به نام خدا

تکلیف دوم 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 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 HW2-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.

دیتاست شماره ۱: مربوط به بیماران مبتلا به دیابت است. با نام diabetes در سئوالات به آن اشاره شده است.

- ## پیش پردازش

۱- (سوال سوم از تکلیف قبلی) از دیتاست Vehicle که در اختیارتان قرار داده شده است برای حل این سوال استفاده نمایید:

a) ابتدا دیتاست Vehicle را با استفاده از کتابخانه pandas خوانده و تبدیل به دیتافریم نمایید.
b) اطلاعات توصیفی دیتاست مانند تعداد و نوع ستون ها، حجم دیتاست و غیره نمایش دهید.
c) سطرهایی که مقدار ستون CIRCULARITY بین دو مقدار ۵۰ تا ۶۰ است را بازیابی کرده و در یک فایل csv ذخیره نمایید.
d) میانگین، کمترین، بیشترین و انحراف از معیار ستون RADIUS_RATIO را بدست آورده و نمایش دهید.
e) 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.

f) مقدار میانگین ستون DISTANCE_CIRCULARITY را به ازای هر یک از مقادیر یکتای ستون Class نمایش دهید

g) وابستگی بین مقادیر ستون های مختلف این دیتاست را به صورت دوبه دو بدست آورید

h) نمودار مقادیر ستون SCATTER_RATIO را به صورت هیستوگرام نمایش دهید.

۲- بررسی پراکندگی مقادیر مفقود در دیتاست

بررسی کنید در دیتاست diabetes مقادیر NULL به چند صورت نمایش داده شده است و تمامی این مقادیر را با مقدار اصلی درنظر گرفته‌شده برای مقادیر مفقود در کتابخانه pandas جایگذاری کنید. سپس تعداد مقدار مقادیر مفقود در هر ستون را محاسبه کنید.

۳- مدیریت مقادیر مفقود در دیتاست

با استفاده از دیتاست اصلاح شده در سوال۲ به این سوال پاسخ دهید

a) مقادیر مفقود را با یک عبارت ثابت جایگزین کنید.

b) مقادیر مفقود را با یک مقدار ثابت از همان ستون جایگزین کنید.

c) مقادیر مفقود را با مقدار میانگین همان ستون جایگزین کنید.

d) مقادیر مفقود را با مقدار مد همان ستون جایگزین کنید.

e) نمودار هیستوگرام ستون Glucose را برای قسمت‌های قبل این سوال رسم کرده و توضیح دهید به نظر شما کدام روش، روشی منطقی‌تر برای جایگذاری مقادیر NULL است.

f) تحقیق کنید چه روشهایی برای imputation داده‌های مفقود وجود دارد؟ به منابعی که مطالعه نموده‌اید اشاره کنید.

g) مقادیر مفقود در هر ستون را با استفاده از متد imputer با میانگین داده‌های آن ستون جایگزین کنید.

۴- نرمال‌سازی و استانداردسازی داده‌های دیتاست

a) دلیل اهمیت نرمال‌سازی و استانداردسازی محدوده مقادیر داده‌ها را بیان نموده و با ذکر مثال مشخص نمایید در صورت عدم انجام این فرایند چه مشکلاتی در فرایندهای بعدی داده کاوی رخ می‌دهد

b) ابتدا همانند قسمت g سوال قبل مقادیر NULL دیتاست diabetes را جایگزین نمایید

c) سپس با استفاده از توابع StandardScaler، Normalize و MinMaxScaler مقادیر ستون‌های این دیتاست را نرمال‌سازی کنید.

d) نمودار هیستوگرام ستون Glucose دیتاست را به ازای روش‌های MinMaxScaler و Normalize در کنار هم رسم نموده و نتایج هر دو نمودار را تفسیر نمایید.
e) تفاوت روش StandardScaler و Normalize را توضیح دهید.

۵- داده های پرت

a) ابتدا همانند قسمت g سوال ۳ ، مقادیر NULL دیتاست diabetes را جایگزین نمایید.
b) با استفاده از روش IQR داده های پرت موجود در دیتاست را به ازای هر ستون شناسایی و نمایش دهید و همچنین نمودار boxplot آن را رسم نمایید.
c) با استفاده از روش z-score داده‌های پرت دیتاست را حذف نمایید و تعداد سطرهای حذف شده در حالت فعلی و قبلی را نمایش دهید.

۶- Binning

a) ابتدا همانند قسمت g سوال ۳ ، مقادیر NULL دیتاست diabetes را جایگزین نمایید
b) با استفاده از روش cut ستون های Pregnancies را در سه دسته و ستون Age را در چهار دسته، دسته بندی نمایید و نمودار هیستوگرام این دو ستون را نمایش دهید.
c) تحقیق کنید برای binning داده‌ها چه روشی علاوه بر روش‌های cut و qcut در پایتون وجود دارد و عملکرد آن را توضیح دهید.