به نام خدا

تکلیف هفتم درس مبانی داده کاوی

ترم بهار ۱۴۰۰

راهنمایی :
زبان برنامه نویسی سئوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج‌های اصلی استفاده شده seaborn،numpy, pandas, sklearn mlxtend 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 HW7-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.

۱. رگرسیون لاجستیک (Logistic Regression)

a) دیتاست bearst_cancer را از کتابخانه sklearn بارگذاری 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)، نرمال‌سازی داده‌های عددی و همچنین encode کردن ستون‌های دسته‌ای بررسی کرده و کارهای لازم برای آماده سازی دیتافریم به منظور ایجاد مدل را انجام دهید.
d) مقادیر همه ستون ها به جز ستون Cancer را در متغیر x قرار داده و ستون Cancer را در متغیر y قرار دهید.
e) مجموعه‌های آموزشی و تست را با نسبت ۰.۸ به ۰.۲ ایجاد کنید.

f) با استفاده از کلاس LogisticRegression مدل رگرسیون لاجستیک موردنظر خود را ایجاد نمایید.

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

h) حال با استفاده از پارامتر solver، دو مدل جدید با solverهای 'liblinear' و 'saga' ایجاد نمایید و دقت مدل‌های جدید را روی این دو مدل بررسی کنید. تفاوت این دو مدل را شرح دهید.

۲. رگرسیون پواسون (Poisson Regression)

a) دیتاست boston را از کتابخانه sklearn بارگذاری 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)، نرمال‌سازی داده‌های عددی و همچنین encode 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 ۰.۸ به ۰.۲ ایجاد کنید.

d) با استفاده از کلاسPoissonRegressor 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) دقت حاصل از این مدل را با دقت بددست آمده از مدل ایجاد شده در **سوال دوم تکلیف ششم** مقایسه کنید.

۳. کاوش الگوهای دنباله‌ای (Sequence Pattern Mining)

a) دیتاست Sequence.csv را خوانده و هر سطر از این دیتاست را به عنوان دنباله‌ی پشت سرهم کالاهای خریداری شده در یک خرید درنظر بگیرید. کالاهای هر خرید با کاما از یکدیگر جدا شده‌اند. (ترتیب کالاها برای حل سوال باید مددرنظر گرفته شود.)

b) الگوهای دنباله‌ای با حداقل ساپورت ۰.۳ را در این مجموعه بدست آورید. (یعنی الگوهایی که حداقل در ۳۰ درصد از تراکنش‌ها تکرار شده‌اند.)

c) الگوی Bread,Sweet بیشتر در ادامه کدام محصول تکرار شده است. پنج مورد از پرتکرارترین محصولات را شناسایی کنید.

۴. کاوش قواعد پیشرفته (Advanced Association Rule Mining)

a) دیتاست Heart را خوانده و به صورت دیتافریم تبدیل نمایید. در صورت وجود مقادیر NULL، این مقادیر را با مقدار مناسب جایگذاری یا حذف نمایید.

b) مجموعه‌های پرتکرار را با استفاده از apriori و حداقل ساپورت ۰.۲ به ازای حالت‌های وقوع یا عدم وقوع سکته قلبی بدست آورید.

c) قوانین وابستگی مجموعه تولیدشده در مرحله b را به ازای حالت‌های وقوع یا عدم وقوع سکته قلبی تولید کنید. استفاده تنها از چند ستون مهم در نتیجه ستون target نیز برای تولید قوانین وابستگی قابل قبول می‌باشد.

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