به نام خدا

تکلیف سوم 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 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 HW3-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 اطلاعات تعدادی خانه است. با نام housing در سئوالات به آن اشاره شده است.
دیتاست شماره ۲: مربوط به اطلاعات تعدادی گوشی همراه است. با نام smartphone در سئوالات به آن اشاره شده است.
دیتاست شماره ۳: مربوط به بیماران مبتلا به تیروئید است . با نام thyroid در سئوالات به آن اشاره شده است.

نکته ۱ : برای نمایش گرافیکی گراف های سوال پنج از دو کتابخانه pydotplus و graphviz استفاده نمایید. برای دانلود graphviz از لینک زیر استفاده نمایید: (https://graphviz.gitlab.io/download/)

• تحلیل کاوشگرانه داده‌ها (EDA)

۱. خلاصه سازی و بصری سازی

a) دیتاست housing را خوانده و سطرهای شامل مقادیر Null را حذف نمایید.

b) برای ستون ocean_proximity تمامی مقادیر یکتا را به همراه تعداد آن‌ها نمایش دهید.

c) نمودار هیستوگرام هر یکی از ویژگی‌های دیتاست housing را نمایش دهید.

d) 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. (با استفاده از طول و عرض جغرافیایی موارد موردنظر را نمایش دهید)

۲. بررسی همبستگی بین متغیرها

a) در دیتاست housing، همبستگی بین متغیرها را با استفاده از نمودار pairplot بررسی کرده و این نمودار را تفسیر کنید.

b) کدام دو متغیر بیشترین correlation را در دیتاست housing دارند؟

c) تحقیق کنید روش محاسبه همبستگی pearsonr به صورت عمل میکند و مقدار p-value به چه معناست.

d) در دیتاست housing، همبستگی بین ستون‌های median_house_value و median_income را با استفاده از تابع pearsonr بدست آورید.

e) در دیتاست housing، همبستگی بین ستون‌های housing_median_age و total_rooms را با استفاده از تابع spearman بدست آورید.

f) با استفاده از متد corr از کتابخانه pandas مقدار همبستگی بین متغیرهای median_house_value و median_income را در دیتاست housing نشان دهید.

g) تمامی همبستگی‌های دوبه دوی دیتاست housing را با استفاده از نمودار heatmat از پکیج seaborn نشان دهید.

۳. جدول همسانی (Contingency Table)

a) جدول همسانی دو متغیر Company و Capacity را در دیتاست smartphone بدست آورده و نمایش دهید.

b) جدول همسانی سه متغیر Company و Weight و inch را در دیتاست smartphone بدست آورده و نمایش دهید.

c) جدول همسانی دو متغیر Company و OS را با استفاده از نمودارهای ستونی نمایش دهید.

۴. تقسیم داده ها

a) داده های دیتاست thyroid را به دو مجموعه آموزشی و تست به نسبت ۰،۸ و ۰،۲ تقسیم کنید و مجموعه های بدست آمده را ذخیره کنید.

b) تحقیق کنید پارامتر stratify در کتابخانه scikit برای تقسیم داده‌ها به چه منظوری استفاده می‌شود.

c) بررسی کنید آیا ویژگی دسته‌بندی نوع تیروئید (Outcome) در هر دو مجموعه‌ی آموزشی و تست به طور یکسان توزیع شده‌اند یا خیر.

d) انواع روش‌های موجود برای پیش پردازش دیتاست‌های Imbalanced را بررسی کرده و شرح دهید. (بیان ۲ روش کافیست)

۵. درخت تصمیم

a) دیتاست thyroid را تبدیل به دیتافریم نموده و در صورت داشتن مقادیر null در این دیتاست، مقادیر null را با مقدار میانگین همان ستون جایگذاری کنید.

b) همبستگی تمامی ستون‌ها را نسبت به یکدیگر محاسبه و نمودار heatmap آن را رسم کنید.

c) آیا این دیتاست از نوع Imbalanced دیتاست میباشد؟ در صورت Imbalanced بودن با روش مناسبی توزیع ستون Outcome را تصحیح نمایید.

d) ابتدا مقادیر همه ستون‌ها به جز ستون Outcome را در متغیر x قرار داده و ستون Outcome را در متغیر y قرار دهید. سپس با استفاده از تابع train_test_split و انتخاب مقدار test_size=0.2 مجموعه های آموزشی و تست را ایجاد کنید.

e) با استفاده از DecisionTreeClassifier داده ها را دسته‌بندی کنید . (راهنمایی : مقدار پارامترهای ورودی را به صورت زیر قرار دهید: criterion='gini')

f) داده های تست را به مدل درخت تصمیم بدهید و میزان دقت را نمایش دهید.

g) بهترین مقدار پارامتر max_depth را با استفاده از مقادیر مختلف ۱ تا ۹ بررسی کرده و بهترین مقدار دقت درخت بدست آمده را نشان دهید.

h) توضیح دهید متد feature_importances نشان دهنده چیست و مقدار آن را برای classifier بدست آورید.

i) خروجی تابع export_graphviz را بر روی classifier ی که با بهترین پارامترهای بدست آمده خواهید ساخت بدست آورده و ذخیره کنید.

j) کتابخانه pydotplus را نصب کنید و با استفاده از آن فایل dot_data را به گراف تبدیل کنید و آن را نمایش دهید.