# Aprendizagem Automática

PROJETO FINAL – CLASSIFICAÇÃO DE CRÍTICAS DE CERVEJAS

DUARTE GONÇALVES – Nº46484

# Introdução

- Este trabalho tem como objetivo lidar com textos de críticas de cerveja, onde o principal problema é treinar vários tipos de classificadores diferentes capazes de classificar de forma automática .

- Para resolver o problema tem-se em mente que é necessário classificar essas criticas de duas formas diferentes:
- Binária, onde cada crítica é classificada como positivo ou negativo;
- Multi-classe, as críticas estão classificadas num intervalo de 0 a 5 ou 0 a 10;

Primeiramente, foi feito uma análise detalhada dos dados fornecidos avaliando os valores de treino e teste conforme o descrito anteriormente, e foram tratados conforme o necessário. Passando a seguir para a classificação desses dados pelos classificadores escolhidos.

# SVM (Support Vectorial Machines )

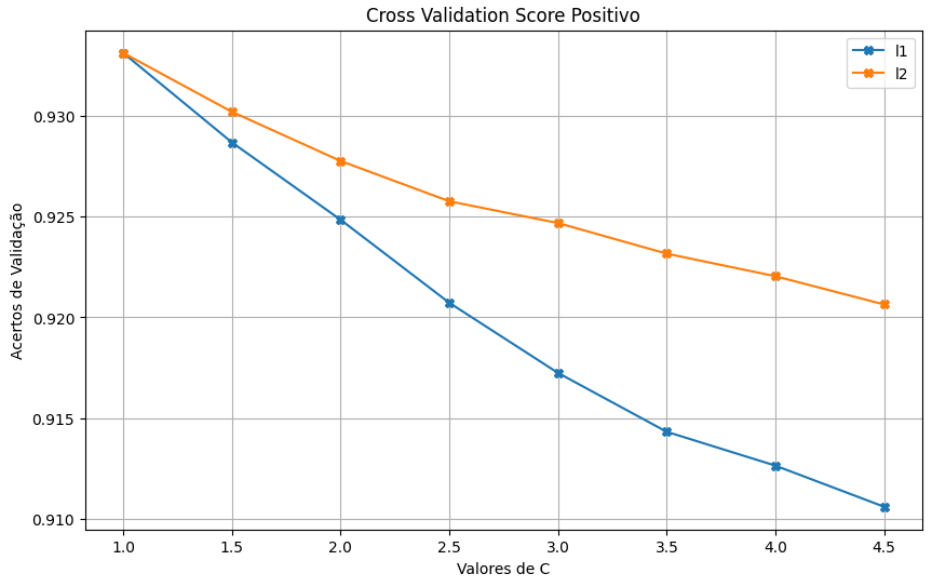
É um algoritmo usado na análise e classificação de dados, que pode ser usado na regressão como na classificação de dados. É um classificador que apresenta bons resultados em problemas de classificação com dados de alta dimensão.

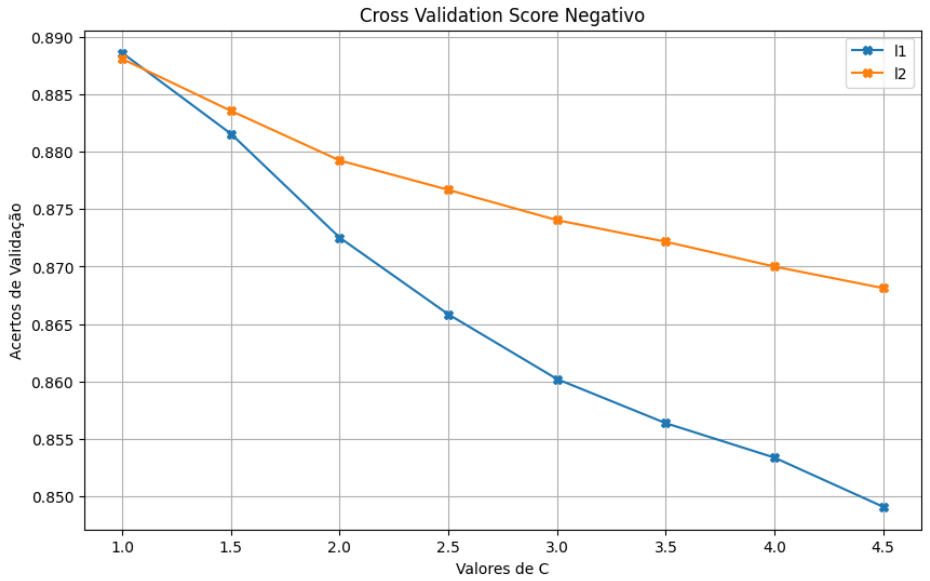
Neste projeto foi utilizado o LinearSVC, por ser um classificador que funciona bastante bem em problemas com dados de alta dimensão, e por ser um classificador bastante rápido na apresentação de resultados e devido também à sua eficienência.

Na obtenção dos seus melhores parâmetros, foi utilizado o Cross Validation Score por ser uma técnica, que apresenta um melhor score de validação, para o par de conjunto necessário, Penalização e Regularização;

# SVM – Classificação Binária

Procura de Melhores Parâmetros




Valores Positivos

- Penalização = L2
- C = 1.0
- Resultado de Acertos = 93,31%

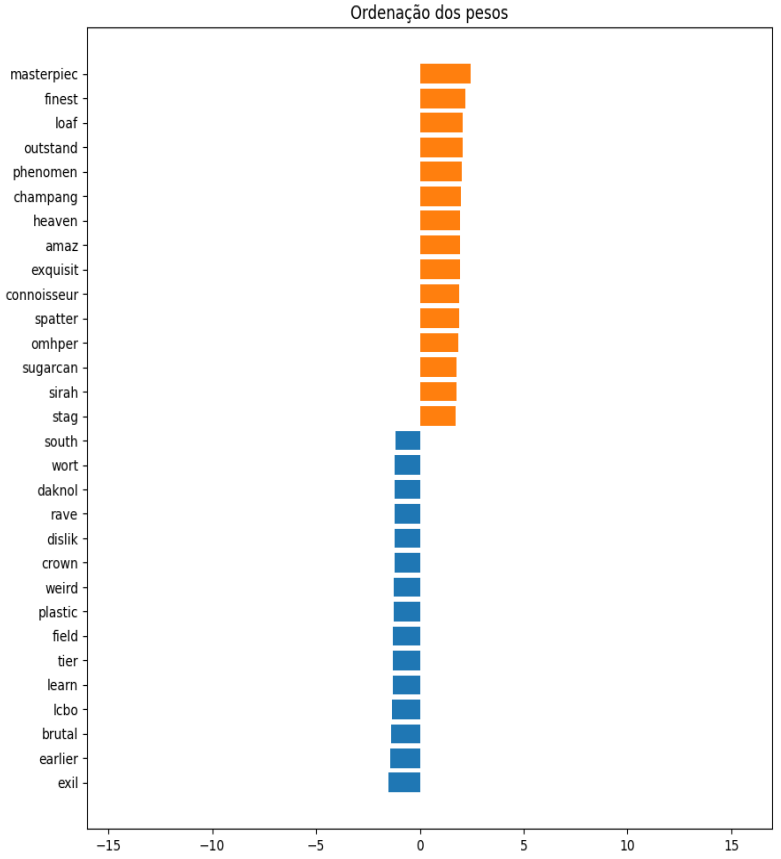
Valores Negativos

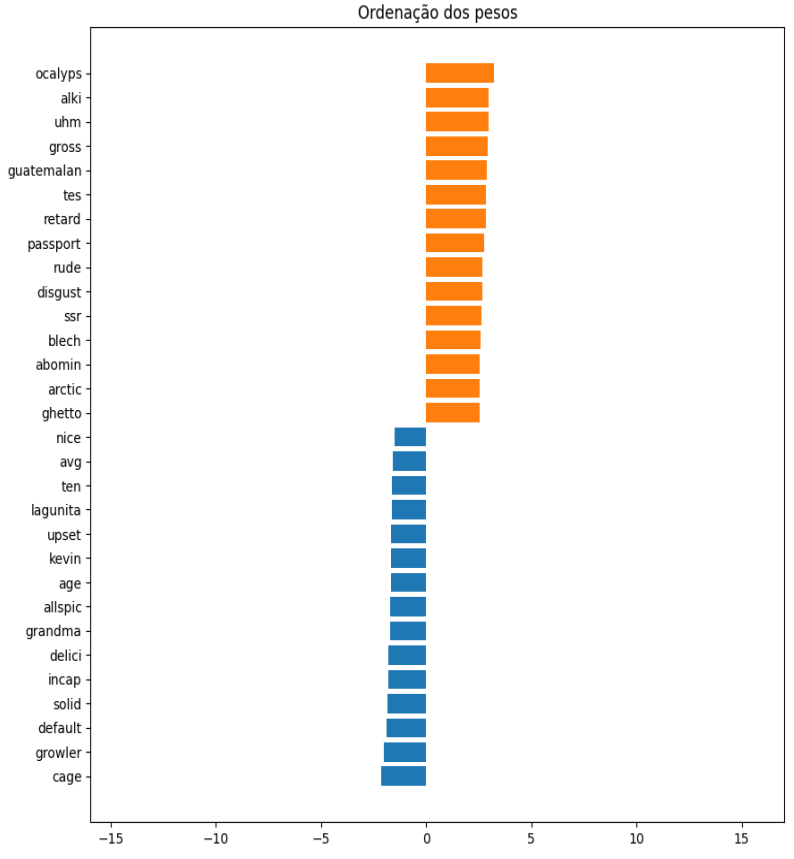
- Penalização = L1
- C = 1.0
- Resultado de Acertos = 88,86%

# SVM – Classificação Binária

## Resultados

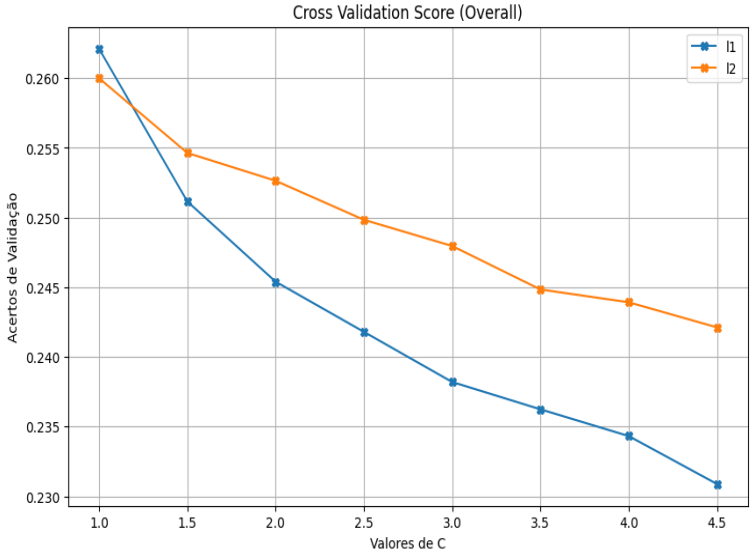
```
Tamanho de Vocabulário (Treino Positivo):  15179 | Score: 95.07%
Número de Erros Teste:  3696
Matriz de Confusão:
 [[69693   255]
 [ 3441  1611]]

Tamanho de Vocabulário (Treino Negativo):  15179 | Score: 96.98%
Número de Erros Teste:  2268
Matriz de Confusão:
 [[71468   186]
 [ 2082  1264]]

Tamanho de Vocabulário (Teste Positivo):  15179 | Score: 95.48%
Número de Erros Teste:  1130
Matriz de Confusão:
 [[23735   154]
 [  976   135]]

Tamanho de Vocabulário (Teste Negativo):  15179 | Score: 91.93%
Número de Erros Teste:  2018
Matriz de Confusão:
 [[22366   205]
 [ 1813   616]]
```



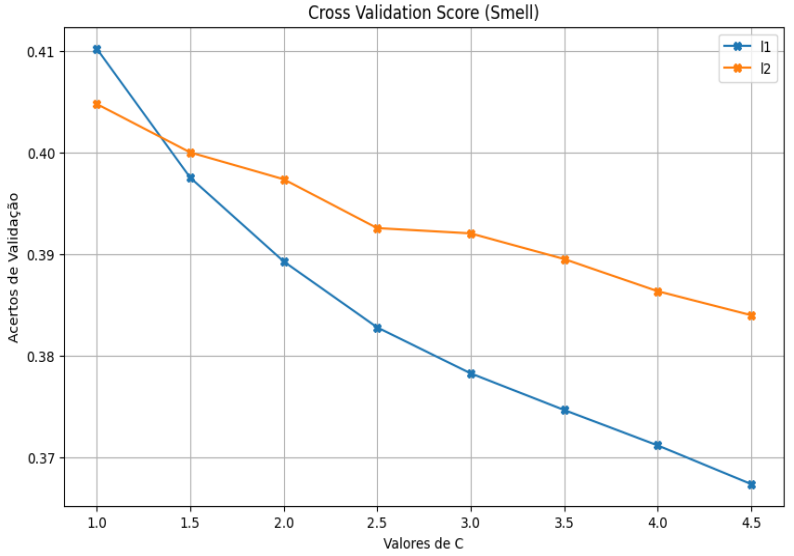



# SVM – Classificação Multi-Classe

Procura de Melhores Parâmetros

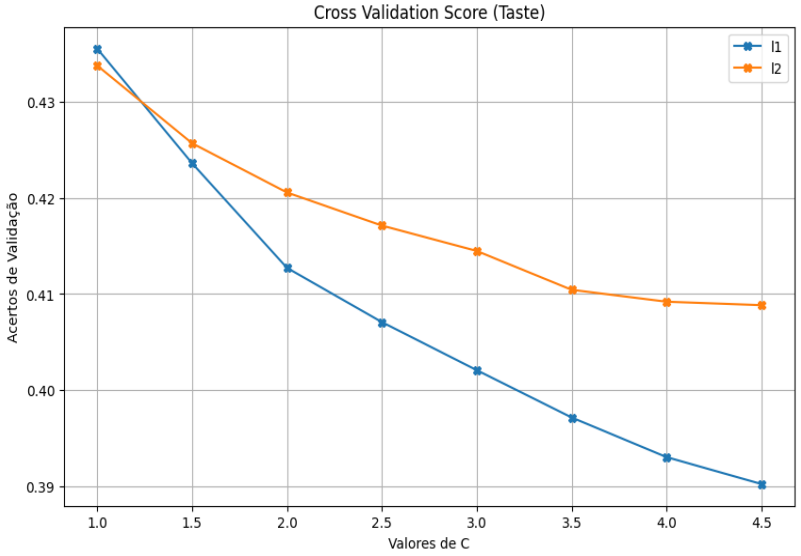



- Valores de Overall
  - Penalização = L1
  - C = 1,0
  - Resultado de Acertos = 26,21%




Valores Smell

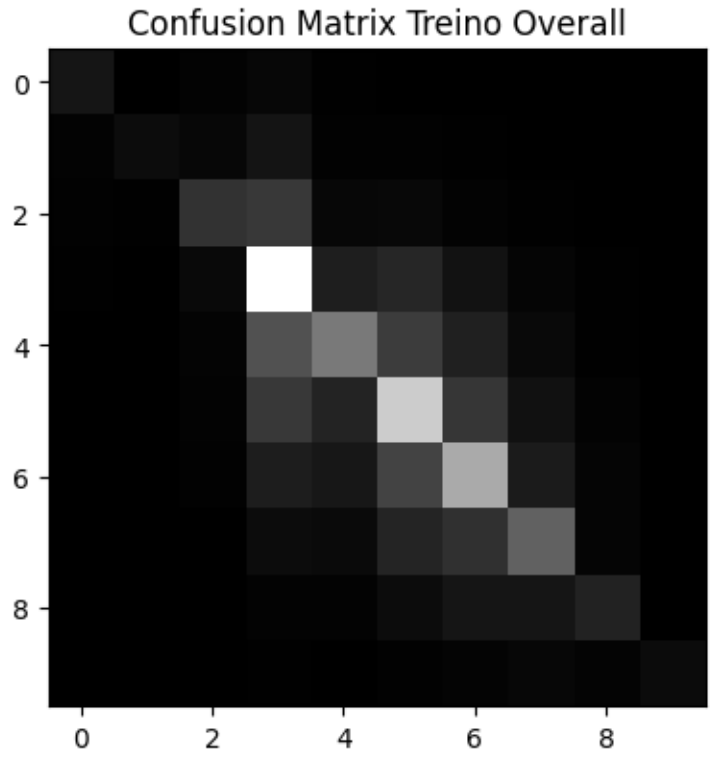
- Penalização = L1
- C = 1.0
- Resultado de Acertos = 41,02%

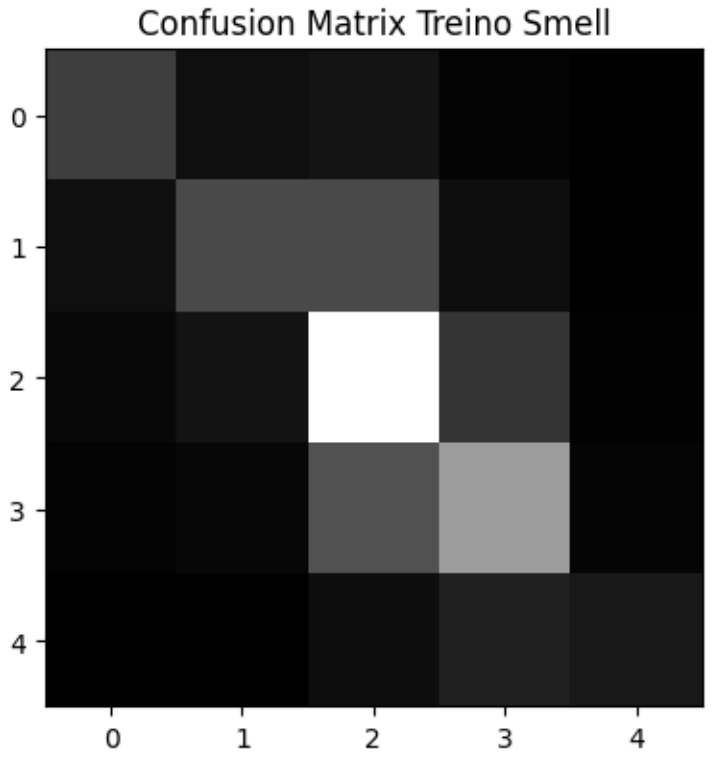


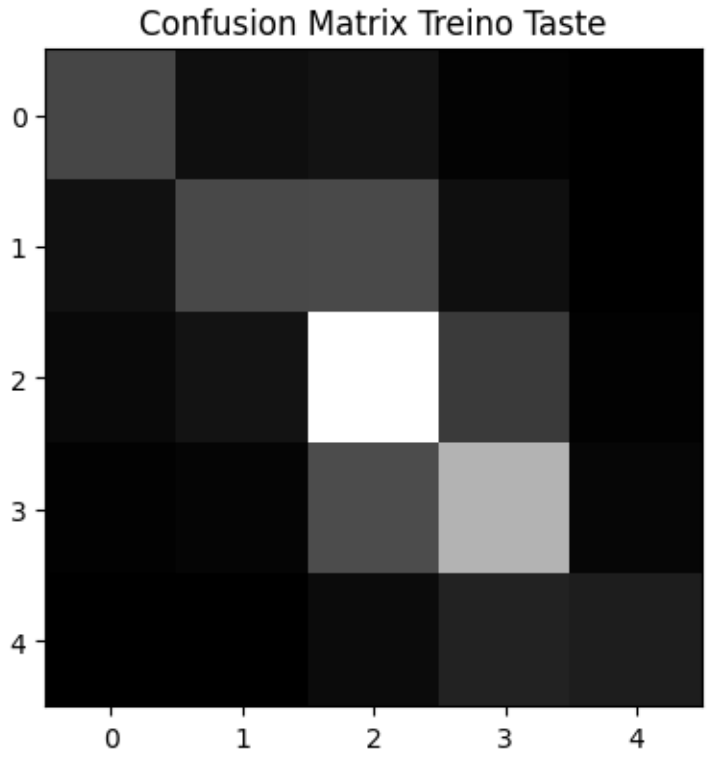

Valores Taste

- Penalização = L1
- C = 1,0
- Resultado de Acertos = 43,55%

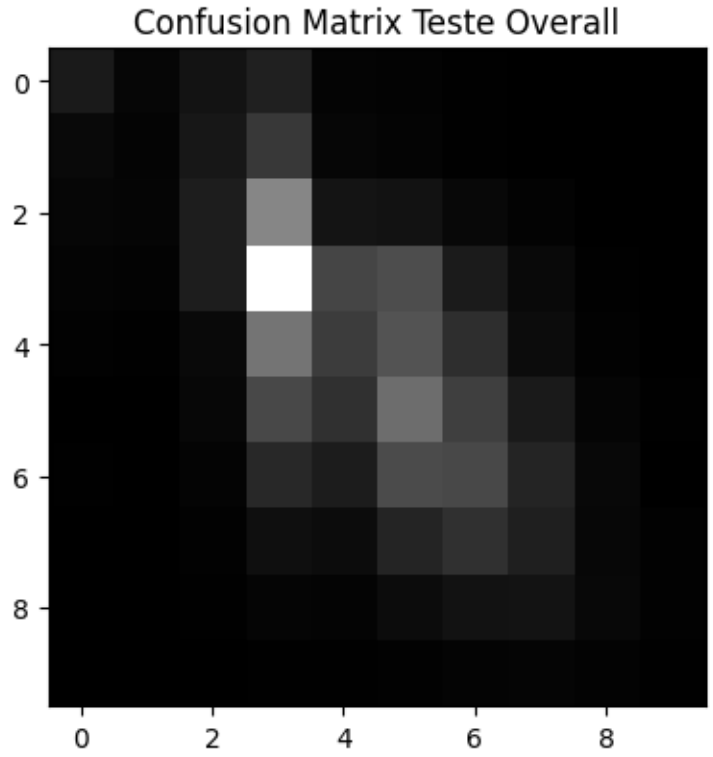
# SVM – Classificação Multi-Classe

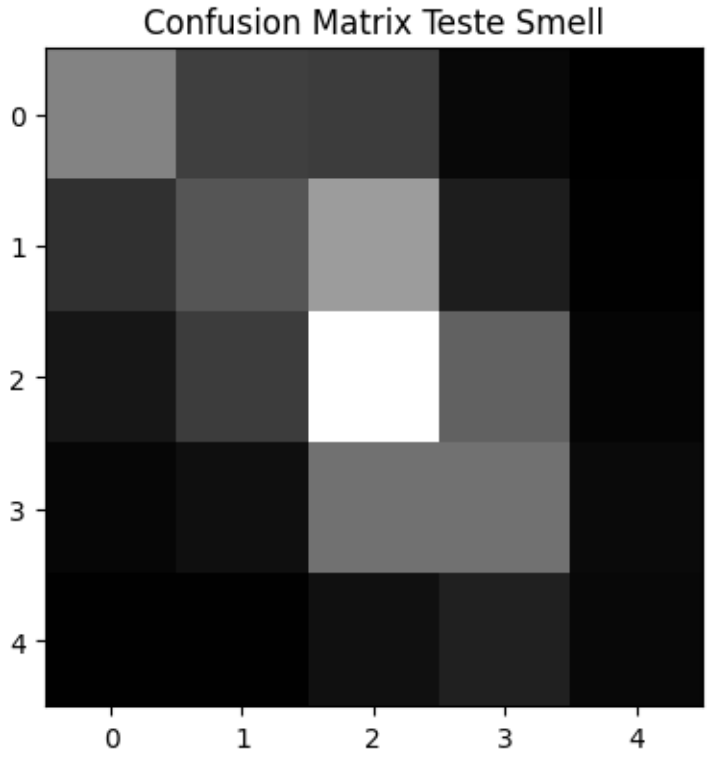


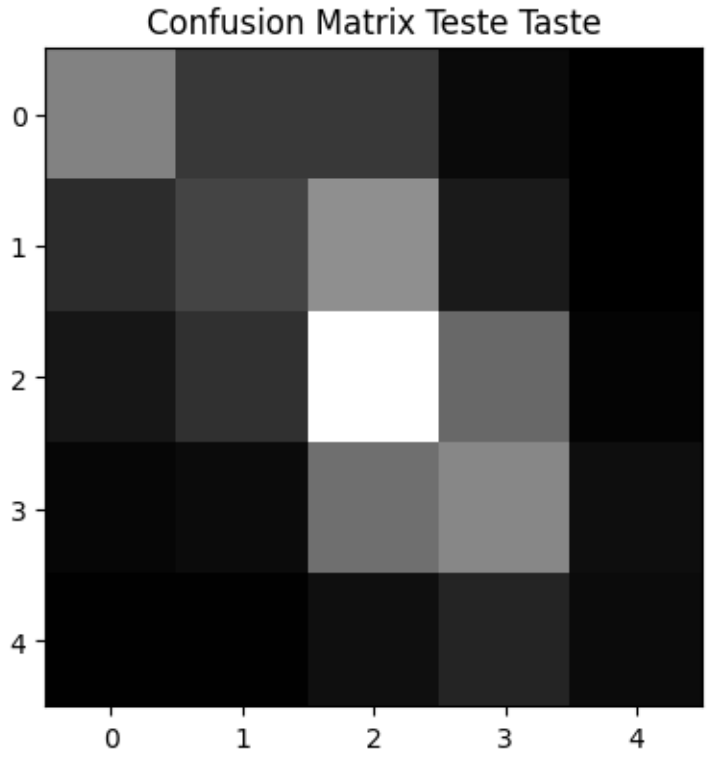



# SVM – Classificação Multi-Classe



# Regressão Logística

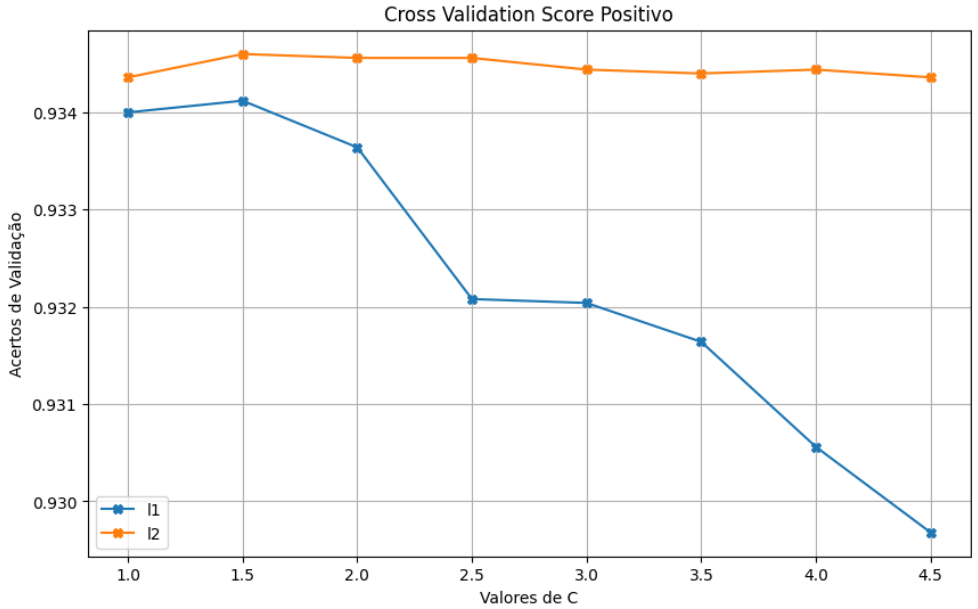
É um algoritmo baseado em observações anteriores de um conjunto de dados, prevendo com base na análise da relação entre uma ou mais variaveis independentes, isto é os dados estão categorizados permitindo saber a sua claassificação.

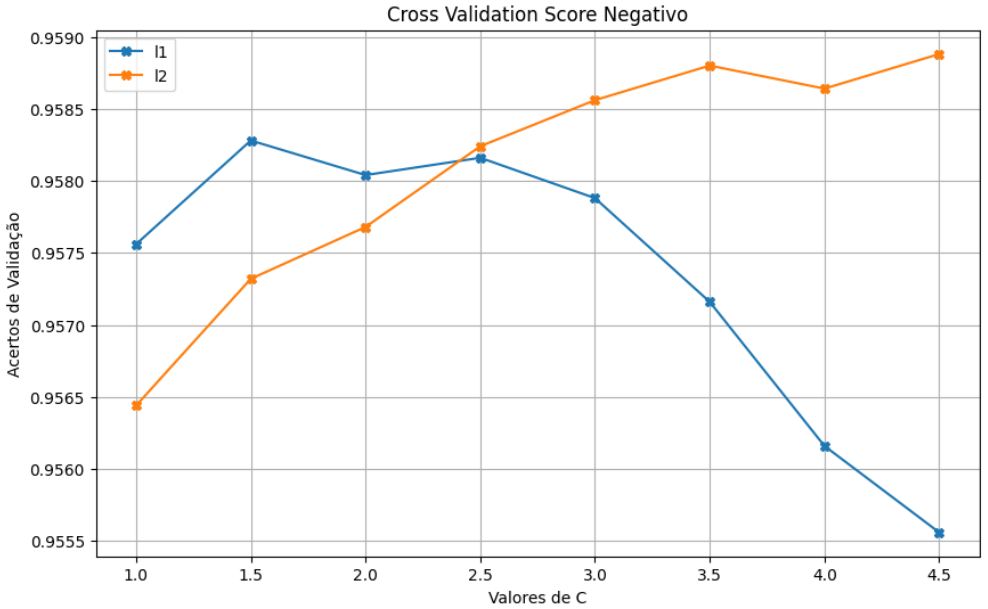
O motivo de utilização deste classificador prevem do facto deste algoritmo, obter bons resultados quando o conjunto de dados forncidos estiver linearmente separados, sendo por isso um bom classificador para a classificação binária, ser fácil de implementar e interepretar tendo um treino bastante eficiente.

Para a utilização deste algoritmo é necessário também obter um bom regulador e penalização, por isso para a obtenção destes valores é novamente utilizado o Cross Validation Score.

# Regressão Logística - Binário

Procura de Melhores




Valores Positivos

- Penalização = L2
- Regularizador C = 1.5
- Resultado de Acertos = 93.46%

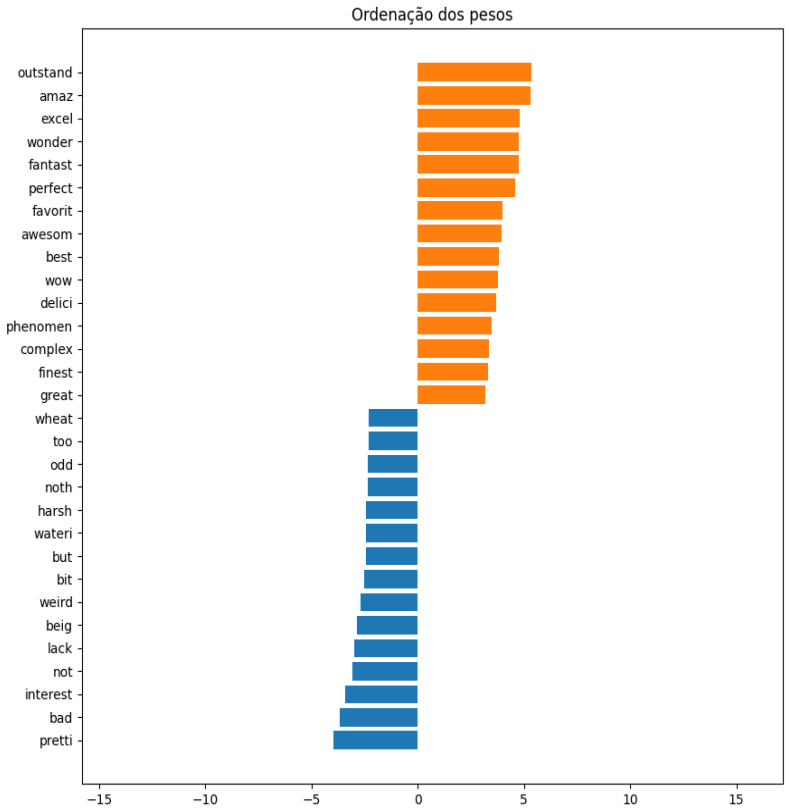
Valores Negativos

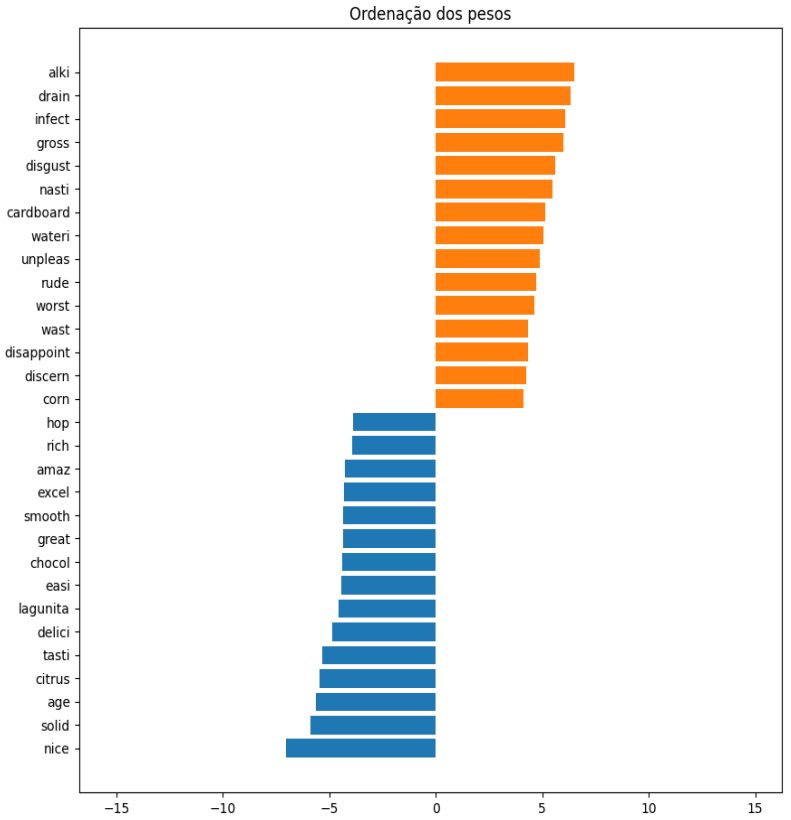
- Penalização = L2
- Regularizador C = 4.5
- Resultados de Acerrtos = 95.88%

# Regressão Logística - Binário

## Resultados Obtidos

```
Tamanho de Vocabulário (Treino Positivo):  15179 | Score: 94.16%
Número de Erros Teste:  4381
Matriz de Confusão:
 [[69595   353]
 [ 4028  1024]]

Tamanho de Vocabulário (Treino Negativo):  15179 | Score: 97.07%
Número de Erros Teste:  2200
Matriz de Confusão:
 [[71434   220]
 [ 1980  1366]]

Tamanho de Vocabulário (Teste Positivo):  15179 | Score: 95.55%
Número de Erros Teste:  1113
Matriz de Confusão:
 [[23763   126]
 [  987   124]]

Tamanho de Vocabulário (Teste Negativo):  15179 | Score: 92.18%
Número de Erros Teste:  1955
Matriz de Confusão:
 [[22319   252]
 [ 1703   726]]
```



# PCA

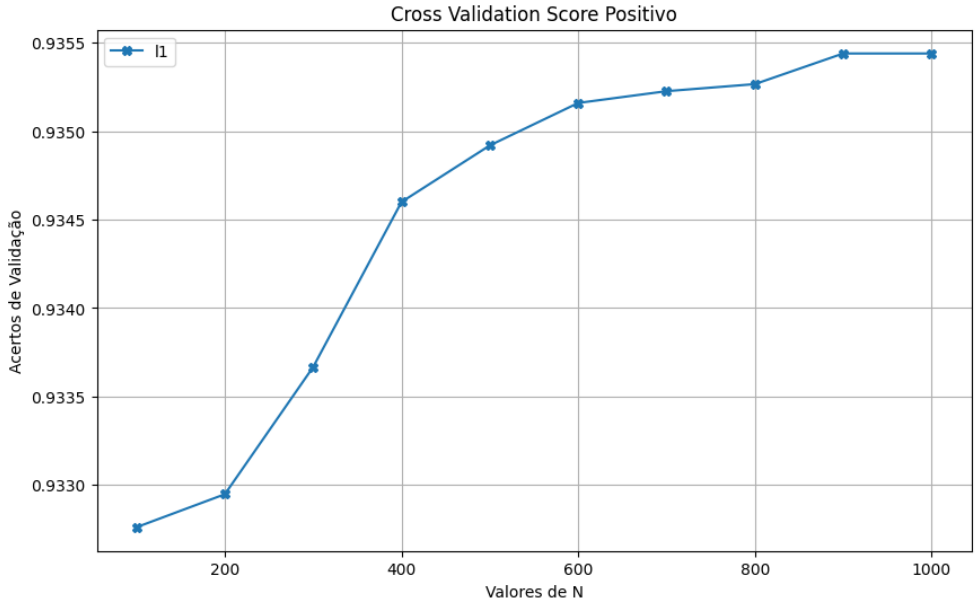
O algoritmo PCA, ao contrário dos outros classificadores utilizados, é do tipo de apredizagem não supervisionado, significamdo que este irá tentar aprender qual o melhor resultado dos dados fornecidos.
O PCA, é utilizado para diminuição da dimensionalidade dos dados fornecidos e também é uma ferramenta que permite identificar padrões ou estruturas de dados.

Utilizou-se este tipo de técnica, devido a sua diferenciação perante os restantes classificadores de tipo de aprendizagem e também devido a sua característica de diminuir os dados fornecidos de modo a obter melhores valores.

Uma vez que será necessário utilizador um outro classificador em conjunto para visualização de diferenças, é utilizado o classificador LinearSVC, por ter sido utilizado tanto para classificação binário como para a classificação multi-classe.

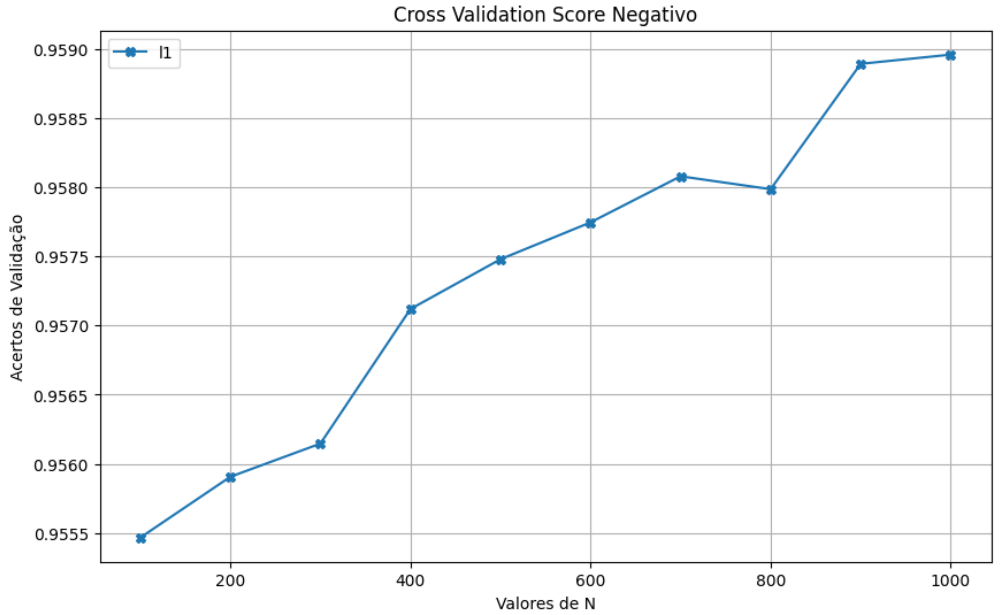
# PCA - Binário

Procura de Melhores Parâmetros




Valores Positivos

- Nº Componentes =  900
- Resultado de Acerto = 93.54%




Valores Negativos

- Nº de Componentes = 1000
- Resultado de Acertos = 95.89%

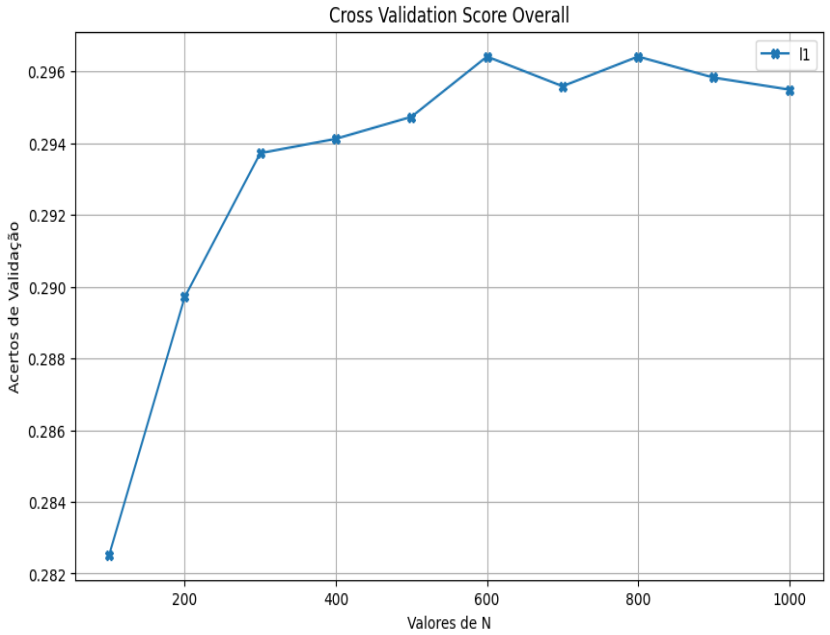
# PCA - Binário

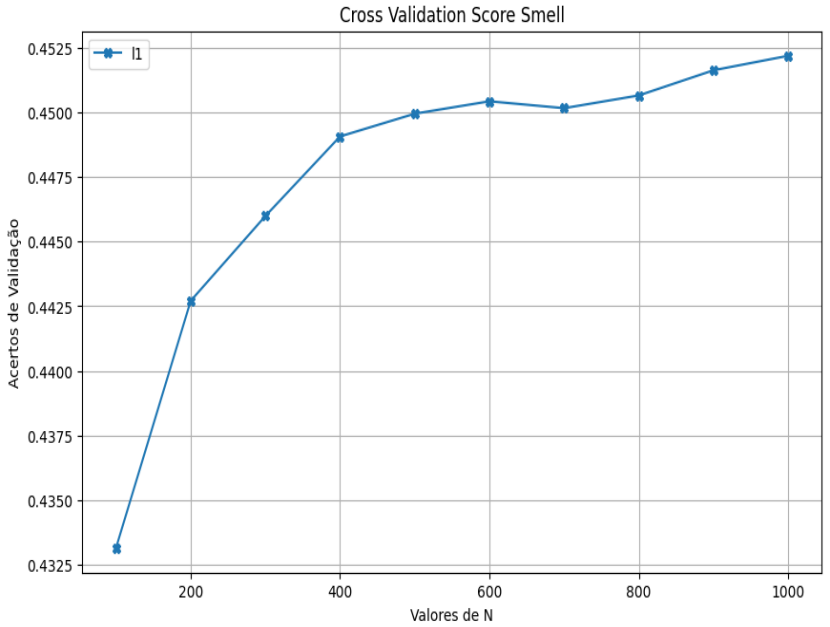
Resultados Obtidos

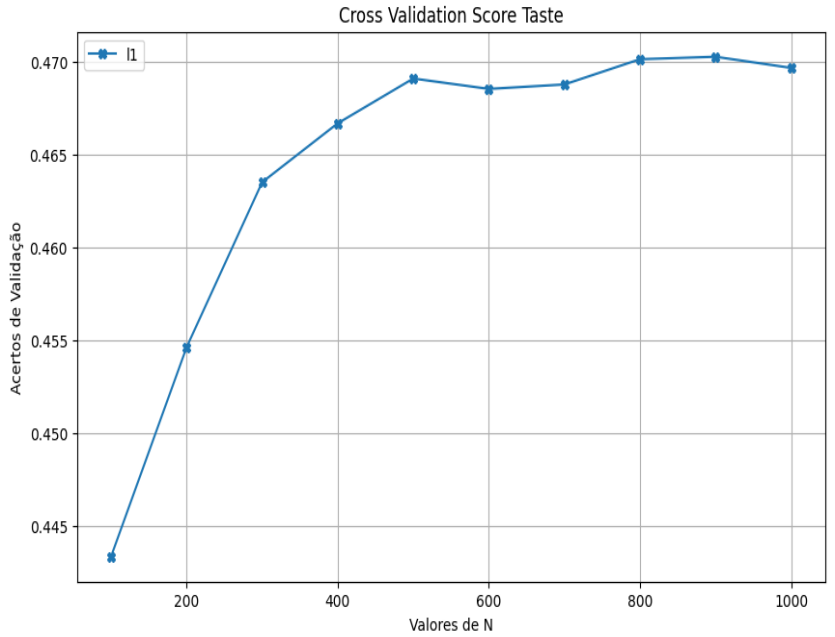
```
Tamanho de Vocabulário (Treino Positivo):  15179 | Score: 93.28%
Número de Erros Teste:  5037
Matriz de Confusão:
 [[69942     6]
 [ 5031    21]]

Tamanho de Vocabulário (Treino Negativo):  15179 | Score: 96.05%
Número de Erros Teste:  2960
Matriz de Confusão:
 [[71487   167]
 [ 2793   553]]

Tamanho de Vocabulário (Teste Positivo):  15179 | Score: 95.55%
Número de Erros Teste:  1113
Matriz de Confusão:
 [[23887     2]
 [ 1111     0]]

Tamanho de Vocabulário (Teste Negativo):  15179 | Score: 91.38%
Número de Erros Teste:  2156
Matriz de Confusão:
 [[22457   114]
 [ 2042   387]]
```

# PCA - Multi-Classe

Procura de Melhores Parâmetros




Valores Overall

- Nº de Componentes = 600
- Resultado de Acertos = 29.64%

Valores Smell

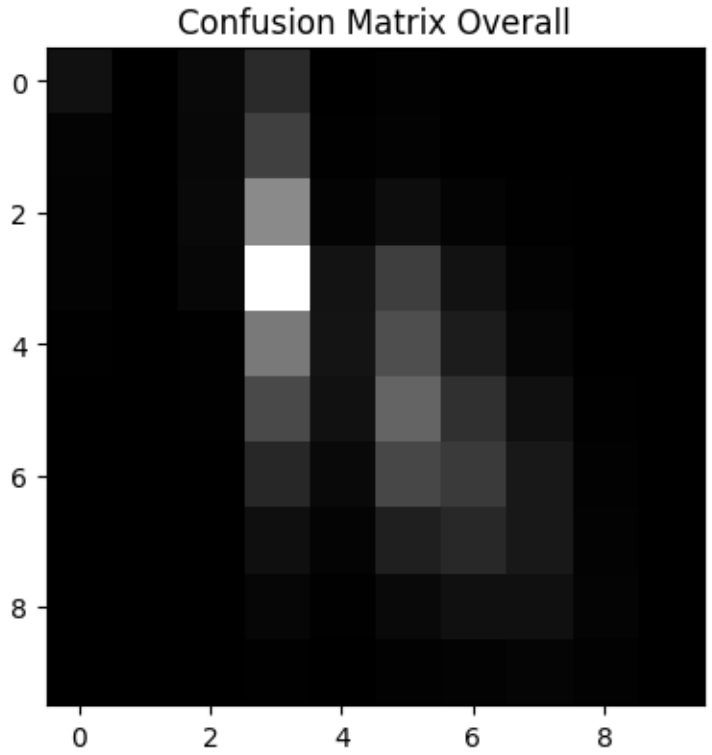
- Nº de Componentes = 1000
- Resultado de Acertos = 45.21%

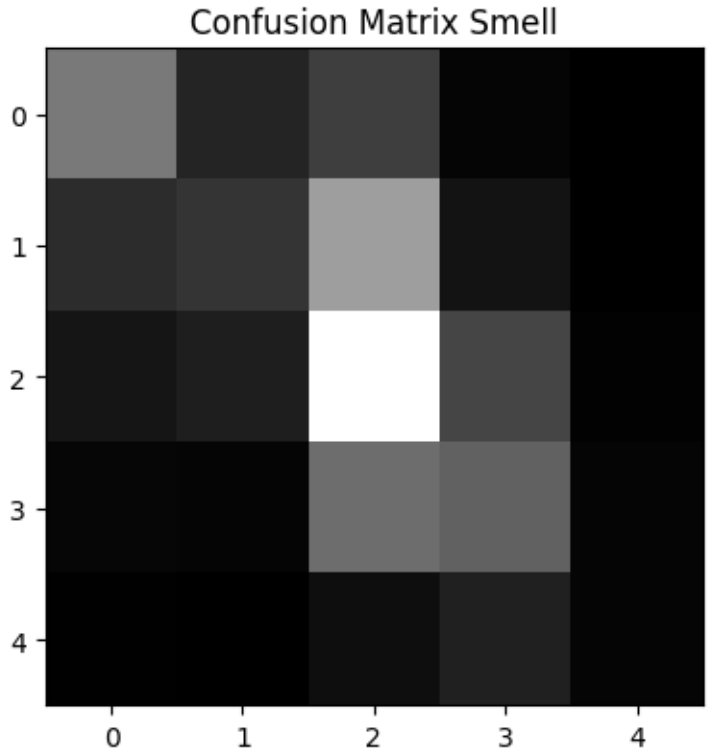
Valores Taste

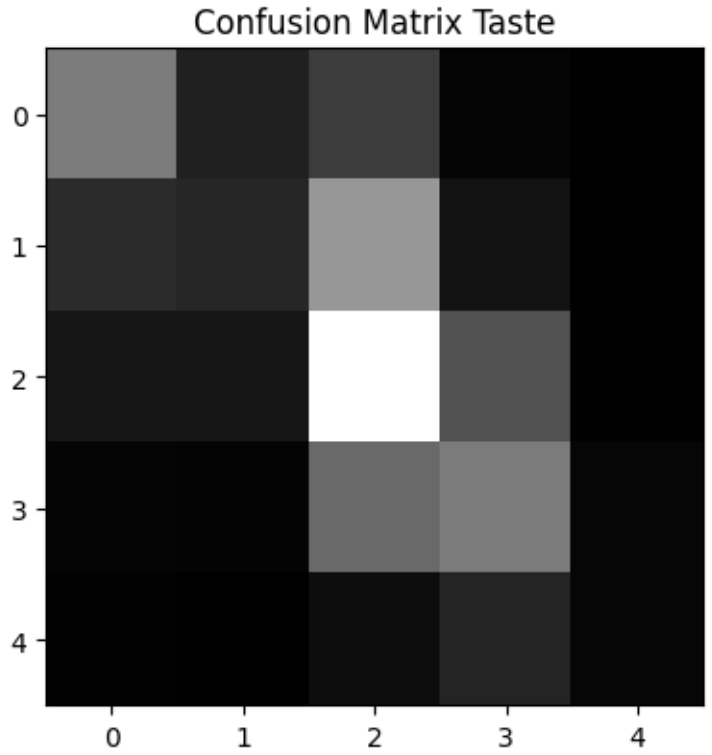
- Nº de Componentes = 900
- Resultado de Acerto = 47.02%

# PCA - Multi-Classe

Resultados Obtidos



# Comparações entre Classificadores

## Binário (Positivo)

| Classificador | Parâmetros | Score Val. (%) | Score Teste (%) | Erros Teste |
|---|---|---|---|---|
| **Regressão logística** | p=L2, C=1.5 | 93.46 | 95.55 | 1113 |
| **LinearSVC** | p=L2, C=1.0 | 93.31 | 95.48 | 1130 |
| **PCA + LinearSVC** | N = 900 | 93.54 | 95.55 | 1113 |

## Binário (Negativo)

| Classificador | Parâmetros | Score Val. (%) | Score Teste (%) | Erros Teste |
|---|---|---|---|---|
| **Regressão logística** | p=L2, C=4.5 | 95.88 | 92.18 | 1955 |
| **LinearSVC** | p=L1, C=1.0 | 88.86 | 91.93 | 2018 |
| **PCA + LinearSVC** | N = 1000 | 95.89 | 91.38 | 2156 |

## Multi-Classe (Overall)

| Classificador | Parâmetros | Score Val. (%) | Score Teste (%) | Erros Teste |
|---|---|---|---|---|
| **LinearSVC** | p=L1, C=1.0 | 26.21 | 28.52 | 17871 |
| **PCA + LinearSVC** | N = 600 | 29.64 | 30.21 | 17448 |

## Multi-Classe (Smell)

| Classificador | Parâmetros | Score Val. (%) | Score Teste (%) | Erros Teste |
|---|---|---|---|---|
| **LinearSVC** | p=L1, C=1.0 | 41.02 | 43.70 | 14076 |
| **PCA + LinearSVC** | N = 1000 | 45.21 | 45.64 | 13590 |

## Multi-Classe (Taste)

| Classificador | Parâmetros | Score Val. (%) | Score Teste (%) | Erros Teste |
|---|---|---|---|---|
| **LinearSVC** | p=L1, C=1.0 | 43.55 | 45.03 | 13743 |
| **PCA + LinearSVC** | N = 900 | 47.02 | 46.75 | 13313 |

# Conclusões

Fazendo uma comparação dos valores anteriormente mostrados onde para o caso binário o melhor classificador a usar será a Regressão Logística uma vez que esta apresenta uma média de scores de teste igual a 93.86% enquanto que o LinearSVC apresenta 92.70% e o PCA 93.47%. A Regressão Logistica ainda apresenta 3068 erros de ambos os valores, um valor inferior comparativamente com os outros classificadores.

Para a multiclasse, uma vez que, se fez apenas para o LinearSVC para os ambas as classificações, tentou-se que com o PCA se pode-se obter melhores valores de LinearSVC quando este dois "trabalhassem" juntos. Por isso, quando aplicado PCA no LinearSVC, obtem-se um valor percentual de Cross Validation superior que o LinearSVC base. Portanto, apesar de não se ter os restantes resultados, podemos prever que com o PCA os valores de LinearSVC serão bastante melhores e por isso uma boa dupla para resultados multi-classe.