دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

تمرین سری دوم

مبانی داده کاوی

بهار ۱۴۰۴

استاد درس: دکتر حمیدرضا حکیم

دستیار آموزشی: فرزانه کوهستانی

موعد تحویل تکلیف: ۸ اردیبهشت

۱ بخش تئوری

سوالات ۳، ۵، ۸ و ۹ از تمرینهای فصل سوم کتاب مرجع را به صورت کامل حل کنید.

۲ بخش عملی

سوالات زیر را در KNIME پیاده سازی کنید.

۱.۲ کلاس‌بندی مجموعه داده کارتهای اعتباری

۱.۱.۲ مجموعه داده‌ای با عنوان **credit cards.csv** در اختیار شما قرار دارد. این مجموعه شامل اطلاعات مربوط به سوابق پرداخت مشتریان بانک در استفاده از کارت‌های اعتباری است. هدف از استفاده از این داده‌ها، پیش‌بینی وضعیت پرداخت مشتریان است تا مشخص شود آیا مشتری در پرداخت بدهی خود دچار مشکل (وضعیت غیرعادی) خواهد شد یا خیر. جزئیات بیشتر درباره این مجموعه داده، در فایل ضمیمه با نام **creditCardReadMe.txt** آورده شده است.

۱. با استفاده از درخت تصمیم وضعیت کارتهای اعتباری را پیش بینی کنید.

۲. چگونه می‌توانید با استفاده از Cross-validation در KNIME عملکرد این مدل دسته‌بندی را ارزیابی کنید؟ مراحل را به صورت عملی نشان دهید.

۳. نحوه بهینه‌سازی تقسیم‌بندی داده‌ها (Splitting Criterion) در مدل‌های درخت تصمیم در KNIME را بررسی کنید و بهترین معیار تقسیم (... ،Entrop ،Gini) را تعیین کنید.

۲.۲ کلاس‌بندی مجموعه داده بیماران قبلی

۱.۲.۲ مجموعه داده‌ای با نام **heart diagnose.csv** در اختیار شما قرار گرفته است که حاوی اطلاعات مربوط به بیماران قلبی (مانند سن، جنسیت و سایر مشخصه‌های فردی) و همچنین داده‌های پزشکی این بیماران (مانند نوع درد و سایر علائم بالینی) است. هدف اصلی از این مجموعه داده، پیش‌بینی و تشخیص بیماری قلبی در بیماران است. اطلاعات تکمیلی درباره این مجموعه داده را می‌توانید در فایل ضمیمه با عنوان **heartDiagnoseReadMe.txt** مشاهده نمایید.

۱. با کمک KNIME نشان دهید که چگونه افزایش پیچیدگی درخت تصمیم باعث بروز پدیده‌ی Overfitting می‌شود. این موضوع را با استفاده از نمودار و داده‌های نمونه تفسیر کنید.

۲. یک مثال عملی در KNIME ارائه دهید تا تأثیر هرس کردن درخت تصمیم (Pruning) را بر کاهش Overfitting به وضوح نشان دهد.

۳. در KNIME تأثیر تغییر اندازه‌ی داده‌های آموزشی را بر خطای تعمیم (Generalization Error) بررسی کرده و نتیجه را با رسم نمودار تفسیر کنید.

۴. چگونه می‌توانید در KNIME با ایجاد یک مجموعه اعتبارسنجی (Validation Set) بهترین مدل را انتخاب کنید؟ مراحل را به صورت عملی و با استفاده از داده‌های نمونه نشان دهید.

۵. با استفاده از روش‌های ارزیابی مدل در KNIME مانند Holdout و Cross-validation بهترین مدل درخت تصمیم را از نظر پیچیدگی و دقت تعمیم انتخاب کرده و دلایل خود را شرح دهید.

روش تحویل

۱. برای سوالات عملی پیاده‌سازی باید در نرم‌افزار KNIME انجام شود. فایل‌های workflow هر سوال باید ضمیمه گردد.

۲. فایل گزارش جامع باید در قالب pdf ارائه شود. در این گزارش باید به طور کامل پیاده‌سازی، تحلیل‌ها و نتایج حاصل از هر سوال ذکر شود.

۳. از ارسال پاسخها به صورت دست نویس اکیدا خودداری نمائید.

۴. در اسم فایل ارسالی، نام و شماره دانشجویی به صورت Lastname-StudentCode نوشته شده باشد.