فاز دوم پروژه

درس مبانی داده کاوی

نیمسال دوم تحصیلی 1403

Clustering & Classification

در این فاز می خواهیم با استفاده از Classification و Clustering به کشف الگو های پنهان دیتاست بپردازیم.

خوشه بندی(Clustering)

1. الگوریتم های خوشه بندی

- داده های خود را بر اساس برخی از ویژگی های عددی و categorical خوشه بندی کنید. توجه کنید که فیچرهایی که انتخاب میکنید باید در کنار هم معنا داشته باشند. در واقع باید فیلم های مشابه در یک خوشه قرار بگیرند. با الگوهای موجود در کلاسترها، شباهت فیلم های هر کلاستر را در داکیومنت خود بررسی کنید.
- یک بار هم خوشه بندی را بر اساس ترکیب سال و فصل فیلم ها انجام دهید. با این کار، الگوهایی از هر کلاستر در بیارید و بررسی کنید در هر فصل از سال چه سبک فیلم هایی تولید شده اند و چه ویژگی هایی داشته اند.
- خوشه بندی را با هر دو الگوریتم Kmeans و DBSCAN انجام دهید.
- برای Kmeans، Hyperparameter tuning انجام دهید و برای پارامترهای DBSCAN حداقل سه حالت را تست کنید.

- برای کلاسترینگ خود یک معیار ارزیابی پیدا کنید و میزان خوبی آن را گزارش دهید.
- الگوهایی که در خوشه‌ها شناخته می‌شوند را مورد بررسی قرار دهید.

2. تصویرسازی الگوهای خوشه‌بندی(Visualization)

- از نمودارها و تصویرسازی‌ها برای نمایش الگوهای خوشه‌بندی بهره ببرید. این تصاویر را در داکیومنت خود بررسی کنید.

دسته بندی(Classification)

فاز اول:

1. ابتدا ویژگی‌هایی را انتخاب کنید که مربوط به محبوبیت و موفقیت فیلم باشد(داده های شما یا همان **X**). ستونی تحت عنوان **is_popular** ایجاد کنید(اگر از فازهای قبل دارید از همان استفاده کنید) و این ستون را به عنوان لیبل در نظر بگیرید. در واقع به دنبال شکست خوردن یا موفقیت هستیم. بر اساس فیچرهای انتخابی خود به دنبال دلایل شکست فیلم ها باشید و در داکیومنت خود این موارد را بررسی کنید.
2. داده های خود را به دو مجموعه **train** و **test** تقسیم کنید.
3. چند مدل از جمله **Decision Tree** و **SVM** و **Naïve Bayes** را بر روی داده های خود آموزش داده و دقت **train** آن را محاسبه کنید.(سه مدل ذکر شده حتما باشند)
4. برای مدل های خود **hyperparameter tuning** انجام دهید تا به بهترین دقت برسید و از **overfit** شدن مدل های خود جلوگیری کنید.
5. در نهایت دقت مدل های خود بر روی داده های تست به دست آورید. همچنین معیار های **Precision** و **Recall** و **F1-Score** را برای داده های تست محاسبه کنید.
6. در این بخش به محبوب بودن یک فیلم پرداختید؛ دلایل عدم موفقیت بعضی فیلم ها را با توجه به ویژگی هایی که انتخاب کردید؛ در داکیومنت خود بررسی کنید.

فاز دوم:

در این بخش ویژگی هایی را انتخاب کنید که مربوط بتواند **Vote Average** فیلم ها را پیش بینی کنند. توجه کنید این ستون باید گسسته سازی شود و به ستونی **Categorical** تبدیل شود.
تمام مراحل بخش قبل را صرفا با ویژگی های متفاوتی و لیبل **Vote Avergae** که کتگوریکال شده انجام دهید.
در واقع در این بخش با انتخاب فیچرهای مناسبی؛ قرار است امتیاز فیلم ها را پیش بینی کنید.

نکات تحویل:

- تمام بخش ها و نتایج خود را در داکیومنت تحلیل و بررسی کنید.(**کد خود را توضیح ندهید!**)
- پروژه در قالب گروه های حداکثر دو نفری پیاده شود.
- فایل نهایی به فرمت zip و با نام studentName1-studentName2-phase2.zip ارسال شوند.

موفق باشید