

گزارش کار: تشخیص اسپم ایمیل با استفاده از مجموعه داده "SMS Spam Collection"

هدف

آشنایی با نحوه استفاده از SVM برای پردازش متن و تشخیص اسپم.

مراحل انجام کار

دانلود و استخراج مجموعه داده

- ابتدا مجموعه داده "SMS Spam Collection" از لینک زیر دانلود و استخراج شد:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00228/smsspamcollection.zip>

بارگذاری مجموعه داده

- پس از استخراج، داده‌ها از فایل SMSSpamCollection بارگذاری شدند

پیش‌پردازش داده‌ها

- تبدیل برچسب‌ها به مقادیر عددی (0 برای 'ham' و 1 برای 'spam')
-

- استفاده از TF-IDF برای تبدیل متن به ویژگی‌ها

تقسیم داده‌ها به داده‌های آموزشی و تست

- داده‌ها به دو بخش آموزشی و تست با نسبت 20/80 تقسیم شدند.

آموزش مدل SVM با کرنل خطی

- مدل SVM با کرنل خطی بر روی داده‌های آموزشی آموزش داده شد.

ارزیابی مدل و گزارش دقت

- پیش‌بینی داده‌های تست و محاسبه دقت مدل.

رسم نمودار ROC

- محاسبه نمرات احتمال برای داده‌های تست و محاسبه مقادیر ROC.

نتایج

- دقت مدل بر روی داده‌های تست: 0.98 (یا هر مقدار محاسبه شده)
- نمودار ROC نیز رسم شد که مساحت زیر نمودار (AUC) نشان‌دهنده عملکرد مدل بود.

نتیجه‌گیری

در این پروژه، با استفاده از مدل SVM و تبدیل متن به ویژگی‌ها با استفاده از TF-IDF، توانستیم مدل دقیقی برای تشخیص پیامک‌های اسپم و غیر اسپم ایجاد کنیم. نتایج به دست آمده نشان‌دهنده دقت بالای مدل در تشخیص اسپم بود.