Classificação de Críticas de Cinema do IMDb

Trabalho Laboratorial

Aprendizagem Automática

Maria Franco A46320

José Siopa A46338

Construção do vocabulário

- Carregamento dos dados
- Divisão dos dados em treino e teste
- Limpeza dos dados de texto:

    - Remoção das mudanças de linha e de todos os caracteres não pertencentes ao alfabeto latino

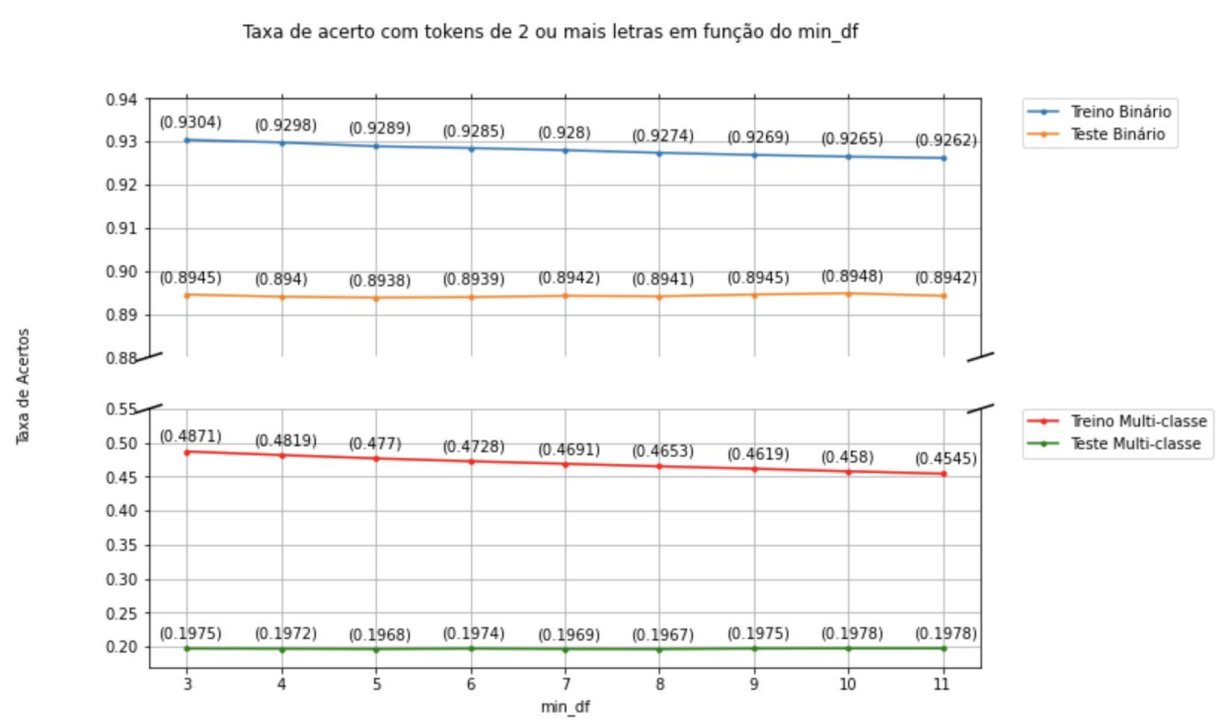
    - Representação *Bag of Words* e *tf-idf*

Função `TfidfVectorizer`

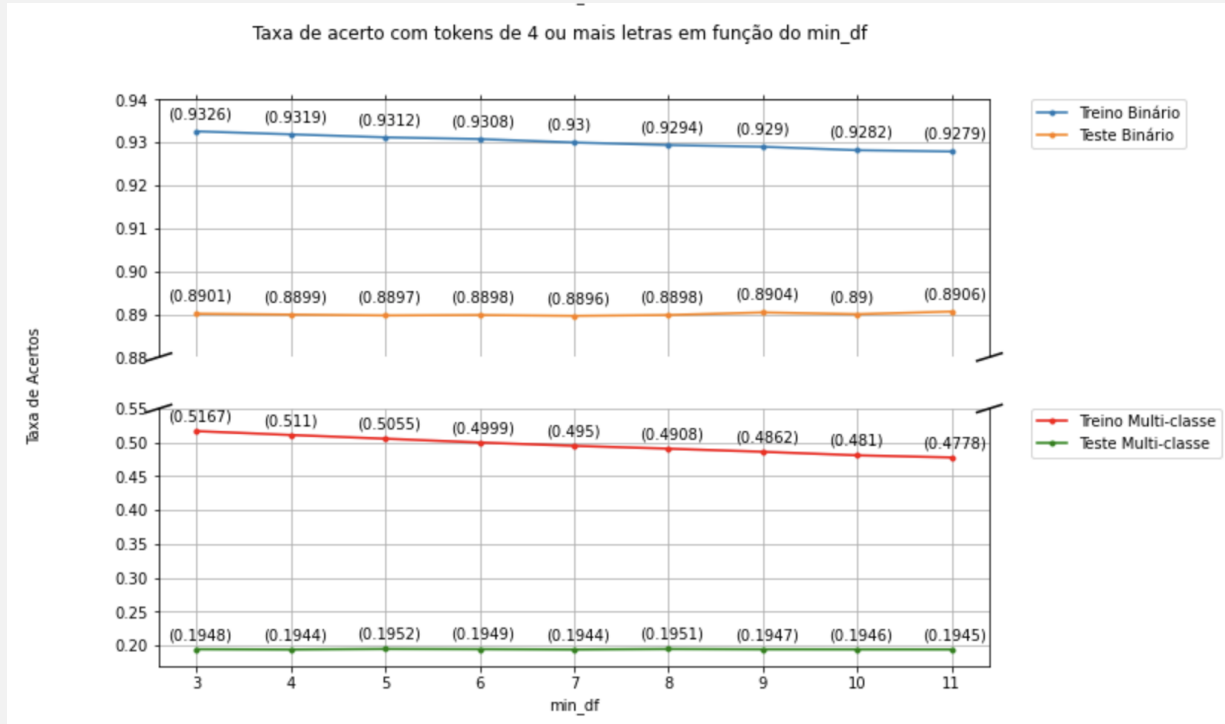
- Permite efetuar, simultaneamente, a representação BoW e tf-idf e coloca todo o vocabulário dos documentos em letra minúscula.
- Contém híper-parâmetros importantes:
  - `min_df` – minimum document frequency – número mínimo de documentos em que um dado *token* é utilizado;
  - `token_pattern` – permite restringir o número mínimo de caracteres presentes numa palavra para ser aceite como *token*.
- Para achar os melhores parâmetros, é necessário experimentar para diferentes valores, quais aqueles que para o mesmo classificador apresentem os melhores resultados.

Variação dos parâmetros da função `TfidfVectorizer`

Independentemente do `token_pattern`, verifica-se uma ligeira diminuição da taxa de acertos com o ligeiro aumento do `min_df`



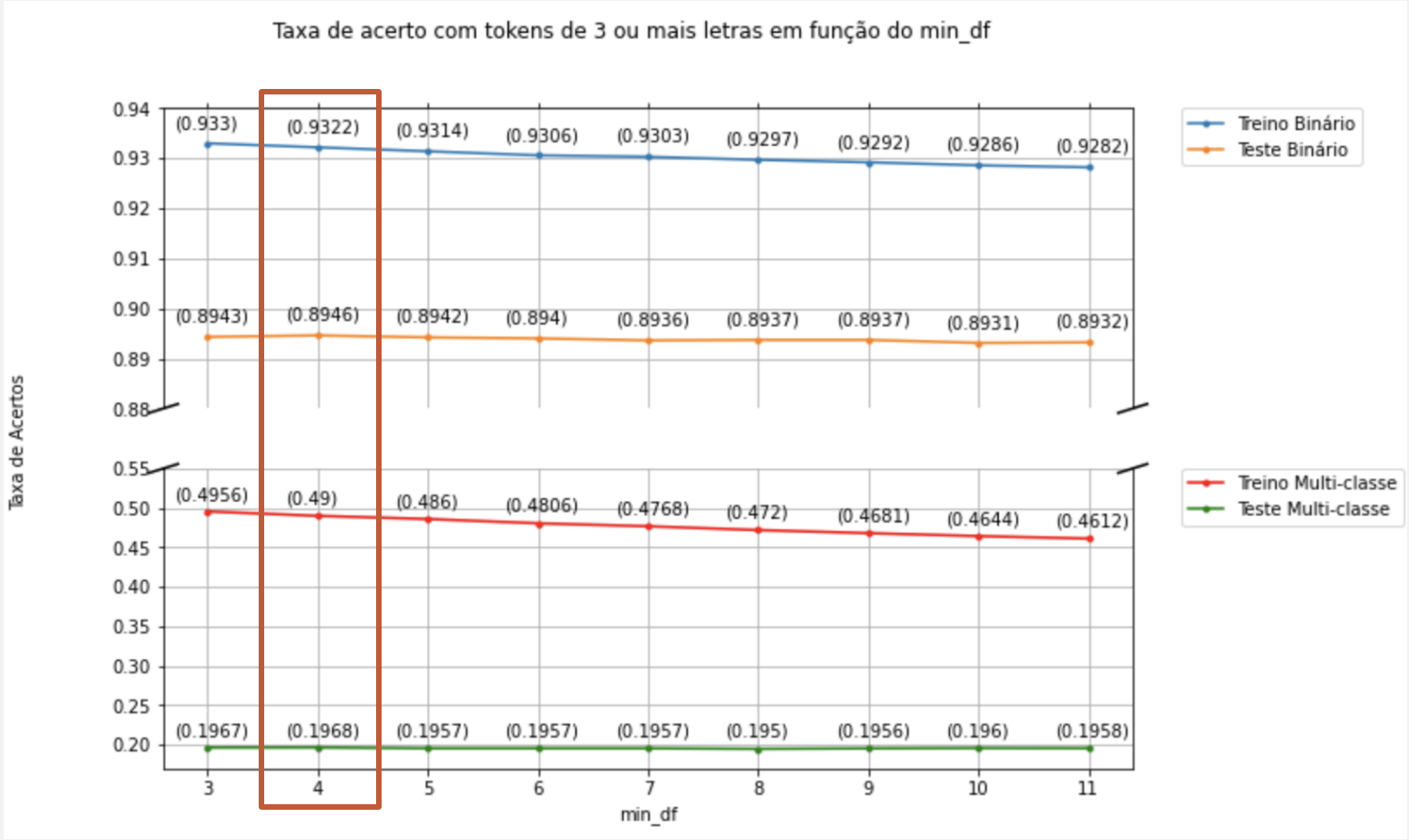
Taxas de acerto utilizando *tokens* de 2 ou mais letras em função de diferentes valores para `min_df`



Taxas de acerto utilizando *tokens* de 4 ou mais letras em função de diferentes valores para `min_df`

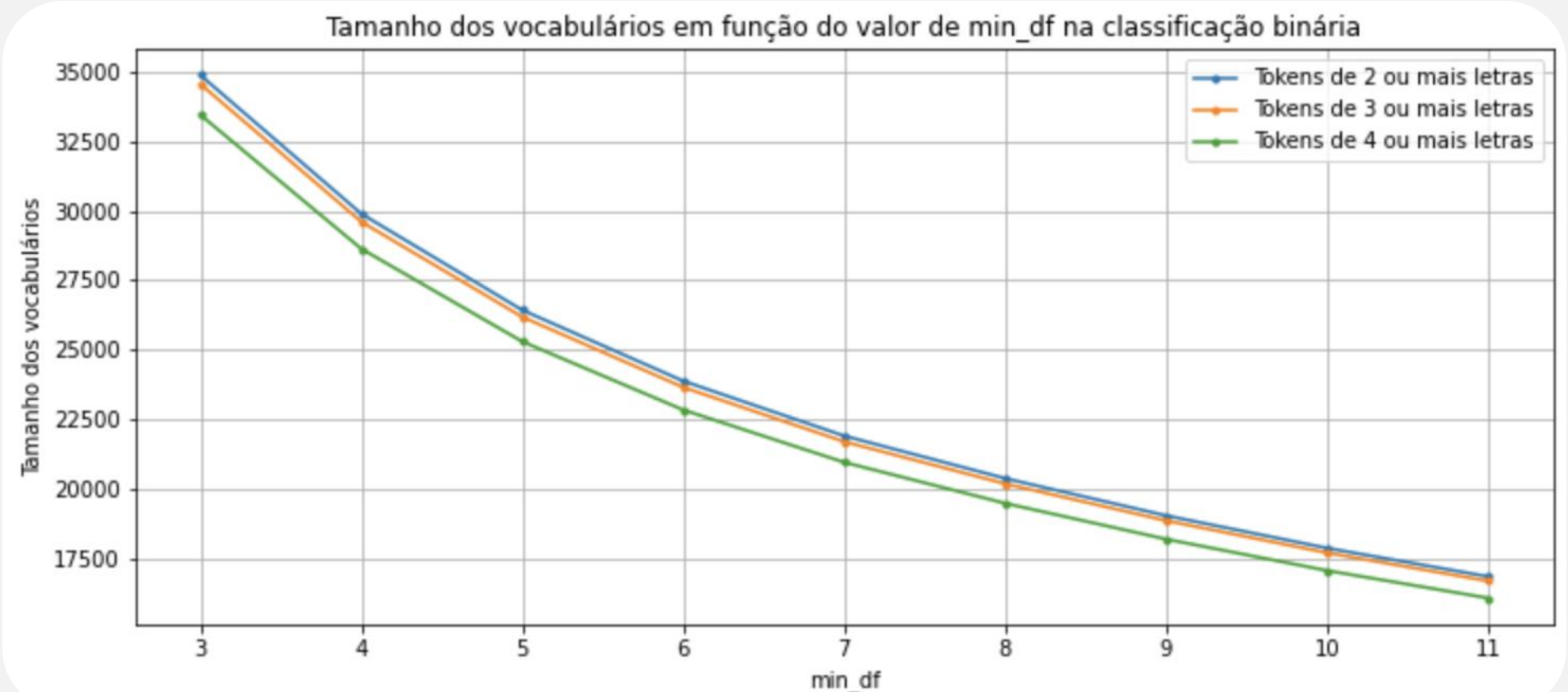
Valores escolhidos para a função `TfidfVectorizer`

- Os melhores parâmetros para a função `TfidfVectorizer` são aqueles que apresentem uma menor diferença entre as taxas de acerto em treino e teste, mas tendo os valores mais altos.
- Assim, o melhor valor para `min_df` é 4 para *tokens* de 3 ou mais letras.

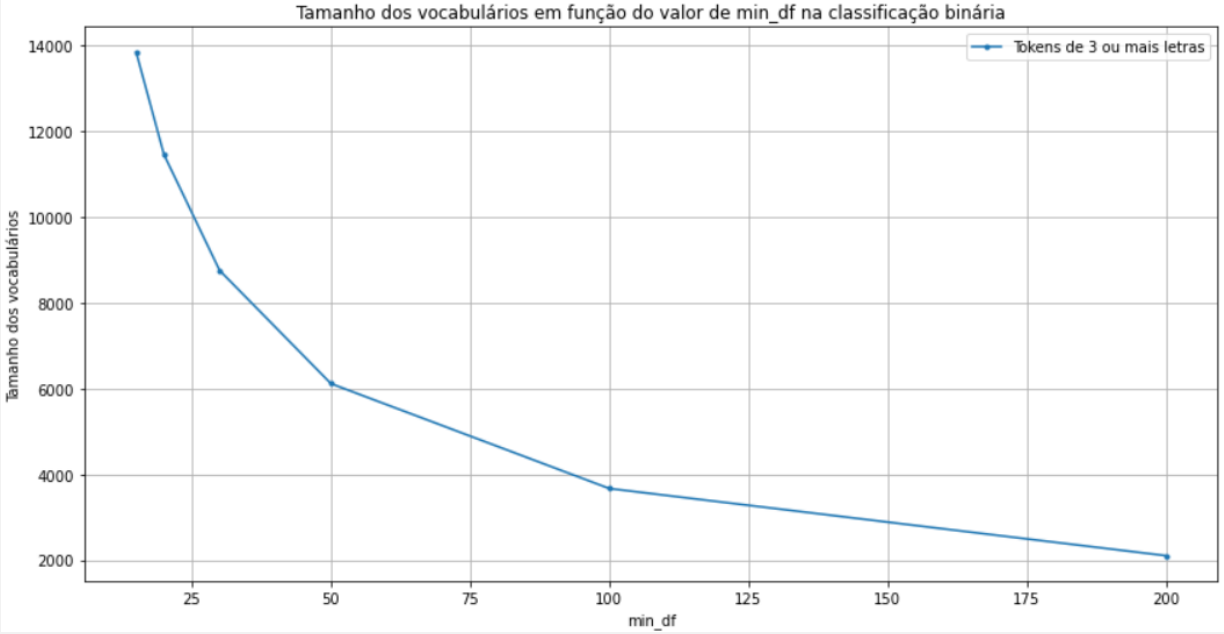


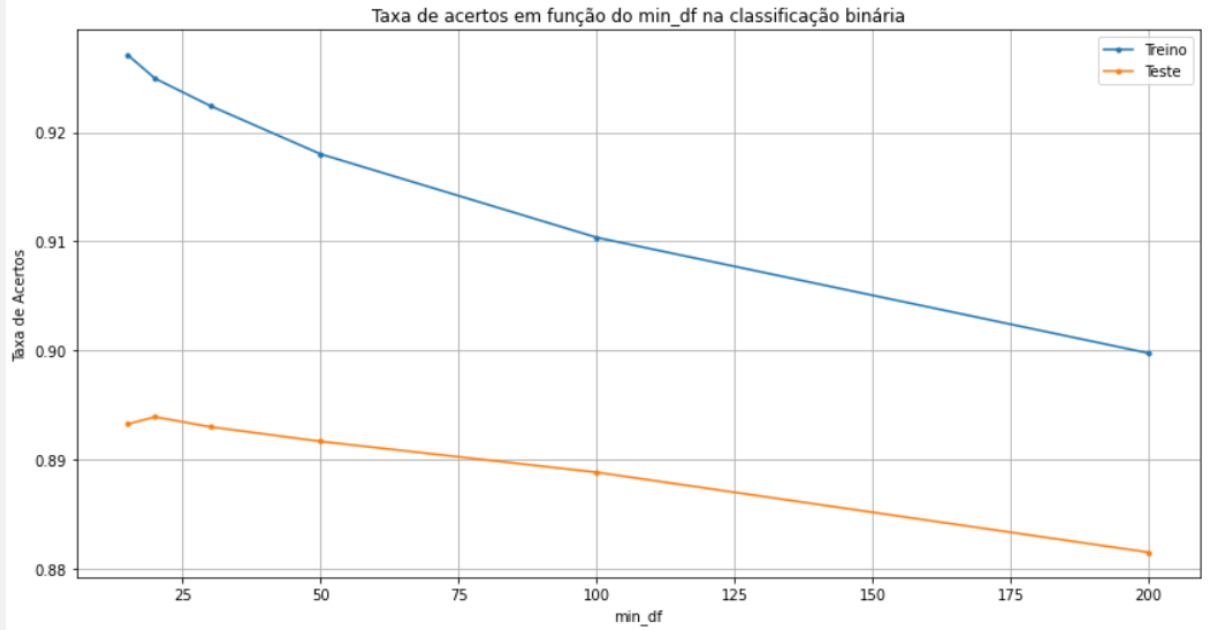
Variação da dimensão do vocabulário com a variação dos parâmetros da função `TfidfVectorizer`

Independentemente do `token_pattern`, verifica-se uma diminuição exponencial da dimensão do vocabulário com o aumento do `min_df`



Dimensão mínima do vocabulário com obtenção de bons resultados





- Melhor resultado obtido, na classificação binária: 0.9322 em treino e 0.8946 em teste, com *token pattern* de 3 ou mais letras.
- Variando o parâmetro `min_df` para 100: apenas **menos 2**% de acertos em **treino** e **1**% em **teste**, com *token pattern* de 4 ou mais letras.
- Dimensão do vocabulário diminuída em 87,6%.

Inclusão de n-gramas

- A inclusão de n-gramas poderia trazer melhores resultados na classificação. Porém, iria aumentar demasiado a dimensão do vocabulário, o que seria incomportável, a nível de complexidade computacional.

	Uni-gramas	Bi-gramas	Tri-gramas	De uni a bi-gramas	De uni a tri-gramas	De uni a quadri-gramas
Treino	93.2	96.3	96.2	94.9	95.1	95.5
Teste	89.5	88.4	82.5	89.4	89.5	90.0
Dimensão do vocabulário	29601	141839	81743	135047	160627	271971
Token pattern	3+	3+	3+	4+	4+	3+

A taxa de acertos encontra-se em percentagem.

- Aumentando o valor de `min_df`, por exemplo, para 30, verificou-se uma ligeira diminuição da dimensão do vocabulário, mas os resultados pioraram.

Metodologias de teste e métricas de desempenho

- Para obtermos os resultados mostrados nos diapositivos anteriores, utilizámos sempre a mesma metodologia de teste:
  - Divisão dos dados: 70% para treino e 30% para teste
  - Classificador de regressão logística com C = 1.
- Não se pretende dar mais importância a críticas positivas mal classificadas ou vice-versa, no contexto deste problema.
- Métrica de desempenho utilizada: *accuracy*, ou taxa de acertos:

$$ACC = \frac{tp + tn}{tp + fp + tn + fn}$$

Classificadores binários

- LogisticRegression

Utilizou-se a regularização Ridge, na função LogisticRegressionCV, que determinou que o melhor valor para C era 3.

Pode-se ver que, com os mesmos parâmetros mas com regularização Lasso, obtêm-se resultados ligeiramente piores.

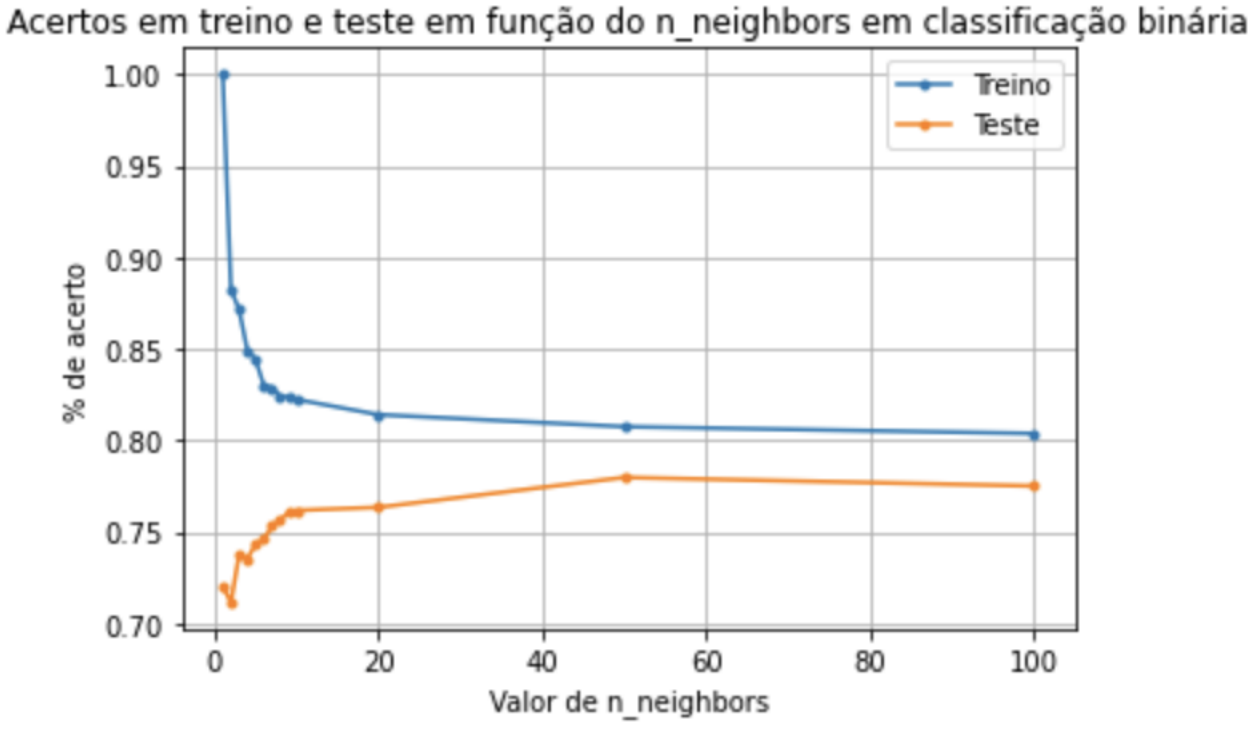
	C=3 c/ valid. cruzada e regularização Ridge	C=3 c/ valid. cruzada e regularização Lasso
Treino	95.8	88.6
Teste	89.9	87.1
Teste com função `accuracy_score`	89.9	-
Teste com função `balanced_accuracy_score`	89.9	-

A taxa de acertos encontra-se em percentagem.

Ainda em relação às métricas de desempenho, pode-se verificar que as funções `score` e `accuracy_score` dão resultados iguais. A `balanced_accuracy_score` dá um resultado muito parecido, o que mostra que a diferença entre o número de críticas positivas e negativas não é significativa para que se tenha de usar esta função em detrimento das anteriores.

Classificadores binários

- K-nearest neighbors

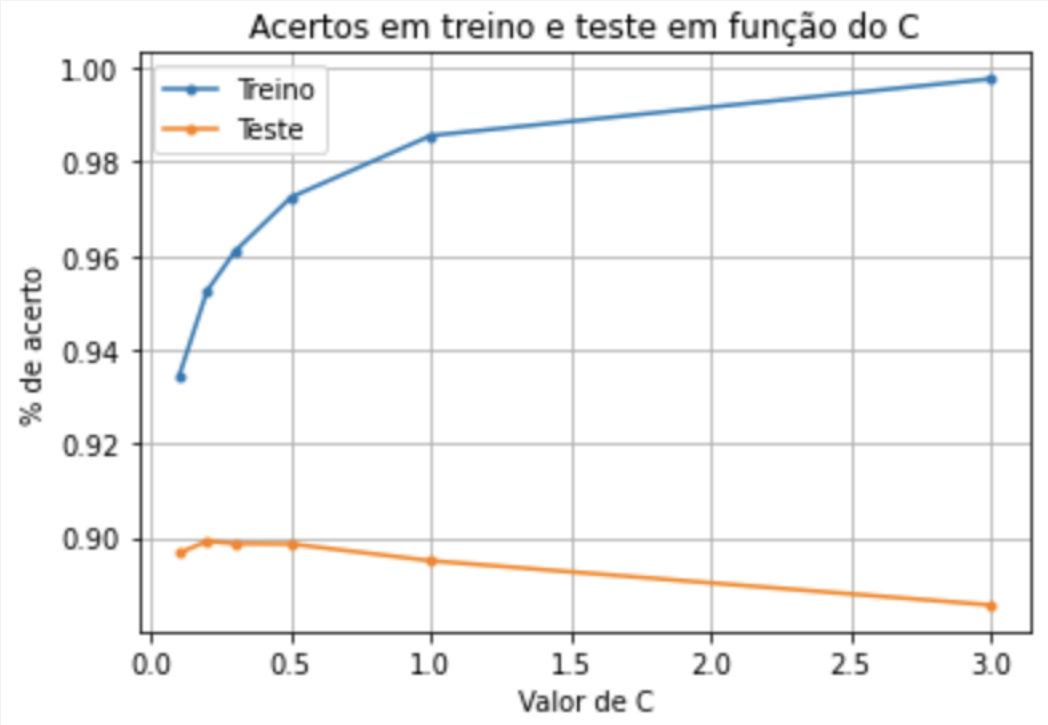


O KNN é um classificador que se mostra suscetível a sobre aprender, caso tenha o parâmetro relativo ao número de vizinhos menor que 10.

- Considerámos que os melhores resultados foram com `n_neighbors` a 50.

Classificadores binários

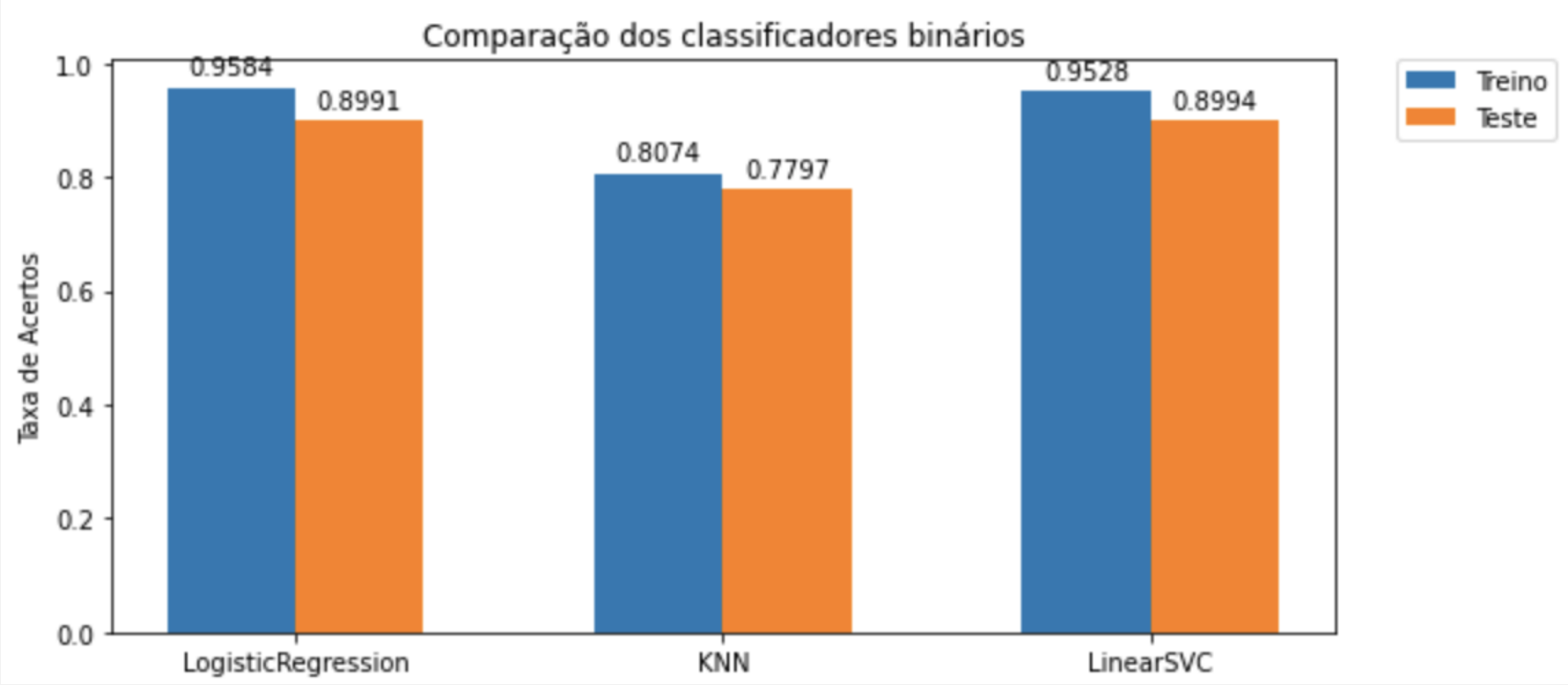
- LinearSVC



O LinearSVC é um classificador que se mostra suscetível a sobre aprender, caso tenha o parâmetro C acima de 1.

- Considerámos que os melhores resultados de entre os valores de C testados, com o mesmo a 0.2.

Melhor dos classificadores binários



- O classificador LinearSVC é aquele que apresenta o melhor valor em teste e tem um menor valor de diferença entre este e o de treino. Logo, este é o melhor classificador binário de entre os testados.

Classificadores multi-classe

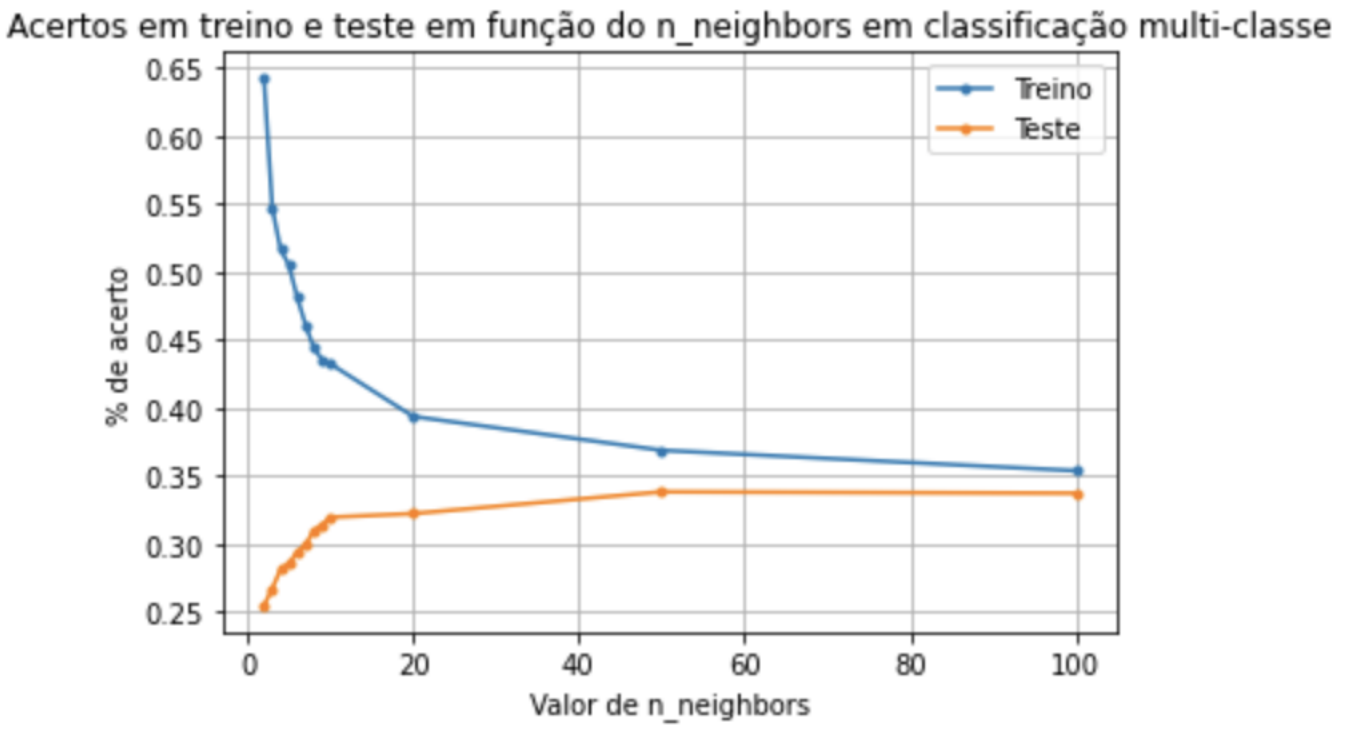
- LogisticRegression
- Com o parâmetro C a 3, que foi valor que apresentou melhores resultados na classificação binária, neste verifica-se sobre aprendizagem.
- O valor ótimo que encontrámos foi de 1, com validação cruzada.

	C=3	C=1 c/ valid. cruzada	C=1.03 s/ validação cruzada	C=1.03 c/ validação cruzada
Treino	85.2	69.5	69.9	43.6
Teste	43.1	44.4	44.5	42.4

A taxa de acertos encontra-se em percentagem.

Classificadores multi-classe

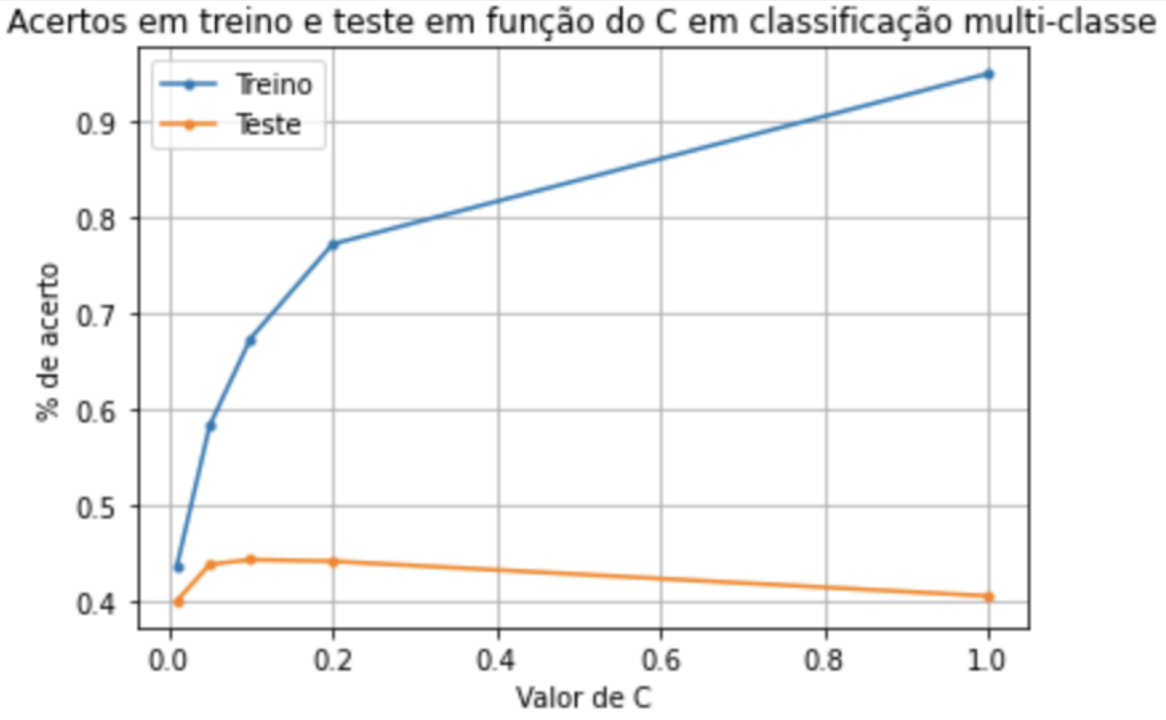
- K-nearest neighbors



- Tal como na classificação binária, o parâmetro `n_neighbors` abaixo de 10 provoca sobre aprendizagem.
- Novamente, consideramos que os melhores resultados foram com `n_neighbors` a 50.

Classificadores multi-classe

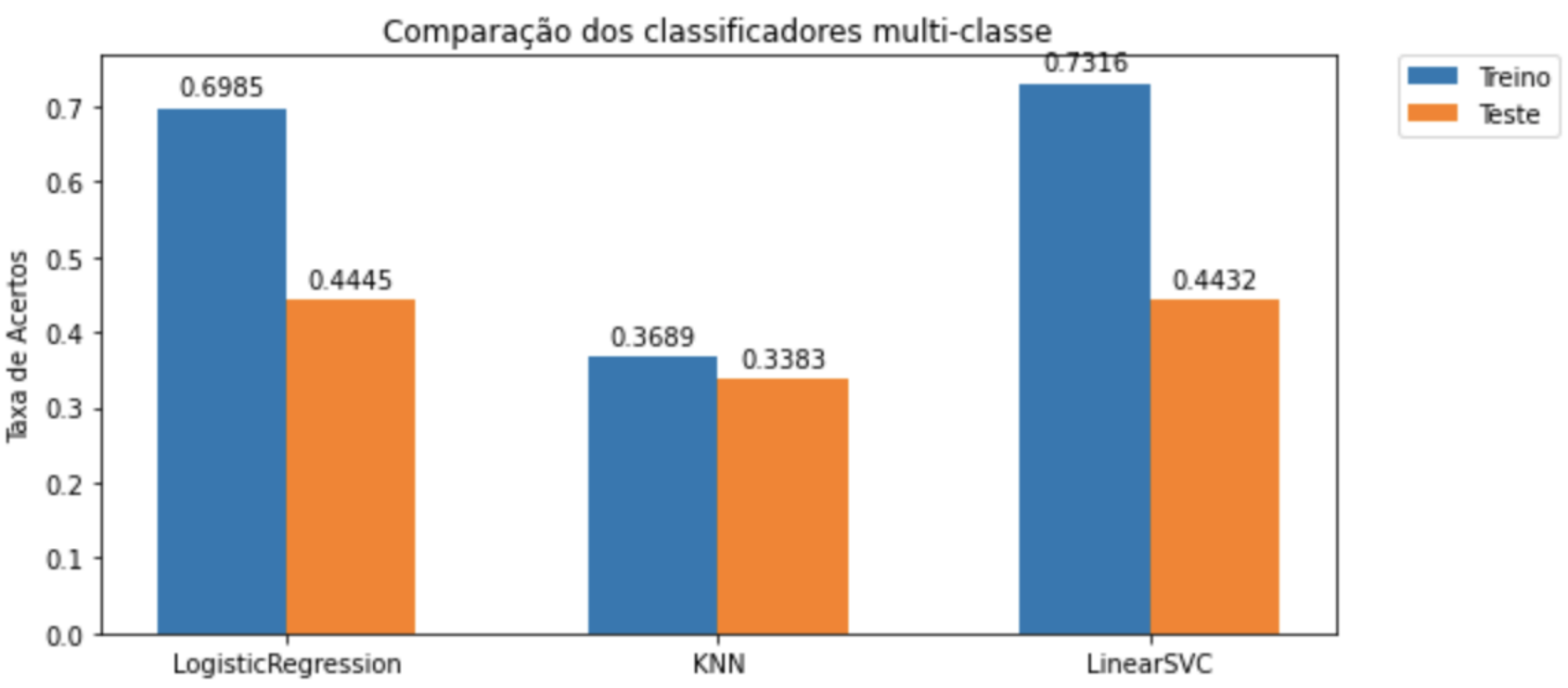
- LinearSVC



O LinearSVC , mais uma vez, é mostra-se suscetível a sobre aprender, caso tenha o parâmetro C acima de 1. Assim, testámos vários valores entre 0 e 1.

- Considerámos que os melhores resultados foram, de entre os valores de C testados, com o mesmo a 0.1.

Melhor dos classificadores multi-classe

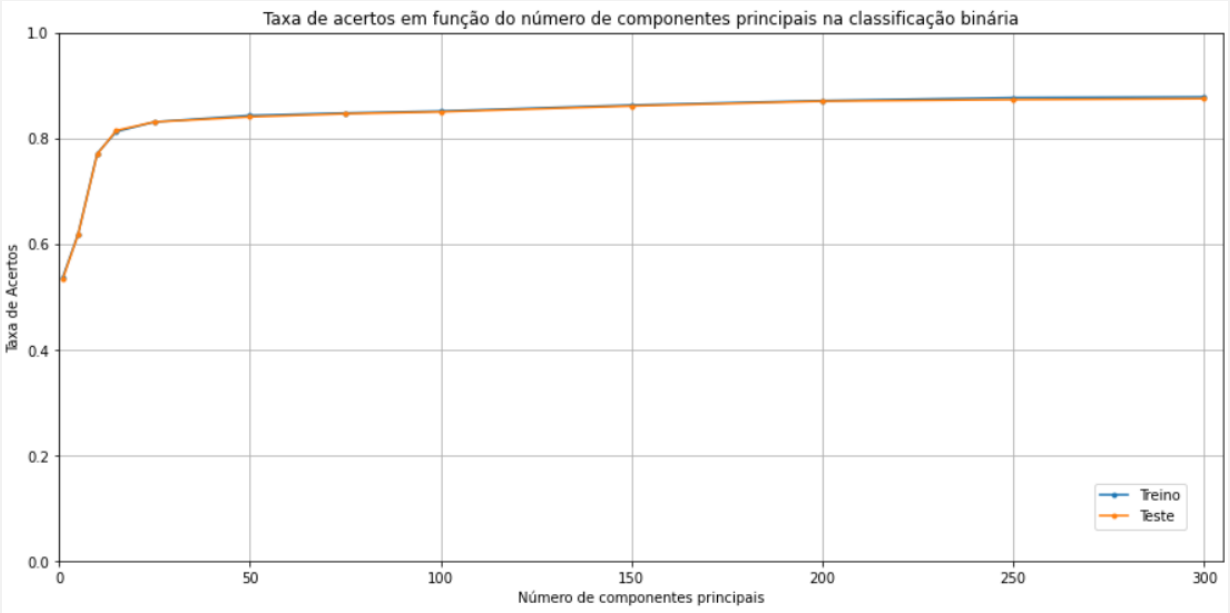


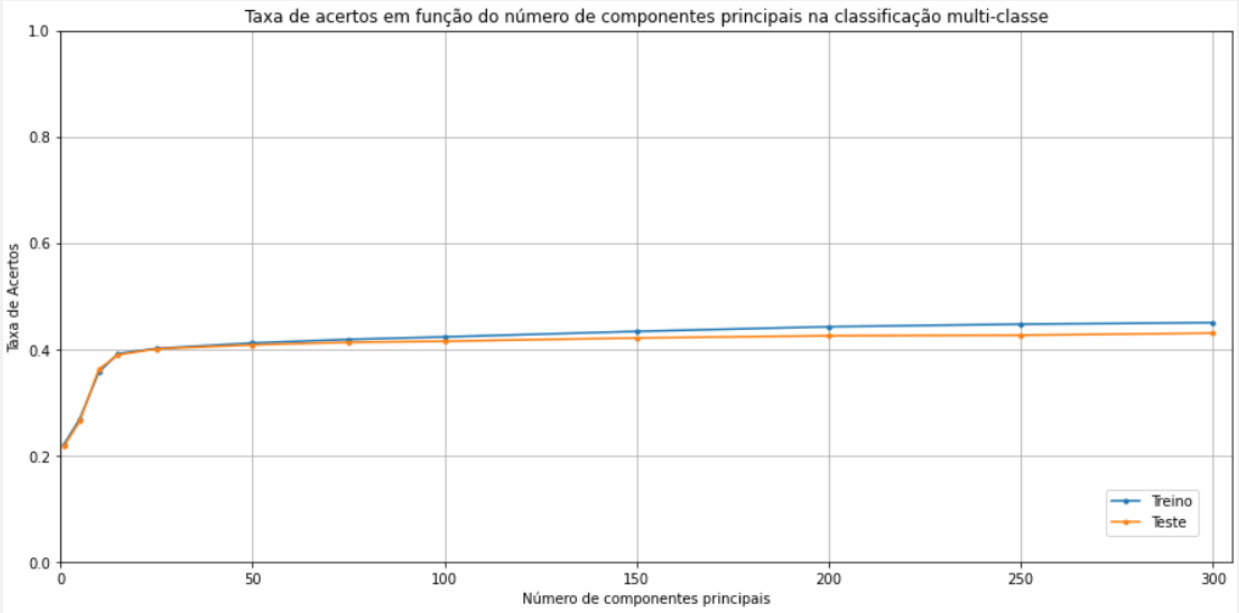
- O classificador LogisticRegression (discriminante logístico) é aquele que apresenta os melhores valores para treino e para teste. Logo, este é o melhor classificador multi-classe entre os testados.

Pré-processamento dos dados com PCA

- Para determinar se o pré-processamento dos dados é benéfico para o desempenho de descriminantes logísticos na tarefa de classificação, é necessário achar o melhor valor para as componentes principais.
- Testam-se diferentes valores de componentes principais e comparam-se os resultados. No instante antes dos resultados da classificação estagnarem, obtém-se o valor ótimo e verifica-se se houve melhorias ao utilizar PCA, tanto na classificação binária como na classificação multi-classe.

Análise da taxa de acertos ao utilizar PCA





Nº Componentes	1	5	10	15	25	50	75	100	150	200	250	300
Treino Binário	0,5375	0,6192	0,7715	0,8127	0,8312	0,8434	0,8480	0,8517	0,8630	0,8715	0,8771	0,8784
Teste Binário	0,5348	0,6182	0,7702	0,8149	0,8312	0,8408	0,8467	0,8503	0,8613	0,8704	0,8734	0,8752
Treino Multi-classe	0,2227	0,2692	0,3576	0,3924	0,4021	0,4124	0,4192	0,4239	0,4343	0,4431	0,4431	0,4507
Teste Multi-classe	0,2191	0,2653	0,3629	0,3903	0,4015	0,4089	0,4138	0,4158	0,4219	0,4260	0,4260	0,4310

Análise da taxa de acertos ao utilizar PCA

- Pelos gráficos e a tabela anteriores, verifica-se que até às 15 componentes principais, a taxa de acertos cresce depressa. A partir daí, tem um crescimento mais demorado e só a partir das 200 componentes é que se verifica que as taxas convergem para um valor. Assim, considera-se que 200 é o valor ótimo para o número de componentes principais no PCA.
- Para a classificação binária com PCA, obteve-se 0,8715 em treino e 0,8704 em teste e sem PCA, obteve-se 0,9528 em treino e 0,8994 em teste.
- Para a classificação multi-classe com PCA, obteve-se 0,4431 em treino e 0,4260 em teste e sem PCA, obteve-se 0,6985 em treino e 0,4445 em teste.
- Apesar de os resultados em teste diminuírem um pouco, conclui-se que fazer pré-processamento dos dados é benéfico pois ir-se-á obter resultados mais previsiveis, devido à diferença entre treino e teste ser menor.

Clustering: K-means

- Para agrupar as críticas de uma forma não supervisionada, usou-se o algoritmo K-médias.
- As críticas foram limpas sem restrições quanto ao `token_pattern` (por *default*, *tokens* de 2 ou mais letras) e ao `min_df` (por *default*, a 1), pela propensão a aparecerem mais *tokens* apreciativos. A restrição foi a retirada das `stop_words` da língua inglesa, de forma a não obtermos conectores de frases comuns.

Modificação do número de clusters

- Ao alterar o parâmetro de número de clusters a formar, no algoritmo K-médias, observámos que obtínhamos *clusters* com diferentes tipos de palavras mais recorrentes.
- À medida que se aumenta o número de clusters, acima de 8 (número de classificações possíveis) as palavras são mais específicas ao género cinematográfico a que as críticas agrupadas se referem, e menos à sua apreciação.
- Por exemplo: 10 *tokens* mais recorrentes com:

  - 2 clusters

  - Cluster 0: movie, bad, just, like, movies, good, really, don, watch, time
  - Cluster 1: film, movie, like, just, good, story, time, great, really, films

  - 10 clusters

  - Cluster 0: funny, movie, comedy, just, jokes, like, really, film, good, laugh
  - Cluster 2: horror, film, movie, gore, good, like, just, films, really, zombie

Clustering: Agrupamento de críticas por tópico

- Com um valor mais alto, como 25 *clusters*, já se obtêm *clusters* de temas mais distinguíveis e agrupáveis por tópico.

Por exemplo,

- Cluster 7: christmas, santa, movie, scrooge, claus, film, grinch, family, story, holiday, ... – filme de Natal
- Cluster 10: game, games, graphics, play, movie, video, mario, ..., levels – filme de videojogo
- Cluster 13: effects, special, movie, sci, fi, ... – filme de ficção científica
- Cluster 19: disney, animation, cinderella, movie, film, story, kids, animated, original ... – filme de animação infantil

Procura de críticas por tópico

- É possível, através dos tópicos de cada *cluster*, observar as críticas relacionadas com o mesmo.
- Por exemplo, pode-se ver comentários acerca de filmes que se inserem nos tópicos do *cluster* 19:

No day passes without a new released computer **animated** movie, so we now really have chances to see more than some nice effects. After watching Ice Age I felt that's it was not that big impact on me than some other films of this genre. But it's because I am a Big Guy now, and I am pretty sure that this is a very enjoyable movie for children (maybe up to 14). The **story** is quite simple, and the "actors" are funny in a cute way, without any crude or complex humour. Even the "evil" is lovely, fluffy big cat with those funny teeth. And the **story** has a happy end, which was a small disappointment for me (knowing that most of the main characters are doomed to extinction in a sad way) but a great thing for children. And apart from some fights nobody dies (not even when he gets stomped on by a mammoth, several times), which made a cartoony feeling. The computer animation part is nice but nothing special, apart from some really nice cartoony feeling scenes, when you feel like walking in a nice painting or pages of a comics. [Which means lots of work nevertheless!] There were some gags which made me smile - I accept, the creators tried to satisfy those grownups - but they are hard to spot and (in my opinion) better left unnoticed, since it does not feel to fit into the **story**. Overall it's a nice movie, but it's rather in the ideal-world-and-fluffy-animals-for-children **disney** cliche. If you don't hate cute animals making funny things, watch it at least once.

I wasn't sure when I heard about this coming out. I was thinking how dumb is **Disney** getting. I was wrong. I found it to be very good. I mean it's not The Lion King but it's cool to see another side from a certain point. It was very funny. Also it wasn't one of those corny **disney** sequels were the **animation** sucks, it was just like The Lion King **animation**. The only thing that eritated me was the whole movie theater thing through out the movie. Not to give anything way but you'll know what I am talking about. I also fun that it was cool to have most of the cast from the **original** to return. It was a very good movie over all.

If you have read the books then forget the characters that Tolkien built in your head. The representation of hobbits, dwarves etc have had the '**disney**' treatment. The dark riders are excellent, and as I had always imagined from the books. Cinematically this is an excellent film, mixing live motion and **animation** to produce amazing effects for the year. I only wish he (Bakshi) had been given the money to complete his epic. It's worth having the video as they will be worth a bit after the 2001 Lord of the Rings !!

Predição de clusters de críticas

- É possível predizer em que *cluster*

novas críticas se inserem.

- Testámos este aspeto com algumas críticas curtas falsas:

```
'I love every single Jackie Chan movie.' => 22
'Vietnam war movie with so many deaths, the soldiers looked so realistic.' => 2
'Such a realistic sci-fi.' => 15
'Comedy movies like this are always terrible, hate them.' => 7
```

```
Cluster 2:
 war
 film
 movie
 world
 soldiers
 story
 time
 vietnam
 german
```

```
Cluster 7:
 comedy
 movie
 film
 funny
 good
 romantic
 like
 just
 laugh
```

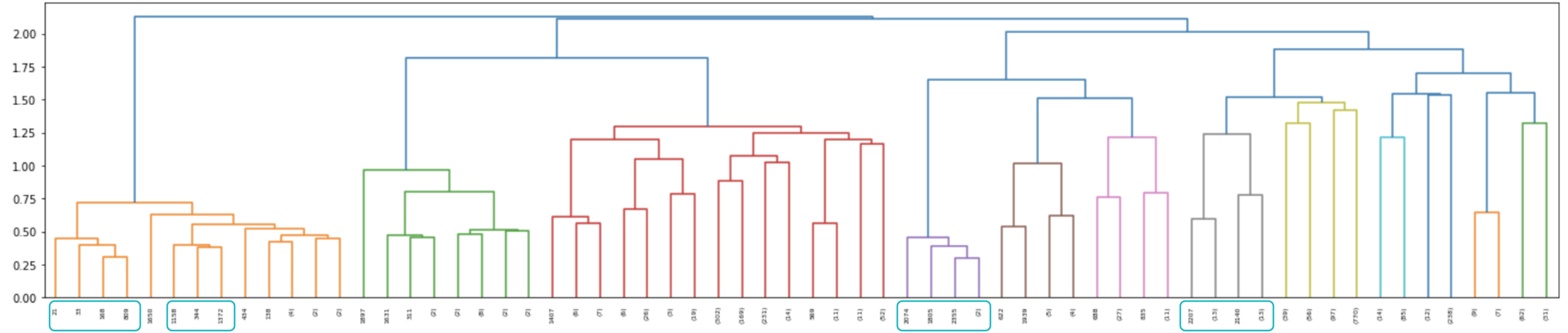
```
Cluster 15:
 sci
 fi
 movie
 film
 like
 effects
 space
 good
```

```
Cluster 22:
 fu
 kung
 martial
 chan
 jackie
 arts
 movie
 fight
 action
```

Clustering hierárquico

- Pré-processamento dos dados com `TruncatedSVD` para reduzir dimensionalidade e obter matriz não esparsa.
- Uso de apenas 5% das críticas
- Função `AgglomerativeClustering` permitiu criar agupamentos através de dendrogramas, que agrupa os *clusters* com menor distância euclidiana entre eles, e com algoritmo de ligação *Ward*, que os une, minimizando a variância *intra-cluster*.

Tópicos abordados por ramo do dendrograma



Jogos

Videojogos

Batman

Infantil

K, o nível de corte do dendrograma, poderia ser um valor igual ou superior a 4, para que os tópicos semelhantes ficassem agrupados.