Airmon: Sistema IoT de Monitorização e Classificação da Ocupação em Salas de Aula

João Silvestre

Faculdade de Ciência e Tecnologia

Universidade Fernando Pessoa

Dissertação submetida para obtenção do grau de

Mestre em Engenharia Informática - Ramo Computação Móvel

2023

Resumo

A gestão dos parâmetros ambientais em espaços fechados é de extrema importância para manter a salubridade e conforto dos recintos, principalmente em espaços públicos muito frequentados. Em particular, em ambientes escolares, a má gestão destes parâmetros, pode impactar negativamente o bem-estar dos alunos e docentes, manifestando-se em dificuldades de concentração, fadiga e/ou dores de cabeça, afetando, consequentemente, o processo de ensino-aprendizagem. A utilização de inteligência artificial para classificar a ocupação desses espaços, oferece a oportunidade de otimizar a sua gestão e planeamento, tornando o processo educacional mais eficiente e adaptado às necessidades dos intervenientes. Este trabalho tem como objetivo contribuir para uma possível solução para a lacuna existente relativamente à gestão de parâmetros ambientais, através da conceção e implementação de um sistema de monitorização de baixo custo e alta escalabilidade, visando a recolha e agregação precisa de dados ambientais, oriundos de diversas salas de aula, distribuídas por diferentes estabelecimentos de ensino e com a capacidade de integrar todos os que se pretendam associar ao projeto. Foram desenvolvidas caixas equipadas com um conjunto diversificado de sensores de baixo custo e consumo energético. O sistema disponibiliza uma interface intuitiva para o acesso e monitorização em tempo real de variáveis como os níveis de CO_2, humidade, temperatura e partículas, referentes a cada sala de aula das escolas monitorizadas. Adicionalmente, os dados ambientais foram complementados com a indicação da ocupação das salas, através da colaboração da comunidade escolar, fornecendo assim uma solução que respeita a privacidade das pessoas envolvidas, não requerendo a utilização de métodos de recolha de dados invasivos, como câmaras. A estes dados foram aplicadas técnicas de inteligência artificial com o intuito de classificar a ocupação das salas de aula, obtendo uma acurácia de, no mínimo, 83% na classificação da ocupação com um modelo geral para todas as salas de aula e de, pelo menos, 85% quando treinadas para uma sala específica.

Abstract

The management of environmental parameters in enclosed spaces is extremely important in order to maintain the health and comfort of the premises, especially in highly frequented public spaces. In particular, in school environments, poor management of these parameters can have a negative impact on the well-being of students and teachers, manifesting itself in concentration difficulties, fatigue and/or headaches, consequently affecting the teaching-learning process. The use of artificial intelligence to classify the occupancy of these spaces offers the opportunity to optimise their management and planning, making the educational process more efficient and adapted to the needs of those involved. The aim of this work is to contribute to a possible solution to the existing gap in the management of environmental parameters by designing and implementing a low-cost, highly scalable monitoring system aimed at accurately collecting and aggregating environmental data from various classrooms distributed across different educational establishments and with the capacity to integrate all those who wish to be associated with the project. Boxes equipped with a diverse range of low-cost, energy-efficient sensors have been developed. The system provides an intuitive interface for accessing and monitoring in real time variables such as CO_2 levels, humidity, temperature and particulates for each classroom in the monitored schools. In addition, the environmental data was complemented with an indication of classroom occupancy, through the collaboration of the school community, thus providing a solution that respects the privacy of the people involved and does not require the use of invasive data collection methods such as cameras. Artificial intelligence techniques were applied to this data in order to classify classroom occupancy, achieving an accuracy of at least 83% when classifying occupancy with a general model for all classrooms and at least 85% when trained for a specific classroom.

Dedico este trabalho à minha namorada e família que me motivaram e acompanharam ao longo de toda esta etapa.

Agradecimentos

A conclusão desta dissertação representa o final de um objetivo pessoal que, ao longo destes dois anos, contribuiu para o aumento do meu conhecimento.

Um especial agradecimento aos Professores-orientadores, Dr. Pedro Sobral e Dr. Rui Moreira, que deram a conhecer o tema da dissertação e que acompanharam todo o processo com os seus altos e baixos, com grande rigor cientifico e opiniões valiosas, contribuindo para que todos os objetivos traçados fossem cumpridos. Agradecer a todos os Professores que participaram ativamente na campanha de recolha de dados, tornando-se uma parte fulcral para que atingíssemos todos os nossos objetivos, um muito obrigado.

Um obrigado muito especial à minha namorada, Maria Eduarda, que sempre me acompanhou nesta e noutras aventuras a que me propus, incentivando-me a continuar mesmo quando me questionei se valeria a pena. Foi uma parte essencial para a conclusão desta etapa e trabalho, tentando sempre ajudar da melhor forma, com paciência, amor e manifestando o seu orgulho a cada etapa concluída.

Um obrigado muito especial aos meus pais, Cláudia e Adriano, por me proporcionarem todas as condições necessárias para que este caminho fosse percorrido sem que nada faltasse, juntamente com a motivação, carinho, orgulho e amor que foram demonstrando ao longo do caminho. Agradecer também à minha irmã, Inês, por me motivar a ser um modelo a seguir para ela. Ao meu avô, Valentim, pela partilha da sua experiência e conhecimento. À minha avó, Elisabete, que fez desta etapa um sonho seu a cumprir e, por isso, a sua conclusão é-lhe também dedicada.

Um grande obrigado ao grupo de colegas e amigos, Pedro Pinheiro, Manuel Palavras, Bernardo Mantas, João Januário e Pedro Castro, que contribuíram no dia-a-dia para que esta etapa fosse ultrapassada mais facilmente, através do seu companheirismo, amizade e motivação. Um agradecimento também ao Filipe, Mara, Mariana e Cristina que também fizeram parte desta etapa que agora se conclui.

Por último, quero expressar a minha gratidão à Universidade Fernando Pessoa, que me transmitiu os mais valiosos princípios e ensinamentos para levar adiante na minha jornada.

Lista de Siglas

IoT *Internet of Things*

HA *HomeAssistant*

IAQ *Indoor Air Quality*

VM *Virtual Machine*

MQTT *Message Queuing Telemetry Transport*

CSV *Comma-Separated Values*

UUID *Universally Unique Identifier*

KNN *K-Nearest Neighbor*

MLP *Multilayer Perceptron*

CNN *Convolutional Neural Network*

RF *Random Forest*

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Atualmente, em Portugal, a maioria das escolas carece de componentes sensoriais para a recolha de informações ambientais no interior das salas de aula. Vários estudos realizados ao longo dos anos indicam que a baixa qualidade do ar interior (IAQ), em espaços fechados, pode ter um impacto direto no bem-estar, causando dificuldades de concentração, fadiga, dores de cabeça e, em casos extremos, a perda de consciência das pessoas. Neste contexto, e face à relevância do tema, o presente trabalho propôs-se a desenvolver um sistema de monitorização modular, de fácil replicação, escalabilidade e baixo custo, que possibilitasse a recolha precisa de dados ambientais em diversas salas de aula, localizadas em diferentes estabelecimentos de ensino. Para alcançar este objetivo, foram concebidas caixas de monitorização equipadas com uma ampla variedade de sensores de baixo consumo energético. Além disso, este sistema inclui a recolha de dados relativos à ocupação de cada espaço monitorizado, graças à colaboração dos docentes de cada escola. Desta forma, é possível criar um mapeamento da ocupação dos espaços sem recorrer a câmaras ou a outros métodos intrusivos que possam comprometer a privacidade dos intervenientes. Este sistema tem como objetivo central disponibilizar, de forma acessível, informações pormenorizadas acerca de cada sala de aula. Por exemplo, permite a monitorização em tempo real de variáveis como os níveis dióxido de carbono (CO_2), humidade, temperatura e partículas. Através da recolha de dados ambientais e da ocupação de cada sala, e com recurso a técnicas de inteligência artificial, procedeu-se à classificação da ocupação desses espaços. Esta abordagem contribui significativamente para um planeamento mais eficiente da utilização dos espaços nas instituições de ensino.

1.1 Motivação do Trabalho

O presente tema foi sugerido e apresentado pelos Professores Orientadores, despertando desde cedo um profundo interesse. Dado que os parâmetros ambientais em espaços públicos fechados raramente são objeto de estudo, torna-se inegável a pertinência de investigar

o seu impacto no quotidiano das pessoas. Assim, seguindo a relevância deste tópico, a pesquisa realizada destacou a necessidade de analisar de que forma as condições ambientais nas salas de aula podem contribuir para uma gestão mais eficiente desses espaços. Caso a gestão eficiente da ocupação dos espaços não seja realizada, isso pode resultar em consequências adversas. Em termos ambientais, o consumo ineficiente de recursos, como a energia, pode levar a um maior desperdício e uma pegada ambiental mais significativa. Além disso, as condições ambientais inadequadas, como níveis elevados de poluentes ou desconforto térmico, podem afetar negativamente o bem-estar das pessoas. Do ponto de vista do conforto, a falta de gestão adequada dos espaços pode resultar em salas superlotadas, falta de ventilação adequada e desconforto térmico, o que pode causar desconforto, fadiga e até mesmo afetar a saúde dos ocupantes. Por fim, em termos de desempenho, um ambiente de aprendizagem insatisfatório devido a condições ambientais inadequadas pode prejudicar a concentração dos alunos e, consequentemente, o seu desempenho académico. Portanto, a gestão eficiente dos espaços é fundamental não apenas para promover a sustentabilidade ambiental, mas também para garantir o conforto e o desempenho ideais dos ocupantes.

1.2 Problema

O ambiente nas salas de aula desempenha um papel crucial na experiência de ensino e aprendizagem. No entanto, é preocupante notar que, em muitas instituições de ensino, não existe uma recolha sistemática de dados ambientais em sala de aula. Um dos desafios está diretamente relacionado com o custo de aquisição dos vários sensores e outros materiais necessários para efetuar a recolha de dados ambientais. Além disso, a complexidade da configuração e interligação dos diversos módulos associados é um obstáculo a considerar, especialmente quando se alarga a recolha a diversas salas de aula em diferentes escolas. Outro fator a ter em conta, é a variedade de parâmetros ambientais que podem ser monitorizados, o que pode tornar ainda mais complexa a escolha de um conjunto que conduza a resultados satisfatórios. A ausência de dados ambientais pode acarretar consequências significativas, como a privação da capacidade de compreender o impacto do ambiente nas salas de aula e no bem-estar dos intervenientes ou uma gestão ineficiente da ocupação dos espaços letivos.

1.3 Objetivos

A presente dissertação tem como seu principal objetivo contribuir em três áreas distintas, embora interligadas, a saber:

- **Desenhar e Implementar um Sistema de Monitorização:** desenvolver um sis-

tema de monitorização abrangente, destinado à instalação em múltiplas salas de aula, abarcando diversas escolas. O principal foco recai na recolha eficaz de informações ambientais, com vista à compreensão das condições nas salas de aula, tudo isto assegurando o devido respeito pela privacidade dos alunos e docentes durante o processo de recolha de dados. Simultaneamente, almeja-se a implementação de um sistema escalável que, por sua vez, incorpore opções de baixo custo, viabilizando a adoção acessível desta tecnologia por parte das escolas.

- **Construir Conjuntos de Dados Representativos:** criar um conjunto de dados abrangente e representativo, consolidando informações provenientes de diversas salas de aula, com o intuito de proporcionar informações valiosas sobre o ambiente presente nestes espaços. Este processo engloba a recolha de dados sobre diversas características ambientais, complementada por informações sobre a ocupação das salas através de métodos não intrusivos. A subsequente organização destes dados é realizada de forma estruturada e acessível, permitindo a disponibilização de uma visão detalhada acerca do ambiente em estudo.

- **Utilizar Técnicas de Inteligência Artificial para a Classificação de Ocupação:** utilizar algoritmos de inteligência artificial para desenvolver modelos com a capacidade de classificar a ocupação das salas de aula, visando aprimorar a eficiência na gestão da ocupação desses espaços.

- **Conjuntos de Dados Públicos:** pretende-se que todos os conjuntos de dados elaborados ao longo da presente dissertação sejam de domínio público, permitindo que outros estudos sejam realizados com recurso aos mesmos. Isto é fundamental para promover a colaboração e o avanço do conhecimento na área. Os dados estão disponíveis em formato *Comma-Separated Values* (CSV), e podem ser acedidos através do *repositório* online.

1.4 Estrutura da Dissertação

Esta dissertação de mestrado está organizada em seis capítulos principais, conforme se descreve abaixo.

No estado da arte, abordam-se conceitos tecnológicos considerados fundamentais para o contexto desta dissertação, nomeadamente, sistemas *Internet of Things* (IoT) e conceitos de inteligência artificial. Além disso, realiza-se uma análise de trabalhos relacionados com o tema, começando por explorar a combinação de sistemas IoT e técnicas de inteligência artificial aplicados à qualidade do ar e, posteriormente, com ênfase na ocupação de espaços, realizando uma análise critica acerca dos mesmos.

O terceiro capítulo descreve pormenorizadamente a arquitetura do sistema *Airmon*, apresentando os seus objetivos, uma visão geral da arquitetura, o conceito de instâncias, a sensorização dos componentes ambientais, o fluxo de dados e, ainda, aborda tópicos como escalabilidade, desempenho, segurança, tolerância a falhas e recuperação.

Na fase de implementação do sistema *Airmon*, procura-se detalhar a concretização da arquitetura previamente definida no capítulo anterior. São abordadas em detalhe as integrações do *HomeAssistant* (HA) utilizadas, bem como todos os sensores que compõem as caixas de monitorização. Também são descritos os vários fluxos de dados entre as diferentes instâncias.

O quinto capítulo incide sobre a componente dos conjuntos de dados, onde se explana o procedimento de recolha de dados adotado, os métodos de pré-processamento e limpeza dos dados recolhidos. Além disso, realiza-se uma análise detalhada dos conjuntos de dados construídos.

O sexto capítulo aborda as experiências realizadas envolvendo técnicas de inteligência artificial aplicadas aos conjuntos de dados elaborados. São explicados em detalhe os parâmetros dos testes, a metodologia adotada e os resultados obtidos, tanto para cada sala monitorizada individualmente, como para o conjunto de todas as salas. É igualmente realizada uma avaliação da acurácia das diversas técnicas utilizadas.

No último capítulo, a conclusão, resume-se todo o trabalho desenvolvido, enfatizando as etapas do desenvolvimento do sistema *Airmon* destacando-se os principais resultados e contribuições. Adicionalmente, são discutidos potenciais objetivos para trabalhos futuros.

Capítulo 2

Estado da Arte

Este capítulo aborda os conceitos tecnológicos que serviram como base para o desenvolvimento desta dissertação, bem como a revisão de trabalhos que seguiram a mesma linha temática e a aplicação de Inteligência Artificial em dados ambientais.

2.1 Conceitos Tecnológicos

2.1.1 Sistemas IoT

Os sistemas IoT resultam da combinação de diferentes tecnologias, que fornecem soluções baseadas na integração de tecnologia de informação, referente ao hardware e software utilizados para armazenar, recuperar e processar dados que, recorrendo a tecnologia de comunicação, como sistemas eletrónicos, os comunica entre indivíduos ou grupos (Patel et al., 2016).

Estes sistemas possuem características como a interconectividade, a deteção inteligente, que ajuda a criar experiências que refletem uma verdadeira consciência do mundo físico entre pessoas e objetos e a economia de energia baseada, por exemplo, em informação de um sensor de movimento para ligar e desligar luzes (Singh and Singh, 2015).

Tipicamente, os sistemas de IoT podem adotar diversas arquiteturas, entre as quais se destacam três tipos: *cloud computing, edge computing e fog computing*. No que diz respeito à *cloud computing*, esta arquitetura foi projetada para permitir uma gestão centralizada e apresenta como principais características a interligação de objetos físicos e informação virtual, baseando-se em tecnologias de comunicação ubíquas, interoperáveis e a pedido. Promove, assim, a utilização e libertação rápida e eficiente dos recursos, com um mínimo de esforço de gestão e interação, permitindo uma partilha de recursos computacionais. Por sua vez, a *fog computing* é uma arquitetura concebida para possibilitar serviços distribuídos complexos próximos dos dispositivos finais. Os seus objetivos incluem a preservação da bateria dos dispositivos finais e a redução da latência de comunicação. Além disso, contribui para aumentar a privacidade dos utilizadores, uma vez que

processa os dados brutos localmente em vez de os enviar para a nuvem. A arquitetura de *edge computing* foi desenvolvida para permitir a computação nos nós finais, também conhecidos como motes ou "coisas", tornando possível o processamento e a transformação dos dados brutos em informações mais elaboradas. Normalmente, esses dispositivos não precisam de transmitir constantemente, o que permite a implementação de uma estratégia de gestão energética mais eficaz em termos de consumo e autonomia (Moreira et al., 2020). Os sistemas IoT podem ser aplicados em vários cenários, no caso de Saxena et al. (2019), o objetivo foi automatizar tarefas do dia-a-dia, contribuindo para uma melhor qualidade de vida das pessoas, automatizando parte das rotinas pessoais e tarefas. No caso de of Technology et al. (2017), foi proposto um sistema IoT que tinha como objetivo auxiliar na gestão de resíduos provocados pelos resíduos domésticos.

2.1.2 Inteligência Artificial

No âmbito da utilização de inteligência artificial, é relevante esclarecer alguns conceitos fundamentais relacionados com o tema. Segundo Farnham et al. (2019) os sistemas de aprendizagem automática podem ser classificados com base na quantidade e no tipo de supervisão que recebem durante o processo de treino. Destacam-se quatro categorias principais: aprendizagem supervisionada, aprendizagem não supervisionada, aprendizagem semi-supervisionada e aprendizagem por reforço.

- **Aprendizagem Supervisionada:** neste tipo de aprendizagem, os dados de treino fornecidos ao algoritmo incluem as soluções desejadas, geralmente designadas por etiquetas. Uma tarefa típica de aprendizagem supervisionada é a classificação, como, por exemplo, classificar um email como *spam*.

- **Aprendizagem Não Supervisionada:** ao contrário do exemplo anterior, na aprendizagem não supervisionada, as soluções não são indicadas nos dados de treino. Este tipo de aprendizagem pode ser aplicado, por exemplo, para detetar grupos de visitantes de um *website* com base nos registos de visitas.

- **Aprendizagem Semi-Supervisionada:** a aprendizagem semi-supervisionada destaca-se pela sua capacidade de lidar com dados que incluem alguns exemplos etiquetados e outros não etiquetados. Serviços de alojamento de fotos representam um bom exemplo de aplicação deste tipo de aprendizagem, uma vez que, após a submissão de algumas fotografias, os modelos conseguem reconhecer a mesma pessoa em diversas imagens e apenas requerem que o utilizador identifique essa pessoa. A partir desse ponto, é possível realizar ações como a pesquisa de fotografias dessa pessoa.

- **Aprendizagem por Reforço:** a aprendizagem por reforço difere significativamente das outras categorias. Neste caso, o sistema de aprendizagem chamado agente, é

capaz de observar o ambiente, selecionar e executar ações e, em troca, receber recompensas ou penalizações, normalmente na forma de recompensas negativas. Um exemplo típico de aplicação desse tipo de aprendizagem é a programação de robôs para aprenderem a caminhar.

Entre os exemplos de algoritmos de aprendizagem supervisionada, explorados na presente dissertação, destacam-se o *Random Forest* (RF) e a *Convolutional Neural Network* (CNN). O algoritmo RF, consiste num conjunto de árvores de decisão que introduzem um grau adicional de aleatoriedade no processo de crescimento dessas árvores. Ao invés de procurar a melhor característica ao dividir um nó, o RF procura a melhor característica num subconjunto aleatório das características disponíveis. Essa abordagem resulta numa maior diversidade de árvores, o que troca um viés potencialmente elevado por uma variância mais baixa, geralmente levando à criação de um modelo globalmente mais eficaz (Farnham et al., 2019). Um exemplo de aplicação do algoritmo RF é apresentado no trabalho de Rodriguez-Galiano et al. (2012), que tinha como objetivo a classificação da cobertura do solo. Já as redes neurais convolucionais (CNN) possuem uma arquitetura de camadas que incorporam conexões espacialmente locais, particularmente nas camadas iniciais e padrões de pesos que são replicados nas unidades de cada camada. Esse padrão de pesos replicados é conhecido como *kernel*, e o processo de aplicação desse *kernel* é denominado como convolução (Norvig and Russell, 2021).

2.2 Trabalho Relacionado

2.2.1 Sistemas IoT e IA na qualidade do ar

Até aos dias de hoje, a monitorização de parâmetros ambientais era uma atividade frequentemente episódica, que implicava a utilização de equipamento dispendioso e complexo, bem como a necessidade de deslocação de especialistas para operá-lo. Contudo, com os avanços proporcionados pelo IoT, assiste-se a uma notável transformação deste cenário. O IoT viabiliza a monitorização contínua e em tempo real destes parâmetros, a um custo substancialmente inferior, tornando-se acessível a um público mais vasto. Esta democratização da monitorização ambiental constitui um marco significativo, abrindo portas para uma compreensão mais abrangente e acessível do ambiente que nos envolve.

Trabalhos como Wang et al. (2023) enfatizam a importância crucial da qualidade do ar em ambientes fechados para a saúde, dado que a maioria das pessoas passa maior parte do seu tempo neles. A medição em tempo real por meio da utilização de sensores torna-se cada vez mais relevante neste contexto. Estes dispositivos proporcionam uma análise dinâmica e abrangente da qualidade do ar, possibilitando a identificação e monitorização contínua de uma variedade de poluentes, sejam eles gasosos (como CO, SO2, O3, NOx

e COVs) ou provenientes de partículas (como PM10, PM2,5, PM1 e BC). Além de monitorizar diretamente a presença destes poluentes, a previsão com recurso a técnicas de inteligência artificial da qualidade do ar, desempenha também um papel importante na implementação de estratégias preventivas eficazes. Modelos estatísticos são normalmente utilizados para prever a concentração destes poluentes, baseados em dados provenientes dos sensores, oferecendo uma visão aprofundada das condições do ar em ambiente fechado. Esta compreensão detalhada é essencial para a existência de um ambiente saudável e seguro, com potencial para impactar positivamente a saúde e o bem-estar dos ocupantes. Em ambiente escolar, como é o caso do trabalho desenvolvido nesta dissertação, este tema é crucial pois pode impactar a capacidade e qualidade de aprendizagem dos alunos. Trabalhos como Bogdanovica et al. (2020) indicam que, uma má qualidade do ar numa sala de aula, pode levar a que, por exemplo, os sintomas de fadiga aumentem entre 21 a 45%. No retorno às aulas durante a pandemia do COVID19, foram definidas novas regras para os alunos seguirem, como ficar dentro das salas durante os intervalos o que pode contribuir para um nível não saudável de CO_2, uma vez que interferem com a ventilação natural das salas de aula (Zemitis et al., 2021).

Uma vez que cada espaço fechado possui as suas próprias características, como por exemplo materiais de construção, capacidade de ocupação, exposição solar, entre outras, por vezes é necessário testar diversas técnicas de inteligência artificial de forma a apurar os melhores resultados. Foi o caso do trabalho Liu et al. (2023) que aplicou técnicas de *machine learning* para estimar de forma precisa as emissões de *VOCs* (compostos orgânicos voláteis), relacionadas com atividades humanas numa sala de aula universitária. A técnica *least squares support vector machine (LSSVM)* obteve melhores resultados quando comparada com as técnicas de RF *regression*, *adaptive boosting (Adaboost)*, *gradient boosting regression tree (GBRT)*, *extreme gradient boosting (XGboost)* na previsão da concentração de *6-methyl-5-hepten-2-one (6-MHO)*, medindo variáveis como o número de ocupantes (registado manualmente), a humidade relativa, a concentração de ozono e a temperatura. Salienta-se, ainda, a alta acurácia na previsão da concentração de 4-OPA, utilizando a mesma técnica destacada anteriormente, com um erro percentual absoluto médio (MAPE) inferior a 5%.

Cada vez mais as técnicas de *deep learning* são utilizadas no contexto de previsão da qualidade do ar em espaços fechados. No trabalho Marzouk and Atef (2022) foram criadas treze caixas de monitorização com sensores para a recolha e construção de conjuntos de dados de treze salas de uma universidade, durante nove dias. As informações ambientais recolhidas dentro das salas foram a temperatura, humidade, pressão do ar, monóxido de carbono, dióxido de carbono e partículas, enquanto que, no exterior, foram recolhidas informações acerca da velocidade e direção do vento, temperatura e humidade. O conjunto de dados construído, foi dividido em duas partes, sendo a primeira destinada ao treino de uma *Artificial Neural Network (ANN)* com 80% dos dados e a segunda destinada

ao teste das previsões efetuadas pelo modelo contando com os restantes 20%. Foi ainda utilizado um conjunto de validação durante a fase de treino do modelo para verificar o desempenho do mesmo. Este processo foi realizado para garantir que o modelo desenvolvia a capacidade de prever dados a partir dos quais não tinha conhecimento. A acurácia obtida pelo modelo no conjunto de testes variou entre os 82% e os 91%. Uma vez que cada sala possui características particulares, como a sua utilização ou número de janelas, os autores concluem que cada sala deveria ser estudada separadamente para uma melhor avaliação do *Indoor Air Quality* (IAQ) em edifícios.

Através da utilização de variáveis como a temperatura, concentração de CO_2, total de *VOCs*, partículas 2.5, partículas 10 e humidade relativa, o trabalho Lee et al. (2023), desenvolveu sete modelos de inteligência artificial com o objetivo de determinar em tempo real ou a curto prazo, a concentração de bactérias, fungos, pólen e partículas 2.5 e 10. As técnicas utilizadas foram *Linear Regression*, *Lasso Regression*, RF, *XgBoost*, *Multilayer Perceptron* (MLP), *Long short-term memory (LSTM)* e *Recurrent Neural Network (RNN)*. Os conjuntos de dados utilizados foram repartidos num rácio de 9:1, ou seja, 90% para treino e 10% para testes. Na realização do estudo, técnicas como RF, *XgBoost* e *Long Short-Term Memory (LSTM)* destacaram-se pelo desempenho superior face às outras técnicas, com a última a obter uma acurácia entre os 60% e os 80% para o conjunto de testes.

2.2.2 Sistemas IoT e IA na ocupação de espaços

Ao longo dos anos, a preocupação com a monitorização de dados ambientais em espaços fechados tem vindo a aumentar cada vez mais, contribuindo para o aparecimento de novos sistemas de monitorização e diferentes abordagens. Os sistemas de monitorização com aplicação de algoritmos de inteligência artificial, podem contribuir para um melhor planeamento e otimização das condições ambientais em salas de aula, muitas vezes baseados em informação recolhida dentro das mesmas, permitindo prever a ocupação com um elevado grau de eficácia (Bockstael and Jadin, 2018). Ao longo desta secção, serão descritas diferentes abordagens, baseadas na análise de trabalhos anteriores, quanto aos protocolos de comunicação entre componentes, os tipos de sensores utilizados para recolha de informação ambiental, as operações de pré-processamento dos dados recolhidos, os algoritmos de inteligência aplicados sobre os mesmos e os resultados obtidos.

Bockstael and Jadin (2018), utilizaram um sensor Elsys ERS para medir informações ambientais de duas salas de aula. Este sensor utiliza o protocolo *Lora* e comunica os dados recolhidos para uma *Raspberry PI* equipada com uma antena *Lora*. Por sua vez, a *Raspberry PI* após receção dos dados, comunica via wi-fi para uma plataforma (*The Things Network*) presente na internet. No estudo dos autores supracitados, o sensor recolheu diversas informações ambientais como os níveis de CO_2, temperatura, humidade,

luminosidade e movimento, sendo que a periodicidade da recolha poderia ser facilmente alterada através de uma aplicação móvel. O conjunto de dados englobava, para além da informação ambiental, o número de pessoas presentes na sala. Para obter esse número, inicialmente foi utilizada uma câmara montada na entrada das salas de aula para registar as entradas e saídas, contudo, este método foi descartado derivado à relutância das pessoas em relação à mesma. A segunda abordagem foi a utilização de um sistema de botões, em que ao entrarem na sala, as pessoas carregariam num botão para incrementar o contador de pessoas e ao sair carregariam noutro botão para decrementar, no entanto, esta abordagem era altamente dependente das pessoas que utilizavam as salas. A terceira abordagem foi a de colocar um código QR na entrada da sala, em que qualquer pessoa poderia submeter o número de pessoas presentes na sala naquele momento. Além destas variáveis, foram ainda criadas mais duas, nomeadamente, a média de CO_2, que corresponde à média dos registos recolhidos nos últimos cinco minutos e a derivada de CO_2, que corresponde à derivada da curva da média de CO_2, em que é utilizada a última leitura de CO_2 e a atual, estas duas variáveis correspondem também a uma técnica de pré-processamento de dados designada por extração de características. Foram ainda aplicadas outras técnicas como balanceamento de classes, em que o objetivo é equilibrar as classes para que uma classe não contenha muitos mais registos em comparação a outra e remoção de *outliers* em que se pretende retirar valores atípicos em relação ao conjunto de dados.

O objetivo deste trabalho era o de estimar a ocupação exata das salas de aula, criando-se um modelo de voto, em que diversos modelos de inteligência artificial fazem a sua previsão da classe objetivo (número de pessoas) e a média é calculada para eleger a previsão final. Os modelos utilizados para compor este "super"modelo foram: o RF Classifier e Regressor, o *K-Nearest Neighbor* (KNN) Regressor e Classifier, que segue o principio de que, se uma nova amostra está próxima a alguns exemplos de treino de uma classe, então esta deve ser a mesma. Utiliza, ainda, um parâmetro K que irá representar o número de vizinhos mais próximos envolvidos na classificação de novas amostras. Utilizou-se também o modelo MLP Regressor e Classifier que funciona sob a premissa de que, aplicando várias transformações lineares umas atrás das outras, pode eventualmente resultar numa função não linear, resultando assim num modelo que se pode ajustar a um limite de decisão não linear. Ademais, os autores procuraram perceber qual a importância que cada variável tinha para cada uma das técnicas, removendo-a e avaliando o seu efeito na acurácia no conjunto de testes. Em relação ao treino do modelo, numa fase inicial, o conjunto de dados de uma sala foi utilizado para treinar o modelo e o conjunto da outra sala utilizado para testes. Conforme esperado, os resultados obtidos foram maus, uma vez que as previsões do modelo foram perto de aleatórias, sendo que a acurácia restrita (sem margem de erro) ficou entre os 5 e os 6%. A segunda abordagem foi a de treinar o modelo com os conjuntos de dados das duas salas, inicialmente com balanceamento, uma vez que diferiam no tamanho e, após isso, sem balanceamento. O modelo conseguiu uma

acurácia restrita de mais de 87% (em média 90%), chegando aos 93% se permitir uma margem de erro de 0.35 vezes o número de pessoas presentes, isto é, se existirem sete pessoas presentes na sala e permitirmos uma margem de erro máxima de duas pessoas.

Outro trabalho que optou por utilizar um sensor já montado, foi o de Zhang et al. (2023) que utilizou uma estação meteorológica para monitorizar uma sala de aula na universidade *Virginia Tech*, em que os dados recolhidos através da mesma, foram enviados para um servidor remoto, através da utilização de wi-fi e acedidos pela aplicação da estação. As variáveis ambientais recolhidas pela estação foram o CO_2, o nível de ruído, a pressão do ar, a temperatura e humidade. O conjunto de dados foi também composto pelo número de pessoas dentro da sala e, nesse sentido, foi necessário contar, no início de cada aula, as pessoas presentes. A aplicação da estação meteorológica enviava também uma notificação quando os níveis de CO_2 ultrapassavam um determinado intervalo, permitindo assim contabilizar as pessoas que estivessem na sala numa visita mais breve.

Após a recolha dos dados e antes de os utilizar na criação do seu modelo, o autor aplicou técnicas de pré-processamento, como a normalização, que tem como objetivo converter todos os valores numéricos para um intervalo entre zero e um, facilitando assim o processamento dos dados por parte do modelo. Foi ainda aplicada a técnica de codificação distribuída *(One-hot encoding)* que transforma números em representação binária.

O principal objetivo deste trabalho era o de estimar a ocupação exata de uma sala de aula, tendo-se utilizado para o efeito, dois modelos, o *Long-Short-Term Memory* (LSTM) que é um tipo único de rede neuronal recorrente que geralmente possui desempenhos superiores às redes neuronais recorrentes convencionais e uma das suas vantagens, é a sua capacidade de processar de forma eficiente grandes sequências de dados. O segundo modelo utilizado foi o *Attention Mechanism* que é frequentemente utilizado em simultâneo com arquiteturas de codificação e descodificação em tarefas como sumarização de textos, análise de sentimento de textos ou tradução. No caso de Zhang et al. (2023), o modelo foi adaptado para o contexto ambiental, uma vez que os seus dados representavam cinco vetores de entrada e o número de pessoas presentes na sala representava um vetor de saída, tornando assim o contexto do trabalho num caso de uso viável para a utilização deste modelo.

Para avaliar o desempenho do modelo, foi utilizada a métrica "*F1 Score*"que consiste num método que avalia para além da precisão, o recordar do modelo. Tipicamente quando a precisão é alta, o valor de recordar é baixo e vice-versa. O cálculo de *F1 Score*, tenta equilibrar estes dois parâmetros de forma a avaliar o desempenho do modelo. O autor realizou diversos testes nas duas técnicas de inteligência artificial, sendo que o *LSTM* quando treinado sem a técnica de pré-processamento de codificação distribuída, obtém um resultado F1 de 77%, este resultado sobe para 93% quando aplicada essa mesma técnica. O melhor resultado obtido neste trabalho foi com a utilização do *Attention Mechanism* quando utilizado com um optimizador do tipo *AdaMax* chegando a um resultado F1 de

95%.

Outros trabalhos optaram por montar os seus próprios conjuntos de sensores, o que permite mais liberdade na escolha das variáveis ambientais a medir, assim como na escolha dos modelos de sensores a utilizar.Foi o caso de Candanedo and Feldheim (2016) que optaram por escolher um sensor DHT22 para medição de temperatura e humidade, um telaire 6613 para medir os valores de CO_2 e um TSL2561 para a medição do nível de luminosidade. Os autores à semelhança de Bockstael and Jadin (2018), decidiram criar uma nova variável a partir de informações recolhidas pelos sensores, calculando o rácio de humidade através da temperatura e humidade. Foi ainda utilizada uma variável numérica que indica se o dia é semanal ou não e uma outra variável que representa o número de segundos desde a meia-noite até ao momento atual, permitindo identificar a que período temporal corresponde o registo atual.

Candanedo and Feldheim (2016) optaram, pela utilização de modelos estatísticos para determinar se um escritório estava ocupado ou não. Foram utilizados quatro modelos, nomeadamente, o *Classification And Regression Trees* (CART) que estratifica a região onde as previsões são feitas e o seu espaço em várias regiões simples, o RF que, de forma a melhorar a sua acurácia, cria muitas árvores de classificação e, ao construí-las, uma amostra aleatória dos preditores é selecionada e as melhores divisões são utilizadas, o *Gradient Boost Machines* (GBM) que utilizam informações de árvores geradas anteriormente para melhorar a sua previsão e o *Latent Dirichlet Allocation* (LDA) que utiliza o teorema de *Baye*, que constrói um classificador com base numa combinação linear das variáveis sob a suposição de que cada uma segue uma distribuição normal.

Após os modelos terminarem a fase de treino, estes foram avaliados recorrendo aos conjuntos de dados de treino, conjunto de testes um e conjunto de testes dois, utilizando uma matriz de confusão para calcular a acurácia dos modelos. A fórmula utilizada para calcular a acurácia das previsões dos modelos foi a seguinte: Acurácia = (A + D) / (A + B + C + D) . Sendo que A representa a soma dos verdadeiros positivos com D que representa os verdadeiros negativos, divididos pelo número total de previsões. De uma forma geral, os modelos do tipo RF revelaram uma maior acurácia nos conjuntos de dados de treino, face aos conjuntos de testes, sendo que os modelos que conseguiram melhores resultados nos conjuntos de testes foram os LDA com 97.9% e 99.33% no primeiro e segundo conjunto respetivamente. Conclui-se, então, que os modelos do tipo LDA fornecem uma previsão consistente e precisa para os casos presentes nos diversos conjuntos de dados do trabalho, contudo, todos os modelos conseguiram bons desempenhos quando treinados com todas as variáveis, oscilando entre os 94% e os 99%.

À semelhança do caso anterior, também no trabalho Kim et al. (2023), optaram por escolher os próprios sensores para realizar a montagem de caixas de monitorização. No caso da variável de humidade, escolheram um sensor DHT22 tal como Candanedo and Feldheim (2016), anteriormente referidos. Para medição dos níveis de CO_2, o sensor uti-

lizado foi um CM1107 e no que diz respeito às partículas, utilizaram um PM2008. Este trabalho distingue-se pela forma como alimentam as caixas de monitorização, fazendo-o com recurso à utilização de quatro baterias de 3.7V, o que permitiu uma recolha contínua de dados de aproximadamente vinte cinco horas e armazenamento interno dos dados a cada minuto. O número de ocupantes foi medido através dos horários das salas, juntamente com a observação, controlando de uma forma não intrusiva e respeitando a privacidade dos ocupantes.

O principal objetivo do trabalho foi desenvolver dois algoritmos para prever o número de ocupantes e a carga de energia nos edifícios universitários. Para isso, foram recolhidas informações ambientais, como temperatura, humidade, concentração de CO_2 e PM2.5 com recurso aos sensores indicados anteriormente, em cada sala. Posteriormente, foi desenvolvido um algoritmo para prever o número de ocupantes, usando um modelo de regressão linear múltipla(MLR) com os dados ambientais recolhidos, obtendo uma alta acurácia na comparação entre os valores previstos e reais. Com base nos resultados do primeiro algoritmo e nos dados dos sensores ambientais, foi criado um segundo algoritmo para prever a carga de energia. Além disso, o estudo estabeleceu a hipótese de que a temperatura, humidade, concentração de CO_2 e partículas aumentariam de acordo com o número de ocupantes. O *software* de análise de energia TRNSYS foi também utilizado para calcular o consumo de energia dos espaços, considerando diversas variáveis, como as estruturas de paredes e janelas, dados dos equipamentos, temperatura interna/externa e número de ocupantes.

As métricas de avaliação utilizadas no trabalho foram: o MSE que representa a média dos erros dos quadrados entre os valores medidos e previstos. Um MSE mais próximo de zero indica que o valor previsto está mais próximo do valor original, correspondendo assim a uma maior acurácia. O RMSD foi também calculado e representa o desvio padrão dos erros entre os valores medidos e previstos, representando o grau de dispersão com base na linha de regressão como um valor. Portanto, um RMSD mais próximo de zero indica também uma maior acurácia. O MAE foi utilizado principalmente como indicador para avaliação de regressão, sendo que um MAE mais próximo de zero indica um modelo de maior qualidade.

O trabalho de Kim et al. (2023), propôs assim um método simples e económico quando comparado com outras soluções existentes para prever indiretamente o número de ocupantes usando a concentração de CO_2, e para prever o consumo de energia usando o número previsto de ocupantes e as temperaturas internas e externas medidas em tempo real.

A deteção de presenças em espaços fechados é também um grande desafio. O trabalho Mohammadabadi et al. (2022) dedica-se a enfrentá-lo através da recolha e análise criteriosa de dados ambientais, como o CO_2, a temperatura e a humidade relativa. Estes foram recolhidos em diferentes cenários, abrangendo um quarto, um escritório de

dimensões reduzidas e um escritório de maior envergadura. Para garantir a privacidade, a replicabilidade e viabilidade da implementação em contextos distintos, foi estrategicamente planeado o uso de um número mínimo de sensores económicos. Estes dispositivos foram distribuídos estrategicamente nos ambientes estudados, transmitindo os dados recolhidos para um *cloud storage*. A variável de presença, elemento central neste estudo, foi registada por dois voluntários ao longo de treze dias, enriquecendo assim o conjunto de dados analisados. A preparação dos dados incluiu a sua divisão em conjuntos de treino (80%) e teste (20%). Dada a discrepância entre os dados de ocupação e não ocupação, optou-se pela técnica de *stratified sampling*, preservando a devida proporcionalidade no conjunto de teste. O cerne da análise reside na aplicação de modelos preditivos, com destaque para uma CNN combinada com *XGBoost*. Os resultados deste modelo foram então comparados com outras técnicas, como *Linear Regression (LR)*, *Decision Trees (DT)*, RF, *Gradient Boosting (GB)*, *K-means Clustering (KMC)*, KNN, *Support Vector Machine (SVM)*, CNN e *XGBoost* isoladas. Para avaliar o desempenho dos modelos, utilizaram-se métricas como o *F1 Score* e a *Mean Absolute Error (MAE)* para cada técnica. Os resultados destacaram o modelo proposto, a CNN combinada com XGBoost, que superou as outras técnicas em todos os ambientes avaliados. No quarto, o modelo proposto atingiu um notável *F1 Score* de 0.986 e a menor MAE de 0.011. Este desempenho foi igualmente evidente no escritório de tamanho reduzido, com um *F1 Score* de 0.876 e MAE de 0.029. Ainda que no escritório maior todas as técnicas tenham apresentado uma diminuição nos *F1 Scores*, o modelo proposto continuou a obter os melhores resultados, com um *F1 Score* de 0.751 e MAE de 0.073. Desta forma, os autores concluíram que, embora o *XGBoost* isolado tenha resultados semelhantes, a combinação das capacidades de aprendizagem da CNN com o *XGBoost* otimizou significativamente os resultados obtidos.

Em trabalhos como o de Tekler and Chong (2022), o objetivo foi o de prever o número de ocupantes de diferentes espaços, com o mínimo de sensores possível. Foram selecionados três locais de uma universidade para monitorização: um escritório exclusivo para investigadores, uma biblioteca acessível a todos os alunos e uma sala de aula. Durante 123 dias, foram recolhidos dados semanalmente, divididos em períodos de 5 minutos. Os dados recolhidos incluíram condições ambientais interiores e exteriores, o número de dispositivos ligados à rede Wi-Fi em cada ponto de acesso, dados de consumo de energia para diversos fins (como *Heating, ventilation, and air conditioning*(HVAC), ventiladores de teto, aparelhos elétricos e iluminação) e operações de HVAC registadas por sensores distribuídos a nível da sala e do edifício. O controlo de presenças foi registado manualmente, através das imagens de câmaras de vigilância localizadas à entrada de cada espaço monitorizado. Antes da aplicação das técnicas de inteligência artificial, procedeu-se ao pré-processamento dos dados. Este incluiu a remoção de dados em falta ou incorretos, resultantes de falhas nos sensores. Foram também utilizadas técnicas de extração e normalização de características nos dados. Além disso, procedeu-se à seleção das melhores

características para simplificar a complexidade e reduzir o tempo de treino dos modelos. Cinco técnicas de *deep learning* foram consideradas, incluindo *deep neural network (DNN)* e modelos sequenciais como LSTM, Bi-LSTM, *Gated recurrent units (GRU)* e *Bidirectional GRU (Bi-Gru)*, devido à sua capacidade de reter a informação temporal em dados de séries temporais. O conjunto de dados de treino representou 80% dos dados, enquanto o conjunto de testes representou 20%. Dada a importância das séries temporais para as técnicas utilizadas e para preservar a sua natureza, os conjuntos de dados não foram randomizados. Tanto os conjuntos de dados de treino como os de teste foram transformados, agrupando os pontos de dados sequenciais com base numa janela deslizante de tamanho 5 para a entrada do modelo. A verdade fundamental foi estabelecida como a contagem de ocupantes no próximo passo temporal (ou seja, uma previsão para 5 minutos à frente). Optaram pela escolha de uma janela de tamanho 5, pois uma janela mais curta limitaria a informação sobre as tendências históricas de ocupação para previsão, enquanto uma janela maior poderia prejudicar o desempenho preditivo do modelo devido à introdução de ruído excessivo. O mesmo raciocínio foi aplicado ao horizonte de previsão, que foi definido como 5 minutos para permitir aplicações em controlo preditivo em tempo real, evitando assim maiores erros de previsão. As métricas utilizadas para avaliação do desempenho foram o MAE e o RMSE. Ao analisar as melhores características, foi observado que os valores de CO_2 e o número de dispositivos ligados à rede Wi-Fi estavam consistentemente presentes na lista de características cruciais para a previsão de ocupação dos espaços. O modelo Bi-GRU apresentou os melhores resultados no escritório, com um MAE de 0.116 e um RMSE de 0.326, e na sala de aula, com um MAE de 0.085 e um RMSE de 0.318. Já o modelo GRU obteve os melhores resultados na biblioteca, com um MAE de 0.16 e um RMSE de 0.331.

Alguns estudos e, de forma a desenvolver um sistema não intrusivo, têm como objetivo indicar se um espaço de encontra ocupado ou não, como se verifica no trabalho Alsmirat et al. (2019). O objetivo foi desenvolver um sistema capaz de recolher e transmitir dados de forma não intrusiva para uma *cloud* privada, com o objetivo de, posteriormente, através do uso da técnica SVM, prever se uma sala de aula se encontra ocupada ou não. As informações recolhidas incluíam dados de movimento (PIR), através do uso do sensor HC-SR501, temperatura com o sensor BMP280, CO_2 com o sensor SGP30 , humidade com o sensor HC1080, som com o sensor KY-038 e luminosidade com o sensor SI1145. Estes dados eram enviados periodicamente (a cada 10 segundos) ou em resposta a eventos desencadeados pelos ocupantes, através de um microcontrolador NodeMCU com ligação Wi-fi. O experimento decorreu ao longo de duas semanas e o controlo de presenças na sala de aula foi realizado através de um módulo colocado à entrada da sala, equipado com sensores de movimento (PIR), som e CO2, projetado especificamente para detetar a entrada de pessoas na sala. Outro módulo foi colocado dentro da sala para recolher informações dos seis sensores durante a aula. Para facilitar a análise, foi aplicada a técnica

de *data summarization*, cujo o propósito é fornecer uma descrição concisa do conjunto de dados. Esta técnica é importante pois reduz o tamanho dos dados processados, diminuindo assim o tempo necessário para a sua análise. Os dados foram consolidados com base no identificador dos sensores e na localização da sala. Cada conjunto de dados recebeu um carimbo de tempo indicando quando chegou ao *gateway* antes de ser enviado para a *cloud*. 80% dos dados foram utilizados no conjunto de treino e o restante para teste. Para avaliação, foram utilizadas métricas como Precisão, *Recall* e *F1 Score*, sendo esta última crucial pois representa tanto a *Recall* quanto a precisão do modelo SVM. A precisão global da previsão foi de 96%.

O trabalho de Yang et al. (2021) é distintivo pela importância atribuída à estação do ano na previsão da ocupação de edifícios residenciais. O estudo foi realizado num apartamento equipado com um sistema *HEMS* que geria a componente energética. Este sistema era dotado de diversos sensores, incluindo os que monitorizavam CO_2 e humidade relativa (AMUN 716), luz e carga nas tomadas (*KNX Energy Module*: EM/S 3.16.1), temperatura (Varia) e movimento (PlanoCentro A-KNX). Foram ainda aplicadas algumas técnicas de pré-processamento no conjunto de dados, como o processamento de valores em falta, por exemplo uma falha contínua de dados, a remoção de valores atípicos (*outliers*) e normalização dos dados.

Uma vez que a estação do ano possui um peso importante no contexto do trabalho, foram treinados modelos para cada estação do ano, com base nos seus respetivos conjuntos de dados. Estes foram construídos tendo em conta características como o horário do dia (1,2,3), dia da semana (1 - Dia da semana, 0 - Fim de Semana), período do dia (1 - Período de pico de ocupação, 0 - Não período de pico de ocupação), temperatura exterior (C○), humidade exterior (%), irradiância solar (W/m2), velocidade do vento (m/s), iluminação exterior (Lux), presença ou ausência de chuva (1 - Chover , 0 - Não Chove), temperatura interior (C○), humidade interior (%), CO_2 interior (ppm), ponto de ajuste térmico (C○), luminosidade interior (Lux), posição das persianas (0 - Completamente abertas, 100 - Completamente fechadas), estado de bloqueio automático das janelas (1 - Auto bloqueio, 0 - Normal), consumo de energia para iluminação (Wh) e consumo de energia das tomadas elétricas (Wh). Foram aplicadas várias técnicas de inteligência artificial, como LR, SVM, DT, *Gradient Boosting Decision Trees (GBDT)*, RF e ANN e, de forma a avaliar o desempenho das mesmas, foram utilizadas as métricas F1 Score e *area under the curve (AUC)*. No decorrer do trabalho foi também realizado um estudo de seleção de características para cada técnica em cada estação do ano, comparando os resultados ao utilizar as características selecionadas versus utilizar todas as características. Foi concluído que a maioria dos modelos beneficiou desta técnica, resultando em melhorias significativas. Por exemplo, o modelo da Primavera obteve um aumento de 4% na métrica F1 e 6% na métrica AUC quando comparado com utilizar todas as características. Outra análise realizada no trabalho comparou o desempenho por estação com o desempenho anual, des-

considerando as estações do ano e os resultados demonstraram que as correlações entre as características e a ocupação podem variar com base nas diferentes estações (de positivas para negativas, coeficientes de grandes para pequenos e vice-versa), o que significa que não existem variáveis ótimas fixas para prever o estado de ocupação em todas as estações. Entre as técnicas de inteligência artificial consideradas, o GBDT, o RF e a ANN apresentaram os melhores desempenhos, atingindo mais de 85% de acurácia no Verão, Outono e Inverno, e mais de 80% na Primavera.

2.3 Análise Crítica

Nesta secção, é realizada uma análise crítica aos diversos trabalhos relacionados com a aplicação de técnicas de inteligência artificial em dados ambientais para a previsão ou classificação de presença de pessoas em espaços fechados. Na Tabela 2.1 são analisadas um conjunto de métricas que se consideram importantes para o contexto da presente dissertação, com especial destaque para:

Trabalho: Identificação do trabalho com a sua referência bibliográfica.

Protocolos de Comunicação: Protocolos utilizados na troca de dados entre os diferentes componentes de *hardware* e *software*.

Tipo de Sensores: Indica se os sensores utilizados no trabalho foram selecionados e montados pelo próprio autor ou se foram adquiridos e implementados a partir de fontes pré-existentes.

Pré-Processamento de dados: Indica se existiu um pré-processamento dos dados e quais as técnicas utilizadas.

Divisão do conjunto de dados: Indica como foi feita a divisão dos dados.

Características utilizadas: Indica quais as características presentes nos conjuntos de dados.

Técnicas de Inteligência Artificial: Indica quais as técnicas de inteligência artificial utilizadas.

Métricas de Avaliação: Indica quais as métricas de avaliação utilizadas para avaliar o desempenho dos modelos.

Objetivo: Indica qual o objetivo do trabalho, se prever ocupação ou classificá-la.

Tabela 2.1: Comparação de Trabalhos: Protocolos, Sensores, Processamento de Dados, Modelos de IA e Objetivos Relacionados à Ocupação em Espaços Fechados

Trabalho	Protocolos de Comunicação	Tipo de Sensores	Pré-Processamento de dados	Divisão do conjunto de dados	Características utilizadas	Técnicas de Inteligência Artificial	Métricas de Avaliação	Objetivo
Bockstael and Jadin (2018)	Lora Wi-fi	Elsys ERS - Pré montado	Sim, sendo elas: extração de características, remoção de valores atípicos e balanceamento de classes	Escolhidos de forma aleatória: 90% para treino 10% para teste	CO2, Humidade, Temperatura, Luminosidade, Movimento, 1º Derivada CO2, 2º Derivada CO2, Número de Ocupantes	Modelo de voto composto por: *RF Classifier,* *RF Regressor,* *KNN Classifier,* KNN Regressor, *MLP Classifier,* *MLP Regressor*	Métricas de Acurácia: Acurácia Estrita, Acurácia com Limite, Acurácia com Limite Proporcional Métricas de erro: *MSE,* *MAD,* *R2 Score*	Classificar presença de forma exata
Zhang et al. (2023)	Wi-fi	Estação Metereológica Netatmo - Pré montada	Sim sendo elas: normalização e codificação distribuida	Não referido	CO2, Nível de Ruído, Pressão do Ar, Temperatura, Humidade, Número de Ocupantes	*LSTM* *Attention Mechanism*	*F1 Score*	Classificar presença de forma exata
Candanedo and Feldheim (2016)	Zigbee	DHT22, Telaire 6613 e TSL2561 - Escolhidos pelos autores	Não indicado	Foram utilizados 3 datasets: Treino, Teste 1, Teste 2	CO2, Temperatura, Humidade, Luuminosidade, Rácio de Humidade, Dia de Semana, Número de Segundos desde a Meia-Noite, Sala Ocupada	*CART,* *RF,* *GBM,* *LDA*	Acurácia	Classificar presença
Kim et al. (2023)	Manual	DHT22, CM1107, PM2008 - Escolhidos pelos autores	Não indicado	Não indicado	CO2, Temperatura, Humidade, Partículas, Número de Ocupantes	*MLR*	*MSE,* *RMSD,* *MAE*	Prever número de ocupantes
Mohammadabadi et al. (2022)	Não indicado	Modelos não indicados - Escolhidos pelos autores	Sim, amostragem estratificada no conjunto de teste	80% para treino, 20% para teste	CO2, Temperatura, Humidade, Variavel de Ocupação	*CNN + XgBoost,* *CNN,* *XgBoost,* *LR,* *DT,* *RF,* *GB,* *KMC,* *KNN,* *SVM*	*F1 Score,* *MAE*	Classificar ocupação
Tekler and Chong (2022)	Não indicado	Modelos não indicados - Escolhidos pelos autores	Sim, remoção ou substituição de dados em falta, extração e normalização de caracteristicas	Escolhidos de forma não aleatória: 80% para treino, 20% para teste	VOC, Nível do som, Humidade relativa, Temperatura, Luminosidade, PM2.5, CO2, Dispositivos ligados por wi-fi, Consumo de Energia, Operações de HVAC, CO2 Exterior, Humidade relativa exterior, Direção do vento, Velocidade do vento, Chuva, Radiação global solar, Pressão baromática, Temperatura exterior	*LSTM,* *BI-LSTM,* *GRU,* *BI-GRU* *DNN*	*MAE* *RMSE*	Prever número de ocupantes
Alsmirat et al. (2019)	Wi-fi	HC-SR501, BMP280, SGP30, HC1080, KY-038, SI1145 - Escolhidos pelos autores	Sim, sumarização dos dados.	80% para treino, 20% para teste	Temperatura, Humidade, Som, Luminosidade, Movimento, CO2, Variável de Ocupação	*SVM*	Precisão, *Recall,* *F1 Score*	Classificar Presença
Yang et al. (2021)	Não indicado	AMUN 716, EM/S 3.16.1, Varia, PlanoCentro A-KNX - Pré-montados	Sim, remoção de valores atípicos, processamento dos dados em falta e normalização	Criados conjuntos de treino e teste para cada estação do ano e anual.	Horário do dia, Dia da semana, Período do dia, Temperatura exterior, Humidade exterior, Irradiância solar, Velocidade do vento, Iluminação exterior, Chuva, Temperatura interior, Humidade interior, CO2 interior, Ponto de ajuste térmico, Luminosidade interior, Posição das persianas, Estado de bloqueio automático das janelas, Consumo de energia para iluminação, Consumo de energia das tomadas elétricas	*LR,* *SVM,* *DT,* *GBDT,* *RF,* *ANN*	*F1 Score,* *AUC*	Prever Ocupação

Nos trabalhos em que são abordados os protocolos de comunicação, observa-se uma clara preferência pelo Wi-Fi, possivelmente justificada pelo facto de, comummente, já existir uma infraestrutura implementada em ambientes académicos ou residenciais que suporta este protocolo. Destacam-se, igualmente, as suas vantagens em termos de velocidade e largura de banda. Contudo, nos casos em que os sensores operam com baterias, a opção pelo Wi-Fi pode não ser a mais adequada devido ao seu elevado consumo energético. Assim, em cenários em que a gestão eficiente do consumo de energia das baterias é primordial, é sensato considerar protocolos como o *Zigbee*, reconhecido pelo seu baixo consumo energético.

A seleção dos sensores constitui também uma divergência entre os autores dos trabalhos analisados. Alguns optam por soluções pré-fabricadas, tal como evidenciado no trabalho de Bockstael and Jadin (2018), que escolheu o sensor *Elsys ERS*, ou no estudo de Zhang et al. (2023), que utilizou uma estação meteorológica. Esta abordagem oferece vantagens, como a facilidade de iniciar a recolha de dados, uma vez que dispensa a montagem individual dos sensores e simplifica a transmissão dos dados recolhidos. No entanto, a seleção dos sensores em si, pode ser mais apropriada em determinados trabalhos, pois permite escolher sensores específicos alinhados com o contexto da investigação, conferindo liberdade para monitorizar apenas os dados pertinentes ao trabalho e possibilitando a adoção de soluções economicamente mais viáveis.

No âmbito das técnicas de pré-processamento de dados, observa-se uma clara prevalência na aplicação da técnica de normalização dos dados em vários trabalhos. A adoção desta técnica permite a uniformização dos dados para um formato comum, simplificando, desta forma, o processo de aprendizagem dos modelos. Outras técnicas de pré-processamento amplamente empregues incluem a remoção ou substituição de valores em falta ou atípicos, abordando possíveis falhas nos sensores ou lacunas na comunicação dos dados. Este procedimento é considerado crucial para assegurar a consistência dos dados fornecidos aos modelos, com o intuito de aprimorar a qualidade do processo de treino e, consequentemente, os resultados futuros.

Dado que o propósito subjacente aos trabalhos analisados é a aplicação de técnicas de inteligência artificial fundamentadas em dados recolhidos, a abordagem à divisão desses dados assume uma importância crucial e necessita de análise aprofundada. Nos estudos em que a divisão dos dados é mencionada, é unânime a opção pela separação entre conjuntos de treino e de teste. A divisão mais frequentemente adotada consiste na atribuição de 80% dos dados para treino e 20% para teste. Contudo, é igualmente relevante observar a aplicação da divisão de 90% para treino e 10% para teste. Importa ressaltar que não há uma percentagem universalmente correta de divisão, sendo imperativo analisar cada caso para determinar a divisão mais eficaz. Salienta-se que, durante a etapa de treino dos modelos, estes não têm acesso aos dados do conjunto de teste, garantindo, desta forma, uma avaliação imparcial do desempenho dos modelos. Em conjuntos de dados de menor di-

mensão, torna-se pertinente considerar a aplicação de técnicas de aumento de dados, tais como a replicação. Adicionalmente, em cenários em que se verifique uma disparidade na quantidade de dados entre diferentes classes, é recomendável proceder ao equilíbrio dos dados.

Embora as características ambientais dos espaços fechados desempenhem um papel significativo no que concerne à deteção de presenças, por vezes outras características também podem emergir dessas condições ou ser levadas em consideração. Alguns estudos optaram por calcular novas características a partir dos dados recolhidos, como derivadas do dióxido de carbono (CO_2) ou o rácio de humidade. Estas variáveis adicionais podem ser pertinentes para descobrir padrões nos dados que não sejam prontamente discerníveis no seu estado bruto. Contudo, existem outros fatores que podem indicar a presença num determinado espaço, como, por exemplo, o número de dispositivos conectados ao Wi-Fi. No entanto, é importante notar que este tipo de variável pode induzir a erros, dado que a mesma pessoa pode possuir vários dispositivos ligados ao Wi-Fi, o que poderia inflacionar a indicação de presença. A inclusão destas variáveis requer uma análise cuidadosa do contexto do trabalho. Dependendo do âmbito, pode ser relevante incorporar variáveis relacionadas com as condições ambientais exteriores. Por exemplo, num contexto de espaços públicos, onde o objetivo é prever a presença de pessoas, condições climatéricas como chuva podem influenciar a afluência ao espaço, devendo tal aspeto ser refletido nos resultados obtidos pelos modelos. No entanto, no contexto da deteção de presenças, variáveis como operações de HVAC podem não ser relevantes, dado que, em geral, não apresentam uma correlação direta com a presença de pessoas.

No que diz respeito às técnicas de inteligência artificial empregues, identificam-se duas abordagens distintas: a utilização de um modelo de voto composto por diversas técnicas ou a adoção de uma única técnica. A abordagem do modelo de voto apresenta vantagens, nomeadamente o aprimoramento da acurácia e robustez, uma vez que a combinação de várias técnicas permite mitigar as limitações inerentes a cada uma. Contudo, é relevante sublinhar que esta abordagem acarreta uma complexidade adicional na implementação dos modelos, sendo, até à data, aplicada em apenas um dos trabalhos analisados. No que concerne à abordagem mais tradicional, destaca-se a preferência por técnicas clássicas, tais como o RF ou o KNN, amplamente presentes em diversos trabalhos. A escolha por estas técnicas pode ser atribuída à sua facilidade de utilização e à flexibilidade que proporcionam. Na análise dos trabalhos, constata-se também a aplicação de técnicas focadas em séries temporais, como o *Long Short-Term Memory (LSTM)* ou *Bidirectional LSTM (BI-LSTM)*, o que se revela pertinente dada a natureza dos projetos em análise e a importância intrínseca destas técnicas. De igual modo, algumas investigações incorporaram técnicas de *deep learning*, tais como CNN, despertando certa curiosidade, uma vez que este tipo de técnicas é habitualmente direcionado para problemas que envolvem imagens ou áudio.

Um aspeto crucial a ter em consideração reside na avaliação dos resultados obtidos pelos modelos. Uma métrica comummente utilizada nos trabalhos analisados é o *F1 Score*, particularmente útil quando se enfrentam problemas de classificação desbalanceados e não se efetua um balanceamento do conjunto de dados. Neste tipo de cenário, a métrica de acurácia não é a mais adequada, uma vez que deve ser aplicada em situações onde é crucial evitar falsos positivos. Uma abordagem interessante no âmbito desta métrica é a utilização da acurácia com um limite, ou seja, a aplicação de uma margem de erro para a acurácia obtida. A pertinência desta abordagem depende do contexto do problema; por exemplo, pode não ser crucial classificar a presença de pessoas de forma exata. Outra métrica adotada em alguns dos trabalhos é o *Mean Absolute Error (MAE)*, que se destaca pela sua facilidade de compreensão e pela menor sensibilidade à presença de valores atípicos.

A abordagem relativa à deteção de presenças em espaços fechados pode seguir diversas vias. Alguns trabalhos concentram-se na previsão da ocupação, o que se revela relevante do ponto de vista do planeamento da utilização de salas de aula ou espaços públicos. Por outro lado, outros estudos concentram-se na classificação da ocupação do espaço, podendo ou não referir o número específico de pessoas. No entanto, é compreensível que a classificação da presença também seja de extrema importância, tal como a previsão. A classificação permite um planeamento mais eficiente da utilização do espaço, ao mesmo tempo que oferece uma visão precisa da sua ocupação. Isto é particularmente valioso em cenários educacionais, permitindo perceber se uma sala específica é adequada, em termos de capacidade, para determinados grupos de estudantes. A diversidade de objetivos encontrados nos trabalhos reflete a existência de várias opções no que diz respeito à deteção de presenças em espaços fechados. A seleção entre estas opções dependerá do contexto específico das necessidades em análise.

Capítulo 3

Especificação e Arquitetura do Sistema Airmon

3.1 Objetivos do sistema Airmon

O principal objetivo do sistema *Airmon* é, através da utilização de caixas sensoriais, recolher dados ambientais de diversas salas de aula, em diferentes escolas, agregando-os num só sistema. A informação recolhida permitirá construir vários conjuntos de dados que possibilitarão, através da utilização de algoritmos de inteligência artificial, classificar a ocupação de salas de aula. Estes dados poderão ser gerais (englobando informação de várias salas) ou restritos a cada sala monitorizada. Pretende-se também que o sistema *Airmon* seja de baixo custo mas com alta escalabilidade para suportar um grande número de salas de diferentes escolas, juntamente com tolerância a falhas e segurança dos dados.

Requisitos funcionais

Na Tabela 3.1, estão representados os requisitos funcionais (RF) do sistema. Na coluna Componente apresentam-se os dispositivos ou serviços do sistema, associado ao respetivo requisito. A Instância Central refere-se ao sistema agregador central que contém os dados de todas as escolas em monitorização, a Instância Local refere-se ao sistema agregador local que contém os dados de várias salas de uma escola e a Caixa de monitorização corresponde às caixas instaladas em cada uma das salas de aula para monitorizar os dados ambientais.

Tabela 3.1: Requisitos Funcionais (RF)

ID	Componente	Descrição
RF1.1	Instância Central	Deve ser capaz de receber os dados enviados pelas instâncias
RF1.2	Instância Central	Deve ser capaz de guardar os dados recebidos em base de dados
RF1.3	Instância Central	Deve permitir o registo de novas instâncias
RF1.4	Instância Central/Local	Deve ser capaz de efetuar cópias de segurança
RF1.5	Instância Central/Local	Deve ser capaz de demonstrar os dados presentes em formato de gráficos
RF1.6	Instância Central/Local	Deve ser capaz de comunicar de forma segura com as caixas de monitorização
RF1.7	Instância Central/Local	Deve ser capaz de permitir a extração dos dados em formato CSV
RF2.1	Instância Local	Deve ser capaz de receber dados dos micro-controladores
RF2.2	Instância Local	Deve ser capaz de guardar os dados dos micro-controladores em base de dados
RF2.3	Instância Local	Deve ser capaz de enviar os dados para a instância central
RF2.4	Instância Local	Deve permitir o registo de novas caixas de monitorização
RF3.1	Caixa de monitorização	Deve ser capaz de recolher informação ambiental das salas de aula
RF3.2	Caixa de monitorização	Deve ser capaz de enviar a informação recolhida para a instância local

Requisitos não funcionais

Na Tabela 3.2, representam-se os requisitos não funcionais (RNF) do sistema. Tal como na tabela anterior, a coluna Componente apresenta o dispositivo associado ao requisito descrito.

Tabela 3.2: Requisitos Não Funcionais (RNF)

ID	Componente	Descrição
RNF1.1	Caixa de monitorização	Deve ser de baixo custo
RNF1.2	Caixa de monitorização	Deve ter baixo consumo energético
RNF1.3	Caixa de monitorização	Deve ser de fácil instalação nas salas de aula
RNF1.4	Caixa de monitorização	Deve estar constantemente ligado a uma fonte de energia elétrica
RNF2.1	Instância Central/Local	Deve oferecer um bom mecanismo de segurança
RNF2.2	Instância Central/Local	Deve ser escalável

RNF2.3	Instância Central/Local	Deve oferecer mecanismos de tolerância a falhas
RNF2.4	Instância Central/Local	Deve oferecer interfaces intuitivas e fácil usabilidade

3.2 Visão geral da arquitetura

A arquitetura do sistema (Figura 3.1) é dividida em três camadas: caixas de monitorização instaladas em cada sala de aula, responsáveis por recolher os dados ambientais das mesmas, uma instância local em cada escola, responsável por agregar a informação das diversas salas e uma instância central que une toda a informação recolhida. As caixas de monitorização, possuem diversos sensores ambientais e são alimentadas através de tomadas presentes nas salas. Posteriormente, a informação é enviada através de um microcontrolador para a instância local, através do uso de wi-fi. A instância local recebe os dados de cada uma das caixas de monitorização e guarda-os numa base de dados local, permitindo também a sua consulta em tempo real. Além disso, é responsável por enviá-los para a instância central, que deverá receber e guardar os dados de todas as escolas monitorizadas, possibilitando assim, uma visão sobre os mesmos em tempo real.

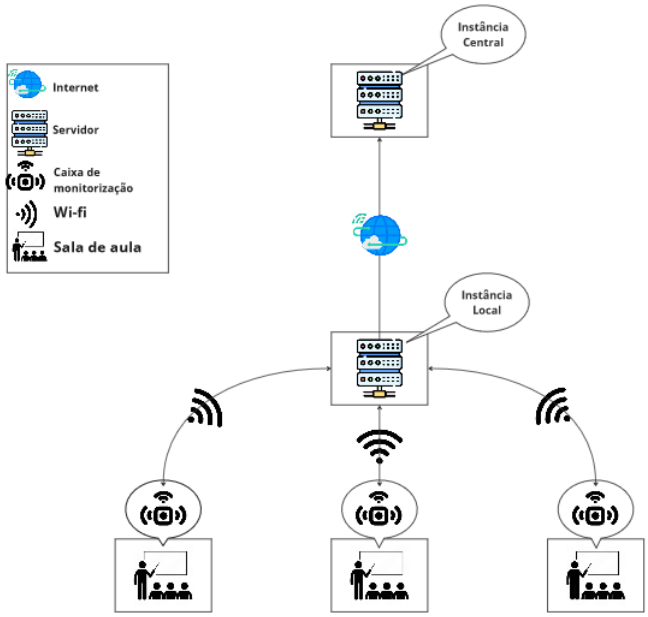


Figura 3.1: Arquitetura do sistema

3.3 Instâncias

Um dos objetivos desta dissertação, é definir um sistema que possa ser facilmente escalável e incluir a monitorização de dados ambientais de novas escolas. Para isso, foi pensada a existência de uma instância central que agrega a informação das mesmas, permitindo assim consultar e construir um conjunto de dados mais completo. Por outro lado, cada escola terá também a sua instância local que agregará apenas a sua informação, possibilitando a consulta dos dados e a construção do seu próprio conjunto de dados.

3.4 Sensorização no sistema Airmon

Os sensores representam um papel crucial no trabalho a desenvolver nesta dissertação, uma vez que, através deles, é possível captar informações precisas em tempo real sobre a componente ambiental de cada uma das salas de aula em monitorização.

3.4.1 Sensor de dióxido de carbono

O dióxido de carbono (CO_2) desempenha um papel fundamental na classificação de ocupação de uma sala de aula, pois os níveis de CO_2 estão diretamente relacionados à ocupação do espaço. A monitorização do CO_2 torna-se relevante para o contexto deste trabalho, uma vez que permite identificar a presença de pessoas na sala de aula com base nas variações dos níveis de CO_2. Espaços ocupados geralmente apresentam concentrações mais elevadas de CO_2, tornando este num indicador útil para a classificação de ocupação.

3.4.2 Sensor de humidade

Por sua vez, a humidade (referente à quantidade de vapor de água presente no ar) pode, em níveis demasiado baixos, contribuir para irritação nos olhos, garganta seca ou desconforto respiratório, enquanto que, níveis elevados, podem contribuir para o aparecimento de bactérias e condensação no espaço. Este é um elemento que é também importante monitorizar, pois influencia o conforto das pessoas dentro de uma sala podendo afetar diretamente a assiduidade da mesma.

3.4.3 Sensor de temperatura

A temperatura é também um fator essencial no bem-estar térmico em espaços fechados. Níveis inadequados de temperatura, podem contribuir para sensações de desconforto, fadiga e dificuldade de concentração. A monitorização deste elemento é significativa, pois em espaços fechados ocupados, a temperatura tende a aumentar.

3.4.4 Sensor de partículas

Um dos aspetos a ter em conta quando se fala acerca da ocupação de salas de aula, é a presença de partículas e as suas oscilações, que podem ser provenientes de diversas fontes, como a movimentação de pessoas na sala de aula, a libertação de partículas no ar ao respirar, falar ou tossir, a circulação de ar feita pelos sistemas de ventilação pode redistribuir as partículas presentes no ambiente, entre outros. Recolher informação acerca das mesmas, poderá contribuir positivamente na classificação de ocupação.

3.5 Fluxo dos dados

Cada sensor recolhe informação ao longo do tempo, enviando-a para o microcontrolador, presente em cada caixa de monitorização. Por sua vez, o microcontrolador, terá a responsabilidade de comunicar os dados recebidos para a instância local da escola. Os dados enviados devem estar identificados com o nome da sala, de forma a ser possível saber a qual pertencem. A instância local armazenará os dados numa base de dados, para que seja possível construir um registo com toda a informação recolhida ao longo do tempo. Além disso, os dados deverão ser enviados, periodicamente, para a instância central, identificando não só a sala, como a escola a que pertencem. Nas instâncias, quer local, quer central, é possível descarregar os dados e, a partir dos seus ficheiros, construir um conjunto de dados.

3.6 Escalabilidade, desempenho , segurança , tolerância a falhas e recuperação

O sistema proposto deverá também ser escalável, seguro, com bom desempenho e tolerante a falhas. Em termos de escalabilidade, deverá suportar diversas escolas, em que cada uma poderá monitorizar diversas salas de aula, não sacrificando o bom desempenho do sistema. Toda a comunicação entre os componentes deve ser feita de forma segura, recorrendo a técnicas de criptografia. Cada uma das instâncias deverá implementar mecanismos que permitam efetuar cópias de segurança do sistema e dos seus dados, contribuindo para que o sistema seja tolerante a falhas e, ao mesmo tempo, seja possível recuperar toda a informação partindo de um espaço temporal.

Capítulo 4

Implementação do sistema Airmon

Neste capitulo, abordar-se-á a implementação do sistema *Airmon* representado na Figura 4.1, descrevendo todo o processo de integração entre os vários componentes de *hardware*, presentes nas caixas de monitorização, e ainda, os componentes de *software*. Com isto, visa-se justificar as escolhas realizadas ao longo da implementação, relativamente às tecnologias utilizadas, sendo também apresentadas algumas das limitações encontradas e o modo como foram ultrapassadas.

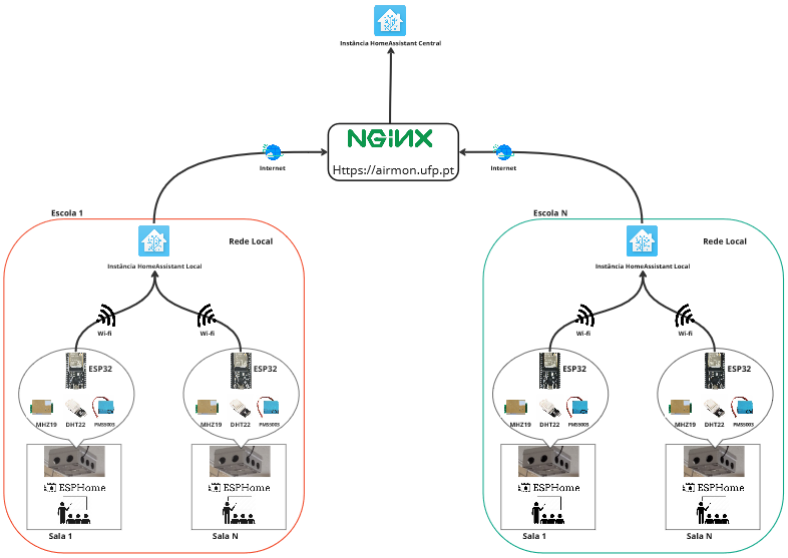


Figura 4.1: Arquitetura da Implementação do Sistema *Airmon*

4.1 HomeAssistant

O HA é uma plataforma de código aberto com foco na automação residencial, projetado para controlar e gerir diversos dispositivos inteligentes numa casa. Possui suporte nativo

para diversos sistemas operativos em diferentes plataformas, podendo ser instalado e executado em *linux*, *windows*, entre outras opções. Tem como premissa a privacidade dos dados, pelo que foi desenhado para que toda a informação seja mantida localmente no dispositivo em que é executado sem envio de informação para a *cloud* (HomeAssistant, 2023).

Apesar do HA ter como foco a automação residencial, diversos fatores levaram a que este fosse escolhido como sistema base para o projeto *Airmon*. Em projetos anteriores, o autor da presente dissertação, teve a oportunidade de explorar e trabalhar com o HA, ganhando conhecimento acerca do seu uso e potencialidades de personalização. Outro fator importante a ter em conta, é a escalabilidade nativa do HA uma vez que foi desenhado para receber e comunicar com um grande número de dispositivos em simultâneo, oferecendo, ao mesmo tempo, recursos avançados de segurança como comunicação criptografada.

4.1.1 Integrações HomeAssistant

As integrações são pacotes pré-configurados que podem ser facilmente instalados e integrados no HA, fornecendo funcionalidades adicionais e expandindo a capacidade do sistema base. De forma a tirar proveito desta potencialidade, foram utilizadas diversas integrações para melhorar a segurança e robustez das instâncias, permitindo a fácil configuração e manutenção das caixas de monitorização e melhorar a forma de visualizar toda a informação recolhida ao longo do tempo. As próximas subsecções visam explicar quais integrações foram escolhidas e o seu papel no sistema *Airmon*.

4.1.1.1 EspHome

EspHome é uma plataforma de código aberto projetada para facilitar a integração e controlo de microcontroladores ESP32 e ESP8266. Uma das principais vantagens, é facilitar a configuração dos microcontroladores através de ficheiros YAML, permitindo assim, com poucas linhas de código, configurar e personalizar o comportamento de diversos sensores, abstraindo a necessidade de programar os mesmos de raiz. Além disto, salienta-se a possibilidade de realizar atualizações *over-the-air* (OTA), facilitando a manutenção e atualização das caixas de monitorização, uma vez que não é necessário retirá-las das salas de aula para efetuar qualquer alteração nas mesmas (EspHome, 2023).

Todas as caixas de monitorização foram configuradas com recurso à utilização do *EspHome*. Na Figura 4.2, é possível visualizar um exemplo de configuração para uma das caixas de monitorização, em que é indicado o tipo de sensor, o nome que será utilizado para criar uma entidade na instância HA, assim como os pinos a serem utilizados no microcontrolador e os tempos de recolha de dados para cada um dos sensores. A primeira configuração do microcontrolador é necessária ser realizada via cabo, contudo, após esse processo, todas as alterações podem ser feitas *over-the-air*.

```yaml
sensor:
  - platform: pmsx003
    type: PMSX003
    pm_1_0:
      name: "UFP SALA 204 Particulate Matter <1.0µm Concentration"
    pm_2_5:
      name: "UFP SALA 204 Particulate Matter <2.5µm Concentration"
    pm_10_0:
      name: "UFP SALA 204 Particulate Matter <10.0µm Concentration"
    update_interval: 300000ms
    uart_id: pmx

  - platform: dht
    pin: GPIO32
    temperature:
      name: "UFP SALA 204 DHT Temperature"
    humidity:
      name: "UFP SALA 204 DHT Humidity"
    model: AM2302
    update_interval: 300s

  - platform: mhz19
    co2:
      name: "UFP SALA 204 CO2 Value"
    temperature:
      name: "UFP SALA 204 CO2 Temperature"
    update_interval: 300s
    automatic_baseline_calibration: false
    id: mhzsensor_ufp_sala_204
    uart_id: uartmhz
```

Figura 4.2: *EspHome*

4.1.1.2 InfluxDB

Um dos principais objetivos desta dissertação é a recolha de dados ao longo do tempo, por isso, as séries temporais possuem uma grande importância. A integração *InfluxDB*, instala na instância HA, uma base de dados local projetada para armazenar e consultar dados de séries temporais. A utilização do *InfluxDB*, é devido à sua eficiência e escalabilidade na manipulação de grandes volumes de dados, que foi um requisito importante dada a quantidade de informação que se pretende agregar na realização deste trabalho (InfluxDB, 2023).

4.1.1.3 Grafana

Recorreu-se, ainda, à integração *Grafana*, que visa facilitar a visualização dos dados recolhidos de forma fácil e intuitiva. Permite realizar diversas pesquisas com base em filtros temporais, assim como a consulta dos dados recolhidos em tempo real. Na Figura 4.3 é possível visualizar os diversos dados ambientais recolhidos durante trinta dias, das três salas monitorizadas na Universidade Fernando Pessoa no presente estudo (Grafana, 2023).

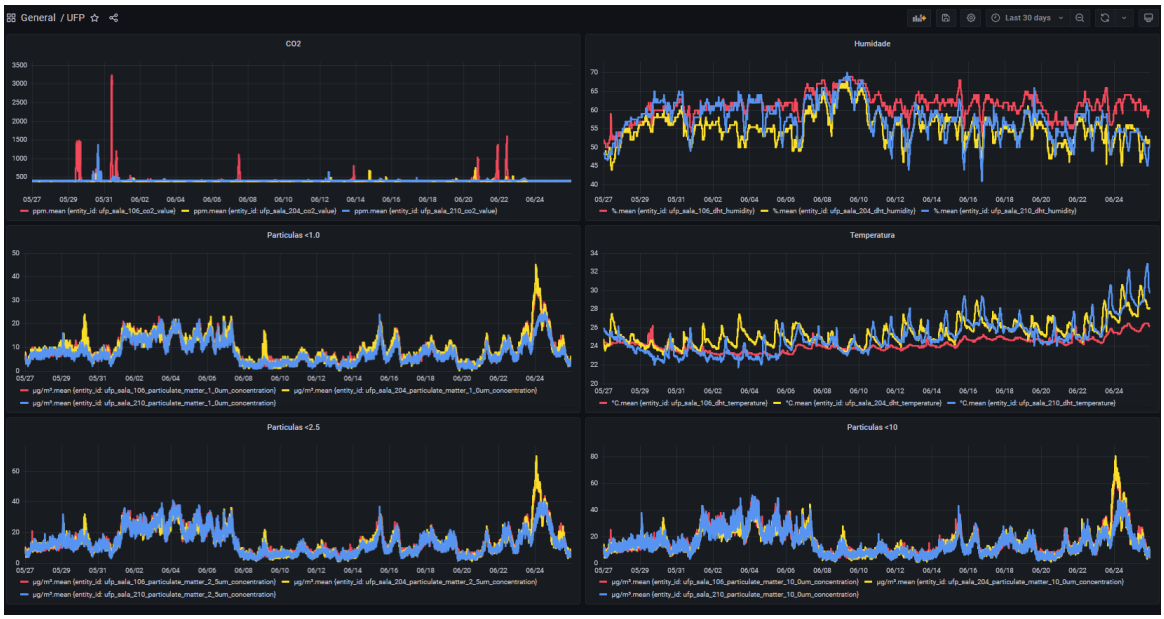


Figura 4.3: *Grafana*

4.1.1.4 Google Drive Backup

Como referido anteriormente, pretende-se recolher informação ao longo do tempo, pelo que, é de extrema importância garantir que estes não são perdidos em circunstâncias como a danificação do servidor que aloja a instância do HA. Para isso, foi utilizada a integração *Google Drive Backup* que permite realizar cópias de segurança de todo o sistema e guardá-las na *drive* da *Google*. Estas cópias são feitas de forma automática, diariamente, em que é guardada uma cópia integral de toda a instância, permitindo assim a recuperação de qualquer informação ou nova instanciação facilmente. Como cada instância presente em cada uma das escolas monitorizadas, envia os dados ao longo do tempo para uma instância central, estes encontram-se guardados em dois servidores, garantindo maior segurança na preservação dos mesmos (Beechen, 2023).

4.1.1.5 NGINX

Esta integração apenas está presente na instância central, uma vez que é a única que se encontra exposta na Internet. Foi instalado para atuar como um intermediário entre os clientes e o HA central, adicionando assim uma camada de segurança à instância, uma vez que garante a utilização de criptografia SSL/TLS para proteger as comunicações entre cliente e servidor. Como se efetua recolha de dados, é essencial garantir a sua segurança quando os mesmos são enviados pela Internet (Nginx, 2023).

4.2 Caixas de Monitorização

Nesta subsecção aborda-se o microcontrolador e os diversos sensores utilizados na montagem das caixas de monitorização, indicando os seus modelos e as suas características, sendo que alguns destes foram utilizados em projetos anteriores e outros adquiridos em *websites* de revendedores especializados.

4.2.1 Micro-controlador

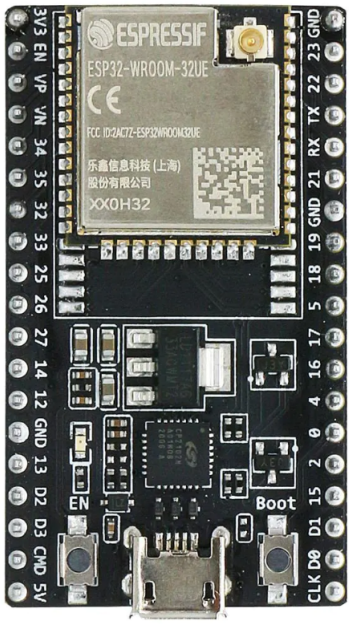


Figura 4.4: Micro-controlador - ESP32-DevKitC-32UE

O microcontrolador possui um papel fundamental na recolha e envio dos dados dos diversos sensores presentes nas caixas de monitorização, por isso, foi escolhido o modelo ESP32-DevKitC-32UE (Figura 4.4), que possui poder de processamento, flexibilidade e uma ampla conectividade, de forma a agregar os diversos sensores. Este modelo em concreto, possui ainda uma ficha do tipo U.FL que permite ligar uma antena externa de forma a melhorar a capacidade de sinal wi-fi do microcontrolador (Electronics, 2022).

4.2.2 Sensor de dióxido de carbono

Figura 4.5: Sensor de CO_2 - MH-Z19

O sensor MH-Z19 (Figura 4.5), utiliza o princípio de deteção não dispersiva de infravermelho (NDIR) para medir os níveis de CO_2, emitindo uma luz infravermelha com uma determinada frequência e, em seguida, mede a quantidade de luz absorvida pelo CO_2 presente no ar. Com base nessa absorção de luz, o sensor determina a concentração de CO_2 presente no espaço. Este modelo é tipicamente utilizado em projetos de monitorização da qualidade do ar em espaços fechados, assim como em projetos de casas inteligentes, devido à sua baixa taxa de erro e longo período de vida (Winsen, 2022). Na Tabela 4.1 estão presentes informações acerca das dimensões, intervalo de medição, margem de erro e período de vida do sensor MH-Z19.

Tabela 4.1: Informações acerca do sensor - MH-Z19

Dimensões (C x L x A)	**Intervalo de Medição (ppm)**	**Margem de Erro**	**Período de Vida (anos)**
33 mm×20 mm×9 mm	0 a 5000	± (50ppm+3% valor lido)	5

4.2.3 Sensor de Temperatura e Humidade

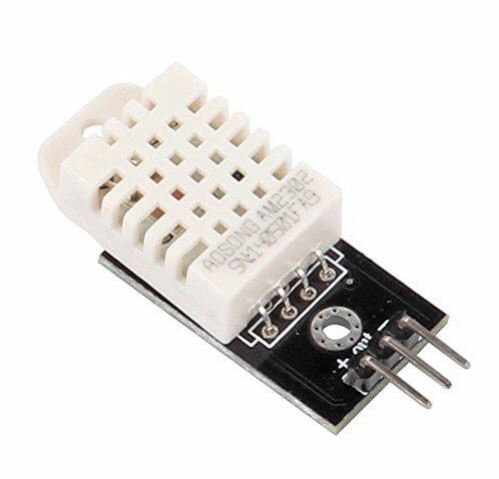


Figura 4.6: Sensor de Temperatura e Humidade - DHT22

O sensor DHT22 (Figura 4.6), também conhecido como AM2302, é um sensor de temperatura e humidade utilizado para medir estes dados ambientais com precisão, possuindo dimensões pequenas e apresentando um baixo consumo energético, sendo amplamente utilizado em projetos de sistemas de automação ou estações meteorológicas (Adafruit, 2022). Na Tabela 4.2 estão presentes informações acerca das dimensões, intervalo de medição, margem de erro e período de vida do sensor DHT22.

Tabela 4.2: Informações acerca do sensor - DHT22

Dimensões (C x L x A)	**Intervalo de Medição**	**Margem de Erro**	**Período de Vida (anos)**
15.3 mm×7.8 mm×25.3 mm	-40 a 80 ºC 0-100% HR	± 0.5 ºC ± 2% HR	2

4.2.4 Sensor de Partículas

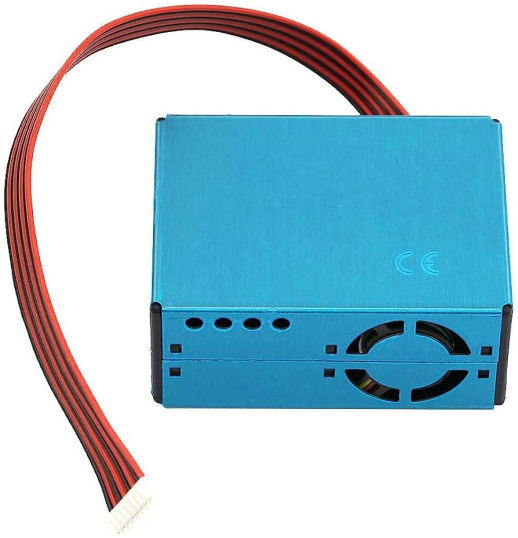


Figura 4.7: Sensor de Partículas - PMS5003

O sensor PMS5003 (Figura 4.7) é um sensor de partículas finas que mede a concentração de partículas suspensas no ar, utilizando um sistema ótico de dispersão a laser para detetar e contar as mesmas. O sensor emite um feixe de laser e mede a quantidade de luz dispersa pelas partículas e, com base nessa dispersão, o sensor determina a concentração de partículas. É capaz de medir partículas com um diâmetro de 1, 2.5 e 10 micrómetros, permitindo assim medir partículas de pó, pólen, fumo, entre outras, que podem ser prejudiciais à saúde humana, fazendo com que este sensor seja comummente utilizado em projetos de monitorização da qualidade do ar (Plantower, 2022). Informações como dimensões, intervalo de medição, margem de erro e período de vida do sensor PMS5003 são apresentadas na Tabela 4.3.

Tabela 4.3: Informações acerca do sensor - PMS5003

Dimensões (C x L x A)	**Intervalo de Medição ($\mu g/m^3$)**	**Margem de Erro**	**Período de Vida (anos)**
50 mm×38 mm×21 mm	0 a 500	±10%	3

4.2.5 Validação das caixas de monitorização

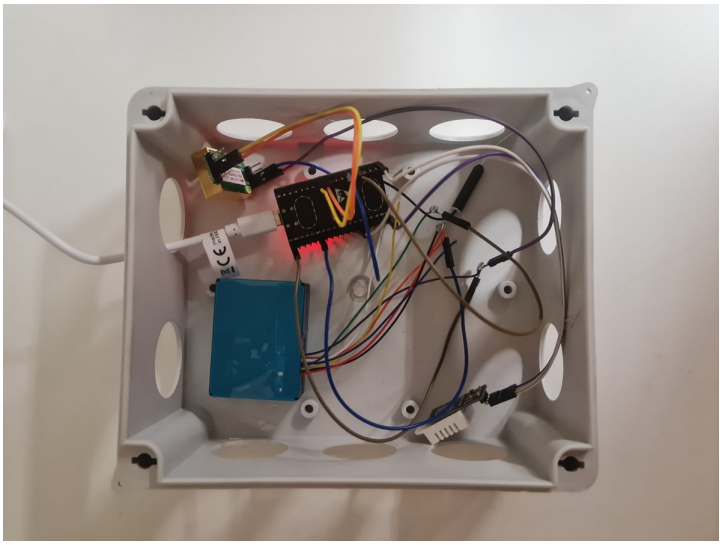


Figura 4.8: Interior Caixa de Monitorização

Na Figura 4.8 é possível observar o interior de uma das caixas de monitorização. O modelo escolhido possui dez entradas nas suas laterais que podem impactar na capacidade de leituras das caixas. Para averiguar esse impacto, um dos testes efetuados no período de validação, foi o de obter leituras com todas as entradas fechadas (incluindo a tampa), com a exceção de uma, para a entrada do cabo de alimentação do microcontrolador. Com isto, verificou-se que esta configuração afetava o fluxo de ar dentro da caixa, não permitindo assim uma leitura correta dos valores ambientais e contribuindo para a formação de microclimas dentro da mesma. Outro teste realizado, foi o de abrir uma entrada adicional e acoplar uma ventoinha de 5v, de forma a puxar o ar para o interior da caixa, contudo, foi verificado que os valores de CO_2 estabilizavam em torno dos 400ppm, quando em comparação com outras caixas que, sem a ventoinha, apresentavam valores superiores. A configuração escolhida para as caixas foi a de todas as entradas laterais estarem abertas e com a tampa colocada, sem ventoinha acoplada nas entradas. Esta configuração permite um fluxo de ar no interior da caixa sem que microclimas sejam formados no interior das mesmas.

Figura 4.9: Caixa de Monitorização em sala de aula

De forma a garantir uma monitorização eficiente, a caixa foi cuidadosamente instalada na sala de aula, seguindo algumas diretrizes importantes. Primeiramente, a caixa foi montada a cerca de 50cm de altura do teto, utilizando um suporte adequado conforme demonstrado na Figura 4.9. Esta posição elevada foi escolhida estrategicamente para evitar que a caixa fosse facilmente alcançada pelas pessoas presentes na sala, prevenindo assim qualquer interferência indesejada no decorrer das medições. Foi também decidido que a caixa deveria ser montada próximo ao centro da sala, proporcionando uma representação mais equilibrada das condições ambientais presentes no espaço.

4.3 Fluxo dos dados

No seguimento da informação apresentada anteriormente, revela-se importante também descrever o fluxo dos dados desde os componentes de *hardware* até à formação dos ficheiros que compõem os conjuntos de dados a serem utilizados mais tarde nos algoritmos de inteligência artificial.

4.3.1 Caixa de Monitorização para Instância Local

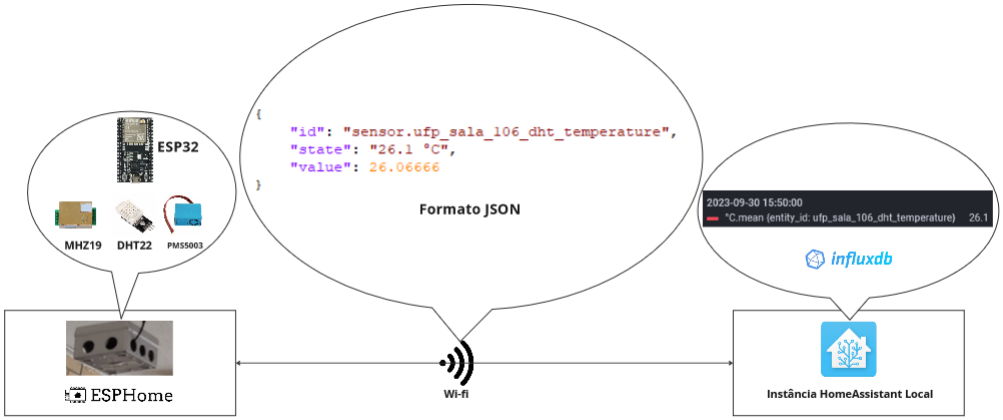


Figura 4.10: Comunicação da Caixa de Monitorização para Instância Local

De forma a que as caixas de monitorização de informação ambiental fossem facilmente compatíveis com a maioria das escolas no território nacional, estas foram desenhadas para serem alimentadas via tomada elétrica, geralmente presente em todas as salas de aula e eliminando, assim, a necessidade de carregar baterias, enquanto se garante perto de 100% de tempo de funcionamento. Outro aspeto importante é a forma como comunicam com a instância local da escola, demonstrado na Figura 4.10, sendo que esta comunicação é feita via wi-fi, uma vez que a infraestrutura já se encontra montada em grande parte das escolas nacionais e não requer investimento extra para a instalação das caixas. As ESP32 presentes nas caixas de monitorização, recolhem a informação dos sensores de cinco em cinco minutos e enviam toda essa informação para a instância local que os guarda na sua base de dados local. Todos os dados recolhidos podem ser consultados em tempo real. Na Figura 4.11 encontra-se um exemplo de uma informação acerca das partículas, enviada pela ESP32, presente numa das salas monitorizadas, sendo possível observar o último valor obtido, há quanto tempo foi enviado e um gráfico acerca do mesmo nas últimas vinte e quatro horas.

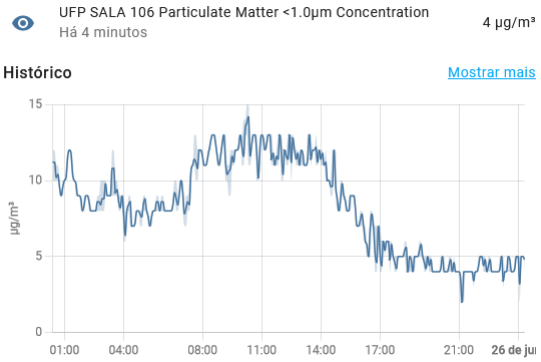


Figura 4.11: Exemplo de informação recebida pela instância local no HA

4.3.2 Instância Local para Instância Central

Uma vez que o HA foi desenvolvido para correr apenas localmente, foi necessário explorar diversas opções para permitir que as instâncias comunicassem entre si. Uma das opções consideradas foi a utilização da tecnologia *Message Queuing Telemetry Transport* (MQTT) em que mensagens seriam enviadas para o *broker* da instância central, contudo, foi validado que este método era complexo de se configurar para o número de sensores envolvidos e não escalava tão bem quanto outras opções disponíveis.

Outro dos testes realizados foi através do uso da integração *Remote HomeAssistant* que permite que duas instâncias comuniquem entre si remotamente. No entanto, revelou-se necessário que todas as instâncias estivessem expostas para a Internet, uma vez que estão presentes em diversas escolas. Esta abordagem tem como limitação as altas regras de segurança presentes nas escolas no que toca à política de abertura de portas de encaminhamento e que impedem a fácil abertura e configuração necessárias para a utilização desta integração.

A abordagem então utilizada foi a de instanciar um HA Central, disponível através do endereço *https://airmon.ufp.pt*, em que a comunicação é apenas de uma via, isto é, apenas as instâncias locais conseguem comunicar com a instância central. O processo de comunicação é realizado através da configuração de uma automação em cada uma das instâncias presentes nas escolas, que de 3 em 3 minutos, recolhe, dinamicamente, com recurso à utilização de um *script*, todos os dados das entidades da integração *EspHome*, isto é, todos os sensores de todas as caixas de monitorização presentes nas salas de aulas e envia-os através do serviço *rest command* para a *application programming interface (API)* exposta da instância central. Cada escola possui um *token* de autenticação válido por 10 anos, permitindo a comunicação entre instâncias durante um longo período de tempo. De forma a organizar os dados recebidos, foi definido ainda um padrão para o nome a ser dado a cada sensor, escolhendo-se a nomenclatura "Nome da Escola-Nome da Sala-Tipo de Sensor", tal como exemplificado na Figura 4.11. A Figura 4.12 demonstra

o fluxo no envio dos dados para a instância central. Todo este processo necessita de ser configurado apenas inicialmente na configuração da instância da escola, podendo ser adicionados novos sensores, sem a necessidade de configurar o envio da informação dos mesmos, uma vez que o sistema tem a capacidade de os incluir e enviar para a instância central automaticamente. A instância central, ao receber dados de um novo sensor cria, de forma automática, os seus registos e guarda em base de dados, sendo que todo este mecanismo é garantido pelas funcionalidades presentes no HA e permite uma fácil adição de novos sensores e também de novas escolas. Na Figura 4.13 encontra-se um exemplo das várias entidades envolvidas no envio dos dados entre a Instância Local e a Instância Central, podendo-se observar o tipo de informação que é trocada, o seu formato e como a mesma fica guardada em base de dados.

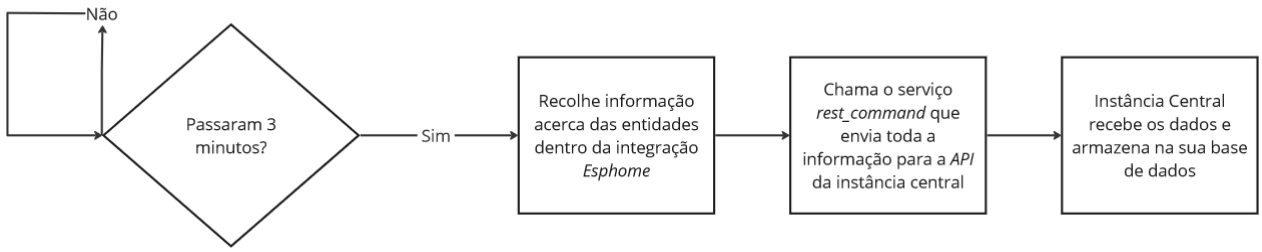


Figura 4.12: Fluxograma da Automação do Envio dos Dados

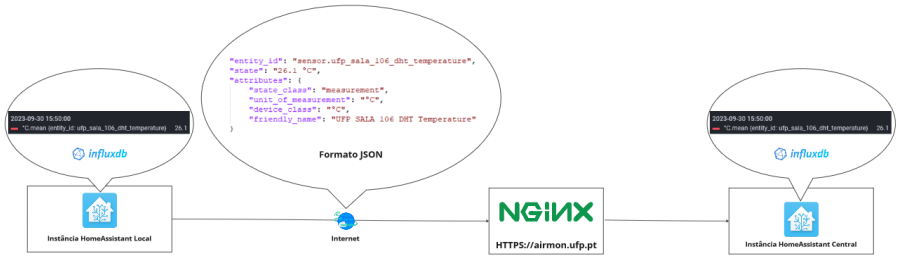


Figura 4.13: Comunicação da Instância Local para Instância Central

4.3.3 Extração dos dados recolhidos

Com recurso à integração *Grafana*, é possível extrair toda a informação recolhida pelos diversos sensores para um ficheiro no formato CSV, pelo que, para isso, é necessário selecionar o período temporal a ter em conta e os diversos sensores. A Figura 4.14 demonstra a extração dos três sensores de CO_2 presentes em salas da Universidade Fernando Pessoa para um único ficheiro CSV, com dados de 30 dias. Estes ficheiros serão mais tarde processados pelos *scripts* desenvolvidos em *Python* para a elaboração de conjuntos de dados. Todo este processo será abordado e explicado em pormenor no próximo capítulo.

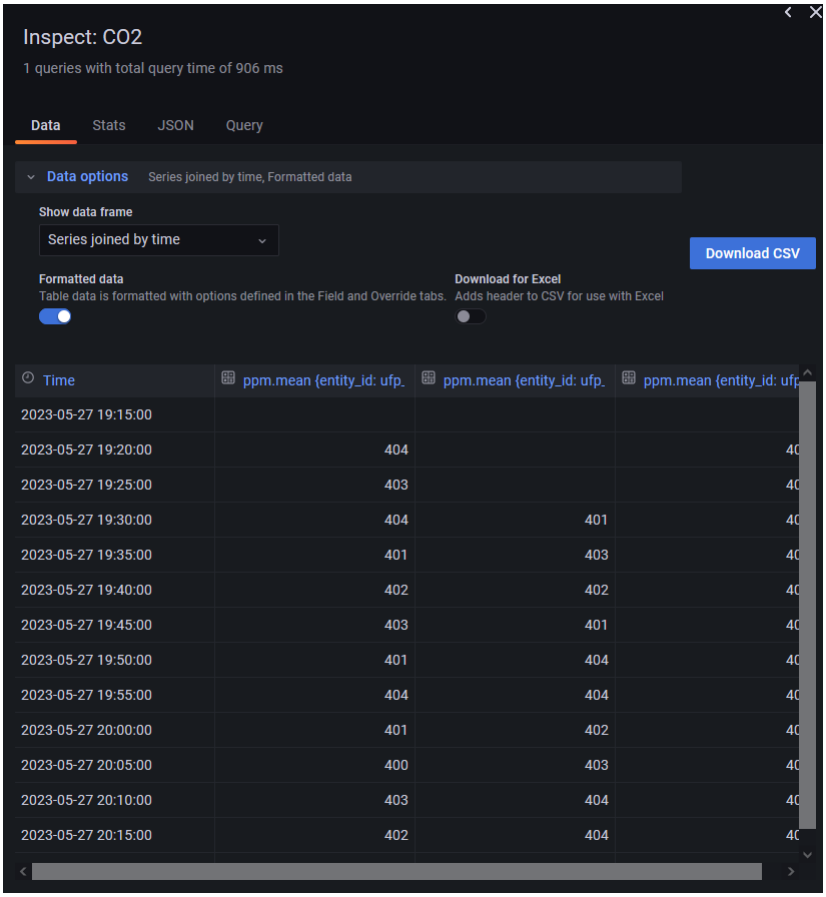


Figura 4.14: Processo de extração dos dados para ficheiro CSV

4.4 Conclusão

Neste capítulo foi descrito todo o processo da implementação do sistema *Airmon*. Inicialmente foi necessário perceber a viabilidade da utilização de um sistema como o HA para o contexto e caso de uso deste trabalho, uma vez que o seu foco é a automação residencial. Decidiu-se utilizar este sistema derivado à sua grande versatilidade, escalabilidade e potencialidade em encurtar o tempo necessário, de forma a ter uma plataforma pronta a recolher dados. Com a possibilidade de utilizar integrações como o *EspHome*, o tempo necessário de codificar os microcontroladores para recolha de dados foram mínimos, uma vez que toda a configuração foi feita com recurso à integração. Importa também reforçar a disponibilização de atualizações *over-the-air*, que permite realizar tudo à distância sem necessidade de deslocação até às caixas e de desmontar as mesmas. Foi também fulcral garantir que todos os modelos de sensores escolhidos eram compatíveis com a integração e possíveis de montar simultaneamente no mesmo microcontrolador.

Numa fase inicial, o sistema foi testado em ambiente fechado com apenas uma instância e uma caixa de monitorização, para verificar o adequado funcionamento dos componentes envolvidos e a comunicação entre si. Numa segunda fase, foram adicionadas mais

duas caixas de monitorização no mesmo espaço físico da primeira fase. Esta etapa permitiu efetuar uma comparação entre os dados recolhidos de cada uma das caixas, de forma a analisar a sua consistência e confirmar se os resultados obtidos eram semelhantes para as mesmas condições ambientais. No caso dos dados de CO_2, estes foram ainda comparados com leituras recolhidas por uma sonda de referência no mesmo espaço físico, para avaliar a fiabilidade dos mesmos. Foi ainda realizada uma análise estatística dos dados para avaliar a sua consistência. Com base no sucesso dos primeiros dois testes efetuados, foi criada uma *Virtual Machine* (VM) no servidor interno da Universidade Fernando Pessoa, para instanciar o servidor central responsável por agregar os dados de diversas escolas. Foi também criada uma instância local com recurso a uma *Raspberry Pi* para recolher os dados das caixas instaladas na Universidade Fernando Pessoa e foram, ainda, instaladas as caixas de monitorização em três salas da Universidade, cada uma com as suas próprias características, como tamanho, número de janelas e condições ambientais especificas, nomeadamente, a exposição solar, índices de humidade, entre outras. Esta abordagem progressiva permitiu uma validação robusta e a obtenção de dados confiáveis para utilização futura em algoritmos de inteligência artificial.

Após uma instalação bem-sucedida na Universidade Fernando Pessoa, foi também feita uma instalação na Escola Secundária de Alpendorada. O principal objetivo desta instalação foi validar a escalabilidade e fiabilidade da arquitetura e implementação realizada no projeto, permitindo assim recolher informação de duas escolas em simultâneo. Esta instalação foi efetuada com recurso a uma *Raspberry Pi*, que ficou responsável por correr a instância HA local. Contudo, foram apresentados alguns obstáculos, como a comunicação dos dados recolhidos pela Escola Secundária de Alpendorada para a instância central. Inicialmente, o problema estava relacionado com a utilização de um certificado não emitido pela empresa que registou o domínio utilizado, sendo que este problema foi solucionado, com a configuração do certificado correto, permitindo assim que o mesmo fosse registado como de confiança na rede presente na Escola Secundária de Alpendorada.

Capítulo 5

Conjunto de dados

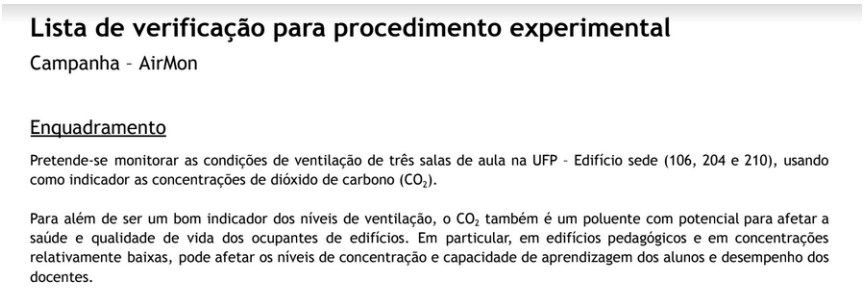
Com o propósito de cumprir o objetivo de utilizar algoritmos de inteligência artificial para a classificação da ocupação em salas de aula, torna-se imperativo a construção e utilização de um conjunto de dados fidedigno, que garanta a obtenção de resultados confiáveis. Neste capítulo, serão delineados os tipos de dados que constituem o conjunto, o procedimento de recolha de dados não ambientais, as técnicas de pré-processamento e limpeza de dados aplicadas, bem como uma análise detalhada do conjunto de dados.

5.1 Procedimento de recolha dos dados

No capítulo anterior, foram pormenorizadamente descritas as metodologias seguidas para a aquisição dos dados ambientais de cada sala de aula, utilizando caixas de monitorização. Além dos dados ambientais, foi delineado um procedimento que, tanto docentes quanto alunos, deveriam seguir durante a ocupação das salas monitorizadas. Este protocolo teve como objetivo principal garantir o controlo da componente ambiental de cada sala sujeita a monitorização. Numa primeira fase, o protocolo foi estritamente aplicado, estabelecendo diretrizes rígidas para as condições ambientais durante o período das aulas. Posteriormente, numa segunda fase, as regras referentes ao ambiente em sala de aula foram flexibilizadas. Permitindo que os docentes tivessem a opção de manter as portas e janelas abertas ou fechadas durante as aulas, com o propósito de aumentar a variabilidade nos dados recolhidos. O protocolo foi afixado em todas as salas de aula monitorizadas e compreendia três etapas distintas:

Enquadramento:

A primeira etapa trata-se do enquadramento do projeto (Figura 5.1) que visa elucidar o docente sobre o propósito do mesmo, os respetivos objetivos e relevância no contexto da sala de aula.



Lista de verificação para procedimento experimental

Campanha - AirMon

Enquadramento

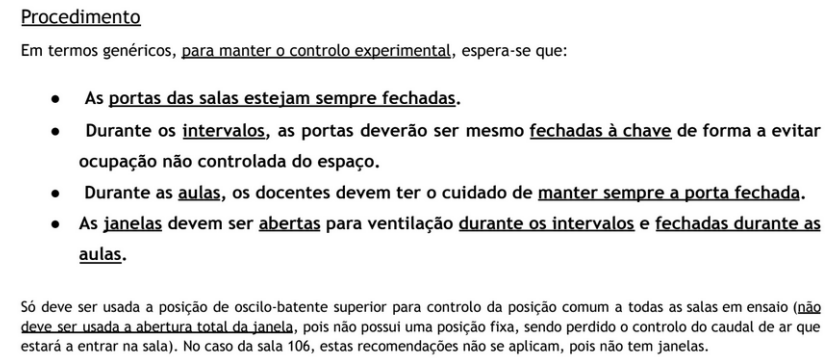
Pretende-se monitorar as condições de ventilação de três salas de aula na UFP - Edifício sede (106, 204 e 210), usando como indicador as concentrações de dióxido de carbono (CO_2).

Para além de ser um bom indicador dos níveis de ventilação, o CO_2 também é um poluente com potencial para afetar a saúde e qualidade de vida dos ocupantes de edifícios. Em particular, em edifícios pedagógicos e em concentrações relativamente baixas, pode afetar os níveis de concentração e capacidade de aprendizagem dos alunos e desempenho dos docentes.

Figura 5.1: Enquadramento

Procedimento:

A segunda fase aborda o procedimento (Figura 5.2), o qual instruí o docente sobre as normas a seguir para manter o controlo. Para isso, foi solicitado aos docentes que mantivessem as portas da sala de aula sempre fechadas. Durante os intervalos entre as aulas, deveriam trancá-las, de modo a evitar a ocupação do espaço fora do período letivo. Além disso, foi indicado que as janelas das salas deveriam estar abertas durante os intervalos, mas fechadas durante as aulas.



Procedimento

Em termos genéricos, para manter o controlo experimental, espera-se que:

- As portas das salas estejam sempre fechadas.
- Durante os intervalos, as portas deverão ser mesmo fechadas à chave de forma a evitar ocupação não controlada do espaço.
- Durante as aulas, os docentes devem ter o cuidado de manter sempre a porta fechada.
- As janelas devem ser abertas para ventilação durante os intervalos e fechadas durante as aulas.

Só deve ser usada a posição de oscilo-batente superior para controlo da posição comum a todas as salas em ensaio (não deve ser usada a abertura total da janela, pois não possui uma posição fixa, sendo perdido o controlo do caudal de ar que estará a entrar na sala). No caso da sala 106, estas recomendações não se aplicam, pois não tem janelas.

Figura 5.2: Procedimento

Lista de Verificação:

Na terceira fase (Figura 5.3), encontra-se a lista de verificação, que orienta o docente a seguir um conjunto de procedimentos no final de cada aula. Nomeadamente, deixar as janelas abertas utilizando a abertura oscilo-batente superior e trancar a sala, devolvendo a chave ao contínuo. Adicionalmente, através de um código QR, o docente era direcionado a preencher um formulário que fornecia informações referentes à aula, tais como o número de pessoas presentes, a sala onde decorreu a aula e o horário da mesma.

Figura 5.3: Lista de Verificação

Questionário:

Como parte do procedimento de recolha de dados, cada docente, no final de cada aula, completou um questionário composto por dez perguntas (conforme apresentado na Tabela 5.1). Este questionário angariou aproximadamente duzentas e oito respostas, das quais foram criadas variáveis para serem integradas no conjunto de dados. Importa salientar que, através do preenchimento deste questionário, foi possível efetuar um registo não intrusivo das presenças, respeitando, assim, a privacidade dos ocupantes nas distintas salas de aula.

Tabela 5.1: Perguntas presentes no questionário

Pergunta
Em que sala vai lecionar?
Ao chegar à sala, a porta estava devidamente fechada? (não aplicável à sala 106)
Ao chegar à sala, as janelas estavam devidamente abertas (não aplicável à sala 106)?
A que horas iniciou a aula?
A que horas terminou a aula?
Durante o período da aula, a porta esteve?
Durante o período da aula, a janela esteve?
Durante a aula, o ar condicionado esteve ligado?
Quantas pessoas estiveram na sala de aula?
Agora que vai sair da sala. Abriu as janelas na posição oscilo-batente superior (não aplicável à sala 106)?

5.2 Pré-processamento e limpeza dos dados

Após a conclusão do período de recolha de dados, tanto dos parâmetros ambientais quanto das respostas dos questionários, foi imperativo proceder a uma fase de pré-processamento e organização dos dados. Este processo foi fragmentado em várias etapas e implementado

utilizando *scripts* desenvolvidos na linguagem *Python* de forma a automatizar e simplificar o processo.

A primeira etapa consistiu no *download* das respostas dos questionários fornecidas pelos docentes, bem como a obtenção dos dados ambientais registados pelos sensores nas salas de aula, através do sistema *Airmon*. Os dados foram descarregados em formato CSV, gerando-se um ficheiro para cada tipo de variável ambiental. A Figura 5.4 apresenta um exemplo do ficheiro gerado para as leituras de CO_2, composto por quatro colunas. A primeira coluna indica a data da leitura, enquanto as colunas subsequentes representam os valores medidos em cada uma das salas monitorizadas.

Time	ppm.mean {en	ppm.mean {entity_id:	ppm.mean {entity_id: ufp_sala_210_co2_value}
27/03/2023 00:15	404	402	401
27/03/2023 00:20	401	404	401
27/03/2023 00:25	401	404	401
27/03/2023 00:30	401	400	400
27/03/2023 00:35	402	404	402
27/03/2023 00:40	402	404	402
27/03/2023 00:45	402	404	402

Figura 5.4: Ficheiro CSV com as leituras de CO_2

Na segunda etapa, o primeiro *script* (Figura 5.5) é responsável por ler os diversos ficheiros que contêm as informações ambientais, combinando-os num único ficheiro. Nesta fase, o ficheiro resultante contém toda a informação ambiental recolhida de todas as salas. Em seguida, o programa gera três ficheiros CSV distintos, um para cada sala de aula, segmentando, assim, a informação.

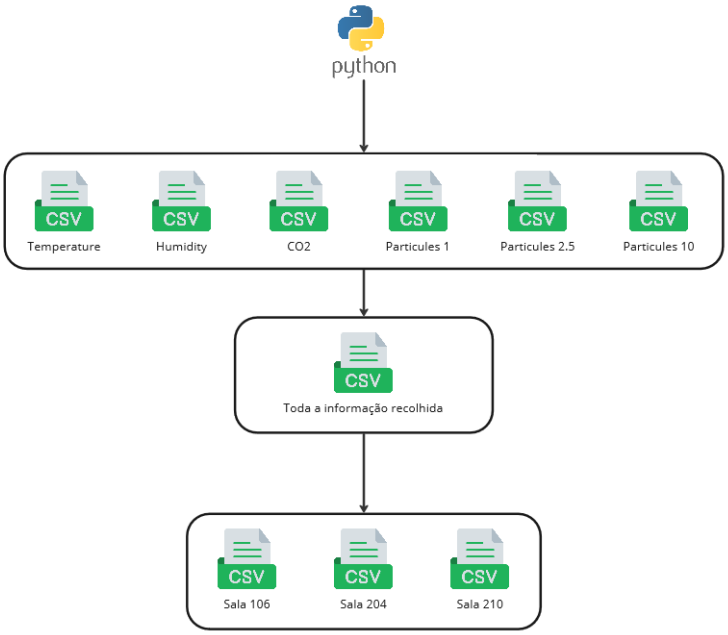


Figura 5.5: Fluxo do *script Python*

A terceira etapa envolve a execução de um segundo *script*. Inicialmente, este *script* altera o nome das colunas no ficheiro de respostas dos questionários, simplificando o formato para facilitar a leitura através do mesmo. Uma vez que o ficheiro de respostas contém apenas uma entrada para cada aula, torna-se necessário gerar os períodos correspondentes a cada uma, dado que as leituras dos sensores ocorrem em intervalos de cinco minutos. Após a modificação dos nomes das colunas, o programa percorre cada entrada no ficheiro de respostas (correspondente a uma aula lecionada) e gera os períodos de cinco em cinco minutos. Adicionalmente, é criado um campo adicional, *"ClassId"*, que corresponde a um identificador único para cada aula. Dado que os questionários são preenchidos após a lecionação de uma aula, os períodos em que não existem registos de aulas, não possuem valores para as variáveis derivadas dos questionários. Estes períodos foram preenchidos com as últimas respostas conhecidas dos questionários, garantindo que as variáveis estejam completas, à exceção do campo *"Persons_in_classroom"* que foi sempre marcado com o valor zero. As Figuras 5.6 e 5.7 apresentam a entrada original no ficheiro de respostas e a nova entrada após esta etapa[1].

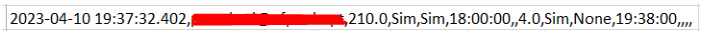


Figura 5.6: Resposta original do questionário

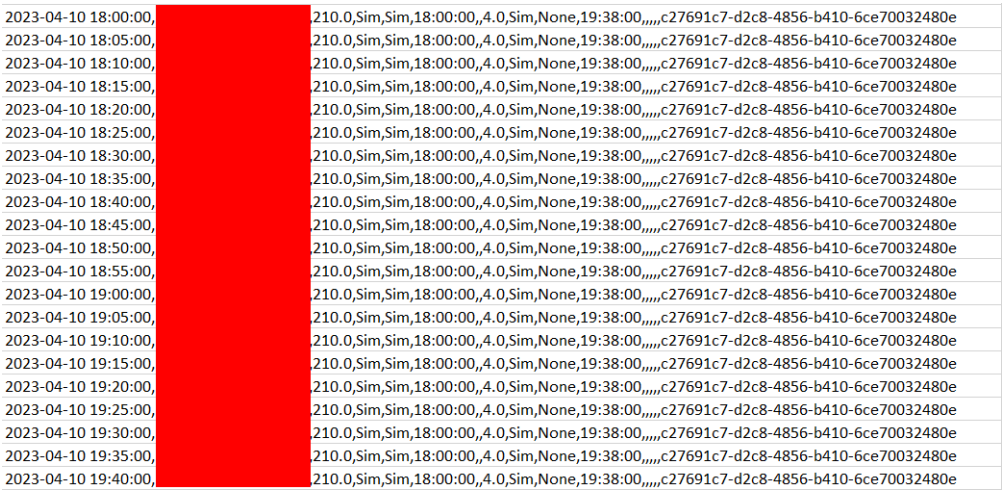


Figura 5.7: Resposta do questionário com os períodos da aula

Posteriormente, o *script* procede à fusão da informação ambiental com as entradas no ficheiro de respostas, originando, desta forma, um novo ficheiro com a informação completa para cada sala. Estes novos ficheiros contêm tanto as leituras ambientais quanto as respostas dos docentes em relação às aulas lecionadas naquela sala.

[1]Nas Figuras 5.6 e 5.7, os dados sensíveis dos intervenientes foram ocultados

A quarta etapa envolve a execução de um terceiro *script*, responsável pela criação dos ficheiros finais dos conjuntos de dados. Inicialmente, para cada um dos ficheiros das salas, é realizada uma regressão linear nos valores das variáveis ambientais, visando eliminar valores atípicos, tais como erros de leitura dos sensores. Este processo é implementado através da função *Linear Regression* da biblioteca *sklearn*, utilizando o parâmetro *"positive"* com o valor *false*, indicando, assim, a aplicação da técnica *Ordinary Least Squares* [1]. Em seguida, são geradas duas novas variáveis, nomeadamente *"CO_2 Velocity"* e *"CO_2 Acceleration"*, correspondentes à primeira e segunda derivada do valor de CO_2 na entrada atual. Estas variáveis foram criadas com o intuito de permitir que os algoritmos de inteligência artificial detetem padrões que poderiam não estar visíveis nos valores brutos de CO_2. Posteriormente, através de um parâmetro configurável, o *script* elimina um número N de entradas iniciais e finais para cada aula. Por exemplo, se N = 3 e uma aula ocorrer entre as catorze e as dezasseis horas, as entradas consideradas serão apenas das catorze e quinze até às quinze e quarenta e cinco. Este procedimento visa remover os períodos nos quais poderia ocorrer a abertura e o fecho de portas para a entrada e saída de alunos, evitando possíveis oscilações nas leituras efetuadas. Por último, são removidas algumas colunas que não serão utilizadas pelos algoritmos, tais como as colunas com os endereços de *e-mail* dos docentes ou a coluna das observações. Adicionalmente, são convertidos os valores "Sim"e "Aberta"para 1, e "Não"e "Fechada"para 0.

5.3 Análise dos conjuntos de dados

Na secção anterior, foram definidas as várias etapas que compreendem, desde os diversos ficheiros que contêm a informação ambiental recolhida das salas de aula, até às respostas do questionário, englobando também o processo de construção dos conjuntos de dados finais. Nesta secção, é apresentada uma análise minuciosa dos dados presentes nos quatro conjuntos de dados, correspondendo cada um a uma sala de aula específica, e da fusão da informação proveniente das três salas. O propósito desta análise é compreender de que forma as informações presentes estão interligadas, como se distribuem em termos da ocupação das distintas salas de aula e divulgar a estrutura de toda a informação recolhida.

5.3.1 Informação presente nos conjuntos de dados

A Tabela 5.2 apresenta as colunas presentes nos diferentes conjuntos de dados construídos no estudo realizado na Universidade Fernando Pessoa. A coluna *"Time"* indica a data e hora em que foi registada a entrada dos dados, seguindo o padrão Ano-Mês-Dia Hora:Minutos:Segundos. A coluna *"Classroom"* representa o número de identificação de cada sala de aula, permitindo uma organização eficaz dos dados. As colunas *"Door_-*

[1] https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

closed_on_arrival" e *"Windows_closed_on_arrival"* são binárias, indicando se a porta e as janelas da sala, respetivamente, estavam fechadas (1) ou abertas (0), no momento da entrada de dados. A coluna *"Persons_in_classroom"* indica o número de pessoas presentes na sala de aula no momento da entrada dos dados, sendo crucial para a classificação da ocupação do espaço. A coluna *"Opened_windows_end_of_class"* (binária), indica se as janelas foram abertas no final da aula. O estado do ar condicionado durante a aula consta na coluna *"Ac_on_during_class"*. A coluna *"During_class_door_state"* informa se a porta esteve aberta ou fechada durante a aula e a coluna *"During_class_window_-state"* evidencia o estado das janelas durante a aula. A coluna *"ClassId"* é um identificador único, no formato *Universally Unique Identifier* (UUID), associado a cada aula, permitindo uma ligação específica entre os dados e a aula correspondente. Os valores da humidade e temperatura da sala de aula estão representados, respetivamente, pelas colunas *"Humidity"* e *"Temperature"*. As colunas *"Particules 1"*, *"Particules 2.5"* e *"Particules 10"* indicam os valores das partículas suspensas no ar com diâmetros de 1, 2.5 e 10 micrómetros, respetivamente. A concentração de dióxido de carbono (CO_2) na sala de aula é apresentada na coluna "CO_2". Além disso, as colunas "*CO_2_Velocity* (Primeira Derivada do CO_2)"e "*CO_2_Acceleration* (Segunda Derivada do CO_2)"representam, respetivamente, a primeira e segunda derivada do valor do dióxido de carbono (CO_2).

Tabela 5.2: Colunas dos conjuntos de dados

Coluna	**Descrição**
Time	Data de registo da entrada de dados
Classroom	Número da sala de aula
Door_closed_on_arrival	Se a porta se encontrava fechada na chegada à sala
Windows_closed_on_arrival	Se as janelas se encontravam fechadas na chegada à sala
Persons_in_classroom	Número de pessoas presentes na sala durante a aula
Opened_windows_end_of_class	Se a janela foi aberta no final da aula
Ac_on_during_class	Se o ar condicionado foi ligado durante a aula
During_class_door_state	Se a porta durante a aula esteve aberta ou fechada
During_class_window_state	Se as janelas durante a aula estiveram abertas ou fechadas
ClassId	Identificador da aula
Humidity	Valor de Humidade
Temperature	Valor da Temperatura
Particules 1	Valor das Partículas com diâmetro 1
Particules 2.5	Valor das Partículas com diâmetro 2.5
Particules 10	Valor das Partículas com diâmetro 10
CO_2	Valor do CO_2
CO_2_Velocity	Primeira derivada do valor de CO_2
CO_2_Acceleration	Segunda derivada do valor de CO_2

5.3.2 Análise da informação dos conjuntos de dados

Correlação entre Variáveis Ambientais e Ocupação

Na Figura 5.8 é apresentado um mapa de calor que representa as correlações entre as variáveis ambientais e o número de ocupantes nas três salas de aula monitorizadas. A correlação foi calculada utilizando a função *corr* da biblioteca Pandas, aplicando o coeficiente de correlação de *Pearson*[1]. Como antecipado, observa-se uma forte correlação entre as variáveis relacionadas com partículas (aproximadamente 100%). Surpreendentemente, esta forte correlação não se reflete nas variáveis derivadas do valor original de CO_2, representando a primeira e segunda derivadas, com correlações de 49% e 42%, respetivamente. Era esperada uma correlação mais robusta entre estas variáveis derivadas. Além disso, não se verifica uma correlação significativa entre o CO_2 e a temperatura, o que é inesperado, pois previa-se que ambas aumentassem de forma semelhante. Observa-se também uma correlação moderada (aproximadamente 25%) entre as partículas e a humidade, o que é coerente, considerando que em ambientes com alguma humidade, há maior

[1] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

acumulação de partículas no ar. No que diz respeito ao número de pessoas, é crucial analisar como cada variável no conjunto de dados está correlacionada com este fator. O CO_2 e a temperatura apresentam uma correlação mais significativa, de 46% e 26%, respetivamente. No entanto, não se observa uma correlação forte entre elas. As variáveis derivadas do CO_2 também apresentam alguma correlação, embora menos pronunciada em comparação com o valor original. Quanto às partículas, a variável que representa partículas com um diâmetro de 1 apresenta a correlação mais expressiva, cerca de 12%.

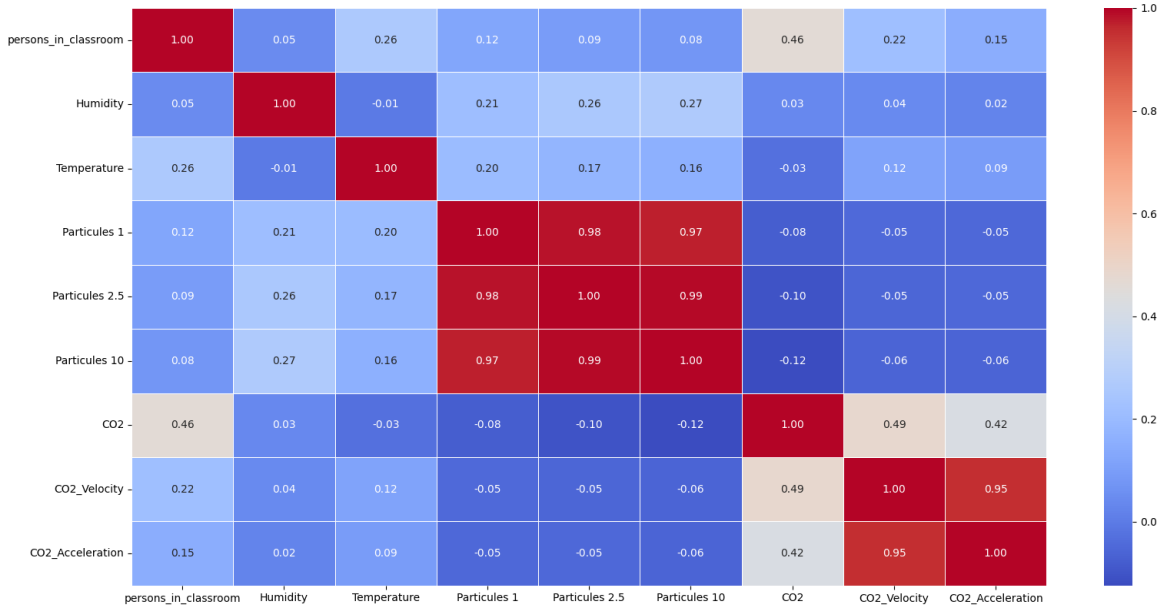


Figura 5.8: Mapa de calor da correlação entre variáveis na junção dos três conjuntos de dados

Conforme evidenciado no mapa de calor, algumas variáveis ambientais apresentam uma correlação mais significativa com a ocupação das salas de aula. Dado que o corpo humano liberta CO_2, é expectável que os seus níveis na sala aumentem com a sua ocupação, o que justifica a correlação encontrada anteriormente. Contudo, no caso da humidade, não é esperado o mesmo padrão. Nas Figuras 5.9, 5.10, 5.11, 5.12, 5.13 e 5.14, visa-se analisar especificamente a informação ambiental recolhida na sala 204 da Universidade Fernando Pessoa, no dia 11 de Abril de 2023, em relação à ocupação registada pelos docentes. Pretende-se demonstrar como os parâmetros ambientais variaram de acordo com a ocupação e a sua correlação com a mesma. Neste dia, foram registadas três aulas: a primeira com dezanove pessoas, a segunda com onze pessoas e a terceira com cinco pessoas.

CO_2 e o Número de ocupantes na sala

Conforme ilustrado na Figura 5.9, verifica-se a correlação esperada: os níveis de CO_2 aumentam em função da ocupação da sala. Antes do início da primeira aula, observou-se, detalhadamente, que os níveis de CO_2 situavam-se por volta dos quatrocentos e cinquenta ppm. Com o decorrer do tempo e a presença de dezanove pessoas, esses níveis elevaram-se para dois mil e oitocentos ppm, ultrapassando consideravelmente os limites classificados como saudáveis para um ambiente fechado. Entre a primeira e a segunda aula, registou-se uma diminuição nos níveis de CO_2, atribuída à abertura das janelas e à ventilação natural. Além disso, constatou-se que os níveis de CO_2 aumentaram durante a segunda e terceira aulas, enquanto a sala esteve ocupada. No entanto, devido à menor ocupação nestas aulas em comparação com a primeira, não atingiram picos tão elevados.

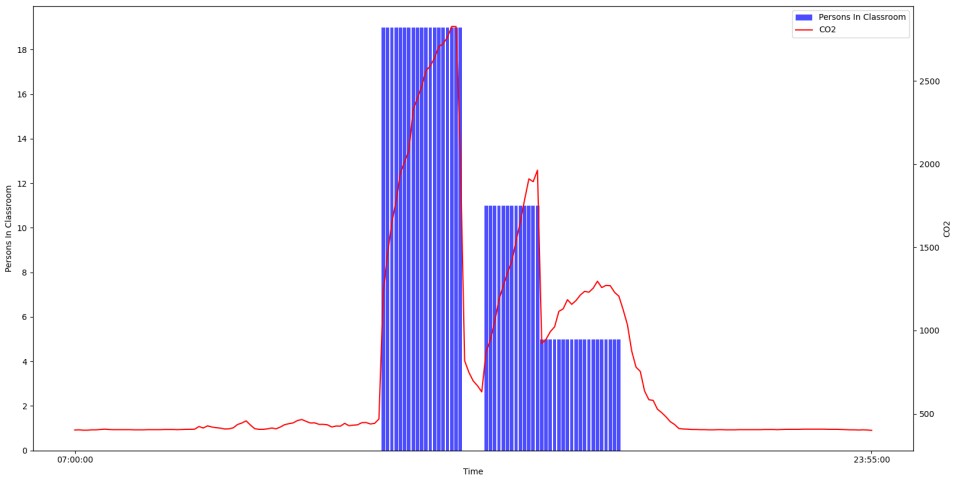


Figura 5.9: Níveis de CO_2 e Número de ocupantes na sala

Temperatura e o Número de ocupantes na sala

A temperatura, quando correlacionada com o número de ocupantes presentes na sala, manifesta um comportamento semelhante ao do CO_2. Conforme ilustrado na Figura 5.10, observa-se que durante a realização de uma aula com a sala ocupada, a temperatura tende a aumentar, embora não tanto quanto os níveis de CO_2. No entanto, constata-se um aumento semelhante da temperatura fora dos períodos de aula, quando a sala se encontra vazia. Este fenómeno poderá ser atribuído, por exemplo, à exposição solar que a sala recebe ou às condições climatéricas do dia em que foram efetuados os registos.

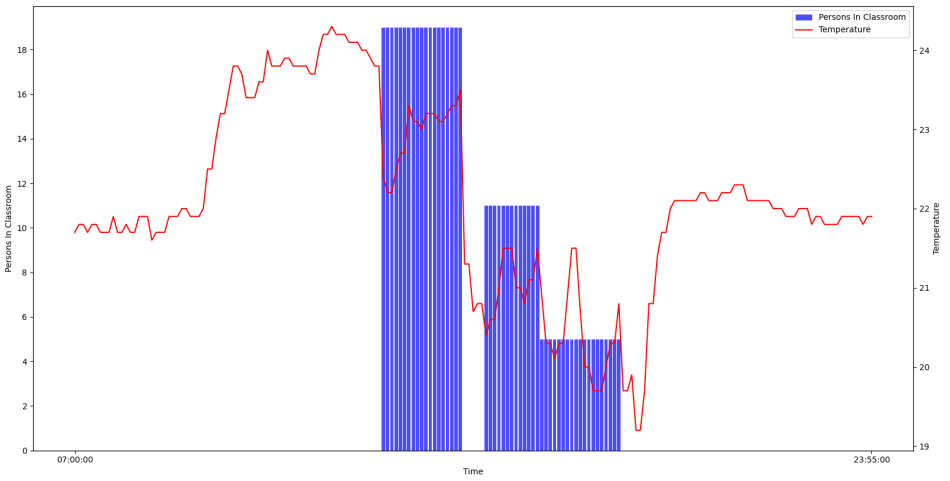


Figura 5.10: Temperatura e Número de ocupantes na sala

Humidade e o Número de ocupantes na sala

No que diz respeito à humidade, ao analisar a Figura 5.11, verifica-se a ausência de uma correlação direta entre os seus níveis e a presença de pessoas na sala de aula. Durante o decorrer das aulas, não se evidenciou uma clara tendência de aumento em função da ocupação. Semelhantemente à temperatura, os níveis de humidade registados na sala podem ser influenciados pelas condições climatéricas do dia ou pela exposição solar que a sala recebe.

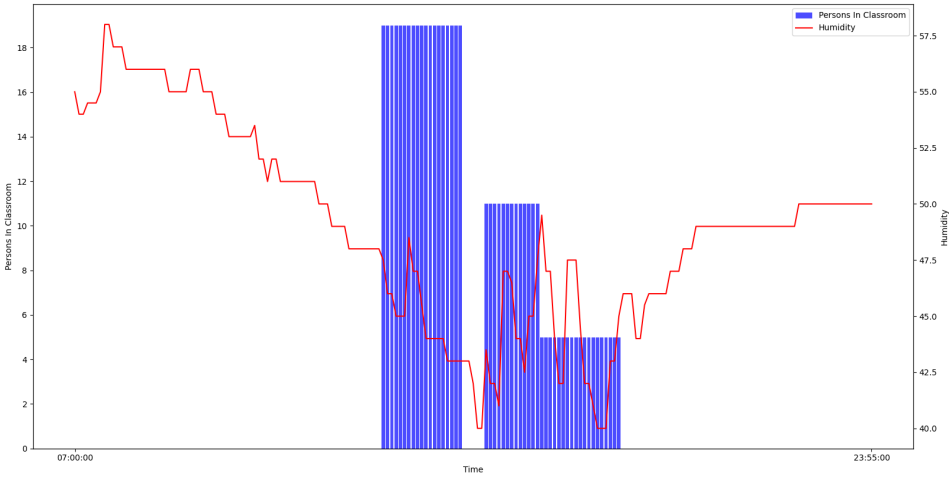


Figura 5.11: Humidade e Número de pessoas na sala

Partículas e o Número de ocupantes na sala

As Figuras 5.12, 5.13 e 5.14 apresentam os gráficos que representam as partículas com diâmetro de 1, 2.5 e 10 micrómetros, respetivamente, em relação ao número de pessoas na sala. Não é possível identificar uma tendência clara de aumento de nenhum dos casos, com base no número de pessoas presentes na sala de aula. Ou seja, constata-se que, apesar das diferentes ocupações das aulas, os gráficos das diversas partículas exibem padrões semelhantes. As oscilações durante o período das aulas podem ser justificadas pelos movimentos das pessoas que ocorrem.

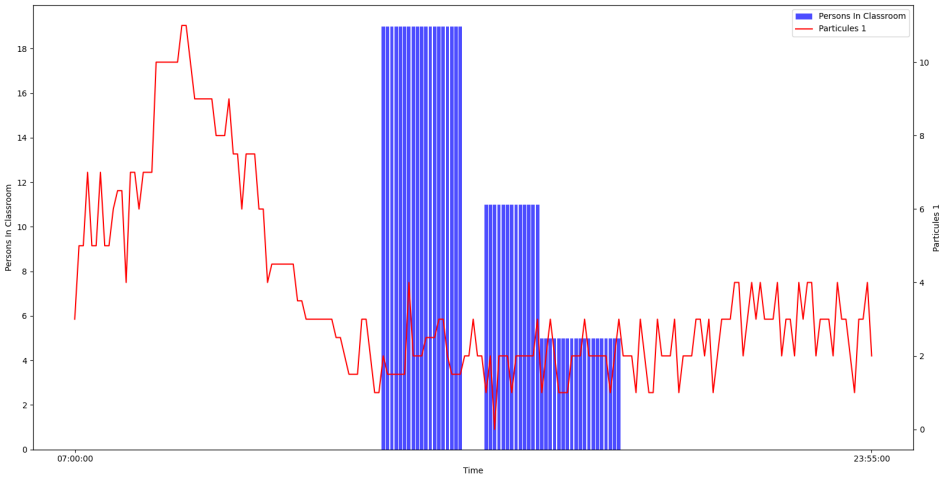


Figura 5.12: Partículas 1 e Número de pessoas na sala

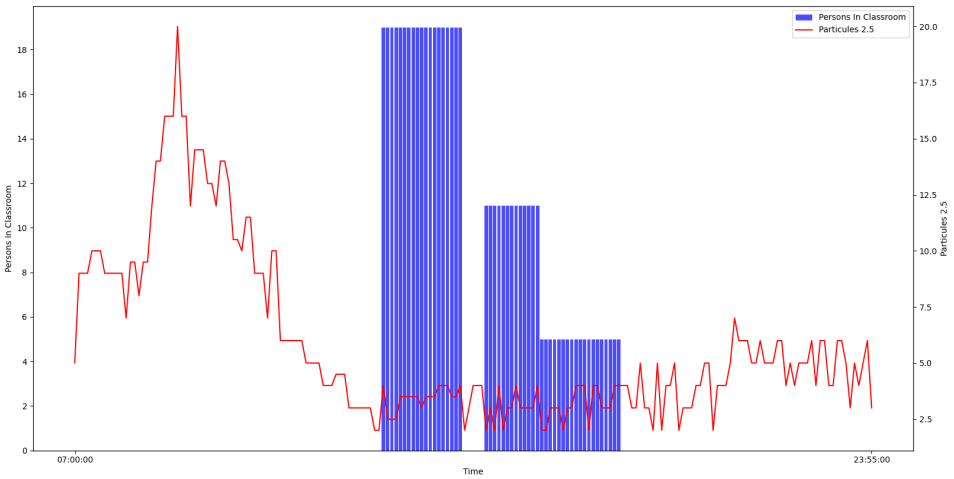


Figura 5.13: Partículas 2.5 e Número de pessoas na sala

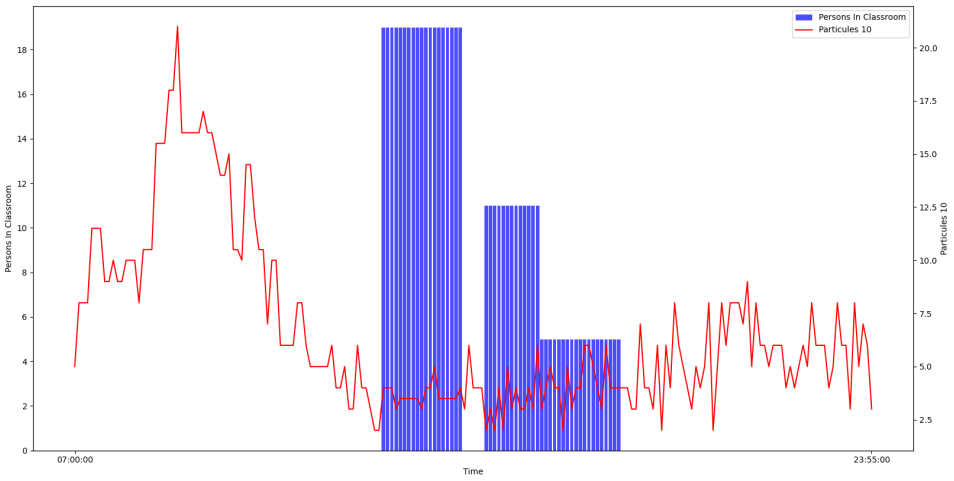


Figura 5.14: Partículas 10 e Número de pessoas na sala

Distribuição da Ocupação nas Salas de Aula

Além da relevância da correlação entre as variáveis ambientais e a ocupação das salas de aula, é essencial compreender a distribuição dessa ocupação em cada uma das salas monitorizadas. Na Figura 5.15, apresenta-se o histograma que ilustra a ocupação registada na sala 106. Pode-se observar que as três ocupações mais frequentes possuem uma considerável diferença entre si, sendo a ocupação de seis pessoas a mais prevalente, seguida pelas de vinte e duas e vinte pessoas, respetivamente. Não foram registadas ocupações de nove e vinte e uma pessoas e a ocupação registada na sala varia de uma a vinte e sete pessoas, com uma média de aproximadamente catorze pessoas.

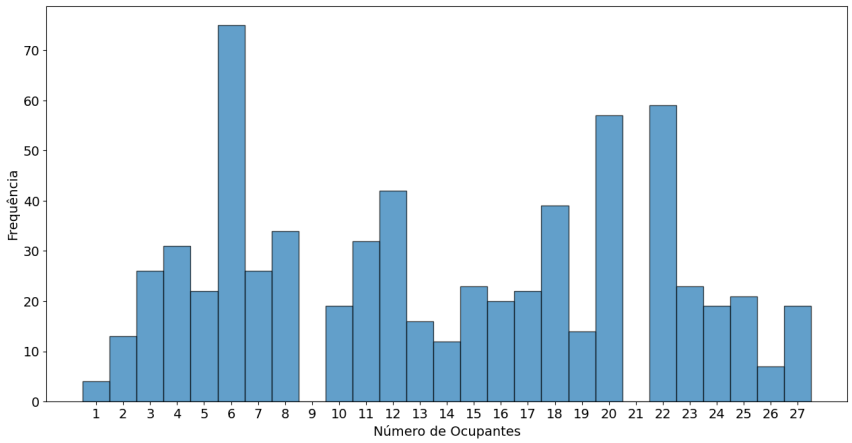


Figura 5.15: Histograma de Ocupação da Sala 106

No que concerne à sala 204, o histograma correspondente, apresentado na Figura 5.16, evidencia uma discrepância no intervalo de ocupação, variando de uma a trinta pessoas, com uma média de ocupação de onze pessoas. As três ocupações mais frequentes diferem igualmente daquelas observadas na sala 106. Na sala 204, a ocupação mais comum é de dez pessoas (mais próxima da média de ocupação), seguida de quinze e sete pessoas, pela segunda e terceira ordem, respetivamente. É notável também um maior registo de ocorrências para uma única ocupação, sendo que, no caso de dez pessoas, este número ultrapassa os duzentos registos. No caso da sala 204, existem diversas ocupações que não se encontram representadas.

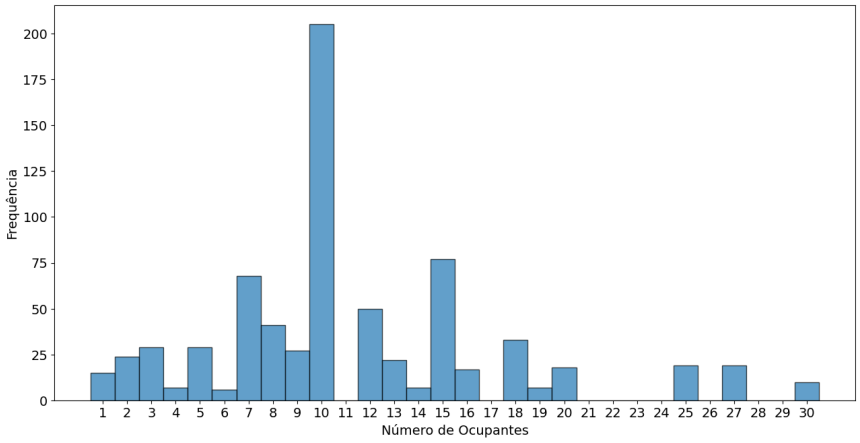


Figura 5.16: Histograma de Ocupação da Sala 204

Na sala 210, ao analisar o histograma correspondente, representado na Figura 5.17, constata-se que as três ocupações mais frequentes situam-se abaixo de dez pessoas, com a média de ocupação aproximando-se das nove pessoas. Destaca-se que o intervalo de ocupação nesta sala é o menor em termos de valor máximo registado, atingindo no máximo vinte e quatro pessoas e no caso da ocupação de vinte e duas pessoas, não existe nenhuma ocorrência.

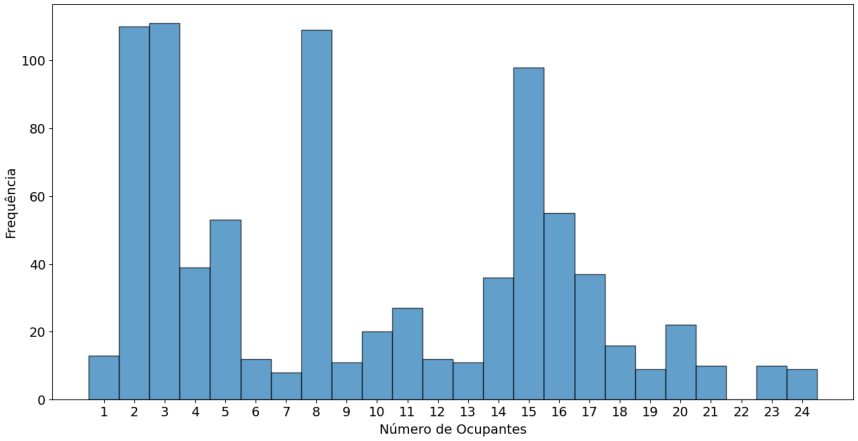


Figura 5.17: Histograma de Ocupação da Sala 210

Além da análise individual de cada sala, é crucial realizar uma análise conjunta das mesmas. O histograma apresentado na Figura 5.18 representa a união dos histogramas das três salas analisadas. É notável a baixa representatividade de algumas ocupações, como vinte e seis pessoas ou trinta pessoas. A média de ocupação das salas é de aproximadamente onze pessoas e as ocupações mais distintivas são de dez, quinze e oito pessoas.

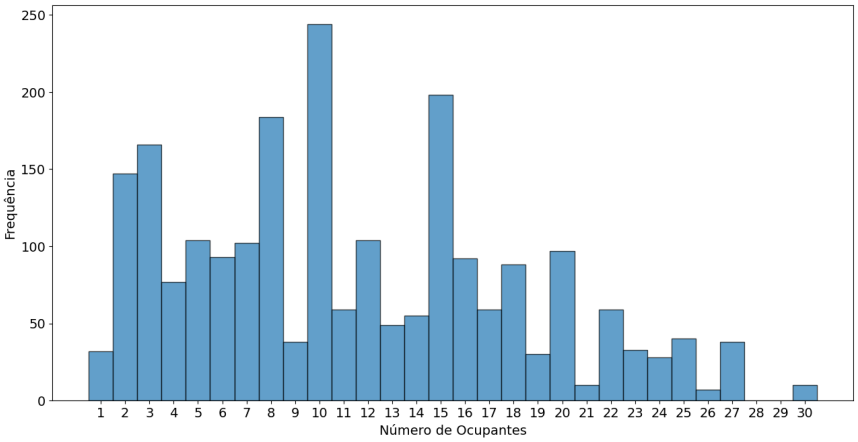


Figura 5.18: Histograma de Ocupação das três salas

5.4 Conclusão

Ao longo deste capítulo foi apresentada uma descrição detalhada do conjunto de dados recolhido para a classificação de ocupação de salas de aula, abordando os procedimentos de recolha, pré-processamento, e limpeza dos dados, bem como uma análise minuciosa dos conjuntos de dados resultantes. A recolha dos dados foi feita através de um protocolo que os docentes e alunos deveriam seguir, visando controlar a componente ambiental das salas monitorizadas. Este protocolo envolveu diferentes etapas, desde o enquadramento do projeto até ao preenchimento de questionários pelos docentes após cada aula. Durante a recolha, foram adotadas estratégias para aumentar a variabilidade do conjunto de dados. Após a recolha dos dados, foi realizado um extenso processo de pré-processamento e limpeza para organizar e preparar os dados para análise. Este processo envolveu o *download* das respostas dos questionários e dos dados ambientais, a combinação dos diferentes ficheiros, geração de variáveis derivadas e eliminação de entradas não relevantes. Foram utilizados *scripts* em *Python* para automatizar estas tarefas. A análise dos conjuntos de dados foi realizada, incluindo uma análise da correlação entre as variáveis ambientais e o número de ocupantes nas salas de aula. Sendo observada uma correlação significativa entre os níveis de CO_2 e a ocupação da sala, indicando que os níveis de CO_2 aumentam com a presença de mais pessoas. No entanto, outras variáveis ambientais como temperatura e humidade não apresentaram uma correlação tão evidente com a ocupação. Foi também feita uma análise da distribuição da ocupação nas salas, revelando padrões distintos em cada uma delas. Todos os conjuntos de dados construídos ao longo deste capítulo encontram-se disponíveis para acesso[1].

[1] Visitar: `https://github.com/jotaSVV/AirmonSystem-Datasets`

Capítulo 6

Inteligência Artificial nos Dados

Neste capítulo, serão explorados pormenores cruciais relativos à aplicação de algoritmos de inteligência artificial para a classificação da ocupação em salas de aula. Este processo baseia-se nos conjuntos de dados previamente elaborados, conforme explicado no capítulo anterior. Serão abordados em detalhe a metodologia adotada, as técnicas utilizadas, a criação de modelos e os testes realizados para avaliar os resultados obtidos.

6.1 Intervenção e Análise dos conjuntos de dados

Embora nos testes a serem realizados existam parâmetros que podem ser sujeitos a ajustes, as seguintes operações foram aplicadas aos conjuntos de dados. A primeira operação a ser executada, consistiu no mapeamento do número de pessoas para uma nova coluna designada por *"Class"*. Nesta coluna, o número de pessoas é categorizado em várias classes com base em intervalos previamente definidos. Por exemplo, se o intervalo for de cinco e a ocupação for de seis pessoas, o valor da classe será 1. A abordagem baseada em classes de ocupação pode oferecer resultados mais eficazes em contraste com a utilização direta das presenças. Outra operação realizada nos dados envolveu a utilização da função *"StandardScaler"* do módulo *sklearn*[1]. Esta função tem como propósito padronizar os dados.

Uma vez que foi criada uma nova coluna para representar a ocupação das salas em classes, é fundamental analisar como a ocupação está distribuída nessas classes. Para efeitos de avaliação dos modelos criados, serão consideradas as classes com intervalo três e intervalo cinco. Nas Figuras 6.1 e 6.2, estão representados os histogramas das classes na sala 106 com intervalo três e intervalo cinco, respetivamente. Na classe três, observa-se que existem nove possíveis classes para atribuir a uma ocupação, sendo que a mais frequente é a classe dois, que compreende um número de pessoas entre quatro e seis. Em relação à classe com intervalo cinco, é expectável que haja menos possibilidades devido

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

ao intervalo maior. Neste caso, existem seis possíveis classes, sendo a mais representada também a classe dois, que abrange um intervalo de seis a dez pessoas.

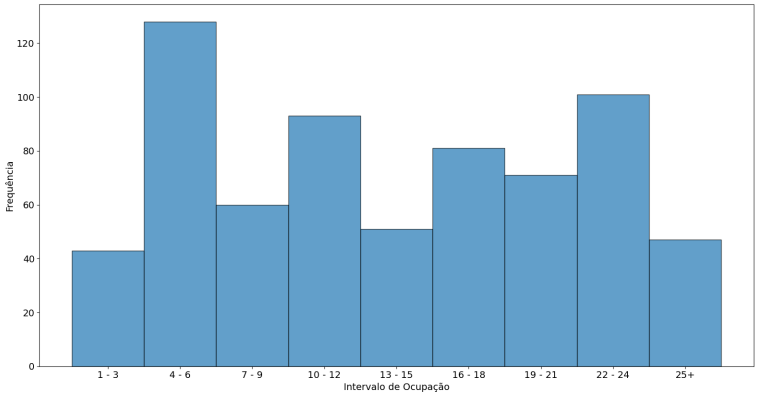


Figura 6.1: Histograma de Ocupação por Classes 3 da Sala 106

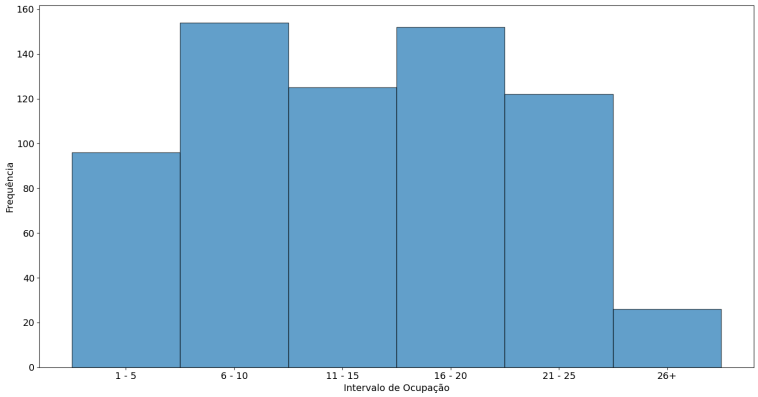


Figura 6.2: Histograma de Ocupação por Classes 5 da Sala 106

Em relação à sala 204, as Figuras 6.3 e 6.4 representam os histogramas das classes com intervalo de três e cinco pessoas, respetivamente. Na análise da primeira classe, observa-se que a mais frequente compreende um intervalo de pessoas entre dez e doze. Relativamente à classe com intervalo de cinco pessoas, a mais comum está entre seis e dez pessoas.

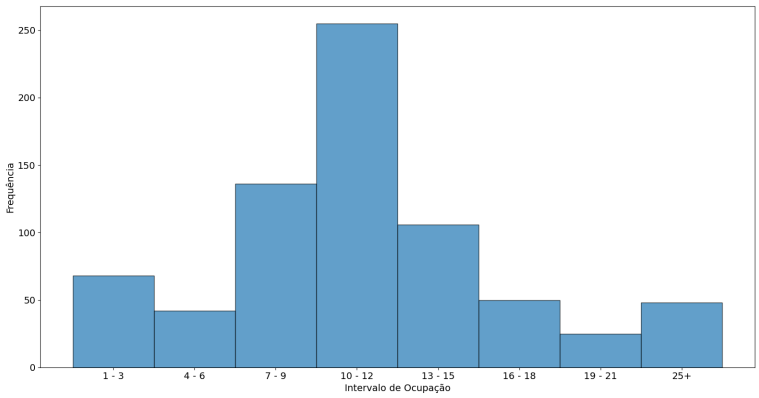


Figura 6.3: Histograma de Ocupação por Classes 3 da Sala 204

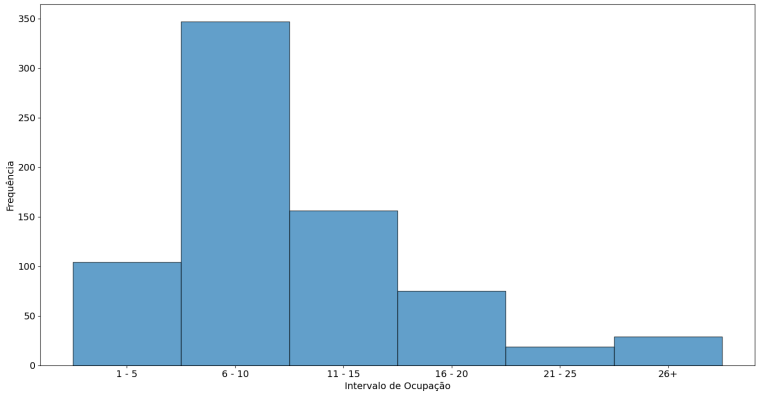


Figura 6.4: Histograma de Ocupação por Classes 5 da Sala 204

Na sala 210, e analisando os histogramas apresentados nas Figuras 6.5 e 6.6, constata-se que, na classe de intervalo três, a representatividade mais significativa situa-se na ocupação de uma a três pessoas, embora seja relativamente menor em comparação com as outras duas salas. Quanto à classe de intervalo cinco, observa-se um padrão semelhante, com a classe mais representada sendo a número um, que abrange uma ocupação entre uma e cinco pessoas.

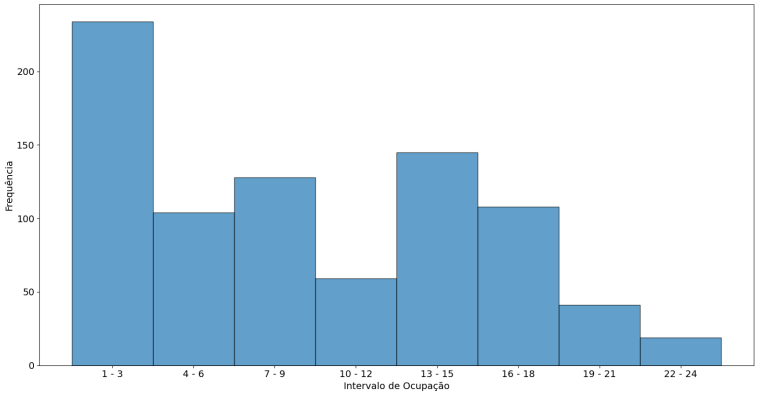


Figura 6.5: Histograma de Ocupação por Classes 3 da Sala 210

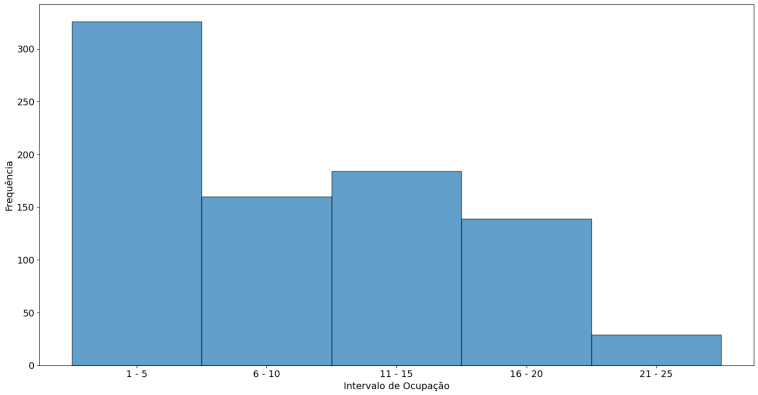


Figura 6.6: Histograma de Ocupação por Classes 5 da Sala 210

Uma vez que nos testes realizados treinou-se um modelo que engloba os dados das três salas de aula monitorizadas, torna-se relevante visualizar a distribuição da ocupação no formato de classes. Nas Figuras 6.7 e 6.8, apresentam-se os histogramas resultantes da combinação dos dados de todas as salas para cada uma das classes. No que concerne à classe de intervalo três, é possível concluir que a ocupação entre as classes um e cinco, está presente em mais de duzentas e cinquenta ocorrências. Contudo, a ocupação na classe nove possui uma frequência inferior a cem, sugerindo que o modelo poderá ter dificuldades em classificar este tipo de classes. Relativamente à classe de intervalo cinco, a situação é semelhante. No caso da classe seis, que representa uma ocupação superior a vinte e seis pessoas, a representatividade é reduzida.

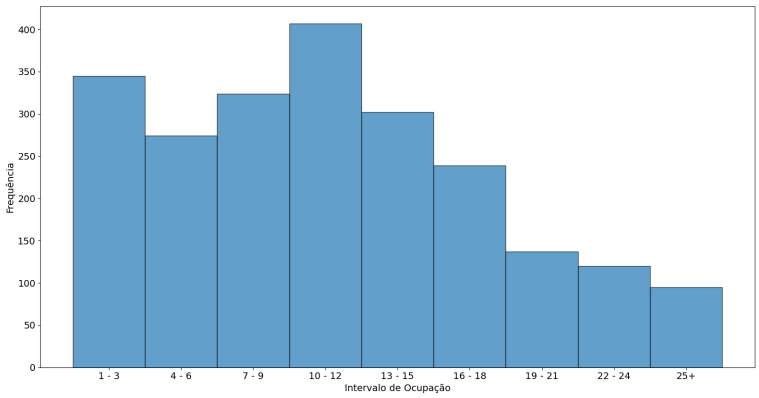


Figura 6.7: Histograma de Ocupação por Classes 3 em todas as salas

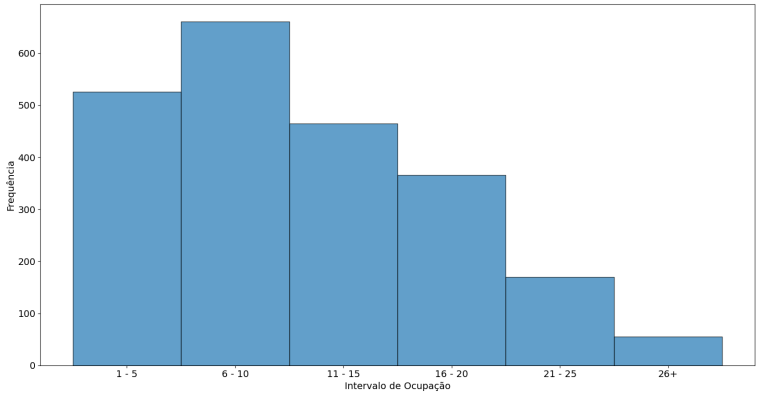


Figura 6.8: Histograma de Ocupação por Classes 5 em todas as salas

6.2 Parametrização dos testes

Com o objetivo de facilitar a alteração dos parâmetros utilizados, foi desenvolvida uma função em *Python* que recebe parâmetros essenciais para serem fornecidos aos modelos. A Tabela 6.1 apresenta cada um desses parâmetros, juntamente com as suas descrições. O parâmetro *file_paths* indica a lista que contém os caminhos dos ficheiros de onde os dados devem ser lidos. *model_name* define o nome a ser utilizado para guardar o ficheiro do modelo criado. *columns_to_drop* representa o conjunto de colunas a ser ignorado no conjunto de dados. Os parâmetros *start_date* e *end_date* indicam o intervalo de datas que os modelos devem considerar. *ambiental_only* determina se devem ser consideradas apenas as variáveis ambientais e *class_interval* indica o intervalo a ser considerado nas classes.

Tabela 6.1: Parâmetros presentes na função de teste

Parâmetro	**Descrição**
file_paths	Lista do caminho dos ficheiros de onde devem ser lidos os dados
model_name	Nome utilizado para guardar o modelo criado
columns_to_drop	Lista de colunas a serem desconsideradas pelo modelo
start_date	Data de inicio a ser considerada
end_date	Data de fim a ser considerada
ambiental_only	Se apenas devem ser utilizadas variáveis ambientais
class_interval	Qual o intervalo a ser utilizado para gerar as classes de ocupação

6.3 Metodologia

Para a realização de testes nos conjuntos de dados construídos, foi necessário seguir uma metodologia coerente, com o objetivo de avaliar diferentes técnicas de inteligência artificial, incluindo CNN, MLP, RF *Classifier* e KNN. As primeiras duas técnicas foram escolhidas pela curiosidade acerca das mesmas, enquanto que as técnicas RF *Classifier* e KNN foram escolhidas por serem técnicas tradicionais da inteligência artificial. No decorrer do processo de análise, o primeiro passo consistiu na realização de testes individuais de cada técnica, considerando todas as características ambientais. Posteriormente, com base na matriz de correlação ilustrada na Figura 5.8, procedeu-se aos testes, removendo as colunas *"Particules 2.5"*, *"Particules 10"* e *"CO2_Acceleration"*. Esta decisão foi tomada devido à sua fraca correlação com a variável de ocupação e à alta correlação com outras características. A realização deste tipo de teste, caso não afete significativamente os resultados, pode contribuir para uma otimização de recursos computacionais e auxiliar na determinação da inclusão ou exclusão de sensores adicionais nas caixas de monitorização. É importante salientar que, no caso do parâmetro *class_interval*, este foi testado com os valores cinco e três, a fim de compreender como a classificação da ocupação das salas de aula, utilizando diferentes intervalos, poderia impactar nos resultados obtidos pelos modelos. Na Tabela 6.2, estão descritos o conjunto de parâmetros que não foram variados ao longo dos testes e os valores fixos que foram utilizados durante os mesmos.

Tabela 6.2: Valores fixos utilizados durante os testes

Parâmetro	**Valor**
start_date	2023-04-13 18:45:00
end_date	2023-06-18 18:55:00
ambiental_only	*True*

Durante os testes, o conjunto de dados foi sistematicamente dividido em dois sub-

conjuntos distintos: um destinado ao treino dos diversos modelos e outro para avaliar o desempenho desses modelos após o treino. Essa divisão seguiu uma proporção de 80% para treino e 20% para testes. É importante ressaltar que, durante a fase de treino, os modelos não têm contacto com os dados de teste. Desta forma, a classificação da ocupação é realizada em dados desconhecidos pelos modelos. Na Figura 6.9 está representada a forma como essa classificação é feita, nomeadamente, os modelos recebem os dados de uma leitura e indicam a classe de ocupação[1].

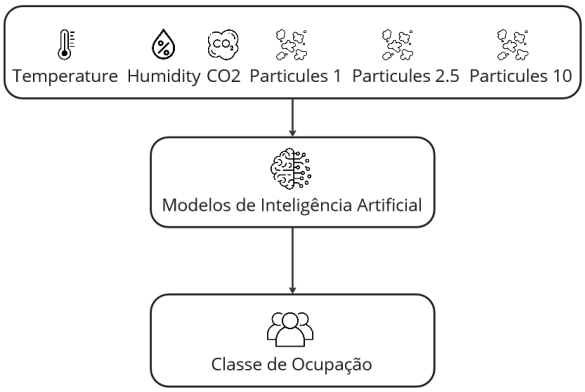


Figura 6.9: Fluxo de Classificação dos Modelos

Para efetuar a divisão dos dados de forma eficaz, utilizou-se a função *'train_test_split'* do módulo *'sklearn'*[2]. Ao aplicar esta função, o parâmetro *'random_state'*[3] foi configurado com o valor oito para garantir a replicabilidade dos resultados em diferentes chamadas e a *flag 'shuffle'* foi configurada como *'True'*, possibilitando a reorganização aleatória dos dados. Esta abordagem foi adotada para mitigar o risco dos modelos inclinarem excessivamente para uma classe específica de ocupação, especialmente quando essa classe está sobrerrepresentada no subconjunto de treino e o balanceamento não foi aplicado.

É de extrema importância ressaltar que este procedimento foi realizado individualmente para cada uma das salas, visando compreender o comportamento de cada técnica em ambientes específicos. Adicionalmente, durante esses testes, foram considerados somente os períodos de aula, excluindo os intervalos entre as mesmas. Esta abordagem possibilitou uma análise mais precisa das características ambientais relevantes durante as aulas, ou seja, nos momentos em que as salas estavam efetivamente ocupadas. Após concluir a análise de cada sala, as técnicas selecionadas foram aplicadas nas três salas em conjunto. Esta etapa foi realizada com o intuito de avaliar a capacidade de generalização dos modelos. Todos os modelos foram avaliados com o conjunto de dados de

[1] As características utilizadas na Figura podem diferenciar das utilizadas nos testes
[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[3] https://scikit-learn.org/stable/glossary.html#term-random_state

teste, permitindo o cálculo da acurácia que se define pela avaliação da taxa de acertos, obtida pela razão entre o número de acertos e o número total de entradas. No caso das técnicas MLP, CNN e RF *Classifier*, foi utilizada a função *accuracy_score*[1] disponível no módulo *sklearn* e, no caso da técnica KNN, foi utilizada a função *score*, fornecida pelo próprio modelo. Esta abordagem na avaliação dos modelos, proporcionou uma compreensão mais aprofundada do desempenho de cada técnica escolhida para cada sala, bem como da capacidade de generalização do modelo treinado para todas as salas.

6.4 Resultados

6.4.1 Sala 106

Resultados obtidos com todas as características

As Tabelas 6.3 e 6.4 apresentam os resultados obtidos no conjunto de dados da sala 106, utilizando todas as características ambientais disponíveis. Observa-se que, nesta sala específica e ao usar todas as características ambientais, a acurácia entre classes de ocupação permanece relativamente constante em algumas técnicas, como é o caso da CNN e do RF *Classifier*. Além disso, a acurácia diminui ligeiramente (1%) no caso da técnica KNN. Um resultado atípico foi observado na técnica MLP, que apresentou uma melhoria de acurácia (+8%) ao utilizar a classe três.

Tabela 6.3: Sala 106 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	97%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
RFClassifier	96%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
CNN	93%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
MLP	60%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Tabela 6.4: Sala 106 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	96%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
RFClassifier	96%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
CNN	93%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
MLP	68%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3

Resultados obtidos retirando características

A remoção de determinadas características pode, por vezes, resultar em melhorias nos resultados obtidos, uma vez que nem todas as características contribuem de forma positiva. Isso ficou evidente no caso da sala 106, como indicado nas Tabelas 6.5 e 6.6. Verificou-se que técnicas como o KNN e MLP aumentaram a sua acurácia quando foram removidas características como *"Particules 2.5"*, *"Particules 10"* e *"CO2_Acceleration"*. No que diz respeito à variação da classe de ocupação, a técnica KNN manteve uma acurácia quase máxima, atingindo 99%, enquanto técnicas como o RF *Classifier* sofreram uma pequena perda de aproximadamente 1%. Técnicas como CNN e MLP aumentaram as suas precisões quando a classe de ocupação foi reduzida, sendo que a segunda técnica apresentou o mesmo comportamento quando utilizou todas as características ambientais.

Tabela 6.5: Sala 106 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	99%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
RFClassifier	96%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
CNN	93%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
MLP	65%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5

Tabela 6.6: Sala 106 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	99%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
RFClassifier	95%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
CNN	95%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
MLP	70%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3

6.4.2 Sala 204

Resultados obtidos com todas as características

Analisando as Tabelas 6.7 e 6.8, observamos que a maioria das técnicas utilizadas obtiveram resultados satisfatórios no conjunto de dados da sala 204, com três delas atingindo uma acurácia de, pelo menos, 90% em ambas as classes de ocupação. No entanto, a técnica MLP, como já verificado na sala 106, manteve a menor acurácia. Notavelmente, neste conjunto de dados específico, quando todas as características ambientais foram consideradas, os modelos apresentaram uma diminuição na acurácia ao reduzir a classe de ocupação. Isto sugere que o modelo teve um desempenho mais desafiador ao lidar com esse conjunto de dados.

Tabela 6.7: Sala 204 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	96%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
CNN	93%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
RFClassifier	92%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
MLP	74%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5

Tabela 6.8: Sala 204 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	93%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
RFClassifier	91%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
CNN	90%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
MLP	68%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3

Resultados obtidos retirando características

De maneira semelhante ao que ocorreu no conjunto de dados da sala 106, as técnicas aplicadas na sala 204, com exceção da MLP, apresentaram benefícios ao remover algumas características ambientais, como demonstrado nas Tabelas 6.9 e 6.10. No entanto, ao contrário do que acontece quando todas as características são consideradas, as técnicas KNN, RF *Classifier* e CNN obtiveram uma diminuição na sua acurácia. Esta diminuição foi mais acentuada na técnica CNN, com uma redução de 6%.

Tabela 6.9: Sala 204 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	97%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
RFClassifier	94%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
CNN	94%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
MLP	63%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5

Tabela 6.10: Sala 204 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	94%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
RFClassifier	92%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
CNN	88%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
MLP	71%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3

6.4.3 Sala 210

Resultados obtidos com todas as características

Nas Tabelas 6.11 e 6.12, encontram-se os resultados dos testes realizados no conjunto de dados da sala 210. Tal como observado nos testes realizados nas outras duas salas, a técnica MLP continua a obter os piores resultados. No caso específico da sala 210, ao considerar todas as características ambientais, registou a menor acurácia obtida, com apenas 54%. Quando todas as características ambientais foram consideradas, nenhuma técnica obteve mais de 90% de acurácia, independentemente da classe de ocupação utilizada. Além disso, notou-se que a acurácia foi semelhante, quer se utilizasse a classe cinco ou três, indicando que os modelos não foram significativamente desafiados quando a classe de ocupação foi diminuída.

Tabela 6.11: Sala 210 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
RFClassifier	89%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
KNN	88%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
CNN	88%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
MLP	54%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5

Tabela 6.12: Sala 210 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
RFClassifier	90%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
KNN	88%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
CNN	87%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
MLP	59%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3

Resultados obtidos retirando características

No conjunto de dados da sala 210, verifica-se um ligeiro aumento da acurácia em algumas técnicas, como o KNN ou a CNN, ao utilizar a classe cinco, como ilustrado na Tabela 6.13. Comparando a acurácia entre a classe cinco e três, após a remoção de algumas características, é possível notar que, no caso da classe três, técnicas como o KNN mantêm a sua acurácia, mas no caso da CNN e RF *Classifier*, estas aumentam ligeiramente, como é demonstrado na Tabela 6.14. Em ambas as classes, o MLP continua a registar a pior acurácia.

Tabela 6.13: Sala 210 com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	91%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
RFClassifier	89%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
CNN	89%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
MLP	59%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5

Tabela 6.14: Sala 210 com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	91%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
RFClassifier	91%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
CNN	90%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
MLP	57%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3

6.4.4 Conjunto das Salas

Resultados obtidos com todas as características

Nas Tabelas 6.15 e 6.16, estão representados os resultados dos testes realizados com todas as características ambientais para as classes de ocupação cinco e três. Estes testes têm como objetivo avaliar a capacidade de generalização das técnicas para as diferentes salas e determinar a adequação dos resultados obtidos. Iniciando com a classe de ocupação cinco, observa-se que a técnica KNN apresenta uma acurácia de 94%, comparando com os resultados anteriores de 97% na sala 106, 96% na sala 204 e 88% na sala 210. Embora tenha havido uma ligeira diminuição no desempenho, ainda mantém uma capacidade satisfatória de generalização. A técnica KNN também regista uma acurácia de 93% quando aplicada à classe de ocupação três. As técnicas RF *Classifier* e CNN obtêm precisões bastante semelhantes em ambas as classes de ocupação, com uma ligeira redução no desempenho, embora ainda mantendo resultados satisfatórios. A técnica MLP continua a apresentar a pior acurácia em ambos os conjuntos de teste, sugerindo que pode não ser a mais adequada para o cenário em questão.

Tabela 6.15: Conjunto de Salas com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	94%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
RFClassifier	90%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
CNN	86%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5
MLP	52%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	5

Tabela 6.16: Conjunto de Salas com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	93%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
RFClassifier	89%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
CNN	85%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3
MLP	40%	*Humidity, Temperature, Particules 1, Particules 2.5, Particules 10, CO2, CO2_Velocity, CO2_Acceleration*	3

Resultados obtidos retirando características

As Tabelas 6.17 e 6.18 apresentam os resultados obtidos após a remoção das mesmas características ambientais usadas nos testes anteriores. Nota-se que a acurácia das técnicas permanece relativamente semelhante, não demonstrando ganhos ou perdas significativas de desempenho, tanto na classificação da classe de ocupação cinco como na classe três. Neste caso de teste específico, as técnicas KNN e RF *Classifier* destacam-se ao manterem uma acurácia de pelo menos 90%. As técnicas CNN e MLP, por outro lado, obtêm resultados ligeiramente inferiores, ficando abaixo desse limiar.

Tabela 6.17: Conjunto de Salas com Classe de Intervalo 5: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	94%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
RFClassifier	90%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
CNN	83%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5
MLP	51%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	5

Tabela 6.18: Conjunto de Salas com Classe de Intervalo 3: Técnicas, Acurácia, Características

Técnica	Acurácia	Características	Intervalo de Ocupação
KNN	93%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
RFClassifier	90%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
CNN	84%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3
MLP	48%	*Humidity, Temperature, Particules 1, CO2, CO2_Velocity*	3

6.5 Conclusão

O presente capítulo discutiu detalhadamente a aplicação de algoritmos de inteligência artificial para a classificação da ocupação em salas de aula, baseando-se nos conjuntos de dados previamente elaborados, conforme abordado no capítulo anterior. Foram explorados os pormenores da intervenção nos conjuntos de dados e a análise dos mesmos, a parametrização dos testes, a metodologia utilizada e os resultados obtidos.

No processo de intervenção e análise dos conjuntos de dados, foram aplicadas operações como mapeamento da ocupação em classes, com base em intervalos predefinidos e padronização dos dados, visando uma melhor preparação para a aplicação dos modelos de inteligência artificial. Na metodologia, a abordagem seguida para testar diferentes técnicas de inteligência artificial, como CNN, MLP, RF *Classifier* e KNN foi descrita, abordando o uso destas técnicas em ambientes escolares. Cada técnica foi testada individualmente e foi explorado ainda o impacto da remoção de características altamente

correlacionadas entre si no desempenho. Além disso, foram considerados diferentes intervalos de classes para compreender como afetariam os resultados dos modelos.

Por fim, os resultados dos testes em cada sala foram apresentados, destacando o desempenho de cada técnica e a influência das características ambientais. Concluí-se que, em geral, as técnicas tradicionais KNN e RF *Classifier* mantiveram uma acurácia acima de 90% e a CNN também apresentou resultados satisfatórios. No entanto, a técnica MLP revelou-se consistentemente com a pior acurácia, indicando que não será a mais indicada para o contexto deste trabalho. Este capítulo proporcionou uma análise detalhada do processo de aplicação de técnicas de inteligência artificial na classificação da ocupação de salas de aula. Isto permitiu uma compreensão mais profunda do desempenho das técnicas e da sua capacidade de generalização ao serem treinadas com diferentes salas de aula. Além disso, constatou-se que a remoção de características não tem um impacto significativo no desempenho, o que pode resultar em economia de recursos computacionais ou na escolha de sensores adequados sem comprometer os resultados.

Capítulo 7

Conclusão

O trabalho desenvolvido nesta dissertação foi concluído na sua totalidade, cumprindo de forma rigorosa os diversos objetivos definidos no âmbito da dissertação. Este percurso abarcou várias etapas, desde a especificação e implementação do sistema *Airmon* até à aplicação de algoritmos de inteligência artificial para a classificação da ocupação em salas de aula. No entanto, é importante reconhecer que existem áreas de aprimoramento e oportunidades de pesquisa futura que merecem destaque. Inicialmente, a fundamentação para o projeto envolveu uma pesquisa aprofundada de conceitos tecnológicos relacionados com sistemas de IoT e inteligência artificial, com base em artigos científicos e trabalhos académicos. Esta fase foi essencial para adquirir conhecimento sobre as componentes fundamentais dessas tecnologias e como poderiam ser aplicadas em contextos práticos. Posteriormente, a revisão da literatura centrou-se em estudos que exploraram a integração de IoT e inteligência artificial, com ênfase na qualidade do ar. Também foram analisados trabalhos relacionados diretamente com a temática de ocupação, com o intuito de compreender abordagens anteriores, identificar falhas e aprender com a montagem de sistemas, seleção de sensores e erros cometidos. Esta revisão permitiu uma base sólida para o desenvolvimento do projeto, evitando começar "do zero". A arquitetura do sistema *Airmon* foi concebida com rigor e clareza, estabelecendo os objetivos do projeto e delineando o processo de recolha de dados nas salas de aula. Foram definidos requisitos funcionais e não funcionais, além de estabelecer as bases para escalabilidade, desempenho, segurança e recuperação de falhas. A fase de implementação do sistema envolveu a seleção de sensores apropriados, como o DHT22, MH-Z19 e PMS5003, que foram escolhidos com base na sua compatibilidade com o HA, especificamente com a integração do *EspHome*. A flexibilidade do HA, que originalmente está mais voltado para a automação residencial, provou ser uma solução versátil e eficaz para o projeto. A configuração e envio de dados foram simplificados, poupando tempo no desenvolvimento. Além disso, a estratégia de criar instâncias locais em cada escola, que posteriormente comunicam com uma instância central, possibilitou a monitorização de várias escolas e a centralização de dados, enriquecendo o conjunto de dados resultante, sem prejudicar a capacidade de cada escola criar

o seu conjunto de dados. A fase de testes iniciais, realizada num ambiente controlado, validou a confiabilidade do sistema e da infraestrutura, demonstrando a viabilidade das escolhas de sensores e protocolos de comunicação. Em seguida, a implantação de caixas de monitorização na Universidade Fernando Pessoa, com foco em três salas de aula distintas e posteriormente na Escola Secundária de Alpendorada, validou a escalabilidade e a robustez da arquitetura e implementação do projeto. Durante a recolha de dados, um procedimento rigoroso foi seguido envolvendo os docentes e protocolos específicos para garantir a precisão das informações recolhidas em condições controladas. Os docentes desempenharam um papel crucial na colaboração, preenchendo questionários que incluíam informações sobre horários de aulas, número de pessoas presentes, estado dos sistemas de ventilação e mais, sem comprometer a privacidade dos intervenientes. Após a fase de recolha de dados, os conjuntos de dados foram elaborados com a devida limpeza e pré-processamento, para prepará-los para a aplicação de técnicas de inteligência artificial. Esta etapa, automatizada com *scripts Python*, permitiu também criar variáveis derivadas dos valores originais de CO_2, procurando padrões não visíveis nos dados originais. A análise dos conjuntos de dados, pela criação e observação de histogramas e matrizes de correlação, proporcionou uma compreensão mais profunda da distribuição da ocupação nas salas de aula e das relações entre as variáveis ambientais. A aplicação de técnicas de inteligência artificial para a classificação de ocupação, foi uma fase crucial do projeto. Foram testadas técnicas, como KNN, CNN, MLP e RF *Classifier*, em cada sala, com intervalos de classes de ocupação de cinco e três. O desempenho de cada técnica foi cuidadosamente avaliado, revelando que o KNN, RF *Classifier* e CNN obtiveram resultados especialmente satisfatórios. Além disso, a possibilidade de treinar um modelo geral para todas as salas, embora com uma ligeira diminuição na acurácia, destacou a capacidade de generalização das técnicas. Embora todos os objetivos tenham sido alcançados, existem várias áreas de melhoria que servem de sugestão para futuras investigações. Uma delas seria a recolha de dados durante as estações de outono e inverno, complementando e enriquecendo o conjunto de dados com informações de todas as estações do ano. Além disso, a possibilidade de classificar a ocupação sem recorrer a intervalos de classes, utilizando dados brutos, representa uma abordagem interessante a explorar. O aumento do número de salas e escolas monitorizadas contribuiria para a expansão dos conjuntos de dados e testaria ainda mais a fiabilidade do sistema *Airmon*. Por último, a aplicação dos conjuntos de dados para estudos relacionados com a previsão da ocupação, classificação da qualidade do ar e outras análises mais abrangentes poderiam ser alvo de investigações posteriores, oferecendo novas perspetivas e aprofundamento.

Referências

Adafruit (2022). Dht22, https://www.adafruit.com/product/385. Accessed: 2022-11-17.

Alsmirat, M., Jaraweh, Y., of Electrical, I., Section, E. E. S., de Valencia, U. P., de Granada, U., Graz, T. U., University, S., of Science & Technology, J. U., of Electrical, I. and Engineers, E. (2019). Indoor occupancy prediction using an iot platform.

Beechen, S. (2023). Google drive backup in homeassistant, https://community.home-assistant.io/t/add-on-home-assistant-google-drive-backup/107928. Accessed: 2023-01-27.

Bockstael, N. and Jadin, A. (2018). Co2 based room occupancy detection: an iot and machine learning application. Prom.: Schaus, Pierre.

Bogdanovica, S., Zemitis, J. and Bogdanovics, R. (2020). The effect of co2 concentration on children's well-being during the process of learning, *Energies* **13**.

Candanedo, L. M. and Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models, *Energy and Buildings* **112**: 28–39.

Electronics, M. (2022). Esp32 mouser, https://pt.mouser.com/ProductDetail/Espressif-Systems/ESP32-DevKitC-32UE?qs=GedFDFLaBXFguOYDKoZ3jA%3D%3D. Accessed: 2022-11-16.

EspHome (2023). Esphome, https://esphome.io/index.html. Accessed: 2023-01-19.

Farnham, B., Tokyo, S., Boston, B., Sebastopol, F. and Beijing, T. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION*, pp. 8–15.

Grafana (2023). Grafana in homeassistant, https://community.home-assistant.io/t/home-assistant-community-add-on-grafana/54674. Accessed: 2023-01-25.

HomeAssistant (2023). Homeassistant, https://www.home-assistant.io/getting-started/. Accessed: 2023-01-10.

InfluxDB (2023). Influxdb in homeassistant, https://www.home-assistant.io/integrations/influxdb/. Accessed: 2023-01-25.

Kim, Y., Park, Y., Seo, H. and Hwang, J. (2023). Load prediction algorithm applied with indoor environment sensing in university buildings, *Energies* **16**.

Lee, J. Y., Miao, Y., Chau, R. L., Hernandez, M. and Lee, P. K. (2023). Artificial intelligence-based prediction of indoor bioaerosol concentrations from indoor air quality sensor data, *Environment International* **174**.

Liu, J., Zhang, R. and Xiong, J. (2023). Machine learning approach for estimating the human-related voc emissions in a university classroom, *Building Simulation* **16**: 915–925.

Marzouk, M. and Atef, M. (2022). Assessment of indoor air quality in academic buildings using iot and deep learning, *Sustainability (Switzerland)* **14**.

Mohammadabadi, A., Rahnama, S. and Afshari, A. (2022). Indoor occupancy detection based on environmental data using cnn-xgboost model: Experimental validation in a residential building, *Sustainability (Switzerland)* **14**.

Moreira, R. S., Soares, C., Torres, J. M. and Sobral, P. (2020). *Combining IoT architectures in next generation healthcare computing systems*, Elsevier, pp. 1–29.

Nginx (2023). Nginx in homeassistant, https://github.com/home-assistant/addons/tree/master/nginx_proxy. Accessed: 2023-02-01.

Norvig, P. and Russell, S. J. (2021). *Artificial Intelligence A Modern Approach Fourth Edition*, p. 811.

of Technology, S. I., of Electrical, I. and Engineers, E. (2017). *Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017) : 7-8, December 2017*, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad.

Patel, S., Salazar, C., Patel, K. K., Patel, S. M. and Scholar, P. G. (2016). Internet of things-iot: Definition, characteristics, architecture, enabling technologies, application & future challenges, *International Journal of Engineering Science and Computing* . **URL:** *http://ijesc.org/*

Plantower (2022). Pms5003, https://www.plantower.com/en/products_33/74.html. Accessed: 2022-11-17.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification, *ISPRS Journal of Photogrammetry and Remote Sensing* **67**: 93–104.

Saxena, S., Jain, S., Arora, D. and Sharma, P. (2019). Implications of mqtt connectivity protocol for iot based device automation using home assistant and openhab.

Singh, S. and Singh, N. (2015). Internet of things (iot): Security challenges, business opportunities & reference architecture for e-commerce, *Internet of Things (IoT): Security challenges, business opportunities & reference architecture for E-commerce*.

Tekler, Z. D. and Chong, A. (2022). Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy, *Building and Environment* **226**.

Wang, J., Du, W., Lei, Y., Chen, Y., Wang, Z., Mao, K., Tao, S. and Pan, B. (2023). Quantifying the dynamic characteristics of indoor air pollution using real-time sensors: Current status and future implication.

Winsen (2022). Mhz-19, `https://www.winsen-sensor.com/product/mh-z19b.html`. Accessed: 2022-11-16.

Yang, B., Haghighat, F., Fung, B. C. and Panchabikesan, K. (2021). Season-based occupancy prediction in residential buildings using machine learning models, *e-Prime - Advances in Electrical Engineering, Electronics and Energy* **1**.

Zemitis, J., Bogdanovics, R. and Bogdanovica, S. (2021). The study of co2concentration in a classroom during the covid-19 safety measures, *E3S Web of Conferences* **246**.

Zhang, Z., Rahman, S., Yu, G. and Ampadu, P. K. (2023). Building occupancy analytics based on deep learning through the use of environmental sensor data.