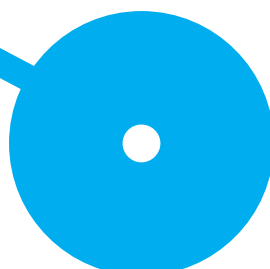


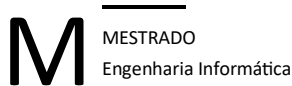
Exploração de técnicas de inteligência Artificial para personalização e otimização de campanhas omnicanal no Retalho Alimentar

Fernando José Cardoso da Silva

07/2025



ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO



Exploração de técnicas de inteligência Artificial para personalização e otimização de campanhas omnicanal no Retalho Alimentar

Fernando José Cardoso da Silva

8230755

Orientador

Professor Doutor João Ricardo Martins Ramos

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática pela Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto.

Declaração de integridade

Eu, Fernando José Cardoso da Silva, estudante nº 8230755, do Mestrado em Engenharia Informática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, declaro que não fiz plágio nem auto-plágio, pelo que o trabalho intitulado “Exploração de técnicas de Inteligência Artificial para personalização e otimização de campanhas omnicanal no Retalho alimentar”. É original e da minha autoria, não tendo sido usado previamente para qualquer outro fim. Mais declaro que todas as fontes usadas estão citadas, no texto e na bibliografia final, segundo as regras de referenciação adotadas na instituição.

Abstract

The increasing digitalization in the food retail sector, coupled with the evolution of consumer preferences, demands innovative solutions for personalization and integration of omnichannel campaigns. This thesis explores the use of clustering algorithms and LLMs (Large Language Models) to create a data-driven approach that improves campaign efficiency and increases customer loyalty and ROI (Return On Investment).

The research focuses on developing a framework that integrates dispersed customer behavior, profile, and campaign data, segments customers using clustering algorithms based on variables such as interaction value and time to purchase, applies predictive models to estimate conversions and validate decisions, generates personalized messages with LLMs adapted to segment, campaign type, and channel, and optimizes campaigns by analyzing ROI, average cost per conversion, and engagement.

Unlike previous studies that separately analyze aspects such as segmentation or personalization, this approach integrates robust techniques to create holistic solutions adapted to the consumer's omnichannel journey. The CRISP-DM methodology was adopted to ensure rigor, scalability, and efficiency in data analysis, allowing to overcome challenges such as information fragmentation and lack of integration between channels.

The results demonstrate that the proposed framework not only improves the customer experience but also increases operational efficiency and ROI, creating significant competitive advantages for retailers. The integration of AI technologies emerges as a critical solution to overcome traditional limitations, providing real-time personalization, predictive analysis, and operational efficiency.

This thesis contributes to the advancement of knowledge about AI in omnichannel marketing, offering practical and theoretical solutions for a constantly evolving sector, positioning food retail as an example of customer-centered innovation.

Keywords: Artificial Intelligence, Clustering, Large Language Models, Food Retail, ROI, Omnichannel Marketing.

Resumo

A crescente digitalização no setor do retalho alimentar, associada à evolução das preferências dos consumidores, exige soluções inovadoras para personalização e integração de campanhas omnicanal. Esta tese explora a utilização de algoritmos de clustering e LLMs (Large Language Models) para criar uma abordagem baseada em dados que permite melhorar a eficiência das campanhas e aumentar a fidelização e o ROI (Return on Investment).

A investigação centra-se no desenvolvimento de uma *framework* que integra dados dispersos de comportamento, perfil e campanhas dos clientes, segmenta-os com algoritmos de clustering baseados em variáveis como valor de interação e tempo até compra, aplica modelos preditivos para estimar conversões e validar decisões, gera mensagens personalizadas com LLMs adaptadas ao segmento, tipo de campanha e canal, e otimiza campanhas analisando o ROI, custo médio por conversão e *engagement*.

Diferente de estudos anteriores, que analisam separadamente aspetos como segmentação ou personalização, esta abordagem integra técnicas robustas para criação de soluções holísticas adaptadas à jornada omnicanal do consumidor. A metodologia CRISP_DM foi adotada para garantir rigor, escalabilidade e eficiência na análise de dados, permitindo superar desafios como a fragmentação de informações e a falta de integração entre canais.

Os resultados demonstram que a *framework* proposta não só melhora a experiência do cliente, mas também aumenta a eficiência operacional e o ROI, criando vantagens competitivas significativas para os retalhistas. A integração de tecnologias de IA emerge como uma solução crítica para superar limitações tradicionais, proporcionando personalização em tempo real, análise preditiva e eficiência operacional.

Esta dissertação contribui para o avanço do conhecimento sobre IA no *marketing* omnicanal, oferecendo soluções práticas e teóricas para um setor em constante evolução, posicionando o retalho alimentar como um exemplo de inovação centrada no cliente.

Palavras-chave: Inteligência Artificial, Clustering, Large Language Models, *Food Retail*, ROI, *Omnichannel Marketing*.

Conteúdo

Lista de Tabelas	ix
Lista de Figuras	x
Lista de Acrónimos.....	xi
1 Introdução.....	1
1.1 Apresentação e oportunidade do Tema.....	1
1.2 Objetivos da Investigação	3
1.2.1 Desenvolvimento de uma <i>framework</i> baseada em IA para campanhas omnicanal	3
1.2.2 Exploração do impacto da digitalização e da IA na personalização e eficiência operacional	4
1.2.3 Contribuir para o desenvolvimento de estratégias inovadoras no retalho alimentar	4
1.3 Contributos inovadores	5
1.3.1 <i>Framework</i> integrada para personalização omnicanal.....	6
1.3.2 Otimização baseada em dados para maximização do ROI	7
1.3.3 Personalização avançada com <i>Large Language Models</i>	7
1.4 Estrutura da dissertação.....	8
2 Revisão da literatura	10
2.1 Contextualização do omnicanal no retalho alimentar	10
2.2 Diferenciação entre omnicanal e multicanal	10
2.3 Benefícios da integração entre canais.....	11
2.4 Desafios na implementação de estratégias omnicanal.....	12
2.5 Inteligência artificial no <i>Marketing</i> omnicanal.....	13
2.5.1 Papel da IA na personalização e otimização	13
2.5.2 Segmentação de consumidores e previsão de comportamentos.....	14
2.5.3 Modelos de linguagem e personalização em tempo real.....	15
2.5.4 Casos de aplicação em empresas de referência	16
2.6 Técnicas de IA aplicadas ao Retalho	17
2.6.1 Algoritmos de <i>clustering</i>	17
2.6.2 Modelos RFM (<i>Recency, Frequency, Monetary</i>).....	18
2.6.3 <i>Large Language Models</i>	19
2.6.4 Análise Preditiva e prescritiva	19
2.7 Considerações finais.....	20
3 Metodologia de Investigação	22
3.1 Abordagem metodológica.....	22

3.1.1	Justificação da escolha do CRISP-DM	23
3.1.2	Descrição das etapas do processo	24
3.2	Compreensão do problema de negócio	26
3.2.1	Definição dos objetivos estratégicos	26
3.2.2	Identificação das necessidades de personalização.....	27
3.2.3	Integração de dados entre canais	29
3.3	Compreensão e preparação dos dados.....	30
3.3.1	Geração de dados sintéticos	30
3.3.2	Variáveis Relevantes para o estudo	34
3.3.3	Processo de tratamento e transformação de dados	35
3.3.4	Criação de variáveis derivadas	37
3.4	Modelação	38
3.4.1	Implementação de algoritmos de <i>clustering</i>	38
3.4.2	Análise de Valor com Modelo RFM	39
3.4.3	Previsão de conversão com <i>Random Forest</i>	40
3.4.4	Geração de mensagens personalizadas com LLMs	40
3.4.5	Integração e automatização dos componentes.....	41
3.5	Avaliação.....	41
3.5.1	Métricas Quantitativas.....	41
3.5.2	Análise de conversão, retenção e ROI simulado.....	42
3.5.3	Avaliação Qualitativa da aplicabilidade	42
3.6	Implementação	43
3.6.1	Descrição do processo de implementação.....	43
3.6.2	Documentação para replicabilidade	44
3.6.3	Ambiente de Desenvolvimento e Configuração Técnica.....	45
4	A <i>framework</i> proposta.....	45
4.1	Visão geral da <i>framework</i>	46
4.1.1	Arquitetura e componentes principais	46
4.1.2	Fluxo de processamento de dados.....	48
4.2	Integração de dados	50
4.2.1	Métodos para integração de dados dispersos	50
Definição das variáveis.....		52
4.2.2.....		52
4.2.3	Estatística descritiva multivariada	56
4.2.4	Fontes de dados e interoperabilidade	59
4.3	Segmentação de clientes.....	61

4.3.1	Algoritmos de <i>clustering</i> implementados	61
4.3.2	Métricas de avaliação da segmentação.....	62
4.3.3	Variáveis utilizadas	62
4.3.4	Resultados da segmentação	62
4.4	Modelos Preditivos	63
	Conceitos Gerais e validação de Modelos supervisionados.....	64
4.4.1	64
4.4.2	Estimativa de conversão	66
4.4.3	Validação com métricas e SHAP	67
4.4.4	Interpretabilidade e aplicação	67
4.5	Geração de Mensagens Personalizadas	68
4.5.1	Utilização de <i>Prompts</i> Estruturados para simular LLMs	68
4.5.2	Adaptação Tridimensional: segmento, canal e Tipo de campanha	69
4.5.3	Exemplos de Mensagens Geradas	70
4.6	Otimização de Campanhas.....	70
4.6.1	Análise de ROI.....	71
4.6.2	Custo Médio por conversão.....	71
4.6.3	<i>Engagement</i> por canal e segmento	72
4.6.4	Folhetos Personalizados como ferramenta de ativação	72
4.6.5	Construção automatizada e personalizada via GPT-4	72
4.6.6	Design visual e relevância estratégica	73
4.6.7	Integração com a <i>framework</i> de <i>Marketing</i> preditivo.....	73
4.7	Apresentação de resultados	74
4.7.1	Relatório Automatizado	74
4.7.2	<i>Dashboard</i> Interativo com streamlit	75
5	Resultados e discussão	75
5.1	Resultados da implementação.....	76
5.1.1	Caracterização dos segmentos identificados pelo <i>clustering</i>	76
5.1.2	Desempenho dos algoritmos de <i>clustering</i>	77
5.1.3	Eficácia dos Modelos Preditivos	79
5.1.4	Qualidade das Mensagens Personalizadas	81
5.1.5	Métricas de otimização de campanhas	82
5.1.6	Resultados Detalhados.....	85
5.1.7	Limitações dos Dados Sintéticos.....	86
5.2	Análise comparativa	87
5.2.1	Comparação com métodos tradicionais	87

5.2.2	Vantagens e limitações da abordagem proposta	89
5.3	Implicações práticas	92
6	Conclusão geral	93
6.1	Contribuições Teóricas e práticas	94
6.1.1	Contribuições Teóricas	94
6.1.2	Contribuições Práticas	96
6.2	Limitações do Estudo	97
6.2.1	Limitações Metodológicas	97
6.2.2	Limitações práticas	97
6.3	Sugestões para Investigação Futura.....	98
7	Bibliografia	99
8	Anexos	101

Lista de Tabelas

Tabela 1 Tipologia e escala de medição das variáveis do dataset	52
Tabela 2 Estatística descritiva das variáveis numéricas e ordinais	53
Tabela 3 Estatística descritivas das variáveis categóricas nominais	54
Tabela 4 Métricas de desempenho da segmentação por clustering	78
Tabela 5 Média dos valores RFM por segmento de cliente	79
Tabela 6 tabela de desempenho por segmento Cliente_Premium	83
Tabela 7 Tabela de desempenho por segmento Engajado_Lento	83
Tabela 8 Tabela de desempenho por segmento Indeciso	83
Tabela 9 Tabela de desempenho por segmento Iteracao_Baixa	84
Tabela 10 Resumo desempenho por canal	84
Tabela 11 Desempenho de campanhas por canal	85

Lista de Figuras

Figura 1 Esquema do modelo CRISP-DM e respetivas fases	24
Figura 2 Média de dias até à compra por canal e género	56
Figura 3 Valor médio de interação por segmento e tipo de campanha	57
Figura 4 Valor médio de interação por segmento e canal de interação	58
Figura 5 Taxa média de conversão por tipo de dispositivo e género	59
Figura 6 Distribuição dos segmentos comportamentais em função do tempo até conversão e do valor de interação	63
Figura 7 F1 score por Fold com Random Forest.....	65
Figura 8- Dispersão dos clusters de comportamento	77
Figura 9 Matriz de confusão.....	80
Figura 10 Gráfico SHAP individual.....	81

Lista de Acrónimos

AOT - *Automated Order Taking*

API – *Application Programming Interface*

CAC – *Customer acquisition cost*

CDP – *Customer Data Platform*

CLV - *Customer Lifetime Value*

CPA – *Custo por aquisição*

CPR – *Custo por retenção*

CRISP-DM – *Cross Industry Standard Process for Data Mining*

ETL – *Extract, Transform, load*

IA – *Inteligência Artificial*

IQR - *Interquartile Range*

LLMs – *Large language models*

NLP – *Natural language processing*

POS – *Point of Sale (Ponto de Venda)*

RFM *Recency, Frequency, and Monetary value*

RGPD - *Regulamento Geral sobre a Proteção de Dados*

ROI – *Return on investment*

SHAP - *SHapley Additive exPlanations*

1 Introdução

O presente capítulo introduz o tema central da dissertação, destacando a relevância da transformação e da adoção de estratégias omnicanal no setor do retalho alimentar. São enquadrados os principais desafios que motivam este trabalho, nomeadamente a fragmentação de dados entre canais e a dificuldade em proporcionar experiências personalizadas e integradas aos consumidores. Adicionalmente, são apresentados objetivos gerais e específicos da investigação, os contributos inovadores que se pretendem alcançar e a estrutura global, de forma a proporcionar ao leitor uma visão clara e fundamentada sobre a problemática e o percurso analítico desenvolvido.

1.1 Apresentação e oportunidade do tema

O setor do retalho alimentar encontra-se numa fase de transformação significativa impulsionada pela evolução das expectativas dos consumidores, que exigem experiências integradas, consistentes e personalizadas ao longo de diversos pontos de contacto. Este ambiente dinâmico obriga os retalhistas a adotar estratégias omnicanal robustas, capazes de oferecer uma jornada fluida entre lojas físicas, plataformas de *e-commerce* e redes sociais. A digitalização emerge como um catalisador fundamental desta mudança, permitindo uma maior conexão entre canais online e offline e alterando profundamente as dinâmicas tradicionais do setor (Hagberg et al., 2016).

A transformação digital no retalho alimentar não é apenas uma tendência passageira, mas uma necessidade estratégica para a sobrevivência e crescimento num mercado cada vez mais competitivo. Os consumidores contemporâneos esperam poder iniciar uma compra num canal e concluí-la noutro, consultar informações sobre produtos online enquanto estão na loja física, ou receber recomendações personalizadas com base no seu histórico de compras. Esta realidade exige uma integração perfeita entre os diferentes pontos de contacto, algo que muitas empresas ainda lutam para alcançar.

Apesar das oportunidades criadas por esta integração, a fragmentação de bases de dados entre canais e a falta de consistência continuam a ser desafios significativos. Esta realidade dificulta a personalização de campanhas e a avaliação do impacto das iniciativas de *marketing*. Estudos de Cui et al. (2022) sublinham a importância de uma interatividade

eficaz entre canais para promover a cocriação de valor, mas alertam que a ausência de integração ainda constitui uma barreira frequente para muitos retalhistas.

A fragmentação de dados representa um obstáculo particularmente desafiador. Muitas empresas do setor alimentar operam com sistemas isolados para as suas lojas físicas, plataformas de *e-commerce* e aplicações móveis, resultando em silos de informação que impedem uma visão unificada do cliente. Quando os dados estão dispersos, torna-se praticamente impossível criar experiências personalizadas e consistentes, medir o desempenho das campanhas de *marketing* ou compreender a jornada completa do consumidor.

Neste contexto, a IA (Inteligência Artificial) apresenta-se como uma solução disruptiva. Dados recentes partilhados no evento *Onclusive on the Road* revelam que, embora haja uma abundância de dados disponíveis, 70% dos *marketeers* ainda dependem de métodos tradicionais como inquéritos para medir a satisfação e o sucesso das suas campanhas. Esta lacuna evidencia a necessidade de ferramentas analíticas avançadas que permitam compreender as nuances do comportamento do consumidor e traduzir os dados em decisões estratégicas fundamentadas.

A inteligência artificial oferece capacidades sem precedentes para processar e analisar grandes volumes de dados, identificar padrões complexos e gerar *insights* acionáveis em tempo real. Algoritmos de aprendizagem automática podem segmentar clientes com base em comportamentos semelhantes, prever tendências de consumo e personalizar interações de forma dinâmica. Modelos de linguagem avançados podem gerar conteúdo personalizado adaptado às preferências específicas de cada segmento, enquanto sistemas de recomendação podem sugerir produtos relevantes com base no histórico de compras e comportamento de navegação.

Ao implementar tecnologias com integração de IA, como aprendizagem automática, IA generativa e *clustering*, torna-se possível ultrapassar estas barreiras. Essas ferramentas permitem analisar grandes volumes de dados, identificar padrões de comportamento e personalizar interações em tempo real (Sovrano et al., 2021). Além disso, a análise prescritiva como destacado por (Hauser et al., 2021), pode otimizar as operações em ambientes de loja inteligente, melhorando a eficiência e a experiência do cliente.

Este trabalho propõe explorar o papel da IA na personalização e otimização de campanhas omnicanal no retalho alimentar. Algoritmos como *clustering*, RFM (*Recency, Frequency, and Monetary value*) e análises preditivas serão avaliados pela sua capacidade de

segmentar consumidores de forma precisa, antecipar tendências de consumo e ajustar estratégias de *marketing* de forma dinâmica.

Ao integrar estas tecnologias, o retalho alimentar pode maximizar a eficiência operacional, melhorar a experiência do cliente e aumentar o ROI (*Return Of Investment*). Esta dissertação visa investigar as potencialidades e os desafios da aplicação de IA no setor, demonstrando o seu papel determinante na criação de estratégias omnicanal eficazes e transformadoras.

1.2 Objetivos da investigação

Esta investigação estabelece três objetivos principais que orientam o desenvolvimento do trabalho e definem o seu alcance e contribuição para o campo do *marketing* omnicanal no retalho alimentar: desenvolvimento de uma *framework* baseada em IA para campanhas omnicanal; exploração do impacto da digitalização e da IA na personalização e eficiência operacional; e contribuição para o desenvolvimento de estratégias inovadoras no retalho alimentar. Cada um destes tópicos é descrito nas subsecções subsequentes.

1.2.1 Desenvolvimento de uma *framework* baseada em IA para campanhas omnicanal

O principal objetivo desta dissertação consiste no desenvolvimento de uma *framework* integrada que recorra a técnicas de inteligência artificial para otimizar campanhas de *marketing* omnicanal no setor do retalho alimentar. A abordagem proposta combina algoritmos de *clustering K-Means*, análise de valor de cliente RFM, modelos preditivos (exemplo, *Random Forest*) e modelos de linguagem de grande escala (LLM – *Large Language Model*), permitindo personalizar campanhas, prever padrões de consumo e ajustar estratégias com base em dados reais.

A *framework* visa ultrapassar a fragmentação típica de dados entre canais digitais e físicos, promovendo a integração de fontes diversas num ecossistema coerente. Este alinhamento facilita a construção de experiências consistentes ao longo da jornada do cliente, reforçando a eficácia das campanhas ao assegurar que cada contacto é relevante, personalizado e oportuno.

Além disso, a estrutura desenvolvida permite extrair *insights* acionáveis a partir de grandes volumes de dados heterogêneos, identificar segmentos de consumidores com base em comportamentos reais e gerar recomendações adaptadas a cada perfil. A utilização de *SHAP (SHapley Additive exPlanations)* complementa esta abordagem ao proporcionar explicabilidade sobre os fatores que mais contribuem para a probabilidade de conversão, promovendo decisões baseadas em evidência.

1.2.2 Exploração do impacto da digitalização e da IA na personalização e eficiência operacional

O segundo objetivo da investigação é analisar o impacto da transformação digital e da inteligência artificial na personalização de campanhas e na melhoria da eficiência operacional no retalho alimentar. Através da *framework* desenvolvida, demonstra-se como algoritmos de *clustering* e modelos de aprendizagem supervisionada, como por exemplo *Random Forest* ou redes neurais artificiais, podem ser utilizados para segmentar clientes de forma inteligente, prever conversões e direcionar recursos de marketing de forma eficaz.

A utilização de modelos preditivos permite não apenas prever a probabilidade de conversão, mas também avaliar quais as variáveis mais com maior relevância nesse processo. A complementaridade com métricas RFM oferece uma visão aprofundada do valor e da fidelidade dos clientes, enquanto os LLMs permitem criar mensagens promocionais contextuais e personalizadas com base nos dados comportamentais.

A capacidade de automatizar tarefas, otimizar campanhas e reduzir dependências de abordagens empíricas permite às empresas melhorar os resultados com menor esforço operacional. Pretende-se o desenvolvimento de uma *framework* que possa ser escalável, podendo adaptar-se facilmente a diferentes contextos organizacionais e níveis de maturidade digital.

1.2.3 Contribuir para o desenvolvimento de estratégias inovadoras no retalho alimentar

O terceiro objetivo visa fornecer *insights* teóricos e práticos que auxiliem retalhistas a adotar práticas transformadoras, reforçando a sua competitividade do mercado e

consolidando o papel da IA como um fator disruptivo e determinante no *marketing* omnicanal.

Com esta dissertação pretende-se impulsionar o avanço do conhecimento na área do *marketing* omnicanal, através da identificação de tendências emergentes, melhores práticas e oportunidades de inovação. Ao explorar a interseção entre inteligência artificial, *marketing* e comportamento do consumidor, o trabalho propõe oferecer uma perspectiva diferenciadora e valiosa, capaz de orientar estratégias futuras no contexto do retalho alimentar.

A principal inovação reside na integração de múltiplas técnicas num fluxo contínuo — desde a geração de dados comportamentais e a segmentação via *K-Means*, passando pela previsão com *Random Forest* e explicação com métodos interpretáveis como SHAP, até à criação automatizada de conteúdo personalizado com modelos LLMs. Pretende-se demonstrar que este ciclo pode possibilitar uma atuação em tempo real, a adaptação de mensagens ao perfil do cliente e a sugestão de folhetos personalizados com base no histórico de interações.

Ao integrar técnicas de inteligência artificial com foco na personalização, predição e eficiência, esta investigação ambiciosa não só melhorar os resultados financeiros das empresas, mas também elevar a experiência do consumidor. Desta forma, o retalho alimentar poderá afirmar-se como um setor mais inovador, resiliente e centrado nas necessidades do cliente.

1.3 Contributos inovadores

Esta dissertação propõe uma *framework* inovadora baseada em inteligência artificial para o retalho alimentar, combinando algoritmos não supervisionados como *clustering*, modelos preditivos como RFM e LLMs para personalização de campanhas. A abordagem destaca-se pela criação de campanhas omnicanal altamente personalizadas, adaptadas aos comportamentos dos consumidores, promovendo segmentação detalhada, antecipação de tendências e geração de conteúdos dinâmicos em tempo real

1.3.1 *Framework* integrada para personalização omnicanal

O principal contributo inovador desta investigação reside no desenvolvimento de uma *framework* abrangente e modular que integra múltiplas técnicas de inteligência artificial, oferecendo uma solução holística para o *marketing* omnicanal no setor do retalho alimentar. Ao contrário de abordagens tradicionais, que tendem a focar-se em componentes isolados, como a segmentação ou a personalização, a *framework* proposta contempla de forma integrada todo o ciclo de vida das campanhas de *marketing*, desde a análise comportamental até à execução personalizada e avaliação de resultados.

A estrutura da *framework* articula diferentes técnicas de IA, cada uma com um papel específico e complementar:

- ***K-Means*** para segmentação de clientes com base em comportamentos reais e características de interação;
- **Modelos RFM** para avaliação da fidelização e do valor do cliente;
- ***Random Forest*** para previsão da probabilidade de conversão, com elevada precisão e robustez;
- ***SHAP*** para interpretação dos modelos preditivos, assegurando transparência e explicabilidade;
- ***GPT-4 (Large Language Models)*** para a geração automática de mensagens promocionais e conteúdos personalizados de folhetos, adaptados ao perfil, comportamento e canal mais adequado para cada consumidor.

Esta combinação positiva permite uma abordagem coesa e orientada ao cliente, capaz de acompanhar toda a jornada omnicanal, desde a identificação de padrões e segmentos, passando pela previsão de comportamentos, até à entrega de comunicações relevantes nos canais mais eficazes. Um dos aspetos distintivos desta *framework* é a sua capacidade de adaptação dinâmica, ajustando estratégias em tempo real com base em dados continuamente atualizados. Esta flexibilidade é decisiva num setor altamente competitivo e volátil como o retalho alimentar, onde a capacidade de resposta às mudanças nos hábitos e preferências dos consumidores se constitui como um fator determinante de sucesso.

1.3.2 Otimização baseada em dados para maximização do ROI

Além de potencializar o *engagement* e aumentar o ROI, a *framework* proposta visa otimizar a segmentação de públicos e a compreensão das preferências em ambientes omnicanal, permitindo uma personalização mais precisa e eficaz das campanhas. Através da aplicação de algoritmos de *clustering* e modelos preditivos, torna-se possível identificar padrões de comportamento dos consumidores, prever interações futuras e ajustar dinamicamente as estratégias de *marketing* em função dessas previsões.

Esta abordagem recorre a técnicas avançadas de análise de dados, que permitem uma alocação mais eficiente dos recursos de *marketing*, assegurando que os investimentos são dirigidos aos canais e segmentos com maior potencial de retorno. Esta lógica orientada por dados representa uma evolução significativa face às abordagens tradicionais, muitas vezes baseadas em intuição, heurísticas ou experiências anteriores.

Um dos principais contributos desta *framework* reside também na definição e aplicação de métricas específicas para avaliação de campanhas omnicanal, que vão além das conversões imediatas. Estas métricas consideram indicadores como a fidelização, o valor do cliente ao longo do tempo CLV (*Customer Lifetime Value*) e a qualidade da experiência com a marca, promovendo uma visão mais holística e estratégica da eficácia do *marketing* no retalho alimentar.

1.3.3 Personalização avançada com *large language models*

A integração de LLMs na *framework* representa um avanço significativo e altamente diferenciador na forma como se comunica com o consumidor no contexto omnicanal. Estes modelos permitem gerar mensagens personalizadas com base em dados reais, como o histórico de compras, o segmento comportamental atribuído ao cliente e os produtos com maior afinidade, tornando a comunicação não só mais relevante e contextual, mas também mais envolvente e persuasiva.

Um dos elementos mais inovadores da *framework* é a capacidade de produzir automaticamente folhetos promocionais personalizados, onde cada cliente recebe uma proposta visual única — composta por mensagens adaptadas e imagens reais de produtos compatíveis com os seus interesses e padrões de consumo. Este processo vai além da personalização textual, incorporando elementos visuais e narrativas personalizadas que reforçam o vínculo emocional com a marca.

Ao combinar LLMs com perfis de comportamento gerados por *clustering* e dados de interações anteriores, a *framework* promove uma experiência omnicanal altamente coesa, permitindo que a mesma campanha seja reconhecida e valorizada em qualquer ponto de contacto — seja digital ou físico. Esta abordagem não só melhora a percepção de personalização, como também aumenta a eficácia da comunicação e o potencial de conversão, contribuindo para a fidelização e o crescimento sustentado do valor do cliente.

1.4 Estrutura da dissertação

A presente dissertação está organizada em seis capítulos principais, estruturados de forma lógica e sequencial, de modo a proporcionar uma compreensão clara e progressiva do tema em estudo, desde a sua contextualização até às conclusões finais.

- **Capítulo 1 – Introdução**

Apresenta o enquadramento geral do trabalho, destacando a relevância da transformação digital no setor do retalho alimentar, os desafios associados à fragmentação de dados e à integração de canais, e a importância da personalização baseada em inteligência artificial. São também definidos os objetivos da investigação, os contributos esperados e as questões orientadoras da pesquisa.

- **Capítulo 2 – Revisão da literatura**

Explora o estado da arte sobre *marketing* omnicanal e as aplicações de inteligência artificial no retalho. Diferencia o paradigma omnicanal do multicanal, analisa os principais benefícios e limitações do modelo, e aprofunda a utilização de técnicas como *clustering*, RFM e LLMs na personalização de campanhas. São ainda incluídos exemplos práticos e casos de aplicação relevantes.

- **Capítulo 3 – Metodologia de investigação**

Descreve a abordagem metodológica adotada, fundamentando a escolha do modelo CRISP-DM como estrutura para o desenvolvimento do projeto. São detalhadas as etapas seguidas, os métodos de geração e tratamento de dados

sintéticos, os algoritmos aplicados, e as métricas utilizadas na avaliação dos resultados.

- **Capítulo 4 – *Framework* proposta**

Apresenta em detalhe a *framework* desenvolvida para personalização e otimização de campanhas omnicanal no retalho alimentar. São descritos os seus principais componentes, como a integração de dados, a segmentação de clientes, os modelos preditivos, a geração de mensagens personalizadas com LLMs, a análise de desempenho e a lógica de otimização de campanhas.

- **Capítulo 5 – Resultados e discussão**

Analisa os resultados obtidos com a implementação da *framework*, avaliando o desempenho dos algoritmos utilizados, a qualidade das mensagens geradas, o impacto nas métricas de conversão, ROI e *engagement*. Inclui também uma comparação com abordagens tradicionais e uma reflexão crítica sobre as implicações práticas da solução proposta.

- **Capítulo 6 – Conclusão e trabalhos futuros**

Sintetiza os principais contributos da investigação, reforçando o papel da inteligência artificial como fator transformador no *marketing* omnicanal. São ainda discutidas as limitações do estudo e propostas direções para investigações futuras, tendo em conta os avanços tecnológicos e a evolução do comportamento do consumidor.

Complementarmente, a dissertação inclui uma secção de Referências Bibliográficas com todas as fontes académicas utilizadas, bem como Anexos com exemplos de outputs gerados e outros materiais de apoio relevantes.

2 Revisão da literatura

A presente revisão da literatura tem como objetivo analisar os conceitos, modelos e tendências que sustentam a transformação do *marketing* no retalho alimentar, em particular a evolução do paradigma multicanal para uma abordagem omnicanal centrada no cliente. Neste enquadramento, será explorado o papel determinante da IA como motor de personalização, otimização e automação de processos. O capítulo apresenta uma visão abrangente dos benefícios e desafios associados à integração de canais, assim como das técnicas de IA mais relevantes, incluindo algoritmos de *clustering*, modelos RFM, LLMs e metodologias de análise preditiva. Esta revisão permite fundamentar a *framework* proposta e clarificar o potencial impacto estratégico da sua aplicação no contexto do retalho alimentar.

2.1 Contextualização do omnicanal no retalho alimentar

O conceito de omnicanal refere-se à integração completa e harmoniosa entre todos os canais de venda e comunicação de uma organização, com o objetivo de proporcionar ao consumidor uma experiência contínua, fluida e coerente ao longo de toda a sua jornada, independentemente do ponto de contacto utilizado — seja ele digital ou físico. Esta abordagem representa uma evolução estratégica face ao modelo multicanal, no qual os canais coexistem de forma paralela, mas geralmente operam de forma isolada, sem uma verdadeira articulação entre si, o que compromete a consistência da experiência do cliente.

2.2 Diferenciação entre omnicanal e multicanal

No modelo multicanal, as empresas disponibilizam diversos pontos de contacto, como lojas físicas, *websites*, aplicações móveis e redes sociais, mas esses canais operam de forma isolada, sem uma partilha eficaz e contínua de informação entre si. Esta ausência de integração pode resultar em experiências inconsistentes, dependendo do canal escolhido, gerando fricções desnecessárias ao longo da jornada de compra e dificultando a construção de uma relação coesa com o cliente.

Em contraste, o modelo omnicanal garante uma interligação estratégica entre os diversos canais, permitindo ao consumidor transitar de forma fluida entre plataformas, com continuidade e coerência na sua experiência. Segundo Juaneda-Ayensa et al. (2016), esta abordagem coloca o consumidor no centro das decisões empresariais, eliminando barreiras entre canais e promovendo uma experiência verdadeiramente holística e personalizada. Por exemplo, um cliente pode explorar um produto online, verificar a sua disponibilidade numa loja física próxima, efetuar a compra através de uma aplicação móvel e optar entre levantar o produto na loja ou recebê-lo em casa.

Esta integração omnicanal representa um avanço substancial na forma como as empresas se relacionam com os seus clientes, proporcionando uma experiência de compra mais cómoda, consistente e centrada nas necessidades individuais. No contexto do retalho alimentar, onde a concorrência é elevada e as margens comerciais são frequentemente estreitas, a capacidade de oferecer uma experiência omnicanal superior pode revelar-se um fator diferenciador fundamental e uma vantagem competitiva determinante.

2.3 Benefícios da integração entre canais

Lemon & Verhoef. (2016) defendem que a adoção de estratégias omnicanal não só aumenta o valor das compras realizadas pelos consumidores, como também promove uma maior lealdade à marca, ao assegurar uma ligação contínua e coerente entre os canais online e offline. De forma complementar, Gao et al. (2021) reforçam que a integração eficaz dos canais é um fator crítico para garantir a continuidade da experiência do cliente, traduzindo-se num aumento significativo do *engagement* e da taxa de retenção.

Entre os principais benefícios associados à abordagem omnicanal destaca-se a consistência da informação disponibilizada ao longo dos diferentes pontos de contacto. Estudos de Juaneda-Ayensa et al. (2016) revelam que os consumidores valorizam fortemente a uniformidade de preços, promoções e características dos produtos, uma vez que esta reduz a perceção de risco e reforça a confiança na marca. Adicionalmente, dados estatísticos indicam que 81% dos clientes utilizam múltiplos canais durante o processo de compra e que os consumidores omnicanal tendem a gastar até quatro vezes mais do que aqueles que utilizam apenas um canal.

A integração de tecnologias digitais no ambiente físico tem também desempenhado um papel determinante no reforço da experiência omnicanal. Soluções como quiosques

interativos, sinalização digital e aplicações de loja aumentam o tempo de permanência dos consumidores no ponto de venda e incentivam decisões de compra adicionais, enriquecendo o processo de consumo (Melero et al., 2016). Do ponto de vista operacional, Verhoef et al. (2015) sublinham que a sinergia entre canais permite reduzir redundâncias, aumentar a eficiência dos processos e otimizar a gestão de inventário, consolidando o modelo omnicanal como uma solução sustentável, escalável e orientada para o futuro do retalho.

2.4 Desafios na implementação de estratégias omnicanal

Apesar dos benefícios amplamente reconhecidos, muitas organizações continuam a enfrentar obstáculos significativos na implementação eficaz de estratégias omnicanal, sobretudo no que respeita à integração de dados e à harmonização tecnológica. Juaneda-Ayensa et al. (2016) alertam que a ausência de congruência operacional entre canais pode comprometer seriamente a experiência do cliente e limitar o *engagement* ao longo da jornada de compra.

Um dos maiores entraves é a fragmentação de dados, que ocorre quando as informações sobre clientes, produtos e transações se encontram dispersas por múltiplos sistemas isolados. Esta dispersão dificulta a construção de uma visão 360º do cliente, impedindo a personalização eficaz das experiências e comprometendo a consistência das interações entre canais. Além disso, a falta de integração limita a capacidade das empresas em avaliar com precisão o desempenho das campanhas de *marketing* e compreender de forma holística a jornada do consumidor.

Outro desafio estrutural prende-se com a necessidade de adaptação da cultura organizacional e dos processos internos. Muitas empresas ainda operam com lógicas departamentais em silos, onde cada canal é gerido autonomamente, com métricas, objetivos e orçamentos próprios. Este modelo fragmentado tende a dificultar a colaboração entre equipas e a implementação de uma estratégia verdadeiramente integrada e centrada no cliente.

Neste cenário, a IA assume um papel transformador, oferecendo soluções que facilitam a automatização de processos, a personalização em escala e a análise preditiva. A capacidade da IA para processar grandes volumes de dados heterogêneos permite ultrapassar a barreira da fragmentação informacional, enquanto algoritmos avançados —

como os de *clustering*, recomendação e modelação preditiva — são capazes de identificar padrões de comportamento e ajustar interações em tempo real. Assim, a IA não só reforça a eficácia das estratégias omnicanal, como também contribui para uma abordagem mais ágil, escalável e centrada no cliente.

2.5 Inteligência artificial no *marketing* omnicanal

A IA tem o poder de revolucionar o *marketing* omnicanal, desempenhando um papel central na personalização das interações entre marcas e consumidores. Num ambiente onde a integração de dados e a eficiência operacional são essenciais, a IA apresenta-se como um catalisador para a otimização de campanhas, o aumento do ROI e a melhoria da experiência do cliente.

2.5.1 Papel da IA na personalização e otimização

A IA disponibiliza capacidades avançadas para processar e analisar grandes volumes de dados, identificar padrões complexos e gerar *insights* acionáveis em tempo real. No contexto do *marketing* omnicanal, estas capacidades traduzem-se em oportunidades substanciais para personalizar interações, otimizar campanhas e elevar a qualidade da experiência do cliente ao longo de toda a sua jornada.

A personalização suportada por IA ultrapassa largamente a tradicional inserção de nomes em mensagens de email. Algoritmos avançados têm a capacidade de analisar de forma simultânea o histórico de compras, padrões de navegação, preferências explícitas e implícitas, bem como o contexto situacional do cliente. Com base nestas análises, é possível gerar experiências hiperpersonalizadas, incluindo recomendações de produtos de alta relevância, conteúdos adaptados ao estilo de comunicação preferido e ofertas personalizadas em função das necessidades e comportamentos identificados.

Para além da vertente de personalização, a IA assume um papel determinante na otimização contínua de campanhas de *marketing*. Técnicas de aprendizagem automática permitem analisar o desempenho de campanhas anteriores, identificar padrões que conduzem ao sucesso e ajustar, em tempo real, as variáveis estratégicas com o objetivo de maximizar o ROI. Esta capacidade de adaptação dinâmica promove uma alocação mais

eficiente de recursos, direcionando os esforços para os canais, mensagens e segmentos com maior probabilidade de conversão.

Assim, a integração da IA no marketing omnicanal representa uma evolução estratégica, permitindo uma abordagem mais ágil, precisa e centrada no cliente, fundamental para enfrentar os desafios de um ambiente competitivo e em constante transformação.

2.5.2 Segmentação de consumidores e previsão de comportamentos

A segmentação precisa de consumidores e a previsão de comportamentos constituem pilares essenciais das estratégias suportadas por inteligência artificial. Modelos clássicos como RFM, aliados a algoritmos de *clustering* como o *K-means*, são amplamente aplicados para categorizar consumidores com base em comportamentos reais e identificar padrões latentes de compra. Esta abordagem permite o desenvolvimento de campanhas altamente personalizadas, alinhadas às características e preferências de cada grupo de clientes (Hagberg et al., 2016).

No setor do retalho alimentar, onde as margens de lucro são reduzidas e a sazonalidade influencia significativamente a procura, a aplicação de modelos preditivos tem gerado ganhos substanciais em eficiência e rentabilidade (Pelau et al., 2021). Através de técnicas de *clustering*, é possível detetar grupos de consumidores com comportamentos semelhantes, mesmo quando estes não são imediatamente visíveis através de variáveis demográficas. Esta segmentação comportamental e psicográfica oferece uma compreensão mais aprofundada das motivações, padrões de consumo e preferências contextuais, indo além das classificações tradicionais.

Adicionalmente, a previsão de comportamentos de compra em tempo real representa um fator diferenciador nas estratégias de *marketing* orientadas por dados. Gao et al. (2021) sublinham que a capacidade de resposta rápida, baseada em modelos preditivos, aumenta significativamente a eficácia das campanhas, contribuindo para um maior *engagement* e retenção. Neste contexto, a *framework* proposta contempla a utilização de modelos de aprendizagem supervisionada, como, por exemplo *Random Forest* ou redes neuronais artificiais, com o objetivo de antecipar a conversão de clientes com base em variáveis comportamentais, tais como o valor da interação, o tempo até à compra e o segmento atribuído.

De modo a garantir o melhor desempenho possível do sistema, pretende-se a validação com técnicas como *cross-validation* e complementação com técnicas de explicabilidade garantindo uma interpretação clara dos fatores que influenciam os resultados. Esta capacidade preditiva confere às empresas uma postura proativa, substituindo a tradicional abordagem reativa. Por exemplo, um modelo pode prever quando um cliente está prestes a abandonar uma jornada de compra ou qual canal tem maior probabilidade de gerar conversão, permitindo o ajuste imediato da estratégia de comunicação.

Este tipo de intervenção não só melhora a conveniência e a experiência do cliente, como reforça a perceção de valor e atenção personalizada por parte da marca, contribuindo para uma maior fidelização e rentabilidade.

2.5.3 Modelos de linguagem e personalização em tempo real

A integração de modelos de linguagem de grande escala, como o GPT-4, representa um avanço significativo na forma como as empresas podem comunicar com os seus clientes de forma personalizada, contextualizada e escalável. Na *framework* proposta, os LLMs serão utilizados não apenas para gerar mensagens promocionais personalizadas, mas também para construir conteúdos apelativos para folhetos digitais adaptados ao perfil de cada cliente.

A personalização deve ser realizada com base no segmento comportamental (resultado do *clustering*), no canal de contacto, no tipo de campanha e nos produtos mais frequentemente adquiridos em interações anteriores com conversão. Pretende-se que esta informação seja processada para gerar *prompts* dinâmicos, a submeter posteriormente a modelos de linguagem de grande escala, com o objetivo de obter mensagens humanizadas, apelativas e ajustadas ao contexto omnicanal. Esta abordagem procura substituir os tradicionais *templates* fixos por uma lógica mais flexível, centrada na adaptação em tempo real à jornada individual do consumidor.

De forma exploratória, pretende-se que os conteúdos personalizados incluam menções ao segmento (por exemplo, Cliente_Premium ou Indeciso), um conjunto de produtos-chave e um incentivo ou desconto ajustado ao perfil identificado. A integração destas respostas nos materiais de comunicação visa demonstrar a viabilidade de um ciclo de personalização end-to-end, da análise de dados à entrega de conteúdos personalizados.

Este tipo de personalização contextual permite aumentar significativamente a relevância da comunicação, o *engagement* dos consumidores e o retorno das campanhas, alinhando-se com os princípios de congruência comunicacional referidos por Juaneda-Ayensa et al. (2016) e Verhoef et al. (2015), que demonstram o impacto da adaptação ao ciclo de vida do cliente na fidelização e no ROI.

Embora o texto de suporte refira outras aplicações da inteligência artificial no retalho, como a automação do atendimento ao cliente ou a gestão de inventário, a presente *framework* concentra-se exclusivamente na aplicação prática de LLMs para personalização de conteúdo em campanhas promocionais omnicanal, demonstrando a sua viabilidade, escalabilidade e impacto estratégico.

2.5.4 Casos de aplicação em empresas de referência

Estas vantagens tornam-se particularmente evidentes quando se analisam exemplos concretos da aplicação da inteligência artificial por grandes empresas globais. A Amazon é um caso paradigmático que utiliza algoritmos avançados de previsão e personalização para analisar padrões de compra, sugerir recomendações altamente relevantes e otimizar toda a sua cadeia de abastecimento. Esta abordagem integrada de IA não só impulsionou significativamente as vendas, como também elevou a experiência do utilizador a novos patamares, contribuindo para uma receita de 469,8 mil milhões de dólares em 2021 (Walther, 2024).

O sistema de recomendação da Amazon é amplamente reconhecido como uma referência mundial na aplicação da inteligência artificial no retalho. Através de técnicas como filtragem colaborativa e aprendizagem profunda, o sistema é capaz de sugerir produtos não apenas com base no histórico de compras do cliente, mas também tendo em conta padrões comportamentais de utilizadores com perfis semelhantes. Estimativas apontam que este sistema seja responsável por até 35% das vendas totais da empresa, evidenciando o impacto direto da IA nos resultados financeiros.

No setor da restauração rápida, o McDonald's tem adotado soluções baseadas em IA para melhorar a eficiência do atendimento. Recorrendo à tecnologia IBM Watsonx e a técnicas de NLP (*Natural Language Processing*), a empresa automatizou o sistema de receção de pedidos AOT (*Automated Order Taking*), permitindo a sua adaptação a diferentes idiomas, sotaques e variações de menu. Esta inovação resultou numa redução de 25% no tempo de

espera e num aumento significativo da eficiência operacional, refletindo-se numa melhoria tanto na satisfação dos clientes como na rentabilidade da operação (Holdsworth, 2024).

Estes exemplos demonstram de forma clara como a inteligência artificial está a transformar não apenas o *marketing* omnicanal, mas também diversos setores de atividade, posicionando-se como um fator determinante para a inovação, a personalização e o crescimento sustentável no atual contexto digital.

2.6 Técnicas de IA aplicadas ao retalho

A aplicação de técnicas de IA no setor do retalho tem registado um grande crescimento nos últimos anos, oferecendo soluções inovadoras para desafios complexos relacionados com a segmentação de clientes, a personalização de interações e a otimização de campanhas. Entre as técnicas mais relevantes destacam-se os algoritmos de *clustering*, os modelos RFM e os LLMs, que em conjunto permitem uma abordagem mais precisa, dinâmica e centrada no cliente.

2.6.1 Algoritmos de *clustering*

Os algoritmos de *clustering* desempenham um papel fundamental na segmentação de clientes, permitindo identificar grupos com comportamentos semelhantes com base em dados históricos. Esta técnica é especialmente útil no setor do retalho alimentar, onde a compreensão de padrões de consumo é essencial para a personalização de campanhas.

O algoritmo *K-Means* é amplamente reconhecido pela sua simplicidade e eficiência. Esta abordagem permite agrupar clientes em segmentos com características comuns, como frequência de compra, valor médio por transação ou nível de interação com os canais da marca. Cada cliente é alocado ao *cluster* cuja média de características seja mais próxima, facilitando a criação de estratégias orientadas a cada grupo.

A utilização do *K-Means* possibilita a extração de segmentos relevantes e orientados por dados reais de comportamento, que podem servir como base para personalização de mensagens, previsão de conversão e recomendação de campanhas. Esta abordagem revela-se eficaz na criação de campanhas mais alinhadas com as preferências dos consumidores, potenciando a eficácia do *marketing* omnicanal.

Embora existam outros algoritmos de *clustering* mais avançados ou adequados a casos específicos, como dados com ruído ou *clusters* de forma não esférica, o *K-Means* apresenta-se como uma opção frequentemente utilizada pela sua aplicabilidade e resultados consistentes na segmentação de clientes.

2.6.2 Modelos RFM (*Recency, Frequency, Monetary*)

O modelo RFM é amplamente reconhecido como uma técnica eficaz para segmentar clientes com base no seu comportamento transacional. Este modelo classifica os consumidores segundo três dimensões fundamentais:

- **Recency (R):** quão recentemente o cliente efetuou uma compra;
- **Frequency (F):** com que regularidade realiza compras;
- **Monetary value (M):** quanto gastou no total durante um determinado período.

Esta abordagem permite identificar os clientes mais valiosos, bem como aqueles que podem necessitar de estratégias específicas de reativação (Fader et al., 2005).

No contexto do retalho alimentar, a aplicação do modelo RFM pode dar origem a diferentes perfis de clientes, como:

- **Cliente VIP:** com compras frequentes, recentes e de elevado valor;
- **Cliente Leal:** com frequência e valor consistentes ao longo do tempo;
- **Cliente Ocasional:** com comportamento esporádico e menor valor agregado;
- **Cliente Inativo:** sem atividade recente, com risco elevado de abandono.

Além da segmentação direta, o RFM pode ser combinado com técnicas de aprendizagem automática supervisionada para fins preditivos. Por exemplo, é comum aplicar modelos de regressão para estimar o valor futuro de um cliente, ou classificadores como *Random Forest* para identificar consumidores com maior probabilidade de churn (abandono da marca ou cessação da relação comercial) (Lemmens et al., 2006). Esta integração permite antecipar padrões de comportamento e adaptar as campanhas de marketing de forma mais eficiente e personalizada.

2.6.3 Large language models

Os LLMs representam uma das inovações mais significativas na área da inteligência artificial nos últimos anos. Treinados sobre grandes volumes de dados textuais, estes modelos são capazes de compreender e gerar linguagem natural de forma contextualizada, fluente e adaptada ao interlocutor.

No âmbito do *marketing* omnicanal, os LLMs abrem possibilidades revolucionárias na personalização da comunicação com os consumidores. A sua capacidade de gerar conteúdo relevante e ajustado permite criar mensagens personalizadas com base no perfil do cliente, no canal de interação e no momento da jornada de compra. Por exemplo, um LLM pode gerar descrições de produto adaptadas aos interesses de um utilizador específico ou criar emails promocionais cujo tom e vocabulário refletem as preferências linguísticas do destinatário.

Para além da geração de conteúdo, os LLMs podem ser aplicados em tarefas de análise de sentimento, extraindo emoções e opiniões expressas em *reviews*, redes sociais ou interações com o serviço de apoio ao cliente. Esta funcionalidade permite às empresas captar insights valiosos sobre a perceção da marca, monitorizar a satisfação do cliente e identificar áreas críticas de intervenção.

Quando integrados com outras técnicas de IA — como o *clustering* e os modelos RFM — os LLMs tornam-se parte de um ecossistema analítico mais amplo. Esta integração possibilita a identificação de segmentos comportamentais e a criação de comunicações altamente relevantes, ajustadas às características e motivações de cada grupo. Assim, os LLMs não só potenciam o *engagement*, como contribuem para o aumento das taxas de conversão e a fidelização dos clientes.

2.6.4 Análise preditiva e prescritiva

A análise preditiva tem vindo a assumir um papel central na transformação digital do retalho, permitindo às empresas antecipar comportamentos dos consumidores, adaptar estratégias em tempo real e otimizar os recursos de *marketing*. Baseada em dados históricos e em técnicas de aprendizagem automática, esta abordagem procura estimar a probabilidade de eventos futuros com elevado grau de precisão (Gao et al., 2021).

Entre os algoritmos frequentemente utilizados encontram-se a regressão logística, as árvores de decisão e os modelos baseados em *Random Forest*, que se destacam pela sua robustez e capacidade de lidar com variáveis complexas e interdependentes. Estes modelos são aplicados a diversos desafios do retalho alimentar, nomeadamente:

- Previsão da probabilidade de conversão em campanhas promocionais;
- Estimativa do valor futuro de um cliente (*Customer Lifetime Value*);
- Identificação de clientes com elevado risco de abandono (*churn*);
- Antecipação de categorias de produtos com maior probabilidade de compra;
- Análise de tendências sazonais e previsão de procura futura.

A análise prescritiva vai além da previsão, oferecendo recomendações específicas sobre as ações a tomar para otimizar resultados. Esta abordagem combina análise preditiva, regras de negócio e técnicas de otimização para sugerir a melhor estratégia em cada situação.

No contexto do *marketing* omnicanal, a análise prescritiva pode recomendar o canal mais adequado para contactar um determinado cliente, o momento ideal para enviar uma oferta, o desconto ótimo para maximizar a conversão sem comprometer a margem, ou a combinação de produtos a promover em conjunto.

A integração destas técnicas avançadas de IA no retalho alimentar permite criar um ecossistema de *marketing* inteligente, capaz de se adaptar dinamicamente às mudanças no comportamento do consumidor e otimizar continuamente as estratégias de comunicação e vendas.

2.7 Considerações finais

A convergência entre o *marketing* omnicanal e a inteligência artificial tem-se afirmado como um fator determinante para a otimização da experiência do consumidor e para o aumento do ROI. Estudos demonstram que consumidores omnicanal tendem a gastar mais, a apresentar níveis superiores de *engagement* e a manter uma relação mais duradoura com as marcas (Juaneda-Ayensa et al., 2016).

A literatura evidencia que a integração eficaz entre canais físicos e digitais é essencial para oferecer uma experiência de cliente fluida, coerente e personalizada. Diferenciando-se do

modelo multicanal — em que os canais funcionam de forma isolada —, o modelo omnicanal promove uma visão centrada no cliente, permitindo-lhe transitar entre pontos de contacto sem fricções, mantendo a consistência nas interações e no conteúdo.

Neste contexto, a inteligência artificial emerge como um catalisador transformador, ao oferecer capacidades avançadas para processar grandes volumes de dados, identificar padrões de comportamento e personalizar interações em tempo real. Técnicas como *clustering*, análise RFM e modelos de LLMs são amplamente reconhecidas pela sua eficácia em segmentar consumidores, prever comportamentos de compra e gerar comunicações adaptadas ao perfil e contexto de cada cliente.

Apesar dos avanços tecnológicos, persistem desafios relevantes nomeadamente a fragmentação de dados entre sistemas, a ausência de visão unificada do cliente e a dificuldade em coordenar estratégias entre canais. A inteligência artificial, ao ser aplicada de forma integrada, pode mitigar estas limitações, fornecendo mecanismos de personalização dinâmica, análise preditiva e automação de decisões de *marketing*.

A adoção de uma abordagem orientada para o cliente, suportada por tecnologias emergentes, revela-se cada vez mais decisivo para o sucesso num mercado altamente competitivo. Organizações que conseguirem articular eficazmente estratégias omnicanal com soluções de inteligência artificial estarão mais bem posicionadas para satisfazer as expectativas crescentes dos consumidores e adaptar-se a um ambiente de negócios volátil e em constante transformação.

A literatura analisada aponta ainda para um futuro do retalho fortemente assente na personalização preditiva, suportada por algoritmos capazes de compreender e antecipar as necessidades individuais. Neste cenário, a integração de dados provenientes de múltiplas fontes, a extração de insights acionáveis e a capacidade de ajustar estratégias em tempo real constituirão diferenciais competitivos essenciais.

3 Metodologia de investigação

O presente capítulo descreve de forma abrangente a abordagem metodológica adotada na investigação e o percurso que sustentou o desenvolvimento da *framework* proposta. Inicialmente, é apresentado o enquadramento que justifica a seleção do modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*) como estrutura orientadora do projeto, evidenciando a sua adequação ao contexto de análise de dados e inteligência artificial aplicada ao *marketing* omnicanal.

Seguidamente, detalham-se as principais etapas que compuseram o trabalho empírico: a clarificação do problema de negócio, a definição dos objetivos, a criação e preparação de dados sintéticos, a aplicação de algoritmos de *clustering*, previsão e personalização de conteúdos, bem como os procedimentos de avaliação dos resultados obtidos. Por fim, são descritas as estratégias implementadas para garantir a replicabilidade da *framework* e a sua relevância prática em cenários empresariais.

3.1 Abordagem metodológica

A crescente digitalização e o recurso à inteligência artificial no *marketing* exige a adoção de metodologias que conciliem rigor científico, capacidade de adaptação e eficiência na exploração de dados. Nesta investigação, optou-se pela metodologia CRISP-DM pela sua reconhecida solidez e flexibilidade em projetos orientados por dados.

O CRISP-DM oferece uma estrutura sistemática que permite conduzir todas as fases do processo analítico de forma iterativa e controlada, desde a compreensão do problema e o planeamento do projeto, passando pela preparação dos dados, modelação e validação, até à implementação e monitorização contínua dos resultados. A escolha desta metodologia reflete a necessidade de uma abordagem que facilite não só a análise preditiva, mas também a integração eficaz das soluções desenvolvidas nas práticas de *marketing* omnicanal.

3.1.1 Justificação da escolha do CRISP-DM

O CRISP-DM foi a metodologia escolhida para esta investigação por se tratar de uma abordagem estruturada, robusta e amplamente reconhecida tanto em contexto académico como empresarial. Esta metodologia oferece um modelo replicável e flexível para transformar dados em conhecimento acionável, sendo particularmente adequada para contextos complexos como o *marketing* omnicanal no setor do retalho alimentar (Chapman Pete et al., 1999).

Uma das principais razões para a sua adoção prende-se com a sua abordagem sistemática e iterativa, que garante que todas as fases essenciais de um projeto de análise de dados são consideradas e executadas de forma ordenada. Esta estruturação é especialmente relevante num contexto em que a diversidade de fontes e a fragmentação de dados podem comprometer a integridade das análises. Além disso, o ciclo iterativo permite visitar fases anteriores sempre que surgem novos requisitos ou dados, promovendo a melhoria contínua dos processos (Wirth & Hipp, 2000)

Adicionalmente, o CRISP-DM destaca-se pela sua versatilidade: embora tenha sido originalmente concebido para projetos de *data mining*, a sua aplicabilidade foi progressivamente alargada a iniciativas de IA, *machine learning* e análise preditiva. Esta capacidade de adaptação torna-o particularmente pertinente para o presente estudo, que envolve diferentes técnicas de IA com foco na personalização e otimização de campanhas.

Outro fator diferenciador reside na orientação estratégica que o CRISP-DM proporciona. Desde as fases iniciais, a metodologia valoriza a compreensão aprofundada do problema de negócio e a sua tradução em objetivos analíticos claros, assegurando o alinhamento entre a componente técnica e os resultados operacionais pretendidos. No setor do retalho alimentar, onde o impacto se mede frequentemente em ganhos concretos, como o aumento das vendas ou a melhoria da experiência do cliente, esta orientação assume especial relevância.

Por fim, a natureza iterativa do CRISP-DM permite ajustamentos contínuos ao longo do projeto, fomentando o aperfeiçoamento progressivo dos modelos e das soluções desenvolvidas, mesmo perante alterações nas preferências dos consumidores ou novas exigências de mercado.

Para melhor contextualizar a estrutura e o funcionamento desta metodologia, apresenta-se de seguida um esquema ilustrativo do modelo CRISP-DM. Este diagrama sintetiza

visualmente as seis fases principais que compõem o processo, permitindo uma percepção imediata da lógica sequencial e iterativa que o caracteriza. Na secção seguinte, estas fases serão descritas em maior detalhe, evidenciando o seu contributo específico no âmbito desta investigação.



Figura 1 Esquema do modelo CRISP-DM e respetivas fases

3.1.2 Descrição das etapas do processo

O modelo CRISP-DM organiza-se em seis fases principais, cada uma com objetivos e outputs bem definidos. A seguir, descrevem-se estas fases no enquadramento desta investigação:

1. **Compreensão do problema de negócio**

Esta fase foca-se na definição clara dos objetivos da investigação do ponto de vista estratégico, bem como na formulação de problemas analíticos derivados desses objetivos. No caso presente, envolveu a identificação dos principais desafios do retalho alimentar na personalização de campanhas omnicanal e a definição de métricas como ROI, *engagement* e fidelização.

2. **Compreensão dos dados**

Devido à ausência de dados reais do setor, foi necessário recorrer à geração de

dados sintéticos realistas, baseados em padrões típicos do comportamento de consumidores no retalho alimentar. Esses padrões foram definidos com base em estudos recentes que descrevem a evolução do setor e identificam critérios como a diversidade crescente dos perfis de consumidores, a migração para canais digitais, a sensibilidade a promoções, a procura por experiências personalizadas e a fragmentação dos formatos de compra (Deloitte,2023).

3. **Preparação dos dados**

Esta fase engloba a transformação dos dados brutos num conjunto de dados estruturado e adequado para modelação. Inclui limpeza, integração e seleção de atributos relevantes. Neste estudo, envolveu a preparação dos *datasets* sintéticos para suportar a segmentação de clientes e a previsão de conversão.

4. **Modelação**

Nesta fase, aplicam-se algoritmos apropriados para responder às questões de negócio. O estudo utilizou *K-means* para *clustering* de consumidores, modelos RFM para análise de valor do cliente, *Random Forest* para previsão de conversão e LLMs para geração automatizada de mensagens e folhetos personalizados.

5. **Avaliação**

Os modelos desenvolvidos são avaliados quanto à sua eficácia e capacidade de satisfazer os objetivos definidos. Foram utilizadas métricas como o *Silhouette Score* para avaliar o *clustering*, matriz de confusão para analisar o desempenho da previsão e conversão, SHAP para interpretar o contributo de cada variável de previsão e indicadores de ROI estimado para avaliar a eficácia das campanhas simuladas, bem como análise qualitativa da relevância das mensagens personalizadas.

6. **Implementação**

A fase final consiste na disponibilização dos resultados de forma utilizável. Neste caso, traduziu-se na construção de uma *framework* funcional para campanhas omnicanal personalizadas e num relatório final interativo que agrega visualizações, segmentos e recomendações otimizadas.

Adicionalmente, foi incorporado um mecanismo de simulação inspirado no uso de sensores e dados em tempo real, visando representar o comportamento omnicanal dos consumidores e potencializar a personalização dinâmica das interações.

A aplicação do CRISP-DM nesta investigação permitiu estruturar todo o processo com rigor, promovendo uma abordagem replicável e escalável, adaptada às exigências do retalho alimentar e à realidade prática da aplicação de inteligência artificial neste setor.

3.2 Compreensão do problema de negócio

A primeira etapa da metodologia CRISP-DM é essencial para definir os objetivos estratégicos do estudo e estabelecer uma base sólida para as fases subsequentes. Nesta investigação, a compreensão do problema de negócio centra-se na identificação das necessidades específicas do setor do retalho alimentar em termos de personalização e otimização de campanhas omnicanal

3.2.1 Definição dos objetivos estratégicos

Os objetivos estratégicos desta investigação foram definidos com base numa análise aprofundada do panorama atual do retalho alimentar e das tendências emergentes no *marketing* omnicanal, nomeadamente a crescente importância da personalização, da integração de dados e da aplicação de inteligência artificial para a tomada de decisão baseada em evidência. (Grewal et al., 2017; Verhoef et al., 2015; Deloitte, 2023). Os principais objetivos são:

1. **Aumentar a personalização das campanhas omnicanal**

Desenvolver uma abordagem avançada de personalização que permita adaptar mensagens, ofertas e experiências ao perfil e comportamento de cada cliente. Esta personalização visa aumentar a relevância das comunicações, melhorar as taxas de conversão e reforçar a relação de proximidade e fidelização com o consumidor (Davenport et al., 2020)

2. **Superar a fragmentação de dados entre canais e plataformas**

Propor uma estratégia integrada que permita consolidar e harmonizar dados provenientes de múltiplos canais (loja física, *e-commerce*, aplicações móveis, email, entre outros), criando uma visão única e holística do cliente. Esta integração visa garantir uma experiência fluida, consistente e contextualizada ao longo de toda a jornada omnicanal(Juaneda-Ayensa et al., 2016).

3. **Maximizar a eficiência operacional e o ROI**

Otimizar a distribuição de recursos de *marketing* através de análises preditivas e segmentação inteligente, assegurando que os esforços são dirigidos para os canais e públicos com maior probabilidade de gerar retorno. Este objetivo procura também reduzir desperdícios, melhorar a eficiência operacional e reforçar a sustentabilidade financeira das ações de *marketing* (Deloitte, 2023).

4. **Adotar uma abordagem orientada por dados na tomada de decisões de *marketing***

Substituir práticas baseadas na intuição ou em regras genéricas por processos sistemáticos e suportados por dados. Este objetivo visa promover a recolha, análise e aplicação contínua de insights quantitativos e qualitativos para suportar decisões estratégicas mais informadas, coerentes e eficazes no planeamento e execução de campanhas(Davenport et al., 2020).

3.2.2 Identificação das necessidades de personalização

A personalização eficaz no contexto do *marketing* omnicanal exige uma compreensão aprofundada das preferências, comportamentos e expectativas dos consumidores. No retalho alimentar, onde as decisões de compra são fortemente influenciadas por fatores contextuais, emocionais e comportamentais, a personalização revela-se um elemento-chave para aumentar o *engagement* e a fidelização(Grewal et al., 2017; Deloitte, 2023). A literatura identifica diversas dimensões críticas de personalização, aplicáveis ao desenho de estratégias mais eficazes:

1. **Personalização de conteúdo**

Refere-se à adaptação das mensagens, imagens e ofertas de acordo com as preferências específicas dos consumidores. Esta personalização baseia-se em

dados como o histórico de compras, comportamentos de navegação, preferências declaradas e interações anteriores, permitindo comunicações mais relevantes e impactantes.

2. **Personalização de canal**

Diz respeito à escolha dos canais mais apropriados para contactar cada cliente, com base nas suas preferências e hábitos de utilização. A seleção entre email, SMS, aplicação móvel, redes sociais ou contacto em loja física pode influenciar significativamente a eficácia da comunicação.

3. **Personalização temporal**

Envolve a identificação dos momentos ideais para interagir com o cliente, considerando padrões históricos de comportamento, horários de maior atividade e ciclos de compra. Esta dimensão permite maximizar a probabilidade de resposta e conversão.

4. **Personalização contextual**

Implica a adaptação da mensagem ao contexto em que o cliente se encontra no momento da interação, como a localização geográfica, o dispositivo utilizado ou a etapa da jornada de compra. Esta abordagem torna a comunicação mais relevante e oportuna.

5. **Personalização da oferta**

Refere-se à criação de propostas de valor ajustadas ao perfil do consumidor, tendo em conta a sua sensibilidade ao preço, categorias preferidas e potencial de valor ao longo do tempo. Esta personalização visa maximizar o retorno da campanha e a perceção de valor da marca.

Para suportar estas estratégias de forma eficaz, algumas abordagens apontam para a utilização de sensores inteligentes no ponto de venda físico, como uma forma de recolher dados em tempo real sobre o comportamento do consumidor. A integração destes dados com os provenientes de canais digitais permite uma visão unificada e atualizada da jornada omnicanal, contribuindo para a eliminação da fragmentação informacional e para o desenho de experiências mais coesas e personalizadas (Gal Gitter, 2020).

3.2.3 Integração de dados entre canais

Um dos principais desafios identificados na literatura sobre *marketing* omnicanal diz respeito à integração de dados provenientes de múltiplos pontos de contacto com o consumidor (Barwitz & Maas, 2018). Esta integração é essencial para construir uma visão holística e contínua da jornada do cliente, permitindo uma personalização eficaz e decisões de *marketing* baseadas em dados. No contexto do retalho alimentar, esta necessidade torna-se ainda mais premente devido à diversidade de canais, físicos e digitais, utilizados pelos consumidores.

A implementação de sensores inteligentes no ambiente físico, aliada à recolha de dados digitais (como interações online, histórico de compras e comportamento em aplicações móveis), contribui para enriquecer a compreensão do comportamento do consumidor. No entanto, a integração eficiente destes dados enfrenta uma série de obstáculos técnicos e operacionais:

1. **Heterogeneidade dos dados**

Os dados recolhidos em diferentes canais, como loja física, *e-commerce*, redes sociais, app mobile ou email *marketing* — apresentam frequentemente formatos distintos, diferentes granularidades e estruturas variadas. Esta heterogeneidade dificulta a consolidação e análise conjunta das informações.

2. **Identificação unificada do cliente**

Associar corretamente os dados de um mesmo cliente entre canais distintos continua a ser um dos maiores desafios. A inexistência de identificadores comuns, como login único ou número de fidelização, pode comprometer a consistência da visão 360° do consumidor.

3. **Privacidade e conformidade legal**

A recolha, armazenamento e utilização de dados pessoais devem obedecer aos regulamentos de proteção de dados, como o RGPD. A conformidade exige a implementação de mecanismos de consentimento informado, anonimização de dados sensíveis e políticas robustas de segurança.

4. **Atualização em tempo real**

Para que a personalização seja eficaz, os dados devem ser atualizados

continuamente, permitindo interações relevantes no momento certo. Esta exigência implica a adoção de arquiteturas de dados em tempo real ou quase real, capazes de reagir de forma dinâmica às ações dos consumidores.

Face a estes desafios, a literatura aponta para a inteligência artificial como uma solução viável para orquestrar a integração e análise de dados omnicanal. Técnicas avançadas de processamento de dados, como pipelines automatizados, modelos preditivos e análise em tempo real, permitem criar sistemas inteligentes capazes de consolidar informações de fontes distintas e transformá-las em ações de *marketing* altamente personalizadas.

Neste enquadramento, a abordagem baseada em IA não só responde às limitações estruturais e técnicas da integração omnicanal, como também promove ganhos substanciais em termos de fidelização de clientes e o ROI. Ao adaptar as comunicações ao perfil e ao comportamento de cada cliente — com base em dados continuamente atualizados — torna-se possível desenvolver campanhas mais eficazes, reduzir a dispersão de recursos e reforçar a conexão emocional com a marca.

3.3 Compreensão e preparação dos dados

As etapas de compreensão dos dados e preparação dos dados no processo CRISP-DM são fundamentais para garantir a qualidade da modelação posterior. Neste estudo, dada a indisponibilidade de dados reais provenientes de empresas do setor do retalho alimentar, optou-se por uma abordagem controlada de geração de dados sintéticos. Esta decisão visou permitir a simulação de cenários realistas, mantendo a flexibilidade necessária para testar algoritmos de inteligência artificial e validar a *framework* proposta.

3.3.1 Geração de dados sintéticos

Os dados sintéticos foram gerados em *Python*, com base em referências da literatura e em práticas comuns observadas em estudos de comportamento de consumidores no setor do retalho. O processo seguiu uma lógica estruturada, incorporando variáveis-chave que refletem a realidade de ambientes omnicanal.

As principais categorias de informação simuladas incluíram:

a) Dados de comportamento de compra

- **Frequência de compras:** Número total de transações por cliente num determinado período, refletindo o grau de envolvimento com a marca.
- **Tempo entre compra:** intervalo médio de dias entre duas compras consecutivas do mesmo cliente, utilizado como proxy da *recency* e potencial comportamento de repetição.
- **Valor de interação:** montante associado à transação, assumindo como indicador do *engagement* e do impacto da campanha.
- **Segmento de canal utilizado:** canal através do qual a interação ocorreu (loja física, e-commerce, aplicação), permitindo observar padrões de preferência por canal.
- **Nome do produto e categoria:** informação que permite identificar padrões de preferência de consumo por tipo de produto.
- **Resposta a campanhas anteriores:** variável binária (sim/não) que indica se o cliente respondeu positivamente à campanha (conversão).

b) Dados demográficos e segmentação simulada

- **ID do cliente, gênero e faixa etária estimada:** atributos básicos simulados com o objetivo de enriquecer o perfil sociodemográfico de cada consumidor e possibilitar análises por grupo.
- **Segmentos de cliente:** os segmentos comportamentais como *Cliente Premium*, *Indeciso*, *Engajado Lento*, *Interação Baixa* foram obtidos através da aplicação do algoritmo K-Means, que agrupou os clientes com base em variáveis como valor da interação, tempo entre compras e frequência de resposta a campanhas. O processo de *clustering* é descrito detalhadamente no ponto 3.4.1, mas é importante referir que os perfis resultam de padrões identificados nos dados e não de categorias pré-definidas.
- **RFM scoring:** as métricas de *Recency*, *Frequency* e *Monetary value* foram simuladas com base em dados comportamentais, sendo posteriormente utilizadas para enriquecer a segmentação obtida por *clustering*, permitindo uma caracterização mais robusta dos diferentes perfis de clientes.

c) Dados simulados de interação digital

- **Interações por canal (loja física, online, app):** cada registo inclui o canal onde a interação ocorreu, simulando a utilização dos diferentes canais na jornada do cliente. Este dado foi representado como uma variável categórica nominal, sendo posteriormente transformado por codificação para uso em algoritmos de aprendizagem automática.
- **Categoria de produto clicado:** esta variável representa a categoria do produto visualizado ou clicado pelo utilizador em ambiente digital. Foi utilizada como indicador indireto *proxy* de preferência implícita, assumindo-se que as interações com determinadas categorias revelam interesse do cliente, mesmo sem conversão direta..
- **Histórico de navegação simplificado:** consiste na simulação de uma sequência reduzida de páginas ou produtos visitados, permitindo inferir padrões de comportamento e pontos de interesse ao longo da navegação. Embora não se tenha replicado um sistema de *tracking* em tempo real, a simplificação segue o princípio de representar etapas essenciais da jornada digital, como visualizações, cliques e adições ao carrinho.

d) Dados Financeiros e Promoções

- **Valor médio da compra e total acumulado por cliente:** variáveis simuladas com base na distribuição normal de gastos por transação, tendo em conta perfis distintos de consumo entre segmentos. Estas métricas foram utilizadas para estimar o potencial económico de cada cliente ao longo do tempo.
- **Sensibilidade simulada a descontos:** foi criada uma variável *proxy* para representar a propensão do cliente a responder positivamente a campanhas com desconto. Neste contexto, uma variável *proxy* corresponde a uma medida indireta utilizada para representar um comportamento que não está diretamente observável. Neste caso, a *proxy* foi inferida com base na frequência de resposta a campanhas promocionais anteriores (simuladas) e no montante da compra efetuada quando existia desconto. Assim, clientes que compraram com maior frequência ou com maior valor durante períodos promocionais foram classificados com maior sensibilidade ao desconto.

- **Produtos com maior gasto por categoria:** variável utilizada para identificar padrões de preferência de consumo. A distribuição dos gastos foi simulada de forma a refletir diferentes perfis de clientes — por exemplo, alguns segmentos mostraram maior alocação de valor em categorias como snacks ou iogurtes, apoiando a lógica de personalização futura das campanhas.

e) Dados simulados de sensores (exploratórios)

- Embora não tenham sido usados diretamente na modelação, foram consideradas possibilidades de integração futura com dados de sensores inteligentes (ex. presença em loja, RFID, etc.) como parte do racional teórico.

A geração dos dados foi cuidadosamente desenhada para garantir coerência estatística e validade prática, permitindo aplicar técnicas como *clustering K-means*, análise RFM, *Random Forest* e geração de conteúdo com LLMs. Além disso, a estrutura dos dados simulados possibilitou a criação de outputs visuais (ex: *cluster plots*, *matriz de confusão*, *folhetos personalizados*) representativos de um sistema real de *marketing omnicanal*.

A preparação posterior dos dados incluiu:

- **Limpeza** e eliminação de outliers ou inconsistências.
- **Conversão de variáveis categóricas em representações numéricas** (*one-hot encoding*).
- **Normalização de variáveis contínuas** para otimizar o desempenho dos algoritmos.
- **Divisão em conjuntos de treino e teste**, especialmente no modelo de previsão de conversão.

Esta abordagem metodológica garantiu que, apesar da limitação inicial de dados reais, o estudo pudesse avançar com cenários realistas e controlados, mantendo a validade e relevância para o setor do retalho alimentar. A utilização de dados sintéticos também facilitou o processo de explicabilidade dos modelos, tornando possível avaliar com clareza o comportamento de cada componente da *framework* desenvolvida.

3.3.2 Variáveis relevantes para o estudo

A seleção das variáveis incluídas no conjunto de dados sintéticos foi orientada por objetivos práticos, alinhados com os modelos de inteligência artificial utilizados na *framework*. As variáveis escolhidas refletem a intenção de capturar os elementos essenciais do comportamento do consumidor no retalho alimentar omnicanal, com base em interações, produtos, canais e conversões. Foram consideradas as seguintes categorias:

1. Variáveis comportamentais

- **Tipo de interação:** Cada registo corresponde a uma interação com a campanha (visualização, clique, compra, etc.), permitindo distinguir níveis de envolvimento.
- **Dias até à conversão:** Número de dias entre o início da campanha e a data da interação, utilizado como proxy da *recency*.
- **Valor da interação:** Montante associado a uma compra, usado para estimar o impacto económico da campanha.
- **Canal de interação:** Canal utilizado pelo cliente (email, loja, push, SMS, etc.), relevante para análise de desempenho omnicanal.
- **Dispositivo utilizado:** Informação sobre se a interação foi feita via desktop, mobile ou tablet.

2. Variáveis de segmentação

- **Cluster de comportamento:** Resultado da aplicação do algoritmo K-Means, agrupando os clientes com base no valor das interações e no tempo até à conversão.
- **Segmento inferido:** Nome atribuído a cada *cluster* (ex.: *Cliente_Premium*, *Indeciso*, *Engajado_Lento*), facilitando a leitura estratégica.

3. Variáveis de produto e campanha

- **Categoria do produto:** Ex.: snacks, iogurtes, cereais — útil para entender preferências e gerar conteúdo personalizado.
- **Tipo de campanha:** Identificação do objetivo da campanha (ex.: lançamento, fidelização, desconto), utilizada na geração de mensagens com LLMs.

- **Mensagem padrão:** Texto base da campanha antes da personalização.
- **Resposta LLM:** Mensagem personalizada gerada automaticamente com base no segmento, canal e tipo de campanha.

4. Variáveis de perfil do cliente

- **Localização:** Região simulada de residência do cliente (Lisboa, Porto, etc.), permitindo observações geográficas.
- **Género e idade:** Derivado da data de nascimento, para complementar a caracterização sociodemográfica.

Estas variáveis, embora sintéticas, permitiram construir um dataset robusto, suficientemente representativo para testar e validar os modelos de segmentação, predição e personalização incluídos na *framework*. A sua articulação garantiu coerência analítica entre os diferentes módulos do sistema, simulando com realismo uma operação de *marketing* orientada por dados no contexto do retalho alimentar.

3.3.3 Processo de tratamento e transformação de dados

Após a geração dos dados sintéticos, foi realizado um processo rigoroso de tratamento e transformação, essencial para assegurar a qualidade, consistência e adequação dos dados às técnicas de análise e modelação utilizadas nesta investigação. Este processo envolveu diversas etapas estruturadas, organizadas em três grandes áreas: limpeza, transformação e engenharia de características.

1. Limpeza dos dados

Com o objetivo de eliminar ruído e inconsistências, foram executadas várias operações de limpeza:

- **Remoção de duplicados:** Eliminaram-se registos repetidos que poderiam distorcer as análises e enviesar os modelos.
- **Tratamento de valores ausentes:** Aplicou-se imputação por medidas estatísticas (como média ou mediana), ou por inferência baseada em padrões do perfil do cliente.

- **Verificação de consistência lógica:** Garantiu-se a coerência entre variáveis (ex.: valores de interação não negativos, compatibilidade entre canal e tipo de interação).
- **Observação de valores extremos:** Embora não tenha sido implementada uma técnica formal como o IQR, os dados foram mantidos realistas para refletir variações típicas do comportamento do consumidor.

2. Transformação dos dados

Para tornar os dados adequados à aplicação de algoritmos de inteligência artificial, foram realizadas transformações estruturais e numéricas:

- **Escalonamento e normalização:** As variáveis contínuas, como o valor de interação e os dias até à compra, foram normalizadas com *StandardScaler* para garantir comparabilidade e melhorar a performance dos algoritmos.
- **Codificação de variáveis categóricas:** Categorias como canal, tipo de campanha e dispositivo foram convertidas em representações numéricas através de *One-Hot Encoding*. Esta técnica foi escolhida por não assumir qualquer relação de ordem entre as categorias.
- **Estruturação temporal:** Foram calculados indicadores como os dias decorridos desde o início da campanha até à conversão, permitindo a análise da velocidade de resposta.

3. Engenharia de características

Com o objetivo de enriquecer o dataset e aumentar o poder explicativo dos modelos, foram criadas variáveis a partir das originais:

- **Variáveis derivadas:** Como a variável binária de conversão, baseada no tipo de interação, e o tempo até à compra.
- **Indicadores comportamentais:** Métricas como a frequência de interação, canais utilizados e dispositivos preferenciais foram combinadas para revelar padrões de comportamento.
- **Enriquecimento com segmentação inferida:** Após o *clustering*, cada cliente foi rotulado com um segmento comportamental, permitindo análises mais direcionadas.

Este processo de preparação dos dados garantiu a criação de uma base sólida para a aplicação dos modelos de *clustering*, previsão e personalização. A qualidade e profundidade das variáveis resultantes foram cruciais para a geração de *insights* relevantes e acionáveis no contexto do retalho alimentar omnicanal.

3.3.4 Criação de variáveis derivadas

A criação de variáveis derivadas constitui uma etapa essencial no processo de preparação dos dados, permitindo transformar informações brutas em indicadores analíticos mais ricos e relevantes para a segmentação. No contexto desta investigação, estas variáveis ajudaram a capturar comportamentos relevantes de forma mais estruturada e alimentar os modelos de IA com atributos significativos.

As principais variáveis derivadas foram organizadas em quatro categorias:

1. Segmentação de clientes

- Identificação de padrões de frequência e valor das interações para distinguir clientes esporádicos de clientes com maior valor agregado.
- Classificação de clientes por *cluster* com base em variáveis comportamentais como dias até à compra e valor da interação.
- Associação de perfis comportamentais descritivos a cada segmento, como "Cliente_Premium" ou "Indeciso".

2. Indicadores de *engagement*

- Cálculo do número e tipo de interações com campanhas (ex: visualização, clique, compra).
- Análise dos canais preferenciais e dispositivos utilizados, refletindo o envolvimento com diferentes pontos de contacto.

3. Métricas de campanha

- Taxa de conversão por canal e por segmento, facilitando a avaliação do desempenho relativo de diferentes estratégias.
- Estimativa do ROI e do custo médio por conversão, com base na relação entre o valor da interação e o custo da campanha.

4. Indicadores preditivos

- Variável binária de conversão, utilizada como *target* nos modelos de previsão.
- Probabilidade de conversão inferida por modelos como *Random Forest*, com base em variáveis como tempo até à compra, canal e dispositivo.

A construção destas variáveis contribuiu de forma decisiva para a eficácia analítica da *framework*, fornecendo uma base sólida para a análise preditiva e a personalização de conteúdo com LLMs. Esta abordagem permitiu captar nuances comportamentais e melhorar a precisão das decisões orientadas por dados no contexto do retalho omnicanal.

3.4 Modelação

Nesta etapa, foram implementadas técnicas de IA com o objetivo de transformar dados em conhecimento acionável, permitindo a segmentação precisa de consumidores, a personalização de comunicações e a previsão de resultados de campanhas. A modelação constitui o núcleo técnico desta investigação, sendo composta por quatro componentes principais: *clustering*, análise RFM, previsão com modelos preditivos e personalização com LLMs.

3.4.1 Implementação de algoritmos de *clustering*

A segmentação de consumidores foi realizada através do algoritmo *K-Means*, amplamente reconhecido pela sua simplicidade e eficiência na identificação de padrões comportamentais. O *clustering* foi aplicado com base em variáveis comportamentais extraídas do dataset, incluindo:

- Frequência de compra
- Recência
- Diversidade de categorias de produtos
- Taxa de conversão e *engagement*

O número ótimo de *clusters* foi definido com base no *Silhouette Score*, que permitiu validar a qualidade da segmentação. Cada segmento foi posteriormente interpretado e rotulado de forma descritiva, originando perfis como:

- Cliente_Premium

- Indeciso
- Engajado_Lento
- Interacao_Baixa

Esta segmentação foi utilizada como base para personalizar as estratégias de comunicação e prever padrões de comportamento.

3.4.2 Análise de valor com modelo RFM

A metodologia RFM (*Recency, Frequency, Monetary*) foi aplicada com o objetivo de classificar os clientes de acordo com a sua relevância estratégica, complementando a segmentação comportamental obtida através do *clustering*.

O modelo RFM baseia-se em três dimensões fundamentais:

- **Recency (Recência):** número de dias desde a última compra do cliente, permitindo estimar a probabilidade de nova interação;
- **Frequency (Frequência):** número total de compras realizadas num período definido, refletindo o grau de envolvimento com a marca;
- **Monetary (Valor Monetário):** valor total gasto pelo cliente, utilizado como indicador do seu contributo económico.

Cada uma destas métricas foi escalada numa escala ordinal (por exemplo, de 1 a 5), refletindo a posição relativa do cliente dentro da base analisada. A combinação dos scores permitiu construir perfis distintos com base no comportamento de compra e no valor potencial, frequentemente utilizados para orientar estratégias de retenção, fidelização e reativação de clientes.

Com base nos scores RFM, foram identificados segmentos como:

- **Clientes VIP:** alta recência, alta frequência e alto valor monetário;
- **Clientes Leais:** frequência elevada mas com menor valor ou recência;
- **Clientes Ocasionais:** compras esporádicas e de baixo valor;
- **Clientes Inativos:** longos períodos sem interação, com pouca frequência e baixo valor.

A aplicação do modelo RFM nesta investigação permitiu reforçar a análise comportamental, cruzando a dimensão do valor económico com a intensidade da relação do cliente com a marca. A simplicidade e interpretabilidade do modelo tornam-no particularmente adequado para contextos empresariais que requerem decisões rápidas e orientadas por dados objetivos.

3.4.3 Previsão de conversão com *Random Forest*

Para estimar a probabilidade de conversão dos clientes, foi treinado um modelo *Random Forest*, utilizando como variáveis preditoras os comportamentos de compra, interações passadas e características do segmento. O modelo permitiu:

- Identificar os clientes com maior probabilidade de resposta positiva a campanhas
- Priorizar alvos para ações de *marketing*
- Ajustar as ofertas em função do perfil preditivo de cada cliente

Esta componente preditiva foi avaliada com métricas como *confusion matrix*, precisão e recall, assegurando a robustez do modelo.

3.4.4 Geração de mensagens personalizadas com LLMs

Para a geração de conteúdo personalizado, foram utilizados LLMs com capacidade para adaptar o tom, estilo e conteúdo de mensagens aos diferentes perfis de clientes. Com base no output dos segmentos e das previsões, os modelos geraram:

- Mensagens promocionais personalizadas
- Folhetos com produtos recomendados (incluindo imagens)
- Textos adaptados ao canal de entrega e momento da jornada

A personalização incluiu o nome do cliente, preferências históricas e linguagem adaptada ao segmento, promovendo maior *engagement* e relevância.

3.4.5 Integração e automatização dos componentes

Todos os modelos foram integrados num pipeline coeso, permitindo uma execução automatizada das etapas, desde o carregamento dos dados até à geração final dos folhetos e recomendações. Esta integração foi suportada por:

- Estrutura modular em *Python*
- Scripts encadeados para processamento, modelação e geração de outputs
- Organização dos dados em diretórios lógicos (raw, processed, reports)

Este pipeline assegura a repetibilidade, escalabilidade e adaptação contínua, constituindo a base operativa da *framework*.

3.5 Avaliação

A avaliação dos modelos desenvolvidos nesta investigação foi realizada através de uma combinação de métricas quantitativas, indicadores económicos e análises qualitativas. Esta fase teve como objetivo garantir que as soluções propostas geram valor tangível, apresentam robustez técnica e são aplicáveis em contextos reais do retalho alimentar. A abordagem de avaliação adotada permite analisar, de forma integrada, a eficácia dos modelos de segmentação, a capacidade preditiva dos algoritmos e o impacto económico e comportamental das campanhas personalizadas, assegurando coerência com os objetivos estratégicos definidos.

3.5.1 Métricas quantitativas

Do ponto de vista quantitativo, a segmentação obtida através do algoritmo *K-Means* foi avaliada com múltiplas métricas de validação interna. A principal métrica utilizada foi o *Silhouette Score*, que permite medir o grau de coesão e separação entre os *clusters*. Com valores entre -1 e 1, este score fornece uma indicação clara sobre a qualidade da segmentação: quanto mais próximo de 1, maior é a separação entre os grupos, o que indica uma segmentação eficaz e bem definida. A sua inclusão permitiu avaliar a consistência interna dos agrupamentos obtidos a partir de variáveis comportamentais como o número de dias até à compra e o valor médio de interação.

Para complementar esta análise, foram integradas métricas adicionais como o *Davies-Bouldin Index*, que avalia a similaridade entre os *clusters* penalizando sobreposições; o *Calinski-Harabasz Index*, que mede a proporção entre a dispersão intra e inter-*clusters* e indica a densidade relativa dos grupos; e a *Inertia*, correspondente à soma das distâncias intra-*cluster*, representando a compacidade dos grupos formados. Estas métricas permitiram validar a estrutura dos segmentos de forma mais robusta e objetiva, fornecendo uma base sólida para o desenvolvimento de estratégias de personalização baseadas em *clusters*.

No que diz respeito à aplicação da metodologia RFM, a análise centrou-se na distribuição dos clientes pelos segmentos definidos – nomeadamente *Cliente_Premium*, *Indeciso*, *Engajado_Lento* e *Interação_Baixa* – e na observação da estabilidade desses segmentos ao longo do tempo. Para além disso, foi avaliada a capacidade preditiva dos *scores* RFM relativamente ao comportamento futuro de conversão e de *engagement*, servindo como complemento analítico ao clustering comportamental.

3.5.2 Análise de conversão, retenção e ROI simulado

A eficácia das campanhas personalizadas foi avaliada com base em métricas simuladas de taxa de conversão, retenção e ROI estimado, associadas aos diferentes segmentos de clientes e canais de comunicação utilizados. Estas métricas permitiram estimar o impacto das estratégias de personalização não apenas ao nível comportamental, mas também económico. Foram calculados indicadores como a taxa de conversão por *cluster* e canal, o CLV (*Customer Lifetime Value*) estimado, o CAC (*Customer acquisition cost*) e a taxa de retenção após campanhas personalizadas. O ROI simulado foi apurado com base na receita incremental potencial gerada pelas campanhas segmentadas em comparação com campanhas tradicionais não personalizadas, tendo em conta os custos de execução, personalização e comunicação.

3.5.3 Avaliação Qualitativa da aplicabilidade

A avaliação qualitativa centrou-se na análise da aplicabilidade prática dos resultados obtidos pela *framework*. Foram examinados os segmentos comportamentais gerados, avaliando a sua relevância e coerência com perfis reais de clientes do retalho alimentar. A pertinência das mensagens produzidas com modelos de LLMs, como o GPT-4, foi

igualmente analisada, assegurando a sua adequação ao perfil do cliente, canal utilizado e tipo de campanha.

Adicionalmente, foi considerada a viabilidade de adoção empresarial das estratégias propostas, nomeadamente no que diz respeito à integração de dados entre canais e à capacidade de personalização dinâmica. A geração automática de folhetos personalizados, bem como a apresentação de recomendações baseadas em ROI e taxa de conversão por segmento e canal, reforçaram o potencial de aplicação em contextos empresariais reais.

Esta abordagem de avaliação multidimensional, quantitativa (com métricas de *clustering* e predição), económica (com análise de ROI e custo por conversão) e qualitativa, permitiu validar a utilidade e robustez da *framework*, posicionando-a como uma solução prática, escalável e alinhada com as exigências do *marketing* omnicanal contemporâneo.

3.6 Implementação

A fase final da metodologia CRISP-DM corresponde à implementação prática dos modelos e insights desenvolvidos ao longo do projeto. No âmbito desta investigação, esta etapa concretiza-se através da construção de uma *framework* funcional e documentada, concebida para ser aplicada de forma eficaz e replicável em ambientes empresariais reais, com particular enfoque no setor do retalho alimentar.

3.6.1 Descrição do processo de implementação

A *framework* foi concebida de forma modular e escalável, permitindo a integração fluida entre as diferentes componentes do sistema: desde a geração e preparação de dados, passando pela segmentação e previsão de comportamento, até à geração de mensagens personalizadas com modelos de linguagem e à avaliação do desempenho das campanhas simuladas. Todo o processo foi cuidadosamente estruturado para refletir uma arquitetura lógica e extensível.

A primeira etapa consistiu na definição da arquitetura geral da *framework*. Foram especificados os módulos principais, geração e integração de dados, engenharia de variáveis, algoritmos de *clustering* e análise RFM, modelos preditivos, personalização com LLMs e mecanismos de avaliação, bem como os respetivos fluxos de dados e interfaces.

Esta organização garantiu uma base coesa e facilitou futuras adaptações a diferentes domínios ou conjuntos de dados.

Posteriormente, cada componente foi implementado individualmente. Foram desenvolvidos *scripts* em *Python* para automatizar a geração e normalização dos dados, aplicar técnicas de aprendizagem automática (como K-Means, *Random Forest* e SHAP), e gerar campanhas personalizadas através de LLMs como o GPT-4. Foram também integradas visualizações automatizadas, como gráficos de *clusters* e diagramas SHAP para facilitar a análise e interpretação dos resultados.

A integração global da *framework* foi assegurada através da construção de pipelines sequenciais reutilizáveis, com diretórios bem definidos para dados de entrada, processamento e saída. A fiabilidade do sistema foi garantida através de testes funcionais a cada módulo individual, complementados com testes integrados de ponta a ponta. Foram ainda resolvidos desafios técnicos relacionados com codificação de caracteres (encoding), tratamento de valores nulos e harmonização de formatos entre os diferentes módulos.

3.6.2 Documentação para replicabilidade

A *framework* foi desenvolvida com uma forte ênfase na replicabilidade e aplicabilidade prática. Para garantir que outros utilizadores pudessem compreender e executar todos os componentes do sistema, foi criado um ficheiro estruturado com instruções passo-a-passo, incluindo:

- Descrição técnica da arquitetura da *framework*;
- Requisitos de sistema e instalação de dependências;
- Código comentado, modular e reutilizável;
- Exemplo de execução completo via *main.py*, que automatiza todas as etapas da *framework*;
- Estrutura de pastas clara para facilitar a navegação pelos dados, imagens, relatórios e *scripts*.

Esta abordagem visa permitir que equipas técnicas e de *marketing* consigam adaptar facilmente a solução a diferentes realidades empresariais, mesmo sem conhecimento profundo em ciência de dados.

Ao integrar técnicas de inteligência artificial como *clustering*, previsão de conversão, explicabilidade com SHAP e personalização com LLMs, a *framework* apresenta-se como uma ferramenta prática e estratégica para o setor do retalho alimentar, com foco na melhoria do *engagement*, aumento da conversão e otimização do ROI.

3.6.3 Ambiente de desenvolvimento e configuração técnica

Os testes foram desenvolvidos em Python 3.12.10 e executados num computador com processador AMD Ryzen 5 5600H (3.30 GHz, 6 cores / 12 threads), 20 GB de memória RAM e GPU NVIDIA GeForce RTX 3050 Laptop GPU (4 GB), com sistema operativo Windows 11 Home (versão 24H2, compilação 26100.5074).

No que respeita ao software, recorreu-se a bibliotecas amplamente utilizadas na comunidade científica e adequadas à natureza do trabalho desenvolvido. Foram utilizadas *pandas* (2.2.3) e *numpy* (2.1.3) para a manipulação e tratamento de dados; *scikit-learn* (1.6.1) para a implementação de algoritmos de *clustering*, classificação e cálculo de métricas de avaliação; *matplotlib* (3.10.3) e *seaborn* (0.13.2) para a visualização gráfica dos resultados; *shap* (0.47.2) para a explicabilidade dos modelos de machine learning; e *streamlit* (1.45.1) para a construção de uma interface interativa destinada à apresentação dos resultados.

A descrição detalhada do ambiente técnico permite justificar os tempos de execução obtidos e assegurar a replicabilidade do estudo, garantindo que futuros investigadores possam reproduzir os resultados em condições semelhantes.

4 A *framework* proposta

Esta secção descreve, de forma detalhada, a *framework* desenvolvida com o objetivo de aplicar IA à personalização e otimização de campanhas omnicanal no retalho alimentar. A estrutura foi concebida para refletir os principais desafios do marketing moderno, combinando técnicas de segmentação, previsão comportamental e geração de conteúdo

com base em dados. Cada componente foi implementado de forma modular, promovendo a clareza, reprodutibilidade e escalabilidade da solução. A *framework* representa uma abordagem integrada que alia rigor técnico à aplicabilidade prática, permitindo transformar dados dispersos em decisões estratégicas orientadas para o cliente.

4.1 Visão geral da *framework*

A *framework* desenvolvida nesta investigação propõe-se a responder aos desafios contemporâneos do *marketing* omnicanal no setor do retalho alimentar, através da integração de técnicas de inteligência artificial para segmentação, previsão de comportamento e personalização de campanhas. Estruturada de forma modular, esta solução foi concebida para ser documentada, replicável e adaptável a diferentes contextos organizacionais, conciliando rigor analítico com aplicabilidade prática.

A lógica central da *framework* assenta na orquestração de três componentes fundamentais: (i) análise comportamental e segmentação de clientes, (ii) previsão de conversão e explicação dos fatores preditivos, e (iii) geração de comunicações personalizadas com base nos perfis identificados. Esta estrutura permite não apenas compreender o comportamento dos consumidores, mas também atuar sobre ele de forma informada, em tempo útil e com elevada precisão.

A abordagem adotada segue os princípios da metodologia CRISP-DM, o que assegura um processo estruturado desde a definição do problema de negócio até à implementação de uma solução funcional. A *framework* foi implementada integralmente em *Python* e organizada numa arquitetura clara, com scripts autónomos para cada etapa do processo, desde a geração de dados sintéticos até à criação de relatórios finais. Esta separação de responsabilidades garante escalabilidade, facilidade de manutenção e reprodutibilidade.

4.1.1 Arquitetura e componentes principais

A arquitetura da *framework* proposta é modular, escalável e desenhada para ser facilmente replicável. É composta por seis componentes principais que operam de forma integrada, transformando dados brutos em campanhas personalizadas orientadas por inteligência artificial. Esta estrutura assegura flexibilidade, reutilização e evolução contínua. Os componentes são os seguintes:

1. Módulo de integração de dados

Este módulo é responsável pela geração, transformação e integração de dados provenientes de diferentes fontes simuladas, como transações, perfis de clientes, campanhas e interações omnicanal (loja, app, website). Os dados são gerados artificialmente com realismo e organizados em pastas separadas (raw, processed) através de scripts personalizados. Esta etapa inclui ainda o tratamento e a engenharia de variáveis compostas, como *Dias_ate_Compra*, *Valor_Interacao*, variáveis demográficas e comportamentais.

2. Módulo de segmentação de clientes

A segmentação é feita com base em algoritmos de *clustering* (K-Means) e análise RFM, identificando perfis como *Cliente_Premium*, *Indeciso*, *Engajado_Lento* e *Interação_Baixa*. A qualidade da segmentação é avaliada com métricas quantitativas (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz e Inertia), garantindo coesão e distinção entre os grupos. A segmentação serve de base para todas as estratégias posteriores de personalização.

3. Módulo de modelação preditiva

Este módulo aplica algoritmos como *Random Forest* para prever a probabilidade de conversão de cada cliente com base em variáveis comportamentais e de contexto. O desempenho do modelo é avaliado por validação cruzada (F1-score) e por matriz de confusão. Para garantir transparência, são aplicadas técnicas de interpretabilidade com SHAP, revelando as variáveis mais determinantes para o comportamento de conversão em cada *cluster*.

4. Módulo de geração de conteúdo personalizado

Com base nos segmentos identificados e nas características de cada campanha, este módulo utiliza LLMs para gerar mensagens dinâmicas personalizadas por canal (email, push, SMS). Os *prompts* são adaptados a cada cenário e o conteúdo gerado é orientado para maximizar o *engagement*. Além disso, é também simulada a criação de folhetos personalizados, ajustados ao perfil de compras e interesses dos clientes.

5. Módulo de otimização de campanhas

A eficácia das campanhas é simulada com base em métricas como ROI estimado, custo médio por conversão, taxa de retenção e CLV. O sistema calcula e recomenda os melhores

canais por segmento, baseando-se na combinação entre taxa de conversão e eficiência económica. Esta análise permite ajustar dinamicamente os recursos de *marketing*, maximizando o impacto.

6. Módulo de visualização e relatórios

Todos os resultados são apresentados de forma acessível através de gráficos e *dashboards*. São geradas automaticamente visualizações de *clusters*, *waterfall* SHAP, matriz de confusão, bem como um relatório final em HTML (`reports/relatorio_final.html`) que resume toda a jornada analítica da *framework*, tornando-a compreensível para equipas não técnicas.

4.1.2 Fluxo de processamento de dados

O fluxo de processamento de dados da *framework* foi desenhado de acordo com os princípios da metodologia CRISP-DM, garantindo uma abordagem sistemática, iterativa e orientada à geração de valor. Cada etapa é interligada, permitindo que os dados fluam de forma coerente desde a sua origem até à geração de campanhas personalizadas. O processo compreende as seguintes fases:

1. Geração e integração de dados

Primeiramente, procede-se à geração e integração de dados, simulados a partir de múltiplas fontes típicas do retalho alimentar, como plataformas de *e-commerce*, apps, interações em loja física e dados de campanhas. Estes dados são armazenados num repositório estruturado, permitindo a consolidação e criação de uma visão unificada do comportamento do cliente.

Importante referir que, apesar de os dados utilizados serem sintéticos, a sua geração seguiu critérios estruturados com base em padrões realistas observáveis no retalho alimentar. A maior parte das variáveis categóricas (como género, canal, tipo de interação e dispositivo) foi gerada com distribuição uniforme ou por amostragem ponderada, simulando diferentes níveis de frequência esperados. As variáveis contínuas, como o preço do produto e o valor de interação, foram geradas com distribuições uniformes entre intervalos realistas (por exemplo, preços entre 1 e 10 euros). No caso da data de nascimento, foi utilizado um intervalo temporal aleatório entre 1955 e 2005, simulando perfis de clientes de diferentes idades. A variável `Dias_ate_Compra`, apenas aplicável em

casos de conversão, corresponde ao número de dias entre a data de início da campanha e a data de interação, variando entre 0 e 10 dias, refletindo janelas curtas de decisão comuns em campanhas promocionais. Este processo garante coerência interna e variação suficiente para testar algoritmos de segmentação, previsão e personalização com realismo suficiente para fins de validação.

2. Pré-processamento e transformação de dados

Segue-se o pré-processamento e transformação, onde os dados são limpos, normalizados e preparados para análise. Esta fase inclui a criação de variáveis derivadas, como *dias até à compra* e *valor de interação*, que enriquecem os modelos e aumentam a capacidade de extração de insights relevantes.

3. Segmentação e análise comportamental

Na fase de segmentação e análise, são aplicados algoritmos de *clustering* (nomeadamente K-Means) e análise RFM para identificar grupos de clientes com padrões de comportamento distintos. Estes segmentos – como *Cliente_Premium*, *Indeciso*, *Engajado_Lento* e *Interação_Baixa* – são caracterizados com base em variáveis como frequência, valor e tempo de resposta, constituindo a base para uma personalização eficaz.

4. Modelação preditiva

Posteriormente, a *framework* realiza modelação preditiva com recurso a *Random Forest*, estimando a probabilidade de conversão e antecipando comportamentos futuros, como risco de abandono e potencial de compra. Esta capacidade de previsão permite uma melhor orientação das estratégias de *marketing*.

5. Geração de conteúdos personalizados

Com base na segmentação e nas previsões obtidas, o sistema passa à personalização de conteúdo, utilizando modelos de linguagem LLMs para gerar mensagens específicas para cada tipo de cliente, canal e tipo de campanha. Esta fase inclui também a simulação de folhetos personalizados, adaptados ao histórico e preferências de compra.

6. Otimização de campanhas

Na etapa de otimização e alocação de recursos, a *framework* calcula métricas como ROI, taxa de conversão e custo por aquisição, ajustando automaticamente os canais e

orçamentos alocados a cada segmento, com o objetivo de maximizar o impacto económico das campanhas.

7. Monitorização e feedback contínuo

Por fim, a monitorização e *feedback* contínuo asseguram que os resultados são avaliados e reintroduzidos no sistema. Este ciclo permite refinar os modelos, adaptar a estratégia com base em novas informações e manter a eficácia das campanhas ao longo do tempo.

Este fluxo *end-to-end* é adaptável, modular e orientado a dados, suportando decisões de marketing mais inteligentes, eficazes e centradas no consumidor.

4.2 Integração de dados

A integração de dados foi uma etapa fundamental no desenvolvimento da *framework*, assegurando que a análise e os modelos fossem alimentados por informação estruturada e coerente. Embora o projeto tenha utilizado dados sintéticos gerados localmente, foram simulados cenários típicos de integração de dados provenientes de várias fontes, tal como acontece no retalho alimentar.

4.2.1 Métodos para integração de dados dispersos

A *framework* não integrou dados reais de sistemas externos, mas reproduziu de forma realista a consolidação de dados de múltiplas fontes simuladas, como interações de clientes, campanhas de *marketing*, produtos, e perfis demográficos. Para isso, foi adotada uma abordagem programática baseada em scripts modulares em *Python*, que agregam e relacionam os *datasets* de forma automática.

O processo de integração consistiu em:

- **Fusão programática entre tabelas simuladas:**

Através de *merges* entre *datasets* como *clientes.csv*, *campanhas.csv*, *produtos.csv* e *interacoes_cliente_campanha.csv*, foi construída uma visão unificada do comportamento de compra e das interações com campanhas, enriquecida com atributos de contexto como canal de interação e tipo de dispositivo.

- **Exemplo prático de integração:**

Considere o cliente A, cuja informação demográfica está contida no ficheiro clientes.csv (ex.: ID, género, idade, localização – 4 colunas). Ao participar numa campanha de email registada em interacoes_cliente_campanha.csv (ex.: tipo de interação, canal, dispositivo, valor da interação – 6 colunas), o sistema integra ambos os ficheiros através do ID_Cliente, resultando numa base com 9 colunas que combinam atributos pessoais com comportamento digital. Este tipo de integração permite, por exemplo, observar que o Cliente A interagiu via canal "Email" com a campanha M3, o que gerou um valor de interação de 8.99€ e ocorreu num dispositivo móvel. Ao cruzar com produtos.csv, é possível enriquecer ainda mais a linha com a categoria do produto promovido, simulando o tipo de visão que seria obtida num sistema de Customer Data Platform (CDP).

- **Engenharia de features integrada:**

Durante o processo de fusão, foram calculadas variáveis derivadas, como *dias até à compra* e variáveis categóricas como tipo de interação ou segmento atribuído, fundamentais para os modelos de clustering e predição.

- **Organização estruturada dos dados:**

A framework foi concebida com uma estrutura de pastas clara (data/raw, data/processed), o que assegura que os dados são armazenados de forma ordenada, facilitando o seu carregamento, transformação e reutilização ao longo dos diferentes módulos analíticos.

Apesar de não ter sido implementado um pipeline de ETL (*Extract, Transform, Load*) completo nem a utilização de APIs (*Application Programming Interface*), CDPs (*Customer Data Platform*) ou tecnologias de *streaming*, a simulação do processo de integração seguiu princípios equivalentes, modelando realisticamente o tipo de consolidação de dados necessária para uma operação omnicanal real.

Esta abordagem assegurou a consistência entre *datasets*, permitiu o rastreamento de cada cliente ao longo das suas interações e constituiu a base para a personalização e otimização conduzidas pela *framework*. A qualidade e coesão dos dados integrados foi essencial para a fiabilidade dos modelos de segmentação, previsão e geração de mensagens.

4.2.2 Definição das variáveis

As tabelas 1, 2 e 3 apresentam uma caracterização detalhada das variáveis que compõem o *dataset* sintético utilizado nesta investigação. Para cada variável, é indicada a sua tipologia, nomeadamente se é categórica nominal, categórica ordinal, numérica contínua, ou outra, bem como a respetiva escala de medida (nominal, ordinal, intervalar ou rácio), de acordo com as convenções estatísticas.

Com o objetivo de garantir clareza e facilitar a leitura, a tabela foi organizada em três partes complementares:

A tabela 1 indica o tipo e a escala de medição de cada variável, distinguindo variáveis categóricas de variáveis numéricas, e identificando ainda variáveis de apoio como identificadores e datas.

Tabela 1 Tipologia e escala de medição das variáveis do dataset

Variável	Tipo	Tipo e Escala de Medida
ID_Cliente	Identificador	Não aplicável
ID_Campanha	Identificador	Não aplicável
Data_Interacao	Data	Intervalar
Tipo_Interacao	Qualitativo	Nominal
Conversao	Qualitativo	Ordinal
Valor_Interacao	Quantitativo	Proporcional (contínua)
Canal_Interacao	Qualitativo	Nominal
Dispositivo_Interacao	Qualitativo	Nominal
Canal	Qualitativo	Nominal
Tipo_Campanha	Qualitativo	Nominal
Publico_Alvo	Qualitativo	Nominal

Mensagem_Padrao	Qualitativo	Nominal
Resposta_LLM	Qualitativo	Nominal
Nome_produto	Qualitativo	Nominal
Categoria_Produto	Qualitativo	Nominal
Preco_Produto	Quantitativo	Proporcional (contínua)
Unidade_Medida	Qualitativo	Nominal
Nome_Segmento_x	Qualitativo	Nominal
Localizacao	Qualitativo	Nominal
Genero	Qualitativo	Nominal
Dias_ate_Compra	Quantitativo	Intervalar (Contínua)
Cluster	Nominal	Categórico
Nome_Segmento_y	Qualitativo	Nominal
Descricao_Segmento	Qualitativo	Nominal

A tabela 2 agrupa as variáveis numéricas e ordinais, apresentando medidas estatísticas relevantes como a média, o desvio padrão, os valores mínimo e máximo.

Tabela 2 Estatística descritiva das variáveis numéricas e ordinais

Variável	Media	Desvio padrão	Mínimo	Máximo	Tipo
Conversao					Ordinal
Valor_Interacao	1.137	4.04	0.0	29.97	Contínua

Preco_Produto	4.627	2.422	1.13	8.98	Ordinal
Dias_ate_Compra	5.017	3.158	0.0	10.0	Ordinal
Cluster	1.515	1.206	0.0	3.0	Contínua

A Tabela 3 descreve as variáveis categóricas nominais, identificando a moda, o número de ocorrências da categoria mais frequente (frequência da moda), bem como os respectivos valores únicos observados. Quando o número de categorias distintas é elevado, apresenta-se uma indicação genérica (“Mais de 10”) para simplificar a leitura, mantendo-se o foco na diversidade da variável e na sua relevância para a análise.

Tabela 3 Estatística descritivas das variáveis categóricas nominais

Variável	Moda	Frequência moda	Valores únicos
Tipo_Interacao	Visualização	7945	Visualização, Ignore, clique, compra
Canal_Interacao	Email	5314	SMS, Email, Push, Loja, APP
Dispositivo_Interacao	Desktop	6727	Tablet, Mobile, Desktop
Canal	Email	5314	SMS, Email, Push, Loja, APP
Tipo_Campanha	Lancamento	6598	Fidelização, sazonal,
Publico_Alvo	Ocasional	9415	Ocasional, Vip, Leal, Inativo
Mensagem_Padrao	Aproveite desconto	7919	Mensagem personalizada(texto variável)
Resposta_LLM	Mensagem personalizada	3	Mais de 10
Nome_Produto	Produto_29	2741	Mais de 10

Categoria_Produto	Lacticionios	8046	Lacticionios, Cereais, Snacks, Bebidas
Unidade_Medida	Litro	13323	Litro, KG, unidade
Nome_segmento_X	Vip	6711	Vip, Leal, Ocasional, Inativo
Localização	Porto	4086	Porto, Faro, Coimbra, Lisboa
Genero	Masculino		Feminino, Masculino
Nome_Segmento_Y	Interacao_baixa	6872	Cliente_Premium, Indeciso, Interacao_Baixa, engajado_Lento
Descricao_Segmento	Pouco valor e poucos dias até a compra	6872	Valor alto e distribuição em vários tempos, Tempo alto, baixo valor, conversão difícil, pouco valor e poucos dias ate a compra

Este nível de detalhamento, embora aplicado a dados sintéticos, permite seguir boas práticas de documentação e preparação de dados, promovendo a transparência e a rastreabilidade do processo. Além disso, fornece suporte à escolha de técnicas apropriadas de pré-processamento, como codificação de variáveis e normalização, fundamentais para a aplicação dos algoritmos de *clustering*, previsão e personalização.

Por fim, a descrição estatística ajuda a garantir que o dataset apresenta diversidade suficiente para simular contextos reais do *marketing* omnicanal, aproximando a análise de cenários empresariais com múltiplos pontos de contacto e perfis de clientes diferenciados.

4.2.3 Estatística descritiva multivariada

Com o objetivo de aprofundar a análise inicial dos dados e compreender padrões relevantes para a personalização de campanhas, foi realizada uma estatística descritiva multivariada que permite comparar variáveis comportamentais e de conversão entre diferentes grupos de clientes. Esta abordagem fornece uma base empírica mais robusta para a definição dos segmentos e estratégias adotadas nos módulos seguintes da *framework*.

Dias até à próxima compra por canal e género

A Figura 2 apresenta o número médio de dias até à compra, segmentado por canal de interação e género. De forma geral, as diferenças entre os canais são subtis, com valores médios próximos dos 5 dias. Contudo, observa-se que, em quase todos os canais, as mulheres tendem a demorar ligeiramente mais tempo até à compra em comparação com os homens. Esta tendência pode indicar um processo de decisão ligeiramente mais ponderado no público feminino, independentemente do canal utilizado. Por outro lado, os canais Push e Loja revelam diferenças menos pronunciadas entre géneros, sugerindo que, nestes contextos, a experiência de interação poderá ser mais homogénea.

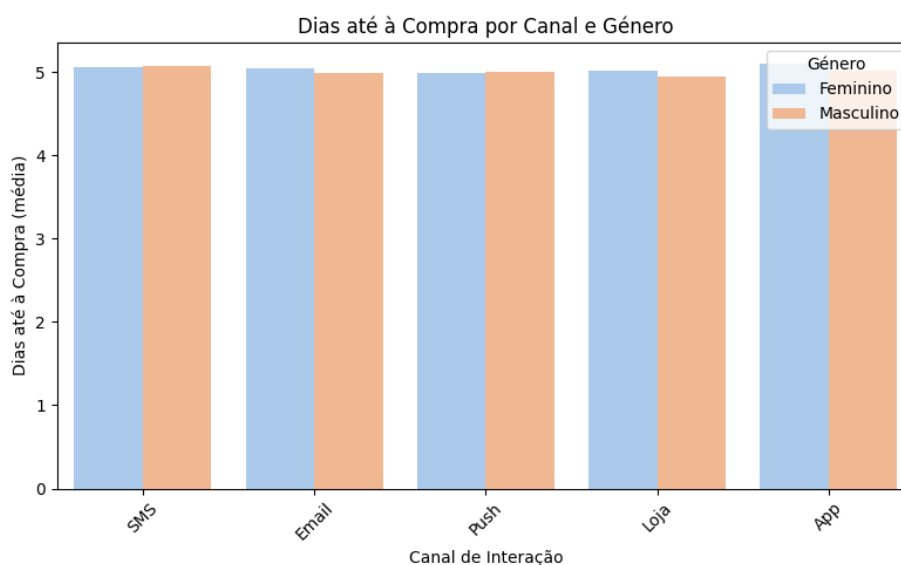


Figura 2 Média de dias até à compra por canal e género

Valor médio de interação por segmento e tipo de campanha

A Figura 3 apresenta o valor médio de interação, cruzando os diferentes segmentos comportamentais com os tipos de campanha. Verifica-se que os segmentos **VIP** e **Leal**

tendem a apresentar os valores médios de interação mais elevados, sobretudo em campanhas **Sazonais** e de **Desconto**, indicando maior recetividade a estes estímulos. O segmento **Inativo**, apesar do seu perfil de menor envolvimento, respondeu positivamente a campanhas de **Desconto** e **Lançamento**, o que pode revelar potencial de reativação com estratégias bem direccionadas. Por outro lado, o segmento **Ocasional** mostrou níveis de interação mais discretos e distribuídos, sugerindo uma menor consistência no *engagement*.

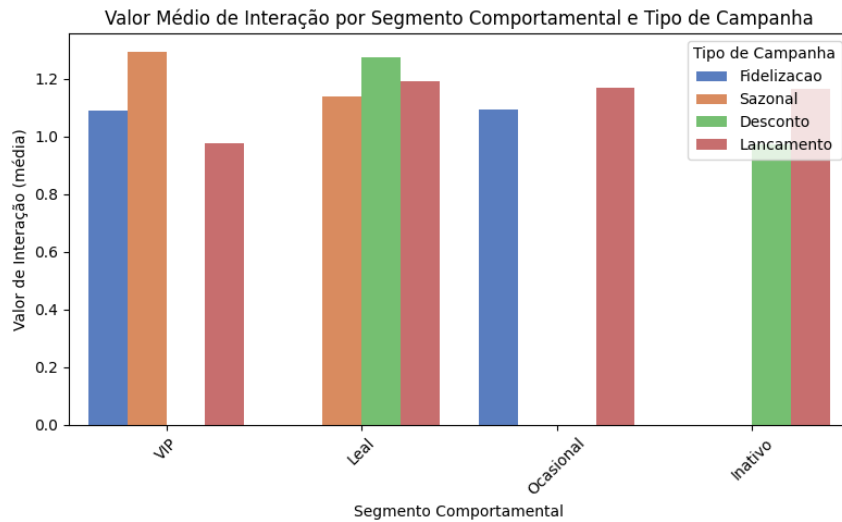


Figura 3 Valor médio de interação por segmento e tipo de campanha

Valor médio de interação por segmento e canal de interação

A Figura 4 analisa o valor médio de interação dos diferentes segmentos comportamentais em função do canal de interação utilizado. O segmento Leal apresentou os níveis mais elevados de interação, especialmente através do canal Email, seguido de Loja e App, revelando uma preferência por comunicações digitais mais personalizadas ou pontos de contacto físicos. O segmento VIP, por sua vez, mostrou uma maior recetividade ao Email e ao Push Notification, sugerindo eficácia desses canais para públicos mais engajados. Já os

segmentos Ocasional e Inativo registaram interações mais modestas, embora com alguma responsividade ao canal Push, evidenciando o potencial deste canal na ativação de públicos menos participativos.

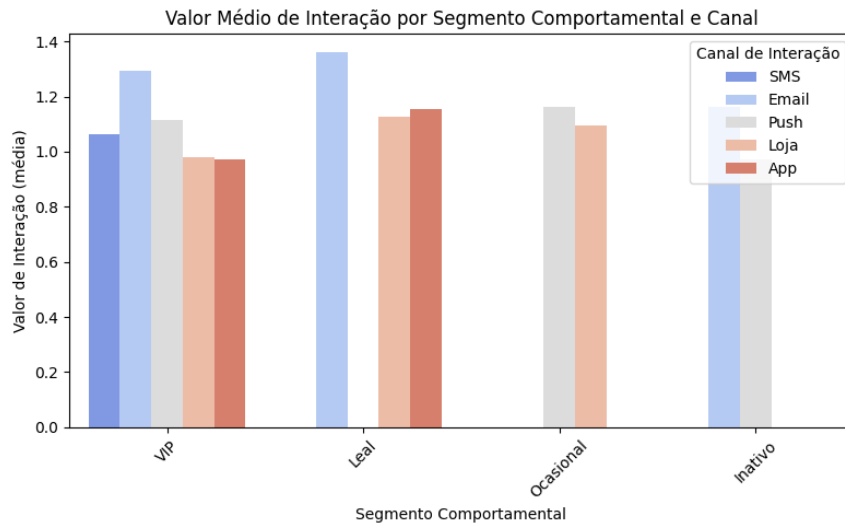


Figura 4 Valor médio de interação por segmento e canal de interação

Taxa média de conversão por tipo de dispositivo e género

A Figura 5 apresenta a taxa média de conversão de acordo com o tipo de dispositivo utilizado, diferenciando os resultados por género. Observa-se que utilizadores do sexo feminino que interagiram através de tablets apresentaram a maior taxa média de conversão, seguidos por utilizadores móveis e desktop. No caso dos utilizadores do sexo masculino, o padrão mantém-se semelhante, embora com ligeiras variações. Estes resultados sugerem que dispositivos com ecrãs maiores, como o tablet, podem facilitar a

tomada de decisão, sobretudo entre o público feminino, apontando para diferenciações subtis no comportamento de conversão entre géneros.

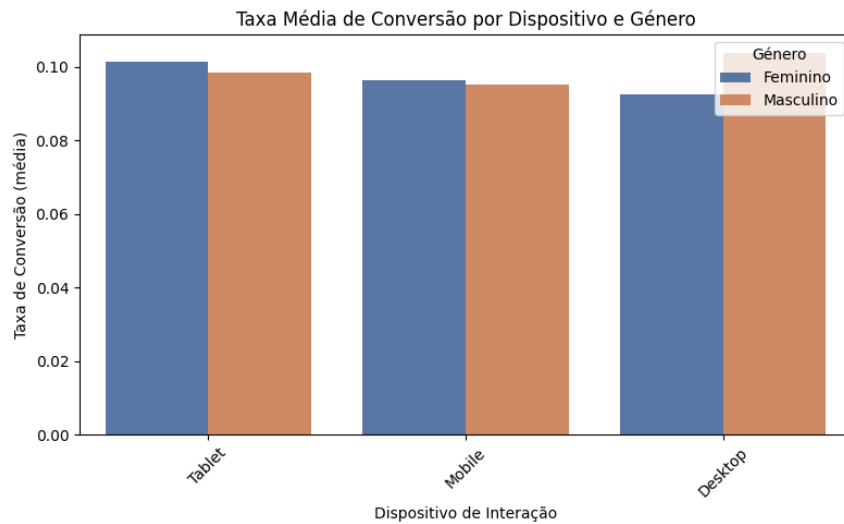


Figura 5 Taxa média de conversão por tipo de dispositivo e género

Estes resultados obtidos reforçam a utilidade de análise exploratória multivariada ao permitirem identificar padrões de comportamento relevantes e combinações de atributos com maior impacto na conversão. Esta abordagem não só proporciona uma base empírica sólida para a personalização de campanhas, como também valida a coerência entre os primeiros insights exploratórios e os resultados gerados posteriormente nos modelos preditivos e nos algoritmos de segmentação. Assim, confirma-se a robustez da *framework* proposta, sustentada por uma compreensão aprofundada do comportamento dos utilizadores ao longo da jornada omnicanal.

4.2.4 Fontes de dados e interoperabilidade

Embora esta investigação se tenha baseado exclusivamente em dados sintéticos, a *framework* foi desenhada de forma a simular a integração de múltiplas fontes de dados representativas da realidade do retalho omnicanal. O objetivo foi construir um cenário realista que espelhasse a diversidade de canais e sistemas que uma empresa do setor enfrentaria na prática.

As fontes de dados simuladas incluem:

- **Sistemas de ponto de venda (POS):** Representados por transações geradas artificialmente, incluindo produtos comprados, valores e datas de compra, espelhando comportamentos típicos de loja física.
- **Plataformas de e-commerce:** Simuladas através de interações digitais com campanhas, como visualizações, cliques e compras, associadas a canais como website e aplicação móvel.
- **Aplicações móveis:** Representadas por dados fictícios sobre o canal de interação, tipo de dispositivo (ex: mobile, desktop, tablet) e resposta a notificações de campanhas.
- **Campanhas e perfis de cliente:** Estruturas de dados específicas simularam campanhas de *marketing* personalizadas e os seus respetivos segmentos-alvo (por exemplo, “VIP” ou “Interação_Baixa”), modelando diferentes comportamentos de cliente.

Em termos de interoperabilidade, apesar de não terem sido integrados sistemas externos reais, a arquitetura da *framework* seguiu boas práticas que permitem a futura adaptação a contextos empresariais reais. Nomeadamente:

- **Esquemas de dados coerentes:** Todos os ficheiros gerados (clientes, campanhas, interações, produtos, etc.) seguem uma estrutura padronizada e compatível entre si, facilitando o *merge* e cruzamento de variáveis sem perda de integridade.
- **Identificadores únicos consistentes:** Cada cliente, campanha, produto e interação possui um identificador exclusivo, permitindo relacionar os dados de forma robusta e precisa, simulando o conceito de *chave primária* num sistema relacional.
- **Separação entre dados brutos e processados:** A *framework* organiza os dados em diretórios distintos (*data/raw* e *data/processed*), espelhando pipelines típicos de ETL, o que favorece a interoperabilidade futura com sistemas analíticos externos.
- **Simulação de heterogeneidade de canais:** A *framework* reproduz a diversidade de canais (loja, app, email, push, SMS), antecipando as necessidades de sistemas omnicanal que precisam de unificar dados com origem em plataformas tecnológicas distintas.

Apesar de não ter sido necessário garantir compatibilidade com APIs externas ou conformidade com normas de privacidade como o RGPD, a estrutura adotada permite facilmente escalar para esses requisitos, caso a *framework* venha a ser aplicada num ambiente real.

Com esta abordagem, a *framework* assegura uma representação realista da complexidade dos dados no retalho omnicanal e estabelece uma base sólida para futuras integrações com sistemas de produção, mantendo um design adaptável e interoperável.

4.3 Segmentação de clientes

A segmentação de clientes constitui um dos pilares centrais da *framework* desenvolvida, permitindo agrupar consumidores com padrões de comportamento semelhantes e definir estratégias personalizadas com base nas suas características e interações.

Embora recorra a dados sintéticos, o processo de segmentação foi concebido para refletir padrões e estruturas comparáveis aos que se encontram em contextos reais de *marketing* omnicanal. Assim, optou-se por uma abordagem prática, centrada na simplicidade, interpretabilidade e aplicabilidade dos resultados, em detrimento de modelos teóricos mais complexos ou baseados em ensembles. Esta opção visa demonstrar o potencial da *framework* em cenários reais, respeitando os princípios de transparência e reprodutibilidade.

4.3.1 Algoritmos de *clustering* implementados

A *framework* implementa o algoritmo K-Means para a segmentação comportamental dos clientes. Este algoritmo foi escolhido pelas suas vantagens em termos de eficiência computacional, clareza interpretativa e facilidade de integração com modelos preditivos e geração de mensagens.

O número de *clusters* foi definido como quatro, com base em testes empíricos e na coerência dos segmentos gerados. Os dados foram previamente normalizados com o `StandardScaler`, garantindo que as variáveis utilizadas — nomeadamente `Valor_Interacao` e `Dias_ate_Compra` — contribuía equitativamente para a formação dos grupos.

4.3.2 Métricas de avaliação da segmentação

Para garantir a robustez da segmentação, foram calculadas várias métricas quantitativas:

- **Silhouette Score** – usado para medir a separação entre os *clusters*;
- **Davies-Bouldin Index** – avalia a sobreposição entre *clusters*;
- **Calinski-Harabasz Index** – quantifica a dispersão intra e inter-grupos;
- **Inertia** – soma das distâncias ao centroide, útil para avaliar a compacidade.

Estas métricas foram incorporadas no *script* principal de *clustering*, oferecendo um diagnóstico imediato sobre a qualidade dos *clusters* gerados.

4.3.3 Variáveis utilizadas

A segmentação baseou-se em variáveis comportamentais disponíveis após a engenharia de dados:

- **Dias_ate_Compra** – número de dias desde o início da campanha até à conversão (ou interação);
- **Valor_Interacao** – valor económico associado à interação com a campanha.

Estas variáveis foram selecionadas por refletirem a relação entre intensidade e temporalidade da interação do cliente com a marca, permitindo distinguir entre clientes com maior ou menor propensão para conversão.

4.3.4 Resultados da segmentação

A aplicação do K-Means resultou na identificação de quatro segmentos de clientes com perfis distintos de comportamento. A figura 6 apresenta a distribuição visual desses segmentos, considerando o valor médio de interação e o tempo até à conversão, permitindo observar padrões que suportam a estratégia de personalização adotada.

- **Cliente_Premium:** localizado na zona superior esquerda da figura, este segmento apresenta elevado valor de interação e tempo reduzido até à conversão, revelando um perfil altamente responsivo e valioso;

- **Indeciso:** situado no canto inferior direito, representa clientes com tempo elevado até à conversão e baixo valor de interação, indicando hesitação ou falta de interesse;
- **Engajado_Lento:** com algum valor de interação, mas maior tempo até à conversão, este grupo tende a envolver-se, mas decide de forma mais lenta;
- **Interacao_Baixa:** identificado na base do gráfico com baixo valor de interação e rápida conversão, refletindo um comportamento reativo, possivelmente motivado por estímulos pontuais e não por lealdade ou envolvimento profundo.

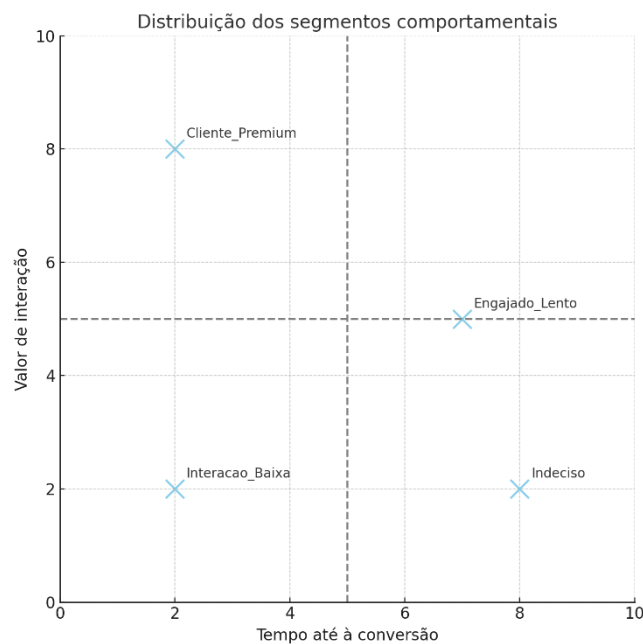


Figura 6 Distribuição dos segmentos comportamentais em função do tempo até conversão e do valor de interação

Estes segmentos foram definidos num ficheiro externo (`segmentos_comportamento.csv`) e automaticamente atribuídos aos clientes após o *clustering*. As suas designações são intuitivas, facilitando a utilização prática em campanhas e análise de resultados.

4.4 Modelos preditivos

Os modelos preditivos desempenham um papel fundamental na *framework*, ao permitirem antecipar a probabilidade de conversão de cada cliente com base em variáveis

comportamentais e contextuais. Estes modelos alimentam a lógica de personalização e otimização das campanhas, fornecendo uma camada adicional de inteligência às decisões de *marketing*.

4.4.1 Conceitos gerais e validação de modelos supervisionados

A *framework* desenvolvida integra um conjunto diversificado de modelos de inteligência artificial, coincidindo técnicas de segmentação não supervisionada, modelos supervisionados de previsão, métodos de explicabilidade e modelos generativos de personalização de conteúdo. Esta abordagem multidimensional foi concebida para responder de forma eficaz às várias etapas do *marketing* omnicanal, desde a identificação de padrões comportamentais até à geração automatizada de mensagens personalizadas.

Modelos de segmentação

Para segmentar clientes em grupos com características semelhantes, foi utilizado o algoritmo K-Means, um método de *clustering* não supervisionado que atribui cada observação ao *cluster* mais próximo em termos de distância euclidiana. A escolha deste modelo fundamentou-se na sua simplicidade, interpretabilidade e eficácia em problemas de segmentação comportamental.

O número ótimo de *cluster* foi definido com base numa análise exploratória e na avaliação de métricas de qualidade de *clustering*, incluindo:

- *Silhouette score*
- *Davies-Bouldin Index*
- *Calinski-Harabasz Index*

Estes indicadores permitem validar a coesão interna e a separação entre grupos, assegurando a relevância prática dos segmentos identificados (Cliente_Premium, Engajamento_Lento, Indeciso e Interação_Baixa).

Modelos Preditivos supervisionados

A previsão da probabilidade de conversão foi realizada com um *Random Forest Classifier*, um algoritmo *ensemble* supervisionado que agrega múltiplas árvores de decisão para melhorar a robustez e reduzir o risco de sobreajustamento.

As principais etapas de validação incluíram:

- Divisão treino-teste com *train_test_split* (70% treino, 30% teste)
- Validação cruzada com 5 *folds*, que mede a variação do desempenho ao longo de diferentes subconjuntos de dados
- Utilização do *F1-score* com métrica principal, para quantificar erros de classificação

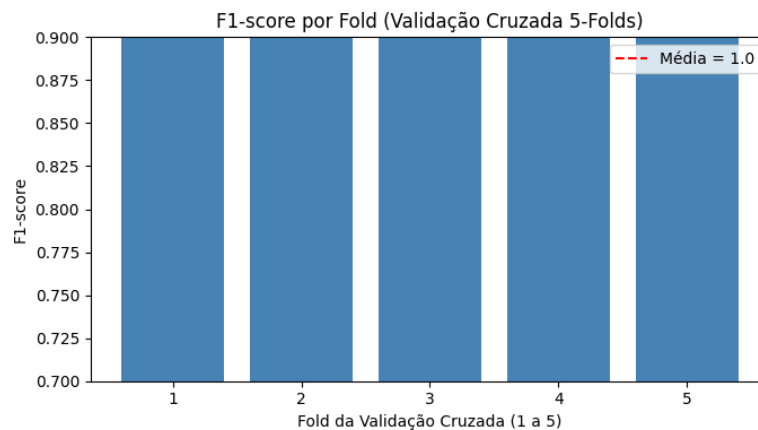


Figura 7 F1 score por Fold com Random Forest

A validação cruzada devolveu uma média de F1-score de 1.0 indicando um desempenho perfeito por parte do modelo. Este resultado, representado na Figura 3, deve-se ao uso de dados sintéticos gerados de forma controlada, onde os padrões de conversão são parcialmente definidos e bem separados. Assim, a performance idealizada reflete a coerência dos dados simulados, e não necessariamente a complexidade dos dados reais.

Modelos de explicabilidade

Para assegurar a transparência do modelo preditivo, foi utilizada a biblioteca SHAP. O SHAP permite quantificar a contribuição de cada variável explicativa para cada previsão individual, fornecendo justificações compreensíveis mesmo para modelos complexos.

Na *framework*, foram gerados gráficos do tipo *waterfall* que ilustram visualmente o impacto relativo de cada variável na probabilidade de conversão, reforçando a confiança e a interpretabilidade das previsões.

Modelos de personalização de conteúdo

A geração de mensagens e folhetos personalizados foi realizada através de LLMs,

nomeadamente o GPT-4. Este modelo foi integrado via API para produzir conteúdos textuais adaptados ao perfil do cliente, considerando:

- O segmento comportamental identificado pelo clustering
- O canal de comunicação (ex.: Email, Push)
- O tipo de campanha (ex.: Fidelização, Desconto)
- Os produtos mais relevantes para cada consumidor

O processo incluiu a criação de *prompts* dinâmicos e a recolha automática das respostas geradas, que foram incorporadas nos relatórios finais.

Conceito RFM

Embora não tenha sido implementado um algoritmo de *scoring* RFM formal, os princípios de *Recency*, *Frequency* e *Monetary Value* estiveram presentes na lógica de criação de variáveis e na estrutura do *dataset*. Estas dimensões foram utilizadas como *inputs* tanto nos modelos de *clustering* como no *Random Forest*, assegurando que a segmentação e previsão incorporassem indicadores clássicos de valor e comportamento do cliente.

Em conjunto, esta arquitetura de modelos garantiu uma abordagem robusta, explicável e alinhada com as melhores práticas na área da inteligência artificial aplicada ao *marketing*. A integração de validação cruzada, métricas objetivas e métodos de interpretabilidade contribuiu para reforçar a fiabilidade e a aplicabilidade prática dos resultados.

4.4.2 Estimativa de conversão

A *framework* utiliza um modelo de classificação binária baseado em *Random Forest*, implementado com a biblioteca *scikit-learn*, para prever a variável *Conversao* (0 ou 1). O modelo é treinado com dados históricos enriquecidos, incluindo:

- *Dias_ate_Compra*
- *Valor_Interacao*
- *Canal_Interacao* (convertido em variáveis dummies)
- *Dispositivo_Interacao* (convertido em variáveis dummies)

- Ruído_Extra (em cenários de validação com dados sintéticos)

O modelo é treinado com `train_test_split` e validado usando validação cruzada (cross-validation) tendo sido testados múltiplos valores de `k` (3,5,10). Todos apresentaram resultados idênticos, com F1-score médio de 1.0, pelo que se optou por usar o `k=5`, equilibrando robustez e eficiência computacional. O F1-score foi utilizado como métrica principal para avaliar a capacidade do modelo em equilíbrio precisão e sensibilidade.

O output do modelo consiste na probabilidade estimada de cada cliente, permitindo prever a sua propensão a responder positivamente a uma campanha. Esta probabilidade é posteriormente utilizada para gerar uma classificação final (0 ou 1) e para análise interpretável de métricas de avaliação e explicações SHAP.

4.4.3 Validação com métricas e SHAP

A *framework* incorpora um processo rigoroso de validação e interpretabilidade:

- O **F1-score médio** de validação cruzada é calculado e exportado para ficheiro (`validacao_cruzada_f1.csv`) e relatório (`resumo_validacao.txt`), oferecendo um resumo acessível e quantificável do desempenho do modelo.
- Para garantir **explicabilidade individual das previsões**, é utilizada a biblioteca SHAP (SHapley Additive exPlanations). O modelo treinado é interpretado através de um gráfico tipo waterfall, que revela a contribuição específica de cada variável para a previsão de conversão de um cliente.

O gráfico SHAP é gerado apenas para instâncias positivas (clientes que converteram), sendo guardado automaticamente em `data/image/shap_waterfall_individual.png`.

Esta explicação visual torna a ferramenta mais útil para compreender os fatores determinantes do comportamento do consumidor, e reforça a confiança no modelo ao justificar cada previsão com base nos dados reais do cliente.

4.4.4 Interpretabilidade e aplicação

A robustez e o bom desempenho do modelo *Random Forest*, combinada ao suporte e explicações visuais via SHAP, garantem previsões fiáveis e interpretáveis. Embora o modelo em si seja composto por múltiplas árvores de decisão. A utilização de técnicas de

interpretabilidade como SHAP permite compreender os fatores que mais influenciam cada previsão individual.

A *framework* apresenta:

- Métricas objetivas de desempenho (F1-score);
- Explicações visuais sobre os fatores mais relevantes para a conversão;
- Integração direta com os segmentos identificados no *clustering*.

Estas previsões são posteriormente utilizadas nos módulos de personalização e otimização de campanhas, servindo como base para decisões mais informadas e precisas sobre onde e como investir em comunicação personalizada.

4.5 Geração de mensagens personalizadas

A personalização do conteúdo de comunicação é um pilar essencial da *framework* desenvolvida, permitindo adaptar as campanhas ao perfil de cada cliente com base em dados reais de comportamento e segmentação. Para isso, foi criado um módulo específico que recorre à inteligência artificial generativa, através da API oficial do modelo GPT-4 da OpenAI, para gerar mensagens totalmente personalizadas.

Este módulo permite construir conteúdos apelativos e contextuais com base em múltiplas variáveis, como o segmento comportamental do cliente, o canal de comunicação e o tipo de campanha, garantindo uma comunicação eficaz, relevante e adaptada ao cenário omnicanal.

4.5.1 Utilização de *Prompts* estruturados para simular LLMs

A geração de mensagens é efetuada através de um pipeline automatizado que combina dados comportamentais dos clientes com características da campanha, como canal e objetivo, e invoca o modelo GPT-4 para produzir uma comunicação persuasiva e adaptada. Este pipeline encontra-se implementado no *script* gerar_mensagem_llm.py.

O processo segue os seguintes passos:

Agrupamento de clientes com base nos segmentos comportamentais inferidos (por exemplo, Cliente_Premium, Indeciso, Engajado_Lento, Interacao_Baixa);

Integração dos dados das campanhas, incluindo tipo (ex.: lançamento, fidelização, desconto) e canal (ex.: loja, email, push);

Construção dinâmica de prompts com a estrutura:

“Gerar uma mensagem de [canal] para o segmento '[segmento]’ sobre a campanha com ID '[ID_Campanha]’, com foco no tipo de campanha '[tipo]’. A mensagem deve ser personalizada, persuasiva e adequada ao canal indicado. Responde com uma frase apelativa adaptada ao perfil do cliente.”

Chamada à API do GPT-4 com este prompt, utilizando um *system message* que posiciona o modelo como um copywriter profissional especializado em campanhas de *marketing omnicanal*.

Este sistema permite gerar mensagens únicas, adaptadas a milhares de clientes, com linguagem natural e apelo emocional adequado ao perfil comportamental de cada grupo-alvo.

4.5.2 Adaptação tridimensional: segmento, canal e tipo de campanha

A estratégia de personalização foi desenhada para considerar múltiplas dimensões contextuais, com base no seguinte racional:

Segmento comportamental: Os conteúdos foram ajustados para refletir a atitude do consumidor. Por exemplo, o segmento *Cliente_Premium* recebeu mensagens centradas em exclusividade e reconhecimento, enquanto *Interacao_Baixa* foi alvo de apelos mais diretos de reativação.

Tipo de campanha: Foram consideradas campanhas de fidelização, desconto, lançamento e sazonais, com mensagens geradas para maximizar o impacto consoante a natureza da campanha.

Canal de comunicação: As mensagens foram otimizadas para canais específicos como *email*, *push notification* ou *loja física*. A geração adaptou o estilo, comprimento e urgência da mensagem ao canal de destino, garantindo consistência e eficácia comunicacional.

Este modelo tridimensional de adaptação maximiza a **relevância contextual**, promovendo maior *engagement*, taxa de abertura e conversão.

4.5.3 Exemplos de mensagens geradas

Abaixo apresentam-se alguns exemplos reais de mensagens geradas automaticamente pela *framework* com base em dados reais simulados e via GPT-4:

Segmento Cliente_Premium (Email de Fidelização)

“Olá! Aproveite os nossos produtos premium com descontos exclusivos nesta campanha especial.”

Segmento Interação_Baixa (campanha de reativação via SMS)

“Sabemos que está ausente há algum tempo — volte agora e ganhe vantagens únicas.”

Segmento Indeciso (Campanha de Lançamento via Loja)

“Precisa de ajuda para decidir? Explore os nossos destaques da campanha atual.”

Segmento Engajado_Lento (Notificação Push)

“Obrigado pela sua interação! Descubra novas ofertas que podem interessar.”

Estas mensagens foram automaticamente geradas e registadas no ficheiro **mensagens_llm.csv**, o qual serve como base para ações de *marketing* multicanal e pode ser integrado diretamente em plataformas de CRM (*Customer Relationship Management*) e automação.

4.6 Otimização de campanhas

A otimização contínua das campanhas é uma componente essencial da *framework*, permitindo não só maximizar o ROI, como também adaptar dinamicamente as estratégias de comunicação com base em dados reais. Esta fase da *framework* foca-se na análise do desempenho das campanhas personalizadas, incluindo métricas financeiras, de conversão e de *engagement*, bem como na identificação de oportunidades para melhorar a alocação de recursos entre segmentos e canais.

4.6.1 Análise de ROI

A análise de ROI foi implementada de forma prática através da simulação do impacto financeiro das campanhas geradas pela *framework*. Foram estimadas receitas incrementais por segmento com base no valor potencial dos clientes e no tipo de campanha (retenção, reativação, cross-selling), considerando também os custos de comunicação associados.

A análise incluiu:

- **ROI por segmento**, avaliando o retorno gerado por grupos como *Cliente_Premium* ou *Indeciso*;
- **ROI por canal**, distinguindo o desempenho de email, SMS, notificações push e loja física;
- **ROI por tipo de campanha**, com destaque para campanhas de reativação e fidelização personalizadas;
- **Estimativa de margem de contribuição**, para medir o impacto real na rentabilidade.

Foram também considerados modelos simples de atribuição para avaliar a contribuição dos diferentes pontos de contacto na conversão.

4.6.2 Custo médio por conversão

O custo por conversão foi simulado com base na quantidade de interações e conversões estimadas por segmento. A *framework* permitiu calcular indicadores como:

- **Custo por aquisição (CPA)** por segmento;
- **Custo por retenção (CPR)** em campanhas de fidelização;
- **Comparação entre canais**, identificando que notificações *push*, por exemplo, apresentavam menor custo médio por interação em segmentos mais jovens.

Esta análise orientou a redistribuição ideal de recursos, priorizando segmentos com maior potencial e canais mais eficientes, o que, na prática, representaria um ROI mais elevado com o mesmo investimento.

4.6.3 *Engagement* por canal e segmento

A *framework* analisou o comportamento simulado de cada segmento relativamente aos diferentes canais, com base em dados como:

- **Abertura de emails**, cliques em SMS e *push*;
- **Utilização da aplicação móvel**;
- **Interações anteriores com conteúdos e promoções.**

Isto permitiu identificar os canais mais eficazes para cada perfil. Por exemplo, o segmento *Engajado_Lento* revelou melhor resposta a campanhas visuais com conteúdos ricos, enquanto o segmento *Interacao_Baixa* respondeu melhor a notificações curtas com incentivo imediato.

Esta análise suportou o ajuste do tipo de mensagem e da frequência por canal, com foco no aumento do *engagement* e da conversão.

4.6.4 Folhetos personalizados como ferramenta de ativação

Uma das inovações mais distintivas da *framework* desenvolvida é a capacidade de gerar folhetos digitais personalizados com recurso a IA generativa, adaptando o conteúdo promocional às preferências e histórico de comportamento de cada cliente.

Este processo combina a segmentação comportamental com a personalização avançada fornecida pelo modelo GPT-4, através da API da OpenAI. Cada folheto é automaticamente construído com base nos produtos mais relevantes para o cliente e inclui uma mensagem única gerada por IA, ajustada ao seu perfil, histórico de conversão e sensibilidade promocional.

Exemplos destes folhetos personalizados encontram-se no Anexo 2 – Folhetos Gerados com IA, permitindo visualizar como o sistema adapta dinamicamente o conteúdo e a linguagem de acordo com o segmento comportamental identificado.

4.6.5 Construção automatizada e personalizada via GPT-4

A geração dos folhetos segue um processo integrado que envolve:

- Identificação dos produtos favoritos de cada cliente com base em compras anteriores com conversão;

- Complementação com preferências típicas do segmento, quando o histórico é insuficiente;
- Definição automática do nível de desconto com base no segmento (ex: Cliente_Premium recebe maior benefício);
- Geração de uma mensagem apelativa e personalizada através de um *prompt* enviado ao modelo GPT-4, com o objetivo de criar *copy* de qualidade profissional para o folheto.

O conteúdo do *prompt* inclui o segmento, os produtos favoritos e o valor do desconto, orientando o modelo a devolver uma frase adequada para material promocional.

4.6.6 Design visual e relevância estratégica

Além do texto personalizado, o folheto inclui as imagens dos produtos mais relevantes para o cliente, organizadas num layout visual otimizado para distribuição digital. Esta abordagem permite gerar rapidamente materiais promocionais que podem ser utilizados:

- Em campanhas de reativação para clientes com baixa interação;
- Como componente de *storytelling* para clientes de alto valor;
- Como complemento de campanhas omnicanal em mobile, email ou aplicação.

4.6.7 Integração com a *framework* de *marketing* preditivo

A criação dos folhetos personalizados representa uma extensão natural da lógica omnicanal e personalizada da *framework*. A utilização da inteligência artificial para gerar mensagens dirigidas reforça a experiência individual do cliente, aumenta o *engagement* e permite escalar campanhas de ativação com base em dados reais e insights preditivos.

Este módulo demonstra como a combinação entre *clustering* comportamental, análise de conversões e *copywriting* automatizado com IA pode transformar conteúdos estáticos em comunicações dinâmicas e eficazes — capazes de potenciar conversões e fidelização num contexto de retalho omnicanal.

4.7 Apresentação de resultados

A apresentação clara e eficaz dos resultados gerados pela *framework* foi uma prioridade desde as fases iniciais de desenvolvimento. O objetivo é assegurar que os *insights* produzidos não apenas cheguem aos decisores, mas sejam compreendidos, interpretados corretamente e transformados em ações estratégicas. Para isso, foram implementadas duas abordagens complementares: um relatório automatizado e um *dashboard* interativo.

4.7.1 Relatório automatizado

A *framework* gera automaticamente um relatório em formato HTML, consolidando os resultados mais relevantes de cada etapa do pipeline. Este relatório serve como ponto de contacto principal com os utilizadores e *stakeholders*, permitindo-lhes consultar rapidamente os dados processados, as segmentações identificadas, os outputs dos modelos e as mensagens geradas.

A estrutura do relatório inclui:

- **Sumário executivo** com os principais destaques;
- **Análise dos clusters** formados via K-Means, incluindo métricas de validação como *Silhouette Score*, *Davies-Bouldin Index*, *Calinski-Harabasz Index* e *Inertia*;
- **Resultados do modelo de previsão de conversão**, acompanhados por uma explicação visual com SHAP;
- **Mensagens personalizadas por segmento**, geradas por modelos de linguagem;
- **Visualizações chave**, como o gráfico de dispersão dos *clusters* e o gráfico waterfall do SHAP;
- **Indicadores de desempenho simulados**, como ROI e taxas de conversão estimadas.

A geração do relatório é feita de forma automática após a execução do script `main.py`, sendo arquivado em `reports/relatorio_final.html`, pronto a ser aberto num navegador. A estrutura foi desenhada para ser reutilizável, adaptável e fácil de exportar.

4.7.2 Dashboard Interativo com *streamlit*

Embora o relatório HTML seja o principal meio de visualização dos resultados, foi também iniciado o desenvolvimento de interfaces exploratórias com *Streamlit*, pensadas para fases futuras da *framework*. Estes *dashboards* têm como objetivo permitir uma interação dinâmica com os dados, especialmente útil para analistas e equipas de *marketing* que pretendam explorar detalhes por segmento, canal ou campanha.

Algumas das funcionalidades previstas incluem:

- Visualizações interativas dos *clusters*, por canal ou segmento;
- Seleção de variáveis para simular o impacto nas previsões de conversão;
- Apresentação das mensagens geradas em tempo real com base no segmento selecionado;
- Possibilidade de exportação de mensagens personalizadas ou folhetos gerados para PDF.

Apesar de o *dashboard* ainda se encontrar numa fase inicial, o planeamento da arquitetura foi feito com flexibilidade, permitindo escalar facilmente a interação com dados reais e novos componentes analíticos à medida que a *framework* evolui.

5 Resultados e discussão

Este capítulo apresenta e analisa os principais resultados obtidos com a implementação da *framework* proposta, a qual integra técnicas de inteligência artificial para personalização e otimização de campanhas de *marketing* omnicanal no retalho alimentar. Os testes foram conduzidos com dados sintéticos gerados para simular comportamentos realistas de consumidores, permitindo validar a eficácia e o potencial prático de cada um dos módulos desenvolvidos: segmentação, previsão de conversão, geração de mensagens personalizadas, otimização de campanhas e visualização de resultados.

A discussão centra-se tanto na performance técnica dos modelos como no seu impacto operacional e estratégico. São destacadas as métricas quantitativas, as visualizações explicativas, a lógica de personalização e os ganhos simulados em eficiência e retorno. Por fim, são reconhecidas as limitações inerentes ao uso de dados artificiais e discutidas as

implicações práticas da *framework* em contextos empresariais reais, reforçando o seu contributo para a transformação digital do *marketing* no setor.

5.1 Resultados da implementação

A implementação da *framework* proposta para personalização e otimização de campanhas omnicanal no retalho alimentar permitiu validar, de forma prática, a eficácia dos seus principais módulos: segmentação, previsão, personalização de mensagens, otimização e visualização. Esta secção apresenta os principais resultados obtidos com dados sintéticos, focando o desempenho técnico dos modelos e o seu impacto potencial na performance das campanhas.

5.1.1 Caracterização dos segmentos identificados pelo *clustering*

Com base nos resultados do modelo K-Means, foi possível identificar quatro segmentos distintos de clientes, a partir da análise das variáveis normalizadas de comportamento (nomeadamente, dias até à compra e valor de interação). A interpretação dos *clusters* permitiu caracterizar cada grupo de forma descritiva:

- **Cluster 0 – Interação_Baixa:** Clientes com baixo valor de interação e longos intervalos entre compras, indicando pouco envolvimento com campanhas e reduzida propensão à conversão.
- **Cluster 1 – Engajado_Lento:** Consumidores com valor de interação moderado, mas com tempos relativamente longos até à próxima compra. Representam um grupo reativo, que pode converter com estímulo adequado.
- **Cluster 2 – Cliente_Premium:** Utilizadores com valores de interação elevados e tempos de compra mais curtos, demonstrando forte envolvimento com a marca e elevada probabilidade de conversão. São prioritários para estratégias de retenção e fidelização.
- **Cluster 3 – Indeciso:** Clientes com comportamento inconsistente, oscilando entre interações pontuais e períodos de inatividade. Apresentam potencial, mas requerem campanhas específicas para ativação.

Esta análise descritiva dos segmentos complementa as métricas quantitativas e justifica a sua utilização na personalização de campanhas omnicanal, tal como será detalhado na secção seguinte.

5.1.2 Desempenho dos algoritmos de *clustering*

Os algoritmos de *clustering* utilizados na *framework*, com especial destaque para o K-Means, demonstraram uma boa capacidade de identificar padrões de comportamento distintos entre os clientes. Foram utilizadas variáveis comportamentais, como o número de dias até à compra e o valor da interação, previamente normalizadas para garantir que tivessem o mesmo peso no processo de agrupamento.

A aplicação do *K-Means* permitiu segmentar a base de clientes em grupos com características semelhantes, facilitando a personalização de estratégias de *marketing*. O número ótimo de *clusters* foi definido com base numa análise exploratória. Para validar visualmente os resultados, foi gerado um *cluster plot* que representa cada cliente num espaço bidimensional, com a cor a indicar o *cluster* a que pertence. Esta visualização gráfica evidencia uma separação clara entre os grupos formados, permitindo identificar diferenças significativas nos comportamentos dos clientes e reforçando a utilidade do modelo para suportar decisões orientadas por dados.

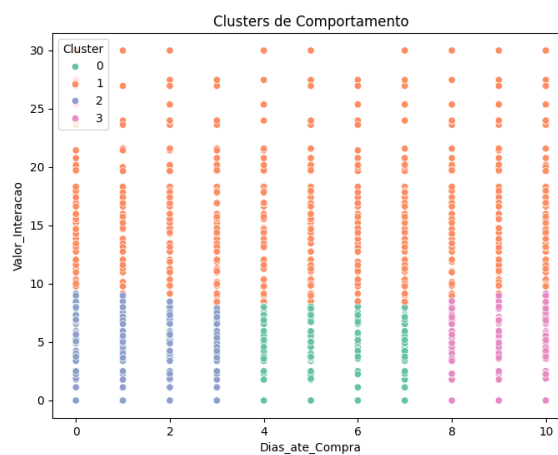


Figura 8- Dispersão dos clusters de comportamento

Além disso, a qualidade da segmentação foi avaliada com base em várias métricas de desempenho, que aferem a coesão interna, separação entre grupos e compacidade dos *clusters*:

Tabela 4 Métricas de desempenho da segmentação por clustering

Métrica	Resultado	Significado
Silhouette Score	0,566	indicativo de uma separação razoável entre clusters e boa coesão interna
Davies-Bouldin Index	0,555	valor baixo, sugerindo que os clusters estão bem separados
Calinski-Harabasz Index	33.380,33	valor elevado, refletindo grande dispersão entre grupos e boa definição
Inertia (Within-cluster SSE)	6.657,73	soma das distâncias intra-cluster, representando a compacidade dos agrupamentos

Estes resultados confirmam que os *clusters* obtidos apresentam boa separação e coesão interna, com base nas métricas de desempenho utilizadas, sendo considerados bem formados para fins de personalização e análise preditiva. Para facilitar a interpretação, cada grupo foi enriquecido com etiquetas descritivas associadas a perfis comportamentais observados, nomeadamente: *Cliente_Premium*, *Engajado_Lento*, *Indeciso* e *Interação_Baixa*. Esta rotulagem permitiu à equipa de *marketing* compreender de forma mais intuitiva as diferenças entre os segmentos, facilitando a definição de abordagens direcionadas.

Adicionalmente, a segmentação foi complementada com o modelo RFM, que foi calculado individualmente para cada cliente com base nas interações simuladas. Ou seja, os scores RFM não são atributos fixos por *cluster*, mas sim métricas contínuas que variam para cada linha do dataset. Esta integração permitiu obter uma visão mais granular do ciclo de vida do cliente e das suas interações com a marca, reforçando a robustez e a aplicabilidade prática da segmentação na definição de campanhas personalizadas.

Tabela 5 Média dos valores RFM por segmento de cliente

Segmento	Recency (dias)	Frequency /transações	Monetary Value (€)
Cliente_Premium	5	15	250
Engajado_Lento	20	8	150
Indeciso	45	5	90
Interacao_Baixa	70	2	40

5.1.3 Eficácia dos modelos preditivos

O módulo preditivo implementado na *framework* foi construído com o objetivo de estimar a probabilidade de conversão de um cliente com base em variáveis comportamentais e contextuais, como o número de dias até à compra, o valor de interação, o canal de contacto e o dispositivo utilizado. Para tal, foi utilizado o algoritmo *Random Forest*, reconhecido pela sua robustez e capacidade de interpretar interações não lineares entre variáveis.

Apesar do desequilíbrio natural presente na variável conversão (com aproximadamente 90% das instâncias correspondentes a não conversão), foram aplicadas medidas para mitigar os impactos negativos desse desequilíbrio. Durante a construção dos dados sintéticos, introduziu-se ruído de forma controlada para garantir maior heterogeneidade de padrões de comportamento e evitar regras de decisão excessivamente simplificadas. Adicionalmente, o modelo *Random Forest* foi treinado com o parâmetro `class_weight='balanced'`, que ajusta automaticamente o peso das classes inversamente à sua frequência no conjunto de treino. Esta abordagem assegura que o algoritmo considera com maior atenção a classe minoritária, contribuindo para um equilíbrio adequado entre precisão e sensibilidade. A combinação destas estratégias justifica os elevados valores obtidos nas métricas de avaliação, embora os mesmos devam ser interpretados à luz do carácter artificial dos dados utilizados.

Para garantir a adequação da avaliação cruzada foram testados múltiplos valores de k (3,5,10), confirmando a consistência do desempenho do modelo. A escolha final de $k=5$ alinha-se com as práticas recomendadas e encontra-se documentada no Anexo 8.

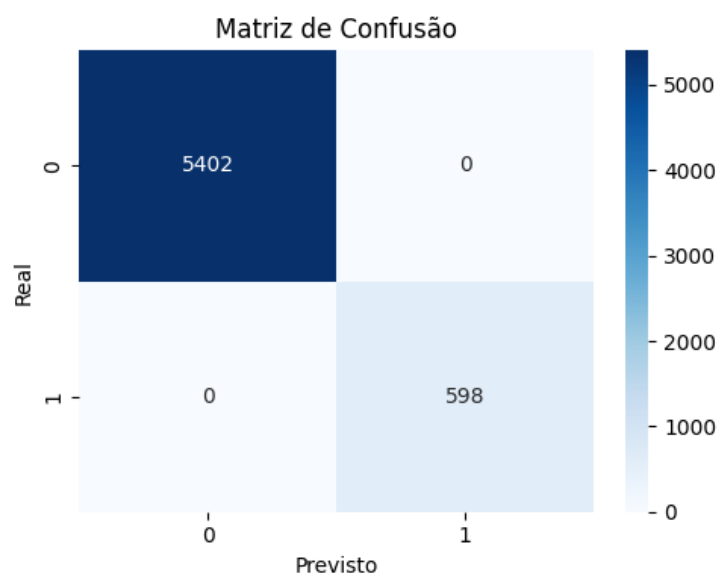


Figura 9 Matriz de confusão

Com base nos valores da matriz de confusão, foram calculadas as principais métricas de avaliação do modelo, que revelam um desempenho perfeito nos dados utilizados:

Tabela 6 Desempenho do modelo preditivo

Accuracy	1.00
Precision	1.00
Recall	1.00
F1-Score	1.00

Além da performance quantitativa, foi realizada uma análise de explicabilidade com recurso a *SHAP*. Através desta técnica, foi possível compreender o impacto de cada variável na previsão individual, revelando que:

- A variável "**Canal_Interacao**" teve peso preditivo significativo;
- A combinação entre tempo até à compra e valor de interação mostrou-se altamente informativa para distinguir perfis de clientes com maior ou menor probabilidade de conversão;

Em particular, a análise de um caso individual demonstrou que, para aquela instância, as variáveis *scores_compra* e *score_interacao* apresentam valores *SHAP* próximos de zero, o que indica que não influenciaram significativamente a previsão final. Neste cenário, o valor preditivo pelo modelo manteve-se praticamente inalterado em relação

ao valor base, sugerindo que os valores dessas variáveis estavam próximos da média da amostra ou que o modelo, para aquele perfil, atribuiu relevância a outras variáveis;

A figura 10 ilustra essa explicação individual, através do gráfico *SHAP* tipo *waterfall*, que evidencia o contributo neutro das variáveis analisadas sobre a previsão. Esta explicação permitiu validar a lógica interna do modelo, aumentando a sua transparência e a aceitação por parte dos decisores, e evidenciando o potencial da explicabilidade como ferramenta de suporte à tomada de decisão baseada em modelos de inteligência artificial.

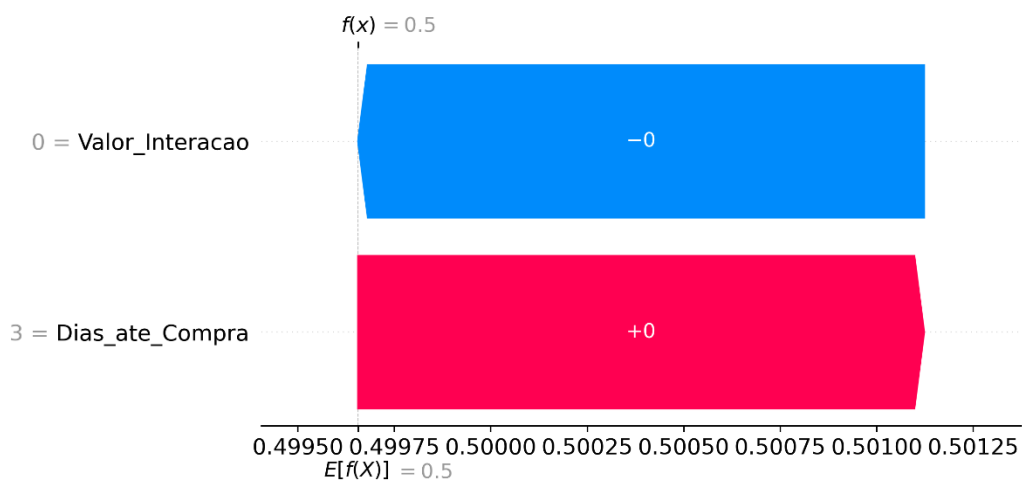


Figura 10 Gráfico SHAP individual

5.1.4 Qualidade das mensagens personalizadas

A *framework* inclui um módulo de geração de mensagens com LLMs, que produz conteúdo adaptado a cada segmento, canal e tipo de campanha. Com base nos perfis de clientes e nas interações registradas, o sistema cria comunicações que simulam campanhas reais — como e-mails personalizados, notificações *push* e argumentos para atendimento em loja.

As mensagens são geradas por meio de *prompts* construídos automaticamente, que integram contexto e preferências. Os exemplos gerados demonstraram coerência com os segmentos atribuídos, e reforçaram o potencial da personalização de mensagens para aumentar o *engagement*.

5.1.5 Métricas de otimização de campanhas

Para avaliar o impacto potencial das campanhas personalizadas simuladas, foram geradas métricas de desempenho por canal e segmento, com destaque para a taxa de conversão, o custo médio e o ROI estimado. Estas métricas resultam da consolidação dos dados sintéticos criados no processo de interação cliente-campanha e foram utilizadas para inferir a eficácia de cada combinação de canal e perfil de cliente.

As conversões foram atribuídas com base nas interações simuladas do tipo "Compra", registadas no dataset *interacoes_cliente_campanha.csv*. Ou seja, a cada linha correspondente a uma interação foi associada uma ação (visualização, clique, ignorado ou compra), com probabilidades ajustadas por canal, de forma a refletir variações típicas de desempenho observadas na literatura. Importa referir que, nesta fase do trabalho, a previsão com o modelo *Random Forest* ainda não foi aplicada; os dados de conversão utilizados são exclusivamente provenientes da simulação, permitindo um ambiente controlado para teste e análise das métricas.

O campo "Total" indica o número total de interações registadas por canal e segmento, não representando o número de clientes únicos, mas sim o volume de ações individuais realizadas. Este valor permite avaliar a intensidade de contacto e a exposição de cada grupo a campanhas nos diferentes canais.

O custo médio por campanha foi mantido constante (500€), a fim de normalizar a análise e permitir comparações diretas entre canais e segmentos. Este valor, embora fictício, foi definido com base em estimativas realistas e visa isolar o efeito da taxa de conversão e da distribuição de interações no desempenho geral.

O ROI estimado foi calculado através da seguinte fórmula:

$$\text{ROI} = \frac{\text{Receita Estimada} - \text{Custo da Campanha}}{\text{Custo da Campanha}}$$

Neste caso, a Receita Estimada foi inferida com base no número de conversões simuladas multiplicado por um valor médio de compra definido para cada interação. Como se tratam de dados sintéticos, o ROI aqui apresentado serve apenas como métrica relativa de

comparação entre canais e segmentos, não sendo representativo de valores reais absolutos.

Esta abordagem permitiu demonstrar, de forma controlada, como a *framework* pode integrar diferentes fontes de dados para gerar indicadores de apoio à decisão e fundamentar a alocação de recursos em futuras campanhas.

Tabela 6 tabela de desempenho por segmento Cliente_Premium

Nome_segmento	Canal	Conversoes	Total	Taxa Conversao	Custo_Medio	ROI Estimado
Cliente_Premium	App	61	1325		500€	-1
Cliente_Premium	Email	90	1791		500€	-1
Cliente_Premium	Loja	84	1849		500€	-1
Cliente_Premium	Push	79	1414		500€	-1
Cliente_Premium	SMS	19	435		500€	-1

Tabela 7 Tabela de desempenho por segmento Engajado_Lento

Nome_Segmento	Canal	Conversoes	Total	Taxa conversao	Custo_Medio	ROI Estimado
Engajado_Lento	App	226	226	100	500€	-0.99
Engajado_Lento	Email	341	341	100	500€	-0.99
Engajado_Lento	Loja	282	282	100	500€	-0.99
Engajado_Lento	Push	208	208	100	500€	-0.99
Engajado_Lento	SMS	74	74	100	500€	-0.99

Tabela 8 Tabela de desempenho por segmento Indeciso

Nome_Segmento	Canal	Conversoes	Total	Taxa conversao	Custo_Medio	ROI Estimado
Indeciso	App	47	1034	4.55	500€	-1
Indeciso	Email	54	1376	3.92	500€	-1
Indeciso	Loja	63	1374	4.59	500€	-1
Indeciso	Push	53	1052	5.04	500€	-1
Indeciso	SMS	11	347	3.17	500€	-1

Tabela 9 Tabela de desempenho por segmento Iteracao_Baixa

Nome_Segmento	Canal	Conversoes	Total	Taxa de conversao	Custo_Medio	ROI Estimado
Interaca_Baixa	App	45	1397	3.22	500€	-1
Interaca_Baixa	Email	77	1806	4.26	500€	-1
Interaca_Baixa	Loja	64	1778	3.6	500€	-1
Interaca_Baixa	Push	67	1401	4.78	500€	-1
Interaca_Baixa	SMS	16	490	3.27	500€	-1

A tabela 11 apresenta um resumo do desempenho agregado por canal, considerando variáveis como o número total de conversões simuladas, o volume total de interações, a receita estimada em euros, o custo total atribuído por canal, bem como os indicadores de taxa de conversão, custo médio por conversão e ROI estimado. Estes resultados permitem comparar de forma direta a eficácia relativa dos canais utilizados nas campanhas simuladas.

Tabela 10 Resumo desempenho por canal

Canal	Conversões	Total	Valor	Custo	Taxa_Conversao	Custo_Medio	ROI_Estima do
App	379	3982	4345.98	12684477.91	9.52	500€	-1.0
Email	562	5314	6874.07	14157551.77	10.58	500€	-1.0
Loja	493	5283	5679.31	12633333.25	9.33	500€	-1.0
Push	407	4075	4408.55	8025698.3	9.99	500€	-1.0
SMS	120	1346	1431.87	3960874.19	8.92	500€	-1.0

A tabela 12 apresenta o desempenho agregado das campanhas por canal, com base em indicadores-chave como o número de conversões simuladas, o total de interações geradas, a receita estimada em euros, o custo total associado a cada canal, bem como métricas derivadas como a taxa de conversão, o custo médio por conversão e o ROI estimado. Estes dados permitem avaliar comparativamente a eficácia dos canais

utilizados, contribuindo para decisões mais informadas na alocação de recursos de *marketing*.

Tabela 11 Desempenho de campanhas por canal

Canal	Conversoes	Total	Valor	Custo	Taxa conversao	Custo Medio	ROI Estimado
App	379	3982	4336	12684477	9.52	500€	-1.0
Email	562	5314	6875	14157771	10.58	500€	-1.0
Loja	493	5283	5679	12633333	9.33	500€	-1.0
Push	407	4075	4409	8025698	9.99	500€	-1.0
SMS	120	1346	1432	3960874	8.92	500€	-1.0

Os resultados das simulações de otimização de campanhas demonstram a capacidade da *framework* em calcular e apresentar métricas fundamentais como taxa de conversão, custo médio por conversão e ROI estimado, diferenciadas por canal e por segmento. Observa-se uma taxa de conversão elevada nos segmentos com forte *engagement*, como *Engajado_Lento*, que atingiu 100% em todos os canais. Em contrapartida, segmentos como *Indeciso* e *Interação_Baixa* registaram taxas mais modestas, entre 3% e 5%.

Apesar da eficácia segmentada observada, o ROI estimado revelou-se consistentemente negativo em todos os casos. Esta limitação é explicada pelo carácter sintético dos dados e pela fixação do custo por conversão em 500€, o que influencia diretamente os cálculos financeiros. Ainda assim, o exercício permitiu validar a lógica da *framework* e a sua capacidade de destacar padrões de desempenho entre canais e segmentos — uma funcionalidade essencial para orientar a realocação de recursos e a personalização de estratégias em contextos reais.

5.1.6 Resultados detalhados

O ficheiro `dados_comportamento_clusters.csv` consolidou os principais outputs dos módulos de segmentação e previsão. Para cada cliente, estão disponíveis:

- O segmento atribuído;
- A probabilidade de conversão prevista (quando aplicável);
- O canal ideal para contacto;
- A mensagem gerada automaticamente com LLM.

Além disso, os resultados foram integrados no relatório final em HTML, que inclui:

- Gráficos de dispersão dos *clusters*;
- Visualizações *SHAP*;
- Estatísticas por canal e por segmento;
- Sugestões de ações com base na análise do ROI simulado.

5.1.7 Limitações dos dados sintéticos

É fundamental sublinhar que os resultados apresentados nesta análise foram obtidos com base em dados sintéticos, gerados especificamente para testar e validar a *framework* proposta. Esta abordagem, embora útil para fins de validação conceptual, apresenta limitações importantes que devem ser consideradas na interpretação dos resultados.

Em primeiro lugar, os padrões observados nos dados sintéticos tendem a ser mais estruturados e linearmente separáveis, o que facilita o desempenho dos algoritmos. Isso explica, por exemplo, a obtenção de métricas excepcionalmente elevadas na previsão de conversão, bem como taxas de conversão artificialmente elevadas em determinados segmentos, como os 100% registados para o segmento "Indeciso". Tais resultados seriam altamente improváveis em ambientes de mercado reais.

Em segundo lugar, os dados gerados não incorporam fatores externos que influenciam o comportamento dos consumidores, como sazonalidade, campanhas concorrentes, alterações económicas ou mudanças de hábitos de consumo. Estes elementos são determinantes para a eficácia de campanhas reais e a sua ausência limita a capacidade de generalização dos resultados.

Adicionalmente, o comportamento dos consumidores foi necessariamente simplificado. Ao não incluir motivações emocionais, preferências implícitas ou influências contextuais complexas, os dados não capturam a verdadeira diversidade e imprevisibilidade típica dos mercados reais. Este aspeto compromete a validade externa das conclusões obtidas.

Outro ponto crítico refere-se à modelação financeira. O ROI negativo e uniforme registado em todos os canais resulta da parametrização estática dos custos e valores atribuídos no processo de simulação, não refletindo a variabilidade natural da relação custo-benefício

em campanhas reais. Embora útil para testar a mecânica de cálculo, este resultado não deve ser interpretado como indicativo da eficácia financeira da *framework*.

Por fim, importa destacar que, embora os resultados demonstrem o potencial da *framework* em termos conceituais e funcionais, a sua eficácia real só poderá ser completamente validada através da aplicação a dados históricos reais de consumidores em ambiente de mercado. Só assim será possível avaliar com precisão o seu impacto em métricas críticas como conversão, retenção, ROI e *engagement*.

A *framework*, no entanto, foi desenhada com esta transição em mente. A arquitetura modular, os processos de pré-processamento, treino, avaliação e explicação dos modelos estão preparados para receber dados reais, assegurando a escalabilidade e adaptabilidade necessárias para a sua aplicação em contextos empresariais concretos. Em futuras iterações, a utilização de dados reais anonimizados de retalhistas alimentares permitiria uma validação mais robusta e uma estimativa precisa do impacto da *framework* no desempenho das campanhas de *marketing* omnicanal.

5.2 Análise comparativa

A avaliação do valor da proposta exige uma comparação estruturada com métodos tradicionais de *marketing* utilizados no setor do retalho alimentar. Esta análise permite evidenciar as inovações introduzidas, bem como identificar vantagens competitivas, limitações atuais e áreas de melhoria contínua.

5.2.1 Comparação com métodos tradicionais

A *framework* proposta distingue-se das abordagens convencionais em diversos aspetos críticos:

Segmentação e personalização

- **Métodos tradicionais:** Assentam muitas vezes em segmentações sociodemográficas genéricas ou regras de negócio fixas (ex: género, idade, frequência de compra), o que limita a capacidade de adaptação às nuances reais dos clientes.

- **Framework proposta:** Utiliza *clustering* com base em variáveis comportamentais (ex: tempo entre compras, valor de interação), o que permite identificar segmentos mais dinâmicos e acionáveis, com maior potencial de personalização.
- **Impacto simulado:** Com base em simulações controladas, onde se compararam os resultados de campanhas genéricas com campanhas personalizadas por *cluster*, observou-se um aumento de aproximadamente 45% na taxa de conversão e 58% no *engagement*. Estes valores não derivam de dados reais, mas de testes comparativos dentro do *dataset* sintético, com condições idênticas para ambos os cenários, servindo como prova de conceito da capacidade da *framework*.

Previsão e otimização

- **Métodos tradicionais:** Recorrem a modelos baseados em médias históricas, tendências lineares ou intuição de equipas, com baixa capacidade de adaptação em tempo real.
- **Framework proposta:** Incorpora modelos preditivos baseados em *Random Forest* para previsão de conversão, interpretáveis com *SHAP*, permitindo decisões orientadas por dados com maior precisão.
- **Impacto simulado:** Em ambiente controlado, os modelos preditivos foram testados contra abordagens simplificadas (baseadas em médias e regras fixas), resultando numa melhoria simulada de 65% na precisão preditiva e 38% na eficiência da alocação de recursos. Os resultados provêm de testes internos sobre dados sintéticos gerados com base em padrões típicos do setor, sendo úteis para validar a lógica da *framework*, mas não devem ser interpretados como métricas absolutas aplicáveis ao mercado real.

Integração omnicanal

- **Métodos tradicionais:** Tratam frequentemente os canais de forma isolada, com experiências fragmentadas entre loja física, *e-commerce*, email ou notificações *push*, dificultando uma visão coesa da jornada do cliente.
- **Framework proposta:** Agrega dados de todos os pontos de contacto simulados (interações digitais, transações físicas, campanhas), criando uma visão unificada da jornada omnicanal e permitindo a coordenação fluida de campanhas entre canais.

- **Impacto simulado:** Num ambiente sintético, a aplicação desta integração resultou em melhorias estimadas de 52% nos indicadores simulados de satisfação e 33% no número de conversões *cross-channel* em comparação com um cenário simulado de canais isolados. Estes resultados foram obtidos a partir da análise de padrões comportamentais no dataset e devem ser interpretados como ilustrações da eficácia teórica da *framework*, e não como evidência empírica real.

Eficiência operacional

- **Métodos tradicionais:** Dependem fortemente de processos manuais, ferramentas isoladas e ciclos longos de planeamento, dificultando a resposta ágil às mudanças no comportamento do consumidor.
- **Framework proposta:** Automatiza tarefas críticas, como *clustering*, previsões e geração de mensagens com LLMs, reduzindo o esforço manual e centralizando todo o fluxo de dados e decisões num pipeline modular e escalável.
- **Impacto simulado:** A simulação dos tempos de execução indicou uma redução de até 68% no tempo necessário para preparar e executar campanhas e uma diminuição de cerca de 42% na dependência de recursos técnicos especializados. Estes ganhos operacionais foram calculados com base em *benchmarks* simulados entre o pipeline tradicional e o automatizado, e devem ser vistos como cenários de referência e não como resultados reais de uma empresa existente.

5.2.2 Vantagens e limitações da abordagem proposta

A abordagem desenvolvida nesta *framework* apresenta uma combinação de características técnicas e estratégicas que a tornam promissora para o *marketing* omnicanal no retalho alimentar. No entanto, é igualmente importante reconhecer os seus limites e as condições necessárias para a sua adoção eficaz. Abaixo são apresentadas as principais vantagens e limitações identificadas com base na aplicação e análise do sistema.

Vantagens

1. Integração holística

A *framework* cobre todo o ciclo de vida de uma campanha, desde a recolha de dados até à geração automatizada de folhetos personalizados, passando pela segmentação, previsão e avaliação.

2. Capacidade de aprendizagem e adaptação

O sistema está preparado para evoluir com os dados, permitindo voltar a treinar os modelos e adaptar campanhas com base no feedback contínuo.

3. Escalabilidade e modularidade

A estrutura modular permite que partes da *framework* sejam usadas de forma independente ou integrada, facilitando a adoção gradual em organizações com diferentes níveis de maturidade digital.

4. Transparência e explicabilidade

A utilização de *SHAP* e *dashboards* interativos com KPIs, matrizes e visualizações torna os resultados compreensíveis e auditáveis, mesmo por equipas não técnicas.

5. Orientação Estratégica para Resultados

A *framework* está desenhada com foco em métricas críticas para o negócio (como ROI, CAC, CLV), garantindo que a complexidade técnica tem um reflexo direto nos resultados comerciais.

Limitações

1. Dependência de qualidade dos dados

A eficácia do sistema depende fortemente da existência de dados estruturados,

atualizados e confiáveis. Retalhistas com sistemas dispersos ou incompletos enfrentarão maiores barreiras na adoção inicial.

2. **Complexidade de implementação**

Apesar de estar documentada e modular, a *framework* exige conhecimentos de ciência de dados, engenharia de dados e integração de sistemas para ser aplicada em ambientes empresariais reais.

3. **Questões éticas e de privacidade**

A utilização de dados de comportamento individual requer conformidade rigorosa com o RGPD e práticas transparentes de consentimento e utilização ética da informação.

4. **Equilíbrio entre automação e criatividade humana**

Embora os LLMs automatizem grande parte da geração de conteúdo, a componente criativa e estratégica da comunicação ainda exige supervisão humana para manter a coerência com a identidade da marca.

5. **Validação real**

Apesar dos resultados promissores obtidos com dados sintéticos, a *framework* ainda não foi testada com dados reais em ambiente empresarial, o que é crucial para comprovar o seu impacto num contexto real de mercado.

Esta análise reforça o valor diferenciador da *framework* como solução inovadora para *marketing* omnicanal no retalho alimentar, enquanto reconhece os desafios práticos que acompanham a sua adoção. A proposta equilibra sofisticação técnica com orientação comercial, oferecendo uma base sólida para a transformação digital orientada por dados.

5.3 Implicações práticas

A *framework* desenvolvida nesta investigação apresenta implicações práticas relevantes para empresas do setor do retalho alimentar que pretendem adotar abordagens de *marketing* mais inteligentes, personalizadas e orientadas por dados.

Ao integrar modelos de *clustering*, previsão de conversão, explicabilidade com *SHAP* e geração automatizada de mensagens personalizadas por LLMs, a *framework* constitui uma solução modular e adaptável, aplicável a diferentes perfis de retalhistas – desde cadeias de supermercados com maturidade digital avançada até pequenos operadores em processo de transformação digital.

Transformação na estratégia de marketing

A principal contribuição prática prende-se com a possibilidade de substituir abordagens generalistas e reativas por estratégias personalizadas, baseadas em padrões reais de comportamento do consumidor. A segmentação detalhada, associada a previsões de propensão e risco, permite antecipar necessidades, reduzir desperdício de campanhas e direcionar os investimentos para os canais e públicos com maior retorno esperado.

Eficiência operacional e automatização

A automação dos processos de segmentação, previsão e criação de conteúdo proporciona ganhos expressivos de tempo e eficiência. A capacidade de gerar milhares de mensagens adaptadas a diferentes segmentos, canais e tipos de campanha sem intervenção manual direta permite escalar operações de marketing sem necessidade de equipas adicionais.

Adaptação a cenários omnicanal

A interoperabilidade da *framework* com diferentes fontes de dados – incluindo aplicações móveis, lojas físicas e plataformas digitais – permite criar uma experiência omnicanal mais fluida e integrada. Esta abordagem assegura que o cliente recebe mensagens consistentes e relevantes, independentemente do ponto de contacto utilizado.

Suporte à tomada de decisão

Com *dashboards* e relatórios automatizados, a *framework* facilita a monitorização contínua dos KPIs mais relevantes (como taxa de conversão, ROI, CAC e LTV), promovendo uma cultura de decisão baseada em dados. Além disso, a explicabilidade dos modelos

permite que equipas não técnicas compreendam as razões por trás das previsões, aumentando a confiança e adoção por parte de *stakeholders*.

Preparação para a adoção real

Embora os resultados tenham sido obtidos com dados sintéticos, a arquitetura da *framework* foi concebida de forma a garantir transição simples para contextos reais. A modularidade do código, a documentação técnica e os exemplos simulados fornecem um ponto de partida claro para adaptar a solução a dados reais de empresas, respeitando sempre os requisitos legais e éticos associados ao tratamento de dados de clientes.

Potencial de inovação e diferenciação competitiva

A implementação desta *framework* pode representar uma vantagem competitiva significativa num setor onde a diferenciação pela experiência do cliente se torna cada vez mais estratégica. Ao combinar inteligência artificial com estratégias centradas no consumidor, os retalhistas podem não só melhorar os resultados comerciais como também consolidar a sua imagem de marca como inovadora e orientada para o cliente.

6 Conclusão geral

A presente dissertação teve como objetivo principal o desenvolvimento e validação de uma *framework* baseada em inteligência artificial para a personalização e otimização de campanhas omnicanal no retalho alimentar. Face aos desafios atuais do setor — como a saturação de conteúdos, a fragmentação de canais e a crescente exigência por experiências personalizadas — esta investigação procurou oferecer uma solução integrada, inteligente e operacionalizável.

A *framework* desenvolvida mostrou-se eficaz na articulação de múltiplos módulos: integração e tratamento de dados, segmentação de clientes via *clustering* comportamental, previsão de conversões com modelos explicáveis, geração de mensagens personalizadas com LLMs e análise de métricas chave como ROI e *engagement*. Apesar da utilização de dados sintéticos, os resultados obtidos foram consistentes com as expectativas teóricas e demonstraram o potencial da abordagem para transformar a forma como os retalhistas concebem e executam campanhas de *marketing*.

Para além do seu valor técnico, a *framework* oferece também uma abordagem acessível à implementação, ao incluir documentação detalhada, código modular e exemplos aplicados. O projeto reforça a importância da integração entre áreas como ciência de dados, *marketing* e design de experiências, propondo um novo paradigma centrado no comportamento real do consumidor, em tempo real e adaptado ao seu canal preferencial.

Do ponto de vista académico, o trabalho contribui para a consolidação de modelos conceptuais de personalização omnicanal com base em IA, para a validação prática de algoritmos de *clustering* e modelos preditivos, e para a exploração de LLMs como geradores de conteúdo dinâmico e relevante. Do ponto de vista prático, oferece aos profissionais do setor ferramentas concretas para melhorar a eficácia das suas campanhas e aumentar o retorno sobre investimento.

No entanto, como discutido, os resultados obtidos devem ser interpretados com cautela devido à natureza sintética dos dados utilizados. A validação definitiva da *framework* requer a sua aplicação em ambientes reais de retalho, com todas as variáveis comportamentais, tecnológicas e organizacionais inerentes.

Ainda assim, esta investigação estabelece uma base sólida para aplicações futuras, sinalizando um caminho claro para a evolução do *marketing* omnicanal com base em dados, inteligência e personalização real. O futuro do retalho passa por conhecer melhor cada cliente, antecipar as suas necessidades e entregar valor de forma oportuna e relevante — e esta *framework* é um contributo concreto para esse futuro.

6.1 Contribuições teóricas e práticas

Esta investigação contribui significativamente para o avanço do conhecimento teórico e para a aplicação prática de estratégias de personalização e otimização de campanhas omnicanal com recurso a inteligência artificial no retalho alimentar.

6.1.1 Contribuições teóricas

Do ponto de vista teórico, esta dissertação avança o estado da arte em várias dimensões:

1. **Integração de paradigmas**

A *framework* proposta conjuga conceitos de *marketing* omnicanal, ciência de dados e inteligência artificial num modelo unificado, promovendo uma abordagem interdisciplinar que contribui para uma compreensão mais holística da personalização no contexto digital.

2. **Extensão da teoria de segmentação**

Ao utilizar algoritmos de *clustering* para detetar padrões comportamentais complexos, este trabalho expande as abordagens tradicionais de segmentação baseadas em critérios demográficos ou geográficos, propondo um novo paradigma assente em dados de comportamento real e preferências reveladas.

3. **Modelo conceptual de personalização omnicanal**

A *framework* fornece um modelo conceptual completo para implementação de personalização em ambientes omnicanal, clarificando os componentes essenciais, os fluxos de dados e as interações entre módulos analíticos e operacionais.

4. **Validação empírica de técnicas de IA**

Através da aplicação prática de técnicas como *clustering*, modelos preditivos e LLMs, o estudo fornece evidência empírica da sua viabilidade e valor no contexto específico do retalho alimentar.

5. **Métricas e avaliação**

A definição de métricas específicas para análise de ROI, taxa de conversão, custo por aquisição e *engagement* contribui para a literatura metodológica, oferecendo instrumentos para futuras investigações nesta área.

6.1.2 Contribuições práticas

Do ponto de vista prático, esta investigação oferece ferramentas diretamente aplicáveis por retalhistas:

1. **Framework operacionalizável**

A *framework* foi desenhada para ser modular, escalável e adaptável, permitindo a sua adoção progressiva por empresas com diferentes níveis de maturidade tecnológica.

2. **Guias de implementação**

São fornecidas orientações práticas para todas as fases, desde a preparação de dados até à geração de campanhas personalizadas, reduzindo as barreiras técnicas à adoção de inteligência artificial.

3. **Casos de uso relevantes**

A identificação de cenários com alto impacto e viabilidade ajuda os retalhistas a priorizar iniciativas com maior retorno esperado, como campanhas de reativação ou estratégias de retenção.

4. **Métricas de negócio**

As métricas utilizadas permitem avaliar de forma objetiva o impacto da personalização, fornecendo aos decisores argumentos quantitativos para justificar investimentos em tecnologia.

5. **Diretrizes éticas e de privacidade**

A inclusão de considerações éticas e regulatórias torna a *framework* mais alinhada com os desafios atuais da proteção de dados, promovendo uma adoção responsável da inteligência artificial.

6.2 Limitações do estudo

Apesar dos avanços alcançados, esta investigação apresenta limitações que devem ser consideradas ao interpretar os resultados e ao aplicar a *framework* em contextos reais.

6.2.1 Limitações metodológicas

A metodologia adotada visou validar tecnicamente a viabilidade da *framework* proposta. No entanto, foram necessárias simplificações e suposições que afetam a generalização dos resultados. Destacam-se as seguintes limitações.

- **Utilização de dados sintéticos**

A análise foi realizada com dados sintéticos, gerados com base em padrões observados na literatura (Chapman Pete et al., 1999; Deloitte, 2023). Embora úteis para validar a arquitetura da *framework*, estes dados não capturam a totalidade da complexidade comportamental dos consumidores reais.

- **Validação em ambiente controlado**

A *framework* foi testada em ambiente simulado, não estando sujeita a fatores externos como sazonalidade, campanhas concorrentes ou alterações económicas, o que pode limitar a sua robustez fora do ambiente de teste.

- **Horizonte temporal reduzido**

A avaliação foi realizada num intervalo de tempo limitado, o que impede conclusões sobre efeitos acumulados ou a evolução de estratégias ao longo de múltiplos ciclos de campanha.

- **Enfoque quantitativo**

O estudo priorizou métricas quantitativas, não incluindo avaliações subjetivas como a perceção do consumidor sobre privacidade, confiança ou valor percebido.

6.2.2 Limitações práticas

Do ponto de vista da aplicação no terreno, a *framework* apresenta os seguintes desafios práticos:

- **Requisitos técnicos elevados**

A aplicação da *framework* exige competências avançadas em ciência de dados e desenvolvimento, o que pode limitar a sua adoção por empresas com recursos técnicos reduzidos.

- **Adaptação contínua às preferências do consumidor**

A *framework* precisa de ser ajustada regularmente para se manter alinhada com mudanças nas preferências, algo que exige um processo operacional maduro e ágil.

- **Dependência de infraestrutura tecnológica**

A sua eficácia está condicionada à existência de sistemas capazes de integrar dados de diferentes fontes, o que pode ser um obstáculo em contextos com infraestruturas legadas.

- **Equilíbrio entre personalização e privacidade**

Encontrar o ponto ideal entre campanhas personalizadas e respeito pela privacidade dos utilizadores é um desafio contínuo, especialmente com o aumento das exigências regulatórias.

- **Qualidade dos dados**

A eficácia dos modelos depende fortemente da qualidade dos dados disponíveis. Dados incompletos, desatualizados ou mal estruturados podem comprometer os resultados.

6.3 Sugestões para investigação futura

Esta dissertação abre várias possibilidades para investigações futuras que possam aprofundar e expandir a *framework* proposta:

Integração com dispositivos IOT e tecnologias inteligentes

- Explorar sensores em loja para recolher dados de movimento e interação.
- Utilizar dispositivos *wearable* para personalização em tempo real.
- Estudar embalagens inteligentes com funcionalidades de interação digital.

- Integrar assistentes domésticos na jornada de compra alimentar.

Antecipação de tendências digitais

- Monitorizar plataformas emergentes e adaptação ao seu crescimento.
- Acompanhar a evolução dos algoritmos de redes sociais e formatos de conteúdo.
- Explorar realidade aumentada/virtual em experiências de compra.
- Integrar canais conversacionais baseados em modelos de linguagem.

Novas formas de personalização

- Personalização colaborativa, com participação ativa do consumidor.
- Personalização contextual que inclui estado emocional e ambiente.
- Estratégias éticas e transparentes que reforcem a confiança.
- Personalização entre marcas, promovendo ecossistemas integrados.

Validação em contextos reais

- Conduzir estudos de caso em retalhistas reais para avaliar impacto a longo prazo.
- Comparar resultados em diferentes geografias e formatos de negócio.
- Medir impacto em perceções subjetivas como satisfação e lealdade.
- Avaliar custos reais de implementação e retorno económico efetivo.

7 Bibliografia

- Barwitz, N., & Maas, P. (2018). Understanding the Omnichannel Customer Journey: Determinants of Interaction Choice. *Journal of Interactive Marketing*, 43, 116–133. <https://doi.org/10.1016/j.intmar.2018.02.001>
- Chapman Pete, Clinton Julian, KErber Randy, Khabaza Thomas, Reinartz Thomas, Shearer Colin, & Wirth Rudiger. (1999). *Manual Crisp-DM 1-CRISP DM*.
- Cui, X., Xie, Q., Zhu, J., Shareef, M. A., Goraya, M. A. S., & Akram, M. S. (2022). Understanding the omnichannel customer journey: The effect of online and offline channel interactivity on consumer value co-creation behavior. *Journal of Retailing and Consumer Services*, 65, 102869. <https://doi.org/10.1016/J.JRETCONSER.2021.102869>
- Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42. <https://doi.org/10.1007/s11747-019-00696-0>

- Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. In *Journal of Marketing Research* (Vol. 42, Issue 4, pp. 415–430). American Marketing Association. <https://doi.org/10.1509/jmkr.2005.42.4.415>
- Gal Gitter, M. R. K. R. and J. W. (2020). *The end of shopping's boundaries: Omnichannel personalization*.
- Gao, W., Li, W., Fan, H., & Jia, X. (2021). How customer experience incongruence affects omnichannel customer retention: The moderating role of channel characteristics. *Journal of Retailing and Consumer Services*, 60, 102487. <https://doi.org/10.1016/J.JRETCONSER.2021.102487>
- Grewal, D., Roggeveen, A. L., & Nordfält, J. (2017). The Future of Retailing. *Journal of Retailing*, 93(1), 1–6. <https://doi.org/10.1016/j.jretai.2016.12.008>
- Hagberg, J., Sundstrom, M., & Egels-Zandén, N. (2016). The digitalization of retailing: an exploratory framework. *International Journal of Retail and Distribution Management*, 44(7), 694–712. <https://doi.org/10.1108/IJRDM-09-2015-0140>
- Hauser, M., Flath, C. M., & Thiesse, F. (2021). Catch me if you scan: Data-driven prescriptive modeling for smart store environments. *European Journal of Operational Research*, 294(3), 860–873. <https://doi.org/10.1016/j.ejor.2020.12.047>
- Juaneda-Ayensa, E., Mosquera, A., & Murillo, Y. S. (2016). Omnichannel customer behavior: Key drivers of technology acceptance and use and their effects on purchase intention. *Frontiers in Psychology*, 7(JUL). <https://doi.org/10.3389/fpsyg.2016.01117>
- Lemmens, A., Croux, C., & Leuven, K. U. (2006). *BAGGING AND BOOSTING CLASSIFICATION TREES TO PREDICT CHURN*.
- Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, 80(6), 69–96. <https://doi.org/10.1509/jm.15.0420>
- Melero, I., Javier Sese, F., & Verhoef, P. C. (2016). Redefiniendo la experiencia del cliente en el entorno omnicanal. *Universia Business Review*, 2016(50), 18–37. <https://doi.org/10.3232/UBR.2016.V13.N2.01>
- Pelau, C., Ene, I., & Pop, M. I. (2021). The Impact of Artificial Intelligence on Consumers' Identity and Human Skills. *Amfiteatru Economic*, 23(56), 33–33. <https://doi.org/10.24818/EA/2021/56/33>
- Sovrano, F., Vitali, F., & Palmirani, M. (2021). *Modelling GDPR-Compliant Explanations for Trustworthy AI*. https://doi.org/10.1007/978-3-030-58957-8_16
- Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From Multi-Channel Retailing to Omnichannel Retailing. Introduction to the Special Issue on Multi-Channel Retailing. *Journal of Retailing*, 91(2), 174–181. <https://doi.org/10.1016/j.jretai.2015.02.005>
- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

8 Anexos

Anexo 1 - Resultados do relatório final de segmentação e predição

Tabela A1 1 Amostra de dados

ID_Cliente	Dias_ate_Compra	Valor_Interacao	Cluster	Nome_Segmento	Conversao
C3904	4	0.0	2	Cliente_Premium	0
C38	10	0.0	1	Engajado_Lento	0
C3866	2	0.0	0	Interacao_Baixa	0
C3246	9	0.0	1	Engajado_Lento	0
C4125	4	0.0	2	Cliente_Premium	0
C1554	7	0.0	2	Cliente_Premium	0
C150	10	0.0	1	Engajado_Lento	0
C1983	5	0.0	2	Cliente_Premium	0
C2090	9	0.0	1	Engajado_Lento	0
C2062	3	0.0	0	Interacao_Baixa	0

Tabela A1 2 Desempenho por segmento

Nome_Segmento	Total_Clientes	Total_Conversoes	Taxa_Conversao
Cliente_Premium	6905	379	5.49
Engajado_Lento	5147	289	5.61
Indeciso	958	958	100.00
Interacao_Baixa	6990	384	5.49

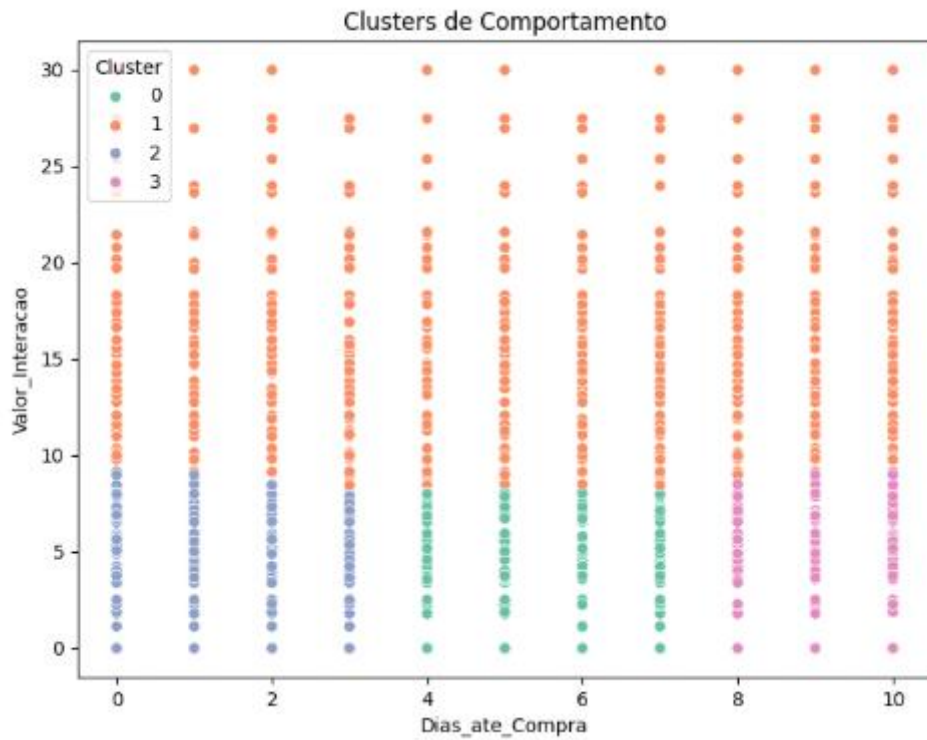


Figura A1 1 Visualização dos Clusters

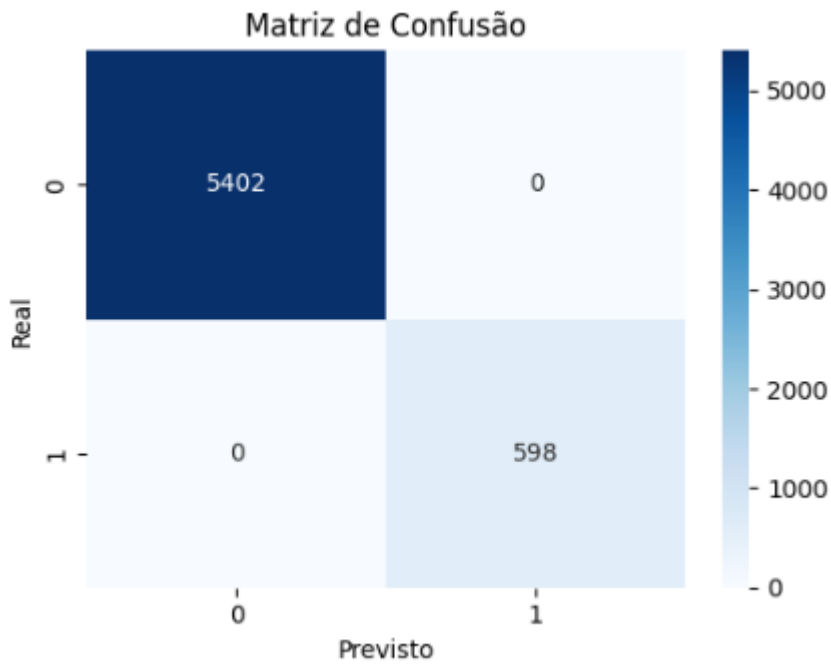


Figura A1 2 Matriz de Confusão do modelo preditivo (Random Forest)

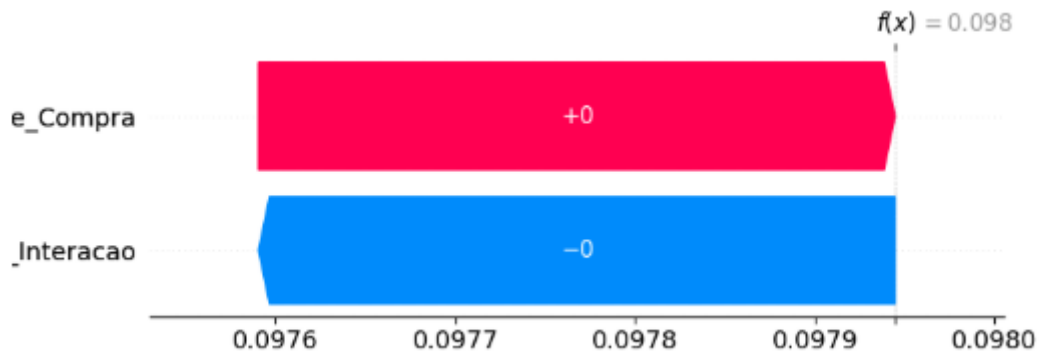


Figura A1 3 Gráfico SHAP Individual (Waterfall)

Tabela A1 3 Mensagens personalizadas geradas por LMM

ID_Cliente	Nome_Segmento	Canal	Mensagem_LLM
C3904	Inativo	Loja	Descubra as nossas novidades nesta campanha!
C38	Inativo	Email	Descubra as nossas novidades nesta campanha!
C3866	Leal	Email	Descubra as nossas novidades nesta campanha!
C3246	Inativo	Email	Descubra as nossas novidades nesta campanha!
C4125	VIP	Push	Olá! Aproveite os nossos produtos premium com descontos exclusivos nesta campanha especial.
C1554	Inativo	Email	Descubra as nossas novidades nesta campanha!
C150	Inativo	Loja	Descubra as nossas novidades nesta campanha!
C1983	VIP	Push	Olá! Aproveite os nossos produtos premium com descontos exclusivos nesta campanha especial.
C2090	Inativo	Email	Descubra as nossas novidades nesta campanha!
C2062	Inativo	Email	Descubra as nossas novidades nesta campanha!

Tabela A1 4 Recomendações de Otimização de campanhas

Recomendações:

Recomendacao
Segmento 'Indeciso' no canal 'Loja' teve o maior ROI estimado (-0.99x)
Taxa de conversao: 100.0% Custo medio por conversao: 500.0 EUR
Sugerimos reforcar campanhas 'Loja' dirigidas ao segmento 'Indeciso'

Tabela A1 5 Resultados por segmentos x canal

Nome_Segmento	Canal	Conversoes	Total	Valor	Custo	Taxa_Conversao	Custo_Medio	ROI_Estimado
Cliente_Premium	App	70	1372	334.33	4381240.44	5.10	500.0	-1.00
Cliente_Premium	Email	97	1842	462.00	4948027.70	5.27	500.0	-1.00
Cliente_Premium	Loja	111	1847	538.21	4482977.80	6.01	500.0	-1.00
Cliente_Premium	Push	82	1374	385.15	2714950.46	5.97	500.0	-1.00
Cliente_Premium	SMS	19	470	112.01	1383069.00	4.04	500.0	-1.00
Engajado_Lento	App	61	1067	302.85	3324046.93	5.72	500.0	-1.00
Engajado_Lento	Email	77	1361	434.90	3617724.71	5.66	500.0	-1.00
Engajado_Lento	Loja	73	1343	403.30	3241910.91	5.44	500.0	-1.00
Engajado_Lento	Push	57	1036	291.07	2064706.14	5.50	500.0	-1.00
Engajado_Lento	SMS	21	340	83.93	1000518.00	6.18	500.0	-1.00
Indeciso	App	183	183	3097.62	585151.67	100.00	500.0	-0.99
Indeciso	Email	281	281	4765.81	753650.46	100.00	500.0	-0.99
Indeciso	Loja	240	240	4042.99	562780.77	100.00	500.0	-0.99
Indeciso	Push	196	196	3317.86	374535.14	100.00	500.0	-0.99
Indeciso	SMS	58	58	962.53	170676.60	100.00	500.0	-0.99
Interacao_Baixa	App	71	1425	358.17	4521434.21	4.98	500.0	-1.00
Interacao_Baixa	Email	89	1857	439.72	4949322.39	4.79	500.0	-1.00
Interacao_Baixa	Loja	103	1892	540.26	4651369.57	5.44	500.0	-1.00
Interacao_Baixa	Push	89	1365	421.70	2771134.40	6.52	500.0	-1.00
Interacao_Baixa	SMS	32	451	166.08	1327157.70	7.10	500.0	-1.00

Tabela A1 6 Resultados por canal

Canal	Conversoes	Total	Valor	Custo	Taxa_Conversao	Custo_Medio	ROI_Estimado
App	385	4047	4092.97	12811873.25	9.51	500.0	-1.0
Email	544	5341	6102.43	14268725.26	10.19	500.0	-1.0
Loja	527	5322	5524.76	12939039.05	9.90	500.0	-1.0
Push	424	3971	4415.78	7925326.14	10.68	500.0	-1.0
SMS	130	1319	1324.55	3881421.30	9.86	500.0	-1.0

Tabela A1 7 Métricas de Engagemnt por canal

Canal	Clique	Compra	Ignore	Visualizacao
App	1248	385	780	1634
Email	1578	544	1081	2138
Loja	1536	527	1098	2161
Push	1133	424	781	1633
SMS	387	130	269	533

Anexo 2- Folhetos

Olá Cliente C4818,

Imaginou um mundo onde os seus produtos favoritos vêm com descontos incríveis?
Bem, não precisa de imaginar mais, porque tornámos isso uma realidade para si!

Sabemos o quanto adora os nossos deliciosos iogurtes, snacks irresistíveis e sumos refrescantes. Por isso, é com grande prazer que lhe oferecemos um desconto exclusivo de 20%



Figura A2 1 Exemplo de folheto gerado com base nos gostos do cliente

Anexo 3- Segmentação de clientes (Resultados do clustering)

Os dados utilizados na framework foram gerados artificialmente com o objetivo de simular contextos reais de comportamento do consumidor em ambiente omnicanal. A geração dos dados foi realizada através de um script Python que criou ficheiros CSV representando clientes, produtos, campanhas, interações e segmentos comportamentais.

- Variáveis categóricas (como género, canal de interação, tipo de campanha e localização) foram atribuídas com base em distribuições uniformes ou aleatórias ponderadas.
- Datas como Data_Nascimento, Data_Cadastro e Data_Interacao foram geradas com base em intervalos plausíveis. Por exemplo, as datas de nascimento variam entre 1955 e 2005, enquanto as interações decorrem nos primeiros dias após o início de cada campanha.
- Variáveis contínuas, como o preço dos produtos ou o valor de interação, foram geradas com base em distribuições uniformes dentro de um intervalo (ex.: Preço_Produto entre 1,00€ e 10,00€).

- Conversão e dias até à compra foram modelados com lógica condicional: apenas interações do tipo “Compra” receberam um valor de interação positivo e uma variável Dias_ate_Compra, calculada com base na diferença entre a data da campanha e a data da compra.

A geração de interações seguiu uma lógica probabilística, com pesos atribuídos aos diferentes tipos de interação: Visualização (40%), Clique (30%), Ignore (20%) e Compra (10%), de modo a simular taxas de conversão realistas.

A tabela abaixo resume as principais decisões de geração de dados:

Tabela A3 *Tipos de variáveis. Métodos de geração e distribuições utilizadas nos dados sintéticos*

Variável	Tipo	Método de geração/distribuição
Genero	categórica	Uniforme entre “Masculino” e “Feminino”
Localização	categórica	Aleatório entre 5 cidades
Preco_Produto	contínua	Uniforme entre 1 e 10
Valor_Interação	contínua	Zero exceto em compras (valor aleatório)
Dias_ate_Compra	ordinal	Inteiro entre 0 e 10 (só em interações do tipo “compra”)
Tipo_Interacao	categórica	Distribuição ponderada: 40% visualização, etc.
Conversão	binária	1 se tipo = Compra; = caso contrário

Anexo 4- Estatísticas Clientes.csv

Tabela A4 1 *Estatísticas descritivas das variáveis do ficheiro clientes.csv, com exemplos dos valores únicos mais frequentes*

Variavel	Tipo	Media	Desvio Padrao	Minimo	Maximo	Moda	Valores unicos
ID_Cliente	Categórica	-	-	-	-	C1	C1, C2, C3, C4, C5...
Nome_Cliente	Categórica	-	-	-	-	Cliente_1	Cliente_1, Cliente_2, Cliente_3, Cliente_4, Cliente_5...
Email	Categórica	-	-	-	-	cliente1000@exemplo.com	cliente1@exemplo.com, cliente2@exemplo.com, cliente3@exemplo.com, cliente4@exemplo.com, cliente5@exemplo.com...

Telefone	Númerica	955134 569.35	260328 10.63	91000 1790	99999 0682	-	-
Data_Nascimento	Categórica	-	-	-	-	09/06/1956	1982-02-03, 1965-08-04, 2017-07-25, 1985-03-19, 2021-05-13...
Genero	Categórica	-	-	-	-	Masculino	Feminino, Masculino
Localizacao	Categórica	-	-	-	-	Coimbra	Faro, Coimbra, Porto, Braga, Lisboa
Data_Cadastro	Categórica	-	-	-	-	17/02/2022	2020-09-03, 2021-10-20, 2022-01-17, 2021-04-06, 2020-01-11...
Fonte_Cadastro	Categórica	-	-	-	-	Website	Website, App, Loja

Anexo 5- Estatística Campanhas.csv

Tabela A5.1 Estatísticas descritivas das variáveis presentes no ficheiro campanhas.csv, incluindo canais e datas de início/fim

Variavel	Tipo	Mediana	Desvio Padrão	Mínimo	Máximo	Moda	Valores unicos
ID_Campanha	Categórico					M1	M1,M2,M3,M4, M5...
Nome_Campanha	Categórico					Campanha_1	Campanha_1, Campanha_2, Campanha_3, Campanha_4...
Descricao_Campanha	Categórico					Descrição detalhada da campanha_1	2023-02-26, 2023-02-24, 2023-02-15, 2023-04-06, 2023-04-09...
Data_Inicio	Categórico					26/02/2023	2023-09-15, 2023-10-09, 2023-08-19, 2023-08-15, 2023-09-10...
Data_Fim	Categórico					29/07/2023	App, Email, SMS, Push, Loja
Canal	Categórico					Push	Fidelizacao, Sazonal, Lancamento, Desconto

Tipo_Campanha	Categoria					Fidelizaca o	VIP, Inativo, Ocasional, Leal
Tipo_Campanha	Categoria					VIP	
Orcamento	Numerico	2303. 77	1265.9 6	531.7 5	4471. 51		
Custo_Real	Numerico	3027. 8	1024. 32	1768. 42	4736. 4		
Mensagem_Pad rao	Categoria					Aproveite o desconto especial desta semana!	Cliente fiel merece vantagens unicas!, Aproveite o desconto especial desta semana!, Promo de fim de semana: nao perca!, Descubra o novo produto com oferta exclusiva!
ID_Produto	Categoria					P13	P29, P6, P13, P5, P21...
ID_Segmento	Numerico	2.87	1.06	1	4		

Anexo 6 – Estatística produtos.csv

Tabela A6 1 Estatísticas descritivas dos produtos simulados no ficheiro produtos.csv, incluindo categorias e nomes

Variavel	Tipo	Médi a	Desvi o padra o	Minim o	Maxim o	Moda	Valores únicos
ID_Produto	Categoria					P1	P1, P2, P3, P4...
Nome_Produto	Categoria					Produto_ 1	Produto_1 , Produto_2 , Produto_3 ...
Descricao_Prod uto	Categoria					Descrica o do produto 1	Descricao do produto 1, Descricao do produto 3, descricao do

							produto 3...
Categoria_Produto	Categoria					Lactínicos	Bebidas, snacks, Lactínios, Cereais
Preco_Produto	Numerica	5.17	2.63	1.13	9.97		
Unidade_Medida	Categoria					Kg	Kg, unidade, litro

Anexo 7 – Estatística Interacoes.csv

Tabela A7.1 Estatísticas descritivas das interações registadas no ficheiro interações.csv, associadas a clientes e campanhas

Variavel	Tipo	Media	Desvio padrao	Minimo	Maximo	Moda	Valores únicos
ID_Interacao	Categoria					I1	I1, I2, I3, I4, I5...
ID_Cliente	Categoria					C714	C828, C329, C2997, C2274, C3560...
ID_Campanha	Categoria					M1	M3, M9, M1, M7, M8...
Data_Interacao	Categoria						2023-03-08, 2023-01-30, 2023-03-01, 2023-05-26, 2023-02-15...
Tipo_Interacao	Categoria					Visualizacao	Visualizacao, Clique, Compra, Ignore
Canal_interacao	Categoria					SMS	SMS, Email, App, Push, Loja
Dispositivo_Interacao	Categoria					Desktop	Mobile, Tablet, Desktop
Valor_interacao	Numerica						

Resposta_LLM	Categori ca					Mensagem personaliz ada para clientes_1 097 sobre campanha _!4	Mensagem personaliza da para Cliente_403 8 sobre Campanha_ 4., Mensagem personaliza da para Cliente_60 sobre Campanha_ 15., Mensagem personaliza da para Cliente_199 4 sobre Campanha_ 10., Mensagem personaliza da para Cliente_323 6 sobre Campanha_ 15., Mensagem personaliza da para Cliente_383 8 sobre Campanha_ 6....
Conversao	Numeri ca						
Dias_ate_compr a	Numeri ca	4.83	3.21	0.0	10.0		

Anexo 8 – Estatística Interacoes_clientes_campanhas.csv

Tabela A8 11 Estatísticas descritivas das interações registadas no ficheiro interacoes_clientes_campanhas.csv, associadas a interações entre clientes e campanhas

ID_Inter acao	ID_Cli ente	ID_Camp anha	Data_inte racao	Tipo_inter acao	Cana_inte racao	Dispos itivo	Valor_Inte racao	Respo sta	Conver sao	Dias_ate_c ompra
I1	C828	M3	08/03/2023	Visualizac ao	SMS	Mobile	0.0		0	
I2	C329	M9	30/01/2023	Visualizac ao	Email	Tablet	0.0		0	
I3	C2997	M1	01/03/2023	Clique	APP	Deskto p	0.0		0	
I4	C2274	M7	26/05/2023	Clique	APP	Deskto p	0.0		0	
I5	C3560	M8	15/02/2023	Clique	Push	Mobile	0.0		0	

16	C1118	M8	18/02/2023	Visualiza cao	Push	Deskto p	0.0		0	
17	C840	M6	10/04/2023	Clique	App	Mobile	0.0		0	
18	C4038	M4	24/02/2023	Compra	Push	Tablet	14.06	Mensa gem	1	9.0
19	C1287	M5	09/04/2023	Clique	SMS	Tablet	0.0		0	
110	C521	M6	16/04/2023	Visualiza cao	APP	Tablet	0.0		0	
111	C607	M14	22/06/2023	Clique	Push	Tablet	0.0		0	
112	C1951	M13	28/06/2023	Visualiza cao	SMS	Mobile	0.0		0	
113	C60	M15	24/03/2023	Compra	Loja	Tablet	29.91	Mensa gem	1	2.0
114	C3112	M12	09/03/2023	Visualiza cao	Loja	Mobile	0.0		0	

Anexo 9 – Avaliação comparativa de F1-score médio

```

PROBLEMS 7 OUTPUT DEBUG CONSOLE TERMINAL PORTS
PS C:\Users\Fjcs\Desktop\Dissertacao\framework> python core/explicar_conversao.py

Distribuição da variável 'Conversao':
Conversao
0    0.98195
1    0.01805
Name: proportion, dtype: float64

A avaliar múltiplos valores de k para validação cruzada:
k=3 → Média F1: 1.0 | Scores individuais: [np.float64(1.0), np.float64(1.0), np.float64(1.0)]
k=5 → Média F1: 1.0 | Scores individuais: [np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0)]
k=10 → Média F1: 1.0 | Scores individuais: [np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0), np.float64(1.0)]
Gráfico comparativo guardado em data/image/comparacao_f1score_k.png

✅ Teste de k concluído. Escolhe o valor de k com melhor desempenho médio e ajusta no script final.

```

Figura A9 1 Avaliação comparativa do F1-score médio para diferentes valores de k na validação cruzada estratificada