دانشگاه صنعتی شریف
دانشکده‌ی مهندسی صنایع

مبانی داده کاوی و کاربردهای آن
بهار ۱۴۰۴

استاد: مانا مس کار
مسئول تمرین: محمدمهدی منتظری هدش

تمرین شماره یک **مهلت تحویل: ۱۸ اسفند ۱۴۰۳**

- دانشجویان محترم، لطفاً حداکثر تا **سه روز آینده** اعضای گروه خود را به ایمیل montazeriyazd1@gmail.com ارسال کنید. در ایمیل ارسالی، نام، نام خانوادگی و شماره دانشجویی اعضای گروه را ذکر کنید. شماره گروه از طریق ایمیل اعلام می‌شود. شایان ذکر است که گروه‌ها باید دو نفره باشند و پس از این موعد امکان تغییر وجود ندارد.
- تمرین را در قالب یک فایل **ZIP** با نام `DataMining_Assignment1_GroupX.zip` ارسال کنید. این فایل باید شامل یک **Notebook** و یک **PDF** باشد. در فایل نوت‌بوک، کدهای اجرایی همراه با خروجی‌ها قرار داده شود و توضیحات مختصر در کنار کدها نوشته شود. خروجی‌ها را ذخیره کنید تا نیازی به اجرای مجدد نباشد. در فایل PDF، تحلیل و تفسیر نتایج، مقایسه روش‌ها، بررسی تأثیر داده‌های پرت و نمودارهای مورد نیاز مانند **Boxplot**، **Histogram** و **Q-Q Plot** ارائه شود. بیشترین نمره به بخش PDF اختصاص دارد، زیرا تحلیل داده‌ها مهم‌تر از اجرای کد است. استفاده از **هوش مصنوعی** برای کدنویسی مجاز است، اما تحلیل باید کاملاً توسط شما انجام شود.

بخش ۱: شناخت انواع داده‌ها

دانلود داده این بخش

۱. (۵ نمره) برای هر ستون، مشخص کنید که نوع داده آن چیست؟ (Nominal, Ordinal, Binary, Numeric)

۲. (۵ نمره) بین متغیرهای Gender و Loan Approval، تفاوت در نوع داده‌ها را توضیح دهید.

۳. (۵ نمره) آیا می‌توان از متغیر Education به‌عنوان داده عددی استفاده کرد؟ چرا؟

بخش ۲: توصیف آماری و نمایش داده‌ها

دانلود داده این بخش

۱. (۳ نمره) برای متغیرهای total_bill و tip، مقادیر زیر را محاسبه کنید:
   - میانگین، میانه، مد، دامنه، واریانس و انحراف معیار.

۲. (۷ نمره) آیا توزیع متغیر tip دارای چولگی است؟ چگونه متوجه می‌شوید؟

۳. (۷ نمره) با استفاده از نمودار هیستوگرام، توزیع متغیر total_bill را نمایش دهید. آیا این متغیر دارای توزیع نرمال است؟

۴. (۷ نمره) نمودار جعبه‌ای (Boxplot) متغیر tip را رسم کنید و نقاط پرت را مشخص کنید.

۵. (۹ نمره) نمودار Q-Q Plot برای متغیر total_bill رسم کنید و تحلیل کنید که آیا این متغیر از توزیع نرمال پیروی می‌کند؟

۶. (۷ نمره) با استفاده از معیار ۱.۵ برابر IQR، داده‌های پرت متغیر total_bill را شناسایی کنید.

بخش ۳: مقایسه شباهت بین داده‌ها

دانلود داده این بخش

دو فرد زیر را در نظر بگیرید:

PersonID	(cm) Height	(kg) Weight	Age
۹۶	۱۷۷	۱۰۰.۲۳	۵۸
۵	۱۹۵	۱۰۲.۶۲	۲۵

۱. (۲ نمره) فاصله اقلیدسی بین شماره ۹۶ و ۵ را به‌صورت دستی محاسبه کنید:

$$d(96,5) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

۲. (۵ نمره) همان فاصله اقلیدسی را مجدداً محاسبه کنید، اما این بار بعد از استانداردسازی داده‌ها با Z-score. آیا نتیجه تغییر کرد؟ چرا؟

۳. (۱۰ نمره) چه زمانی باید داده‌ها را استاندارد کنیم و چه زمانی نیاز نیست؟

۴. (۳ نمره) محاسبه شباهت کسینوسی (Cosine Similarity) بین این دو فرد را به‌صورت دستی انجام دهید.

دو فرد زیر را در نظر بگیرید:

PersonID	Color Eye	Color Hair	Type Body
۹۶	Blue	Black	Average
۵	Green	Black	Average

۵. (۵ نمره) فاصله جاکارد را برای ویژگی‌های رنگ چشم، رنگ مو و نوع بدن بین شماره ۹۶ و ۵ محاسبه کنید.

بخش ۴: آزمون‌های آماری

دانلود داده این بخش

۱. (۲۰ نمره) بررسی کنید که آیا رابطه‌ای بین جنسیت و نوع بدن وجود دارد؟

- یک جدول توافقی (Contingency Table) بین دو متغیر Gender و Body Type بسازید.
- آزمون کای‌ـ دو (Chi-Square Test) را روی این داده‌ها انجام دهید.
- مقدار p-value را تفسیر کنید: آیا این دو متغیر وابسته هستند یا مستقل؟