عنوان تمرین: تشخیص اسپم در ایمیل با استفاده از الگوریتم‌های یادگیری ماشین

توضیح تمرین:

در این تمرین، هدف شما طراحی و ارزیابی یک مدل یادگیری ماشین برای تشخیص اسپم بودن ایمیل‌ها است. دیتاست ارائه‌شده شامل دو ستون می‌باشد:

- **متن ایمیل** :محتوای ایمیل
- **برچسب اسپم بودن یا نبودن**(Spam / Not Spam) :

مراحل انجام تمرین:

۱ .پیش‌پردازش داده‌ها:

- متن ایمیل‌ها را به کلمات تقسیم کنید.(Tokenization)
- **استخراج ویژگی‌های عددی** :ویژگی‌های عددی را با استفاده از دو روش زیر استخراج نمایید:
  - تعداد تکرار کلمات (**Count Vectorization**)
  - وجود یا عدم وجود کلمات (**Binary Vectorization**)

۲ .آموزش مدل‌های طبقه‌بندی:

از الگوریتم‌های زیر برای آموزش مدل استفاده کنید:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Naive Bayes

۳ .تنظیم ابرپارامترها:

با استفاده از Randomized Cross-Validation ، تنظیمات بهینه برای هر مدل را پیدا کرده و عملکرد آن را بهبود دهید.

۴ .ارزیابی مدل‌ها:

- **ماتریس کانفیوژن** :محاسبه و رسم ماتریس کانفیوژن برای هر مدل.
- **نمودار ROC** : رسم نمودار ROC و محاسبه مساحت زیر منحنی.(AUC)
- **مقایسه عملکرد مدل‌ها** : مقایسه‌ی عملکرد مدل‌ها بر اساس دقت, پرسیژن و ریکال

۵ .گزارش نهایی:

گزارشی تهیه کنید شامل:

- مراحل تمیزکاری و پیش‌پردازش داده‌ها.
- تحلیل عملکرد مدل‌های مختلف.
- نمودارها و نتایج ارزیابی.
- انتخاب بهترین مدل و دلیل انتخاب آن.

فرمت تحویل:

- فایل گزارش با فرمت Word (docx)
- فایل کدهای اجرایی ترجیحاً به صورت Jupyter Notebook

مهلت تحویل:

پنجشنبه ۱۴۰۴/۰۴/۰۵