به نام خدا

تکلیف اول درس مبانی داده کاوی

ترم دوم ۱۴۰۰-۱۴۰۱

راهنمایی :
زبان برنامه نویسی سوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج های اصلی مورد نیاز شامل pandas،numpy می باشند.
مجموعه داده 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 شما است، زیپ شده و روی سامانه تا زمان مشخص شده آپلود شوند.
ب) گزارش نهایی باید شامل پاسخ تمامی سوالات (سوالات تشریحی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی درمورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد (گزارش سوالات پیادهسازی را میتوانید در همان محیط notebook jupyter بنویسید).
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.

۱. فرض کنید مجموعه داده ای شامل اطلاعات مربوط به خودروهای مختلف در اختیار شما قرار گرفته است. این اطلاعات شامل ویژگیهای مختلف خوردوها مانند وزن، استایل، میزان سوخت مصرفی بر حسب لیتر بر کیلومتر، نوع موتور خودرو و ... است. روی این مجموعه داده مسئله پیش بینی قیمت خودرو است. فازهای مختلف فرآیند CRISP-DM را برای این مسئله بطور کامل بیان کنید.

۲. برای دانش آموزان یک مدرسه در طول سال سه بار ارزشیابی انجام می شود و نمرات ارزشیابی به ترتیب G1 و G2 و G3 نامیده می شود. در مجموعه داده student.csv اطلاعات هر دانش آموز از جمله ساعات مطالعه در روز، وضعیت خانوادگی ، جنسیت و دیگر موارد وجود دارد که جزئیات آن در فایل StudentsReadMe موجود است. براساس این مجموعه داده به سوالات زیر پاسخ دهید:
   a. شکل داده ها به چه صورتی است و چند مقدار missing وجود دارد.
   b. هیستوگرامی رسم کنید که فراوانی نمرات G1و G2 و G3 در یک شکل نشان دهد.
   c. خصوصیت جدیدی بنام Grade ایجاد کنید که متوسط سه نمره ارزشیابی سالیانه باشد.
   d. با استفاده از نمودار Pie ، فراوانی مقادیر خصوصیات جنسیت، internet ، آدرس ، Pstatus، reason و schoolsup را در یک شکل نشان دهد.
   e. نمودار میله ای برای جنسیت و نمره (Grade) رسم کنید. نمرات پسران بیشتر بوده یا دختران؟
   f. نمودار میله ای برای وضعیت سلامت health و نمره (Grade) به تفکیک جنسیت رسم کنید و آنرا تفسیر کنید.
   g. نمرات را به چهار دسته A و B و C و D و E تقسیم کنید و سپس فراوانی هر دسته را به نمودار Pie رسم کنید.
   h. با استفاده از heatmap میزان همبستگی دو به دو خصوصیات را رسم کنید. کدامیک بیشترین همبستگی را به هم دارند؟

با توجه به مجموعه داده machines.csv که شامل اطلاعات دستگاههای مختلف برای تخمین کارآیی CPU آنها است (جزئیات آن در فایل MachinesReadMe نوشته است)، به سوالات ۳ تا ٦ پاسخ دهید:

۳. آشنایی با pandas

a. به کمک کتابخانه pandas مجموعه داده را بخوانید، آن را در یک دیتافریم ذخیره کنید و سپس ستونهای آن را به ترتیب زیر نام گذاری کنید:

```
Vendor_name،model،MCYT،MMIN،MMAX،CACH،CHMIN،CHMAX،PRP،ERP
```

b. نوع دادهای هر یک از فیلدهای مجموعه داده را نشان داده و فیلد vendor name را به نوع رشته ای تبدیل و ذخیره کنید.

c. برای ستون vendor_name مقادیر یکتای موجود را نشان دهید

d. برای دستگاههایی که کمینه حجم حافظه اصلی آنها بین ۱۰۰۰ تا ۲۰۰۰۰ است، ستونهای vendor_name، model، MMIN، MMAX، PRP و ERP را در یک دیتافریم جدید با نام df1 ذخیره کرده و ۱۰ تاپل آخر آن را نمایش دهید. df1 را در یک فایل csv نیز ذخیره کنید.

e. اطلاعات مربوط به دستگاههایی که مدل آنها از سری b است (با حرف b شروع میشوند) را نمایش دهید.

f. برای فیلد ERP مقادیر میانگین و انحراف از معیار را نشان دهید. هم چنین رکوردی که مقدار فیلد ERP آن بیشینه است را نشان دهید.

g. تمامی رکوردها را بر اساس فیلد vendor_name گروه بندی کنید، برای هر گروه تعداد رکوردهای آن گروه و نیز min و max فیلد ERP آن گروه را نمایش دهید.

h. همبستگی pearson را بین تمام فیلدهای عددی مجموعه داده به دست آورده و نتایج را تفسیر کنید.

٤. بررسی balance/imbalance بودن مجموعه داده

a. ابتدا توضیح دهید که imbalance بودن مجموعه داده چه مشکلاتی میتواند بوجود بیاورد؟

b. یک فیلد جدید تحت عنوان 'labeled_PRP' به مجموعه داده اضافه کنید. مقدار این فیلد را برای تاپل هایی که فیلد PRP آنها کمتر و یا مساوی ۳۰۰ است برابر با 'A' و برای سایر تاپلها برابر با 'B' در نظر بگیرید.

c. بر اساس 'labeled_PRP' مشخص کنید که مجموعه داده دادهشده imbalance است یا خیر؟

d. در صورت imbalance بودن به کمک تکنیک up sampling آنرا balance کنید و نتیجه را در یک دیتافریم جدید با نام df_upSampled ذخیره کنید و به اختصار در مورد این تکنیک توضیح دهید.(میتوانید از تابع resample کتابخانه sklearn استفاده کنید).

e. در صورت imbalance بودن به کمک تکنیک down sampling آنرا balance کنید و نتیجه را در یک دیتافریم جدید با نام df_downSampled ذخیره کنید. به اختصار در مورد این تکنیک توضیح دهید.

٥. بررسی داده های پرت

a. به کمک رسم boxplot برای مجموعه داده وجود داده outlier را بررسی کنید.

b. به کمک روش z-score دادههای پرت را مشخص کنید، سپس آنها را از مجموعه داده حذف کرده و مجموعه داده جدید را در یک مجموعه داده با نام cdf1 ذخیره کنید و بررسی کنید چه تعداد رکورد از مجموعه داده اولیه حذف شده است.

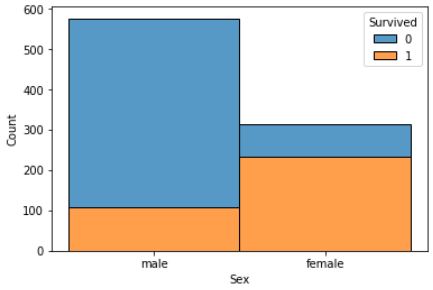
c. به کمک روش IQR دادههای پرت را مشخص کنید، سپس آنها را از مجموعه داده حذف کرده و مجموعه داده جدید را در یک مجموعه داده با نام cdf2 ذخیره کنید و بررسی کنید چه تعداد رکورد از مجموعه داده اولیه حذف شده است.

۶. Binning

a. به کمک روش فاصله و تابع cut در پایتون، مقادیر ستون CACH را به ۳ دسته تقسیم بندی کرده و نمودار هیستوگرام آنرا رسم کنید و این روش را به اختصار توضیح دهید.
b. به کمک روش فاصله و تابع qcut در پایتون، مقادیر ستون CACH را به ۳ دسته تقسیم بندی کرده و نمودار هیستوگرام آنرا رسم کنید و این روش را به اختصار توضیح دهید.
c. از کتابخانه jenkspy برای پیداکردن natural breaks روی ستون CACH استفاده کرده و آنرا به ۳ دسته تقسیم کرده و نمودار هیستوگرام آنرا رسم کنید و این روش را به اختصار توضیح دهید.

۷. در مجموعه داده Kaggle Titanic ، اطلاعات مربوط به تعدادی از مسافرین کشتی تایتانیک وجود دارد. این کشتی پس از برخورد به کوه یخ غرق شد و تعداد از مسافران آن نجات پیدا کردند. در این مجموعه اطلاعات مسافران از جمله ID ، سن ، تعداد فرزند و ... وجود دارد که در فایل TitanicReadMe توضیح داده شده است. با استفاده از فایل train.csv به سوالات زیر پاسخ دهید:

a. از طریق pandas داده ها را خوانده و اطلاعات آماری آن شامل میانگین، واریانس حداقل و حداکثر و صدک های آن را نمایش دهید.
b. در داخل یک شکل ، میانگین feature ها را بصورت مرتب شده نشان دهید.
c. Boxplot همه feature ها را داخل یک شکل نشان دهید.
d. با استفاده از heatmap ، همبستگی دو به دو feature ها را نمایش دهید و مشخص کنید کدام دو خصوصیت بیشترین همبستگی را با یکدیگر دارند.
e. داده های missing را توسط heatmap نمایش دهید. سطرها passengerID و ستون ها features باشند. درصد missing هر خصوصیت را مشخص کنید. بیشترین missing در کدام ستون قرار دارد.
f. خصوصیت جدیدی بنام alone ایجاد کنید که مشخص کند آیا مسافر وابستگانی در داخل کشتی داشته است یا خیر.
g. هیستاگرام برای سن ایجاد کنید و آنچه از این نمودار متوجه می شوید شرح دهید.
h. مقادیر missing مربوط به سن را به روش mode پرکنید.
i. در مورد مقادیر missing در خصوصیت cabin چه روشی بهتر است انجام شود؟
j. کدی برای ایجاد نمودار زیر بنویسید:



i.

k. خصوصیت جدیدی به نام who ایجاد کنید که اگر سن مسافر کمتر از ۱۰ سال باشد برابر با child و در غیر اینصورت برابر با جنسیت فرد باشد(male/female).
l. از طریق دستور cut خصوصیت کرایه (fare) را به چهار دسته ارزان ، متوسط گران و خیلی گران تقسیم کنید و نمودار هیستوگرام آن را ترسیم کنید.