1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Reconhecimento de Padrões

First Project Milestone

Heart Disease Prediction (Scenario A)

2021/2022

Mestrado em Engenharia Informática

PL2	Gabriel Fernandes	2018288117	gabrielf@student.dei.uc.pt
PL2	Miguel Rabuge	2018293728	rabuge@student.dei.uc.pt
PL2	Pedro Rodrigues	2018283166	pedror@student.dei.uc.pt

Índice

1. Data Preprocessing

Neste capítulo detalhamos como foi feito o pré-processamento dos dados. Abordamos os temas do carregamento, *scaling*, visualização dos dados através de uma *scatter matrix* e *box plots* das *features*. Encontra-se também neste capítulo uma secção sobre redução e seleção de features, recorrendo à *Principal Component Analysis* e à *Linear Discriminant Analysis*. Pode observar-se também as features de um ponto de vista de redundância e relevância, através de uma matriz de correlação, de discriminância, através do teste não paramétrico *Kruskal-Wallis*, bem como as suas distribuições, analisando a normalidade através do teste de Kolmogorov-Smirnov.

1.1. Carregamento

Os dados foram carregados através do ficheiro CSV fornecido "*heart_2020_cleaned"*, sendo divididos em dois *sets*: 200.000 *entries* para treino e o restante para validação, como referido no enunciado. As features são armazenadas na variável X da estrutura, enquanto que os labels são guardados na y, de acordo com o padrão das estruturas do **stprtool**.

1.2. Scaling

Depois de carregados os dados, é necessário normalizá-los devido à sensibilidade que o PCA tem sobre a variância. 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.

1.3. Visualização dos Dados

Nesta secção apresentamos visualizações sobre os dados originais, bem como o nosso racional.

1.3.1. Scatter Matrix

Com o objetivo de visualizar os dados, produzimos uma *scatter matrix*, recorrendo ao módulo **pandas** do Python, pandas.plotting.scatter_matrix:

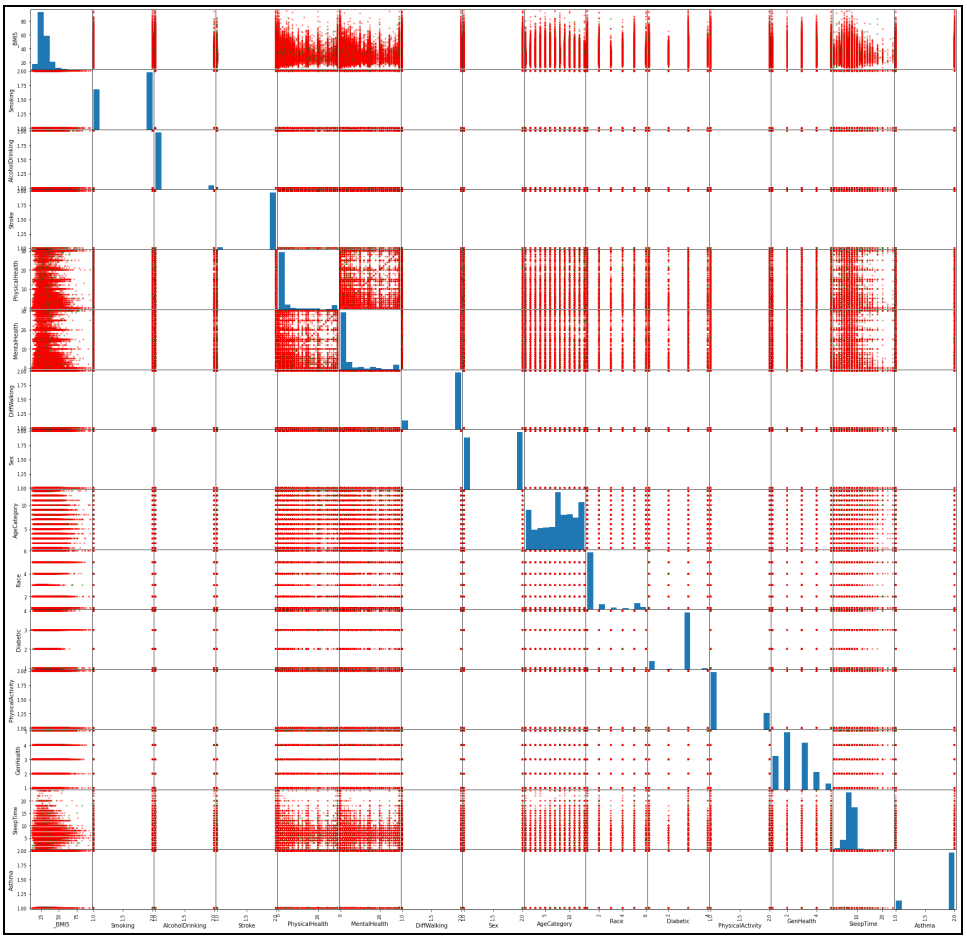

Figura x - Scatter matrix CHD

Através desta matriz de plots, apesar da baixa resolução devido à combinatória, podemos observar que não se destaca nenhuma combinação de features a duas dimensões que discrimine bem as duas classes: Coronary Heart Disease (CHD) a vermelho, e sem Coronary Heart Disease (no-CHD) a verde. Na diagonal, encontram-se os histogramas de cada feature onde podemos observar as distribuições e sensivelmente identificar quais são as variáveis categóricas.

1.3.2. Box Plots

Com a finalidade de alcançar um ponto de vista mais detalhado sobre cada feature, recorremos à utilização de *box plots* para a visualização de cada uma das features.

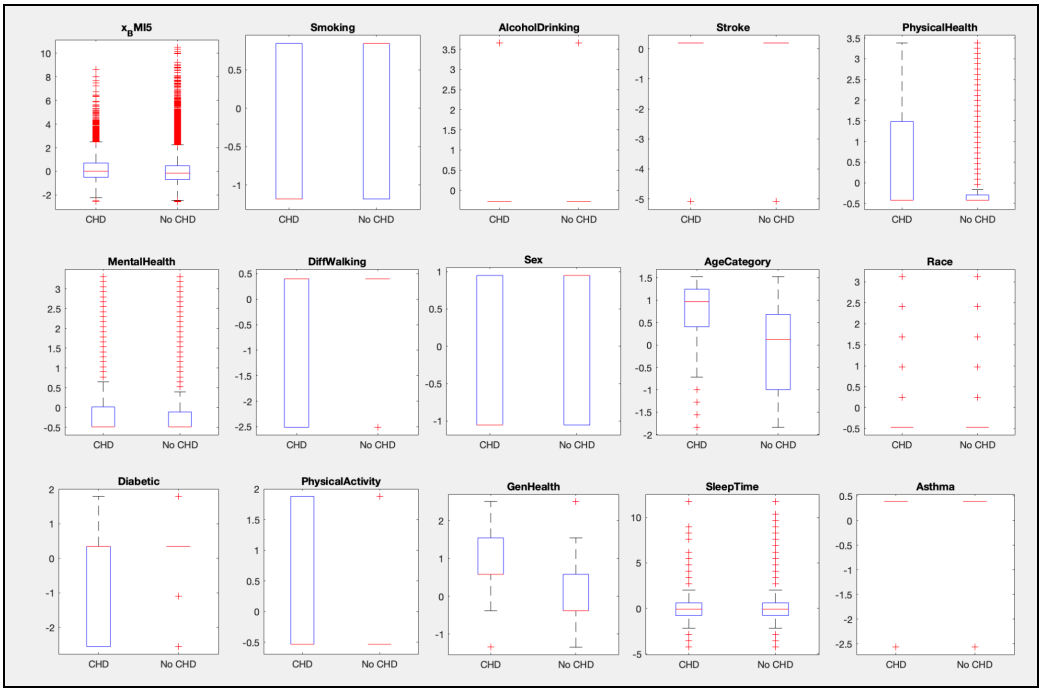

Figura X - *Scaled Features Box Plots*

Como podemos observar na figura acima, existem algumas features que revelam alguma discriminância sobre Coronary Heart Disease, como a *AgeCategory* e *GenHealth*, utilizando apenas uma mera análise visual. Iremos proceder a uma análise mais profunda para estas features, de modo a quantificar a discriminância das mesmas.

1.4. Redução e Seleção de Features

Nesta secção explicamos o procedimento efetuado na escolha das features mais discriminantes para os nossos classificadores.

1.4.1. Análise de Redundância e Relevância Linear

Com vista a inspecionar a redundância e relevância das nossas features, recorremos a uma matriz de correlação (pearson) para esse efeito. Abaixo, consta essa matriz onde cada célula corresponde à correlação de Pearson entre cada feature:

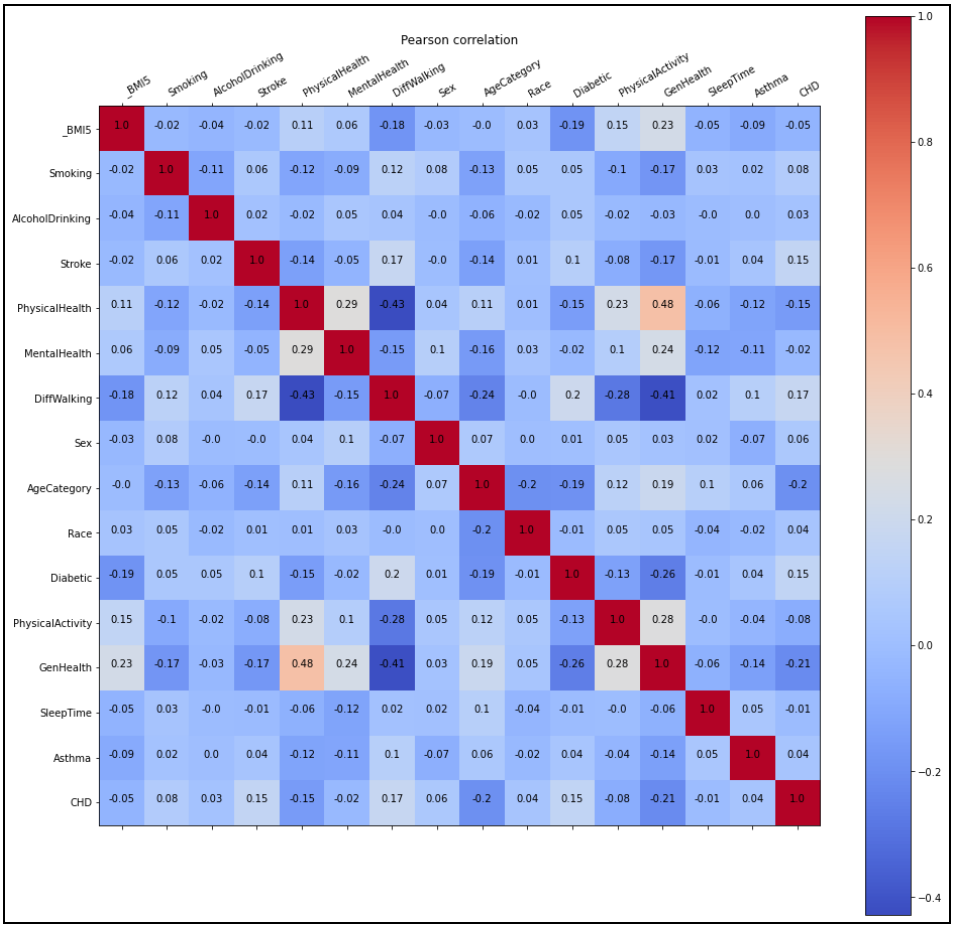

Figura X - Matriz de Correlação de Pearson - Features + Label

Importa notar que a última coluna/linha não é uma feature, mas sim o label booleano CHD (Coronary Heart Disease) cujas features estamos a investigar para produzirmos classificadores robustos.

Podemos observar que determinadas features apresentam alguma correlação com outras, nomeadamente *BMI*, *Stroke*, *PhysicalHealth*, *MentalHealth*, *DiffWalking*, *AgeCategory*, *Diabetic*, *PhysicalActivity*, e *GenHealth*, o que revela redundância destas features com as que estas se correlacionam mais fortemente. Por outro lado, é possível observar também que as features *Stroke*, *PhysicalHealth*, *DiffWalking*, *AgeCategory*, *Diabetic* e *GenHealth* se correlacionam bem com o label, o que sustenta a relevância destas features para o problema em mãos.

1.4.2. Análise de Discriminância - *Kruskal-Wallis*

Com a finalidade de inspecionar e quantificar as features sobre a sua discriminância para as classes “CHD” (Coronary Heart Disease) e “No CHD” (No Coronary Heart Disease), aplicamos o teste não paramétrico

Kruskal-Wallis e ordenamos as features pelo valor do qui-quadrado. Abaixo estão apresentados os resultados:

Feature	Qui-quadrado
Menos Relevante	
MentalHealth	0.28
SleepTime	14.54
AlcoholDrinking	155.32
Asthma	240.65
Race	406.54
x_BMI5	500.73
Sex	687.29
PhysicalActivity	1255.28
Smoking	1375.21
PhysicalHealth	3203.73
Diabetic	3833.52
Stroke	4732.90
DiffWalking	5790.09
GenHealth	7518.71
AgeCategory	8292.66
Mais Relevante	

Tabela 3 - Resultados Kruskal-Wallis

Como podemos observar, as features que identificamos informalmente nos passos anteriores, recorrendo à matriz de correlação e aos *box plots*, encontram-se entre as mais discriminantes segundo este teste, corroborando assim as nossas hipóteses.

1.4.3. 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.

1.4.3.1. Scree Plot

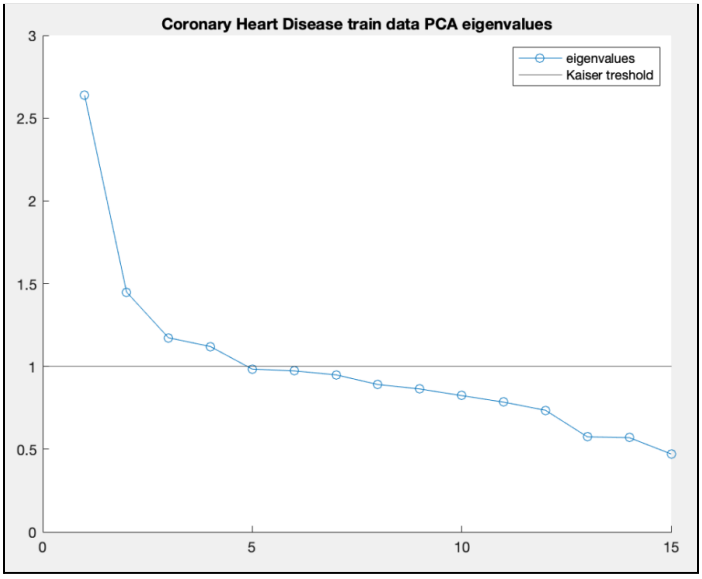

Figura X - *Scree Plot* PCA

1.4.3.2. *Kaiser Criterion* & *Scree Test*

Com base no *plot* acima, podemos observar que:

- Há 4 *principal components* cujos valores próprios são maiores do que 1
- Não existe uma clara estabilização dos valores próprios

Deste modo:

- Seguindo o *Kaiser Criterion* deveremos escolher as 4 primeiras *principal components* (43% da variância).
- Seguindo o Scree test, deveremos escolher todas as *principal components*, dado que o plot não estabiliza (100% da variância).

1.4.3.3. Variância

Principal Component	Valores Próprios	Variância	
		Individual	Cumulativa
1	2.64	0.18	0.18

2	1.45	0.10	0.27
3	1.17	0.08	0.35
4	1.12	0.07	0.43
5	0.98	0.07	0.49
6	0.97	0.06	0.56
7	0.95	0.06	0.62
8	0.89	0.06	0.68
9	0.86	0.06	0.74
10	0.82	0.05	0.79
11	0.78	0.05	0.84
12	0.74	0.05	0.89
13	0.57	0.04	0.93
14	0.57	0.04	0.97
15	0.47	0.03	1.00

Tabela 1 - *Principal Components*

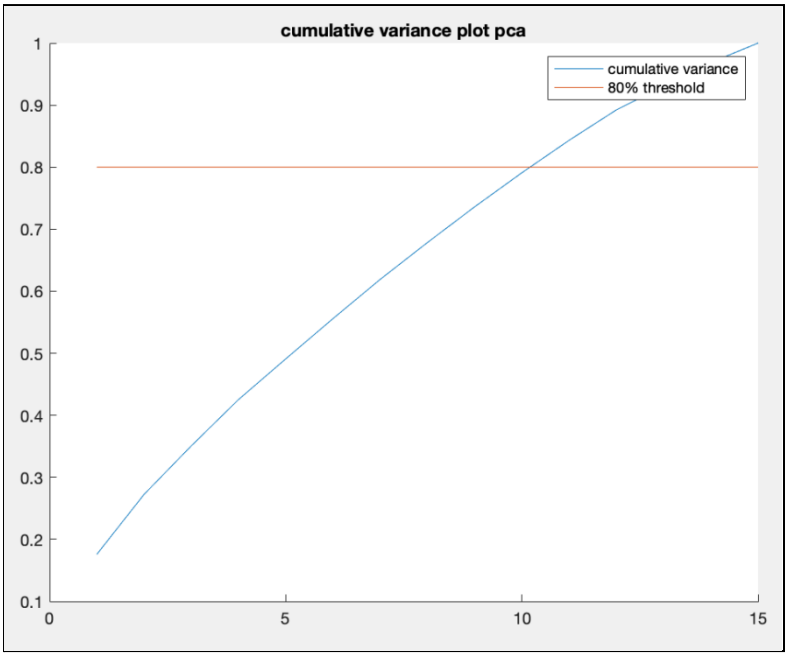

Figura X - Variância cumulativa PCA

Dado que as abordagens discutidas acima não funcionam muito bem (*Kaiser Criterion* e *Scree Test*), decidimos calcular a variância cumulativa para nos ajudar na escolha da nova dimensão a usar no PCA. Utilizando um *threshold* de 0.8 (80%), temos de utilizar uma dimensão de 11 para assegurar que ficamos com uma variância superior a este.
Com base neste estudo, optamos por reduzir a dimensão do dataset para 11 utilizando o PCA.

1.4.4. Linear Discriminant Analysis

Analogamente à secção anterior, iremos repetir a mesma ideia, 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* cujo valor próprio é 11303,36.

1.5. Feature Assessment

Nesta secção iremos analisar as distribuições das features do ponto de vista da sua normalidade. Para tal, numa primeira parte iremos apresentar as suas distribuições, e no final a comparação das funções de repartição destas com a função de repartição normal padrão. Este procedimento será efetuado para a totalidade dos dados, e para cada uma das duas possíveis labels.

1.5.1. Distribuições

Analisando as distribuições, apresentamos abaixo os *plots*.

1.5.1.1. Total

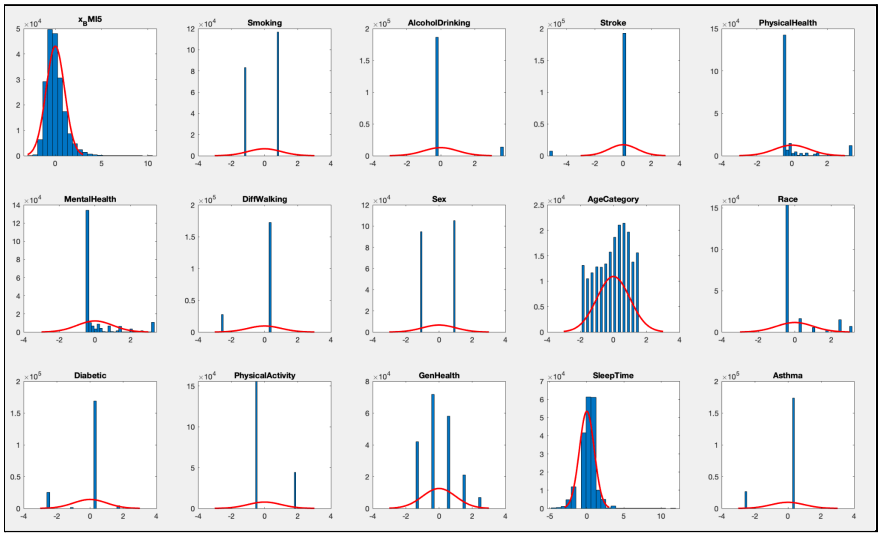

Figura X - Distribuições das features com distribuição normal padrão

1.5.1.2. Coronary Heart Disease (CHD)

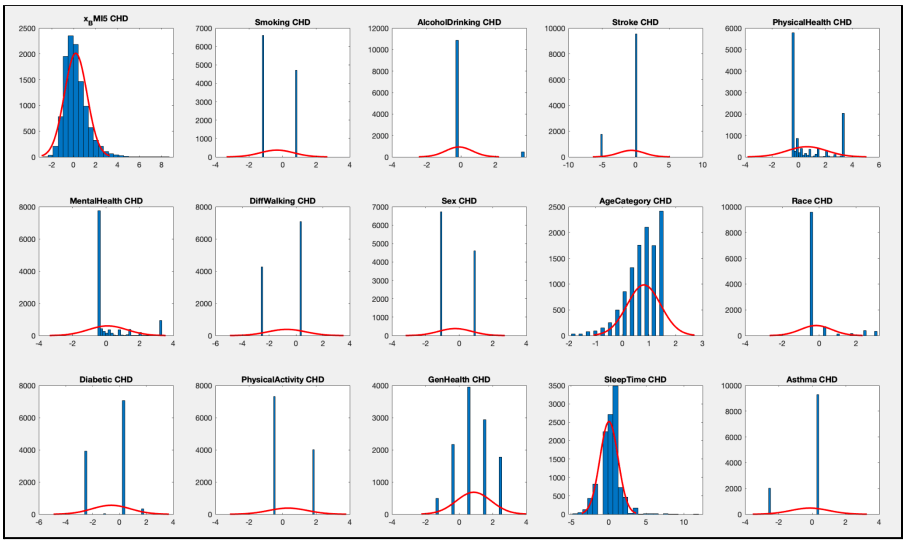

Figura X - Distribuições das features com distribuição normal padrão

1.5.1.3. No Coronary Heart Disease (No CHD)

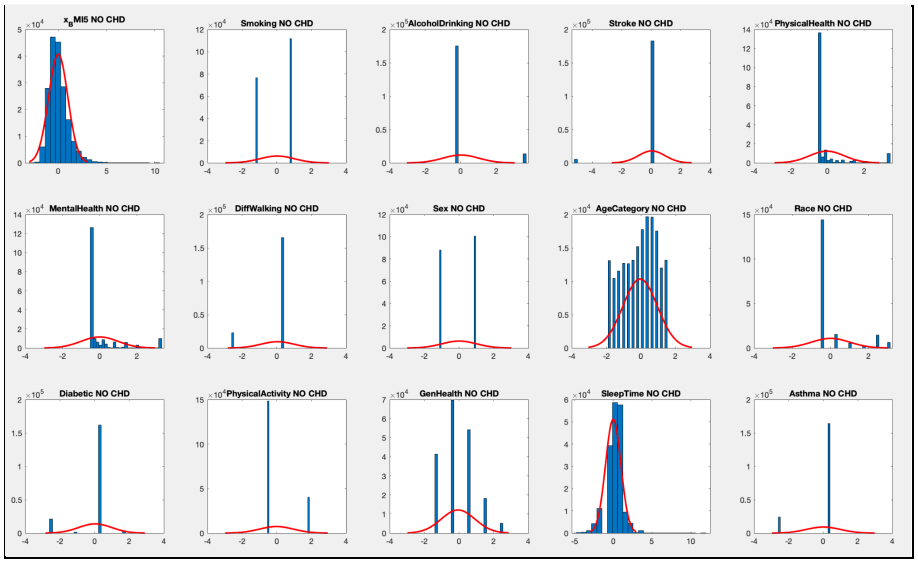

Figura X - Distribuições das features com distribuição normal padrão

Podemos observar que os histogramas diferem da distribuição normal padrão, exceto o *BMI* e o *SleepTime*. Vejamos melhor através da comparação das funções de repartição abaixo.

1.5.2. Função de Repartição Empírica Vs Normal Padrão

Agora, apresentamos a comparação entre as funções de repartição.

1.5.2.1. Total

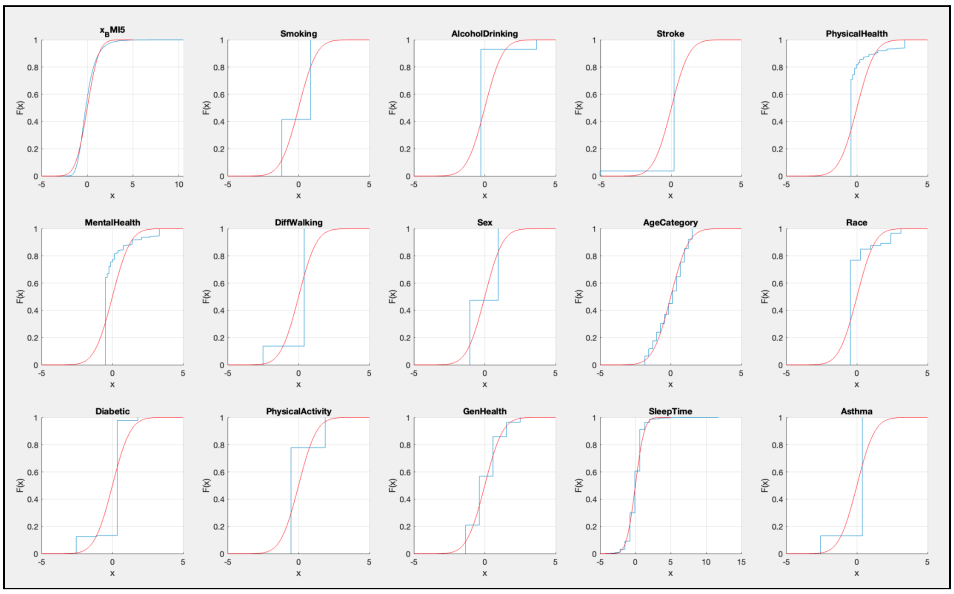

Figura X - Função de Repartição Empírica Vs Normal Padrão

1.5.2.2. Coronary Heart Disease (CHD)

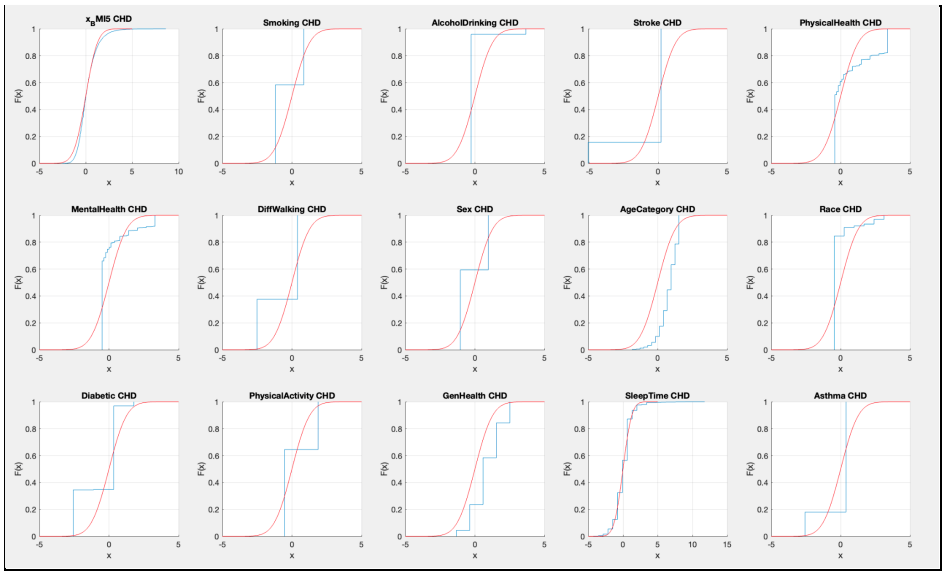

Figura X - Função de Repartição Empírica Vs Normal Padrão

1.5.2.3. No Coronary Heart Disease (No CHD)

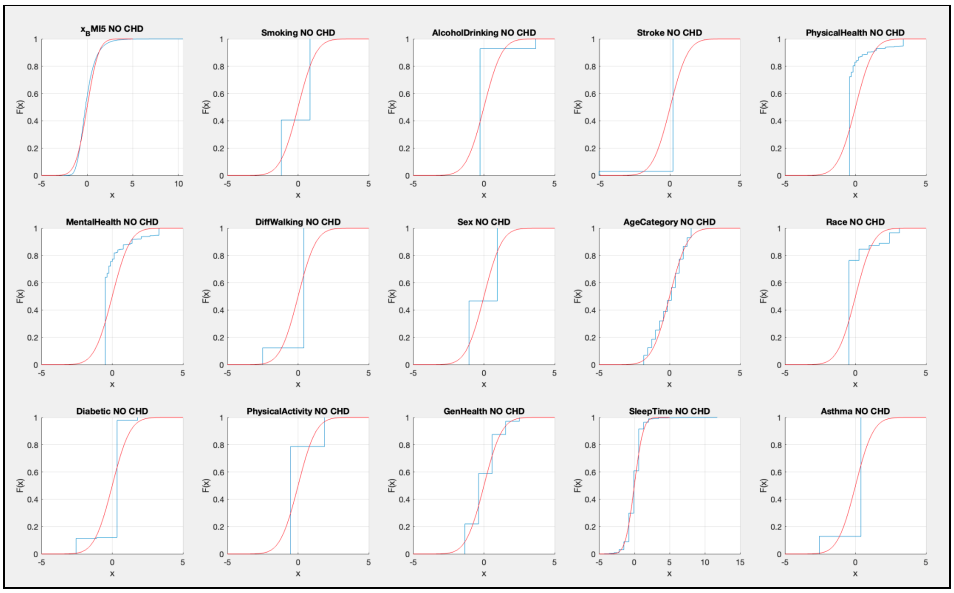

Figura X - Função de Repartição Empírica Vs Normal Padrão

Podemos observar que as comparações diferem, exceto o *BMI* e o *SleepTime*. As *features* *AgeCategory* e *GenHealth* aproximam-se também da função de repartição da distribuição normal no caso de No CHD. Deste modo, surge a necessidade de quantificar a normalidade destas distribuições. Para tal, iremos recorrer ao teste não paramétrico Kolmogorov-Smirnov.

1.5.3. Análise de Normalidade - Kolmogorov-Smirnov

Aplicando o teste KS, através da função kstest do Matlab, para o nível de confiança de 95%, às várias *features*, obtivemos os seguintes resultados:

Feature	P-Valor Total	P-Valor CHD	P-Valor No CHD
BMI	0	0	0
Smoking	0	0	0
AlcoholDrinking	0	0	0
Stroke	0	0	0
PhysicalHealth	0	0	0
MentalHealth	0	0	0
DiffWalking	0	0	0

Sex	0	0	0
AgeCategory	0	0	0
Race	0	0	0
Diabetic	0	0	0
PhysicalActivity	0	0	0
GenHealth	0	0	0
SleepTime	0	0	0
Asthma	0	0	0

Deste modo, para um nível de confiança de 95%, rejeitamos sempre a hipótese nula, o que nos diz que os dados não seguem a distribuição normal padrão.

2. Classificadores

Neste capítulo apresentamos os resultados obtidos ao aplicar os seguintes classificadores aos dados pré-processados. Estes serão analisados através da média das métricas obtidas ao aplicar k-fold Cross-Validation a estes classificadores. Estas métricas incluem a *accuracy*, *sensitivity* e *specificity*. No final, é indicado também o resultado final dos modelos, treinando com a totalidade do *set* de treino, para o *set* de teste.

2.1. Minimum Distance Classifiers

Nesta secção abordamos o ELD e o MLD enquanto classificadores de CHD.

2.1.1. *Euclidean Linear Discriminant*

Para o ELD, aplicamos o PCA e o LDA como técnicas de redução de *features* antes destas serem classificadas, tal que:

2.1.1.1. PCA + ELD

	Accuracy	Sensitivity	Specificity
Avg 10 Folds	0.79	0.63	0.80

Teste	0.79	0.63	0.80

2.1.1.2. LDA + ELD

	Accuracy	Sensitivity	Specificity
Avg 10 Folds	0.74	0.80	0.73
Teste	0.74	0.80	0.73

2.1.2. *Mahalanobis Linear Discriminant*

De modo análogo ao ELD, aplicamos o PCA e o LDA como técnicas de redução de *features* antes destas serem classificadas, de modo que:

2.1.2.1. PCA + MLD

	Accuracy	Sensitivity	Specificity
Avg 10 Folds	0.77	0.71	0.77
Teste	0.77	0.72	0.77

2.1.2.2. LDA + MLD

	Accuracy	Sensitivity	Specificity
Avg 10 Folds	0.74	0.80	0.73
Teste	0.74	0.80	0.73

2.2. Fisher Linear Discriminant

Nesta secção abordamos o FLD enquando classificador de CHD.

2.2.1. PCA + FLD

	Accuracy	Sensitivity	Specificity

Avg 10 Folds	0.80	0.65	0.81
Teste	0.80	0.64	0.81

2.2.2. LDA + FLD

	Accuracy	Sensitivity	Specificity
Avg 10 Folds	0.74	0.80	0.73
Teste	0.74	0.80	0.73

3. Conclusões

- Tendo em conta a natureza da assimetria no número de *samples* por classe (18085 CHD, 300873 No CHD), a *accuracy* pouco diz sobre a performance do modelo.

- Considerando a média dos 10 *folds*, podemos observar que a *accuracy*, quando utilizado o LDA como técnica de redução e seleção de *features*, o sistema apresenta resultados inferiores do que com o PCA.

- Também nos *folds*, observamos que o LDA consegue muito mais *sensitivity* (TPR) do que o PCA. No caso da *specificity* (TNR), verifica-se o contrário. Encontramo-nos perante uma situação de *tradeoff*. Considerando que é mais relevante acertar naqueles que realmente têm CHD do que nos que não têm, os modelos que recorrem ao LDA enquanto técnica de redução e seleção de *features*, aparentam ser mais vantajosos.

- Comparando os resultados da média de 10 *folds* e teste, verificamos que os resultados são semelhantes. Deste modo, podemos afirmar que o modelos são robustos, tendo as métricas chave entre 70% a 80%.

4. Bibliografia

- Slides RP @ 2022
- de Sá, J. M. P. (2001). Pattern Recognition: Concepts, Methods and Applications (1st ed.). Springer.
- An illustrative introduction to Fisher's Linear Discriminant - Thalles' blog
- Factor analysis - Wikipedia
- Kaiser Rule - Displayr
- Discriminant Analysis Essentials in R - Articles - STHDA
- https://sthalles.github.io/fisher-linear-discriminant/
- Proportion of explained variance in PCA and LDA - Cross Validated
- Linear Discriminant Analysis
- sklearn.metrics.classification_report — scikit-learn 1.0.2 documentation
- Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are.