به نام خدا

تکلیف پنجم 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 :
a) 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 HW5-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 پیاده سازی را میتوانید در همان محیطJupyter notebook بنویسید).
ج) زمان و نحوه تحویل تکلیف روی سامانه و در فایل راهنمای ترم مشخص شده است.
د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

۱. **Association Rules** (زمان تقریبی: ۴۵ دقیقه)

a. کتابخانه mlxtend را نصب کنید.
b. داده Hackathon_Working_Data را خوانده و مقادیر null را حذف کنید.
c. ID جدیدی برای تراکنش ایجاد کنید که شامل کد قروشگاه و Bill_ID باشد.
d. سبد خرید را با گروه بندی بر روی ID و گروه محصول GRP بدست آورید و مجموع تعداد QTY را به ازای هر تراکنش و محصول را بدست آورید. مقادیر null را با صفر پر کنید.
e. در سبد خرید مجموع خرید بیشتر از ۱ را با ۱ و کمتر از آنرا با صفر جایگزین کنید و مقادیر null را حذف کنید.
f. با استفاده از الگوریتم apriori مجموعه frequent itemset هایی با حداقل support برابر ۰.۱ بدست آورید.
g. با معیار lift مجموعه قوانین وابستگی را از بین frequent itemset مرحل قبل بدست آورید.
h. یکی از قوانین بدست آماده را تفسیر کنید و مشخص کنید منظور از conviction و leverage چیست؟
i. قوانین با lift بیشتر از ۲ و confidence بیشتر از ۰.۳ کدام هستند؟

۲. **Advanced Association Rule** (زمان تقریبی: ۴۵ دقیقه)

a. مجموعه داده diabetes را خوانده و مقادیر null را از آن حذف کنید
b. هیستوگرام همه ستونها را رسم کنید
c. از طریق qcut مقادیر ستونهای Age و BMI را گسسته سازی کنید
d. از طریق get_dumies خصوصیات BMI و AGE و outcome را بصورت one hot تبدیل کنید و بعنوان داده اصلی استفاده کنید
e. از کتابخانه mlxend همه frequent itemset را توسط الگوریتم apriori با حداقل پشتیبانی ۰.۰۱ بدست آورید. نام ستونها را برای itemset ها در نظر بگیرید
f. Frequent itemset ها را توسط الگوریتم fpgrowth با حداقل پشتیبانی ۰.۰۱ بدست آورید.
g. قوانین بدست آمده از دو روش بالا را با یکدیگر مقابسه کنید.(confidence ۰.۵)

---

۳. **Sequence Pattern mining** (زمان تقریبی: ۴۵ دقیقه)

با توجه به فایل Sequences.csv که در اختیارتان قرار گرفته است به مجموع سوالات زیر پاسخ دهید:

a. مجموعه داده را بخوانید و هر سطر از آنرا دنباله ای از اعداد مرتبط در نظر بگیرید.
b. به کمک کتابخانه gsppy تعداد تکرار دنباله های عددی با حداقل ساپورت ۲۰ درصد از کل مجموعه داده را به دست آورید.
c. بررسی کنید که در مجموعه داده چه اعدادی قبل از دنباله ۴۴,۴۵ ظاهر شده اند؟