VETOR

# Analisando Dados de Qualidade do Ar por Machine Learning☆

## Analyzing Air Quality Data by Machine Learning

Ardson dos Santos Vianna Jr.[1,†], Fernando de Come[1]

[1]*Departamento de Engenharia Química, Escola Politécnica, Universidade de São Paulo, São Paulo, Brasil*
[†]**Autor correspondente:** ardson@usp.br

**Resumo**

A aprendizagem de máquina (Machine Learning - ML) possibilita a análise de grande quantidade de dados. Estes dados são gerados a todo instante, como dados de consumo, saúde pública e processos industriais. Um exemplo destes *datasets* são os conjuntos de parâmetros gerados no monitoramento da qualidade do ar. Neste trabalho, são utilizadas ferramentas de ML para avaliar a qualidade do ar em uma estação da CETESB em Cubatão, estação 66 - Parisi. Os dados de um ano, 1/1/2022 a 1/1/2023, para Partículas Inaláveis ($MP_{10}$), óxidos de nitrogênio (NO, $NO_2$ e $NO_x$) e $SO_2$ foram considerados. Foram desenvolvidas engenharia de atributos, agrupamento e classificação. O dendrograma indicou quatro clusters, o que foi confirmado pelo método do K-médios. O classificador com melhor performance foi o algoritmo de k vizinhos mais próximos, com um coeficiente igual a 0,953138. As ferramentas resultaram em análises interessantes que contribuem para o melhor controle sobre os poluentes na atmosfera. O cuidado com o meio ambiente deve ser uma atitude de todos. Mesmo pequenas iniciativas podem contribuir para criar movimentos e políticas públicas.

**Palavras-chave**

Qualidade do ar • Agrupamento • Classificação

**Abstract**

Machine learning (ML) allows for the continuous analysis of large volumes of data, including information on consumption, public health, and industrial processes. One example of such datasets is the parameters produced by air quality monitoring. This study utilized ML tools to assess air quality at CETESB station 66 - Parisi in Cubatão, São Paulo, Brazil. Data from a one-year period, from 1/1/2022 to 1/1/2023, for Inhalable Particulate Matter (PM10), nitrogen oxides (NO, $NO_2$, and NOx), and $SO_2$ were examined. Feature engineering, clustering, and classification were conducted, resulting in valuable analyses that improve pollutant control in the atmosphere. The dendrogram indicated the presence of four clusters, which was confirmed by the K-mean method. The k nearest neighbor algorithm emerged as the classifier with the best performance, with a coefficient of 0.953138. Protecting the environment should be a collective responsibility; even small initiatives can significantly contribute to movements and public policies.

**Keywords**

Air quality • Clustering • Classification

☆ Este artigo é uma versão estendida do trabalho apresentado no XXVII ENMC Encontro Nacional de Modelagem Computacional e XV ECTM Encontro de Ciência e Tecnologia de Materiais, ocorridos em Ilhéus – BA, de 1 a 4 de outubro de 2024.

# 1 Introdução

O mundo passa por grandes transformações devido à quantidade sem precedentes de dados gerados, armazenados e compartilhados. Uma enxurrada de dados foi avaliada [1] por intermédio de aplicações como *big data*, inteligência artificial e aprendizado de máquina. A big data é usada em vários campos como negócios (tomada de decisão baseada em dados), ciência política (discursos no congresso americano), esportes (para identificar *performance* de jogadores), publicidade (Google e Facebook) e saúde pública. Mas também aplicada na Internet das Coisas ou na Internet Industrial, por exemplo, no uso de sensores digitais em equipamentos industriais, indústria de processos químicos.

As tecnologias emergentes têm um impacto significativo na prática da engenharia, especialmente no contexto de planejamento de produção e adaptação às mudanças externas [2,3]. De acordo com Schwab [4], algumas particularidades de tais transformações são: velocidade (transformações não lineares em um mundo interconectado); amplitude, profundidade e impacto nos sistemas (amplo intercâmbio entre países, empresas e sociedade). Possibilitam novas oportunidades para inovação e competitividade em um cenário econômico global interconectado e dinâmico. Exemplos específicos são:

1. Internet das Coisas (IoT): permite a conexão de equipamentos e dispositivos industriais, possibilitando monitoramento em tempo real de variáveis de processo, condições ambientais e desempenho de ativos. Isso ajuda na tomada de decisão baseada em dados precisos e atualizados.
2. Computação em Nuvem: facilita o acesso a recursos computacionais escaláveis e armazenamento de grandes volumes de dados de maneira econômica. Isso é significativo para análises avançadas e modelos de simulação que apoiam o planejamento de produção.
3. Inteligência Artificial (IA): utilizada para prever demandas de mercado, otimizar processos produtivos e identificar padrões ocultos em grandes conjuntos de dados. Algoritmos de IA podem indicar decisões para adaptar rapidamente a mudanças nas condições de mercado.
4. Sistemas Robóticos: automatizam tarefas repetitivas e perigosas, melhorando a eficiência operacional e reduzindo riscos de segurança no ambiente industrial.
5. Impressão 3D: possibilita flexibilidade na prototipagem rápida e fabricação de peças sob demanda, permitindo ajustes ágeis na produção conforme necessário.
6. Análises Avançadas e Big Data: Permitem análises preditivas e prescritivas baseadas em grandes conjuntos de dados históricos e em tempo real, ajudando a antecipar mudanças no mercado, melhorar a eficiência energética e otimizar processos para redução de custos e impacto ambiental.

Especificando em números, foram criados **2,5 quintilhões de bytes de dados diariamente em 2020,** sendo que 1 quintilhão = 1 exabyte = $\mathbf{10^{18}}$ **bytes** [5]. **Em 2020, cada ser humano criou pelo menos 1,7 MB de dados por segundo.** As origens desses dados são diversas: sensores, sites, imagens e vídeos digitais, registros de transações comerciais, sinais de GPS de telefones celulares, etc.

O Aprendizado de Máquina (*Machine Learning*- ML) faz parte de um grande item que é a Ciência de Dados, que combina estatística, métodos científicos, inteligência artificial (IA), análise de dados, algoritmos, etc. [6]. Pode estar relacionado à mineração de dados e big data [7].

A qualidade do ar é um aspecto fundamental da saúde pública e do meio ambiente, sendo diretamente influenciada pelas atividades industriais, veiculares e urbanas. No Brasil, a Companhia Ambiental do Estado de São Paulo (Cetesb) desempenha um papel fundamental na regulamentação e monitoramento da qualidade do ar [8].

A Cetesb é responsável por estabelecer normas e padrões de qualidade do ar que visam proteger a saúde da população e preservar o meio ambiente [8]. Os parâmetros monitorados são a concentração de poluentes atmosféricos como dióxido de enxofre ($SO_2$), monóxido de carbono (CO), óxidos de nitrogênio (NOx), ozônio ($O_3$), material particulado (MP) e compostos orgânicos voláteis (VOCs).

A legislação ambiental da Cetesb estabelece limites máximos permitidos para esses poluentes, com base em diretrizes da Organização Mundial da Saúde (OMS) e normativas nacionais [8]. A fiscalização e o monitoramento contínuo são realizados por meio de uma ampla rede de estações de monitoramento distribuídas pelo estado de São Paulo, que coletam dados em tempo real e fornecem informações essenciais para análises e tomadas de decisão.

No presente trabalho, é analisado um conjunto de dados que foi gerado no período de 1/1/2022 a 1/1/2023, na estação 66 - Parisi da CETESB em Cubatão [8]. Os dados foram analisados por ferramentas de ML, seguindo uma progressão. Primeiro, os dados foram analisados por histogramas individuais dos parâmetros, dados das distribuições (média, desvio-padrão e quartis) e correlações (*heat map*). Depois, foram encontrados os *clusters* por

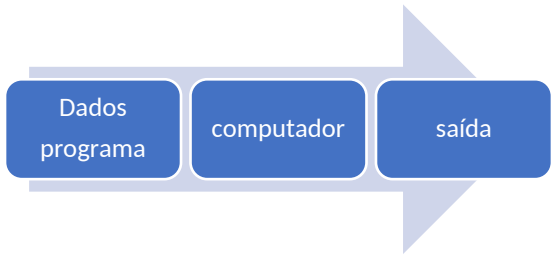
dois métodos: o dendrograma e o do joelho (*elbow*). Ao fim, uma classificação é avaliada por árvore de decisão. Foi possível concluir algo sobre um conjunto de dados reais. As ferramentas resultaram em análises interessantes que contribuem para o melhor controle sobre os poluentes na atmosfera.

## 2 Fundamentos

Aprendizado de máquina – ML – é uma das mimetizações do pensamento humano englobadas pela Inteligência Artificial (AI – *Artificial Intelligence*). Outros exemplos são o raciocínio, a busca, o planejamento, o processamento de linguagem natural (NLP – *Natural Language Processing*) e mover e manipular objetos [9].

A construção de um modelo matemático segue os mesmos passos de uma teoria científica. Primeiro são observados os fenômenos físicos; depois, são adotadas hipóteses que possibilitam gerar um modelo matemático a partir das equações fundamentais. A solução das equações matemáticas concebe soluções – modelagem – que, por sua vez, são comparadas aos dados experimentais de forma que o modelo seja validado.

Portanto, quando são utilizadas ferramentas matemáticas na modelagem, há um fluxo de programação: dados são introduzidos (propriedades físicas, químicas), ocorrem cálculos para a solução das equações matemáticas e resultados são gerados (Fig. 1(a)).




(a)

(b)

Figura 1: Sentido da programação: a) modelo fenomenológico; b) ML.

Em ML acontece uma inversão deste sentido de programação. São os dados de saída, os resultados, que constroem o modelo, Fig. 1(b). Daí, pode se colocar o paradigma básico, que é:

1. Observar um conjunto de exemplos: treinamento de dados;
2. Inferir algo sobre o processo que gerou os dados;
3. Usar a inferência para fazer predições sobre dados que não foram avaliados: teste de dados.

Os algoritmos básicos em ML são: regressão, agrupamento (*clustering*) e a classificação [10,11]. Outras possiblidades são a decomposição em valores singulares (SVD), a análise por componentes principais (PCA), a análise de componentes independentes (ICA) e os algoritmos bayseanos [12,13].

As etapas a serem seguidas na aplicação de ML são [14]:

1. Definir o problema;
2. Preparar os dados;
3. Avaliar os atributos;
4. Avaliar algoritmos;
5. Melhorar resultados;
6. Apresentar resultados.

## 2.1 Preparar os dados

O conjunto de dados recebido pode conter números e alfanuméricos. As entradas são os atributos (*features*) e as saídas são os rótulos (*labels*). As manipulações que os dados sofrem são matemáticas. Por isso, os atributos devem ser numéricos e tratáveis. Isto é feito pela engenharia de atributos [10]. Algumas ferramentas a serem usadas:

1. ***Cleaning***: é o processo de preparação dos dados por remoção ou modificação de dados incorretos, incompletos, irrelevantes e duplicados.
2. **Escalonamento** (*scaling*): converter números de sua faixa natural a faixas padrões, 0 a 1, -1 a +1
3. Tratar *outliers* extremos.
4. ***Binning***: empacotamento – faixas.
5. **Representação esparsa**: se muitos zeros, só guardar os valores diferentes de zero.

## 2.2 Avaliar os atributos

Os atributos formam distribuições que devem ser avaliadas previamente. O Pandas e o Sklearn [15] são bibliotecas que possuem diversas ferramentas estatísticas que analisam estas distribuições. Podem ser construídos histogramas; os parâmetros da distribuição são facilmente obtidos: média, desvio-padrão, valor máximo, valores dos quartis.

Os quartis ajudam na interpretação de distribuições. Também podem ser usados para retirar *outliers* à esquerda e à direita. Primeiro calcula-se a faixa interquartil (FIQ), que é a diferença entre o valor para o quartil 25%, Q1, e valor para o quartil 75%, Q3. Daí utilizam-se as faixas de valores menores que Q1 – 1,5 FIQ e maiores que Q3 + 1,5 FIQ. Estas faixas determinam os diagramas de caixa (*box plots*) [11].

As correlações entre os atributos também fazem parte das bibliotecas. Uma forma bastante interessante são os mapas de calor, os *heat maps*.

## 2.3 Agrupamento (clusterização)

Agrupamento pode ser definido como o processo de organização de objetos em grupos onde os membros são similares de alguma forma. Um ponto chave é a definição do que é similar**.** Uma forma disto ser feito é pelo cálculo de distância entre fatores [16].

Algoritmos de agrupamento podem ser hierárquicos e baseados em centroides, distribuição ou densidade. Um exemplo de algoritmo hierárquico é o dendrograma, baseado em conectividades, *single-linkage clustering*. É um algoritmo lento, por ter tempo computacional proporcional ao número de fatores elevado ao cubo ($o^3$).

Um algoritmo baseado em centroides é o k-médio (k-*means*). É muito mais rápido, é dito um algoritmo guloso (*greedy*). Em contrapartida, precisa que seja pré-definido o número de *clusters*, que é justamente o que deve ser resgatado.

## 2.4 Classificações

Classificação é uma aprendizagem supervisionada onde a saída y é um conjunto finito de valores (como ensolarado, nublado ou chuvoso). Caso a saída seja dois valores, a classificação é dita booleana [16]. Um bom exemplo é verificar se um e-mail é ou não é um *spam*.

O algoritmo mais simples é o k vizinhos mais próximos (K *nearest neighbors* - *KNN*) [16]. É uma classificação não paramétrica, onde k é o número de elementos mais próximos que deve ser usado. A classificação é obtida pelo cálculo de distâncias do novo ponto aos outros próximos, indicando um rótulo.

A análise discriminante linear (*Linear Discriminant Analysis* - LDA) possibilita encontrar uma reta ou plano que melhor separa classe de objetos, portanto, usa uma superfície de decisão linear. São inerentemente multiclasse, apresentaram bons resultados e não têm hiperparâmetros [15].

O algoritmo Gaussian Naive Bayes - NB é uma simplificação da aplicação do teorema de Bayes. É considerado que os atributos seguem a distribuição gaussiana e que há independência entre as variáveis.

A regressão logística aplica uma função sigmoide (modelo logístico, logit) para classificar conjunto de dados. É mais aplicada a casos binários, mas pode ser usada para casos multinominais [15].

Dentre estas diversas possibilidades, alguns algoritmos funcionam melhor que outros. Isto é mensurado pelos parâmetros de desempenho. Uma ferramenta bastante utilizada é a matriz de confusão [10], que quantifica o verdadeiro positivo (TP), o falso positivo (FP), o falso negativo (FN) e o verdadeiro negativo (TN). Outros parâmetros

de desempenho são: acurácia, Eq. (1); precisão, Eq. (2); especificidade, Eq. (3); sensibilidade (*recall*), Eq. (4) e Score F1, Eq. (5):

$$Acurácia = \frac{TP + TN}{TP + TN + FN + FP}. \tag{1}$$

$$Precisão = \frac{TP}{TP + FP}. \tag{2}$$

$$Especificidade = \frac{TN}{TN + FP}. \tag{3}$$

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

$$Score\ F1 = \frac{2}{\frac{1}{precisão} + \frac{1}{recall}}. \tag{5}$$

O Sklearn também avalia a performance por validação cruzada (*cross-validation*). O conjunto de dados é particionado em subconjuntos independentes, sendo um subconjunto usado para o treinamento do modelo e o outro é usado como teste [15].

# 3 Resultados e discussão

## 3.1 Avaliação dos atributos

A primeira etapa de uma construção por Machine Learning, e, até mesmo no processo de modelagem de forma geral, é avaliar os dados disponíveis. A ferramenta Pandas gera a matriz de dispersão (*scatter_matrix)*, Fig. 2.

As diagonais são os histogramas de distribuição para cada atributo e os outros elementos são as correlações entre cada par de variáveis. A distribuição de material particulado inalável ($PM_{10}$), menor que 10 μm, é assimétrica à esquerda (assimetria positiva). A média é 66 μg/m$^3$ e o valor máximo é 1.171 μg/m$^3$. A distribuição de NO também é assimétrica à esquerda (assimetria positiva). A média é 95 μg/m$^3$ e o valor máximo é 726 μg/m$^3$. De novo, a distribuição de $NO_2$ também é assimétrica à esquerda. A média é 44 μg/m$^3$ e o valor máximo é 247 μg/m$^3$. Seguindo a tendência, a distribuição de NOx é assimétrica à esquerda (assimetria positiva) com média de 67 μg/m$^3$ e o valor máximo é 621 μg/m$^3$. A concentração média de $SO_2$ é 8 μg/m$^3$ e o valor máximo é 1.309 μg/m$^3$, com praticamente uma única coluna à esquerda.

Todas as distribuições se concentram ao lado esquerdo, mas com valores máximos ordens de grandeza maiores que a média. Em se tratando de poluentes lançados na atmosfera, estes valores máximos não devem ser tratados como *outliers*, já que trazem grande prejuízo a saúde. Contudo, foram construídos alguns *boxplots*, retirando-se o quartil superior, para se observar melhor as distribuições em torno da moda, Fig. 3.

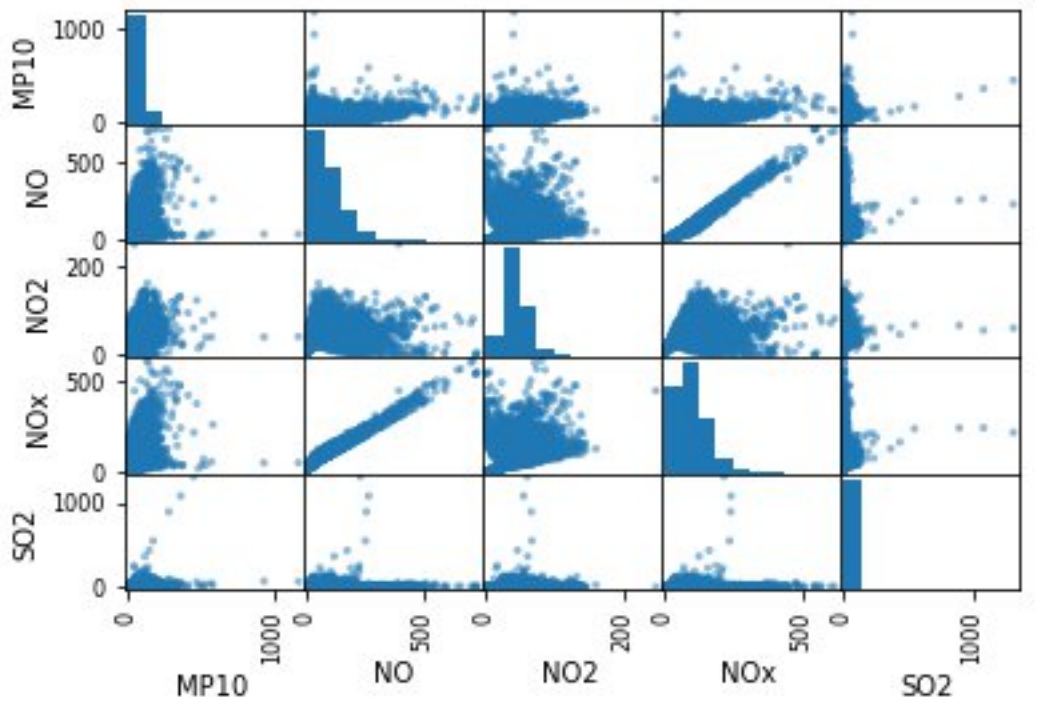



Figura 2: Matriz de dispersão (*scatter_matrix)* dos dados, concentrações em $\mu g/m^3$.

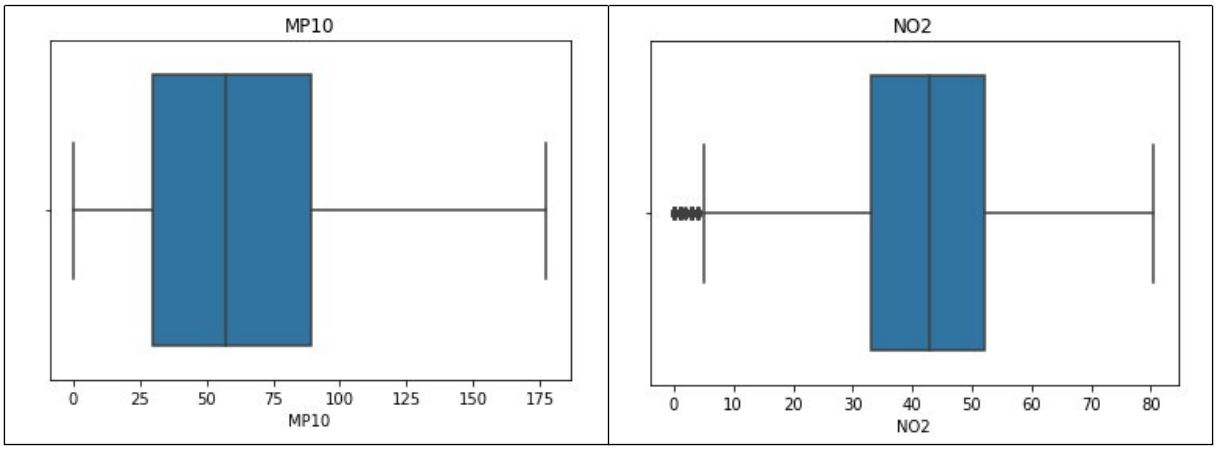



Figura 3: *Boxplots* para MP10 e $NO_2$ em $\mu g/m^3$.

Observa-se uma forte correlação entre as concentrações de NO e NOx, o que era esperado. Mas não se observa esta mesma correlação entre as concentrações de $NO_2$ e NOx. Todas as correlações estão apresentadas no *heat map* na Fig. 4.

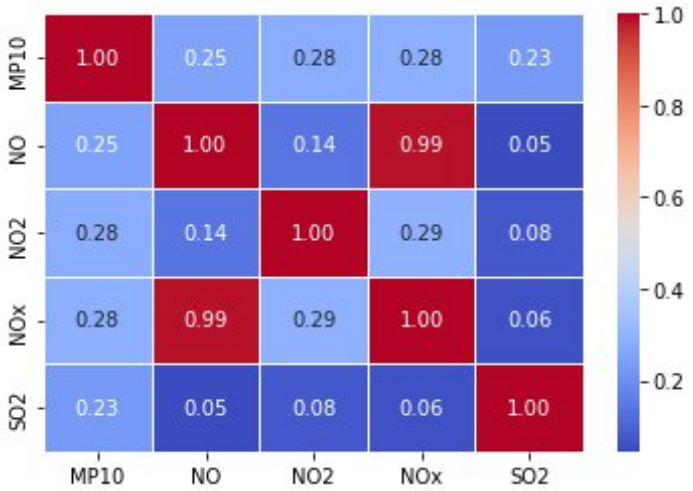



Figura 4: Matriz de correlações (*heat map)* dos dados.

## 3.2 Agrupamentos

Foram aplicadas duas ferramentas para agrupamento: o dendrograma, Fig. 5 e o método do K-médio (*K-means*), Fig. 6. A leitura para o dendrograma é realizada traçando-se retas horizontais, paralelas ao eixo x. O número de vezes que esta reta auxiliar é cortada pelas linhas verticais aponta um número de *clusters*. Uma interpretação do dendrograma indica agrupamentos com três ou quatro *clusters*.

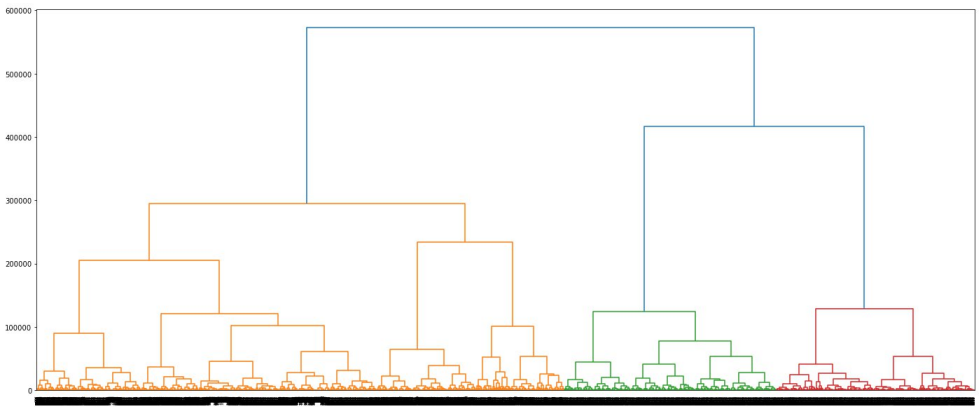



Figura 5: Dendrograma para os dados.

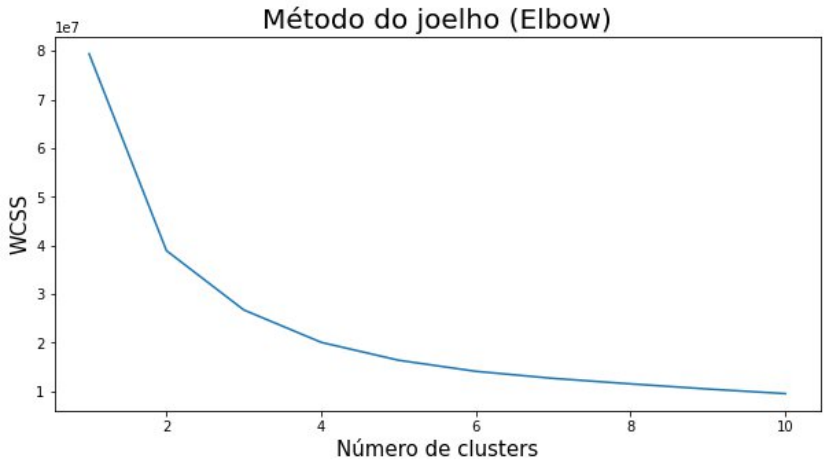



Figura 6: WCSS versus número adotado de clusters para os dados.

A Figura 6 relaciona a soma dos quadrados intra-*cluster* (*within-cluster-sum-of-square* – WCSS) com um número de clusters previamente escolhido. O ponto de inflexão indica o número de clusters, por isso, é chamado de método do joelho ou cotovelo. Portanto, o método do K-médio indica um agrupamento com 4 *clusters*, que foi o adotado aqui. À princípio, um agrupamento em *clusters* bom, moderado e ruim acompanharia a classificação utilizada pela Cetesb para o público, o que pode ser visto em relógios nas ruas da Cidade de São Paulo, capital. Contudo, quatro clusters foi considerado o mais adequado.

## 3.3 Classificações

Aqui foi utilizado o mesmo *data set* para agrupamento e considerando o critério de qualidade do ar: 1- Boa; 2- Moderada; 3- Ruim; 4- Muito ruim e Péssimo, estes dois itens reunidos em um só. Estes são critérios adotados pela Cetesb [8]. Para indicar um único índice, foi adotado o mais elevado dentre todas as variáveis.

Os algoritmos de classificação aplicados foram: i- regressão logística (RL); ii- análise discriminante linear (LDA); iii- k vizinhos mais próximos (KNN); iv- Gaussian Naive Bayes (NB).

O conjunto de dados foi separado em 80% para treinamento e 20% para teste. As métricas para validação cruzadas (*cross validation score*) para cada algoritmo estão na Tabela 1:

Tabela 1: Métricas para validações cruzadas para diferentes algoritmos

| Algoritmo | Média | Desvio-padrão |
|---|---|---|
| RL | 0,846888 | 0,008721 |
| LDA | 0,887568 | 0,017583 |
| KNN | 0,953138 | 0,007425 |
| NB | 0,865115 | 0,023393 |

O algoritmo por K vizinhos mais próximos apresentou melhor performance já que apresentou o melhor *cross validation score*, como pode ser na Tabela 1. A Figura 7 também indica que o algoritmo por K vizinhos mais próximos apresenta a menor faixa de erro e, com isso, a melhor performance. Portanto, pode ser usada para estimar a qualidade do ar por este algoritmo.

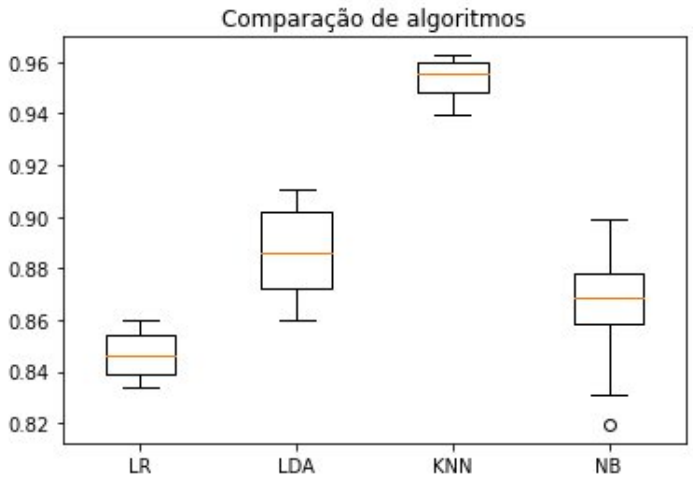



Figura 7: *Cross validation score* para diferentes classificações.

A matriz de confusão para o classificador KNN é: $\begin{pmatrix} 656 & 12 & 0 & 0 \\ 20 & 520 & 10 & 0 \\ 0 & 22 & 211 & 4 \\ 0 & 0 & 10 & 72 \end{pmatrix}$

São quatro atributos, por isso uma matriz 4x4. A diagonal principal indica os verdadeiros positivos. O total de valores é 1537, logo, total de todos os positivos = (656+520+211+72) = 1459, chegando a uma porcentagem de acerto de 1459/1537 = 0,95.

Os parâmetros de desempenho para o KNN estão na Tabela 2.

Tabela 2: Parâmetros de desempenho para o KNN

| Classe | Precisão | recall | F1-score |
|---|---|---|---|
| 1- Boa | 0,97 | 0,98 | 0,98 |
| 2- Moderada | 0,94 | 0,95 | 0,94 |
| 3- Ruim | 0,91 | 0,89 | 0,90 |
| 4- Muito ruim ou péssima | 0,95 | 0,88 | 0,91 |

# 4 Conclusões

A Cetesb disponibiliza conjunto de dados sobre a qualidade do ar de diversas estações em todo o Estado do São Paulo. Alguns parâmetros são disponibilizados. Aqui foram considerados dados de concentrações de dióxido de enxofre ($SO_2$), óxidos de nitrogênio (NOx) e material particulado inalável ($MP_{10}$). O conjunto de dados precisou ser tratado, foram retiradas linhas que não continham todas as concentrações dos poluentes avaliados. Foi inserida uma linha com o índice de qualidade do ar considerando as concentrações de dióxido de enxofre ($SO_2$), dióxido de nitrogênio ($NO_2$) e material particulado inalável ($MP_{10}$). A própria Cetesb indica:1- Boa; 2- Moderada; 3- Ruim; 4- Muito ruim ou Péssima, mas, estes dois itens foram reunidos em um só. Para indicar um único índice, foi adotado o mais elevado.

Os algoritmos de agrupamento (clusterização), baseado em conectividades (dendrograma) e método do k médio (método do joelho com diversos Ks), indicaram quatro clusters, o que está de acordo com o índice de qualidade utilizado pela Cetesb: 1- Boa; 2- Moderada; 3- Ruim; 4- Muito ruim ou Péssima, estes dois itens reunidos em um só.

Foram aplicados quatro classificadores. O que apresentou melhor performance foi o algoritmo de k vizinhos mais próximos, com um *Cross Validation Score* igual a 0,953138. Os box-plots também revelam, Fig. 7, que o algoritmo KNN apresenta o menor espalhamento.

# Agradecimentos

O presente trabalho foi realizado com apoio parcial da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

# Referências

[1] S. Lohr, “The age of big data,” *New York Times*, 11, 2012. Disponível em: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[2] National Academies of Sciences, Engineering, and M. and others, *Data science for undergraduates: opportunities and options. National Academies of Sciences, Engineering and Medicine* Tech. rep, 2018(a). Disponível em: http://nap.edu/25104

[3] National Academies of Sciences, Engineering, and M. and others (2018b) *Data science: opportunities to transform chemical sciences and engineering: proceedings of a workshop in brief. National Academies of Sciences, Engineering and Medicine Tech. rep*., 2018(b). Disponível em: https://doi.org/10.17226/25191

[4] K. Schwab, *The fourth industrial revolution*. Currency, 2017. Disponível em: http://voicebucketvoitto.s3.amazonaws.com/pdf/ingles/%5BENG%5D%20A%20Quarta%20Revolucao%20Industrial.pdf

[5] Techjury. Acesso em 23 de março de 2022. Disponível em: https://techjury.net/blog/how-much-data-is-created-every-day/#gref,

[6] V. Dhar, “Data science and prediction,” *Communications of the ACM,* vol. 56, no. 12, pp. 64–73, 2013. Disponível em: https://doi.org/10.1145/2500499

[7] J. Leek, “The key word in Data Science is not Data, it is Science”. *Simply Statistics*, vol. 12, 2013. Disponível em: https://www.linkedin.com/pulse/keyword-data-science-kanika-garg-85o3c

[8] Cetesb. Acesso em 02 de maio de 2024. Disponível em: https://cetesb.sp.gov.br/ar/padroes-de-qualidade-do-ar/

[9] P. Norvig, S. Russell, *Inteligência Artificial*, tradução da 3a ed., Elsevier, 2013. Disponível em: https://www.grupogen.com.br/livro-inteligencia-artificial-uma-abordagem-moderna-stuart-russell-e-peter-norvig-9788595158870

[10] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow”*, O'Reilly Media, Inc., 2022. Disponível em: https://anayamultimedia.es/primer_capitulo/aprende-machine-learning-con-scikit-learn-keras-y-tensorflow-tercera-edicion.pdf

[11] A. Kadiwal. Acesso em 09 de maio de 2024. Disponível em: https://www.kaggle.com/datasets/adityakadiwal/water-potability

[12] J. D. Hurwitz, Kirsch, *Machine Learning for Dummies*, IBM Limited Edition. New Jersey: Jhon Wiley, 2018. Disponível em: https://www.wiley.com/en-mx/Machine+Learning+For+Dummies-p-9781119245513

[13] E. Alpaydin, *Introduction to machine learning*, 4th edition, MIT press, 2020. Disponível em: https://www.bme.ufl.edu/wp-content/uploads/2018/07/Fall-2015-Syllabus-BME6938-Machine-Learning.pdf

[14] S. Marsland, *Machine learning: an algorithmic perspective*, 2nd edition, Chapman and Hall/CRC, 2018. Disponível em: http://2.180.2.83:801/opac/temp/11623.pdf

[15] J. Guttag, *Introduction to Computation and Programming Using Python: With Application to Understanding Data,* 2nd ed. MIT Press, Cambridge, 2016. Disponível em: https://thuvienso.hoasen.edu.vn/bitstream/handle/123456789/8846/Contents.pdf?sequence=3

[16] Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011. Disponível em: https://doi.org/10.1002/hbm.25822