Aprendizagem Automática

LICENCIATURA EM ENGENHARIA INFORMÁTICA E MULTIMÉDIA

Trabalho realizado por:
- António Luís Ferreira, 47500
- Tomás Gomes, 48614

Classificação de críticas IMDB

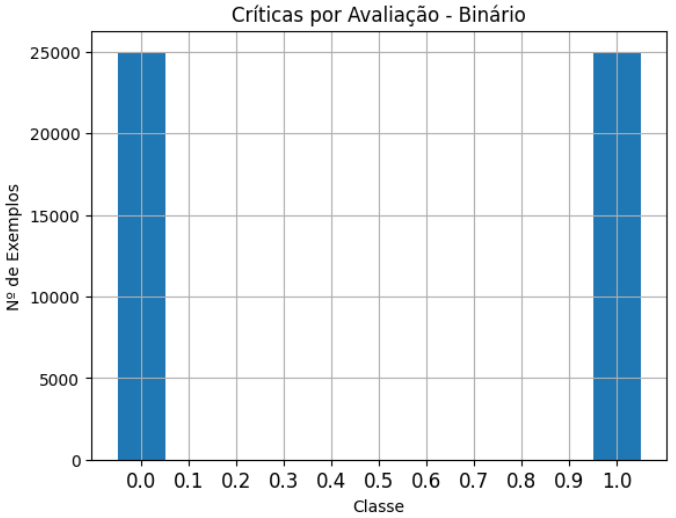
Este trabalho teve como objetivo classificar críticas de texto da base de dados IMDB, apresentando-se como um problema de classificação, neste caso, supervisionada.

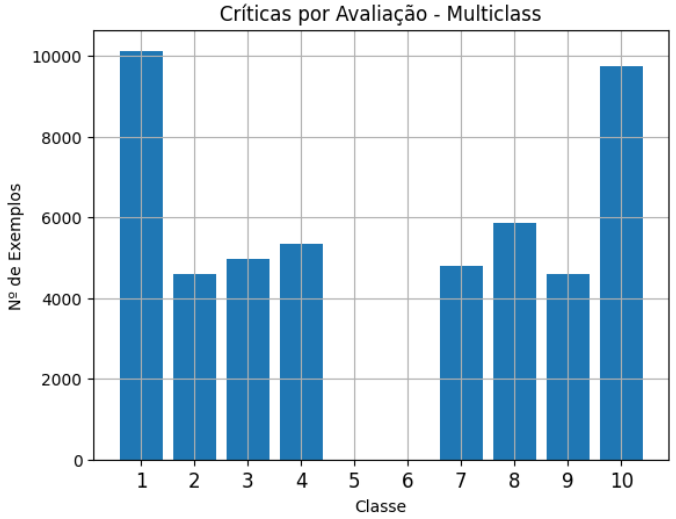
Foram produzidos dois tipos de classificadores de críticas:

- Classificadores Binários, em que a crítica é positiva ou negativa;
- Classificadores Multiclasse, em que a crítica é classificada por um número compreendido entre os intervalos 1-4 (Negativa) e 7-10 (Positiva)

Dados Fornecidos

Divisão das críticas num problema binários, 0 sendo uma crítica negativa e 1 uma crítica positiva.



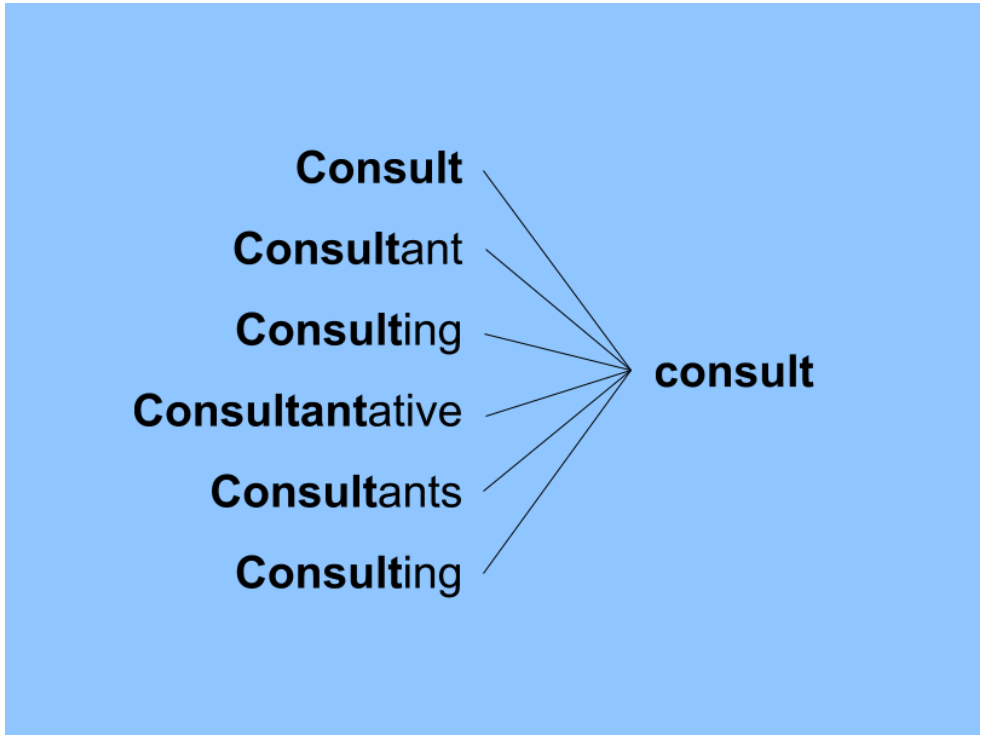


Divisão das críticas nos intervalos anteriormente referidos, 1 a 4 para críticas negativas e 7 a 10 críticas positivas.

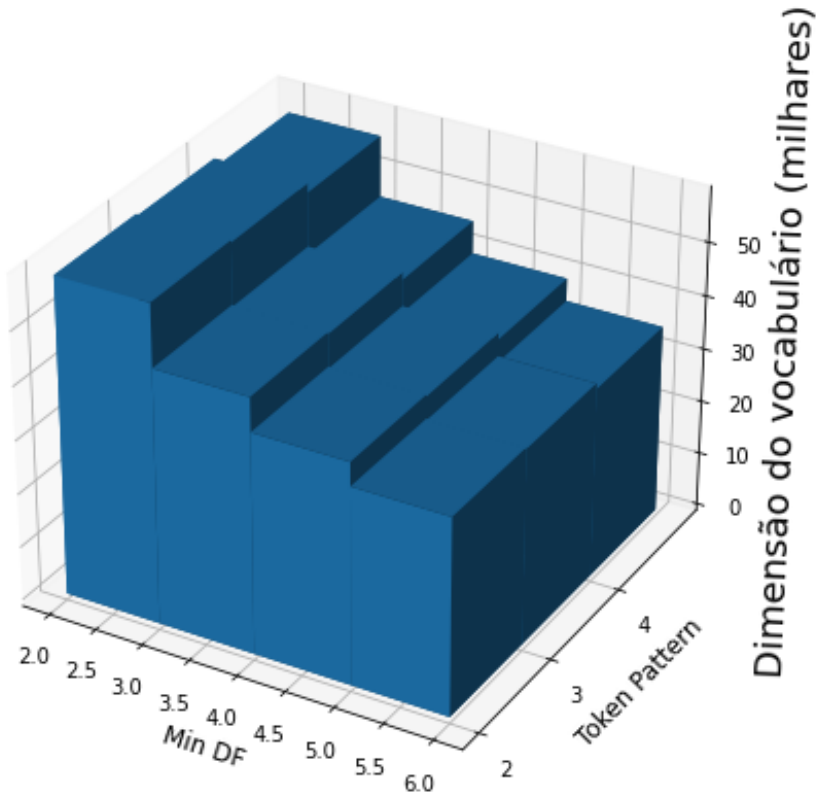
Stemming

O *Stemming* consiste num processo de transformar uma qualquer palavra no seu radical, ou seja, palavras como por exemplo *studies*, *studying* e *studied* iriam ser transformadas no radical *studi*

Este processo é utilizado com o objetivo de reduzir o vocabulário e assim o tamanho dos dados.



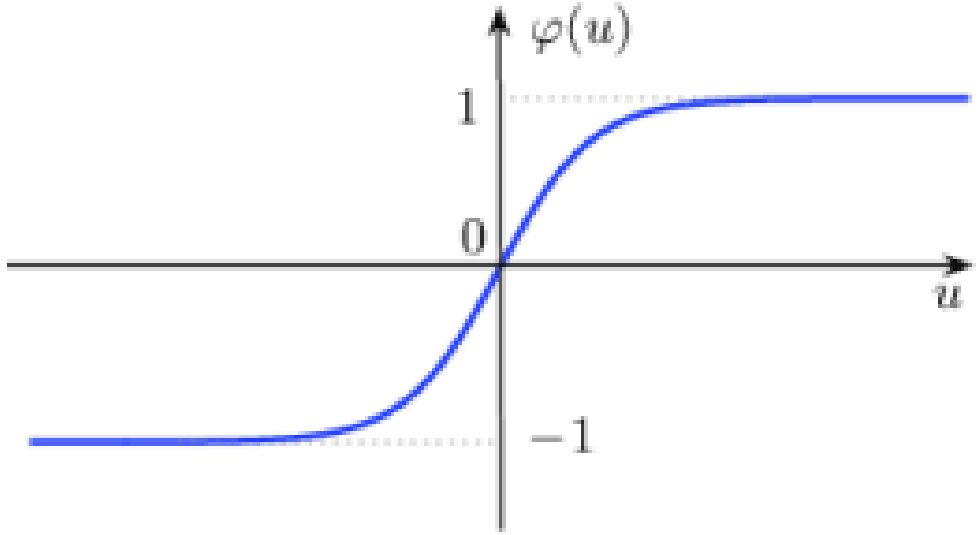
TF-IDF – Term Frequency-inverse document frequency



O TF-IDF vectoriza os dados de texto, sendo assim possível usá-los para o treino de modelos de classificação. Este recebe variados parâmetros, tais como o min_df, token_patern e n_gramas.

Foram variados estes parâmetros de maneira a influênciar o tamanho dos dicionários de palavras resultantes desta vetorização.

Logistic Regression

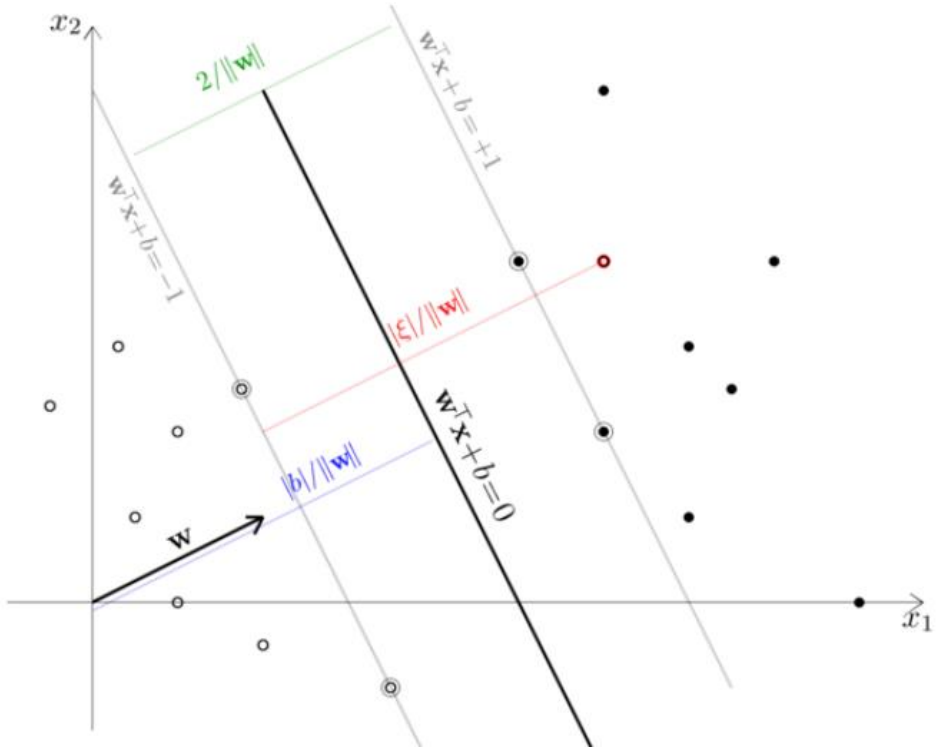


Apesar do nome, *logistic regression* ou em português discrimante logístico, não é um método de regressão, mas sim, um método de classificação.

O objetivo da regressão logística é encontrar o melhor modelo que possa prever a probabilidade de um resultado dado um conjunto de entradas, geralmente binário.

$$u_k = \mathbf{w}_k^\top \mathbf{x} = w_{0k} + w_{1k}x_1 + \ldots + w_{dk}x_d, \quad k = 1, \ldots, c$$

Linear SVC



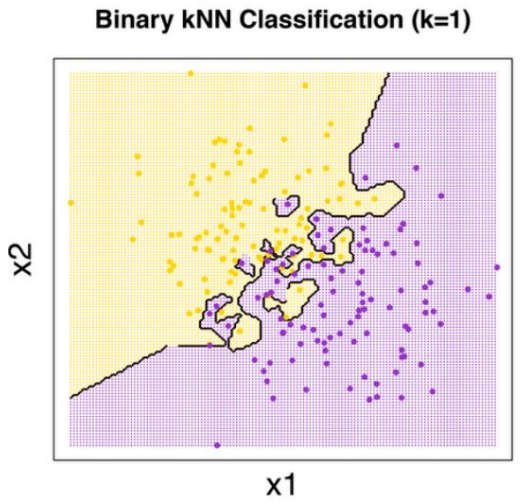
Este modelo de classificação pertence à família de classificadores supervisionados, SVM(Máquinas de Suporte Vetorial), sendo aplicável a problemas de classificação e regressão.

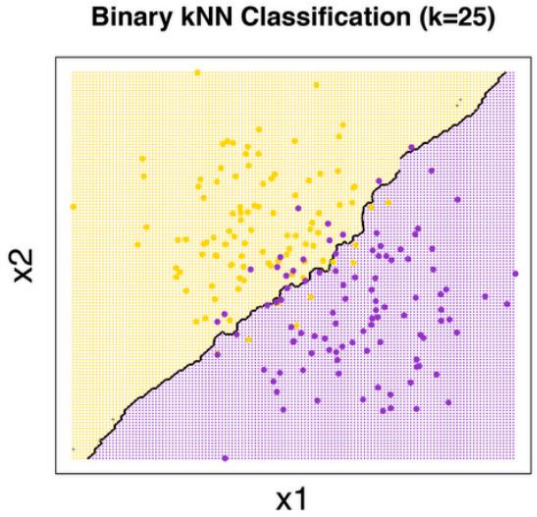
As principais vantagens na utilização deste modelo é que, lida bem com problemas complexos e com dados de alta dimensão, sendo perfeito para trabalhar com uma base de dados grande como a deste projeto.

K-Neighbours

Este classificador é um classificador baseado em distâncias, não-paramétrico em que a classificação é baseada nos exemplos de treino. A classe estimada vai ser aquela que possuí mais *neighbours* próximos.

O valor ótimo para *k* depende do problema. Valores pequenos tendem a originar fronteiras de decisão irregulares ao contrário de valores elevados que fazem o oposto, criando fronteiras mais regulares.

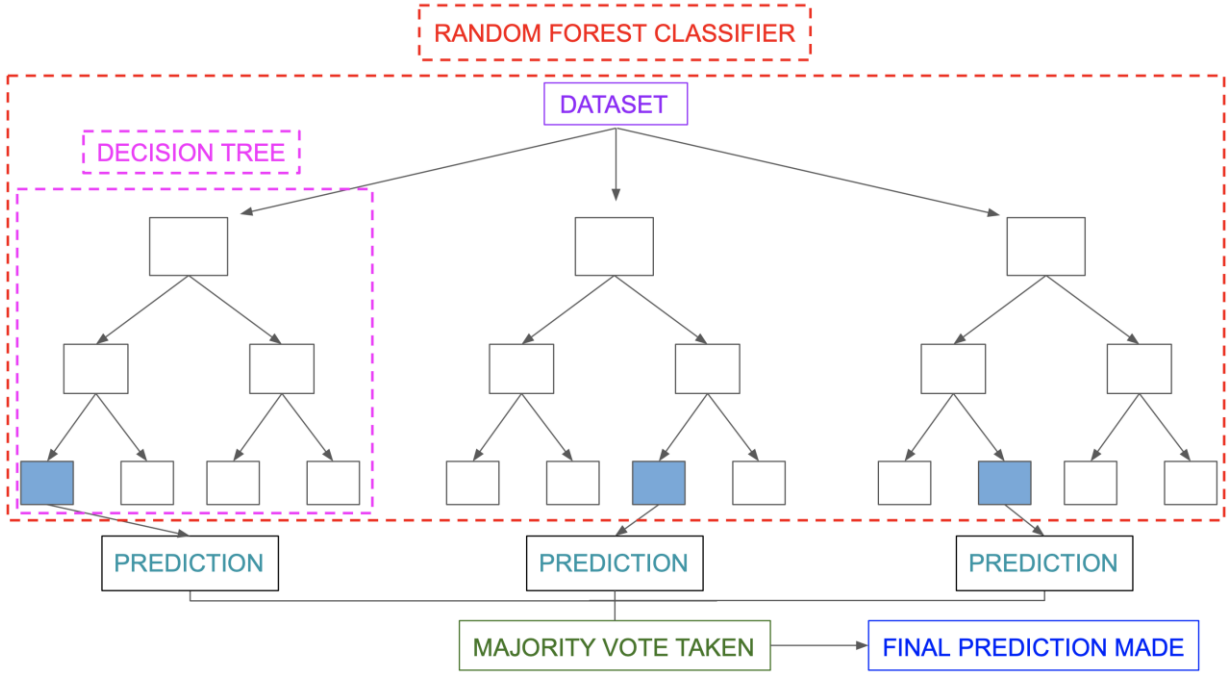




Random Forest

O algoritmo *random forest* consiste na construção de várias árvores de decisão, cujos *nodes* representam uma "questão" baseada nas características dos dados de treino.

É escolhida aleatoriamente uma amostra de dados para treinar cada árvore pelo que irão ser combinadas e irá ser deliberado por voto da maioria, a previsão final feita.



Hiper parâmetros

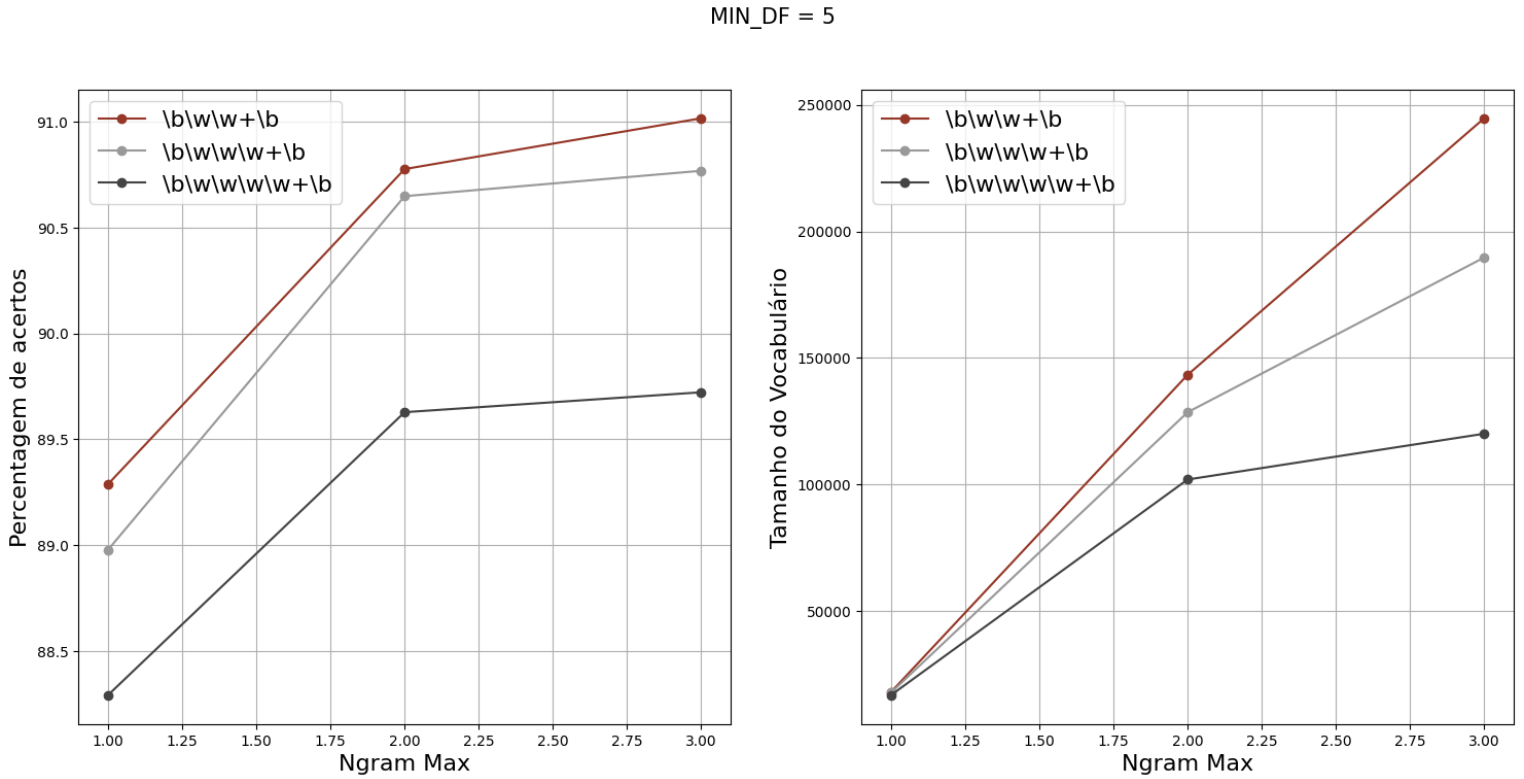
Para cada classificador, é importante denotar que este receberá vários hiper parâmetros, que neste projeto, foram ajustados de maneira a obter melhores resultados de classificação.

Para cada classificador existem hiper parâmetros diferentes.

Para a vetorização, também foram escolhidos os parâmetros com o maior número de acertos (No Linear Regression)

Logistic Regression	Linear SVC	K-neighbors	Random Forest	TF-IDF
C	C	K	N_estimators	Min_df
Penalty	Penalty	Weights	-	Token_pattern
-	-	-	-	N_grams

TF-IDF - Binário

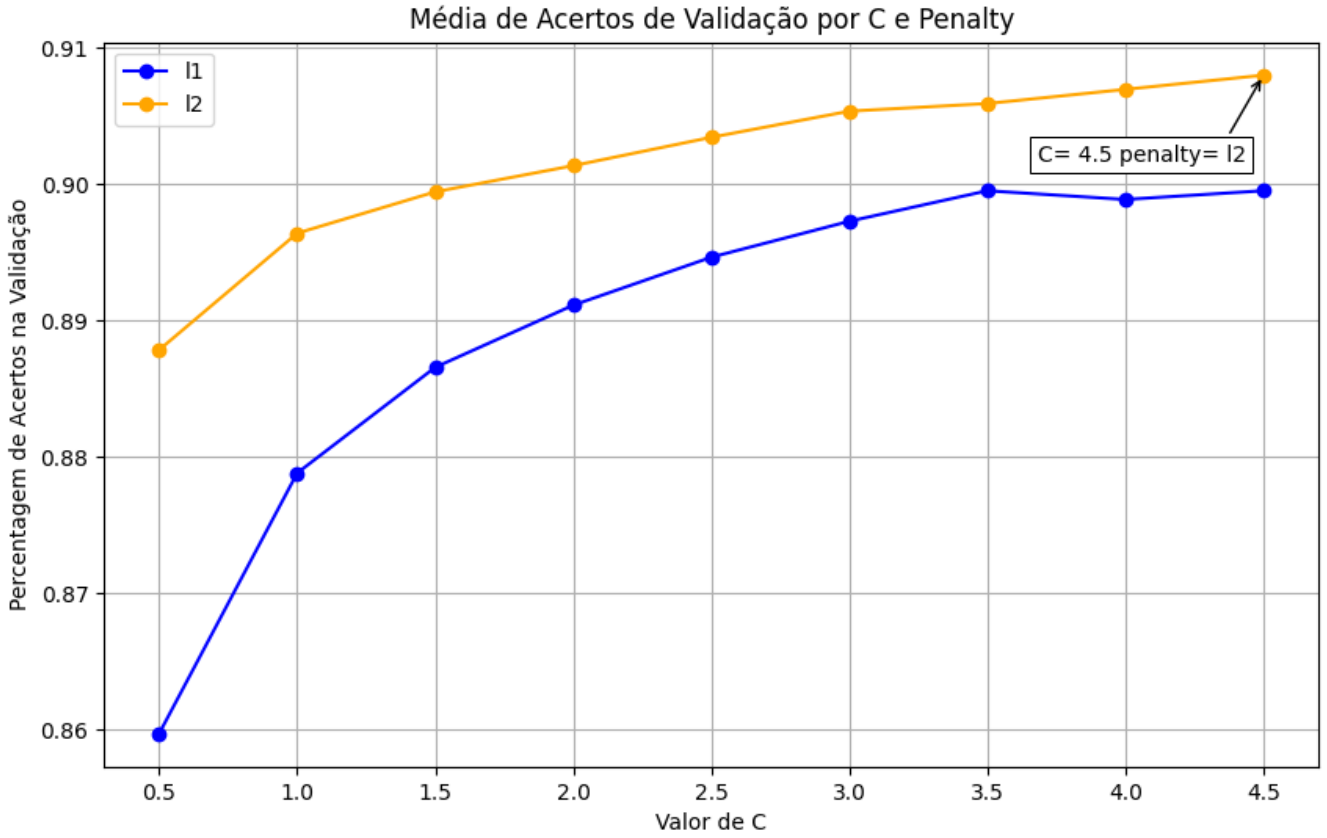


Para a vectorização do texto, foram testados os hiperparametros que melhor fossem classificados no modelo *logistic regression.*

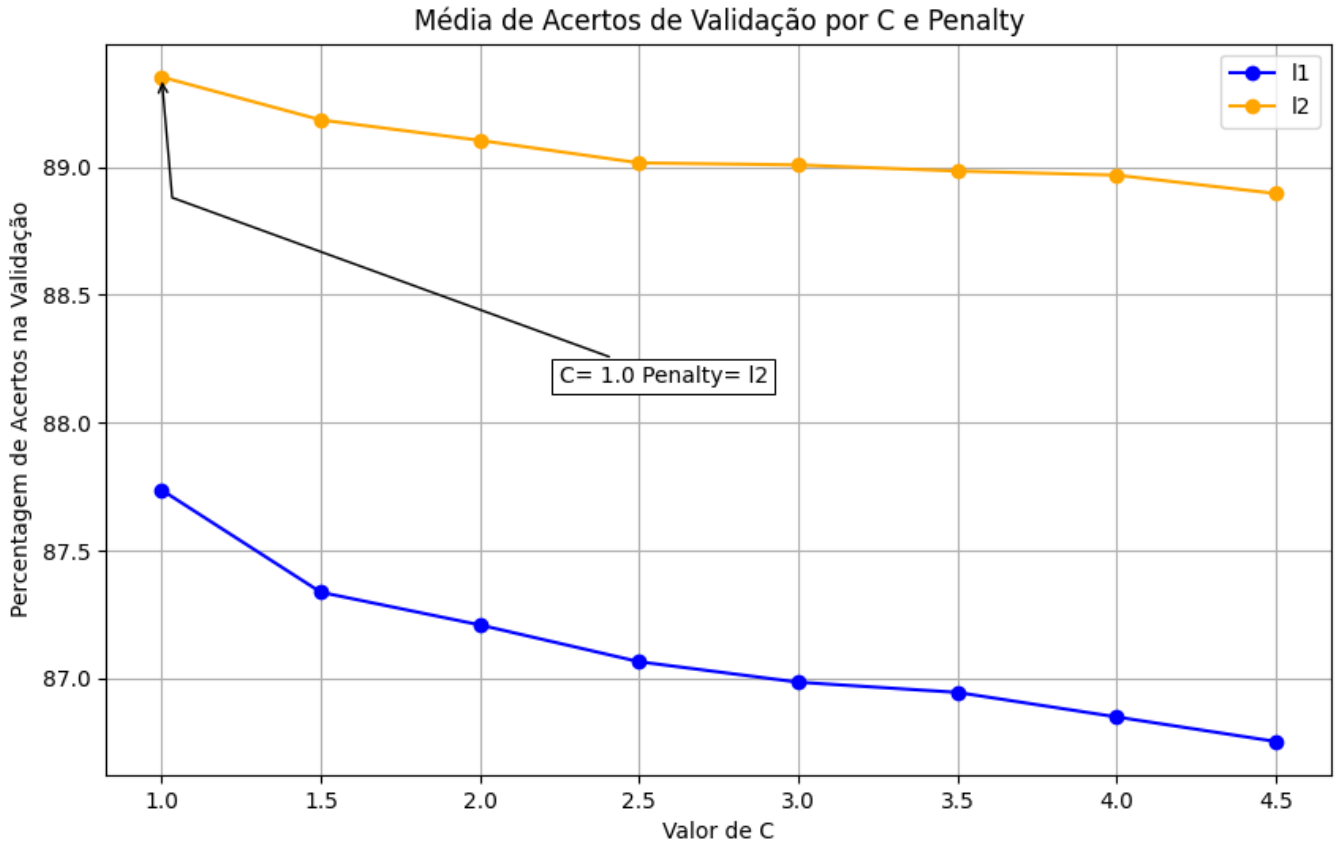
Não foram testados para cada modelo, apenas para o *logistic regression*.

É preciso ter em atenção que o número de acertos não foi o único critério usado para a eleição da melhor combinação pois foi também tido em conta a dimensão do vocabulário.

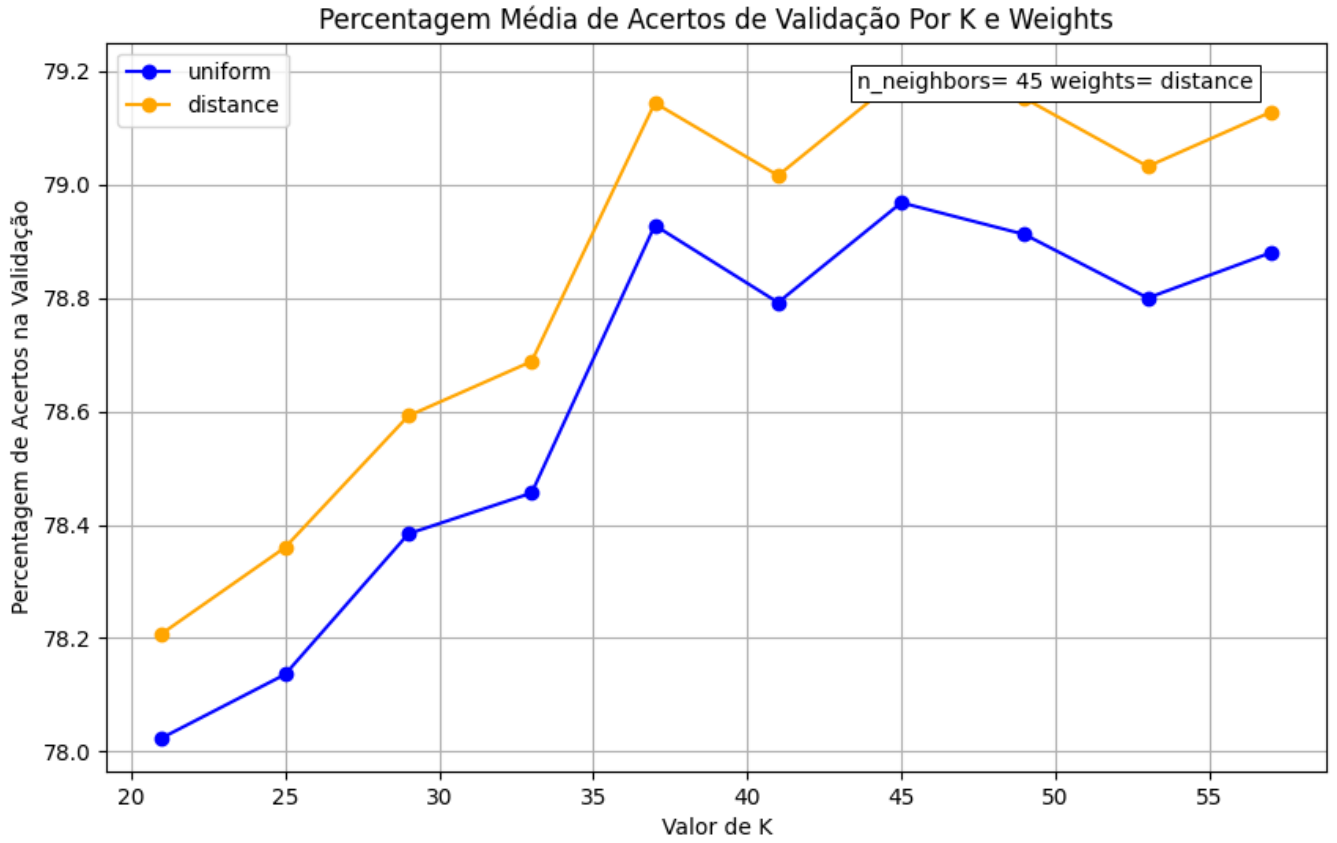
Logistic Regression - Binário



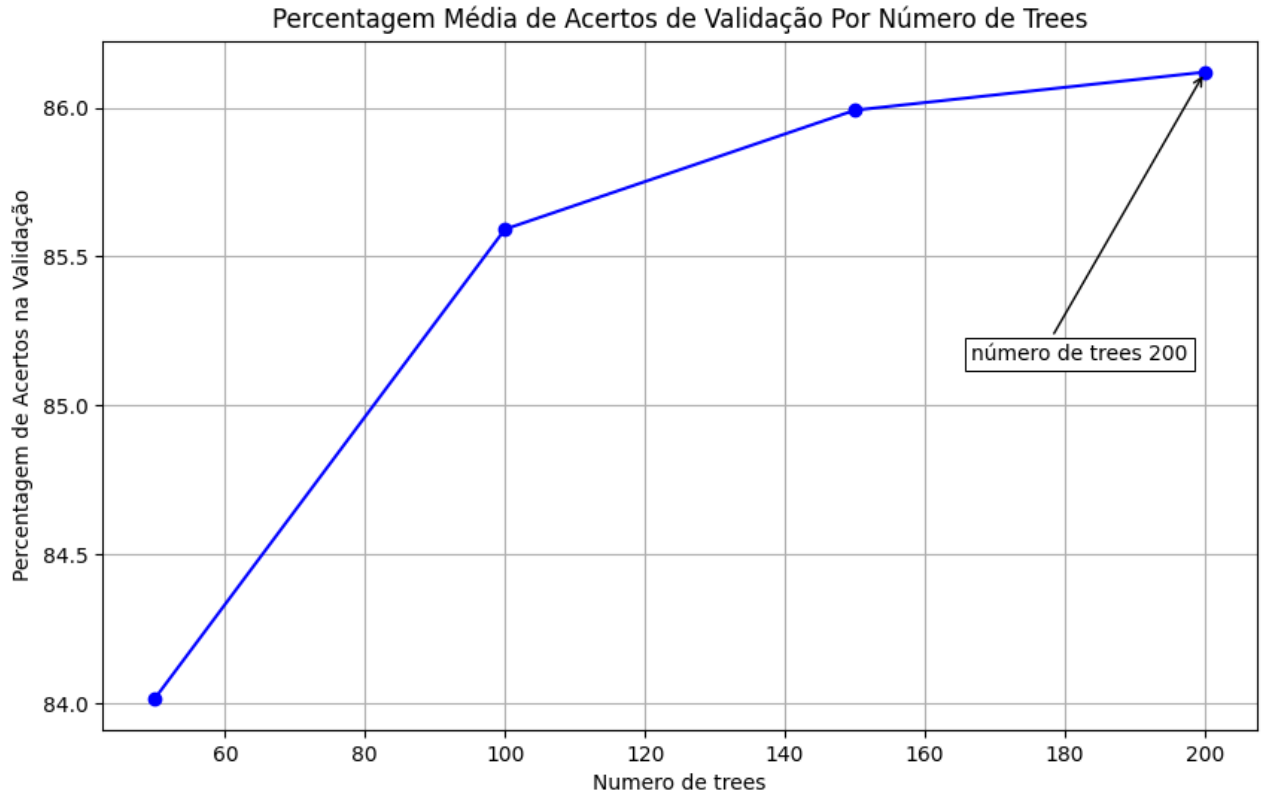
Linear SVC - Binário



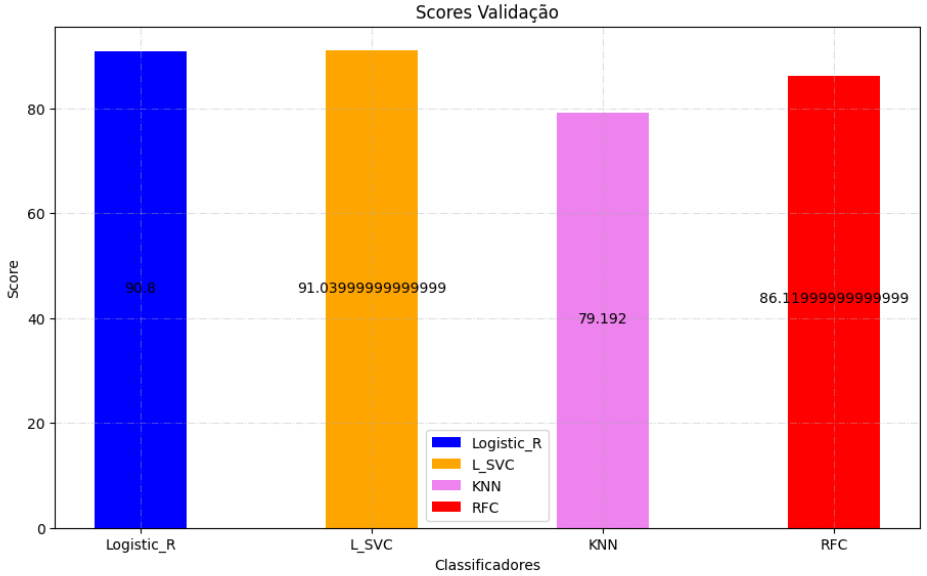
K-Neighbors - Binário

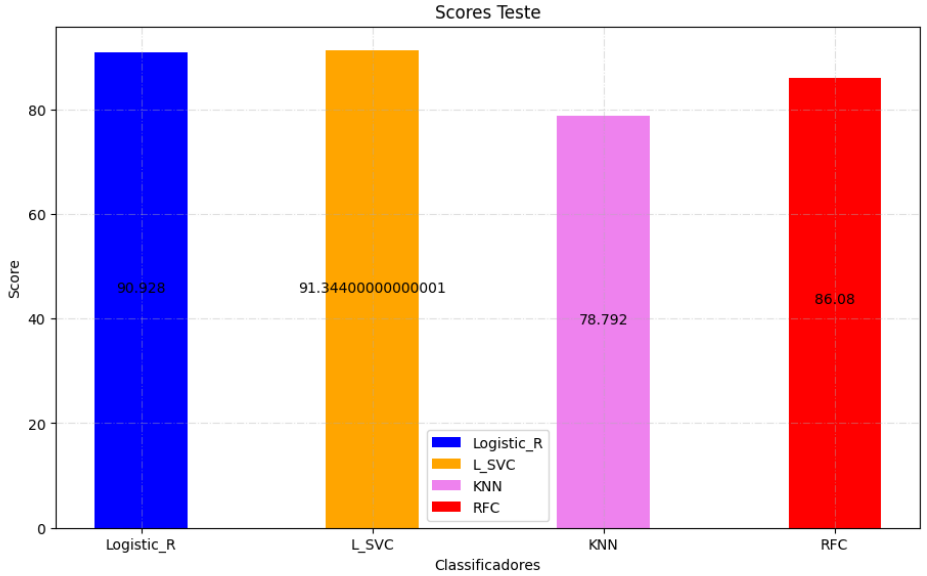


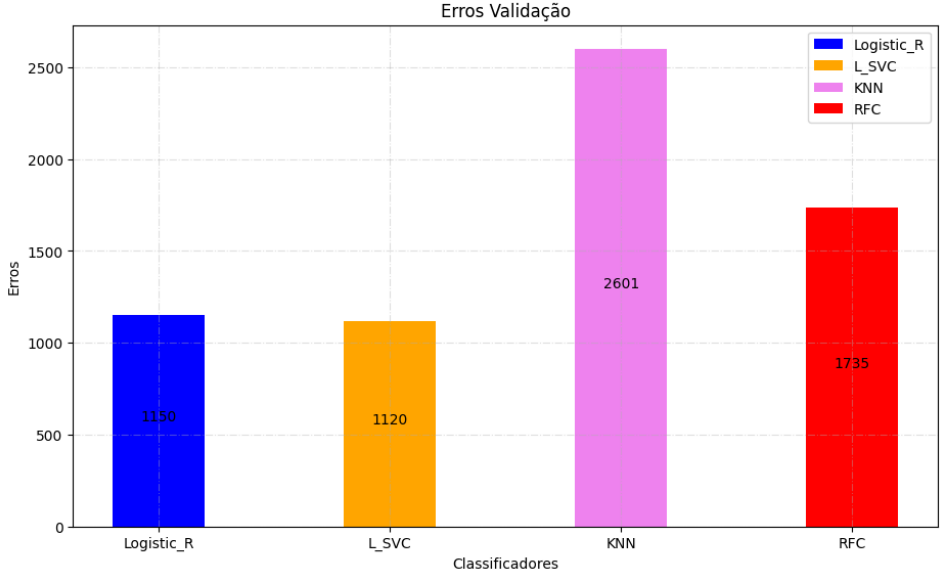
Random Forest - Binário

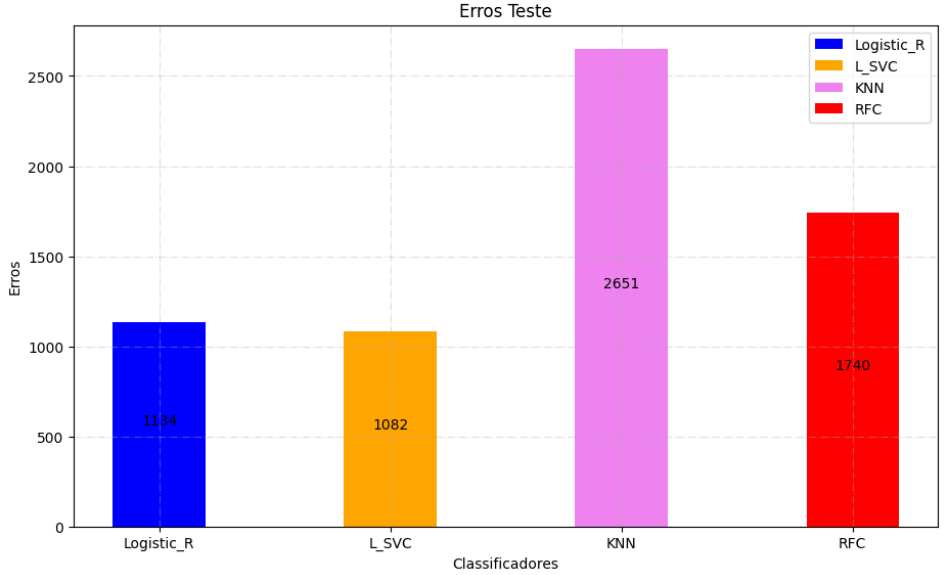


Comparação de Resultados


Scores Validação
Score
Classificadores
Logistic_R
L_SVC
KNN
RFC
90.8
91.03999999999999
79.192
86.11999999999999

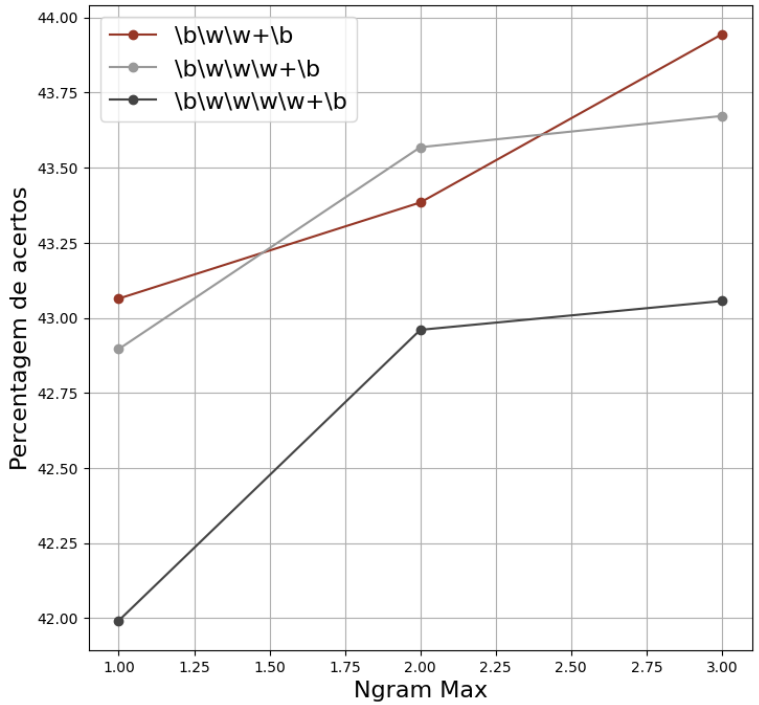

Scores Teste
Score
Classificadores
Logistic_R
L_SVC
KNN
RFC
90.928
91.34400000000001
78.792
86.08

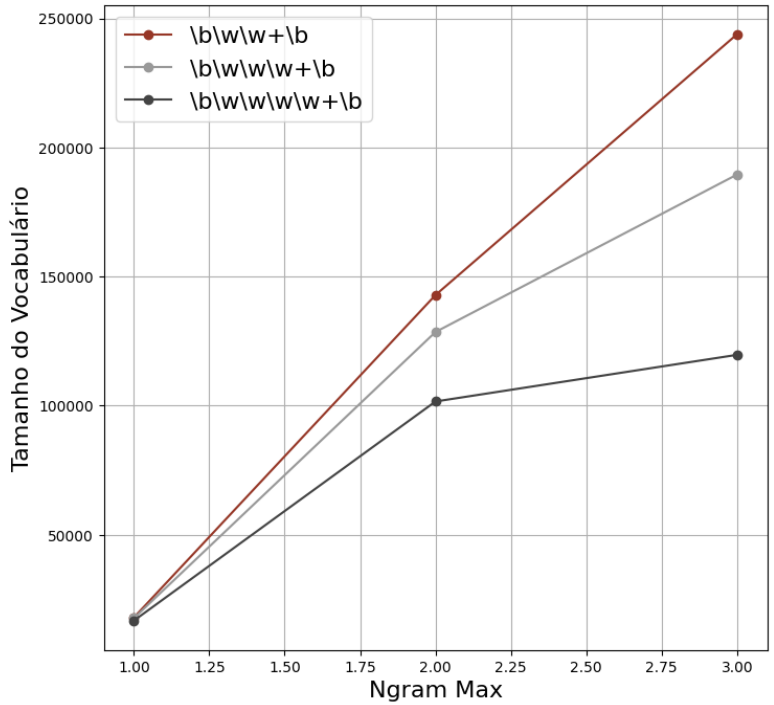

Erros Validação
Erros
Classificadores
Logistic_R
L_SVC
KNN
RFC
1150
1120
2601
1735


Erros Teste
Erros
Classificadores
Logistic_R
L_SVC
KNN
RFC
1134
1082
2651
1740

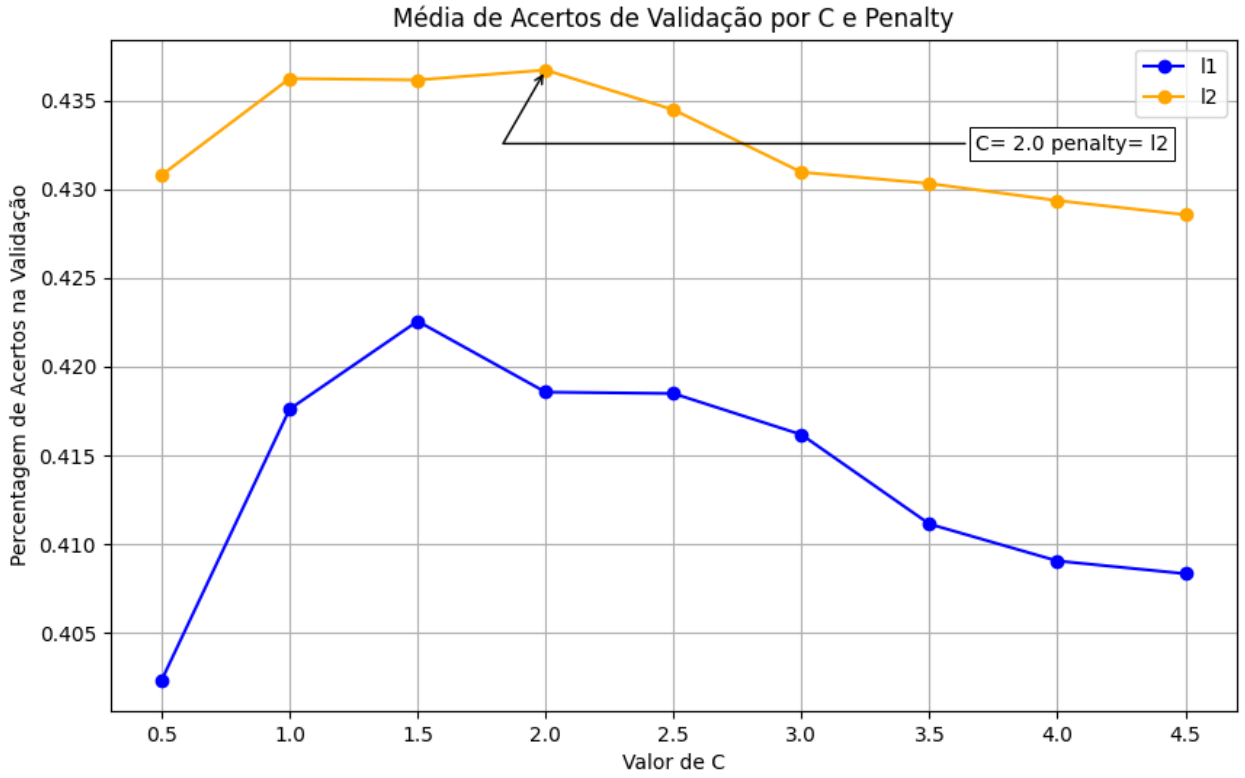
TF-IDF – Multiclasse

MIN_DF = 5

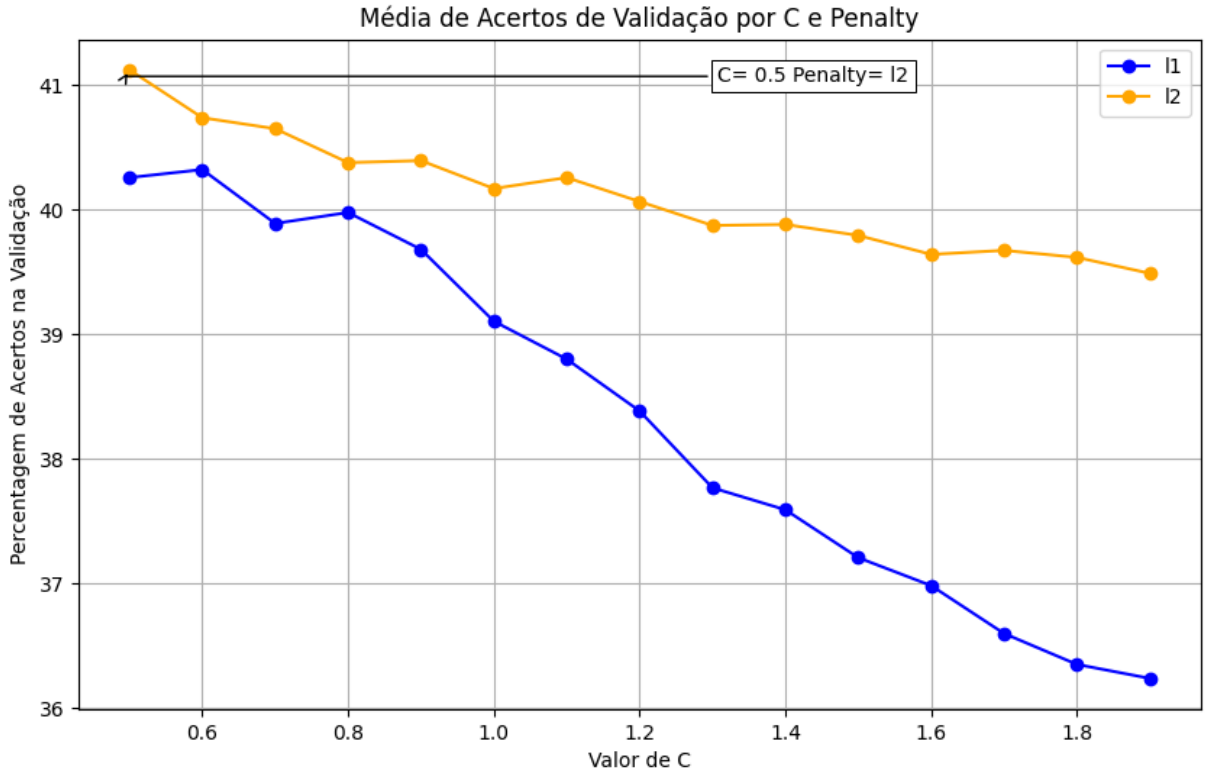




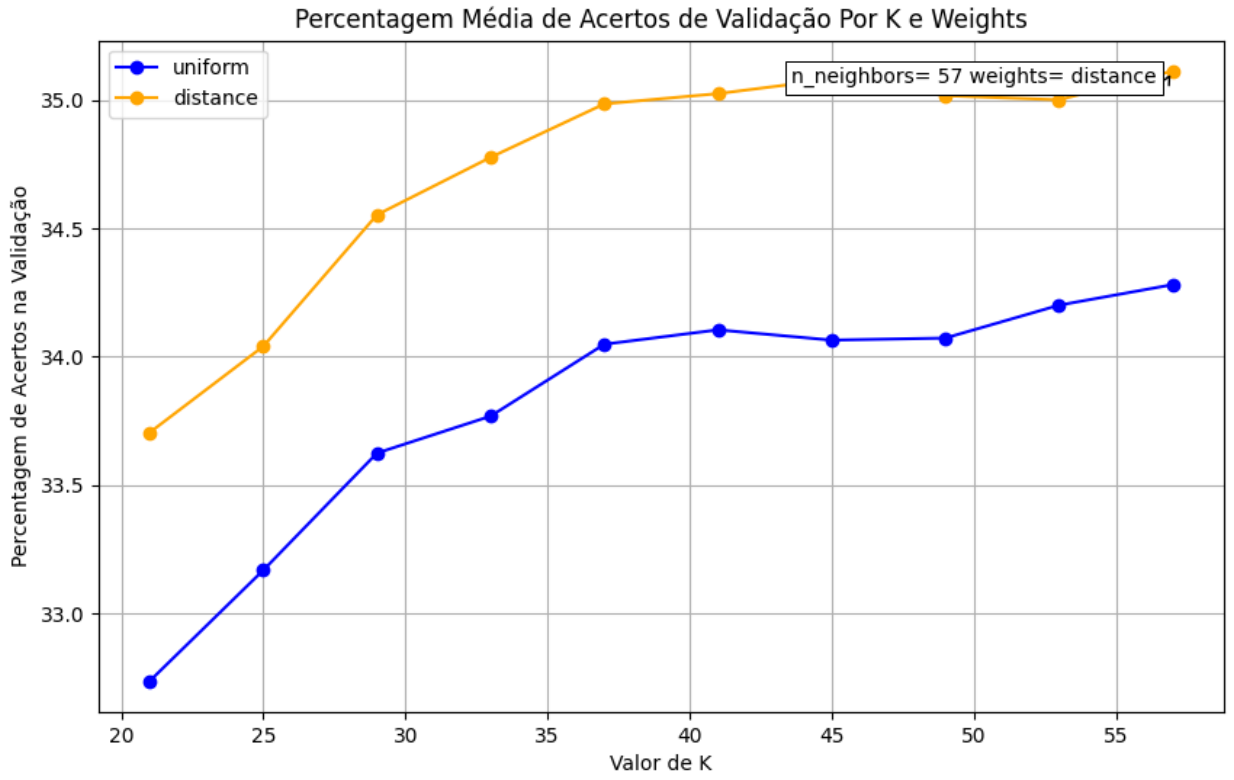
Logistic Regression – Multiclasse



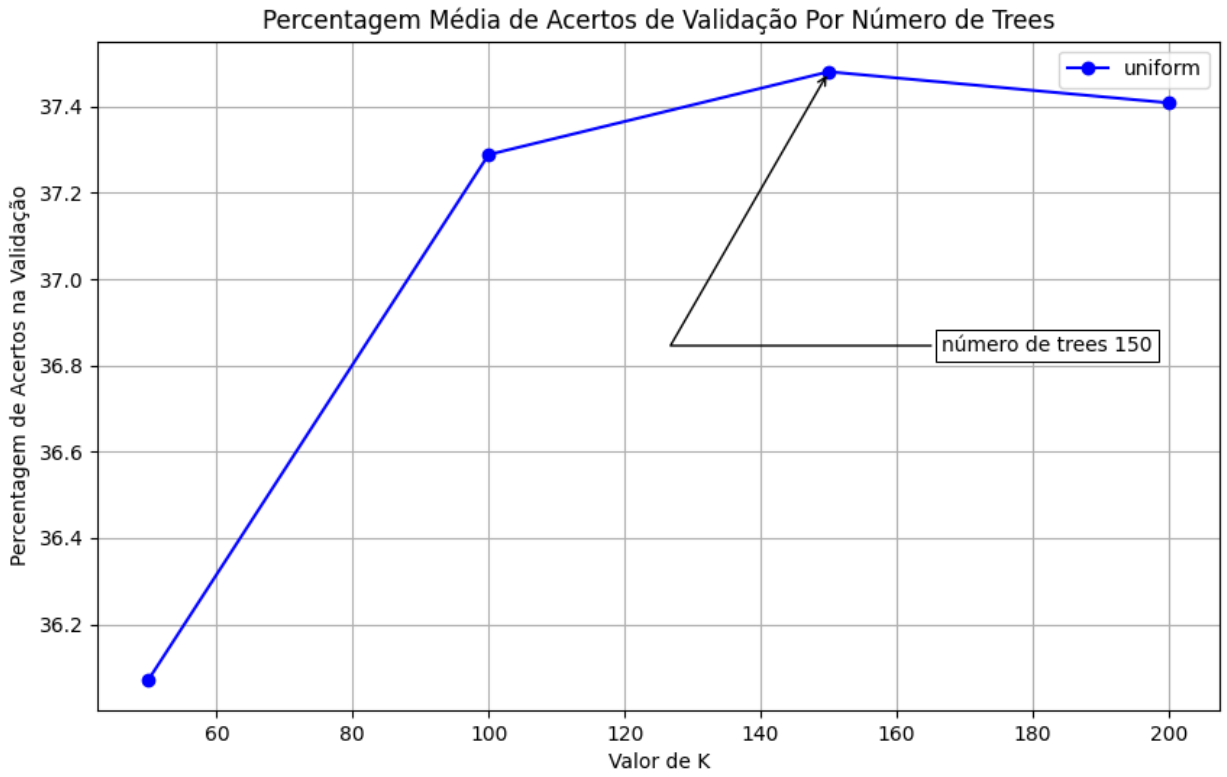
Linear SVC – Multiclasse



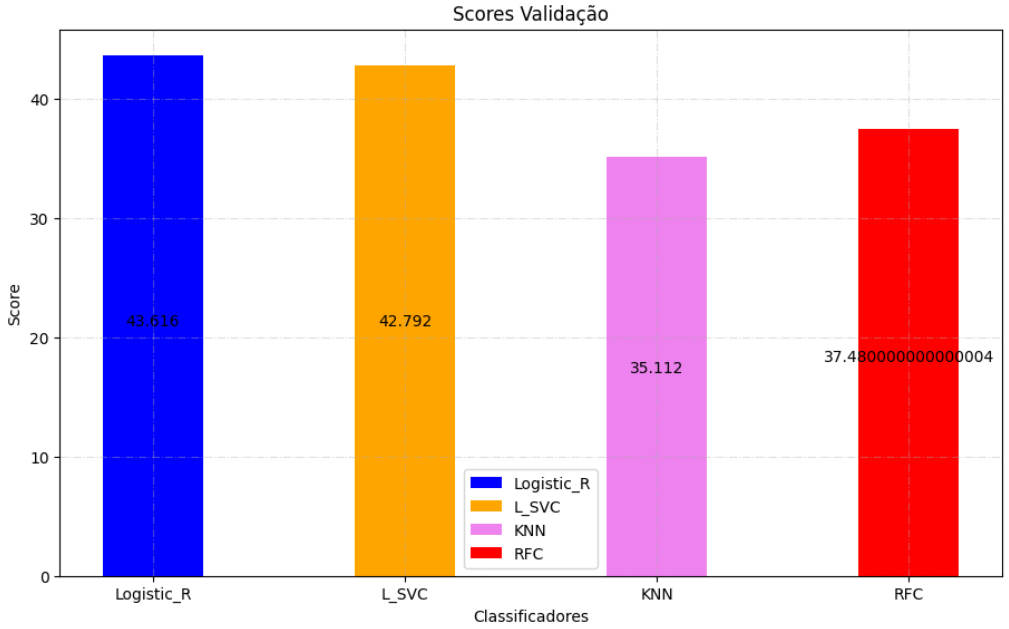
K-Neighbors – Multiclasse

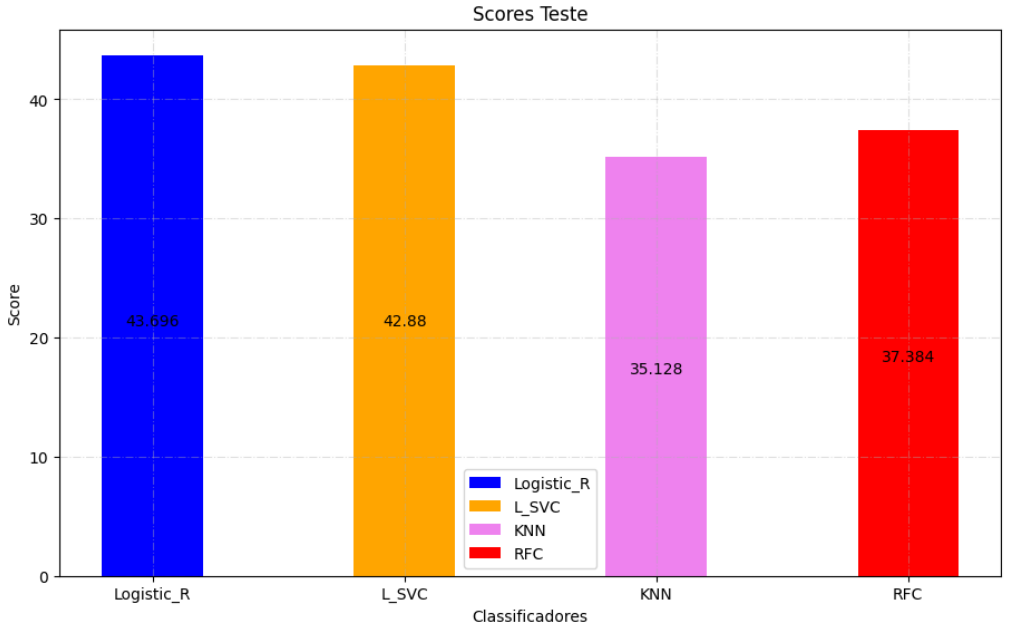


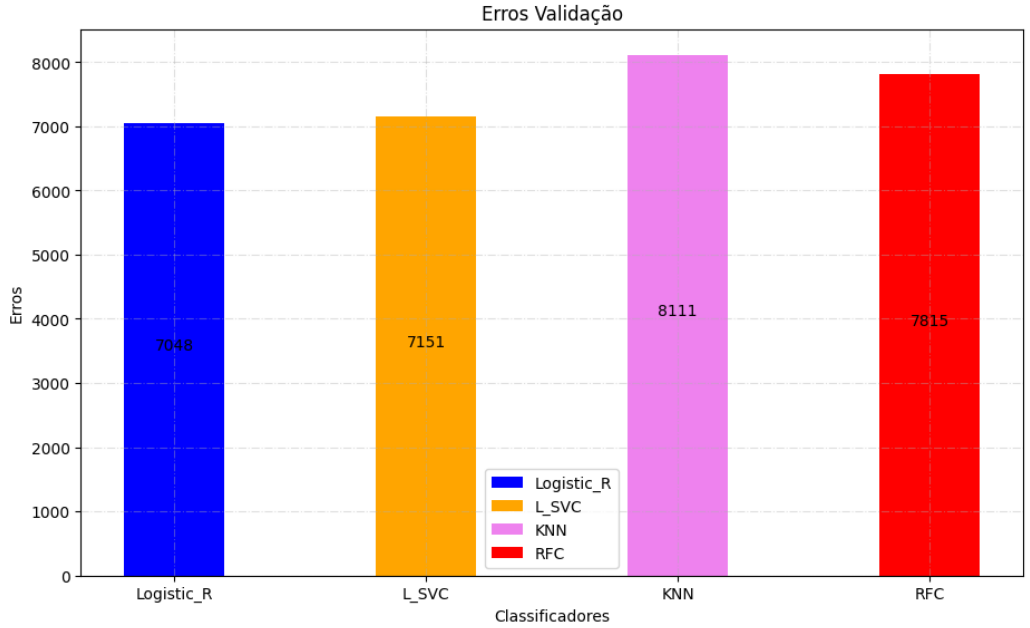
Random Forest – Multiclasse

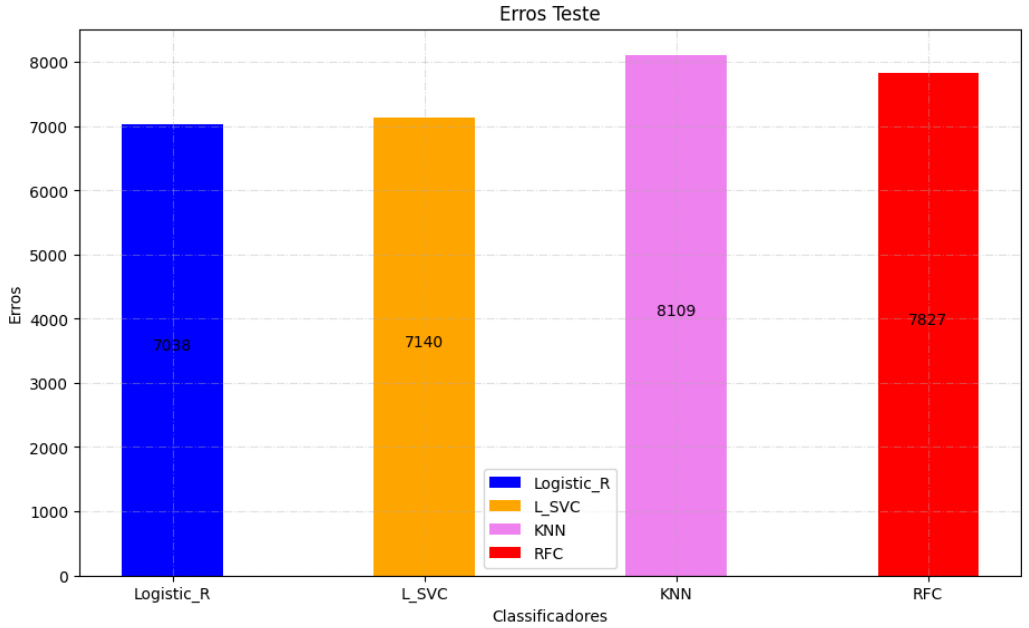


Comparação de Resultados

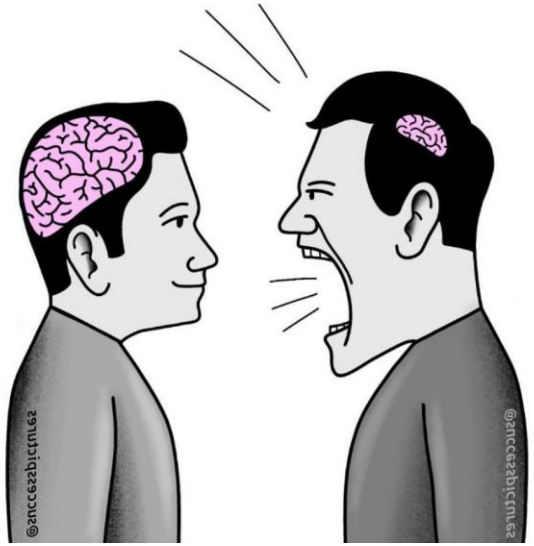

Scores Validação
Score
Classificadores
Logistic_R
L_SVC
KNN
RFC
43.616
42.792
35.112
37.480000000000004


Scores Teste
Score
Classificadores
Logistic_R
L_SVC
KNN
RFC
43.696
42.88
35.128
37.384


Erros Validação
Erros
Classificadores
Logistic_R
L_SVC
KNN
RFC
7048
7151
8111
7815


Erros Teste
Erros
Classificadores
Logistic_R
L_SVC
KNN
RFC
7038
7140
8109
7827

Considerações Multiclasse

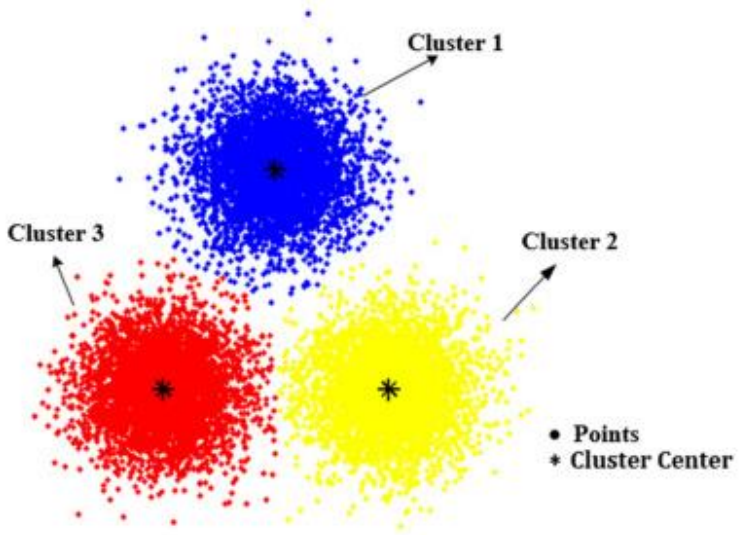


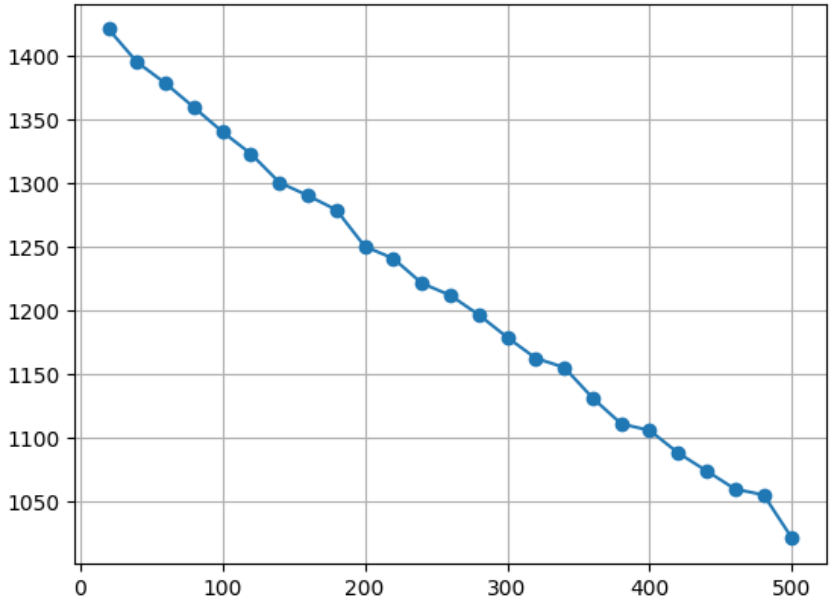
Como podemos ver analisando os diapositivos anteriores, os classificadores multiclasse deram taxas de erro muito elevadas. Este acontecimento deve-se ao facto de ser impossível uma máquina analisar os tipos de expressão linguística como por exemplo a ironia e o sarcasmo, entre outros.

Cada ser humano tem os seus critérios na avaliação de um filme, por exemplo, uma avaliação de 8 para mim pode equivaler a uma avaliação de 4 de outra pessoa, mesmo que tenhamos escrito críticas idênticas.

Clustering – K-médias

O *clustering* é um método de agrupamento. É uma técnica de aprendizagem não supervisionada, em que o objetivo é dividir os dados em *clusters*(grupos) de modo a que os seus constituintes sejam mais semelhantes entre si do que a dados de outros grupos.





Clustering – Temas

Alvin e os esquilos:

```
Cluster  136  =>  animation,cartoon,provide,movie,voices,kids,adventure,cgi,alvin,chipmunk
```

Pulp Fiction:

```
Cluster  135  =>  mobster,fiction,film,dialogue,movie,cool,pulp,travolta,shorty,russian
```

World War II:

```
Cluster  386  =>  point,heroism,human,film,convincing,reason,films,wwii,germans,war
```

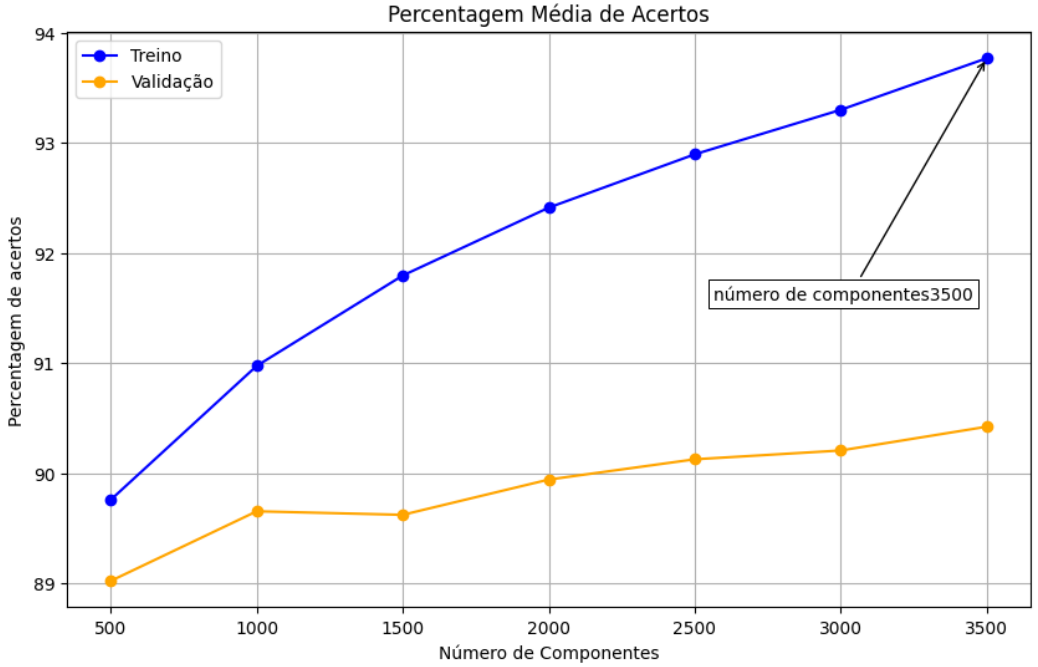
Rambo:

```
Cluster  249  =>  anti,allies,propaganda,partly,war,mission,vietnam,american,blood,rambo
```

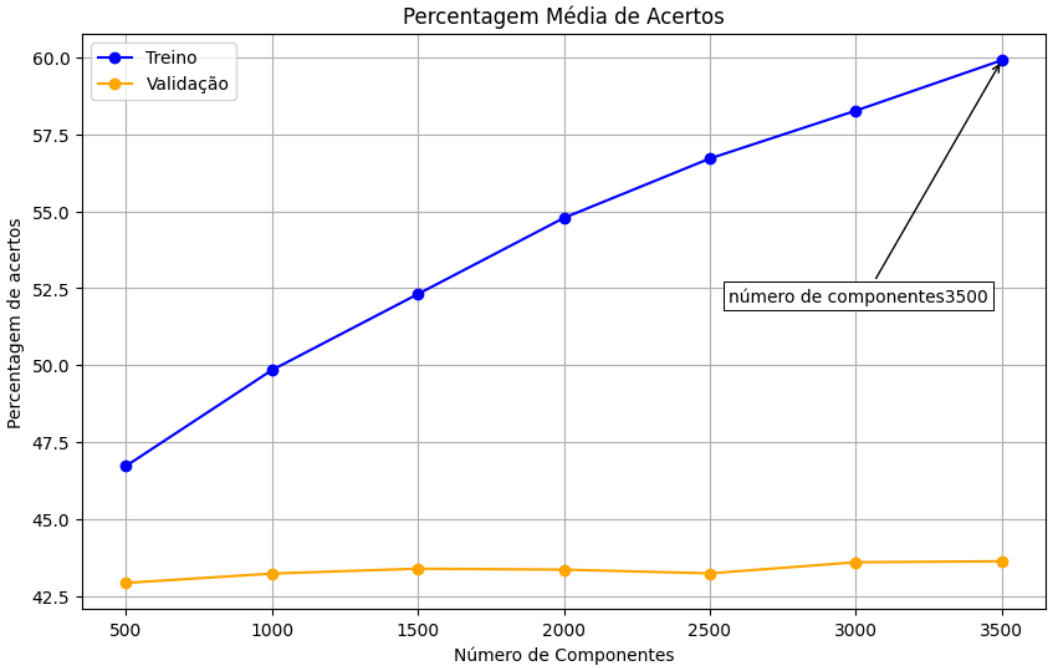
Star Trek:

```
Cluster  115  =>  spock,shatner,conveyed,star,episode,episodes,william,trek,vanishes,campbell
```

PCA – Análise de componentes principais



LR – Caso binário



LR – Caso multiclasse

Bibliografia

Gonçalo Marques, Métricas de distância e classificadores em distâncias, https://2223moodle.isel.pt/pluginfile.php/1180197/mod_resource/content/1/AP-Distancias.pdf

Gonçalo Marques, Trabalhar com dados de texto, https://2223moodle.isel.pt/pluginfile.php/1187034/mod_resource/content/1/AP-DadosTexto.pdf

Gonçalo Marques, Generalização de modelos lineares aplicados à classificação, https://2223moodle.isel.pt/pluginfile.php/1187053/mod_resource/content/1/AP-GenLMods.pdf