دانشگاه صنعتی اصفهان

دانشکده برق و کامپیوتر

# مبانی داده‌کاوی

## سوالات مباحث پیش‌پردازش (تکلیف سری 6)

بهار 1403

فهرست مطالب

**۱- سوالات** ..... 3

1- ..... 3

2- ..... 3

3- ..... 3

4- ..... 4

**۲- نکات پاسخ‌دهی** ..... 5

# ۱- سوالات

1- با استفاده از الگوریتم Apriori و جدول زیر مجموعه آیتم‌های رایج را بیابید، از آنها قوانین ارتباطی را استخراج کنید و برای هر کدام معیار confidence را محاسبه کنید (minsup = 3).

| ID | آیتم |
|---|---|
| $T_1$ | شکر، شیر |
| $T_2$ | شیر، نان، کره |
| $T_3$ | شکر، کره |
| $T_4$ | شیر، نان، شکر |
| $T_5$ | شیر، نان، کره |
| $T_6$ | نان، شکر |
| $T_7$ | نان، شیر |
| $T_8$ | شیر، کره |
| $T_9$ | شیر، نان، کره |
| $T_{10}$ | شیر، شکر، کره |
| $T_{11}$ | نان، کره |

2- اعداد زیر را در نظر بگیرید و الف) با استفاده از روش equal frequency این داده‌ها را سبد بندی کنید.

ب) با mean, median وboundary داده‌ها را smooth نمایید.

3, 9, 11, 14, 20, 30, 32, 36, 45

3- با توجه به دیتاست Cereals.csv به سوالات زیر پاسخ دهید.

الف) کدامیک زا متغیرها عددی، کدامیک کیفی ترتیبی و کدامیک کیفی اسمی هستند؟

ب) مقادیر میانگین، میانه، کمینه، بیشینه و انحراف استاندارد را برای متغیرهای عددی محاسبه نمایید.

پ) برای هر یک از متغیرهای عددی یک هیستوگرام رسم نمایید. براساس هیستوگرام های رسم شده به پرسش‌های زیر پاسخ دهید:
1- کدام متغیر(ها) دارای بالاترین مقدار تغییرپذیری (Variability) است؟

2- کدام متغیر(ها) دارای چولگی است؟

ت) با رسم نمودارهای جعبه‌ای در کنار یکدیگر، کالری‌های غلات سرد و گرم را مقایسه کنید. این نمودار چه چیزی را نشان می‌دهد؟
ث) یک نمودار جعبه‌ای از رتبه‌بندی مصرف‌کننده به عنوان تابعی از ارتفاع قفسه رسم کنید. اگر مایل به پیش‌بینی مصرف‌کننده از ارتفاع قفسه بودید، آیا لازم بود تا تمای سه طبقه مربوط به ارتفاع را نگهداری کنید؟

ج) جدول همبستگی متغیرهای عددی را محاسبه کنید و سپس با استفاده از کتابخانه‌ی seaborn یک نمودار ماتریسی برای این متغیرها تولید کنید. به سوالات زیر پاسخ دهید:

1- کدام زوج(ها) بیشترین مقدار همبستگی را دارند؟

2- چگونه می‌توان براسا این همبستگی‌ها تعداد متغیرها را کاهش داد؟

3- اگر در ابتدا داده‌ها نرمال سازی شوند آنگاه چه تغییراتی در همبستگی‌ها دیده می‌شود؟

4- با توجه به دیتاست diabetes.cvs مراحل زیر را انجام دهید.
الف) دیتاست را بخوانید و گزارشی از آن بگیرید.
ب) مقادیر گمشده‌ی هر ستون را با استفاده از ()isnull چاپ کنید.

پ) با استفاده از تابع describe در pandas مقادیر مینیمم را برای هر ستون بدست بیاورید. با بررسی این مقادیر برای ستون‌هایی مانند فشار خون، انسولین و ... چه نتیجه‌ای می‌توان گرفت؟ چه اتفاقی برای داده‌های گمشده افتاده است؟
ت) روند برعکس آنچه در قسمت قبل فهمیدید را طی کنید (آن مقادیر را با NaN جایگزین کنید) و سپس سعی کنید با استفاده از SimpleImputer مقادیر گمشده را با استراتژی های مختلفی که تابع دارد جایگزین کنید.

## ۲- نکات پاسخ‌دهی

- تمرینات به صورت مرتب و خوانا بارگذاری شود.
- برای تمرینات غیر عملی که به صورت تایپی ارسال شوند امتیاز تشویقی درنظر گرفته می شود.
- کدهای خود را حتماً در فایل PDF نیز قرار دهید.
- در سوالات توضیحی، قدرت تحلیل افراد ملاک مقایسهٔ پاسخ ها خواهد بود.
- فایل پایتون و یا Notebook برای تمرینات ضمیمه شود و همه به صورت یک فایل zip بارگذاری شوند. فایل zip را با فرمت DM4022_HW5_[StudentNumber].zip نام‌گذاری کنید.
- درصورت وجود ابهام خاص می توانید موارد را با دستیار آموزشی مطرح کنید.