پروژه درس «مبانی داده‌کاوی و کاربردهای آن»

پیش‌بینی وایرال شدن یک توییت

تاریخ تحویل:۱۰ تیر

نحوه ارسال پروژه:

دانشجویان گرامی در خصوص گزارش پروژه درس مبانی داده کاوی و کاربردهای آن موارد زیر را لحاظ فرمایید:

- ✓ گزارش پروژه با صفحه بندی مناسب با فونت **B Nazanin 12** برای متن فارسی و **Times New Roman12** برای متن انگلیسی، به صورت تایپ شده در ابعاد **A4** و در فرمت **Pdf** تهیه گردد. همچنین در گزارش فهرست و منابع و مراجع مورد استفاده را مشخص نمایید.
- ✓ متن گزارش باید پیوسته و دارای انسجام باشد. رعایت توالی منطقی در گزارش‌نویسی از معیارهای اساسی نگارش می‌باشد.
- ✓ خلاقیت در به کارگیری روش‌ها، نمودارها و تفسیر و تحلیل نتایج ملاک‌های اصلی نمره دهی محسوب خواهند شد.
- ✓ تمامی نمودارها و جداول باید دارای شماره و عنوان مختصر بوده و هیچ نمودار یا جدولی بدون تحلیل و بررسی در گزارش قرار داده نشود. محورهای افقی و عمودی نمودارها باید نام گذاری شده باشند.
- ✓ در صورتی که روش مورد استفاده شما مستلزم بررسی پاره‌ای از مفروضات می‌باشد، این مرحله در قسمت نتایج ارائه گردد.
- ✓ در صورتی که تکنیک **cross_validation** جهت تنظیم پارامترها مورد استفاده قرار گیرد، جزئیات خروجی های این فرآیند در قسمت نتایج تشریح گردد.
- ✓ فایل pdf گزارش و فایل کد پایتون را در یک فایل zip به آدرس **iedatamining2@gmail.com** ارسال کنید. موضوع ایمیل و فرمت نام‌گذاری فایل حتماً به‌صورت **Project_StudentNumbers** باشد.
- ✓ در صورت داشتن تأخیر در ارسال پروژه، به ازای هرروز تأخیر ۳۰٪ نمره کسر می‌شود.
- ✓ در صفحه اول گزارش اسامی اعضای گروه به همراه شماره دانشجویی را ذکر کنید.
- ✓ رعایت نکردن هر یک از موارد بالا منجر به کسر نمره خواهد شد.

تعریف مسئله

پروژه تعریف شده مربوط به اطلاعات توییت‌های نوشته شده در یک بازه زمانی کوتاه است. هدف از اجرای این پروژه ارائه مدلی برای پیش‌بینی وایرال شدن یا نشدن یک توییت است. این پروژه دارای دو قسمت اصلی است.

قسمت اول: مقدمه و مرور ادبیات

در این مرحله پس از توضیح مختصر موضوع پروژه، پژوهش‌ها و مدل‌های معرفی شده جهت بررسی پیام‌ها در شبکه‌های اجتماعی را مورد بررسی قرار دهید و عوامل و پارامترهایی که بیشترین تاثیر را در محبوبیت این پیام‌ها دارند تشریح کنید.

قسمت دوم: ارائه مدل پیش‌بینی

مرحله اول: شناسایی و استخراج داده‌های مورد نظر جهت داده‌کاوی (Feature Engineering)

در این مرحله با استفاده از دیتاست موجود و متغیرهای دیکشنری موجود در آن باید متغیرهای جدیدی را استخراج کنید که در ارائه مدل پیش‌بینی به شما کمک کنند.

راهنمایی: برای مثال می‌توانید تعداد کاراکترهای یک توییت را به عنوان یک متغیر جدید از ستون text استخراج و استفاده کنید.

مرحله دوم: تشریح داده‌ها

برای دست یافتن به درک و دید مناسب نسبت به داده‌ها لازم است تا از توابع مناسب توصیفی استفاده کنید. سعی کنید در خروجی گزارش این بخش خود از نمودارها و تصویرسازی مناسب استفاده کنید.

مرحله سوم: پیش‌پردازش داده‌ها

در این مرحله با استفاده از ابزارهای مناسب، داده را برای برازش مدل آماده کنید.

مرحله چهارم: ارائه مدل جهت پیش بینی و ارزیابی آن

در این مرحله از حداقل سه روش متفاوت جهت پیش‌بینی استفاده کنید. سپس دقت هر یک از مدل‌ها را بر اساس دو شاخص Recall و Accuracy بسنجید و با یکدیگر مقایسه کنید.

مدل‌های مورد نظر خود را با استفاده از داده‌های train_data برازش نمایید. از بهترین مدل خود جهت پیش‌بینی داده‌های test_data استفاده نموده و مقادیر پیش بینی شده را در یک ستون جدید با نام Viral_prediction اضافه کنید. فایل test_data را به همراه گزارش پایانی خود تحویل دهید.

مرحله پنجم:نتیجه گیری و تحلیل

الگوی نهایی مدل و دانش کسب شده را توضیح داده و تحلیل کنید. سپس بر اساس آن‌ها نتیجه خود را در گزارش بنویسید.