فاز اول پروژه

بخش اول - شناخت مجموعه داده

با توجه به مجموعه داده ای که در اختیار دارید، سعی کنید موارد ذیل را برای آن انجام دهید.

1. ویژگی های مجموعه داده را طبق جدول زیر بر اساس هر جدول توصیف نمایید .

نام جدول

نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean	Mode	Median	Outlier

2. با رسم نمودار Box Plot مقادیر پرت هر ویژگی را شناسایی کنید.
3. به عنوان تحلیلگر داده و متناسب با مجموعه داده مورد نظر سعی کنید پیشنهاداتی برای تحلیل، ارائه نمایید. به عنوان مثال یك پیشنهاد میتواند پیش بینی حمله قلبی باشد.

بخش دوم - ارزیابی کیفیت داده

با توجه به شناختی که از مجموعه داده در فاز اول به دست آوردید، برای این مرحله سعی کنید موارد ذیل را برای آن انجام دهید.

1. با توجه به مدل کیفیت ISO 25012 و بعد ذاتی آن، برای هر ویژگی در هر جدول، کیفیت آن را با توجه به فاکتورهای کیفیت مربوطه ارزیابی نمایید. مشخص کنید کدام فاکتورهای کیفیت را میتوانید ارزیابی کرده وبرای هر کدام چه درصدی از کیفیت حاصل میشود.

نام ویژگی	تعداد رکورد	تعداد مقدار Null	Accuracy	Completeness	Validity	Currentness	Consistency

2. با توجه به موارد زیر در جدول، برای هر کدام حداقل ۳ مورد از اشکالات در دیتاستی که در اختیار دار ید را مشخص کنید وبه صورت مختصر درباره هر کدام توضیح دهید.

Single-Schema	Single-Instance	Multi-Schema	Multi-Instance

3. برای بهبود کیفیت داده مورد نظر، راهکارهای خود را ارائه نمایید.

بخش سوم - پیش پردازش

در این مرحله، هدف ما پیش پردازش داده های موجود در مجموعه داده است. در این مرحله نیازمند آن هستیم تا داده های خام را تمیز کرده و در یک قالب مناسب برای تجزیه و تحلیل تبدیل کنیم.

موارد زیر برخی از اقداماتی است که در این گام می بایستی انجام دهید.

1. مدیریت missing value

با توجه به مقادیر ستون های دیتاست، با استفاده از روش هایی مانند میانگین، مد، میانه و یا رگرسیون مقادیر ناموجود را مقداردهی کنید. در صورتی که ستونی بیش از میزان مجاز مقدار ناموجود داشت می توانید آن ستون را حذف کنید.

2. تبدیل داده (data conversion)

برای برخی از داده های موجود در دیتاست عملیات نرمالسازی را انجام دهید.

3. ساخت ویژگی های جدید

با استفاده از ترکیب ستون های موجود می توانید برای دستیابی به دانش بیشتر برخی از ستون های را ترکیب کرده و به عنوان ستونی جدید در دیتاست نگه داری کنید.

4. برای دادها های عددی outlier را شناسایی کنید و از دیتاست حذف کنید.
5. در صورت نیاز از تکنیک های data reduction استفاده کنید.
6. برای داده های متنی در صورت نیاز عملیات stemming،lemmetizing و حذف stopwords انجام شود. (برای این کار می توانید از کتابخانه nltk استفاه کنید)

7. مصور سازی دیتاست بر اساس مقادیر موجود الزامی است.

نکات تحویل:

- پروژه در گروه‌های حداکثر دو نفره پیاده‌سازی شود.
- فایل‌ها باید در قالب studentOneName-studentTwoName-phase1.zip ارسال شود.
- ارسال فایل تنها از طریق سامانه VU مورد قبول بوده و فایل‌های ارسال شده در تلگرام و ... تصحیح نخواهند شد.

موفق باشید