هدف این تمرین آشنایی شما با دسته‌بندی KNN و Naive bayes و موارد مرتبط با آن است. به این منظور لازم است شما با مجموعه داده breast_cancer کار کنید. این مجموعه داده با تابع datasets.load_breast_cancer() قابل استفاده خواهد بود. این مجموعه داده شامل داده‌های بیماران مبتلا به سرطان سینه است و داده‌ها در دو دسته «خوش خیم» و «بدخیم» دسته‌بندی شده‌اند.

در این تکلیف باید به موارد زیر پاسخ دهید.

۱- بالاترین میزان صحت (accuracy) که هر یک از این دو دسته‌بند دارند چه میزان است. روش ارزیابی 10-fold cross validation است. برای محاسبه صحت از تابع score (مطابق با موارد بحث شده در کلاس) استفاده کنید. لازم بر روی نحوه مقداردهی هر پارامتر در حل مسئله بحث کنید.
۲- به نظر شما کدام یک از ویژگی‌های این مجموعه داده در نتایج تاثیر بیشتری دارند؟ در هر دو دسته‌بند بررسی کنید.
۳- معیارهای precision و recall و F1 و accuracy را برای نتایج نهایی مورد ۱ محاسبه کنید. به نظر شما کدام یک از این ۴ معیار برای این مسئله اهمیت بیشتری دارد؟ حال تلاش کنید بالاترین نتایج را در معیار انتخابی (با تغییر پارامترها) بدست آورید.

نکات:

۱- در خصوص پیاده سازی و پارامترهای Naive bayes می‌توانید با یک جستجوی ساده مطالب مفیدی پیدا کنید.
۲- خروجی کار را در قالب یک فایل Jupiter تحویل دهید که جواب سوالات کدها و نتایج اجرا در آن آمده است.

موفق باشید.