نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

-مقدمه:

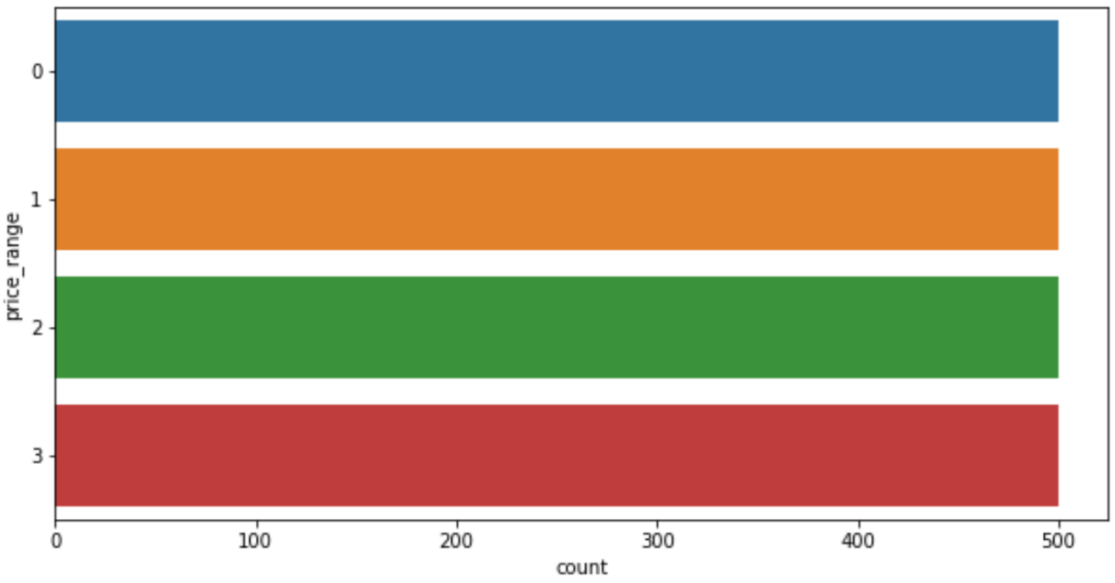
در این تمرین داده هایی از اطلاعات یک سری تلفن همراه به ما داده شده است و استخراج ویژگی های موثر در قیمت و... از ما خواسته شده است که در ادامه در بخش های مختلف به بررسی آنها می پردازیم .

خواندن داده و پیش پردازش :

در ابتدا ما داده ها را می خوانیم و ستون هایی که دارای دیتای null هستند را مورد بررسی قرار می دهیم .

مشاهده می کنیم که همه ی داده ها عددی و ناتهی هستند .

تعداد نمونه های هر کلاس برای محدوده قیمتی را بررسی می کنیم



مشاهده می شود که تعداد همه با هم برابر و 500 عدد می باشد .

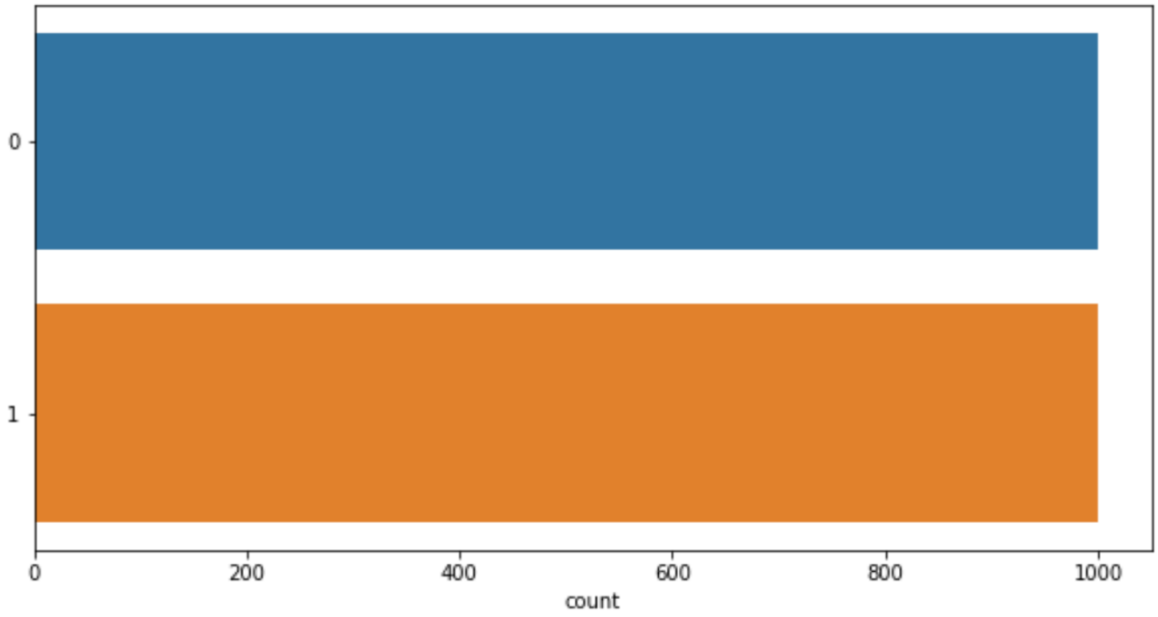
نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش 1 :

ابتدا داده ها را به دو دسته گران(1) و ارزان(0) تقسیم میکنیم به اینصورت که داده های محدوده قیمتی 0و1 را ارزان و نحدوده قیمتی 2و3 را گران در نظر میگیریم .



در ادامه متد انتخاب پیشرو را پیاده سازی کردیم که بر اساس AUC فیچر ها را بر میگزینند . نتیجه به شکل زیر می باشد . 6 ویژگی ابتدایی را انتخاب می کنیم و به مدل می دهیم که عبارتند از :

'ram', 'battery_power', 'px_height','px_width', 'mobile_wt', 'int_memory'

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

	feature to add	ROC AUC
0	ram	0.975568
1	battery_power	0.987810
2	px_height	0.996815
3	px_width	0.999420
4	mobile_wt	0.999740
5	int_memory	0.999785
6	sc_h	0.997565
7	m_dep	0.974665
8	wifi	0.972285
9	three_g	0.971895
10	blue	0.970560
11	touch_screen	0.969415
12	dual_sim	0.970585
13	four_g	0.968710
14	clock_speed	0.968100
15	fc	0.964600
16	talk_time	0.964870
17	n_cores	0.964090
18	pc	0.965150
19	sc_w	0.964540

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش 2 :

در این بخش با استفاده از ویژگی های استخراج شده بخش 1 مدل رگرسیون لاجستیک خود را اجرا کردیم و نتایج زیر حاصل شد :

```
score 0.9975
f1_score 0.9974992341404556
precision_score 0.9974619289340101
recall_score 0.9975490196078431
```

اما با استفاده از همه ی ویژگی ها نتایج به شکل زیر می باشد :

```
score 0.9
f1_score 0.899877349753448
precision_score 0.900501253132832
recall_score 0.8996598639455782
```

بخش 3و4 :

در این قسمت از طریق اجرای الگوریتم PCA بر روی داده ها اجرا می کنیم و تعداد ویژگی ها را کم میکنیم (برابر تعداد ویژگی ها در حالت استخراج شده از بخش 1) و به 6 ویژگی می رسانیم و سپس رگرسیون لاجستیک را پیاده سازی می کنیم . نتایج زیر بدست می آید که نتایج فوق العاده ای است :

```
score 0.995
f1_score 0.9949998749968749
precision_score 0.995
recall_score 0.995049504950495
```

بخش 5 :

کرنل ها برای تفکیک کردن داده ها کاربرد دارند و به ما برای این موضوع کمکگر هستند و با افزایش ابعاد فضای ویژگی تفکیک پذیری داده ها را برای ما راحت تر می کنند . در زیر به چند نمونه کرنل و موارد غالب استفاده آنها اشاره می کنیم .

الف) کرنل خطی : برای توابع خطی به کار می رود .

ب) کرنل های گاوسی : این کرنل ها از توزیع گاوسی پیروی می کنند و بسیار پرکاربردند .

ج) کرنل های RBF : مشابه کرنل های گاوسی هستند با این تفاوت که شعاعی را نیز برای تفکیک داده ها در نظر می گیرند . این نوع کرنل هم بسیار پرکاربرد است .

د) کرنل سیگموید : در شبکه های عصبی و مباحثث مربوط به یادگیری عمیق بیشتر کاربرد دارد .

ه) کرنل های چند جمله ای : این گونه از کرنل ها بیشتر در پردازش تصویر کاربرد دارند .

و) Linear splines kernel in one-dimension : این کرنل بیشتر برای برای داده هایی با پراکندگی بالا (مانند تشخیص متن) کاربرد دارد .

بخش 6 :

در این تمرین از ما خواسته شده که بر روی دیتای موبایل SVM اجرا کنیم که پس از بررسی عدم تهی بودن و بررسی پراکندگی و دیگر پیش پردازش های مقدماتی مدل SVM پیاده سازی و اجرا شد که به ما دقت خوب 97٪ را داد .

بخش 7 :

در این بخش از تمرین از ما خواسته شده بود که کرنل های مختلف را با پارامتر های مختلف روی دیتاست بررسی کنیم .

کرنل خطی : دقت در این حالت 96.67٪ بود .

کرنل rbf : دقت 97٪ بود .

کرنل سیگموید : دقت در حالت سیگموید بسیار پایین و حدود 17 درصد بود و با تغییر پارامتر coef0 به 0.2 دقت به حدود 14 درصد کاهش یافت .

کرنل چند جمله ای : در حالت های مختلف با پارامتر های مختلف برسی شد . نتیجه ان بود که با افزایش درجه چند جمله ای دقت کاهش پیدا می کند و به ترتیب دقت برای درجه های 2و3و4و10و100 برابر 96.67 و 96.33 و 96 و 94.67 و 69.67 درصد بود .

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

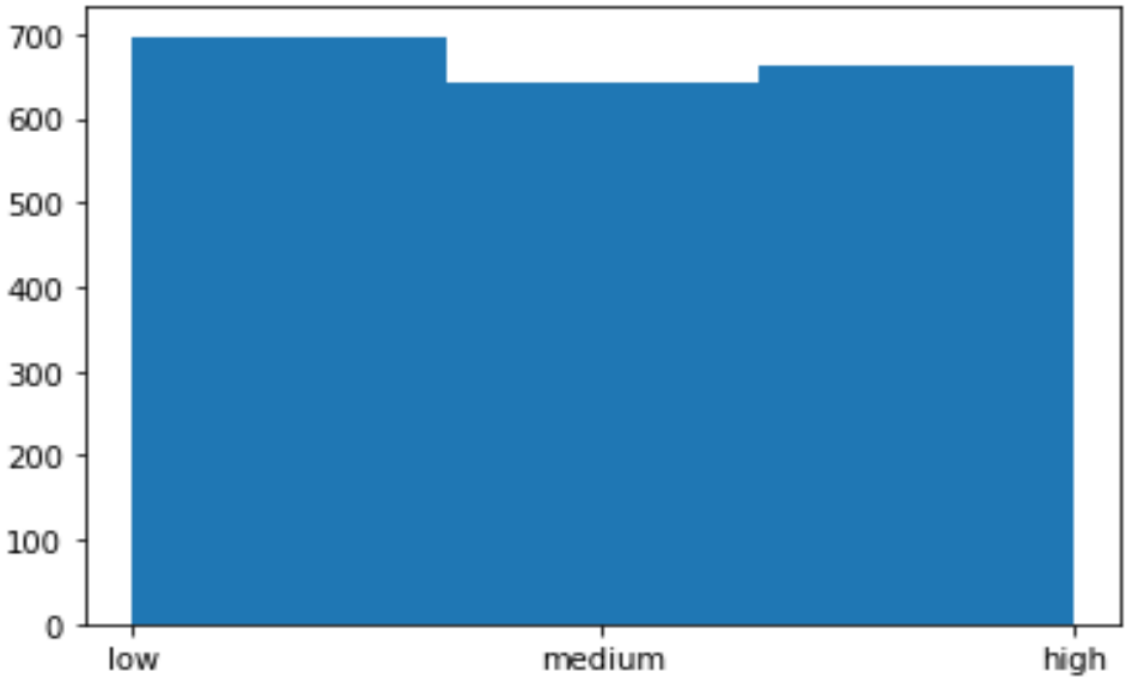
بخش 8 :

هنگامی که از هارد مارجین استفاده می کنیم به دلیل عدم miss classification دقت داده های آموزشی بالاست اما از طرفی چون به گونه ای داده ها اورفیت می شود ، دقت تست پایین تر از حالت سافت مارجین است . (دقت soft margin برابر 98.33٪ و دقت hard margin برابر 96.67٪)

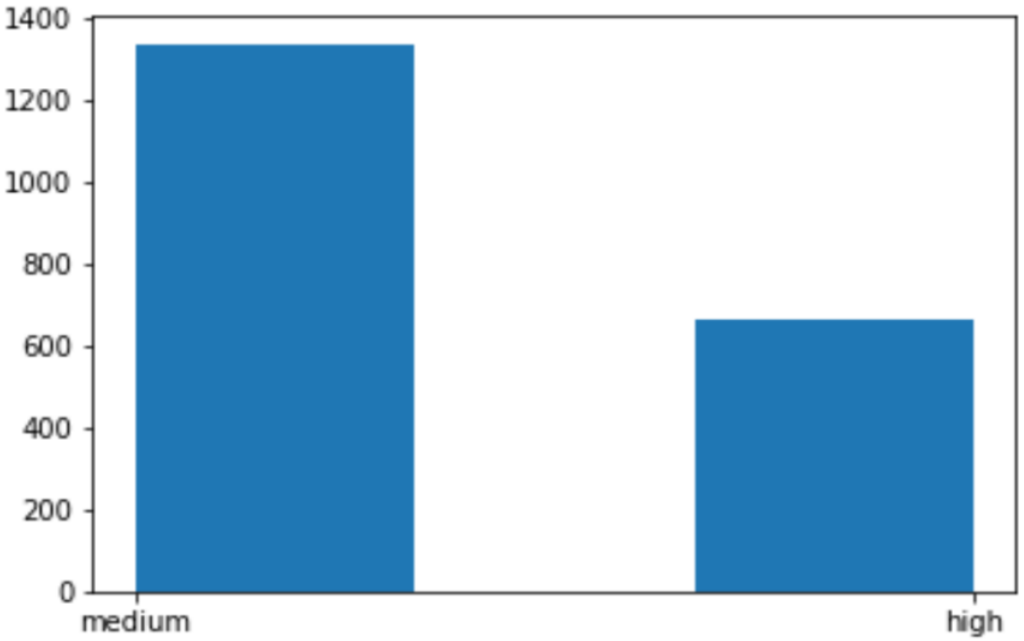
بخش 9 الف :

سه اندازه مختلف ، binهای مختلف را در نظر می گیریم.

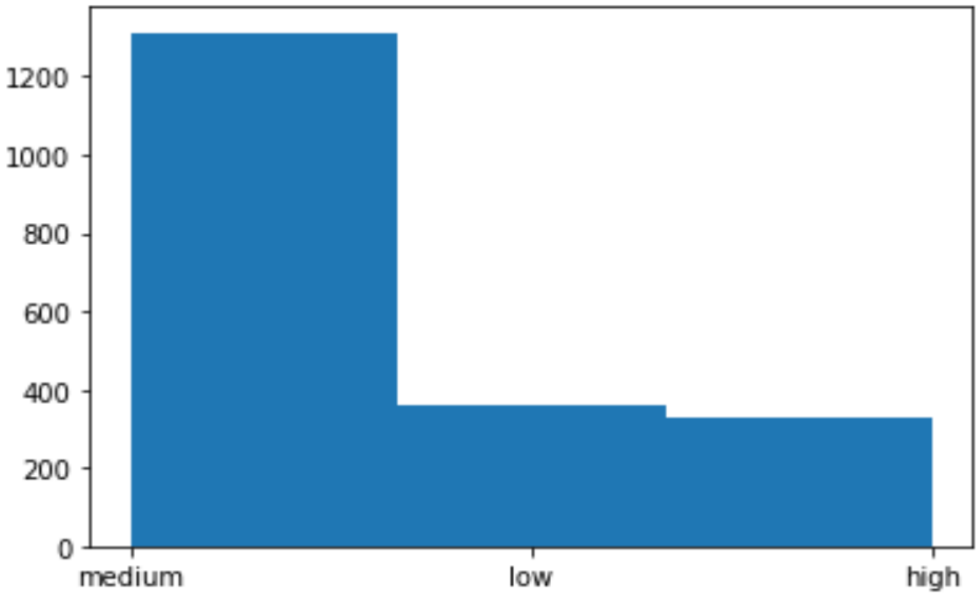
الف) سه قسمت مساوی در دامنه اعداد : که توزیع آن به شکل زیر است :



ب) bins = [0, 499, 1499, 2000] :



ج)bins = [0, 750, 1750, 2000] :



پس از اجرای SVM دقت برابر 83٪ شد .

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش9 ب :

این دیتا ست داده ی کتگوریکال ندارد و تمامی داده ها عددی هستند .

این روش برای عددی کردن داده ها است و با جدا سازی همه متغیر های یک ستون به ستون های مجزا باعث میشود اگر مقداری در ستون مهم بود ، مدل ما راحت تر آن را کشف کند.

بخش 9 ج :

استفاده از تبدیل ها باعث می شود که داده ها توزیع بهتری داشته باشند .

Log transform بیشتر مواقعی به کار میرود که داده ها گوناگونی زیادی دارند و از توزیع نرمال دور هستند (توزیع نمایی دارند) و توزیع آنها چولگی داشته باشد .پس از انجام این تبدیل داده ها به توزیع نرمال نزدیک میشوند . همچنین میزان تاثیر داده های پرت را نیز با اینکار کم میکند و مدل را بهتر میکند.

در ابتدا مینیمم داده ها را برای هر ویژگی میگیریم تا چک کنیم ببینیم نا منفی نباشد .

که مشاهده می کنیم کمترین مقدار صفر است . به این دلیل همه را یک واحد به راست شیفت میدهیم .

دقت مدل svm در این حالت برابر 89.25٪ بود .

بخش 9 د :

با ضرب طول در عرض مساحت رابدست می آوریم و به جای این دوستون جایگذاری میکنیم و مدل را اجرا می کنیم . با این تغییر دقت به 31% کاهش یافت .

بخش 10 :

مدل SVM به طور جداگانه برای هر کدام از بخش های تمرین 5 اجرا شده است و گزارش آن نوشته شده است . حال یک مدل SVM کلی را پیاده سازی میکنیم .

دقت در این حالت حدود 87 درصد بود.

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش 11 :

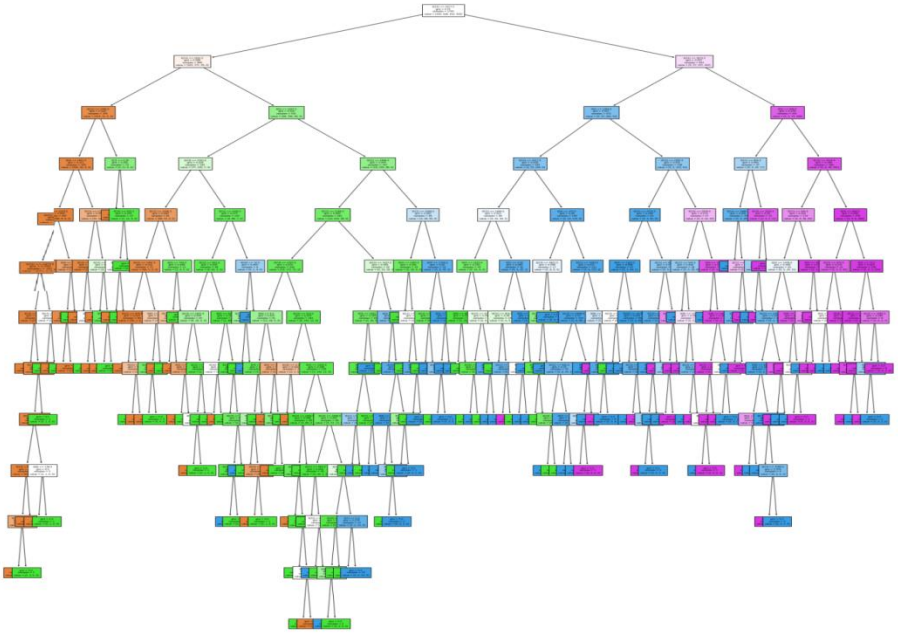
الگوریتم های مختلفی برای درخت تصمیم وجود دارد که برای مثال در برخی از آنها از هرس کردن استفاده میشود و... .

الگوریتم اول ID3است یا نام کامل آن Iterative Dichotomiser 3 که این الگوریتم از information gainبرای ساخت درخت استفاده میکند. این معیار تعیین میکند که کدام ویژگی ها اطلاعات بیشتری دارند و برای ساخت درخت لازم هستند . آنهایی که اطلاعات بیشتری دارند در راس درخت قرار می گیرند و بدین ترتیب زیر درخت های دیگر نیز ساخته میشوند. این الگوریتم مخصوص داده های پیوسته است .

الگوریتم دیگر ، الگوریتم CHAIDاست. در این الگوریتم برای ساخت درخت، داده ها را متناوبا به زیر مجمعه های یکسان تقسیم میکند تا جایی که هر زیرمجموعه دارای تعداد مشخصی نمونه شود. این الگوریتم از آزمون Chi squaredبرای تصمیم گیری در هر تقسیم برای مشخص کردن زیر درخت ها استفاده می کند .

بخش 12 :

با استفاده از پکیج sklearn یک مدل درخت تصمیم می سازیم و آنرا پیاده سازی می کنیم . دقت در این حالت برابر84.66% بود .



نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش 13 :

در ابتدا عمق درخت را زیاد میکنیم و مشاهده می کنیم با افزایش عمق درخت ، دقت نیز افزایش می یابد .

در حالتی که عمق درخت 5 بود ، دقت برابر 89.33٪ ولی هنگامی که عمق را به 10 و 50 افزایش دادیم دقت به ترتیب برابر 98.67٪ و 98٪ شد .

همچنین با افزایش تعداد ویژگی نیز دقت بالا می رفت . برای مثال برای ماکسیمم تعداد ویژگی 2 و 5 و 10 ، دقت به ترتیب برابر 90٪ و 95.67٪ و 96٪ شد .

بخش 14 :

هرس کردن یعنی اینکه ما بخشی از درخت تصمیممان را در نظر نگیریم .

هرس کردن به گونه ای مانند drop out است و از پیچیدگی مدل ما کم میکند و مانع از اورفیت شدن می شود.

بخش 15 :

روشbootstrappingاز روشهای نمونه گیری مجدد است وروش کارآمدی برای محاسبه میزان دقت و خطای استاندارد(standard error)متغیر تخمین زده شده است.در روشbootstrappingبه صورت تصادفی nعضو با جایگذاری انتخاب میکنیم و در هر مرحله سعی بر کم کردن خطا داریم .

از بوت استرپینگ می‌توان در ساخت آزمون فرض آماری استفاده کرد. از این روش معمولاً به عنوان جایگزینی برای روش‌های استنباطی بر پایه فرضهای پارامتری هنگامی که در مورد این فرضها شک داشته باشیم استفاده می‌شود. همچنین در استنباط پارامتری زمانی که محاسبه خطای استاندار فرمول محاسباتی پیچیده شود از بوت استرپینگ استفاده می‌کنیم.

در روشcross validationهر بار یک زیرمجموعه از دادههای آموزشی را کنار میگذاریم،مدل را با استفاده از بقیه دادهها آموزش میدهیم واز دادههای کنار گذاشته شده به عنوان دادههای تست استفاده میکنیم.

این روش برای نمونه گیری از داده اصلی است که با جایگذاری انجام میشود یعنی هر داده ای که وارد نمونه گیری میشود دوباره پس از آنکه کار آن تمام شد به دیتاست برمیگردد و باز هم میتواند در نمونه گیری های بعدی شرکت داده شود و سپس میانگین دقت و خطا را برای سنجش در نظر می گیریم . تفاوت این روش با cross validation این است که در cross validation ما بخشی از دیتا که برای سنجش است را به کل از مجموعه داده جدا می کنیم اما در BootStrapping اینگونه نیست.

بخش 16 :

در 2-fold cross-validationداده ها را بصورت رندوم به 2بخش هم اندازه تقسیم می کنیم. یک بخش به عنوان داده ی تست و دیگری بعنوان داده ی ترین استفاده می شود و بار دیگر همین کار را با بخش دیگر انجام مید هیم در نتیجه دو معیار خطای و دقت خواهیم داشت که می توانیم از آن ها میانگین بگیریم و بعنوان خطای و دقت مدل گزارش دهیم. حال همین کار را 5 بار مکررا انجام می دهیم و برای سنجش میانگین نتایج خطا و دقت این 5 مرتبه را گزارش می دهیم . این کار احتمالا باعث می شود که خطا و دقت بهتر و دقیق تری را گزارش دهیم .

بخش 17 :

به طور کلی تا به حال در متد های مختلف از این روش استفاده کرده ایم مثلا در فرایند توقف زود هنگام در آموزش شبکه های عصبی هنگامی که خطا بر روی تست روبه افزایش میرود فرایند آموزش متوقف میشود که همان نمایان گر متدelbowکه در سوال گفته شده است میباشد و به طور کلی میدانیم که هر تابع خطایی را میتوان به حاصل جمع دو خطای بایس و واریانس تبدیل کردو همواره خطای تست برابر جمع این دو خطا می باشد پس میتوان نقطهelbowرامحلی که از آنجاواریانس روبه افزایش است در نظر گرفت و میتوان بایس و واریانس کمی داشت، پس میتوان همواره از این متد برای انتخاب مدل استفاده کرد.

نام و نام خانوادگی: محمدرضا ضیالاری

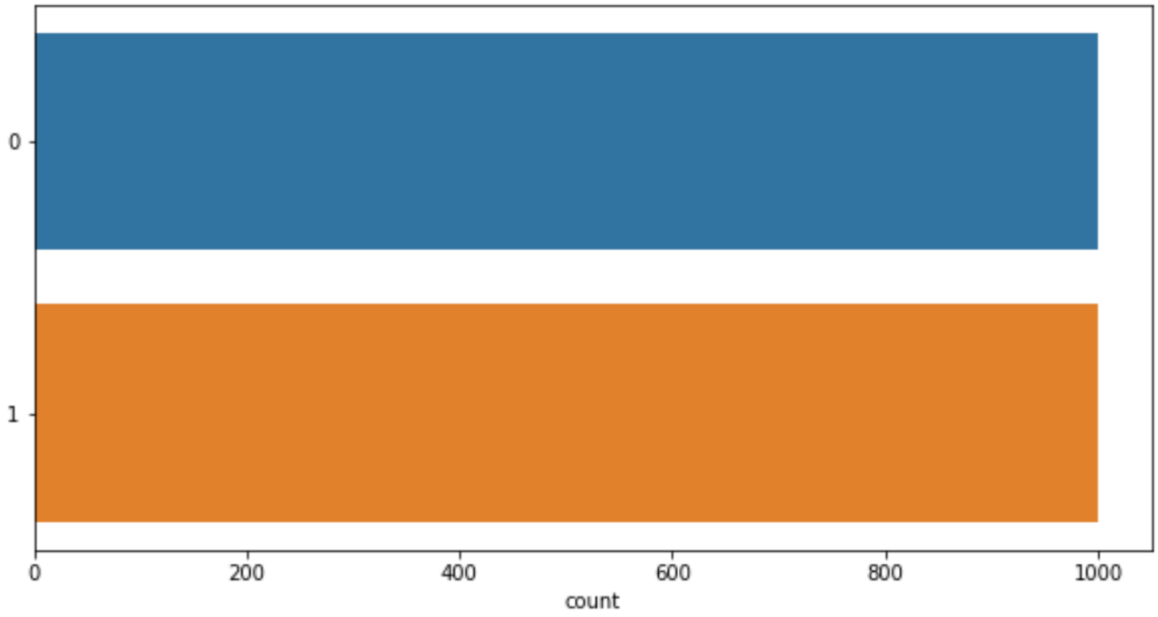
شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

تسک های امتیازی :

بخش 1 :

ابتدا داده ها را به دو دسته گران(1) و ارزان(0) تقسیم میکنیم به اینصورت که داده های محدوده قیمتی 0و1 را ارزان و نحدوده قیمتی 2و3 را گران در نظر میگیریم .



در ادامه متد حذف پسرو را پیاده سازی کردیم که بر اساس AUC فیچر ها را بر حذف می کند . نتیجه به شکل زیر می باشد . 6 ویژگی ابتدایی را انتخاب می کنیم و به مدل می دهیم که عبارتند از :

'ram', 'battery_power', 'px_height','mobile_wt', 'sc_h', 'int_memory'

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

	feature removed	ROC AUC
0	ram	0.000000
1	battery_power	0.975568
2	px_height	0.987810
3	mobile_wt	0.996815
4	sc_h	0.996815
5	int_memory	0.996815
6	pc	0.996865
7	talk_time	0.968440
8	n_cores	0.963430
9	clock_speed	0.964020
10	wifi	0.964745
11	blue	0.964785
12	px_width	0.964625
13	touch_screen	0.965165
14	m_dep	0.964875
15	four_g	0.964995
16	dual_sim	0.965440
17	three_g	0.965660
18	fc	0.965380
19	sc_w	0.965175

در این بخش با استفاده از ویژگی های استخراج شده مدل رگرسیون لاجستیک خود را اجرا کردیم و نتایج زیر حاصل شد :

```
score 0.9675
f1_score 0.9674900438259217
precision_score 0.9674551774149184
recall_score 0.9675370148059224
```

که نسبت به حالت انتخاب پیشرو دقت کمتری دارد اما نسبت به حالتی که همه ویژگی ها به عنوان ورودی به مدل داده شوند دقت بالا تر است .

در زیر نتایج مربوط به انتخاب پیشرو را مشاهده می کنیم :

```
score 0.9975
f1_score 0.9974992341404556
precision_score 0.9974619289340101
recall_score 0.9975490196078431
```

نام و نام خانوادگی: محمدرضا ضیالاری

شماره دانشجویی: 97222057

شماره تمرین:بخش اول تمرین شماره 2 درس داده کاوی

بخش 2 :

یکی از متد های معروف در این بخشcv 5x2است(نه آن متدی که در سوالات قبل اشاره شد) در این روش برای دو مدلی که قصد مقایسه آن ها راداریم5بارcv 2-foldرا انجام میدهیم و نتایج آن ها را نگه میداریم چون این نتایج به هم وابسته اند میتوان بر روی آن ها یکt-testزد و اگرpvalueکمتر از 0.05باشد میگوییم که فرقی به دو مدل وجود ندارد در غیر این صورت یکی از مدل ها موفق تر عمل کرده است .

بخش 3 :

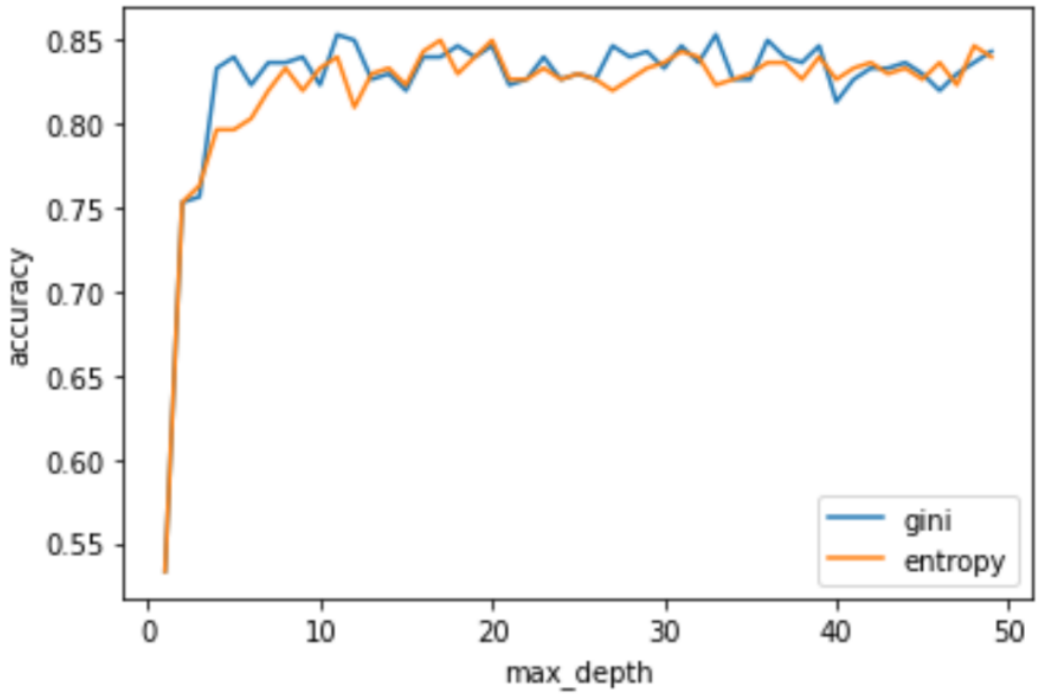
Matthews correlation coefficientیک ملاک برای پیدا کردن دقت مدل هادر تسک طبقه بندی دو کلاسه یا چند کلاسه میباشد و مانندf1-scoreمیتواند در شرایطی که اندازه کلاس ها برابر نیستند نیز خوب عمل کند، در این ملاک مدل رندوم مقدار0 بهترین مدل مقدار1 و مدل کاملابرعکس مقدار -1 را میگیرد، و از طریق فرمول زیر محاسبه میشود:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

تفاوت MCC با f1 –score این است که f1-scoreدر فرمول خود به مقادیرTNتوجه نمیکند اما این ملاک این کار را انجام میدهد،در جاهایی که نیازی به تفسیر ملاک نداری مو میخواهیم مدل بر روی هر دو کلاس یا همه کلاس ها خوب عمل کند و به طور کلی در دیتا ست های غیر بالانس ملاک خوبی برای استفاده است.

بخش 4 :

در این بخش به ازای عمق های1تا50 درخت های تصمیم با معیار Index Giniو نیزEntropy را تشکیل دادیم و نتایج بصورت زیر بود:



میبینیم که معیار gini عملکرد بهتری را داشته است اما باز هم هرس کردن تاثیر مثبتی نداشته زیرا دقت در حوالی 80 تا 85 درصد تغییر می کند اما بدون این کار دقت های ما بالای 90 درصد بود .