پروژه درس داده کاوی ترم دوم سال تحصیلی ۱۴۰۰- ۱۳۹۹

سجاد ابراهیمی ۹۷۱۲۷۶۲۴۶۵

فاز سوم پروژه: خوشه بندی

برای یک مجموعه داده که مربوط به دیتاست دیوار و دیجی کالا می باشد. موارد ذیل را برای این دیتاست ها انجام دهید.

در این فاز نیز چهار بخش اساسی باید صورت بپذیرد که به شرح ذیل می باشند:

توضیح درباره کلیات خوشه بندی صورت گرفته: در این فاز خوشه بندی ها از طریق الگوریتم Kmeans صورت گرفته است که نحوه عملکرد آن بدین صورت است که ابتدا تعداد کلاستر های مورد نظر را در ورودی از ما گرفته و با تعیین n نقطه تصادفی و سپس میانگین گیری از دیتاهای داخل هر کلاستر مراکز هر کلاستر و اعضای آن مشخص میگردند.

در اینجا ما از دو تابع find_cluster_labels(kmeans, actual_labels) برای پیدا کردن label اعضای داخل هر کلاستر و find_data_labels(X_labels, cluster_labels) برای پیدا کردن لیبل هر یک از داده ها استفاده میکنیم. سپس با مقایسه این label ها به accuracy خوشه بندی دست پیدا میکنم.

- **خوشه بندی شهرها براساس کالاهایی که در آن ها پست فروش گذاشته می‌شود(این خوشه بندی باید براساس همه ی فیلدهای مورد نیاز صورت بپذیرد)(دیوار)**

در این بخش محصولات به وسیله cat1, cat2 خود شناسایی می شوند. سپس این ستون ها را به عنوان داده ورودی و ستون city را به عنوان label در نظر گرفته و پس از انجام عمل خوشه بندی شهر هایی که از طریق خوشه بندی تعییین گردیده و شهرهای واقعی هر ایتم را مقایسه کرد و به دقت خوشه بندی دست پیدا میکنیم. به علت اینکه تعداد شهرهای موجود در دیتاست برابر ۹ بود تعداد کلاستر ورودی به الگوریتم را برابر همین عدد در نظر میگیریم.

- **خوشه بندی شهرها براساس محصولاتی که در آن ها به فروش می‌رسد.(دیجی کالا)**

در این بخش محصولات به وسیله ID_Item خود شناسایی می شوند. سپس این ستون ها را به عنوان داده ورودی و ستون city_name_fa را به عنوان label در نظر گرفته و پس از انجام عمل خوشه بندی شهر هایی که از طریق خوشه بندی تعییین گردیده و شهرهای واقعی هر ایتم را مقایسه کرد و به دقت خوشه بندی دست پیدا میکنیم. به علت اینکه تعداد شهرهای موجود در دیتاست برابر ۹۰۶ بود امکان قرار دادن این عدد به عنوان ورودی های کلاستر وجود نداشت به همین دلیل تعداد ۲۰ برای تعداد خوشه های ورودی به الگوریتم در نظر گرفته شد.

- **مقایسه خوشه بندی ها صورت گرفته در دو فاز قبلی**

با توجه به اینکه در هر دو دیتاست تعداد فروش آیتم ها در شهر تهران بسیار بالاتر از شهر های دیگر بوده و هنگام توزیع این آیتم ها در کلاستر های مختلف نیز تعداد آیتم های مربوط به تهران با اختلاف از دیگر شهرها بالاتر بوده و نحوه تعیین label هر کلاستر نیز با استفاده از پر تکرار ترین شهری که در هر خوشه وجود داشته تعیین میگردد، label همه کلاستر ها در هر دو حالت برابر شهر تهران تعیین میگردد که در هنگام محاسبه accuracy به دقت های ۴۶٪ و ۵۴٪ برمیخوریم که نشان دهنده این است که این میزان از آیتم های هر دیتاست مربوط به شهر تهران بوده اند.

- **خوشه بندی محصولات براساس قیمت آن‌ها(دیوار)**

در این بخش محصولات به وسیله cat1, cat2 خود شناسایی می شوند. سپس این ستون ها را به عنوان داده ورودی و ستون price را به عنوان label در نظر گرفته و پس از انجام عمل خوشه بندی قیمتی که برای هر کلاستر در نظر گرفته شده را با قیمت واقعی محصول با اختلاف ۲۰ درصد بررسی میکنیم.