به نام خدا

تکلیف اول درس داده کاوی

ترم دوم ۹۹-۹۸

راهنمایی : زبان برنامه نویسی برای هر سئوال در ابتدای سئوال مشخص شده است.
است برای سئوالاتی که با Python باید پاسخ داده شود از کتابخانه های numpy, pandas, sklearn , seaborn استفاده نمایید. سایر کتابخانه ها مورد نظر در هر سئوال اشاره شده است. دیتاست های مورد نیاز در ادامه معرفی شده اند.

روش تحویل:

الف) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Ax.zip که x شماره سوال است زیپ شوند، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW1-Lastname.zip که Lastname فامیل شماست، زیپ شده و روی سامانه تا **مهلت اعلام شده** اپلود شوند.

ب) تحویل بصورت حضوری (در صورت رفع شرایط کرونا) یا بصورت تماس ویدئویی در مهلت اعلام شده خواهد بود. برای هر سئوال کد نوشته شده و نتیجه اجرا را در فایل نهایی وارد کنید. فایل نهایی باید به صورت pdf باشد.

دیتاست ها:

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

دیتاست شماره ۲: دیتاست داده‌های مربوط به مشخصات خانه‌های ساحلی است که با نام housing به آن اشاره شده است.

دیتاست شماره ۳: مربوط به بیماران مبتلا به دیابت است . با نام diabetes در سئوالات به آن اشاره شده است.

دیتاست شماره ۴: مربوط به بیماران مبتلا به کرونا است . با نام corona در سئوالات به آن اشاره شده است.

۱. خلاصه سازی داده‌ها:

۱,۱. R: با استفاده از متد summary نمایی از دیتاست thyroid را نشان دهید.

۱,۲. Python : با استفاده از کتابخانه pandas دیتاست thyroid را به دیتافریم تبدیل کرده و سپس اطلاعات کلی در مورد این دیتاست و ویژگی‌های آن را نمایش دهید.

۱,۳. Python : مقادیر یکتای ویژگی‌های دسته‌ای دیتاست thyroid را همراه با تعداد موجود در هر دسته نمایش دهید.

۲. **Missing Value ها را در دیتاست thyroid شناسایی کنید.**

۲,۱. R: تابعی بنویسید که با دریافت دیتاست و با استفاده از متد is.na تعداد Missing Value های هر ستون را برگرداند.

۲,۲. R: در مورد روش استفاده از کتابخانه های VIM و MICE برای شناسایی Missing Values تحقیق نموده و نتایج حاصل از اجرای دستور aggr را تفسیر کنید. راهنمایی : https://rpubs.com/sediaz/na_aggr

۲,۳. Python : بررسی کنید در دیتاست thyroid مقادیر null به چند صورت نمایش داده شده است و تمامی این مقادیر را با مقدار صحیح null جایگذاری کنید. سپس با استفاده از دستور isNull مقدار مقادیر null در هر ستون را محاسبه کنید.

۳. روشهای جایگزینی Missing Value – دیتاست thyroid

۳,۱. R: تابعی بنویسید که مقادیر مفقود را با عبارت ثابت جایگزین کند. هیستوگرام داده ها را قبل و بعد از اجرای این روش با استفاده از تابع barMiss نشان دهید. دیتاست جدید تولید شده را ذخیره کنید.

۳,۲. R: تابعی بنویسید که مقادیر مفقود در یک ستون را با یک مقدار تصادفی از داده‌های همان ستون جایگزین کند. دیتاست جدید تولید شده را ذخیره کنید.

۳,۳. R : تابعی بنویسید که مقادیر مفقود در هر ستون را با مقدار Mode داده های آن ستون جایگزین کند. دیتاست جدید تولید شده را ذخیره کنید.

۳,۴. Python: مقادیر مفقود را با عبارت ثابت جایگزین کنید.

۳,۵. Python : مقادیر مفقود در هر ستون را با استفاده از کتابخانه imputer با میانگین داده های آن ستون جایگزین کنید.

۳,۶. تحقیق: چه روشهایی برای imputation داده های مفقود وجود دارد؟ به منابعی که مطالعه نموده اید اشاره کنید.

۴. دسته بندی داده‌ها – دیتاست housing

۴,۱. Python : داده های موجود در دیتاست housing را با استفاده از متد groupby و بر اساس مقدار ocean_proximity دسته‌بندی کنید و برای نمایش داده‌ها در هر دسته از میانگین استفاده کنید

۵. نرمال‌سازی داده های دیتاست

۵,۱. Python: با استفاده از توابع StandardScaler ، normalize و minmax_scale مقادیر دیتاست diabetes را نرمال‌سازی کنید

۵,۲. R: با استفاده از روش Min-Max و Z-score داده های ستون های عددی دیتاست thyroid را نرمال سازی کنید. نمودار های Side-by-Side ی رسم کنید که داده ها را قبل و بعد از نرمال سازی به روشهای فوق نشان دهد

۶. داده های پرت:

۶,۱. Python : با استفاده از روش IQR داده های پرت را در دیتاست thyroid شناسایی و boxplot آن را رسم نمایید

۶,۲. R: تابعی بنویسید که با استفاده از روش IQR داده های پرت در یک ستون را مشخص کند. نتیجه را برای ستون T3_resin از دیتاست thyroid شناسایی و boxplot آن را رسم نمایید .

۷. خلاصه سازی و بصری سازی- دیتاست housing

۷,۱. Python : نمودار هیستوگرام هر یکی از ویژگی‌های دیتاست housing را نمایش دهید.

۷,۲. Python : دو ویژگی longitude و latitude در دیتاست housing را با استفاده از داده‌های مکانی بر روی نقشه نمایش دهید. تراکم مناطقی که تعداد خانه‌های بیشتری در آنجا وجود دارد را نیز روی نقشه مشخص نمایید.

۸. بررسی همبستگی بین متغیرها در دیتاست housing

۸,۱. Python : در دیتاست housing، همبستگی بین متغیرها را با استفاده از نمودار pairplot بررسی کرده و این نمودار را تفسیر کنید.

۸,۲. Python : در دیتاست housing، همبستگی بین متغیرهای median_house_value و median_income را با استفاده از تابع pearsonr بدست آورید.

۸,۳. Python : با استفاده از متد corr از کتابخانه pandas مقدار همبستگی بین متغیرهای median_house_value و median_income را در دیتاست housing نشان دهید.

۸,۴. Python : دیتافریم بدست آمده در مرحله قبل را با استفاده از نمودار heatmat از پکیج seaborn نشان دهید. این نمودار برای چه مواقعی مناسب است؟

۸,۵. R: با استفاده از دیتاست housing برای ستون‌های housing_median_age, total_rooms, median_house_value نمودار pairplot رسم کنید.

۸,۶. R: به نمودار رسم شده در مرحله قبل correlation ها را اضافه کنید. کدام دو متغیر بیشترین correlation را دارند؟

۸,۷. R: برای داده های housing نمودار heatmap را رسم کنید. این نمودار را تفسیر کنید.

۹. Chi-Square دیتاست diabetes

۹,۱. در خصوص رابطه بین گلوکز خون و ابتلا به دیابت در دیتاست diabetes آن فرض H0 و H1 را تعیین کنید. هدف بررسی وابستگی یا استقلال این دو پارامتر است.

۹,۲. Python : جدول observed را برای متغیرهای گلوکز خون و ابتلا به دیابت در دیتاست diabetes ایجاد کنید.

۹,۳. Python : با استفاده از متد chi2_contingency از کتابخانه scipy مقادیر مربوط به chi-square ، p-value ، درجه آزادی و جدول expected را نشان دهید.

۹,۴. R : با استفاده از متد chisq.test مقادیر مربوط به chi-square ، p-value و جدول expected را نشان دهید.

۹,۵. با توجه به مقادیر بدست آمده آیا شواهد کافی برای رد فرض صفر وجود دارد؟

۱۰. رگرسیون

۱۰,۱. Python: دیتاست diabetes را تبدیل به دیتافریم نموده و در صورت داشتن مقادیر null در این دیتاست، این مقادیر را با روش مناسب جایگزین کنید.

۱۰,۲. Python: در دیتاست diabetes برای پیش‌بینی اینکه فردی دیابت دارد یا خیر، یک مدل Regression ایجاد و آموزش دهید. در این قسمت مقادیر ستون‌ها را بدون نرمال سازی آموزش دهید و سپس مقدار خطای مجموعه تست را محاسبه و نمایش دهید.

۱۰,۳.Python: یک بار دیگر مدلی جدید آموزش دهید و برای نرمال سازی مقادیر ستون‌های عددی از تابع MinMax استفاده کنید. مقدار خطای مدل جدید را با مدل قبلی مقایسه نمایید.

۱۰,۴.Python: یک بار دیگر مدلی جدید آموزش دهید و برای نرمال سازی مقادیر ستون‌های عددی از تابع StandardScaler استفاده کنید. مقدار خطای مدل جدید را با مدل قبلی مقایسه نمایید.

۱۰,۵.تحقیق کنید کدوم تابع نرمال‌سازی خطای کمتری برای مدل ایجاد می‌کند.

۱۱. تقسیم داده ها دیتاست diabetes

۱۱,۱.R: داده های دیتاست diabetes را به مجموعه آموزشی و تست به نسبت ۰,۸ و ۰,۲ تقسیم کنید و مجموعه های بدست آمده را با اسامی متناسب نام گذاری کنید.

۱۱,۲.Python: داده های دیتاست diabetes را به مجموعه آموزشی و تست به نسبت ۰,۸ و ۰,۲ تقسیم کنید و مجموعه های بدست آمده را با اسامی متناسب نام گذاری کنید.

۱۱,۳.Python: تحقیق کنید پارامتر stratify در کتابخانه scikit برای تقسیم داده‌ها به چه منظوری استفاده می‌شود.

۱۱,۴.بررسی کنید آیا ویژگی ابتلا به دیابت در هر دو مجموعه آموزشی و تست به طور یکسان توزیع شده‌اند یا خیر.

۱۲. kNN دیتاست Corona

۱,۱. ابتدا داده‌ها را با استفاده از کتابخانه pandas به فرمت دیتافریم تبدیل کنید.

۱,۲. متد ()describe را در مورد داده‌هایی که به صورت دیتافریم تبدیل شده اند اجرا نمایید.

۱,۳. دستور ()value_counts را در مورد فیلد outcome اجرا کنید. نتیجه اجرا چه اطلاعاتی در بردارد؟

۱,۴. مقادیر مختلف ستون outcome را به گونه‌ای با یکدیگر تجمیع کنید تا مقادیر مشابه در یک تارگت قرار گیرند و دسته‌های نهایی شامل مقادیر **فوت‌شده، بهبودیافته، تحت درمان** باشد.

۱,۵. مقادیر null در این دیتاست را با روشی مناسب جایگزین یا حذف نمایید.

۱,۶. ستون‌های دارای مقادیر دسته‌ای را به گونه‌ای مناسب تبدیل به مقادیر عددی کنید. برای این کار میتوان از تابع OneHotEncoder استفاده نمایید.

۱,۷. برای تقسیم داده ها به مجموعه تست و آموزش ، تابع train_test_split را مقداردهی و اجرا نمایید.

۱,۸. ابعاد مجموعه های X_train ، X_test، y_train و y_test را نشان دهید.

۱,۹. دسته بند KNeighborsClassifier را با مقدار ۵ روی داده های آموزشی اجرا نموده (مدل را آموزش دهید) و دقت دسته بندی را روی داده های تست با تابع score نشان دهید.

۱,۱۰. مقدار هدف را برای مجموعه X_test با استفاده از تابع predict بدست آورید.

۱،۱۱. به زبان ساده عملکرد predict را توضیح دهید.

۱،۱۲. از پکیج preprocessing تابع MinMaxScaler را ایمپورت کرده و با استفاده از آن داده های X_train و X_test را نرمال سازی کنید.

۱،۱۳.بار دیگر مدل را با استفاده از داده های آموزشی نرمال سازی شده ، آموزش دهید.

۱،۱۴.دقت مدل را روی داده های آموزشی و روی داده های تست با استفاده از تابع score بدست آورید.

۱،۱۵.برای بررسی تاثیر وزن همسایه‌های هر نقطه میتوان از پارامتر weights در مدل استفاده نمود. تحقیق کنید این پارامتر چه مقادیری میپذیرد و میزان تاثیر این پارامتر در دقت مدل را بررسی نمایید. (برای نمونه تابع وزن را براساس فاصله اقلیدسی درنظر گرفته و دقت مدل را بررسی نمایید.)

۱،۱۶.برای بررسی تاثیر نوع الگوریتم محاسبه نزدیکترین همسایه میتوان از پارامتر algorithm در مدل استفاده نمود. تحقیق کنید این پارامتر چه مقادیری میپذیرد و میزان تاثیر این پارامتر در دقت مدل را بررسی نمایید.

۱،۱۷.برای بررسی اثر تعداد همسایه ها ، یک آرایه به نام train_accuracy و یک آرایه به نام test_accuracy ایجاد نموده ، سپس در یک حلقه for مقدار همسایگی را از ۱ تا ۱۰ افزایش داده و هر بار دقت مدل را روی داده های آموزشی و تست در ایندکس مورد نظر از آرایه های مربوطه ذخیره کنید. (دقت مدل روی داده های آموزشی در آرایه train_accuracy و دقت مدل روی داده های تست در آرایه test_accuracy ذخیره شود.)

۱،۱۸.با استفاده از کتابخانه matplotlib.pyplot روند تغییرات دقت بدست آمده روی داده های آموزشی و تست را که در قسمت قبل در آرایه های مورد نظر ذخیره نمودید به صورت نمودار نشان داده و جزئیات نمودار را مشخص کنید.

۱،۱۹.تفسیر خود را از نمودار بنویسید.