

**گزارش تمرین شماره 3**

**درس داده کاوی**

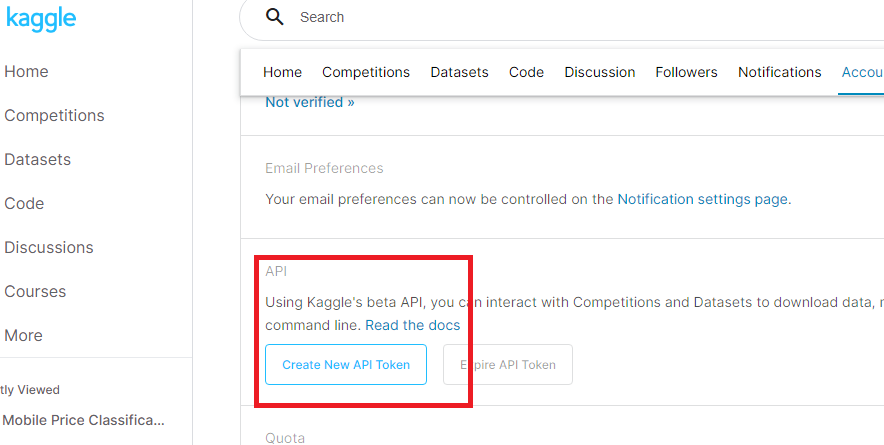
**جناب آقای دکتر فراهانی**

**جناب آقای دکتر خرد پیشه**

**نسرین هداوند سوری**

**10/03/1400**

**در ارتباط قسمت اول سوالات استفاده از دیتاست کگل در کولب و اجرا کردم ولی خوب با Jupiter Notebook راحت تر می تونم کار کنم ادامه تمرین ها را روی سیستم خودم انجام دادم**

****

****

**بررسی سوال 1 »**

**دلیل استفاده از کرنل در ماشین بردار پشتیبان:**

هدف اول در طبقه بندی با ماشین بردار پشتیبان، ایجاد بیشترین فاصله بین بردارهای پشتیبان هر کلاس می باشد. حال گاهی اوقات ممکن است فضای ویژگی به شکلی باشد که با نگاشت داده ها به فضای با ابعاد بالاتر، رسیدن به این هدف به بهترین شکل ممکن محقق شود

این عمل با استفاده از توابع کرنل و صرفا با استفاده از توابع کرنل انجام شده و نیازی به نگاشت مختصات داده ها نیست.

**کرنل های پرکاربرد:**

**کرنل خطی**

پایه ای ترین نوع کرنل محسوب می شود که معمولا  زمانی بکار می رود که تعداد فیچر بالا داشته باشیم. معمولا عملکرد آنها در دیتای متن از سایر کرنل ها بهتر است. تابع این کرنل به شکل زیر است:

F(x, xj) = sum( x.xj)

**کرنل چند جمله ای :**

حالت عمومی تری از کرنل خطی است. بطور کلی استفاده کمتری از سایر کرنل ها دارد. تابع این کرنل به شکل زیر است :

F(x, xj) = (x.xj+1)^d

**کرنل RBF**  :

یکی از پرکاربردترین کرنل ها می باشد که برای دیتاهای غیر خطی بکار می رود. مهمترین ویژگی این کرنل نگاشت فضای داده به یک بازه محدود می باشد. تابع تعریف کننده این کرنل به شکل زیر است:

F(x, xj) = exp(-gamma \* ||x - xj||^2)

**کرنل گوسی :**

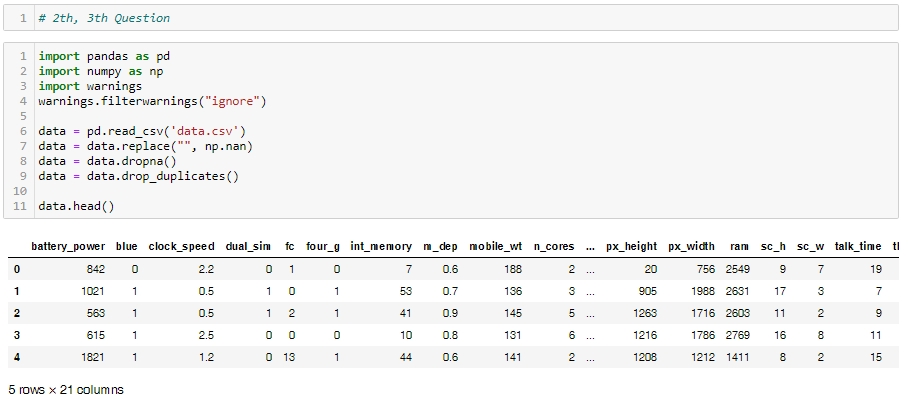
از این کرنل زمانیکه دانش پیشین از دیتا وجود ندارد استفاده می شود.

F(x, xj) = exp(-(x - xj)/2\*sigma\*\*2)

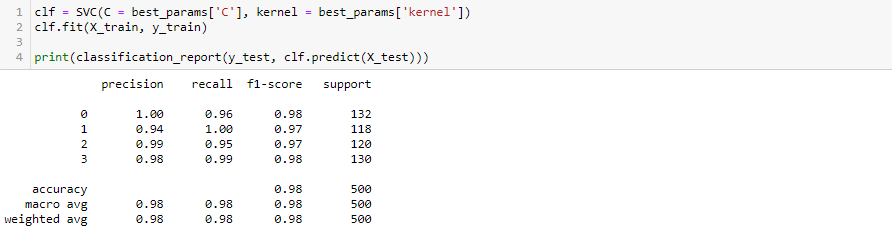
**آیا میتوان در خصوص استفاده از کرنل ها حکم کلی داد؟**

در خصوص  استفاده از آنها میتوان حکم کلی در یک موضوع کلی داد نه در یک موضوع جزیی. بطور مثال می توان گفت در صورت عدم نبود دانش قبلی از داده کرنل آر بی اف می تواند گزینه خوبی باشد. و این ادعا ها برمبنای خواص موجود در توابع این کرنل ها می باشد.

**بررسی سوال 2 و 3 >>**

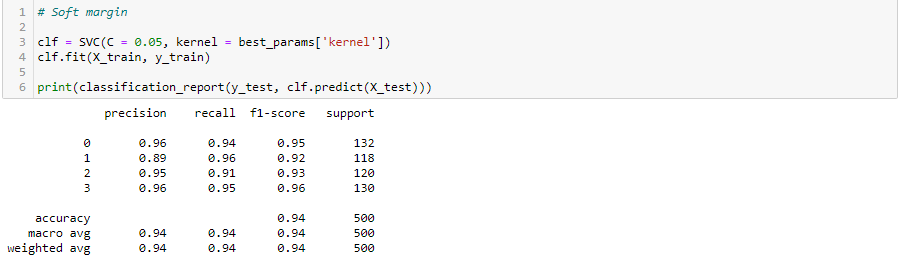
****

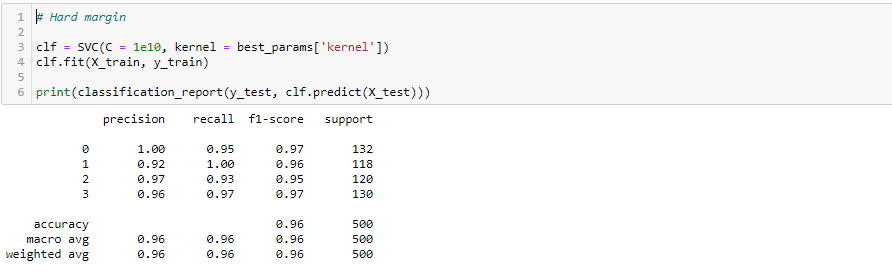




**بررسی سوال 4 >>**

برای حالت حاشیه نرم مقدار متغیر سی را کم و برای حاشیه سخت این مقدار را یک عدد بالا قرار دادم.

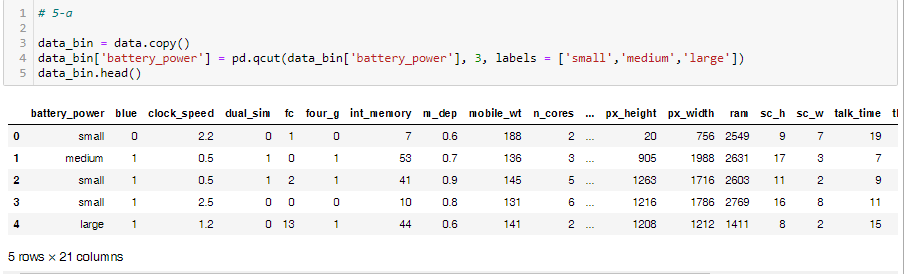
'''

****

**بررسی سوال 5 >>**

**قسمت الف )**

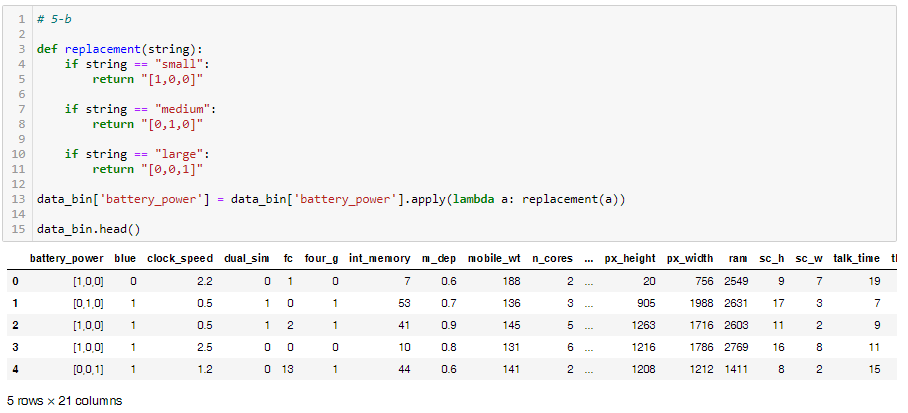
روش binning : داده ها را به سه قسمت نا مساوی تقسیم کردم از فانکشن qcut از پکیج pandas استفاده کردم.

****

**قسمت ب )**

با اعمال one-hot encoding هر کدام از داده های کتگوریکال به یک بردار به طول n که هرکدام از داده های کتگوریکال از داده ها یک و بقییه صفر هست .

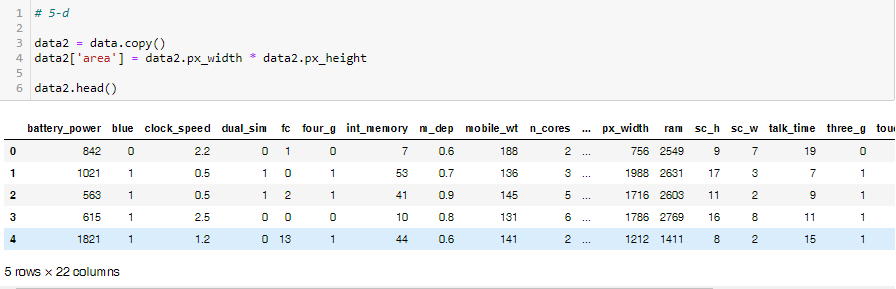
دلیل استفاده از این انکدینگ از آنجاییکه مدل های یادگیری ماشین صرفا با اعداد سرو کار دارند، باید داده هایی که به شکل کتگوریکال هستند را با استتفاده از یک تکنیک انکدینگ به عدد تبدیل کردن. او از آنجاییکه این انکدینگ نمایش قابل فهم تری از یک داده کتگوریکال ارایه می دهد جزو پرکاربردترین انواع انکدر ها می باشد.



**قسمت ج )**

گاهی اوقات داده های موجود از یکدیگر پرت هستند و برخی الگوریتم های یادگیری ماشین مثل ماشین بردار پشتیبان به پراکنده بودن داده ها حساس هستند. استفاده از تبدیلاتی مثل تبدیل لگاریتمی یا مین مکس اسکالر موجب نزدیک شدن سمپل ها به یکدیگر از نظر مقدار می شود این تبدیل پیش از طبقه بندی داده ها در سوال دو مورد استفاده قرار دادم و سوال 6 هم مورد استفاده قراردادم

**قسمت د )**



بررسی سوال 6 >>

