

۷	روانشناسی
۸	روانشناسی، مذهبی، رمان
۹	روانشناسی، مذهبی، تاریخی
۱۰	روانشناسی، مذهبی

الف) با استفاده از روش FP-Growth تمامی frequent itemset هایی که به کتاب نوع رمان ختم می شود را بیابید. (support = ۲۰٪).

ب) با در نظر گرفتن confidence = ۵۰٪ قواعد معتبر قابل استخراج از frequent itemset های به دست آمده از قسمت الف را بیابید.

پیاده سازی:

سوال ۸:

در این بخش هدف استفاده از جنگل تصادفی برای کلاس بندی می باشد. مسئله تایتانیک (سوال ۶ در تمرین ۲) را این بار با به جای درخت تصمیم با استفاده از جنگل تصادفی پیاده سازی کرده و دقت پیش بینی را اندازه گیری کنید (۲ حالت برای عمق درخت ها و ۲ حالت برای تابع تقسیم gini یا entropy در مجموع ۴ حالت مختلف). سپس مقایسه های زیر را انجام دهید:

۱- بالاترین دقت به دست آمده توسط مدل جنگل تصادفی را با مدل درخت تصمیم در تمرین قبل مقایسه کنید.

۲- سرعت یادگیری و تست جنگل تصادفی و درخت تصمیم را با هم مقایسه کنید.

توجه ۱- همانطور که در تمرین قبل اطلاع رسانی شد، برای محاسبه دقت پیش بینی روی مجموعه داده تایتانیک یکی از دو روش زیر را می توانید انتخاب کنید:

۱. تقسیم بندی داده های فایل train.csv به دو قسمت train set و test set (با نسبت ۸۰ به ۲۰) و

محاسبه دقت مدل با استفاده از test set

۲. ثبت نام در سایت [kaggle.com](https://www.kaggle.com) و بارگذاری نتایج پیش‌بینی شده برای فایل `test.csv` و دریافت امتیاز (که همان دقت پیش‌بینی شما می‌باشد) از سایت [kaggle](https://www.kaggle.com). نحوه بارگذاری و فرمت فایل ارسالی برای سایت در این [لینک](#) توضیح داده شده است.

روش انتخابی شما برای محاسبه دقت باید در تمرین دوم و سوم یکسان باشد تا مقایسه بین درخت تصمیم و جنگل تصادفی به درستی صورت بگیرد.

توجه ۲- نیازی به مصورسازی جنگل تصادفی نیست.

سوال ۹:

در این قسمت هدف آشنایی بیشتر و کلاس بندی با SVM (Support Vector Machine) می‌باشد. مجموعه داده مورد استفاده در اینجا همان مجموعه داده تایتانیک و هدف پیش‌بینی زنده ماندن یا نماندن مسافر خواهد بود. برای پیاده‌سازی این قسمت می‌توانید از کتابخانه‌های موجود استفاده کنید. مراحل زیر را انجام داده و نتایج خواسته شده را در گزارش خود بیاورید:

۱- پیش‌پردازش مناسب روی مجموعه داده انجام دهید. دقت کنید که نحوه ارزیابی مدل به یکی از دو روش ذکر شده در قسمت قبل خواهد بود (تقسیم داده‌های `train.csv` به دو قسمت `train set` و `test set` و یا استفاده از سایت [kaggle](https://www.kaggle.com))، لذا در صورت استفاده از روش اول داده‌های `train.csv` را در این مرحله به هم زده و تقسیم کنید.

۲- مدل خود را با SVM با هسته خطی (Linear kernel function) آموزش داده و دقت آن را محاسبه کنید.

۳- دقت به دست آمده در بخش قبل را با دقت درخت تصمیم (در تمرین قبل) و جنگل تصادفی مقایسه کنید. علت پایین‌تر (یا بالاتر) بودن دقت را از نظر خود بیان کنید.

۴- مدل خود را با SVM با یک هسته غیر خطی (به انتخاب خودتان) آموزش داده و دقت آن را به دست آورید.

۵- دقت پیش‌بینی در بخش ۲ و ۴ را با هم مقایسه کرده و علت تغییر آن را ذکر کنید.