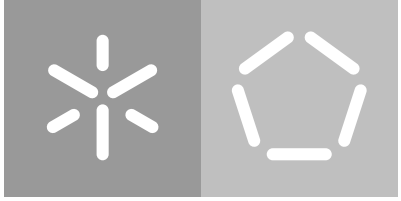
Universidade do Minho
Escola de Engenharia

Márcia Maria Fernandes Costa

Modelos preditivos para monitorização do Índice da Qualidade do Ar

Abril, 2024



Universidade do Minho
Escola de Engenharia

Márcia Maria Fernandes Costa

Modelos preditivos para monitorização do Índice da Qualidade do Ar

Dissertação de Mestrado

Mestrado em Engenharia de Sistemas

Trabalho efetuado sob a orientação do(a)

Paulo Novais

Abril, 2024

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Assinado por: Márcia Maria Fernandes da Costa
Identificação: BI14661428
Data: 2024-04-25 às 08:31:08
Local: Braga

(Márcia Maria Fernandes Costa)

Agradecimentos

Gostaria de expressar a minha mais sincera gratidão a todas as pessoas e instituições que contribuíram de forma significativa para a realização desta dissertação de mestrado.

Em primeiro lugar, pretendo expressar a minha profunda gratidão ao meu orientador, o professor Paulo Novais, cuja orientação sábia, apoio constante e incentivo foram fundamentais para a conclusão desta dissertação. Agradeço também à Universidade do Minho, por fornecer o ambiente propício para a realização deste estudo. Adicionalmente, gostaria de expressar a minha profunda gratidão ao Pedro Oliveira pelo notável suporte, incansável empenho e valiosa ajuda proporcionados ao longo deste percurso académico.

Agradeço ainda aos meus colegas de mestrado, cuja colaboração e troca de ideias enriqueceram a qualidade deste trabalho.

Além dos agradecimentos institucionais, quero estender o meu reconhecimento à minha família, cujo amor, paciência e apoio incondicional foram a força motriz por trás desta jornada académica. Agradeço aos meus amigos, que partilharam não apenas momentos de lazer, mas também ofereceram palavras de estímulo nos momentos desafiadores.

Por último, dedico um agradecimento especial ao meu namorado que esteve ao meu lado durante este período, compreendendo os momentos de ausência e celebrando os sucessos.

A todos os mencionados e àqueles que, porventura, tenham contribuído de alguma forma, o meu mais sincero obrigada. Este projeto foi parcialmente apoiado por Fundos Nacionais através da agência de financiamento portuguesa, FCT - Fundação para a Ciência e a Tecnologia, dentro do projeto AIM4WATER, com a referência 2022.06822.PTDC (https://doi.org/10.54499/2022.06822.PTDC).

Resumo

Modelos preditivos para monitorização do Índice da Qualidade do Ar

Atualmente, observa-se uma crescente preocupação acerca dos impactos prejudiciais que as gerações futuras poderão enfrentar em virtude da insustentabilidade ambiental do planeta. Diante dessa realidade, a presente dissertação tem como objetivo analisar a forma como a sociedade está a utilizar os recursos naturais disponíveis. Nesse sentido, esta dissertação visa estudar esse fenómeno através da conceção de modelos preditivos sobre o Índice da Qualidade do Ar (IQA), efetuando previsões para dois *timesteps* futuros, recursivamente.

O IQA é um indicador que traduz o estado da qualidade do ar. Para além da compreensão e visualização de dados relacionados com o IQA, é também objeto de estudo a integração de dados relacionados com o clima (temperatura, vento, etc.) para entender que impacto podem ter nesse índice.

A presente dissertação passa pela conceção de modelos de Deep Learning (DL), como o Long Short Term Memory (LSTM), Multilayer Perceptron (MLP), Gated recurrent unit (GRU) e Convolutional Neural Networks (CNN) para a previsão do IQA, incluindo o treino e o *tuning* de vários modelos candidatos. Pretende-se que seja possível tomar decisões de forma proativa minimizando, assim, situações de risco, e melhorando a qualidade de vida da população mundial.

Os resultados demonstraram que o modelo com melhor *performance* dependia da abordagem adotada. No caso da abordagem *Univariate*, o modelo candidato baseado no modelo CNN apresentou a melhor performance, com um valor aproximado de Root Mean Square Error (RMSE) de 4.46. Por outro lado, no que concerne à abordagem *Multivariate*, o modelo candidato com melhor *performance*, foi baseado no modelo MLP, também com um valor aproximado de RMSE de 4.46.

Palavras-chave: Deep Learning, Índice da qualidade do ar, Machine learning, Sustentabilidade Ambiental

Abstract

Predictive Models for Air Quality Index Monitoring

There is a growing concern about the detrimental impacts that future generations may face due to the environmental unsustainability of the planet. Faced with this reality, this dissertation analyses how society utilises available natural resources. This dissertation seeks to study this phenomenon by designing predictive models for the IQA, making forecasts for two future *timesteps* recursively.

The IQA is an indicator that reflects air quality. In addition to understanding and visualising data related to the IQA, the integration of climate-related data (temperature, wind, etc.) is also studied to understand their impact on this index.

This dissertation involves the design of DL models such as LSTM, MLP, GRU, and CNN for predicting the IQA, including the training and tuning of various candidate models. The goal is to make proactive decisions, thus minimising risk situations and improving the quality of life for the global population.

The results demonstrated that the model with the best performance depended on the adopted approach. In the case of the *Univariate* approach, the candidate model based on the CNN model showed the best performance, with an approximate RMSE value of 4.46. On the other hand, regarding the *Multivariate* approach, the candidate model with the best performance was based on the MLP model, also with an approximate RMSE value of 4.46.

Keywords: Air Quality Index, Deep Learning, Environmental Sustainability, Machine Learning

Índice

Lista de Figuras

Lista de Tabelas

Listagens

Siglas

AD	Árvore de Decisão
ANS	Aprendizagem não Supervisionada
APA	Agência Portuguesa do Ambiente
AQHIc	Air Quality Health Index - children
AR	Aprendizagem de Reforço
AS	Aprendizagem Supervisionada
BP	Back Propagation
C_6H_6	Benzeno
CAQI	Composite Air Quality Index
CART	Classification and Regression Trees
CCDR	Comissões de Coordenação e Desenvolvimento Regional
CH_4	Metano
Cl	Cloro
CNN	Convolutional Neural Networks
CO	Monóxido de Carbono
CO_2	Dióxido de Carbono
ConvLSTM	Convolutional LSTM Network
COV	Compostos Orgânicos Voláteis
CRISP-DM	Cross Industry Standard Process for Data Mining
DALY	Disability-Adjusted Life Years DALY
DI	Departamento de Informática
DL	Deep Learning
DMA	Desvio Médio Absoluto
DRA	Direções regionais do ambiente

GAM	Generalized additive model
GLCM	Gray Level Co-occurrence Matrix
GONM	Gases Orgânicos não-metano
GRU	Gated recurrent unit
H	Hidrogénio
IA	Inteligência Artificial
ID3	Iterative Dichotomiser 3,C4.5
IDE	Integrated Development Environment
IQA	Índice da Qualidade do Ar
LSTM	Long Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MES	Mestrado de Engenharia de Sistemas
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
NAAQS	National Ambient Air Quality Standards
NAQI	National Air Quality Index
NH3	Amónia
NO_2	Dióxido de Azoto
NO_x	Óxidos de Azoto
NxOy	Óxidos de Nitrogénio
O_3	Ozono
OMS	Organização mundial de saúde
Pb	Poluentes chumbo
PM_{10}	Partículas em Suspensão

$PM_{2,5}$	Partículas em Suspensão Finas
PSO	Particle Swarm Optimization
ReLU	Unidade Linear Retificada
RF	Random Forest
RFMC	Redes Feedforward Multicamada
RFSC	Redes Feedforward de uma só camada
RMSE	Root Mean Square Error
RNA	Redes Neuronais Artificiais
RNR	Rede Neuronal Recorrente
S	Enxofre
SARSA	State–action–reward–state–action
SO_2	Dióxido de Enxofre
SVMs	Support vector machines
SVR	Support Vector Regression
TanH	Tangente Hiperbólica
UM	Universidade do Minho
UV	Ultra Violeta

Capítulo 1

Introdução

O primeiro capítulo desta dissertação proporciona uma abordagem teórica ao tema, tendo sido promovida no âmbito da Unidade Curricular de Dissertação do Mestrado de Engenharia de Sistemas (MES), desenvolvida no Departamento de Informática (DI) na Universidade do Minho (UM). Neste contexto, o alcance primordial deste projeto centra-se na aplicação de temáticas da ciência da computação, nomeadamente no campo de ML. Especificamente, pretende-se implementar modelos de DL, para prever o IQA.

O propósito subjacente a este projeto é capacitar a tomada de decisões pro ativas, contribuindo para a minimização de situações de risco e promovendo a melhoria da qualidade de vida da população global.

Desta forma, o presente capítulo será composto por várias secções. Na secção 1.1, serão delineados o enquadramento e a motivação que fundamentam a escolha deste tema. Na Secção 1.2, será apresentada a metodologia de trabalho que guiará o desenvolvimento desta dissertação. Os objetivos essenciais para a consecução deste projeto serão expostos na Secção 1.3. Por fim, a Secção 1.4 caracterizará a estrutura do documento, proporcionando uma visão abrangente do percurso a ser trilhado.

1.1 Enquadramento e Motivação

A evolução do ser humano desde os primórdios da sua existência tornou possível o mundo tal como o conhecemos hoje. As sociedades estão cada vez mais dotadas de conhecimentos e ferramentas que aceleram ainda mais esta evolução, algo que se repercute no estilo de vida de cada um de nós [1]. Os veículos que usamos para nos movermos de um local para outro, os edifícios onde passamos grande parte do nosso dia a dia ou os dispositivos eletrónicos que nos acompanham constantemente, são produtos dessa evolução [2]. Todos esses fatores conjugados contribuem para que o ser humano em geral tenha uma melhor qualidade de vida, algo que é notório com o aumento da esperança média de vida [3].

Apesar de todos os prós, é inevitável referir também que nem tudo é positivo neste processo evolutivo [4]. Nomeadamente, o crescimento de população mundial cria necessidades que apenas podem ser

supridas com métodos pouco ou nada naturais. As construções de grandes cidades extinguem a natureza que outrora lá existia [4]. A atual industrialização de produtos, por exemplo, contrasta com a normal morosidade com que os mesmos eram produzidos há alguns anos atrás, criando excedentes cuja degradação demora muito tempo [1]. Para além disso, todos estes fatores implicam a exploração, alteração e transformação de recursos naturais a um ritmo que não é suportável [5]. Prova disso é, por exemplo, o declínio das populações de peixes, anfíbios, aves, répteis e mamíferos que diminuiu 68% entre 1970 e 2016 [6].

Tudo isto demonstra que é urgente tomar atitudes que moderem o impacto dos humanos no ambiente de forma a não prejudicarem a evolução de outras espécies [7]. O próprio ser humano beneficiará dessas medidas, reduzindo a poluição que, para além, de já afetar as gerações mais recentes, prevê-se que afete também as gerações futuras expostas a essa poluição através dos seus progenitores [8, 9].

É necessário compreender que para evitar que as gerações futuras, venham a sofrer represálias as entidades responsáveis devem tomar no imediato decisões que diminuam o risco da população. Uma forma de o fazerem é recorrendo ao ML [10]. O ML é uma tecnologia de IA, que permite que os sistemas aprendam e melhorem a partir da experiência [11]. Uma vez que o ML é capaz de fornecer várias soluções através das informações dos dados disponíveis, as entidades que operam nas mais diversas áreas podem assim, adquirir a capacidade de agir antes que algo ocorra, evitando impactos negativos [12].

A escolha do tema desta dissertação, centrada em ML, surge num contexto global de crescente relevância. A expansão exponencial da tecnologia de ML têm impactado diversos setores, despertando em mim um interesse significativo.

A crescente urbanização nas grandes cidades é um fenómeno contemporâneo que, embora possibilite avanços económicos e tecnológicos, está inseparavelmente associado a uma preocupação cada vez mais imperativo, a poluição atmosférica [1]. Este fenómeno, amplificado pelo aumento da atividade industrial, traz consigo uma série de consequências nefastas para o meio ambiente e para a saúde humana [13].

Globalmente falando, a qualidade do ar tornou-se uma prioridade, uma vez que afeta de forma direta e indireta diversas esferas da sociedade [14]. A relevância do problema reside no facto de que a poluição do ar não conhece fronteiras, sendo um desafio transnacional que sugere abordagens holísticas e soluções inovadoras [1].

As implicações da poluição atmosférica para a saúde humana são vastas e profundas. Estudos científicos, como os destacados por Goodfellow et al. [15] e outros pesquisadores [14], evidenciam que o crescimento descontrolado das cidades, especialmente em regiões economicamente dinâmicas como as urbanas chinesas, está correlacionado diretamente com o aumento dos níveis de poluentes atmosféricos. Esta associação não só compromete a qualidade do ar respirado pela população, mas também está intrinsecamente ligada a uma série de problemas de saúde, desde doenças respiratórias até complicações cardiovasculares [16].

Além dos impactos diretos na saúde, a poluição atmosférica tem implicações sócio económicas consideráveis [17]. Custos associados a despesas médicas, perda de produtividade e degradação ambiental

representam uma carga significativa para a sociedade como um todo. Para além do mencionado anteriormente, a reputação das cidades afetadas pode ser comprometida, influenciando negativamente setores como o turismo e investimentos [18].

Nesse contexto, a aplicação de tecnologias de ML para controlar e prever padrões de poluição do ar emerge como uma ferramenta crucial. Ao integrar sistemas de ML a redes de sensores e dados ambientais, é possível não apenas compreender os padrões de poluição, mas também desenvolver estratégias eficazes de mitigação [11].

Desta forma, a presente pesquisa almeja contribuir para a compreensão mais profunda dos desafios associados à poluição atmosférica em ambientes urbanos, destacando a interseção vital entre a tecnologia de ML e a preservação ambiental. A busca por soluções inovadoras, baseadas em dados precisos e análises preditivas, não apenas promove uma abordagem pro ativa para mitigar os impactos da poluição do ar, mas também sinaliza um compromisso com a construção de cidades mais sustentáveis e saudáveis para as gerações futuras [19].

1.2 Metodologia

A abordagem metodológica empregada nesta dissertação segue o Cross Industry Standard Process for Data Mining (CRISP-DM), um modelo padrão que esboça o processo de exploração de dados. Este modelo oferece uma estrutura abrangente e orienta passo a passo a condução de projetos de exploração de dados de maneira sistemática e eficaz [20].

Esta metodologia é composta por seis etapas principais: *Business Understanding*, *Data understanding*, *Data preparation*, *Modeling*, *Evaluation* e *Deployment*, estando estas representadas no esquema da Figura 1.1:

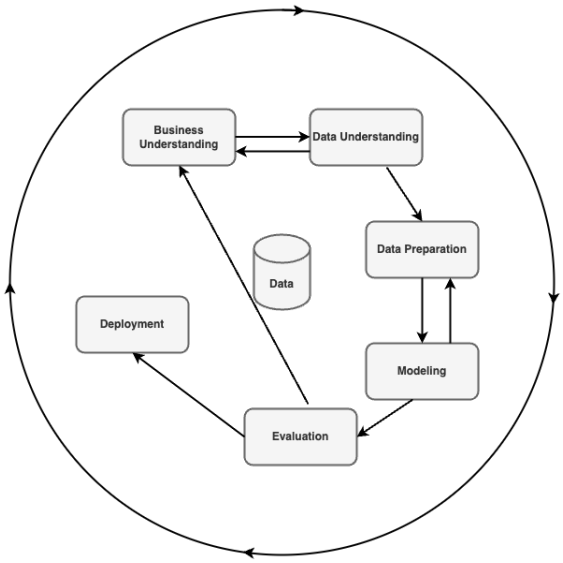


Figura 1.1: Metodologia CRISP-DM

Business Understanding: Nesta fase inicial, o objetivo é compreender os objetivos e requisitos do

projeto de exploração de dados em relação ao contexto e necessidades do negócio. É importante identificar os problemas a serem resolvidos e definir metas claras.

Data Understanding: Esta fase, inicia-se com a obtenção dos dados, a sua exploração e subsequente análise. Esses processos incluem a identificação da qualidade dos dados, a familiarização com o seu conteúdo e a deteção de possíveis problemas ou limitações.

Data Preparation: Nesta etapa, os dados são preparados para análise, o que pode envolver a seleção de atributos relevantes, a limpeza e transformação dos dados, bem como a integração de diferentes fontes de dados, se necessário. O objetivo é obter um conjunto de dados apropriado e de qualidade para ser utilizado nas etapas seguintes do processo.

Modeling: Nesta fase, os modelos são criados usando os dados preparados. Vários parâmetros podem ser ajustados para otimizar o desempenho do modelo. Diferentes algoritmos podem ser testados para encontrar o que melhor se adequa ao problema em questão.

Evaluation: Neste momento, os modelos construídos são avaliados de acordo com critérios previamente estabelecidos. Nesse sentido, existe uma análise dos resultados obtidos, a validação dos modelos e a verificação se atendem aos objetivos iniciais do projeto.

Deployment: Por fim, os resultados da modelação dos dados são apresentados aos utilizadores finais. Isso pode incluir a implementação de sistemas, a criação de relatórios ou ações baseadas nas descobertas obtidas.

Pode-se ainda destacar que CRISP-DM é um processo iterativo, o que significa que as fases podem ser analisadas e repetidas conforme necessário. Este método oferece uma abordagem flexível e adaptável para lidar com projetos de exploração de dados em diferentes indústrias e contextos [21].

1.3 Objetivos

A presente dissertação visa implementar e criar modelos de ML, levando em conta o bem-estar das gerações presentes e futuras. Dessa maneira, o projeto de pesquisa apresentado tem como objetivo contribuir de maneira favorável para a tomada de decisões mais sustentáveis, tanto para o meio ambiente quanto para a humanidade. Os principais objetivos deste projeto são:

- Revisão da literatura existente sobre o uso de modelos preditivos sobre o IQA;
- Compreensão e visualização de dados relacionados às diferentes substâncias que constituem o IQA;
- Integração de dados relacionados ao clima (temperatura, vento, etc.) para entender que impacto podem ter no IQA;
- Pré-processamento de dados considerando o problema de séries temporais;

- Conceção de modelos de *Deep Learning (LSTM, GRU, MLP e CNN)* para previsão dO IQA, incluindo treinar e fazer o *tuning* de vários modelos com o objetivo de obter previsões melhores e mais precisas;

- Análise dos resultados obtidos nos diferentes modelos.

1.4 Estrutura da Dissertação

A presente dissertação encontra-se dividida em 6 capítulos. O capítulo 1 aborda o enquadramento e a motivação do tema, de forma a integrar o leitor à temática da dissertação. É também neste capítulo que é apresentada a metodologia adotada, seguida dos objetivos da dissertação. Por fim, a estrutura global do documento é esboçada, proporcionando uma visão geral do seu *layout*.

No capítulo 2 é apresentada a revisão bibliográfica dos temas da Sustentabilidade ambiental, do IQA e da IA, temas que serão abordados ao longo da dissertação.

No capítulo 3 deste projeto de dissertação, são abordados os materiais e métodos empregados. Aqui, ocorre a investigação dos dados, examinando-se os diversos conjuntos de dados a serem utilizados, bem como o processo de preparação dos mesmos. Adicionalmente, são descritas as modificações pelas quais esses conjuntos foram submetidos, culminando na exposição das tecnologias empregadas para a execução desta dissertação.

No capítulo 4, são detalhadas as experiências computacionais da pesquisa. São expostos os diversos cenários criados, juntamente com o desenvolvimento dos modelos e a otimização dos hiperparâmetros correspondentes.

No capítulo 5, são expostos os resultados alcançados neste estudo, abrangendo todos os modelos e conjuntos de dados utilizados. Além disso, é realizada uma discussão abrangente sobre esses resultados.

Por último, no capítulo 6, são apresentadas as conclusões finais decorrentes deste estudo, juntamente com as perspetivas e direções para trabalhos futuros.

Capítulo 2

Estado da Arte

Este capítulo será feita uma análise do estado atual da pesquisa, visando identificar os artigos pertinentes na literatura relacionados ao tema central desta dissertação. Simultaneamente, procura-se identificar e analisar os trabalhos desenvolvidos nesse contexto, explorando os avanços mais recentes em áreas fundamentais para compreender o tema em questão.

Neste sentido, na secção 2.1 é abordado o tema da Sustentabilidade Ambiental, destacando a preocupação com a preservação dos recursos naturais e a qualidade de vida das gerações presentes e futuras. Para além disso, serão examinados minuciosamente os diferentes poluentes atmosféricos, como o NO_2, SO_2, PM_{10}, Poluentes chumbo (Pb), CO, Benzeno (C_6H_6) e o O_3, no âmbito da análise da poluição do ar.

Na secção 2.2, é analisado o IQA, explorando cada um dos seus constituintes. De seguida na secção 2.3 é explorada a relação entre a IA, o ML e o DL. Nesta secção serão minuciosamente explorados vários modelos tanto de ML como de DL.

A secção 2.5 será dedicada à Revisão da Literatura, apresentando uma análise crítica ao trabalho.

2.1 Sustentabilidade Ambiental

A sustentabilidade ambiental diz respeito à habilidade de utilizar os recursos naturais de maneira equilibrada e responsável, assegurando a preservação do meio ambiente a longo prazo, como apontado por *Meadows* e outros colaboradores em 1972 [22]. Este conceito visa harmonizar o progresso da espécie humana com a preservação dos recursos naturais e a salvaguarda dos ecossistemas, conforme delineado pelo *World Commission on Environment and Development* em 1987 [23].

A interdependência entre os sistemas naturais e sociais é reconhecida pela sustentabilidade ambiental, que leva em consideração não apenas as necessidades presentes, mas também as necessidades futuras das gerações subsequentes [24]. Este conceito envolve práticas e políticas que tentam minimizar

os impactos negativos no meio ambiente, promovendo a conservação dos ecossistemas, a redução da poluição, o uso eficiente dos recursos naturais e a promoção da biodiversidade [25].

Para alcançar a sustentabilidade ambiental, é necessário adotar abordagens como a utilização de energias renováveis, o desenvolvimento de tecnologias limpas, a promoção da reciclagem e da economia circular, a conservação dos ecossistemas naturais e o incentivo à agricultura sustentável [26].

Atualmente, seja por meio da rádio, televisão ou Internet, somos regularmente expostos a assuntos que abordam a busca por lares mais eco amigáveis, práticas de reciclagem, a diminuição do uso de plástico, métodos de agricultura sustentável e optar por transportes públicos em vez de veículo individual (sustentabilidade na locomoção), entre outros tópicos [27]. Somos incentivados a parar e a contemplar essas questões, e é crucial refletir sobre os nossos comportamentos e considerar maneiras de promover mudanças positivas [28].

Com o constante crescimento da população mundial, a preservação de recursos naturais torna-se uma tarefa extremamente importante, mas em simultâneo, exaustiva e complexa [29]. Uma compreensão intuitiva da sustentabilidade ambiental está associada ao conceito de viabilidade ao longo do tempo, alinhando-se com a própria definição da palavra, que sugere algo ”que pode ser mantido”, ”que possui condições para perdurar ou preservar”[30].

A constante inovação no campo científico e tecnológico tem proporcionado à humanidade a capacidade de transformar o ambiente e a sociedade [31]. No entanto, é importante ressaltar que nem sempre se tem conhecimento dos efeitos completos dessas mudanças. Essa falta de compreensão pode resultar em consequências imprevistas ou indesejadas para o meio ambiente e para a sociedade em geral [32].

Os últimos dois séculos assistiram a fenómenos como a industrialização e a revolução tecnológica, que culminaram nas crises de sustentabilidade [33]. Assim, o desenvolvimento sustentável procura harmonizar as necessidades presentes e futuras de forma a construir um futuro inclusivo, sustentável e resiliente para a população e para o planeta [30]. Resumidamente, a sustentabilidade ambiental desempenha um papel crucial ao assegurar um equilíbrio entre as necessidades humanas e a capacidade do planeta de sustentar a vida. Este conceito é fundamental para enfrentar os desafios ambientais, promover a justiça social e construir um futuro mais resistente e próspero para todos [33].

A poluição do Ar

Ao explorar o conceito de sustentabilidade, é inevitável abordar a preocupante questão da poluição ambiental, um problema que exige atenção urgente [34]. A crise ambiental alarmante que enfrentamos hoje é resultado de uma série de transformações no ambiente e na ecologia, causadas principalmente pelo rápido avanço da economia e da tecnologia promovido pela humanidade ao longo deste século [35]. Embora se tenha alcançado um progresso socioeconómico, científico e tecnológico, é inegável que essas conquistas vêm acompanhadas de consequências devastadoras para o nosso planeta [36].

A complexidade do problema é incontestável, exigindo uma análise cuidadosa das principais causas que contribuíram para o estado atual. Embora haja diversas perspetivas sobre o tema, é lógico reconhecer

que não existe uma única causa isolada, mas sim uma combinação de diversos fatores interligados [37]. A poluição ambiental é resultado não apenas de uma única fonte, mas de uma teia complexa de influências, incluindo a atividade industrial, o uso desenfreado de recursos naturais, a urbanização descontrolada e as práticas insustentáveis de consumo [27].

Para enfrentar efetivamente essa crise, é essencial compreender a interconexão entre esses fatores e adotar medidas abrangentes e integradas. É fundamental repensar o atual modelo de desenvolvimento, criando soluções inovadoras que considerem não apenas o crescimento económico, mas também a preservação dos recursos naturais e a saúde do meio ambiente [38].

Em suma, a poluição ambiental é um dos grandes desafios da sustentabilidade, resultado das alterações ambientais e ecológicas causadas pelo desenvolvimento económico e tecnológico [39]. A complexidade do problema requer uma abordagem holística e ações concretas para mitigar os impactos negativos [38].

O conceito de poluição ambiental e de poluição do ar estão intimamente relacionadas, sendo a poluição do ar um dos componentes da poluição ambiental [39]. A presente dissertação terá como foco primordial a poluição do ar. A poluição do ar, é um aspeto específico da poluição ambiental que se concentra na contaminação da atmosfera por substâncias nocivas [40]. A queima de combustíveis fósseis, a atividade industrial, o transporte e outras fontes de emissões libertam poluentes na atmosfera, como Dióxido de Carbono (CO_2), Óxidos de Azoto (NO_x), SO_2, PM_{10} e Compostos Orgânicos Voláteis (COV) [41].

Esses poluentes do ar têm efeitos adversos na saúde humana, contribuindo para doenças respiratórias, cardiovasculares e outros problemas de saúde [39]. Além disso, a poluição do ar também afeta negativamente os ecossistemas, causando danos à flora, fauna e aos recursos naturais [31].

Portanto, a poluição do ar é uma das principais preocupações da poluição ambiental, e a redução das emissões de poluentes atmosféricos é essencial para minimizar os impactos negativos na saúde humana e no meio ambiente como um todo [38].

Neste sentido, pode dizer-se que a qualidade do ar está diretamente relacionada à poluição, uma vez que a presença de poluentes na atmosfera afeta negativamente a sua qualidade [42]. A qualidade do ar tem um enorme impacto na qualidade de vida e bem-estar da população [40]. Um pouco por todo o mundo o aumento da população nas áreas urbanas, a queima de combustíveis fósseis para produção de energia, os grandes processos industriais, como a indústria metalúrgica ou a indústria de construção e transporte têm contribuído para o aumento da poluição do ar, colocando assim, em risco a saúde humana [17].

A poluição do ar é a causa de um conjunto de problemas, nomeadamente [13] :

- Deterioração da qualidade do ar;
- Exposição humana e dos ecossistemas a substâncias tóxicas;
- Danos na saúde humana;
- Danos nos ecossistemas e património construído;

- Acidificação;
- Destruição da camada de ozono estratosférico;
- Alterações climáticas.

Os danos nos ecossistemas podem incluir a oxidação de estruturas da vegetação, que entre muitas outras consequências podem originar a queda prematura das folhas em algumas espécies ou o apodrecimento precoce de alguns frutos [43]. Quando se trata de culturas agrícolas estes danos podem também significar grandes prejuízos económicos [44]. Entre os efeitos na saúde humana são de destacar os problemas ao nível dos sistemas respiratório e cardiovascular [17].

A poluição do ar é um dos principais fatores de risco de morte no mundo, atribuído a milhões de mortes a cada ano [13]. Esta desenvolve-se em dois contextos:

- Poluição do ar *Indoor* (doméstico)
- Poluição do ar *Outdoor*.

A poluição do ar *Indoor* resulta da queima de combustíveis sólidos, como resíduos de colheitas, estrume, carvão para cozinhar e aquecer as casas [13]. Por sua vez, a queima desses combustíveis produz material particulado – um grande risco para a saúde, principalmente para doenças respiratórias [39]. Por conseguinte, a combustão dessas partículas em espaços fechados, como pequenas residências, é um importante fator de risco para a exacerbação dessas doenças. As taxas de mortalidade causadas por poluição do ar *Indoor* são mais altas em países menos desenvolvidos [45].

Globalmente, nas últimas décadas as taxas de mortalidade por poluição total do ar diminuíram, desde 1990, as taxas de mortalidade caíram quase para metade [46]. Esse declínio foi impulsionado principalmente por melhorias na poluição do ar interno, tal como se pode observar na Figura 2.1.

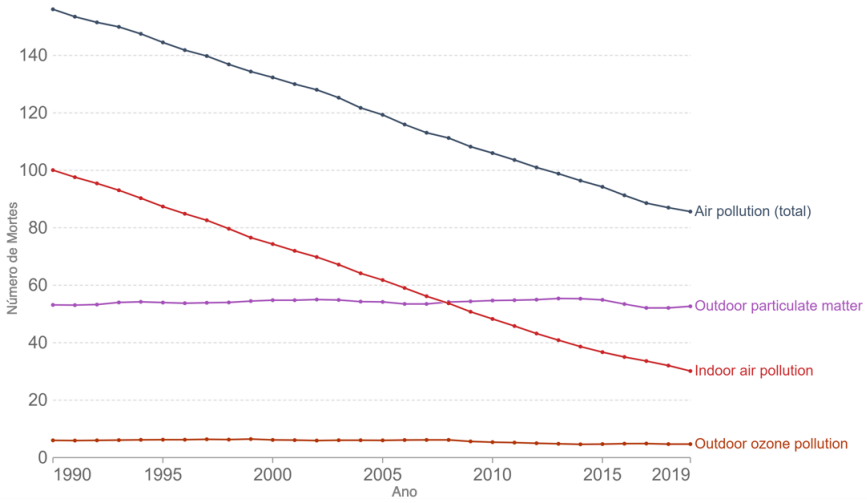


Figura 2.1: Taxas de mortalidade por poluição do ar no Mundo de 1990 a 2019 (Extraído de [1])

As taxas de mortalidade por poluição do ar *indoor* tiveram um declínio impressionante, enquanto as melhorias na poluição *outdoor* foram menos notórias.

Os impactos da poluição do ar vão ainda mais longe, sendo também um dos principais contribuintes para a carga global de doenças[46].

A carga global de doenças leva em consideração não apenas os anos de vida perdidos por morte precoce, mas também o número de anos vividos com saúde precária[42]. A carga de doenças é medida pelo número de anos de vida ajustados por incapacidade, Disability-Adjusted Life Years DALY (DALY). Um DALY representa um ano perdido de vida saudável [1].

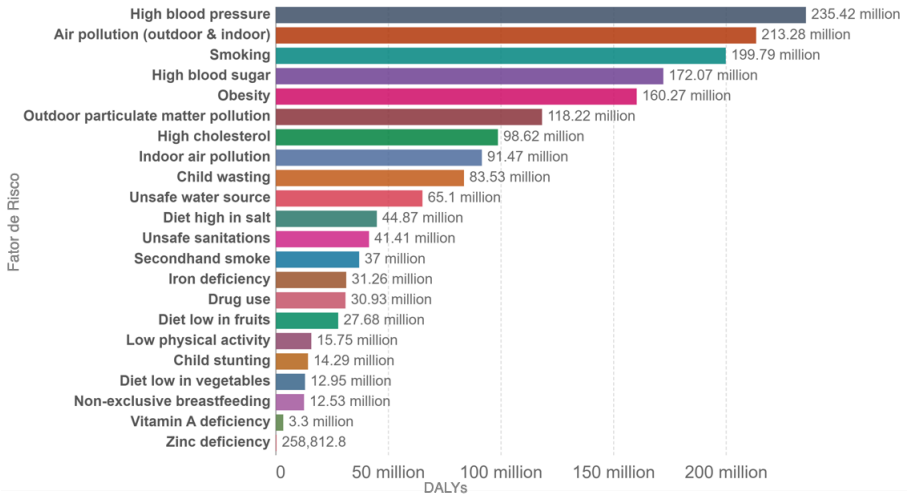


Figura 2.2: Carga da doença por fator de risco, Mundo, 1990 a 2019 (Extraído de [1])

Na Figura 2.2, vemos os fatores de risco classificados em ordem de DALY. Mais uma vez, a poluição do ar está perto do topo da lista, tornando-se um dos principais fatores de risco para problemas de saúde em todo o mundo [47].

As atividades industriais são cada vez mais pronunciadas nos países desenvolvidos, aumentando assim a necessidade de encontrar um equilíbrio entre sustentabilidade ambiental e económica [33]. Tudo isso gera emissões antrópicas, que colocam em risco o meio ambiente, causando impactos visíveis como problemas de saúde pública, mudanças climáticas, entre outros [48].

A poluição, em particular as PM_{10}, foram reconhecidas como uma séria ameaça à saúde humana e estima-se que cause entre 3 e 7 milhões de mortes por ano, causando principalmente doenças cardio-respiratórias [49].

As substâncias químicas que são introduzidas no meio ambiente e alteram a sua constituição natural e podem ser responsáveis por efeitos nocivos na saúde humana e/ou no ambiente são denominados de poluentes [50]. Os poluentes podem ser emitidos diretamente para a atmosfera, designando-se por isso poluentes primários, ou podem resultar de reações químicas que ocorrem na atmosfera entre os poluentes primários ou entre os poluentes e os seus constituintes naturais, designando-se por poluentes secundários [51].

Os efeitos nocivos dos poluentes na saúde humana e no ambiente dependem da sua concentração na atmosfera e do tempo de exposição [52]. Do ponto de vista ecológico pode-se distinguir vários tipos de poluentes atmosféricos, dependendo de como afetam o ecossistema [53].

- Poluentes que terão impacto direto na saúde humana, como NO_x, CO, SO_2 ou COV [1];
- Poluentes com impacto direto na vegetação, como NO_x, SO_2 e compostos químicos de Cloro (Cl) e Hidrogénio (H) [1];
- Poluentes que formarão ácidos, como NO_x e SO_2, principalmente quando encontrados em altas concentrações e ao mesmo tempo em atmosfera húmida [1];
- Poluentes persistentes com um longo ciclo de vida e acumulados no solo e por transferência através da cadeia biológica planta-animal-humano ou acumulações no corpo humano terão graves consequências para a saúde humana [1];
- Poluentes com influência direta sobre o clima, como CO_2 e Metano (CH_4) com grande impacto nas questões do aquecimento global [1].

Na Tabela 2.1 encontra-se sistematizada informação relativa a cada poluente, assim como as respetivas fontes emissoras e alguns dos efeitos que esses mesmos poluentes podem ter na saúde humana [13]. De um modo geral os poluentes evoluem de acordo com o ciclo sazonal, isto é, de forma periódica de acordo com as quatro estações do ano. Distinguem-se em duas partes, associadas, nomeadamente às alturas do ano mais quentes e mais frias [1].

- **Outono/Inverno:** Os poluentes primários como os NO_x, SO_2 e $PM_{2,5}$, apresentam por norma concentrações mais elevadas durante os períodos de Inverno, devido as condições atmosféricas de anticiclone que favorecem o fenómeno de inversão térmica e, por sua vez, acentua o efeito de acumulação dos poluentes ao nível do solo, sendo a sua dispersão e transporte vertical desfavorecidos [13]
- **Primavera/Verão:** As concentrações de O_3 são mais elevadas durante a Primavera e Verão em comparação com as de Outono e Inverno. Trata-se de um poluente secundário cuja produção está essencialmente relacionada com a intensidade da radiação solar [13].

2.2 O Índice da qualidade do Ar

O ar é um recurso vital e principal para a sobrevivência da vida humana, assim como a água e a terra [54]. O IQA é um indicador amplamente utilizado que traduz o estado da qualidade do ar ambiente num determinado local e permite, através de uma classificação expressa segundo uma escala de cores, dar à população uma orientação sobre o quão poluído o ar se encontra atualmente ou o quão poluído está previsto ficar [44]. Desta forma, permite que a população consiga adequar os seus comportamentos e ações de forma a proteger a saúde pública [49].

Tabela 2.1: Principais fontes e efeitos na saúde humana dos poluentes atmosféricos (adaptado de [13])

Poluente	Fontes emissoras	Efeitos na saúde humana
NO_2	O dióxido de azoto resulta da queima de combustíveis nas unidades industriais e da combustão, nos motores dos veículos automóveis.	Este poluente inibe algumas funções dos pulmões.
SO_2	O dióxido de enxofre provém da combustão dos combustíveis fósseis que contêm enxofre. É um gás que é emitido principalmente por fontes industriais e também pelo tráfego rodoviário.	Os seus efeitos encontram-se associados a doenças respiratórias (como a bronquite crónica e asma) e cardiovasculares.
PM_{10}	As partículas em suspensão provêm das cinzas e de outras partículas produzidas principalmente pela combustão de carvão e fuel-óleo na indústria e nos automóveis.	As partículas mais finas, conseguem penetrar no sistema respiratório, com consequências graves para a saúde.
Pb	Antes da utilização da gasolina sem chumbo, esta era a fonte responsável por 80% deste poluente na atmosfera.	Quando inalado distribui-se por todo o organismo, acumulando-se no tecido ósseo. É um metal pesado que produz envenenamento enzimático.
CO	O monóxido de carbono provém das emissões geradas pelos veículos a gasolina, e por alguns processos industriais.	Este poluente reduz a capacidade de transporte de oxigénio até aos tecidos vitais pelo sangue, afetando os sistemas cardiovascular e nervoso.
C_6H_6	O benzeno é utilizado como matéria-prima para síntese de compostos orgânicos e como aditivo nos combustíveis para veículos, substituindo, em parte, o chumbo.	Quando inalado, afeta principalmente o fígado, a placenta e a medula óssea. Causa também leucemia, cancro da pele e pulmão.
O_3	O ozono é um poluente secundário, resultando da transformação fotoquímica de certos poluentes primários na atmosfera.	O ozono é um poderoso oxidante, uma exposição crónica pode provocar dificuldades respiratórias.

A formulação do IQA apresenta-se como uma ferramenta importante, quer pelo seu poder de agregação de informação técnica, quer pela possibilidade de informar e consciencializar o público sobre os riscos da poluição e fazer com que as entidades competentes tomem as medidas necessárias [54].

É importante existir um acompanhamento deste valor, uma vez que, os riscos para a saúde aumentam na mesma proporção com que o IQA aumenta. Cada país assume diferentes padrões e limites da qualidade do ar, e consequentemente diferentes valores de IQA considerados aceitáveis [13].

Em Portugal, estão estabelecidos objetivos de qualidade do ar e limiares de informação e alerta para os níveis de qualidade do ar a curto e a longo prazo. Entende-se de curto prazo aqueles que são de hora em hora ou diários, e de longo prazo os anuais [13].

Em 2002 foi estabelecida a forma, o conteúdo e o método de cálculo do índice de qualidade do ar. Este, é calculado diariamente e disponibilizado pela Agência Portuguesa do Ambiente (APA), tendo por base a informação recolhida pelas Comissões de Coordenação e Desenvolvimento Regional (CCDR) e pelas Direções regionais do ambiente (DRA) nas regiões autónomas dos Açores e da Madeira a partir dos valores médios de concentração dos seguintes poluentes [13]:

- NO_2 – através das médias horárias;
- SO_2 – através das médias horárias;
- O_3 – através das médias horárias;

- CO – através das médias de 8 horas consecutivas;
- PM_{10} – através das médias diárias.

O IQA para um determinado dia e para uma determinada zona resulta da média aritmética calculada para cada um dos poluentes medidos em todas as estações da rede dessa área [53]. Os valores assim determinados são comparados com gamas de concentrações associadas a uma escala de cores sendo os poluentes com a concentração mais elevada os responsáveis pelo IQA [14].

O índice tem cinco classes que variam de "Muito Bom"a "Mau". Para que o índice possa ser calculado numa determinada zona/aglomeração é necessária a verificação das seguintes condições:

- Zonas - é obrigatória a medição dos poluentes ozono (O_3) e partículas PM_{10} ou partículas $PM_{2,5}$ (partículas de diâmetro igual ou inferior a 10 µm e 2,5 µm);
- Aglomerações - é obrigatória a medição dos poluentes NO_2 e partículas PM_{10} ou partículas $PM_{2,5}$ (partículas de diâmetro igual ou inferior a 10 µm e 2,5 µm); podendo incluir, quando disponível, o poluente SO_2.

Os intervalos de classificação do IQA estão alinhados com os valores estabelecidos na legislação e ainda com o referencial dos valores recomendados pela Organização mundial de saúde (OMS) [13]. A Tabela 2.2 evidencia a classificação das concentrações dos vários poluentes do ar (expressos em $\mu g/m^3$):

Classificação	PM_{10}	$PM_{2,5}$	NO_2	O_3	SO_2
Muito bom	0-20	0-10	0-40	0-80	0-100
Bom	21-35	11-20	41-100	81-100	101-200
Médio	36-50	21-25	101-200	101-180	201-350
Fraco	51-100	26-50	201-400	181-240	351-500
Mau	101-1200	51-800	401-1000	241-600	501-1250

Tabela 2.2: Classificação das concentrações de poluentes do ar (expressos em $\mu g/m^3$)

2.2.1 Material Particulado – PM_{10}

O ar que respirámos impacta bastante a nossa saúde e bem-estar, estando a qualidade do mesmo diretamente relacionada com o surgimento de alguns problemas de saúde, principalmente doenças cardio-respiratórias [55].

O material particulado, também designado muitas vezes por partículas inaláveis ou em suspensão têm por norma, um diâmetro inferior a 10 µm, e são consideradas partículas de elevado risco para a saúde, uma vez que, conseguem penetrar profundamente ao nível dos pulmões e atingir os alvéolos pulmonares, causando desta forma, perturbações no sistema respiratório [56]. Podem ser emitidas diretamente para o ar, designando-se assim, de partículas primárias, ou serem formadas na atmosfera por precursores

gasosos como SO_2, Óxidos de Nitrogénio (NxOy), Amónia (NH3) e Gases Orgânicos não-metano (GONM) [13].

Algumas das fontes mais prováveis para a existência destas partículas são as Instalações de Combustão, as atividades agrícolas, o tratamento de madeiras, as atividades industriais metalúrgicas, a exploração de minas, a construção civil e as refinarias [57].

A determinação da concentração mássica das partículas atmosféricas baseia-se no Decreto-Lei n.º 102/2010, de 23 de setembro. Este decreto estabelece valores limite para a concentração dessas partículas no ar ambiente e define as regras de gestão da qualidade do ar, alinhadas com as diretrizes da OMS, para evitar, prevenir ou reduzir as emissões de poluentes atmosféricos [58].

Os valores limite referenciados do Decreto-Lei anterior estão apresentados na Tabela 2.3:

Poluente	Período de Referência	Valor Limite	Margem Tolerância
PM_{10}	1dia(não exceder 35vezes/ano)	50 µg/m³	50%
	1 Ano Civil	40 µg/m³	20%

Tabela 2.3: Valores limite para PM_{10} (Anexo XII do Decreto-Lei n.º 102/2010, de 23 de setembro

2.2.2 Dióxido de Azoto - NO_2

O NO_2 é um dos principais poluentes atmosféricos. A exposição a altas concentrações pode traduzir-se em graves problemas na saúde, nomeadamente ao enfraquecimento da função pulmonar [16]. Ao ser oxidado na atmosfera, pode mesmo produzir o ácido nítrico, que é um dos componentes que aumenta a acidez da chuva, e causa vários estragos na natureza e materiais, por ser corrosivo [16]. Trata-se de um gás altamente tóxico, que resulta da queima de combustíveis fósseis a temperaturas bastante elevadas. É essencialmente, proveniente da circulação dos automóveis e também do sector industrial [59].

O ano de 2020 foi caracterizado por ter sido um ano atípico, devido à pandemia da COVID-19, e por ter sido um período marcado por uma forte diminuição das emissões de poluentes atmosféricos. Consequentemente, foi verificada uma melhoria significativa da qualidade do ar sobretudo nas zonas de tráfego mais intenso, como nos grandes aglomerados urbanos [51]. A redução das concentrações foi principalmente sentida nas concentrações de NO_2, poluente que tem como principal fonte o tráfego rodoviário e que é responsável pela generalidade das situações de poluição atmosférica verificadas nos últimos anos nas cidades de Lisboa, Porto e Braga [60]. Este comportamento pode ser observado através da Figura 2.3 que se segue:

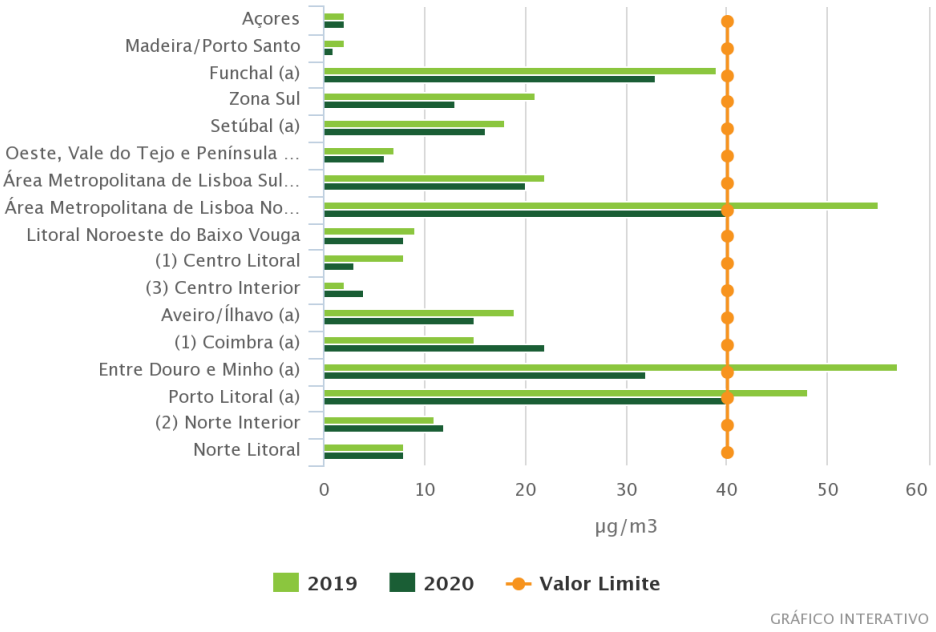


Figura 2.3: Excedências ao valor limite anual de NO_2 nas zonas e aglomerações que as monitorizam (estações de fundo, tráfego e industriais, em 2019 e 2020 (Extraído de [60]))

Verifica-se, pela primeira vez, a inexistência de excedências ao valor limite anual nas grandes aglomerações. É ainda possível observar, uma diminuição generalizada na ordem de grandeza dos níveis medidos, salientando-se contudo, na aglomeração de Coimbra um ligeiro acréscimo dos níveis medidos de NO_2 em relação a 2019 (de 15 µg/m3 para 22 µg/m3). Esta situação pode ser resultante da utilização, na avaliação de 2019, da estação urbana de fundo em vez da habitual estação de tráfego, que nesse ano, devido a avaria no equipamento, não esteve operacional [60]. De acordo com a APA, os valores de referência para a concentração de NO_2 no ar são os apresentados na Tabela 2.4.

Poluente	Período de Referência	Valor Limite	Margem Tolerância
NO_2	1 Hora	200µg/m³(sem exceder 18 vezes/ano)	50%
	1 Ano	40µg/m³	50%

Tabela 2.4: Valores limite para NO_2 (Anexo XII do Decreto-Lei n.º 102/2010, de 23 de setembro

2.2.3 Dióxido de enxofre - SO_2

O (SO_2) ou anidrido sulfuroso, é um gás proveniente, principalmente, da atividade humana, como processos de combustão de combustíveis contendo Enxofre (S), a produção de ácido sulfúrico, a produção de eletricidade ou a combustão de suporte a processos industriais, comerciais e residenciais [61].

Do ponto de vista natural o SO_2 tem origem sobretudo nas atividades vulcânicas e também na decomposição da matéria orgânica [62]. Para além disso, a emissão deste poluente na atmosfera está diretamente ligada à chuva ácida, que afeta o meio ambiente de diversas formas, particularmente, na reprodução e crescimento das plantas e no pH da água dos rios [60]. Paralelamente a estas situações,

também é bastante corrosivo, danificando assim, estruturas e monumentos nos grandes centros urbanos [13].

O SO_2 tem vários efeitos negativos na saúde e no ambiente que podem ir desde a irritação dos olhos, nariz e garganta a problemas mais graves do foro respiratório como lesões pulmonares, tosse e bronco-constrição [62]. Para além dos problemas acima mencionados, alguns estudos têm demonstrado que as plantas também sofrem com a exposição a este poluente, pois podem perder as suas folhas e consequentemente, ficarem menos produtivas, podendo, mesmo morrer prematuramente [60].

Desta forma, a exposição ao SO_2, mesmo que em alguns casos, em poucas concentrações pode comprometer o normal funcionamento do ecossistema [62]. De acordo com a APA, os valores de referência para a concentração de SO_2 no ar são os apresentados na Tabela 2.5.

Poluente	Período de Referência	Valor Limite	Margem Tolerância
SO_2	1 Hora	350µg/m³ (sem exceder 24vezes/ano)	43%
	1 Dia	125µg/m³ (sem exceder 3vezes/ano)	Nenhuma

Tabela 2.5: Valores limite para SO_2 (Anexo XII do Decreto-Lei n.º 102/2010, de 23 de setembro

2.2.4 Monóxido de Carbono - CO

O CO tem origem antropogénica na combustão incompleta de combustíveis fósseis ou de outras matérias orgânicas que contêm carbono na sua composição [13]. Emerge nas emissões da produção de eletricidade, da combustão nas indústrias, no comércio e nas residências de cada um de nós e ainda se encontra presente nos transportes com motores a combustão [63]. Para além de fontes ligadas ao ser humano, o CO pode surgir na natureza sob a forma de erupções vulcânicas e fogos florestais [1].

Em áreas urbanas, os transportes rodoviários são a principal fonte de CO e, como tal, as concentrações deste poluente vão oscilando de acordo com as variações de tráfego [63]. É ainda de realçar que este poluente é emitido em maiores quantidades quando os motores se encontram em rotações elevadas, ou seja, nas situações de pára-arranca ou de baixa velocidade de circulação [13], também bastante característico das grandes cidades. Do ponto de vista da saúde humana, destaca-se o efeito secundário mais característico do CO a hipóxia tecidual, que se traduz numa deficiência de oxigénio nos tecidos, devido à sua capacidade de se ligar à hemoglobina [64].

Para além dos pontos já mencionados, existem vários estudos que evidenciam que a exposição ao CO pode causar sequelas neurológicas nos seres humanos, e comprometer as capacidades neuro cognitivas e comportamentos anormais em crianças [56]. Este gás, quando se apresenta em concentrações altas, também apresenta outros efeitos, como o stress de oxigénio, a inflamação e a disfunção endotelial [65].

Assim, uma vez que este gás em concentrações elevadas, pode causar vários problemas para a população, surge constantemente a necessidade de tentar controlar os seus níveis no meio ambiente. Segundo a APA, os valores de referência para a concentração de CO no ar são os apresentados na Tabela 2.6

Poluente	Período de Referência	Valor Limite	Margem Tolerância
CO	Máx. Diário média 8H	10µg/m³	60%

Tabela 2.6: Valores limite para CO (Anexo XII do Decreto-Lei n.° 102/2010, de 23 de setembro

2.3 Inteligência Artificial

A IA representa um campo em constante evolução que utiliza sistemas computacionais com a capacidade de realizar tarefas que normalmente exigiriam inteligência humana. Essas tarefas incluem o reconhecimento de padrões, a tomada de decisões, a resolução de problemas e a interação com o ambiente [66]. Através de algoritmos avançados, a IA é aplicada numa ampla gama de setores, desde assistentes virtuais e reconhecimento de voz até diagnósticos médicos e veículos autónomos, transformando significativamente a forma como interagimos com a tecnologia e abordamos problemas complexos [67]. Os sistemas de IA são capazes de adaptar o seu comportamento, até certo ponto, através de uma análise das ações passadas e de um trabalho autónomo [66].

Tanto o ML como o DL são métodos da IA usados para treinar modelos, que conseguem ensinar máquinas como aprender a classificar dados [11]. No entanto, o rápido avanço da IA também levanta questões éticas e sociais. A busca por um equilíbrio entre a inovação tecnológica e a consideração de implicações éticas é crucial para maximizar os benefícios da IA e mitigar potenciais desafios [68]. A Figura 2.4 evidencia a relação entre a IA, ML e o DL.

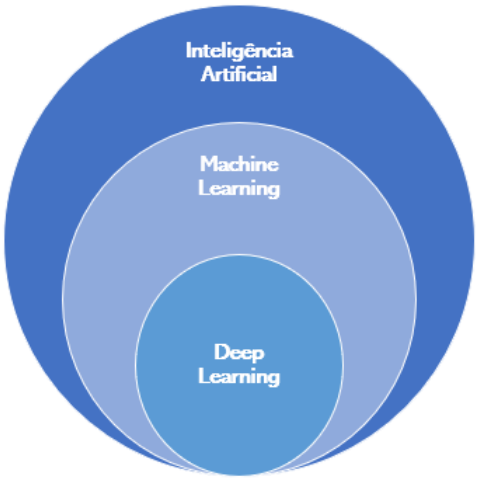


Figura 2.4: Relação entre Inteligência Artificial, IA, ML e o DL

2.3.1 Machine Learning

Num mundo cada vez mais digitalizado, o ML emerge como uma ferramenta essencial que permite aos sistemas aprenderem padrões a partir de dados e, com base nesse conhecimento, realizar tarefas complexas sem programação explícita [11]. Essa capacidade de automação e adaptação está cada vez mais a transformar radicalmente diversas indústrias, desde diagnósticos médicos até ao reconhecimento de voz e previsão de comportamentos [69].

As técnicas de ML não procuram reproduzir automaticamente o comportamento humano, mas sim utilizá-lo para complementar a inteligência humana, permitindo o desenvolvimento de tarefas que vão além das capacidades humanas [70].

Os sistemas de aprendizagem caracterizam-se por paradigmas da computação com capacidade de aprender em modo autónomo e independente [70]. Existem três tipos de aprendizagem, tal como ilustrado na Figura 2.5:

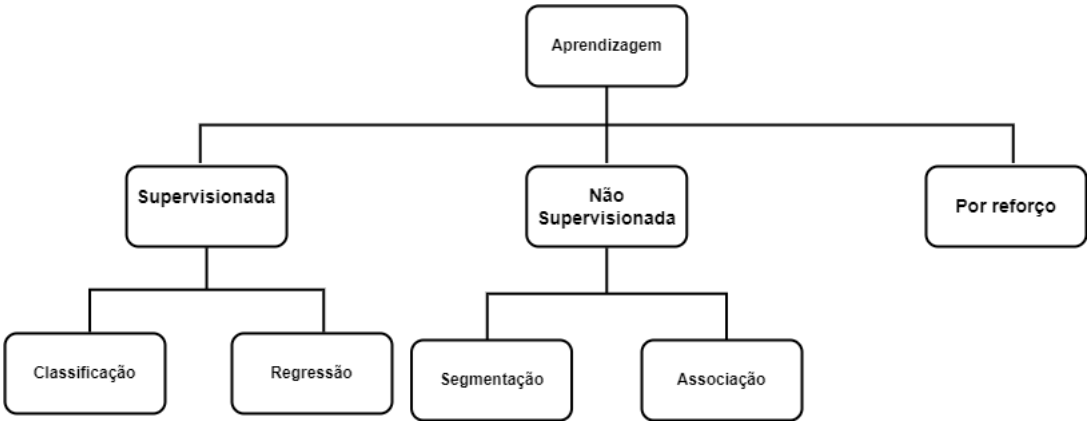


Figura 2.5: Tipos de aprendizagem

2.3.1.1 Aprendizagem Supervisionada

O termo de Aprendizagem Supervisionada (AS), é justificado pelo facto de este tipo de mapeamento ser acompanhado por um professor ou treinador que, tal como o próprio termo indica, supervisiona o processo de aprendizagem [71]. É um paradigma de aprendizagem em que os casos que se usam para aprender contém informação acerca dos resultados pretendidos, possibilitando assim, uma relação entre os valores pretendidos e produzidos pelo sistema [70]. Normalmente são utilizados quando os dados históricos estão disponíveis e o objetivo é prever resultados futuros [72].

Os algoritmos de ML utilizam maioritariamente este tipo de abordagem, onde os dados de entrada (x) e os respetivos resultados (y), possibilitam que o algoritmo aprenda com a função (f) de mapeamento dos dados nos resultados [19].

$$y = f(x) \tag{2.1}$$

Existem essencialmente duas categorias, para este tipo de abordagem [12]:

- Classificação: quando os resultados são discretos.
- Regressão: quando os resultados são contínuos.

Alguns tipos comuns de problemas construídos sobre classificação e regressão incluem recomendação e previsão de séries temporais, respetivamente. Alguns dos exemplos mais conhecidos de algoritmos de AS são de regressão linear para problemas de regressão, RF para problemas de classificação e regressão e Support vector machines (SVMs) para problemas de classificação [70].

2.3.1.2 Aprendizagem Não Supervisionada

Ao contrário do que acontece com AS, esta caracteriza-se por ser um paradigma onde á priori são desconhecidos os resultados sobre os casos. Apenas são conhecidos os resultados enunciados nos problemas, tornando necessário a escolha de técnicas de aprendizagem que avaliem o funcionamento interno do sistema [19]. Tem como objetivo a modelação da estrutura ou a distribuição dos dados do problema [70].

Assim, na Aprendizagem não Supervisionada (ANS) existem dados de entrada (x), mas não existem os resultados correspondentes. Por norma são divididos em duas categorias [70]:

- Segmentação: quando se pretende organizar os dados em grupos coerentes (Ex.:agrupar clientes que compram determinados artigos).
- Associação: quando se pretende conhecer regras que associem o comportamento.

2.3.1.3 Aprendizagem de Reforço

A Aprendizagem de Reforço (AR) caracteriza-se por um paradigma de aprendizagem que, apesar de não possuir informação sobre os resultados pretendidos, permite efetuar uma avaliação sobre a qualidade dos resultados produzidos [70]. O agente aprende a atingir um objetivo num ambiente complexo e incerto. Esta abordagem utiliza técnicas de auto-alimentação de sinais, com o objetivo de melhorar os resultados perante diversas situações. As soluções são encontradas por tentativa erro, uma vez que a máquina entende quais são os objetivos e recebe recompensas ou penalização pelas ações que realiza [19].

Existem dois grandes tipos de modelos caracterizados por esta abordagem [70]:

- Q-Learning: assume que está a seguir uma política ótima e usa-a para atualização dos valores das ações.
- State–action–reward–state–action (SARSA): considera a política de controlo que está a ser seguida e atualiza o valor das ações.

2.3.2 Deep Learning

O DL é um método de ML e funciona de maneira semelhante, no entanto, as suas capacidades são diferentes [73]. Embora os modelos básicos de ML se tornem progressivamente melhores no desempenho das suas funções específicas à medida que recebem novos dados, eles ainda precisam de alguma intervenção humana [74]. Se um algoritmo de IA retornar uma previsão imprecisa, um engenheiro precisará intervir e fazer ajustes. Com um modelo de DL, um algoritmo pode determinar se uma previsão é precisa ou não por meio de uma rede neuronal, não sendo necessária nenhuma ajuda humana [75].

Uma forma de compreender a diferença entre ML e DL é que o ML usa algoritmos para analisar dados, aprender com esses dados e tomar decisões informadas com base no que aprendeu [76]. Em contra partida, o DL estrutura algoritmos em camadas para criar uma Redes Neuronais Artificiais (RNA) que pode aprender e tomar decisões inteligentes de forma independente [77].

2.3.3 Modelos de Machine Learning e Deep Learning

Tal como foi anteriormente mencionado, existem vários tipos de modelos de ML e DL. De seguida serão abordados alguns desses modelos [78]. Dentro do ML, serão abordados problemas de AS, nomeadamente algoritmos de regressão e classificação, como o modelo de AD. Relativamente a modelos de DL, serão abordados modelos de RNA e RNR [79].

Existem diferentes tipos de modelos de ML, como Regressão Linear, AD, SVMs, Naive Bayes, K-Nearest Neighbors (K-NN), RNA, RF, Gradient Boosting, CNN e RNR.

Por outro lado, o DL é uma subárea do ML que se baseia em RNA profundas [80]. Essas redes são compostas por várias camadas de unidades de processamento interligadas, que aprendem automaticamente a partir dos dados [81]. O DL tem se vindo a destacar em várias areas, tais como em tarefas de visão computacional, processamento de linguagem natural e reconhecimento de voz [80]. Alguns dos modelos de DL mais conhecidos incluem CNN e RNR, como a LSTM e a GRU.

Este tipo de modelos são capazes de lidar com grandes volumes de dados, identificar padrões complexos e gerar *insights* preciosos. Além disso, estes modelos têm a capacidade de aprender e se adaptar a novos dados, tornando-os extremamente poderosos em problemas de previsão, classificação, agrupamento e geração de recomendações [79]. No entanto, é importante ressaltar que a construção e o treino desses modelos requerem um conhecimento profundo em algoritmos de ML, bem como uma compreensão dos dados e das características específicas do problema em questão [82].

2.3.3.1 Árvores de decisão

Os algoritmos mais conhecidos para construção de AD são o Iterative Dichotomiser 3,C4.5 (ID3) e o Classification and Regression Trees (CART). O algoritmo ID3 foi desenvolvido por *J. Ross Quinlan*, durante os anos 70, caracterizado por uma construção simples de árvores desde a raiz até às folhas [83]. Este algoritmo apresentou uma limitação, uma vez que apenas era capaz de lidar com variáveis nominais e, consequentemente, ser utilizado em problemas de classificação. Neste sentido, *Quinlan* criou um novo algoritmo, o C4.5, que seria, assim, uma melhoria do ID3 [83]. Portanto, através da junção de variáveis numéricas contínuas, era então possível resolver problemas de regressão.

Outro algoritmo desenvolvido por um grupo de estatísticos, como o *L. Breiman, J. Friedman, R. Olshen e C. Stone*, em paralelo com o mencionado anteriormente, é o modelo CART [84]. Este é um algoritmo com uma abordagem muito semelhante ao C4.5, usado também para modelos de classificação e regressão. Normalmente, uma AD adota a estrutura representada em 2.6:

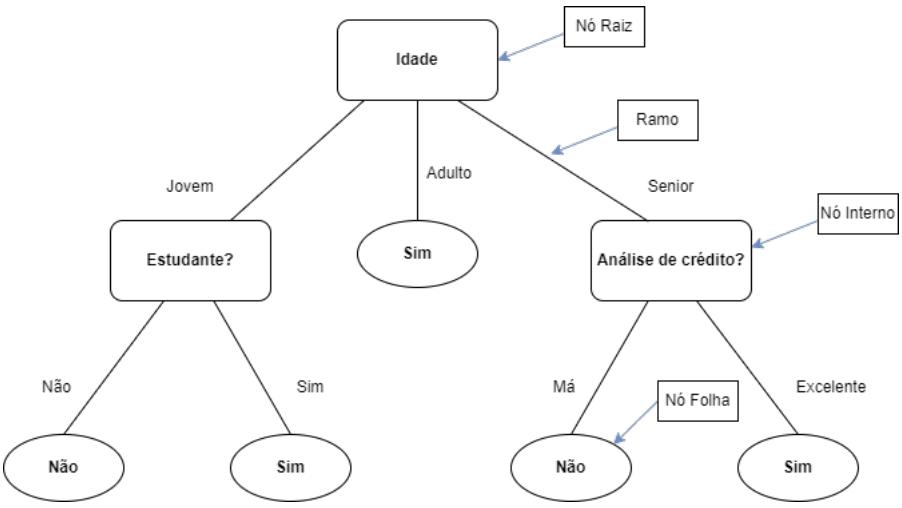


Figura 2.6: Exemplo de AD, referente a atribuição de crédito (adaptado de [85])

Dos principais constituintes de uma árvore, destacam-se:

- **Nó raiz**: representa uma variável e é o primeiro nó da AD.
- **Ramo/Ramificação**: Representam os resultados de cada teste executado num nó. Cada teste é feito tendo por base as condições "se-então (if-then)"e o resultado de cada teste é usado para decidir que ramo se segue.
- **Nó interno**: representa também uma variável e envolve todos os nós da árvore, exceto o nó raiz e as folhas.
- **Nós Folha**: Estes são os nós terminais da AD. Quando um recurso atinge o nó terminal, é-lhe atribuída uma classe.

Tal como foi mencionado anteriormente, o modelo de AD adota uma construção *top-down*, da raiz para as folhas. Cada nó de decisão contém um teste num atributo. Cada ramo descendente corresponde a um possível valor desse atributo. Por sua vez, cada folha está associada a uma classe e cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação [86].

O procedimento para a construção de uma árvore é descrito abaixo:

1. A árvore começa com um único nó, *D*, que representa todo o conjunto de dados de treino. As instâncias (exemplos) são representadas por pares atributo-valor.
2. Se todos os exemplos de *F* pertencem à mesma classe *C*, o nó *D* torna-se uma folha com essa classe atribuída.
3. Caso contrário, o algoritmo define o critério de divisão de acordo com o método de seleção de atributos escolhido. O critério de divisão define:
    a) Que atributo no nó *D* define a melhor maneira de dividir o conjunto de dados de treino em classes individuais;

b) Que ramos crescem do nó *D* de acordo com os resultados dos testes realizados no nó anterior;

4. O nó *D* é marcado com o critério de divisão. Cada ramo que cresce a partir deste nó é um dos resultados do critério de divisão. Seja *X* o atributo de divisão, com *v* valores distintos,x_1,x_2,x_3,x_v ,consideremos três possíveis cenários de divisão:

   a) **X como um valor discreto:** Os resultados do teste no nó *D* são todos os valores diferentes possíveis de *X*. A partição F_j é o subconjunto de *F* que contém todas as instâncias com o valor x_v, de *X*. Uma vez que todas as instâncias em cada partição F_j têm o mesmo valor para *X*, não há mais necessidade de particionar. Por esta razão, *X* é removido da lista de atributos. As Figuras 2.7 e 2.8 evidenciam um exemplo de variáveis discretas:

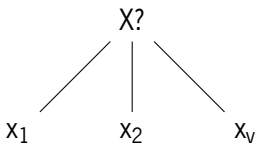


Figura 2.7: Cenário - Variável Discreta

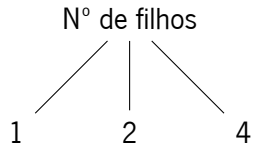


Figura 2.8: Exemplo - Variável Discreta

   b) **X como valor contínuo:** Neste caso, existem duas possibilidades como resultados do teste: *X ≤ valor e X > valor* . O ponto de divisão é definido pelo método de seleção de atributo escolhido. Dois ramos resultam do nó *D* e cada um representa o resultado possível. As instâncias são divididas de tal forma que F_1 é um subconjunto que respeita a condição *X ≤ valor* e o subconjunto F_2 as restantes, como mostram as Figuras 2.9 e 2.10.

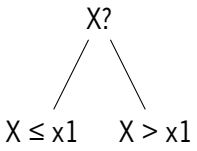


Figura 2.9: Cenário - Variável Contínua

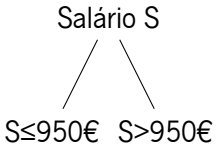


Figura 2.10: Exemplo - Variável Contínua

   c) **X como um valor discreto que pode gerar uma árvore binária:** Neste caso específico o teste realizado no nó *D* segue a condição $X \in S_x$, onde S_x representa um subconjunto de *X*. Um valor de *X*,x_v, é testado da seguinte maneira: se $x_v \in S_x$, então o teste no nó D é positivo. São criados dois ramos a partir do nó e, por convenção, o ramo esquerdo fica designado de "sim"e o direito, de "não". Como resultado, existem dois subconjuntos, F_1, correspondendo ao teste com designação "sim" e F_2, designado de "não". Podemos observar os exemplos das Figuras 2.11 e 2.12.

5. O processo é interrompido e as classes são atribuídas quando (casos de paragem):

   - Todos os exemplos de *F* pertencem à mesma classe *C*;

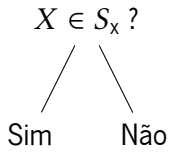


Figura 2.11: Cenário - Variável Binária

Figura 2.12: Exemplo - Variável Binária

- Não há atributos esquerdos nos quais as instâncias podem ser particionadas. Neste caso específico, utiliza-se a votação maioritária, e o nó torna-se uma folha e é designado como a classe mais frequente;
- Não há instâncias restantes, ou seja, a partição F_j está vazia.

As medidas de seleção de atributos determinam como as instâncias de um determinado nó devem ser divididas e, por isso, também são denominadas por regras de divisão [87]. Assim, existem três medidas de seleção de atributos, descritas a seguir:

- **Ganho de Informação**

  O ganho de informação deve ser calculado para cada atributo. O atributo que resultar no maior ganho de informação é selecionado como atributo de teste [87].

  Trata-se de uma medida de seleção de atributo que tem em consideração a entropia no seu cálculo. A entropia é uma medida de incerteza que permite identificar o grau de desordem dos dados, e descobrir que atributo deve ocorrer em cada posição da árvore [87].A fórmula matemática para Entropia é apresentada pela Equação 2.2:

$$E(F) = -\sum_{i=1}^{m} \pi log_2 \pi \tag{2.2}$$

  Onde *p* representa a probabilidade de cada classe aparecer num conjunto de dados, calculado sobre um total de *m* classes. A entropia varia num intervalo entre 0 e 1, onde valores mais próximos de 1 significam um alto nível de desordem de dados e um baixo nível de pureza [87]. Em contrapartida, valores mais próximos de 0 significam o oposto. Se os valores estiverem mais próximos de 0 isso representa, também, que os dados estão distribuídos de forma homogénea, não havendo predominância de classes nesse conjunto de dados. O gráfico da Figura 2.13 mostra como a entropia varia de acordo com a probabilidade *(p)* de cada classe aparecer no conjunto de dados [86].

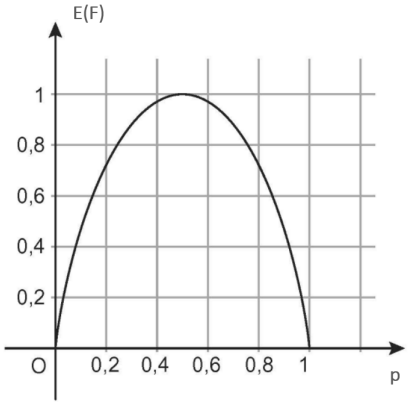


Figura 2.13: Visão gráfica da entropia

O atributo com maior redução de entropia e, consequentemente, maior ganho de informação é considerado aquele que deve ocupar a raiz da árvore, uma vez que produz um efeito melhor. Ocupar a posição de raiz permite reduzir a profundidade da árvore [72]. O cálculo do ganho de informação é dado pela Equação 2.3:

$$Ganho(X) = E(F) - \sum_{j=1}^{v} \frac{|F_j|}{|F|} \times E(F_j) \tag{2.3}$$

Assim, pela análise da fórmula, o ganho de informação resulta da diferença entre o requisito original de informação (antes da divisão) e o novo requisito (após a divisão). A medida de ganho de informação tende a ser distorcida em relação a testes com vários resultados. Neste caso, cada partição é pura, uma vez que cada uma contém apenas uma instância de dados e, consequentemente, o ganho de informação por ser maximizado [72]. Certamente, esta situação seria inútil para problemas de classificação.

- **Razão de ganho**

Esta medida de seleção de atributos procura resolver algumas das limitações e especificidades do ganho de informação, sendo, assim considerada uma extensão do ganho de informação. A Equação 2.4, mostra a informação potencial que é criada através da divisão do conjunto de dados de treino [84].

$$PotencialX(F) = -\sum_{j=1}^{v} \times log_2 \times \frac{|F_j|}{|F|} \tag{2.4}$$

Finalmente, a taxa de ganho é calculada de acordo com a seguinte Equação 2.5

$$RazaoGanho(X) = \frac{Ganho(X)}{PotencialX(F)} \tag{2.5}$$

- **Índice Gini**

  Esta medida é aplicada no algoritmo CART. É usado para medir a impureza de um conjunto ou subconjunto de dados. O índice de Gini é calculado através da Equação 2.6 [86]:

$$Gini(F) = 1 - \sum_{i=1}^{m} \times p_i{}^2 \tag{2.6}$$

  Onde p_i representa a probabilidade de uma instância *D* em pertencer à classe *C* , num total de m classes. Esta medida contempla uma divisão binária para cada atributo. Há duas possibilidades a serem consideradas: o atributo como valor discreto ou como valor contínuo [86].

  Considerando o caso em que *X* é um valor discreto e existem valores distintos para esse atributo, analisamos todos os subconjuntos potenciais que podem ser gerados usando os valores de X. Cada subconjunto, S_x , pode ser considerado como um teste binário para o atributo *X* na forma "$X \in SX$?".

  Assim, este teste é satisfeito se o valor de *X* para a instância estiver entre os valores de S_x. Ao considerar uma divisão binária, calcula-se uma soma ponderada da impureza de cada partição resultante. O Índice de Gini de *D*, passa a ser representado pela Equação 2.7 [86]:

$$GiniX(D) = \frac{|F_1|}{|F|} \times Gini(D_1) + \frac{|F_2|}{|F|} \times Gini(D_2) \tag{2.7}$$

  Por outro lado e, para atributos de valor contínuo, os pontos de divisão devem ser considerados. A estratégia do ponto médio, é usada como um possível ponto de divisão. Para um possível ponto de divisão de *X*, D_1 é o conjunto de instâncias em D que satisfaz a condição $X \leq$ valor e D_2 é o conjunto de instâncias em *D* que satisfaz $X > valor$. O índice de Gini é apresentado na Equação 2.8 [86]:

$$\Delta Gini(X) = Gini(D) - GiniX(D) \tag{2.8}$$

  O atributo que maximiza o valor mínimo do índice de Gini, ou seja, aquele que maximiza a redução de impureza, é designado de atributo de divisão [86].

  Outra questão importante sobre AD é a poda de árvores. Este processo é caracterizado pela exclusão de nós internos de uma AD. Os métodos de poda foram desenvolvidos para permitir que um modelo de AD pare quando não é possível a sua divisão [72]. Isso acontece uma vez que pode ocorrer *overfitting*. Esta situação ocorre quando o modelo aprende demasiado sobre os dados, mostrando-se adequado apenas para os dados de treino, como se o modelo tivesse apenas decorado os dados de treino e não fosse capaz de generalizar para outros dados nunca antes vistos [86].

Assim, a poda das árvores permite que dados com ruído ou redundâncias sejam eliminados do modelo. O processo de poda de árvores pode utilizar duas técnicas diferentes: a pré-poda e a pós-poda [86]. A primeira diz respeito ao facto de que o processo de construção da árvore é interrompido quando se considera que não há vantagens em ter mais divisões na árvore. No entanto, a pós-poda baseia-se em deixar a árvore crescer o máximo possível, e de seguida podar os nós desnecessários [72]. Cortar um nó significa remover a sub-árvore a partir daquele nó, tornando-o num nó-folha [86].

As AD são um modelo rápido, simples, intuitivo e barato [72]. Ademais, o modelo fornece uma boa visão sobre que atributos são mais significativos. Em contrapartida este é um modelo que não produz bons resultados se houver muitas variáveis no conjunto de dados. Além disso, conjuntos de dados mais complexos geram árvores também mais complexas. Assim, apesar de simples de implementar, árvores grandes possuem inúmeras divisões e muitas sub-árvores, o que leva muito tempo e pode ser computacionalmente caro [86].

2.3.3.2 Random Forest

O próprio nome do algoritmo explica bastante bem o funcionamento do mesmo, isto é, a criação de muitas AD, de forma aleatória, formando o que podemos chamar de floresta. Cada árvore será utilizada na escolha do resultado final. É considerado um modelo com resultados e previsões de alta precisão [88].

Ao contrário do que acontece na criação de uma AD simples, ao utilizar o método RF, o primeiro passo a executar será a seleção aleatória de algumas amostras de dados de treino, e não a sua totalidade [88]. Nesta etapa é utilizado o *bootstrap*, que é um método de re-amostragem onde as amostras selecionadas podem ser repetidas na seleção [86]. Com esta primeira seleção de amostras será construída a primeira AD.

Conforme verificado nos detalhes sobre construções de AD, para iniciar o processo é necessário definir o primeiro nó da árvore, aquele que tipicamente é designado de nó raiz [86]. Esta será a primeira condição verificada, originando assim, os dois primeiros ramos. Seguidamente, é utilizando o método de entropia ou o método do índice de Gini, será escolhida a melhor variável para definir o nó raiz, variando de acordo com o método utilizado [86]. A Figura 2.14 evidencia o funcionamento do modelo de RF, para o caso de um exemplo de Classificação:

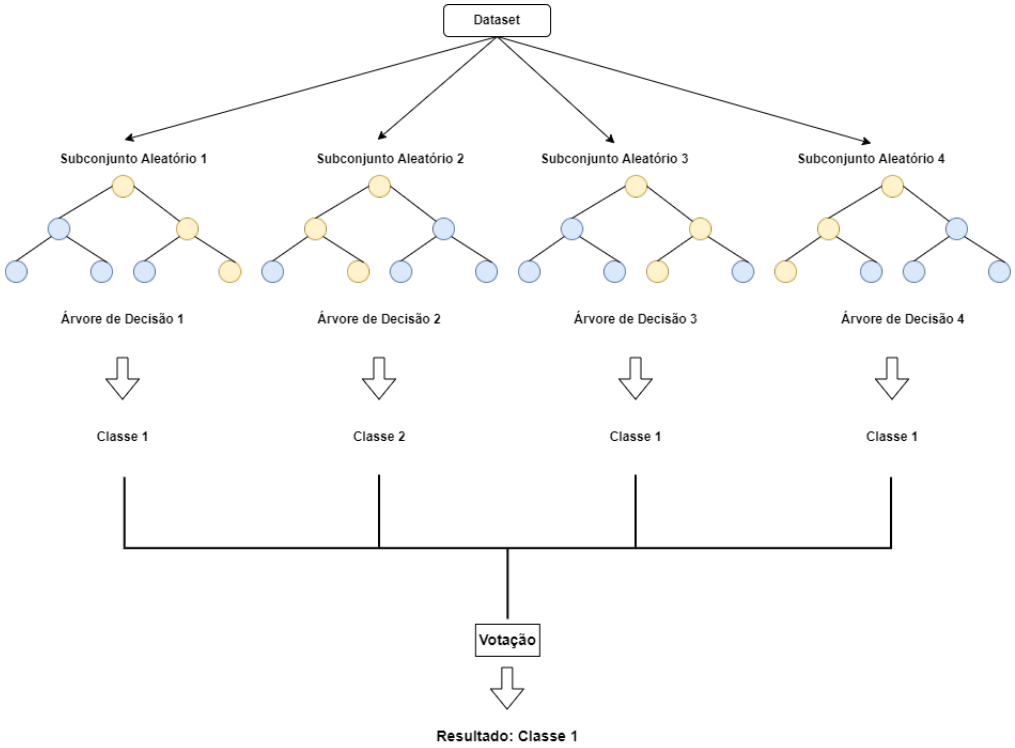


Figura 2.14: Exemplo do algoritmo RF

No algoritmo de RF a definição desta variável não acontece com base em todas as variáveis disponíveis. O algoritmo escolhe de maneira aleatória duas ou mais variáveis, e dessa forma, efetua os cálculos com base nas amostras selecionadas, para definir qual dessas variáveis será utilizada no primeiro nó. Para a escolha da variável do próximo nó, serão novamente escolhidas duas, ou mais variáveis, excluindo as já selecionadas anteriormente, e o processo de escolha vai se repetindo [88]. Desta forma, a árvore será construída até ao seu último nó. Este não é o melhor método para construção de uma AD, pois, o algoritmo pode, selecionar as duas piores variáveis na primeira seleção, escolhendo uma variável péssima para o primeiro nó. Por outro lado, como vão ser construídas várias árvores, essa estratégia acaba por se tornar bastante eficiente, e normalmente consegue evitar o *overfitting* [86].

Na construção da próxima árvore, os dois processos anteriores voltam a repetir-se, levando à criação de uma nova árvore. Provavelmente essa árvore será diferente da primeira, pois tanto na seleção das amostras, como na seleção das variáveis, o processo acontece de maneira aleatória [88].

Podem ser construídas as árvores que se pretender, sendo que quanto mais árvores criadas, melhor serão os resultados do modelo, até determinado ponto. Em contrapartida, quanto mais árvores forem criadas, maior será também o tempo de criação do modelo [88].

Cada árvore criada apresenta o seu respetivo resultado, sendo que em problemas de regressão será realizada a média dos valores previstos, e em problemas de classificação o resultado que mais vezes for apresentado será o escolhido, dizendo assim respeito à moda. [72].

2.3.3.3 Redes Neuronais Artificiais

As RNA, também designadas por sistemas conexionistas, são modelos simplificados do sistema nervoso central do ser humano [89].Trata-se de uma estrutura extremamente conectada de unidades computacionais, frequentemente designadas por neurónios ou nodos, com capacidade de aprendizagem [89].

Nos modelos de RNA, os neurónios são responsáveis por processar as informações. Assim, os sinais gerados a partir de cada neurónio são transmitidos para os próximos neurónios, através de uma ligação de comunicação [90]. Essas ligações possuem um peso, responsável por multiplicar os sinais transmitidos. Para criar uma saída, cada neurónio aplica uma função de ativação em cada entrada [71]. O processo de treino da RNA consiste na atualização dos pesos ao longo de iterações até produzir resultados significativos. Essas iterações são chamadas de épocas e são um parâmetro importante nos modelos de RNA [72]. Assim, uma RNA é constituída por camadas de neurónios, que estão conectadas entre si [89].

Um nodo, termo usado para distinguir um neurónio natural e artificial, é a unidade de processamento chave para a operação de uma RNA [76]. Embora existam vários tipos de nodos, este, comporta-se como um comparador que produz uma saída quando o efeito cumulativo das entradas excede um dado valor limite [91].

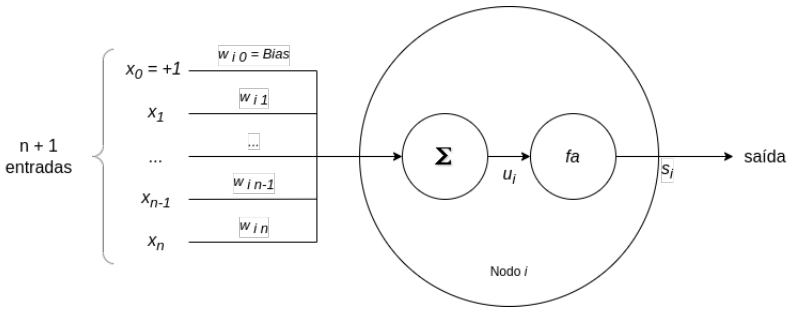


Figura 2.15: Estrutura geral de um nodo

Um nodo pode ser dissecado de acordo com o exposto na Figura 2.15, que contém:

- Um conjunto de conexões (w_{ij}),cada uma identificada com um peso, ou seja, um número real ou binário, que tem um efeito excitatório para valores positivos e inibitório para valores negativos. Assim o sinal ou estimulo (x_i), como entrada da conexão, é multiplicado pelo seu peso correspondente w_{ij}, onde i denota o nodo objeto de estudo e j o nodo de onde partiu o sinal [91]. Adicionalmente, pode também existir uma conexão extra, denominada de *bias*, cuja entrada toma o valor +1, que estabelece uma certa tendência ou inclinação no processo computacional, ou seja, adiciona uma constante (w_{i0}) para que se estabeleçam as corretas condições operacionais para o nodo [91].

- Um integrador (g) que reduz os *n* argumentos de entrada (estímulos) a um único valor. Frequentemente é utilizada a função de adição $\sum$,pesando todas as entradas numa combinação linear.

- Uma função de ativação (f), que pode condicionar o sinal de saída, introduzindo uma componente de não linearidade no processo computacional.

Formalmente, este nodo pode ser descrito através das Equações 2.9 e 2.10:

$$u_i = g(1 \times w_{i0}, x_1 \times w_{i1}, x_2 \times w_{i2}, ..., x_n \times w_{in}) \tag{2.9}$$

$$S_i = f(u_i) \tag{2.10}$$

Onde para cada nodo *i* com *n* entradas e uma saída, u_i representa o ganho do nodo i e s_i a saída do nodo.

A arquitetura ou topologia das RNA diz respeito à forma como os nodos se interligam numa estrutura de rede. Existem inúmeros tipos de arquiteturas de RNA, que normalmente se subdividem em duas categorias [91]:

- RFSC: na sua forma mais simples uma rede é composta por uma camada de entrada, cujos valores de saída são fixados externamente, e por uma camada de saída, tal como pode ser observado na Figura 2.16 [91]:

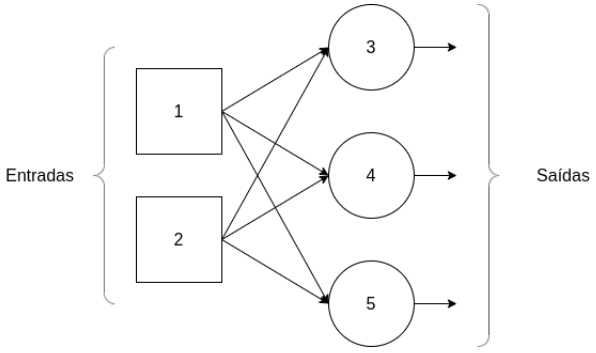


Figura 2.16: Exemplo de uma RFSC

- RFMC: estas redes distinguem-se das primeiras, devido ao facto que possui uma ou mais camadas intermédias. A função destes nodos intermédios é intervir de forma útil entre a entrada e a saída da rede [91]. Ao se acrescentarem camadas intermédias esta-se a aumentar a capacidade da rede em modelar funções de maior complexidade. Contudo, o tempo de aprendizagem aumenta de forma exponencial, tal como evidenciado na Figura 2.17 [76].

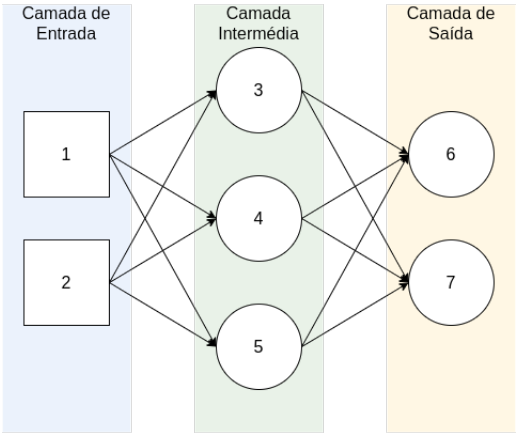


Figura 2.17: Exemplo de uma RFMC

O processo de *feedforward* envolve o fluxo de informações da camada de entrada para a camada de saída, enquanto que o *backforward* ocorre da camada de saída para a camada de entrada. Matematicamente o processo *feedforward* pode ser representado pela Equação 2.11 [72].

$$\begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} \times \begin{bmatrix} w11, w12, ..., w1i, ..., w1n \\ w21, w22, ..., w2i, ..., w2n \\ ..... \\ wj1, wj2, ..., wji, ..., wjn \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{bmatrix} = \begin{bmatrix} x_1 \times w11 + x_2 \times w12 + ... + x_n \times wj1 + b1 \\ x_1 \times w12 + x_2 \times w22 + ... + x_n \times wj2 + b2 \\ ..... \\ x_1 \times w1n + x_2 \times w2n + ... + x_n \times wjn + bj \end{bmatrix} \tag{2.11}$$

Onde a primeira matriz diz respeito aos dados de entrada,a segunda corresponde aos pesos, e a terceira ao *bias*. A combinação das três resulta no processo descrito acima. Onde para cada j neurónios:

$$O_j = \sum_{i=1}^{n} w_j i x_i + b_j, \forall j \tag{2.12}$$

O b_j corresponde ao *bias*, w_{ij} diz respeito aos pesos das ligações e x_i representa os dados de entrada. O *bias* funciona como uma adição, permitindo que a saída (que vem da função de ativação) mova os seus valores para a esquerda ou para a direita, no eixo das abcissas [72].

2.3.3.4 Redes Neuronais Recorrentes

As redes neuronais tradicionais não utilizam informações do passado para entender as situações futuras. Nessas circunstâncias, quando uma entrada chega a um neurónio, é aplicada uma função de ativação, dando origem a uma saída. Então, essa saída será transmitida para o neurónio da próxima camada, sem armazenar nenhuma informação no neurónio atual [89]. As RNR possuem uma particularidade, usando *loops* [91]. Além disso, neste tipo de redes neuronais, os dados sequenciais de comprimento diferente podem ser processados. Aqui, em vez de gerar a saída e apenas transmiti-la ao próximo neurónio, o

neurónio RNR transmitirá essa informação para o neurónio da próxima camada, mas também para si mesmo novamente [91].

Ao contrário das redes neuronais *feedforward*,as RNR podem usar o seu estado interno (memória) para processar sequências de entradas. Este fenómeno potencia tarefas como o reconhecimento de texto manuscrito conectado e não segmentado ou reconhecimento de fala [91].

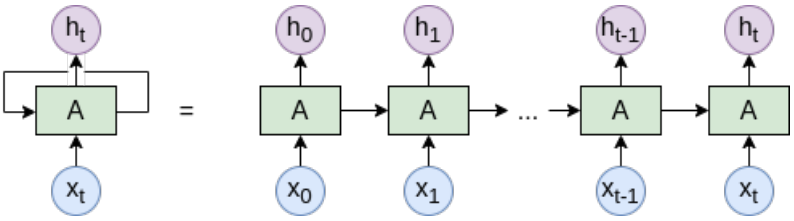


Figura 2.18: Exemplo de uma RNR

Analisando a Figura 2.18, verifica-se que inicialmente se assume que X_0 é a sequência de entrada que produz h_0, que por sua vez em conjunto com X_1, forma a entrada para a etapa seguinte. Da mesma forma, h_1 é a entrada com X_2 para a próxima etapa, e assim sucessivamente. A fórmula para o estado atual é dada pela expressão:

$$h_t = f(h_{t-1}, x_t) \tag{2.13}$$

Onde a função de ativação:

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{2.14}$$

Aqui, W é o peso , h é o único vetor oculto, W_{hh} é o peso no estado oculto anterior, W_{hx} é o peso no estado de entrada atual, ?? é a função de ativação, que implementa uma não linearidade e comprime as ativações para o intervalo [- 1,1]:

$$y_t = W_{hy}h_t \tag{2.15}$$

Onde, Y_t é o estado de saída.

A recorrência existe em sistemas dinâmicos quando uma saída de um elemento influencia, de algum modo, a entrada para esse mesmo elemento, criando, assim, um ou mais circuitos fechados tal como apresentado na Figura 2.18. Ao conter ciclos, as saídas não são função exclusivamente das conexões entre nodos, mas também de uma dimensão temporal, ou seja, esta-se na presença de um cálculo recursivo, que obedecerá naturalmente a uma certa condição de paragem, com a última iteração a ser dada como saída para o nodo [91].

Quando comparada com os problemas de regressão e classificação, as séries temporais adicionam a dependência temporal entre as observações [92]. Além disso, é importante referir que os modelos RNR podem dar boas respostas quando há uma pequena lacuna entre as informações relevantes e o que é necessário no momento. Pelo contrário, quando a lacuna entre a informação relevante e o momento em que é necessária é muito grande, as RNR tornam-se um modelo ineficiente [92]. Para corrigir isso, em

1997, *Hochreiter* e *Schmidhuber* introduziram o modelo LSTM, capaz de lidar com "dependências de longo prazo".

LSTM

Um exemplo de RNR é a arquitetura LSTM. Este tipo de RNR é um método de DL, capaz de aprender dependências de longo prazo entre entradas e saídas [93]. A LSTM é adequada para classificar, processar e prever séries temporais com intervalos de tempo de duração desconhecida. Nas células LSTM, existem estruturas chamadas portas. Estas são as estruturas que permitem remover ou adicionar informações à célula [94]. Estas, por sua vez, utilizam uma função sigmoide e, consequentemente, os seus valores resultantes variam entre 0 e 1. Um valor igual a 1 representa a informação que pode passar e é importante, por outro lado, 0 representa a informação que não pode ser armazenada, uma vez que não é útil [95].

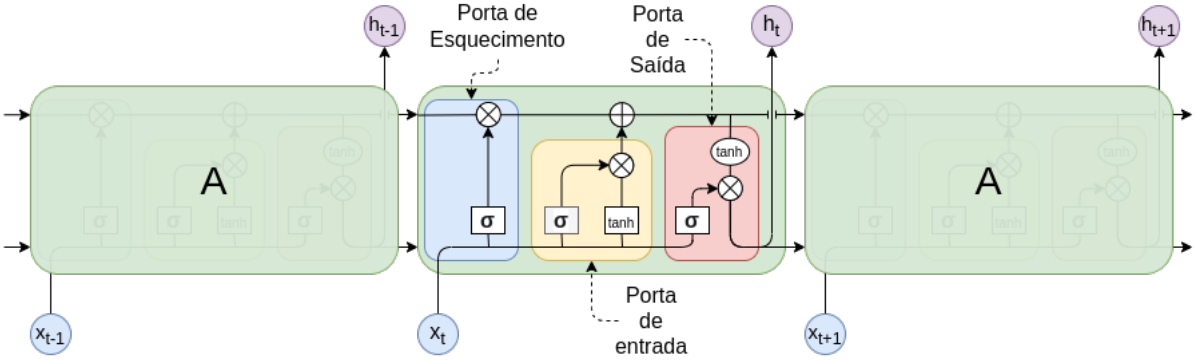


Figura 2.19: Exemplo de uma LSTM

As três portas são essencialmente neurónios com pesos que controlam o fluxo de informação dentro da célula LSTM [96]. Especificamente, a porta de esquecimento determina que estado deve ser retido na célula [97]. Por sua vez, a porta de entrada controla a extensão em que uma nova entrada flui para a célula e, por fim, a porta de saída decide que quantidade de memória é usada para produzir a saída [93]. Em conjunto, estas portas permitem que o modelo LSTM aprenda com os dados de sequência de longa duração com mais eficiência que o RNR convencional [98].

Assim, uma célula LSTM funciona seguindo quatro passos fundamentais. Primeiro, usa-se h_{t-1} e x_t como entradas, e é aplicada a função sigmóide, criando um valor entre 0 e 1, onde, 0 representa que nenhuma informação é adicionada à memória daquele bloco de células. Este processo pode ser descrito segundo a Equação 2.16 [80].

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{2.16}$$

Depois disso, usando h_{t-1} (saída anterior) e o x_t, a entrada deste estado, é aplicada uma função Tangente Hiperbólica (TanH), criando uma entrada candidata a ser adicionada, C_t, como a seguinte equação, evidencia:

$$\hat{c}_t = \tanh(W_C \times [h_{t-1}, x_t] + b_C) \tag{2.17}$$

Assim, são determinadas as informações que são mantidas na célula e as que são eliminadas, de acordo com a Equação 2.18.

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{2.18}$$

Seguidamente, o modelo LSTM precisa definir qual é a saída. Primeiro usando h_{t-1} e o x_t, uma função TanH é aplicada, na última porta, a porta de saída, seguindo a Equação 2.19.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{2.19}$$

Por fim, para criar a saída, é usado o C_t, (o estado real da célula por meio de uma função TanH, para fornecer uma saída entre -1 e 1, e multiplicado pela porta de saída, resultando em:

$$h_t = o_t \tanh \otimes C_t \tag{2.20}$$

GRU

As redes GRU foram propostas como uma versão simplificada das redes LSTM, uma vez que requerem menos tempo de treino e obtêm um desempenho maior face as redes anteriormente abordadas [99].

Em termos de execução, os modelos GRU e o LSTM funcionam de forma bastante semelhante, contudo, as GRU utilizam um estado oculto que forma em conjunto com a porta de esquecimento e a porta de entrada uma porta de atualização [100]. Portanto, o número total de portas em GRU é metade do número total de portas no LSTM, tornando assim, as GRU mais eficiente [101].

A estrutura específica da célula de uma GRU é ilustrada na Figura 2.20.

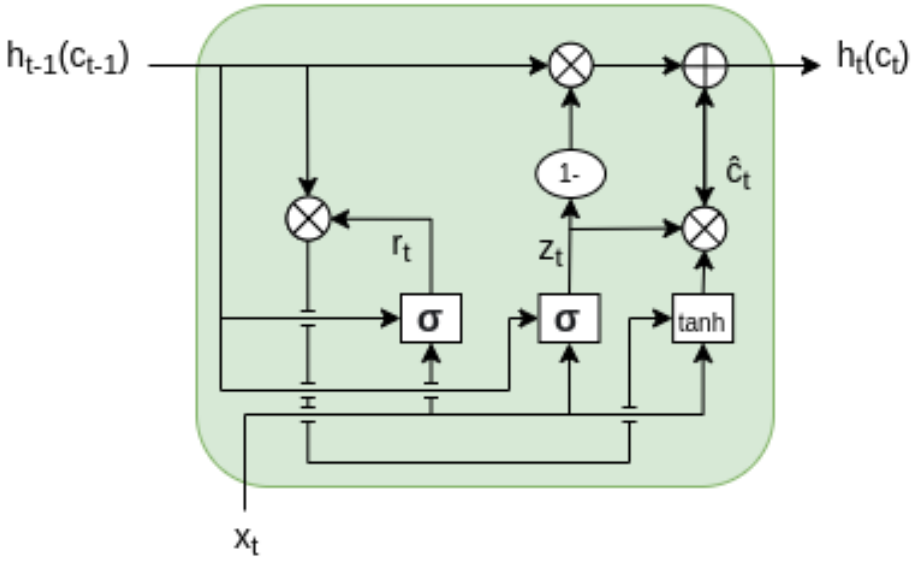


Figura 2.20: A estrutura geral de GRU

No modelo GRU, existem duas portas de controlo: a porta de atualização (z_t) e a porta de *reset* (r_t). Essencialmente, a porta de *reset* (r_t) é usada para decidir que quantidade de informação será para esquecer. Para a calcular, é usada a Equação 2.21 [91]:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \tag{2.21}$$

Onde, W_{xr} e W_{hr} são as matrizes de pesos da rede e b_r é um vetor de polarização.

A porta de atualização (z_t) é usada para controlar o grau em que a informação de estado $h_{t-1}(c_{t-1})$ no espaço de tempo anterior t-1 será trazida para o espaço de tempo atual t. Para calcular a porta de atualização (z_t) é usada a Equação 2.22 [91]:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \tag{2.22}$$

Onde W_{xz} e W_{hz} são as matrizes de pesos da rede e b_z é o vetor de polarização. Quanto maior o valor da porta de atualização, mais informações de estado no espaço de tempo anterior são trazidas.

Para verificar como as portas afetam a saída final, é utilizada a porta de *reset* para armazenar as informações relevantes do passado. É calculado da seguinte forma [91]:

$$\hat{c}_t = \tanh(W_{xc}x_t + W_{hc}(r_t \otimes h_{t-1})b_c) \tag{2.23}$$

Por fim, para calcular o estado da célula (c_t), o vetor que contém informação para a unidade atual, passa-a para a rede. Para isso, é necessário a porta de atualização. Isto é feito da seguinte forma [91]:

$$c_t = (1 - z_t) \otimes c_{t-1} + z_t \otimes \hat{c}_t \tag{2.24}$$

2.3.3.5 CNN

As CNN são essencialmente estruturas de classificação que se aplicam ao processamento de imagem, processamento de linguagem e a outros tipos de tarefas cognitivas [102]. Tal como as RNA, uma CNN tem uma camada de entrada, uma camada de saída e várias camadas ocultas. Algumas dessas camadas são convolucionais, usando um modelo matemático para transmitir os resultados às camadas sucessivas [102].

Normalmente, estas redes apresentam uma camada de entrada, seguida por uma ou mais camadas convolucionais e camadas de *pooling*, uma camada totalmente conectada e uma camada de saída. A camada de *pooling* conecta o *cluster* de neurónios de saída, num único neurónio, reduzindo assim, a dimensão dos sinais [102]. A camada totalmente conectada é tipicamente uma rede neuronal do tipo *feedforward*, tal como se ilustra na Figura 2.21 [80]:

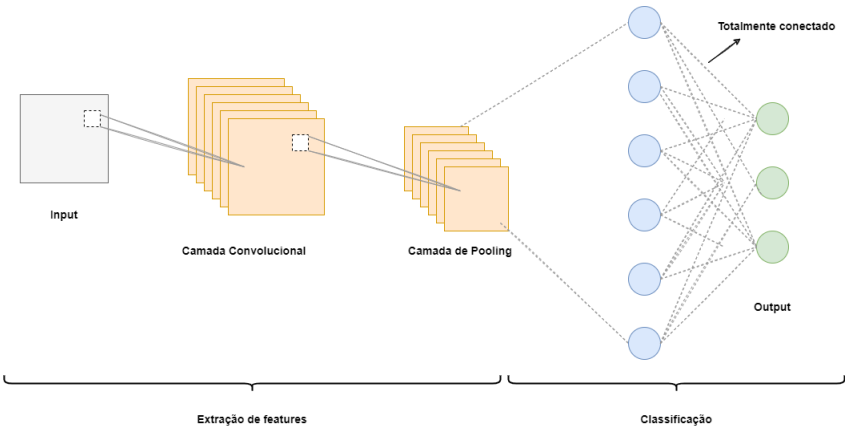


Figura 2.21: Esquema geral de uma CNN

De forma a compreender o processo de uma CNN, uma imagem RGB pode ser visualizada como uma matriz de valores de pixeis com três planos, enquanto uma imagem em tons de cinzento, tem um único plano [90]. A Figura 2.22, evidencia a matriz de cores. O papel da CNN é reduzir as imagens num formato mais fácil de processar [102].

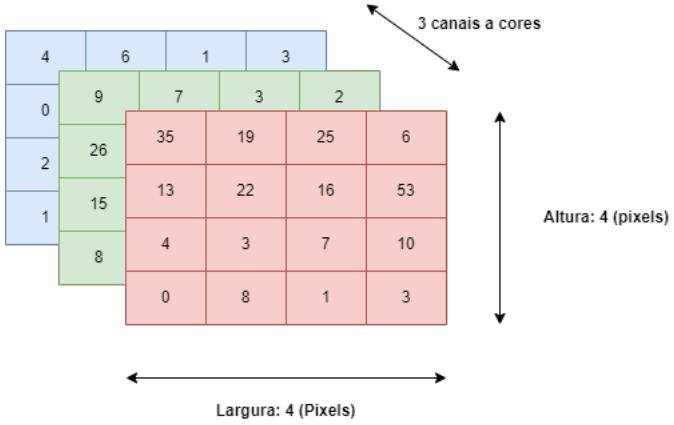


Figura 2.22: Matriz a cores

Por simplicidade, as imagens a preto e branco como as que são evidenciadas na Figura 2.23, mostram o que é uma convolução [103].

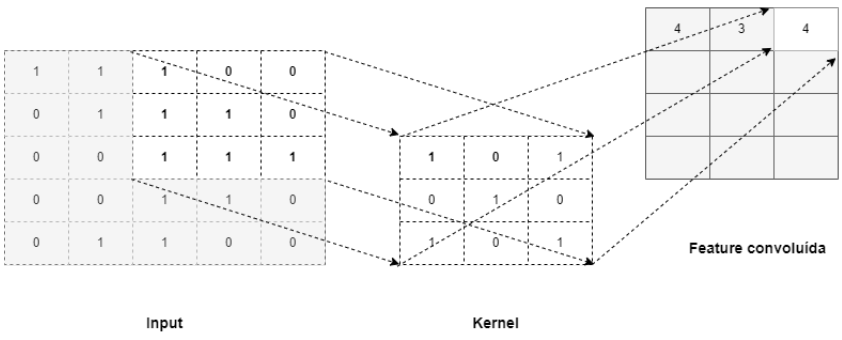


Figura 2.23: Matriz preto e branco

Ou seja, ao analisar o esquema observa-se um filtro/kernel (matriz 3×3) que é aplicado à imagem de entrada para obter a *feature* convoluída, que por sua vez, é passada para a camada seguinte [103].

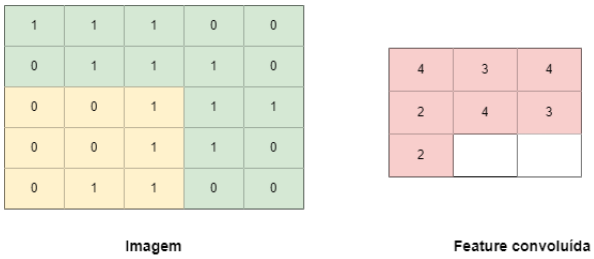


Figura 2.24: Visualização de convolução. Através de uma imagem 5x5 e um kernel de 3x3 pixels

Na Figura 2.24, a matriz da esquerda representa uma imagem a preto e branco. Cada entrada corresponde a um pixel, 0 para a cor preta e 1 para a branca. A janela que se vai deslizando é chamada de *kernel*, filtro ou detetor de características. A região da matriz da imagem de *input* do mesmo tamanho do filtro é designada de campo recetivo [104]. Em cada etapa são multiplicados os valores do filtro pelos do campo elemento a elemento e são, de seguida, somados, assim, o filtro desliza-se um passo sobre o seguinte campo recetivo da matriz e repete-se a mesma operação até que a imagem tenha sido passada por completo [105]. O resultado desta operação é um número inteiro do volume do *output*. O *output* será o *input* da camada seguinte.

As CNN são essencialmente usadas para lidar com o reconhecimento de imagens e vídeos, classificação de imagens, análise de imagens médicas e até processamento de linguagem natural, contudo, também se podem aplicar convoluções unidimensionais para lidar com séries temporais [106]. Desta forma, as operações passam a ser através de vetores em vez de serem através de matrizes.

Ao realizar convoluções sem preenchimento, o número de *timesteps* em toda a rede será reduzido, isto é:

- O tamanho do kernel define a janela de *timesteps* afetada por cada filtro.
- Considerando que o modelo CNN utiliza sequências dos seus *timesteps* (it), ele deve garantir, para todas as camadas convolucionais:
    1. Quanto mais profunda a CNN, menor o número de *timesteps* de entrada
    2. Quanto maior o tamanho do kernel, menos profunda a CNN será

$$(it - ks) + 1 > 0 \tag{2.25}$$

De seguida é evidenciado um exemplo na Figura 2.25, de forma a explicar melhor o modelo, através de séries temporais:

Considerando um exemplo, onde na camada de *Input Layer*,com N = 1609, *timesteps* = 7 e *features* = 4.

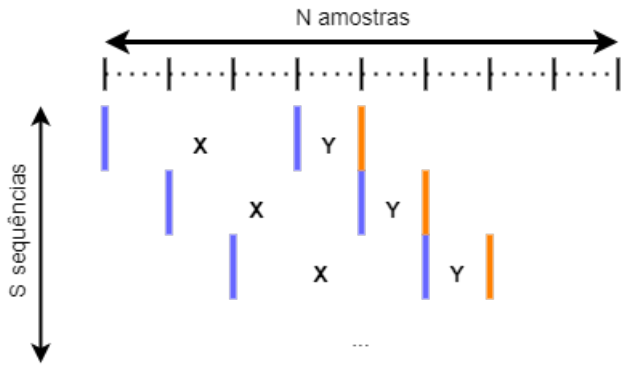


Figura 2.25: Input Layer

Assim, partindo de um *Input Shape* (1601, 7, 4):

$$S = N - (size(X) + size(Y)) \tag{2.26}$$

Sendo a sequência final:

$$S = 1609 - (7 + 1) \tag{2.27}$$

$$S = 1601 \tag{2.28}$$

Desta forma, a camada que se segue tomará como *Input Shape*, o *output Shape* da camada anterior, ou seja, (1601, 7, 4). Seguindo para a camada de Conv1D(Multi-Variate), com um *Kernel* = 5, um *stride* de 1 e com 16 filtros, tal como ilustrado na Figura 2.26:

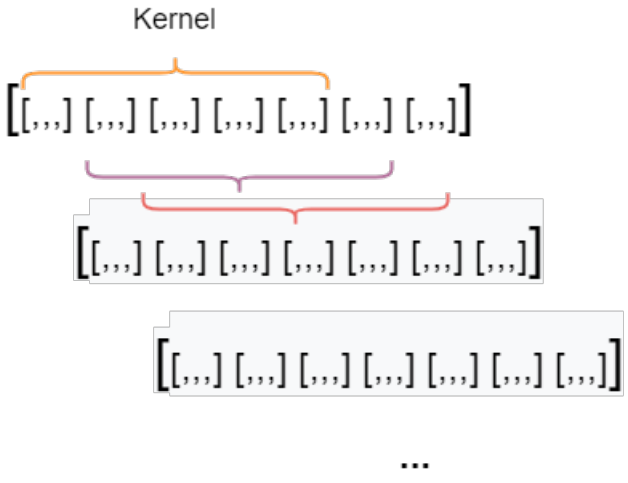


Figura 2.26: Conv1D (Multi-Variate)

De acordo com a Equação 2.29, iremos obter o novo valor pata *timesteps*:

$$output_{\text{timesteps}} = (it - ks) + 1 \tag{2.29}$$

Substituíndo os valores, resulta em:

$$output_{\text{timesteps}} = (7 - 5) + 1 \tag{2.30}$$

O número de *output* final é, assim, diminuído para 3:

$$output_{\text{timesteps}} = 3 \tag{2.31}$$

Assim, partindo de um *input Shape* de (1601, 7, 4), gera um *output Shape* de (1601, 3, 16).

Seguindo o mesmo raciocínio, o *output* gerado na camada de Conv1D será utilizado como entrada na camada de Max Pool1D.

Em relação às operações de *pooling*, estas podem seguir duas abordagens distintas:

- *Channels' First* - não afeta o *Input Layer*, aliás, reduz o número de *features*, como se pode observar na Figura 2.27, com um pool size de 3:

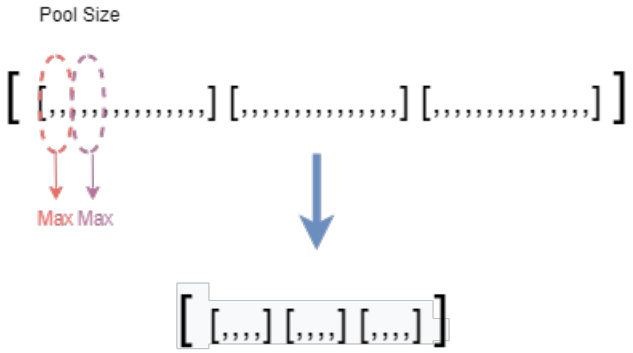


Figura 2.27: Channels' First (redução de features)

Assim, partindo de um *input shape* de (1601, 3, 16) resulta num *output shape* de (1601, 3, 5), uma vez que agrupa de 3 elementos em 3 elementos, e seleciona o maior valor desses 3 elementos.

- *Channels' Last* - não afeta o número de *features*, pelo contrário, reduz o número de *timesteps*, como se pode observar na Figura 2.28:

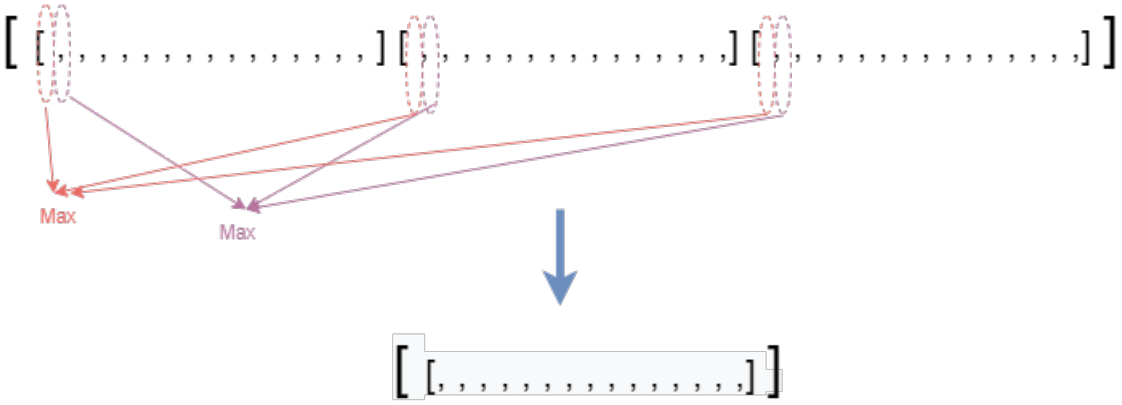


Figura 2.28: Channels' Last

Neste caso, é selecionado o elemento maior de cada *Pool Size*, formando apenas 1 *timestep*. Desta forma, partindo do *input shape* de (1601, 3, 16) o resultado é (1601, 1, 16).

2.4 Métricas de Avaliação

Com o propósito de avaliar e por, conseguinte comparar o desempenho dos modelos aplicados foram levadas em consideração duas métricas de avaliação, o RMSE e o MAE. Esta escolha advém do facto de que estas métricas são, bastante utilizadas para avaliar a *performance* de modelos preditivos que têm como alvo variáveis contínuas [17]. Essas medidas são usadas principalmente para ajustar, validar, selecionar, comparar e avaliar o modelo de previsão [107].

A raiz do erro quadrático médio é uma métrica de avaliação amplamente utilizada e reconhecida na comunidade para medir o desempenho de modelos de ML. O cálculo advém da raiz quadrada da média dos quadrados dos erros, onde o erro bruto é a diferença entre o valor previsto pelo modelo e o valor real [107]. A Equação 2.32 demonstra o cálculo do RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i)^2} \tag{2.32}$$

Onde, x_i, para i = 1,2,...,n são o conjunto dos valores observados e $\widehat{x_i}$, para i = 1,2,...,n são os seus valores de previsão.

Por outro lado, existe uma outra métrica que calcula o erro absoluto médio dos erros entre valores observados, isto é, reais e as suas predições, ou seja, as suas hipóteses. De seguida ilustramos o cálculo do MAE, com a equação 2.33:

$$MAE = \frac{1}{n}\sum_{x=1}^{n}|x_i - \widehat{x_i}| = \frac{1}{n}\sum_{x=1}^{n}|e_i| \tag{2.33}$$

Nesta ilustração verifica-se que , $e_i = x_i - \widehat{x_i}$ para i = 1,2,...,n, corresponde ao conjunto de erros e n designa o tamanho da amostra. O valor de e_i tanto pode ser positivo como negativo.

2.5 Revisão da literatura

Na presente secção é feita uma análise a várias soluções e trabalhos de investigação existentes na literatura. Os principais objetivos desta análise serão a identificação de aspetos relevantes nas diferentes soluções existentes e a reflexão crítica sobre o conteúdo mais importante no domínio desta dissertação. A partir deste levantamento de artigos e trabalhos, pretende-se identificar a melhor estratégia a aplicar em termos de técnicas de ML.

2.5.1 Revisão da literatura

Os autores Liu et al. [17] fizeram uma previsão da qualidade do ar em três cidades da China: Pequim, Tianjin a Shijiazhuang. O estudo incidiu num total de 851 dias desde o dia 1 de janeiro de 2014 a 30 de abril de 2016. Foi comparado o mesmo algoritmo em vários conjuntos de dados tendo em consideração o ar das três cidades, a informação da qualidade do ar e as condições meteorológicas como dados de *input* e o IQA como *output*. Para além disso, também foi analisada a correlação entre as mesmas áreas de poluição do ar urbano. O objetivo seria apresentar um novo modelo de previsão usando o algoritmo Support Vector Regression (SVR) [17] e minimizar o erro de previsão dos atuais algoritmos de ML. Para tal, foi usada a técnica de *cross validation*, 4 vezes. Os critérios aplicados para medir o desempenho deste modelo foram o Mean Absolute Percentage Error (MAPE), o Mean Square Error (MSE), RMSE e o MAE. Os autores concluíram que o SVR se superou face às RNA. Assim, os resultados mostraram uma diminuição do MAPE, quando há uma forte interação e correlação dos atributos da qualidade do ar com o IQA. Também, a localização geográfica desempenhou um papel significativo na previsão do IQA [1]. Os valores de RMSE e os conjuntos de dados de teste foram <12, podendo assim, aos autores inferir que o SVR é forte e confiável para prever os valores de IQA [17]. Também o MAPE para todos os casos apresentou valores entre 0,05 e 0,09, o que indica um resultado de previsão altamente preciso [17].

Os autores Kumar et al. [44] realizaram um estudo para comparar o estado da qualidade do ar antes e depois do confinamento, ou seja, entre 25 de dezembro de 2019 e 25 de junho de 2020, numa cidade da Índia, em Meerut. Foram adotadas várias metodologias para avaliar o IQA, tendo em consideração quatro poluentes atmosféricos: PM_{10}, $PM_{2,5}$, SO_2 e o NO_2. Para além disso, foram também tidos em consideração as condições meteorológicas, como a velocidade e direção do vento, a humidade do ar e a temperatura. Assim, foram estimados diferentes IQA para os seis meses do estudo e foram observados resultados variados.O primeiro método utilizado teve como base a média aritmética da razão de concentração de poluentes para o valor padrão de cada poluente. No segundo método, o processo de cálculo do IQA teve em consideração a média geométrica da razão de concentração de poluentes com o respetivo valor padrão. Por outro lado, o terceiro método calculou o IQA através da sua derivada. Por fim, o IQA foi determinado com base na quantidade de poluentes para adquirir a concentração do ponto de rutura. Os autores não indicaram nenhuma análise de correlação às substâncias, nem nenhuma técnica de *cross-validation* nos seus modelos. Os autores concluíram que desde o início da pandemia, Dezembro(M1), ainda em pré-confinamento, a cidade de Meerut tinha um M1 de (IQA :230) , que comparado com os valores padrão era classificada como um estado mau. No segundo M2, os valores de (IQA:212) também foram considerados pobre. Em M3 com valores de(IQA:165) considerado moderado e assim sucessivamente até M6. Demonstrando assim, uma queda significativa do IQA [44].

Os autores Shah et al. [54] realizaram um estudo para comparar o IQA em seis locais distintos na cidade de Ahmedabad, na Índia, entre maio de 2019 e julho de 2019 para obter resultados no período de verão e de Dezembro de 2019 a março de 2020 para o período de inverno. Foi monitorizada a concentração de seis poluentes: PM_{10}, $PM_{2,5}$, SO_2, NO_2, O_3 e o CO. Para tal, o estudo comparou dois métodos

de cálculo, o National Air Quality Index (NAQI) baseado na função operacional máxima e o Composite Air Quality Index (CAQI) que se baseia na agregação de poluentes. Foi também considerado o índice de Ultra Violeta (UV), as condições meteorológicas como a temperatura, a humidade, a velocidade e direção do vento. Os autores não indicaram nenhuma análise de correlação às substâncias, nem nenhuma técnica de *cross-validation* nos modelos. Os autores concluíram que no verão, a concentração média de todos os poluentes em todos os locais selecionados estava dentro dos padrões definidos pelo National Ambient Air Quality Standards (NAAQS),considerando dados de monitorização de todos os locais, 95,60% dos valores NAQI e 92,60% dos valores CAQI. Já no inverno, a concentração média de partículas PM_{10} e $PM_{2,5}$ excederam os padrões em quatro dos seis locais selecionados durante a fase de monitorização e 20,25% dos valores do NAQI e 6,33% dos valores do CAQI ficaram, portanto, abaixo dos padrões considerados normais.Por fim, os autores concluíram que o CAQI estima de forma mais eficiente a exposição a poluentes da população em comparação ao NAQI [54].

Os autores Janarthanan et al. [108] realizaram a combinação dos modelos SVR com o modelo de DL, o LSTM para estimar e prever o IQA,na cidade metropolitana de Chennai, na Índia. A medição da qualidade do ar dependeu de oito poluentes, nomeadamente o PM_{10}, $PM_{2,5}$, SO_2, NO_2, O_3, CO e o NH_3. A extração de características foi realizada com o uso da técnica Gray Level Co-occurrence Matrix (GLCM) para extrair as características como a média, o erro quadrado médio e o desvio padrão. Os critérios aplicados para medir o desempenho dos modelos foram o MSE e o R^2. O modelo LSTM proposto, os dados de treino e os dados de teste permitiram avaliar o desempenho do IQA para NO_2,SO_2,CO e o O_3, tendo os autores verificado que o modelo obteve uma boa *performance* no ajuste dos dados de treino e teste, uma vez que o valor de MSE e os valores de R^2 foram obtidos para NO_2 como 0,908 e 0,092, respetivamente, para SO_2 o valor de MSE e R^2 de 1,005 e -0,005, para o CO o valor MSE foi de 0,920 e o valor R^2 de 0,080. Para o O_3 o valor MSE foi 0,971 e o valor de R^2 de 0,029. Os autores não indicaram nenhuma análise de correlação às substâncias, nem nenhuma técnica de *cross-validation* nos modelos.

Os autores Abirami et al. [109] propuseram um modelo de DL chamado DL-Air que incorpora três componentes para a previsão da qualidade do ar. A primeira componente é um codificador que codifica todas as relações espaciais nos dados. A segunda componente trata-se de uma variante do modelo LSTM. A terceira componente é um descodificador, que descodifica adequadamente as relações para obter a previsão real. A estrutura proposta foi extensivamente avaliada para prever os dados de qualidade do ar em Delhi, na Índia, sendo que, os valores incluem 8 concentrações de poluentes atmosféricos PM_{10}, $PM_{2,5}$, SO_2, NO_2, O_3, CO e o NH3 e Pb e vários parâmetros meteorológicos, como a temperatura do ar, direção do vento, pressão atmosférica, humidade relativa e a velocidade do vento. Este artigo considerou os dados no período de 1 de dezembro de 2018 a 30 de novembro de 2019. De todo o conjunto de dados, 80% foi utilizado para treinar o modelo e 20% foi utilizado para teste. A performance do modelo foi avaliada através dos valores de Desvio Médio Absoluto (DMA), RMSE, MAE e R^2 para todos os poluentes previstos. O DL-Air mostrou melhor desempenho com cerca de 30% de RMSE e MAE reduzidos, 37% de DMA, 11% de R^2 e 8% de precisão aprimorada na previsão de IQA. Além disso, o desempenho da previsão do DL-Air apresentou-se consistente em todas as estações de Delhi [109].

2.5.2 Análise crítica

Tal como se verifica na secção anterior, existe um volume considerável de estudos e artigos que abordam a área do ML e DL para prever o IQA. Cada vez mais existe esta necessidade, uma vez que a monitorização, modelação e a previsão da qualidade do ar pode ser uma forma prudente de consciencializar e defender o ser humano das adversidades da poluição do ar [109].

Ao dissecar cada um dos estudos científicos, observa-se que existem imensas abordagens distintas para prever o IQA. Uma vez que o objetivo da presente dissertação é também utilizar modelos preditivos para a monitorização do IQA, tornou-se bastante relevante entender que prever a qualidade do ar é uma tarefa complexa devido à natureza dinâmica, volatilidade e alta variabilidade no espaço e no tempo dos poluentes e partículas [108]. Desta forma, e como na maior parte dos estudos analisados não foi mencionada a técnica de *cross-validation* que, tal como se sabe previne a ocorrência de *overfitting*, neste estudo será aplicada a técnica. Os estudos em questão foram também avaliados em relação à análise de correlação entre as variáveis de forma a identificar as variáveis de *input* de maior importância para as previsões. Assim, verificou-se que apenas 1 estudo incluiu este procedimento nas suas pesquisas [17], enquanto que os restantes não o fizeram, ou omitiram. É de salientar a enorme importância que este procedimento tem, pois este permite selecionar os *inputs* consoante a sua correlação com o *output*, permitindo assim obter melhores desempenhos nos modelos, quer a nível de *performance*, quer a nível de tempo de computação. Por conseguinte, adotou-se também este fator na presente dissertação com o intuito de obter melhores resultados.

Capítulo 3

Materiais e métodos

Os materiais e métodos que foram adotados no desenvolvimento desta dissertação são apresentados no presente capítulo. Primeiramente, é apresentada a análise dos *datasets* dos poluentes do ar, assim como a primeira exploração efetuada em torno dos mesmos, em 3.1. Em seguida, na secção 3.2, é explorado o *dataset* com várias condições meteorológicas que podem ou não impactar os resultados dos algoritmos a implementar nas fases mais avançadas desta dissertação. Na secção 3.3, é explicado como os dados foram devidamente tratados. Por fim, na secção 3.4 são mostradas as tecnologias utilizadas para desenvolver o código fonte desta dissertação.

3.1 Análise dos Dados Dos Poluentes do Ar

A primeira etapa da fase prática da presente dissertação é a exploração dos dados. Nesta etapa, começam-se a entender os dados, a descobrir como estão distribuídos e a compreender a relação entre as diferentes variáveis. Inicialmente, será feita a análise de seis *datasets*, correspondentes aos seis poluentes, onde serão explorados e analisados valores como as médias, mediana, desvio padrão, entre outras métricas de estatística descritiva. Desta forma, será possível ter uma melhor visualização dos dados, garantindo assim, um melhor processamento dos mesmos.

3.1.1 Análise do Monóxido de Carbono

A análise inicia-se pelo *dataset* referente aos valores de CO. Este *dataset* apresenta 10 atributos, nomeadamente: a *location*, *city*, *country*, *utc*, *local*, *parameter*, *value*, *unit*, latitude e a longitude. Sendo que a sua descrição se encontra na Tabela 3.1.

Atributo	Descrição	Tipo de dados	Unidade de medida
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Registos do valor de CO	*float64*	$\mu g/m^3$
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.1: Constituição do dataset do CO

De forma a compreender melhor o conjunto de dados, foi decidido descartar todos os atributos, com exceção do *value* e do *local*. Desta forma obtém-se uma precisão maior no estudo desta variável. O *value*, tal como se pode inferir da Tabela 3.1, corresponde ao valor registado de CO. O atributo *local* dita a data em que tal registo sucedeu. Este *dataset* é constituído por 23071 observações, que vão desde o dia 2017-09-26 até ao dia 2021-04-30.

Pela análise dos registos, observa-se que os dados não apresentam um padrão rigoroso, uma vez que, existem dias sem qualquer registo e outros dias em que há mais que um registo. Para além disso a quantidade de registos em cada ano também é aleatória, tendo no ano de 2017 apenas 151 registos e por exemplo, em 2019, um total de 8102 registos. Neste caso em concreto, não existe nenhum *missing value*, e os seus valores variam entre o valor mínimo 0 e um máximo de 5782, tal como se pode verificar na Tabela 3.2. O *dataset* apresenta 437 zeros, o que corresponde a cerca de 1.9% dos valores registados.

Para testar a distribuição dos dados, foi usado o teste de *Kolmogorov-Smirnov*. Como se obteve *p-value* > 0.05, assume-se que os dados não seguem uma distribuição normal. A partir dos valores de Skewness conclui-se que os dados se encontram inclinados positivamente (Skewness >1) e a partir do valor de Kurtosis, também este positivo (Kurtosis > 1), pode-se inferir que a distribuição tem um pico acentuado e é chamada de distribuição leptocúrtica [110].

Foi realizada a análise estatística dos dados do valor de CO com os parâmetros mínimo, máximo, média, desvio padrão, variância, mediana, *Skewness* e *Kurtosis*. O parâmetro Skewness representa a medida de assimetria da curva da função de distribuição de probabilidade [111].

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	5782	550.43	355.0	126022.25	451	3.3	20.82

Tabela 3.2: Análise estatística do CO

De forma a explorar os dados, foram também gerados 2 gráficos com as médias mensais das observações por cada ano e na sua totalidade por estação do ano. A partir da análise da Figura 3.1, é possível observar que os valores dos meses de janeiro apresentam valores mais elevados que vão decaindo sensivelmente até maio, sendo que o ano de 2020 apresenta um pico com valores mais elevados entre março e abril. Em contrapartida o ano de 2017 apenas tem registos a partir do mês de setembro. Pode-se ainda inferir que, analisando os restantes meses, entre maio e dezembro os valores de CO apresentam-se com várias oscilações, nos cinco anos em análise, não tendo, assim um padrão concreto.

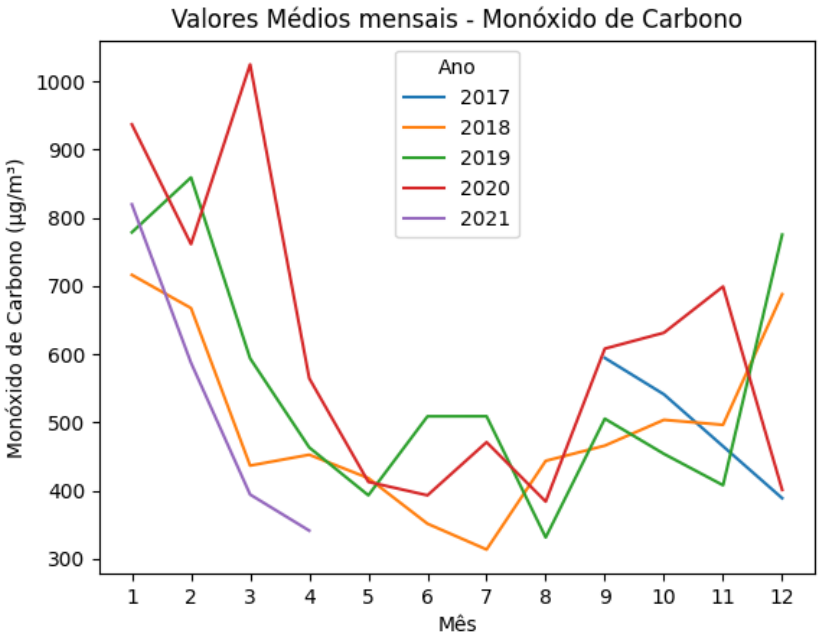


Figura 3.1: Valores médios mensais do CO por ano

A partir da análise da Figura 3.2, é inferível que o valor do CO vai variando de estação para estação, apresentando nas estações de Primavera e Verão, níveis mais baixos de CO. Tendencialmente nos meses de Inverno, os valores são mais elevados. No ano de 2017 observa-se que o Outono tem valores mais elevados face ao Inverno, situação esta, que diverge dos restantes anos em análise. Foi no Inverno de 2020 que se registou o valor mais elevado de CO. Tanto o ano de 2017 como o ano de 2021 apresentam apenas duas estações do ano.

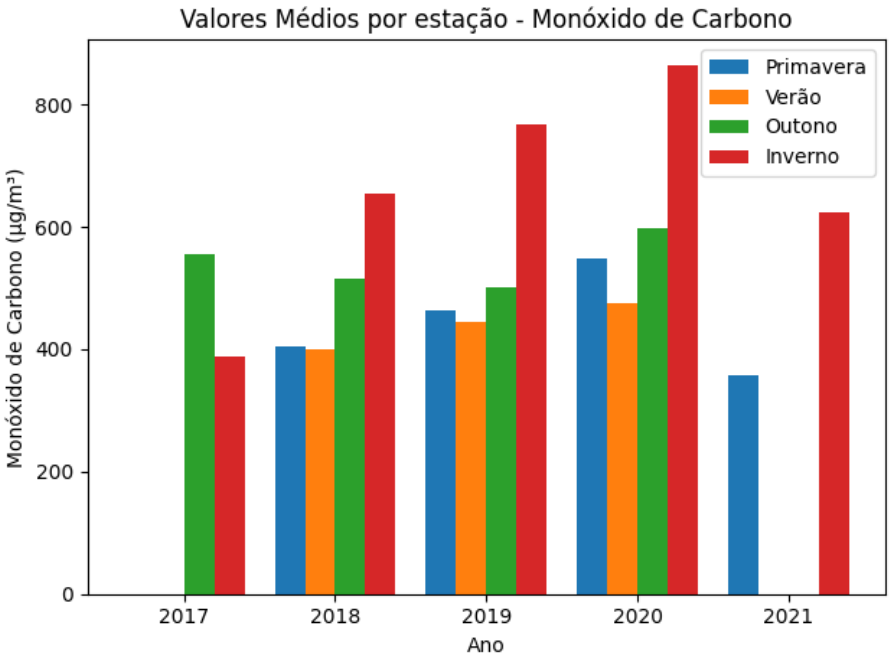


Figura 3.2: Análise dos valores médios do CO por estação

3.1.2 Análise do Dióxido de nitrogénio

O segundo poluente em análise é o NO_2. Este *dataset* é constituído pelos mesmos 10 atributos referidos anteriormente. Da mesma forma que é destacado a maioria dos atributos, este também não foi exceção. Os valores com maior relevância foram os atributos *value* e *local*. O detalhe deste *dataset* encontra-se na Tabela 3.3.

Atributo	**Descrição**	**Tipo de dados**	**Unidade de medida**
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Registos do valor de NO_2	*float64*	µg/m³
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.3: Constituição do *dataset* de NO_2

Este *dataset* é constituído por 21394 observações, que variam desde o dia 2017-09-26 até 2021-04-30. Da análise deste conjunto de dados, pode-se observar que também não apresenta um padrão específico. Não contém qualquer *missing value* e os seus valores variam entre o valor 0 e um máximo de 555.6. Apesar de não conterem *missing values*, existem 1563 zeros, representando cerca de 7.3% das

observações totais. A análise estatística aos dados encontra-se na Tabela 3.4 com os mesmos parâmetros mencionados anteriormente.

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	555.6	44.56	37.42	1400.04	38.5	3.44	28.57

Tabela 3.4: Análise estatística do NO_2

Relativamente à distribuição dos dados, no caso do NO_2, estes também não seguem uma distribuição normal. A partir dos valores de Skewness é possível concluir que os dados se encontram inclinados positivamente e a partir do valor de Kurtosis, também este positivo, pode-se inferir que também se trata de uma distribuição leptocúrtica.

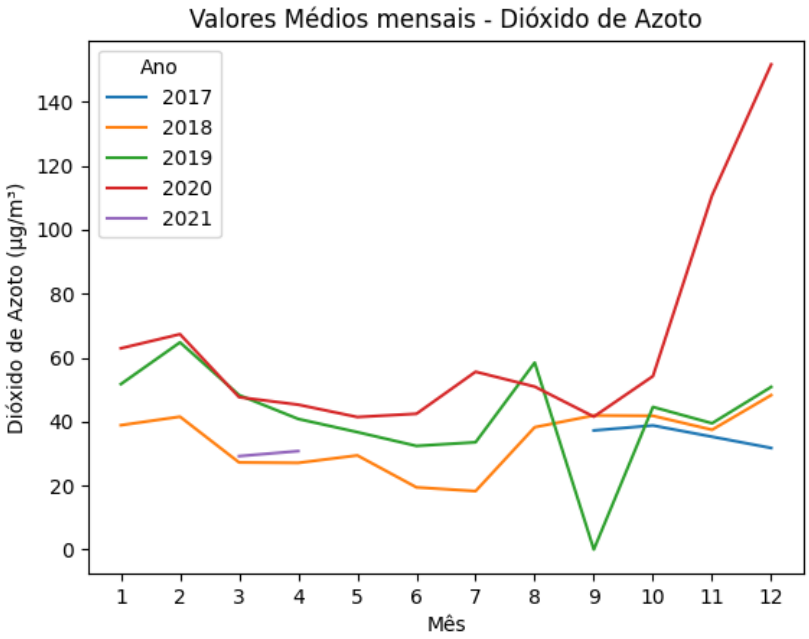


Figura 3.3: Valores médios mensais do NO_2 por ano

A partir da análise da Figura 3.3, é possível observar que os anos de 2017 e 2021 têm poucas observações registadas. O ano de 2020 foi o ano que apresentou valores de NO_2 mais elevados e com comportamento moderadamente constante entre janeiro e setembro, e com tendência crescente de setembro até ao final do ano, atingindo os valores máximos no final do mesmo. Em contrapartida, o ano de 2018 acaba por ter valores quase constantes ao longo de todo o ano, e 2019 sofre uma queda abrupta entre agosto e setembro, mantendo-se parcialmente constante no restante ano. A partir da análise da Figura 3.4, é percetível que o valor do NO_2 vai oscilando bastante de estação para estação, apresentando nas estações do ano mais quentes, valores mais baixos. No ano de 2017, onde apenas são visíveis as estações de Outono e Inverno, é no Outono que os valores são mais elevados. Em 2018 existe um crescimento contínuo desde a Primavera até ao Outono, voltando a descer no Inverno.

Em 2019 os valores mais baixos fazem-se sentir no Verão e os mais elevados no Inverno. É no Outono de 2020 que os níveis de NO_2 se apresentam mais elevados, de todos os anos em análise. Em 2021 apenas são visualizados valores da Primavera e do Inverno, sendo que no Inverno os valores se apresentam mais baixos.

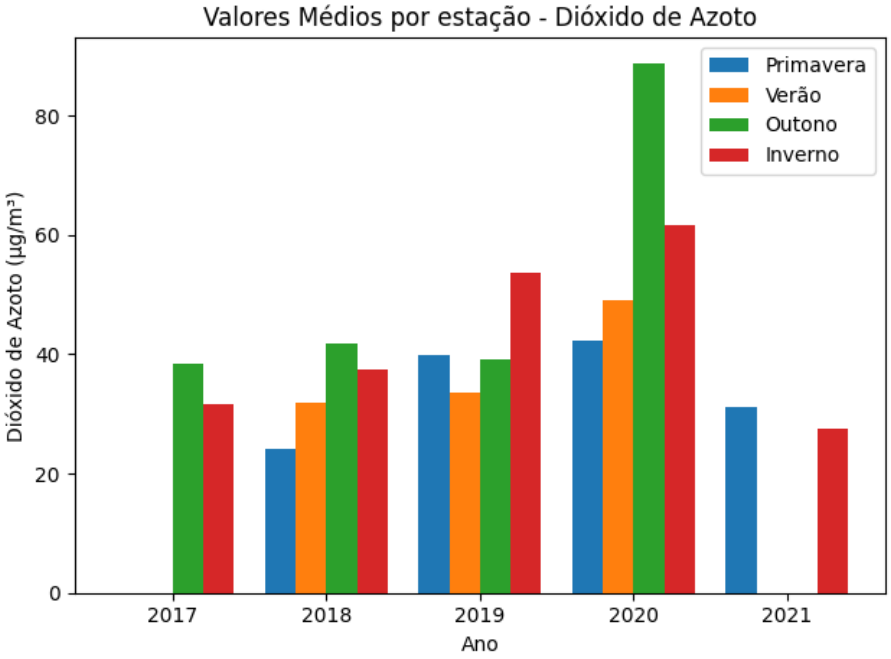


Figura 3.4: Análise dos valores médios do NO_2 por estação

3.1.3 Análise do Ozono

O conjunto de dados relativo ao O_3 é o terceiro *dataset* a ser explorado. Este é também constituído pelos mesmos 10 atributos. Aqui, também foram descartados a maioria dos atributos, ficando apenas para o estudo os atributos *value* e *local*. A descrição do mesmo se encontra na Tabela 3.5.

Atributo	**Descrição**	**Tipo de dados**	**Unidade de medida**
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Registos do valor de O3	*float64*	µg/m³
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.5: Constituição do *dataset* de O_3

Este *dataset* é constituído por 16998 observações, que variam de 2019-01-02 a 2021-04-30. Os dados

não apresentam um padrão concreto e a quantidade de registos por ano também é aleatória. Igualmente aqui, não existe nenhum *missing value*, e os seus valores variam entre o valor mínimo de 0 e um máximo de 160. Apesar de não conterem *missing value*, existem 3076 zeros, o que representa cerca de 18.1% das observações. A análise estatística dos dados encontra-se na Tabela 3.6 com os mesmos parâmetros que anteriormente foram referidos.

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	160.0	38.2	27.93	779.95	39.0	0.19	-0.77

Tabela 3.6: Análise estatística do O_3

Em relação à distribuição dos dados, estes não seguem uma distribuição normal. A partir dos valores de Skewness é possível concluir que os dados se encontram inclinados negativamente (Skewness < 1) e a partir do valor de Kurtosis, também este negativo (Kurtosis < 1), podendo-se assim, inferir que a distribuição tem um pico plano e é chamada de distribuição platicúrtica.

A partir da análise da Figura 3.5, é possível observar uma particularidade face aos restantes poluentes analisados. Este *dataset* apenas apresenta três anos de observações. Os anos de 2017 e 2018 não entram nesta análise. O ano de 2019 tem um comportamento ascendente nos primeiros meses do ano, sendo este o responsável pelo valor máximo registado. O ano de 2020 apresenta pequenas oscilações mas em média sempre acima dos 30 µg/m³ e inferior a 60 µg/m³. Em 2021 apenas foram registados valores até abril, e de janeiro a abril o comportamento dos mesmo é tendencialmente crescente.

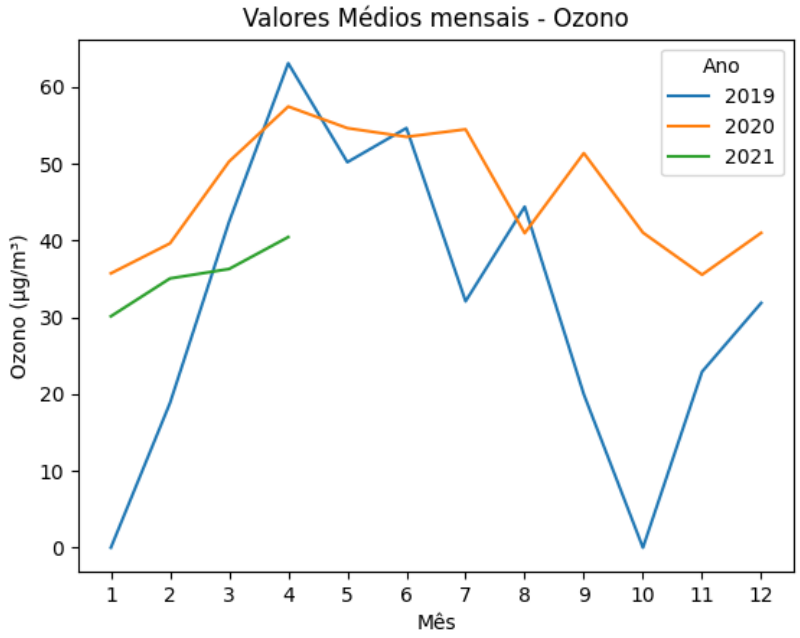


Figura 3.5: Valores médios mensais do O_3 por ano

Ao considerar a Figura 3.6, verificámos que a estação que possui os registos mais elevados é a Primavera, em cada um dos três anos em análise. Em 2019 é visível que da Primavera ao Outono os valores

vão caindo bastante, voltando a aumentar no Inverno. Em 2018 observa-se o mesmo comportamento que em 2019, mas com descidas menos abruptas. Por fim, em 2021, apesar da ausência de registos nas estações de Verão e Outono, vê-se a mesma tendência, valores que vão subindo do Inverno para a Primavera.

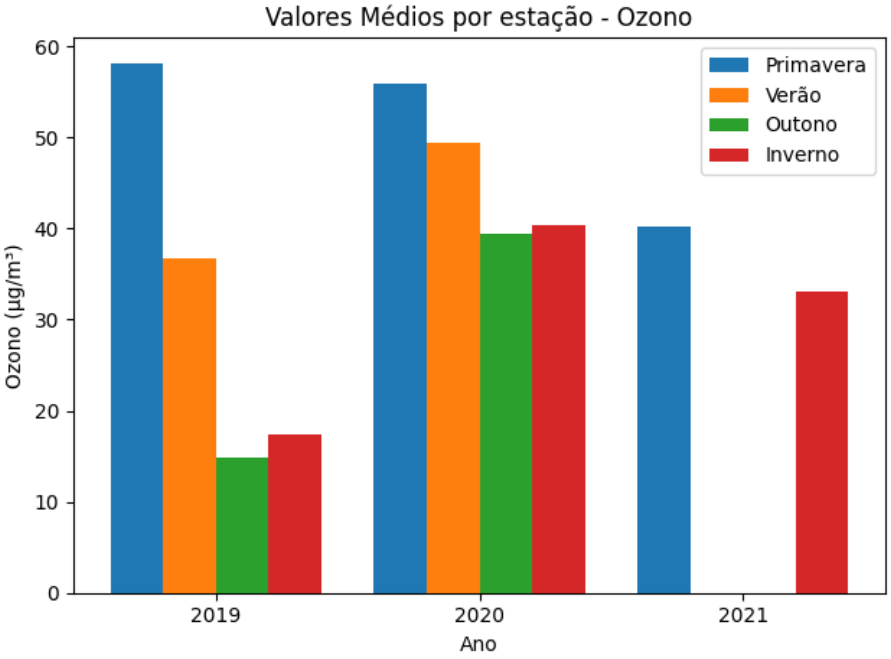


Figura 3.6: Análise dos valores médios do O_3 por estação

3.1.4 Análise de Partículas Inaláveis do tipo PM_{10}

O *dataset* que se segue diz respeito aos valores de PM_{10}, sendo que a descrição do mesmo, uma vez mais completa, encontra-se na Tabela 3.7.

Atributo	Descrição	Tipo de dados	Unidade de medida
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Valor de Partículas Inaláveis (< 10 µm)	*float64*	µg/m³
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.7: Constituição do *dataset* de PM_{10}

Este *dataset* é constituído por 20310 observações, que variam entre o dia 2017-09-26 e 2021-04-30. Os dados não apresentam um padrão concreto e a quantidade de registos por ano também é aleatória.

Além disso, não existe nenhum *missing value*, e os seus valores variam entre o valor mínimo de 0 e um máximo de 234. Apesar de não conterem *missing values*, existem 6731 zeros, o que representa cerca de 33.1% das observações. A análise estatística dos dados encontra-se na Tabela 3.8.

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	234.0	15.69	17.66	311.98	11.0	1.61	5.38

Tabela 3.8: Análise estatística do PM_{10}

Em relação à distribuição dos dados, estes não seguem uma distribuição normal. A partir dos valores de Skewness é possível concluir que os dados se encontram inclinados negativamente e a partir do valor de Kurtosis, afirmando assim, que se trata de uma distribuição leptocúrtica.

A partir da análise da Figura 3.7, é possível observar que em todos os anos em análise existem oscilações. Observa-se que nos anos de 2021 e 2017 existem ausências de informação, isto é, não apresentam registos para todo o ano. O ano de 2019 é o responsável pelo valor máximo, atingido em fevereiro, que vai decaindo gradualmente até outubro, voltando a subir em novembro. Os anos de 2018 e 2020 vão oscilando ao longo do ano, apresentando valores maiores e menores alternadamente. É ainda de notar que para a média mensal dos meses de maio a setembro apenas contribuíram os anos de 2018, 2019 e 2020.

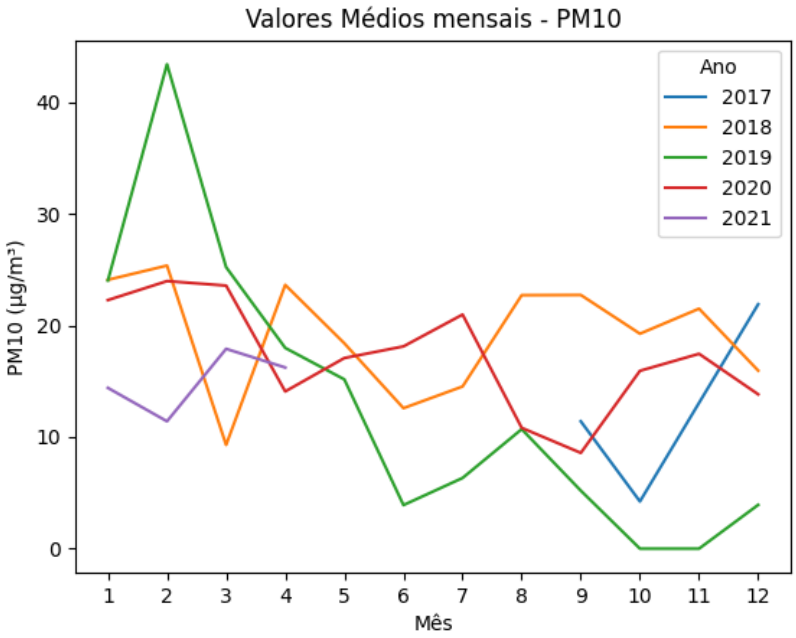


Figura 3.7: Valores médios mensais do PM_{10} por ano

Com base na observação da Figura 3.8, verificam-se várias oscilações de ano para ano. Em 2017 observa-se um salto gigante de valores da estação do Outono para o Inverno, sendo, no Inverno que se observa o valor mais alto. Em contra partida em 2018, o Outono é o responsável pelo registo mais elevado de todo o ano. Observa-se uma subida gradual desde a Primavera até ao Outono, voltando a decair no

Inverno. Uma vez mais, no Outono de 2019 o valor de partículas PM_{10}, é bastante inferior ao que é registo no Inverno desse mesmo ano.

Desta forma, destaca-se que o ano de 2019 emerge como responsável pelos valores extremos, representando tanto o menor quanto o maior registo em todo o conjunto de dados analisado. Já em 2020, a estação do Outono mantém a consistência ao apresentar os valores mais baixos, enquanto o Inverno continua a ser caracterizado pelos índices mais elevados. Em 2021, face a ausência de valores, é na Primavera que os valores são mais elevados, quando comparados com o Inverno.

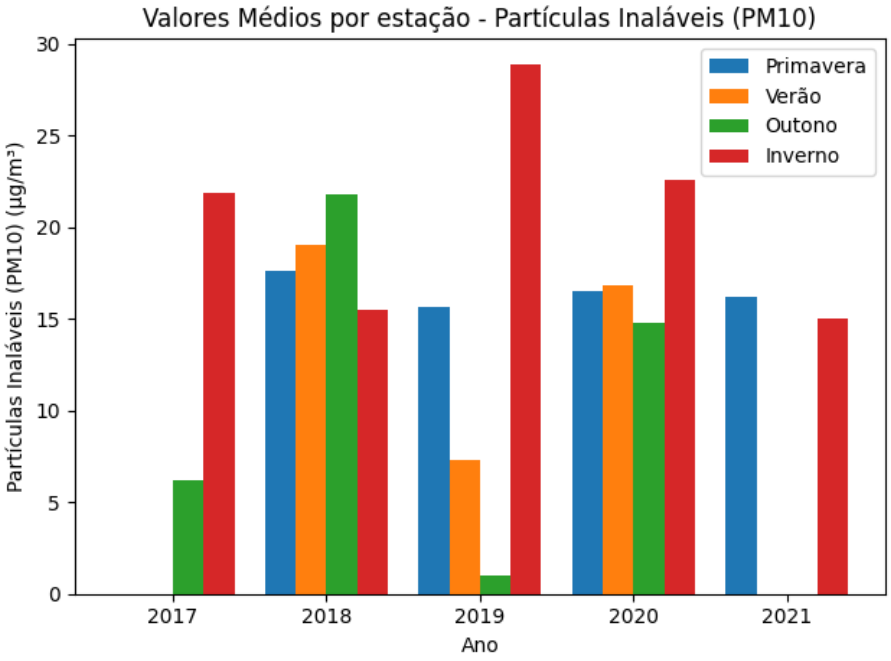


Figura 3.8: Análise dos valores médios do PM_{10} por estação

3.1.5 Análise de Partículas Inaláveis do tipo $PM_{2,5}$

Os dados a seguir são relativos aos valores de $PM_{2,5}$, sendo que uma descrição detalhada desses dados encontra-se disponível na Tabela 3.9.

Atributo	Descrição	Tipo de dados	Unidade de medida
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Valor de Partículas Inaláveis (< 2.5 µm)	*float64*	µg/m³
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.9: Constituição do *dataset* de $PM_{2,5}$

Este *dataset* é constituído por 14861 observações, que variam de 2019-03-01 a 2021-04-30. Os dados não apresentam um padrão rigoroso, onde a quantidade de registos por ano é aleatória, e não há registos no ano de 2017 e 2018. Neste *dataset* não existe nenhum *missing value*, e os seus valores variam entre o valor mínimo de 0 e um máximo de 745. Apesar de não conter *missing value*, existem 7232 zeros, o que representa cerca de 48.7% das observações. A análise estatística dos dados encontra-se na Tabela 3.10.

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	745.0	9.3	22.73	516.72	1.0	11.61	256.42

Tabela 3.10: Análise estatística do $PM_{2,5}$

Em relação à distribuição dos dados, estes também não seguem uma distribuição normal. A partir dos valores de Skewness é possível concluir que os dados se encontram inclinados negativamente e a partir do valor de Kurtosis, concluíndo que se trata de uma distribuição leptocúrtica.

Tomando como ponto de partida a Figura 3.9, é possível observar que, não existem registos dos anos de 2017 e 2018. O ano de 2019 mantém valores médios mais baixos face aos restantes anos em análise, sendo abrangido apenas pelos meses de março a dezembro. O ano de 2020 evidencia várias oscilações nos valores, atingindo valores mais elevados entre agosto e setembro. O ano de 2021 também não apresenta valores no ano inteiro, apenas até ao mês de abril. Os valores de $PM_{2,5}$ mais elevados, foram visualizados entre janeiro e fevereiro.

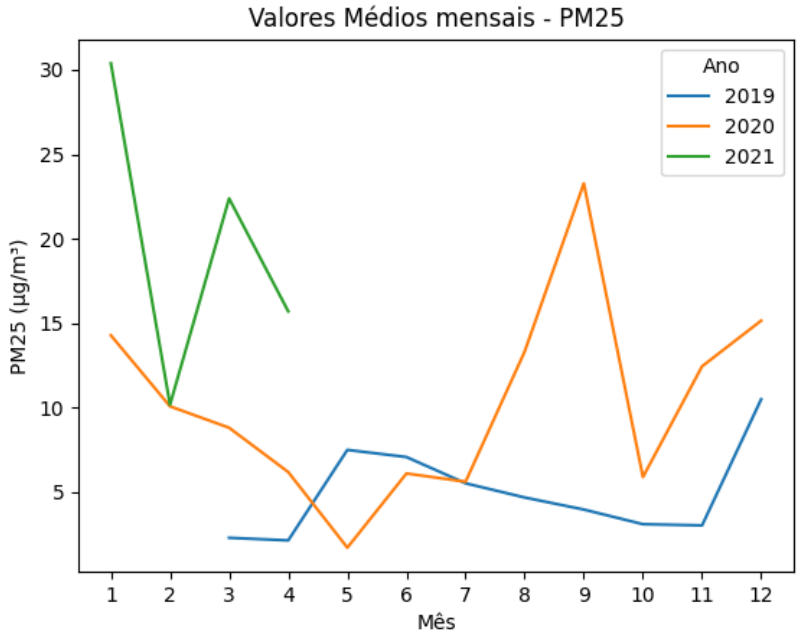


Figura 3.9: Valores médios mensais do $PM_{2,5}$ por ano

A partir da interpretação da Figura 3.10, verifica-se que os valores de $PM_{2,5}$ mais elevados ocorrem, exceção feita no ano de 2019, na estação mais fria do ano, no Inverno. Neste caso observa-se uma

diferença grande de ano para ano, uma vez que de 2019 a 2021 os valores, na sua generalidade, triplicam. Em 2019 os valores apresentam-se com poucas oscilações de estação em estação, e relativamente baixos. Em 2020 já se observa uma diferença considerável nas estações de Verão, Outono e Inverno face ao ano anterior, sendo no Verão que os seus valores são mais elevados. De novo, em 2021 os valores crescem bastante, observando-se um aumento significativo na Primavera face ao ano anterior.

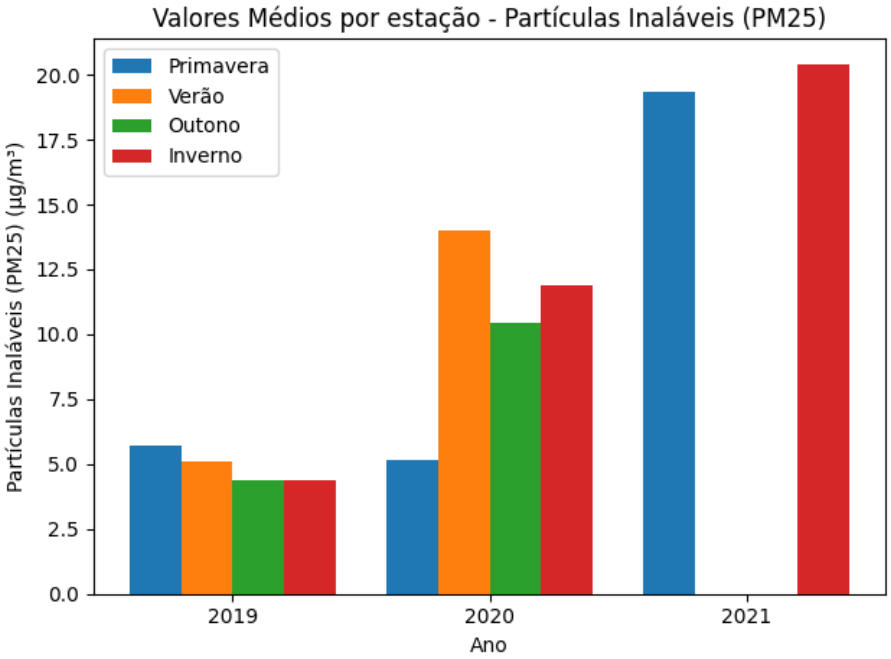


Figura 3.10: Análise dos valores médios do $PM_{2,5}$ por estação

3.1.6 Análise do SO_2

Por último, o poluente que se segue é o SO_2, apresenta 10 atributos, dos quais apenas dois têm relevância. A descrição detalhada do mesmo encontra-se na Tabela 3.11:

Atributo	Descrição	Tipo de dados	Unidade de medida
location	Identificador de localização	*object*	texto
city	Identificador da cidade	*object*	texto
country	Identificador do País	*object*	texto
utc	Registos de Tempo Universal Coordenado	*Timestamp*	*datetime*
local	Registos de data e hora	*Timestamp*	*datetime*
parameter	Identificador do poluente	*object*	texto
value	Valor de Dióxido de Enxofre	*float64*	µg/m³
unit	Identificador da unidade de medida	*object*	texto
latitude	Registos de latitude	*float64*	°C
longitude	Registos de longitude	*float64*	°C

Tabela 3.11: Constituição do *dataset* de SO_2

Este *dataset* é constituído por 15436 observações, que variam de 2019-01-02 a 2021-04-30. Os dados

não apresentam um padrão concreto e a quantidade de registos por ano é, uma vez mais, aleatória. Aqui também não existe nenhum *missing value*, e os seus valores variam entre o valor mínimo de 0 e um máximo de 110. Embora não existam valores ausentes, há 483 zeros, correspondendo a aproximadamente 3.1% das observações. A análise estatística dos dados encontra-se na Tabela 3.12.

Mínimo	Máximo	Média	Desvio Padrão	Variância	Mediana	Skewness	Kurtosis
0	110.0	10.7	7.15	51.11	8.0	1.8	9.58

Tabela 3.12: Análise estatística do SO_2

Quanto à distribuição dos dados, estes não apresentam uma distribuição normal. Os valores negativos de Skewness indicam uma inclinação negativa nos dados, enquanto o valor negativo de Kurtosis sugere que se trata de uma distribuição leptocúrtica.

Mediante a análise gráfica apresentada na Figura 3.11, é possível observar que tanto o ano de 2017 como o ano de 2018 não fazem parte da análise. Neste gráfico observa-se, uma vez mais, uma grande aleatoriedade nos registos dos anos de 2019 e 2020. O ano de 2019 acaba por assumir valores mais constantes e os mais baixos ao longo de todo o período em análise. Por outro lado, em 2020 observam-se valores que oscilam entre os extremos máximo e mínimo registados. Por fim, o ano de 2021 apenas contribui com valores nos primeiros 4 meses do ano.

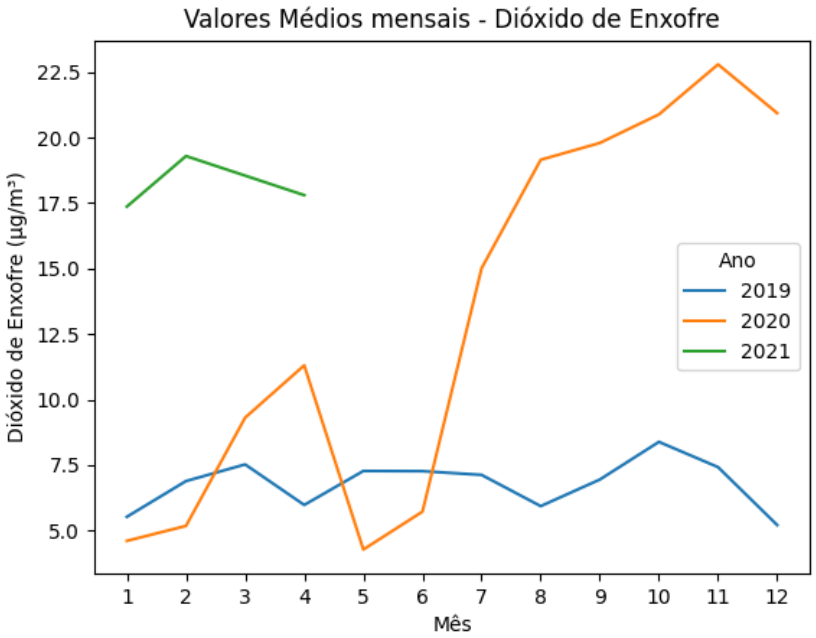


Figura 3.11: Valores médios mensais do SO_2 por ano

Considerando a análise visual da Figura 3.12, verificámos que os valores de SO_2 são mais baixos na Primavera, em todos os anos. Durante os períodos do Verão, os valores apresentam uma tendência crescente até ao Outono, voltando a diminuir com os dias mais frios do Inverno. Este comportamento é visível nos anos de 2019 e 2020, sendo neste último mais incisivo. Em 2021, apesar da ausência de

valores nas estações de Verão e Outono, os valores são mais elevados que em 2020, e em contrapartida é no Inverno que os valores mais elevados são registados.

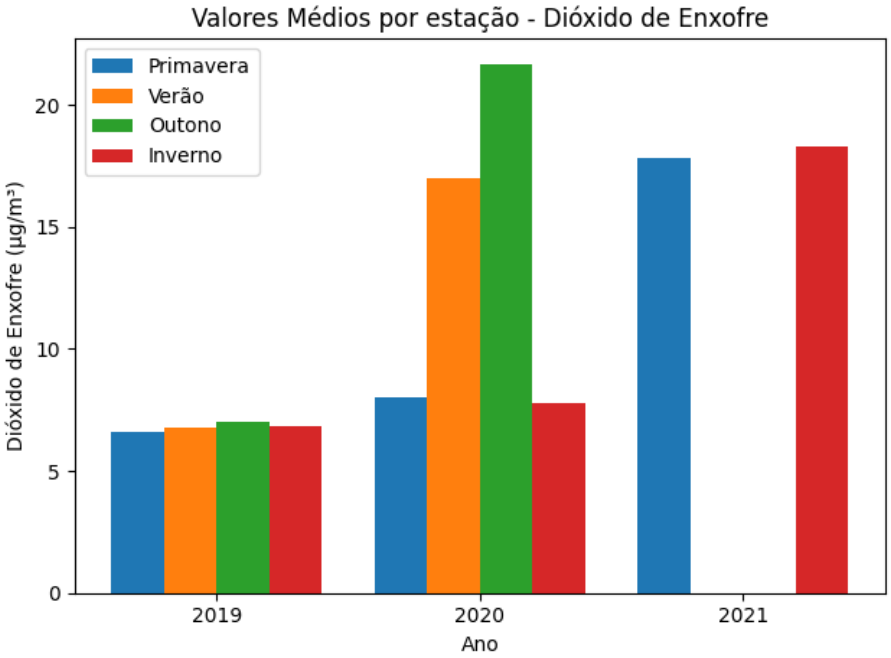


Figura 3.12: Análise dos valores médios de SO_2 por estação

3.2 Análise dos Dados Climatológicos

De forma a estabelecer possíveis relações entre dados climáticos e os poluentes analisados, foi utilizado um *dataset* com informação relativa as condições climatológicas da cidade do Porto. Este *dataset*, apresenta 28 atributos, alguns dos quais são irrelevantes devido à sua redundância. Contudo, nesta secção é apresentada a constituição completa e detalhada do *dataset* original, na Tabela 3.13. O conjunto de dados climatológicos é bastante complexo, uma vez que contém muitos atributos. Este *dataset* contém 102839 observações, com registos da informação meteorológica relativa à cidade do Porto desde 2011-01-01 até 2022-02-01.

Atributo	Descrição	Tipo de dados	Unidade de medida
city_name	Nome da cidade	*string*	string
lat	Coordenadas geográficas do local (latitude)	*float64*	°C
lon	Coordenadas geográficas do local (longitude)	*float64*	°C
temp	Valor da temperatura	*float64*	°C
temp_min	Valor da temperatura mínima	*float64*	°C
temp_max	Valor da temperatura máxima	*float64*	°C
feels_like	Valor da sensação térmica atmosférica	*float64*	hPa
feels_like	Valor da sensação térmica atmosférica	*float64*	hPa
pressure	Valor médio da pressão atmosférica	*float64*	hPa
Humidity	Percentagem média da humidade	*float64*	%
dew_point	Valor do ponto de condensação da água	*float64*	kelvin
wind_speed	Velocidade do vento	*float64*	°C
wind_deg	Grau do vento	*float64*	mm
wind_gust	Rajada de vento	*float64*	m/s
clouds_all	Percentagem de Nebulosidade	*float64*	%
rain_1h	Volume de chuva na última hora	*float64*	mm
rain_3h	Volume de chuva nas últimas 3 horas	*float64*	mm
snow_1h	Volume de neve na última hora	*float64*	mm
snow_3h	Volume de neve nas últimas 3 horas	*float64*	mm
weather_id	ID da condição meteorológica	*float64*	mm
weather_main	Grupo de parâmetros climáticos	*float64*	mm
weather_description	Condições meteorológicas dentro do grupo	*float64*	mm
visibility	Valor da visibilidade média	*float64*	m
dt	Tempo de cálculo de dados	*float64*	*utc*
dt_iso	Data e hora	*float64*	*utc*
timezone	Deslocamento em segundos	*float64*	*utc*
sea_level	Nível do mar	*float64*	*utc*
grnd_level	Nível do solo	*float64*	*utc*

Tabela 3.13: Constituição do *dataset* do dados climatológicos

Ao contrário do que sucede com os *datasets* dos poluentes atmosféricos, este apresenta vários atributos com *missing values*. Na Tabela 3.14, encontra-se evidenciado os atributos com *missing value* e respetiva percentagem.

Atributo	Total de *Missing values*	Percentagem total
visibility	1177	1.1%
sea_level	102839	100%
grnd_level	102839	100%
wind_gust	67434	65.6%
rain_1h	81607	79.4%
rain_3h	102839	100%

Tabela 3.14: Análise dos atributos com *Missing values*

Da informação verificada na tabela, é possível verificar que existem pelo menos 3 atributos sem qualquer registo, pelo que não terão qualquer importância para este estudo. Para além desta informação, existem 4 atributos que contém zeros, neste caso todos os atributos relacionados com o vento, bem como um atributo relativo à nebulosidade. O atributo *wind_speed* tem 5629 zeros, o que equivale a 5.5% dos

valores. O atributo *wind_deg*tem 12465 zeros, ou seja, 12.1%. O *wind_gust* 1442 zeros, perfazendo 1.4% dos valores observados e por fim, o *clouds_all* tem 38083 zeros, ou seja, 37% das observações registadas.

3.3 Preparação dos dados

Nesta secção, os diferentes conjuntos de dados foram trabalhados de forma a garantirem a sua aplicabilidade nos diferentes modelos a implementar nas fases seguintes da dissertação. Estes processos de tratamento e preparação de dados foram aplicados a todos os parâmetros anteriormente mencionados.

3.3.1 Feature Engineering

Os valores observados em cada conjunto de dados disponibilizado, estavam associados a um *timestamp*, constituído pela data, hora e *Timezone* de cada observação. Para uma melhor análise foi decidido extrair apenas a data deste parâmetro. A hora e o *Timezone* foram descartados. Assim, trabalhando apenas com a data, o primeiro passo foi extrair o ano, o número associado a cada mês, o dia e a estação do ano, criando assim 3 novos atributos. A partir da análise das novas informações, foi possível inferir que utilizar um *timestamp* diário seria o mais adequado, uma vez que a maior parte das medições ocorriam com periodicidade diária, havendo dias com mais do que uma observação e até mesmo dias em falta, isto é, dias onde não houve registo de observações. Seguidamente, e com base no supracitado, foi realizado um *group by*, de forma a agrupar as observações por data, calculando a média das observações da substância nesse mesmo período, ficando assim os *datasets* com as suas observações bastante reduzidas, uma vez que, neste ponto existia apenas uma entrada por dia, com o seu respetivo valor médio. Cada conjunto de dados de cada poluente passou por este processo.

3.3.2 Identificação da captura de dados

Com os resultados da alteração da subsecção anterior, foi possível observar que existiam vários dias com observações nulas, isto é, com valor 0, em cada *dataset*. Quando este tipo de situação acontecia, foi considerado que houve falhas por parte do registo das substâncias nesses dias. Assim sendo, para resolver tal problema, atribuiu-se o valor de *NaN* a cada uma dessas entradas, utilizando a função de *replace*, que substituiu os 0 por *NaN*. Também foi identificado como mencionado anteriormente que determinados dias não tinham entradas. De forma a lidar com esta situação foi criado uma espécie de calendário que abrange a data mínima de todos os *datasets* e a data máxima, como limites. Dessa forma, agregaram-se os valores pela sua média. Assim, todas as entradas desde 2017 até 2021 foram preenchidas, e nos casos em que não havia registos, foram atribuídos valores de *NaN*. Este processo, foi, uma vez mais aplicado a cada conjunto de dados de cada poluente.

3.3.3 Identificação e tratamento dos *missing values*

Nesta fase, quando não existe registo de observações em determinada data ou quando esse registo é 0, pressupõe-se agora, que se trata de um *missing value*. Para se resolverem estes problemas de *NaN*, foi decidido substituir o valor de cada *NaN* pela média das últimas 7 observações. Contudo, quando o valor de *NaN* se encontrava nos primeiros 7 dias , o método usado teve que ser ligeiramente ajustado. Existiu a necessidade de criar um ciclo *while* que verifica se nas primeiras sete posições ainda existem *NaN* e este ciclo termina quando as primeiras sete posições estão preenchidas. Para este procedimento, foi necessário criar uma função, denominada *impute*, que numa primeira análise verifica a existência de *NaN* nas primeiras 7 linhas, através da instrução de *if*. Foram adicionalmente definidas algumas variáveis auxiliares que ajudaram a construir a função, tais como uma variável *j* responsável para iterar sobre os índices, uma outra variável *c* que conta quantos valores estão a ser considerados na média até ao momento, a variável *média* e a variável *k* que indica o índice da posição onde se encontra o valor *NaN*. Assim, através desta função para cada *NaN* que surge nas primeiras sete posições, onde se vai buscar os primeiros 7 valores que surgem que não são *NaN* nem são 0, e é colocado o valor da média dessas sete observações no lugar do *NaN*. Depois, para o segundo valor a *NaN*, volta-se a utilizar os valores próximos a ele, incluíndo o valor anterior calculado, ou seja, a função capta o primeiro valor, resultado da média anteriormente mencionada e os seis próximos valores para o cálculo do novo valor, e assim sucessivamente. Caso o *dataset* em questão seja composto por um conjunto de dados onde as primeiras oito observações são *NaN*, o procedimento usado é o descrito anteriormente, e o oitavo elemento é composto pela média dos sete registos seguintes. Desta forma, é garantido que os primeiros sete valores são preenchidos com os valores mais próximos a ele, seja inferiormente ou superiormente.

3.3.4 Tratamento dos dados do *dataset* do clima

Dentro deste conjunto de dados, observam-se vários atributos num intervalo de tempo bastante superior ao dos *datasets* dos poluentes. Em relação aos atributos foram selecionados 8:

1. *Date*
2. Temperatura
3. Pressão
4. Humidade
5. Velocidade do vento
6. Nebulosidade
7. Chuva
8. Visibilidade

Portanto, o primeiro passo foi efetivamente, filtrar as colunas que de facto foram consideradas relevantes pela informação que acrescentavam e pela análise da sua correlação. Seguidamente foi novamente necessário fazer o devido tratamento à coluna responsável pela data. Para isso, foi selecionada a coluna "dt_iso"e converteu-se a mesma numa variável do tipo *date*, excluíndo a hora e a informação relativa ao UTC, usando as funções *datetime*, *strptime* e *date*. Por fim, foi renomeada a coluna através da função *rename* para *date*.

Também neste conjunto de dados foi aplicada a estratégia do calendário, utilizando para isso a função *date_range* com início na data mínima dos *datasets* dos poluentes e terminando na última data dos dados do clima. Foi feita uma vez mais a substituição dos zeros, por valores *NaN*.

3.3.5 Concatenação dos *datasets*

Uma vez que neste momento já se tinham reunidas todas as informações sob a forma que nos era mais favorável, foi então o momento de agregar toda a informação num único *dataset*. Assim, foi necessário fazer a leitura e filtro de cada *CSV*, através das funções *read_csv* e *filter*, respetivamente. De cada *dataset* de cada poluente foram extraídas apenas a informação relativa à data e ao valor do poluente. No *dataset* do clima, foi extraída a data e a informação referida na secção anterior. Cada valor de cada poluente é renomeado pela representação quimíca de cada um deles : CO, NO_2, O_3, PM_{10}, $PM_{2,5}$ e SO_2. Posto isto, foi então possível concatenar as várias colunas com os vários *dataframes*, utilizando um ciclo *for* e a função *concat*.

3.3.6 Cálculo do IQA

Após obter o novo conjunto de dados ordenado pela data e com as colunas de cada poluente, foi necessário criar um novo atributo. O atributo IQA, que tal como a designação sugere, é a coluna que contem o cálculo do índice da qualidade do ar para cada dia. Assim, tendo como base as concentrações dos poluentes e das faixas de concentração definidas para cada poluente, o cálculo do IQA é dado pela seguinte equação [112]:

$$I_p = \frac{I_f - I_i}{C_f - C_i}(C - C_i) + I_i \tag{3.1}$$

onde, I_p = índice para o poluente p

I_f = valor do IQA máximo do intervalo onde o poluente p se encontra

I_i = valor do IQA mínimo do intervalo onde o poluente p se encontra

C_f = valor máximo do intervalo de concentração onde o poluente p se encontra

C_i = valor mínimo do intervalo de concentração onde o poluente p se encontra

C = concentração média do poluente p

O valor I_p é calculado para cada poluente, e a qualidade do ar será classificada a partir do maior índice, ou seja, será determinada pelo poluente que apresentar o pior resultado para I_p, ou seja:

$$IQA = max(I_{p_1}, I_{p_2}, ..., I_{p_n}) \tag{3.2}$$

Para que fosse possível implementar esta formula em *python*, foi necessário definir o valor máximo e mínimo associado ao IQA, respetivamente de 0 e 456. Também foi necessário criar um dicionário com os valores máximo e mínimo de cada poluente recorrendo aos valores mencionados na revisão da literatura, e que correspondem aos valores de C_f e C_i.

3.3.7 Análise dos Outliers

De forma a compreender melhor o *dataset* foi feita a análise dos *outliers* de cada poluente. Os *outliers* são dados que se diferenciam drasticamente de todos os outros, ou seja, corresponde a um valor que foge da normalidade e que pode causar anomalias nos resultados a obter nas fases mais avançadas desta dissertação [113].

A forma através da qual se fez esta análise, foi através da ferramenta gráfica do *boxplot* que, de uma forma geral, faz a comparação entre os diferentes grupos face à posição, à dispersão e à distribuição dos dados [114].

Através da Figura 3.13 é possível observar seis caixas, correspondentes aos seis poluentes em estudo, e é também através desta figura que conseguimos comparar os seis casos simultaneamente. A verde está desenhada a linha interna que designa o segundo quartil ou mediana. Ao analisar esta linha, percebeu-se que a sua posição é bastante variável de caixa para caixa. No caso do CO a mediana encontra-se praticamente na zona central, tal como acontece com os poluentes NO_2 e SO_2. Este comportamento indica que o conjunto de dados apresenta uma distribuição simétrica, pois a mediana é próxima da média. Por outro lado, o $PM_{2,5}$ apresenta a sua mediana muito próxima do terceiro quadril, tratando-se assim, de dados que são assimétricos negativamente. Quanto à análise da mediana dos poluentes O_3 e PM_{10}, estas por sua vez, encontram-se mais próximas ao primeiro quartil, revelando desta forma que os dados são assimétricos positivamente.

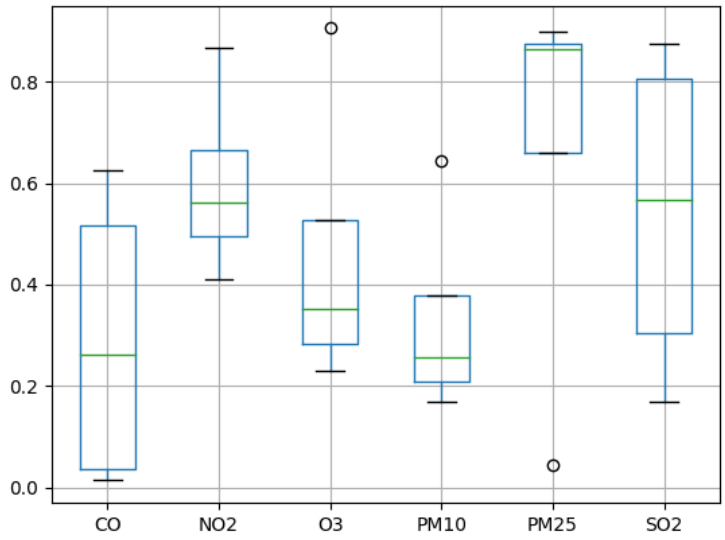


Figura 3.13: Boxplot de todos os poluentes após tratamento dos dados

Uma outra característica a analisar é a variabilidade dos dados. Na figura acima verifica-se que o CO e o SO_2 são os poluentes que apresentam maior variabilidade, uma vez que as suas caixas apresentam os seus extremos mais afastados entre si, tornando, desta forma, as caixas maiores. Observa-se também que a concentração média e mediana do $PM_{2,5}$ é maior do que as restantes, uma vez que apresenta valores mais elevados. Por fim, é possível verificar a existência de um *outlier* nos poluentes O_3, PM_{10} e $PM_{2,5}$, respetivamente. Sendo que nos dois primeiros casos mencionados, os *outlieres* surgem acima do valor máximo e o terceiro, abaixo do valor mínimo [113].

3.3.8 Análise de Correlação dos dados

De forma a perceber quais os atributos que têm um maior impacto sobre os poluentes foi realizado um estudo de correlação. Devido à distribuição dos dados não seguir uma distribuição normal, usualmente designada por gaussiana, foi aplicado o método da correlação de *Spearman* [115]. A Figura 3.14 apresenta a matriz de correlação dos dados do *dataset*. Note que, esta correlação não leva em consideração a distribuição dos dados.

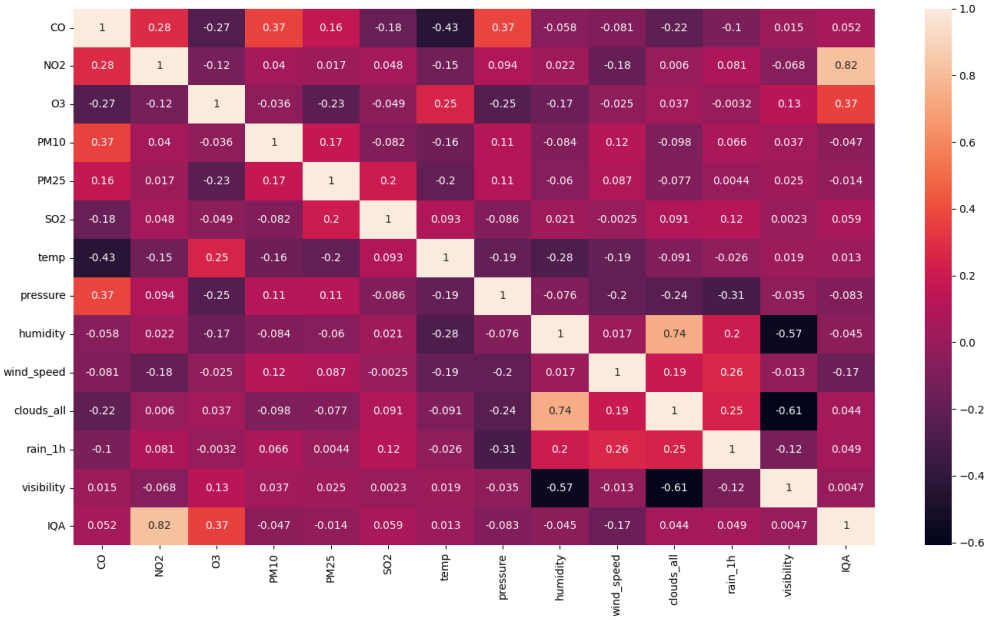


Figura 3.14: Matriz de correlação dos dados

A análise da matriz revela que não existe uma correlação relevante entre os recursos que não são utilizados como destinos. Além disso, a única correlação notável é uma relação positiva entre NO_2 e IQA, enquanto há uma correlação negativa entre as características CO e *Temperature*. Esta análise sugere que estes recursos podem desempenhar um papel significativo para treinar o modelo.

3.4 Tecnologias usadas

Esta dissertação foi apoiada pela linguagem de programação *python*(versão 3.9), tendo esta sido implementada no Integrated Development Environment (IDE) *PyCharm*. Através desta linguagem de programação foi possível desenvolver todo o processo desde a análise dos dados brutos até ao tratamento dos resultados.

Este processamento e implementação tiveram como alicerce diversas bibliotecas como o *pandas*, *matplotlib*, *seaborn*, *NumPy* e *scikit-learn*. Todo este processo foi desenvolvido num *MackBook Air (M1, 2020)*, com um processador Processador M1 da *Apple*, contendo uma 8 GB de memória unificada e um *SSD* de 256 GB.

Capítulo 4

Experiências computacionais

No presente capítulo, são minuciosamente detalhadas todas as abordagens, decisões e técnicas adotadas nas experiências realizadas. Inicialmente, são delineadas as configurações experimentais, conforme identificado na secção 4.1. É importante destacar que são abordados dois cenários, para cada modelo, o *Univariate* e *Multivariate*. Na secção 2.4, são apresentadas as métricas de avaliação adotadas. Posteriormente, na secção 4.2, é discutida a modelação e otimização dos hiperparâmetros para cada modelo.

4.1 Configuração Experimental - Cenários

No que concerne à conceção dos modelos de DL, as diversas experiências desenvolvidas tiveram como base dois cenários. O cenário 1 constituí aquele que é o mais simples, por apenas incluir uma variável, designando-se assim, por *Univariate*. Neste caso foi considerado a própria variável, no caso a IQA. Por outro lado, o cenário 2, considera também a variável NO_2 para prever o IQA, designando-se por *Multivariate*. Este último valor foi considerando no segundo cenário, uma vez que, na análise anteriormente feita em 3.3.8, o NO_2 correlacionou-se fortemente com o IQA, tornando-se assim num dado importante para a sua previsão. Com o intuito de compreender o impacto que estes dois cenários têm na previsão do IQA foram considerados para cada cenário, quatro modelos: LSTM, MLP, CNN e GRU. Dentro destes quatro modelos foram utilizadas duas métricas de avaliação que serão também abordadas nesta secção. Os *datasets* de cada cenário foram tratados da mesma forma, com variação apenas nas *features*. Todas as *features* foram normalizadas usando a função *MinMaxScaler* para melhorar a qualidade dos dados antes de serem usados nos modelos. No caso, para a LSTM, o intervalo escolhido foi de -1 a 1, enquanto para MLP, GRU e CNN o intervalo adotado foi de 0 a 1. Essa normalização foi feita para garantir que as diferentes arquiteturas de redes neuronais recebem dados normalizados de acordo com as especificações apropriadas para cada uma. A Figura 4.1 evidencia de forma esquemática a configuração das experiências da presente dissertação.

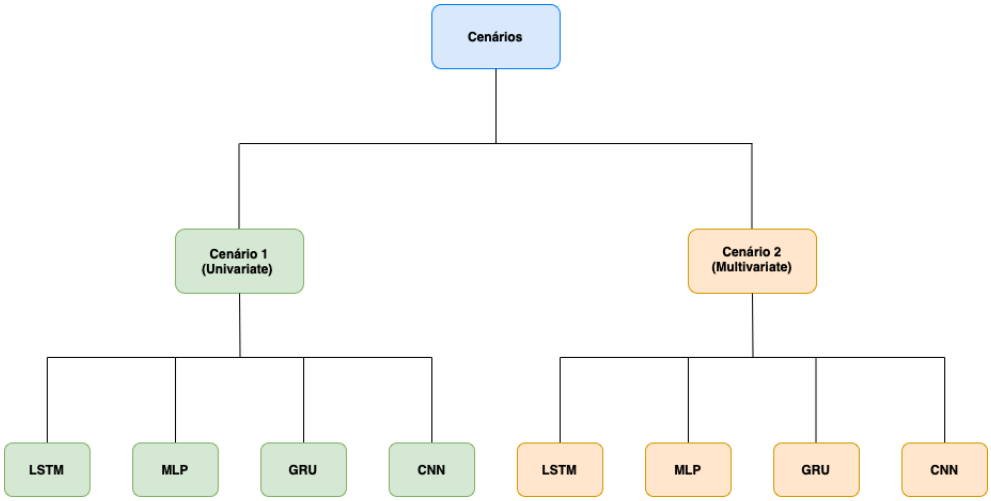


Figura 4.1: Esquematização dos cenários das experiências computacionais

4.2 Modelação e otimização dos hiperparâmetros

O foco primordial desta secção é o de explicar o processo de *tunning*, isto é, o processo de ajuste dos modelos a utilizar. Esta secção está dividida em duas partes. A primeira parte diz respeito à análise dos hiperparâmetros dos modelos LSTM, GRU e MLP. A segunda parte faz a análise aos hiperparâmetros do modelo CNN. Esta divisão advém do facto de que os três primeiros modelos referidos partilham o mesmo conjunto de hiperparâmetros e o último tem algumas particularidades, como veremos ao longo desta secção.

Para além dos hiperparâmetros que foram considerados para cada modelo, também foi importante encontrar o número ideal de *epochs*. As *epochs* são outro parâmetro que todos os modelos têm em comum. Contudo, cada modelo de cada cenário, obteve diferentes valores. Esses valores foram determinados ao analisar os gráficos das *learning curves*, identificando o ponto de interseção entre as curvas de treino e teste. A Tabela 4.1 contém a informação relativa ao número de *epochs* do cenário *Univariate*.

Modelo	Número de epochs
LSTM	50
MLP	25
GRU	75
CNN	45

Tabela 4.1: Valores de *epochs* para cada modelo no cenário *Univariate*

Relativamente ao cenário *Multivariate*, os valores são os que se encontram na Tabela 4.2

Modelo	**Número de *epochs***
LSTM	80
MLP	45
GRU	75
CNN	50

Tabela 4.2: Valores de *epochs* para cada modelo no cenário *Multivariate*

4.2.1 Hiperparâmetros dos modelos LSTM, GRU e MLP

Tal como referido na secção anterior, os modelos LSTM, GRU e MLP partilham o mesmo conjunto de hiperparâmetros, que são o número de camadas, o número de neurónios, a *dropout_rate*, as funções de ativação, o número de *timesteps* e o tamanho do *batch*.

Hiperpâmetro	**Intervalo de valores**
layers	[3, 4, 5]
neurons	[32, 64, 128]
dropout_rate	[0.0, 0.5]
activation	['relu', 'tanh']
timesteps	[14, 21, 28]
batch_size	[5, 10, 20]

Tabela 4.3: Valores considerados para os hiperparâmetros dos modelos

A Tabela 4.3 inclui os seis parâmetros que usualmente são utilizados neste tipo de modelos [95]:

- *layers*: Número de camadas do modelo.
- *neurons*: Número de neurónios por camada.
- *dropout_rate*: Valor decimal entre 0 e 1 que representa a fração de valores a descartar.
- *activation*: função de ativação a utilizar em cada camada.
- *timesteps*: Número de observações que são passadas ao modelo.
- *batch_size*: Número pelo qual o número de observações é divisível. Define após quantas amostras os pesos serão atualizados.

4.2.2 Hiperparâmetros dos modelos CNN

De forma a verificar os diversos hiperparâmetros considerados nos modelos CNN, a Tabela 4.4 apresenta os intervalos de valores utilizados.

Hiperpâmetro	Intervalo de valores
layers	[3, 4, 5]
dropout_rate	[0.0, 0.5]
activation	['relu', 'tanh']
timesteps	[14, 21, 28]
batch_size	[5, 10, 20]
filters	[16, 32]
kernel_size	[3, 4, 5]
pool_size	[2, 3]

Tabela 4.4: Valores considerados para os hiperparâmetros do modelo CNN

Dos hiperparâmetros adicionados a este modelo destacam-se:

- *filters* : Número de filtros que visam aprender padrões na sequência
- *kernel_size*: Tamanho do *Kernel* utilizado no modelo
- *pool_size*: Tamanho da *pool* na camada de *pooling*

4.3 Implementação dos Modelos Candidatos

Considerando a variedade de modelos utilizados nas experiências desta dissertação, este subcapítulo apresenta as abordagens para a implementação dos modelos, bem como os valores dos hiperparâmetros utilizados na configuração das experiências. Nesta dissertação, os modelos de DL selecionados para a previsão do IQA, foram LSTM, MLP, GRU e CNN. Os *datasets* utilizados passaram por um processo de *cross-validation* com o *TimeSeriesSplit*, garantindo assim a preservação da ordem temporal dos dados. Este *cross-validation* foi configurado com *k* igual a 3, e foi a partir do mesmo que foram criados os conjuntos de treino, teste, e validação. Adicionalmente, o processo de previsão consistiu numa abordagem *multi-step recursive forecasting*, utilizando 2 *multi-steps*. Isto significa que, na etapa de validação dos modelos candidatos, são avaliadas previsões para 2 *timesteps* futuros, recursivamente. Ou seja, após efetuar a previsão do primeiro *timestep* futuro, esta é utilizada como *input* da previsão seguinte. Para assegurar a consistência e reprodutibilidade dos resultados, uma *SEED* foi especificada em todos os modelos, com o valor fixado em 91195003.

4.3.1 LSTM

A listagem 4.1 evidencia um excerto do código utilizado na implementação do modelo LSTM. O código utiliza um *loop* que itera no intervalo de camadas, ou seja, no número de camadas LSTM na rede. Dentro do *loop*, existem declarações condicionais que determinam o comportamento de cada camada LSTM com base na sua posição:

- Se i (a camada atual) for 0, significa que é a primeira camada LSTM. Se, i+1 for igual ao valor de *layers*, significa que é também a última camada LSTM e, se assim for, é criada uma camada LSTM com neurónios/2 e *return_sequences=False*, ou seja, não retorna sequências para a próxima camada. Em contrapartida, se não for a última camada, cria uma camada LSTM com determinado número de neurónios e *return_sequences=True*, retornando assim, sequências para a próxima camada.

- Se i+1 (a próxima camada) for igual a *layers*, é criada uma camada LSTM com neurónios*2 e *return_sequences=False*.

- Se nenhuma das condições anteriormente referidas for válida, isso significa que a camada atual não é a primeira nem a última. E neste caso é criada uma camada LSTM com neurónios e *return_sequences=True*.

Fora do *loop*, é adicionada uma camada *Dense* com determinado número de neurónios e função de ativação. Uma camada *dropout* com um determinado *dropout_rate* é aplicada para introduzir a regularização e evitar o *overfitting*. Por fim, é adicionada uma camada de saída com uma única unidade (*Dense(1)*).

Listagem 4.1: Modelo LSTM

```
for i in range(layers):
    #if first LSTM layer
    if i == 0:
        #if also the last LSTM layer then return_sequences=False
        if i+1 == layers:
            x = LSTM(neurons/2, return_sequences=False)(inputs)
        #it has more layers! So return_sequences=True
        else:
            x = LSTM(neurons, return_sequences=True)(inputs)
    #if last LSTM layer then return_sequences=False
    elif i+1 == layers:
        x = LSTM(neurons*2, return_sequences=False)(x)
    #if not the last LSTM layer then return_sequences=True
    else:
        x = LSTM(neurons, return_sequences=True)(x)

x = Dense(neurons, activation=activation)(x)
x = Dropout(dropout_rate)(x)
outputs = Dense(1)(x)
```

4.3.2 MLP

O segundo modelo abordado é o MLP. Este modelo, tal como se pode visualizar na listagem 4.2 utiliza um *loop* que itera sobre o intervalo de camadas na rede. Uma vez mais, dentro do *loop*, existem determinadas condições que determinam o comportamento de cada camada conforme a sua posição:

- Se i (a camada atual) for 0, significa que é a primeira camada. Se, por outro lado, (i+1 == *layers*), então esta é a última camada e, nesse caso, é criada uma camada *Dense* com um valor de neurónios e a função de ativação. Caso contrário, é criada uma camada *Dense* com neurónios/2 e a função de ativação.
- Se i+1 (a próxima camada) for igual a *layers*, é criada uma camada *Dense* com neurónios*2 e a função de ativação.
- Se nenhuma das condições acima for verdadeira, isso significa que a camada atual não é a primeira nem a última. E nesse caso, é criada uma camada *Dense* com determinado número de neurónios e uma vez mais a função de ativação.

Após o *loop*, é aplicada uma camada *dropout* com um determinado *dropout_rate*. Por fim, é adicionada uma camada de saída com uma única unidade (*Dense(1)*).

Listagem 4.2: Modelo MLP

```
for i in range(layers):
     #if first Dense layer
     if i == 0:
        #and also the last
        if i+1 == layers:
           x = Dense(neurons, activation=activation)(f)
        #it has more layers
        else:
           x = Dense(neurons/2, activation=activation)(f)
     #if last Dense layer
     elif i+1 == layers:
        x = Dense(neurons*2, activation=activation)(x)
     #if middle layers
     else:
        x = Dense(neurons, activation=activation)(x)

  x = Dropout(dropout_rate)(x)
  outputs = Dense(1)(x)
```

4.3.3 GRU

A listagem 4.3 evidencia um excerto do código utilizado na implementação do modelo GRU. Equiparado aos modelos anteriores, o excerto de código que se segue também faz uso de um *loop* que itera sobre o intervalo de camadas na rede. No conteúdo do *loop*, existem declarações condicionais que determinam o comportamento de cada camada da GRU, de acordo com a sua posição:

- Se i (a camada atual) for 0, significa que é a primeira camada GRU. Se (i+1 == *layers*) então trata-se da última camada GRU e, se assim for, cria uma camada GRU com neurónios/2 e *return_sequences=False*, ou seja, não retorna qualquer sequência para a próxima camada. Se não for a última camada, cria uma camada GRU com unidades de neurónios e *return_sequences=True*, ou seja, retorna sequências para a próxima camada.

- Se i+1 (próxima camada) for igual a *layers*, significa que é a última camada da GRU. Neste caso, é criada uma camada GRU com neurónios*2 e *return_sequences=False*.

- Se nenhuma das condições acima for verificada, isso significa que a camada atual não é nem a primeira nem a última. E assim, é criada uma camada GRU com unidades de neurónios e *return_sequences=True*.

Fora do *loop*, é adicionada à rede uma camada *Dense* com unidades de neurónios e a função de ativação. Para além disto, é também aplicada uma camada *dropout* com um *dropout_rate* especificado. Por fim, uma camada de saída com uma única unidade (*Dense(1)*) é adicionada.

Listagem 4.3: Modelo GRU

```
for i in range(layers):
     #if first GRU layer
     if i == 0:
        #if also the last GRU layer then return_sequences=False
        if i+1 == layers:
           x = GRU(neurons/2, return_sequences=False)(inputs)
        #it has more layers! So return_sequences=True
        else:
           x = GRU(neurons, return_sequences=True)(inputs)
     #if last GRU layer then return_sequences=False
     elif i+1 == layers:
        x = GRU(neurons*2, return_sequences=False)(x)
     #if not the last GRU layer then return_sequences=True
     else:
        x = GRU(neurons, return_sequences=True)(x)

  x = Dense(neurons, activation=activation)(x)
  x = Dropout(dropout_rate)(x)
  outputs = Dense(1)(x)
```

4.3.4 CNN

Por fim, é apresentado o modelo CNN. Neste último modelo são inseridos alguns parâmetros para além dos que os modelos anteriormente abordados incluíam, como *filters*, *kernel size* e *pool size* e parâmetro neurónios foi descartado. Na listagem 4.4 verificámos que o código utiliza um *loop* que itera sobre o intervalo de camadas CNN na rede. Dentro do *loop*, existem declarações condicionais que determinam o comportamento de cada camada da CNN com base em sua posição:

- O código verifica se i (a camada atual) é igual a 0, isto é, se se trata da primeira camada CNN. Além disso, efetua a verificação da última camada CNN através da condição (i+1 == *layers*) e, em caso afirmativo, cria uma camada convolucional 1D, com o nome de *Conv1D* com *filters*/2, um determinado *kernel_size* e uma função de ativação especificada. Adicionalmente é aplicada a *AveragePooling1D* com *pool_size* e *data_format* específicos. Por fim, utiliza a função *Flatten()*, responsável por aplanar o *output*.

- Se i+1 (a próxima camada) for igual a layers (i+1 == layers), ou seja, quando se refere à última camada, é criada uma camada convolucional 1D (Conv1D) com filters*2, um determinado *kernel_size* e uma função de ativação especificada. É, ainda, aplicado um *pool* médio, a *AveragePooling1D* com *pool_size* e *data_format* especificados. Por fim, é utilizada a função *Flatten()*.

- Se nenhuma das condições acima for verificada, então, a camada atual não é a primeira nem a última. Desta forma, é criada uma camada convolucional 1D, com *filters=filters*, um *kernel_size*

específico e a função de ativação. Além disso, é aplicado um *pool* médio, designadamente o *AveragePooling1D* com *pool_size* e *data_format* especificados.

Após a conclusão do *loop*, é adicionada uma camada densa de *filters* e também uma camada *dropout* com um *dropout_rate* especificado, de forma a evitar o *overfitting*. Por fim, uma camada de saída com uma única unidade (*Dense(1)*) é adicionada.

Listagem 4.4: Modelo CNN

```

    for i in range(layers):
        #if first CNN layer
        if i == 0:
            #and also the last
            if i+1 == layers:
                x = Conv1D(filters=filters/2, kernel_size=kernel_size, activation=activation,
                    ↪ data_format='channels_last')(inputs)
                x = AveragePooling1D(pool_size=pool_size, data_format='channels_first')(x)
                x = Flatten()(x)
            #it has more layers
            else:
                x = Conv1D(filters=filters, kernel_size=kernel_size, activation=activation,
                    ↪ data_format='channels_last')(inputs)
                x = AveragePooling1D(pool_size=pool_size, data_format='channels_first')(x)
        #if last layer
        elif i+1 == layers:
            x = Conv1D(filters=filters*2, kernel_size=kernel_size, activation=activation,
                ↪ data_format='channels_last')(x)
            x = AveragePooling1D(pool_size=pool_size, data_format='channels_first')(x)
            x = Flatten()(x)
        #if middle layers
        else:
            x = Conv1D(filters=filters, kernel_size=kernel_size, activation=activation,
                ↪ data_format='channels_last')(x)
            x = AveragePooling1D(pool_size=pool_size, data_format='channels_first')(x)

    x = Dense(filters)(x)
    x = Dropout(dropout_rate)(x)
    outputs = Dense(1)(x)
```

Capítulo 5

Discussão dos resultados obtidos

O presente capítulo apresenta os resultados das experiências identificadas no Capítulo 4. Importa destacar que as métricas de avaliação foram calculadas durante a validação, seguindo a abordagem *multi-step recursive forecasting*. Como previamente verificado, a análise contempla dois cenários de teste, sendo que para cada cenário foram escolhidos e aplicados quatro modelos distintos para efetuar as previsões, culminando em várias experiências.

Posteriormente, os resultados e discussões das previsões do IQA *Univariate* e *Multivariate*, estão apresentados nas secções 5.1, e 5.2, respetivamente. Cada secção contém as subsecções correspondentes a cada modelo aplicado, onde os resultados são exibidos no formato tabular. Finalmente, é apresentada a secção 5.3 onde se faz uma análise comparativa da *performance* dos modelos candidatos.

5.1 Previsão IQA - Cenário *Univariate*

Primeiramente, são apresentados os resultados de todas as experiências consideradas na previsão do IQA *Univariate*. Desta forma, é mostrada uma tabela com o Top-5 dos melhores modelos candidatos obtidos para cada modelo de DL utilizado neste cenário.

5.1.1 LSTM

Na presente subsecção são apresentados os resultados dos Top-5 modelos candidatos LSTM, num cenário *Univariate*.

Observando os dados da Tabela 5.1, é evidente que o melhor modelo candidato LSTM, para este cenário, foi avaliado com um valor de RMSE de aproximadamente 4.84 e um valor de MAE aproximado de 4.44. O modelo aplicou 14 *timesteps* de *input* e o valor 0.5 de *dropout rate* para obter o melhor

a	b	c	d	e	f	g	h	i
3	128	0.5	tanh	14	20	50	4.44	4.84
3	64	0.5	tanh	28	20	50	4.58	4.86
3	32	0.5	relu	28	20	50	4.71	5.04
3	128	0.5	relu	14	20	50	4.71	5.07
3	32	0.5	tanh	28	20	50	4.77	5.10

Tabela 5.1: Top-5 dos melhores modelos candidatos LSTM (As letras representam: a. *layers*; b. *neurons*; c.*dropout rate*; d.*activation*; e. *timesteps*; f. *batch size*; g. *epochs*; h. MAE; i. RMSE)

desempenho. Relativamente à função de ativação utilizada, verificou-se que a TanH resultou numa melhor *performance*. Além disso, o número de neurónios, *layers*, e *batch size*, foram 128, 3, e 20, respetivamente.

Analisando o top-5 da tabela, de uma forma generalizada, verifica-se a existência de alguns hiperparâmetros homogéneos, nomeadamente os *layers* (3), o número de *dropout rate* (0.5) e o número de *batch size* (20). Em contrapartida, os restantes parâmetros apresentam alguma heterogeneidade, nos seus valores como a função de ativação TanH e o número de *timesteps* (28).

5.1.2 MLP

Apresentam-se nesta subsecção os resultados dos Top-5 modelos candidatos MLP, para o cenário *Univariate*.

a	b	c	d	e	f	g	h	i
3	64	0.5	tanh	28	5	25	4.27	4.58
4	64	0.5	tanh	28	10	25	4.36	4.65
3	64	0.5	tanh	21	10	25	4.40	4.74
5	64	0.5	tanh	28	10	25	4.47	4.77
3	128	0.5	tanh	28	5	25	4.68	4.98

Tabela 5.2: Top-5 dos melhores modelos candidatos MLP (As letras representam: a. *layers*; b. *neurons*; c. *dropout rate*; d. *activation*; e. *timesteps*; f. *batch size*; g. *epochs*; h. MAE; i. RMSE)

De acordo com os resultados apresentados na Tabela 5.2, verifica-se que o melhor modelo candidato apresenta um valor de RMSE de aproximadamente de 4.58 e um valor de MAE aproximado de 4.27. Foram utilizados 28 *timesteps* como dados de entrada, a função de ativação TanH, e 3 camadas na arquitetura deste modelo candidato. Cada camada foi suportada por 64 neurónios e a *dropout rate* considerada foi 0.5. Além disso, o número de *batch size* foi de 5.

Ao analisar o top-5 da tabela conclui-se que os valores são totalmente homogéneos na função de ativação (TanH). Para além disto, verifica-se que o número de *dropout rate* é o mesmo em todas as experiências (0.5). Contudo, alguns dos outros hiperparâmetros apresentam valores com bastante predominância, como é o caso de *layers* (3), neurónios (64) e *batch size* (10). Ao examinar o Top-5 na tabela, destaca-se que o modelo com a melhor performance revela-se com o menor número de neurónios, especificamente, 3.

5.1.3 GRU

Os resultados dos Top-5 modelos candidatos GRU, para o cenário Univariate, são apresentados na presenta subsecção.

a	b	c	d	e	f	g	h	i
3	32	0.5	relu	21	20	75	4.21	4.54
4	32	0.5	tanh	28	20	75	4.22	4.54
3	64	0.5	tanh	28	20	75	4.21	4.57
3	64	0.5	tanh	28	20	75	4.29	4.61
3	128	0.5	tanh	28	5	75	4.29	4.63

Tabela 5.3: Top-5 dos melhores modelos candidatos GRU (As letras representam: a. *layers*; b. *neurons*; c. *dropout rate*; d. *activation*; e. *timesteps*; f.*batch size*; g. *epochs*; h. MAE; i. RMSE)

Como se pode observar através da Tabela 5.3, verifica-se que o melhor modelo candidato apresenta um valor de RMSE de aproximadamente de 4.54 e um valor de MAE aproximado de 4.21. Foram utilizados 21 *timesteps* como dados de entrada, a função de ativação Unidade Linear Retificada (ReLU), e 3 camadas na arquitetura deste modelo candidato. Cada camada foi suportada por 32 neurónios, com *dropout rate* de 0.5 (indicando que 50% dos neurónios foram desativados em cada camada, ao longo do processo de treino) e um *batch size* de 20.

Através da análise do top-5 da tabela nota-se que os valores são totalmente homogéneos no hiperparâmetro de *dropout rate* (0.5). Por outro lado, não sendo homogéneos, alguns dos restantes hiperparâmetros apresentam predominantemente os mesmos valores, como é o caso de *layers* (3), *batch size* (20) e função de ativação TanH. Para além disto, ao analisar os valores deste top-5, verifica-se que as duas primeiras entradas da tabela alcançaram o mesmo valor de RMSE (4.54), embora tenham utilizado diferentes configurações de *layers*, diferentes funções de ativação e valores de *timesteps* distintos. Curiosamente, estas duas entradas da tabela utilizaram 32 neurónios na sua configuração. Este resultado sugere que um desempenho igual pode ser alcançado com diferentes configurações. Além disso, ao analisar a tabela de forma abrangente, nota-se uma tendência, um maior número de neurónios parece ter uma correlação inversa com o desempenho, indicando que modelos com um número menor de neurónios tendem a apresentar melhor *performance*.

5.1.4 CNN

Seguem-se os resultados dos Top-5 modelos candidatos CNN, para o cenário *Univariate*.

a	b	c	d	e	f	g	h	i	j	k
3	0.0	tanh	28	20	32	5	2	45	4.10	4.46
3	0.0	tanh	28	20	32	3	2	45	4.22	4.52
3	0.5	tanh	28	5	32	5	2	45	4.18	4.58
3	0.0	tanh	21	20	32	4	2	45	4.25	4.59
3	0.5	tanh	28	20	32	3	3	45	4.29	4.63

Tabela 5.4: Top-5 dos melhores modelos candidatos CNN (As letras representam: a.*layers*; b. *dropout rate*; c. *activation*; d. *timesteps*; e.*batch size*; f. *filters*; g. *kernel size*; h. *pool size*; i.*epochs*; j. MAE; k. RMSE)

Ao analisar a Tabela 5.4, o melhor modelo candidato baseado em CNN, obteve um valor de RMSE de aproximadamente de 4.46 e um valor de MAE aproximado de 4.10. O número de camadas utilizadas foi 3, *dropout rate* de 0.0 e a função de ativação TanH. Para além destes hiperparâmetros, outros foram tidos em consideração, nomeadamente, o número de *timesteps* de entrada igual a 28, um valor de filtros igual a 32, um tamanho de *kernel* igual a 5, e ainda um valor de 2 para a *pool size*.

Relativamente à homogeneidade de valores, quando se visualiza os resultados de todo o top-5, verifica-se que essa característica está presente no número de camadas igual a 3, na função de ativação TanH e no número de filtros igual a 32. Ainda assim, os restantes hiperparâmetros possuem valores predominantes, nomeadamente a *dropout rate* (0.0), *timesteps* (28), *batch size* (20) e *pool size* (2).

5.2 Previsão IQA - Cenário *Multivariate*

De seguida, são apresentados os resultados de todas as experiências consideradas na previsão do IQA *Multivariate*. Desta forma, é apresentada uma tabela com o Top-5 dos melhores modelos candidatos obtidos para cada modelo de DL utilizado neste cenário.

5.2.1 LSTM

Serve a presente subsecção para apresentar os resultados dos Top-5 modelos candidatos LSTM, para o cenário *Multivariate*.

a	b	c	d	e	f	g	h	i
3	32	0.5	tanh	28	20	80	4.87	5.19
3	64	0.5	relu	28	10	80	4.98	5.34
3	64	0.5	relu	28	5	80	4.98	5.35
3	128	0.5	tanh	28	20	80	5.01	5.36
3	64	0.5	relu	28	20	80	5.00	5.37

Tabela 5.5: Top-5 dos melhores modelos candidatos LSTM (As letras representam: a. *layers*; b. *neurons*; c.*dropout rate*; d. *activation*; e. *timesteps*; f. *batch size*; g. *epochs*; h. MAE; i. RMSE)

Através da análise dos valores da Tabela 5.5, observa-se que o melhor modelo candidato LSTM, para este cenário, foi avaliado com um valor de RMSE de aproximadamente 5.19 e um valor de MAE aproximado

de 4.87. O modelo candidato aplicou 28 *timesteps* de entrada e uma *dropout rate* igual a 0.5 para obter o melhor desempenho. Relativamente à função de ativação utilizada, verificou-se que a TanH resultou numa melhor *performance*. Além disso, o número de neurónios, *layers* e *batch size*, no melhor modelo candidato, foram 32, 3, e 20, respetivamente.

No geral, ao analisar o top-5 da tabela, verifica-se a existência de alguns hiperparâmetros homogéneos, nomeadamente as *layers* (3), o número de *dropout rate* (0.5) e o número de *timesteps* (28). Em contrapartida, os restantes hiperparâmetros não sendo homogéneos, apresentam alguma predominância nos seus valores, como é o caso do número de neurónios (64), na função de ativação ReLU e o número de *batch size* (20). Adicionalmente, olhando para a tabela de forma geral, percebe-se que na sua maioria, quanto menor o número de neurónios, melhor o desempenho dos modelos.

5.2.2 MLP

Os resultados dos Top-5 modelos candidatos MLP, para o cenário *Multivariate*, são exibidos na subsecção atual.

a	b	c	d	e	f	g	h	i
3	32	0.0	relu	14	10	45	4.04	4.46
3	32	0.0	relu	28	20	45	4.29	4.69
3	64	0.0	relu	21	10	45	4.61	4.99
3	128	0.0	relu	14	10	45	4.83	5.25
4	32	0.0	relu	28	20	45	4.94	5.29

Tabela 5.6: Top-5 dos melhores modelos candidatos MLP (As letras representam: a. *layers*; b. *neurons*; c.*dropout rate*; d.*activation*; e. *timesteps*; f. *batch size*; g. *epochs*; h. MAE; i. RMSE)

A Tabela 5.6 indica que o melhor modelo candidato apresenta um valor de RMSE de aproximadamente de 4.46 e um valor de MAE aproximado de 4.04. Foram utilizados 14 *timesteps* como dados de entrada, a função de ativação ReLU, e 3 *layers* na arquitetura deste modelo candidato. Cada camada foi suportada por 32 neurónios e a *dropout rate* considerada foi de 0.0. Além disso, o número de *batch size* foi de 10.

Consolidando o top-5 dos modelos candidatos, utilizados para prever o IQA numa abordagem *Multivariate* com modelos MLP , verifica-se uma homogeneidade nos valores relativos à função de ativação (ReLU). Adicionalmente, verifica-se que o número de *dropout rate* é o mesmo em todas as experiências (0.0). Por outro lado, alguns dos outros hiperparâmetros são apresentados de forma bastante predominante, como é o caso de *layers* (3), *neurons* (32) e *batch size*(10). Relativamente à *performance* do melhor modelo candidato, quando se compara a melhor e a pior *performance* dentro deste Top-5, nota-se uma tendência inversa entre o desempenho e o número de *layers*. O modelo candidato que alcançou um desempenho superior utilizou 3 *layers* e 32 *neurons* na sua configuração, enquanto que, quando configurado com 4 *layers* e 32 *neurons*, apresentou uma *performance* inferior. Essa observação sugere uma relação inversamente proporcional entre o número de *layers* e o desempenho do modelo candidato.

5.2.3 GRU

De seguida são apresentados os resultados dos Top-5 modelos candidatos GRU, para o cenário Multivariate.

a	b	c	d	e	f	g	h	i
4	64	0.5	tanh	21	20	75	4.20	4.59
3	32	0.0	tanh	28	20	75	4.37	4.65
3	32	0.5	tanh	28	20	75	4.44	4.73
3	64	0.0	tanh	28	20	75	4.37	4.74
3	32	0.5	relu	28	20	75	4.46	4.80

Tabela 5.7: Top-5 dos melhores modelos candidatos GRU (As letras representam: a. *layers*; b. *neurons*; c. *dropout rate*; d. *activation*; e. *timesteps*; f. *batch size*; g. *epochs*; h. MAE; i. RMSE)

O conteúdo da Tabela 5.7, indica que o melhor modelo candidato apresenta um valor de RMSE de aproximadamente de 4.59 e um valor de MAE aproximado de 4.20. Utilizaram-se 21 *timesteps*, a função de ativação TanH e 4 camadas na arquitetura do presente modelo candidato. Cada camada foi suportada por 64 neurónios com uma *dropout rate* de 0.5 (indicando que 50% dos neurónios foram desativados em cada camada, ao longo do processo de treino) e um *batch size* igual a 20.

Quando se analisa o top-5 da tabela verifica-se que os valores são totalmente homogéneos no hiperparâmetro de *batch size* (20).

Ao analisar o comportamento dos restantes hiperparâmetros , verifica-se uma certa predominância nos seus valores, nomeadamente nos valores de *layers* (3), no número de neurónios (32), na função de ativação TanH e nos valores de *timesteps* (28). No contexto deste Top-5, observa-se uma redução no desempenho associada à diminuição do número de *layers* e ao aumento de *timesteps*. Adicionalmente, constata-se uma diminuição da *performance* correlacionada à redução da quantidade de neurónios, exceto no caso do quarto melhor modelo candidato.

5.2.4 CNN

Seguidamente são apresentados os resultados dos Top-5 modelos candidatos CNN, para o cenário *Multivariate*.

a	b	c	d	e	f	g	h	i	j	k
3	0.5	tanh	21	20	32	3	3	50	4.18	4.52
3	0.5	tanh	21	20	16	3	2	50	4.30	4.64
3	0.0	tanh	21	20	32	3	2	50	4.29	4.67
4	0.5	tanh	14	20	32	3	3	50	4.34	4.71
3	0.0	tanh	28	10	32	5	2	50	4.35	4.74

Tabela 5.8: Top-5 dos melhores modelos candidatos CNN (As letras representam: a.*layers*; b.*dropout rate*; c.*activation*; d. *timesteps*; e. *batch size*; f. *filters*; g. *kernel size*; h. *pool size*; i. *epochs*; j. MAE; k. RMSE)

Ao analisar a Tabela 5.8 do melhor modelo candidato CNN, este obteve um valor de RMSE de aproximadamente de 4.52 e um valor de MAE aproximado de 4.18. O número de camadas utilizadas foi 3,

dropout rate de 0.5, a função de ativação TanH, *timesteps* de entrada igual a 3, um valor de filtros igual a 32, um tamanho de *kernel* igual a 3, e *pool size* de 3.

Quanto à homogeneidade dos valores observados, considerando os resultados globais do top-5, verifica-se que essa particularidade está visível na função de ativação TanH. Contudo, os restantes hiperparâmetros apresentam-se predominantemente iguais, como no caso do número de *layers* (3), *dropout rate* (0.5), *timesteps* (21), *batch size* (20), *filters* (32), *Kernel size* (3) e *pool size* (2). Atendendo aos valores observados na tabela, destaca-se uma redução na *performance* quando os valores de *batch size* diminuem e os do *kernel size* aumentam.

5.3 Análise Comparativa

Após a análise completa dos resultados obtidos a partir das experiências aplicadas aos diferentes modelos na previsão do IQA, torna-se possível realizar uma minuciosa comparação entre eles. Essa comparação permitirá uma melhor compreensão das diferenças e semelhanças entre os desempenhos dos modelos, contribuindo para uma avaliação mais abrangente das técnicas utilizadas. Desta forma, será apresentada uma análise comparativa entre os diferentes modelos com base na métrica de avaliação RMSE.

A análise comparativa visa examinar e comparar o desempenho dos modelos em relação à capacidade de reduzir o erro entre os valores previstos e os valores observados. Para garantir a robustez e a validade da análise, foi fundamental considerar fatores como a consistência dos conjuntos de dados utilizados, o seu correto pré-processamento, a configuração adequada dos hiperparâmetros e a utilização de métodos de validação adequados. Com base nessa análise comparativa, será possível obter perceções valiosas para a seleção do modelo mais adequado ou para avanços no conhecimento da área em questão. Posto isto, foi criada a Figura 5.1, cujo conteúdo consiste em várias barras representativas do RMSE obtido pelos melhores modelos candidatos do top-5, para cada cenário experimental considerado.

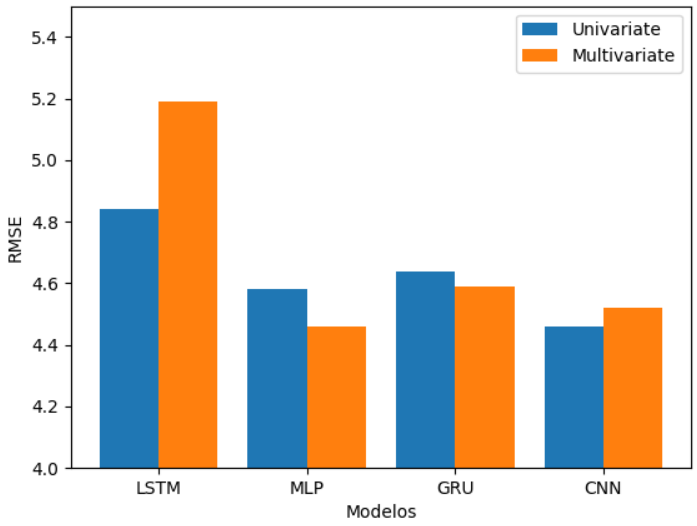


Figura 5.1: Gráfico com os resultados dos diferentes cenários e dos diferentes modelos aplicados.

Através da análise da Figura 5.1, pode-se analisar a comparação dos melhores modelos candidatos, nos dois cenários aplicados. Na figura, é possível visualizar a diferença de desempenho entre as duas abordagens para cada modelo, através da medida RMSE, representado pelo eixo do y.

Numa primeira instância, ao observar a Figura 5.1, há duas inferências evidentes. Primeiramente, é notável que o modelo LSTM, nos dois cenários experimentados, exibe os valores mais elevados de RMSE. Esta constatação sugere uma tendência de desempenho menos precisa em comparação com os restantes modelos avaliados. Além disso, é possível perceber que o modelo LSTM apresenta uma discrepância maior entre os cenários (diferença de 0.35), indicando uma sensibilidade diferenciada às condições experimentais. Ainda dentro desta linha de pensamento, o contrário acontece no modelo GRU que apresenta a menor discrepância entre cenários com o valor de 0.05. Os restantes modelos MLP e CNN apresentam 0.12 e 0.06 respetivamente.

Detalhando mais a análise comparativa dos melhores modelos candidatos, verifica-se que no contexto *Univariate*, o modelo CNN destaca-se ao alcançar o menor valor de RMSE, aproximadamente 4.46. Esse resultado sugere um desempenho superior em comparação com os restantes modelos avaliados nesse cenário especificamente. Por outro lado, quando se considera a abordagem *Multivariate*, foi o modelo MLP que se destacou, apresentando o menor RMSE de aproximadamente 4.46. Este resultado indica uma *performance* mais eficiente dentro desta cenário, revelando a adaptabilidade do modelo MLP aos dados multivariados. Essas observações fornecem *insights* preciosos para a escolha do melhor modelo candidato, dependendo da natureza do cenário do problema em questão. Essa distinção nos resultados destaca a importância de selecionar estratégias de modelação específicas para diferentes contextos, otimizando assim o desempenho preditivo nos cenários considerados.

Posto isto, seguem-se as Figuras 5.2 e 5.3 para ilustrar as previsões para os próximos 6 intervalos de tempo, realizadas pelos melhores modelos candidatos de cada cenário. Estas duas figuras evidenciam as previsões do IQA para os cenários *Univariate* e *Multivariate*, utilizando 28 e 14 *timesteps* como dados de entrada, respetivamente.

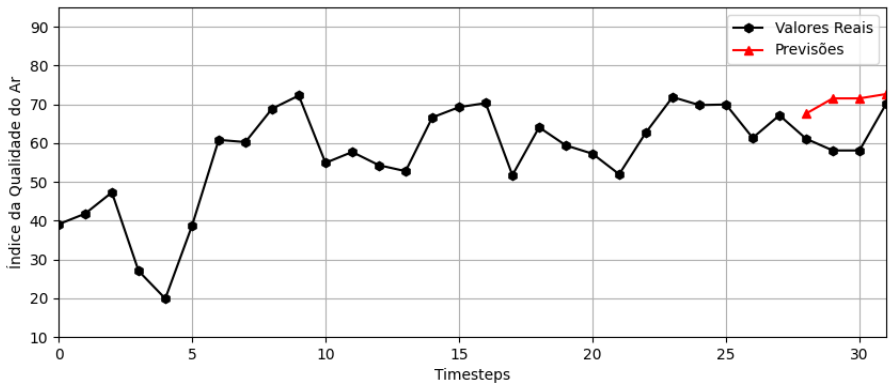


Figura 5.2: Previsões do IQA para os próximos 4 *timesteps*, realizadas pelo modelo CNN no cenário 1

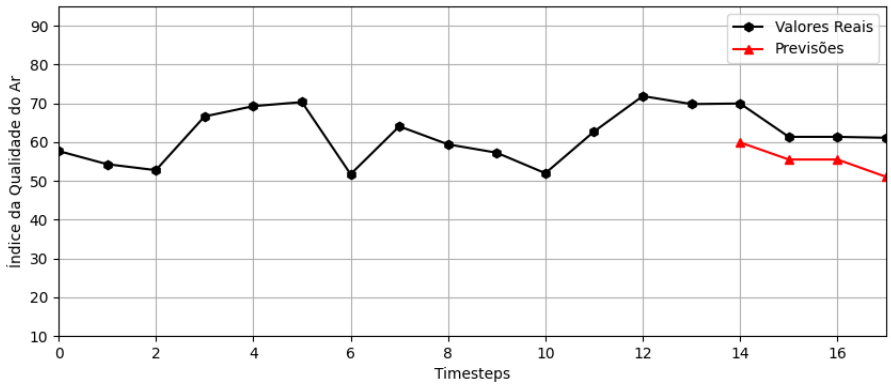


Figura 5.3: Previsões do IQA para os próximos 4 *timesteps*, realizadas pelo modelo MLP no cenário 2

Capítulo 6

Conclusão e Trabalho futuro

Neste capítulo são apresentados os resultados alcançados no desenvolvimento desta dissertação face aos seus objetivos primordiais, bem como, destacar o valor e a contribuição do estudo. Além disso, são abordadas as principais limitações identificadas na pesquisa e são mencionadas possibilidades de trabalhos futuros e investigações.

A integração de sistemas avançados de previsão da qualidade do ar não apenas promove um ambiente saudável, mas também desempenha um papel crucial na formulação de políticas ambientais eficazes. A capacidade de antecipar mudanças nos níveis de poluentes atmosféricos permite uma alocação eficiente de recursos e a implementação de medidas específicas para lidar com episódios críticos de poluição. Além disso, a previsão precisa da qualidade do ar pode ser um instrumento valioso para orientar iniciativas de adaptação e resiliência, especialmente face às mudanças climáticas, que podem afetar significativamente a distribuição e intensidade dos poluentes.

A consciencialização pública sobre a qualidade do ar também é fortalecida por meio desses sistemas de previsão avançados, incentivando uma participação mais ativa da comunidade em busca de práticas sustentáveis. A educação ambiental, apoiada por informações preditivas, tem o potencial de influenciar comportamentos individuais e coletivos, resultando em redução de emissões e maior responsabilidade ambiental. Assim, ao investir em tecnologias e métodos que aprimorem a previsão da qualidade do ar, está-se a contribuir não apenas para a saúde imediata da população, mas também para a construção de sociedades mais resilientes e conscientes do meio ambiente.

Posto isto, no estágio inicial desta dissertação, o foco foi a compreensão teórica dos temas em análise, abrangendo tópicos como a sustentabilidade ambiental, a poluição atmosférica e os impactos dos poluentes na saúde humana. Esta fase de aquisição de conhecimento foi fundamental para estabelecer uma abordagem apropriada em relação aos dados disponíveis, visando uma análise mais informada e contextualizada.

No contexto desse processo inicial, a revisão da literatura desempenhou um papel crucial. Ao longo

desta revisão, foram explorados diversos estudos significativos que proporcionaram *insights* valiosos sobre as abordagens e metodologias utilizadas no domínio da previsão do IQA. Estes estudos destacaram diferentes perspetivas, através do uso de algoritmos de ML para a previsão do IQA.

Entre os estudos explorados, o trabalho de Liu et al. [17] centrou-se na aplicação do algoritmo SVR na previsão do IQA em cidades chinesas, destacando a importância da correlação entre os atributos da qualidade do ar e a localização geográfica. Kumar et al. [44] investigaram o impacto do confinamento na Índia, enquanto Shah et al. [54] compararam métodos de cálculo do IQA em diferentes estações. Por outro lado, Janarthanan et al. [108] propõem uma abordagem combinada de SVR e DL em Chennai, Índia, para aproveitar a capacidade do SVR para lidar com características não lineares, e do DL para encontrar padrões temporais complexos, resultando, desta forma, num modelo mais abrangente e adaptável, potencialmente melhorando a precisão da previsão do IQA em Chennai. Abirami et al. [109] desenvolveu o modelo *DL-Air* para prever a qualidade do ar em Delhi, Índia, devido à complexidade dos dados na cidade. O *DL-Air* é especialmente adequado para lidar com a variedade de poluentes, as condições atmosféricas dinâmicas e as variantes sazonais em Delhi, proporcionando previsões mais precisas.

No entanto, ao analisar criticamente esses estudos, observa-se uma carência significativa na incorporação de técnicas fundamentais, como *cross-validation*, específico para um problema de séries temporais e a análise da correlação. Estas abordagens metodológicas são indispensáveis para assegurar a robustez e a confiabilidade dos modelos preditivos, especialmente considerando a complexidade intrínseca à previsão do IQA. A ausência dessas práticas em grande parte dos estudos analisados pode comprometer a generalização e a precisão dos modelos, visto que a *cross-validation* é essencial para evitar o *overfitting* e a análise de correlação proporciona *insights* valiosos sobre a relação entre as variáveis de entrada e o IQA. Além disso, ao compreender as relações entre as variáveis, é possível tomar decisões mais ponderadas sobre a escolha e interpretação do modelo.

Diante dessa lacuna identificada, a presente dissertação visa preencher esse vácuo metodológico, incorporando de forma sistemática tanto o *cross-validation*, específico para séries temporais (*TimeSerieSplit*) quanto a análise de correlação.

Além disso, após a identificação dessa lacuna metodológica, a investigação aprofundou-se em diferentes vertentes de algoritmos de ML. Essa exploração abrangente dos conceitos de IA permitiu uma compreensão mais holística das diversas técnicas disponíveis, identificando as mais adequadas para a abordagem da previsão do IQA. Ao integrar as práticas metodológicas essenciais e explorar as várias facetas dos algoritmos de ML, esta dissertação visa proporcionar uma contribuição substancial para o avanço do conhecimento e práticas na monitorização eficaz da qualidade do ar.

Na etapa subsequente, foram realizadas atividades de exploração e preparação dos dados com o intuito de alcançar previsões do IQA utilizando os conjuntos de dados disponíveis. A preparação cuidadosa dos dados desempenha um papel essencial em qualquer análise de dados. Dados bem preparados contribuem para análises mais precisas, modelos mais robustos e eficiência computacional aprimorada. Além disso, a preparação de dados é crucial para facilitar a interpretação dos resultados.

Ao garantir a qualidade e integridade dos dados, a preparação não apenas impacta positivamente

a eficácia dos modelos e análises, mas também fortalece a confiança nos *insights* derivados desses dados. Essa etapa pró-ativa não só otimiza o desempenho dos modelos, mas também assegura que as descobertas resultantes sejam fundamentadas e transparentes, contribuindo assim para a credibilidade e aplicabilidade das conclusões obtidas a partir dos dados preparados.

O processo de exploração e preparação dos dados envolveu a interpretação meticulosa dos mesmos, a compreensão da sua distribuição e a análise das relações entre as diversas variáveis envolvidas. Para cada um dos seis poluentes considerados, realizou-se uma análise detalhada dos *datasets* correspondentes, abrangendo métricas essenciais de estatística descritiva, como a média, mediana e desvio padrão.

Ao explorar cada um dos *datasets*, identificou-se uma presença significativa de observações nulas, indicando dias em que não foram registadas medições para determinadas substâncias. Para abordar essa questão, procedeu-se à substituição dos valores nulos por *NaN*, utilizando a função de *replace*. Paralelamente a isso, adotou-se uma abordagem pró-ativa para lidar com os *missing values* e realizou-se um tratamento cuidadoso dos dados climatológicos, filtrando as colunas que efetivamente foram consideradas relevantes pela análise da sua correlação.

Adicionalmente, a etapa de preparação dos dados incluiu a concatenação dos *datasets* dos poluentes com os do clima, garantindo uma visão unificada e abrangente. Além disso, foi essencial a criação de um novo atributo, o *IQA*, que contém o cálculo do índice da qualidade do ar para cada dia.

Posteriormente, a análise da presença de *outliers* e a inspeção detalhada da matriz de correlação foram analisadas para identificar padrões incomuns nos dados e avaliar as relações entre as variáveis, contribuindo assim, para uma compreensão mais aprofundada do conjunto de dados. A análise da correlação desempenhou um papel crucial na seleção das variáveis a serem incluídas para a previsão do IQA. Através dessa análise, foi identificada uma correlação positiva significativa entre os níveis de NO_2 e o IQA. A relação positiva sugere que o aumento nos níveis de NO_2 está associado a um aumento dos valores do IQA. Portanto, reconhecendo a importância dessa variável na explicação das variações na qualidade do ar, optou-se por incluir o NO_2 no conjunto de atributos utilizado nos modelos preditivos. No entanto, vale destacar que, apesar da relevância do NO_2, a análise revelou que nenhum dado climatológico apresentou uma correlação significativa com o IQA. Dessa forma, os dados climatológicos, como a temperatura, não foram incorporados ao estudo, uma vez que não contribuíram de maneira substancial para a previsão da qualidade do ar, de acordo com a análise de correlação. Essa decisão visa garantir que apenas variáveis relevantes foram consideradas, otimizando assim a precisão e eficácia dos modelos preditivos desenvolvidos.

Posto isto, foram criados vários cenários, implementando e ajustando os modelos selecionados para análise. Relativamente à previsão do IQA, foi possível obter várias conclusões. Considerando os resultados alcançados, torna-se evidente que a escolha do melhor modelo candidato está intrinsecamente ligado à abordagem adotada. Esta constatação ressalta a importância de uma abordagem perspicaz e adaptativa na escolha do modelo, garantindo que este esteja alinhado com as particularidades e requisitos específicos de cada cenário experimental. Ao longo desta dissertação, foram exploradas duas abordagens, a *Univariate* e a *Multivariate*, para analisar e prever a qualidade do ar. A abordagem *Univariate* concentra-se no uso

de uma única variável, enquanto a *Multivariate* incorpora mais que uma variável para uma análise mais abrangente. Estas duas abordagens fornecem uma perspetiva abrangente e adaptável, permitindo uma compreensão mais profunda da complexidade dos dados em diferentes contextos experimentais.

Assim, perante a abordagem *Univariate*, o modelo CNN demonstrou o melhor desempenho, com um menor valor de RMSE de aproximadamente 4.46. Este valor indica que o modelo CNN é capaz de fazer previsões mais precisas com base em dados univariados utilizados relacionados à qualidade do ar do que os restantes.

Por outro lado, na abordagem *Multivariate*, o modelo MLP apresentou o melhor desempenho, também com um valor de RMSE de 4.46. Assim, o modelo MLP é mais eficiente ao lidar com conjuntos de dados multivariados utilizados nesta dissertação, onde mais que uma variável é considerada para prever a qualidade do ar.

Na análise comparativa realizada na presente dissertação, o foco foi examinar e comparar o desempenho dos modelos, avaliando a capacidade de redução do erro entre os valores previstos e observados, ou seja, entender a eficácia de cada modelo em prever corretamente os resultados desejados, comparando-os com os dados reais disponíveis. Nessa análise crítica, foram considerados elementos como a consistência dos conjuntos de dados, o adequado pré-processamento, a configuração precisa dos hiperparâmetros e a utilização de métodos de validação apropriados. Essa abordagem assegurou a robustez e a validade dos resultados obtidos.

No entanto, é importante ressaltar que a escolha do melhor modelo também depende do contexto específico, dos requisitos do projeto e da disponibilidade de dados. É fundamental considerar outras métricas de avaliação, além do RMSE, e realizar experiências e validações adicionais para obter uma avaliação mais abrangente e precisa do desempenho dos modelos. Assim, a análise e desenvolvimento de modelos de previsão desempenham um papel crucial, especialmente quando se trata de antecipar a qualidade do ar. Essa abordagem torna-se indispensável para compreender padrões, tendências e variações nos níveis de poluentes, permitindo uma resposta mais eficaz às mudanças na qualidade do ar.

Considerando as descobertas e limitações deste estudo, é relevante ressaltar que subsistem lacunas e oportunidades que podem ser aprofundadas em investigações futuras, ampliando assim o campo de conhecimento. Uma estratégia com grande potencial para aprimorar as previsões e análises em séries temporais consiste na incorporação de modelos *Transformers*. Esses modelos, que inicialmente se destacaram em tarefas sequenciais, como processamento de linguagem natural, têm demonstrado eficácia notável quando aplicados a séries temporais. A vantagem crucial dos modelos *Transformers* consiste na sua capacidade de capturar dependências de longo alcance em dados temporais.

Ao implementar os modelos *Transformers*, é possível potencializar significativamente a capacidade preditiva em séries temporais complexas. Essa potencialização é atribuída à habilidade intrínseca desses modelos de entender e aprender relações temporais distantes, superando as limitações de abordagens mais tradicionais. Assim, a utilização de modelos *Transformers* representa uma estratégia promissora para obter previsões mais precisas e *insights* mais profundos em contextos temporais dinâmicos.

Além disso, uma direção promissora para pesquisas subsequentes consiste na exploração de modelos híbridos. A utilização desses modelos oferece a oportunidade de aprimorar consideravelmente a precisão das previsões e análises em séries temporais complexas. A combinação de diversas técnicas poderá proporcionar uma abordagem mais eficiente para lidar com padrões temporais variados, podendo contribuir para uma melhor capacidade de captura de informações contextuais e identificação de relações não lineares entre as variáveis. A integração de diferentes métodos em modelos híbridos representa uma estratégia que pode elevar significativamente a qualidade das previsões, abrindo caminho para avanços mais substanciais na compreensão e predição de séries temporais complexas.

Outra possibilidade interessante de trabalho futuro é o desenvolvimento de uma aplicação móvel para disponibilizar os resultados e *insights* obtidos neste estudo aos utilizadores finais. A implementação de uma aplicação móvel abriria caminho para um acesso simplificado às informações, permitindo que profissionais e interessados na área acompanhassem as previsões e análises em tempo real. Essa abordagem agilizaria significativamente o processo de tomada de decisão, promovendo eficiência. Além disso, a aplicação móvel poderia oferecer recursos interativos, gráficos dinâmicos e notificações personalizadas, tornando a experiência do utilizador mais envolvente e prática.

As direções futuras abordadas podem oferecer um potencial significativo para expandir e melhorar o trabalho realizado, enriquecendo a pesquisa com outras abordagens e a aplicação prática em dispositivos móveis. Ao implementar melhorias tangíveis nestas frentes, antecipa-se um progresso neste campo. A intenção é fornecer soluções mais eficientes, sintonizadas com os desafios enfrentados pelos profissionais nesta área particular. Essa evolução não apenas impulsionará o avanço do conhecimento, mas também se traduzirá em alternativas práticas e relevantes, atendendo às necessidades da comunidade interessada.

Bibliografia

[1] H. Ritchie e M. Roser. “Air Pollution”. Em: *Our World in Data* (2017). https://ourworldindata.org/air-pollution.

[2] G. Clark. “A farewell to Alms: A brief economic history of the world”. Em: *Journal of Economic Literature* 45.4 (2007), pp. 1065–1119.

[3] E. O.-O. Max Roser e H. Ritchie. “Life Expectancy”. Em: *Our World in Data* (2013). url: https://ourworldindata.org/life-expectancy.

[4] M. Kremer. “Population growth and technological change: One million B.C. to 1990”. Em: *The Quarterly Journal of Economics* 108.3 (1993), pp. 681–716.

[5] D. M. Cutler e A. Lleras-Muney. “Education and health: Evaluating theories and evidence”. Em: *National Bureau of Economic Research* (2006).

[6] R. Almond, M. Grooten e T. Peterson, eds. *Living Planet Report 2020 - Bending the curve of biodiversity loss*. Gland, Switzerland: World Wildlife Fund, 2020. url: http://pure.iiasa.ac.at/id/eprint/16870/.

[7] T. Piketty. “Capital in the twenty-first century”. Em: *Belknap Press* (2014).

[8] J. Colmer e J. Voorheis. *The grandkids aren’t alright: the intergenerational effects of prenatal pollution exposure*. CEP Discussion Papers dp1733. Centre for Economic Performance, LSE, 2020. url: https://ideas.repec.org/p/cep/cepdps/dp1733.html.

[9] T. C. Russ, M. P. C. Cherrie, C. Dibben, S. Tomlinson, S. Reis, U. Dragosits, M. Vieno, R. Beck, E. Carnell, N. K. Shortt, G. Muniz-Terrera, P. Redmond, A. M. Taylor, T. Clemens, M. van Tongeren, R. M. Agius, J. M. Starr, I. J. Deary e J. R. Pearce. “Life course air pollution exposure and cognitive decline: modelled historical air pollution data and the Lothian Birth Cohort 1936”. Em: *medRxiv* (2020). doi: 10.1101/2020.10.16.20163691. eprint: https://www.medrxiv.org/content/early/2020/11/17/2020.10.16.20163691.full.pdf. url: https://www.medrxiv.org/content/early/2020/11/17/2020.10.16.20163691.

[10] T. Hastie, R. Tibshirani e J. Friedman. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”. Em: *Springer* (2009).

[11] M. Hino, E. Benami e N. Brooks. “Machine learning for environmental monitoring”. Em: *Nature Sustainability* 1 (2018). url: https://doi.org/10.1038/s41893-018-0142-9.

[12] J. Qiu, Q. Wu, G. Ding, Y. Xu e S. Feng. “A survey of machine learning for big data processing”. Em: *EURASIP Journal on Advances in Signal Processing* (2016). doi: https://doi.org/10.1186/s13634-016-0355-x.

[13] APA. “WHO introduces ambitious new air quality guidelines”. Em: (2013). url: https://apambiente.pt/.

[14] Y. Liu, Q. Zhu, D. Yao e W. Xu. “Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rule”. Em: *Atmosphere* 6.7 (2015), pp. 891–907. issn: 2073-4433. doi: 10.3390/atmos6070891. url: https://www.mdpi.com/2073-4433/6/7/891.

[15] I. Goodfellow, Y. Bengio e A. Courville. “Deep Learning”. Em: *MIT Press* (2016).

[16] Z. Zhang e Y. Wang. “Air pollution and public health: Emerging hazards and improved understanding of risk”. Em: *Environmental Science and Pollution Research* 26.15 (2019), pp. 15100–2. doi: 10.1007/s11356-019-05399-w.

[17] B. Liu, A. Binaykia, P. C. Chang, M. K. Tiwari e C.-C. Tsao. “Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang”. Em: *PLoS ONE* 12 (2017).

[18] X. Yang e X. Liu. “A review of traffic-related air pollution: Exposure, health effects, and control”. Em: *Traffic Injury Prevention* 18.7 (2017), pp. 694–700. doi: 10.1080/15389588.2017.1316461.

[19] R. Bekkerman, M. Bilenko e J. Langford. “Scaling up Machine Learning: Parallel and Distributed Approaches”. Em: (2011). url: https://doi.org/10.1145/2107736.2107740.

[20] G. Mariscal, Ö. Marbán e C. Fernández. “A survey of data mining and knowledge discovery process models and methodologies”. Em: *The knowledge engineering review* 25.2 (2010), pp. 137–166. issn: 0269-8889. doi: 10.1017/s0269888910000032. url: http://dx.doi.org/10.1017/s0269888910000032.

[21] R. Wirth. “CRISP-DM: Towards a standard process model for data mining”. Em: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (2000), pp. 29–39.

[22] D. H. Meadows, D. L. Meadows, J. Randers e W. W. Behrens. “The limits to growth: A report for the Club of Rome’s project on the predicament of mankind”. Em: *Universe Books* (1972).

[23] World Commission on Environment and Development (WCED). “Our common future”. Em: *Oxford University Press* (1987).

[24] J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. Lambin, ... e B. Nykvist. “A safe operating space for humanity”. Em: *Nature* 461.7263 (2009), pp. 472–475.

[25] S. M. Lele e R. B. Norgaard. “Sustainability and the scientist’s burden”. Em: *Conservation Biology* 10.2 (1996), pp. 354–365.

[26] W. N. Adger, K. Brown e E. L. Tompkins. “The political economy of cross-scale networks in resource co-management”. Em: *Ecology and Society* 10.2 (2005), p. 9.

[27] T. Kuhlman e J. Farrington. “What is Sustainability?” Em: *Sustainability* 2.11 (2010), pp. 3436–3448. issn: 2071-1050. doi: 10.3390/su2113436. url: https://www.mdpi.com/2071-1050/2/11/3436.

[28] N. Arora. “Environmental Sustainability—necessary for survival”. Em: *Environmental Sustainability* 1 (2018). doi: 10.1007/s42398-018-0013-3.

[29] B. Ki-moon. “Sustainability—engaging future generations now”. Em: *The Lancet* 387.10036 (2016), pp. 2356–2358. doi: 10.1016/S0140-6736(16)30271-9.

[30] H. Backer. “Bosselmann, Klaus: The Principle of Sustainability: Transforming Law and Governance (Book review)”. Em: *Helsinki Law Review* (2010), p. 117.

[31] H. Bové e N. A. H. Janssen. “Public health impacts of urban air pollution on children and adolescents: A multicountry study”. Em: *Environment International* 109 (2017), pp. 61–70. doi: 10.1016/j.envint.2017.07.020.

[32] S. Eriksen, E. L. F. Schipper, M. Scoville-Simonds, K. Vincent, H. N. Adam, N. Brooks, B. Harding, D. Khatri, L. Lenaerts, D. Liverman, M. Mills-Novoa, M. Mosberg, S. Movik, B. Muok, A. Nightingale, H. Ojha, L. Sygna, M. Taylor, C. Vogel e J. J. West. “Adaptation interventions and their effect on vulnerability in developing countries: Help, hindrance or irrelevance?” Em: *World Development* 141 (2021), p. 105383. issn: 0305-750X. doi: https://doi.org/10.1016/j.worlddev.2020.105383. url: https://www.sciencedirect.com/science/article/pii/S0305750X20305118.

[33] A. Jorgenson e R. Dunlap. “Environmental Problems”. Em: *The Wiley-Blackwell Encyclopedia of Globalization* (2012). doi: 10.1002/9780470670590.wbeog174.

[34] J. Cleland. “World Population Growth; Past, Present and Future”. Em: *Environmental and Resource Economics* 55 (2013). doi: 10.1007/s10640-013-9675-6.

[35] M. Kummu, J. H. Guillaume, H. de Moel, S. Eisner, M. Flörke, M. Porkka, S. Siebert, T. I. Veldkamp e P. Ward. “The world’s road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability”. Em: *Scientific Reports* 6.1 (2016), pp. 1–16. doi: 10.1038/srep38495.

[36] R. Carr, U. Blumenthal e D. Mara. “Guidelines for the Safe Use of Wastewater in Agriculture: Revisiting WHO Guidelines”. Em: *Water Science and Technology: A Journal of the International Association on Water Pollution Research* 50.Feb. (2004), pp. 31–8. doi: 10.2166/wst.2004.0081.

[37] A. Mukherjee, N. H. Kamarulzaman, G. Vijayan e S. Vaiappuri. “Sustainability: A Comprehensive Literature”. Em: *Jan. 2016* (), pp. 248–268. doi: 10.4018/978-1-4666-9639-6.ch015.

[38] J. Barnes, J. Bender, L. TM e B. AM. “Natural and man-made selection for air pollution resistance”. Em: *Journal of Experimental Botany* 50 (1999). doi: 10.1093/jexbot/50.338.1423.

[39] M. Sierra-Vargas e L. Teran. “Air pollution: Impact and prevention”. Em: *Respirology (Carlton, Vic.)* 17 (2012), pp. 1031–8. doi: 10.1111/j.1440-1843.2012.02213.x.

[40] C. Rao e B. Yan. “Study on the interactive influence between economic growth and environmental pollution”. Em: *Environmental Science and Pollution Research* 27.31 (2020), pp. 39442–39465. doi: 10.1007/s11356-020-10017-6.

[41] A. Inyinbor, B. Adebesin, A. Oluyori, T. Adelani-Akande, A. O. Dada e O. A. *Water Pollution: Effects, Prevention, and Climatic Impact*. 2018. isbn: 978-953-51-3893-8. doi: 10.5772/intechopen.72018.

[42] R. Chen e C. A. Pope. “Air pollution and cardiovascular disease: A window of opportunity”. Em: *Journal of the American Heart Association* 6.12 (2017), e006582. doi: 10.1161/JAHA.117.006582.

[43] Q. Zhang e X. Liang. “Attribution of anthropogenic influence on atmospheric patterns conducive to recent most severe haze over eastern China”. Em: *Geophysical Research Letters* 43.8 (2016), pp. 4287–96. doi: 10.1002/2016GL068924.

[44] G. Kumar, S. Kumar e Suman. “Air quality index – A comparative study for assessing the status of air quality before and after lockdown for Meerut”. Em: *Materials Today: Proceedings* 49 (2022). National Conference on Functional Materials: Emerging Technologies and Applications in Materials Science, pp. 3497–3500. issn: 2214-7853. doi: https://doi.org/10.1016/j.matpr.2021.05.575. url: https://www.sciencedirect.com/science/article/pii/S2214785321042024.

[45] A. J. Cohen, M. Brauer, R. Burnett, H. R. Anderson, J. Frostad, K. Estep, K. Balakrishnan, B. Brunekreef, L. Dandona, R. Dandona et al. “Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015”. Em: *The Lancet* 389.10082 (2017), pp. 1907–1918. doi: https://doi.org/10.1016/S0140-6736(17)30505-6.

[46] Y. O. Khaniabadi, R. Nabizadeh, M. S. Hassanvand, S. Yunesian, M. H. Zare, S. H. Abbasi e K. Naddafi. “Health impact assessment of air pollution in megacity of Tehran, Iran”. Em: *Journal of Environmental Health Science and Engineering* 15.1 (2017), p. 27. doi: 10.1186/s40201-017-0282-4.

[47] T. Burki. “WHO introduces ambitious new air quality guidelines”. Em: *The Lancet* (2021). url: https://doi.org/10.1016/S0140-6736(21)02126-7.

[48] S. A. Sarkodie. “Environmental performance, biocapacity, carbon & ecological footprint of nations: Drivers, trends and mitigation options”. Em: *Science of The Total Environment* 751 (2021), p. 141912. issn: 0048-9697. doi: https://doi.org/10.1016/j.scitotenv.2020.141912. url: https://www.sciencedirect.com/science/article/pii/S0048969720354413.

[49] R. T. Burnett e C. A. Pope. “Epidemiological evidence of cardiovascular effects of particulate air pollution”. Em: *Environmental Health Perspectives* 108.4 (2004), pp. 361–76. doi: 10.1289/ehp.0201081.

[50] C. A. P. III e D. W. Dockery. “Health Effects of Fine Particulate Air Pollution: Lines that Connect”. Em: *Journal of the Air & Waste Management Association* 56.6 (2006), pp. 709–742. doi: 10.1080/10473289.2006.10464485. url: https://doi.org/10.1080/10473289.2006.10464485.

[51] L. C. McCabe. “ATMOSPHERIC POLLUTION”. Em: *Industrial and Engineering Chemistry (U.S.) Formerly J. Ind. Eng. Chem. Superseded by Chem. Technol.* 42 (1950). url: https://www.osti.gov/biblio/4391289.

[52] X. Wu e J. Zheng. “Impacts of air pollution on children’s and adolescents’ respiratory health in China: Recent advances”. Em: *Current Opinion in Pediatrics* 31.2 (2019), pp. 256–62. doi: 10.1097/MOP.0000000000000745.

[53] F. Popescu e I. Ionel. “Anthropogenic Air Pollution Sources”. Em: *Air Quality*. Ed. por A. Kumar. Rijeka: IntechOpen, 2010. Cap. 1. doi: 10.5772/9751. url: https://doi.org/10.5772/9751.

[54] D. P. Shah e D. P. Patel. “A comparison between national air quality index, india and composite air quality index for Ahmedabad, India”. Em: *Environmental Challenges* 5 (2021), p. 100356. issn: 2667-0100. doi: https://doi.org/10.1016/j.envc.2021.100356. url: https://www.sciencedirect.com/science/article/pii/S2667010021003309.

[55] P. Romer e S. H. Schneider. “Understanding and managing the impacts of climate change”. Em: *Science* 262.5134 (1993), pp. 1065–71. doi: 10.1126/science.262.5134.1065.

[56] M. L. Block, A. Elder, R. L. Auten, S. D. Bilbo, H. Chen, J.-C. Chen, D. A. Cory-Slechta, D. Costa, D. Diaz-Sanchez, D. C. Dorman, D. R. Gold, K. Gray, H. A. Jeng, J. D. Kaufman, M. T. Kleinman, A. Kirshner, C. Lawler, D. S. Miller, S. S. Nadadur, B. Ritz, E. O. Semmens, L. H. Tonelli, B. Veronesi, R. O. Wright e R. J. Wright. “The outdoor air pollution and brain health workshop”. Em: *NeuroToxicology* 33.5 (2012), pp. 972–984. issn: 0161-813X. doi: https://doi.org/10.1016/j.neuro.2012.08.014. url: https://www.sciencedirect.com/science/article/pii/S0161813X12002100.

[57] R. K. Pachauri e L. A. Mayer. “Integrating environment and development: Striving for sustainability in policy making”. Em: *Environmental Management* 20.3 (1996), pp. 337–50. doi: 10.1007/s002679900038.

[58] U. Ogbonnaya e A. A. Abia. “Air pollution and respiratory health: A review”. Em: *Biochemistry and Molecular Biology Reports* 8.3 (2019), pp. 268–75. doi: 10.1007/s41048-019-00102-7.

[59] S. Gokhale e M. Khare. “Air pollution and cardiovascular diseases: A review”. Em: *Urban Climate* 29 (2019), p. 100548. doi: 10.1016/j.uclim.2019.100548.

[60] REA. “Portal do estado do ambiente - Portugal”. Em: (2020-2021). url: https://rea.apambiente.pt/.

[61] L. Gimeno, E. Marín, T. del Teso e S. Bourhim. “How effective has been the reduction of SO2 emissions on the effect of acid rain on ecosystems?” Em: *Science of The Total Environment* 275.1 (2001), pp. 63–70. issn: 0048-9697. doi: https://doi.org/10.1016/S0048-9697(00)00854-8. url: https://www.sciencedirect.com/science/article/pii/S0048969700008548.

[62] N. P. Cheremisinoff. “Chapter 1 - Introduction to Air Quality”. Em: *Handbook of Air Pollution Prevention and Control*. Ed. por N. P. Cheremisinoff. Woburn: Butterworth-Heinemann, 2002, pp. 1–52. isbn: 978-0-7506-7499-7. doi: https://doi.org/10.1016/B978-075067499-7/50002-X.

[63] R. S. Sokhi. “Urban air quality”. Em: *Atmospheric Environment* 42.17 (2008). Fifth International Conference on Urban Air Quality, pp. 3909–3910. issn: 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2007.09.035. url: https://www.sciencedirect.com/science/article/pii/S1352231007008394.

[64] K. K. Lee, N. Spath, M. R. Miller, N. L. Mills e A. S. Shah. “Short-term exposure to carbon monoxide and myocardial infarction: A systematic review and meta-analysis”. Em: *Environment International* 143 (2020), p. 105901. issn: 0160-4120. doi: https://doi.org/10.1016/j.envint.2020.105901. url: https://www.sciencedirect.com/science/article/pii/S0160412020318560.

[65] Thom, S. R. Y. Anne, Ischiropoulos e Harry. “Vascular Endothelial Cells Generate Peroxynitrite in Response to Carbon Monoxide Exposure”. Em: *Chemical Research in Toxicology* 10 (1997). url: https://doi.org/10.1021/tx970041h.

[66] M. Janssen, M. Hartog, R. Matheus, A. Y. Ding e G. Kuk. “Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers’ Experience on AI-supported Decision-Making in Government”. Em: *Social Science Computer Review* (). doi: 10.1177/0894439320980118. url: https://doi.org/10.1177/0894439320980118.

[67] Y. Bengio. “Learning deep architectures for AI”. Em: *Foundations and Trends® in Machine Learning* 2.1 (2009), pp. 1–127. doi: 10.1561/2200000006.

[68] M. Attaran e P. Deb. “Machine Learning: The New ’Big Thing’ for Competitive Advantage”. Em: 5 (2018), pp. 277–305. doi: 10.1504/IJKEDM.2018.10015621.

[69] T. Hastie, R. Tibshirani e J. Friedman. "The elements of statistical learning: Data mining, inference, and prediction". Em: *The American Statistician* 51.1 (2017), pp. 143–5. doi: 10.1080/00031305.1997.10473540.

[70] J. Brownlee. "Discover How They Work and Implement Them From Scratch". Em: 1 (2016). url: https://datageneralist.files.wordpress.com/2018/03/master_machine_learning_algo_from_scratch.pdf.

[71] G. E. Hinton e R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". Em: *Science* 313.5786 (2006), pp. 504–507. doi: 10.1126/science.1127647.

[72] C. Silva, P. Novais e B. Dias. "A machine learning approach to environmental sustainability". Em: *Cognitive Systems Research* (2020). url: http://hdl.handle.net/1822/71012.

[73] J. Schmidhuber. "Deep learning in neural networks: An overview". Em: *Neural Networks* 61 (2015), pp. 85–117.

[74] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink e J. Schmidhuber. "LSTM: A search space odyssey". Em: *IEEE Transactions on Neural Networks and Learning Systems* 28.10 (2016), pp. 2222–2232.

[75] D. Gigante, P. Oliveira, B. Fernandes, F. Lopes e P. Novais. "Unsupervised Learning Approach for pH Anomaly Detection in Wastewater Treatment Plants". Em: *International Conference on Hybrid Artificial Intelligence Systems* (2021), pp. 588–599. doi: 10.1007/978-3-030-86271-8_49.

[76] D. Rumelhart, G. Hinton e R. Williams. "Learning representations by back-propagating errors". Em: *Nature* 323.6088 (1986), pp. 533–536.

[77] D. Bzdok, G. Varoquaux, O. Grisel, M. Eickenberg, C. Poupon e B. Thirion. "Data mining and machine learning for neuroimaging: an overview". Em: *NeuroImage* 61.2 (2012), pp. 623–638.

[78] V. Vapnik. "The nature of statistical learning theory". Em: *Springer Science Business Media* (1995).

[79] I. H. Sarker. "Machine learning: Algorithms, real-world applications and research directions". Em: *SN Computer Science* 2.3 (2021), pp. 1–21. doi: 10.1007/s42979-021-00592-x.

[80] A. K. Kumar, M. Ritam, L. Han, S. Guo e R. Chandra. "Deep learning for predicting respiratory rate from biosignals". Em: *Computers in Biology and Medicine* 144 (2022), p. 105338. issn: 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2022.105338. url: https://www.sciencedirect.com/science/article/pii/S0010482522001305.

[81] A. Halevy, P. Norvig e F. Pereira. "Controlling machine learning: The rise of human-algorithmic ecosystems". Em: *AI Magazine* 37.3 (2016), pp. 49–59.

[82] W. Cukierski, D. J. Foran e J. R. Smith. "Statistical prediction of protein functional regions". Em: *PloS one* 6.3 (2011), e18508.

[83] S. L. Salzberg. “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993”. Em: *Machine Learning* 16 (1993). url: https://doi.org/10.1007/BF00993309.

[84] J. R. Quinlan. “Induction of decision trees”. Em: *Machine Learning* 1 (1986). url: https://doi.org/10.1007/BF00116251.

[85] J. Ramos. “Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional”. Tese de doutoramento. 2016.

[86] S. N. Shorabeh, N. N. Samany, F. Minaei, H. K. Firozjaei, M. Homaee e A. D. Boloorani. “A decision model based on decision tree and particle swarm optimization algorithms to identify optimal locations for solar power plants construction in Iran”. Em: *Renewable Energy* 187 (2022), pp. 56–67. issn: 0960-1481. doi: https://doi.org/10.1016/j.renene.2022.01.011. url: https://www.sciencedirect.com/science/article/pii/S0960148122000118.

[87] S. Agarwal. “Data Mining: Data Mining Concepts and Techniques”. Em: (2013), pp. 203–207. doi: 10.1109/ICMIRA.2013.45.

[88] S. Qi, K. Jin, B. Li e Y. Qian. “The exploration of internet finance by using neural network”. Em: *Journal of Computational and Applied Mathematics* 369 (2020), p. 112630. issn: 0377-0427. doi: https://doi.org/10.1016/j.cam.2019.112630. url: https://www.sciencedirect.com/science/article/pii/S0377042719306351.

[89] F. Bre, J. M. Gimenez e V. D. Fachinotti. “Prediction of wind pressure coefficients on building surfaces using artificial neural networks”. Em: *Energy and Buildings* 158 (2018), pp. 1429–1441. issn: 0378-7788. doi: https://doi.org/10.1016/j.enbuild.2017.11.045. url: https://www.sciencedirect.com/science/article/pii/S0378778817325501.

[90] Y. LeCun, L. Bottou, Y. Bengio e P. Haffner. “Gradient-based learning applied to document recognition”. Em: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. doi: 10.1109/5.726791.

[91] S. Gao, Y. Huang, S. Zhang, J. Han, G. Wang, M. Zhang e Q. Lin. “Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation”. Em: *Journal of Hydrology* 589 (2020), p. 125188. issn: 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2020.125188. url: https://www.sciencedirect.com/science/article/pii/S002216942030648X.

[92] B. Fernandes, F. Silva, H. Alaiz-Moreton, P. Novais, J. Neves e C. Analide. “Long Short-Term Memory Networks for Traffic Flow Forecasting: Exploring Input Variables, Time Frames and Multi-Step Approaches”. Em: *Informatica* 31.4 (2020), pp. 723–749. doi: 10.15388/Informatica.2020.251.

[93] P. Oliveira, B. Fernandes, C. Analide e P. Novais. “Forecasting Energy Consumption of Wastewater Treatment Plants with a Transfer Learning Approach for Sustainable Cities”. Em: *Electronics* 10.10 (2021), p. 1149. doi: 10.3390/electronics10101149.

[94] B. Fernandes, F. Silva, H. Alaiz-Moreton, P. Novais, C. Analide e J. Neves. “Traffic Flow Forecasting on Data-Scarce Environments Using ARIMA and LSTM Networks”. Em: *World Conference on Information Systems and Technologies*. Springer International Publishing, 2019, pp. 273–282. doi: 10.1007/978-3-030-16187-3_26.

[95] F. Wang, J. Xia, L. Zou, C. Zhan e W. Liang. “Estimation of time-varying parameter in Budyko framework using long short-term memory network over the Loess Plateau, China”. Em: *Journal of Hydrology* 607 (2022), p. 127571. issn: 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2022.127571. url: https://www.sciencedirect.com/science/article/pii/S0022169422001469.

[96] F. A. Gers, J. Schmidhuber e F. Cummins. “Learning to forget: Continual prediction with LSTM”. Em: *Neural Computation* 12.10 (2000), pp. 2451–2471. doi: 10.1162/089976600300015015.

[97] S. Hochreiter e J. Schmidhuber. “Long short-term memory”. Em: *Neural computation* 9.8 (1997), pp. 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

[98] A. Graves e J. Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. Em: *Neural Networks* 18.5-6 (2005), pp. 602–610. doi: 10.1016/j.neunet.2005.06.042.

[99] K. C. et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. Em: *arXiv preprint arXiv:1406.1078* (2014).

[100] R. Fu, Z. Zhang e L. Li. “Using LSTM and GRU neural network methods for traffic flow prediction”. Em: (2016), pp. 324–328. doi: 10.1109/YAC.2016.7804912.

[101] J. C. et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. Em: *arXiv preprint arXiv:1412.3555* (2014).

[102] J. Teuwen e N. Moriakov. “Chapter 20 - Convolutional neural networks”. Em: The Elsevier and MICCAI Society Book Series (2020). Ed. por S. K. Zhou, D. Rueckert e G. Fichtinger, pp. 481–501. doi: https://doi.org/10.1016/B978-0-12-816176-0.00025-9. url: https://www.sciencedirect.com/science/article/pii/B9780128161760000259.

[103] A. Krizhevsky, I. Sutskever e G. E. Hinton. “ImageNet classification with deep convolutional neural networks”. Em: *Communications of the ACM* 60.6 (2017), pp. 84–90. doi: 10.1145/3065386.

[104] K. H. et al. “Deep residual learning for image recognition”. Em: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[105] K. Simonyan e A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. Em: *arXiv preprint arXiv:1409.1556* (2014).

[106] V. K. Shaojie Bai J. Zico Kolter. “Convolutional Sequence Modeling Revisited”. Em: (2018). url: https://openreview.net/forum?id=rk8wKk-R-.

[107] D. S. K. Karunasingha. “Root mean square error or mean absolute error? Use their ratio as well”. Em: *Information Sciences* 585 (2022), pp. 609–629. issn: 0020-0255. doi: https://doi.org/10.1016/j.ins.2021.11.036. url: https://www.sciencedirect.com/science/article/pii/S0020025521011567.

[108] R. Janarthanan, P. Partheeban, K. Somasundaram e P. Navin Elamparithi. “A deep learning approach for prediction of air quality index in a metropolitan city”. Em: *Sustainable Cities and Society* 67 (2021), p. 102720. issn: 2210-6707. doi: https://doi.org/10.1016/j.scs.2021.102720. url: https://www.sciencedirect.com/science/article/pii/S2210670721000159.

[109] S Abirami e P Chitra. “Regional air quality forecasting using spatiotemporal deep learning”. Em: *Journal of Cleaner Production* 283 (2021), p. 125341. issn: 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2020.125341. url: https://www.sciencedirect.com/science/article/pii/S0959652620353865.

[110] T.-H. Kim e H. White. “On more robust estimation of skewness and kurtosis”. Em: *Finance Research Letters* 1.1 (2004), pp. 56–73. issn: 1544-6123. doi: https://doi.org/10.1016/S1544-6123(03)00003-5. url: https://www.sciencedirect.com/science/article/pii/S1544612303000035.

[111] J. Bai e S. Ng. “Tests for Skewness, Kurtosis, and Normality for Time Series Data”. Em: *Journal of Business & Economic Statistics* 23.1 (2005), pp. 49–60. doi: 10.1198/073500104000000271.

[112] K. Xu, K. Cui, L.-H. Young, Y.-F. Wang, Y.-K. Hsieh, S. Wan e J. Zhang. “Air Quality Index, Indicatory Air Pollutants and Impact of COVID-19 Event on the Air Quality near Central China”. Em: *Aerosol and Air Quality Research* 20.6 (2020), pp. 1204–1221. doi: 10.4209/aaqr.2020.04.0139. url: https://doi.org/10.4209%2Faaqr.2020.04.0139.

[113] K. Andrea, G. Shevlyakov e P. Smirnov. “Detection of outliers with boxplots”. Em: (2013), pp. 141–144.

[114] D Williamson, R. Parker e J. Kendrick. “The box plot: A simple visual method to interpret data”. Em: *Annals of internal medicine* 110 (1989), pp. 916–21. doi: 10.1059/0003-4819-110-11-916.

[115] P. Sedgwick. “Spearman’s rank correlation coefficient”. Em: *BMJ: British Medical Journal* 349 (2014), g7327. doi: 10.1136/bmj.g7327.