این پروژه مربوط به طرح کارآموزی دیتاشیپ است.

لطفاً برای حمایت بیشتر از ما شبکه‌های اجتماعی ما را دنبال کنید.

- Website
- Youtube
- Github
- Linkedin
- Telegram Channel: @**data_hub_ir**
- Instagram Page: @**data_hub_ir**
- Telegram Group: @**data_jobs**

تحلیل جامع کاربران توییتر فارسی

این تمرین هم جنبه تحلیل و هم پیاده‌سازی دارد و نیازمند آشنایی با پایتون و یادگیری ماشین است. هدف تمرین هر چه بیشتر قدرت تجزیه‌وتحلیل و البته مهارت کدنویسی است. تهیه گزارش کار ضروری است.

فاز اول:

توجه: توصیه می‌شود برای اجرا و پیاده‌سازی از colab استفاده کنید.
ابتدا به آدرس زیر مراجعه کرده و داده‌ها را دانلود کنید.
https://github.com/mohamad-dehghani/persian-twitter-user-information

این دیتاست نسبتاً قدیمی شامل اطلاعات 3000 هزار کاربر توییتر فارسی است که ویژگی‌هایی مثل تعداد فالوور، فالوییینگ‌ها، تعداد توییت‌ها و موارد دیگر است. به دلیل حفظ حقوق کاربران، آی دی افراد حذف شد.
ابتدا داده‌ها را دانلود کرده و فایل را باز کنید و باهدف آشنایی بیشتر با ویژگی‌ها، برای تمامی ستون‌ها در حد یک جمله یادداشت کنید. سپس داده‌ها را فراخوانی کرده و سپس باهدف شناخت بهتر داده‌ها به سراغ eda بروید و سعی کنید به کمک نمودارهای مناسب دید خوبی نسبت به پراکندگی داده‌ها به دست آوردید؛ نمودارهای مختلفِ مثل توزیع فراوانی تعداد فالوورها.

سپس داده‌های نال بررسی کرده و حتماً روش مناسبی برای پرکردن آن‌ها استفاده کنید. در ادامه وابستگی بین ویژگی‌های مختلف به کمک نمودارهایی مثل heatmap ترسیم شود.

در قسمت بعدی به بحث پیش‌پردازش داده‌ها پرداخته و برای هر پیش‌پردازشی که انجام می‌دهید در حد کوتاه کامنت بنویسید. مثلاً آیا برای ویژگی‌هایی مثل تعداد فالوور نیاز به نرمال‌سازی مقادیر هست یا خیر.

فاز دوم:

یکی از اهداف پروژه، خوشه‌بندی کاربران است. سپس به سراغ مهندسی ویژگی رفته و مهم‌ترین ویژگی‌ها را انتخاب کنید؛ باید تلاش کنید از حداکثر ویژگی‌ها برای پیاده‌سازی آن استفاده کنید.

نکته: انجام پیش‌پردازش و مهندسی ویژگی خوب، در تسک خوشه‌بندی تضمین‌کننده نتیجه عالی است.

در ادامه به سراغ حداقل 2 الگوریتم از لیست زیر رفته و کاربران را خوشه‌بندی کنید. حتماً باید دلیل انتخاب الگوریتم بیان شود.

- استفاده از الگوریتم kmeans
- استفاده از الگوریتم dbscan
- استفاده از تکنیک rfm
- استفاده از یادگیری عمیق و شبکه عصبی

در نهایت به کمک معیارهای رایج مثل silhouette، خروجی‌ها را به‌خوبی تحلیل کنید و با هم مقایسه کنید و راه‌هایی که برای بهبود بیشتر داده‌ها به ذهنتان می‌رسد را بنویسید. ضمناً به سؤالات زیر حتماً باید پاسخ مناسب داده شود.

- تعداد بهینهٔ خوشه‌ها چند است؟ چرا؟
- پراکندگی داده‌ها زیاد بود یا کم؟
- اگر داده‌ها پراکنده بودند از چه الگوریتمی باید استفاده کرد؟ چرا؟
- چگونه از شبکه عصبی در خوشه‌بندی می‌توان استفاده کرد؟

وقتی مدل نهایی انتخاب شد، حتماً خوشه‌ها به‌صورت نموداری و با رنگ‌های مختلف رسم شوند. حتماً باید برای هر خوشه اسم مناسب انتخاب شود مثل خوشه "افراد کم فعالیت".

فاز سوم:

در ادامه وقتی هر کاربر، به یک خوشه تعلق گرفت، باید عملیات دسته‌بندی[1] انجام شود. هدف پیش‌بینی خوشه آنهاست. یعنی از نتیجه‌ای که در مرحله قبل به دست آمد به‌عنوان ویژگی هدف دسته‌بند استفاده کنید. باید حداقل سه الگوریتم دسته‌بندی استفاده شود. سپس به کمک معیارهای رایج مثل f1-score، خروجی‌ها را به‌خوبی تحلیل کنید و با هم مقایسه کنید و راه‌هایی که برای بهبود بیشتر داده‌ها به ذهنتان می‌رسد را بنویسید. ضمناً به سؤالات زیر حتماً باید پاسخ مناسب داده شود.

- وقتی دیتاست بدون برچسب داشتیم می‌توان ابتدا از خوشه‌بندی استفاده کرده و برچسب خودکار تولید کنیم؟ مزایا و معایب آن چیست؟
- مزیت استفاده از الگوریتم‌های یادگیری ماشین نسبت به شبکه عصبی در تسک دسته‌بندی چیست؟
- اگر دیتاست برچسب نداشت چه روش ساده‌ای برای حل آن وجود دارد؟

فاز چهارم:

ستون "tweets_likes " را به‌عنوان ویژگی هدف انتخاب کرده و به کمک یک الگوریتم دلخواه مثل رگرسیون خطی، رگرسیون انجام دهید و خروجی‌ها به کمک پارامترهایی مثل rmse تحلیل و بررسی کنید.

فاز آخر:

مدلی طراحی کنید که به‌ازای هر خوشه، یک رگرسیون مجزا روی ستون "tweets_likes " انجام دهد یعنی مثلاً اگر 4 خوشه داشتیم باید به‌ازای هر خوشه، یک رگرسیون مجزا آموزش داده شود و در نهایت 4 مدل با if و else با هم ترکیب شوند.

- کدها ترجیحاً به زبان پایتون و به کمک sklearn باشد.
- مرتب و منظم نوشتن کدها و همچنین تکه‌تکه نوشتن امتیاز بالایی دارد.
- وجود کامنت‌های کامل و جامع توصیه می‌شود.
- مثل همیشه استفاده از کدهای آماده قابل‌قبول نیست.

[1] classification