

شرح تمرین

- هدف از دومین تمرین این دوره، گرد هم آوردن تمام مهارت‌های لازم برای انجام سه مرحله اول یک پروژه یادگیری ماشین (Data ingestion, Data Validation, Data Transformation) است.
- مجموعه داده در نظر گرفته شده برای این تمرین را می‌توانید از آدرس زیر دانلود نمایید. از این مجموعه داده برای آموزش مدل پیش‌بینی نوع پوشش جنگل بر اساس متغیرهای ارائه شده استفاده کنید.
<https://archive.ics.uci.edu/dataset/31/covertime>
- ابتدا تعداد ویژگی‌های مجموعه داده را با حداقل سه روش متفاوت کاهش دهید (Feature Selection). این کار به کاهش پیچیدگی مدل شما و صرفه‌جویی در منابع در حین آموزش کمک می‌کند. این نکته را در نظر بگیرید که کاهش ویژگی‌ها نباید بیشتر از ۵ درصد از دقت مدل را کاهش دهد.
- سپس با آماده شدن زیرمجموعه انتخاب شده از ویژگی‌ها، پایپ‌لاین دیتا که شامل validating.ingesting و transform مجموعه داده می‌باشد را طراحی کنید. انجام مراحل زیر در طراحی پایپ‌لاین ضروری می‌باشد.

- ExampleGen
- StatisticsGen
- SchemaGen
- ExampleValidator
- Preprocessing function
- Transform