پیاده سازی:

هدف از این بخش از تمرین آشنایی با کتابخانه های مورد استفاده در پایتون برای داده کاوی می‌باشد. در قسمت اول این تمرین عملیاتی جهت پیش پردازش داده ها و visualization انجام میشود.
پیشنهاد میشود از Jupytr Notebook برای پیاده سازی کد های پایتون خود استفاده کنید.
در قسمت دوم با رگرسیون خطی کار خواهید کرد.
کتابخانه های مورد استفاده:

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit-learn

قسمت اول:

فایل csv موجود در پوشه data با نام covid.csv شامل اطلاعات افراد مبتلا به COVID-19 در کره جنوبی میباشد.

۱- این فایل را خوانده و در یک جدول نمایش دهید.

۲- داده ها را با مشاهده سطر و ستون های آن شرح دهید. تعداد داده ها و نام ستون ها را نمایش دهید.

۳- مقادیر max, mean و std را در ستون birth_year به دست آورده و نمایش دهید.

۴- بررسی کنید که مقدار null در داده ها وجود دارد یا خیر. در صورت وجود با استفاده از متد مناسب آن را از بین ببرید.

۵-در این بخش مصور سازی داده ها را انجام می‌دهید. با انتخاب ستون مناسب از داده ها، scatter plot, matrix plot و histogram plot را نمایش دهید.

۶- بررسی کنید که آیا این مجموعه داده دارای outlier هست یا خیر. در صورت وجود علت خود را بیان کنید و برای روشی برای حل آن ارائه دهید.

قسمت دوم (رگرسیون خطی):

برای این بخش یک مجموعه داده از تعدادی دانش آموز در پوشه data با نام student.csv قرار دارد. هدف از این قسمت پیش بینی نمره نهایی دانش آموز (G3 attribute) با استفاده از رگرسیون خطی میباشد. اطلاعات مربوط به این مجموعه داده را میتوانید در این لینک مشاهده کنید.

داده ها باید به دو بخش train و test تقسیم کنید (نسبت تقسیم ۸۰ به ۲۰ باشد و می‌توانید از متد های آماده استفاده کنید) و روی داده های train رگرسیون خطی انجام دهید. برای سادگی این قسمت فقط ستون هایی که مقادیر عددی دارند را استفاده کنید (در حالت کلی میتوان ستون هایی که مرتبط هستند و مقدار عددی ندارند را به عدد تبدیل کرد).
سپس نمره نهایی را (G3) برای داده های test پیش بینی کنید و دقت (accuracy) مدل آموزش داده شده را به دست آورید.