به نام خدا

# تکلیف چهارم 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 HW4-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.

Clustering

به کمک مجموعه داده customers که در اختیار شما قرار گرفته است به سوالات زیر پاسخ دهید:

۱. Kmeans (زمان تقریبی: ۴۵ دقیقه)

a. مجموعه داده را در یک دیتافریم ذخیره کنید و بررسی کنید که اگر مقدار null در آن وجود دارد با روشی مناسب این مقادیر را جایگزین کنید.
b. به کمک روش z score داده های پرت دیتافریم را حذف کنید.
c. دیتافریم را به کمک تابعهای آماده پایتون نرمال سازی کنید.
d. به کمک تابع PCA از کتابخانه sklearn تعداد فیچرهای دیتافریم را به ۲ فیچر کاهش داده و انرا در یک دیتافریم جدید ذخیره کنید. برای ادامه ی سوال از مجموعه داده جدید استفاده کنید.
e. برای مقادیر k از ۱ تا ۱۵, الگوریتم k-means را روی دیتافریم اجرا کنید. و نمودار SSE بر حسب k را رسم کنید. بر اساس این نمودار و روش elbow بگویید که k مناسب برای این مجموعه داده کدام است؟
f. این بار به کمک تابع KneeLocator از کتابخانه kneed مشخص کنید که k مناسب کدام است؟ آیا برداشت شما از نمودار قسمت قبل درست بوده است؟
g. بر اساس بهترین k که در قسمت قبل به دست آوردید خوشه بندی را روی داده ها انجام دهید و نتایج را به کمک scatter plot نشان داده و تحلیل کنید.
h. برای هر یک از خوشه های به دست آمده به کمک تابع describe خصوصیات انرا توصیف کره و نتایج را تحلیل کنید.

## ۲. Agglomerative Clustering (زمان تقریبی: ۴۵ دقیقه)

مشابه سوال ۱ قسمت های a تا d را تکرار کرده و به سوالات زیر پاسخ دهید:

a. Dendrogram داده ها را با روش ward رسم کنید. ایا میتوانید از روی آن تحلیل کنید که په تعداد کلاستر مناسب است؟
b. برای مقادیر k از ۲ تا ۱۰, خوشه بندی به روش Agglomerative و متد ward را انجام داده و برای هر حالت معیار silhouette را محاسبه کرده, و نمودار میله ای آن را رسم کنید و بر اساس آن بگویید که کدام مقدار k برای خوشه بندی به این روش مناسبتر است؟
c. خوشه بندی را با بهترین k که در قسمت قبل به دست آوردید انجام داده و نتیجه خوشه بندی را به کمک scatter plot رسم کنید.
d. برای هر یک از خوشه های به دست آمده به کمک تابع describe خصوصیات آن را توصیف کرده و نتایج را تحلیل کنید.
e. این روش را با روش سوال اول مقایسه و نتایج را تحلیل کنید.

## ۳. DBSCAN (زمان تقریبی: ۳۰ دقیقه)

داده های customers را خوانده و پس از نرمالایز کردن توسط tSNE در دو بعد نشان دهید. سپس از الگوریتم DBSCAN استفاده کنید و نتیجه خوشه بندی را مجددا توسط tSNE نمایش دهید. گزارش دهید تاثیر پارامتر eps بر نتیجه خوشه بندی چگونه است و بهترین نتیجه با چه مقداری از eps بدست می آید؟