Universidade dos Açores
Faculdade de Ciências e Tecnologias

2020/2021
1º Semestre
Informática

Laboratório de Ciência de Dados

RELATÓRIO FINAL

Pedro Sousa (2019101451@uac.pt)

Salif Faustino (20172005@uac.pt)

Resumo

Este trabalho foi realizado no âmbito da disciplina de Laboratório e ciências de dados e teve como objetivo a analise e exploração de dados. Este projeto pretende descrever de forma sucinta os resultados obtidos no processo de analise e exploração de dados sobre o momento pandémico que atravessamos concretamente em Brasil.

Índice

BREVE INTRODUÇÃO

Na cadeira de Laboratório e ciências de dados, do 3º Ano da Licenciatura em Informática, foi-nos proposto a interpretação e implementação de algoritmos aprendidos durante o semestre, utilizando dados escolhidos por nós. Os dados escolhidos nesse trabalho correspondem aos registos acerca do vírus cov-19 no ano de 2020, entre os meses de março e junho. O objetivo desse trabalho é através dos algoritmos estudados durante o semestre recolher informação que possa ser útil e identificar padrões. Com base nesse objetivo iremos explorar e interpretar todos os resultados obtidos.

A versão final do projeto encontra-se disponível no github, através do seguinte link:
https://github.com/SalifNTC/ Projeto-analise-de-dados

SUMÁRIO EXECUTIVO

Com a situação pandémica que o mundo atravessa, os alunos resolveram fazer uma investigação e através dela recolher dados sobre o número de infeções pelo novo coronavírus no brasil no ano de 2020, entre os meses de março e junho.

Cada registo contém a informação sobre o **ID, CONTADOR, DATA NOTIFICAÇÃO CLASSIFICAÇÃO, SEXO, IDADE, FAIXA ETÁRIA, MUNICIPIO_RESIDENCIA_COD, MUNICIPIO RESIDENCIA, COMORBIDADE, INTERNACAO, UTI, DATA_ATUALIZACAO**. Com base na informação dos dados os alunos têm o intuito de aplicar os conhecimentos adquiridos ao longo do semestre, identificar padrões ou relações que possam ser uteis na discussão.

As adversidades que surgiram no trabalho foram várias. Os dados apresentam uma estrutura complexa e os alunos entenderam que poderiam ter outros campos para permitir uma análise mais detalhada. Os dados também apresentam um défice de atributos com valores numéricos, esses fatores contribuíram para que houve uma maior dificuldade na identificação de padrões.

De um modo geral os objetivos do trabalho foram parcialmente cumpridos, ficando por limar alguns gráficos e conclusões em relação a aplicabilidade dos algoritmos. Todavia, os conhecimentos adquiridos durante as aulas foram todos aplicados de forma adequada, onde também se explorou os conhecimentos para que houvesse um maior enriquecimento do trabalho.

Descrição dos dados

Com a situação pandémica que o mundo atravessa, os alunos resolveram fazer uma investigação e através dela recolher dados sobre o número de registo de casos do novo coronavírus no brasil no ano de 2020, entre os meses de março e junho.

Descrição dos dados na tabela dos dados

Atributo	Descrição
ID	número identificador do caso
CONTADOR	número de pessoa testada no Brasil
DATA_NOTIFICACAO	data de notificação do resultado do teste
CLASSIFICACAO_CASO	se o teste é confirmado ser positivo ou não
SEXO	género da pessoa (masculino ou feminino)
IDADE	idade da pessoa que realizou o teste
FAIXA_ETARIA	faixa etária da pessoa testada
MUNICIPIO_RESIDENCIA_COD	município de residência da pessoa
MUNICIPIO_RESIDENCIA	município de residência da pessoa
COMORBIDADE	se a pessoa tem comorbidade prognóstica
EVOLUCAO	evolução do estado da pessoa no momento da data de obtenção dos dados
INTERNACAO	se a pessoa esteve internada ou não
UTI	se a pessoa esteve numa unidade de tratamento intensivo
DATA_ATUALIZACAO	data de obtenção dos dados anteriores (único valor, 25/06/2020)

Figura 1- Descrição dos dados na tabela de dados

Escalas de medida

Atributo	Escala de medida
ID	ordinal
CONTADOR	ordinal
DATA_NOTIFICACAO	ordinal
CLASSIFICACAO_CASO	nominal
SEXO	nominal
IDADE	razão ou rácio
FAIXA_ETARIA	intervalar
MUNICIPIO_RESIDENCIA_COD	nominal
MUNICIPIO_RESIDENCIA	nominal
COMORBIDADE	nominal
EVOLUCAO	nominal
INTERNACAO	nominal
UTI	nominal
DATA_ATUALIZACAO	ordinal

Figura 2- Escalas de medida dos dados da tabela

Conforme é possível perceber na **Figura 1** acima, os atributos apresentam uma variedade no que diz respeito ao tipo de atributos (numéricos, caracteres), embora haja um défice de atributos numéricos que possam ser relevantes para aplicabilidade dos algoritmos na regressão linear.

Na **Figura 2**, é possível perceber que os atributos apresentam uma variedade no que diz respeito às escalas de medida (escala nominal, escala ordinal e escala métrica), o que irá facilitar na escolha dos atributos para aplicar os algoritmos.

De um modo geral, os atributos apresentam uma boa variedade de dados, todavia alguns atributos irão precisar de algum tratamento para facilitar a aplicabilidade dos algoritmos.

ANÁLISE EXPLORATÓRIA DOS DADOS

O objetivo desse trabalho é, através dos algoritmos estudados durante o semestre recolher informação e identificar padrões que possam ser uteis. Para este efeito, foram aplicados conhecimentos e técnicas de aprendizagem usadas na área da Ciência de Dados, que foram estudados ao longo do semestre.

Para que houvesse uma boa prática dos conhecimentos optamos por tratar alguns atributos, onde tivemos que ter em conta o tipo de aprendizagem e os algoritmos utilizados. Para algoritmos que apenas aceitam atributos numéricos como por exemplo o algoritmo de regressão linear foi escolhido o atributo *target* a **IDADE,** que representa a idade de cada individuo que realizou o teste.

Este atributo foi escolhido com objetivo de verificar que relação pode existir entre a IDADE e outros atributos, com por exemplo a possibilidade de um individuo com uma idade elevada que testou positivo a CVD-19 ser internado, obtido, fazer parte dos cuidados intensivos, etc...
Para algoritmos de aprendizagem que permitem atributos categóricos como target, como por exemplo as árvores de classificação e modelos de classificação, foi escolhido o atributo INTERNACAO_SIM como o atributo *target*, que resultado de um tratamento (***dummys***) e representa se um determinado individuo foi/não foi internado.

DESCRIÇÃO DOS DADOS

Para identificar melhor os atributos e obter uma informação mais detalhada dos mesmos foram aplicadas algumas funções disponibilizadas pelos pandas, tais como o ***describe***. Essas Informações são uteis para saber o tipo de atributo, variedade das informações dos atributos, etc…

Descrição dos dados numéricos

	ID	CONTADOR	IDADE	MUNICIPIO_RESIDENCIA_COD
count	500.000000	500.000000	500.000000	493.00000
mean	252.130000	296546.500000	45.936000	312459.64503
std	146.508945	144.481833	16.634582	2327.08741
min	1.000000	296297.000000	0.000000	310160.00000
25%	125.750000	296421.750000	34.000000	310620.00000
50%	250.500000	296546.500000	42.000000	310620.00000
75%	380.250000	296671.250000	58.000000	314480.00000
max	507.000000	296796.000000	94.000000	317100.00000

Figura 3- Discrição dos dados numéricos

Descrição dos dados não numéricos

	DATA_NOTIFICACAO	CLASSIFICACAO_CASO	SEXO	FAIXA_ETARIA	MUNICIPIO_RESIDENCIA	COMORBIDADE	EVOLUCAO	INTERNACAO	UTI
count	500	500	500	500	500	500	500	500	500
unique	33	1	2	10	58	3	3	3	3
top	04/04/2020	Caso Confirmado	Masculino	30 a 39 anos	BELO HORIZONTE	Não Informado	RECUPERADO	NÃO	NÃO
freq	65	500	291	139	253	338	479	299	365

Figura 4- Discrição dos dados não numéricos

A informação apresentada na **Figura 3 e 4** permite recolher uma serie de informações bastante uteis, como o número de registos, número de vezes que um determinado valor aparece em cada coluna e o tipo de atributo.

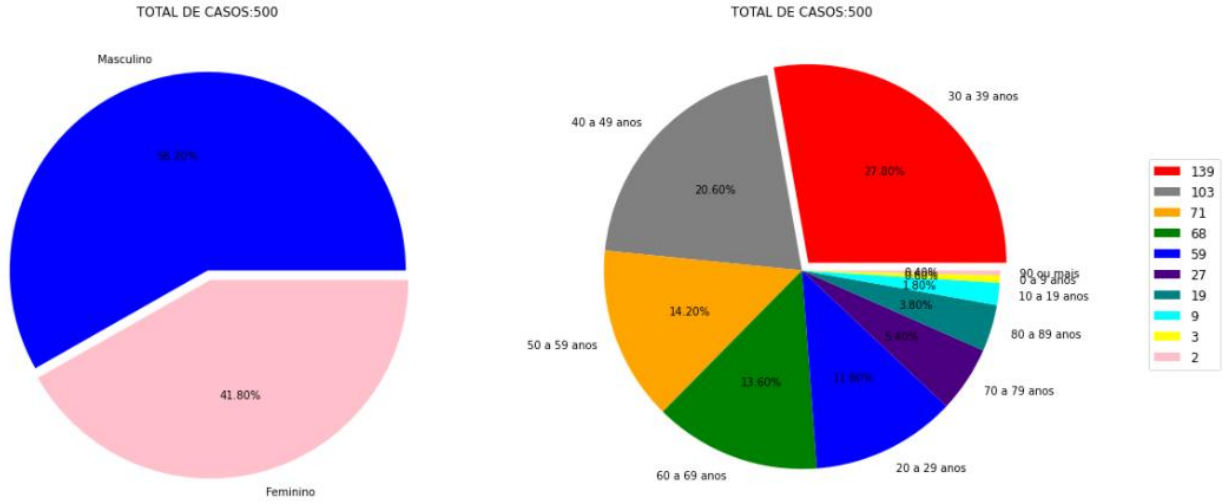

Figura 5- Taxa de infeção por sexo e por grupo etário

Na **Figura 5** acima, o primeiro gráfico mais a esquerda ilustra as percentagens das pessoas infetadas com o coronavírus, onde podemos observar que os indivíduos do sexo masculino apresentam maior percentagem de infeção em relação aos indivíduos do sexo feminino.

No segundo gráfico mais a direita ilustra as percentagens de pessoas infetadas por grupo etário, onde podemos observar que o grupo compreendido entre as idades dos 30 aos 39 apresenta maior percentagem de infeção

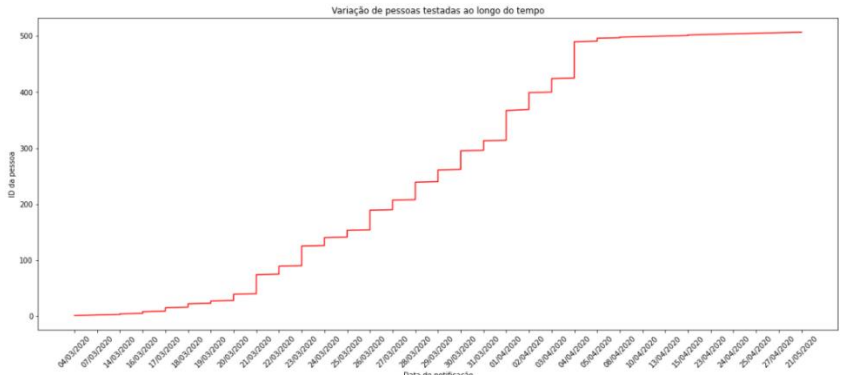

Figura 6 - Número de pessoas testadas ao logo do tempo.

Na **Figura 6** acima, gráfico mostra o número de pessoas testadas ao logo do tempo, onde podemos observar que o crescimento é exponencial, ou seja, na medida que o tempo passa mais testes são realizados devido ao aumento de pessoas infetadas. É importante realçar que chega uma dada altura que o crescimento dos casos deixa de ser exponencial, mantendo-se constante e essa informação reflete-se na **Figura 6** acima conforme podemos ver.

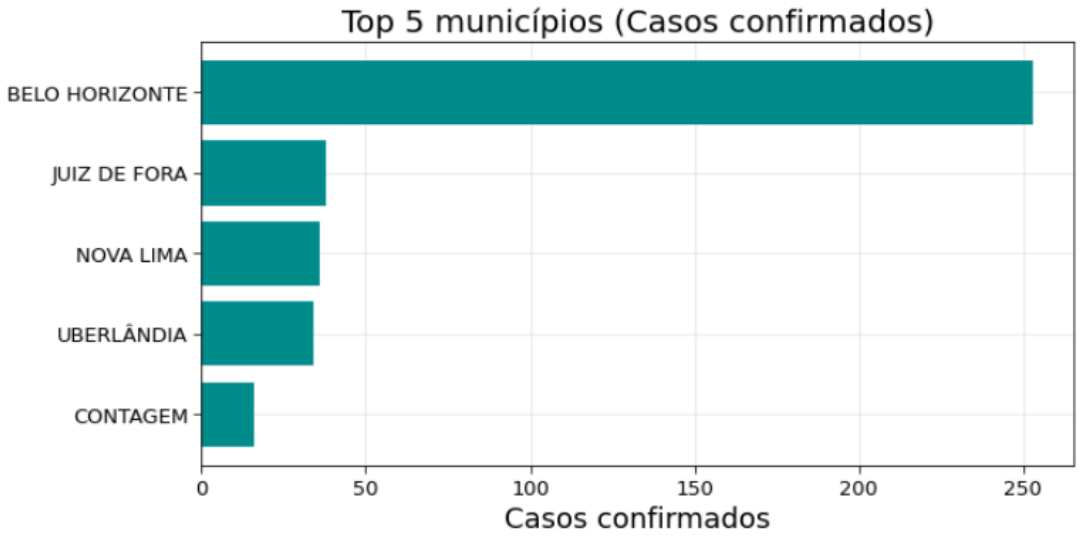

Figura 7- Top 5 Municípios (Casos Confirmados)

Na **Figura 7** acima, o gráfico mostra o top 5 das regiões com maior número de casos, onde o município de BELO HORIZONTE apresenta maior número de casos.

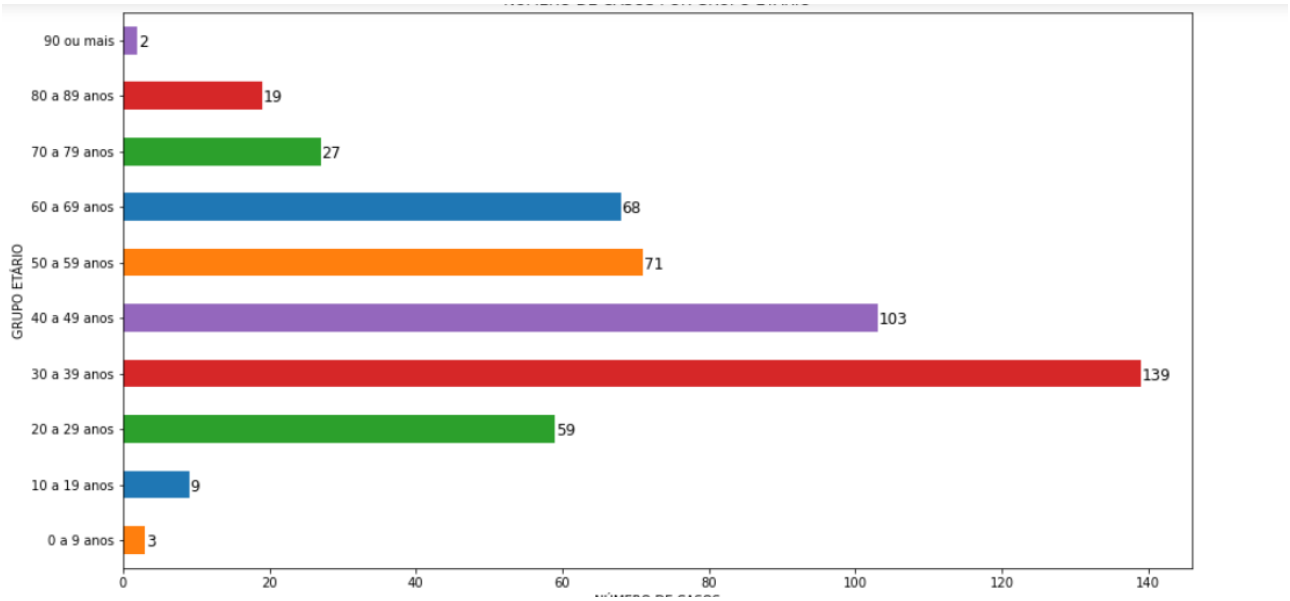

Figura 8 - Número de casos por grupo etário

Na **Figura 8** acima, gráfico ilustra o número de casos registados por faixa etária, podemos observar que o grupo com as idades compreendidas entre os 30 e os 39 apresentam um maior número de casos em relação às demais faixas etárias e o grupo com as idades compreendidas entre os 0 e 9 apresenta um menor número de casos.

É importante salientar que apesar dos grupos etários com idades superiores a 60 terem o menor número de casos, pertencem ao grupo de risco.

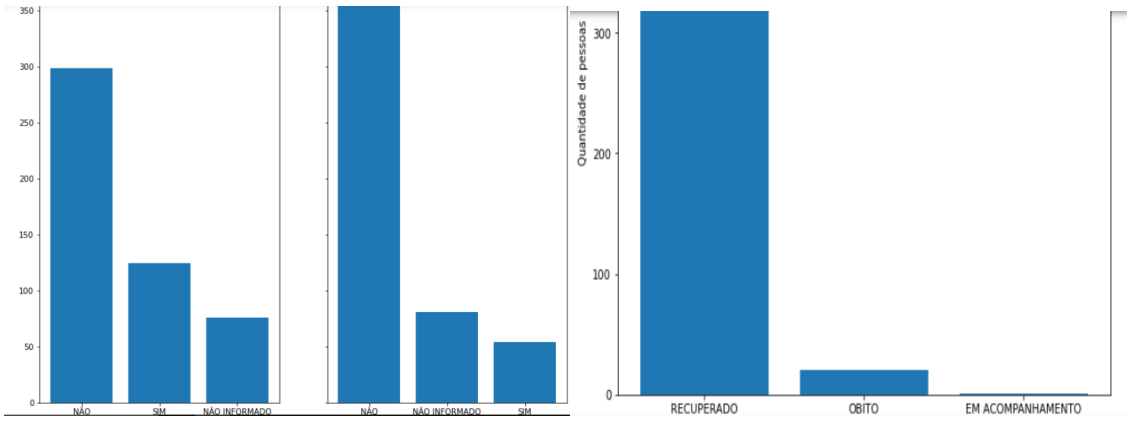

Figura 9- Evolução do vírus

Na **Figura 9** acima, o gráfico a direita compara a quantidade de pessoas das quais se sabe ou não (ou não foi informado) que estiveram internadas com as que estiveram nos cuidados intensivos e o gráfico a esquerda ilustra a evolução dos vírus, onde podemos observar que temos um número de recuperados bastante considerável quando comparado com os óbitos e casos em acompanhamento.

Qualidade dos dados

De um modo geral podemos considerar que os dados apresentam uma boa qualidade embora seja preciso fazer um pré-processamento conforme iremos ver nas seções …

Identificação de outliers

Os principais outliers foram detetados durante a aplicação da regressão multivariada. Como a diferença consequente à remoção de alguns outliers se revelou pouco significante para a qualidade do modelo de regressão, decidimos manter os dados originais. Como o modelo aprende sobre

20475 instâncias, a remoção de algumas baseadas no gráfico de influência terá pouco importância.

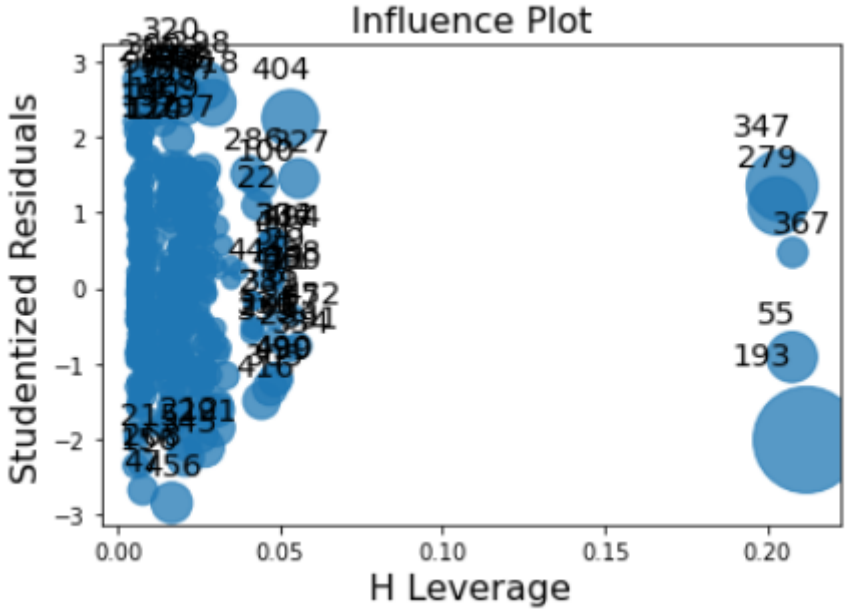

Figura 10 - Gráfico de influência

Pré-processamento

Os atributos por vezes apresentaram informações pouco relevantes para a aprendizagem na nossa perspetiva. Para ultrapassar esse problema optamos por excluir alguns atributos, tais como: DATA_NOTIFICACAO e MUNICIPIO_RESIDENCIA, os atributos foram excluídos por possuírem diversos valores distintos (data de notificação possui 33 valor distintos e o município de residência possui 58 valores destintos)

Nesta etapa também foram criados dummys para os atributos com valores de um domínio inferior a 4. Após criar os dummys foram eliminados os respetivos atributos "originais". A **Figura 11** abaixo ilustra o resultado obtido.

	ID	IDADE	SEXO_Feminino	SEXO_Masculino	COMORBIDADE_Não Informado	COMORBIDADE_SIM	INTERNACAO_NÃO	INTERNACAO_SIM	UTI_NÃO	UTI_SIM
1	1	38	1	0	1	0	1	0	1	0
2	2	47	1	0	1	0	1	0	1	0
3	3	65	0	1	0	1	0	1	0	1
4	4	37	1	0	1	0	1	0	1	0
5	5	45	0	1	1	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...
496	503	64	1	0	0	1	0	1	1	0
497	504	78	0	1	0	1	0	1	0	1
498	505	67	1	0	0	1	0	1	1	0
499	506	36	0	1	0	0	0	1	1	0
500	507	82	0	1	0	1	0	1	1	0

Figura 11 - Criação de dummys

APRENDIZAGEM NÃO-SUPERVISIONADA

Algoritmos Hierárquicos

Na **Figura 12,** foi aplicado o método clustering hierárquico, onde através do mesmo é apresentado um dendrograma com 2 *clusters,* para verificar melhor recorremos o método de Cotovelo **(Figura 13).**

Após verificar os *clusters* através do metódo de cotovelo confirmou-se que o número de clusters *é 2* **(Figura 13).**

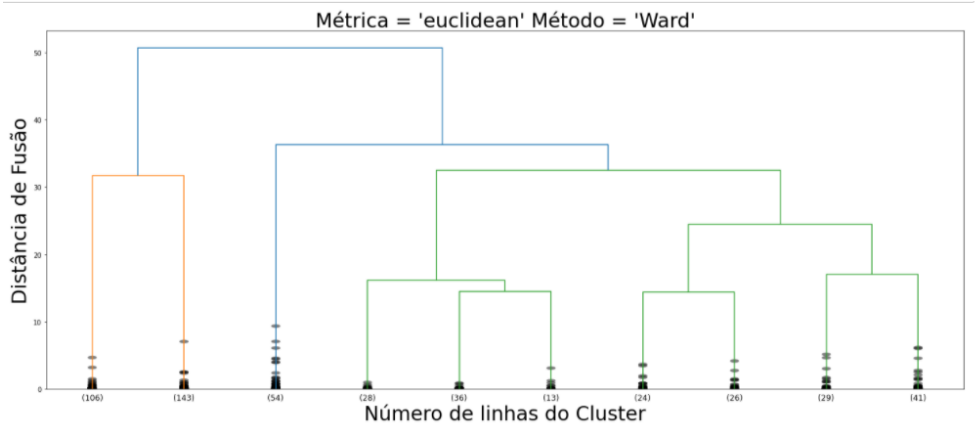

Figura 12 – Número de linhas do cluster hierárquico

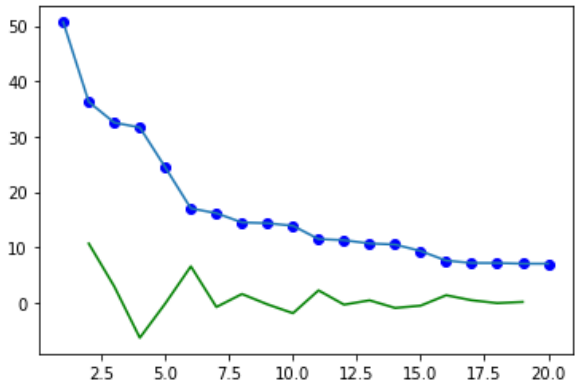

Figura 13 – Gráfico de cotovelo dos dois clusters

Algoritmos não hierárquicos

Erro Quadrado: 2692799.080795635

	ID	SEXO_Feminino	SEXO_Masculino	COMORBIDADE_Não Informado	COMORBIDADE_SIM	INTERNACAO_NÃO	INTERNACAO_SIM	UTI_NÃO	UTI_SIM
0	380.299595	0.453441	0.546559	0.607287	0.303644	0.360324	0.331984	0.546559	0.133603
1	127.000000	0.383399	0.616601	0.743083	0.229249	0.830040	0.169960	0.909091	0.083004

Figura 14 – Posição dos centroides

Aprendizagem Supervisionada

Estas técnicas baseiam-se na aprendizagem automática de um modelo a partir de dados de treino. A máquina infere, a partir de exemplos já classificados, consistindo de dados de entrada e dados de saída, uma função para mapear novos exemplos para o seu respetivo valor previsto (regressão) ou para a sua classe prevista (classificação).

Regressão linear

Através da análise dos resultados obtidos na **Figura 15** conseguimos, através do valor 0.244 de R-squared, perceber que o modelo que temos não é mau e que todas as instâncias são significativas com exceção das instâncias "UTI_NÃO" e "INTERNACAO_NÃO" uma vez que estas possuem P > |t| superior a 5%. Analisando uma instância mais significativa, por exemplo, "INTERNACAO_SIM", podemos ver que, à medida que a idade aumenta, existe um acréscimo de aproximadamente 13.96 em relação ao número de internados. Este resultado realmente faz sentido visto que pessoas com idade mais avançada realmente pertencem a um grupo de risco.

OLS Regression Results

Dep. Variable:	IDADE	**R-squared:**	0.244
Model:	OLS	**Adj. R-squared:**	0.231
Method:	Least Squares	**F-statistic:**	19.79
Date:	Sun, 07 Feb 2021	**Prob (F-statistic):**	6.13e-26
Time:	23:39:03	**Log-Likelihood:**	-2044.9
No. Observations:	500	**AIC:**	4108.
Df Residuals:	491	**BIC:**	4146.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	16.7652	2.842	5.899	0.000	11.181	22.349
ID	0.0220	0.006	3.978	0.000	0.011	0.033
SEXO_Feminino	8.4251	1.605	5.251	0.000	5.272	11.578
SEXO_Masculino	8.3402	1.533	5.442	0.000	5.329	11.351
COMORBIDADE_Não Informado	6.6406	2.972	2.234	0.026	0.801	12.480
COMORBIDADE_SIM	8.6476	3.037	2.847	0.005	2.680	14.615
INTERNACAO_NÃO	2.8005	7.225	0.388	0.698	-11.396	16.997
INTERNACAO_SIM	13.9590	6.811	2.050	0.041	0.577	27.341
UTI_NÃO	3.1878	6.814	0.468	0.640	-10.201	16.577
UTI_SIM	9.0656	6.866	1.320	0.187	-4.424	22.556

Omnibus:	15.200	**Durbin-Watson:**	1.931
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	15.737
Skew:	0.413	**Prob(JB):**	0.000383
Kurtosis:	3.268	**Cond. No.**	2.17e+17

Figura 15 – Regressão Linear com variável dependente IDADE

Árvores de decisão

Uma árvore de decisão é uma representação de uma tabela de decisão sob a forma de árvore. Trata-se de uma forma alternativa de expressar as mesmas regras que são obtidas quando se constrói a tabela.
Foi realizada uma árvore de decisão utilizando o critério da entropia para tentar prever o valor da variável UTI_SIM, ou seja, a necessidade de internamento de uma pessoa com o vírus (**Figura 16**).
No entanto, foi notado que a idade era a variável com maior impacto (**Figura 17**) e daí procedeu-se ao uso de uma árvore de regressão (**Figura 18**) para obter uma previsão da evolução do vírus com as idades. Tal resultado, como esperado, pode ser visto na **Figura 19** onde se mostra a visualização textual da árvore que prevê o vírus ser contraído em pessoas de maior idade.

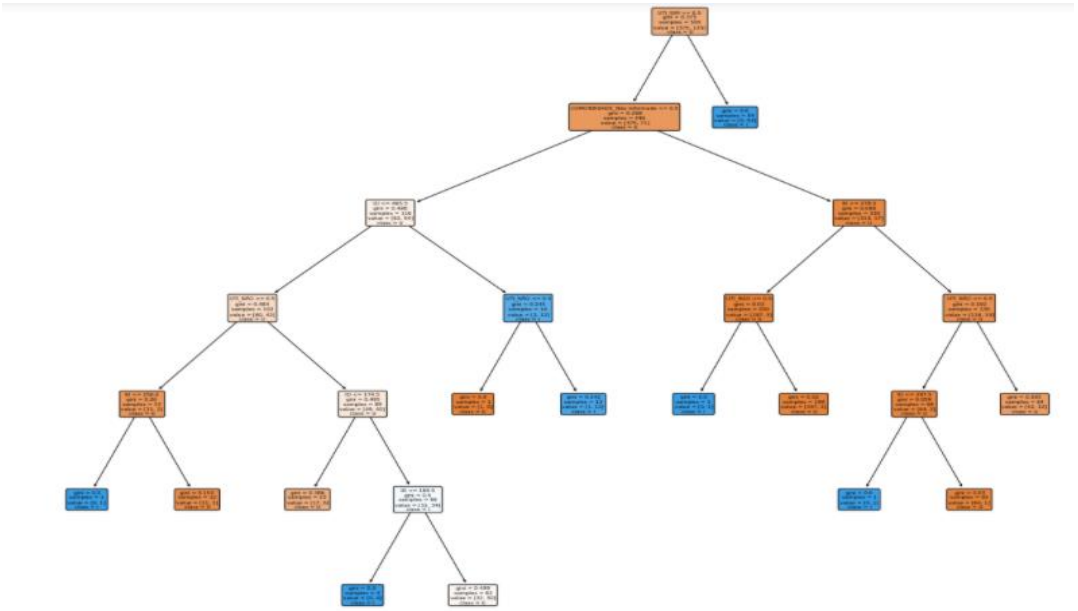

Figura 16 – Árvore de decisão usando o critério da entropia

	ID	IDADE	SEXO_Feminino	SEXO_Masculino	COMORBIDADE_Não Informado	COMORBIDADE_SIM	INTERNACAO_NÃO	INTERNACAO_SIM	UTI_NÃ
Importância	0.000329	0.998944	0.0	0.0	0.0	0.000019	0.0	0.0	0.00062

Figura 17 – Impacto das amostras

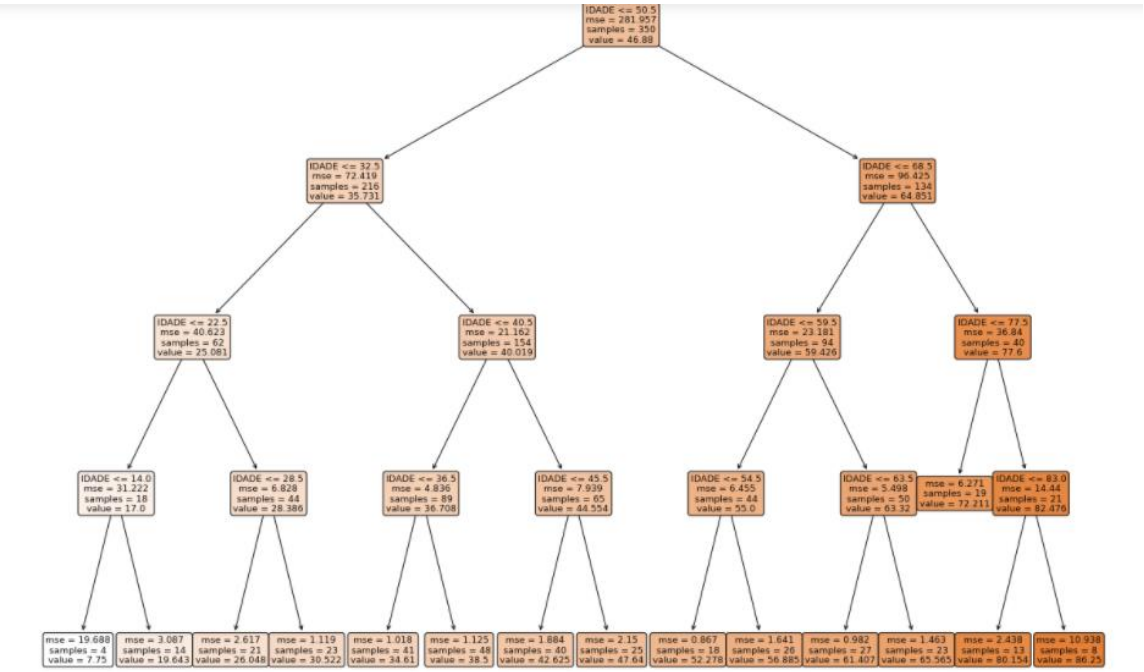

Figura 18 – Árvore de regressão pós-poda

```
|--- IDADE <= 50.50
|   |--- IDADE <= 32.50
|   |   |--- IDADE <= 22.50
|   |   |   |--- IDADE <= 14.00
|   |   |   |   |--- value: [7.75]
|   |   |   |--- IDADE >  14.00
|   |   |   |   |--- value: [19.64]
|   |   |--- IDADE >  22.50
|   |   |   |--- IDADE <= 28.50
|   |   |   |   |--- value: [26.05]
|   |   |   |--- IDADE >  28.50
|   |   |   |   |--- value: [30.52]
|   |--- IDADE >  32.50
|   |   |--- IDADE <= 40.50
|   |   |   |--- IDADE <= 36.50
|   |   |   |   |--- value: [34.61]
|   |   |   |--- IDADE >  36.50
|   |   |   |   |--- value: [38.50]
|   |   |--- IDADE >  40.50
|   |   |   |--- IDADE <= 45.50
|   |   |   |   |--- value: [42.62]
|   |   |   |--- IDADE >  45.50
|   |   |   |   |--- value: [47.64]
|--- IDADE >  50.50
|   |--- IDADE <= 68.50
|   |   |--- IDADE <= 59.50
|   |   |   |--- IDADE <= 54.50
|   |   |   |   |--- value: [52.28]
|   |   |   |--- IDADE >  54.50
|   |   |   |   |--- value: [56.88]
|   |   |--- IDADE >  59.50
|   |   |   |--- IDADE <= 63.50
|   |   |   |   |--- value: [61.41]
|   |   |   |--- IDADE >  63.50
|   |   |   |   |--- value: [65.57]
|   |--- IDADE >  68.50
|   |   |--- IDADE <= 77.50
|   |   |   |--- value: [72.21]
|   |   |--- IDADE >  77.50
|   |   |   |--- IDADE <= 83.00
|   |   |   |   |--- value: [80.15]
|   |   |   |--- IDADE >  83.00
|   |   |   |   |--- value: [86.25]
```

Figura 19 – Representação textual da árvore de regressão

Algoritmos RNA para regressão

Na **Figura 20** notamos que, apesar de não ter sido usada estratificação, as distribuições nas três amostras são semelhantes, como é importante que o sejam.
Através **da Figura 21** podemos comparar o modelo inicial com o modelo SGD, notando que o último é melhor como podemos também ver através da comparação entre as matrizes de confusão da Figura 23. Através da análise da **Figura 22** notamos que os existem grandes diferenças entre os pesos dos neurónios, sendo assim um bom resultado a obter

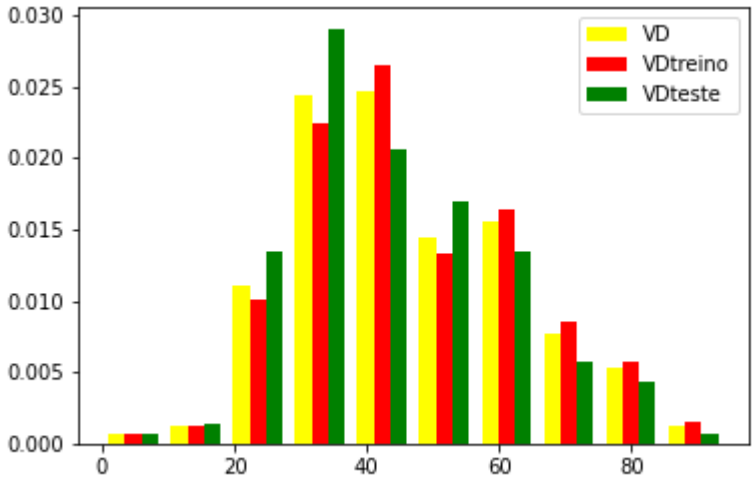

Figura 20 – Modelo de regressão

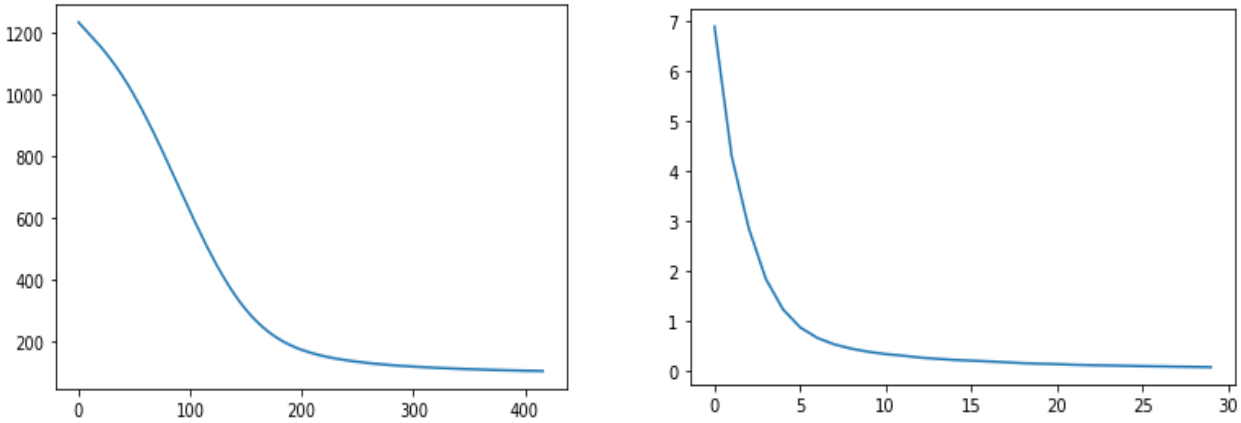

Figura 21 – Modelo loss curve (inicial vs final)

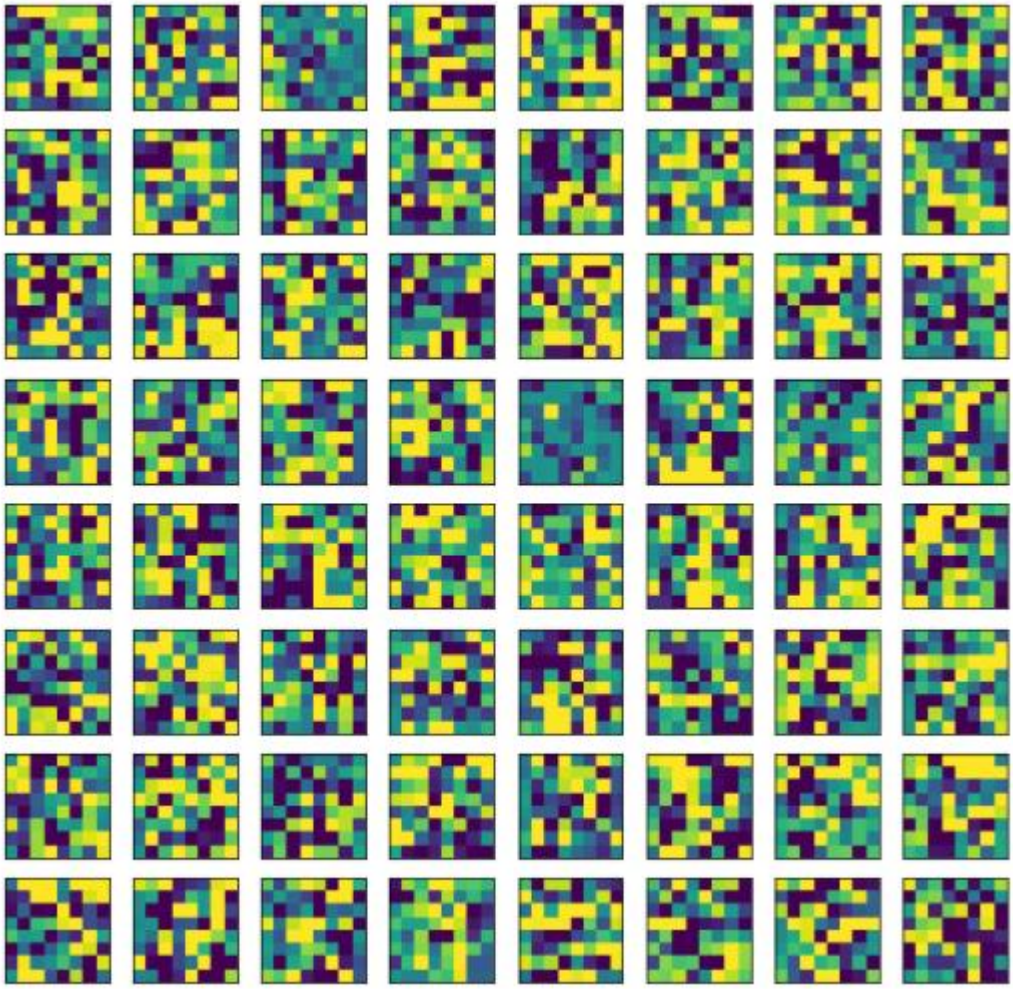

Figura 22 – Neurónios e respetivos pesos (final)

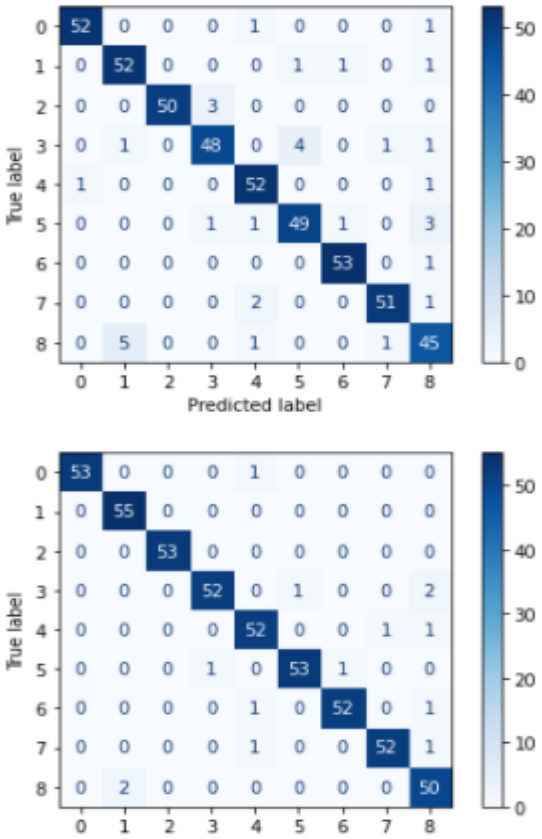

Figura 23 – Comparação das matrizes de ambos os modelos