Data Quality Analysis

&

Preprocessing

Phase 1:
Data Exploration and Initial Analysis

1.1. بررسی اولیه داده

1.1.1. بارگذاری دیتاست (فایلCSV) و بررسی ساختار کلی (شامل ستون‌ها، تعداد سطرها و ...)

1.1.2. استخراج آمار توصیفی اولیه از هر ستون (مانند تعداد مقادیر گمشده، میانگین و انحراف معیار در ستون‌های عددی).

1.2. تجسم داده‌ها

1.2.1. رسم هیستوگرام برای ستون‌های عددی به منظور بررسی توزیع داده‌ها.

1.2.2. رسم نمودار باکس‌پلات برای شناسایی داده‌های پرت.(outlier)

1.2.3. شناسایی و مستندسازی مقادیر گمشده (missing values) با استفاده از نمودارهای مناسب.

1.3. بررسی روابط میان متغیرها

1.3.1. رسم نمودار scatter (پراکنده) میان چند ستون منتخب به منظور بررسی وجود رابطه بین آن‌ها.

1.3.2. تحلیل و تفسیر روابط مشاهده شده (مثلاً همبستگی مثبت، منفی یا عدم وجود رابطه).

1.3.3. نمایش heatmap ماتریس شباهت‌سنجی روش‌های استفاده شده و مقایسه آنها.

Phase 2:
Data Quality Assessment

2.1. تعریف و محاسبه معیارهای کیفیت داده

2.1.1. تعریف معیارهای کیفیت داده شامل:

- ✓ (Accuracy) دقت
- ✓ (Completeness) کامل بودن
- ✓ (Validity) معتبر بودن
- ✓ (Currentness) به‌روز بودن
- ✓ (Consistency) سازگاری

2.1.2. استخراج آمارهای لازم جهت محاسبه هر یک از این معیارها بر اساس داده‌های موجود.

2.2. شناسایی مشکلات کیفیت داده

2.2.1. شناسایی موارد نادرست، ناسازگار، مقادیر گمشده و داده‌های پرت در دیتاست.

2.2.2. تهیه گزارش کتبی شامل مشکلات موجود در هر ستون به همراه نمودارها و آمارهای توضیحی.

2.3. ارائه راهکارهای بهبود کیفیت داده

2.3.1. پیشنهاد راهکارهایی جهت رفع مشکلات شناسایی شده (مانند استفاده از روش‌های مختلف جایگزینی مقادیر گمشده).

2.3.2. تهیه یک چک‌لیست جهت نظارت بر کیفیت داده پس از انجام پیش‌پردازش.

Phase 3:
Data Preprocessing

3.1. پاکسازی و اصلاح داده‌های اشتباه

3.1.1. پر کردن یا حذف مقادیر گمشده **(missing values)**

3.1.2. شناسایی و اصلاح یا جایگزینی داده‌های پرت **(outliers)**

3.2. تبدیل داده‌ها و مهندسی ویژگی‌ها

3.2.1. تبدیل نوع داده‌ها **(Data Conversion)** در صورت نیاز

3.2.2. استخراج ویژگی‌های جدید از داده‌های موجود.

3.2.3. انجام کاهش ابعاد **(Data Reduction)** در صورت وجود ویژگی‌های تکراری یا زائد با روش‌های **PCA**، **kernel trick** و **SNE** و مقایسه نتایج..

3.3. تجسم داده‌های پیش‌پردازش شده

3.3.1. رسم نمودارهای جدید برای ارزیابی تاثیر پیش‌پردازش بر داده‌ها.

3.3.2. مقایسه نتایج قبل و بعد از پیش‌پردازش و تحلیل بهبودهای حاصل.

Phase 4:
Analysis of Results and Reporting

4.1. پاسخ به سوالات کلیدی

4.1.1. تحلیل نهایی روابط بین متغیرهای کلیدی با استفاده از نمودار **scatter** و نمودارهای همبستگی.

4.1.2. پاسخ به سوالاتی نظیر:

- کدام ویژگی بیشترین تاثیر را دارد؟
- آیا بین ستون ها روابط معناداری وجود دارد؟
- با توجه به روابط بین داده‌ها، بهتر است ستون‌هایی را نگه داریم که بیشترین **correlation** را باهم دارند یا کمترین ؟ چرا ؟

4.2. مقایسه و تحلیل بهبودهای حاصل در هر یک از معیارهای کیفیت داده.

4.3. تدوین گزارش نهایی

4.3.1. تهیه یک گزارش جامع شامل مراحل انجام‌شده، تحلیل‌های آماری و بصری، مشکلات شناسایی‌شده و راهکارهای پیشنهادی.

Additional Notes

استفاده از روش‌های دقیق‌تر و خلاقانه در تحلیل و پیش‌پردازش داده با نمره اضافه همراه است.

دانشجویان باید نتایج پیش‌پردازش را به صورت تصویری (با نمودارها) مقایسه کنند تا تاثیر هر مرحله به وضوح نمایش داده شود.

پروژه تنها به صورت انفرادی قابل انجام بوده و در صورت وجود هرگونه تشابه بین دو کد یا عدم تسلط به روند، نمره منفی به دانشجویان داده میشود.

کپی و استفاده کورکورانه از منابع اینترنتی، یا ابزار هایی نظیر **Chat-GPT** ممنوع است.

فایل‌های نهایی پروژه شامل کد و گزارش را با فرمت **StudentNumber_Name_1.zip** زیپ کرده و در **VU** بارگذاری کند.