پیش‌بینی عملکرد دانش‌آموزان

هدف شما ساخت و ارزیابی چندین مدل رگرسیون برای پیش‌بینی نمره نهایی دانش‌آموزان (G3) است. شما از مجموعه داده‌های عملکرد دانش‌آموزان استفاده خواهید کرد. این مجموعه داده‌ها حاوی اطلاعاتی در مورد دانش‌آموزان در یک دوره ریاضی است. متغیر هدفی که باید پیش‌بینی کنید G3 (نمره نهایی، که از ۰ تا ۲۰ متغیر است) است.

لینک دیتاست: لینک

بخش اول: تحلیل اکتشافی داده‌ها

همانطور که در کلاس توضیح داده شد، این مرحله را روی داده‌های خود پیاده‌سازی کنید. حتما و حتما، تحلیل‌های خود را برای هر نمودار نوشته و به طور صریح آن‌ها را بیان کنید و از توضیحات گنگ و مفید و مختصر خودداری کنید.

کارهایی که باید در این مرحله انجام دهید شامل تحلیل تک متغیره و رسم نمودار حرارتی می‌باشد. نیازی به تحلیل دو متغیره و بیشتر نیست. نمودارهای مناسب را برای تحلیل تک متغیره ترسیم و تحلیل کنید. این مرحله برای مرحله‌ی بعدی که پیش‌پردازش داده‌ها است ضروری می‌باشد.

بخش دوم: پیش‌پردازش داده‌ها

حتما در ابتدا داده‌های خود را به داده‌های train و test تقسیم کنید. از فایلی که برای این بخش فرستاده شد استفاده کنید و این مرحله را تکمیل کنید. شما باید در گزارش خود ذکر کنید که از چه روش و به چه علت، در هر کدام از قدم های این مرحله استفاده کرده‌اید. سعی کنید از چندین روش استفاده کنید و نتیجه را گزارش کنید. مثلا برای encoding داده‌های categorical از چند روش موجود استفاده کنید و نتیجه‌ی نهایی مدل را با هم مقایسه کنید.

نکته:

قبل از ساخت مدل، باید بررسی کنیم که آیا ویژگی‌های ما با یکدیگر همبستگی بالایی دارند یا خیر. این پدیده، همخطی چندگانه نامیده می‌شود و می‌تواند ضرایب مدل ما را ناپایدار و غیرقابل تفسیر کند. ما از عامل تورم واریانس (VIF) استفاده خواهیم کرد. امتیاز ۱ به معنای عدم همبستگی، ۱-۵ به معنای متوسط و > ۵ (یا >

۱۰) به معنای همخطی زیاد است. دیتافریم VIF را چاپ کنید. آیا ویژگی با امتیاز VIF بالای ۵ می‌بینید؟ (توجه: ویژگی با VIF بالا، می‌تواند نادیده گرفته شود).

بخش سوم: مدلسازی

مدل خود را آموزش دهید و با استفاده از معیارهای ارزیابی گفته شده، آن را ارزیابی کنید.

بررسی کنید که آیا مدل ما فرضیات کلیدی رگرسیون خطی را برآورده می‌کند یا خیر. ما این کار را با رسم باقیمانده‌های آن (خطاها) انجام می‌دهیم. یک نمودار باقیمانده خوب باید شبیه نویز تصادفی پراکنده در اطراف خط صفر باشد.

برای انجام:

باقی مانده‌ها را محاسبه کنید: باقیمانده‌ها = y_test - y_pred.

یک نمودار پراکندگی (seaborn.scatterplot) با y_pred (مقادیر پیش‌بینی‌شده) روی محور x و باقیمانده‌های خود روی محور y ایجاد کنید.

یک خط قرمز افقی در y=0 اضافه کنید تا دیدن آن آسان شود.

سوال (خطی بودن): به نمودار خود نگاه کنید. آیا نقاط به طور تصادفی در اطراف خط قرمز پراکنده به نظر می‌رسند یا یک منحنی یا الگوی واضح (مانند شکل U) می‌بینید؟ این چه چیزی در مورد اینکه آیا رابطه واقعاً خطی بوده است به شما می‌گوید؟

سوال (هم‌واریانسی): به پراکندگی نقاط از چپ به راست نگاه کنید. آیا پراکندگی ثابت است (نویز تصادفی)، یا اینکه گسترده‌تر یا مخروطی شکل است (پهن‌تر یا باریک‌تر می‌شود)؟ این پهن شدن ناهم‌واریانسی نامیده می‌شود و به این معنی است که خطای مدل ثابت نیست. در نمودار خود چه می‌بینید؟

بخش چهارم: بهبود

با استفاده از روش‌های گفته شده در کلاس، مدل خود را بهبود دهید.

نکته: فرمت نهایی گزارش شما باید به این شکل باشد:

- **مقدمه:** شرح مسئله و معرفی دیتاست
- **روش شناسی:** توضیح روش کار به طور کامل، شامل مراحل EDA، پیش‌پردازش و مدل‌سازی. هر مرحله را به صورت زیربخش از بخش روش‌شناسی توضیح دهید. نیازی نیست که نتایج خود را در این مرحله وارد کنید، صرفا بگویید که در روش خود چه‌کارهایی را و به چه علت استفاده کرده‌اید.
- **نتایج و آزمایشات تجربی:** در این بخش، روش خود را ارزیابی و سپس نتایج به دست آمده را تحلیل نمایید (در این بخش، تحلیل شما اهمیت دارد-بررسی تأثیر بخش‌های مختلف، مقایسه با استفاده از معیارهای ارزیابی). سعی کنید از نمودار و یا جدول و ساختارهایی که نتایج شما را واضح نشان می‌دهند و با معیارهای دیگر مقایسه می‌کنند بهره ببرید.
- **نتیجه گیری:** خلاصه سازی یافته‌ها

- ارسال تمام کدهای شما از پیاده سازی باید در فایل های ipynb. و به همراه تمام نتایج باشد. زبان برنامه‌نویسی شما باید پایتون باشد و حتما باید در فایل ipynb. کدهای خود را بنویسید.
- در گزارش های ارسالی از شبه‌کد می‌توانید استفاده کنید ولی نمی‌توانید مستقیما کد را وارد کنید.