

Projeto 1, Classificação

Cristian Martins

Rodrigo Coffani

Vinícius Guimarães

Vitor Orsin

Dataset

Water quality

Lista de "ingredientes" que compõem a água em ambiente Urbano, podendo torná-la própria para consumo, ou não.

Características do Dataset

20 Atributos

Quantidade de X elemento presente na água, sendo entre eles: alumínio, amônia, mercúrio, bactérias, virus, entre outros...

Classificação

A água é classificada como: Segura ou insegura para consumo, representado na coluna **is_safe**

Sobre o Dataset

8000 linhas, possui uma quantidade satisfatória de amostras, e a característica que a maioria de suas amostras são classificadas como Segura para consumo

Atributos do dataset



coluna	(tradução) descrição
aluminium	(alumínio) perigoso se maior que 2.8
ammonia	(amônia) perigoso se maior que 32.5
arsenic	(arsênio) perigoso se maior que 0.01
barium	(bário) perigoso se maior que 2
cadmium	(cádmio) perigoso se maior que 0.005
chloramine	(cloraminas) perigoso se maior que 4
chromium	(crómio) perigoso se maior que 0.1
copper	(cobre) perigoso se maior que 1.3
fluoride	(fluoreto) perigoso se maior que 1.5
bacteria	(bactérias) perigoso se maior que 0
viruses	(vírus) perigoso se maior que 0
lead	(chumbo) perigoso se maior que 0.015
nitrates	(nitratos) perigoso se maior que 10
nitrites	(nitritos) perigoso se maior que 1
mercury	(mercúrio) perigoso se maior que 0.002
perchlorate	(perclorato) perigoso se maior que 56
radium	(rádio) perigoso se maior que 5
selenium	(selênio) perigoso se maior que 0.5
silver	(prata) perigoso se maior que 0.1
uranium	(urânio) perigoso se maior que 0.3
is_safe	(seguro) atributo de classe (0 - não seguro, 1 - seguro)

Problemas a serem enfrentados

1. Dados Desbalanceados

Está na hora de vermos e nos envolvermos com o mundo de forma diferente. No momento, experimentamos a realidade aumentada nos jogos, mas há muito mais do que entretenimento.

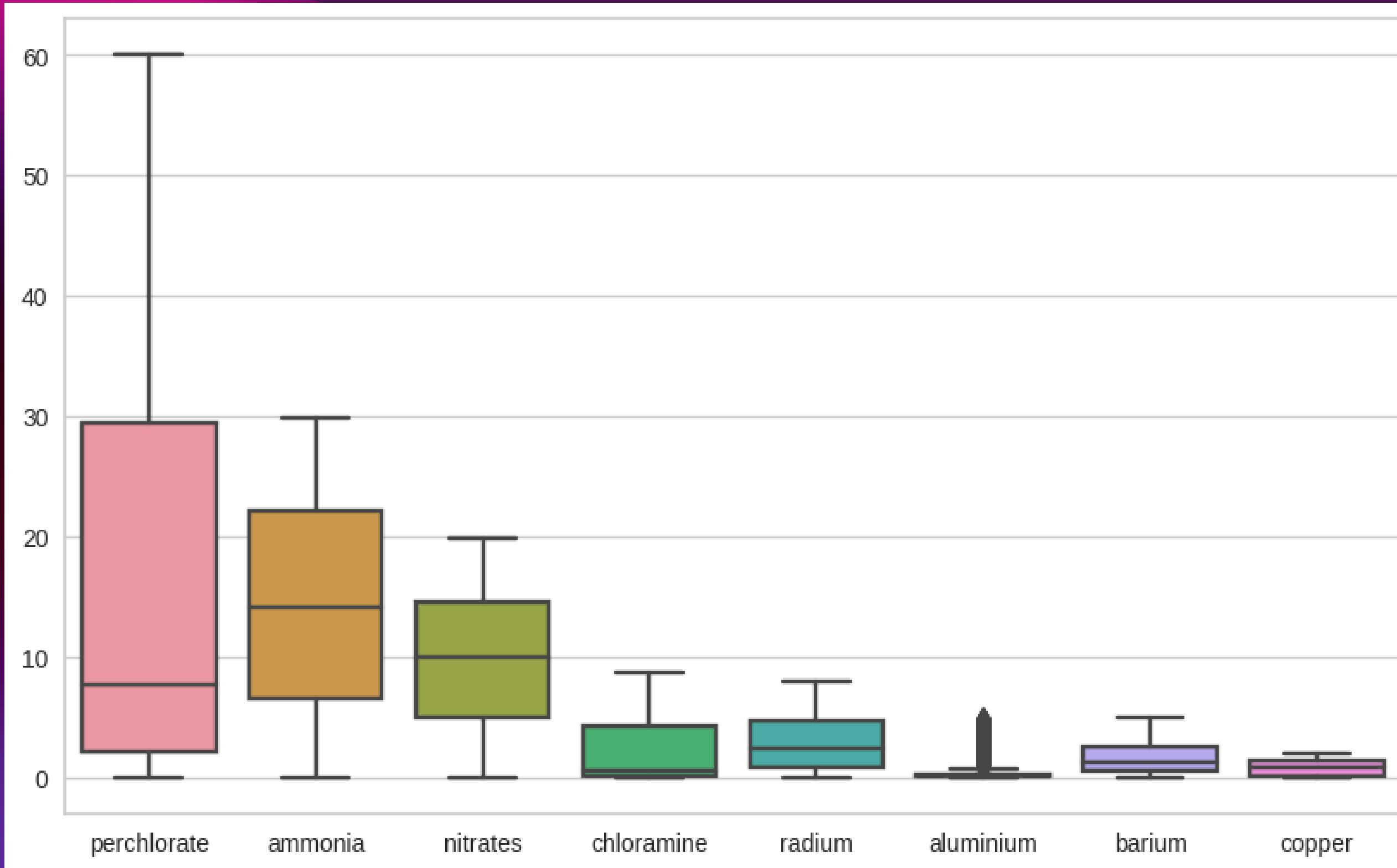
2. Seleção e redução de atributos

Casos reais de uso da realidade aumentada incluem treinamento médico, educação em sala de aula, logística de negócios e turismo.

3. Aplicação de Técnicas

Escolher métodos tanto de visualização para dos dados, quanto de como tratá-los para evitar vieses nos classificadores

Visualização dos dados



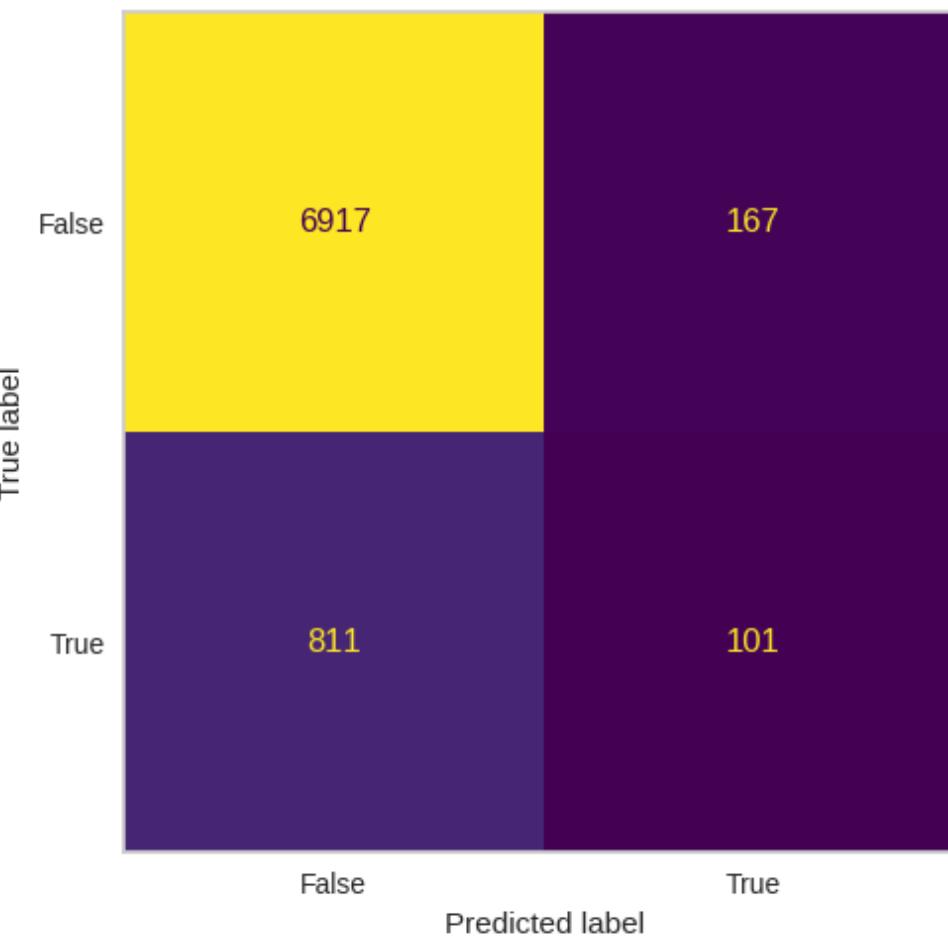
Dados desbalanceados



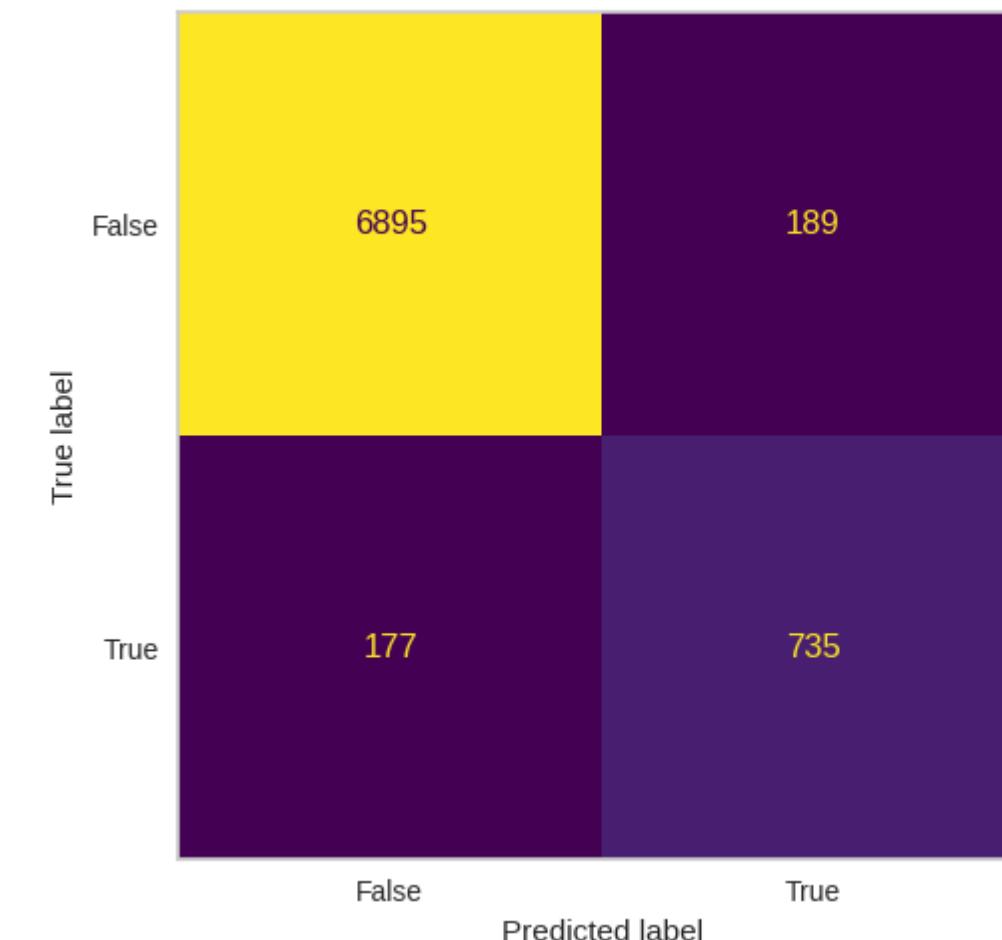
Desempenho sem pré-processamento

Característica comum entre os Classificadores sem Pré-processamento: Acurácia alta e Recall Baixa! Exceto a Rede neural com 0.63 de Recall

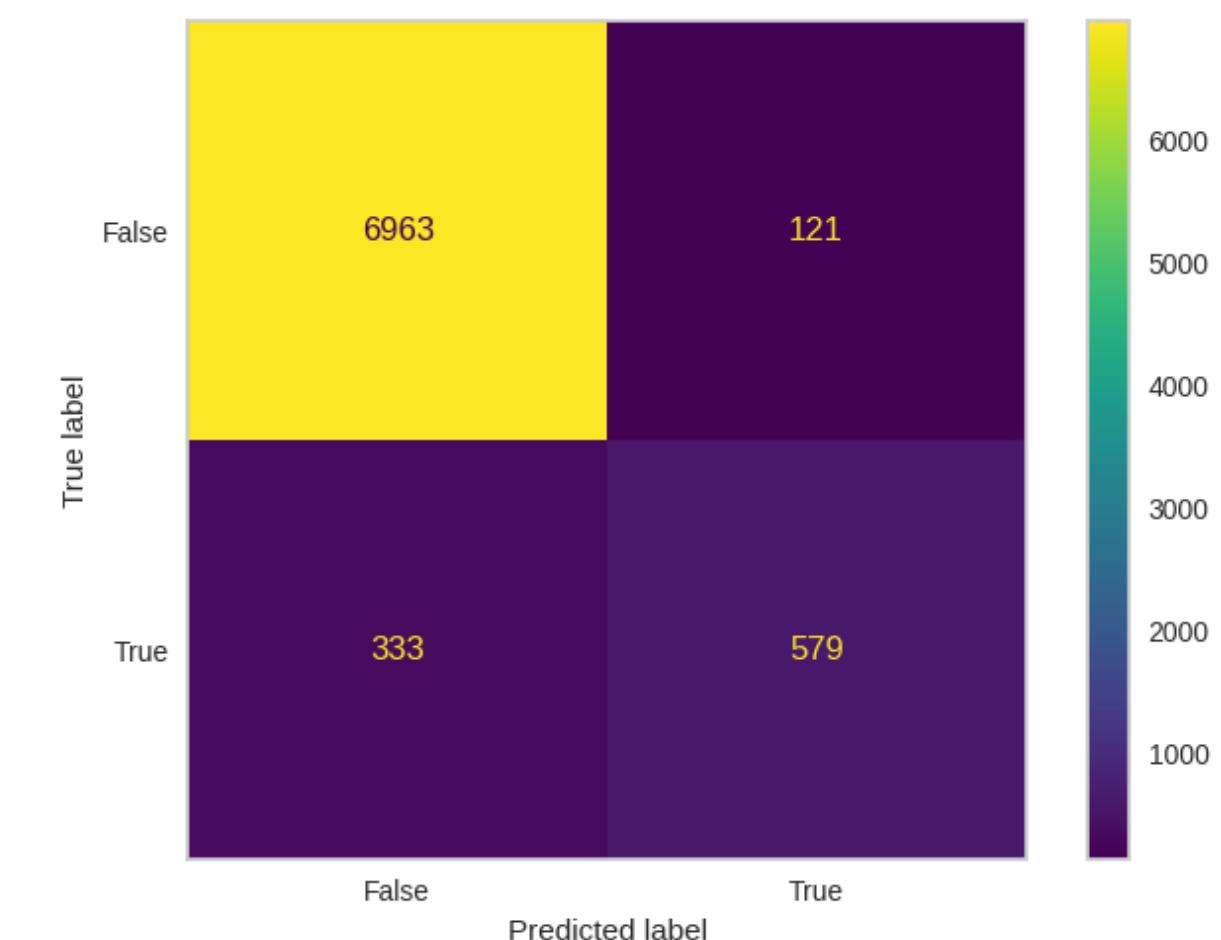
KNN



Árvore de Decisão



Rede Neural



Seleção e Redução de Atributos

Antes de Realizar a seleção, deve ser feito a Normalização e Amostragem.

Normalização

Z-Score

Escolhido o Z-Score pois ele lida bem com diferentes conjuntos de dados com diferentes médias e desvios padrão. Reduz o impacto de outliers na escala dos dados.

Seleção de Atributos

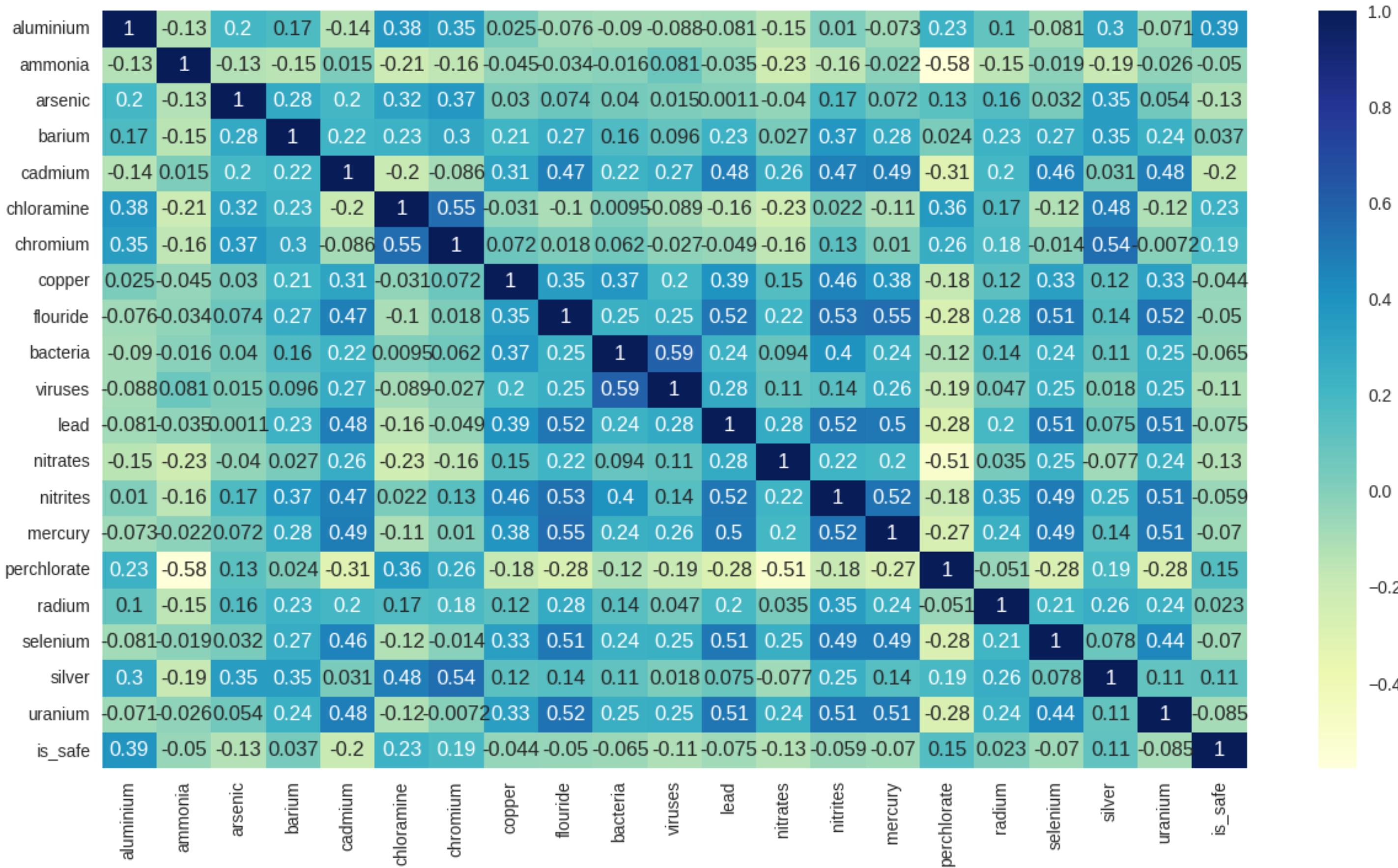
Considerar os 15 Atributos mais relevantes.

Usando Árvore de Decisão para selecionar os atributos.

Foi considerado a Redução dos atributos Virus e Bactérias para Microorganismos, mas, pela dispersão dos dados não foi usado essa técnica que não apresentava melhores resultados.

Correlação

Maiores correlações:
Vírus e Bactérias 0,59. Chromium e Cloramina 0,55



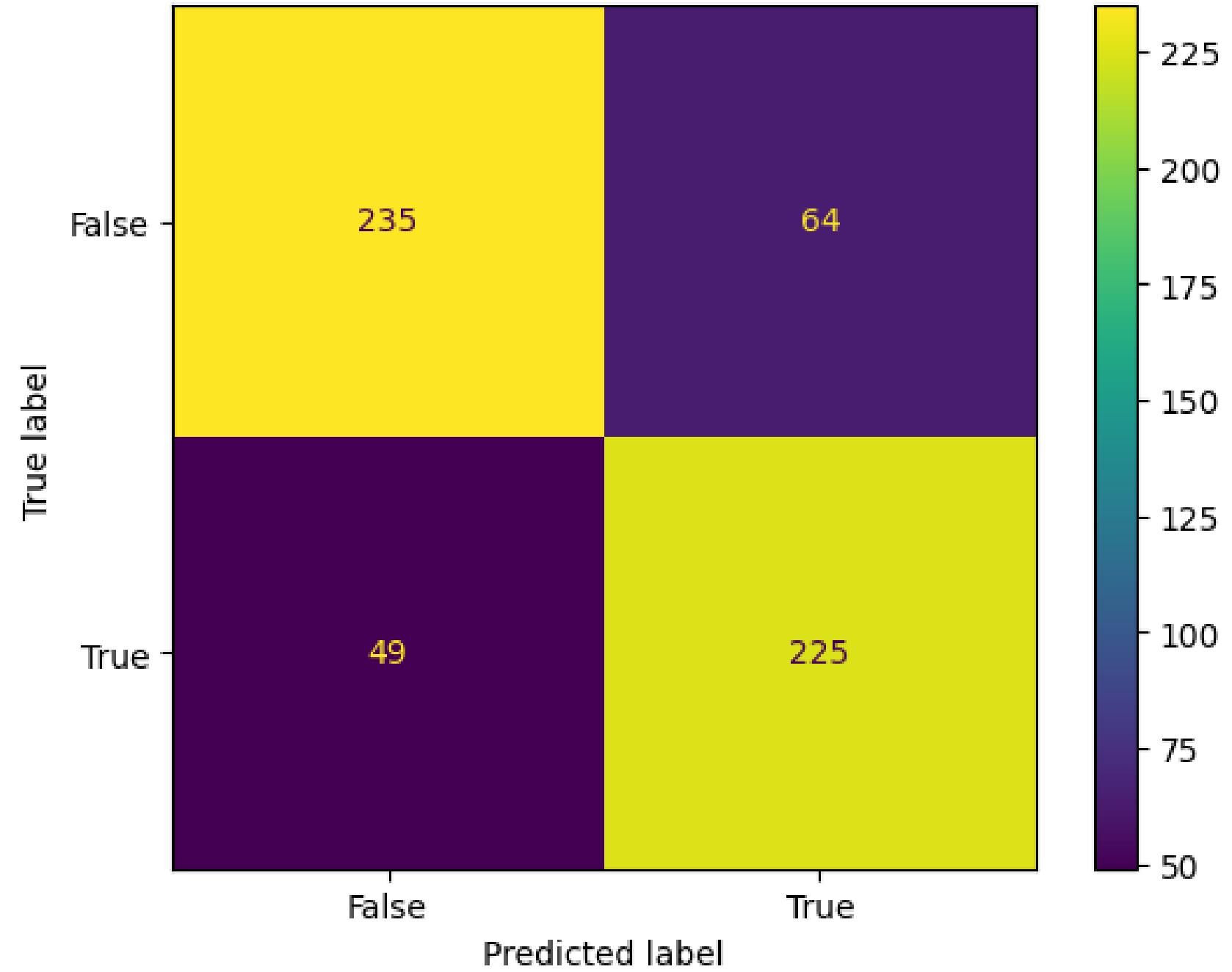
Amostragem

Devido ao Desbalanceamento, para deixar balanceado, pegaremos uma quantidade igual tuplas que são com classe 1 e 0 de forma aleatória, para evitar overfit ao Dataset.

	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	fluoride	bacteria	...	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0	0.274125	0.688501	0.003188	0.034425	0.000064	0.388875	0.044625	0.062475	0.050363	0.045900	...	0.007331	0.077775	0.073313	0.000128	0.486413	0.193163	0.003188	0.018488	0.004463	1.0
1	0.001435	0.457778	0.000179	0.008251	0.000018	0.028163	0.002153	0.004126	0.005561	0.007713	...	0.001830	0.022064	0.026907	0.000018	0.883267	0.089869	0.000359	0.003588	0.000538	0.0
2	0.018895	0.398539	0.004049	0.071918	0.000193	0.080402	0.013690	0.030078	0.012147	0.008676	...	0.002738	0.137281	0.034899	0.000135	0.898111	0.034127	0.001157	0.000578	0.000193	1.0
3	0.002783	0.501566	0.003897	0.039524	0.001113	0.001670	0.002227	0.047874	0.071811	0.000000	...	0.000891	0.670796	0.005010	0.000445	0.537750	0.013917	0.000557	0.004453	0.002227	0.0
4	0.032463	0.278025	0.000928	0.074666	0.000209	0.025507	0.015072	0.003014	0.024116	0.006029	...	0.003641	0.259938	0.034550	0.000046	0.919407	0.021797	0.002087	0.001855	0.000000	1.0
...	
1904	0.077653	0.891090	0.000696	0.105510	0.000035	0.053277	0.029947	0.022982	0.034822	0.023331	...	0.006338	0.407067	0.040045	0.000139	0.112823	0.044224	0.002438	0.013929	0.000000	1.0
1905	0.003279	0.840097	0.003826	0.034435	0.001093	0.001093	0.001640	0.001640	0.060124	0.000000	...	0.009019	0.438359	0.041540	0.000164	0.291328	0.086907	0.004919	0.000547	0.000000	0.0
1906	0.001352	0.622957	0.001352	0.020735	0.002254	0.025694	0.000902	0.088801	0.027497	0.023440	...	0.004282	0.600419	0.078433	0.000225	0.476909	0.083392	0.000451	0.003606	0.000902	0.0
1907	0.057189	0.062960	0.000017	0.047045	0.000087	0.100561	0.002623	0.010493	0.022561	0.000700	...	0.002536	0.148131	0.021861	0.000105	0.968887	0.136414	0.000874	0.005771	0.001049	0.0
1908	0.004513	0.561601	0.031089	0.165472	0.001003	0.386602	0.001504	0.002507	0.069197	0.013037	...	0.001655	0.378078	0.124354	0.000351	0.573134	0.121346	0.004513	0.013037	0.001003	0.0

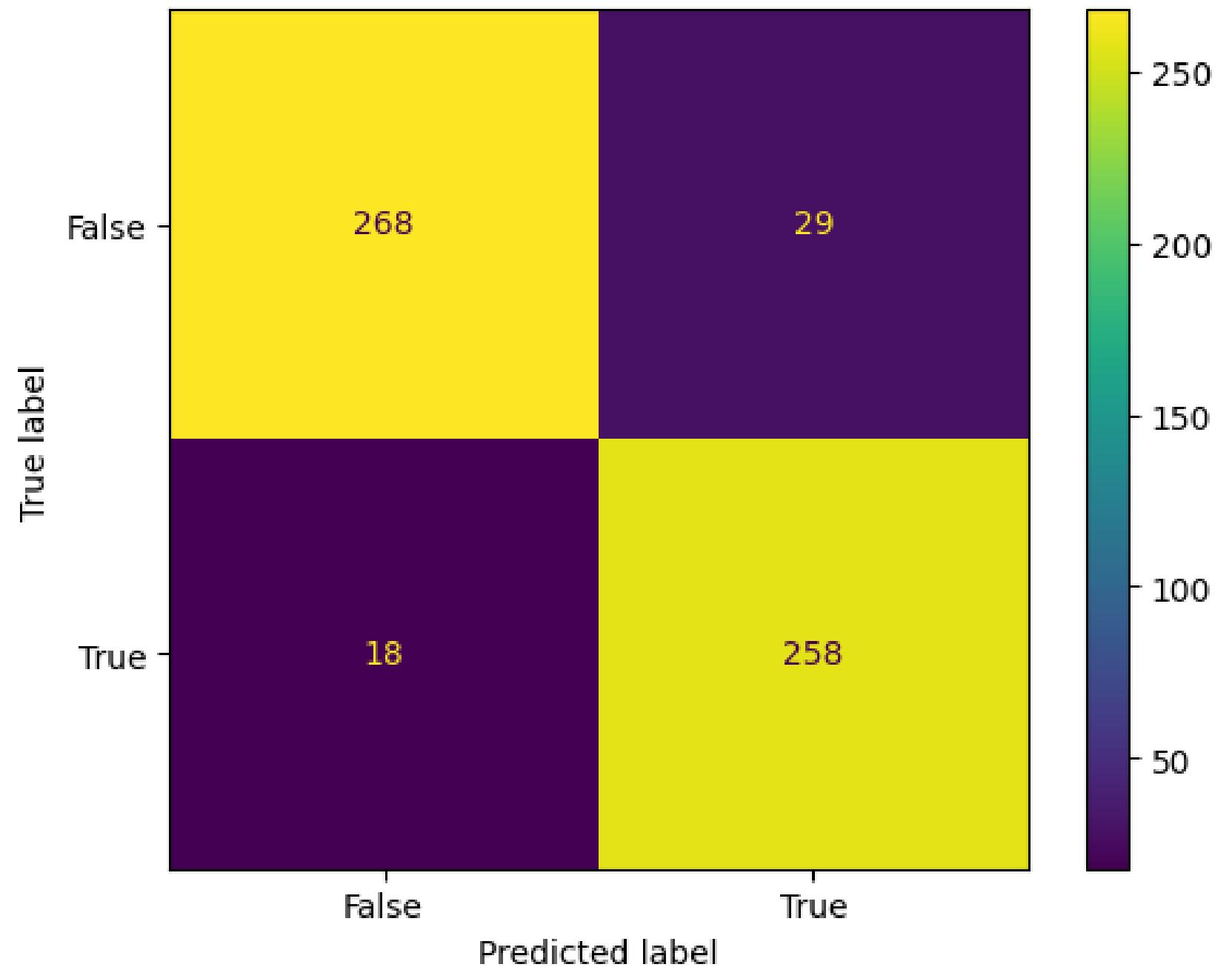
Classificação

KNN



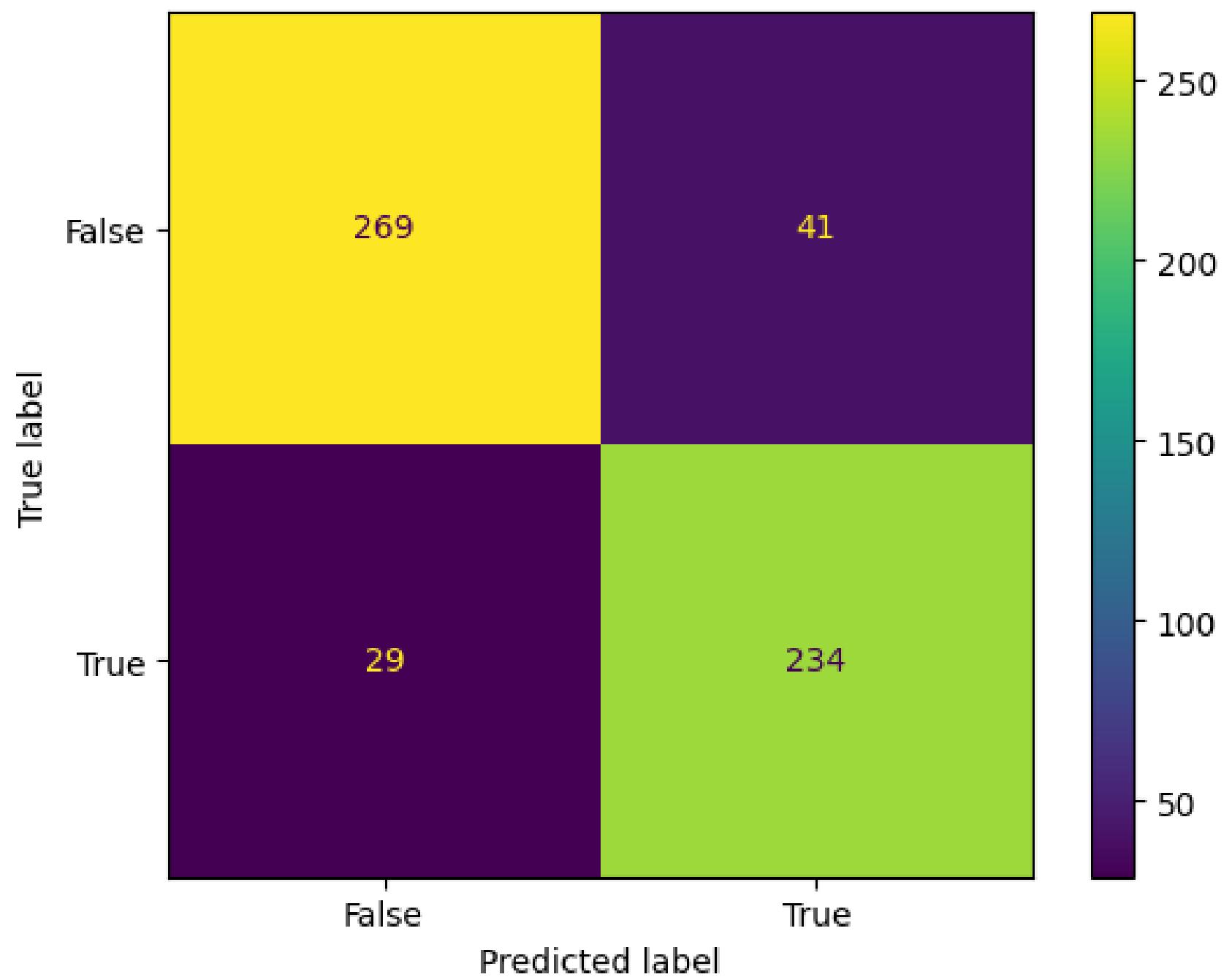
	precision	recall	f1_score	support
0	0.83	0.79	0.81	299
1	0.78	0.82	0.80	274
macro avg	0.80	0.80	0.80	573
weighted avg	0.80	0.80	0.80	573
accuracy		0.80		573

Decision Tree



	precision	recall	f1_score	support
0	0.94	0.90	0.92	297
1	0.90	0.93	0.92	276
macro avg	0.92	0.92	0.92	573
weighted avg	0.92	0.92	0.92	573
accuracy	0.92			573

Rede Neural



	precision	recall	f1_score	support
0	0.90	0.87	0.88	310
1	0.85	0.89	0.87	263
macro avg	0.88	0.88	0.88	573
weighted avg	0.88	0.88	0.88	573
accuracy	0.88			573

Conclusão

algoritmo	acurácia	f1_score
Árvore de Decisão	0.95	0.87
Árvore de Decisão (norm.)	0.92	0.92
Árvore de Decisão com Validação Cruzada	0.95	0.89
Árvore de Decisão com Validação Cruzada (dados norm.)	0.91	0.91
KNN	0.86	0.54
KNN (dados norm.)	0.80	0.80
KNN com Validação Cruzada	0.88	0.55
KNN com Validação Cruzada (dados norm.)	0.79	0.79
Rede Neural	0.88	0.55
Rede Neural (dados norm.)	0.88	0.88
Rede Neural com Validação Cruzada	0.88	0.55
Rede Neural com Validação Cruzada (dados norm.)	0.89	0.89