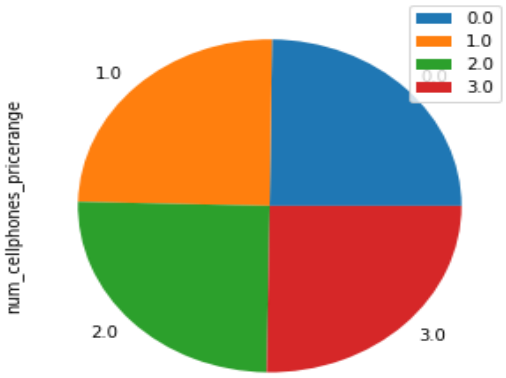
غزل دانایی-97222034

گزارش مجموعه داده اول

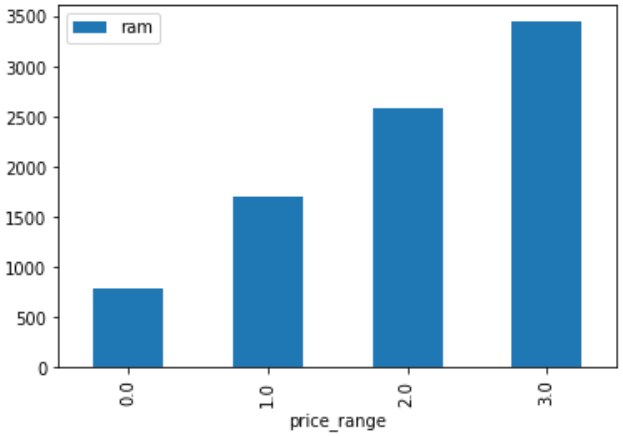
(توضیحات برای سوال های 2 تا 4 با جزییات بیشتر در نوت بوک)

سوال2-

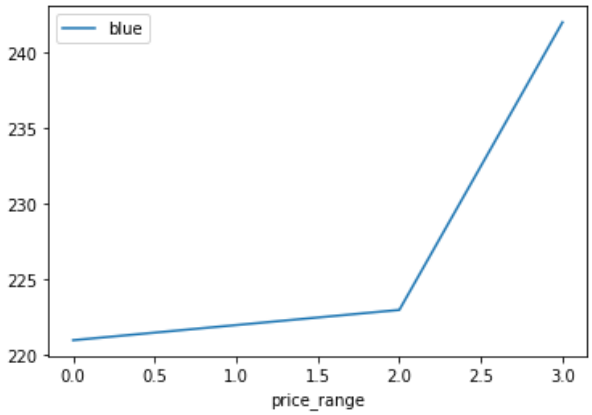
- داده ها از نظر رنج قیمتی پراکندگی متناسبی دارند.



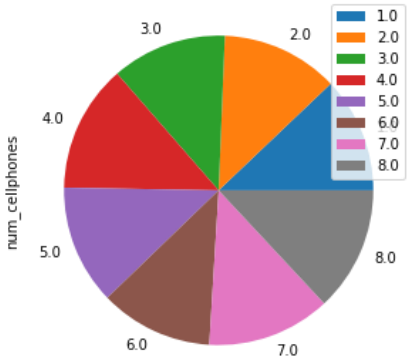
- تناسب قیمت بیشتر با مقدار رم بیشتر در موبایل ها



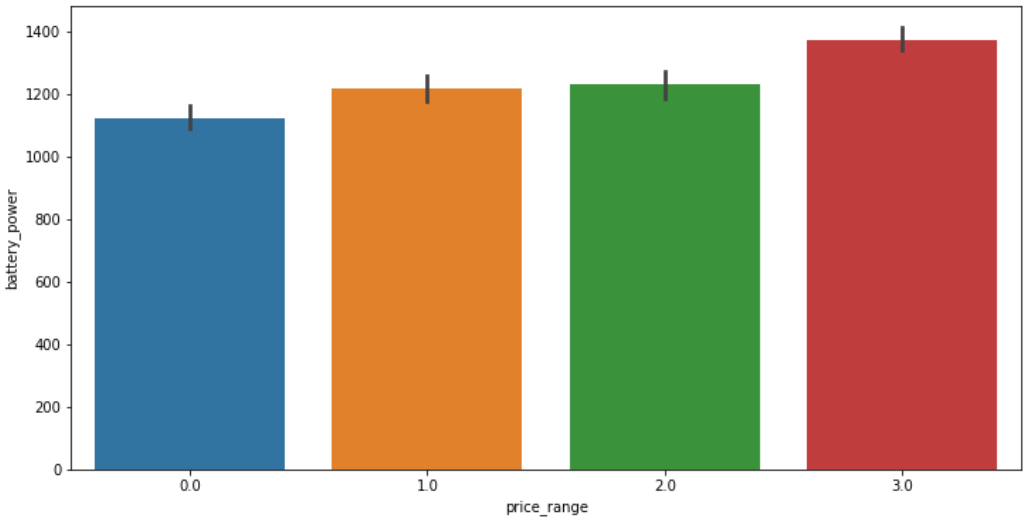
- در رنج های قیمتی بالاتر تعداد گوشی هایی که بلوتوث دارند بیشتر است.



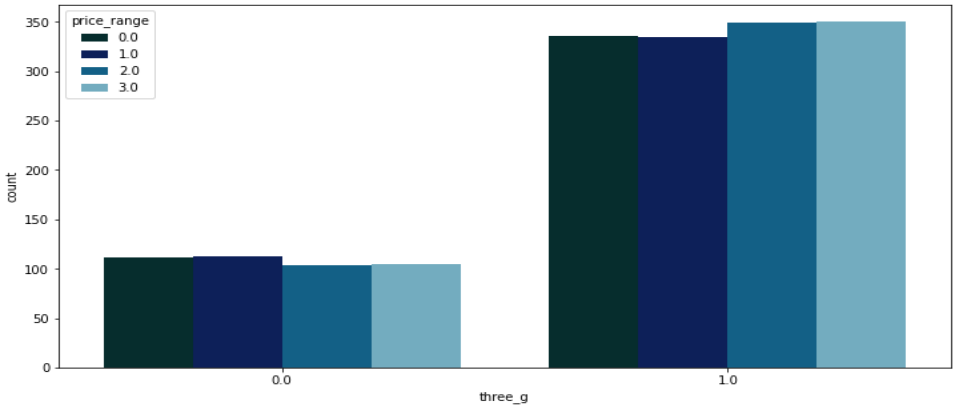
- موبایل های 4 هسته ای بیشترین تعداد را در بین سایرین دارند.



- موبایل هایی که قدرت باتری بیشتری دارند، در گروه های قیمتی بالاتری قرار دارند.



- موبایل هایی که 3g دارند بیشتر از آنهایی هستند که فاقد آن هستند و تعدادشان در گروه های قیمتی تقریبا متناسب است.



سوال 3-

سوال اول>>فرض صفر:میانگین رم در بین موبایل های رنج قیمتی بالاتر برابر با میانگین جامعه است.

فرض 1:میانگین رم در بین موبایل های با رنج قیمتی بالاتر برابر نیست با میانگین جامعه.

در t-test مقدارp-value کمتر از 0.05 است بنابراین، فرض صفر رد و فرض 1 قبول می شود.

سوال دوم>> فرض صفر:هیچ ارتباطی بین وجود دو سیم و داشتن four-g وجود ندارد.

فرض 1: ارتباطی بین وجود دو سیم و داشتن four-g وجود دارد.

در chi-2 مقدارp-value بیشتر از 0.05 است بنابراین، فرض صفر درست است و ارتباطی بین این دو ویژگی وجود ندارد.

سوال سوم>> آیا ارتباطی بین حافظه داخلی و قدرت باتری وجود دارد.

در pearson test مقدارضریب همبستگی کم است بنابراین رابطه ای وجود ندارد.

سوال چهارم>>فرض صفر:میانگین زمان صحبت در بین موبایل های رنج قیمتی بالاتر برابر با میانگین جامعه است.

فرض 1:میانگین زمان صحبت در بین موبایل های با رنج قیمتی بالاتر برابر نیست با میانگین جامعه.

در t-test مقدارp-value بیشتر از 0.05 است بنابراین، فرض صفر درست است.

سوال پنجم>>فرض صفر:میانگین تعداد هسته ها در بین موبایل های رنج قیمتی بالاتر برابر با میانگین جامعه است.

فرض 1:میانگین تعداد هسته ها در بین موبایل های با رنج قیمتی بالاتر برابر نیست با میانگین جامعه.

در t-test مقدارp-value بیشتر از 0.05 است بنابراین، فرض صفر درست است.

سوال 4-

در نوت بوک

سوال 5-

در مدل پیش بینی با روش knn تعداد داده هایی که در کلاس 0 هستند بیشتر از سایر کلاس ها درست پیش بینی شده اند اما در بقیه روش ها این مورد درست نیست و در همه کلاس ها به میزان یکسان و قابل قبولی پیش بینی درست انجام شده است. عملکرد ضعیف KNN روی این دیتاست به دلیل scale نشده بودن دیتاست می باشد.

سوال 6-

اگر داده ها متوازن نباشند می توان از تکنیک SMOTE استفاده کرد که به کمک این روش کلاس با تعداد دیتای کوچکتر را oversample می کنیم تا تعداد دیتا در همه کلاس ها متناسب شود.

یک روش دیگر برای متوازن کردن داده ها Resampling می باشد مثلا می توان از upsample یا downsample استفاده کرد. در upsample برای مثال میتوان تعداد دیتا از کلاسی که تکرار کمتری در بین داده ها دارد را افزایش داد تا نسبت تعداد داده ها در هر کلاس متناسب شود. نمونه های جدید از داده های موجود ساخته می شوند.

راه دیگر استفاده از BalancedBaggingClassifier می باشد. زمانی که از کلاسیفایر عادی استفاده میکنیم بخاطر تعداد بیشتر داده در کلاس غالب، تاثیر زیادی روی عملکرد کلاسیفایر دارد.
BalancedBaggingClassifier به ما متوازن سازی بیشتری ارایه می دهد زیرا در زمان فیت کردن دیتا یک گام اضافه برای متوازن کردن داده ها در نظر می گیرد.

سوال7-

- با اضافه کردن standard scaling در پیش بینی با روش KNN تغییر مشهودی حاصل شد به طوری که تعداد داده بیشتری به طور درست در گروه های به جز 1 پیش بینی شدند و عناصر روی قطر ماتریس confusion مربوط به کلاس های بجز 1 افزایش داشتند.
- در ادامه در پیش بینی به کمک decision tree داده ها را به کمک minmax scaling نیز علاوه بر standard scaling اسکیل شدند اما نتایج در این روش در مقایسه با ماتریسconfusion قبلی بهتر نشدند.

- در پیش بینی به کمک random forest داده ها را به کمک robust scaling نیز علاوه بر standard scaling اسکیل شدند نتایج در این روش در مقایسه با ماتریسconfusion قبلی بهتر شدند و تعداد در درایه های دیگر ماتریس بجز روی قطر کاهش یافت.

سوال 8-

میانگین درستی در test data در حالت اول (10 %) برابر است با 0.906 و در حالت دوم (20%) برابر است با 0.889 که با کمک متود score محاسبه شد. در حالتی که تعداد داده اموزشی کمتر است مدل کمتر این شانس را دارد که به بیشترین قابلیت خودش برسد.

سوال 9-

خیر در حالتی که با استفاده از pca تعداد فیچر ها کاهش یافتند، در پیش بینی با روش random forest نتایج ضعیف تری در ماتریسconfusion و درصد score مشاهده شد. استفاده ازpca برای کاهش تعداد فیچرها در

زمانی که باعث پیچیدگی بیش از حد مدل می شوند و نیز برای پیدا کردن فیچرهای با اثرگذاری بیشتر در پیش بینی مدل می باشد.

سوال 10-

بله زمانی که داده ها نامتوازن شدند عملکرد مدل بخاطر تعدد داده در کلاس غالب اثر می بیند اما چون برای بیشتر پیش بینی ها(که کلاس غالب هست) درست پیش بینی می کند درصد درستی بالایی می گیرد اما از نظر محتوایی مدل خوب کار نمی کند و در تست های دیگر خوب عمل نمی کند.

بعد از upsample کردن داده ها و اسکیل کردن همانند حالت قبل از روش random forest پیش بینی روی داده های تست انجام شد. پیش بینی های مدل در این حالت برای کلاسی که قبلا ماینور بود بیشتر شد و درایه های ماتریس confusion تعداد بیشتری برای کلاس ماینور سابق پیش بینی می کند اما تعداد حالت های اشتباهی که برای این کلاس پیش بینی می کند نیز بیشترشده است.