پروژه چهارم درس داده کاوی

عنوان :

رتبه‌بندی ویژگی‌های مؤثر در دسته‌بندی مجموعه داده‌ها

استاد:

دکتر حسین رحمانی

زمستان ۱۴۰۲

راهنمای پروژه

- مهلت ارسال پروژه تا **ساعت ۲۳:۵۹ تاریخ ۱۳ /۱۱ / ۱۴۰۲** است و قابل تمدید نخواهد بود.
- **حداکثر یک روز با تاخیر می‌توانید ارسال کنید و** از نمره پروژه، ۲۵٪ کسر خواهد شد.
- پاسخ به سوالات این پروژه باید در قالب یک گزارش با **فرمت PDF** یا به همراه **توضیحات فایل نوتبوک (Markdown)** ارائه شود.
- در صورت ارائه گزارش در قالب **توضیحات فایل نوتبوک**، توضیحات باید کامل، جامع و شفاف باشد.
- در صورت ارائه گزارش با **فرمت PDF**، فایل **کدهای اجراشده** نیز پیوست شود.
- تمامی فایل‌های این پروژه (گزارش و کدها) در قالب یک **فایل فشرده rar یا zip** با نام‌گذاری زیر ارسال شود .

StudentNumber_FirstName_LastName_Prj04.zip

- فایل تمرین را **حتما** در سامانه LMS آپلود نمایید. بدیهی است که تحویل از طریق ایمیل و یا سایر راه‌های ارتباطی قابل پذیرش نخواهد بود.
- **رعایت نکات نگارشی** در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این پروژه **حتما** باید از زبان برنامه‌نویسی **پایتون** استفاده شود.

مباحث تحت پوشش: Unsupervised Learning, Supervised Learning

هدف پروژه: می‌خواهیم با توجه به مجموعه داده برچسب گذاری شده، ویژگی‌های موجود را براساس میزان تاثیرگذاری، رتبه‌بندی کنیم.

مراحل انجام پروژه به ترتیب در ادامه آمده است:

۱- فایل ورودی

مجموعه داده متشکل از ۸۳ ستون می‌باشد که شامل ستون Index و ۸۱ ویژگی مربوط به هر رکورد Feature1, Feature2,…., Feature81 و یک ستون حاوی برچسب (Class Label) هر رکورد است.

۲- پیش‌پردازش‌داده

در این بخش فقط نیاز است که **نرمال سازی** مقادیر موجود صورت بگیرد.

۳- تحلیل ویژگی

تجزیه و تحلیل همبستگی (Correlation) را برای شناسایی روابط بین ویژگی‌ها در مجموعه داده انجام دهید. ضریب همبستگی بین هر جفت ویژگی را برای درک وابستگی آن‌ها محاسبه کنید. سپس برای نمایش این همبستگی‌ها بین جفت ویژگی از روش نمایشHeat Map استفاده نمایید. هدف از این مرحله یک تحلیل کلی از وابستگی بین ویژگی‌ها می‌باشد.

۴- یادگیری Unsupervised

حال می‌خواهیم ویژگی‌ها را بر اساس میزان تاثیرشان بر روی برچسب‌گذاری ها رتبه‌بندی کنیم. مثلا اگر ویژگی x تاثیر زیادی بر روی برچسب ۱ رکوردها داشته باشد، رتبه‌ی ویژگی x بالا خواهد بود و بالعکس.

الف) می‌خواهیم با روش خوشه بندی(K-means) ویژگی‌ها را با توجه به رویکرد خودتان، به صورت نزولی براساس تاثیرگذاری، رتبه بندی کنید.

ب) حال رتبه بندی را این بار در هر برچسب‌گذاری، جداگانه انجام دهید. یعنی ابتدا رکوردها را بر اساس برچسبشان جدا کرده، و سپس در هر دسته، رویکرد رتبه‌بندی خود به وسیله خوشه بندی K-means را بر روی ویژگی‌ها اعمال کنید.

ج) الگوریتم Chameleon را که نوعی روش خوشه‌بندی سلسله مراتبی است بر روی کل مجموعه‌داده اعمال کنید. سپس در تصویر گراف نمایش داده شده، برای هر گره برچسبش(Label) را (که مقادیر ۰ یا ۱ دارد) نیز در تصویر نشان دهید و تصویر گراف و خوشه‌های ایجاد شده حاصل از آن را تحلیل کنید.

در این مرحله واضح است که در پایان، سه لیست رتبه‌بندی شده از هر قسمت به عنوان خروجی نمایش داده شوند.

۵- یادگیری Supervised

در ابتدا به بررسی متعادل بودن مجموعه داده بپردازید و اگر داده‌ها نامتعادل (Imbalanced) بود، روش‌های متعادل کردن داده‌ها به جهت جلوگیری از Overfitting را انجام دهید. حال با توجه به ویژگی Label که به عنوان ویژگی طبقه‌بندی استفاده می‌شود، از دو الگوریتم Random Forest و یک روش انتخابی دیگر، استفاده کنید تا داده‌ها طبقه‌بندی شوند. سپس نتایج الگوریتم Random Forest را نمایش دهید و هم‌چنین ویژگی‌های به دست آمده از Random Forest را براساس میزان تاثیرگذاری‌شان رتبه بندی کنید.

۶- ارزیابی

در ارزیابی مدل‌های آموزش داده شده از معیارهایی همچون Accuracy , F-score , Confusion Matrix , Recall , AUC-ROC و Distortion Score , Trustworthiness , Silhouette Coefficient (معیارهای دیگری که مختص به روش خاصی هستند نیز می‌توانید استفاده کنید همچون Calinski-Harabasz Index و Davies-Bouldin Index و RMSR و ...) را که برای هر الگوریتم (چه Supervised و چه Unsupervised) مناسب است، به دست آورید.

در پایان نیز هر معیاری را که محاسبه کرده‌اید، نیاز است که در جدولی همانند جدولی که در ادامه آمده است، در یک فایل xlsx و یا csv با نام metrics ذخیره نمایید.

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
Algorithm 1				
Algorithm 2				
Algorithm 3				
Algorithm 4				

۷- نتیجه‌گیری

الف) آیا ویژگی‌هایی وجود دارند که در هر دو نوع یادگیری Supervised و Unsupervised از نظر تاثیرگذاری همپوشانی داشته باشند؟

ب) به نظر شما کدام یک از رتبه‌بندی‌ها منطقی‌تر است؟ دلیل خود را شرح دهید.

ج) اگر اختلاف زیادی در معیارهای ارزیابی روش‌ها مشاهده شد، دلیل آن را بنویسید.

نکته: هرگونه ابتکار عمل، خلاقیت، تحلیل‌های دیگر و استفاده از روش‌های بیشتر نمره‌ی اضافه خواهد داشت.

موفق باشید.