Relatório - 1ª Atividade de Aprendizado de Máquina Análise de Dados

Erick Santos do Nascimento

Curso de Ciência da Computação - UECE

erick.nascimento@aluno.uece.br

Resumo

Relatório da atividade sobre análise de dados da disciplina de Aprendizado de Máquina (2022.1). O objetivo é a exploração do comportamento dos dados. A seguir descrição dos resultados.

1 Introdução

Utilizando o dataset *Water Quality Dataset* as seguintes atividades foram feitas:

Seleção de 4 atributos, cálculo das seguintes medidas do dataset: (Média, Moda, Mediana, Percentis 20, 50 e 70, Quartis Q1, Q2 e Q3, Variância, Desvio padrão), geração de boxplots dos atributos selecionados, geração de scatter plots dois a dois dos atributos selecionados, cálculo da matrizes de Covariância e Correlação, 3 perguntas sobre o conjunto de dados

Essas são medidas estatísticas importantes na exploração dos dados, para identificar características como a presença de outliers usando boxplot, variância, a distribuição dos dados usando média, variância, quartis, entre outras características da base de dados.

Para saber mais sobre o dataset:
https://www.kaggle.com/datasets/adityakadiwal/water-potability

2 Análise de Dados

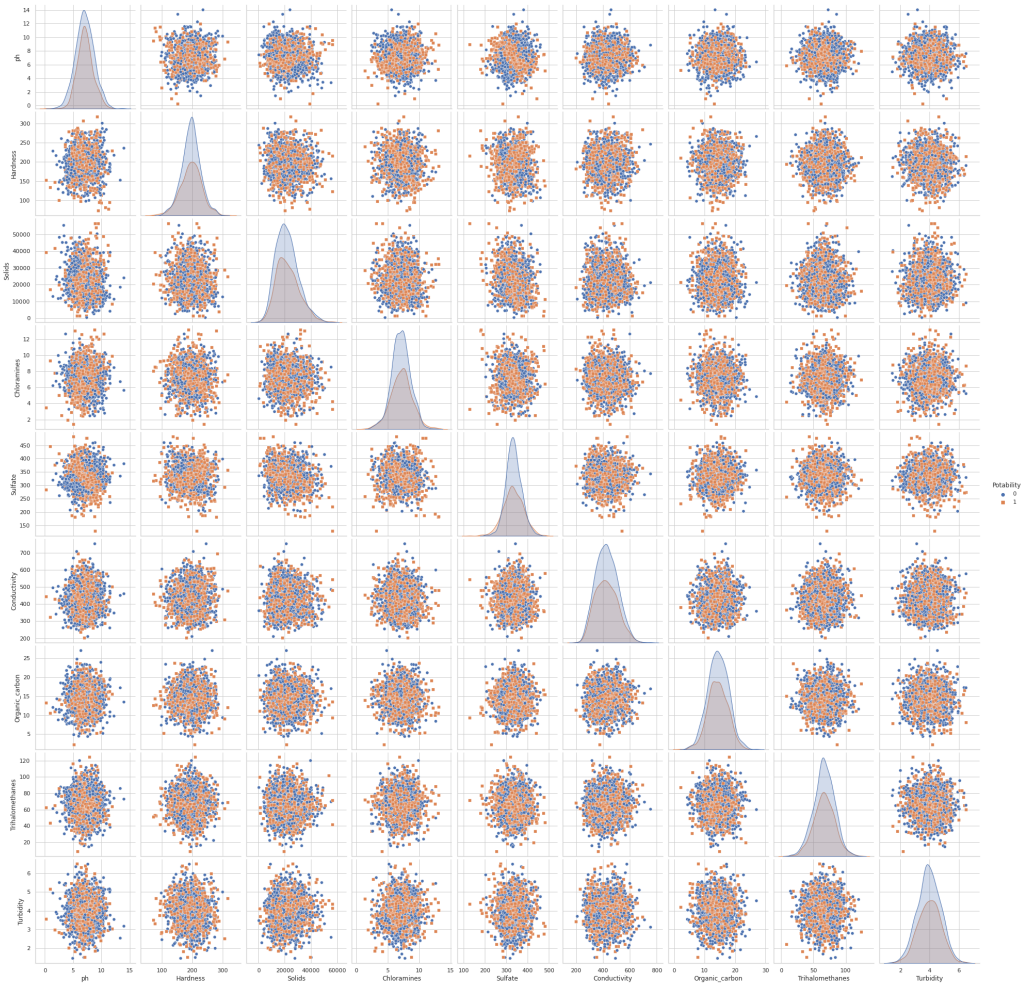
A análise de dados é um procedimento que visa transformar números e informações em insights para a tomada de decisão. Apesar de ser usada em diferentes áreas, mas em aprendizado de máquina ele é essencial já que qualquer modelo de aprendizado de máquina exige dados e a falta de tratamento deles pode prejudicar como no caso de bases com classes desbalanceadas, ou até impossibilitar o processo de aprendizagem de máquina como no caso de dados faltantes em colunas de features ou caracteres não reconhecidos.

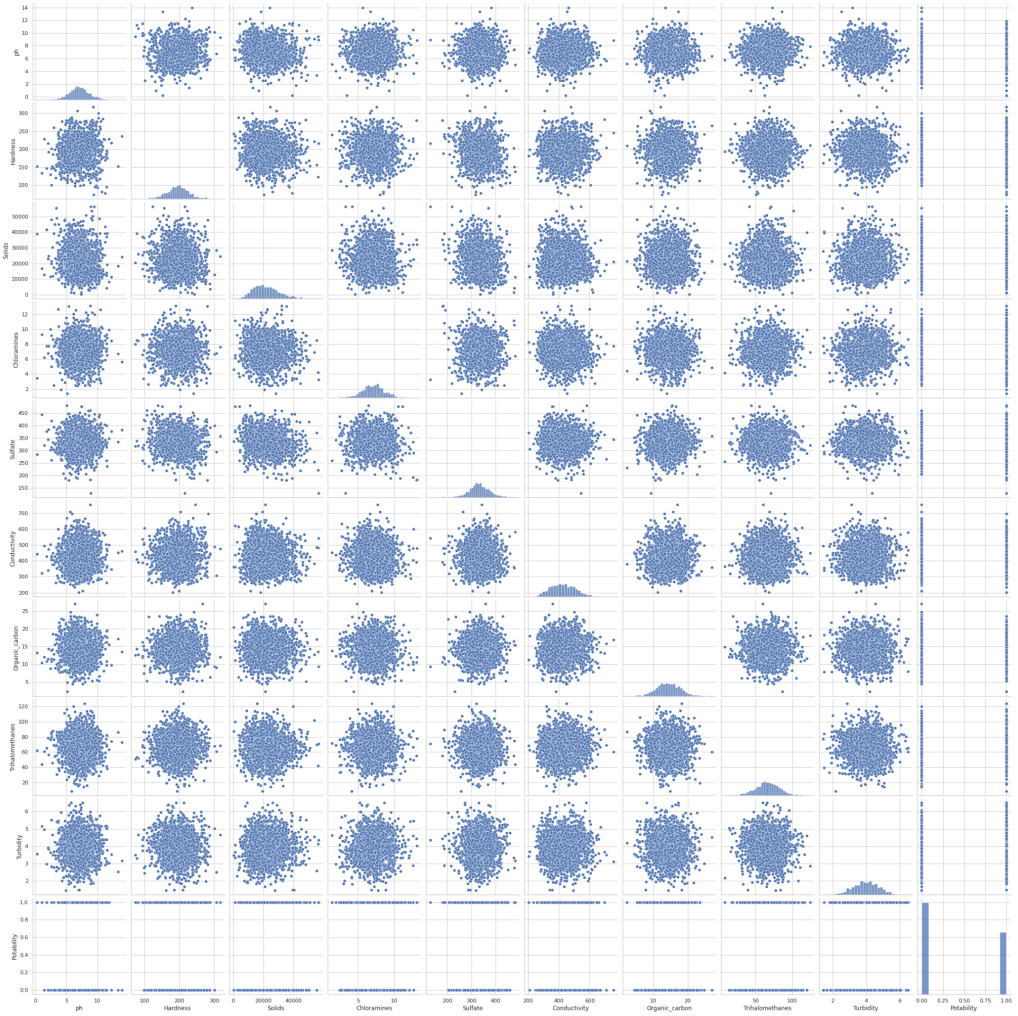
3 Metodologia

Como primeiro **passo eu desconsidero linhas com dados faltantes**, saindo de 3276 para 2011 linhas como exercício de análise dados considero uma perda aceitável, mas caso considerasse importante manter o máximo de linhas normalizaria os dados e copiaria as informações faltantes das amostras do vizinho mais próximo (menor distancia) de mesma classe que tivesse a informação faltante

A escolha dos atributos.

A escolha dos atributos no início foi feita de forma quase aleatória por 2 motivos falta de correlação entre as features e a coluna de labels como pode ser visto nas imagens a seguir:


ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity
Potability
0
1



Aqui (com foco na última coluna) que não há uma relação clara entre as variáveis (excluindo algumas situações onde parece ter uma pequena relação inversa entre algumas colunas como em ph x Sulfate, principalmente no caso de Potability=0, mais nenhuma delas envolve a coluna alvo e parece ser uma relação bem fraca), que é corroborado pela matriz de correlação

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
ph	1.000000	0.108948	-0.087615	-0.024768	0.010524	0.014128	0.028375	0.018278	-0.035849	0.014530
Hardness	0.108948	1.000000	-0.053269	-0.022685	-0.108521	0.011731	0.013224	-0.015400	-0.034831	-0.001505
Solids	-0.087615	-0.053269	1.000000	-0.051789	-0.162769	-0.005198	-0.005484	-0.015668	0.019409	0.040674
Chloramines	-0.024768	-0.022685	-0.051789	1.000000	0.006254	-0.028277	-0.023808	0.014990	0.013137	0.020784
Sulfate	0.010524	-0.108521	-0.162769	0.006254	1.000000	-0.016192	0.026776	-0.023347	-0.009934	-0.015303
Conductivity	0.014128	0.011731	-0.005198	-0.028277	-0.016192	1.000000	0.015647	0.004888	0.012495	-0.015496
Organic_carbon	0.028375	0.013224	-0.005484	-0.023808	0.026776	0.015647	1.000000	-0.005667	-0.015428	-0.015567
Trihalomethanes	0.018278	-0.015400	-0.015668	0.014990	-0.023347	0.004888	-0.005667	1.000000	-0.020497	0.009244
Turbidity	-0.035849	-0.034831	0.019409	0.013137	-0.009934	0.012495	-0.015428	-0.020497	1.000000	0.022682
Potability	0.014530	-0.001505	0.040674	0.020784	-0.015303	-0.015496	-0.015567	0.009244	0.022682	1.000000

Onde todos os valores (excluindo os da diagonal principal que uma feature com ela mesma) são próximos de zero (módulo menor que 0.2 o que é considerado uma correlação no máximo fraca).

E o segundo motivo é a minha falta de conhecimento técnico para saber quais características seriam as mais importantes nessa situação.

Neste caso selecionei os 4 primeiros atributos: ph (sei que um ph próximo de 7, ou seja neutro é o ideal para água potável), Hardness, Solids, Chloramines

1. pH: O pH é um parâmetro importante na avaliação do equilíbrio ácido-base da água. É também o indicador da condição ácida ou alcalina do estado da água. A OMS recomendou o limite máximo permitido de pH de 6,5 a 8,5. Os intervalos de investigação atuais foram de 6,52 a 6,83, que estão na faixa dos padrões da OMS.

2. Hardness (dureza): A dureza é causada principalmente por sais de cálcio e magnésio. Esses sais são dissolvidos a partir de depósitos geológicos através dos quais a água viaja. O período de tempo em que a água está em contato com o material produtor de dureza ajuda a determinar quanta dureza existe na água bruta. A dureza foi originalmente definida como a capacidade da água de precipitar sabão causada por cálcio e magnésio.

3. Solids (Sólidos) (total de sólidos dissolvidos - TDS): A água tem a capacidade de dissolver uma ampla gama de minerais ou sais inorgânicos e alguns orgânicos, como potássio, cálcio, sódio, bicarbonatos, cloretos, magnésio, sulfatos, etc. Esses minerais produziram sabor indesejado e cor diluída na aparência da água. Este é o parâmetro

importante para o uso da água. A água com alto valor de TDS indica que a água é altamente mineralizada. O limite desejável para TDS é de 500 mg/le o limite máximo é de 1000 mg/l prescrito para beber.

4. Chloramines (Cloraminas): Cloro e cloramina são os principais desinfetantes usados em sistemas públicos de água. As cloraminas são mais comumente formadas quando a amônia é adicionada ao cloro para tratar a água potável. Níveis de cloro de até 4 miligramas por litro (mg/L ou 4 partes por milhão (ppm)) são considerados seguros na água potável.

Na tabela são representados como valores reais em escalas diferentes, com algumas colunas variando apenas em algumas unidades e outras variando na casa do milhares com as classes levemente desbalanceadas com 1200 amostras em uma e 811 em outra

4 Resultados

Utilizando o dataset *Water Quality Dataset* as seguintes atividades foram feitas:

1. Seleção de 4 atributos:
    a. ph, Hardness, Solids, Chloramines
2. O cálculo das seguintes medidas do dataset: Média, Moda, Mediana, Percentis 20, 50 e 70, Quartis Q1, Q2 e Q3, Variância, Desvio padrão

	ph	Hardness	Solids	Chloramines	Potability
count	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000
mean	7.085990	195.968072	21917.441374	7.134338	0.403282
std	1.573337	32.635085	8642.239815	1.584820	0.490678
min	0.227499	73.492234	320.942611	1.390871	0.000000
20%	5.841119	170.132446	14450.175747	5.899676	0.000000
50%	7.027297	197.191839	20933.512750	7.143907	0.000000
70%	7.830608	211.783235	25780.059077	7.872038	1.000000
max	14.000000	317.338124	56488.672413	13.127000	1.000000

Mediana

ph	7.027297
Hardness	197.191839
Solids	20933.512750
Chloramines	7.143907
Potability	0.000000

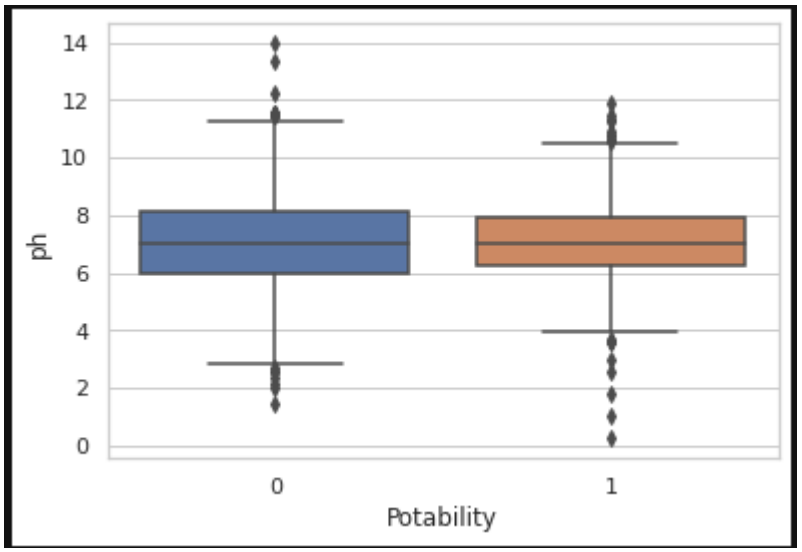
Variância

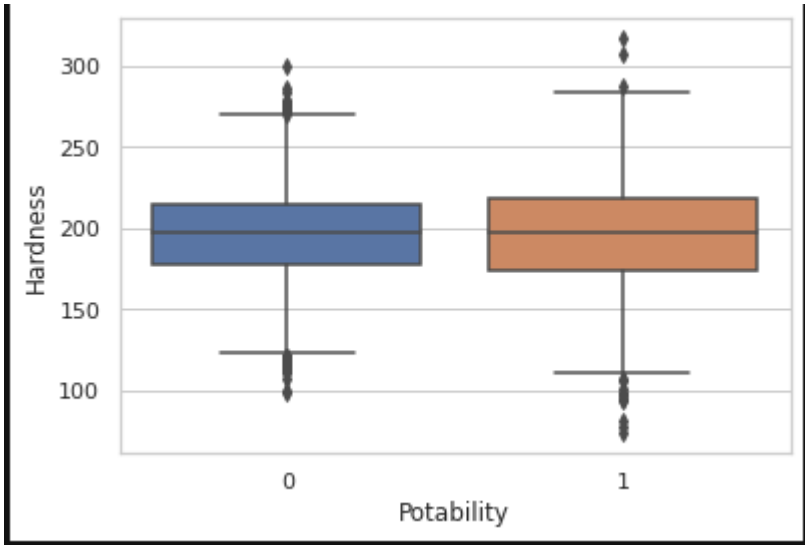
ph	2.475388e+00
Hardness	1.065049e+03
Solids	7.468831e+07
Chloramines	2.511654e+00
Potability	2.407653e-01

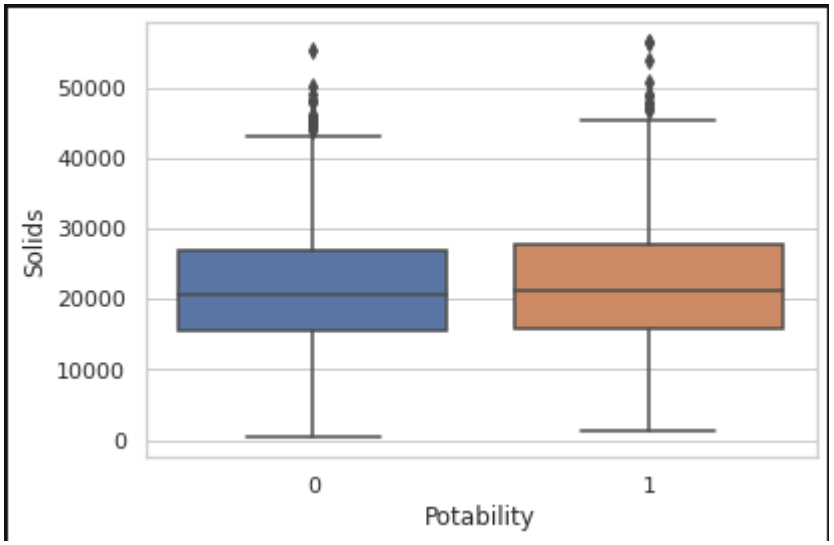
25%	6.089723	176.744938	15615.665390	6.138895	0.000000
50%	7.027297	197.191839	20933.512750	7.143907	0.000000
75%	8.052969	216.441070	27182.587067	8.109726	1.000000

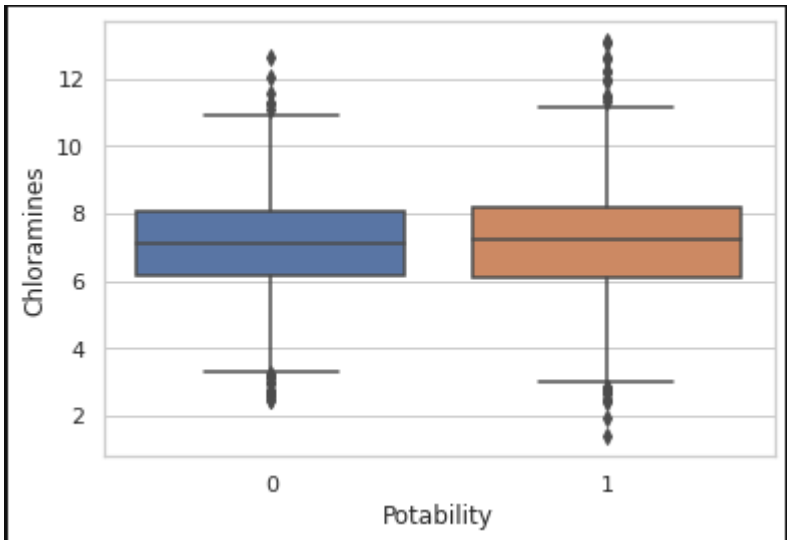
Observando podemos ver que os valores de média e mediana são próximos o que indica dados simétricos algo que também pode ser visto nos boxplots, além disso com os com os percentis e quartis pode ser visto que os dados se reúnem em torno da média e diminuem dos dois lados (valores maiores e menores que a média) quando se afastam dela como em uma distribuição em forma de sino.

3. Geração de boxplots dos atributos selecionados e identifique a presença ou não de outliers.



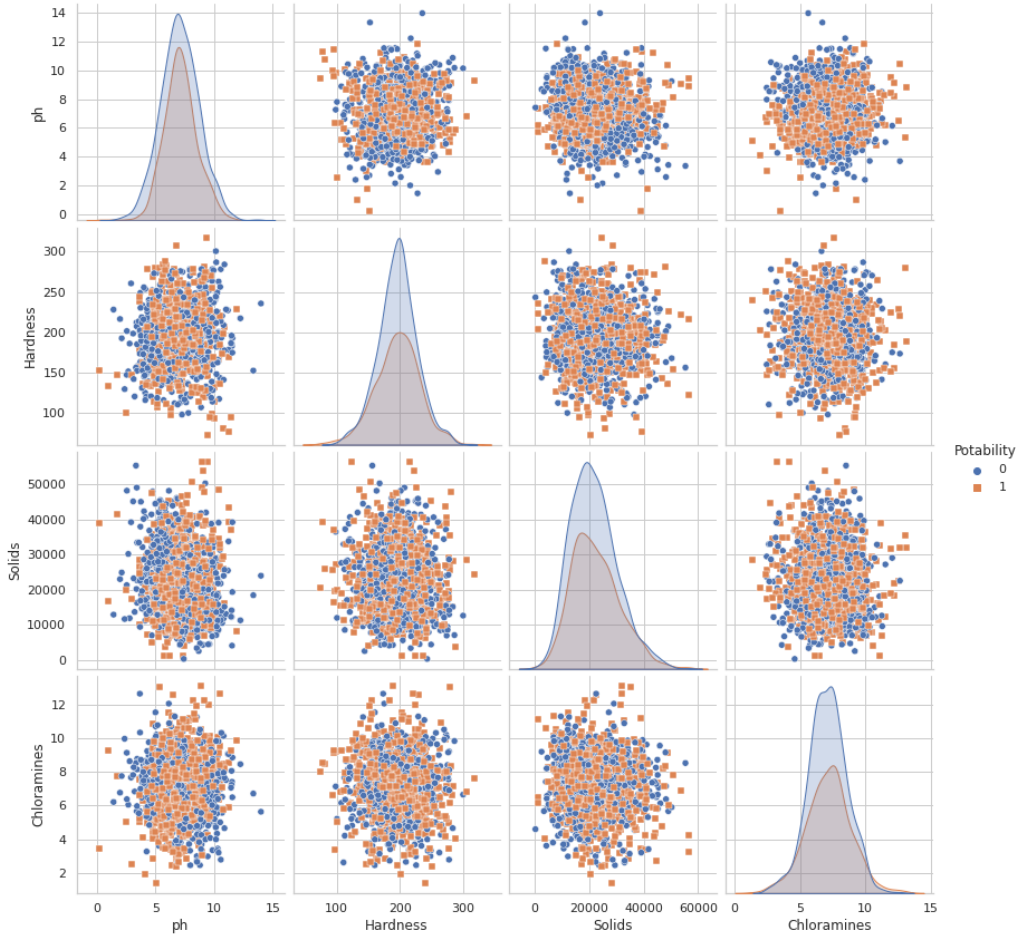

300
250
200
150
100
Hardness
0
1
Potability


50000
40000
30000
20000
10000
0
Solids
0
1
Potability



Outlier é um conceito pouco formal logo é difícil definir o que é e o que não é um outlier e também qual seu “grau de outlier”, mas sendo bem liberal no conceito de outlier e usando os boxplots como base, sim a base tem outliers pois observamos valores bem distantes da média e do ponto de maior concentração de amostras. Observando que no caso de solids os outliers parecem estar presentes apenas nos pontos de maior valor

4. Geração de scatter plots dois a dois dos atributos selecionados



5. Cálculo da matrizes de Covariância e Correlação

Covariância

	ph	Hardness	Solids	Chloramines	Potability
ph	2.475388	5.594047	-1.191314e+03	-0.061759	0.011217
Hardness	5.594047	1065.048742	-1.502397e+04	-1.173283	-0.024100
Solids	-1191.314479	-15023.968352	7.468831e+07	-709.323403	172.481345
Chloramines	-0.061759	-1.173283	-7.093234e+02	2.511654	0.016162
Potability	0.011217	-0.024100	1.724813e+02	0.016162	0.240765

Correlação

	ph	Hardness	Solids	Chloramines	Potability
ph	1.000000	0.108948	-0.087615	-0.024768	0.014530
Hardness	0.108948	1.000000	-0.053269	-0.022685	-0.001505
Solids	-0.087615	-0.053269	1.000000	-0.051789	0.040674
Chloramines	-0.024768	-0.022685	-0.051789	1.000000	0.020784
Potability	0.014530	-0.001505	0.040674	0.020784	1.000000

6. 3 perguntas sobre o conjunto de dados?
    a. Existe forte correlação entre alguma das características com a coluna de labels
        i. não
    b. Os dados estão desbalanceados entre as classes?
        i. Sim, mas não muito
    c. A normalização é necessária nesta base?
        i. Para métodos onde uma grande diferença entre as escalas dos dados é um problema, sim
    d. Há dados faltantes?
        i. Sim. 1265 linha tem dados faltantes
    e. Qual a distribuição dos dados?

i. As features parecem seguir uma distribuição em forma de sino

Referências

[1]http://dainf.pg.utfpr.edu.br/lesic/site/projeto/6

[2]https://www.cortex-intelligence.com/blog/inteligencia-de-mercado/o-que-e-analise-de-dados

[3]https://www.kaggle.com/datasets/adityakadiwal/water-potability

*Além das aulas ministradas na disciplina