به نام خدا

نمونه کار پروژه تشخیص اسپم ایمیل – (Spam Email Detection Project)

DO NOT COPY

خلاصه پروژه تشخیص اسپم ایمیل(Spam Detection)

این پروژه با هدف تشخیص ایمیل‌های اسپم، با استفاده از مدل Multinomial Naive Bayes و تکنیک TF-IDF توسعه داده شده است.

به دلیل عدم تعادل داده‌ها از SMOTE برای Oversampling استفاده شد.

مدل پس از تنظیم Hyperparameter با مقدار alpha = 0.204 توانست عملکرد بسیار خوبی ارائه کند:

نتایج نهایی:(Test Set)

- Accuracy : 97.13%
- (Spam) Precision : 0.90
- (Spam) Recall : 0.883
- F1-score : 0.891
- ROC-AUC : 0.934

مقدمه (introduction)

در این پروژه یک سیستم تشخیص اسپم ایمیل پیاده سازی شده است که با استفاده از روش های پردازش متن ، بردارهای TF-IDF و مدل های یادگیری ماشین، ایمیل های اسپم و غیر اسپم را طبقه بندی میکنه. هدف پروژه، بررسی مدل انتخاب های برتر و ارزیابی عملکرد آن براساس داده های واقعی است.

داده ها (Dataset)

داده ها شامل پیام های متنی ایمیل هستند که در دو دسته Spam و Ham برچسب گذاری شده است مراحل پیش پردازش داده شامل موارد زیر است:

- تبدیل متن به حروف کوچک
- حذف علائم نگارشی
- حذف stopword های انگلیسی مانند the, is , at , for , on , and ,to ,of و ... یعنی کلمات که معنی خاصی ندارند.
- تمیز سازی و ساخت ستون Clear Text
- محاسبه طول متن خام و متن پاکسازی شده

تبدیل متن به ویژگی (Feature Extraction)

برای تبدیل متن خام به ویژگی عددی از روش TF-IDF Vectorizer استفاده شده است.

این روش وزن کلمات را براساس اهمیت آن ها در متن تعیین میکنه و ورودی های مناسبی برای الگوریتم های یادگیری ماشین تولید میکند.

DO NOT COPY

مشکل عدم تعادل داده (imbalanced Data)

در داده ها تعداد ایمیل های ham بسیار تعداد از spam بود. برای رفع این مشکل از تکنیک SMOTE (Synthetic Minority Oversampling Technique) استفاده شده است تا داده های اسپم به مصنوعی افزایش یابد و مدل در تشخیص اسپم عملکرد بهتری داشته باشد.

مدل انتخاب شده (Model Selection)

چندین مدل بررسی و مقایسه شد، از جمله:

1. Multinomial Native Bayes
2. Linear SVC

در نهایت، مدل Multinomial Naive Bayes **با** پارامتر alpha = 0.204 بهترین تعادل بین Precision ، Recall و F1-score را ارائه داد.

ارزیابی مدل (Model Evaluation)

ارزیابی سه مرحله انجام شد:

1. Training set
2. Validation set
3. Test set (برای نتیجه گیری نهایی)

DO NOT COPY

متریک های استفاده شده است:

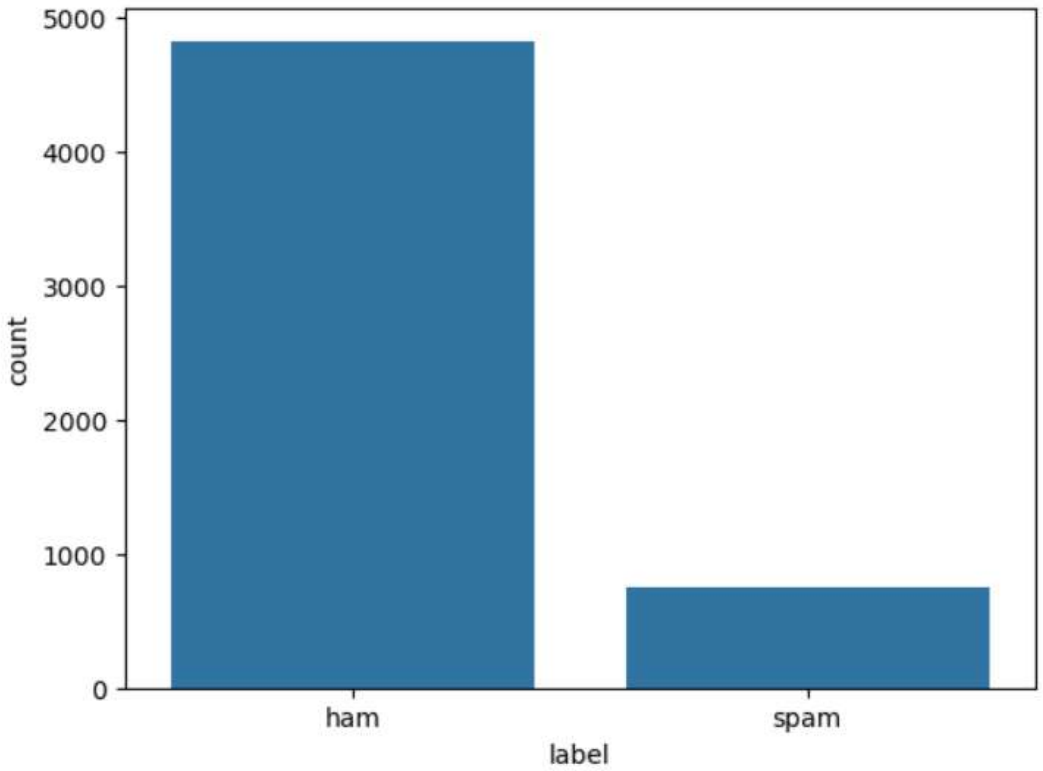
Classification report و Accuracy , Precision, Recall , f1-score, ROC-AUC, Confusion Matrix

نتایج مدل نهایی روی داده Test:

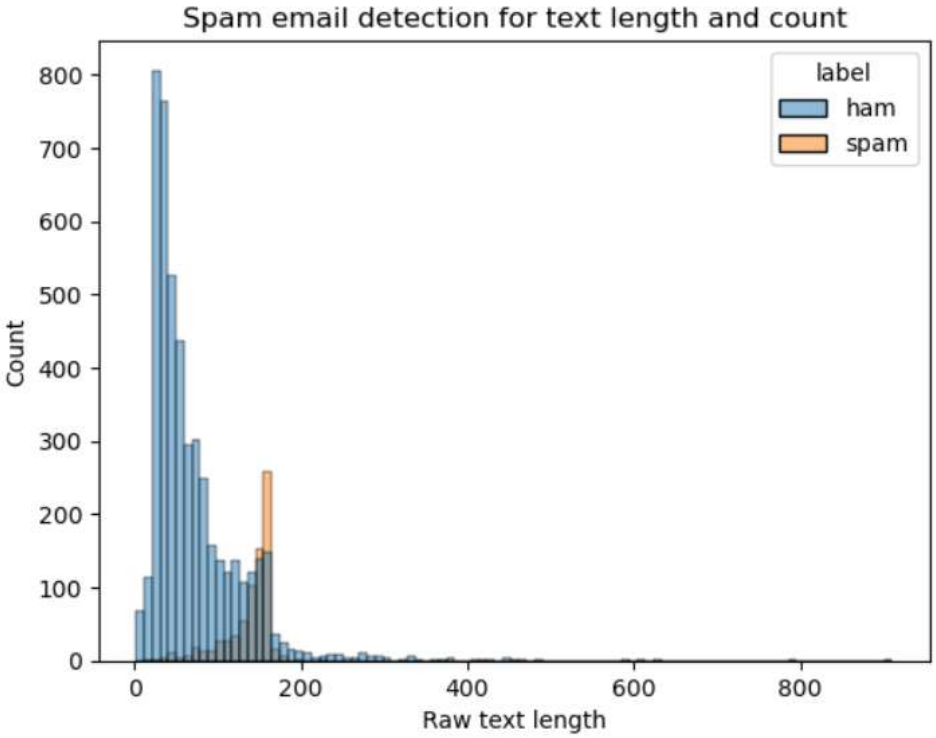
```
========================Testing metrics========================
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       724
           1       0.90      0.88      0.89       112

    accuracy                           0.97       836
   macro avg       0.94      0.93      0.94       836
weighted avg       0.97      0.97      0.97       836

Accuracy: 97.13%
precision: 0.9
f1: 0.8918918918918919
recall: 0.8839285714285714
Confusion matrix:
 [[713  11]
 [ 13  99]]
Roc score: 0.9343676006314128
```

این مقادیر نشان میدهد که مدل توانایی بالایی در شناسایی اسپم دارد.

نمودارها (Visualizations)



این نمودار نشان میدهد که داده های خام دریافت شده به صورت نمودار نمایش میدهد.



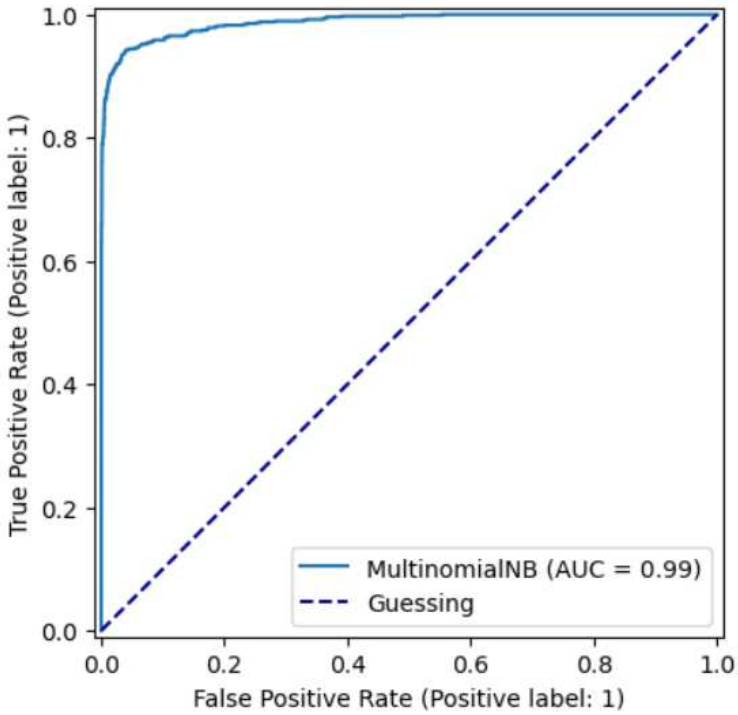
این نمودار براساس طول متن دارای اسپم و غیراسپم به صورت نمودار هسیتوگرام نمایش میدهد.

- **مدل استفاده شده: Multinomial Naive Bayes همراه با SMOTE برای مقابله با عدم تعادل کلاس ها**

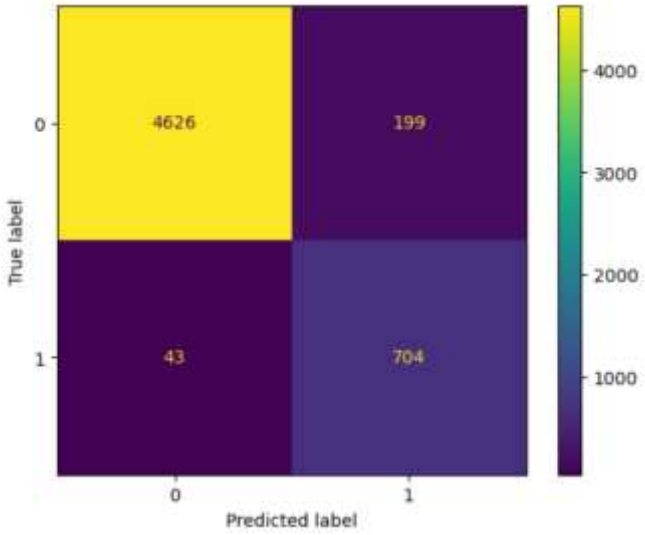
```
The model's accuracy on training set:98.83%
The model's accuracy on testing set:98.83%
              precision    recall  f1-score   support

           0       0.99      0.95      0.97      3868
           1       0.75      0.93      0.83       590

    accuracy                           0.95      4458
   macro avg       0.87      0.94      0.90      4458
weighted avg       0.96      0.95      0.95      4458

Accuracy: 94.86%
precision: 0.7462482946793997
f1: 0.8269085411942555
recall: 0.9271186440677966
Confusion matrix:
 [[3682  186]
 [  43  547]]
Roc score: 0.9395158887350359
```

کلاس 1 دقت پایین تر داره یعنی 75% اما Recall بالایی دارد، یعنی مدل بیشتر پیام های اسپم را شناسایی میکنه.

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x26ae88e9c70>
```



این نمودار ROC نشان میدهد برابر 0.94 معادل 94% است که نشان دهنده عملکرد بسیار خوب مدل در تشخیص صحیح نمونه هاست.



مدل تعداد زیادی از ایمیل‌های غیر اسپم را درست شناسایی کرده (4626 نمونه) و تنها 199 نمونه را اشتباه تشخیص داده است. همچنین، مدل بیشتر ایمیل‌های اسپم را درست شناسایی کرده (704 نمونه) و 43 نمونه اسپم را به اشتباه غیر اسپم تشخیص داده است.

این نشان می‌دهد که مدل در شناسایی کلاس اسپم عملکرد خوبی دارد، مخصوصاً با استفاده از SMOTE که به تعادل داده‌ها کمک کرده است.

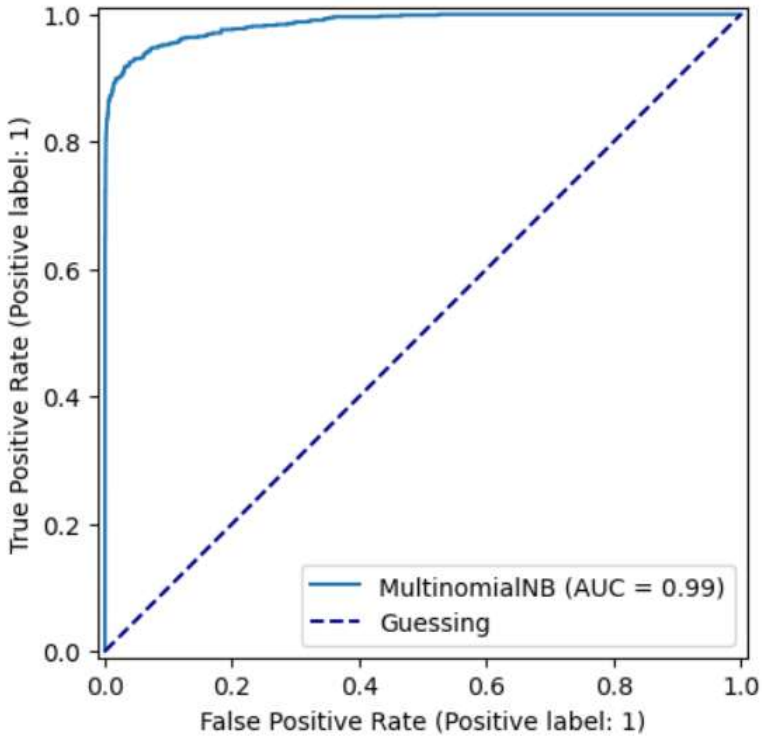
DO NOT COPY

مدل استفاده شده: با مقدار بهینه alpha = 0.204 همراه با SMOTE برای مقابله با عدم تعادل کلاس SMOTE نمونه های کلاس کمتر (ایمیل های اسپم) را افزایش داد تا مدل بهتر یاد بگیرد.

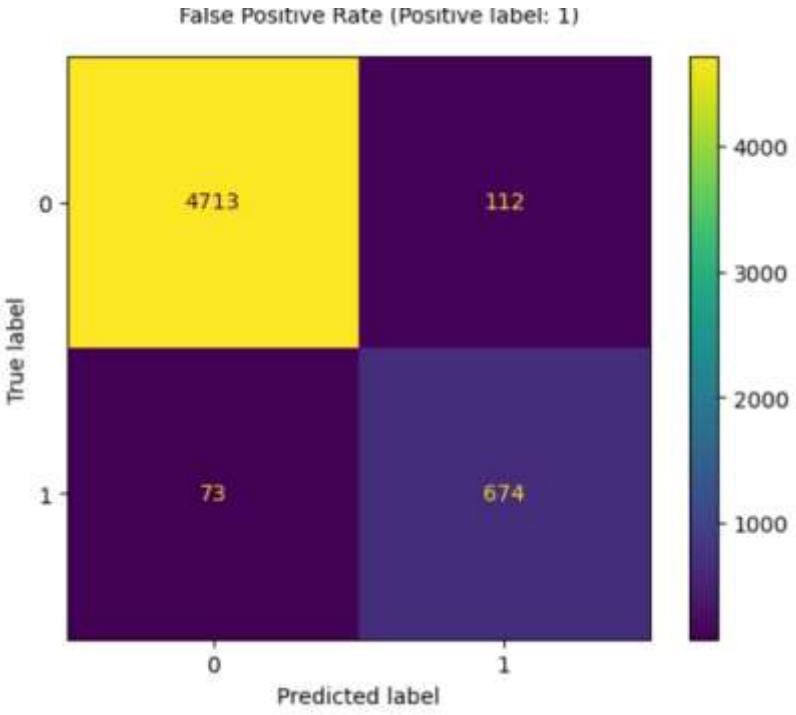
```
The model's accuracy on training set:99.37%
The model's accuracy on testing set:99.37%
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      3868
           1       0.83      0.88      0.85       590

    accuracy                           0.96      4458
   macro avg       0.91      0.92      0.92      4458
weighted avg       0.96      0.96      0.96      4458

Accuracy: 96.01%
precision: 0.8311897106109325
f1: 0.8531353135313532
recall: 0.8762711864406779
Confusion matrix:
 [[3763  105]
 [  73  517]]
Roc score: 0.9245626873258199
```

مدل کلاس 1 (اسپم) را با دقت و فراخوانی مناسب شناسایی کرده، که به کمک SMOTE بهبود یافته است.

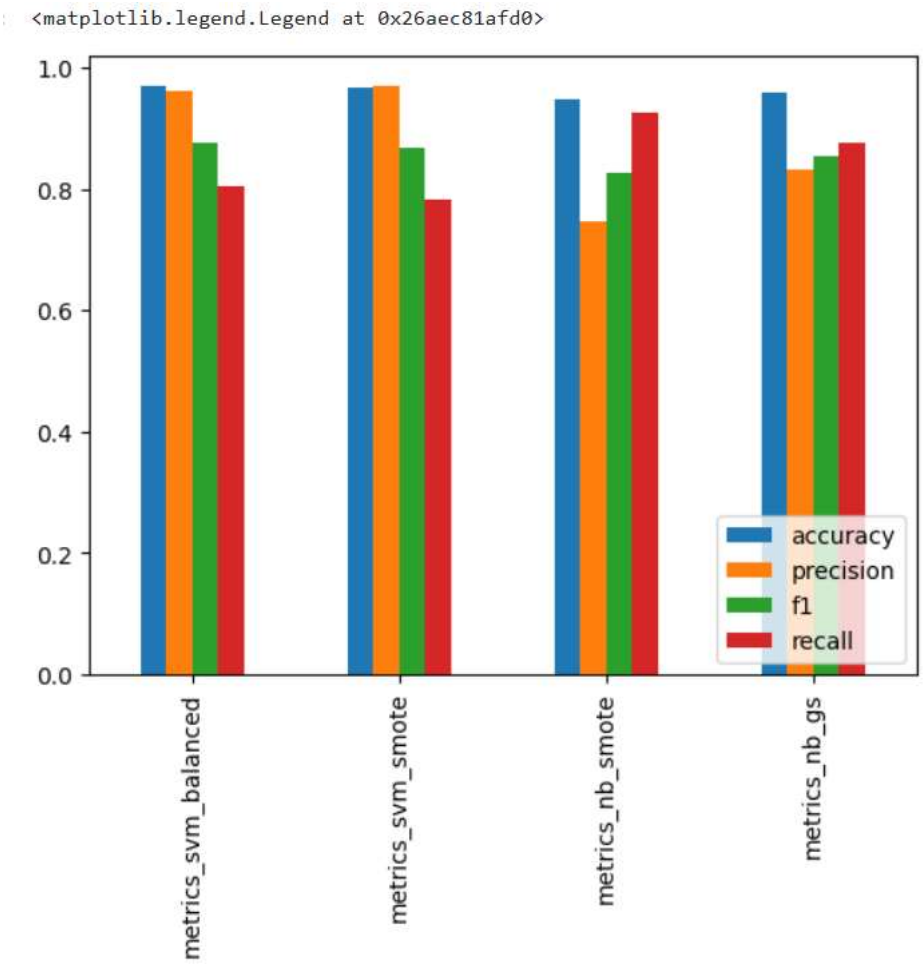
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x26ae4396250>
```



مدل توانایی خوبی در تمایز بین ایمیل اسپم و غیر اسپم دارد و نمودار ROC نشان می‌دهد مدل تقریباً تمامی نمونه‌ها را درست از هم تفکیک می‌کند، خط آبی نقطه‌چین نماینده حدس تصادفی است.



مدل بیشتر ایمیل‌های غیر اسپم را درست تشخیص داده (4713 نمونه) و تنها 112 نمونه را اشتباه تشخیص داده است و همچنین، 674 ایمیل اسپم درست شناسایی شده و 73 نمونه اسپم اشتباه غیر اسپم تشخیص داده شده است. نشان می‌دهد مدل عملکرد خوبی در شناسایی اسپم‌ها دارد.



هدف ما شناسایی دقیق‌تر ایمیل‌های اسپم بوده است. برای این منظور از مدل Multinomial Naive Bayesبا Hyperparameter Tuning استفاده شد.

با توجه به مقادیر اندازگیری به صورت آمار به‌دست آمده در metrics_nb_gs بین 3 تا مدل ، این مدل بهترین عملکرد را برای شناسایی اسپم‌ها ارائه می‌دهد و مناسب‌ترین مدل برای استفاده‌کننده نهایی است.

گزارش مدل نهایی برای شناسایی اسپم مدل

1- **Training Accuracy** = 99.38%

2- **Validation metrics:**

Metric	Value
Accuracy	97.37%
Precision (spam)	0.90
Recall (spam)	0.887
F1-score (spam)	0.896
ROC	0.937

3- **Validation Confusion Matrix:**

Actual \ Predicted	Ham (0)	Spam (1)
Ham (0)	719	10
Spam (1)	12	95

مدل توانست بیشتر ایمیل‌های اسپم را شناسایی کند و نرخ خطا برای اسپم پایین است.

DO NOT COPY

4- Test Metrics (Model Evaluation After Saving with Joblib):

Metric	Value
Accuracy	97.13%
Precision (spam)	0.90
Recall (spam)	0.883
F1-score (spam)	0.891
ROC	0.934

5- Test Confusion Matrix:

Actual \ Predicted	Ham (0)	Spam (1)
Ham (0)	713	11
Spam (1)	13	99

عملکرد مدل روی داده‌های تست مشابه Validation است، نشان می‌دهد مدل **پایدار**، جنرالایز شده و قابل استفاده عملی است.

DO NOT COPY

تکنولوژی ها و ابزارهای استفاده شده:

- Python
- Scikit learn
- Numpy
- Pandas
- Matplotlib
- Seaborn
- SMOTE
- Jupyter notebook

نتیجه گیری (Conclusion)

مدل Multinomial Naive Bayes پس از اعمال SMOTE و تنظیم Hyperparameter توانست عملکرد بسیار مناسبی در تشخیص اسپم ارائه دهد.

این پروژه نشان می‌دهد که حتی با مدل‌های ساده اما به‌درستی تنظیم‌شده می‌توان به دقت بالا در مسائل تشخیص اسپم ایمیل دست پیدا کرد.

ارتباط با ما (Contact us)

Email: www.mohamadnazari771998@gmail.com

Phone number: 09397196427

GitHub: https://github.com/MohammadNazari98/MyPortfolio

DO NOT COPY