

# فاز اول پروژه

## درس مبانی داده کاوی

### نیمسال دوم تحصیلی 1402

#### بخش اول - شناخت مجموعه داده

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

- ویژگی های مجموعه داده را طبق جدول زیر برای داده های عددی بدست بیاورید.  
\* در صورتی که داده ها شامل حروف و علائم می باشند ( مانند ستون Platforms Info ) ولی در دسته داده های عددی قرار می گیرند ، حروف و علائم آن ها را حذف کنید و مقادیر زیر را برای آن دسته از داده ها نیز بدست آورید.

نام ویژگی	نوع	بازه مقادیر	Min	Max	Mean	Mode	Median	Outlier

- با رسم نمودار Box Plot مقادیر پرت هر ویژگی را شناسایی کنید.

مهلت تحویل این بخش ( شناخت مجموعه داده ) تا هفته آینده چهارشنبه 16 اسفند خواهد بود.

#### بخش دوم - ارزیابی کیفیت داده

با بررسی مقادیر گم‌شده ( missing ) ، داده‌های پرت ( outlier )، ناهمسانی‌ها و خطاهای موجود، کیفیت هر دو مجموعه داده را ارزیابی کنید و 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 داده‌هایی که در اختیار دارید، کیفیت آن را با توجه به فاکتورهای کیفیت مربوطه ارزیابی نمایید و برای هر کدام چه درصدی از کیفیت حاصل میشود.

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

3. با توجه به موارد زیر در جدول، اشکالات در دیتاست‌ها وجود دارد، را مشخص کنید و به صورت مختصر درباره هر کدام توضیح دهید.

\* برای بخش Multi از ترکیب دو دیتاست که در "بخش چهارم" توضیح داده شده است، کمک بگیرید.

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

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

### بخش سوم - پیش پردازش (Preprocessing)

در این بخش شما باید داده‌هایی که در اختیار دارید را به فرمی ساختار یافته و تمیز تبدیل نمایید و در یک قالب مناسب برای تجزیه و تحلیل تبدیل کنید.

موارد زیر برخی از اقداماتی است که در این بخش باید انجام دهید.

1. Missing value ها همدل شوند.

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 داده های عددی outlier را شناسایی کنید و از دیتاست حذف کنید.

5. در صورت نیاز از تکنیک های data reduction استفاده کنید.

6. در صورت نیاز داده های عددی به داده های categorical تبدیل شوند .

7. برای داده های متنی در صورت نیاز عملیات stemming, lemmetizing و حذف stopwords انجام شود. (برای این کار می توانید از کتابخانه nltk استفاده کنید)

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

9. مقایسه آماری بین product rating هر video game با ژانر action با متوسط (میانگین) product rating , video game ها با ژانر action انجام شود.

**بخش چهارم – ترکیب دو دیتاست موجود با یکدیگر**

در این بخش با ترکیب دیتاست فعلی با دیتاست دیگری که در اختیار تان قرار گرفته که اندازه آن بزرگتر و کاملتر می باشد ، موارد زیر را پیاده سازی و گزارش کنید.

1. برای هر داده در دیتاست فعلی بررسی کنید که آیا در دیتاست دیگر از آن وجود دارد و در صورت وجود، ستون هایی که در دیتاست فعلی وجود ندارد را به آن اضافه کنید . همچنین اگر داده های ستون اضافه شده شامل مقادیر null یا missing می باشد ، از گام 1 در بخش سوم برای هندل کردن این مورد کمک بگیرید .
2. ستون های دیتاست فعلی را که مشابه ستون های دیتاست دیگر است انطباق دهید ، و در صورت ناسازگاری بین هر یک از داده ها آن را گزارش و روشی را برای جایگزین کردن مقدار فعلی و رفع ناسازگاری ارائه دهید.
3. با بررسی ستون های هر دو دیتاست بررسی کنید که آیا امکان ترکیب داده ها برای بدست آوردن اطلاعات بیشتر برای تحلیل قوی تر وجود دارد یا خیر و در صورت امکان ستون جدیدی با ترکیب انجام شده اضافه کنید .

موفق باشید