توضیحات کد ارسالی مینی پروژه دوم دوره‌ی یادگیری ماشین مکتب خونه

سجاد عابد

در سلول ۱ تمام کتابخانه‌های مورد نیاز ایمپورت شده اند که در ادامه در زمان استفاده از هر کدام در مورد آن بحث خواهیم کرد.

در سلول ۲تا۴ دیتافریم‌ها لود می‌شوند و نگاهی به تعداد و ستون‌های آن‌ها می‌کنیم.

در سلول ۵ با توجه به اینکه ستون weather ۴ حالت را به شکل عددی بین ۱ تا ۴ نشان می‌دهد، لازم است تا با دستور get_dummies این ستون را به ۴ ستون جدای ۰و۱ تبدیل کنیم.

سلول ۷ تنها جهت اطمینان از برابر بودن مجموع ستون‌های registered و casual با ستون count است.

برای دیتافریم تست نیز کاری که در سلول ۵ انجام دادیم را انجام میدهیم.

در سلول ۱۰ الی ۱۵ با توجه به اینکه تاریخ به شکل رشته است و برای اینکه به عنوان فیچر در نظر گرفته شود می‌تواند به شکل time_epoch آن را به یک عدد تبدیل کنیم. همچنین ساعت و روز هفته را نیز به عنوان فیچر جدید از آن ستون استخراج می‌کنیم زیرا میتواند کمک کننده باشد. همچنین اگر با گذشت زمان علاقه مردم به اجاره دوچرخه بیشتر یا کمتر شده باشد، time_epoch یک فیچر مهم برای ما خواهد بود.

در سلول ۱۶ میفهمیم که هیچ دیتای از دست رفته ای نداریم

در سلولهای بعدی نمودار کورلیشین نمایش داده میشود که طبق آن hour و temp بیشترین همبستگی را با فیچرهای registered و casual دارند.

برای learn شدن مدل‌ها با توجه به اینکه کورلیشن فیچرها با registered و casual متفاوت است، هر دو مقدار را پیش بینی میکنیم سپس آن دو را با هم جمع میکنیم تا count به دست بیاید.

در سلول ۱۹ تا ۲۱ train و تست را در دیتافریم train جدا میکنیم. زیرا میخواهیم متریک‌ها را برای هر مدل اندازه بگیریم سپس روی داده های اصلی خود پیاده کنیم.
در سلول‌های بعد ابتدا به صورت خطی تک متغیره (با فیچر temp که بیشترین همبستگی را دارد) سپس با مدل multi_linear و سپس با مدل Polynomial درجه ۲ و ۳ پیشبینی میکنیم که با توجه به متریکهای ثبت شده بهترین حالت با مدل multi_linear ثبت شده است.

در نهایت داده‌ی تست را با مدل multi_linear پیشبینی میکنیم و داده‌های خروجی منفی را صفر در نظر میگیریم و به صورت فایل csv ذخیره میکنیم.