دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

تمرین سری اول
مبانی داده کاوی
زمستان ۱۴۰۳

استاد درس: دکتر حمیدرضا حکیم
دستیار آموزشی: فرزانه کوهستانی
موعد تحویل تکلیف: ۱۴ اسفند ماه

۱ تحلیل و درک مفاهیم داده

۱.۱ انتزاع و تعمیم داده‌ها

فرض کنید شما مسئول طراحی یک سیستم داده‌کاوی برای پیش‌بینی احتمال ترک تحصیل دانشجویان هستید.

۱. کدام نوع از ویژگی‌ها (Nominal، Ordinal، Interval، Ratio) را برای مدل‌سازی این مسئله انتخاب می‌کنید؟ چرا؟

۲. آیا همه ویژگی‌ها دارای اهمیت یکسانی هستند؟ چگونه می‌توان اهمیت ویژگی‌ها را سنجید؟

۲.۱ اثر نوع ویژگی بر تحلیل داده‌ها

فرض کنید دو ویژگی "درآمد سالانه" و "کدپستی" را در یک مجموعه داده در اختیار دارید.

۱. آیا این دو ویژگی را می‌توان در تحلیل‌های آماری به یک صورت در نظر گرفت؟ چرا؟

۲. اگر قرار باشد داده‌های این دو ویژگی را نرمال‌سازی کنید، چه تفاوت‌هایی در روش نرمال‌سازی آنها وجود دارد؟

۳.۱ نقد معیارهای توصیفی داده‌ها

فرض کنید یک مجموعه داده دارید که میانگین و میانه آن برابر هستند، اما واریانس آن بسیار بالا است.

۱. این چه چیزی درباره توزیع داده‌ها به شما می‌گوید؟

۲. آیا در چنین شرایطی میانگین معیار مناسبی برای توصیف داده‌ها است؟ اگر نه، چه معیار دیگری پیشنهاد می‌کنید؟

توجه: سوالات ۲ و ۳ باید در نرم‌افزار KNIME پیاده‌سازی شوند.

۲ تحلیل مجموعه داده تایتانیک

در مجموعه داده Titanic Kaggle، اطلاعات مربوط به تعدادی از مسافران کشتی تایتانیک وجود دارد. این کشتی پس از برخورد به کوه یخ غرق شد و تعداد از مسافران آن نجات پیدا کردند. این مجموعه اطلاعات مسافران از جمله ID، سن، تعداد فرزند، و ... را شامل می‌شود. وجود داده‌ها در فایل TitanicReadMe توضیح داده شده است. با استفاده از فایل train.csv، تحلیل‌های زیر را انجام دهید:

۱.۲ بصری‌سازی

۱. اطلاعات مربوط به ویژگی‌ها و داده‌ها را نمایش دهید. داده‌ها شامل میانگین، واریانس، حداقل و حداکثر مقدار هر ویژگی را بررسی کنید.

۲. Boxplot همه ویژگیها را داخل یک شکل نشان دهید.

۳. همبستگی بین ویژگیها را با نمودار heatmap بررسی کنید و مشخص کنید کدام دو ویژگی بیشترین همبستگی را با یکدیگر دارند.

۴. داده‌های missing را توسط heatmap نمایش دهید. سطرها passengerID و ستون‌ها featuresها که دارای مقدار missing هستند را مشخص کنید. بیشترین مقدار missing در کدام ستون قرار دارد؟

۵. یک متغیر جدید به نام alone ایجاد کنید که مشخص کند آیا مسافر وابستگی به خانواده داشته است یا خیر.

۶. نمودار هیستوگرام را برای سن ایجاد کنید و آنچه را که از نمودار متوجه می‌شوید شرح دهید.

۷. مقادیر missing در خصوصیت age را به روش mode پر کنید.

۸. روش‌های مختلف مقداردهی به مقادیر missing در خصوصیت cabin چه روشی بهتر است انجام شود؟

۲.۲ نقد محدودیت‌های روش‌های بصری‌سازی

یک مجموعه داده دارای ۵۰ ویژگی (Attributes) و ۱۰۰۰۰ نمونه (Instances) است.

۱. آیا می‌توان این داده‌ها را با روش‌های سنتی بصری‌سازی (مانند هیستوگرام یا Plot Scatter) تحلیل کرد؟ چرا؟

۲. چه پیشنهاداتی برای بصری‌سازی چنین داده‌هایی پیشنهاد می‌کنید؟

۳ تحلیل فاصله‌ها در مجموعه داده

مجموعه داده‌ای شامل ۳ ویژگی (x1، x2، x3) برای چهار نقطه داده به شرح زیر دارید:

نقطه	x1	x2	x3
نقطه ۱	2	3	4
نقطه ۲	1	5	6
نقطه ۳	4	6	7
نقطه ۴	3	2	8

۱.۳ الف) فاصله‌ها را با استفاده از معیارهای زیر محاسبه کنید:

۱. Distance Euclidean

۲. Distance Minkowski با r=3

۳. Distance Mahalanobis (برای سادگی، فرض کنید ماتریس کوواریانس برابر با ماتریس همبستگی باشد و همه ویژگی‌ها همبسته هستند)

۲.۳ ب) نتایج فاصله‌های محاسبه شده را در جدول زیر نمایش دهید و بررسی کنید که کدام معیار فاصله بیشتر شبیه به تفاوت بین نقاط را نشان می‌دهد:

از نقطه	به نقطه ۱	به نقطه ۲	به نقطه ۳	به نقطه ۴
نقطه ۱	فاصله	فاصله	فاصله	فاصله
نقطه ۲	فاصله	فاصله	فاصله	فاصله
نقطه ۳	فاصله	فاصله	فاصله	فاصله
نقطه ۴	فاصله	فاصله	فاصله	فاصله

۳.۳ با توجه به چهار زمینه داده‌ای بالا:

۱. در داده‌های پراکنده و گسسته، کدام معیار فاصله (Euclidean، Minkowski یا Mahalanobis) بهترین عملکرد را دارد؟

۲. در داده‌های همبسته و وابسته، کدام معیار فاصله می‌تواند عملکرد دقیق‌تری ارائه دهد؟

۳. در داده‌های با مقیاس‌های مختلف، کدام معیار فاصله نیاز به پیش‌پردازش (مانند نرمال‌سازی یا مقیاس‌بندی) دارد؟

۴. در داده‌های با توزیع‌های غیرنرمال، کدام معیار فاصله مناسب‌تر است؟

روش تحویل

۱. برای سوالات ۲ و ۳ 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. در اسم فایل ارسالی، نام و شماره دانشجویی به صورت Lastname-StudentCode نوشته شده باشد.