

به نام خدا

داده کاوی

نیمسال اول ۱۴۰۴-۱۴۰۵



مدرس: فاطمه کریمی

مهلت تحويل: ۱۴۰۳/۱۰/۲۷

تعريف پروژه

نکته: انجام پروژه به صورت گروههای جداکثراً دو نفری قابل قبول است.

هدف این پروژه، آشنایی عملی با فرآیند کامل داده کاوی از مرحله انتخاب داده تا تحلیل و تفسیر نتایج است. در این پروژه دانشجو با

مفاهیم زیر به صورت عملی کار خواهد کرد:

- تحلیل اکتشافی داده‌ها (EDA)
- پیش‌پردازش و تمیز کاری داده
- ساخت و ارزیابی مدل‌های طبقه‌بندی Ensemble
- استفاده از روش‌های خوش‌بندی داده‌ها و ارزیابی نتایج
- مقایسه نتایج با و بدون اعتبارسنجی (Cross-Validation)

استفاده از ابزارهایی مانند Weka, Python (scikit-learn), RapidMiner آزاد است و انتخاب ابزار بر عهده دانشجو می‌باشد.

✓ هر دانشجو باید بر اساس شماره دانشجویی خود یک دیتابست از مخزن [UCI](#) انتخاب کند:

نکته مهم: چنانچه پروژه در گروه ۲ نفری ارائه می‌شود برای انتخاب دیتابست و تنظیمات پروژه به توضیحات صفحه آخر مراجعه کنید.

دیتابست	رقم سمت راست شماره دانشجویی
Iris	0, 1, 4
Badges	2, 3
Glass Identification	7, 9
Ionosphere	5, 6, 8

در ادامه گزارش باید سه سؤال زیر را پاسخ دهید و جداول خواسته شده را پر کنید. برای گرفتن نمره کامل باید برای هر سؤال، گزارش انجام کار را تهیه کنید: روند انجام تمرین در ابزار، تصاویر انجام مراحل مختلف، نحوه محاسبه مقادیر، تنظیمات و پارامترهای هر روش، فرضی که دانشجو برای انجام الگوریتم داشته است، هر توضیحی که برای فهم روش نیاز است.

۱. تأثیر Decision Tree بر کارایی Cross Validation 10-folds

بر اساس مجموعه داده مربوط به خود، جدول زیر را کامل کنید.

		TP Rate	FP Rate	Precision	Recall	F-measure
1	Decision tree with use training set					
2	Decision tree with Cross-validation					

۲. بررسی روش‌های ترکیبی (Ensemble)

		TP Rate	FP Rate	Precision	Recall	F-measure
1	Ada boost					
2	Random Forest					

۳. خوشه‌بندی (Clustering)

داده‌های خود را بدون برچسب کلاس در نظر بگیرید و آن‌ها را خوشه‌بندی کنید. حال بر اساس برچسب موجود کلاس، درصد نمونه‌های اشتباه را به ازای مقادیر مختلف تعداد خوشه ۲، ۳، ۴ و ۵ در جدول زیر گزارش کنید.

Method \ K	2	3	4	5
K-Means				
K-Medoids				

✓ فایل‌های ارسالی

- شما می‌بایست نحوه انجام تمامی مراحل فوق به همراه تمامی نمودارها و تحلیل‌های صورت گرفته را در قالب یک گزارش علمی ارائه دهید.
- رعایت اصول نگارش یک گزارش علمی (متن ساده و روان، استفاده از فونت‌های استاندارد، بخش‌بندی، استفاده از عنوان برای تصاویر و جداول و ...) در نمره نهایی تأثیرگذار می‌باشد.
- به همراه فایل گزارش، می‌بایست کدهای مربوطه نیز در قالب یک فایل فشرده در موعد مقرر در سامانه LMS بارگذاری شود.

قیدهای شخصی‌سازی برای گروههای ۲ نفره

- قاعده کلی

در پروژه‌های دو نفره، گروه‌ها موظفاند:

۱. یک نفر را به عنوان سرگروه انتخاب کنند.
۲. دیتاست اصلی را بر اساس شماره دانشجویی سرگروه انتخاب کنند (طبق جدول پروژه)
۳. قید شخصی‌سازی را بر اساس شماره دانشجویی عضو دوم گروه اعمال کنند. رقم سمت راست شماره دانشجویی عضو دوم گروه را برابر d در نظر بگیرید.

قید شخصی‌سازی در هر بخش پروژه

◆ بخش ۱: درخت تصمیم

یکی از موارد زیر را بر اساس مقدار d و حاصل $d \bmod 3$ اعمال کنید:

$d \bmod 3$	قید اجباری
0	$\text{max_depth} = d + 2$
1	$d+1$
2	حداقل تعداد نمونه در هر برگ = هرس درخت (Pruning) و گزارش اثر آن

نکته: تنظیمات انتخاب شده باید در گزارش توضیح داده شود.

◆ بخش ۲ Cross Validation :

- اگر d زوج است \leftarrow 10-fold CV
- اگر d فرد است \leftarrow 5-fold CV