به نام خدا

تمرین چهارم داده‌کاوی

خرداد ۱۴۰۴

دوره	استاد	دانشگاه	TA
کارشناسی مهندسی کامپیوتر	دکتر منصوره اژه‌ای	دانشگاه اصفهان	گلکارنور (@adgolkar)

به مجموعه سوالات زیر پاسخ دهید. استفاده از هوش مصنوعی در این تمرین **مجاز نمی‌باشد.** مطمئن باشید اگر با هوش مصنوعی بصورت کپی-پیس بنویسید نمره این تمرین را از دست خواهید داد و بنده جوابگو نخواهم بود. جواب‌ها باید به نحوی باشد که ثابت کنید مباحث را درک کرده‌اید.

۱. تحلیل ارتباطات (Association Analysis)

۱. انگیزه اصلی پشت تحلیل ارتباطات (Association Analysis) چیست و این روش چگونه به وجود آمد؟

۲. محدودیت‌های استفاده از همبستگی (correlation) مانند ضریب پیرسون (Pearson's coefficient) برای یافتن ارتباطات در داده‌های تراکنشی چیست؟

۳. یک قانون ارتباطی مانند $X \rightarrow Y$ چیست؟ و چطور باید قانونی مانند $\{Milk, Diaper\} \rightarrow \{Beer\}$ را تفسیر کرد؟

۴. پشتیبانی (support) و اطمینان (confidence) چه هستند و چگونه قوانین ارتباطی را ارزیابی می‌کنند؟ فرمول هر کدام را بنویسید و درمورد تفاوت‌هایشان توضیح دهید.

۵. اصل آپریوری (Apriori Principle) چیست؟ و مراحل الگوریتم آپریوری (Apriori algorithm) برای تولید مجموعه اقلام پرتکرار (frequent itemsets) را طبق مباحث تدریس شده در کلاس بنویسید.

---

۲. خوشه‌بندی (Clustering)

۱. منظور از "شباهت زیاد درون خوشه‌ای" (high intra-cluster similarity) و "شباهت کم بین خوشه‌ای" (low inter-cluster similarity) چیست؟ یک مثال بزنید.

۲. سه خانواده اصلی الگوریتم‌های خوشه‌بندی (clustering algorithms) را نام ببرید (آن‌هایی که در اسلایدها آمده‌اند) و برای هرکدام یک توضیح یک‌خطی بنویسید و یک نمونه الگوریتم نیز مثال بیاورید.

۳. مراحل محاسبات الگوریتم K-Means را به‌صورت گام‌به‌گام با یک مثال ساده (مثلا یک مثال سه خوشه‌ای) توضیح دهید.

۴. مجموع خطای مربعات (Sum of Squared Error یا SSE) در K-Means چیست و چگونه استفاده می‌شود؟