

1 2



9 0

FACULDADE DE  
CIÊNCIAS E TECNOLOGIA  
UNIVERSIDADE DE  
**COIMBRA**

## Heart Disease Prediction

### Final Project Milestone



Mestrado em Engenharia Informática

Pattern Recognition  
2021 / 2022

PL2	Gabriel Fernandes	2018288117	<a href="mailto:gabrielf@student.dei.uc.pt">gabrielf@student.dei.uc.pt</a>
PL2	Miguel Rabuge	2018293728	<a href="mailto:rabuge@student.dei.uc.pt">rabuge@student.dei.uc.pt</a>
PL2	Pedro Rodrigues	2018283166	<a href="mailto:pedror@student.dei.uc.pt">pedror@student.dei.uc.pt</a>

# Conteúdo

<b>Introdução</b>	2
1 Aplicação	2
2 Data Preprocessing	2
2.1 Carregamento	2
2.2 Scaling	4
3 Pesquisa de hyperparameters	4
3.1 K - Nearest Neighbors	4
3.2 Support Vector Machines	7
<b>Cenário A</b>	8
4 Redução e Seleção de Features	8
4.1 Seleção de Features	8
4.1.1 Kruskal-Wallis	8
4.2 Redução de Features	10
4.2.1 Principal Component Analysis	10
4.2.2 Linear Discriminant Analysis	11
5 Classificadores	12
5.1 Minimum Distance Classifiers	12
5.1.1 Euclidean Linear Discriminant	12
5.1.2 Mahalanobis Linear Discriminant	12
5.2 Fisher's Linear Discriminant	13
5.3 Bayes Classifier	13
5.4 K-Nearest Neighbors	14
5.5 Support Vector Machines	15
<b>Cenário B</b>	17
6 Redução e Seleção de Features	17
6.1 Seleção de Features	17
6.1.1 Kruskal-Wallis	17
6.2 Redução de Features	17
6.2.1 Principal Component Analysis	17
6.2.2 Linear Discriminant Analysis	19
7 Classificadores	19
7.1 Minimum Distance Classifiers	19
7.1.1 Euclidean Linear Discriminant	19
7.1.2 Mahalanobis Linear Discriminant	19
7.2 Fisher's Linear Discriminant	20
7.3 Bayes Classifier	21
7.4 K-Nearest Neighbors	21
7.5 Support Vector Machines	22

<b>Cenário C</b> .....	23
8 Redução e Seleção de Features .....	23
8.1 Seleção de Features .....	23
8.1.1 Kruskal-Wallis .....	23
8.2 Redução de Features .....	23
8.2.1 Principal Component Analysis .....	23
8.2.2 Linear Discriminant Analysis .....	23
9 Classificadores .....	25
9.1 Minimum Distance Classifiers .....	25
9.1.1 Euclidean Linear Discriminant .....	25
9.1.2 Mahalanobis Linear Discriminant .....	25
9.2 Fisher's Linear Discriminant .....	26
9.3 Bayes Classifier .....	27
9.4 K-Nearest Neighbors .....	27
9.5 Support Vector Machines .....	28

## Lista de Figuras

1 Ecrã principal da aplicação .....	3
2 Ecrã de resultados da aplicação .....	3
3 Pesquisa KNN para o cenário A .....	5
4 Pesquisa KNN para o cenário B .....	5
5 Pesquisa KNN para o cenário C .....	6
6 Pesquisa SVM 2D .....	7
7 Pesquisa SVM 3D .....	7
8 Box Plots dos atributos para as classes (0) NO CHD e (1) CHD .....	8
9 Valores ordenados das estatísticas de teste .....	9
10 Valores normalizados cumulativos das estatísticas de teste .....	9
11 Scree plot com kaiser criterion .....	10
12 Variância Cumulativa .....	11
13 Box Plots dos atributos para as classes (0) NO HD e (1) HD .....	17
14 Valores ordenados das estatísticas de teste .....	18
15 Valores normalizados cumulativos das estatísticas de teste .....	18
16 Box Plots dos atributos para as classes (0) NO HD, (1) HD e (2) HDC .....	23
17 Valores ordenados das estatísticas de teste .....	24
18 Valores normalizados cumulativos das estatísticas de teste .....	24

## Lista de Tabelas

1 KW + ELD test scores .....	12
2 PCA + ELD test scores .....	12
3 LDA + ELD test scores .....	12
4 KW + MLD test scores .....	12
5 PCA + MLD test scores .....	12

6	LDA + MLD test scores .....	13
7	KW + FLD test scores .....	13
8	PCA + FLD test scores .....	13
9	LDA + FLD test scores .....	13
10	KW + BC test scores .....	13
11	PCA + BC test scores .....	14
12	LDA + BC test scores .....	14
13	KW + KNN test scores .....	14
14	PCA + KNN test scores .....	14
15	LDA + KNN test scores .....	14
16	KW + SVM test scores .....	15
17	PCA + SVM test scores .....	15
18	LDA + SVM test scores .....	15
19	KW + ELD test scores .....	19
20	PCA + ELD test scores .....	19
21	LDA + ELD test scores .....	19
22	KW + MLD test scores .....	19
23	PCA + MLD test scores .....	20
24	LDA + MLD test scores .....	20
25	KW + FLD test scores .....	20
26	PCA + FLD test scores .....	20
27	LDA + FLD test scores .....	20
28	KW + BC test scores .....	21
29	PCA + BC test scores .....	21
30	LDA + BC test scores .....	21
31	KW + KNN test scores .....	21
32	PCA + KNN test scores .....	21
33	LDA + KNN test scores .....	22
34	KW + SVM test scores .....	22
35	PCA + SVM test scores .....	22
36	LDA + SVM test scores .....	22
37	KW + ELD test scores .....	25
38	PCA + ELD test scores .....	25
39	LDA + ELD test scores .....	25
40	KW + MLD test scores .....	25
41	PCA + MLD test scores .....	26
42	LDA + MLD test scores .....	26
43	KW + FLD test scores .....	26
44	PCA + FLD test scores .....	26
45	LDA + FLD test scores .....	26
46	KW + BC test scores .....	27
47	PCA + BC test scores .....	27
48	LDA + BC test scores .....	27
49	KW + KNN test scores .....	27
50	PCA + KNN test scores .....	28

51	LDA + KNN test scores .....	28
52	KW + SVM test scores .....	28
53	PCA + SVM test scores .....	28
54	LDA + SVM test scores .....	28

# Introdução

## 1 Aplicação

Primeiramente, começamos por explicar o funcionamento da aplicação. Esta contém 6 painéis, enumerados de 1 a 6, cujo processamento é sequencial, como podemos observar na figura 1. Importa salientar, que ao carregar no botão “Classify” no painel 6, a aplicação salta o painel 5, não fazendo a pesquisa de parâmetros.

O primeiro painel, permite ao utilizador escolher o cenário A, B ou C.

O segundo, permite ou não normalizar os dados.

O terceiro, permite fazer feature assessment, carregando no botão “Assess” originando C figuras, onde C é o número de classes do cenário, que contém histogramas com a função normal sobreposta, bem como a função de repartição empírica vs a normal padrão, por forma a inspecionar graficamente a normalidade dos dados, de cada classe.

O quarto painel, permite escolher o método de seleção / redução de dados, nomeadamente Kruskal-Wallis, PCA e LDA. O campo N features permite definir o número das “features” mais discriminantes que o utilizador pretende passar aos classificadores no painel 6. O botão “Plot” mostra, graficamente, ao utilizador o valor dos eigenvalues ordenados, no caso do PCA e do LDA, e o valor das estatísticas de teste ordenadas no caso do Kruskal-Wallis, bem como a versão cumulativa destas.

O quinto painel, permite ao utilizador fazer uma pesquisa dos hyperparameters para determinados modelos, podendo escolher o número de folds a utilizar na cross-validation, bem como os parâmetros que pretende testar, que se alteram dinamicamente na interface com base no modelo escolhido. O botão “Search” inicia esta pesquisa, produzindo plots específicos no final da execução.

Por fim, o sexto painel permite ao utilizador escolher o classificador para dados. Analogamente ao painel 5, este irá mostrar dinamicamente os parâmetros a configurar, para modelos parametrizáveis. Ao carregar no botão “Classify”, no final da classificação, irá aparecer uma interface com os resultados, como se pode observar na figura 2.

## 2 Data Preprocessing

### 2.1 Carregamento

Os dados foram carregados através do ficheiro CSV fornecido “heart\_2020\_cleaned”, sendo divididos aleatoriamente em dois sets: 200.000 entries para treino e o restante para teste. O carregamento é feito através da função “load\_dataset” que se encarrega de gerar as estruturas e classes definidas para cada um dos cenários.



Pattern Recognition @2022 Project

### Heart Disease

1 - Scenarios  
Select Scenario A ▼

2 - Pre-processing  
☐ Scale

3 - Feature Assessment  
Assess

4 - Feature Selection / Reduction

Selection  
☐ Kruskal-Wallis

Reduction  
Model None ▼

N Features

Plot

5 - Hyper-parameters Tunning

Model K - Nearest Neighbors ▼

N Folds

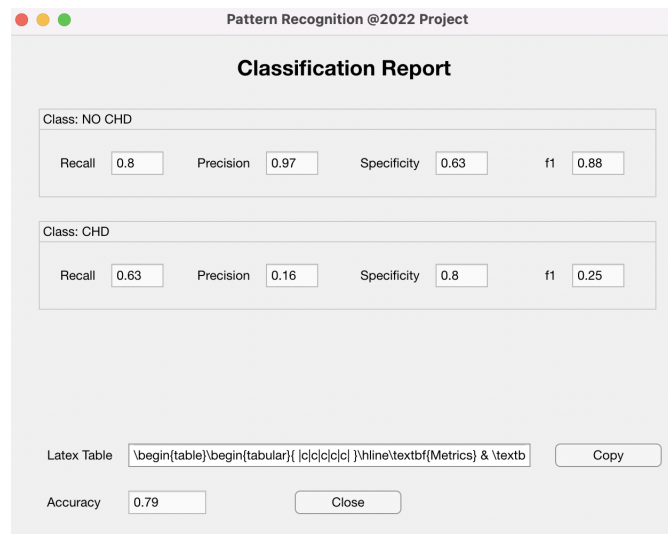
Neighbors

Search

6 - Classification

Model Euclidean Linear Discriminant ▼

Classify

**Figura 1.** Ecrã principal da aplicação

Pattern Recognition @2022 Project

### Classification Report

Class: NO CHD

Recall  Precision  Specificity  f1

Class: CHD

Recall  Precision  Specificity  f1

Latex Table

Copy

Accuracy  Close

**Figura 2.** Ecrã de resultados da aplicação

- Classes Cenário A:
  - No Coronary Heart Disease (*NO CHD*)
  - Coronary Heart Disease (*CHD*)
- Classes Cenário B:
  - No Heart Disease (*NO HD*)
  - Heart Disease (*HD*)
- Classes Cenário C:
  - No Heart Disease (*NO HD*)
  - Heart Disease (*HD*)
  - Heart Disease with Comorbidities (*HDC*)

## 2.2 Scaling

Depois de carregados os dados, é necessário normalizá-los devido à sensibilidade que certos modelos, como o PCA, têm sobre a variância. No caso do PCA, a não normalização destes dados poderia provocar a atribuição de quase toda a variância a uma determinada direção devido a ter uma escala de variância muito maior, relativamente às restantes, apesar de poderem até ser semelhantes. Ao normalizar, subtraindo a média e dividindo pelo desvio padrão, garantimos que isto não acontece.

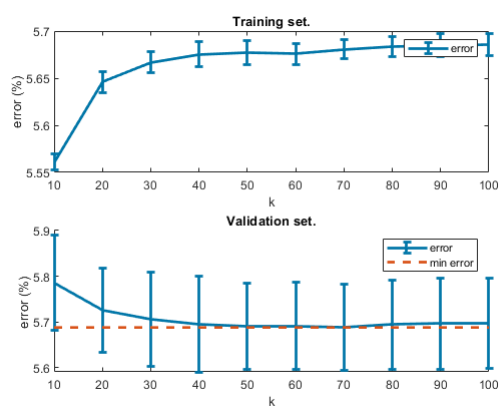
## 3 Pesquisa de hyperparameters

Nesta secção explicamos a forma como fizemos a pesquisa dos hyperparameters, dada a impossibilidade temporal de fazermos a análise como desejávamos, isto seria: k-fold cross-validation com grid-search dos parâmetros. Assim, optamos por alternativas mais simplistas, menos corretas, que consideramos "adequadas", dado o contexto.

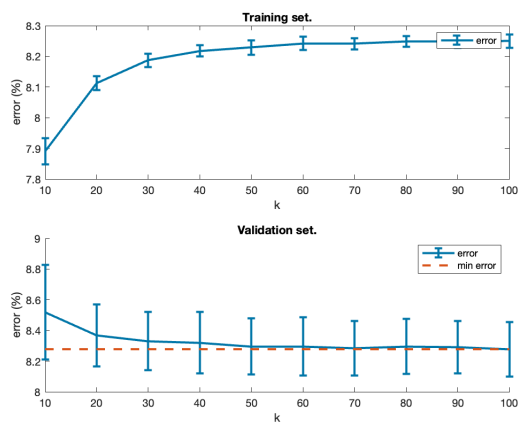
### 3.1 K - Nearest Neighbors

Para o KNN, fizemos uma pesquisa de neighbors entre 10 e 100, com intervalos de 10 em 10, para cada um dos três cenários, com k-fold cross-validation de 10 folds. Os dados foram normalizados e pré-processados com PCA, do qual escolhemos 11 "features", que correspondem a mais de 80% de variância, como justificado mais à frente. Os três resultados, para o cenário A, B e C são apresentados nas figuras 3, 4 e 5. Daqui retiramos que para o cenário A, utilizaremos 70 neighbors, para o B serão 100, e para o C serão 50.

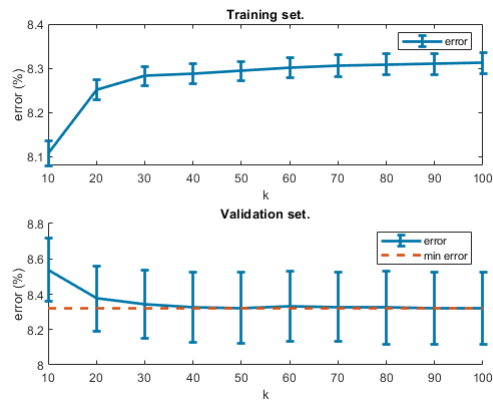




**Figura 3.** Pesquisa KNN para o cenário A



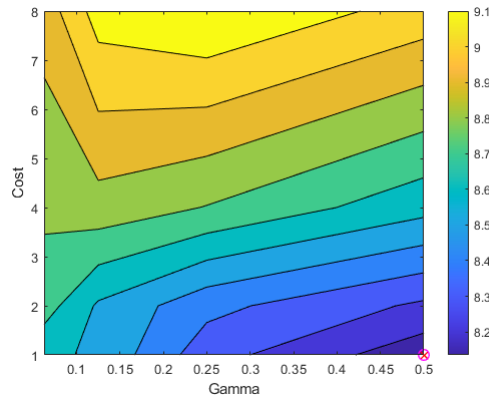
**Figura 4.** Pesquisa KNN para o cenário B



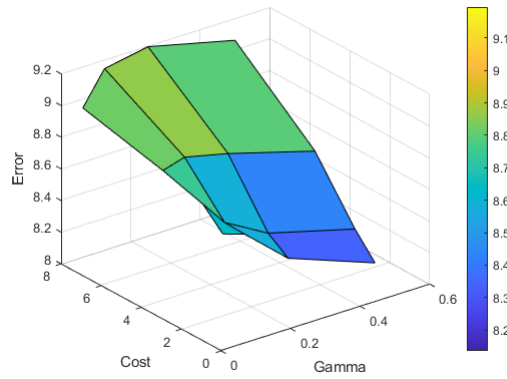
**Figura 5.** Pesquisa KNN para o cenário C

### 3.2 Support Vector Machines

Relativamente às SVMs, considerando que o tamanho do dataset é demasiado grande para fazer k-fold cross-validation com grid-search, para vários parâmetros de Cost e Gamma, dada a combinatória, optamos por reduzir o dataset de treino/validação em 10%, ficando com 20.000 entries. Deste modo, estamos agora em condições de aplicar o método acima. Escolhemos aplicar ao cenário C, com os valores de Cost  $2^0, 2^1, 2^2$  e  $2^3$  e de Gamma  $2^{-4}, 2^{-3}, 2^{-2}$  e  $2^{-1}$ , com 10 folds. Desta pesquisa resultou que os melhores parâmetros são Cost = 1 e Gamma = 0.5, como podemos observar nas figuras 6 e 7.



**Figura 6.** Pesquisa SVM 2D



**Figura 7.** Pesquisa SVM 3D

## Cenário A

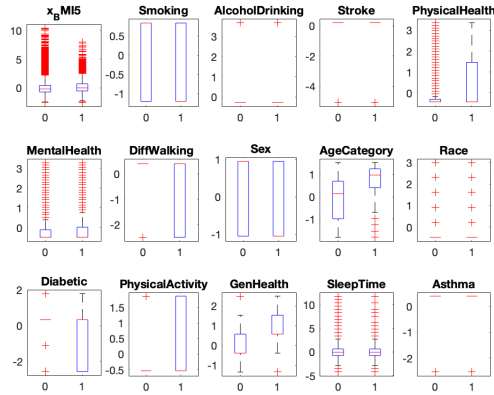
### 4 Redução e Seleção de Features

Nesta secção apresentamos formas de reduzir/selecionar as features do dataset original, com o objetivo de identificar features que realmente importam na classificação, reduzindo assim o tempo que os modelos mais complexos levam a correr, bem como o espaço para armazenar os dados. De igual modo, de um ponto de vista abstrato, utilizando apenas os dados mais relevantes, obrigamos os modelos a focarem-se naqueles atributos que importam, reduzindo as chances de os induzir em erro.

#### 4.1 Seleção de Features

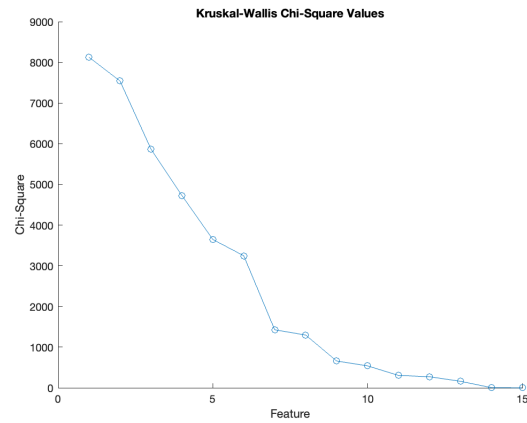
##### 4.1.1 Kruskal-Wallis

Em termos de seleção de features, apresentamos o método de análise de discriminação dos dados recorrendo ao teste não paramétrico Kruskal-Wallis. Para tal, iremos analisar a estatística de teste do mesmo, para cada atributo do dataset scaled. Aqueles com maiores estatísticas de teste, serão os mais relevantes.

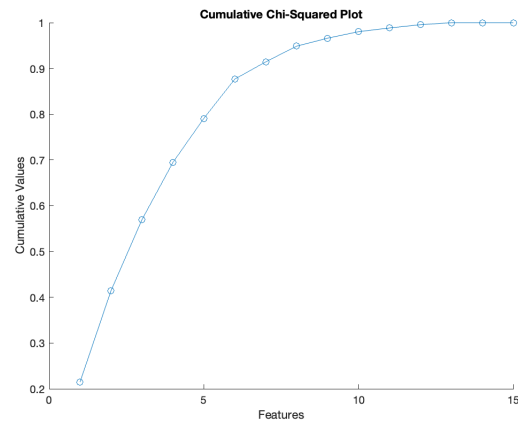


**Figura 8.** Box Plots dos atributos para as classes (0) NO CHD e (1) CHD

Como podemos observar nos plots, há claramente atributos mais relevantes que outros. Deste modo, é possível selecionar quantos dos atributos mais relevantes queremos utilizar para os modelos. Assim, faremos a análise dos classificadores utilizando as 6 features mais relevantes, quando recorrendo à seleção



**Figura 9.** Valores ordenados das estatísticas de teste



**Figura 10.** Valores normalizados cumulativos das estatísticas de teste

pelo teste de Kruskal-Wallis, dado que conseguimos obter cerca de 88% da soma cumulativa, como podemos observar pelo gráfico 10.

## 4.2 Redução de Features

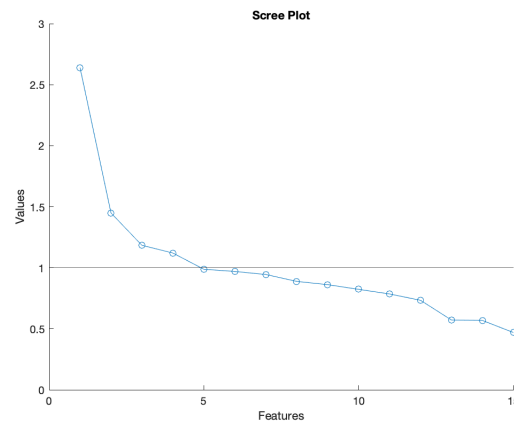
### 4.2.1 Principal Component Analysis

Nesta subsecção fazemos uma Principal Component Analysis dos dados com o objetivo de reduzir a dimensionalidade do nosso problema, enquanto procuramos manter sensivelmente a mesma informação que tínhamos originalmente.

Para o caso do PCA, há várias formas de escolhermos a nova dimensão para a qual queremos reduzir o dataset. De seguida apresentamos 3 métodos.

#### 4.2.1.1 Scree Test

Os valores próprios retornados pelo PCA indicam a importância que o respetivo vetor próprio tem na projeção dos dados. Podemos visualizar graficamente os valores próprios, ao que se deu o nome de Scree Plot. O Scree Test baseia-se no Scree plot e defende que apenas as componentes antes da estabilização dos valores próprios devem ser usadas. Neste caso, os valores próprios nunca estabilizam(ver figura 11), levando a que, segundo o Scree Test, todas as componentes sejam escolhidas, não levando a uma redução de dimensão do dataset.



**Figura 11.** Scree plot com kaiser criterion

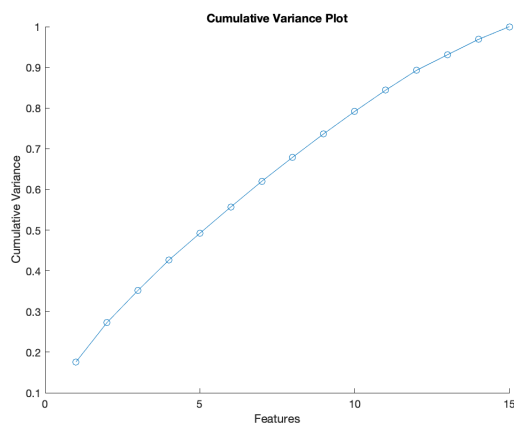
#### 4.2.1.2 Kaiser Criterion

O Kaiser Criterion, diz-nos que apenas devemos considerar as componentes cujos valores próprios tenham um valor de pelo menos 1. Segundo este método,

iríamos reduzir a dimensão do dataset de 15 para 4 (ver figura 11). As 4 componentes mais importantes agregariam cerca de 0.43 da variância cumulativa das 15 componentes, o que é pouco.

#### 4.2.1.3 Variância Cumulativa

Visto que ambos os métodos anteriores não apresentam os melhores resultados para este caso, decidimos utilizar a variância cumulativa como método de escolha da nova dimensão do dataset. Utilizando um *threshold* de 0.8, a nova dimensão do dataset será de 11, como pode ser visto na figura 12.



**Figura 12.** Variância Cumulativa

#### 4.2.2 Linear Discriminant Analysis

Analogamente ao PCA, iremos repetir a mesma ideia, de redução da dimensão dos dados, porém fazendo uma Linear Discriminant Analysis. Ao contrário do PCA, o LDA reduz sempre para  $C - 1$  features, onde  $C$  é o número de classes, ou seja  $2 - 1 = 1$ . Deste modo, aplicando o LDA aos dados, obtemos um único linear discriminant.

## 5 Classificadores

### 5.1 Minimum Distance Classifiers

#### 5.1.1 Euclidean Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.80	0.97	0.62	0.88
<i>Class CHD</i>	0.62	0.16	0.80	0.25

**Tabela 1.** KW + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.80	0.97	0.62	0.88
<i>Class CHD</i>	0.62	0.16	0.80	0.25

**Tabela 2.** PCA + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.68	0.99	0.83	0.80
<i>Class CHD</i>	0.83	0.13	0.68	0.23

**Tabela 3.** LDA + ELD test scores

#### 5.1.2 Mahalanobis Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.72	0.98	0.79	0.83
<i>Class CHD</i>	0.79	0.15	0.72	0.25

**Tabela 4.** KW + MLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.77	0.98	0.71	0.86
<i>Class CHD</i>	0.71	0.16	0.77	0.26

**Tabela 5.** PCA + MLD test scores



Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.68	0.99	0.83	0.80
<i>Class CHD</i>	0.83	0.13	0.68	0.23

**Tabela 6.** LDA + MLD test scores

## 5.2 Fisher's Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.79	0.98	0.68	0.87
<i>Class CHD</i>	0.68	0.16	0.79	0.27

**Tabela 7.** KW + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.81	0.97	0.64	0.88
<i>Class CHD</i>	0.64	0.16	0.81	0.26

**Tabela 8.** PCA + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.68	0.99	0.83	0.80
<i>Class CHD</i>	0.83	0.13	0.68	0.23

**Tabela 9.** LDA + FLD test scores

## 5.3 Bayes Classifier

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.92	0.96	0.37	0.94
<i>Class CHD</i>	0.37	0.21	0.92	0.27

**Tabela 10.** KW + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	0.94	0.96	0.30	0.95
<i>Class CHD</i>	0.30	0.22	0.94	0.25

**Tabela 11.** PCA + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.00	0.97
<i>Class CHD</i>	0.00	NaN	1.00	0.00

**Tabela 12.** LDA + BC test scores

## 5.4 K-Nearest Neighbors

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.01	0.97
<i>Class CHD</i>	0.01	0.50	1.00	0.02

**Tabela 13.** KW + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.01	0.97
<i>Class CHD</i>	0.01	0.53	1.00	0.01

**Tabela 14.** PCA + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.00	0.97
<i>Class CHD</i>	0.00	0.39	1.00	0.00

**Tabela 15.** LDA + KNN test scores

## 5.5 Support Vector Machines

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.01	0.97
<i>Class CHD</i>	0.01	0.55	1.00	0.02

**Tabela 16.** KW + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.01	0.97
<i>Class CHD</i>	0.01	0.37	1.00	0.02

**Tabela 17.** PCA + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class NO CHD</i>	1.00	0.94	0.00	0.97
<i>Class CHD</i>	0.00	NaN	1.00	0.00

**Tabela 18.** LDA + SVM test scores

## Conclusões

Refletindo sobre o problema que procuramos resolver, queremos que o classificador classifique bem os casos positivos e os negativos. No entanto, dada a complexidade do problema, a existência de falsos positivos e de falsos negativos é inevitável. Deste modo, salientamos desde já a importância que damos a cada uma destas falsas classificações: uma pessoa com uma doença que é classificada como não a tendo (falso negativo) é mais grave do que uma que não tem doença e é classificada como tendo (falso positivo). Assim, relativamente ao tradeoff precision / recall, um valor de recall elevado será preferido a um de precision elevado. Tendo em consideração que a classe positiva, no contexto desta análise, é a classe CHD, iremos focar-nos nos valores dessas mesmas linhas nas tabelas.

Relativamente às métricas selecionadas, a accuracy não nos irá fornecer informação muito valiosa dado que as classes não estão equilibradas. Assim, as métricas escolhidas foram o recall, para medir a significância de falsos negativos, a specificity, para quantificar os falsos positivos, a precision para avaliarmos a capacidade de previsão acertada da classe positiva e o f1 score, como uma métrica de média harmónica entre a precision e o recall.

Com base no *Recall* conseguimos ver que mais de metade das amostras de *CHD* estão a ser classificadas como tal pelos classificadores lineares, no entanto, olhando para a *Precision* conseguimos ver que estes estão a classificar muitas amostras de *NO CHD* como *CHD*. Há, portanto, um claro enviesamento devido à desproporção do número de elementos das duas classes, que se reflete nos modelos, o que, no que toca a falsos positivos e falsos negativos, é algo que podemos tolerar.

O *Bayesian classifier* quando usado em conjunto com KW ou PCA consegue obter melhores valores de *Precision* do que os classificadores lineares, mas não classifica como *CHD* cerca de 70% das amostras que pertencem a essa classe (pior que os classificadores lineares). Quando usado com o LDA, este limitou-se a classificar tudo como *NO CHD*, levando a que a *Precision* seja NaN na tabela 5.3. Assim sendo, esta última combinação não é minimamente boa.

A *SVM* não obteve bons resultados, visto que não conseguiu classificar quase nenhuma amostra da classe *CHD* como sendo pertencente a esta. No melhor dos testes, obteve um valor mediano de *Precision*, mas este apenas nos diz que nos poucos que considerou serem *CHD*, 55% pertenciam de facto à classe *CHD*.

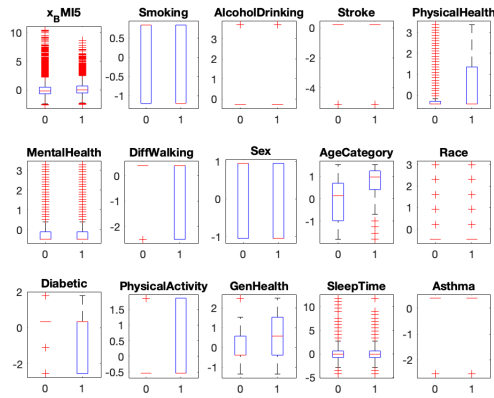
O *KNN* é bastante parecido à *SVM*, não conseguindo classificar quase nenhuma das amostras da classe *CHD*. Assim, admitimos que possa faltar uma melhor parametrização destes últimos dois modelos, com a finalidade de obter resultados mais satisfatórios.

## Cenário B

### 6 Redução e Seleção de Features

#### 6.1 Seleção de Features

##### 6.1.1 Kruskal-Wallis



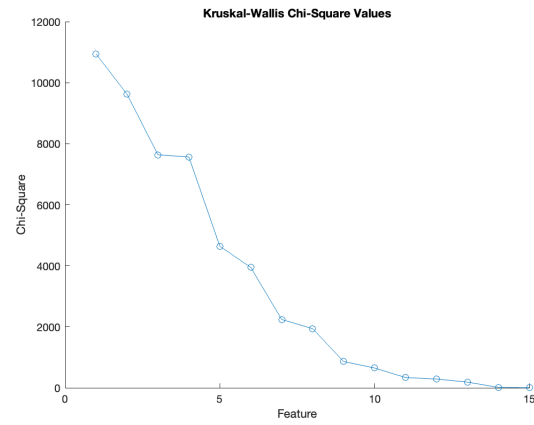
**Figura 13.** Box Plots dos atributos para as classes (0) NO HD e (1) HD

Tendo em conta os resultados, faremos a análise dos classificadores utilizando as 6 features mais relevantes, quando recorrendo à seleção pelo teste de Kruskal-Wallis, dado que conseguimos obter cerca de 87% da soma cumulativa, como podemos observar pelo gráfico 15.

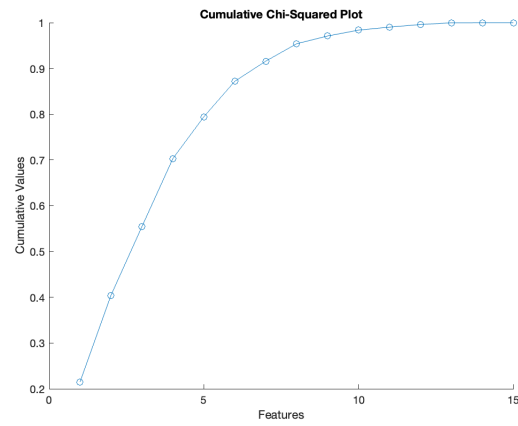
### 6.2 Redução de Features

#### 6.2.1 Principal Component Analysis

Dado que o PCA é um método unsupervised, a análise feita para o cenário A aplica-se também neste caso, dado que os dados são os mesmos, reduzindo a dimensionalidade dos dados para 11 componentes, de modo a superar o threshold de 80% de variância.



**Figura 14.** Valores ordenados das estatísticas de teste



**Figura 15.** Valores normalizados cumulativos das estatísticas de teste

### 6.2.2 Linear Discriminant Analysis

Dado que o LDA é um método unsupervised, a análise feita para o cenário A aplica-se também neste caso, dado que os dados são os mesmos, de modo que se irá reduzir a dimensionalidade dos dados para 1.

## 7 Classificadores

### 7.1 Minimum Distance Classifiers

#### 7.1.1 Euclidean Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.80	0.96	0.62	0.87
<i>Class HD</i>	0.62	0.22	0.80	0.32

**Tabela 19.** KW + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.80	0.96	0.63	0.88
<i>Class HD</i>	0.63	0.23	0.80	0.33

**Tabela 20.** PCA + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.68	0.98	0.83	0.80
<i>Class HD</i>	0.83	0.19	0.68	0.31

**Tabela 21.** LDA + ELD test scores

#### 7.1.2 Mahalanobis Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.73	0.97	0.76	0.83
<i>Class HD</i>	0.76	0.20	0.73	0.32

**Tabela 22.** KW + MLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.77	0.97	0.71	0.86
<i>Class HD</i>	0.71	0.22	0.77	0.33

**Tabela 23.** PCA + MLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.68	0.98	0.82	0.80
<i>Class HD</i>	0.82	0.19	0.68	0.31

**Tabela 24.** LDA + MLD test scores

## 7.2 Fisher's Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.80	0.96	0.66	0.87
<i>Class HD</i>	0.66	0.23	0.80	0.34

**Tabela 25.** KW + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.81	0.96	0.64	0.88
<i>Class HD</i>	0.64	0.24	0.81	0.35

**Tabela 26.** PCA + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.68	0.98	0.82	0.80
<i>Class HD</i>	0.82	0.19	0.68	0.30

**Tabela 27.** LDA + FLD test scores



### 7.3 Bayes Classifier

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.91	0.94	0.37	0.93
<i>Class HD</i>	0.37	0.27	0.91	0.32

**Tabela 28.** KW + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.93	0.94	0.31	0.94
<i>Class HD</i>	0.31	0.30	0.93	0.31

**Tabela 29.** PCA + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.00	0.96
<i>Class HD</i>	0.00	NaN	1.00	0.00

**Tabela 30.** LDA + BC test scores

### 7.4 K-Nearest Neighbors

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.04	0.96
<i>Class HD</i>	0.04	0.53	1.00	0.07

**Tabela 31.** KW + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.04	0.96
<i>Class HD</i>	0.04	0.54	1.00	0.07

**Tabela 32.** PCA + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.02	0.96
<i>Class HD</i>	0.02	0.49	1.00	0.04

**Tabela 33.** LDA + KNN test scores

## 7.5 Support Vector Machines

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.04	0.96
<i>Class HD</i>	0.04	0.52	1.00	0.08

**Tabela 34.** KW + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.03	0.96
<i>Class HD</i>	0.03	0.44	1.00	0.06

**Tabela 35.** PCA + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.00	0.96
<i>Class HD</i>	0.00	0.61	1.00	0.01

**Tabela 36.** LDA + SVM test scores

## Conclusões

Como no cenário anterior, a *SVM* e o *KNN* obtiveram valores muito baixos de *Recall*, obtendo um elevado número de falsos negativos, o que, na nossa visão é o que mais se deveria evitar neste contexto.

O *Bayesian classifier* obteve resultados melhores que os dois mencionados anteriormente, mas continua a obter uma grande quantidade de falsos negativos.

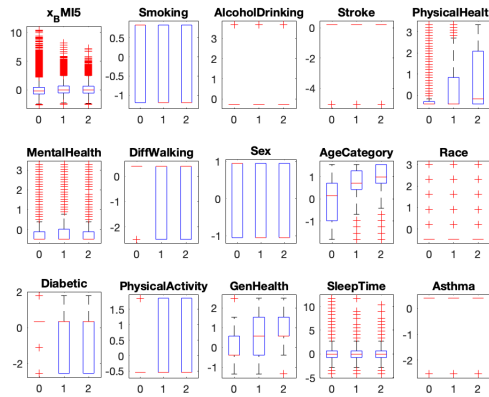
De facto os melhores classificadores que obtivemos foram novamente os lineares, conseguindo classificar como *HD*, nos piores casos, mais de 60% das amostras pertencentes a esta classe. A precisão dos classificadores para esta classe continua a ser bastante baixa, dizendo-nos que estamos a obter um elevado número de falsos positivos.

## Cenário C

### 8 Redução e Seleção de Features

#### 8.1 Seleção de Features

##### 8.1.1 Kruskal-Wallis



**Figura 16.** Box Plots dos atributos para as classes (0) NO HD, (1) HD e (2) HDC

Tendo em conta os resultados, faremos a análise dos classificadores utilizando as 6 features mais relevantes, quando recorrendo à seleção pelo teste de Kruskal-Wallis, dado que conseguimos obter cerca de 88% da soma cumulativa, como podemos observar pelo gráfico 18.

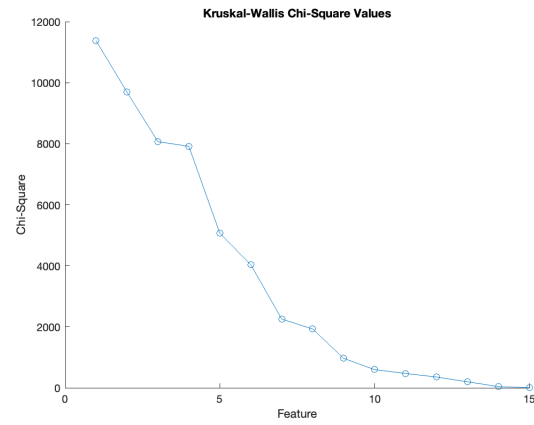
#### 8.2 Redução de Features

##### 8.2.1 Principal Component Analysis

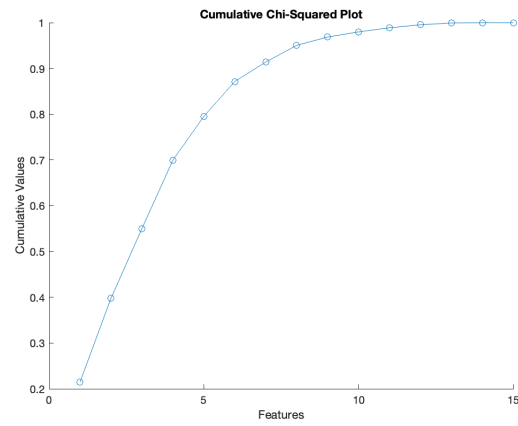
Dado que o PCA é um método unsupervised, a análise feita para o cenário A aplica-se também neste caso, dado que os dados são os mesmos.

##### 8.2.2 Linear Discriminant Analysis

Analogamente ao PCA, iremos repetir a mesma ideia, de redução da dimensão dos dados, porém fazendo uma Linear Discriminant Analysis. Ao contrário do PCA, o LDA reduz sempre para  $C - 1$  features, onde  $C$  é o número de classes,



**Figura 17.** Valores ordenados das estatísticas de teste



**Figura 18.** Valores normalizados cumulativos das estatísticas de teste

ou seja  $3 - 1 = 2$ . Deste modo, aplicando o LDA aos dados, obtemos dois linear discriminants.

## 9 Classificadores

### 9.1 Minimum Distance Classifiers

#### 9.1.1 Euclidean Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.79	0.96	0.63	0.87
<i>Class HD</i>	0.23	0.11	0.89	0.15
<i>Class HDC</i>	0.49	0.09	0.89	0.15

**Tabela 37.** KW + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.79	0.96	0.64	0.87
<i>Class HD</i>	0.23	0.11	0.88	0.15
<i>Class HDC</i>	0.51	0.10	0.89	0.16

**Tabela 38.** PCA + ELD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.65	0.98	0.84	0.78
<i>Class HD</i>	0.37	0.10	0.80	0.16
<i>Class HDC</i>	0.66	0.09	0.84	0.16

**Tabela 39.** LDA + ELD test scores

#### 9.1.2 Mahalanobis Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.68	0.98	0.81	0.80
<i>Class HD</i>	0.37	0.10	0.79	0.16
<i>Class HDC</i>	0.61	0.10	0.87	0.18

**Tabela 40.** KW + MLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.73	0.97	0.74	0.84
<i>Class HD</i>	0.33	0.11	0.84	0.17
<i>Class HDC</i>	0.56	0.10	0.88	0.17

**Tabela 41.** PCA + MLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.65	0.98	0.84	0.78
<i>Class HD</i>	0.37	0.10	0.79	0.16
<i>Class HDC</i>	0.65	0.09	0.84	0.16

**Tabela 42.** LDA + MLD test scores

## 9.2 Fisher's Linear Discriminant

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.79	0.96	0.67	0.87
<i>Class HD</i>	0.63	0.15	0.78	0.25
<i>Class HDC</i>	0.00	0.03	1.00	0.00

**Tabela 43.** KW + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.81	0.96	0.64	0.88
<i>Class HD</i>	0.59	0.16	0.80	0.25
<i>Class HDC</i>	0.00	0.01	1.00	0.00

**Tabela 44.** PCA + FLD test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.68	0.98	0.82	0.80
<i>Class HD</i>	0.77	0.13	0.66	0.22
<i>Class HDC</i>	0.00	0.01	1.00	0.00

**Tabela 45.** LDA + FLD test scores

### 9.3 Bayes Classifier

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.92	0.94	0.37	0.93
<i>Class HD</i>	0.28	0.17	0.92	0.21
<i>Class HDC</i>	0.10	0.20	0.99	0.14

**Tabela 46.** KW + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	0.95	0.94	0.29	0.94
<i>Class HD</i>	0.16	0.20	0.96	0.18
<i>Class HDC</i>	0.18	0.16	0.98	0.17

**Tabela 47.** PCA + BC test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.00	0.96
<i>Class HD</i>	0.00	NaN	1.00	0.00
<i>Class HDC</i>	0.00	NaN	1.00	0.00

**Tabela 48.** LDA + BC test scores

### 9.4 K-Nearest Neighbors

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.01	0.96
<i>Class HD</i>	0.01	0.37	1.00	0.01
<i>Class HDC</i>	0.01	0.35	1.00	0.02

**Tabela 49.** KW + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.01	0.96
<i>Class HD</i>	0.01	0.31	1.00	0.02
<i>Class HDC</i>	0.00	0.43	1.00	0.00

**Tabela 50.** PCA + KNN test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.00	0.96
<i>Class HD</i>	0.00	0.09	1.00	0.00
<i>Class HDC</i>	0.00	0.20	1.00	0.00

**Tabela 51.** LDA + KNN test scores

## 9.5 Support Vector Machines

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.01	0.96
<i>Class HD</i>	0.00	0.29	1.00	0.01
<i>Class HDC</i>	0.00	0.13	1.00	0.01

**Tabela 52.** KW + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.01	0.96
<i>Class HD</i>	0.01	0.33	1.00	0.02
<i>Class HDC</i>	0.01	0.27	1.00	0.01

**Tabela 53.** PCA + SVM test scores

Metrics	Recall	Precision	Specificity	f1
<i>Class No HD</i>	1.00	0.92	0.00	0.96
<i>Class HD</i>	0.00	0.29	1.00	0.00
<i>Class HDC</i>	0.00	0.18	1.00	0.00

**Tabela 54.** LDA + SVM test scores



## Conclusões

Analogamente aos cenários anteriores, existe um claro desequilíbrio no número de labels para cada classe, sendo que a *No HD* domina este espaço. Assim, observamos novamente uma tendência, em todos os classificadores, de classificarem mais esta mesma classe em deterioramento das outras. Relativamente às *SVMs* e *KNNs*, estas optam claramente por classificar basicamente todos como *No HD*. O *bayesian classifier* apresenta resultados ligeiramente melhores, apesar de ainda insatisfatórios, especialmente quando nos focamos no *Recall* das classes *HD* e *HDC*. Ao contrário dos cenários anteriores, o *FLD* apresenta resultados não satisfatórios no que toca à classe *HDC*. Acerca dos restantes classificadores lineares, estes apresentam os melhores resultados de todos os modelos para este último cenário. Não são precisos, mas são os que conseguem classificar mais amostras das classes *HD* e *HDC* como tal.

## Notas Finais

Claramente não conseguimos obter bons resultados. Uma melhor parametrização dos classificadores *KNN* e *SVM* seria necessária. As classes nos diferentes cenários são muito desequilibradas em termos de número de amostras, o que, a nosso ver, compromete bastante o desempenho dos vários classificadores.