Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Aprendizagem automática aplicada na administração pública - Caso da gestão de refeitórios escola

Stelmo Abel da Fonseca Ferreira Barbosa

Orientador(es) | Teresa Gonçalves

Luís Rato

Évora 2025

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Aprendizagem automática aplicada na administração pública - Caso da gestão de refeitórios escola

Stelmo Abel da Fonseca Ferreira Barbosa

Orientador(es) | Teresa Gonçalves

Luís Rato

Évora 2025

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Lígia Maria Ferreira (Universidade de Évora)
Vogais	|	José Saias (Universidade de Évora) (Arguente) Luís Rato (Universidade de Évora) (Orientador)

Évora 2025

Ao Lucas - o que traz luz.

Agradecimentos

Nunca estamos sozinhos, há sempre outros que contribuem para atingirmos as nossas metas. É certo que esta meta sucedeu outras que contribuíram para que chegasse até aqui. No entanto, este tempo que dediquei ao Mestrado e a esta dissertação, tiveram intervenientes mais diretos, pessoas que contribuíram muito para que eu pudesse concretiza-la. Quero agradecer em primeiro lugar a estas pessoas que, generosamente, dispensaram o seu tempo e partilharam o seu conhecimento e muitas vezes me motivaram a ir mais além.

À Professora Teresa Gonçalves e ao Professor Luís Rato agradeço pelo entusiasmo, disponibilidade e pela paciência. Aprendi tanto com as reuniões que mantivemos. A sua orientação foi realmente decisiva para que pudesse concluir o trabalho. Contagiaram-me com a vontade de fazer ciência, de questionar mais e procurar mais respostas.

Ao Sr. João Inês e ao Professor Eduardo Cunha agradeço pelo apoio prestado na disponibilização de informação. Este trabalho é também feito por vós e na expetativa de que possa contribuir de alguma forma para criar inovação nos serviços. Esta colaboração foi preciosa e exemplar, porque, enquanto estudantes ou investigadores, nem sempre se consegue ter acesso à informação de forma ágil.

À Vanessa Flórido também devo dizer obrigado, porque sem ela, não conseguia. O tempo que me deste permitiu realizar isto.

Termino como comecei, há um trajeto educativo muito maior pelo que passei e que me fez ser capaz. Deixo aqui a homenagem à minha mãe Rosalina, ao meu pai Abel (trago-vos em mim) e a todos os professores que tive.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AA Aprendizagem Automática

AD Árvore de Decisão

GB Gradient Boosting

IA Inteligência Artificial

KNN K-nearest neighbors

MAE Mean Absolute Error

MLP Multi-Layer Perceptron

MSE Mean Squared Error

R2 Coeficiente de Determinação

RF Random Forest

RL Regressão linear

RMSE Root Mean Squared Error

RNA Redes Neuronais Artificiais

SVM Support Vector Machine

SVR Support Vector Regression

VRSA Vila Real de Santo António

XGB Extreme Gradient Boosting

Sumário

A administração pública Portuguesa percorreu um caminho de digitalização de serviços que hoje já permite a adoção de métodos de aprendizagem automática, aproveitando os repositórios de dados que mantêm.

Com base nos dados recolhidos em dois refeitórios escolares, de Tavira e de Vila Real de Santo António, entre 2022 a 2023, esta dissertação investiga a aplicação de aprendizagem automática nos domínios da administração pública, focando-se na previsão do número de refeições servidas em refeitórios. O objetivo principal é melhorar a gestão e minimizar desperdícios, utilizando dados históricos para prever a procura diária.

Foram analisados outros trabalhos com o mesmo âmbito e diferentes algoritmos de regressão, tais como *Support Vector Regression, Random Forest* ou *Extreme Gradient Boosting*. Entre os algoritmos utilizados, concluiu-se que Random Forest e *Extreme Gradient Boosting* tiveram o melhor desempenho com um erro médio absoluto de cerca de 38 e 41 refeições, respetivamente.

Ficam evidenciados os benefícios da tomada de decisões baseada em Aprendizagem Automática, utilizando dados da própria administração pública, demonstrando que as técnicas de aprendizagem automática podem melhorar a eficiência dos refeitórios escolares e este pode ser também um caminho para ajudar na redução do desperdício alimentar.

Palavras chave: Aprendizagem automática, refeitórios, administração pública, inteligência artificial

Abstract

Machine learning in the public administration

The case of public catering

The Portuguese public administration has embarked on a path toward digitalizing services that now allows the adoption of machine learning methods, leveraging the data repositories it maintains.

Based on data collected from two school cafeterias, in Tavira and Vila Real de Santo António, between 2022 and 2023, this work investigates the application of machine learning in public administration, focusing on predicting the number of meals served in cafeterias. The main objective is to improve management and minimize waste by using historical data to forecast daily demand.

Other studies with the same scope and different regression algorithms, such as Support Vector Regression, Random Forest, and Extreme Gradient Boosting, were analyzed. Among the algorithms used, Random Forest and Extreme Gradient Boosting, performed best, with mean absolute error of approximately 38 and 41 meals, respectively.

The benefits of Machine Learning based decision-making using data from the public administration itself are evident, demonstrating that machine learning techniques can improve the efficiency of school cafeterias and can also be a way to help reduce food waste.

Keywords: Machine learning, canteen, public administration, artificial intelligence

Capítulo 1

Introdução

Este capítulo, para além de analisar a evolução da adoção das tecnologias de informação na administração pública nos últimos 20 anos, também descreve, de forma sucinta, a gestão dos refeitórios escolares, apresenta os fatores que motivaram a realização deste trabalho e os seus objetivos. Por fim descreve as principais contribuições que resultam deste trabalho.

1.1 Inteligência Artificial na administração pública Portuguesa

Anualmente a Direção-Geral de Estatísticas da Educação e Ciência realiza um inquérito dirigido à administração pública, com o objetivo de conhecer o estado da adoção e utilização de tecnologias de informação [9]. Desde o ano de 2021, incluíram no inquérito questões sobre a utilização de Inteligência Artificial (IA).

No relatório publicado em maio de 2024 [10], 25% das Câmaras Municipais, 20% dos Organismos da Administração Central, 13% dos Organismos da Região Autónoma da Madeira e 7% dos Organismos da Região Autónoma dos Açores, indicaram ter utilizado tecnologias de IA. As tarefas em que foi utilizada passam por identificação de objetos ou pessoas, análise de linguagem escrita, conversão de áudio em texto e até Aprendizagem Automática (AA) [10].

As competências da administração pública englobam as mais diversas atividades que vão muito além da criação e aplicação de legislação. Por exemplo uma câmara municipal por si só tem competências sobre a gestão urbanística, trânsito, transportes, recursos hídricos, proteção civil ou educação, entre outros [5]. E, para cada uma dessas áreas, há necessariamente atividades de gestão corrente que obrigam ao tratamento de dados das mais diversas tipologias como geográficos, financeiros ou dados pessoais.

Nos últimos 20 anos assistimos a várias iniciativas governamentais e programas de

financiamento da União Europeia, que promoveram a adoção de tecnologias de informação a todos os níveis (local, regional e central), como são os casos do Programa Simplex [12] e do Portugal 2020 [11]. Em 2013 a utilização do correio eletrónico e o acesso à *internet* eram de uso generalizado na administração central [4] e 87% dos organismos da administração central tinham sítios eletrónicos para divulgação de informação, mas não dispunham de serviços online. Nesse mesmo ano, nas autarquias o panorama não era muito distinto; a *internet* era utilizada sobretudo para consulta de informação ou correio eletrónico e 84% das Câmaras Municipais tinham sítios eletrónicos na *internet*, ainda sem serviços eletrónicos transacionais. O panorama em 2023 teve mudanças significativas [10]. Na administração central 95% dos organismos têm presença online [9] e 39% têm balcão online. Todos os municípios têm sítio eletrónico e 67% tem balcão virtual. Além disso, 25% dos municípios já adotaram ferramentas de IA [10], e 37% têm redes de sensores para gestão urbana (IOT-*Internet Of Things*). Hoje a administração pública, de forma generalizada, gere os seus serviços com meios informáticos e por isso concentra grandes volumes de dados em formato digital que abrangem não só documentação mas também bases de dados das várias áreas de atuação dos serviços. Todos estes dados são potencialmente utilizáveis para criar modelos de aprendizagem automática quer seja para classificação ou regressão, e que sirvam de apoio à gestão.

Foi com esta premissa que foi lançado pela Fundação para a Ciência e Tecnologia (FCT) em 2018 o Programa de Investigação em Ciência de Dados e Inteligência Artificial na Função Pública [6], para apoiar projetos de investigação e desenvolvimento entre instituições científicas e organismos públicos. Foram financiados 32 projetos que abrangem áreas na gestão de recursos naturais, educação, segurança pública, saúde, mobilidade e transportes, todos baseados em dados detidos ou geridos por entidades públicas. Nestes projetos foram utilizadas técnicas de regressão para prevenção de acidentes, deteção de avarias, prever o risco de complicações cirúrgicas, serviços de apoio com processamento de linguagem natural, sistemas de apoio à decisão na medicina, entre outros. Fica assim evidente que a administração pública é fonte de dados para diversos projetos do seu próprio interesse. A vontade de continuar a explorar as possibilidades da IA na administração pública ficou espelhada na resolução do Conselho de Ministros n.°131/2021 [8], que aprovou a Estratégia para a Transformação Digital da Administração Pública 2021-2026 alicerçada em seis linhas estratégicas de atuação:

- Serviços públicos digitais;
- Valorização dos dados;
- Arquiteturas de referência;
- Competências TIC;
- Infraestruturas e serviços TIC; e
- Segurança e confiança.

Todas estas iniciativas têm promovido a transformação digital dos serviços e a valorização dos dados, como elementos fundamentais para os processo de decisão.

1.2 Refeitórios escolares

Os refeitórios das escolas do ensino básico em Portugal são uma responsabilidade das autarquias. As Câmaras Municipais optam por um dos seguintes modelos de gestão dos serviços de refeições: 1) as refeições são confecionadas na cozinha da escola, ou 2) as refeições são adquiridas pré-confecionadas. Esta escolha depende dos recursos humanos disponíveis nas autarquias e também da vantagem financeira que daí advém. Nas escolas referidas neste estudo, os refeitórios confecionam as refeições.

No dia a dia, a gestão enfrenta o problema de providenciar o número de refeições necessárias para todos os estudantes que vão almoçar, minimizando o desperdício alimentar. Para atenuar esta situação, incentivam a comunidade escolar a comprar senhas de refeição antecipadamente. No entanto, existem sempre alunos que fazem a compra no próprio dia e, por outro lado, há alunos que apesar de terem senhas pré-compradas, não vão almoçar ao refeitório, o que representa sempre um grau de incerteza. Este grau de incerteza é ultrapassado pelo conhecimento do histórico de consumo e pela intuição dos responsáveis pela gestão do refeitório.

1.3 Motivação

Prever o comportamento humano é um exercício desafiante. As condicionantes para a tomada de uma decisão podem ser inúmeras e variadas dependendo da situação. E cada situação tem necessariamente condicionantes mais pertinentes que outras. Ir à praia, escolher um filme para ver ou optar por almoçar no refeitório, são todos exemplos de decisões que se tomam com base em condicionantes que podem estar relacionadas com o nosso estado de espírito, meteorologia, capacidade financeira ou o dia da semana. E estas condicionantes podem estar conjugadas, o que aumenta o grau de imprevisibilidade do comportamento humano.

Num contexto escolar, existem também diversas variáveis que influenciam os alunos a utilizar os refeitórios. A meteorologia, as épocas festivas, eventos decorrentes do calendário escolar, tais como exames, ou greves, podem ser fatores determinantes para afluência aos refeitórios. Para criar um modelo de AA que consiga prever o número de refeições num determinado dia, é necessário avaliar o grau de correlação entre diversos fatores e as refeições consumidas, o que de certo modo é tentar encontrar alguma previsibilidade do comportamento humano. Este desafio foi a principal motivação para este trabalho. Evidentemente o acesso a um historial de dados relativos a refeitórios escolares e aos fatores correlacionados, foi indispensável para este trabalho, para além das fontes disponíveis na *internet*, e dos dados gerados e

mantidos nos organismos públicos.

A administração pública em Portugal passou um período de grande investimento em tecnologias de informação e hoje produz e gere dados que por si só poderão servir para implementar sistemas de apoio à decisão que tornem a gestão pública mais eficiente e fundamentada. Este trabalho, pretende também, exemplificar esta possibilidade, de gerir um serviço público, neste caso refeitórios escolares, com base em dados reais gerados pela dinâmica própria do serviço.

1.4 Objetivos

Nos refeitórios das escolas que forneceram os dados para este trabalho, não é aplicada nenhuma metodologia formal para prever o número de refeições a servir. A venda de senhas de refeições está informatizada, mas, os dados servem apenas para a gestão financeira e contabilística. Estes registos diários constituem uma fonte de dados apropriada para modelos de AA. Este trabalho alicerça-se no estudo deste conjunto de dados e pretende procurar fatores correlacionados que incorporem um modelo para prever as refeições a servir nos refeitórios.

Além do modelo de AA resultante, pretende-se seguir uma metodologia que sirva de guia para trabalhos similares, deixando evidências dos resultados que facilitem as decisões de outros estudos.

Por fim, pretende-se também possibilitar a aplicação prática de AA num contexto que é real, na administração pública, demonstrando os seus benefícios e identificando as dificuldades inerentes ao processo.

1.5 Estrutura

Esta dissertação foi estruturada para descrever todo o trabalho realizado mas também para prestar alguma informação sobre AA que permita guiar o leitor acerca da metodologia utilizada. O Capítulo 2 introduz conceitos de AA, sobretudo sobre modelos de regressão e métodos para medir o seu desempenho. O Capítulo 3 analisa trabalhos que versam sobre a mesma temática e que deram pistas e ideias para o presente trabalho. O Capítulo 4 descreve a metodologia e ferramentas utilizadas. O quinto Capítulo analisa os dados, os atributos e relata o tratamento dado em cada fase de experiências. O Capítulo 6 analisa os resultados obtidos e por último, o Capítulo 7 expõe as conclusões, tentando indicar alguns caminhos a desenvolver a partir daqui.

1.6 Contribuições

O culminar deste trabalho, deixa um modelo de AA que poderá ser aplicado como ferramenta de apoio à gestão de refeitórios escolares, podendo este ser melhorado e dimensionado com novos atributos. Os estudos realizados sobre os atributos, contribuem com evidências que facilitarão a construção de outros modelos. O conjunto de dados utilizado está disponível publicamente na plataforma KAGGLE em https://www.kaggle.com/datasets/stelmoabelbarbosa/refeitoriosescolares.

Demonstram-se, ainda, as vantagens da aplicação de AA para gestão de refeitórios escolares, como uma oportunidade de melhoria para administração de serviços públicos.

Capítulo 2

Aprendizagem Supervisionada

Este capítulo introduz os conceitos teóricos, inerentes à AA, que sustentam a metodologia utilizada ao longo do trabalho. Resume a definição de três grandes áreas de AA e desenvolve os conceitos de alguns algoritmos de regressão.

2.1 Conceitos

Arthur Samuel um dos pioneiros da Inteligência Artificial definiu em 1959 a Aprendizagem Automática como "área de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programados" [39]. Esta capacidade de aprendizagem está baseada em métodos estatísticos para criar modelos capazes de aprender a partir de conjuntos de dados.

A Aprendizagem Automática, subdivide-se em três grandes áreas [40]:

- Supervisionada

A aprendizagem supervisionada é utilizada quando se pretende prever um resultado a partir de determinados dados de entrada e tendo à partida um conjunto de exemplos entrada/saída que servirão para treinar o modelo [27]. Os dados de saída, as soluções associados aos exemplos, designam-se por etiquetas. Na aprendizagem supervisionada, o modelo aprende relacionando os exemplos e interpretando padrões para os resultados. Nesta tipologia encontram-se dois problemas comuns: classificação e regressão. A classificação tem por objetivo descobrir uma etiqueta ou classe, enquanto a regressão tenta realizar uma previsão de valores reais.

- Não Supervisionada

A aprendizagem não supervisionada realiza-se com conjuntos de dados não etiquetados. Ou seja, não existe um resultado associado aos diversos exemplos [24]. Na aprendizagem não supervisionada, são apresentados dados ao algoritmo, para ser extraída informação [27].

Uma aplicação comum é a redução de dimensão. Considera-se que um conjunto de dados de alta dimensão é um conjunto com um número elevado de atributos. A redução de dimensão encontra uma nova maneira de representar esses dados, que resume as características essenciais com menos recursos. Outra aplicação é o agrupamento [22], onde os dados são agrupados conforme as suas semelhanças ou diferenças.

- Por Reforço

Na aprendizagem por reforço os modelos são continuamente treinados. E, mais importante, é que os modelos também são responsáveis por tomar decisões pelas quais recebem uma recompensa periódica. [26].

O sistema de aprendizagem, denominado agente, pode observar o ambiente, selecionar e executar ações e, com isso, obter recompensas [24]. O algoritmo avalia os resultados de modo a encontrarem a melhor estratégia para obter sempre a maior recompensa.

2.1.1 Treino-validação-teste

Para que os algoritmos de AA funcionem corretamente é necessário reunir milhares ou milhões de exemplos [24]. Ao construir um modelo de AA, não podemos utilizar todo o conjunto de exemplos recolhidos para o treinar, uma parte do conjunto de dados deve ser reservada para avaliar o desempenho do modelo. De outro modo, ao testar o modelo, este iria sempre devolver a etiqueta de dados correta e não teria um bom desempenho com novos dados [27].

A execução de um modelo de AA passa, genericamente, por três fases [26]:

- Fase de Treino: Esta é a fase em que os dados de treino são usados para treinar o modelo;
- Fase de Validação e Teste: Esta fase serve para medir a qualidade do modelo com um conjunto de dados de validação e teste;
- Fase de Aplicação: Nesta fase, o modelo é sujeito aos novos dados do mundo real.

Conjunto de treino é um subconjunto dos dados recolhidos, utilizado para o algoritmo aprender. Com base no conjunto de treino, o algoritmo induz uma função que mapeia novos exemplos à sua correspondente previsão ou classe [22].

O conjunto de validação é um subconjunto dos dados de treino utilizado para ajustar a precisão de saída do algoritmo, através da variação dos seus hiperparâmetros [26].

Conjunto de testes é um subconjunto de dados, utilizado para o testar o desempenho do algoritmo. Consoante esta avaliação, o algoritmo será considerado bem sucedido ou não [22].

Por vezes a divisão entre exemplos para teste e treino pode resultar num conjunto de testes insuficientemente representativo do resto do conjunto, e nestes casos avaliação de desempenho pode dar resultados demasiado pessimistas ou demasiado otimistas. Para evitar esta situação, na aprendizagem supervisionada, utiliza-se o processo de validação cruzada [29]. Em validação cruzada, o conjunto de dados de treino é dividido em n partes iguais e o algoritmo é treinado com $n-1$ partes e testado com a parte sobrante. Este procedimento repete-se n vezes, sempre com uma parte de teste diferente. A média dos resultados em cada teste, é a avaliação de desempenho global do algoritmo [22].

2.1.2 Métodos de regressão

O problema de previsão do número de refeições a serem confecionadas num refeitório escolar enquadra-se na regressão. No âmbito da AA, define-se regressão como a tarefa para realizar uma previsão de um valor numérico [24]. Para o âmbito deste trabalho optou-se por avaliar o desempenho dos seguintes algoritmos de aprendizagem supervisionada:

- Regressão linear
- Árvores de decisão
- Máquina de Vetores de Suporte
- Floresta aleatória
- *Gradient Boosting*

2.2 Algoritmos

2.2.1 Regressão linear

Este é o método mais antigo utilizado para fazer previsões baseadas em conjuntos de dados etiquetados [27]. O termo Análise de Regressão define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e prever o valor de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (ou predictoras) [31]. Denomina-se, portanto, variável dependente y, aquela cujo comportamento será explicado pela variável (ou variáveis) X. O algoritmo de regressão linear procura minimizar o erro entre as previsões e os valores reais. Graficamente o modelo representa uma reta e o algoritmo procura estimar a inclinação dessa reta utilizando uma amostra aleatória de dados de X e y. Isto implica que haja de facto uma relação entre X e y. O resultado da regressão linear, obtém-se através da soma valorada dos atributos associados ao modelo.

A regressão linear (RL) simples é descrita pela Equação 2.1. Se não houver uma relação entre a variável dependente e a independente, isto significa que β=0. Por outras palavras, quando mais β divergir de zero, maior o grau de correlação com as variáveis independentes.

$$y = \alpha + \beta X + \epsilon \tag{2.1}$$

y: Variável dependente, que será prevista pelo modelo;

α: Constante, que representa a interceção da reta com o eixo vertical;

β: Coeficiente angular, representando a taxa de variação em y para a variação de uma unidade em X;

X: Variável explicativa (independente);

ϵ: Termo de erro que representa a diferença entre o valor observado e o valor previsto.

A finalidade da regressão linear é estimar os valores de α e β que minimizam a soma das diferenças quadradas entre os valores observados e previstos de y, usando, por norma, o método dos mínimos quadrados [34].

2.2.2 Árvores de decisão

As árvores de decisão (AD) servem para resolver problemas de classificação ou regressão [40]. Este algoritmo começa por dividir o conjunto de dados em segmentos menores com base nos atributos (variáveis independentes). A cada passo procura o melhor ponto de divisão. O processo é repetido recursivamente para cada subconjunto de dados gerado, até que algum critério de paragem seja alcançado, por exemplo, o número máximo de divisões estabelecido. A estrutura resultante é uma árvore, onde cada caminho da raiz até à folha é uma sequência de decisões que podem ser descritas por regras lógicas. Através da árvore pode-se mapear todas as conjugações possíveis de valores baseados nos dados recolhidos. Em problemas de classificação, as variáveis dependentes assumem um númeroă discreto de classes, sem significado métrico intrínseco, no caso de regressão as variáveis dependentes assumem valores numéricos nos quais diferenças e distâncias são semanticamente relevantes. [30].

Tomemos como exemplo o conjunto de dados na Tabela 2.1, que reúne informação ficcional sobre os preços da habitação. Uma árvore de decisão gerada com base nestes dados poderia ser a demonstrada na Figura 2.1. Transpondo a tabela para uma árvore, as diferentes opções, de cada coluna, ramificam-se até à sua terminação que deverá ser o atributo a classificar ou prever. Para determinar o preço de uma habitação, conforme os dados da Tabela 2.1, temos de passar por 2 questões, a primeira é identificar a localidade, que tem 2 opções possíveis. Cada ponto de decisão, é um nó que deriva para outra questão com outras opções. Percorrendo todas as

questões, a localização e a tipologia, e optando por um dos possíveis nós, obtemos o preço correspondente, que identificamos como as terminações da árvore (as folhas).

Localização	Tipo	Preço
Centro	T1	150.000
Praia	T1	145.000
Praia	T2	176.000
Centro	T2	160.000

Tabela 2.1: Preço médio de habitação (dados ficcionados)

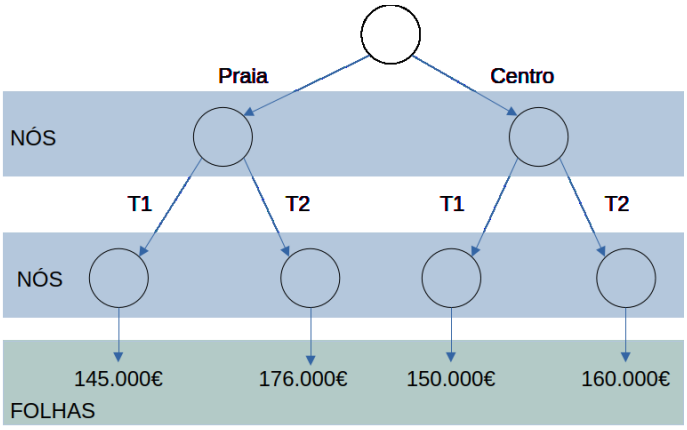

Figura 2.1: Árvore de decisão.

Um dos algoritmos heurísticos utilizados na construção de árvores de decisão é o CART (*Classification And Regression Tree*) [20], que nos casos de classificação, a métrica utilizada para escolher o melhor ponto de divisão é dada pelo índice de Gini [30], que mede a impureza dos conjuntos de dados. Entende-se aqui por impureza, o grau em que os exemplos de treino não pertencem a uma mesma classe [29]. A impureza da divisão é uma média ponderada da impureza de cada nó filho. Em problemas de regressão este grau de impureza é dado pela soma dos quadrados dos erros (SSE[1]). Para um nó t que é dividido em dois nós filhos, tL (esquerda) e tR (direita), a redução da impureza é medida pela diferença na SSE do nó pai e a soma das SSEs dos nós filhos. Este processo de divisão para quando todos os nós atingem um nível suficiente de pureza, quando o número de pontos por folha se torna muito pequeno para uma divisão posterior ou com base em alguma outra heurística semelhante [32].

Uma das vantagens que tem a sua utilização é a facilidade em lidar com diversos tipos de dados, outra é que a escolha dos atributos mais relevantes é feita automaticamente em cada caso [41].

[1]Do Inglês Sum Squared Error

2.2.3 Máquina de Vetores de Suporte

O algoritmo Máquina de Vetores de Suporte (SVM[2]) é recomendado para conjuntos de dados médios ou pequenos [24] e pode ser utilizado em classificação ou regressão. O algoritmo trata cada amostra do conjunto de dados como um ponto num hiperplano. A partir daqui procura definir o plano que melhor separa as diversas classes de dados num mesmo hiperplano [38] como se retrata na Figura 2.2.

Para cada exemplo de um conjunto de dados, o algoritmo identifica, de duas possiveis, a classe a que pertence. O que faz do SVM um classificador linear binário não probabilístico. Um modelo resultante é uma representação de exemplos de dados como pontos no espaço, mapeados de maneira que os exemplos de cada classe estejam divididos por um espaço tão amplo quanto possível [19].

Quando estamos perante classes linearmente separáveis a margem é fixa. No entanto, quando as classes não são completamente separáveis esta margem tem de ter alguma tolerância e nestas circunstancias denomina-se *soft-margin* SVM [19]. Isto é, a maioria dos dados são classificados corretamente, mas é permitido que o modelo classifique incorretamente alguns pontos na vizinhança do limite de separação, como se exemplifica na Figura 2.3.

Esta flexibilidade de *soft-margin* SVM, pode não ser capaz de separar as classes minimizando o número de pontos classificados incorretamente. Nestes casos utiliza-se uma função (kernel) para transformar os dados num espaço de dimensão superior, onde possam ser linearmente separados. Algumas das funções mais utilizadas para este efeito são *Linear Kernel*, Polinomial, Tangente Hiperbólica ou *Gaussian radial* [19].

A regressão por vetores de suporte (SVR [3]) é uma adaptação das SVMs, direcionada a problemas de regressão. Tal como as SVMs lineares, o SVR encontra um hiperplano com a margem máxima entre os pontos de dados e é normalmente usado para a previsão de séries temporais [21].

2.2.4 Floresta aleatória

O algoritmo Floresta Aleatória (RF[4]) enquadra-se nos métodos de comité. Este método procura as vantagens individuais de cada classificador, combinando-as de forma a obter a melhor solução. No caso concreto de RF, são combinadas várias árvores de decisão para reduzir a variância dos modelos individuais [37]. O método começa por criar várias árvores de decisão, em que cada árvore é treinada com uma amostra, extraída aleatoriamente, dos dados. Na divisão de cada nó da árvore, é selecionado, aleatoriamente, um subconjunto de variáveis para realizar a divisão, o que aumenta a diversidade das árvores.

[2]Do Inglês Support Vector Machine
[3]Do Inglês Support Vector Regression
[4]Do Inglês Random Forest

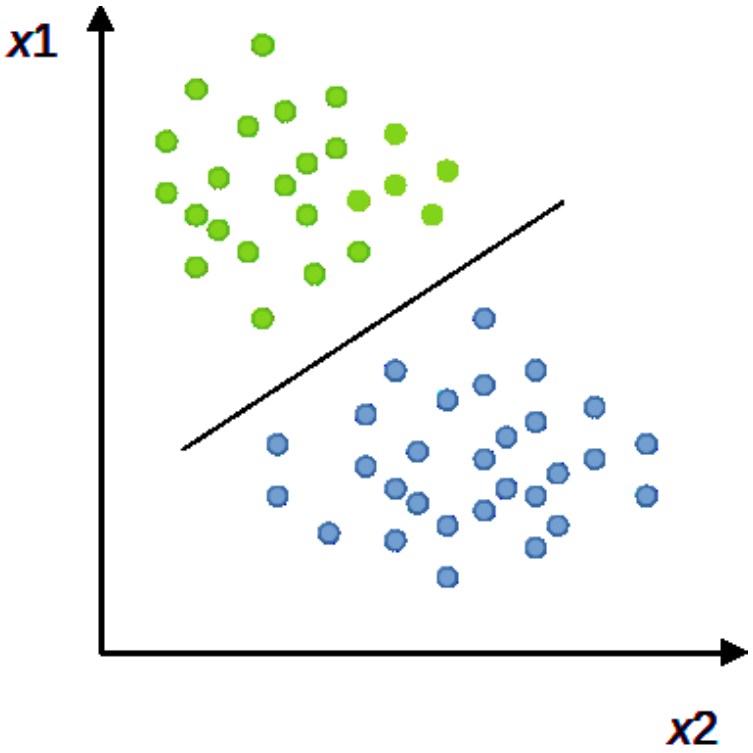

Figura 2.2: Representação gráfica do resultado de SVM.

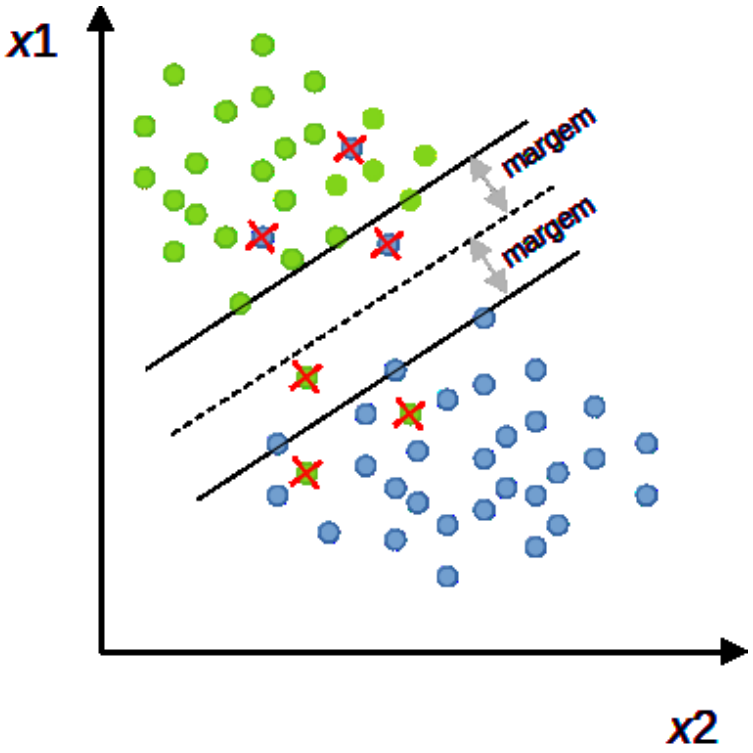

Figura 2.3: Classificação errada de alguns pontos como parte da margem.

Como existem tantos resultados como árvores de decisão combinadas, então o resultado final, num problema de regressão, é obtido pela média dos vários resultados obtidos em cada árvore. Num problema de classificação, o resultado mais frequente corresponderá à classe prevista [28].

É um método mais resistente a *overfitting* (treinar para além dos valores mais otimizados) do que quando se utiliza uma única árvore de decisão, especialmente com grandes conjuntos de dados e não requer demasiados ajustes de parâmetros para obter bons resultados [37].

2.2.5 Gradient Boosting

O algoritmo *Gradient Boosting* (GB) também combina um conjunto de árvores de decisão para construir o modelo e para este efeito utiliza a técnica denominada *Boosting*. Pode ser aplicado a problemas de regressão ou classificação e é robusto mesmo com *outliers* (dados que saem da norma), apesar disso é mais indicado para conjuntos de dados pequenos [29].

Do mesmo modo que o algoritmo RF, *Boosting* combina os resultados de várias AD, mas em vez de decidir com base numa média de resultados ou na classe mais recorrente, esta técnica cria árvores sequencialmente que vão melhorando o resultado das árvores anteriores [33].

O processo divide-se pelas seguintes etapas [23]:

Inicialização: Numa primeira fase é treinada uma única árvore de decisão e de seguida vão sendo adicionados mais modelos para tentar corrigir os erros anteriores.

Cálculo de resíduos: Para cada exemplo de treino, é calculado o erro residual, subtraindo o valor previsto ao valor real. A função utilizada para efetuar este cálculo denomina-se de função de perda. Esta etapa identifica as áreas onde as previsões do modelo podem ser melhoradas.

Regularização: Após o cálculo dos resíduos e antes de se iniciar o treino de um novo modelo, ocorre a regularização. Esta etapa tem por objetivo reduzir a influência de cada nova árvore de decisão (denominadas de modelos fracos) integrada no conjunto. Com isto é possível controlar a rapidez com que o algoritmo de reforço avança otimizando o seu desempenho.

Treino do modelo sucessor: Através dos erros residuais calculados na etapa anterior é treinado um novo modelo, com o propósito de corrigir os erros cometidos pelos modelos anteriores e melhorar a previsão geral.

Atualização do conjunto: Nesta etapa, o desempenho do conjunto atualizado é avaliado através de um conjunto de testes separado. Se o desempenho neste conjunto de dados de teste for satisfatório, o conjunto pode ser atualizado incorporando este novo modelo; caso contrário, pode ser necessário ajustar os hiperparâmetros .

Repetição: As etapas anteriores repetem-se até atingir os critérios de paragem.

Critérios de paragem: O processo é interrompido quando algum dos critérios de paragem definido, for atingido, estes podem ser; um número máximo de iterações, precisão ou retornos decrescentes.

O algoritmo de *boosting* pode também ser interpretado como uma forma de *Gradient Descendent* (gradiente descendente) que otimiza a função de perda em cada iteração [33].

A função de perda utilizada, depende do tipo de problema. Para problemas de regressão usa-se a erro quadrático médio, e para classificação a perda logarítmica (*Logarithmic loss*).

2.3 Medidas de desempenho

O processo de construção de um modelo de AA passa necessariamente por treina-lo com um subconjunto de dados e avaliar o seu desempenho com outros dados, diferentes dos utilizados na fase de treino [39]. Para fazer esta avaliação e medir o desempenho, existem diversas métricas [27]. É importante escolher a métrica correta conforme a finalidade do modelo; cada métrica devolve determinado tipo de informação e pode ser útil analisar várias métricas em conjunto.

Num problema de regressão pretende-se prever um valor numérico. Estas previsões terão uma diferença com os valores reais e é este desfasamento que se pretende analisar. Destacam-se as seguintes métricas para avaliar modelos de regressão, uma vez que são as utilizadas no âmbito deste trabalho.

Erro Quadrático Médio (MSE[5]): Refere-se à média do quadrado dos desvios entre os valores da previsão e os valores reais [36]. O modelo terá um melhor comportamento quanto mais baixo for este valor. A Equação 2.2 mostra a função respetiva.

$$MSE(y,p) = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \tag{2.2}$$

Em que y_i é o valor real, p_i o valor previsto para uma amostra e N o número de amostras. Esta função penaliza mais os valores mais altos, o que significa que devemos utilizar sobretudo quando se pretende sancionar os erros mais elevados.

Raiz do Erro Quadrático Médio (RMSE[6]): É a raiz quadrada de MSE. O modelo terá um melhor comportamento quanto mais baixo for este valor.

Erro Absoluto Médio (MAE[7]): Também designado *Median Absolute Deviation*

[5] Do Inglês Mean Squared Error
[6] Do Inglês Root Mean Squared Error
[7] Do Inglês Mean Asolute Error

[36], mede a média dos erros absolutos da previsão do modelo (Equação 2.3), em que x_i é o valor real, y_i o valor previsto e N o número de amostras. Neste caso todos os erros são penalizados do mesmo modo, ou seja, têm a mesma importância no âmbito do problema.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{N} \tag{2.3}$$

Em problemas de regressão, pode também utilizar-se o coeficiente de determinação R2, que varia de 0 a 1. Quanto mais próximo de 1, melhor está ajustado o modelo. R2 indica a proporção da variância na variável dependente que é previsível a partir da(s) variável(is) independente(s) [2].

Capítulo 3

Revisão de literatura

A análise de trabalhos relacionados teve como propósito conhecer o estado da arte e ajudou a definir os passos iniciais para o trabalho desenvolvido, sobretudo na escolha de algoritmos e atributos do conjunto de dados.

3.1 Análise de outros trabalhos

Santos et al. [2021] realizaram um estudo com a base de dados do refeitório da Universidade Federal de Uberlândia (UFU), Brasil, onde existem registos de pequenos almoços, almoços, jantares e ainda os menus dos vários campus escolares, num total de 6690 amostras. Para os objetivos pretendidos, foram utilizados dados relativos a um único campus para os anos letivos de 2015 e 2016; As 450 amostras do ano letivo de 2015 foram utilizadas para treino do modelo e as 609 amostras de 2016, como teste. Os atributos recolhidos foram:

- Mês
- Dia da semana
- Antecedência do feriado
- Sucede feriado
- Início semestre
- Número de dias início semestre
- Fim semestre
- Número de dias antes fim semestre
- Férias

Aos atributos recolhidos juntaram a estimativa do número de refeições a confecionar, feitas pela nutricionista responsável para o período de 2016. Estas estimativas foram utilizadas posteriormente para comparar com as previsões do modelo.

Os algoritmos escolhidos para realizar experiências foram K-vizinhos mais próximos (*K-nearest neighbors* - KNN), Perceptrão multicamadas (*Multi-Layer Perceptron* - MLP) e RF.

Foram comparados modelos com os 3 algoritmos para prever o número de refeições servidas aos pequenos almoços, almoços em dias letivos, almoços em dias de férias e jantares em dias letivos. A métrica escolhida para comparar os resultados dos vários modelos foi a RMSE. Concluíram que o algoritmo MLP teve o melhor desempenho entre os modelos propostos para os pequenos almoços (RMSE de 27,15), para almoços em dias de aulas (RMSE de 79,94) e jantares (RMSE de 102,85). Para a previsão dos almoços de todos os dias, o modelo com o algoritmo RF teve melhor performance, com RMSE de 100,07. Em comparação com as previsões realizadas pela nutricionista, os modelos com MLP, só não tiveram melhor resultado, na previsão de jantares.

Em Rocha et al. [2011] o conjunto de dados tem 254 amostras, obtidas do refeitório da Faculdade de Ciências e Letras situada em Assis (FCL/Assis) pertencente à Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Os atributos associados ao modelo foram selecionadas pelos autores e pelo nutricionista responsável pelo restaurante universitário conforme se descreve:

- Média dos últimos 30 dias
- Mês
- Dia da semana
- Número de refeições do dia anterior
- Feriado no dia anterior
- Média dos últimos 5 dias
- Feriado no dia posterior

De referir que o atributo nível de aceitação das refeições, foi aferido por um questionário aplicado a 350 alunos para avaliar a aceitação dos pratos principais de todos os menus servidos pelo restaurante universitário.

Estas amostras foram recolhidas no período de maio de 2008 a dezembro de 2009. A média de refeições servidas diariamente é de 366 e a amplitude de atendimento varia entre 100 e 500 refeições.

Foram testadas redes neuronais artificiais (RNA) para realizar a previsão de refeições a servir diariamente neste refeitório.

Comparando as amostras previstas pelo modelo com as amostras registadas no conjunto de dados, foi obtido um erro médio de 9,5%. Do total de dias analisados, verificou-se que o erro causado pelo método da média aritmética simples, de todo o conjunto, foi maior que o obtido por este modelo, levando os autores a concluírem que as redes neuronais são um método mais adequado para a previsão do número de refeições do que a simples ponderação pela média.

Holmberg et al. [2018], recolheram dados de três restaurantes situados em três cidades no sul da Suécia, com uma dimensão demográfica similar, cuja gestão de aprovisionamento é feita com base em médias aritméticas de consumos e consequentemente as escalas de pessoal são determinadas por estas previsões.

Os autores propunham-se testar modelos de aprendizagem automática e melhorar as previsões dos volumes de vendas, em relação aos métodos estatísticos utilizados. A expectativa inicial era de criar um modelo que servisse globalmente para gerar previsões em qualquer restaurante. Após as primeiras análises dos dados, tornou-se evidente que tal não seria possível devido às diferenças de volumes de vendas entre os restaurantes, por esse motivo, foram desenvolvidos modelos distintos para os três restaurantes. Para além de fazer uma previsão do volume de vendas, os autores pretendiam criar uma ferramenta útil para a gestão de pessoal, que possibilitasse organizar, com alguns dias de antecedência, as escalas de pessoal.

As amostras inicialmente recolhidas tinham, como atributos iniciais, apenas a data, hora e valor da venda. Estes dados foram agrupados por dia e para complementar a informação foram adicionados atributos relativos à meteorologia, histórico de vendas e feriados, ficando assim definidos:

- Data
- Ano
- Dia no ano
- Dia no mês
- Mês
- Número da Semana
- É feriado/férias
- Quadrimestre
- Número do dia da semana
- Tendência

Os algoritmos testados foram o *Extreme Gradient Boosted Trees* (XGB) e *Long Short Term Memory Neural Network* (LSTM). Os conjuntos de dados foram divididos em

70% para dados de treino e 30% para dados de teste, em que o total de amostras dos conjuntos de dados dos restaurantes 1, 2 e 3 é de 1268, 1494 e 3959 amostras, respetivamente. A métrica de avaliação escolhida para comparar os modelos foi RMSE, onde para os restaurantes 1, 2, 3 se obteve 24.377 SEK, 21.655 SEK e 49.114 SEK, respetivamente. Por fim, estes resultados foram ainda comparados com a metodologia de previsão utilizada então, com recurso às médias aritméticas de dias do ano correspondentes. A melhor performance foi obtida com *Extreme Gradient Boosted Trees* (ver Tabela 3.2).

Malefors et al. [2021] focaram o seu trabalho no desperdício alimentar e economia de custos durante o período da pandemia COVID 19, gerando um modelo de regressão para o número de refeições a confecionar, e também, com base nestes dados, realizar uma análise financeira. Foram recolhidos dados de escolas da Suécia, no período inicial da pandemia, correspondente a março 2020 até junho 2020, quando o governo introduziu a estratégia de combate à COVID 19 que impunha distanciamento social e restrições na frequência de aulas. Este período foi estudado em conjunto com dados recolhidos antes da pandemia, no período de 2010 a 2019.

Os dados referem-se a 18 refeitórios de escolas primárias e 16 de pré-escolas com um total de 5314 amostras, com os seguintes atributos:

- Quantidade de refeições 1 dia antes
- Quantidade de refeições 2 dias antes
- Quantidade de refeições 3 dias antes
- Dia da semana
- Semana
- Feriado no dia anterior
- Feriado no dia seguinte

Foram criados dois subconjuntos com informação de duas escolas apenas. Com estes subconjuntos é que se desenvolveram as experiências. Nestas duas escolas o volume de refeições servidas diariamente era entre 154 e 209.

Para avaliar o impacto dos modelos e a possível redução de custos, foram comparados 3 cenários distintos para os refeitórios estudados:

- Sem planificação: Em que são sempre confecionadas refeições correspondentes ao número total de alunos inscritos na escola;
- Registos atuais: Número de refeições atualmente servidas e o custo associado à sua confeção, para fazerem uma análise financeira;
- Utilização do modelo: Projeção de refeições baseada no modelo desenvolvido em AA.

A projeção de desperdício, neste contexto, é dada pelas refeições confecionadas subtraindo as efetivamente consumidas. O modelo, baseado em RF, foi treinado e aperfeiçoado sobre os dados iniciais e depois então, foi aplicado sobre as duas escolas selecionadas, correspondendo a 2 conjuntos com dados referentes ao período de 1 de agosto de 2019 até 28 de junho de 2020. Como métrica de avaliação foi utilizado MSE.

Os autores concluíram que a aplicação de técnicas de aprendizagem automática neste contexto são eficazes. Como foi observado o período de entrada em confinamento, concluíram também que nos primeiros dias desta mudança, é extremamente difícil ter previsões eficazes, mas com algumas semanas de dados, tornam-se viáveis. Testaram também a inclusão de atributos relativos à evolução da pandemia, concluindo que não foram determinantes para as previsões do modelo.

Em termos de análise financeira, para o Refeitório 1, com base no modelo criado, o custo total previsto das refeições no período testado seria de 4461 Euros, os custos reais foram de 4451 Euros. Caso não houvesse nenhum planeamento das refeições o custo seria de 5382 Euros. Para o Refeitório 2 o custo total previsto das refeições no período testado seria 774 Euros, o custo real foi de 532 Euros e se não houvesse nenhum método de gestão os custos seriam de 2072 Euros.

O último trabalho analisado, Camba [2023], refere-se à aplicação de um modelo de RNA para realizar a previsão da quantidade de refeições no refeitório da Universidade Federal de Ouro Preto (UFOP) e comparar os resultados obtidos com o método então em uso que consistia na previsão pela média da soma do dia anterior e do último dia da semana anterior.

O conjunto de dados contém 71 amostras relativas a um ano letivo completo. O refeitório funciona de segunda a sexta com almoços e jantares. Para este trabalho os almoços e jantares foram agregados. Os atributos do conjunto de dados são:

- Vendas Sem Impostos
- Venda média do mês
- Média de vendas nos últimos 7 dias
- Venda média do dia no mês
- Venda média do dia da semana
- Venda no ano passado no dia
- Dia no conjunto de dados
- Número do dia da semana
- Data
- Dia no mês

- Mês n°
- Dia no ano
- Semana n°
- Ano
- É feriado
- É noite
- Temperatura
- Temperatura média nos últimos 7 dias
- Chuva
- Minutos de sol
- Velocidade do vento
- Cobertura de nuvens
- Profundidade da neve

Foram utilizadas 56 amostras para treino do modelo e 15 para teste. Como métricas de avaliação de desempenho, optaram por R2, MAE e RMSE.

Através do modelo obtido por RNA obtiveram um MAE de 162,9 refeições e RMSE de 224,4. Verificou-se que os resultados são ligeiramente piores quando comparados com o método de previsão utilizado, uma vez que o MAE deste método é de 145,7 refeições e RMSE de 208,6. Concluíram que apesar dos resultados obtidos, existe espaço para melhorar o modelo, com mais amostras de dados e variáveis independentes.

3.2 Análise comparativa

Entre os trabalhos analisados, constatamos que a utilização de atributos relacionados com a data são comuns (Tabela 3.1), tentando nalguns casos identificar sazonalidades. Rocha et al. [2011], Malefors et al. [2021] e Camba [2023] procuram identificar algum tipo de tendência através da média de períodos temporais anteriores. Camba [2023] utiliza atributos relacionados com meteorologia.

A relação entre o erro obtido e a média dos conjuntos de dados varia entre 9,5% e 41%, sendo que Rocha et al. [2011], por esta avaliação, obteve o modelo mais eficiente através de redes neuronais. Camba [2023], obteve um RMSE de 224 num conjunto de dados com uma média de 1600, mesmo utilizando apenas 71 amostras.

Foram utilizadas diferentes técnicas de AA e abordagens que originaram resultados díspares. Santos et al. [2021], aplicaram um mesmo modelo a 3 conjuntos de dados originários da mesma fonte; como resultado obtiveram um nível de desempenho

Estudo	**Atributos**
Santos et al. (2021)	Mês; Dia da semana Antecedência do feriado Sucede feriado Início semestre Número de dias início semestre Fim semestre Número de dias antes fim semestre Férias
Rocha et al. (2011)	Nível de aceitação das refeições Média dos últimos 30 dias Mês; Dia da semana Número de refeições do dia anterior Feriado no dia anterior Média dos últimos 5 dias Feriado no dia posterior
Holmberg et al. (2018)	Data; Ano; Dia no ano; Dia no mês; Mês; Número da Semana É feriado/férias Quadrimestre Número do dia da semana Tendência
Malefors et al. (2021)	Quantidade de refeições 1 dia antes Quantidade de refeições 2 dias antes Quantidade de refeições 3 dias antes Dia da semana Semana Feriado no dia anterior Feriado no dia seguinte
Camba (2023)	Vendas Sem Impostos Venda média do mês Média de vendas nos últimos 7 dias Venda média do dia no mês Venda média do dia da semana Venda no ano passado no dia Dia no conjunto de dados Número do dia da semana Data; Dia no mês; Mês n°; Dia no ano; Semana n°; Ano É feriado É noite Temperatura Temperatura média nos últimos 7 dias Chuva Minutos de sol Velocidade do vento Cobertura de nuvens Profundidade da neve

Tabela 3.1: Atributos dos trabalhos analisados.

	Santos et al-(2021)	**Rocha et al. (2011)**	**Holmberg et al. (2018)**	**Malefors et al. (2021)**	**Camba (2023)**
Dimensão do conjunto de dados	1059	245	Conj.1 1268 Conj.2 1494 Conj.3 959	5314	71
Separação	Treino 50% Teste 50 %	*Sem informação*	Treino 70% Teste 30%	Treino 88% Teste 12%	Treino 80% Teste 20%
Algoritmos testados	KNN, MLP, RF	RNA Back-propagation	XGB , LSTM	RF, RNA	RNA
Medidas de desempenho	RMSE	MAE	RMSE	CMAPE	RMSE
Melhor modelo	MLP	RNA Back-propagation	XGB	RF	RNA
Média	Pequeno almoço 65 Almoço 758 Jantar 218	366	Conj.1 - 114000SEK Conj.2 - 68000SEK Conj.3 - 144000SEK	181	1500
Desempenho obtido	Pequeno almoço 27,15 Almoço 103,95 Jantar 79,94	34,77	Conj.1 - 24377SEK Conj.2 - 21655SEK Conj.3 - 49114SEK	0,15	224

Tabela 3.2: Resumo dos trabalhos analisados.

distinto para os 3 subconjuntos de dados estudados, com o RMSE corresponde a 13%, 35% e 41% da média de refeições servidas ao pequeno-almoço, almoço e jantar, respetivamente.

Capítulo 4

Método e ferramentas

Para o desenvolvimento de um modelo de aprendizagem automática orientado para a predição diária do número de refeições a confecionar nos refeitórios, seguiram-se metodologias consolidadas na literatura e privilegiou-se a utilização de software de código aberto.

4.1 Processo de trabalho

Este trabalho passou pelas seguintes etapas conforme proposto por Dutt [2018]:

1. Preparação dos dados: Conhecer os dados disponíveis. Descobrir as relações entre atributos e quais as suas potencialidades. Se necessário, aplicar técnicas de normalização, redução de dimensão ou gerar subconjuntos. Suprir falhas nos dados.
2. Treinar o modelo com base nos dados existentes: Treinar um algoritmo, utilizando um subconjunto de dados.
3. Avaliar a performance do modelo criado: Examinar os resultados obtidos através de métricas apropriadas para problemas de regressão, por exemplo, MSE.
4. Melhorar a performance: Com base na análise dos primeiros resultados, realizar afinações em hiperparâmetros do modelo ou aplicar os métodos de *ensemble*, *bagging* ou *boosting*.

De um modo geral foi seguida a metodologia, que se resume na Figura 4.1, dividindo as experiências realizadas em duas fases:

Fase 1 Com o conjunto de dados total, extraído dos sistemas de gestão dos refeitórios;

Fase 2 Com um conjunto de dados restrito ao ano letivo 2022/2023.

A figura 4.2 descreve o fluxo de tarefas desenvolvidas para atingir os resultados finais.

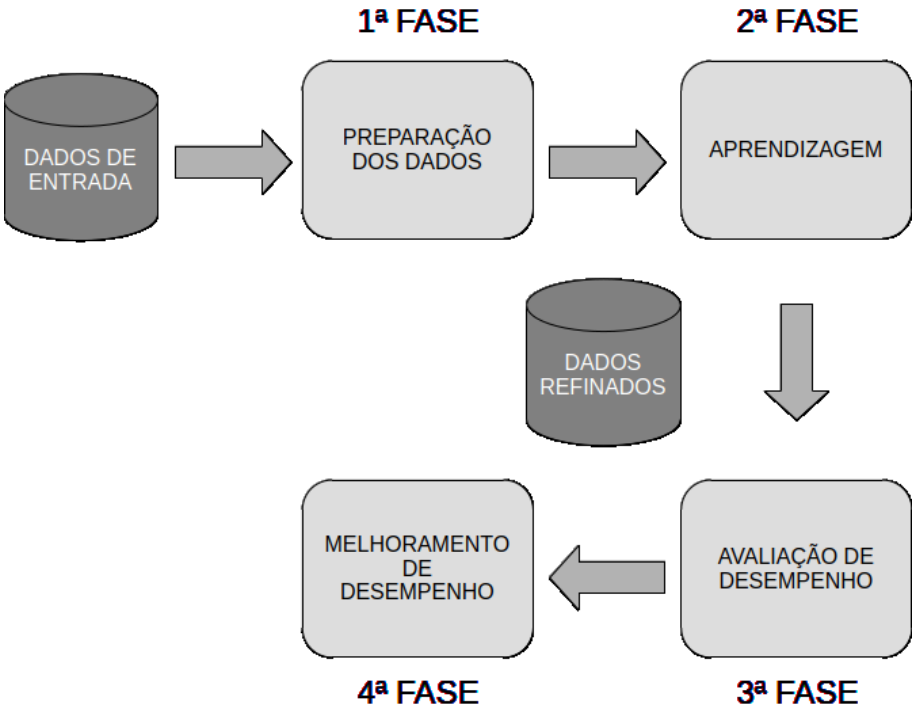

Figura 4.1: Processo de aprendizagem automática. Fonte: Dutt et al [2018]

4.2 Ferramentas utilizadas

O desenvolvimento das várias etapas deste trabalho, incluindo o tratamento de dados e a análise estatística, assentou no uso de ferramentas informáticas de utilização gratuita. Estas aplicações são descritas de seguida:

- Jamovi (www.jamovi.org) Jamovi é um *software* de análise estatística que disponibiliza as mais recentes metodologias de análise estatística [15], como uma interface intuitiva e de fácil aprendizagem.

- Calc (www.libreoffice.org) Os dados foram exportados dos sistemas de gestão dos refeitórios, em formato PDF e transformados em folhas de cálculo. Foi utilizado o Calc [16] para editar as tabelas, separar atributos e inserir novas colunas. O *software* Calc está incluído no pacote LibreOffice, é de código aberto e de utilização gratuita. Está vocacionado para produzir folhas de cálculo, dispondo de funções, criação de gráficos, ferramentas estatísticas, entre outras funcionalidades.

- Tableau (www.tableau.com) O Tableau é um *software* para análise de dados e geração de gráficos ou *dashboards* (visualizações) interativos. Está disponível *online* e permite um acesso gratuito[1] ainda que com funções limitadas. É possível usar, como fonte de dados, formatos CSV, para além de outros formatos e bases de dados incluindo dados geográficos (por exemplo ficheiros *Shapefile*). Como formatos de gráficos inclui, gráficos de barras, linhas, área, dispersão, gráfico de GANTT ou

[1] http://public.tableau.com

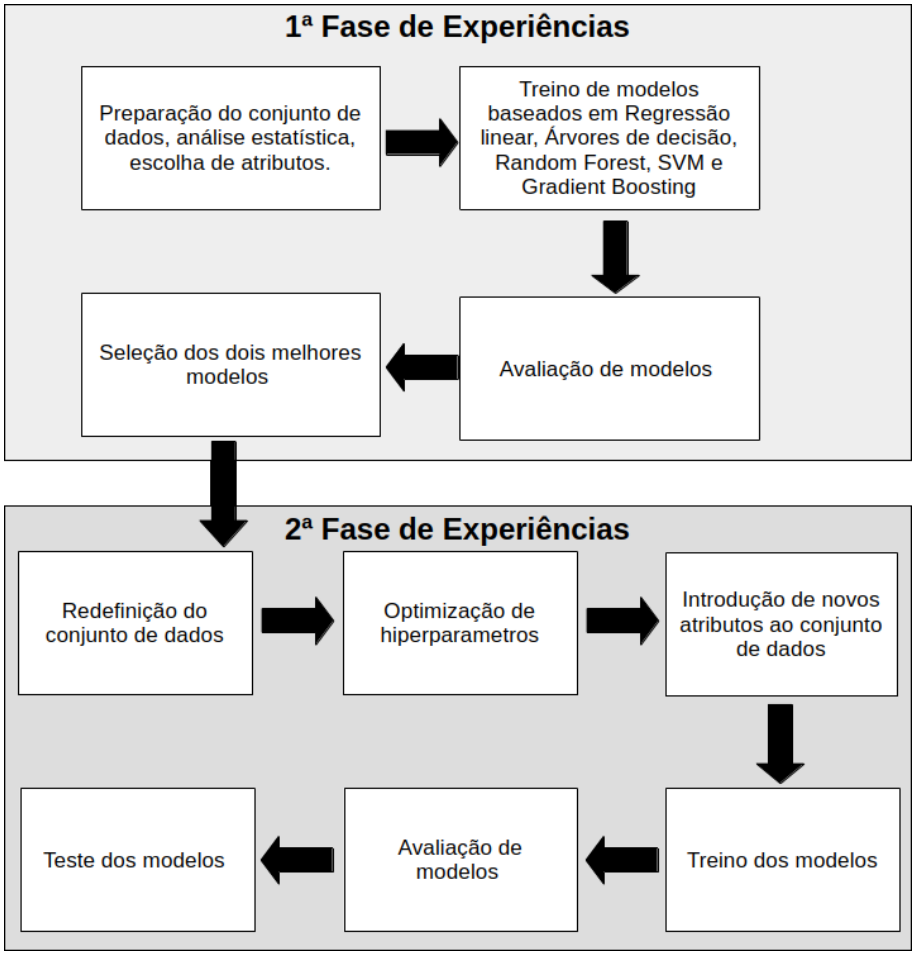

Figura 4.2: Diagrama das tarefas desenvolvidas.

densidade [18]. Foi utilizado para gerar gráficos, correlacionando atributos consoante a análise pretendida.

- Python (www.python.org) O *Python* é uma linguagem de programação de alto nível lançada em 1991 [17], hoje em dia muito utilizada para aprendizagem automática pela sua gama de bibliotecas associadas. É desenvolvida em código aberto e está disponível sob a licença OSI [17]. *Python* foi a linguagem escolhida para desenvolver os modelos e como interface de desenvolvimento utilizou-se o *Jupyter Notebook*[2]. Esta interface integra com *Python* o que agiliza a escrita e execução de código. As bibliotecas utilizadas no código foram:

- ***Numpy***[3]: Destina-se a computação científica. Disponibiliza funções para realizar operações sobre matrizes e outras funções de matemática, lógica, manipulação de formatados, classificação, seleção, transformadas discretas de Fourier, álgebra linear básica, operações estatísticas básicas e simulação aleatória.

[2] www.jupyter.org
[3] www.numpy.org

- ***Scikit-learn***[4]: Biblioteca para desenvolvimento de código de aprendizagem automática. Dispõe de algoritmos para classificação, regressão e agrupamento, incluindo SVM, RF, GB, *k-means* e *DBSCAN*, tem interoperabilidade com as bibliotecas numéricas e científicas *Python NumPy* e *SciPy* [3].

- ***Pandas***[5]: Biblioteca para análise de dados. Reúne um conjunto de funcionalidades para manipular dados, fazer leitura e escrita de ficheiros, dividir ou agregar e transformar as estruturas dos conjuntos de dados.

- Modelo multidimensional de bases de dados

O modelo multidimensional é uma técnica de modelação de bases de dados que tem como objetivo facilitar a leitura intuitiva e maleável da informação [25]. Estes modelos pretendem representar atividades específicas de uma organização tais como, aquisições ou vendas. A representação é feita por dois tipos de tabelas ligadas em estrela [35]:

- Tabela de Factos: Tabela central com os dados a analisar ;

- Tabela de Dimensão: Contém os atributos dos factos.

Neste âmbito, Facto é uma medida da atividade ou ocorrências numa organização e as tabelas de Dimensão contêm a descrição textual da atividade representada [25].

No contexto do modelo multidimensional, a estrutura de dados representativa de uma atividade, ou processo de negócio, denomina-se de um *data mart* [35]. Um *data mart* contêm informações específicas de uma unidade de negócios da organização que por sua vez se enquadra num sistema de armazenamento mais vasto. Este conjunto de dados mais vasto quando organizado em *data marts*, denomina-se de *data warehouse*.

Esta técnica foi aqui utilizada para modelar um *data mart* sobre a venda de refeições nos refeitórios, que foi utilizado ao longo de toda a análise de dados.

[4] www.scikit-learn.org
[5] http://pandas.pydata.org

Capítulo 5

Análise e tratamento de dados

Os dados que servem de base para este trabalho foram recolhidos de duas escolas básicas do segundo ciclo; a escola D. Paio Peres Correia de Tavira[1] e a escola D. José I[2] de Vila Real de Santo António (VRSA). O concelho de Tavira têm 27.536 habitantes [13] e a escola D. Paio Peres Correia cerca de 520 alunos. O concelho de VRSA tem 19.314 habitantes [14] e a escola D. José I, cerca de 430 alunos [7].

5.1 Descrição dos dados

O conjunto de dados recolhido refere-se ao período de 10-1-2022 a 15-12-2023, correspondendo, portanto, a metade do ano letivo de 2021/2022, o ano letivo completo de 2022/2023 e metade do ano letivo de 2023/2024. Este conjunto de dados tem 634 amostras, sendo 322 de Tavira e 312 de VRSA.

Os refeitórios servem apenas almoços e funcionam de segunda-feira a sexta-feira. Os alunos são incentivados a comprar senhas antecipadamente para os almoços, no entanto não está vedada a compra de senhas para refeições no próprio dia. O prato principal da ementa varia diariamente entre peixe e carne, ou seja o prato principal é alternadamente à base de carne ou peixe, além de uma alternativa vegetariana.

Os relatórios do software dos refeitórios escolares, que serviram de fonte para o conjunto de dados, indicam as quantidades de refeições vendidas por dia e por grau escolar. Encontra-se também a descrição das vendas com apoio escolar, pelos 3 níveis: escalão A, B e C (Figura 5.1). Considerando os objetivos estabelecidos, foram recolhidos dos relatórios, a data, a quantidade total de refeições vendidas por cada dia e o prato principal da ementa.

Estes foram os três atributos fundamentais para o desenvolvimento de todo o conjunto de dados.

[1] http://www.aejactavira.pt
[2] http://www.aedji.pt

FOLHA DE CAIXA N.º _____ (ANEXO REFEIÇÕES)

Setor: Refeitório

Tipo Preço	07-06-2023			
	Qtd.	Líquido	I.V.A.	Total (€)
Alunos - 1.º Ciclo	-	-	-	-
Alunos - 2.º Ciclo	16	23,36	0,00	23,36
Alunos - 2.º Ciclo - Escalão A	9	0,00	0,00	0,00
Alunos - 2.º Ciclo - Escalão B	4	2,92	0,00	2,92
Alunos - 2.º Ciclo - Escalão C	6	8,76	0,00	8,76
Alunos - 3.º Ciclo	22	32,12	0,00	32,12
Alunos - 3.º Ciclo - Escalão A	14	0,00	0,00	0,00
Alunos - 3.º Ciclo - Escalão B	11	8,03	0,00	8,03
Alunos - 3.º Ciclo - Escalão C	-	-	-	-
Alunos - Secundário	-	-	-	-
Outros Utentes	2	8,20	0,00	8,20
Taxa	1	0,30	0,00	0,30
TOTAL	**85**	**83,69**	**0,00**	**83,69**

Figura 5.1: Extrato do relatório de vendas

A Tabela 5.1 descreve todos os atributos do conjunto de dados usados nas experiências.

Uma vez que estavam aglomerados dados de duas escolas fez-se uma análise dos dados em conjunto e também separadamente relativamente a cada escola. Isto para se observar se existem diferenças substanciais nas tendências e nos atributos mais influentes.

5.2 Análise dos dados

Globalmente a média de refeições servidas diariamente é de 158 com um desvio padrão de 60,7 (ver Tabela 5.2). No refeitório da Escola D. José I de VRSA a média

Atributo	Descrição
Tavira	Indica se os dados se referem a escola básica D. Paio Peres Correia de Tavira.
VRSA	Indica se os dados se referem a escola básica D. José I de Vila Real de Santo António.
Dia da semana	Números de 2 a 6 em que 2 corresponde a segunda-feira e 6 a sexta-feira
Carne	Indica se o prato principal é a base de carne
Peixe	Indica se o prato principal é a base de peixe
Temperatura	Temperatura média registada entre as 8 e as 14 horas
Precipitação	Indica se houve chuva entre as 8 e as 14 horas
Mês	Mês do ano, valor entre 1 e 12
Semana	A semana do ano comercial. valor de 1 a 52
Dia do mês	Indica o dia do mês. valor de 1 a 31
Dia do ano	Dia sequencial de 1 a 365 num ano.
Semana escolar	Número sequencial da semana dentro do ano letivo.
Precede feriado	Indica se o dia precede um feriado
Sucede feriado	Indica se o dia sucede um feriado
Pontes	Segundas que antecedem feriados ou sextas que sucedem Feriados são marcados como ponte
Quantidade	Quantidade de refeições servidas

Tabela 5.1: Descrição dos atributos do conjunto de dados

é de 137 e para a Escola D. Paio Peres Correia de Tavira é de 178, com desvio padrão de 59,2 e 55,1 respetivamente.

	Tavira	VRSA	Total
Amostras	322	312	634
Média	178	137	158
Mediana	177	140	157
Desvio-padrão	55,1	59,2	60,7
Mínimo	0	6	0
Máximo	288	639	639

Tabela 5.2: Descrição estatística dos dados.

A distribuição das refeições por dia da semana (Figura 5.2), indica que as terças-feiras são os dias com maior afluencia ao refeitório, com uma média de 187 refeições servidas, enquanto que as sextas-feiras são os dias com menor afluência, registando 116 refeições em média. Este é um padrão que se repete em cada escola.

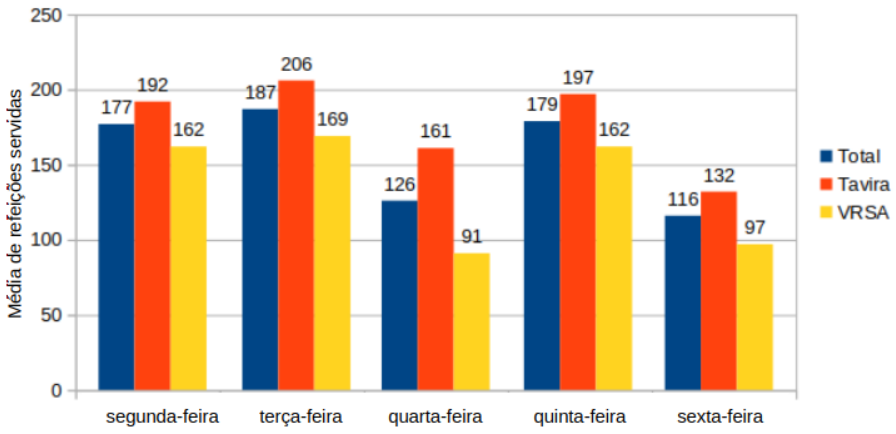

Figura 5.2: Média de refeições servidas por dia da semana.

Encontra-se outra semelhança quando analisamos as refeições servidas ao longo do ano letivo (Figura 5.3): ao longo do ano há uma tendência descendente no número de refeições servidas.

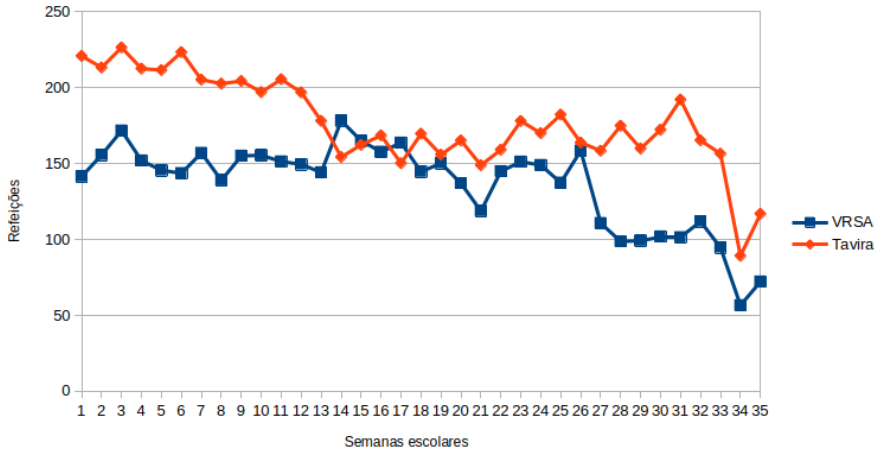

Figura 5.3: Média do número de refeições servidas por semana ao longo do ano letivo.

Durante o período em análise registaram-se 63 dias de chuva. O valor máximo de precipitação registada foi 6,8 mm/h, sendo a média, nestes dias de chuva, 0,4 mm/h. Na escola de Tavira, em média, os dias de chuva registaram, 173 refeições servidas e os dias sem chuva 179 refeições. Na escola de VRSA a média de refeições servidas nos dias com chuva foi de 142 refeições e nos dias sem chuva 136 refeições (ver Figura 5.4).

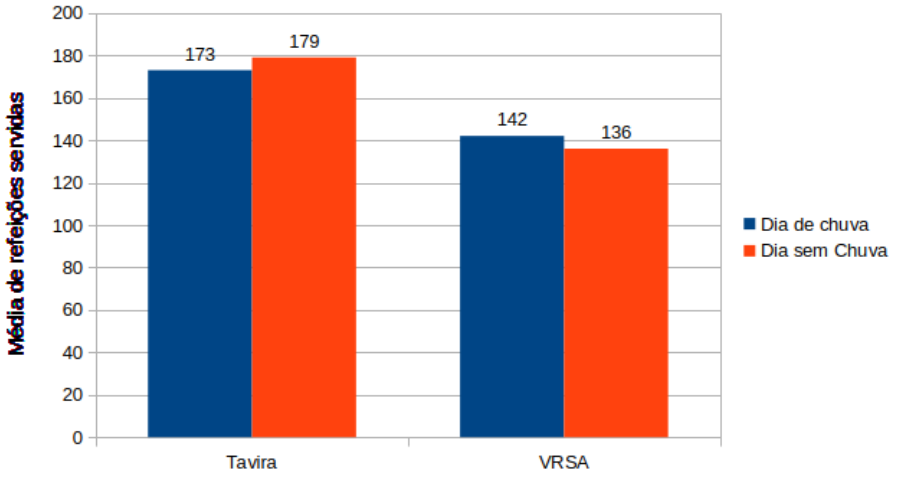

Figura 5.4: Influência dos dias de chuva no número de refeições servidas.

A Figura 5.5 mostra o gráfico da média de refeições servidas, conforme a temperatura ambiente. As temperaturas registadas variam entre os 9 °C e 31 °C, com uma média de 18 °C. No total do conjunto de dados, 77% das amostras têm uma temperatura igual ou superior a 15 °C.

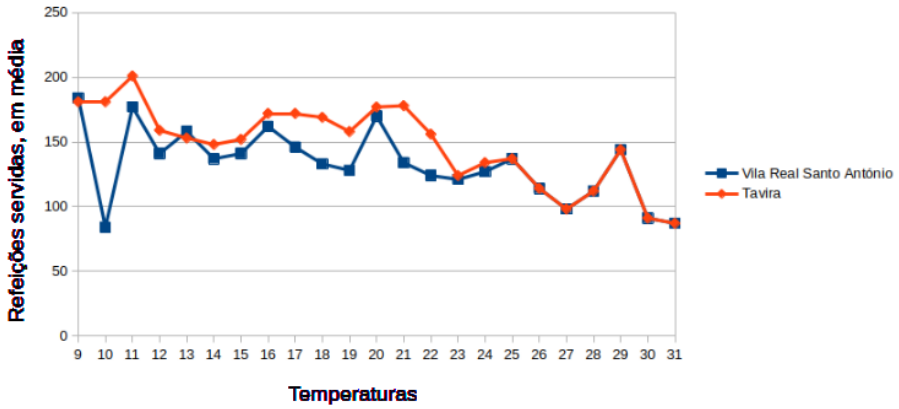

Figura 5.5: Número de refeições servidas segundo a temperatura ambiente.

Globalmente, o número de refeições servidas à base carne é de 52230, o que dá uma média de 184 e 172 refeições, entre os dias com este tipo de ementa, nas escolas de Tavira e VRSA, respetivamente (Tabela 5.3). Relativamente à ementa à base de peixe, foram servidas 47863, o que dá, em média, 144 e 131 refeições, entre os dias com este tipo de ementa, nas escolas de Tavira e VRSA, respetivamente.

Ementa	**Tavira**	**VRSA**
Carne	184	144
Peixe	172	131

Tabela 5.3: Média de refeições servidas por ementa.

Para avaliar a influência dos feriados na utilização dos refeitórios, foram utilizados atributos que indicam se o dia antecede ou sucede um feriado; para avaliar o efeito das *pontes* foram utilizados atributos que registam se o dia é uma segunda-feira anterior a um feriado ou sexta-feira posterior a um feriado. No período correspondente ao conjunto de dados houve 25 feriados e 10 *pontes*. De referir que, somente os dias denominados "ponte", registam o mesmo efeito em ambas as escolas, ou seja, em média há um ligeiro aumento na afluência aos refeitórios nos dias de ponte nas duas escolas (Figura 5.6).

A Tabela 5.4 mostra o coeficiente de correlação de Pearson entre as variáveis independentes e a variável dependente do conjunto de dados. O coeficiente de Pearson mede o grau da correlação (e a direção dessa correlação, se positiva ou negativa) entre duas variáveis de escala métrica [1]. Este coeficiente assume valores entre -1 e 1 em que os valores negativos indicam uma correlação negativa e os valores positivos, uma correlação positiva.

Os atributos com maior coeficiente de correlação são a Semana Escolar e o Dia da Semana.

A distribuição de refeições por dia da semana (Figura 5.2) mostra que existe alguma semelhança entre as duas escolas. As quartas-feiras e sextas-feiras são os dias que, em média, se servem menos refeições.

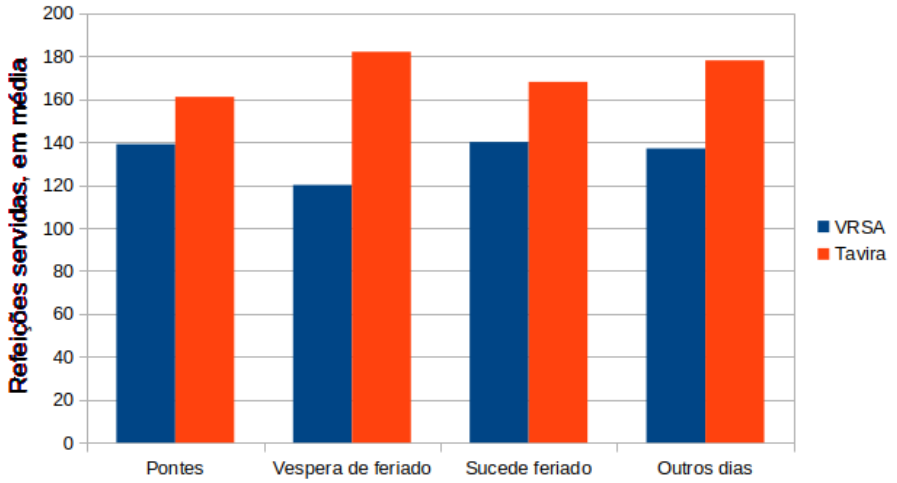

Figura 5.6: Média do número de refeições servidas nos dias anteriores aos feriados, posteriores a feriados e *pontes*.

Atributo	***p***
Semana escolar	-0,334
Dia da semana	-0,295
Dia do ano	0,167
Semana	0,166
Mês	0,158
Temperatura	-0,152
Dia do mes	0,111
Carne	0,089
Peixe	-0,089
Precede_feriado	-0,025
Pontes	-0,016
Precipitação	0,006
Sucede feriado	-0,006

Tabela 5.4: Coeficiente de correlação de Pearson entre as variáveis independentes e a variável dependente, Quantidade.

Encontramos outra semelhança quando analisamos as refeições servidas no decorrer do ano letivo (Figura 5.3). Verificamos que ao longo do ano há uma tendência descendente no número de refeições servidas. O primeiro trimestre de aulas regista uma média de 178 refeições diárias e o último trimestre 124.

As diferenças do número de refeições servidas em dias de chuva e sem chuva não são significativas. Em Tavira a média de refeições servidas em dias sem chuva é ligeiramente maior que nos dias de chuva. Em VRSA acontece o inverso conforme se pode ver na Figura 5.4.

Em relação à temperatura ambiente (Figura 5.5), nota-se um ligeiro decréscimo conforme as temperaturas se elevam. Naturalmente, as temperaturas são mais elevadas nos meses de abril, maio e junho, ou seja mais próximo do final do ano letivo.

Ambas as escolas apresentam uma diferença na média de refeições servidas quando a ementa é de carne ou peixe (Tabela 5.3), embora a diferença não seja significativa.

Os dias feriado, Figura 5.6, mostram ter mais impacto na escola de Tavira do que em VRSA. Em Tavira, nos dias que sucedem feriados e *pontes* há um ligeiro decréscimo no número de refeições servidas. Em VRSA nas vésperas de feriados é que se nota o decréscimo no número de refeições servidas

De forma geral os atributos denotam uma correlação baixa. Os atributos Precipitação e Sucede Feriado mostram mesmo uma correlação muito fraca. Apesar disto, optou-se por manter todos os atributos no conjunto de dados.

5.3 Pré-processamento de dados na Fase 1

Partindo de exemplos tomados de outros trabalhos, decidiu-se adicionar atributos com a seguinte informação:

- Temperatura ambiente;
- Pluviosidade;
- Dias feriado e ponte.

Para refletir o impacto da sazonalidade, proximidade de períodos de férias ou outras festividades importantes, a data foi dividida em diversos atributos temporais, tais como dia da semana, dia do ano letivo ou dia do mês.

Para avaliar particularmente o efeito dos feriados, foram identificados os dias que antecedem e sucedem feriados e as segundas e sextas-feiras que antecedem ou sucedem feriados, respetivamente (denominados vulgarmente de "pontes").

Para transformar a ementa num atributo numérico, introduziram-se dois atributos binários para indicar se o prato principal foi à base de carne e a base de peixe.

Como verificado em Santos et al. [2021], Rocha et al. [2011] ou Holmberg et al. [2018], os dados climatéricos são influentes para a dinâmica dos refeitórios, e assim, foram retirados do sítio WindGuru[3] dados meteorológicos relativos ao período estudado. Neste sítio podemos encontrar registos históricos de temperatura, pluviosidade e ventos, de diversos lugares do mundo, incluindo Tavira e Vila Real de Santo António.

Sobre a pluviosidade, e porque os níveis registados para Tavira e Vila Real de Santo António, foram baixos durante o período em estudo, optou-se por inserir um atributo que indicasse apenas se durante o horário letivo choveu ou não, ficando assim um atributo binário. A temperatura média registada diariamente entre as 8 e as 14 horas, foi utilizada como atributo de números inteiros.

[3] www.windguru.com

Após estas tarefas os atributos ficaram com as tipologias que se descrevem na Tabela 5.6.

Atributo	**Tipo**
Tavira	boleano
VRSA	boleano
Dia da semana	número inteiro
Carne	boleano
Peixe	boleano
Temperatura	número inteiro
Precipitação	boleano
Mês	número inteiro
Semana	número inteiro
Dia do mês	número inteiro
Dia do ano	número inteiro
Semana escolar	número inteiro
Precede feriado	boleano
Sucede feriado	boleano
Pontes	boleano
Quantidade	número inteiro

Tabela 5.5: Tipologia de dados dos atributos

Foi criada uma base de dados com um modelo multidimensional que permitisse gerar tabelas e gráficos de modo interativo, através da plataforma *Tableau* e assim agilizar a análise de dados (Figura 5.7). Foram criadas as seguintes tabelas:

- Refeições: Registas as refeições servidas e os atributos associados;
- Escalão: Indica o tipo de escalão de apoio social atribuíveis aos alunos;
- Temporal: Tabela com as datas;
- Ementa: Indica os tipos de ementa (Carne ou Peixe);
- Ciclos: Indica os ciclos letivos.

A tabela referente às refeições servidas é a tabela de factos. E as tabelas Escalão, Temporal, Ementa e Ciclos são dimensões.

Esta fase de análise foi fundamental para se tomar as primeiras decisões com relação à criação do modelo de AA.

Iniciou-se o treino dos modelos com os dados agregados das duas escolas. Primeiro com hiperparâmetros pré-definidos, com o objetivo de comparar e avaliar os algoritmos RL, AD, RF, SVM e XGB.

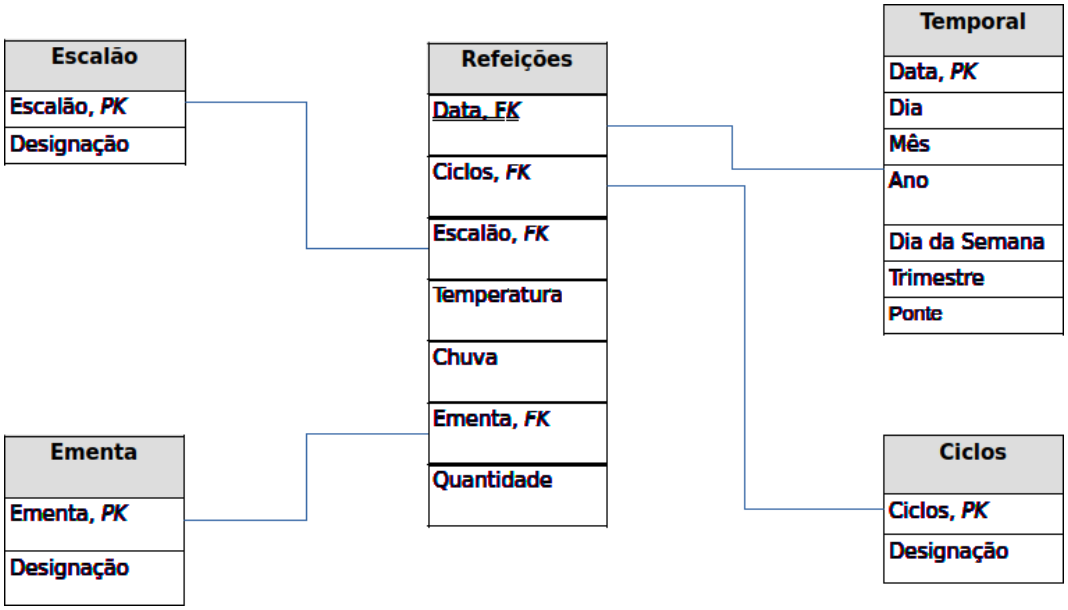

Figura 5.7: Modelo multidimensional de refeições servidas nos refeitórios

De seguida realizou-se um ajuste de hiperparâmetros para escolher os melhores para passar a uma nova fase de experiências.

O apêndice A mostra uma listagem de código em *Python* que exemplifica estes passos. Este código foi criado e executado na consola virtual de desenvolvimento, *Jupyter Notebook*.

O conjunto de dados foi dividido aleatoriamente, com 75% das amostras (475) para treino de modelos e 25% (159) para teste.

5.4 Pré-processamento de dados na Fase 2

Na segunda fase de experiências, optou-se por restringir os dados a um ano letivo, isto é de 15-9-2022 a 16-6-2023, para concentrar a análise num período temporal coerente com a duração do ano letivo, mantendo a sua sequência temporal correspondente. Ao conjunto de dados foram adicionados 2 novos atributos, descritos na Tabela 5.6, que registam tendências de consumo semanais e diárias.

Atributo	Tipo	Descrição
Semana anterior	Número inteiro	Quantidade de refeições servidas no dia homólogo da semana anterior
Dia anterior	Número inteiro	Quantidade de refeições servidas no dia anterior

Tabela 5.6: Atributos adicionados na segunda fase de experiências

Este conjunto ficou com 287 amostras em que a média se situa nas 170 refeições servidas diariamente (ver Tabela 5.7).

Com estes novos atributos foram realizadas 3 experiências distintas. A primeira com o conjunto de atributos original e o novo atributo quantidade de refeições servidas

Dimensão	287
Média	170
Mediana	172
Desvio Padrão	62,6
Mínimo	0
Máximo	288

Tabela 5.7: Descrição estatística do conjunto de dados utilizado na segunda fase de experiências.

no dia da semana anterior. Uma segunda experiência com o conjunto original e a quantidade de refeições servidas no dia anterior e uma terceira com os dois atributos em simultâneo. Estas experiências incluíram a afinação de hiperparâmetros com treino, validação cruzada, e testes dos modelos.

Nestas experiências a separação não foi automática para garantir a sequência temporal dos dados, mas manteve-se a proporção de 75% das amostras (214) para treino e 25% (72) para teste.

Capítulo 6

Resultados obtidos

Neste capítulo analisam-se os resultados obtidos nas duas fases, em que se dividiram as experiências, procurando responder a qual o melhor algoritmo a aplicar e se há vantagem na sua aplicação.

6.1 Primeira fase de experiências

Para a primeira fase de experiências foi utilizado todo o conjunto de dados inicialmente recolhido, com os atributos conforme a Tabela 5.1.

Compararam-se os algoritmos de RL, XGB, RF, AD e SVR. Os algoritmos foram aplicados com os hiperparâmetros por omissão.

Não existindo uma regra para a separação do conjunto de dados para treino e teste, Duamé [2015], sugere que os dados para teste devem ser divididos de modo a representarem todo o conjunto. O conjunto de dados foi dividido de forma aleatória em 75% das amostras para treino e 25% para teste. Esta divisão foi decidida para permitir que o conjunto de testes tenha uma representação suficientemente alargada de vários períodos temporais, tais como, semanas, dias da semana e meses. A primeira experiência avaliou os 5 algoritmos pelas métricas MSE e MAE com validação cruzada, obtendo-se os resultados expressos na Tabela 6.1.

Algoritmo	**MAE**	**MSE**	**RMSE**
XGB	**27,172**	**1273,428**	**35,094**
RF	28,231	1349,965	36,177
DT	35,380	2361,165	47,657
LR	36,362	2030,206	44,702
SVR	44,039	2898,286	53,602

Tabela 6.1: Resultados da primeira experiência.

Os valores de MAE mostram um mínimo de 27,172, obtido por XGB e o máximo de 44,039 por SVR, havendo uma diferença de aproximadamente 17 pontos. Os dois modelos com melhor performance, na avaliação por esta métrica, são RF e XGB com erro médio absoluto nas previsões de 27,172 e 28,231 refeições respetivamente. Constata-se que são valores muito aproximados. Inclusivamente o RMSE de XGB e RF é muito próximo (35,094 e 36,177 respetivamente).

De forma genérica, os modelos baseados em árvores de decisão tiveram melhor prestação. Este conjunto de experiências permitiu escolher os algoritmos mais adequados para uma nova fase de experiências, que foram centradas apenas em XGB e RF.

6.2 Segunda fase de experiências

Para a segunda fase de experiências, centrada nos algoritmos RF e XGB, o conjunto de dados foi reduzido ao ano letivo 2022/2023 e realizou-se ajuste dos hiperparâmetros. A separação das amostras para treino (75%) e testes (25%) manteve a sequência temporal e, portanto, não foi aleatória. As 213 amostras de treino correspondem aos meses de setembro até final de março, e as 74 amostras de teste, aos meses de abril até junho. O conjunto de dados, com 287 amostras, tem uma média de 171,6 refeições servidas diariamente com um desvio padrão 62,6.

Para realizar o ajuste de hiperparâmetros para os modelos, utilizou-se *GridSearchCV*, que é uma classe em *Python* que determina, de forma automatizada, quais os melhores hiperparâmetros, partindo de intervalos definidos pelo programador. A Tabela 6.2 mostra os intervalos de valores dos hiperparâmetros testados para os modelos.

RF	XGB
criterion: squared_error max_depth : 8, 10, 12, 14, 16, 18, 20 random_state : 100, n estimators : 20, 40, 60, 80, 90, 100, 110, 120	n estimators : 20, 40, 60, 80, 90, 100, 110, 120, 130, 140, 150, criterion :squared_error , random_state : 100, loss: squared_error, learning_rate : 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2, 0.22, 0.2 max_depth : 2, 4, 6, 8, 10

Tabela 6.2: Intervalos de hiperparâmetros testados com *GridsearchCV*.

A Tabela 6.3 apresenta os hiperparâmetros que devolveram os melhores modelos e foram utilizados ao longo destas experiências.

Para tentar melhorar a performance do modelo acrescentaram-se novos atributos ao conjunto de dados, conforme a Tabela 5.6 e realizaram-se as seguintes experiências:

A . Utilizando o conjunto de dados com os atributos originais;

B . Utilizando o conjunto de dados com atributos originais e um atributo com o

RF	**XGB**
criterion: squared_error max_depth: 9 random_state: 100 n_estimators: 40	criterion: squared_error loss: squared_error max_depth: 6 n_estimators: 40 random_state: 100 learning_rate= 0.11

Tabela 6.3: Hiperparâmetros utilizados.

	MAE		MSE		RMSE	
Experiências	RF	XGB	RF	XGB	RF	XGB
A	42,521	40,666	7144,004	3340,258	84,522	57,794
B	41,614	44,617	3513,568	3964,694	59,275	62,965
C	**38,851**	41,471	**3012,645**	3371,860	**54,887**	58,067
D	40,866	43,337	3433,988	3779,427	58,600	61,477

Tabela 6.4: Resultados das experiências com os algoritmos RF e XGB.

número de refeições servidas no dia anterior;

C . Utilizando o conjunto de dados com atributos originais e um atributo com o número de refeições servidas no dia homologo da semana anterior;

D . Utilizando o conjunto de dados com atributos originais e os atributos adicionais das experiências C e D;

O desempenho dos modelos sobre o conjunto de testes, é apresentado na Tabela 6.4.

A experiência C, em que se adicionou o atributo com o número de refeições servidas no dia homologo da semana anterior, teve a melhor performance com o algoritmo RF. O erro médio absoluto obtido foi de 38,8515 e RMSE de 54,8875. Este conjunto de dados tem uma média de 170 refeições diárias, pelo que o valor de MAE corresponde a 22% da média e o RMSE, 32%.

A Figura 6.1 mostra, graficamente, o desempenho do modelo RF, ao longo dos dias do conjunto de testes, onde se nota que os valores previstos seguem a tendência dos valores reais ainda que com alguns erros consideráveis, sobretudo nas últimas amostras. Nas amostras de teste №1 até à №59, os valores reais situam-se entre as 55 refeições como mínimo e 266 como máximo. Entre a amostra №60 e à №74, os valores situam-se entre as 19 refeições como mínimo e 164 como máximo. Neste período, em que houve uma alteração abrupta na afluência aos refeitórios, nota-se um maior desfasamento entre os valores previstos e reais. Esta alteração da utilização dos refeitórios não foi devidamente assimilada pelo modelo.

A Tabela 6.5 apresenta o número de previsões por percentagem de erro do modelo RF. O modelo RF obteve 20 previsões com um erro superior a 50% do valor real, mas cerca metade das previsões (36) tiveram um erro igual ou inferior a 15% do valor real.

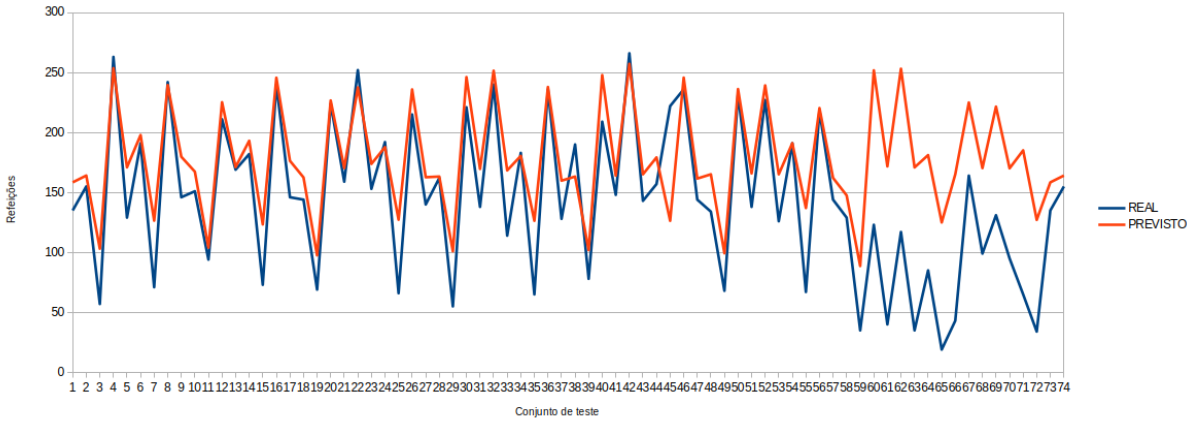

Figura 6.1: Previsões pelo modelo RF.

Erro	$\leq$ **10%**	$\leq$ **15%**	$\leq$ **25%**	$\geq$ **50%**
Amostras	27	36	46	20

Tabela 6.5: Número de amostras por percentagem de erro, do modelo da experiência C.

Santos et al [2021] obtiveram um RMSE de 103 num modelo para previsão de almoços a servir, cujo conjunto de dados tem em média 758 refeições diárias o que corresponde a cerca de 13% do valor médio (Tabela 6.6). Camba [2023], obteve um RMSE de 224 refeições num conjunto de dados com uma média de aproximada de 1600 refeições, o que corresponde a 14% do valor médio de refeições servidas. O melhor desempenho obtido neste estudo possui um RMSE correspondente a 31% da média real de refeições servidas.

	Santos et al	**Camba**	**Modelo RF**
RMSE/MÉDIA	13%	14%	31%

Tabela 6.6: Comparação com outros trabalhos.

Estes resultados, apesar de terem margem para melhorar, são encorajadores porque aproximadamente 50% das previsões tiveram um erro inferior a 15% do valor real. Comparativamente a outros trabalhos, a performance do modelo obtido neste estudo é menos eficiente, sobretudo perante a mudança brusca da quantidade de refeições servidas, como aconteceu no mês de junho derivado da realização de exames nacionais. A divisão sequencial do conjunto de dados, para manter a sua linha temporal, levou a que o modelo fosse treinado com dados que não incluíram o mês de junho, isto pode também explicar a maior taxa de erro neste período.

Retirando as amostras referentes ao mês de junho do conjunto de testes, com o mesmo modelo obtido pela experiência C, verifica-se uma melhoria de desempenho substancial (ver Tabela 6.7). O MSE foi reduzido de, aproximadamente, 38 para 20, ou seja reduziu cerca de 47% e o RMSE de 54 para 31 (redução de 42%).

	MAE	MSE	RMSE
Inicial	38,851	3012,645	54,887
Sem amostras de junho	**20.900**	**1241.883**	**31.377**

Tabela 6.7: Desempenho do modelo com dados de teste sem amostras do mês de junho

Se os gestores dos refeitórios escolares se baseassem na média de refeições servidas diariamente (171), para decidir qual o número a confecionar, deveriam considerar o desvio padrão (66,6) como intervalo de segurança para que não houvesse falta de refeições face à procura. Então, 66,6 seria, em média, o erro máximo de previsão esperado. Comparativamente à utilização do modelo obtido neste estudo, e considerando o RMSE (54,8) como esse intervalo de segurança, podemos dizer que a utilização do modelo RF seria mais exato e que o erro máximo seria reduzido em 17,7%.

6.2.1 Trabalho suplementar

Realizou-se um trabalho suplementar ainda com o conjunto de dados e o modelo da experiência C, em que se separaram os dados relativos às escolas para comparar com os resultados obtidos. A tabela 6.8 apresenta estes resultados. Constatamos que ao treinar o modelo com os dados separadamente, o erro médio absoluto diminui para 33 refeições, no modelo relativo a escola de Tavira, mas aumenta ligeiramente o RMSE para 55 refeições. No modelo da escola de VRSA estes indicadores pioram: MAE de 45 refeições e o RMSE 55.

Escola	**MAE**	**MSE**	**RMSE**
Tavira	33,435	3038,671	55,492
VRSA	45,188	3079,438	55,124

Tabela 6.8: Resultado da experiência de separação dos dados das escolas.

Capítulo 7

Conclusão

7.1 Considerações finais

Foram reunidos dados de dois refeitórios escolares de duas escolas, em diferentes concelhos, que têm as mesmas faixas etárias e, pela análise dos dados, tendências muito similares no que se refere à utilização dos refeitórios.

O modelo aqui criado com RF, apresenta um potencial na sua utilização para resolver o problema proposto de prever o número de refeições num refeitório. Obteve-se um MAE de aproximadamente 38 e RMSE de 54 e uma potencial redução no desperdício alimentar em cerca de 17,7%.

Não existe um registo histórico acerca das previsões do número de refeições a confecionar diariamente, nem o número de alunos que não conseguiram almoço ou quantos almoços sobraram. Por esse motivo não foi possível fazer uma comparação factual entre o processo de decisão atual e a aplicação de métodos de aprendizagem automática. Isto poderia dar uma indicação sobre a vantagem, ou não, de aplicar o modelo num contexto real, mesmo com o MAE obtido.

Considera-se que foi cumprido o objetivo de implementar um modelo de aprendizagem automática baseada nos dados gerados e mantidos por um organismo público, apesar de existir ainda algum trabalho para melhorar o sistema. Com a digitalização dos serviços públicos, existem hoje uma quantidade de dados, de diversas áreas, que podem ser utilizados para gerar modelos de aprendizagem automática e complementar a gestão dos serviços, sendo este trabalho apenas um exemplo disso.

7.2 Trabalhos futuros

A divisão dos dados das duas escolas para treinar e testar o modelo separadamente, num caso não melhorou o desempenho significativamente e noutro piorou. Este facto perspetiva que se poderá testar um modelo com dados aglomerados de mais escolas

e de mais anos letivos, para o melhorar.

Há ainda a possibilidade de encontrar atributos com maior grau de correlação, o que abre aqui vias de trabalho futuro para identificar, por exemplo a sazonalidade ou tendências temporais e/ou preferências gastronómicas.

Com este estudo terminado, guardo a convicção de que é possível aplicar um modelo de AA, com êxito, num caso concreto para prever as refeições a confecionar num refeitório escolar. Esta hipótese poderá ser vantajosa para reduzir o desperdício alimentar e também racionalizar custos.

Apêndice A

Listagem de programação em Python

A listagem de código seguinte é um exemplo do código criado para treinar os modelos de AA. Neste caso adaptada para Random Forest.

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split,
     KFold,  cross_val_score, cross_validate, GridSearchCV
from sklearn import preprocessing, metrics, tree
from sklearn.ensemble import RandomForestRegressor
from sklearn.utils import shuffle
from sklearn.ensemble import GradientBoostingRegressor
import pandas as pd

######################
# LEITURA DO DATASET
######################
metricas = ["neg_root_mean_squared_error","neg_mean_squared_error",
   "neg_mean_absolute_error"]
resultados=[['Algoritmo','MAE','MSE','RMSE']]
resultados_cv=[['Algoritmo','Score-Média']]
data_dir  = './refeicoes.csv'
df = pd.read_csv(data_dir)
df.head()

######################
# BARALHAR O DATASET
######################
```

```
df = shuffle(df, random_state=0)
df.head()

##################################
# SEPARA AS VARIAVEIS DEPENDENTES
##################################

X = df.iloc[:, :-1]
Y = np.ravel(pd.DataFrame(df.loc[:,['Quantidade']]))

#####################################
# DIVIDE O DATASET EM TREINO E TESTE
#####################################

X_train, X_test, y_train, y_test = train_test_split(X, Y,test_size=0.25,
 random_state = 0)

#X_train = X
#y_train = Y
#X_test = Xt
#y_test = Yt

print(X_train.head())
print(X_train.shape)

################################################
# DEFINIR O ALGORITMO A UTILIZAR - RANDOM FOREST
################################################
clf = RandomForestRegressor(n_estimators=15, random_state=0)


# ENCONTRAR OS MELHORES HIPERPARAMETROS

clf_params=[{'criterion': ['squared_error'],
 'max_depth': [6],
 'random_state': [100],
 'n_estimators': [150,160,170,180,190,200,210,220,230,240,]}]

clf_grid = GridSearchCV(estimator=clf,
param_grid=clf_params,
scoring='neg_mean_absolute_error',
cv=10,
return_train_score=True,
verbose=1)
```

```
clf_grid.fit(X_train, y_train)
print(clf_grid.best_params_)
print(clf_grid.score(X_train, y_train))

# VALIDAÇÃO CRUZADA
scores = cross_validate(clf, X_train, y_train, cv=10)
model_clf = clf.fit(X_train,y_train)


# AVALIA EFICÁCIA RANDOM FOREST
y_pred = model_clf.predict(X_test)

MAE = metrics.mean_absolute_error(y_test, y_pred)
MSE = metrics.mean_squared_error(y_test, y_pred)
RMSE = metrics.root_mean_squared_error(y_test, y_pred)

#print('### RANDOM FOREST ###')
print('Mean Absolute Error:', MAE)
print('Mean Squared Error:', MSE)
print('Root Mean Squared error: ', RMSE)

resultados_cv.append(['RF',np.mean(scores['test_score'])])
resultados.append(['RF',MAE,MSE,RMSE])

##########################################
# APLICAR HIPERPARAMETROS EM RANDOM FOREST
##########################################

rf_hiper = RandomForestRegressor(criterion = 'squared_error', max_depth= 5,
random_state = 100, n_estimators = 450)

# VALIDAÇÃO CRUZADA

scores = cross_validate(rf_hiper, X_train, y_train, cv=10, scoring=metricas,
  return_train_score=True )

# Treino com os hiperparametros encontrados

model_rf_hiper = rf_hiper.fit(X_train,y_train)


# AVALIA EFICÁCIA RANDOM FOREST

y_pred = model_rf_hiper.predict(X_test)
```

```
MAE = metrics.mean_absolute_error(y_test, y_pred)
MSE = metrics.mean_squared_error(y_test, y_pred)
RMSE = metrics.root_mean_squared_error(y_test, y_pred)

print('### RANDOM FOREST ###')
print('Mean Absolute Error:', MAE)
print('Mean Squared Error:', MSE)
print('Root Mean Squared error: ', RMSE)

print(scores['test_neg_mean_absolute_error'].mean())
print(scores['test_neg_mean_squared_error'].mean())
print(scores['test_neg_root_mean_squared_error'].mean())
```

Bibliografia

[1] Coeficiente de correlação de Pearson. in Wikipedia. (Acedido em 20-05--2025). https://pt.wikipedia.org/wiki/Coeficiente_de_correla%C3%A7%C3%A3o_de_Pearson.

[2] Coeficiente de determinação. in Wikipedia, (Acedido em 20-09-2025). https://pt.wikipedia.org/wiki/Coeficiente_de_determina%C3%A7%C3%A3o,.

[3] "Scikit-learn - site oficial". (Acedido em 20-05-2025). https://scikit-learn.org/stable/.

[4] Direção-Geral de Estatísticas da Educação e Ciência. "IUTIC na administração pública central, regional e câmaras municipais 2013", 2013. (Acedido em 26-05-2025). https://www.dgeec.medu.pt/api/ficheiros/6571dfdd51625bcaa2a0c0c5/.

[5] República Portuguesa, "Regime jurídico das autarquias locais", Out. 2013. (Acedido em 25-06-2024). https://diariodarepublica.pt/dr/legislacao-consolidada/lei/2013-56366098-56359624.

[6] Fundação para a Ciência e a Tecnologia. "Research in data science and artificial intelligence applied to public administration", 2020. (Acedido em 25-06-2024). https://www.fct.pt/wp-content/uploads/2022/06/Brochura_ResearchinDataScienceandAIappliedtoPA.pdf.

[7] Câmara Municipal de Vila Real de Santo António. "Carta educativa 2021", 2021. (Acedido em 20-05-2025). https://cms.cm-vrsa.pt/upload_files/client_id_1/website_id_1/Atividade%20Municipal/Educacao/DEJS_2021/Caderno_Carta%20Educativa%202021.pdf.

[8] Conselho de Ministros, "Resolução n.º 131/2021" Diário da República n.º 177/2021, série I de 2021-09-10. https://diariodarepublica.pt/dr/detalhe/resolucao-conselho-ministros/131-2021-171096337, 2021.

[9] Direção-Geral de Estatísticas da Educação e Ciência. "Inquérito à utilização das tecnologias de informação e comunicação.", 2023. (Acedido em 26-05-2025). https://www.dgeec.medu.pt/p/ciencia-e-tecnologia/estatisticas/.

[10] Direção-Geral de Estatísticas da Educação e Ciência. "IUTIC na administração pública central, regional e câmaras municipais 2023", 2023.

(Acedido em 26-05-2025). https://www.dgeec.medu.pt/api/ficheiros/66420700af3410403fec6274/.

[11] "Portugal 2020 - site oficial", 2024. (Acedido em 14-06-2024). https://www.portugal2020.gov.pt/.

[12] "Simplex - site oficial", 2024. (Acedido em 14-06-2024). https://www.simplex.pt/.

[13] Câmara Municipal de Tavira. "Conhecer Tavira", 2025. (Acedido em 20-05--2025). https://cm-tavira.pt/site/conhecer-tavira/.

[14] Câmara Municipal de Vila Real de Santo António. "Concelho : Vila Real de Santo António", 2025. (Acedido em 20-05-2025). https://www.gee.gov.pt/pt/lista-publicacoes/estatisticas-regionais/distritos-concelhos/faro/vila-real-de-santo-antonio/3122-vila-real-de-santo-antonio/file.

[15] "Jamovi - site oficial". https://www.jamovi.org, 2025. (Acedido em 20-05--2025).

[16] "Libreoffice - site oficial", 2025. (Acedido em 26-09-2025). https://www.libreoffice.org.

[17] "Python - site oficial", 2025. (Acedido em 26-09-2025). https://www.python.org.

[18] "Tableau - site oficial", 2025. (Acedido em 26-09-2025). https://www.tableau.com.

[19] M. Awad and R. Khanmeduna. Machine learning. In *Efficient Learning Machines*, pages 1–18. Apress, Berkeley, CA, 2015.

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, Out. 2017.

[21] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[22] H. Duamé. *A Course in Machine Learning*. Self-published, 2015.

[23] J. Friedman. Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38:367–378, 02 2002.

[24] A. Geron. *Hands-on machine learning with scikit-learn, keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA, 2 edition, Out. 2019.

[25] M. Golfarelli and S. Rizzi. *Data warehouse design: Modern principles and methodologies*. Osborne/McGraw-Hill, New York, NY, Jun. 2009.

[26] S. Gollapudi. *Practical Machine Learning*. Packt Publishing, Birmingham, England, Abr. 2023.

[27] S. Guido and A. C. Mueller. *Introduction to machine learning with python.* O'Reilly Media, Sebastopol, CA, Out. 2016.

[28] T. K. Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, page 278, USA, 1995. IEEE Computer Society.

[29] D. Julian. *Designing machine learning systems with python.* Packt Publishing, Birmingham, England, Abr. 2023.

[30] W.-Y. Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:14–23, 01 2011.

[31] J. Maroco. *Análise Estatística Com utilização do SPSS, 2ª edição.* Edições Sílabo, 2003.

[32] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning.* Adaptive Computation and Machine Learning series. MIT Press, London, England, 2 edition, Dez. 2018.

[33] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, Dez. 2013.

[34] A. S. Nyamawe, M. M. Mjahidi, N. E. Nnko, S. A. Diwani, G. G. Minja, and K. Malyango. *Practical machine learning.* Chapman & Hall/CRC, Philadelphia, PA, feb 2025.

[35] P. Ponniah. *Data warehousing fundamentals for IT professionals.* Wiley--Blackwell, Hoboken, NJ, 2 edition, Mai. 2010.

[36] S. Raschka. *Python Machine Learning.* Packt Publishing, Birmingham, England, Abr. 2023.

[37] H. Salman, A. Kalakech, and A. Steiti. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69–79, Jun. 2024.

[38] B. Sjardin, L. Massaron, and A. Boschetti. *Large scale machine learning with python.* Packt Publishing, Birmingham, England, Jan. 2016.

[39] M. Swamynathan. *Mastering machine learning with python in six steps.* APRESS, New York, NY, 1 edition, Jun. 2017.

[40] O. Theobald. *Machine learning for absolute beginners.* Independently Published, Abr. 2017.

[41] B. Ville. *Decision Trees - What Are They? In Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner.* Sas Inst, 2006.

UNIVERSIDADE DE ÉVORA
INSTITUTO DE INVESTIGAÇÃO
E FORMAÇÃO AVANÇADA

Contactos:
Universidade de Évora
Instituto de Investigação e Formação Avançada — IIFA
Palácio do Vimioso | Largo Marquês de Marialva, Apart. 94
7002 - 554 Évora | Portugal
Tel: (+351) 266 706 581
Fax: (+351) 266 744 677
email: iifa@uevora.pt