FACULDADE DE COIMBRA
[LICENCIATURA EM
ENGENHARIAE CIÊNCIA DE DADOS]

RELATÓRIO PREVISÃO DA QUALIDADE DA ÁGUA

REALIZADO POR:
CATARINA PIMENTA SIMÕES

DEPARTAMENTODE ENGENHARIA
INFORMATICA

Índice

Conteúdo

Descrição do Dataset

[1] A acessibilidade à água potável é essencial para a saúde e um direito humano básico, sendo muito importante como uma questão de saúde e desenvolvimento a nível nacional, regional e local. Em algumas regiões, foi demonstrado que os investimentos em abastecimentos de água e saneamento podem gerar um benefício económico, visto que as reduções nos efeitos adversos à saúde e nos custos de saúde superam os custos de realização das intervenções.

No ficheiro water_potability.csv contém métricas acerca da qualidade da água para 3276 amostras de água diferentes:

Valor pH: O pH é um parâmetro importante na avaliação do equilíbrio ácido-base da água. É também o indicador da condição ácida ou alcalina da água. A OMS recomendou que o limite máximo permitido de pH fosse 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

Dureza: A dureza é principalmente causada 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 contacto 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 pelo cálcio e magnésio.

Sólidos (total de sólidos dissolvidos - TDS): A água tem a capacidade de dissolver uma ampla gama de minerais ou sais inorgânicos e alguns orgânicos, tais como potássio, cálcio, sódio, bicarbonatos, cloretos, magnésio, sulfatos, etc. Esses minerais produzem um 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 é altamente mineralizada. O limite desejável para TDS é de 500 mg/l e o limite máximo é de 1000 mg/l prescrito para beber.

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.

Sulfato: Os sulfatos são substâncias naturais encontradas em minerais, solo e rochas. Eles estão presentes no ar ambiente, águas subterrâneas, plantas e alimentos. O principal uso comercial do sulfato é na indústria química. A concentração de sulfato na água do mar é de cerca de 2.700 miligramas por litro (mg/L). Varia de 3 a 30 mg/L na maioria dos suprimentos de água doce, embora concentrações muito mais altas (1000 mg/L) sejam encontradas em algumas localizações geográficas.

Condutividade: A água pura não é um bom condutor de corrente elétrica, mas sim um bom isolante. O aumento da concentração de iões aumenta a condutividade elétrica da água. Geralmente, a quantidade de sólidos dissolvidos na água determina a condutividade elétrica. A condutividade elétrica (EC) realmente mede o processo iônico de uma solução que lhe

permite transmitir corrente. De acordo com os padrões da OMS, o valor EC não deve exceder 400 µS/cm.

Carbono orgânico: O Carbono Orgânico Total (TOC) nas águas de origem vem da decomposição da matéria orgânica natural (NOM), bem como de fontes sintéticas. TOC é uma medida da quantidade total de carbono em compostos orgânicos em água pura. De acordo com a US EPA < 2 mg/L como TOC em água tratada/potável, e < 4 mg/Lit em fontes de água e que é usada para tratamento.

Trihalometanos: THMs são produtos químicos que podem ser encontrados em água tratada com cloro. A concentração de THMs na água potável varia de acordo com o nível de matéria orgânica na água, a quantidade de cloro necessária para tratar a água e a temperatura da água que está a ser tratada. Níveis de THM de até 80 ppm são considerados seguros na água potável.

Turbidez: A turbidez da água depende da quantidade de matéria sólida presente no estado suspenso. É uma medida das propriedades de emissão de luz da água e o teste é usado para indicar a qualidade da descarga de resíduos em relação à matéria coloidal. O valor médio de turbidez obtido para Wondo Genet Campus (0,98 NTU) é inferior ao valor recomendado pela OMS de 5,00 NTU

Potabilidade: Indica se a água é segura para consumo humano onde 1 significa Potável e 0 significa Não potável.

Exploração dos Dados

Inicialmente após realizar todos os *imports* das bibliotecas necessárias para este projeto, visualizei os dados existentes e calculei alguns dados estatísticos:

Figura 1. Dados do Dataset

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	nan	204.890455471...	20791.3189807...	7.30021187318...	368.516441349...	564.308654172...	10.3797830780...	86.9909704615...	2.96313538063...	0
1	3.71608007538...	129.422920514...	18630.0578579...	6.635245883862	nan	592.885359134...	15.1800131163...	56.3290762845...	4.50065627494...	0
2	8.09912418929...	224.236259393...	19909.5417322...	9.27588360269...	nan	418.606213064...	16.8686369295...	66.4200925117...	3.05593374966...	0
3	8.31676588421...	214.373394085...	22018.4174407...	8.05933237743...	356.886135643...	363.266516164...	18.4365244954...	100.341674365...	4.62877053683...	0
4	9.09222345629...	181.101509236...	17978.9863389...	6.54659997420...	310.135737524...	398.410813381...	11.5582794434...	31.9979927274...	4.07507542543...	0
5	5.58408663845...	188.313323769...	28748.6877390...	7.54486878877...	326.678362911...	280.467915933...	8.39973464015...	54.9178618419...	2.55970822755...	0
6	10.2238621645...	248.071735270...	28749.7165435...	7.51340846583...	393.663395515...	283.651633507...	13.7896953175...	84.6035561740...	2.67298873693...	0
7	8.63584871850...	203.361522584...	13672.0917639...	4.56300868559...	303.309771159...	474.607644942...	12.3638166987...	62.7983089629...	4.40142471544...	0
8	nan	118.988579090...	14285.5838542...	7.80417355307...	268.646940746...	389.375565871...	12.7060489686...	53.9288457675...	3.59501718095...	0
9	11.1802844707...	227.231469237...	25484.5084909...	9.07720001691...	404.041634684...	563.885481481...	17.9278064112...	71.9766010322...	4.37056193665...	0
10	7.36064010583...	165.520797259...	32452.6144091...	7.55070090670...	326.624353455...	425.383419495...	15.5868104380...	78.7400156643...	3.66229178285...	0

3276 rows × 10 columns

Figura 2. Dados estatísticos do Dataset

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.0	3276.0	3276.0	3276.0	2495.0	3276.0	3276.0	3114.0	3276.0	3276.0
mean	7.08079450427...	196.369496017...	22014.0925260...	7.12227679342...	333.775776610...	426.205110682...	14.2849702476...	66.3962929467...	3.96678616979...	0.3901098901
std	1.59431951870...	32.8797614762...	8768.57082778...	1.58308488903...	41.4168404616...	80.8240640511...	3.30816199912...	16.1750084222...	0.78038240848...	0.4878491696
min	0.0	47.432	320.942611274...	0.35200000000...	129.000000000...	181.483753985...	2.19999999999...	0.73799999999...	1.45	0.0
25%	6.09309191422...	176.850537877...	15666.6902969...	6.12742075549...	307.699497834...	365.734414118...	12.0658013336...	55.8445356209...	3.43971086961...	0.0
50%	7.03675210383...	196.967626863...	20927.8336065...	7.13029897388...	333.073545745...	421.884968280...	14.2183379372...	66.6224850980...	3.95502756299...	0.0
75%	8.06206612314...	216.667456214...	27332.7621274...	8.11488703210...	359.950170384...	481.792304487...	16.5576515438...	77.3374729087...	4.50031978728...	1.0
max	13.9999999999...	323.124	61227.1960077...	13.1270000000...	481.030642305...	753.342619558...	28.3000000000...	124.0	6.739	1.0

8 rows × 10 columns

Na Figura 2 podemos visualizar por exemplo a quantidade de dados em cada uma das colunas, concluindo que estes não são iguais em todas provando que existem valores em falta. Porém conseguimos também observar a média dos seus valores, desvios padrões, valores máximo, etc. A Figura 1 para além de nos mostrar os dados em detalhe mostra-nos as colunas onde existem os valores em falta podendo assim começar a tratá-los mais tarde.

Nesta etapa estudei mais a fundo a distribuição da água (potável/não potável) no Dataset e em cada coluna. No último referido utilizei o método *kdeplot* da biblioteca *seaborn* que consistenum gráfico de estimativa de densidade do *kernel* (KDE). Um método para visualizar a distribuição de observações num conjunto de dados, análogo a um histograma. O KDE representa os dados usando uma curva de densidade de probabilidade contínua em uma ou mais dimensões [2] :

Figura 3. Distribuição da água no Dataset

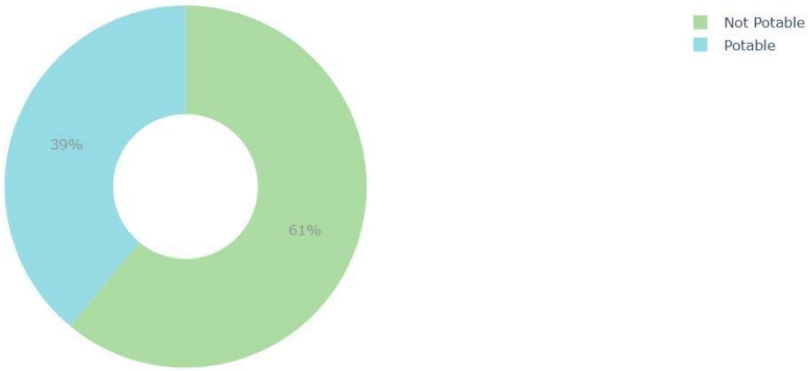
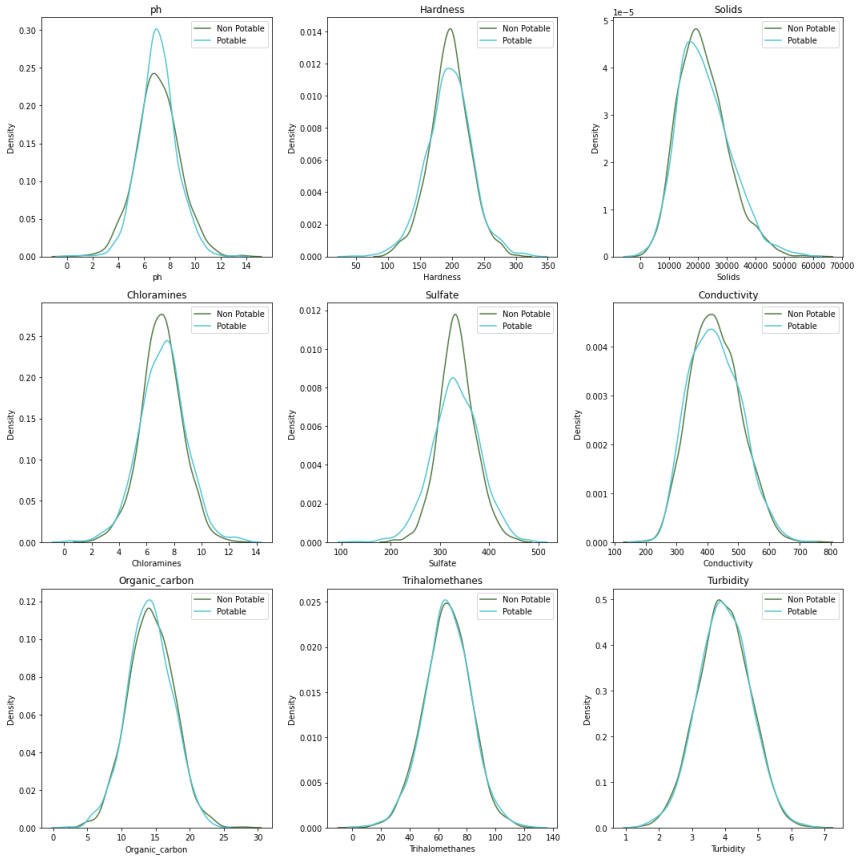


Figura 4. Distribuição da água em cada coluna



Podemos assim concluir que no Dataset existe 39% de água potável e 61% de água não potável. Na figura 4 podemos observar que a água potável e a água não potável na estimativa de densidade évisualmente distinguível onde por exemplo no ph a água não potável alcança valores de densidade menores.

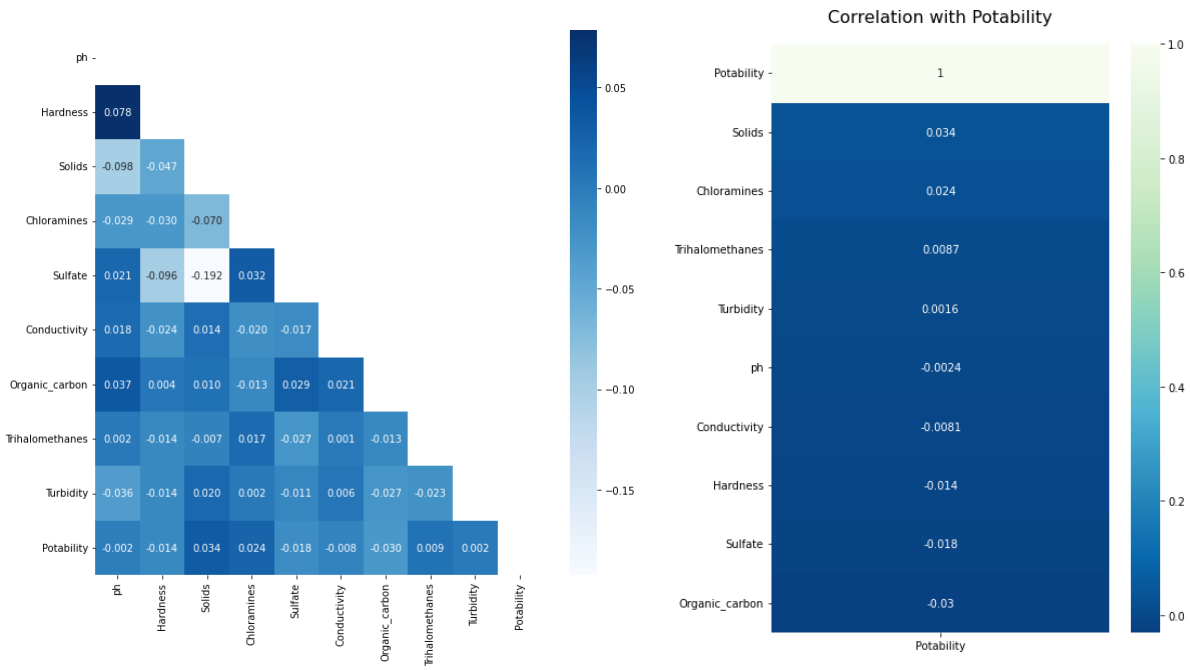
- **Correlação dos dados**

Os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam.

O que a correlação procura entender é como uma variável se comporta num cenário onde outravaria, de forma a identificar se existe alguma relação entre a variabilidade de ambas. Embora nãoimplique em causalidade, o coeficiente de correlação exprime em números essa relação, ou seja,quantifica a relação entre as variáveis.

Neste Dataset como podemos observar na figura abaixo não existem dados muito relacionados mantendo todos um valor muito próximo de 0 como tal, optei por não retirar nenhuma variável:

Figura 5. Correlações



Pré processamento

- **Valores em falta**

Anteriormente descobri a existência de valores em falta no nosso Dataset, posto isto, analisei em precisão as colunas onde se encontram e a quantidade de valores nulos existentes nelas:

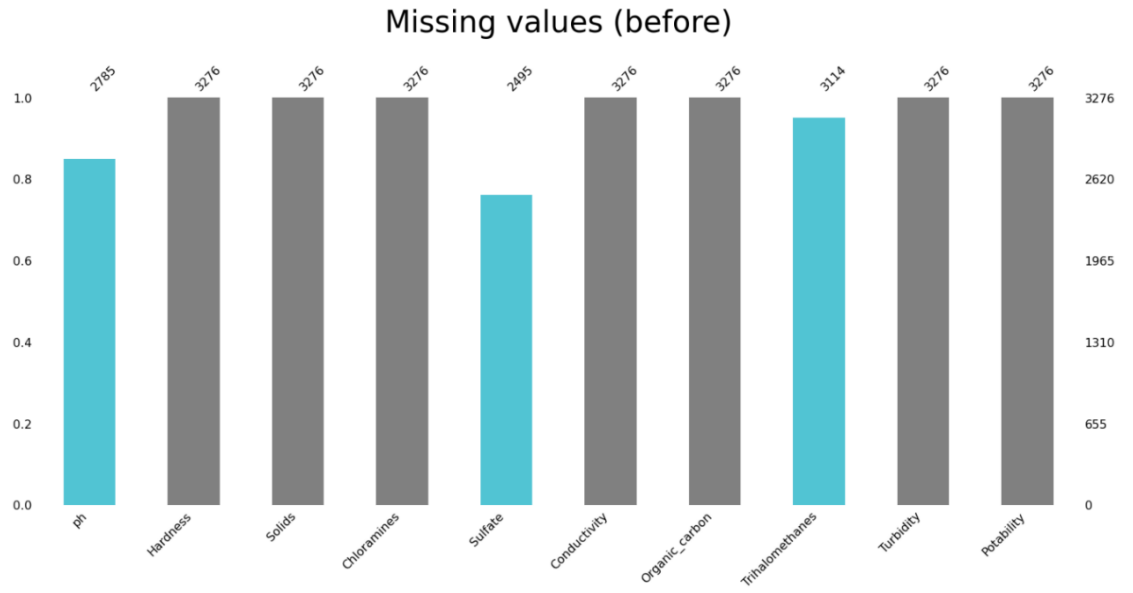
Figura 6. Especificação dos valores em falta

ph	491
Hard...	0
Solids	0
Chlor...	0
Sulfate	781
Cond...	0
Orga...	0
Trihal...	162
Turbi...	0
Potab...	0

10 rows × 1 columns

De seguida, ilustrei o observado, representado a azul estão as colunas com valores em falta e a cinzento o oposto. O número representado em cima de cada barra corresponde à quantidade de dados em cada variável:

Figura 7. Ilustração dos dados em falta



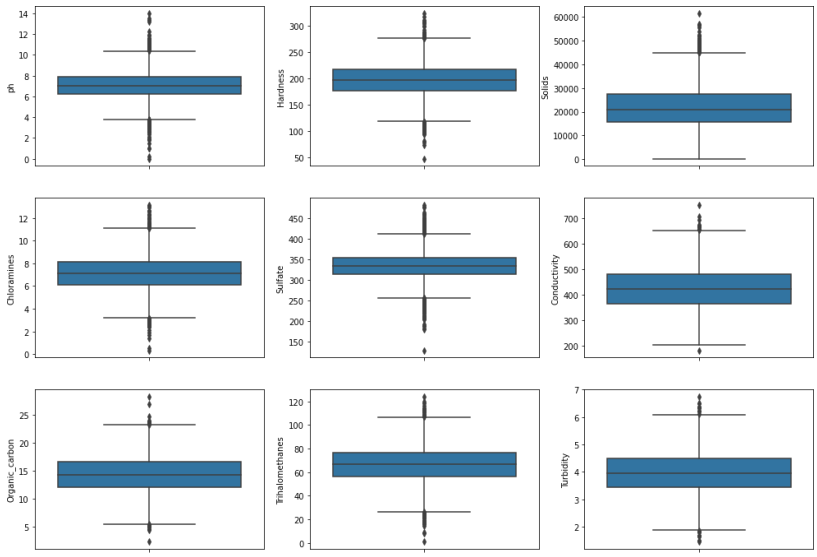
Para tratar estes dados utilizei o método *KNNImputer* da biblioteca *scikit-learn*. Este método identifica os pontos vizinhos por meio de uma medida de distância e os valores ausentes podem ser estimados usando valores completos de observações vizinhas. A ideia deste método é identificar 'k'amostras no conjunto de dados que são semelhantes ou próximas no espaço. Em seguida, usamos essas amostras 'k' para estimar o valor dos pontos de dados ausentes. Os valores ausentes de cada amostrasão colocados através do valor médio dos 'k'-vizinhos encontrados no conjunto de dados. [3]
Neste projeto optei por utilizar k=10 substituindo o resultado do *KNNImputer* no Dataset, ilustrando o resultado:

Figura 8. Dados em falta após KNNImputer

- **Outliers**

Como podemos observar na figura 9 existem Outliers. No entanto, optei por não os retirar visto que estamos a estudar água potável e não potável sendo por exemplo plausível a existência de água não potável com valores de pH superiores a 12 ou inferiores a 4.

Figura 9. Outliers



Modelos

Nesta etapa comecei por dividir os nossos dados em *X_train*, *X_test*, *y_train* e *y_test* deixando 30% dos dados para teste (*X_test* e *y_test*). De seguida apresentei o alcance decada variável de forma a perceber se será necessário escalar os dados.

Através da figura 10 concluí que o alcance para asdiferentes variáveis é bastante distinto, como por exemploao alcance dos sólidos ao do ph. Sendo assim optei por escalar os *X_train* e o *X_test* utilizando o *Standard Scaler*, vistoque as colunas são padronizadas removendo a média e

Figura 10. Alcance das variáveis

	Name	Range
0	ph	0 to 14
1	Hardness	47 to 324
2	Solids	320 to 61228
3	Chloramines	0 to 14
4	Sulfate	129 to 482
5	Conductivity	181 to 754
6	Organic_carbon	2 to 29
7	Trihalomethanes	0 to 124
8	Turbidity	1 to 7

escalando a variância a uma unidade , tornando-as mais manejáveis para os nossos modelos.[4]

Feito isto, segui para a realização dos modelos

- **SVC**

As máquinas de vetores de suporte (*SVMs*) são um conjunto de métodos de aprendizagem supervisionado usados para classificação, regressão e deteção de Outliers. Algumas vantagens das *SVMs* é que são eficazes em espaços de alta dimensão e em casos em que o número de dimensõesé maior que o número de amostras, usa um subconjunto de pontos de treinamento na função dedecisão (chamados vetores de suporte), portanto, também é eficiente em termos de memória e éversátil pois diferentes funções do *Kernel* podem ser especificadas para a função de decisão. Sãofornecidos *Kernels* comuns, mas também é possível especificar *Kernels* personalizados.[5]

Ao realizar este modelo optei por utilizar um método *GridSearchCV*, que consiste numa pesquisa exaustiva sobre os valores de parâmetros especificados para um estimador. Assim sendopara este modelo especifiquei os parâmetros de C e Kernel obtendo por fim uma *accuracy* de 0.674.

Por fim optei por recorrer ao *cross validation*, um método de reamostragem que usa diferentes partes dos dados para testar e treinar um modelo em diferentes iterações, utilizando os melhores parâmetros resultantes da *GridSearchCV*, acabando com uma *accuracy* de 0.61.

Figura 11. Resultados do SVC com GridSearchCV e Cross-validation

```
{'C': 1.0, 'kernel': 'rbf'}
Test accuracy: 0.674
              precision    recall  f1-score   support

         0.0       0.67      0.93      0.78       603
         1.0       0.71      0.27      0.39       380

    accuracy                           0.67       983
   macro avg       0.69      0.60      0.58       983
weighted avg       0.68      0.67      0.63       983
```

```
[0.6097561  0.61068702 0.61068702 0.60916031 0.60916031]
0.61 accuracy with a standard deviation of 0.001
```

- **MLP Classifier**

MLPClassifier significa classificador *Multi-layer Perceptron*. Ao contrário de outros algoritmos de classificação, como Vetores de Suporte ou Classificador *Naive Bayes*, o *MLPClassifier* depende de uma Rede Neural subjacente para realizar a tarefa de classificação.[6]

Ao realizar este modelo optei por utilizar um método GridSearchCV tal como no modelo anterior. No entanto, especifiquei os parâmetros de *Hidden_Layer_Sizes*, *Activation*, *Alpha* e *Learning_Rate*, obtendo por fim uma accuracy de 0.677.

Por fim optei novamente por realizar o cross validation com os melhores parâmetros resultantes da GridSearchCV, acabando com uma accuracy de 0.52.

Figura 12. Resultados do MLP com GridSearchCV e Cross-validation

```
{'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes':
Test accuracy: 0.677
              precision    recall  f1-score   support

         0.0       0.70      0.83      0.76       603
         1.0       0.62      0.43      0.51       380

    accuracy                           0.68       983
   macro avg       0.66      0.63      0.63       983
weighted avg       0.67      0.68      0.66       983
```

```
[0.43445122 0.60763359 0.39541985 0.58015267 0.56946565]
0.52 accuracy with a standard deviation of 0.09
```

- **Random Forest Classifier**

Por fim optámos por usar o *Random Forest Classifier*, generalização da operação Árvore de Decisão, em que se utiliza um conjunto de árvores de decisão (aleatórias) a fim de minimizar o sobreajuste(overfitting) de cada modelo individual de árvore gerado para os dados de entrada.[7] Usei este modelo pois fornece maior precisão por meio de cross validation e manipula os valores em faltamantendo a precisão de uma grande proporção de dados.

Tal como nos modelos anteriores recorri ao GridSearchCV e ao cross validation. No primeiro especifiquei os parâmetros *Max_Features* e *Max_Depth*. Por fim obtive a accuracy de 0.678 e 0.64 respetivamente.

Figura 13. Resultados do Random Forest com GridSearchCV e Cross-validation

```
{'max_depth': 10, 'max_features': 6}
Test accuracy: 0.678
              precision    recall  f1-score   support

         0.0       0.67      0.92      0.78       603
         1.0       0.70      0.29      0.41       380

    accuracy                           0.68       983
   macro avg       0.69      0.61      0.59       983
weighted avg       0.68      0.68      0.64       983
```

```
[0.6097561  0.64885496 0.65648855 0.61526718 0.66564885]
0.64 accuracy with a standard deviation of 0.02
```

Resultados e Discussão

Existem diversos projetos já realizados com o Dataset que estudei acima, com diversos resultados. Por exemplo no estudo "*Water Potability Analysis*"[8] foram estudados vários modelossendo o melhor deles o SVC que com o GridSearchCV conseguiu alcançar uma accuracy de 69%. Outro estudo "*Water Potability Prediction (Best Accuracy 69.5%)"* [9] alcançou uma accuracy de 69,5% com o Random Forest Classifier.

No projeto a melhor accuracy conseguida foi no modelo Random Forest com uma accuracy de 0.678, podendo assim concluir que dentro dos trabalhos existentes o resultado foi dentro do esperado.

Referências

[1] Water Quality Dataset de https://www.kaggle.com/datasets/adityakadiwal/water-potability?datasetId=1292407&searchQuery=mlp

[2] Seaborn.kdeplot de https://seaborn.pydata.org/generated/seaborn.kdeplot.html

[3] KNNImputer: A robust way to impute missing values (using Scikit-Learn) de https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/

[4] Técnicas Standardscaler, Minmaxscaler E Robustscaler – ML de https://acervolima.com/tecnicas-standardscaler-minmaxscaler-e-robustscaler-ml/

[5] Support Vector Machines de https://scikit-learn.org/stable/modules/svm.html

[6] A Beginner's Guide To Scikit-Learn's MLPClassifier de https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/

[7] Random Forest de https://docs.lemonade.org.br/pt-br/spark/aprendizado-de-maquina/classificacao-random-forest.html

[8] Water Potability Analysis de https://www.kaggle.com/code/neesha12/water-potability-analysis

[9] Water Potability Prediction (Best Accuracy 69.5%) de https://www.kaggle.com/code/sinansaglam/water-potability-prediction-best-accuracy-69-5/notebook