پروژه علم داده: پیش‌بینی درآمد فیلم‌ها و تحلیل اکتشافی داده‌ها

استاد: دکتر نادری

دستیاران آموزشی: سجاد مهرپیما، مبین آزادانی، امین نورمحمودی، علی کنتراتچی

مهلت ارسال: 10 بهمن

۱ .مقدمه

این پروژه با هدف تحلیل داده‌های مرتبط با فیلم‌ها و پیش‌بینی درآمد آن‌ها طراحی شـــده اســـت. داده‌های مورد استفاده شامل اطلاعات حدود ۵۰۰۰ فیلم از پایگاه داده Rotten_Tomatoes5000 و اطلاعات اعضای تیم تولید (Credits)است. پروژه به دو بخش اصلی تقسیم می‌شود:

- **تحلیل اکتشافی داده‌ها (EDA)** برای پاسخ به سوالات مشخص.
- **پیش‌بینی درآمد فیلم‌ها** با استفاده از مدل‌های یادگیری ماشین.

۲ .اهداف پروژه

- تحلیل الگوهای موجود در داده‌های فیلم‌ها (مانند رابطه ژانر، کشور، زبان، و هزینه تولید).
- پیش‌بینی درآمد فیلم‌ها بر اساس ویژگی‌های استخراج‌شده.
- بررسی تأثیر عوامل مختلف (مانند بازیگران، کارگردان، ژانر) بر موفقیت مالی فیلم.

۳ .ساختار پروژه

پروژه را به صورت زیر سازماندهی کنید:

```
|── Name_SID.zip
|   |── data
|   |   |── rotten_tomatoes_5000_movies.csv
|   |   |── rotten_tomatoes_5000_movies.csv
|   |── *.ipynb
|   |── Document.pdf
```

۴ .تحلیل اکتشافی داده‌ها(EDA)

در این بخش، سوالات زیر بررسی می‌شوند:

۴.۱ .سوالات تحلیل

1. متوسط هزینه برای هر ژانر فیلم چقدر است؟
2. سهم هر کشور در مجموع هزینه هر ژانر فیلم چقدر است؟(برای 5 تا از پر خرج ترین ژانر ها بدست بیاورید)
3. تعداد فیلم های ساخته شده در 3 ژانر را در 10 سال گذشته مقایسه کنید.
4. به طور متوسط کدام کشور ها طولانی ترین فیلم ها و کوتاه ترین فیلم ها را میسازند؟
5. به غیر از انگلیسی، پر تکرار ترین زبان ها در فیلم ها چه هستند؟
6. آمریکا در 10 سال گذشته، به طور متوسط در هر سال چقدر در صنعت فیلمسازی هزینه کرده است؟(به تفکیک سال)
7. روند قبلی را بدون در نظر گرفتن کشور برای 10 سال گذشته مقایسه کنید.
8. Johnny Depp در چه فیلم هایی بازی کرده است؟
9. به طور متوسط چند درصد نقش اول تا پنجم فیلم ها(به تفکیک برای هر نقش) مرد، و چند درصد زن هستند؟
10. توزیع سنی بازیگر های خانم و آقا را مقایسه کنید.
11. محبوب ترین ژانرهای فیلم در 10 سال گذشته به چه ترتیب بوده است؟(یکبار بر اساس تعداد review و یکبار بر اساس critics_score مقایسه کنید)

۵ .پیش‌بینی درآمد فیلم‌ها

۵.۱ .آماده‌سازی داده‌ها

۵.۲ .ایجاد ویژگی‌های جدید

- **پیش‌پردازش(موارد پیشنهادی):**
  - تبدیل متغیرهای کیفی به عددی.
  - مدیریت مقادیر گم‌شده (حذف یا جایگزینی).
- **استفاده از داده‌هایCredits(موارد پیشنهادی):**
  - تأثیر بازیگران مشهور بر درآمد.
  - رابطه کارگردان‌های شناخته‌شده با موفقیت مالی.

- ترکیب ژانرها و تأثیر آن بر جذب مخاطب.

۵.۳ .آموزش مدل‌ها

- **انتخاب و امتحان مدل های مختلف و مقایسه آن ها**
- **بهینه‌سازی هایپرپارامترها**

check MSE

۵.۴ .ارزیابی مدل‌ها

- **معیارهای ارزیابی:**
  - Mean Squared Error (MSE)
  - R-squared (R^2)
- **نمودارها:**
  - مقایسه پیش‌بینی با مقادیر واقعی (نمودار خطی).
  - اهمیت ویژگی‌ها در مدل نهایی (Feature importance)

۶ .نتیجه‌گیری(مواردی که باید در گزارش آورده شوند):

- گزارش بهترین مدل بر اساس دقت و قابلیت تعمیم.
- بررسی جنبه های مختلف داده با نمودار های قابل فهم.
- تحلیل عوامل کلیدی تأثیرگذار بر درآمد فیلم‌ها.
- پیشنهاداتی برای بهبود عملکرد.

۷ . موارد امتیازی:

- مهندسی ویژگی با کیفیت تر و بهتر.
- استفاده از معیار های ارزیابی بیشتر و مقایسه آن ها در داده های train, validation.
- ارائه بصری بهتر و کامل تر.
- کد نویسی تمیز.