UNIVERSIDADE FEDERAL DE SANTA MARIA
ELC1098 - MINERAÇÃO DE DADOS

Lorenzo Moreira Donatti
Ramon Panazollo

TRABALHO FINAL

SANTA MARIA, 17 DE DEZEMBRO DE 2022

INTRODUÇÃO

Para a resolução deste trabalho, foi utilizada a plataforma Google Colab e a linguagem Python para encontrar as principais características e padrões no DataSet. As bibliotecas utilizadas foram: Pandas, Numpy, Sci-Kit Learn, Mlxtend e dtreeviz.

O dataset utilizado (saudeRS_2022.csv) consiste em dados coletados pela Secretaria Estadual da Saúde de pacientes com covid-19 no estado do Rio Grande do Sul. Possui diversos registros, desde informações pessoais, datas, doenças prévias, até a evolução do quadro do paciente.

PRÉ PROCESSAMENTO

Na etapa do pré-processamento dos dados, o dataset disponibilizado foi baixado. A primeira etapa foi a escolha de quais colunas e informações seriam relevantes para uma análise inicial. Esta primeira análise consistiu na busca por padrões em pacientes que vieram a óbito. As principais colunas utilizadas foram: *FAIXAETARIA, EVOLUCAO, HOSPITALIZADO, FEBRE, TOSSE, GARGANTA, DISPNEIA, CONDICOES e SRAG.*

A seguir, foi necessário verificar dados faltantes e seu impacto no dataset. Dados faltantes da coluna *CONDICOES* foram preenchidos com a string '*Nada',* representando que o paciente não possuía doenças prévias. Além disso, valores categóricos, como SIM e NÃO, foram substituídos por booleanos True e False.

Para a verificação e criação de regras de associação, via algoritmo apriori, foi necessário subdividir o dataset por 'óbitos' e 'recuperados'. Além disso, foi necessário manipular os datasets para que ficassem aptos a serem processados pela biblioteca MLxtend.

Para isso, colunas que possuíam diversos valores categóricos possíveis foram transformados em novas colunas com apenas True e False como valores possíveis, através da função TransactionEncoder.

DESENVOLVIMENTO

Primeira Análise

Com os datasets preparados e limpos, foi possível a aplicação do algoritmo apriori, onde diversas regras de associação foram encontradas. Tanto para o dataset com pacientes que vieram a óbito, quanto para o dataset de pacientes recuperados. O foco da primeira análise foi visualizar regras relacionadas ao dataset com óbitos.

O padrão para análise de regras foi: support = 0.2 e confidence = 0.8. Visualizando o suporte dos itemsets, é possível inferir que 100% dos pacientes que vieram a óbito tiveram SRAG (Síndrome respiratória aguda grave), e que 96% dos pacientes foram hospitalizados.

```
obito = apriori(df_final_obito, min_support = 0.2, use_colnames=True)
obito.sort_values(by=['support'],ascending=False)
```

	support	itemsets
4	1.000000	(SRAG)
0	0.966322	(HOSPITALIZADO)
14	0.966322	(HOSPITALIZADO, SRAG)
29	0.825633	(DISPNEIA, SRAG)
3	0.825633	(DISPNEIA)
...	...	...
68	0.212504	(Outros, SRAG, TOSSE)
88	0.212409	(HOSPITALIZADO, Doença Cardiovascular Crônica,...
60	0.212409	(HOSPITALIZADO, Doença Cardiovascular Crônica,...
82	0.207760	(HOSPITALIZADO, Outros, SRAG, TOSSE)
47	0.207760	(HOSPITALIZADO, Outros, TOSSE)

95 rows × 2 columns

Figura 1 - Valores de suporte para itemsets presentes no dataset óbito

Com isso, as principais regras de associação que resultaram em SRAG e consequentemente óbito dos pacientes foram encontradas:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(HOSPITALIZADO)	(SRAG)	0.966322	1.0	0.966322	1.0	1.0	0.0	inf
16	(DISPNEIA)	(SRAG)	0.825633	1.0	0.825633	1.0	1.0	0.0	inf
45	(HOSPITALIZADO, DISPNEIA)	(SRAG)	0.804288	1.0	0.804288	1.0	1.0	0.0	inf
15	(TOSSE)	(SRAG)	0.608386	1.0	0.608386	1.0	1.0	0.0	inf
40	(HOSPITALIZADO, TOSSE)	(SRAG)	0.590646	1.0	0.590646	1.0	1.0	0.0	inf
13	(FEBRE)	(SRAG)	0.537994	1.0	0.537994	1.0	1.0	0.0	inf
33	(HOSPITALIZADO, FEBRE)	(SRAG)	0.522626	1.0	0.522626	1.0	1.0	0.0	inf
89	(DISPNEIA, TOSSE)	(SRAG)	0.520254	1.0	0.520254	1.0	1.0	0.0	inf
129	(HOSPITALIZADO, DISPNEIA, TOSSE)	(SRAG)	0.508775	1.0	0.508775	1.0	1.0	0.0	inf
24	(Doença Cardiovascular Crônica)	(SRAG)	0.496442	1.0	0.496442	1.0	1.0	0.0	inf
67	(HOSPITALIZADO, Doença Cardiovascular Crônica)	(SRAG)	0.482971	1.0	0.482971	1.0	1.0	0.0	inf
85	(DISPNEIA, FEBRE)	(SRAG)	0.447965	1.0	0.447965	1.0	1.0	0.0	inf
117	(HOSPITALIZADO, DISPNEIA, FEBRE)	(SRAG)	0.437435	1.0	0.437435	1.0	1.0	0.0	inf
98	(DISPNEIA, Doença Cardiovascular Crônica)	(SRAG)	0.414572	1.0	0.414572	1.0	1.0	0.0	inf

Figura 2 - Regras de associação ordenadas por maior suporte.

Além disso, buscou-se entender o impacto da idade dos indivíduos que vieram a óbito. Para isso a coluna ‘antecedents’ foi preenchida com as faixas de idade presentes no dataset. Foi possível inferir que a confiança acima de 80% esteve presente apenas em indivíduos idosos, isso reafirma a constatação que idosos são um grupo de risco. Os seguintes resultados foram obtidos:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
27	(80 e mais)	(SRAG)	0.289062	1.000000	0.289062	1.000000	1.000000	0.000000	inf
11	(80 e mais)	(HOSPITALIZADO)	0.289062	0.966322	0.276065	0.955038	0.988322	-0.003262	0.749029
78	(80 e mais)	(HOSPITALIZADO, SRAG)	0.289062	0.966322	0.276065	0.955038	0.988322	-0.003262	0.749029
21	(80 e mais)	(DISPNEIA)	0.289062	0.825633	0.231382	0.800459	0.969510	-0.007277	0.873841
106	(80 e mais)	(DISPNEIA, SRAG)	0.289062	0.825633	0.231382	0.800459	0.969510	-0.007277	0.873841

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
26	(70 a 79)	(SRAG)	0.291149	1.000000	0.291149	1.000000	1.000000	0.000000	inf
10	(70 a 79)	(HOSPITALIZADO)	0.291149	0.966322	0.284129	0.975888	1.009899	0.002785	1.396728
75	(70 a 79)	(HOSPITALIZADO, SRAG)	0.291149	0.966322	0.284129	0.975888	1.009899	0.002785	1.396728
20	(70 a 79)	(DISPNEIA)	0.291149	0.825633	0.240964	0.827631	1.002420	0.000582	1.011591
103	(70 a 79)	(DISPNEIA, SRAG)	0.291149	0.825633	0.240964	0.827631	1.002420	0.000582	1.011591
58	(70 a 79)	(HOSPITALIZADO, DISPNEIA)	0.291149	0.804288	0.236885	0.813620	1.011603	0.002717	1.050070
163	(70 a 79)	(HOSPITALIZADO, DISPNEIA, SRAG)	0.291149	0.804288	0.236885	0.813620	1.011603	0.002717	1.050070

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
25	(60 a 69)	(SRAG)	0.230623	1.000000	0.230623	1.000000	1.00000	0.000000	inf
9	(60 a 69)	(HOSPITALIZADO)	0.230623	0.966322	0.224931	0.975319	1.00931	0.002075	1.364521
72	(60 a 69)	(HOSPITALIZADO, SRAG)	0.230623	0.966322	0.224931	0.975319	1.00931	0.002075	1.364521

Figura 3 - Regras de associação para diferentes idades no dataset óbitos.

Segunda Análise

A segunda análise será uma tentativa de replicar os parâmetros da primeira análise no dataset de indivíduos recuperados. É importante frisar que o suporte mínimo terá que ser altamente reduzido, devido ao fato de que provavelmente a ocorrência de SRAG e hospitalizações seja menor. Além disso, o dataset de recuperados é muito maior que o de óbitos.

Se utilizássemos os mesmos parâmetros da primeira análise, estes seriam os itemsets com maior suporte:

```
recup = apriori(df_final_recup, min_support = 0.2, use_colnames=True)
recup.sort_values(by=['support'],ascending=False)
```

	support	itemsets
3	0.871215	(Nada)
1	0.428634	(TOSSE)
7	0.346673	(Nada, TOSSE)
0	0.319920	(FEBRE)
2	0.304594	(GARGANTA)
6	0.261859	(Nada, FEBRE)
8	0.261195	(Nada, GARGANTA)
4	0.227931	(30 a 39)
9	0.212993	(Nada, 30 a 39)
5	0.207192	(FEBRE, TOSSE)

Figura 4 - Valores de suporte para itemsets presentes no dataset recuperados

É possível notar uma grande diferença entre os itemsets mais frequentes dos dois datasets. ‘Nada’ passa a ser o maior suporte, que representa que o indivíduo não possui doenças crônicas, seguido de Tosse.

As seguintes regras de associação foram obtidas:

```
recup_ = association_rules(recup, metric="confidence", min_threshold=0.8)
recup_.sort_values(by=['confidence'],ascending=False)

#verificar consequents: SRAG, HOSPITALIZADO, DISPNEIA, Doença Cardiovascular Crônica,
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(30 a 39)	(Nada)	0.227931	0.871215	0.212993	0.934462	1.072597	0.014416	1.965050
2	(GARGANTA)	(Nada)	0.304594	0.871215	0.261195	0.857519	0.984280	-0.004172	0.903876
0	(FEBRE)	(Nada)	0.319920	0.871215	0.261859	0.818515	0.939510	-0.016860	0.709619
1	(TOSSE)	(Nada)	0.428634	0.871215	0.346673	0.808785	0.928343	-0.026759	0.673513

Figura 5 - Regras de associação ordenadas por maior confiança.

Para a obtenção de regras de associação que envolvam SRAG e Hospitalizações, foi necessário diminuir o suporte para 0.01, obtendo os seguintes resultados:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(HOSPITALIZADO)	(SRAG)	0.048015	0.048071	0.048015	1.000000	20.802605	0.045707	inf
50	(HOSPITALIZADO, FEBRE, TOSSE)	(SRAG)	0.023087	0.048071	0.023087	1.000000	20.802605	0.021977	inf
66	(HOSPITALIZADO, Nada, DISPNEIA)	(SRAG)	0.010865	0.048071	0.010865	1.000000	20.802605	0.010342	inf
63	(HOSPITALIZADO, DISPNEIA, Doença Cardiovascula...	(SRAG)	0.012191	0.048071	0.012191	1.000000	20.802605	0.011605	inf
61	(HOSPITALIZADO, Nada, TOSSE)	(SRAG)	0.011380	0.048071	0.011380	1.000000	20.802605	0.010833	inf
58	(HOSPITALIZADO, Doença Cardiovascular Crônica,...	(SRAG)	0.011165	0.048071	0.011165	1.000000	20.802605	0.010628	inf
56	(HOSPITALIZADO, DISPNEIA, TOSSE)	(SRAG)	0.026598	0.048071	0.026598	1.000000	20.802605	0.025319	inf
54	(HOSPITALIZADO, Nada, FEBRE)	(SRAG)	0.010322	0.048071	0.010322	1.000000	20.802605	0.009826	inf
52	(HOSPITALIZADO, DISPNEIA, FEBRE)	(SRAG)	0.021968	0.048071	0.021968	1.000000	20.802605	0.020912	inf
46	(DISPNEIA, Doença Cardiovascular Crônica)	(SRAG)	0.012201	0.048071	0.012201	1.000000	20.802605	0.011614	inf
30	(HOSPITALIZADO, 60 a 69)	(SRAG)	0.010680	0.048071	0.010680	1.000000	20.802605	0.010166	inf
28	(HOSPITALIZADO, 50 a 59)	(SRAG)	0.010111	0.048071	0.010111	1.000000	20.802605	0.009625	inf
26	(HOSPITALIZADO, Nada)	(SRAG)	0.015246	0.048071	0.015246	1.000000	20.802605	0.014513	inf
23	(HOSPITALIZADO, Doença Cardiovascular Crônica)	(SRAG)	0.015948	0.048071	0.015948	1.000000	20.802605	0.015181	inf
20	(HOSPITALIZADO, DISPNEIA)	(SRAG)	0.035550	0.048071	0.035550	1.000000	20.802605	0.033841	inf
17	(HOSPITALIZADO, TOSSE)	(SRAG)	0.034673	0.048071	0.034673	1.000000	20.802605	0.033007	inf
15	(HOSPITALIZADO, FEBRE)	(SRAG)	0.029638	0.048071	0.029638	1.000000	20.802605	0.028213	inf
87	(HOSPITALIZADO, DISPNEIA, FEBRE, TOSSE)	(SRAG)	0.017688	0.048071	0.017688	1.000000	20.802605	0.016837	inf
6	(Doença Cardiovascular Crônica)	(SRAG)	0.017020	0.048071	0.015962	0.937858	19.509884	0.015144	15.318543
37	(Doença Cardiovascular Crônica, TOSSE)	(SRAG)	0.012006	0.048071	0.011171	0.930452	19.355827	0.010594	13.687381

Figura 6 - Regras de associação para SRAG do dataset recuperados.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
59	(SRAG, Doença Cardiovascular Crônica, TOSSE)	(HOSPITALIZADO)	0.011171	0.048015	0.011165	0.999466	20.815603	0.010628	1783.067462
86	(SRAG, DISPNEIA, FEBRE, TOSSE)	(HOSPITALIZADO)	0.017702	0.048015	0.017688	0.999214	20.810349	0.016838	1210.924619
22	(DISPNEIA, Doença Cardiovascular Crônica)	(HOSPITALIZADO)	0.012201	0.048015	0.012191	0.999185	20.809754	0.011605	1168.466099
64	(DISPNEIA, SRAG, Doença Cardiovascular Crônica)	(HOSPITALIZADO)	0.012201	0.048015	0.012191	0.999185	20.809754	0.011605	1168.466099
24	(SRAG, Doença Cardiovascular Crônica)	(HOSPITALIZADO)	0.015962	0.048015	0.015948	0.999128	20.808565	0.015182	1091.926524
53	(DISPNEIA, SRAG, FEBRE)	(HOSPITALIZADO)	0.021988	0.048015	0.021968	0.999096	20.807892	0.020912	1052.895149
16	(SRAG, FEBRE)	(HOSPITALIZADO)	0.029666	0.048015	0.029638	0.999062	20.807183	0.028213	1014.679760
51	(SRAG, FEBRE, TOSSE)	(HOSPITALIZADO)	0.023109	0.048015	0.023087	0.999054	20.807014	0.021978	1005.988263
55	(Nada, FEBRE, SRAG)	(HOSPITALIZADO)	0.010332	0.048015	0.010322	0.999038	20.806685	0.009826	989.492963
57	(DISPNEIA, SRAG, TOSSE)	(HOSPITALIZADO)	0.026624	0.048015	0.026598	0.999029	20.806505	0.025320	980.690766
29	(SRAG, 50 a 59)	(HOSPITALIZADO)	0.010121	0.048015	0.010111	0.999018	20.806268	0.009625	969.310886
21	(DISPNEIA, SRAG)	(HOSPITALIZADO)	0.035586	0.048015	0.035550	0.998994	20.805779	0.033842	946.695959
18	(SRAG, TOSSE)	(HOSPITALIZADO)	0.034711	0.048015	0.034673	0.998912	20.804059	0.033007	874.823893
67	(Nada, DISPNEIA, SRAG)	(HOSPITALIZADO)	0.010877	0.048015	0.010865	0.998903	20.803882	0.010342	868.051440
31	(SRAG, 60 a 69)	(HOSPITALIZADO)	0.010692	0.048015	0.010680	0.998884	20.803487	0.010166	853.295676
1	(SRAG)	(HOSPITALIZADO)	0.048071	0.048015	0.048015	0.998842	20.802605	0.045707	822.106842
27	(Nada, SRAG)	(HOSPITALIZADO)	0.015268	0.048015	0.015246	0.998568	20.796892	0.014513	664.658453
62	(Nada, SRAG, TOSSE)	(HOSPITALIZADO)	0.011397	0.048015	0.011380	0.998430	20.794027	0.010832	606.414294
2	(Doença Cardiovascular Crônica)	(HOSPITALIZADO)	0.017020	0.048015	0.015948	0.937040	19.515473	0.015131	15.120485
19	(Doença Cardiovascular Crônica, TOSSE)	(HOSPITALIZADO)	0.012006	0.048015	0.011165	0.929955	19.367921	0.010588	13.591102

Figura 7 - Regras de associação para Hospitalizados do dataset recuperados.

Mesmo dentre os pacientes recuperados, é possível inferir, com 100% de confiança, que pacientes hospitalizados tiveram SRAG. Além disso, diversas doenças crônicas presentes no dataset de óbitos se confirmam como grupos de risco, como por exemplo, Doença Cardiovascular Crônica, onde em mais de 93% dos pacientes, foram hospitalizados.

Na análise da influência da idade dos pacientes, apenas pacientes menores de 60 anos tiveram retorno com confiança dentro do esperado. É possível notar que a associação se baseia que esses pacientes que se recuperaram da covid estão propícios a não possuírem doenças crônicas.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
32	(50 a 59)	(Nada)	0.151352	0.871215	0.123175	0.813834	0.934137	-0.008685	0.691778
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
31	(40 a 49)	(Nada)	0.190737	0.871215	0.170194	0.8923	1.024202	0.004022	1.195774
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
30	(30 a 39)	(Nada)	0.227931	0.871215	0.212993	0.934462	1.072597	0.014416	1.96505
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
29	(20 a 29)	(Nada)	0.185646	0.871215	0.175878	0.947388	1.087433	0.014141	2.447815
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
28	(15 a 19)	(Nada)	0.034896	0.871215	0.033037	0.946733	1.086682	0.002635	2.417724
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
27	(10 a 14)	(Nada)	0.01794	0.871215	0.017071	0.951574	1.092238	0.001442	2.659406
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
26	(05 a 09)	(Nada)	0.015586	0.871215	0.014908	0.956505	1.097898	0.001329	2.960933
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
25	(01 a 04)	(Nada)	0.014127	0.871215	0.013652	0.966366	1.109218	0.001344	3.829078
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction

Figura 8 - Regras de associação para diferentes idades no dataset recuperados.

Análise pela Árvore de decisão

A técnica escolhida, além das regras de associação, foi a árvore de decisão. A escolha se deve ao fato de que, além de um algoritmo simples que calcula entropias para a tomada de decisão, árvores de decisão possuem uma estrutura de fácil compreensão e montagem. Utilizou-se o dataset com a manipulação pré-processada, porém sem a divisão de óbitos e recuperados, sendo missão da árvore classificar corretamente. As bibliotecas utilizadas foram scikit-learn para treinar o modelo, e Dtreeviz para a visualização da estrutura hierárquica.