به نام خدا

تکلیف اول درس مبانی داده کاوی

ترم دوم ۱۴۰۰-۱۳۹۹

راهنمایی :
زبان برنامه نویسی سئوالات پایتون است.
پیشنهاد می شود از محیط jupyter notebook استفاده کنید.
پکیج های اصلی استفاده شده numpy, pandas 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، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW1-Lastname-StudentCode.zip که Lastname نام خانوادگی و StudentCode شماره دانشجویی فرد است، زیپ شده و روی سامانه تا **زمان مشخص شده** اپلود 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.

- **آشنایی با فرایندهای داده‌کاوی**

۱- فرض کنید مسئله‌ای تحت عنوان "پیش بینی روند مصرف دارو در فصل زمستان در بیمارستان ها" تعریف شده است. همچنین داده هایی از مراجعات سالیان گذشته به بیمارستان ها و وضعیت آب و هوایی در اختیار شما قرار گرفته است. شما به عنوان دانشمند داده می‌خواهید مسیر حل مسئله را با استفاده از فرایند CRISP-DM مشخص نمایید. فازهای فرایند CRISP-DM را برای این مسئله شرح دهید و در هر فاز مواردی که به نظر شما باید در آن فاز مورد توجه قرار گیرد را توصیف نمایید.

۲- فرض کنید داده هایی در مورد دانشجویان یک دانشگاه مانند سوابق تحصیلی (مقطع، رشته، نمرات دروس و ...) و اطلاعات هویتی (سن، جنسیت، خوابگاهی بودن یا نبودن، شهر محل سکونت و غیره) در دسترس میباشد. با توجه به داده های موجود، چهار نمونه مسئله با استفاده از تسک های داده کاوی (مانند توصیفی، پیش بینی، دسته بندی و ...) معرفی و توصیف نمایید.

(مثال: پیش بینی رشد یا افت تحصیلی دانشجو بر اساس پارامتر خوابگاهی بودن یا نبودن)

- **شروع کار با کتابخانه‌های پایتون**

۳- : از دیتاست Vehicle که در اختیارتان قرار داده شده است برای حل این سوال استفاده نمایید:

۳-۱- ابتدا دیتاست Vehicle را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.

۳-۲- اطلاعات توصیفی دیتاست مانند تعداد و نوع ستون ها، حجم دیتاست و غیره نمایش دهید.

۳-۳- سطرهایی که مقدار ستون CIRCULARITY بین دو مقدار ۵۰ تا ۶۰ است را بازیابی کرده و در یک فایل csv ذخیره نمایید.

۳-۴- میانگین، کمترین، بیشترین و انحراف از معیار ستون RADIUS_RATIO را بدست آورده و نمایش دهید.

۳-۵- ستونی با نام ElonBin به دیتافریم اضافه نمایید که مقدار این ستون در هر سطر به ازای مقدار ستون ELONGATEDNESS در آن سطر محاسبه می‌شود. اگر مقدار ELONGATEDNESS کمتر از ۳۰ بود مقدار LOW، اگر بین ۳۰ تا ۴۵ بود مقدار MEDIUM و اگر بیش از ۴۵ بود مقدار HIGH را در ستون ElonBin جایگذاری نمایید.

۳-۶- مقدار میانگین ستون DISTANCE_CIRCULARITY را به ازای هر یک از مقادیر یکتای ستون Class نمایش دهید

۳-۷- وابستگی بین مقادیر ستون های مختلف این دیتاست را به صورت دوبه دو بدست آورید

۳-۸- نمودار مقادیر ستون SCATTER_RATIO را به صورت هیستوگرام نمایش دهید.