به نام خدا

تکلیف پنجم درس مبانی داده کاوی

ترم بهار ۱۴۰۰

راهنمایی :
زبان برنامه نویسی سوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج‌های اصلی استفاده شده seaborn،numpy, pandas, sklearn می باشند.
دیتاست های مورد نیاز در ادامه معرفی شده اند.
روش تحویل:
الف) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Qx.zip که x شماره سوال است زیپ شوند (برای نمایش خروجی دستورات، هرجا مقدور است نام دیتافریم را بزنید تا خلاصه آن را نشان دهد و در سایر حالات از دستور head استفاده کنید)، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW6-Lastname-StudentCode.zip که Lastname نام خانوادگی و StudentCode شماره دانشجویی شما است، زیپ شده و روی سامانه تا **زمان مشخص شده** آپلود شوند.
ب) گزارش نهایی باید شامل پاسخ تمامی سوالات (سوالات تحقیقی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی درمورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد.
ج) زمان و نحوه تحویل تکلیف در فایل راهنمای ترم مشخص شده است.
د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

نکته ۱ : برای پاسخ به سوالات **تحقیقی** و **تفسیری**، پس از مطالعه منابع موردنیاز فقط **برداشت خود** از مسئله را توضیح دهید.
نکته۲ : برای حل سوال چهارم از کتابخانه MLxtend استفاده نمایید.

۱. رگرسیون خطی (Linear Regression)

a) فایل csv دیتاست Camera را خوانده و تبدیل به دیتافریم نمایید. (مقادیر null در این دیتاست با ۰ مشخص شده‌اند)

b) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین اینکد کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

c) درصورت نیاز مقادیر داده‌های پرت را نیز بررسی کرده و در صورت صلاح دید خود حذف نمایید.

d) نمودار pairplot این دیتاست را رسم نموده و وابستگی‌های بین ستون‌های این دیتاست را توضیح دهید.

e) مقادیر همه ستون ها به جز ستون Price را در متغیر x قرار داده و ستون Price را در متغیر y قرار دهید.

f) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید.

g) با استفاده از کلاس LinearRegression مدل رگرسیون موردنظر خود را ایجاد نمایید.

h) مقادیر "_intercept" و "_coef" از مدل ایجاد شده را تحلیل نمایید.

۲. رگرسیون خطی (Linear Regression)

a) دیتاست boston را از کتابخانه sklearn لود کرده و مقادیر موجود در دیکشنری این دیتاست را بررسی نمایید.

b) داده‌های مربوط به feature های آن را به صورت دیتافریم تبدیل نمایید.

c) به انتهای دیتافریم یک ستون به نام Price اضافه کرده و مقدار target این دیتاست را در این ستون قرار دهید و دیتاست جدید را ذخیره نمایید.

i) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین اینکد کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

j) مجموعه‌های آموزشی و تست را با نسبت ۰.۷ به ۰.۳ ایجاد کنید.

d) با استفاده از کلاس LinearRegression مدل رگرسیون موردنظر خود را ایجاد نمایید و داده های آموزشی را به مدل fit کنید. سپس داده های تست را با استفاده از متد predict مدل پیش بینی کنید.

e) به منظور ارزیابی مدل ایجاد شده، مقادیر نتایج Mean Squared Error، Mean Absolute Error، Root Mean Squared Error را براساس مقادیر واقعی قیمت‌ها و مقادیر پیش بینی شده بدست آورید و تحلیل کنید.

f) همچنین در ادامه برای ارزیابی مدل از روش k-Fold Cross Validation استفاده خواهیم کرد. بدین منظور از متد cross_val_score استفاده کنید . مقدار cv را ۵ قرار دهید . (۵ بار مدل را آموزش داده و هر بار با داده تست جدید آن را ارزیابی خواهید کرد.) مقادیر مربوط به score های اجرا های مختلف را نشان داده و از آن میانگین بگیرید.

۳. کاهش ابعاد ویژگی‌ها (PCA)

a) دیتاست bearst_cancer را از کتابخانه sklearn لود کرده و داده‌های مربوط به feature های آن را به صورت دیتافریم تبدیل نمایید.

b) به انتهای دیتافریم یک ستون به نام Cancer اضافه کرده و مقدار target این دیتاست را در این ستون قرار دهید و دیتاست جدید را ذخیره نمایید.

c) دیتافریم موجود را از نظر داده‌های گم شده (Missing Value)، نرمالایز داده‌های عددی و همچنین اینکد کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.

d) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید.

e) با استفاده از کلاس MLPClassifier مدل شبکه عصبی موردنظر خود را ایجاد نمایید و داده های آموزشی را به مدل fit کنید. سپس داده های تست را با استفاده از متد predict مدل پیش بینی کنید.

f) میزان دقت مدل (Accuracy) روی داده‌های تست را نمایش دهید.

g) با استفاده از الگوریتم PCA از کتابخانه sklearn تعداد ویژگی‌های این دیتاست را به ۲ کامپوننت کاهش داده و با استفاده از نمودار scatter داده‌های جدید را به همراه کلاس مربوطه (وخیم و خوش‌خیم) نمایش دهید.

h) یکبار دیگر مدل MLPClassifier را با استفاده از داده‌های جدید آموزش داده و دقت مدل جدید را بدست آورید.

i) دقت مدل قبلی و مدل جدید را که با استفاده از PCA بدست آمده است را تحلیل کنید.

۴. کاوش قواعد (Association Rule)

a) ابتدا کتابخانه mlxtend را نصب نمایید.

b) دیتاست Basket خوانده و به صورت دیتافریم تبدیل نمایید. در این دیتاست هر سطر یک مجموعه فروش است و آیتم‌های هر مجموعه با کاما از یکدیگر جدا شده‌اند.

c) مجموعه‌های پرتکرار (frequent item sets) را با استفاده از کلاس apriori و حداقل ساپورت ۰.۱ بدست آورده تحلیل نمایید هر سطر از خروجی نمایانگر چه نتیجه‌ای است.

d) عملیات فوق را این بار با استفاده از کلاس fpgrowth و حداقل ساپورت ۰.۱ انجام دهید و تفاوت روش F-P Growth با روش قبلی را ذکر کنید.

e) قوانین وابستگی مجموعه تولید شده در مرحله c را تولید کنید. (راهنمایی: metric را برابر با lift و مقدار min_threshold را برابر ۰.۸ یا ۱ قرار دهید.)

f) یک مورد از نتایج بدست آمده را تفسیر نمایید.