Universidade do Rio de Janeiro
Pós-Graduação em Modelagem Computacional
Aprendizagem de Máquina
Trabalho 1

Nome do Autor: Ítalo Rosa Gonçalves

Data: 15 de Outubro de 2025

Sumário

1 Descrição dos Dados

A base de dados utilizada neste trabalho é a ”Heart Failure Clinical Records”, que contém informações clínicas de 299 pacientes com insuficiência cardíaca. O conjunto de dados possui 13 atributos, sendo 12 características preditoras e uma variável alvo binária, `DEATH_EVENT`, que indica se o paciente faleceu (1) ou sobreviveu (0) durante o período de acompanhamento.

Uma análise inicial revelou que a base de dados é comportada, não contendo valores nulos ou amostras duplicadas. A variável alvo é desbalanceada, com 203 amostras da classe 0 e 96 da classe 1.

A Figura 1 apresenta os boxplots de todas as variáveis, permitindo uma análise da distribuição e a identificação de outliers. Nota-se a presença de outliers em atributos como `creatinine_phosphokinase` e `serum_creatinine`, o que reforça a importância da padronização dos dados para algoritmos sensíveis à escala.

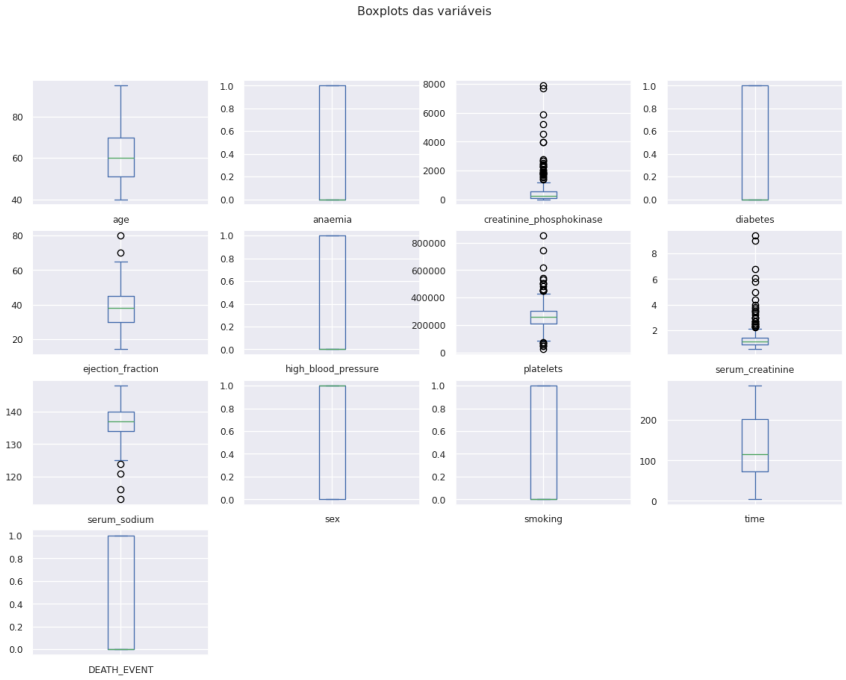

Figura 1: Boxplot das variáveis numéricas da base de dados.

Tabela 1: Resumo das estatísticas descritivas (Parte 1 de 2).

	age	anaemia	creatinine _phosphokinase	diabetes	ejection _fraction	high_blood _pressure
count	299.00	299.00	299.00	299.00	299.00	299.00
mean	60.83	0.43	581.84	0.42	38.08	0.35
std	11.89	0.50	970.29	0.49	11.83	0.48
min	40.00	0.00	23.00	0.00	14.00	0.00
25%	51.00	0.00	116.50	0.00	30.00	0.00
50%	60.00	0.00	250.00	0.00	38.00	0.00
75%	70.00	1.00	582.00	1.00	45.00	1.00
max	95.00	1.00	7861.00	1.00	80.00	1.00

Tabela 2: Resumo das estatísticas descritivas (Parte 2 de 2).

	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
count	299.00	299.00	299.00	299.00	299.00	299.00	299.00
mean	263358.03	1.39	136.63	0.65	0.32	130.26	0.32
std	97804.24	1.03	4.41	0.48	0.47	77.61	0.47
min	25100.00	0.50	113.00	0.00	0.00	4.00	0.00
25%	212500.00	0.90	134.00	0.00	0.00	73.00	0.00
50%	262000.00	1.10	137.00	1.00	0.00	115.00	0.00
75%	303500.00	1.40	140.00	1.00	1.00	203.00	1.00
max	850000.00	9.40	148.00	1.00	1.00	285.00	1.00

As Tabelas 1 e 2 apresentam um resumo estatístico detalhado de todos os atributos da base de dados. Geradas a partir do método describe() da biblioteca Pandas, elas consolidam informações essenciais como média, desvio padrão e quartis, fornecendo uma visão geral da distribuição e escala de cada variável.

Uma inspeção preliminar do conjunto de dados revelou que este não possui valores faltantes ou duplicados, e todas as suas variáveis já se encontram em formato numérico, o que simplifica as etapas de pré-processamento. A análise dos boxplots, apresentada anteriormente na Figura 1, complementa esta descrição ao evidenciar a presença de outliers em atributos como creatinine_phosphokinase e serum_creatinine.

2 Análise dos Dados

As principais medidas estatísticas dos atributos, como média, desvio padrão e quartis, estão resumidas na Tabela 1 e 2. Esta tabela fornece uma visão geral da escala e distribuição de cada variável. Assim, pode-se partir para a matriz de correlação que irá nos dizer como uma característica influencia na outra. Para ficar mais fácil a visualização, foi gerado a partir da matriz de correlação um heatmap.

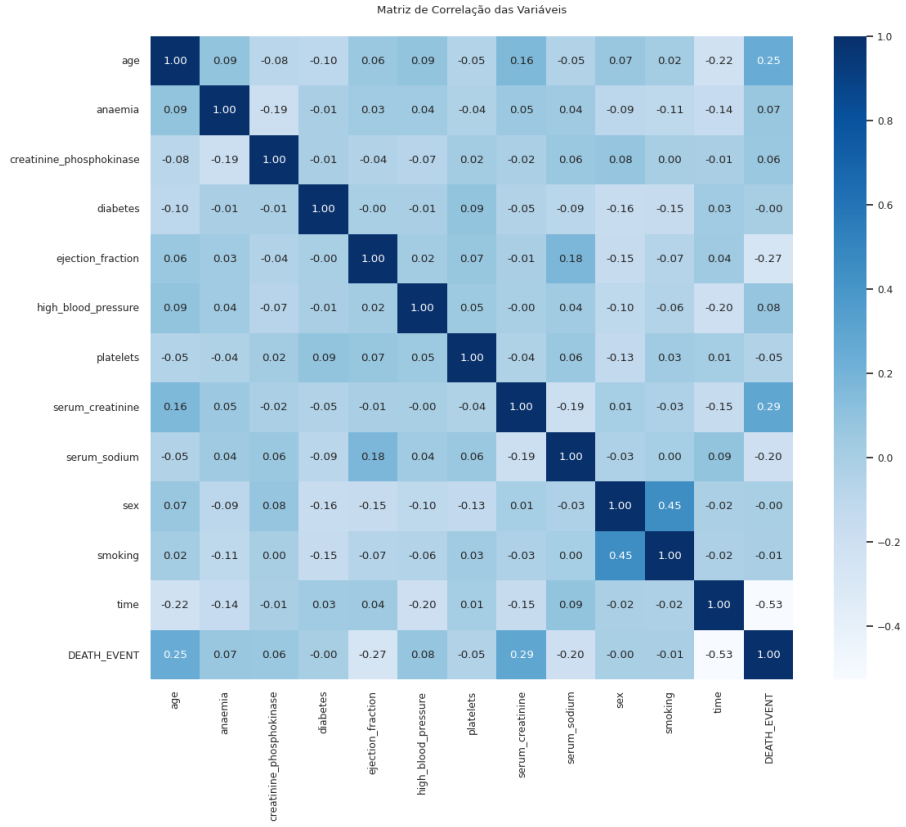

Figura 2: Matriz de correlação entre as variáveis.

A Figura 2 apresenta a matriz de correlação entre todas as variáveis do conjunto de dados. A diagonal principal da matriz possui valor 1, pois representa a correlação de cada característica consigo mesma.

Observa-se que, de modo geral, as correlações lineares entre os atributos são fracas. As correlações mais expressivas com a variável alvo, `DEATH_EVENT`, são as do tempo de acompanhamento (`time`), da creatinina sérica (`serum_creatinine`) e da fração de ejeção (`ejection_fraction`). Ainda assim, por serem correlações de magnitude moderada a fraca, a análise do heatmap sugere que o desfecho do paciente não é determinado por uma única variável isoladamente, mas sim por uma combinação mais complexa de fatores.

2.1 Etapa de pré-processamento

O pré-processamento dos dados foi fundamental para garantir que os algoritmos de agrupamento e validação pudessem funcionar de forma eficaz.Assim após a fase de analise dos dados e verificar a ausência de valores nulos ou categóricos, aplicamos a padronização utilizando o StandardScaler do scikit-learn.Esta etapa é fundamental, pois na base de dados há atributos como o `creatinine_phosphokinase` com valores com grande diferença entre seu mínimo e máximo contidos na tabela 1. A padronização transforma os dados de modo que cada atributo tenha média zero e desvio padrão unitário, permitindo que todos os atributos contribuam de forma equilibrada para o cálculo das distâncias entre as amostras.

3 Metodologia

A metodologia adotada para a analise do trabalho seguiu algumas etapas, primeiramente a verificação de valores faltantes, dados duplicados e a presença de dados no formatado caractere, seguindo pela fase de seleção de melhores característica e redução de dimensionalidade sempre que for necessário a visualização, e por fim, a etapa de clusterização usando o agglomerarive clustering e o parameter grid com o criterio de validação sendo o silhouette a fim de buscar os melhores clusteres.

3.1 Seleção de Características com o Select k-best

O heatmap da figura 2 nos mostra que não existe uma grande correlação, assim é inútil manter todas essas características e se faz necessário usar um método de seleção para identificar os atributos mais relevantes em relação a variável alvo `DEATH_EVENT` além de separar a variável alvo do resto dos atributos .

Tabela 3: Scores dos Atributos

Atributo	**Score**
age	20.44
anaemia	1.31
creatinine_phosphokinase	1.17
diabetes	0.00
ejection_fraction	23.09
high_blood_pressure	1.88
platelets	0.72
serum_creatinine	28.16
serum_sodium	11.77
sex	0.01
smoking	0.05
time	114.18

Resultando na seleção das seguintes características com base nos 6 melhores scores da tabela 3: time,`resultando na seleção das seguintes características: time, serum_creatinine, ejection_fraction, age, serum_sodium e creatinine_phosphokinase. Estas variáveis apresentaram maior poder discriminativo em relação ao evento de morte, confirmando os resultados observados na análise de correlação inicial`. Estas variáveis apresentaram maior poder discriminativo em relação ao evento de morte, confirmando os resultados observados na análise de correlação inicial

3.2 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) foi aplicada tanto no conjunto completo de atributos quanto no conjunto reduzido após a seleção. O PCA funciona fazendo combinações lineares das variáveis originais a fim de reduzir a dimensão dos dados. para essa análise, o PCA foi aplicado nos atributos originais, reduzindo de 12 colunas para 3 componentes principais na tabela 4 e no gráfico 3, e novamente foi aplicado no conjunto apos a seleção das características mais relevantes, indo de 6 colunas para 3 na tabela 5 e no gráfico 4, a fim de permitir a visualização em 3D.

- PCA na base de dados padrão:

Tabela 4: Variância nos PCAs

Variância explicada por cada componente:	
PC1:	13.86%
PC2:	13.16%
PC3:	10.57%
Total:	37.59%

- PCA na base com seleção:

Tabela 5: Variância nos PCAs

Variância explicada por cada componente	
PC1	24.73%
PC2	20.32%
PC3	16.89%
Total:	61.94%

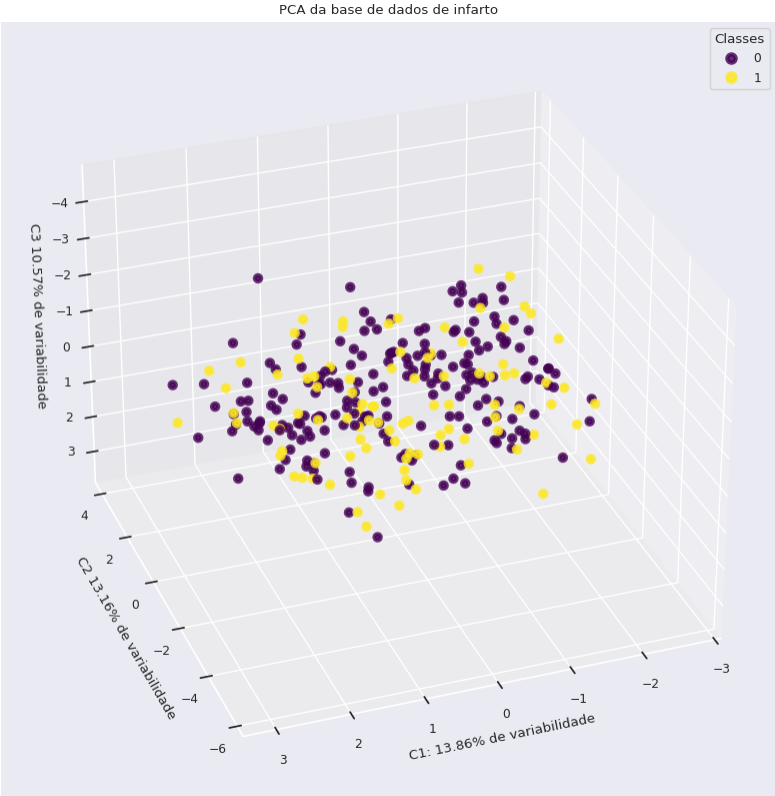

Figura 3: —Gráfico do PCA na base de dados padrão

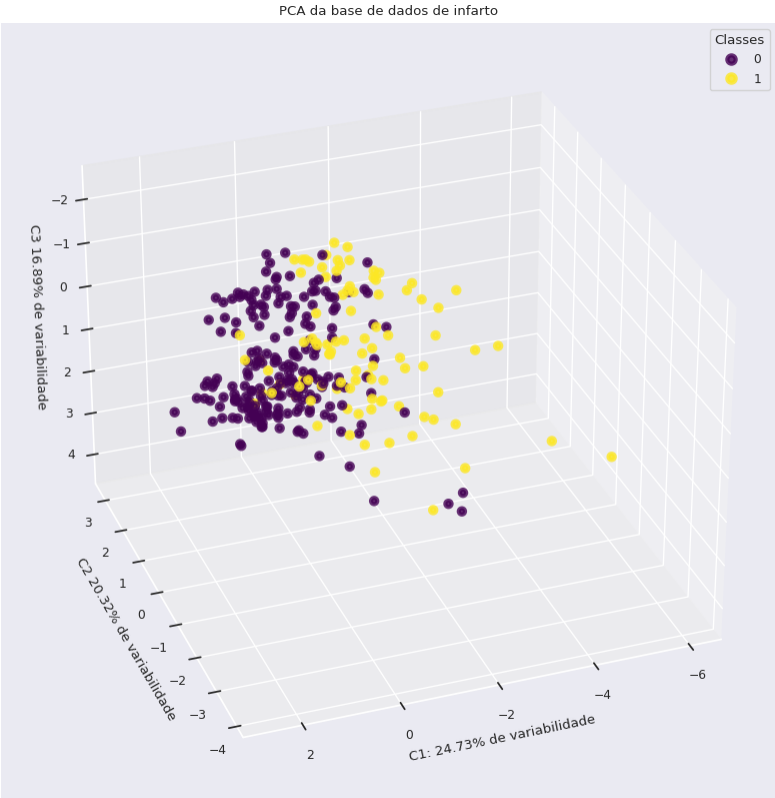

Figura 4: PCA da base de dados com seleção de característica

3.3 Método de Agrupamento

Para explorar a estrutura dos dados, foi aplicado o algoritmo de Agrupamento Hierárquico Aglomerativo. Este método constrói uma hierarquia de clusters de forma ”bottom-up”, onde, inicialmente, cada ponto de dado é seu próprio cluster. Em seguida, os clusters mais próximos são sucessivamente fundidos até que permaneça apenas o número desejado de clusters. O número de clusters foi definido como $n = 2$, alinhando-se à natureza binária da variável alvo.

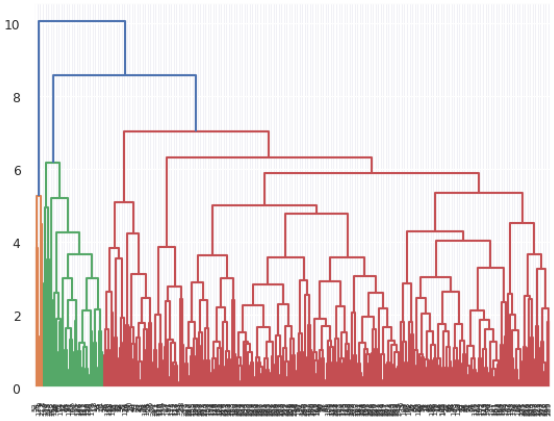

Figura 5: Dendrograma de formação dos clusters(Dados selecionados)

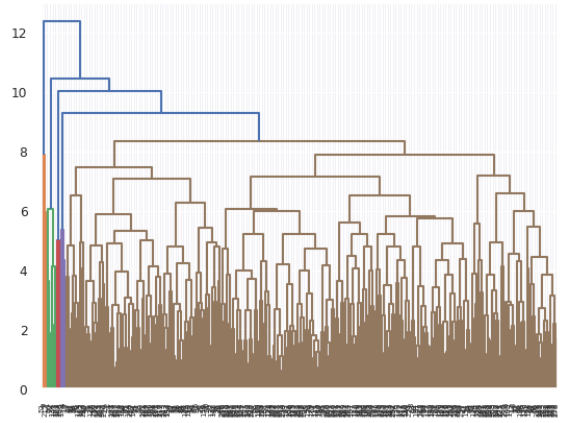

Figura 6: Dendrograma de formação dos clusters(Dados sem seleção)

A figura 5 e a figura 6 nós mostram como foi feita a formação dos clusters.Para melhor funcionamento do algoritmo de aglomeração, os dados utilizados sempre serão escalados usando o StandardScaler.

Avaliando ambas as figuras, podemos ver que os dados após a seleção das características ajudou na criação dos clusters, reduzindo o numero de junções feitas, mostrando como a seleção das melhores características foi essencial.

3.4 Critério de Validação

A validação do modelo de agrupamento foi realizada utilizando o **Coeficiente de Silhueta (Silhouette Score)**. Esta é uma métrica interna que avalia a qualidade da estrutura dos clusters sem a necessidade de rótulos verdadeiros. O objetivo foi encontrar a combinação de número de clusters (n) e método de linkage que maximizasse essa métrica. O Coeficiente de Silhueta mede quão similar um ponto é ao seu próprio cluster em comparação com os outros clusters, e seu valor varia de -1 a 1, onde valores mais altos indicam clusters mais densos e bem definidos.

4 Experimentos Computacionais

4.1 Resultados do Agrupamento

Usando o parameter grid para otimizar os melhores parâmetros para o método de agrupamento. Com o objetivo de encontrar a melhor combinação, foi realizada 20 combinações entre o número de clusters e 4 opções para o **linkage** sendo eles **'ward', 'complete', 'average', 'single'**.

Tabela 6: Resultados do Silhouette(Dados com seleção)

ID	Parâmetros	Silhouette Score
0	linkage='ward', n_clusters=2	0.150731
1	linkage='ward', n_clusters=3	0.174044
2	linkage='ward', n_clusters=4	0.183874
3	linkage='ward', n_clusters=5	0.194151
4	linkage='ward', n_clusters=6	0.149817
5	linkage='complete', n_clusters=2	0.538165
6	linkage='complete', n_clusters=3	0.139826
7	linkage='complete', n_clusters=4	0.128182
8	linkage='complete', n_clusters=5	0.093420
9	linkage='complete', n_clusters=6	0.077693
10	linkage='average', n_clusters=2	0.592976
11	linkage='average', n_clusters=3	0.397862
12	linkage='average', n_clusters=4	0.393962
13	linkage='average', n_clusters=5	0.334329
14	linkage='average', n_clusters=6	0.332698
15	linkage='single', n_clusters=2	0.614052
16	linkage='single', n_clusters=3	0.603218
17	linkage='single', n_clusters=4	0.512094
18	linkage='single', n_clusters=5	0.433127
19	linkage='single', n_clusters=6	0.413224

Primeiramente vamos mostrar como o parameter grid atuou na base de dados que teve suas características selecionadas pelo método **select K best** que pode ser visto na tabela 3. A tabela 6 mostra as 20 combinações agrupadas pelo **linkage** determinado, analisando a tabela 6 podemos extrair quais são os melhores parâmetros para esse agrupamento.

Tabela 7: Melhor Parâmetro Encontrado pela Análise do Silhouette Score.

Parâmetros	Silhouette Score
linkage='single', n_clusters=2	0.614052483059734

O valor de clusters $n = 2$ mostrado na 7 faz sentido, tendo em vista que a variavel alvo é `DEATH_EVENT`, tendo duas possibilidades, morto ou vivo.

Tabela 8: Resultados do Silhouette(dados sem seleção.

ID	Parâmetros	Silhouette Score
0	linkage='ward', n_clusters=2	0.088189
1	linkage='ward', n_clusters=3	0.098763
2	linkage='ward', n_clusters=4	0.105056
3	linkage='ward', n_clusters=5	0.091834
4	linkage='ward', n_clusters=6	0.087379
5	linkage='complete', n_clusters=2	0.468611
6	linkage='complete', n_clusters=3	0.353493
7	linkage='complete', n_clusters=4	0.329549
8	linkage='complete', n_clusters=5	0.274969
9	linkage='complete', n_clusters=6	0.069459
10	linkage='average', n_clusters=2	0.400827
11	linkage='average', n_clusters=3	0.349332
12	linkage='average', n_clusters=4	0.342607
13	linkage='average', n_clusters=5	0.252384
14	linkage='average', n_clusters=6	0.218627
15	linkage='single', n_clusters=2	0.416394
16	linkage='single', n_clusters=3	0.416581
17	linkage='single', n_clusters=4	0.415614
18	linkage='single', n_clusters=5	0.349754
19	linkage='single', n_clusters=6	0.275151

Para melhor comparação e verificar se a metodologia empregada estar funcionando, foi aplicado os métodos de agrupamento e parameter grid com o silhouette na base padrão sem a seleção de características, que os resultados podem ser vistos na tabela 8.

Tabela 9: Melhor Parâmetro Encontrado pela Análise do Silhouette Score.

Parâmetros	Silhouette Score
linkage='complete', n_clusters=2	0.46861144710828245

Novamente, a tabela 9 mostra a melhor combinação de parâmetros foi usando 2 clusters, a escolha correta para a base de dados.

4.2 Visualização do clusters

A figura 7 e 8 mostra como ficou a divisão entre os clusters de forma gráfica. Para a analise dos dados com a seleção de características a divisão ficou:

Tabela 10: Divisão dos clusters

Cluster	Contagem
0	298
1	1

A análise da tabela 13 e da figura 7 revela um resultado de agrupamento extremamente desbalanceado: 298 amostras foram alocadas ao cluster 0, enquanto apenas uma foi designada ao cluster 1. Este resultado sugere que o algoritmo isolou um outlier em vez de identificar um grupo com características distintas, tornando a clusterização ineficaz para segmentação. Para visualizar foi utilizado novamente o PCA, com a redução de dimensionalidade para 3, a tabela 11 mostra como cada característica influenciou nas componente principais.

A primeira componente principal sofreu maior influencia da caracterisitica `time` a segunda componente da caracteristica `serum_sodium` e a terceira `ejection_fraction`.

Tabela 11: Influência das Variáveis nos Componentes Principais (PC).

Variável	**PC1**	**PC2**	**PC3**	**Influência Total**
high_blood_pressure	0.285768	0.449440	0.565788	1.300995
ejection_fraction	0.112310	0.559708	0.576172	1.248190
serum_creatinine	0.472807	0.212897	0.464060	1.149765
age	0.490998	0.265238	0.280748	1.036984
time	0.562955	0.191808	0.231813	0.986576
serum_sodium	0.352390	0.576424	0.001032	0.929845

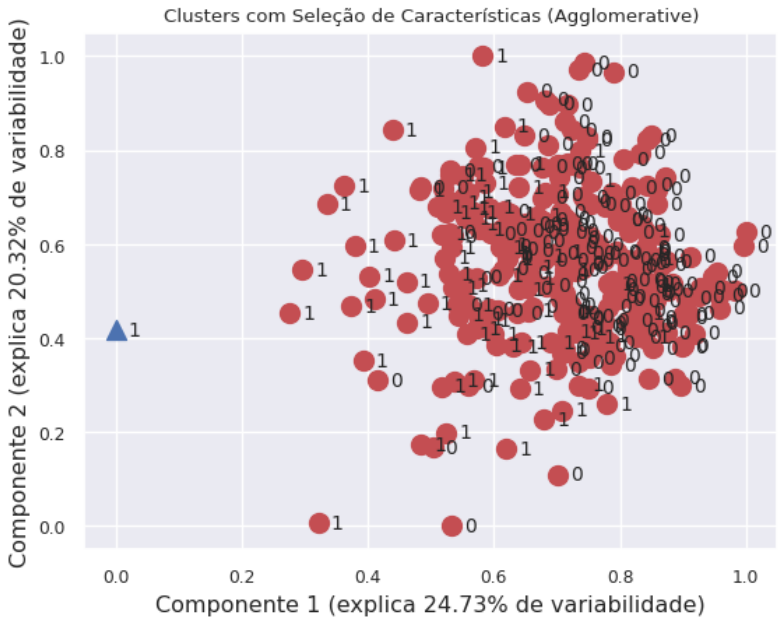

Figura 7: Clusters (com seleção de características)

Tabela 12: Distribuição de Eventos de Morte (DEATH_EVENT) por Cluster.

Evento de Morte	**Cluster 0**	**Cluster 1**
0 (Sobreviveu)	203	0
1 (Faleceu)	95	1

A tabela 12 nos mostra como os quem faleceu e sobreviveu ficou distribuído em cada cluster da base com seleção, informação importante para a analise e conclusão. o mesmo se repete na tabela 15, sendo a diferença é que ela se refere a base de dados padrão.

Tabela 13: Divisão dos clusters

Cluster	**Contagem**
0	296
1	3

O mesmo ocorre em relação os dados sem a seleção de características, são produzidos 3 outlier que podem ser considerados casos isolados e não devem ser considerados .

A tabela 14 mostra a influencia de cada característica para a base padrão, a primeira componente `sex`, a segunda tem-se novamente a caracteristica tempo `time`, e a terceira novamente a `serum_sodium`.

Tabela 14: Influência das Variáveis nos Componentes Principais (PC).

Variável	**PC1**	**PC2**	**PC3**	**Influência Total**
serum_sodium	0.093538	0.240871	0.625528	0.959937
diabetes	0.284697	0.151578	0.389240	0.825514
smoking	0.553870	0.008550	0.216302	0.778722
ejection_fraction	0.261812	0.046662	0.443315	0.751789
sex	0.622376	0.033969	0.092957	0.749303
age	0.049101	0.490932	0.182108	0.722141
serum_creatinine	0.037199	0.403600	0.269003	0.709802
anaemia	0.228183	0.320464	0.150436	0.699083
high_blood_pressure	0.190256	0.265189	0.229776	0.685221
time	0.001982	0.510580	0.041592	0.554154
creatinine_phosphokinase	0.162765	0.230922	0.083985	0.477671
platelets	0.170213	0.157016	0.124557	0.451787

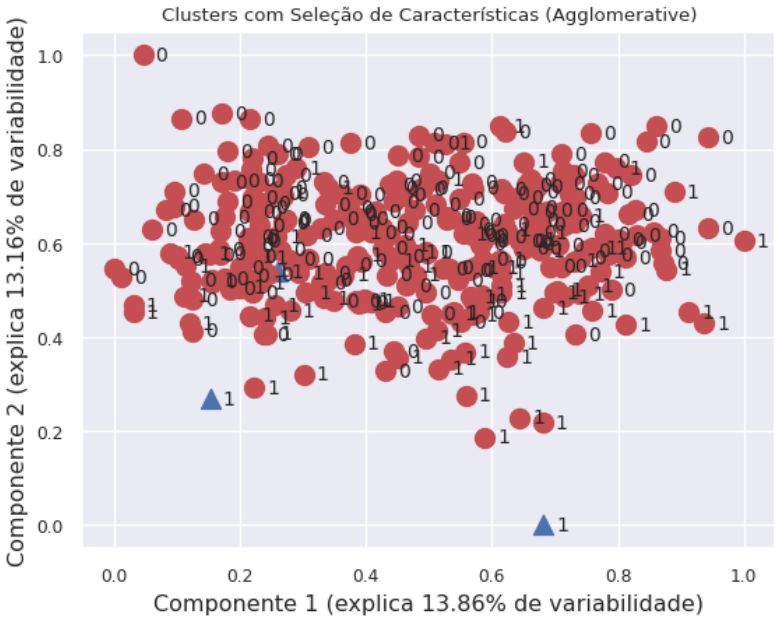

Figura 8: Clusters (base padrão)

Tabela 15: Distribuição de Eventos de Morte (DEATH_EVENT) por Cluster.

Evento de Morte	Cluster 0	Cluster 1
0 (Sobreviveu)	203	0
1 (Faleceu)	93	3

5 Conclusão

Nesta analise aplicamos vários métodos, seleção de característica (Select k Best) redução de dimensionalidade (PCA),além do agglomerative clustring sobre a base dade dados `Heart Failure Clinical Records`.A seleção de características reduziu o conjunto para seis atributos mais informativos (time, serum creatinine, ejection fraction, age, serum sodium, creatinine phosphokinase) e aumentou a variância explicada pelas três primeiras componentes principais de 37,59% para 61,94%, facilitando a visualização e interpretação dos dados.

Assim otimizar o agrupamento via parameter grid e usar o critério de validação do Silhouette tivemos as melhores combinações de parâmetros(tabelas 7 e 9), no entanto a clusterização produzida foi feita numa proporção desigual (tabelas 12 e 15), além da distribuição desigual interna do cluster 1, que nos mostra a ausência de uma separação consistente dos pacientes que morreram e sobreviveram, ou seja os métodos não produziu resultados satisfatórios para analises clinicas.

Para melhor analises é necessário testar outro métodos de agrupamento menos sensíveis a outliers ou mesmo partir para técnicas mais robusta de aprendizado supervisionado de classificação, regressão ou utilizar um rede neural.

Referências

[1] O QUE é a análise de componentes principais (PCA)?. *IBM*, [s.d.]. Disponível em: https://www.ibm.com/br-pt/think/topics/principal-component-analysis. Acesso em:

10 out. 2025.

[2] SCIKIT-LEARN. *scikit-learn: machine learning in Python*. [s.l.]: scikit-learn, [s.d.]. Disponível em: https://scikit-learn.org/stable/. Acesso em: 10 out. 2025.

[3] AMARAL, Fernando. *Inteligência Artificial e Machine Learning*. Udemy, [s.d.]. Disponível em: https://www.udemy.com/course/inteligencia-artificial-e-machine-learning/?couponCode=MT251015G4. Acesso em: 10 out. 2025.