تمرین اول داده کاوی

با توجه به دیتاست مشخص شده برای گروه تان، مراحل و گام های زیر را برای انجام تمرین انجام دهید.

۱. بررسی اولیه داده ها

- بارگذاری دیتاست: بارگذاری داده‌ها و بررسی ساختار کلی دیتاست (شامل تعداد سطرها و ستون‌ها، نوع داده‌ها).
- آمار توصیفی: استخراج انواع مختلف معیارهای آمار توصیفی از هر ستون.
- شناسایی مشکلات داده‌ها: شناسایی داده‌های از دست رفته (missing values)، داده‌های پرت (outliers)، داده‌های نویزی و دیگر مشکلات.
- کیفیت داده‌ها: بررسی کیفیت داده‌ها از جنبه‌های مختلف نظیر دقت (accuracy)، کامل بودن (completeness)، انسجام (consistency) و اعتبار (Validity).
- نمایش داده‌ها: نمایش داده‌های اولیه با استفاده از نمودارهای مختلف مانند هیستوگرام، باکس پلات، اسکتر پلات و اسکتر پلات متقابل برای شناسایی بهتر ویژگی‌های داده‌ها.
- توزیع داده‌ها: بررسی توزیع داده‌ها با استفاده از ابزارهای معرفی‌شده نظیر هیستوگرام یا روش‌های دیگر بر اساس دانش پیشین.
- چک‌لیست مشکلات: تهیه چک‌لیستی از مشکلات شناسایی‌شده و آماده کردن پیشنهاد روش‌های مناسب برای بهبود هر یک از مشکلات

۲. پیش پردازش داده ها

- جایگزینی مقادیر گمشده: شناسایی و جایگزینی مقادیر گمشده با حداقل دو روش مختلف مانند استفاده از معیارهای توصیفی آماری و توزیع داده‌ها.
- حذف داده‌های نویزی و پرت: شناسایی و حذف داده‌های نویزی و پرت با حداقل دو روش مختلف.
- اصلاح داده‌های نادرست و ناسازگار: شناسایی و اصلاح داده‌های نادرست و ناسازگار با استفاده از روش‌های مناسب.
- استانداردسازی/نرمال‌سازی: ایجاد یک دیتابیس جدید از دیتابیس قبلی که داده کاملا استانداردسازی (یا نرمال‌سازی) شده باشد.
- حذف داده‌های تکراری: شناسایی و حذف داده‌های کاملاً تکراری و شبه‌تکراری (با درصد شباهت زیاد).
- مقایسه توزیع داده‌ها: مقایسه توزیع داده‌ها حداقل بین دو ویژگی با استفاده از دیورژانس کولبک-لایبر.
- نمایش داده‌های پاکسازی‌شده: نمایش داده‌های پس از پیش‌پردازش با استفاده از نمودارهای جدید برای ارزیابی تأثیر پیش‌پردازش بر داده‌ها.

- مقایسه قبل و بعد از پیش‌پردازش: مقایسه نتایج پیش‌پردازش با نتایج اولیه و تحلیل بهبودهای حاصل از پیش‌پردازش.

۳. نتیجه گیری و جمع بندی

- ارائه آمار بهبود: ارائه آماری از میزان بهبود در دیتاست پس از انجام مراحل پیش‌پردازش.
- بررسی میزان وابستگی بین داد های دیتاست با ابزارهای معرفی شده
- تهیه یک گزارش جامع شامل مراحل انجام‌شده، تحلیل‌های آماری و بصری، مشکلات شناسایی‌شده و راهکارهای پیشنهادی.

۴. قوانین و مقررات انجام تمرین:

- 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.
- 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.
- فایل نهایی پروژه را با فرمت StudentNumber_Name_G#.zip (در قسمت G# به جای # شماره گروه خود را بگذارید.)زیپ کرده و در VU بارگذاری کند.