تمارین دیتا ساینس

هفته ششم

پاییز ۱۴۰۴

قبل از این که به بحث شیرین تمارین برسیم، لازمه که چند تا نکته رو در رابطه با تمارین بدونید:

- در این تمرین، تعدادی سوال در چند دسته بندی متفاوت در اختیار شما قرار گرفته است.
- اولویت اول ما و شما، یادگیری شماست.
- زمان کافی برای حل سوالات و تمارین اختصاص دهید.
- **زمان تحویل دوشنبه ۵ آبان ماه میباشد.(ساعت ۱۱:۵۹:۵۹ شب)**
- ددلاین تمرین به هیچ عنوان تغییر نخواهد کرد.
- تمرینا رو برای منتور(نوا) ارسال کنید. (فایل های ارسالی رو با اسم خودتون بفرستید.)

Telegram: @n80a90zh94

- در صورت داشتن سوال در رابطه با تمرین یا نیاز به راهنمایی، می توانید به آدرس تلگرام زیر پیام بدهید.

Telegram: @mahdavian_a
Telegram: @mqod1996
Telegram: @alistelegramid
Telegram: @n80a90zh94

سوالات پیاده سازی:

۱. یک بانک به‌صورت دوره‌ای کمپین‌های بازاریابی تلفنی اجرا می‌کند تا مشتریان موجود خود را به افتتاح حساب سپرده‌گذاری مدت‌دار ترغیب کند. در هر کمپین، هزاران تماس تلفنی با مشتریان بالقوه برقرار می‌شود، اما نرخ تبدیل (Conversion Rate)پایین است (حدود ۱۱ تا ۱۲ درصد). لینک دیتاست :

https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing

هدف شما این است که به تیم بازاریابی بانک کمک کنید که پیش‌بینی کنند آیا مشتری اشتراک سپرده‌گذاری مدت‌دار را قبول می‌کند یا خیر(y: yes/no) و عوامل کلیدی مؤثر بر تصمیم مشتریان را درک کند.

مراحل:

۱. پیش‌پردازش داده‌ها :
- هر کاری که لازم است را انجام دهید.
- (چرا duration را باید حذف کنید؟)

۲. مدل‌سازی : (مدل‌های زیر را آموزش داده و ارزیابی کنید.)
- درخت تصمیم(Decision Tree)
  - با استفاده از sklearn.tree.plot_tree ساختار درخت را رسم کنید (لینک راهنما:

https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html

  - نام ویژگی‌ها را به صورت خوانا در رسم درخت نمایش دهید و کلاس‌ها را مشخص کنید.
  - یک تصویر از درخت (PNG/PDF) خروجی بگیرید و سپس این درخت را به ChatGPT بدهید و بخواهید که اطلاعاتی که میتواند را از آن استخراج کند.

- XGBoost(اختیاری)
- جنگل تصادفی(Random Forest) (اختیاری)
- شبکه عصبی چندلایه(MLP)
- Ensemble (اختیاری)

برای ارزیابی دقیق‌تر، از 5-Fold Cross Validation استفاده کنید.

۳. ارزیابی مدل‌ها : (عملکرد مدل‌ها را با معیارهای زیر مقایسه و تحلیل کنید.)

- Accuracy
- Precision
- Recall
- F1-Score

- ROC-AUC

همچنین:

- ازConfusion Matrix برای نمایش خطاهای مدل استفاده کنید.
- برای تفسیر مدل از Feature Importance بهره بگیرید.

۴. گزارش : (یک گزارش کوتاه یک صفحه ای از مراحل و شهودات و نتایج خود بنویسید.)

---

۲. داده‌ی مورد استفاده شامل۵۹۴۳ جمله است که هر جمله به‌عنوان ویژگی (feature) در نظر گرفته شده و احساس یا هیجان (emotion) مربوط به آن به‌عنوان برچسب (label) مشخص شده است. هدف این پروژه، آموزش یک مدل طبقه‌بندی‌کننده برای تشخیص احساسات (Sentiments) بر اساس داده‌های متنی است.

https://www.kaggle.com/datasets/abdallahwagih/emotion-dataset

۱. پیش‌پردازش(Pre-Processing)

- داده را بارگذاری کرده و برای مرحله‌ی استخراج ویژگی آماده کنید.
- جملات را توکنایز (Tokenize) کنید و کلمات توقف(Stop Words) ، علائم نگارشی و سایر نویزهای متنی را حذف نمایید.

۲. تقسیم داده‌ها(Train-Test Split)

داده‌ها را به سه بخش تقسیم کنید:

- داده‌ی آموزش :(Train) ۷۰٪
- داده‌ی اعتبارسنجی :(Validation) ۱۵٪
- داده‌ی آزمون :(Test) ۱۵٪

۳. استخراج ویژگی‌ها(Feature Engineering)

از جملات توکنایزشده برای استخراج ویژگی‌های متنی استفاده کنید. روش‌های پیشنهادی عبارت‌اند از:

- بررسی کنید که آیا کلمات خاصی در جمله وجود دارند یا خیر (Binary Features)
- محاسبه‌ی تعداد تکرار (Frequency) هر واژه در جملات.
- استفاده از دوواژه‌ای‌ها (Bigrams) و تکرار روش‌های بالا برای آن‌ها.
- تعیین حد بالایی برای تعداد کلمات (Vocabulary Limit)مقدار را با دقت انتخاب کنید.

- از مهارت‌های شخصی خود برای استخراج ویژگی‌های جدید و خلاقانه بهره ببرید.

۴. آموزش و ارزیابی مدل(Training and Evaluation)

- از Naive Bayes و SVM استفاده کنید تا جملات را بر اساس احساسات به کلاس مناسب تخصیص دهید.
- عملکرد مدل را با معیار مناسب ارزیابی (Evaluation Metric) کنید.
- بهترین مدل را معرفی کنید و دلیل انتخاب خود را به‌صورت منطقی و تحلیلی توضیح دهید.
- تحلیل کنید که چرا یک مدل ممکن است برای این مسئله بهتر یا ضعیف‌تر عمل کند.

۵. گزارش : (یک گزارش کوتاه یک صفحه ای از مراحل و شهودات و نتایج خود بنویسید.)

KELAASOR