به نام خدا

پروژه دوم درس هوش مصنوعی

سوال اول

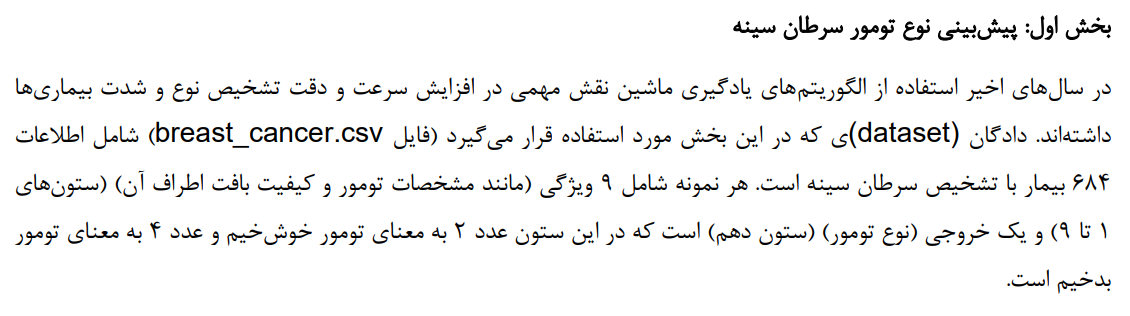
محمد امین سلطانی چم حیدری

810601081

استاد مربوطه:

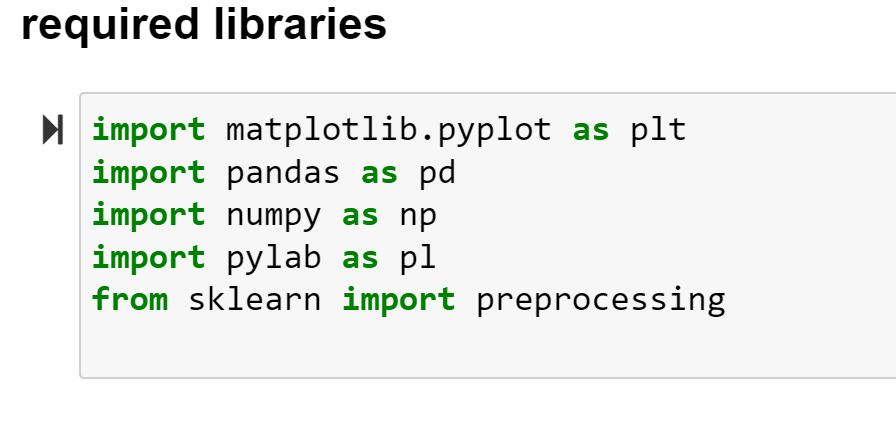
دکتر مسعود شریعت پناهی

بهار1402



در این سوال برای نوشتن کد از jupyter notebook استفاده شده است. همچنین با توجه به نکته ذکر شده در پانوشت سوال از توضیح جزئیات کد خودداری شده است.

برای حل موارد الف تا ه ابتدا کتابخانه های مورد نیاز در پایتون فراخوانی می شوند.



matplotlib 🡨 نمایش data و رسم نمودار

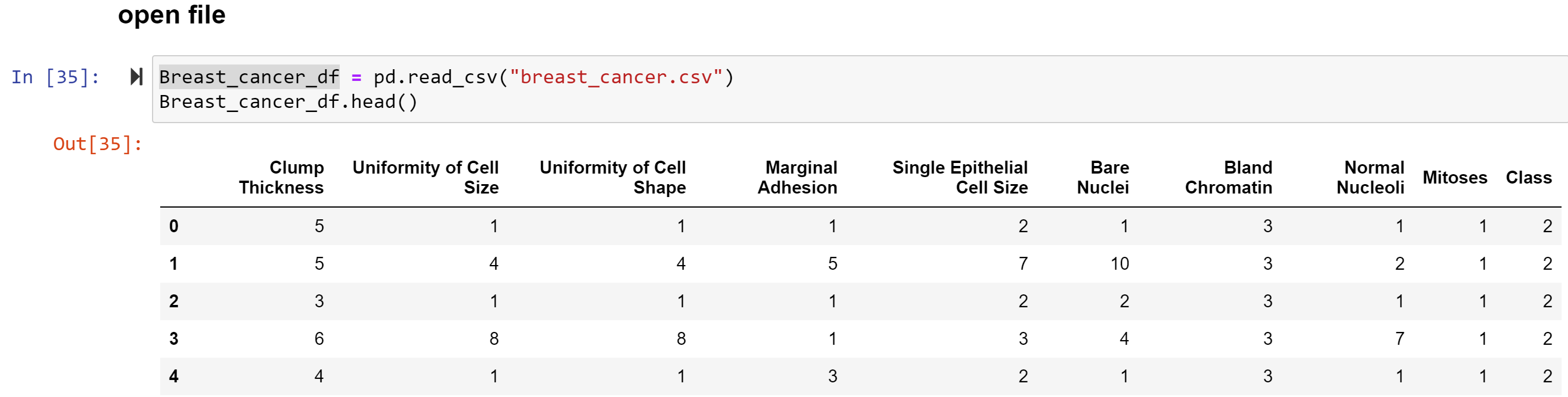
pandas 🡨 کار با داده های ساختار یافته مانند جداول،فایل csv و excel و...

Numpy 🡨 انجام عملیات ریاضی،آماری و عددی

Pylab 🡨 تجسم داده های علمی و عملی.ترکیبی از matplotlib و numpy می باشد

Sickit\_learn 🡨 یادگیری ماشین و تحلیل داده

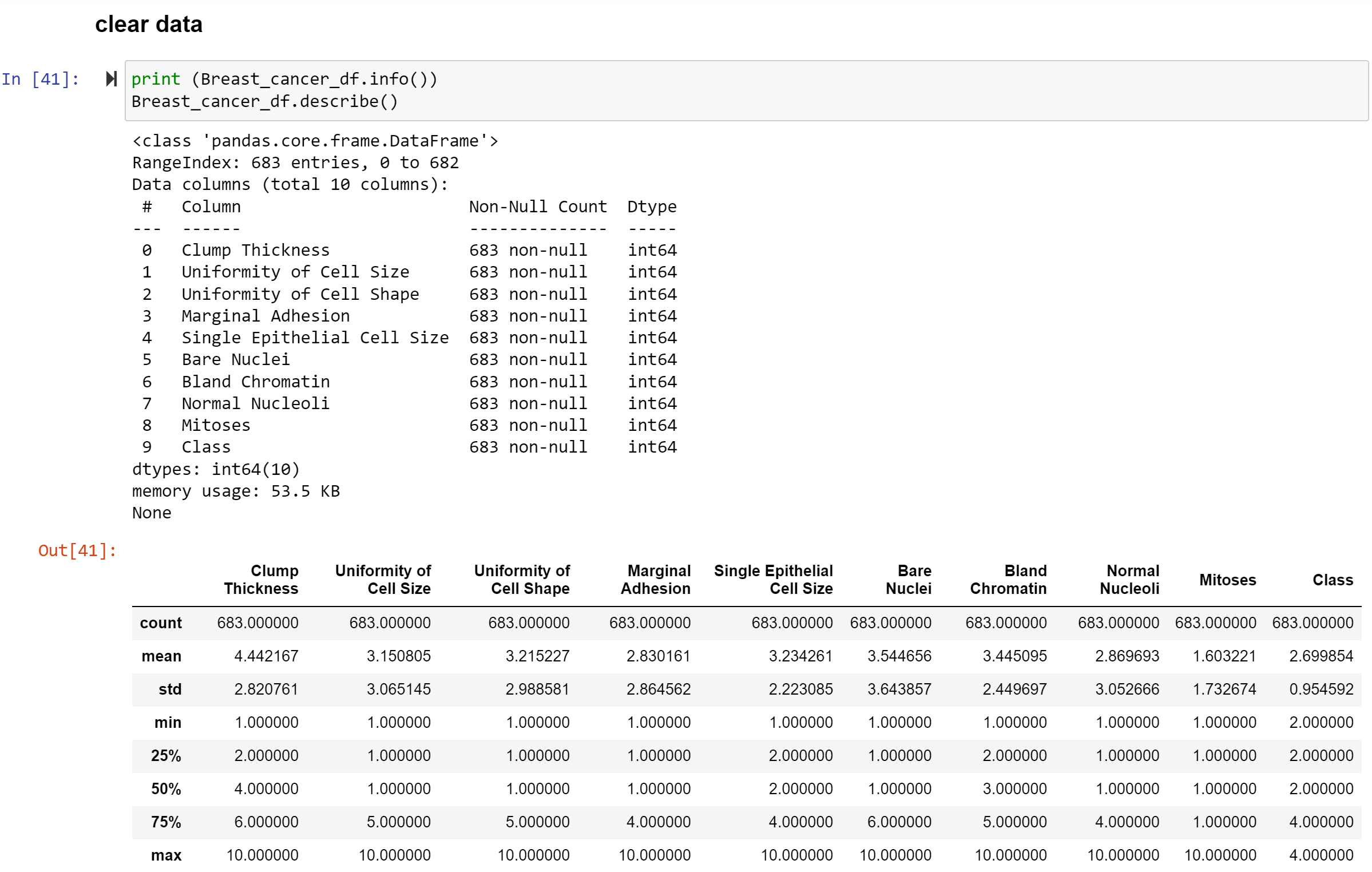
در قدم بعدی فایل excel مورد نیاز شامل اطلاعات بیماران فراخوانی می شود.



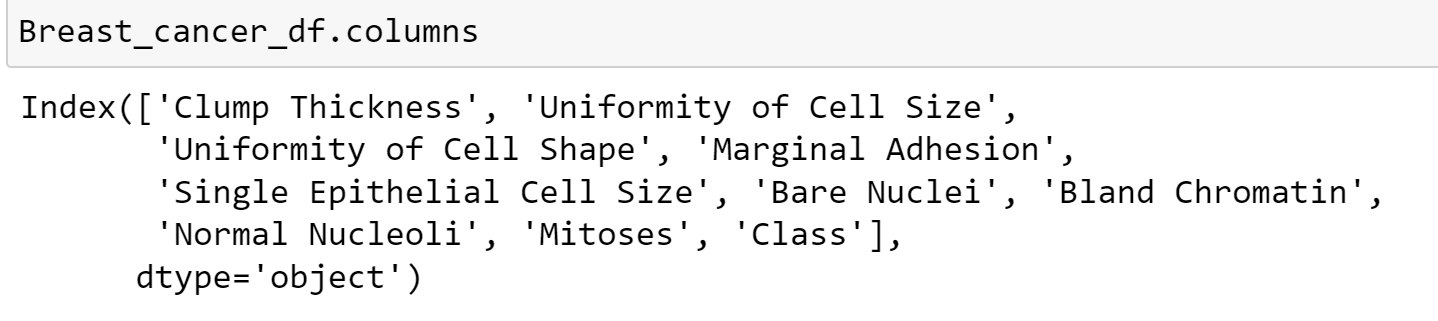
در مرحله بعد برای اثبات عدم وجود نقصی و پرتی در داده ها اطلاعات مربوط به آن ها فراخوانی می شود.از هر دو دستور زیر جداگانه می توان به این نتیجه رسید که در داده ها نقص و پرتی وجود ندارد.

در دستور اول با توجه به اینکه همه داده ها non-null و از نوع int64 هستند می توان نتیجخه گرفت که این دسته از داده قبلا پاک سازی شده اند و مشکلی ندارند.

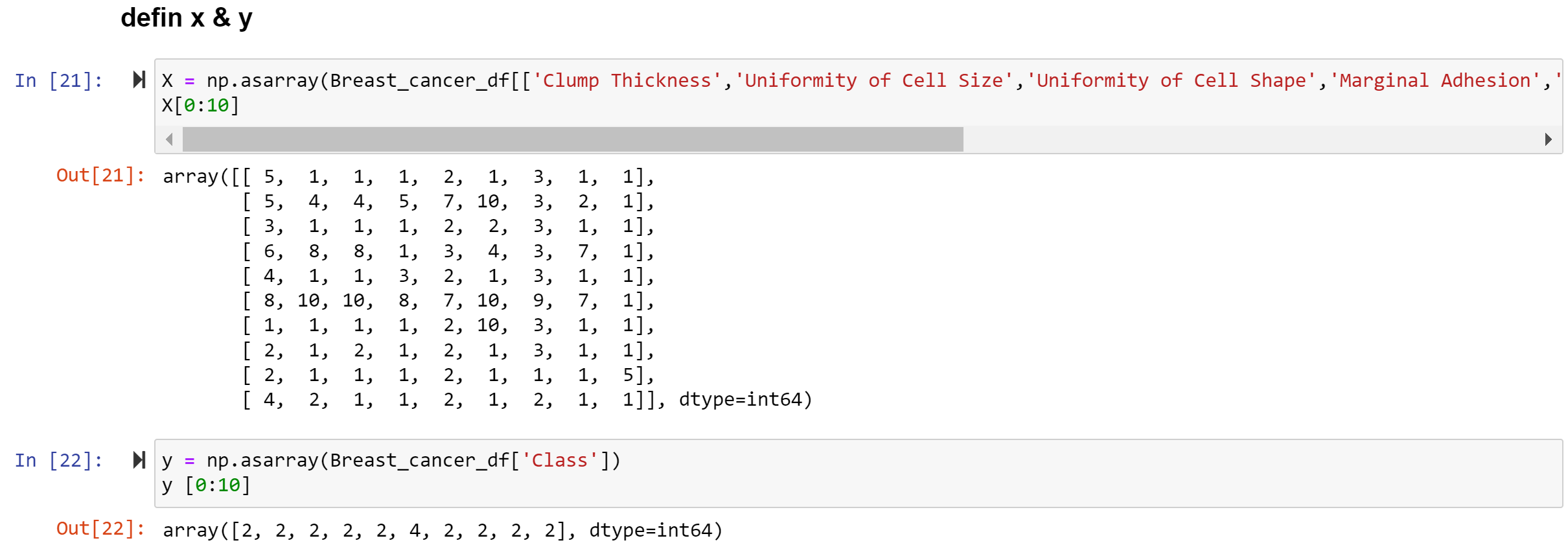
در دستور دوم نیز باتوجه به اینکه در همه ستون های جدول اعداد وجود دارند می توان این نتیجه را گرفت.



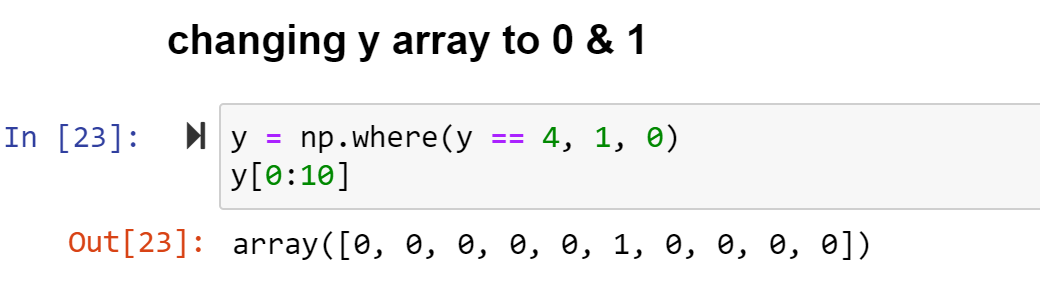
در این قدم نام ستون های جدول استخراج می شود تا در مراحل بعد از آن متغیر ها و خروجی مسئله به سادگی تعریف شود.



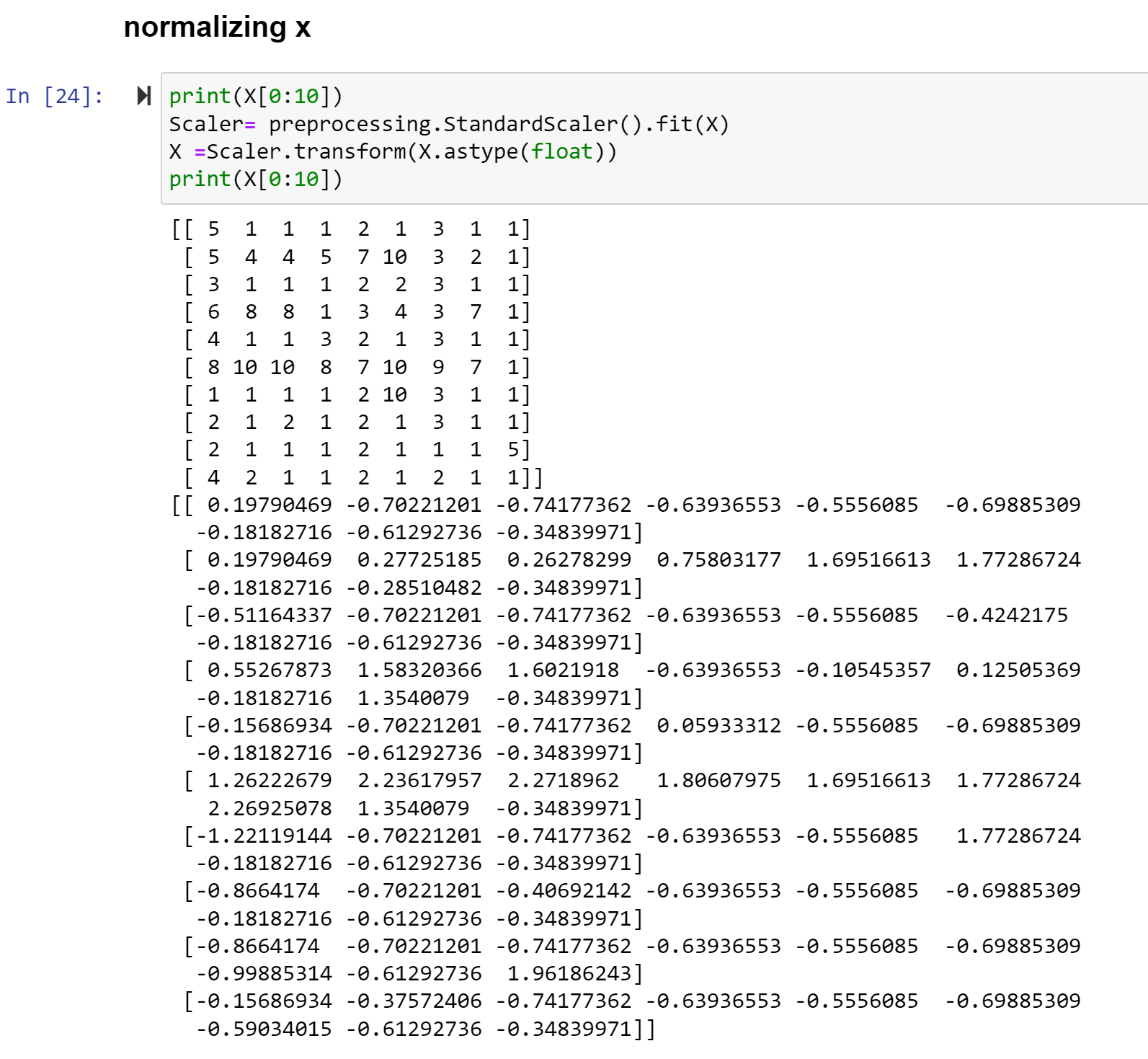
با توجه به اینکه کتابخانه pandas آرایه ها را به صورت str فراخوانی می کند و سوال نیاز به داده های ریاضی دارد در این قسمت برای تعریف آرایه های ورودی و خروجی در متغیرهای x , و y از دستور numpy مطابق شکل زیر استفاده می شود.

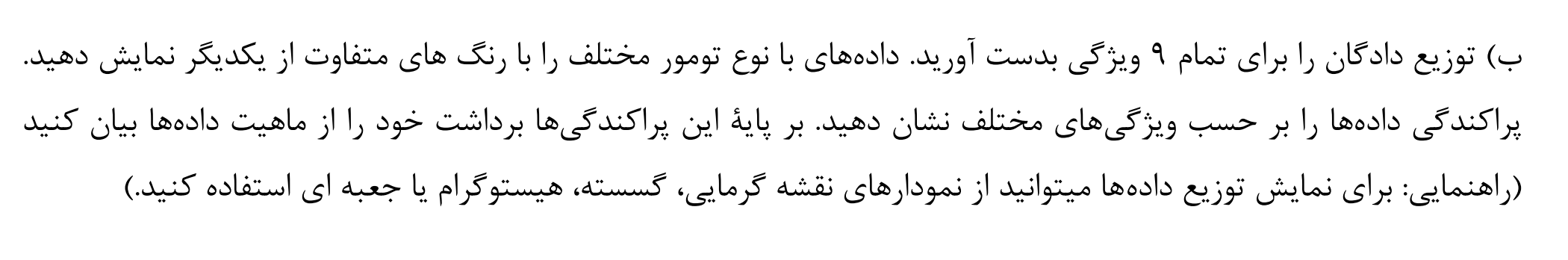


با توجه به خواسته ی قسمت الف داده های ستون class که شامل اعداد 2 و 4 بودند و در متغیر y تعریف شدند، مطابق دستور زیر به آرایه های صفر و یکی تبدیل می شوند.

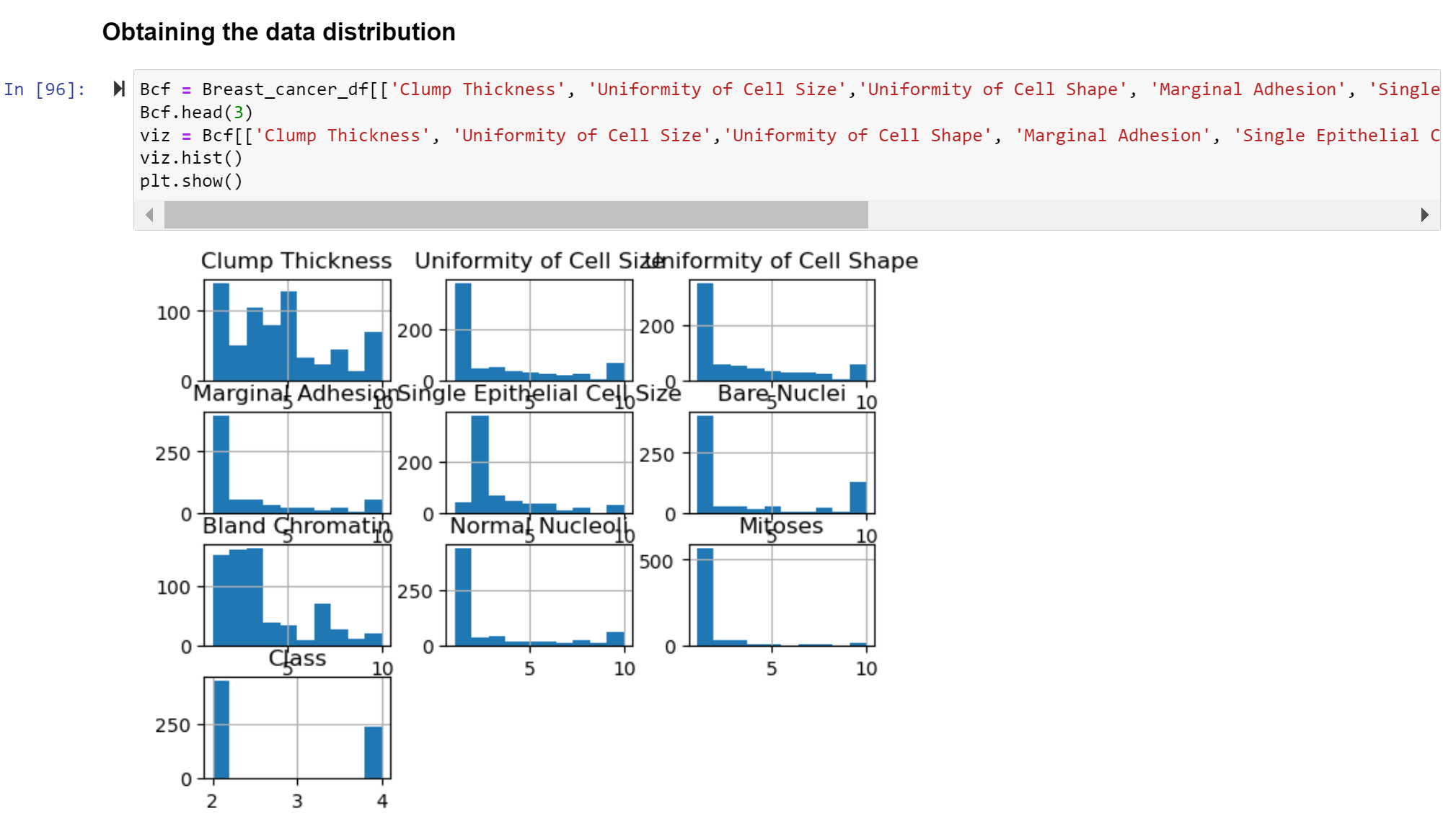


همچنین خواسته ی دیگر قسمت الف normalize کردن یا عادی سازی داده های ورودی یا همان آرایه های متغیر x می باشد که مطابق دستور زیر داده های متغیر x عادی سازی یا همان normalize شده اند.



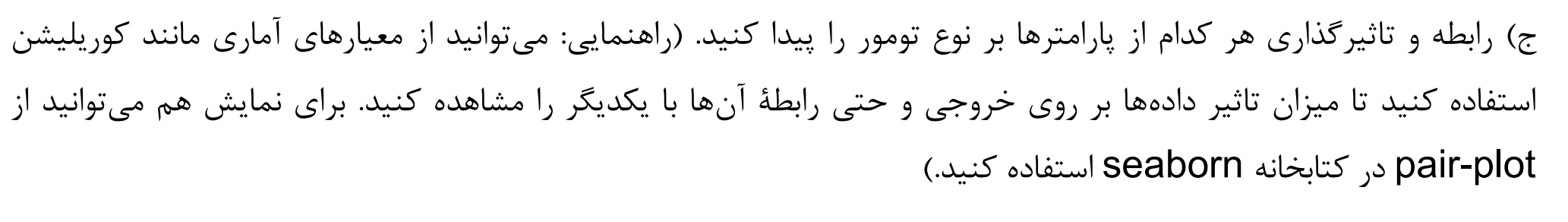


برای بدست آوردن توزیع دادگان برای هر ویژگی ابتدا تابعی شامل ویژگی های مد نظر تعریف کرده و سپس با استفاده از دستور رسم هیستوگرام برای هر ویژگی تابع هیستوگرام رسم می شود.



با توجه به نمودارهای بدست آمده می توان برای هر ویژگی پراکندگی داده ها را از روی شکل بالا مشاهده کرد. مثلا برای ویژگی Clump Thickness پراکندگی داده ها بین 1 تا 10 به طور نسبتا یکنواخت تری از ویژگی Bare Nuclei می باشد. می توان از این موضوع نتیجه گرفت که ویژگی Clump Thickness کمتر از Bare Nuclei مدل ما را دچار خطا میکند.

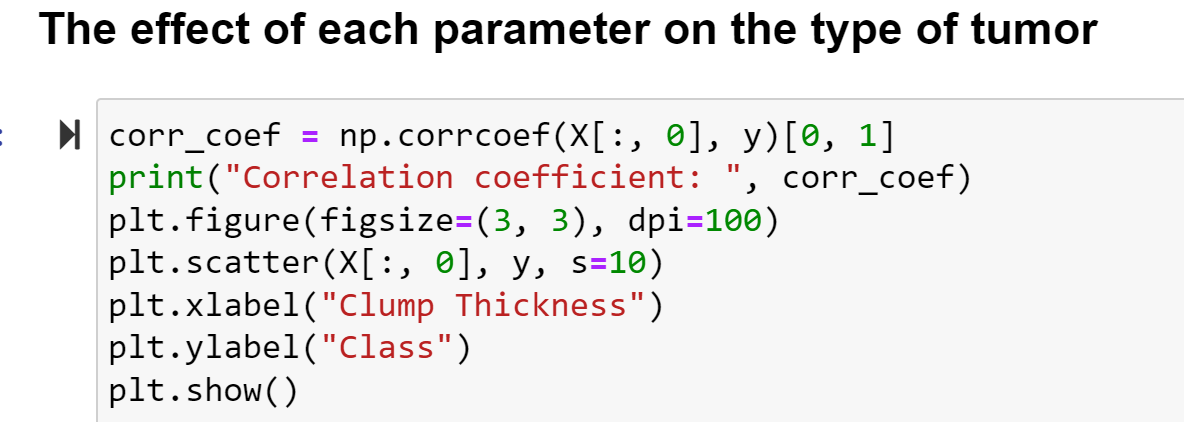
البته داده های مشاهده شد پس از normalize کردن نشان داده شده اند.اگر قبل از آن نمودار ها را رسم می کردیم پراکندگی بیشتر بود و همچنین دامنه تغییرات ویژگی ها نیز با یکدیگر متفاوت می بود.



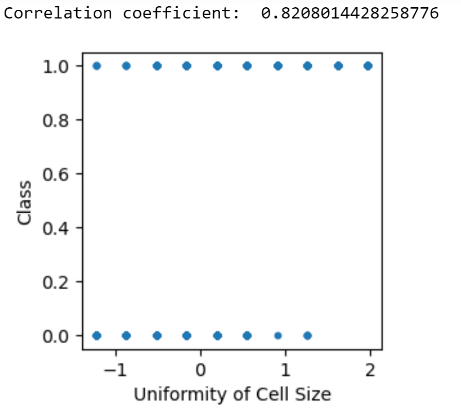
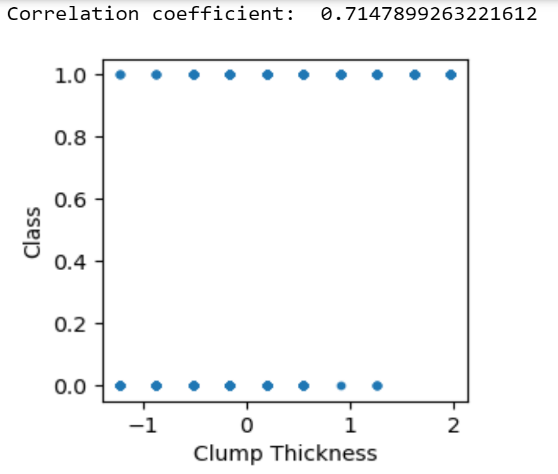
برای محاسبه تاثیر هرپارامتر بر نوع تومور مقدار کورلیشن هر پارامتر بدست آورده شده است.

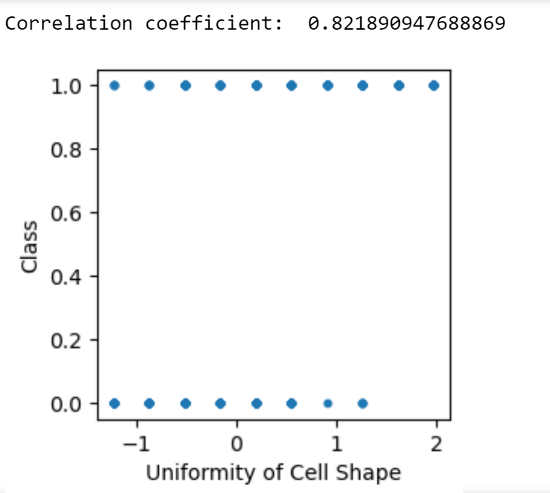
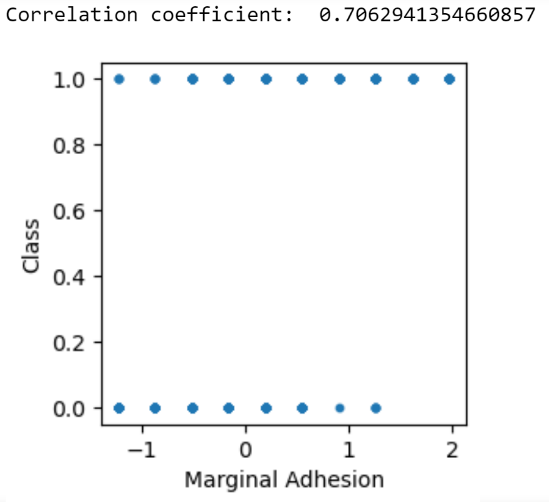
ضریب (correlation coefficient) یا ضریب همبستگی نشان دهنده قدرت و جهت رابطه بین دو متغیر می باشد و مقدار آن بین 1 و-1 است. مقدار یک نشان دهنده رابطه مستقیم کامل بین دو متغیز است.یعنی با افزایش یک متغیر متغیر دیگر نیز افزایش می یابد و مقدار -1 نشان دهنده رابطه معکوس بین دو متغیر است.یعنی با افزایشیک متغیر،متغیر دیگر کاهش می یابد.مقدار صفر نشان دهنده عدم وجود هرگونه رابطه بین دو متغیر می باشد.

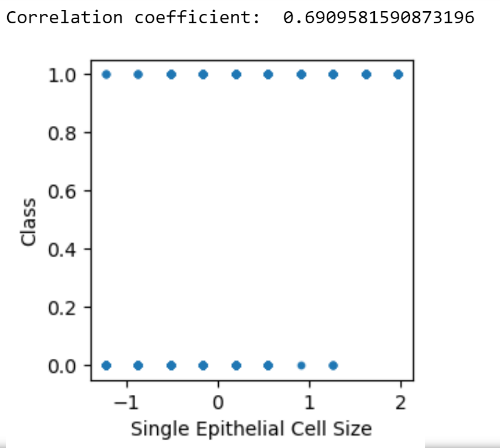
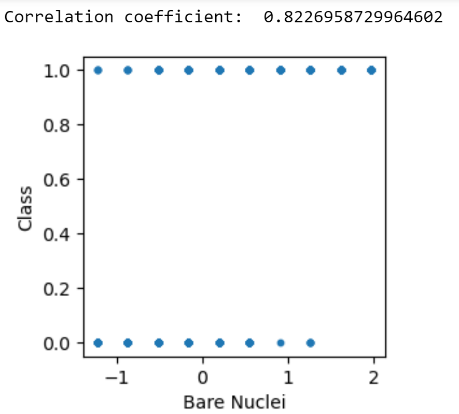
برای محاسبه ضریب همبستگی و نمایش نمودار تومور بر حسب هر متغیر مطابق کد زیر عمل می شود.

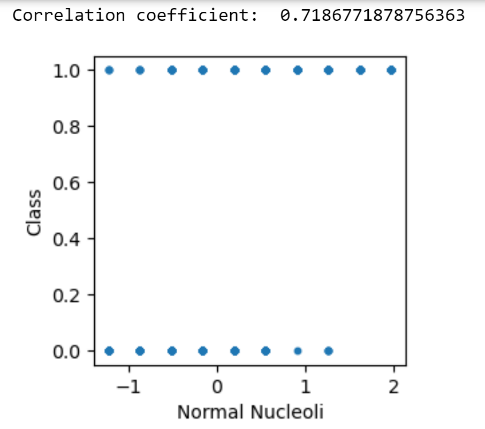
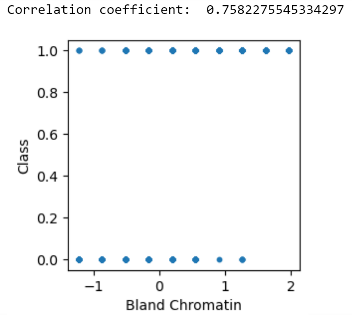


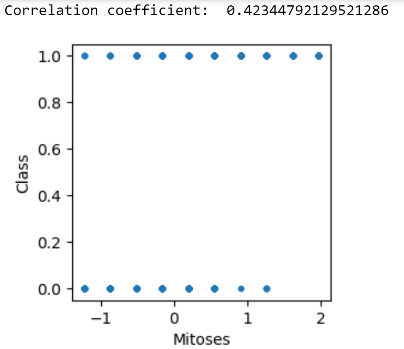
در ادامه نمودار تاثیرگذاری و مقدار correlation coefficient (که برای هر نمودار بالای آن نوشته شده است) برای هر متغیر آورده شده است.







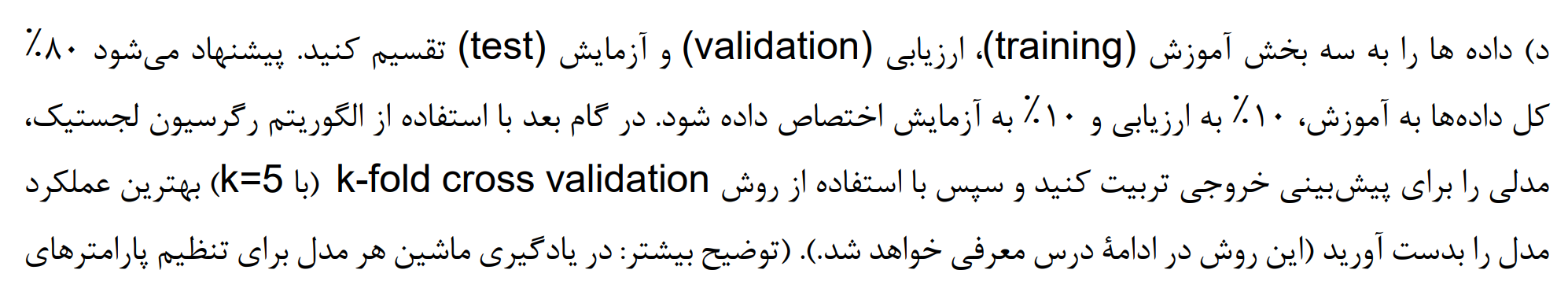


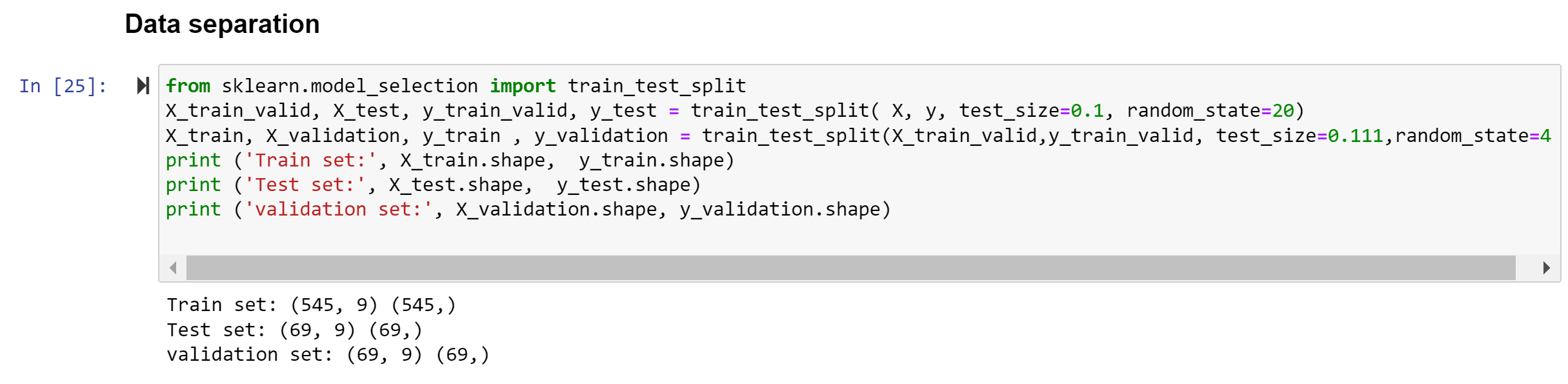


همانگونه که مشخص است ضریب Bare Nuclei باتوجه به اینکه بیشترین correlation coefficient را دارد بیشترین تاثیر مستقیم را بر کلاس تومور می گذارد و ضریب Mitoses با کمترین correlation coefficient در بین متغیرها کمترین تاثیر را در کلاس تومور خواهد داشت.

نکته دیگر اینکه باتوجه به اینکه داده های خروجی بر حسب صفر و یک منظم شده و متغیر ها نیز normalize شدند لا نمودارها به خوبی قابل مقایسه هستند.

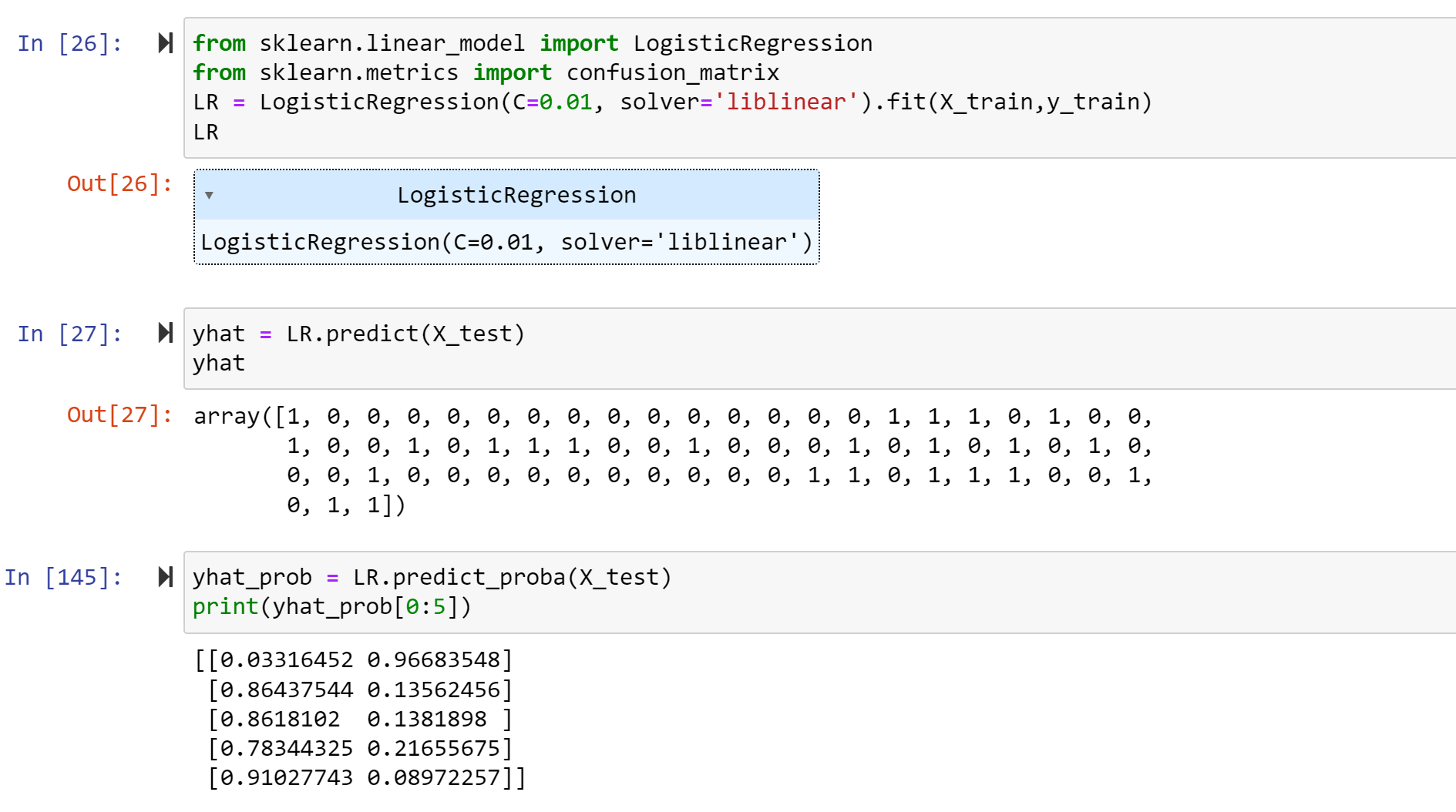
همچنین با توجه به اینکه تعداد data ما برای سوال بیشتر از نقاط مشخص شده در شکل بود می توان نتیجه گرفت نقطه ها روی هم افتاده اند و هر نقطه بیان گر چندنقطه ی روی هم می باشد.





مطابق کد بالا و با استفاده از دستور Train-Test-split از کتابخانه scikit-learn داده های سوال شامل ورودی و خروجی ها به سه بخش تقسیم می شوند.مطابق خواسته سوال 80 درصد داده ها را برای تمرین و تعریف مدل،10 درصد برای تست و 10 درصد نیز برای ارزیابی در نظر گرفته می شود.

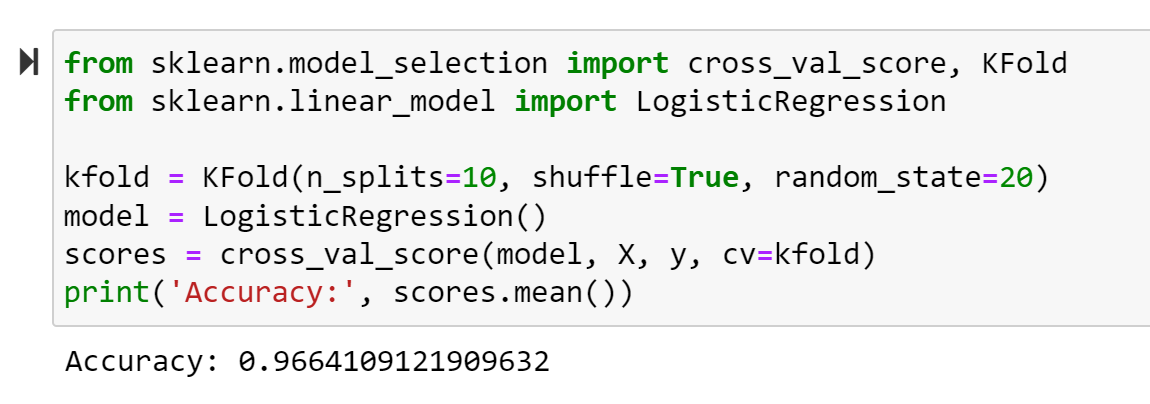
همچنین برای اینکه نتایج ذکر شده در فایل ارائه با نتایج فایل کد تفاوت نداشته باشد در اینجا حالت random state ثابت در نظر گرفته شده تا نتایج باهم همخوانی داشته باشند. در صورت تمایل برای تصادفی در نظر گرفته شدن داده ها در هر مرحله می توان حالت random state را پاک کرد.



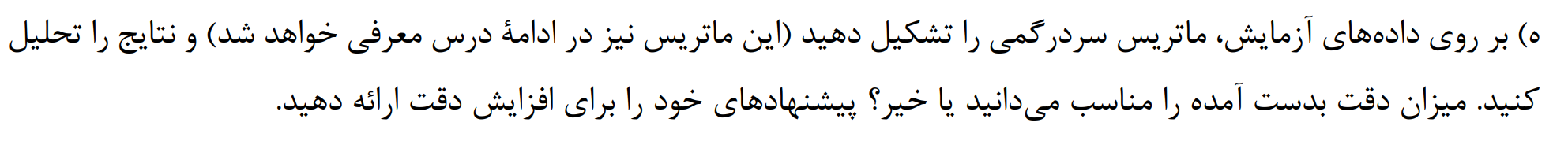
مطابق خواسته صورت سوال با استفاده است رگرسیون لجستیک مدلی برای پیش بینی خروجی تعریف می شود.

در تابع y hat prob به محاسبه احتمال خروجی توسط رگرسیون لجستیکی پرداخته شده.اگر مقدار این تابع بزرگتر از نیم بود برابر یک و اگر کوچک تر از نیم بود برابر صفر قرار داده می شود.

در شکل بالا تابع y hat پیش بینی از حالت خروجی به صورت صفر و یک است. باتوجه به ماهیت رگرسیون لجستیکی که در دسته بنده classification قرار میگیرد و برای نتایج گسسته مورد استفاده قرار می گیرد در اینجا نیز خروجی سوال با توجه به مدل تربیت شده برابر صفر و یک قرار داده شده است.



در پایان نیز صحت مدل تربیت شده با استفاده از روش k fold cross validation محاسبه شده که برابر 0.966 می باشد.

در این قسمت تابع سردرگمی تعریف شده و نتایج به صورت قابل مشاهده در آمده است.

در این ماتریس عدد 21 بیانگر TP می باشد.یعنی 21 مورد از داده های تست دارای تومورخوش خیم بوده اند و مدل آن را درست پیش بینی کرده است.

عدد 44 TN می باشد.یعنی 44 نفر توموربدخیم داشته اند که مدل به درستی آن را پیش بینی کرده است

عدد 3 بیانگر FP می باشد.یعنی سه مورد را مدل دارای توموربد خیم پیش بینی کرده در صورتی که تومورآنها خوش خیم بوده است.

عدد 1 نیز معرف FN می باشد.یعنی یک نفر دارای تومور بدخیم بوده است درحالی که مدل آن را خوش خیم پیش بینی کرده است.

در موارد حساس مانند این سوال اهمیت FN برای ما از همه بیشتر است زیرا درصورت که تومور بدخیم،خوش خیم پیش بینی شود خطر جانی دارد.لذا میزان دقت ما در این مورد نیاز به افزایش دارد.

برای این مورد میتوان داده های تست را افزایش داد.

همچنین می توان مدل های دیگر را برای پیش بینی امتحان کرد و بهترین مدل از لحاظ دقت را استفاده کرد.مثلا اگر از SVM برای تربیت این مدل استفاده شود در پایان می توان با به کار بردن Regularization دقت را افزایش داد.

و.....