بیتا آذری جو 839598007

**گزارش کار تمرین اول**

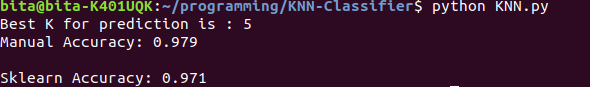
* 1. **تشریح الگوریتم KNN :**

این الگوریتم یکی از الگوریتم های یادگیری با ناظر می باشد. به دلیل سادگی پیاده سازی به شدت مورد استفاده قرار می گیرد. این الگوریتم non-parametric بوده به این معنی که هیچ فرض اولیه ای راجع به داده ها نمی کند. این الگوریتم lazy learning هم می باشد به این معنی که هیچ فاز آموزش(train) ندارد و مستقیما با داده های تست کار می شود.

در این الگوریتم داده ها به صورت رندوم به دو بخش آموزش و تست تقسیم می شود. هر داده برداری از ویژگی های مربوط به آن نمونه دارد. برای هر داده ی تست فاصله ی اقلیدسی آن با سایر نقاط را به دست آورده و به صورت صعودی مرتب می کنیم. سپس در k نمونه ی اول می بینیم جواب نهایی متعلق به کدام کلاس بوده است. هر کلاسی که تعدادش بیشتر بود می گوییم داده ی تست متعلق به آن کلاس خواهد بود.

در تمرین برای مقایسه دقت با کتابخانه ی sklearn و انتخاب فاصله ی minkowski با پارامتر 2 جواب زیر به دست آمد که نشان دهنده ی دقت خوب پیاده سازی می باشد.

پ.ن : در کد اصلی قسمت های مربوط به کتابخانه ی sklearn کامنت شده اند.

****

* 1. **تشریح مفهوم میانگین و واریانس و چولگی :**

میانگین : مقدار مرکزی تعدادی داده ی گسسته می باشد. در واقع جمع مقادیر داده تقسیم بر تعداد کل داده ها. اگر توزیع داده پیوسته باشد. به جای جمع انتگرال می گیریم و امید ریاضی را محاسبه می کنیم.



حالت پیوسته

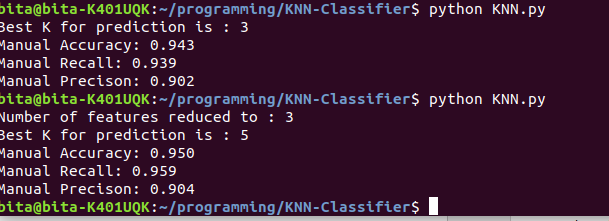
حالت گسسته

به دلیل پراکندگی در مقادیر داده میانگین معیار خوبی برای مقایسه نمی باشد. معمولا از میانه استفاده می کنند.

واریانس : نوعی سنجش پراکندگی می باشد. مقدار واریانس با میانگین‌گیری از مربع فاصله مقدار محتمل یا مشاهده شده با [مقدار موردانتظار](https://fa.wikipedia.org/wiki/%D8%A7%D9%85%DB%8C%D8%AF_%D8%B1%DB%8C%D8%A7%D8%B6%DB%8C) محاسبه می‌شود. در مقایسه با [میانگین](https://fa.wikipedia.org/wiki/%D9%85%DB%8C%D8%A7%D9%86%DA%AF%DB%8C%D9%86) می‌توان گفت که میانگین مکان توزیع را نشان می‌دهد، در حالی که واریانس مقیاسی است که نشان می‌دهد که داده‌ها حول میانگین چگونه پخش شده‌اند. واریانس کمتر بدین معنا است که انتظار می‌رود که اگر نمونه‌ای از توزیع مزبور انتخاب شود مقدار آن به میانگین نزدیک باشد. واحد واریانس مربع واحد کمیت اولیه می‌باشد.



اگر ویژگی ای در داده دارای واریانس بالا باشد اتفاقا خوب است چون نشان دهنده ی تغییرات زیاد می باشد و این تغییرات می تواند در تصمیم گیری نهایی اثرگذار باشد.

هنگام استفاده از PCA، من تعیین کردم که 80 درصد واریانس داده ها حفظ شود. نتیجه این شد که فقط سه مولفه ی اصلی تعیین کننده برای یادگیری پیدا شد و خطا هم خیلی کمتر شد. سرعت بالاتر رفت و دقت مانند حالت بدون استفاده از PCA شد.

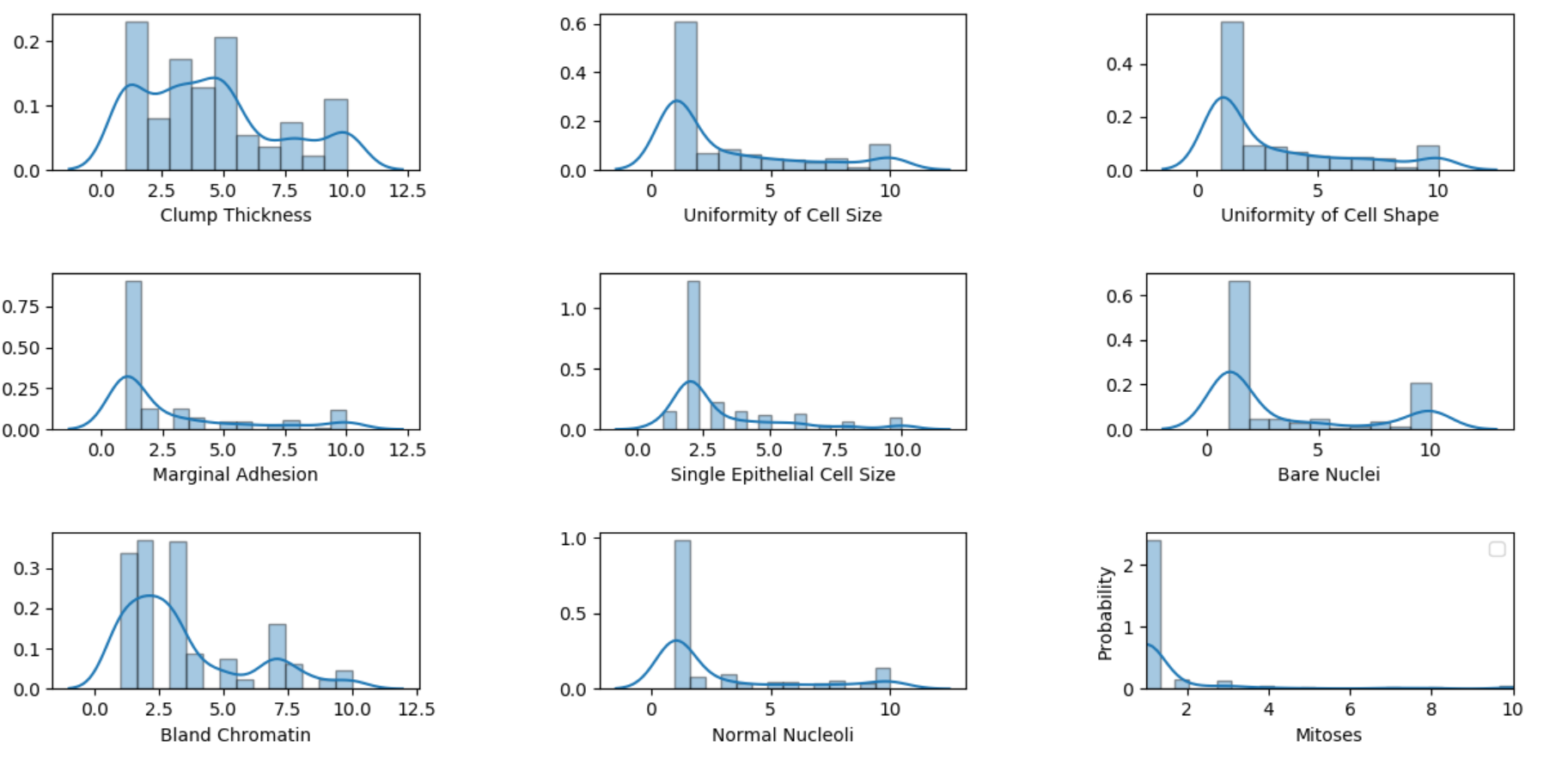
*شکل اول بدون PCA و شکل دوم دقت ها با استفاده از PCA به ازای همان داده ها می باشد*

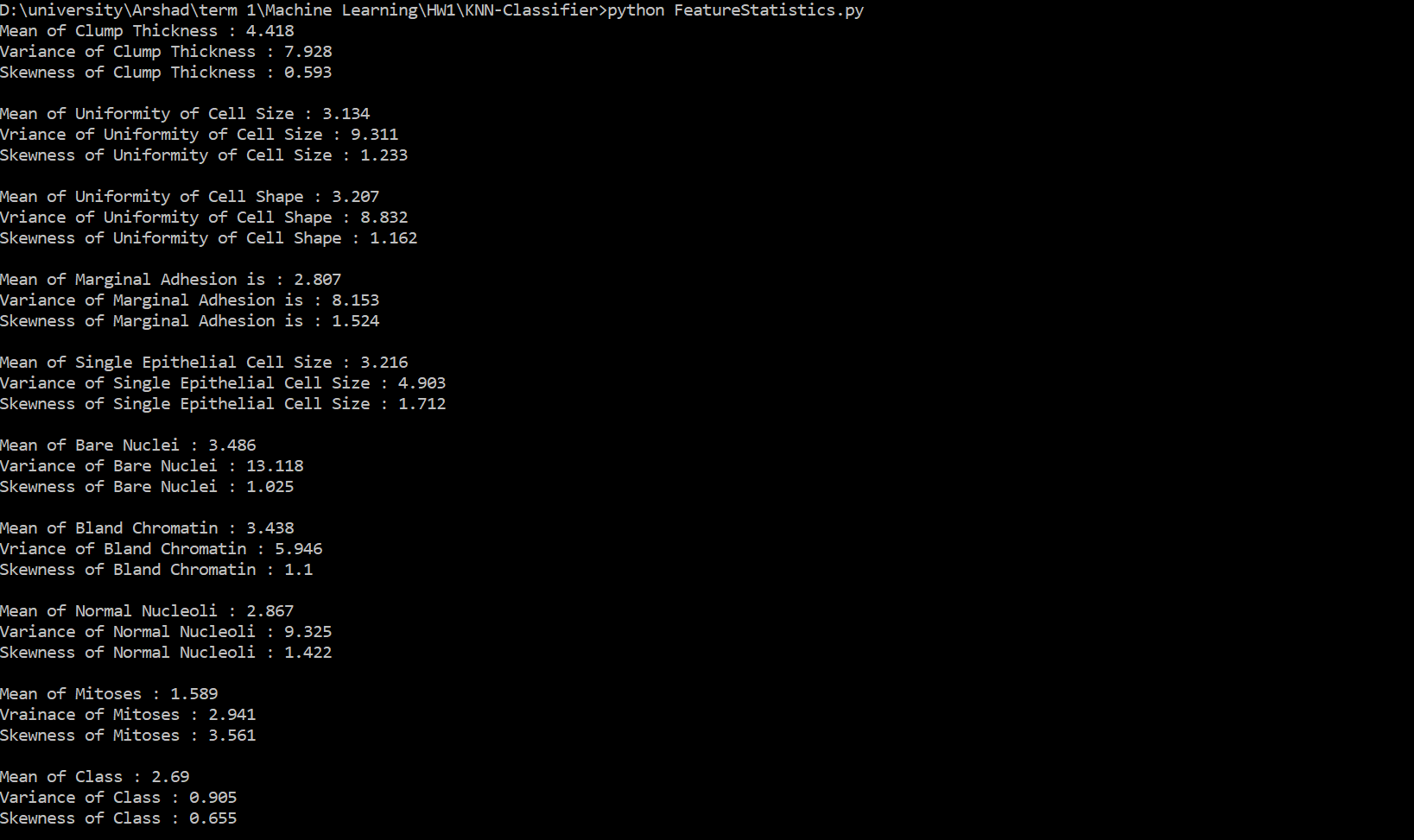
می بینیم که دقت ها حفظ شده اند.

چولگی : نشان دهنده ی میزان عدم تقارن توزیع احتمالی است. اگر داده ها نسبت به میانگین متقارن باشد، چولگی صفر است. در واقع فرمول چولگی گشتاور مرتبه ی سوم به انجراف معیار است.

در داده های موجود پیک داده ها به سمت چپ متمایل است بنابراین چولگی مثبت به دست می آید.

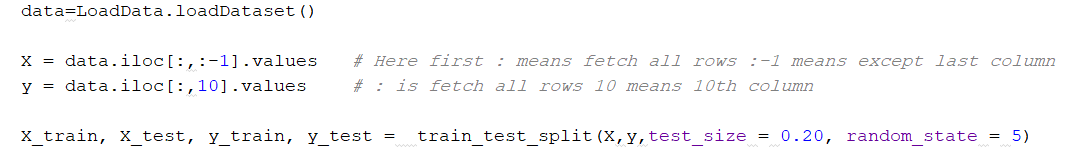
شکل توزیع احتمالی ویژگی ها به شرح زیر است :





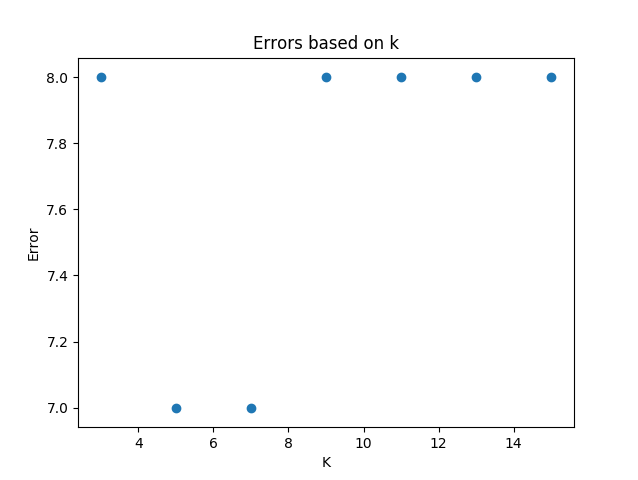
مقدار واریانس بزرگ در داده ها به این معنی است که پراکندگی بیشتری وجود دارد و هنگام استفاده از الگوریتم PCA داده هایی که واریانس بیشتری دارند شانس بیشتری هم برای انتخاب به عنوان principal component دارند.

* 1. **تقسیم داده ها به صورت 80-20 برای داده و تست :**

****

* 1. **انتخاب بهترین k و رسم خطای مدل :**

به ازای random seed=5 و استفاده از PCA در کاهش بعد داده ی آموزش و تست، نتایج زیر به دست آمد:



* 1. **محاسبه ی متریک های precision و recall با استفاده از بهترین K :**

ابتدا به تعریف های confusion matrix می پردازیم که برای محاسبه ی کارایی مدل کلاس بندی به کار می رود. سطر آن واقعیت و ستون مقدار پیش بینی است :

**True Positive :** اینکه فرد سرطان بدخیم داشته باشد و الگوریتم درست تشخیص دهد.

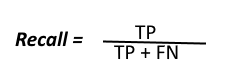
**True Negative :** اینکه فرد سرطان بدخیم داشته باشد و الگوریتم تشخیص ندهد.

**False Positive :** فرد سرطان بدخیم نداشته باشد ولی الگوریتم تشخیص دهد که سرطان بدخیم دارد.

**False Negative :** فرد سرطان بدخیم نداشته باشد و الگوریتم هم سرطان را شناسایی نکند.

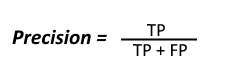
دقت مقدار تشخیص درست سرطان (چه داشته باشد چه نداشته باشد)

به تعداد کل نمونه های تست می باشد.



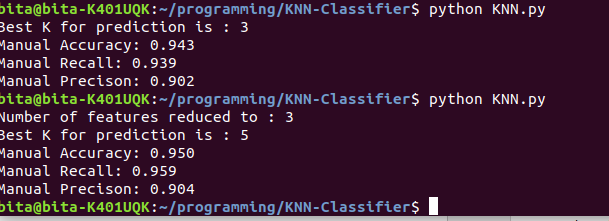
Recall میزان تشخیص درست و مثبت سرطان به کل تعداد نمونه هایی است

که سرطان داشته اند.

Precision میزان تشخیص درست و مثبت سرطان به کل نمونه هایی است که الگوریتم

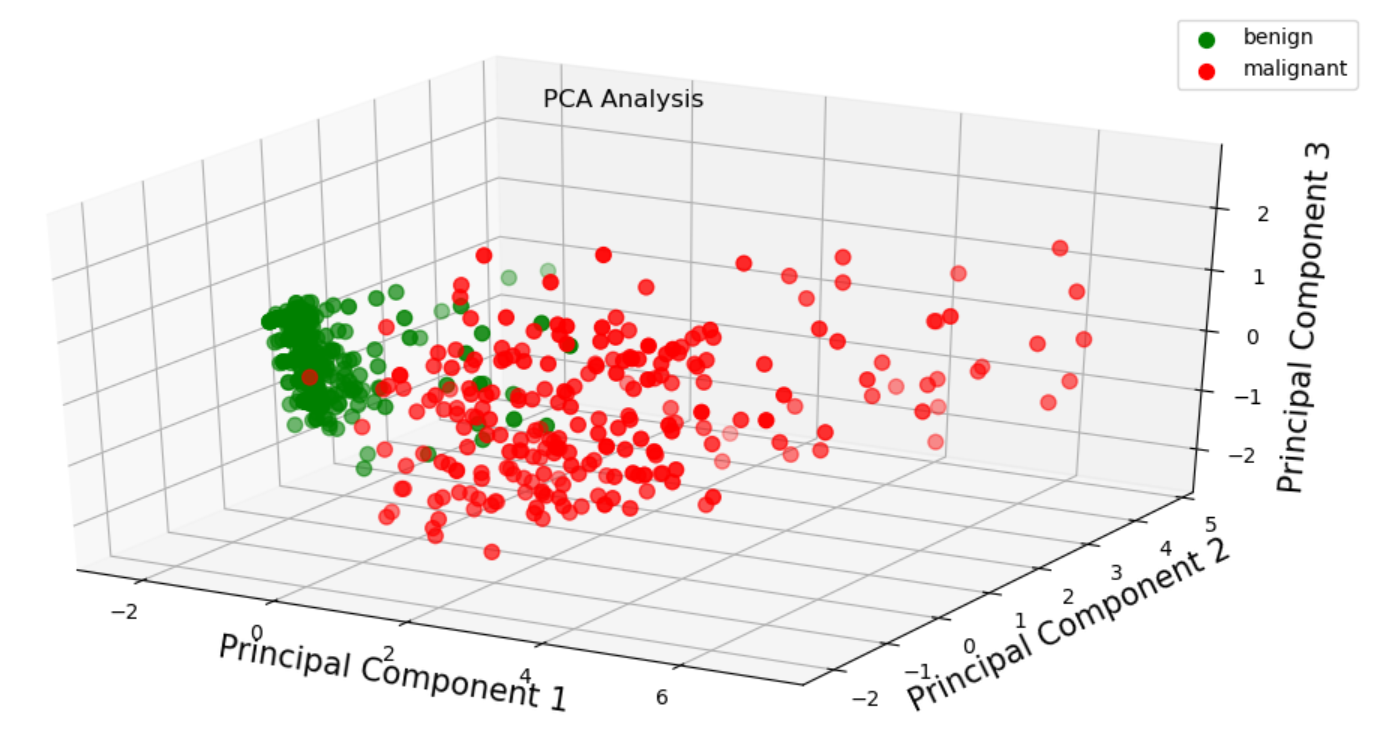
مثبت تشخیص داده.

مقادیر به دست آمده بدون استفاده از PCA و با استفاده از آن :

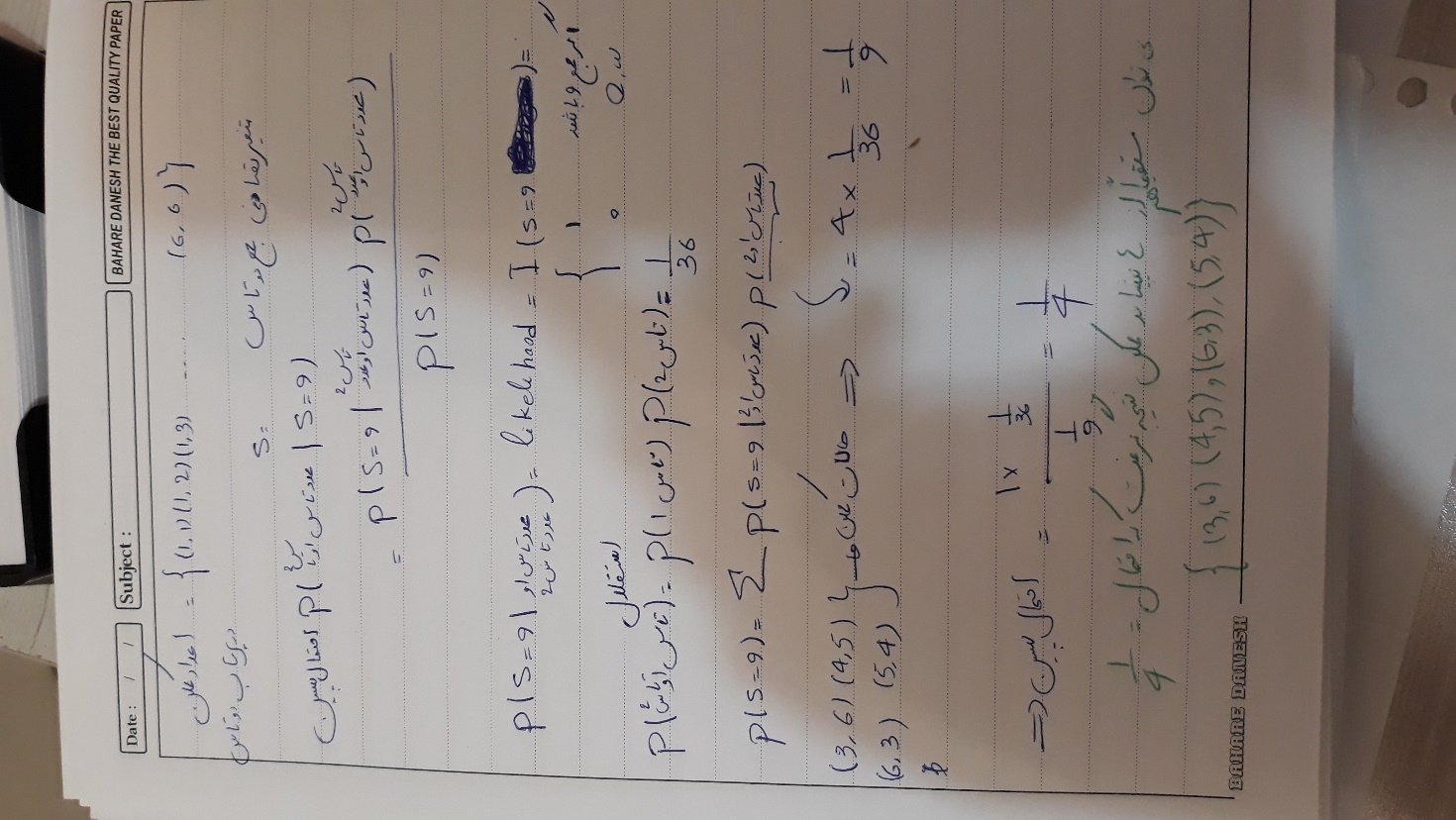


**رسم نمودار PCA:**

ویژگی ها از 9 عدد به 3 عدد کاهش یافت و پیچیدگی محاسباتی کم شد.



**2. محاسبه ی احتمال پسین**

****