده اند: گروه چهارم، گروه سوم، گروه دوم و گروه اول

۲.۳ بخش دوم سوال سوم



این نمودار نشان می دهد که درصد افرادی که نوع درد آنها ازنوع اول، دوم و یا سوم بوده است و دچار بیماری قلبی شده اند حدودا باهم برابر است، اما همانطور که در قسمت قبل ذکر شد، تعداد افرادی که درد نوع چهارم را داشته اند و دچار بیماری قلبی شده اند، بسیار از بقیه دسته ها بیشتر است.

۳.۳ بخش سوم سوال سوم



نمودار بالایی ، حالت معمولی و نمودار دوم، حالت استاندارد شده را نشان می دهد. تعداد افراد در هر گروه که به بیماری قلبی دچار شده بودند، به ترتیب زیر است: گروه چهارم، گروه سوم، گروه دوم، گروه اول تعداد افرادی در هر گروه که بیماری قلبی دچار نشده بودند نیز، به ترتیب زیر است: گروه سوم، گروه چهارم، گروه دوم و گروه اول

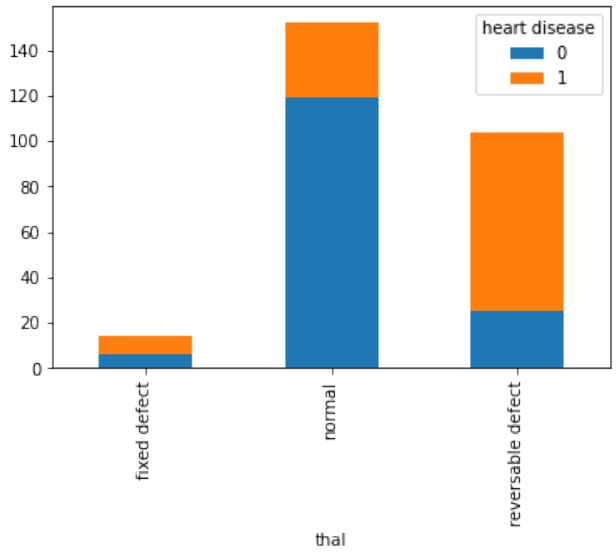
۴.۳ بخش چهارم سوال سوم

به طور کلی، از ۲۷۰ نفر، ۱۲۰ نفر به بیماری های قلبی دچار شده و ۱۵۰ نفر دچار نشده اند.همچنین فراگیر ترین درد، از نوع چهارم بوده و ۱۲۹ نفر از ۲۷۰ نفر، دچار این درد هستند. که ۹۱ نفر از این ۱۲۹ نفر، بیماری قلبی دارند. به طور کلی ، درد گروه های دیگر، کم خطرتر از گروه چهارم بوده و نشان دهنده بیماری قلبی نیست.

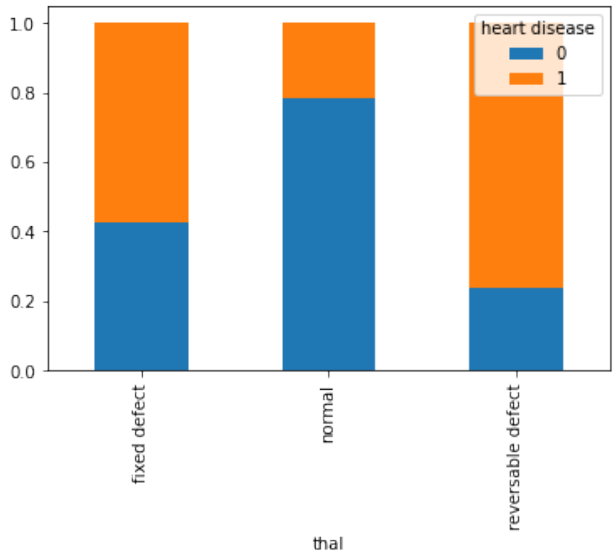
۵.۳ بخش پنجم سوال سوم

در این نمودار مشاهده می شود که ۴۷ درصد افراد، درد نوع چهارم، ۲۹ درصد آنها، درد نوع سوم، ۱۵ درصد نوع دوم و ۷ درصد درد نوع اول را تجربه می کنند. درصد کسانی که درد نوع چهارم را تجربه کرده اند، و دچار بیماری قلبی هستند برابر ۷۵ است. و تنها ۴ درصد افراد، کسانی هستند که درد نوع اول را تجربه می کنند، و دچار بیماری قلبی هستند. این نشان می دهد که درد نوع اول، کم خطر ترین درد در بین بقیه درد هاست.

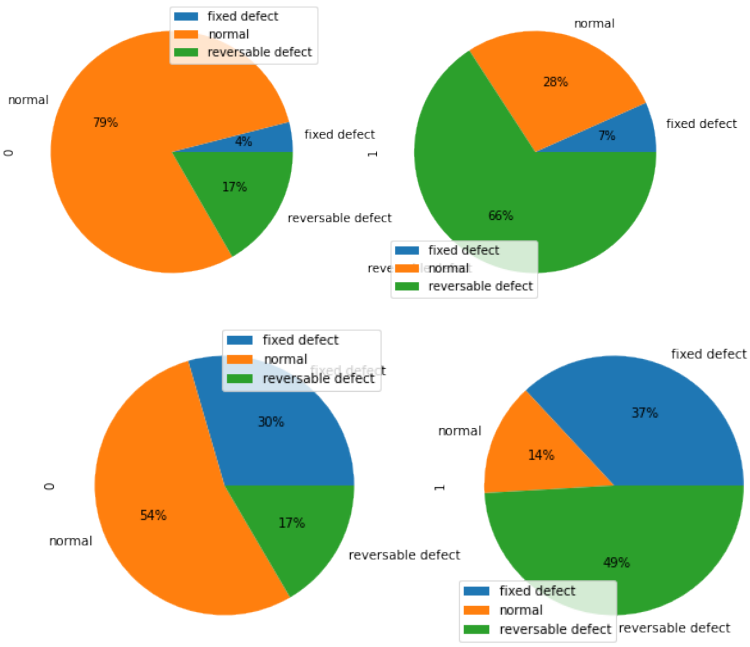
۶.۳ بخش ششم سوال سوم



نمودار اولی، تعداد افراد با وضعیت های thal مختلف و بیماری های قلبی را نشان می دهد. در این نمودار مشخص است که بیشرین تعداد افراد در دسته normal قرار می گیرند و سپس دسته reversable defect و دسته fixed defect به ترتیب از نظر تعداد قرار می گیرند. بیشترین تعداد افراد مبتلا به بیماری های قلبی در دسته reversable defect و بعد از آن در دسته normal قرار دارند.

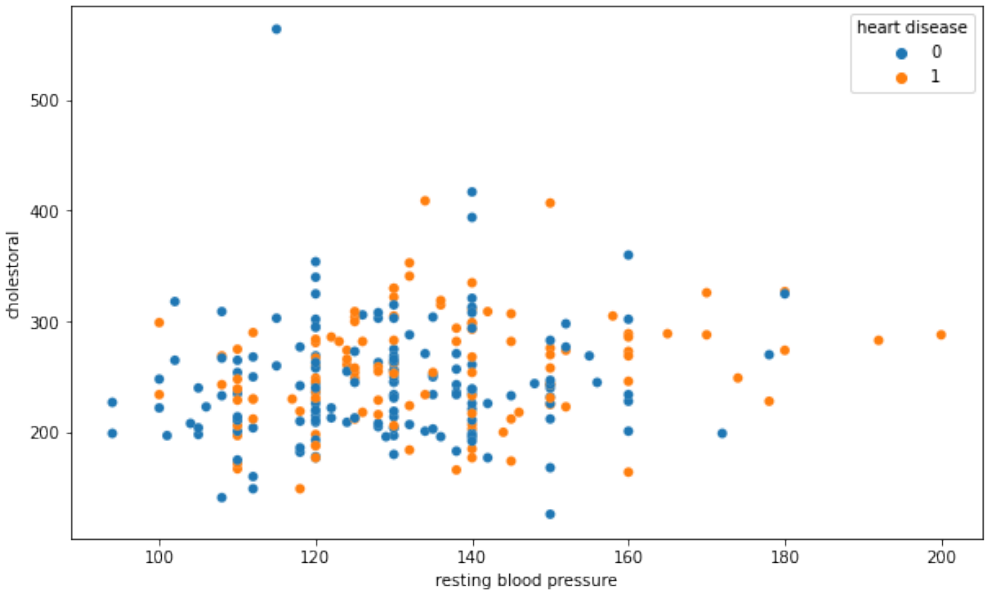


نمودار دوم، اسکیل شده حالت اول را نشان می دهد. مشخص است که درصد افراد مبتلا در دسته reversable defect بیشترین مقدار است. همچنین در مورد افرادی که وضعیت thal آنها، normal بوده است، درصد افرادی که به بیماری قلبی مبتلا نشده اند، از بقیه دسته ها بیشتر است. همچنین می توان نتیجه گرفت حدود نیمی از افرادی که وضعیت thal آنها ، fixed defect بوده است، مبتلا شده و نحدود نیمی از آنها مبتلا نشده اند.



نمودار بالا، حالت معمولی و نمودار پایین، نشان دهنده حالت اسکیل شده است. از نمودار بالا می توان نتیجه گرفت که وضعیت thal اکثر (حدود ۷۹ درصد) افرادی که مبتلا به بیماری های قلبی نشده اند، وضعیت normal بوده است. همچنین وضعیت thal اکثر افرادی (حدود ۶۶ درصد) که به بیماری قلبی مبتلا شده اند، وضعیت reversable defect بوده است. همچنین درصد افرادی که وضعیت thal آنها، fixed defect بوده است و مبتلا شده اند، بسیار ناپیز است و این نشان می دهد که این نسبتا کم خطر است.

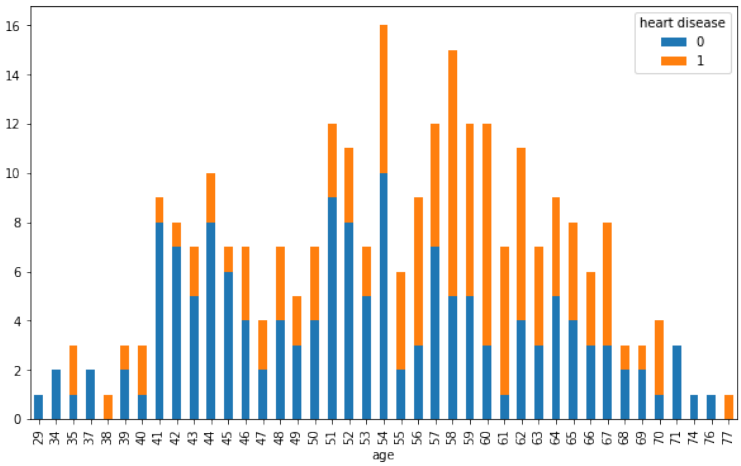
۷.۳ بخش هفتم سوال سوم



خیر، بنده رابطه خاصی بین این دو ویژگی و ویژگی هدف نیافتم.

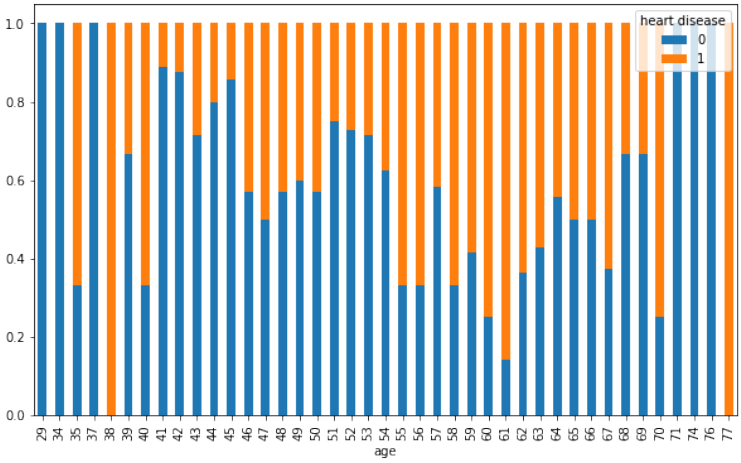
۴ سوال چهارم

۱.۴ بخش اول سوال چهارم



به طور کلی سنی که بیشترین فرد را در خود دارد، سن ۵۴ سالگی است که در مجموعه دادگان، ۱۶ نفر این سن را دارند که ۶ نفر آنها به بیماری مبتلا شده اند. می توان نتیجه گرفت که حدود نیمی از افراد ۶۳ تا ۶۷ سال، به بیماری قلبی مبتلا شده اند. و کمتر از ۴۰ درصد افراد زیر ۵۰ سال، به بیماری قلبی مبتلا شده اند.

۲.۴ بخش دوم سوال چهارم



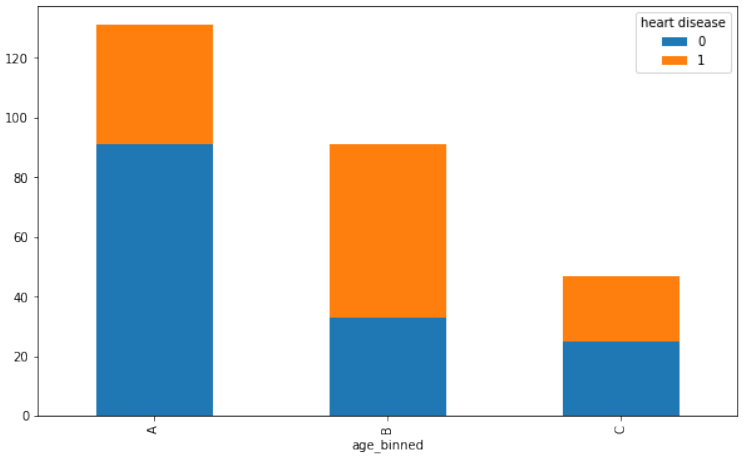
این نمودار، حالت اسکیل شده نمودار قسمت اول است. در این نمودار متوجه می شویم که چند درصد افراد هر سن به بیماری مبتلا شده اند و چند درصد مبتلا نشده اند. بنابراین می توان نتیجه بگیریم که بیشتر افراد زیر ۵۴ سال، مبتلا نشده اند. البته به جز سن های ۳۸ ، ۲۹ ، ۳۴ و ۳۷ . زیرا در هرکدام از این سن ها، یک یا دو نفر در دیتاست بوده اند که این تعداد کم قابل استناد نیست. بنابراین به همان نتیجه اولیه اکتفا می کنیم. و همچنین می توانیم نتیجه بگیریم که بیشتر افراد بین سن ۵۴ تا ۶۳ سال، به بیماری دچار شده اند. و افراد بالاتر از سن ۶۳ سال، به طور میانگین کمتر به بیماری دچار شده اند.

سال ، زیرا تنها یک بیمار ۳۸ ساله در دیتاست موجود بوده است که مبتلا نشده

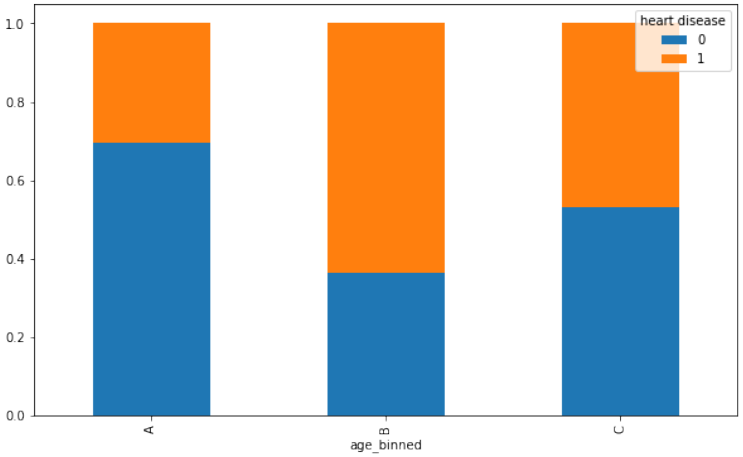
۳.۴ بخش سوم سوال چهارم

$$A = (29, 54] \; B = (54, 63] \; C = (63, 77]$$

۴.۴ بخش چهارم سوال چهارم



۵.۴ بخش پنجم سوال چهارم

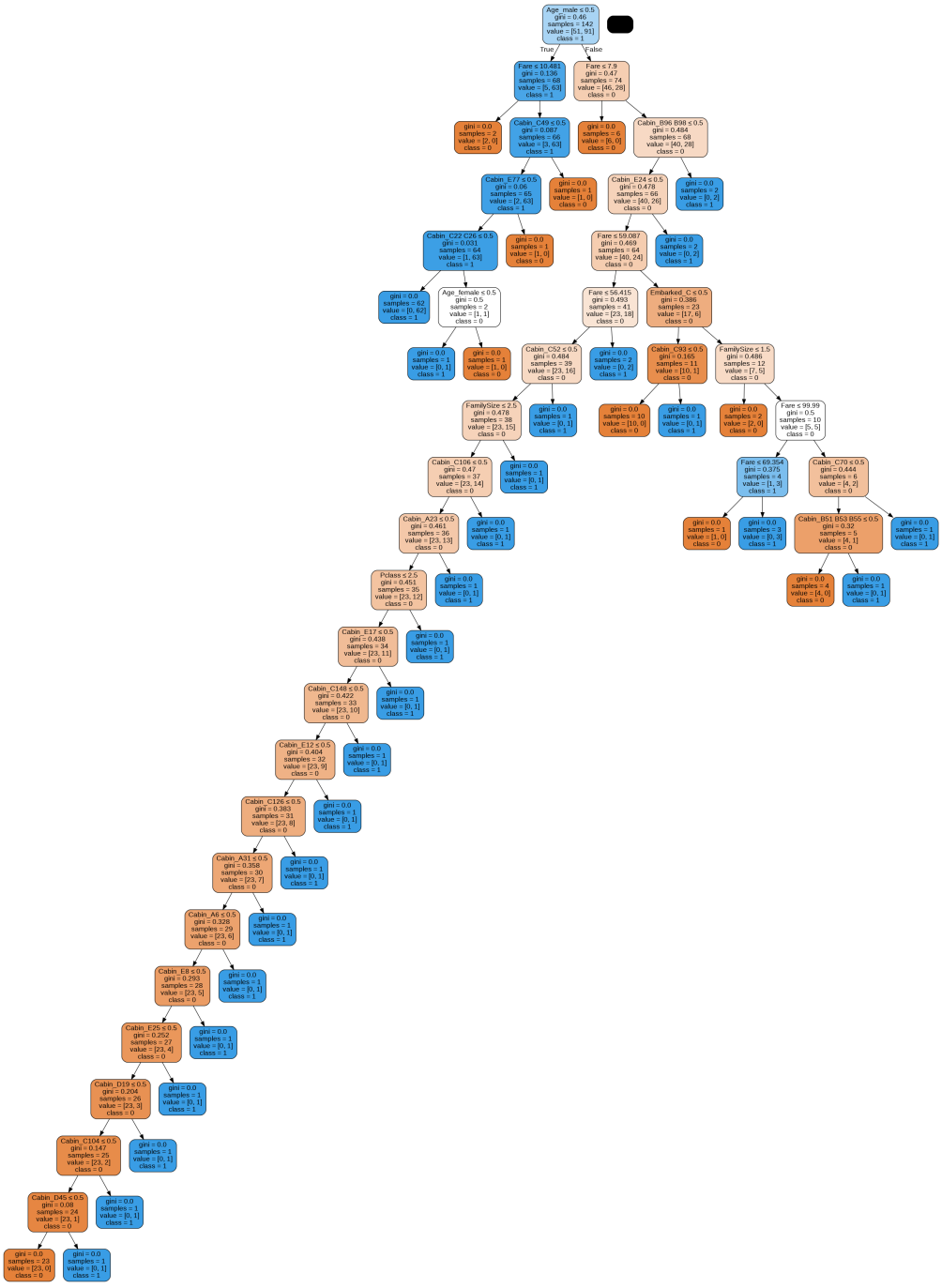


از هر دو نمودار بالا می توان نتیجه گرفت که طبق حدسی که در قسمت b زدیم، افراد در دسته اول که سن بین ۲۹ تا ۵۴ داشته اند، کمتر به بیماری مبتلا شده اند و ۶۵ درصد آنها به بیماری مبتلا نشده اند که حدود ۹۰ نفر بوده اند. یعنی ۹۰ نفر از حدود ۱۳۰ نفر افرادی که بین ۲۹ تا ۵۴ سال سن داشته اند، به بیماری مبتلا نشده اند. اما در مورد دسته دوم که افراد بین ۵۴ تا ۶۳ سال بوده اند، مشاهده می کنیم که اکثر آنها به بیماری قلبی مبتلا شده اند. یعنی حدود ۶۳ درصد آنها که حدودا ۷۰ نفر بوده اند، دچار بیماری های قلبی شده اند. در مورد دسته سوم که بالای ۶۳ سال هستند، می توان گفت که نیمی از آنها مبتلا شده و نیمی از آنها مبتلا نشده اند.

۵ سوال پنجم

۱.۵ Decision Tree

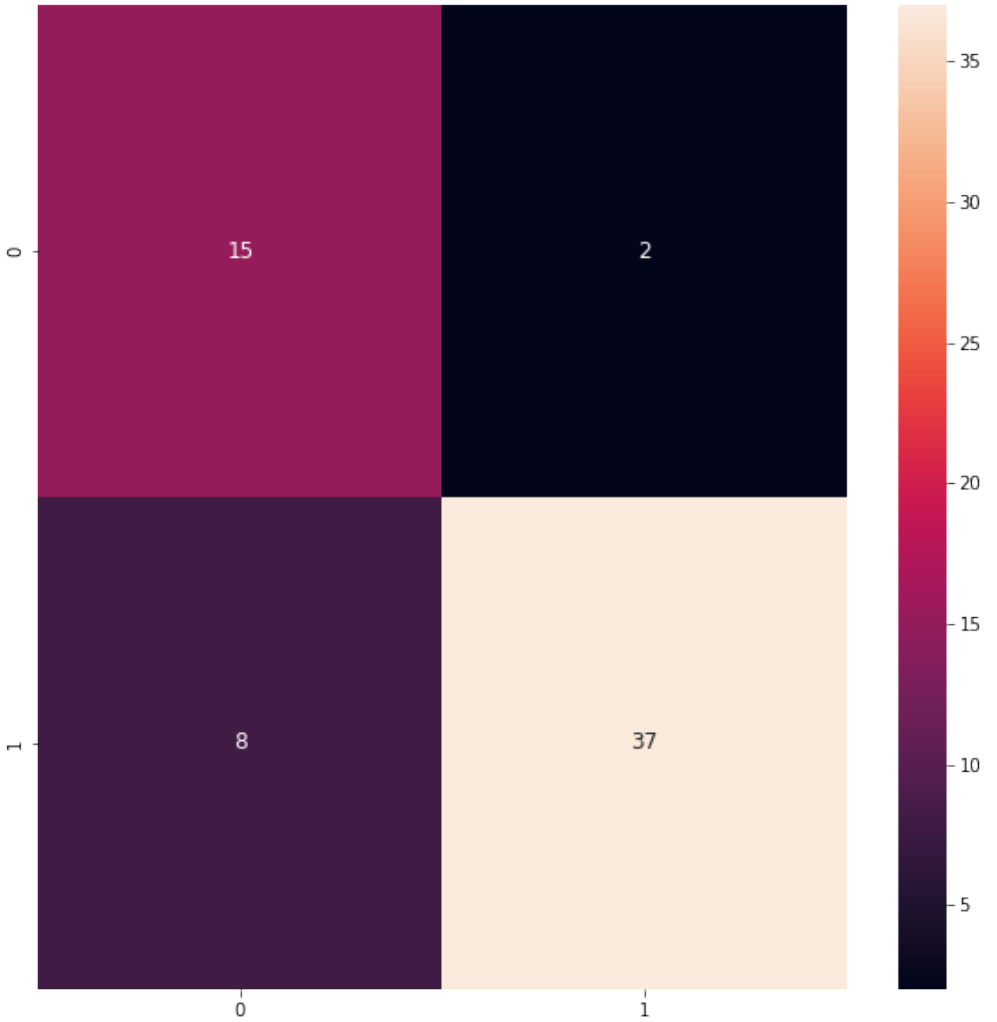
۱.۱.۵ بخش **l**



۲.۱.۵ بخش **m**

بله overfitting رخ داده است. زیرا ارتفاع درخت بسیار زیاد است و معلوم است که قدرت تعمیم پذیری ندارد. همچنین دقت training برابر ۱۰۰ درصد شده است. یعنی این درخت تمامی دادگان قسمت train را به درستی دسته بندی می کند که این تعمیم پذیری مدل را به شدت کاهش می دهد. راه حل این مشکل استفاده از ویژگی max depth است که اجازه نمی دهد ارتفاع درخت بیشتر از مقدار تعیین شده شود. همچنین می توان از min samples leaf هم استفاده کرد. و به طور کلی می توان با استفاده از hyper parameter tuning این مشکل را برطرف نمود.

۳.۱.۵ بخش **n**



تعداد کسانی که نجات یافته بودند و مدل درست تشخیص داده برابر ۳۷ نفر بوده است. تهداد افرادی که نجات نیافته بودند و مدل نیز درست تشخیص داده برابر ۱۵ نفر بوده اند. همچنین تعداد افرادی که در حقیقت نجات نیافته بودند ولی مدل آنها را نجات یافته تخمین زده برابر ۲ تفر بوده اند و در نهایت تعداد افرادی که نجات یافته بودند ولی مدل قادر به تشخیص درست آنها نشده است، برابر ۸ نفر بوده اند.

۲.۵ Hyper Parameter Tuning

۱.۲.۵ بخش o

خیر مدل بهتر نشده است.

۳.۵ Random Forest

۱.۳.۵ بخش r

تفاوتی با قسمت decision tree نکرده است.

۲.۳.۵ بخش t

با توجه به خروجی best-params معیار gini موثرتر از معیار entropy بوده است.