

بنام خدا

تمرین اول

درس داده کاوی

- به زبان برنامه نویسی پایتون برنامه ای بنویسید که موارد خواسته شده زیر را انجام دهد.
- موارد خواسته شده را به دقت انجام دهید.
- از خروجی های بدست آمده در هر مرحله عکس بگیرید و در فایل پاسخ قرار دهید. پاسخ های خود را به طور کامل به صورت صوتی شرح دهید.
- فایل نهایی توضیحات به همراه کدها و خروجی های بدست آمده را به صورت یک فایل فشرده (rar یا zip) در لینک زیر آپلود کنید. اگر ویدیو قرار داده می شود بایستی حجم آن تا حد امکان کاهش داده شود.

<https://cloud.vru.ac.ir/s/CpHf6A9TG48zBZ7>

- به کسی که بهترین و کاملترین محتوا برای هر تمرین را تولید کند (هم از نظر صحت پاسخ و هم از نظر گرافیکی و هم ارزش آموزشی برای دیگران) نمره اضافه نیز در نظر گرفته خواهد شد.

تبدیل داده ها

۱. فایل social_network_ads.csv را که به همراه تمرین ارسال شده است را خوانده و ۱۰ رکورد داده اولیه آن را در خروجی نمایش دهید.
۲. دو ستون داده "Age" و "Estimated Salary" را انتخاب کرده مقادیر هر ستون را و به صورت نرمال استاندارد سازی نمایید و سپس ۱۰ سطر اول را در خروجی نمایش دهید. (ابتدا بدون استفاده از جعبه ابزار sklearn) و سپس با استفاده از جعبه ابزار sklearn و متدهای بخش preprocessing).
۳. مقادیر استاندارد بدست آمده از دو ستون در بخش دوم تمرین را به بازه ۰ تا ۱۰ اسکیل نمایید. (برای این کار می توانید از متد MinMaxScaler() در زیر کلاس preprocessing از کتابخانه sklearn استفاده کنید) و سپس ۱۰ سطر اول را در خروجی نمایش دهید.
۴. متد RobustScaler() در جعبه کتابخانه sklearn و زیر کلاس preprocessing وجود دارد. کارکرد آن را مشخص کنید و بر روی دو ستون بخش اول تمرین بکار ببرید و نتایج بدست آمده را توضیح دهید.

۵. در خصوص تبدیل one-hot جستجو کنید و کارکرد آن را شرح دهید. سپس با متد `LabelEncoder()` بر روی ستون Gender تبدیل one-hot را اعمال کنید. سپس مقادیر ۱۰ سطر اول را در خروجی نمایش دهید. سپس تبدیل معکوس را استفاده کنید و دوباره مقادیر Gender را به فرم اولیه برگردانید.

کاهش ابعاد

۶. فایل `Mall_Customers.csv` را که به همراه این تمرین ارسال شده است را خوانده و ۱۰ رکورد اول آن را نمایش دهید.

۷. مجدد مقادیر ستون Genre را با تبدیل one-hot به فرم عددی تبدیل کنید.

۸. با استفاده از متد PCA که در زیر کلاس `decomposition` از کتابخانه `sklearn` قرار دارد، چهار صفت `genre`، `age`، `Annual income` و `spending Score` را به سه بعد کاهش دهید.

۹. بعد از اعمال `pca` با متد `explained_variance_ratio_` توانایی و ارزش سه صفت داده (بعد) جدید بدست آمده را با هم مقایسه کنید.

موفق باشید