Aprendizagem Automática

Classificação de Críticas de Cinema do IMDb

Docente: Gonçalo Marques

Desenvolvido por: Ana Oliveira - A39275

Limpeza dos Dados

Limpeza dos Dados
(exemplo 1)

- Crítica Original *(182 palavras)*

This interesting Giallo boosts a typical but still thrilling plot and a really sadistic killer that obviously likes to hunt his victims down before murdering them in gory ways.

Directed by Emilio P. Miraglia who, one year earlier, also made the very interesting "La Notte che Evelyn Usci della Tomba" (see also my comment on that one), the film starts off a little slow, but all in all, no time is wasted with unnecessary sub plots or sequences.

This film is a German-Italian coproduction, but it was released in Germany on video only in a version trimmed by 15 minutes of plot under the stupid title "Horror House". At least the murder scenes, which will satisfy every gorehound, are fully intact, and the viewer still gets the killer's motive at the end. But the Italian version containing all the footage is still the one to look for, of course.

A convincing Giallo with obligatory twists and red herrings, "La Dama Rossa Uccide Sette Volte" is highly recommended to Giallo fans and slightly superior to Miraglia's above mentioned other thriller.

- Crítica Sem Mudanças de Linha *(179 palavras)*

This interesting Giallo boosts a typical but still thrilling plot and a really sadistic killer that obviously likes to hunt his victims down before murdering them in gory ways. Directed by Emilio P. Miraglia who, one year earlier, also made the very interesting "La Notte che Evelyn Usci della Tomba" (see also my comment on that one), the film starts off a little slow, but all in all, no time is wasted with unnecessary sub plots or sequences. This film is a German-Italian coproduction, but it was released in Germany on video only in a version trimmed by 15 minutes of plot under the stupid title "Horror House". At least the murder scenes, which will satisfy every gorehound, are fully intact, and the viewer still gets the killer's motive at the end. But the Italian version containing all the footage is still the one to look for, of course. A convincing Giallo with obligatory twists and red herrings, "La Dama Rossa Uccide Sette Volte" is highly recommended to Giallo fans and slightly superior to Miraglia's above mentioned other thriller.

Limpeza dos Dados
(exemplo 2)

- Crítica Original *(182 palavras)*

This interesting Giallo boosts a typical but still thrilling plot and a really sadistic killer that obviously likes to hunt his victims down before murdering them in gory ways.

Directed by Emilio P. Miraglia who, one year earlier, also made the very interesting "La Notte che Evelyn Usci della Tomba" (see also my comment on that one), the film starts off a little slow, but all in all, no time is wasted with unnecessary sub plots or sequences.

This film is a German-Italian coproduction, but it was released in Germany on video only in a version trimmed by 15 minutes of plot under the stupid title "Horror House". At least the murder scenes, which will satisfy every gorehound, are fully intact, and the viewer still gets the killer's motive at the end. But the Italian version containing all the footage is still the one to look for, of course.

A convincing Giallo with obligatory twists and red herrings, "La Dama Rossa Uccide Sette Volte" is highly recommended to Giallo fans and slightly superior to Miraglia's above mentioned other thriller.

- Limpeza alfanumérica *(112 palavras)*

They really can t get stupider than this film dealing with **3** losers who try to capture the college spirit during the annual spring break festivities at many of our higher schools of learning The problem is that these losers try to do this **15** years after their college years when one is assigned to watch over the daughter of a woman senator being groomed to be the next vice president Trouble is that her daughter is anything but popular but of course she comes out of all that The girls go through drunken rages exotic dancing and other absolute nonsense It really can t get much worse than this awful film

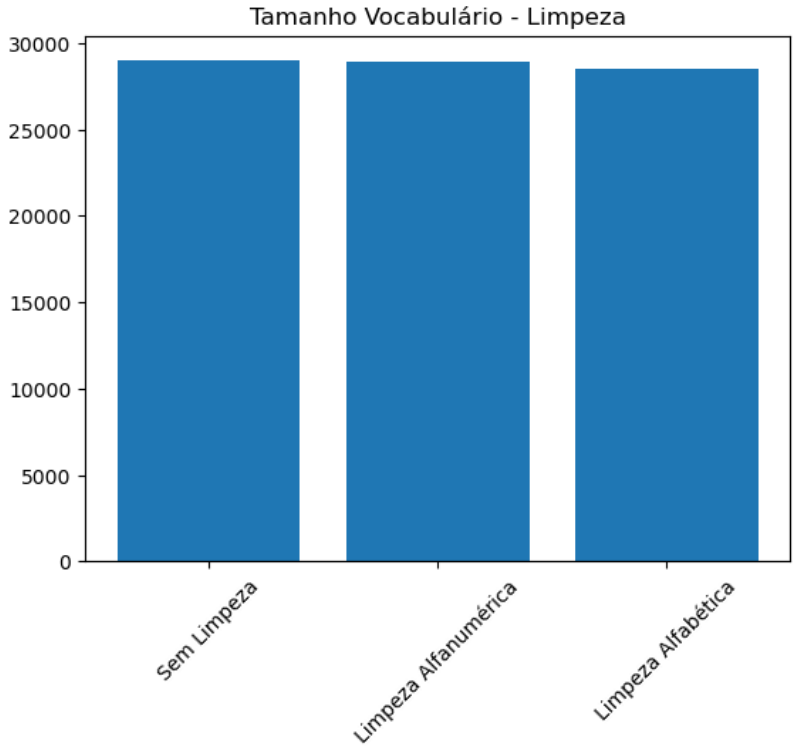
- Limpeza alfabética (110 palavras)

They really can t get stupider than this film dealing with losers who try to capture the college spirit during the annual spring break festivities at many of our higher schools of learning The problem is that these losers try to do this years after their college years when one is assigned to watch over the daughter of a woman senator being groomed to be the next vice president Trouble is that her daughter is anything but popular but of course she comes out of all that The girls go through drunken rages exotic dancing and other absolute nonsense It really can t get much worse than this awful film

Limpeza dos Dados
(comparação)

Maior vocabulário:
Sem limpeza

Menor vocabulário:
Limpeza Alfabética



Stop Words

- Palavras com maior frequência na utilização do inglês.
  - Ex: “a” e “the”
- Usadas num contexto genérico e não específico;
- Não são úteis para a análise de críticas.

```
Tamanho do dicionário: 28521
Mean cross-validation accuracy: 0.83
```

```
Tamanho do dicionário: 28509
Mean cross-validation accuracy: 0.85
```

- Impacto vocabulário: reduzido;
- Impacto esperado na classificação: elevado.

Tokenização

Porter Stemmer
Snowball Stemmer
Lancaster Stemmer

- Processo que reduz palavras à sua raiz.
- Não considera a linguística. Ou seja, pode gerar tokens linguisticamente incorretos ou palavras que não existem.
  - Ex: “car” e “care” ambas são convertidas para “car”.

- Foram avaliados 3 formas de tokenização:
  - Porter Stemmer
    - Mais usado
    - Simples
    - Rápido de executar
    - Suporta apenas inglês
  - Snowball Stemmer
    - Versão melhorada do Porter
    - Mais complexo
    - Mais lento
  - Lancaster Stemmer
    - Abordagem mais agressiva
    - Pode levar a *over-stemming*
    - Suporta apenas inglês

```
Tamanho do dicionário: 19613
Mean cross-validation accuracy: 0.84
```

```
Tamanho do dicionário: 19316
Mean cross-validation accuracy: 0.84
```

```
Tamanho do dicionário: 16222
Mean cross-validation accuracy: 0.85
```

Lematização
WordNetLemmatizer

- Processo identifico à tokenização
- Considera a composição natural das palavras
  - Ex: “car” e “care” são convertidos para a sua forma canónica.
- Palavras não perdem o contexto.

```
Tamanho do dicionário: 24272
Mean cross-validation accuracy: 0.85
```

- WordNetLemmatizer é o módulo da biblioteca NLTK que é utilizado e que permite efetuar a lematização.

Tokenização Vs Lematização

Maior vocabulário:

Sem tokenização/ lematização

Menor vocabulário:

Lancaster Stemmer

Melhor performance:

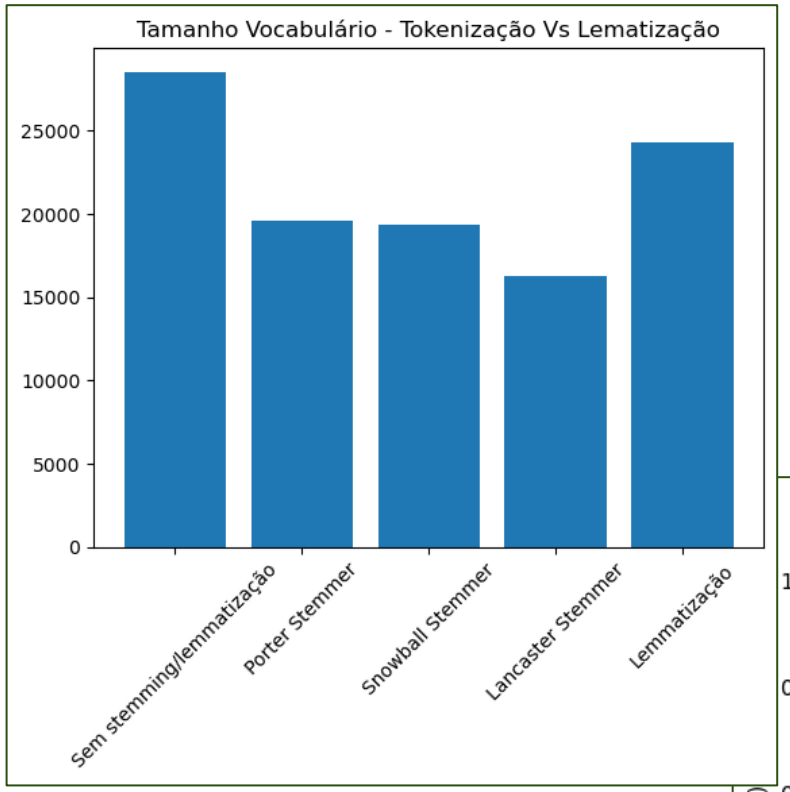
Porter Stemmer

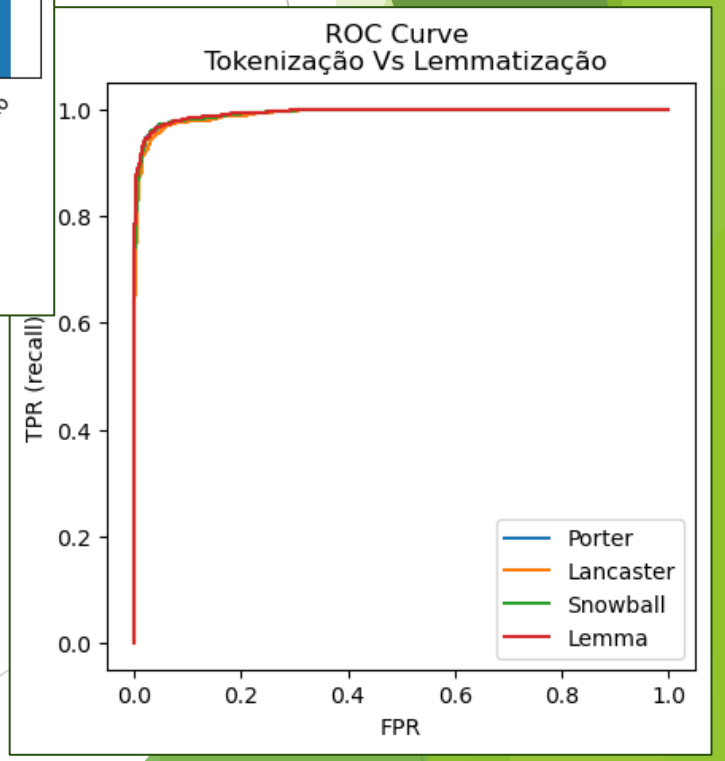
Lematização

Pior Performance:

Lancaster Stemmer

Snowball Stemmer





Limpeza dos Dados (conclusões)

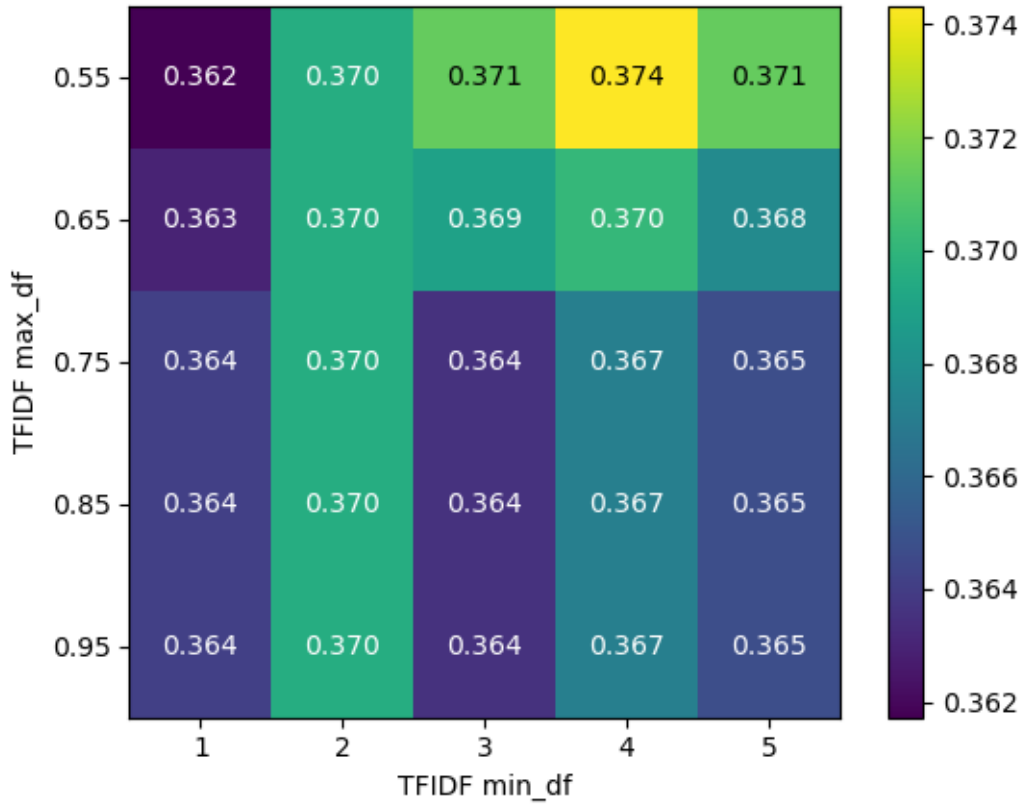
- Informação numérica nas críticas não é a mais importante a ser mantida;
- Limpeza alfabética diminui o tamanho do vocabulário;
- Significado associado ao contexto da crítica é importante para discriminar as várias críticas.
- Embora lematização não diminua muito o tamanho do vocabulário mantém o contexto.

Processo escolhido:

1. Remoção de mudanças de linha
2. Limpeza alfabética
3. Stop Words
4. Lematização – NetWordLemmatizer

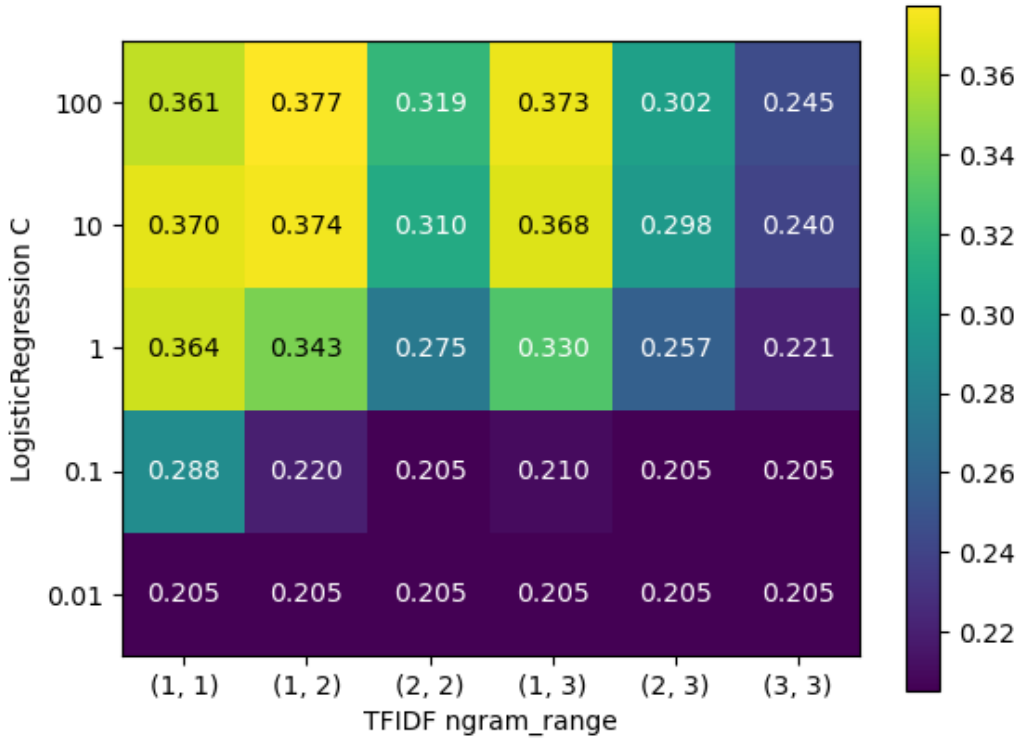
Extração de Características

TFIDF
Min_df e Max_df



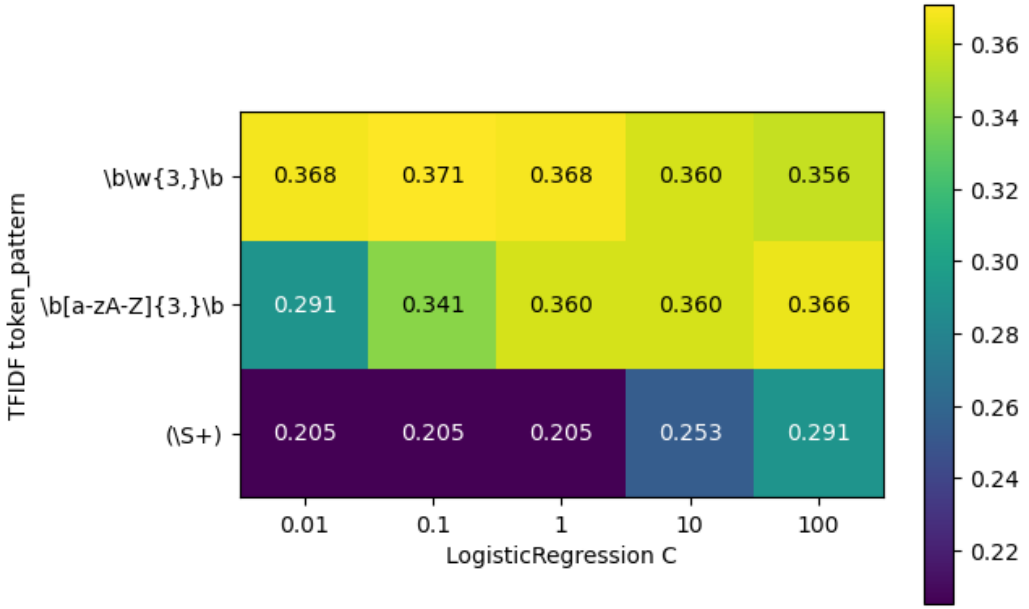
- Melhores valores para min_df = [3, 4, 5]
- Melhores valores para max_df = [0.55, 0.65, 0.75]

TFIDF ngram_range



- Melhores valores para ngram_range = [(1, 1) , (1, 2), (1, 3)]

TFIDF token_pattern



- Melhores valores para token_pattern = r'\b\w{3,}\b'

TFIDF (conclusões)

- TFIDF com parâmetros default:

```
Score Treino:  0.87
Score Teste:  0.362
```

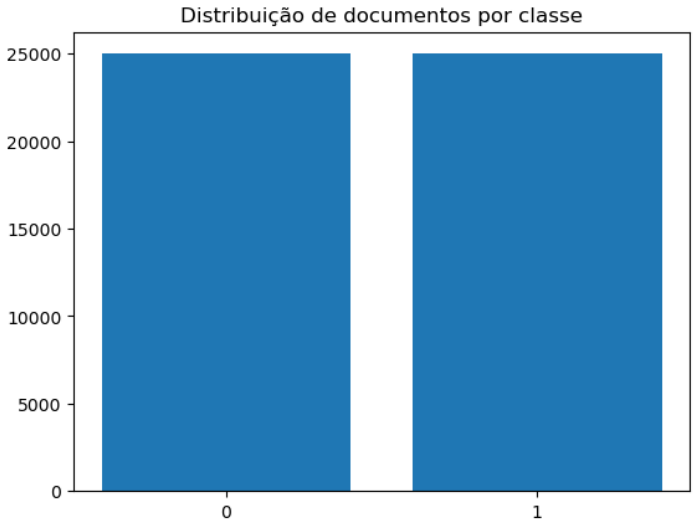
Parametros escolhidos:

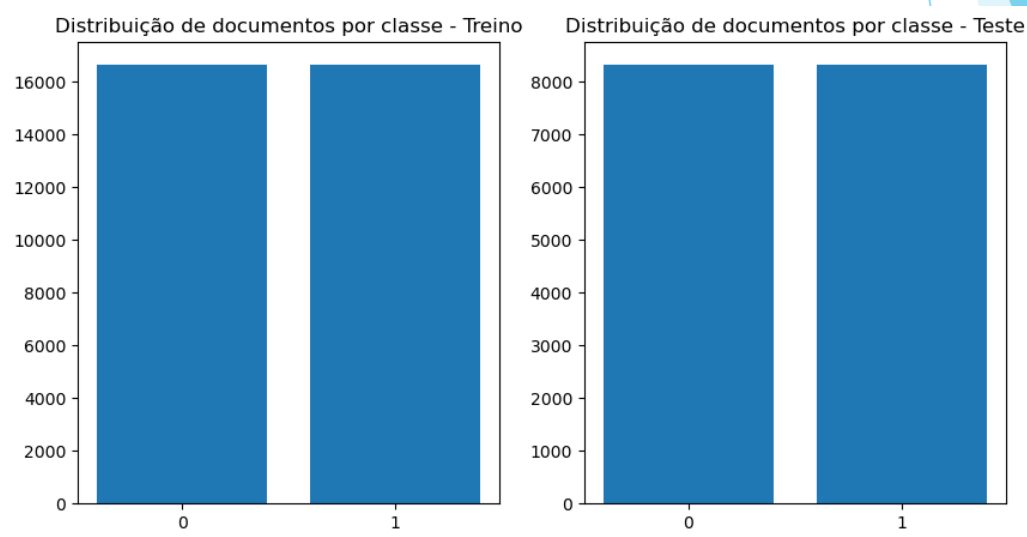
- min_df=3
- max_df=0.75
- ngram_range=(1,3)
- token_pattern=r'\b\w{3,}\b'

- TFIDF parametrizado:

```
Score Treino:  0.894
Score Teste:  0.36
```

Classificação Binária





Distribuição uniforme das críticas positivas e negativas

<u>Dados de treino</u>

2/3 das críticas

<u>Dados de teste</u>

1/3 das críticas

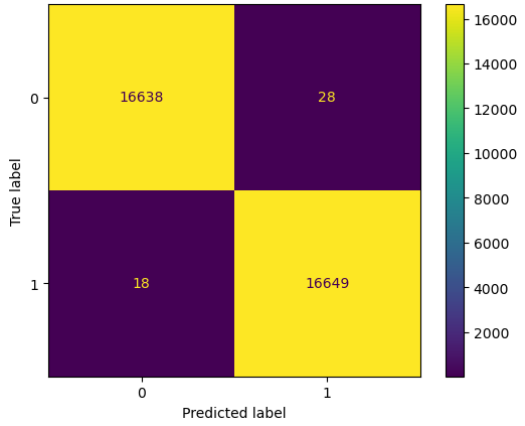
Logistic Regression

- Logistic Regression com parâmetros default:

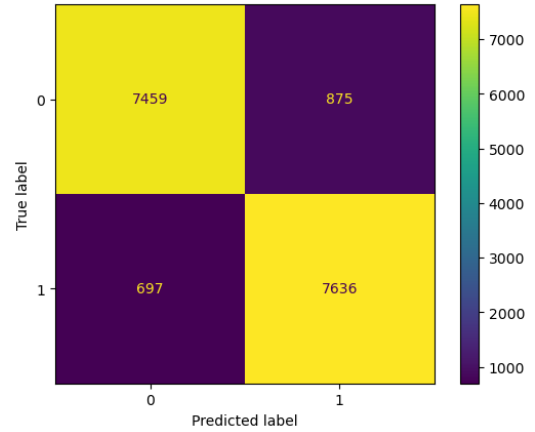
```
Mean cross-validation accuracy: 0.84
```

- Parametrização:
  - C=10
  - penalty='l2'
  - solver='liblinear'
  - tol=1e-3

- Treino
  - Score: 0.999
  - Número de erros: 46
  - F1-Score: 1.00



- Teste
  - Score: 0.905
  - Número de erros: 1572
  - F1-Score: 0.91



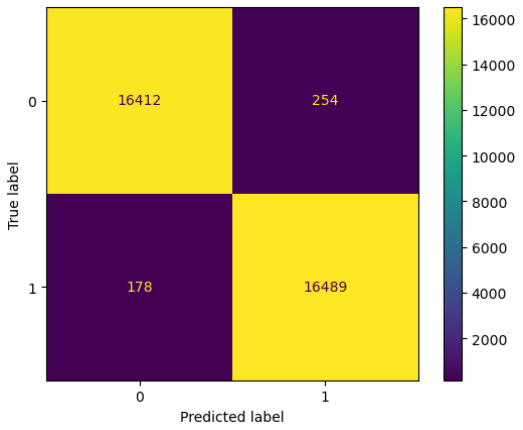
Máquina de Suporte Vetorial Linear (LinearSVC)

- LinearSVC com parâmetros default:

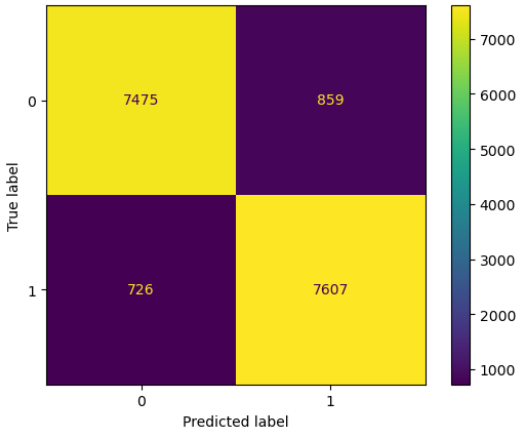
```
Mean cross-validation accuracy: 0.86
```

- Parametrização:
  - C=0.01
  - loss='hinge'
  - tol=0.001

- Treino
  - Score: 0.987
  - Número de erros: 432
  - F1-Score: 0.99



- Teste
  - Score: 0.905
  - Número de erros: 1585
  - F1-Score: 0.90



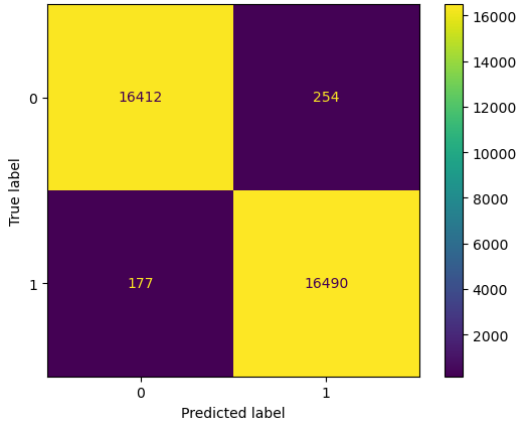
Máquina de Suporte Vetorial (SVC)

- SVC com parâmetros default:

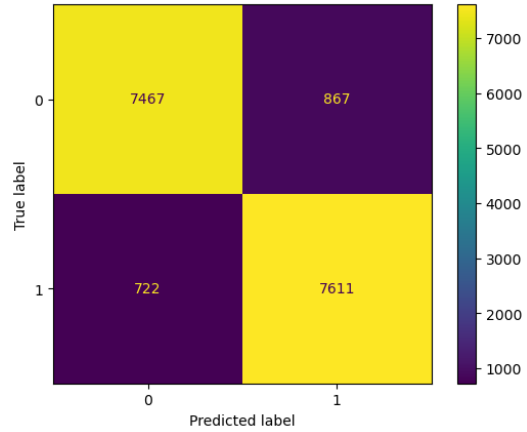
```
Mean cross-validation accuracy: 0.85
```

- Parametrização:
  - C=10
  - degree=1
  - gamma=10
  - kernel='poly'

- Treino
  - Score: 0.987
  - Número de erros: 431
  - F1-Score: 0.99



- Teste
  - Score: 0.905
  - Número de erros: 1589
  - F1-Score: 0.90



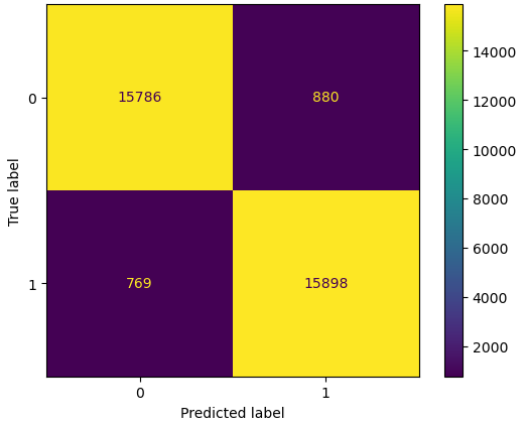
Naive Bayes (MultinomialNB)

- MultinomialNB com parâmetros default:

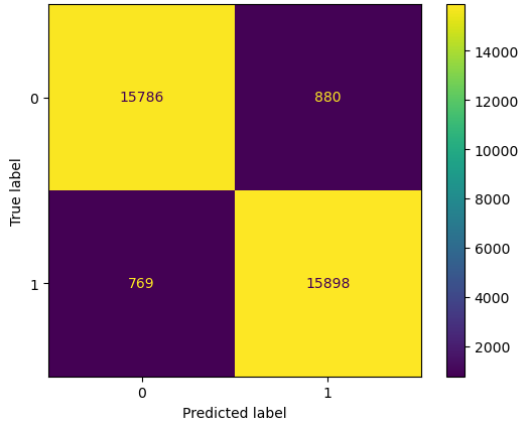
```
Mean cross-validation accuracy: 0.86
```

- Parametrização:
  - alpha=0.1

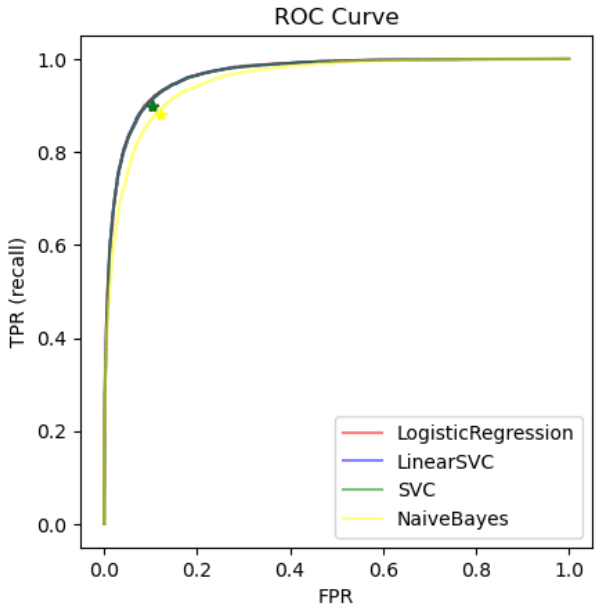
- Treino
  - Score: 0.951
  - Número de erros: 1649
  - F1-Score: 0.95



- Teste
  - Score: 0.885
  - Número de erros: 1915
  - F1-Score: 0.89



Classificador Binário (conclusões)



★★★

Melhor Performance:

Logistic Regression

LinearSVC

SVC

Menos Erros:

Logistic Regression

★☆☆

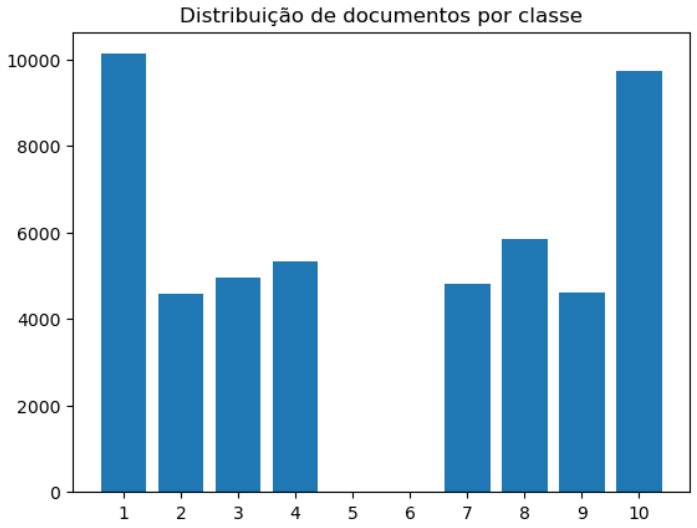
Pior Performance:

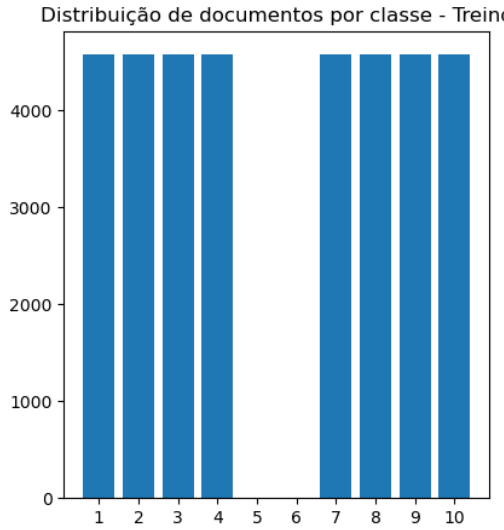
Naive Bayes

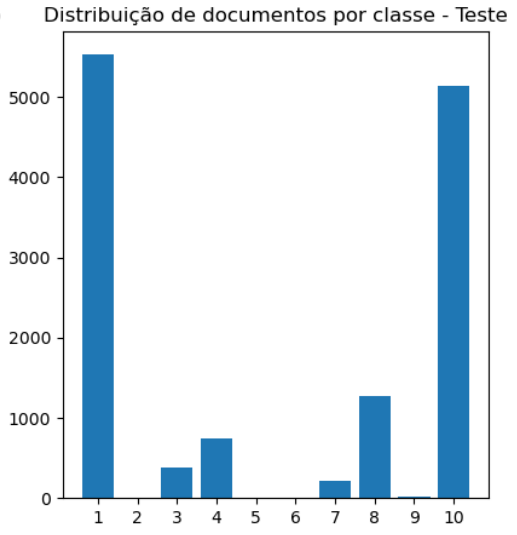
Mais erros:

Naive Bayes

Classificação Multiclasse







Mais amostras de umas classes do que de outras

<u>Dados de treino</u>

1/2 número mínimo amostras de todas as classes

<u>Dados de teste</u>

Restantes críticas

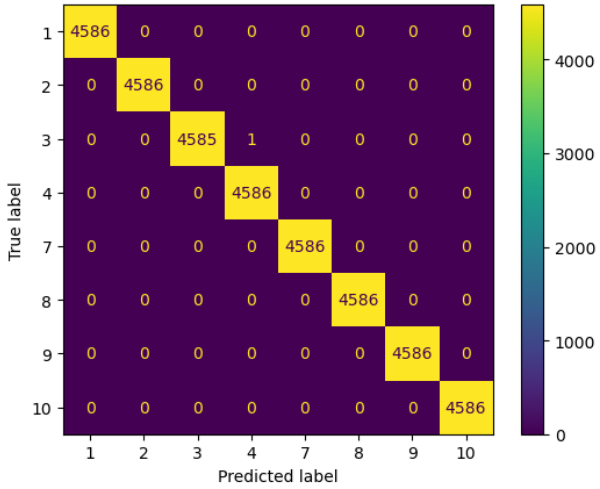
Logistic Regression

- Logistic Regression com parâmetros default:

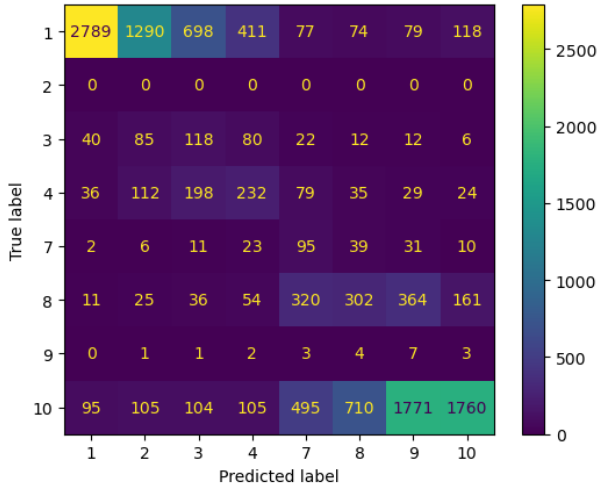
```
Mean cross-validation accuracy: 0.38
```

- Parametrização:
  - C=100
  - multi_class='ovr'
  - solver='saga'
  - tol=0.001

- Treino
  - Score: 1.0
  - Número de erros: 1



- Teste
  - Score: 0.398
  - Número de erros: 8009



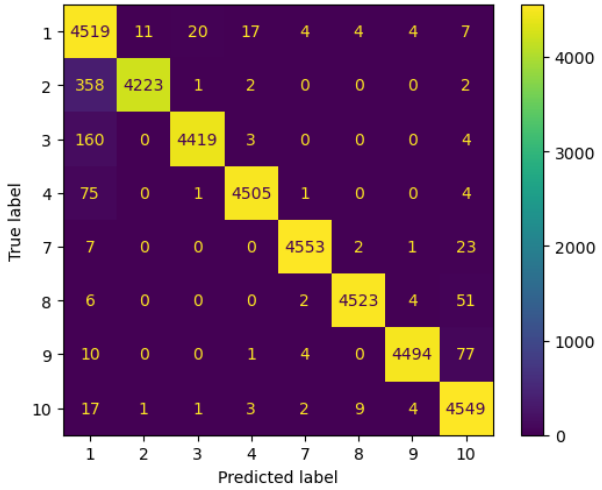
Máquina de Suporte Vetorial Linear (LinearSVC)

- LinearSVC com parâmetros default:

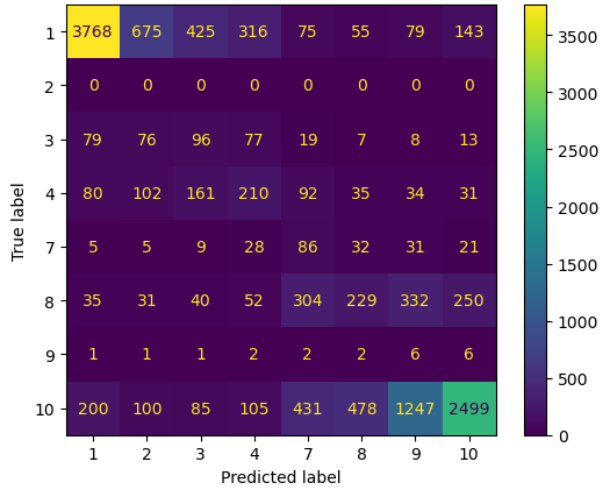
```
Mean cross-validation accuracy: 0.37
```

- Parametrização:
  - C=0.01
  - loss='hinge'
  - tol=0.001

- Treino
  - Score: 0.975
  - Número de erros: 903



- Teste
  - Score: 0.518
  - Número de erros: 6418

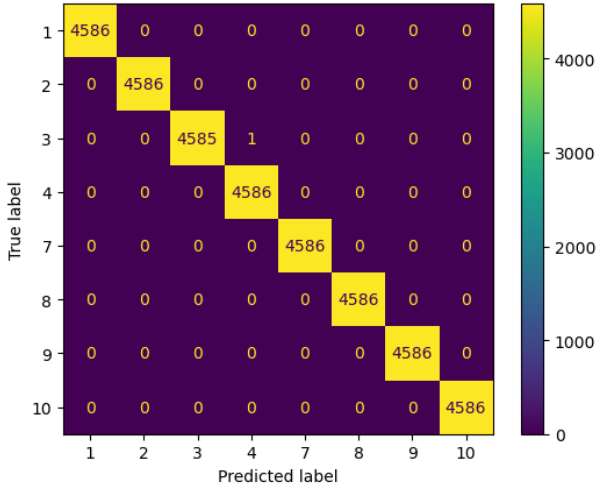


Máquina de Suporte Vetorial (SVC)

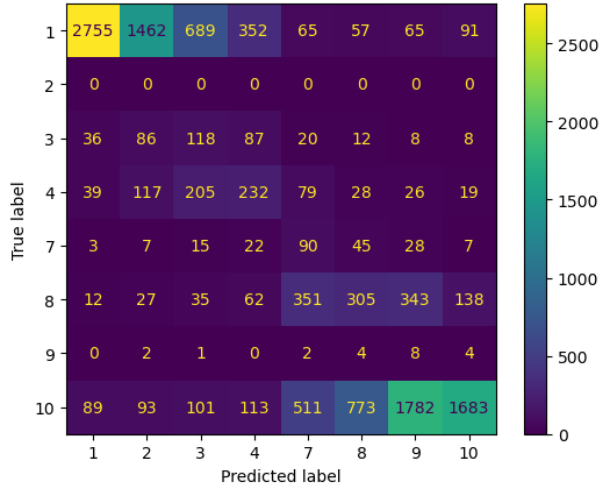
- SVC com parâmetros default:

```
Mean cross-validation accuracy: 0.36
```

- Parametrização:
  - C=10
  - degree=1
  - gamma=10
  - kernel='poly'
- Treino
  - Score: 1.0
  - Número de erros: 1



- Teste
  - Score: 0.39
  - Número de erros: 8121



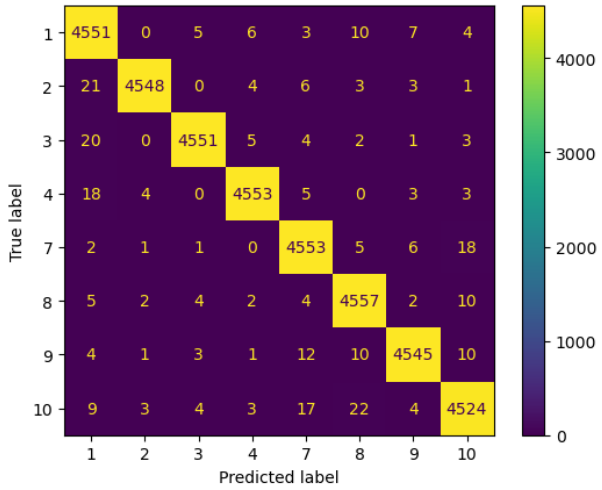
Naive Bayes (MultinomialNB)

- MultinomialNB com parâmetros default:

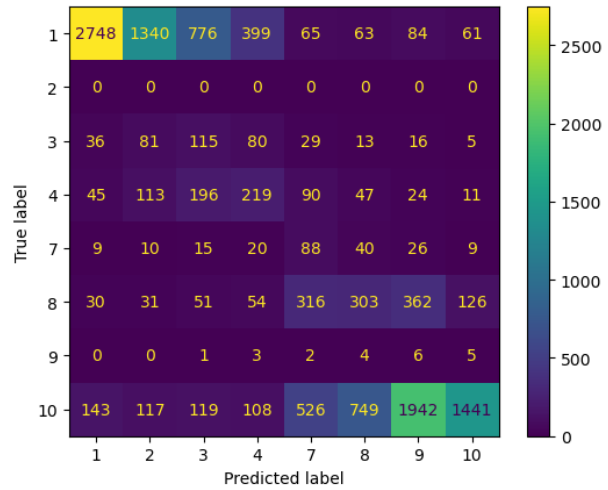
```
Mean cross-validation accuracy: 0.33
```

- Parametrização:
  - alpha=0.1

- Treino
  - Score: 0.992
  - Número de erros: 306



- Teste
  - Score: 0.37
  - Número de erros: 8392



Classificador Multiclasse (conclusões)

Melhor Performance:

LinearSVC

Menos Erros:

LinearSVC

Pior Performance:

Naive Bayes

Mais erros:

Naive Bayes