

Universidade Federal de Minas Gerais (UFMG)

Departamento de Ciência da Computação

Guilherme Mendes de Oliveira

guilhermemendes@ufmg.br

**CLASSIFICAÇÃO DE POTABILIDADE
DE AMOSTRAS DE ÁGUA**

Belo Horizonte, MG – Brasil

2022

Guilherme Mendes de Oliveira

CLASSIFICAÇÃO DE POTABILIDADE
DE AMOSTRAS DE ÁGUA

**Trabalho prático de classificação da
matéria de Mineração de Dados
ministrada pelo professor Wagner
Meira Jr.**

Belo Horizonte, MG – Brasil

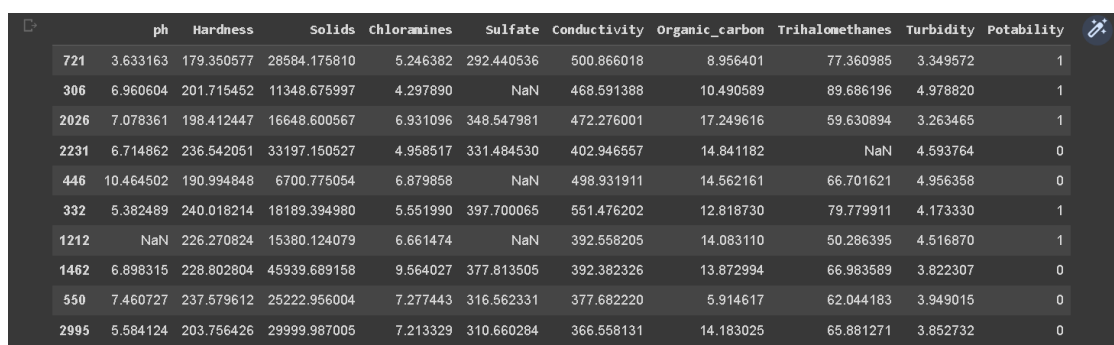
2022

1. Introdução

Este trabalho tem como objetivo a implementação de uma técnica de mineração de dados para classificação em uma base de dados. Para tal atividade foram utilizados os algoritmos *KNN*, *SVM* e *Árvore de Decisão*.

Além disso, a escolha para base de dados foi feita a partir do Kaggle, um portal que disponibiliza diversas bases de dados para a prática de técnicas relacionadas à ciência de dados.

A base apresenta diversos atributos físicos e químicos em diferentes faixas de valores e escalas, juntamente com a classificação da potabilidade da água..



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
721	3.633163	179.350577	28584.175810	5.246382	292.440536	500.866018	8.956401	77.360985	3.349572	1
306	6.960604	201.715452	11348.675997	4.297890	NaN	468.591388	10.490589	89.686196	4.978820	1
2026	7.078361	198.412447	16648.600567	6.931096	348.547981	472.276001	17.249616	59.630894	3.263465	1
2231	6.714862	236.542051	33197.150527	4.958517	331.484530	402.946557	14.841182	NaN	4.593764	0
446	10.464502	190.994848	6700.775054	6.879858	NaN	498.931911	14.562161	66.701621	4.956358	0
332	5.382489	240.018214	18189.394980	5.551990	397.700065	551.476202	12.818730	79.779911	4.173330	1
1212	NaN	226.270824	15380.124079	6.661474	NaN	392.558205	14.083110	50.286395	4.516870	1
1462	6.898315	228.802804	45939.689158	9.564027	377.813505	392.382326	13.872994	66.983589	3.822307	0
550	7.460727	237.579612	25222.956004	7.277443	316.562331	377.682220	5.914617	62.044183	3.949015	0
2995	5.584124	203.756426	29999.987005	7.213329	310.660284	366.558131	14.183025	65.881271	3.852732	0

Figura 1 – Amostra dos dados da base de amostras de água.

2. Motivação/ justificativa

A aplicação das técnicas de modelagem envolvidas no trabalho juntamente com a utilização de um algoritmo em uma base de dados que se aproxima de problemas reais encontrados no cotidiano é além de uma forma de avaliação uma oportunidade de desenvolver todas as competências técnicas necessárias para aplicar futuramente esse conhecimento no “mundo real” ou seja uma forma de ganhar experiência analítica dos resultados e competências técnicas de programação para exploração, limpeza dos dados , aplicação do algoritmo e extração de métricas para avaliação dos resultados.

3. Objetivo

A aplicação das técnicas de classificação nessa base de dados de amostras de águas, busca avaliar qual dos três algoritmos selecionados terá um melhor desempenho para prever a potabilidade da água.

Esse entendimento por exemplo poderia ser utilizado ainda que de forma genérica para a tomada de decisão na avaliação de estações de tratamento de água sendo uma espécie de controle de qualidade, no entanto por se tratar de uma questão de saúde dizer qual é o limiar

aceitável de erro e as possíveis implicações do erro do modelo principalmente no que diz respeito ao falso positivo, isto é, dizer que uma amostra é potável e na verdade ela não é, é um ponto muito sensível. Portanto, uma avaliação mais criteriosa de um ou mais especialistas de negócio seria fundamental para que o modelo de fato fosse implementado.

4. Metodologia

A metodologia utilizada foi inspirada no CRISP-DM, inspirada porque uma das etapas, mais especificamente a do entendimento do negócio, é realizada de forma simulada uma vez que a proposição de um problema só é possível após uma verificação preliminar da base de dados, ou ela é proposta e partir disso que os dados são buscados. Além disso, ela está direcionada a ser um problema que seja possível explorar e propor soluções com base em uma técnica específica, nesse caso a classificação.

Além disso, a etapa de *deployment* também não é pertinente uma vez que é uma prática e nesse caso não há um problema real a ser tratado, para as demais etapas todas foram possíveis e foram aplicadas para a conclusão do trabalho prático.

O fluxo de entendimento, preparação, modelagem e avaliação podem ser vistos [aqui](#).

5. Desenvolvimento

5.1. Business Understanding

Essa etapa compreende o entendimento do problema, como foi propositivo a partir da base de dados e já direcionado para a aplicação de classificação esta etapa não se aplica.

5.2. Data Understanding

A base utilizada possui um total de 3276 linhas, 10 colunas e possui amostras com valores nulos em alguns de seus atributos, em uma quantidade que impacta o prosseguimento do trabalho caso fossem expurgados na próxima etapa. Portanto, no entendimento da base de dados verificamos que a mesma está apta para a aplicação da técnica de mineração uma vez que possui uma quantidade considerável de registros, mas que será necessário atuar nos casos de dados faltantes e também avaliar e tratar as diferentes escalas de medição dos atributos.

```
dfAguas.isnull().sum()
ph          491
Hardness    0
Solids       0
Chloramines  0
Sulfate     781
Conductivity 0
Organic_carbon 0
Trihalomethanes 162
Turbidity   0
Potability  0
dtype: int64
```

Figura 2 – Atributos com valores nulos.

Com o objetivo de entender se a distribuição de atributos pode ocasionar algum viés dados alguns valores foi realizado uma plotagem do histograma de cada atributo estratificado pela potabilidade da água. A distribuição dos valores não indica viés uma vez que não há faixas dominantes, intervalos ausentes ou muito concentrados, a princípio as amostras parecem muito heterogêneas.

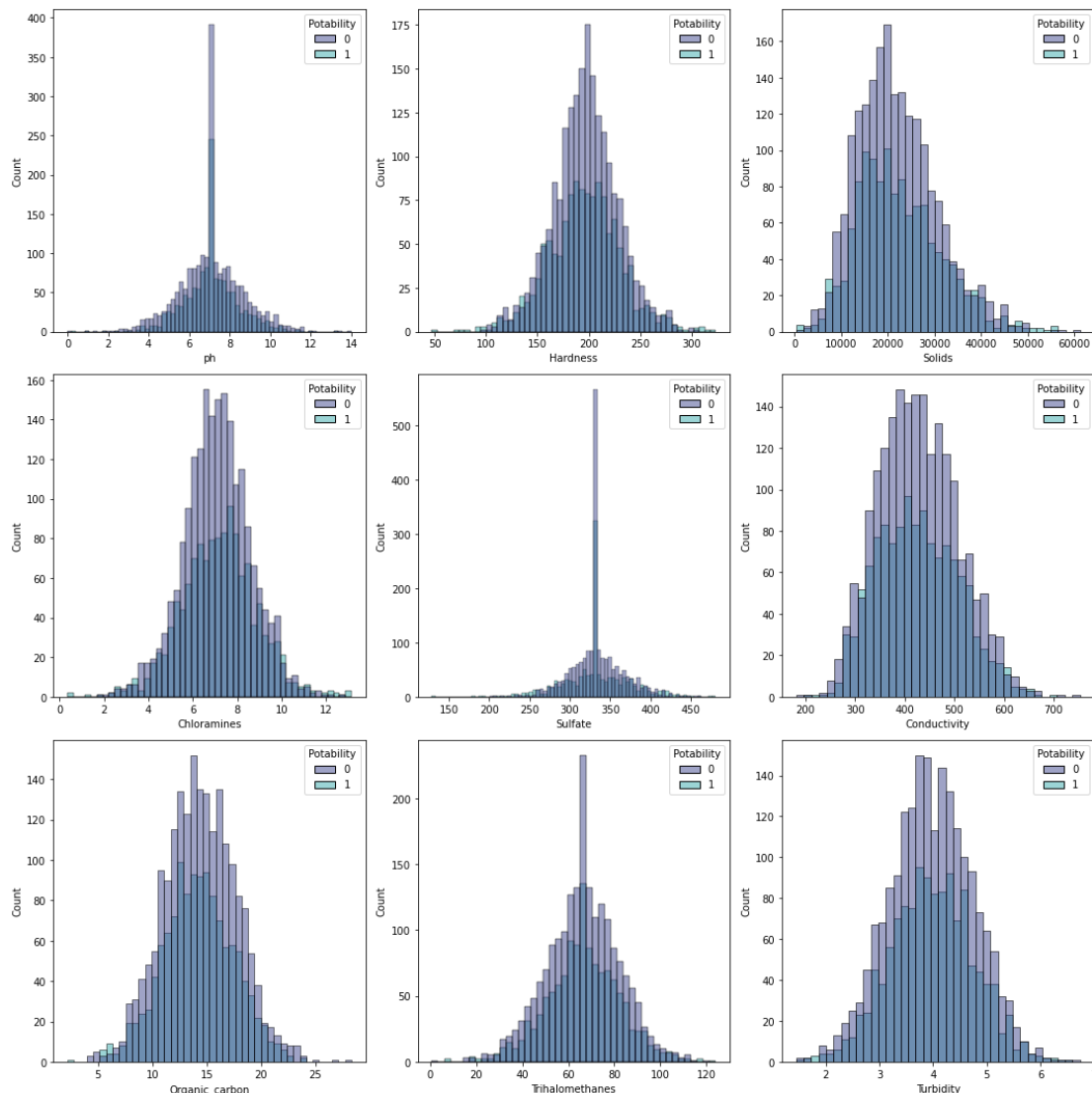


Figura 3 – Histograma dos atributos estratificados pela potabilidade da água.

Com o objetivo de entender se há possibilidade de reduzir a dimensionalidade, a matriz de correlação foi plotada também, no entanto a correlação dos atributos uns com os outros é fraca e não indica que é possível suprimir algum atributo a ser passado para o modelo.

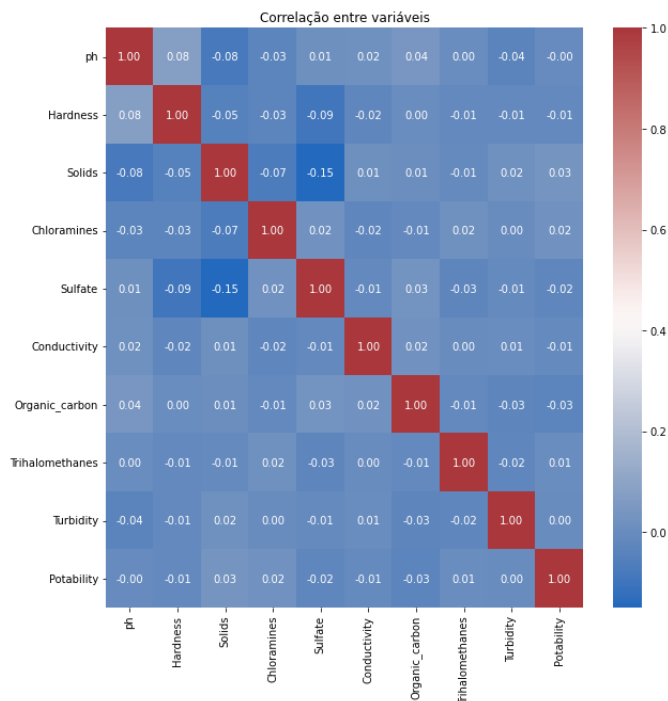


Figura 4 – Correlação entre os atributos da base de dados.

5.3. Data Preparation

Na etapa de preparação inicialmente foi feito o tratamento dos registros que não tem os valores dos atributos que serão utilizados para a aplicação das técnicas de classificação. Para tal etapa os dados nulos foram substituídos pela mediana de cada atributo.

```
#Eliminando as linhas com valores nulos impacta muito a quantidade de registros portanto substituindo os valores pela mediana
dfAguas['sulfate'] = dfAguas['sulfate'].fillna(value=dfAguas['sulfate'].median())
dfAguas['ph'] = dfAguas['ph'].fillna(value=dfAguas['ph'].median())
dfAguas['Trihalomethanes'] = dfAguas['Trihalomethanes'].fillna(value=dfAguas['Trihalomethanes'].median())
```

Figura 5 – Processo de tratamento de valores nulos das amostras.

5.4 Modelagem

Na etapa de modelagem da base para a aplicação dos algoritmo de classificação , foi realizado o procedimento de Normalização nos campos numéricos a fim de tratar problemas

relacionados a escala das ordens de grandeza dos valores que serão utilizados. Após a etapa de normalização dos atributos a base foi dividida para que fosse realizado o treino de cada algoritmo, para tal a base de treino foi de 75% e a de teste 25%

Para a calibragem do número de vizinho do KNN foram avaliadas a acurácia e a precisão do teste para uma quantidade de 1 a 15. Escolhido como parâmetro 11, pois foi o valor que apresentou a melhor acurácia (índice de acerto) e uma precisão (índice de pureza) bem próxima ao máximo obtido para todos os outros valores do parâmetro.

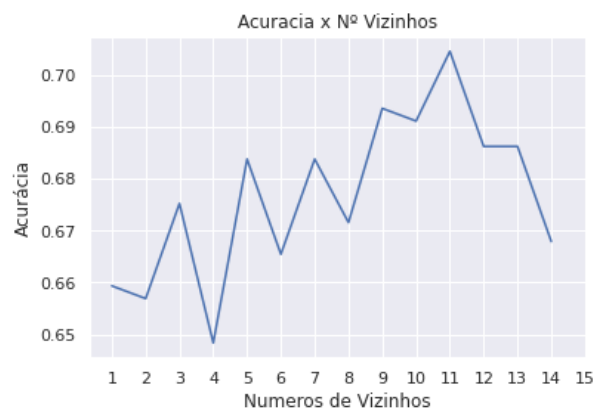


Figura 6 – Gráfico de acurácia x número de vizinhos testados.

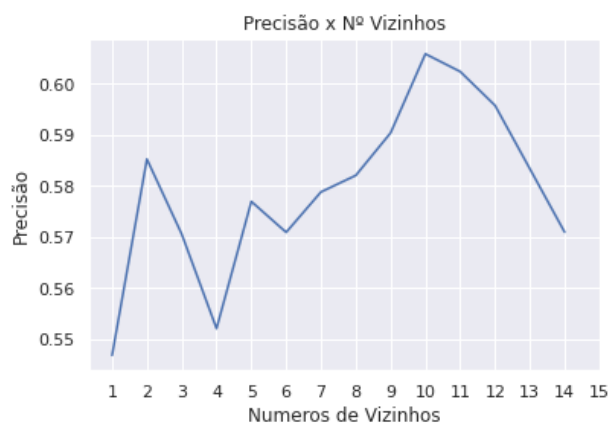


Figura 7 – Gráfico de precisão x número de vizinhos testados.

Para os demais classificadores, por falta de conhecimento aprofundado do comportamento da troca de parâmetros, não foi feita nenhuma mudança e todos foram utilizados com seus respectivos parâmetros padrões.

6. Resultados experimentais e análise

6.1. Evaluation

Foi considerada a ideia de utilizar validação cruzada para avaliação de overfit, no entanto não houve sucesso na implementação em tempo hábil para a entrega do trabalho, apenas o SVM cujo parâmetro que habilita a melhor medição da curva ROC realiza internamente para um valor de 5 subconjuntos no treino.

De modo geral os três classificadores apresentam um comportamento bem próximo, com algumas peculiaridades:

Em relação à acurácia que indica o índice de acertos na classificação em relação ao total de amostras o KNN apresenta um percentual de 70%, o SVM 75% e a Árvore de Decisão 70%.

Em relação à precisão que indica o quanto o modelo acertou na classificação de amostras em termos de classificações corretas frente ao total da classe. Na média o KNN apresenta um valor de 69%, o SVM um valor de 78% e a Árvore de Decisão 70%. Vale ressaltar que a precisão do SVM para a classe de Não Potável tem um desempenho bem expressivo 95%, isto é, o modelo consegue classificar corretamente 95% das amostras que não são próprias para consumo.

Em relação à revocação que indica o quanto o modelo captura das classificações corretas de cada classe. Na média o KNN apresenta um valor de 69%, o SVM 79% e a Árvore de Decisão 69%

	precision	recall	f1-score	support
0	0.78	0.74	0.76	508
1	0.60	0.65	0.63	311
accuracy			0.70	819
macro avg	0.69	0.69	0.69	819
weighted avg	0.71	0.70	0.71	819

Figura 8 – Avaliação do KNN.

	precision	recall	f1-score	support
0	0.95	0.63	0.76	599
1	0.62	0.95	0.75	384
accuracy			0.75	983
macro avg	0.78	0.79	0.75	983
weighted avg	0.82	0.75	0.75	983

Figura 9 – Avaliação do SVM.

	precision	recall	f1-score	support
0	0.76	0.76	0.76	599
1	0.62	0.62	0.62	384
accuracy			0.70	983
macro avg	0.69	0.69	0.69	983
weighted avg	0.70	0.70	0.70	983

Figura 10 – Avaliação da Árvore de Decisão.

Verificando a análise ROC e a AUC que especifica o quanto o modelo é bom em distinguir amostras em potáveis (1) e não potáveis (0) através da plotagem da taxa de verdadeiros positivos em relação aos falsos positivos para determinados valores de limiares. O KNN apresenta um valor de AUC 0.6945, o SVM 0.7967 e a Árvore de Decisão 0.6910.

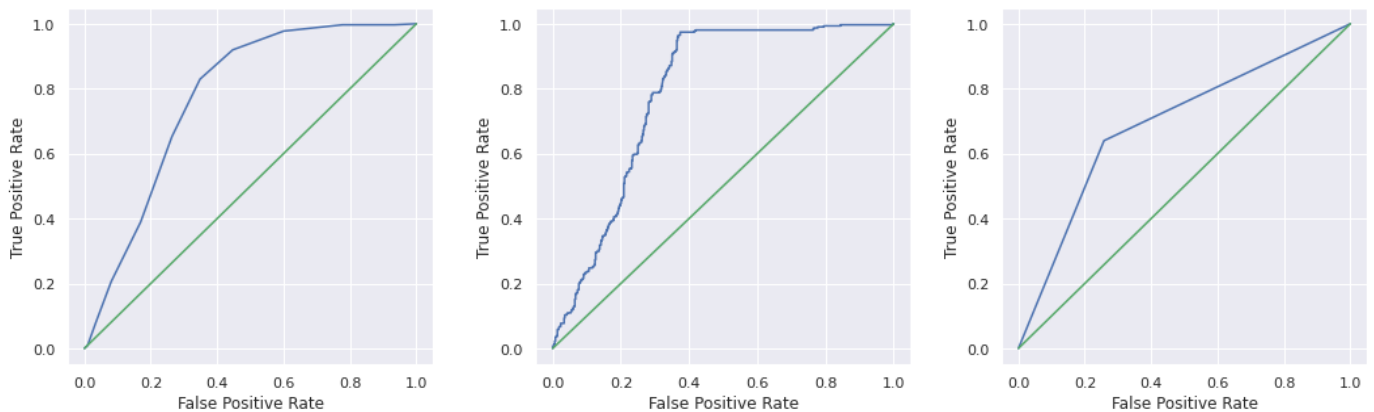


Figura 11 – Curva ROC do KNN, SVM e Árvore de Decisão respectivamente.

Portanto o SVM apresenta uma melhor capacidade de classificação para esta base analisando a AUC juntamente com o desempenho do algoritmo nas outras métricas avaliadas.

7. Conclusões e perspectivas

O resultado do trabalho prático foi bem satisfatório, foi possível explorar algumas das técnicas de tratamento e preparação de dados, calibração do parâmetro k que diz respeito ao número de vizinhos do KNN, experimentar a implementação de três algoritmos de classificação além de práticas de métricas de avaliação a forma de analisá-las.

8.Referências

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 18: Probabilistic Classification;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 19: Support Vector Machines;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 21: Decision Tree Classifier;

Zaki, M and Meira, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Capítulo 22: Classification Assessment;

<https://scikit-learn.org/stable/supervised_learning.html> . Acesso em: 20 de Novembro de 2022.

<<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>> . Acesso em: 25 de novembro de 2022.