# تمرین سری دوم درس یادگیری ماشین

|  |
| --- |
| در محیط برنامه نویسی پایتون برنامه‌ای بنویسید که از دیتاست data.csv که حاوی اطلاعات درباره بیماری سرطان سینه می‌باشد استفاده کند و داده‌ها را طبقه بندی کرده و نتیجه accuracy را در حالت‌های مختلف زیر گزارش کنید. به عنوان پیش پردازش از روش standardization استفاده کند.  الف) استفاده از الگوریتم KNN و در نظر گرفتن 70 درصد داده‌ها به عنوان داده آموزشی و 30 درصد داده‌ها به عنوان داده تست.  ب) استفاده از الگوریتم KNN و درنظرگرفتن روش K-fold (K=10)  ج) استفاده از الگوریتم SVM و درنظرگرفتن 70 درصد داده‌ها به عنوان داده آموزشی و 30 درصد داده‌ها به عنوان داده تست د) استفاده از الگوریتم SVM و درنظر گرفتن روش K-fold (K=10) |

## ایمپورت کردن کتابخانه‌های مورد نیاز

در قدم اول کتابخانه‌های اصلی که برای این تمرین به آن‌ها نیاز داشتم (numpy, sklearn و pandas) را import کردم.

## لود کردن و خواندن دیتاست

در مرحله دوم، فایل اکسل دیتاست را فراخوانی کردم و آن را با دستور pd.read\_csv لود کردم. سپس اطلاعات ستون لیبل‌ها را در L و اطلاعات سطرها و ستون‌های دیگر را در X ریختم.

## قسمت الف تمرین و استفاده از مدل KNN

در این مرحله ابتدا دستور train\_test\_split را import کردم سپس با استفاده از آن داده‌ها را به صورت 70 درصد به عنوان داده آموزشی و 30 درصد داده تست تقسیم کردم. از دستوری StandardScaler() تحت عنوان sc برای پیش پردازش داده‌ها استفاده کردم.

سپس با دستور from sklearn.neighbors import KNeighborsClassifier مدل KNN را فراخوانی کردم و آن را تحت عنوان Model در کدم استفاده کردم. با دستور Model.fit مدلم را آموزش دادم سپس با دستور Model.predict آن را تست کردم.

در آخر با دستور Model.score مقدار Accuracy را حساب کردم. این مقدار برای قسمت الف تمرین 0.9707602339181286بدست آمد.

## قسمت ب تمرین استفاده از KNN و K-fold

ابتدا با دستور from sklearn.model\_selection import KFold K-fold را فراخوانی کردم و آن را در ادامه kf نام گذاری کردم. مقدار k را 10 تعیین کردم. سپس یک لیست خالی تعریف کردم تا مقادیر Accuray را در آن ذخیره کنم.

در قدم بعدی یک حلقه for نوشتم تا برای هر fold مراحل train و test را انجام بدهد و مقادیر Accuracy را در لیست بالا ذخیره کند. در آخر مقادیر Accuracy , Loss تمامی fold ها را پرینت کردم اما میانگین آن‌ها را هم حساب کردم. میانگین این مقادیر برابر با 0.9665726817042606 شد.

## قسمت ج تمرین استفاده از الگوریتم SVM

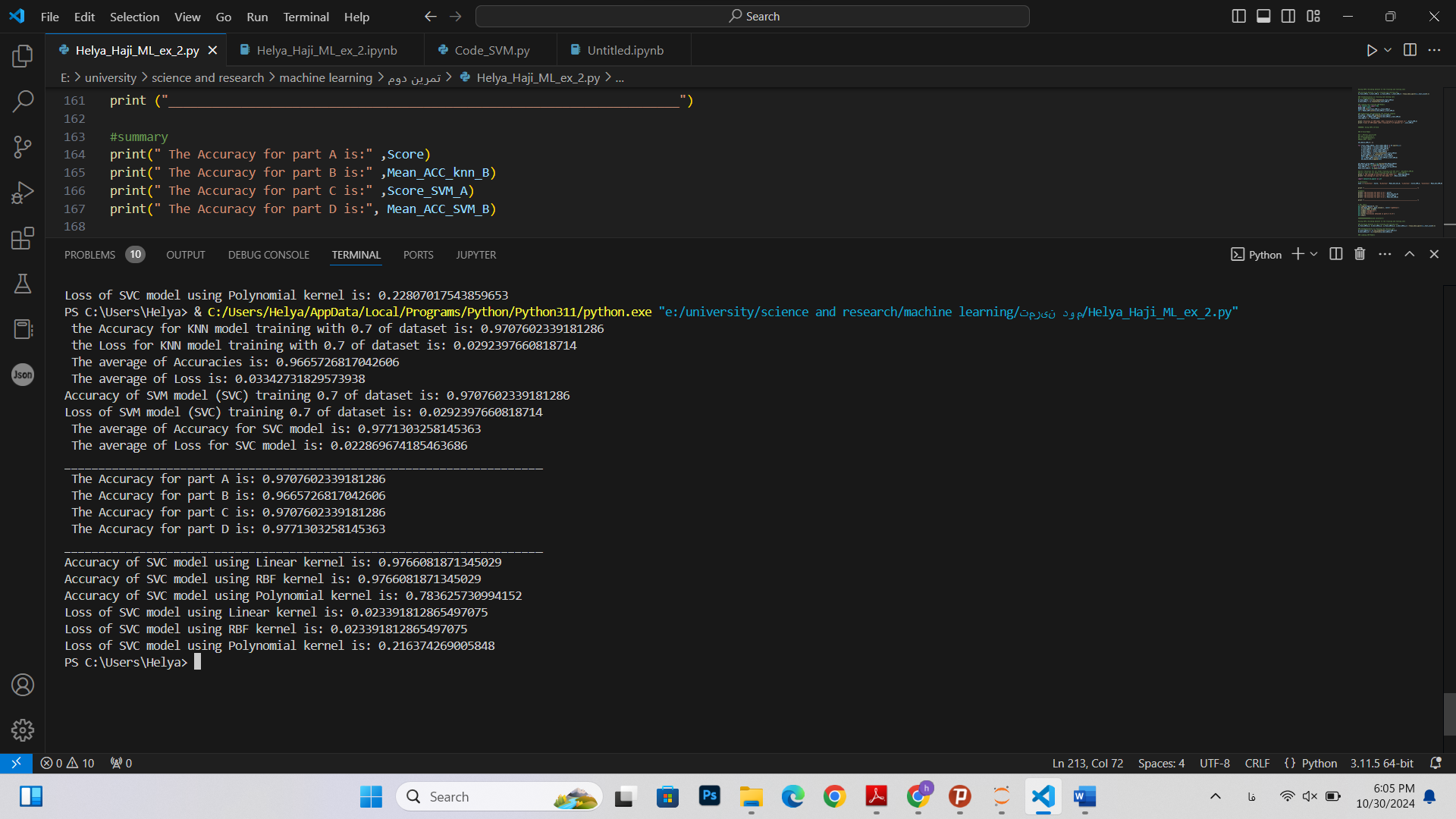
در این مرحله هم مانند قسمت الف، داده ها را به دو قسمت train و test تقسیم کردم. سپس با دستور from sklearn.svm import SVC مدل SVC (همان SVM) را فراخوانی کردم و مدلم را با دستور fit آموزش دادم. در آخر با دستور Model.score Accuracy مدلم را بررسی کردم که مقدار آن برابر با 0.9707602339181286 است.

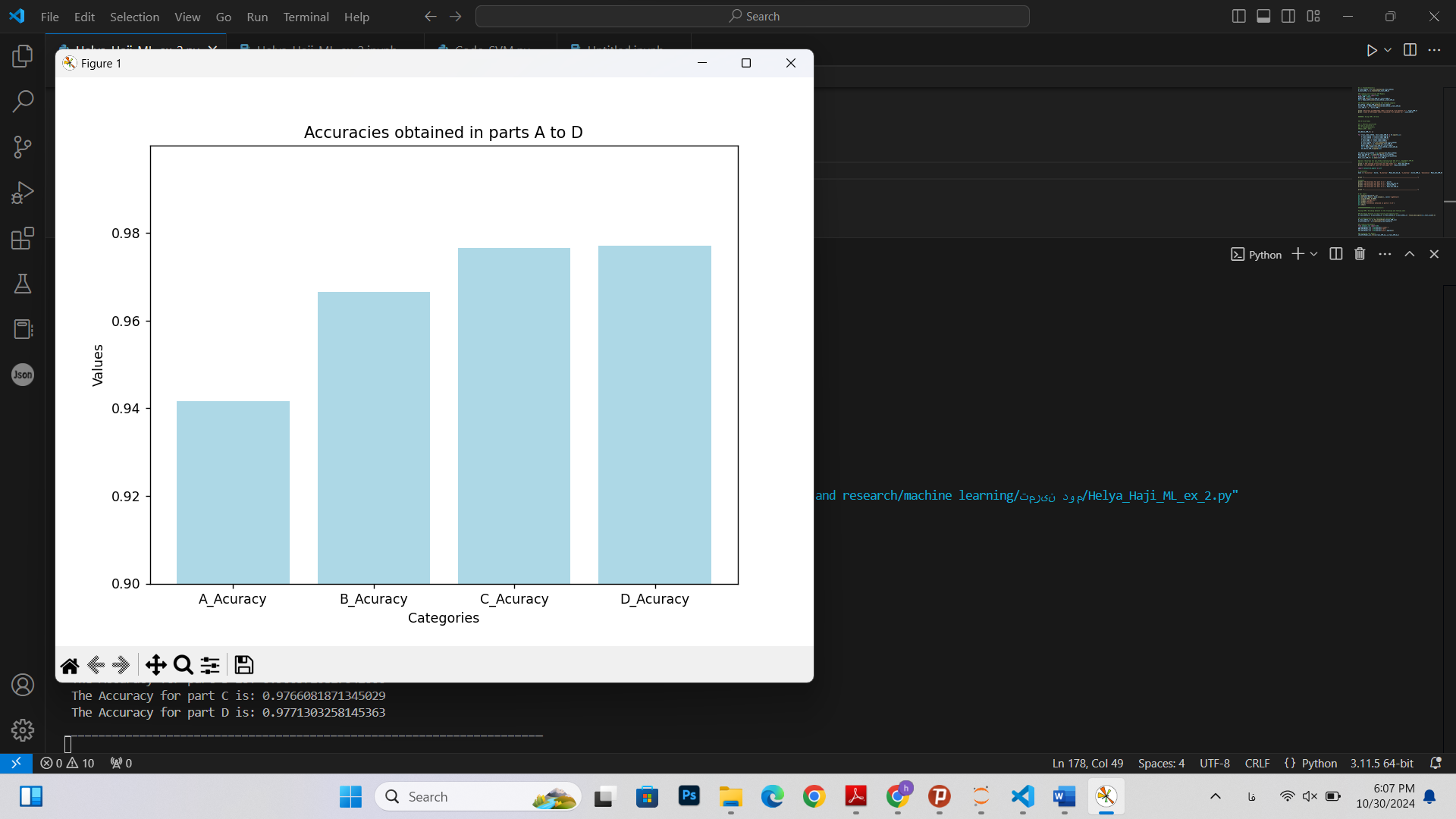
## قسمت د تمرین استفاده از SVM و K-fold

از k-fold استفاده کردم و مقدار k را برابر با 10 قرار دادم. سپس یک حلقه نوشتم تا Accuray را برای هر fold محاسبه کند. در آخر میانگین Accuray را با دستور np.mean حساب کردم که برابر با 0.9771303258145363 شد.

## نتیجه گیری و خلاصه کردن نتایج

در آخر مقدار Accuray بدست آمده در چهار بخش الف تا د را حساب کردم و با یک نمودار نتایج را نشان دادم. به طور کلی با هر بار ران کردن کدها عدد Accuracy تغغیر کرد اما به طور کلی بیشترین دقت برای قسمت ج و قسمت د مسئله بدست آمد. پس می‌توانیم این طور نتیجه گیری کنیم که دقت مدل SVM از مدل KNN بیشتر است.





## راه حل دوم برای قسمت ج

برای قسمت‌های ج و سوال راه حل دیگری هم به ذهنم رسید. در روش دوم من از هر سه کرنل SVM برای آموزش دادن مدل استفاده کردم. در این حالت دقت مدل SVM با کرنل Linear از سایر کرنل‌ها بالاتر بود.

