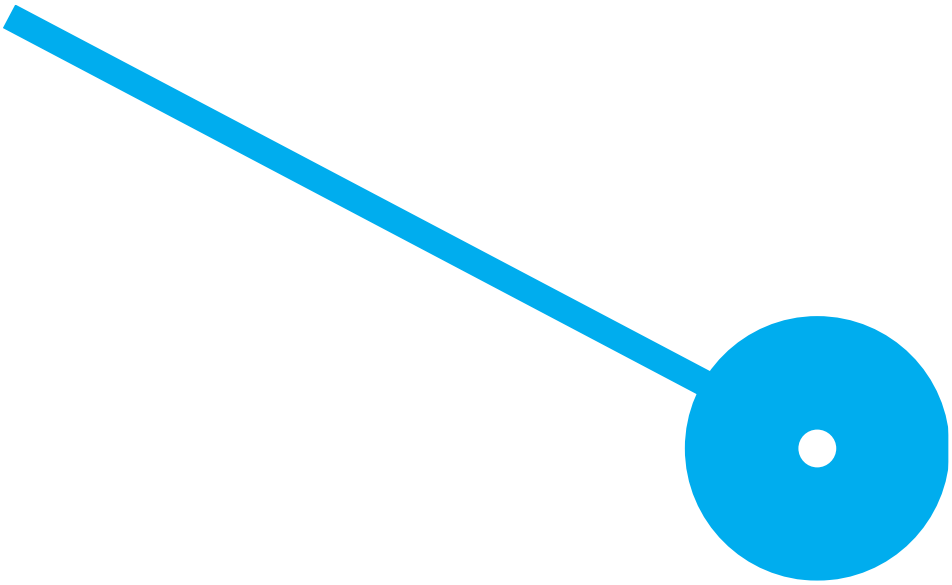
ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

MESTRADO
Engenharia Informática

Análise do processo de *checkout* e carrinhos abandonados

Pedro Miguel Carneiro Silva

07/2024



ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

MESTRADO
Engenharia Informática

Análise do processo de *checkout* e carrinhos abandonados

Pedro Miguel Carneiro Silva
8180239

Orientador

Professor Doutor Davide Carneiro

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.

07/2024

Declaração de Integridade

Eu, Pedro Miguel Carneiro Silva, estudante nº 8180239, do Mestrado de 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 "Análise do processo de *checkout* e carrinhos abandonados de uma loja virtual" 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.

Agradecimentos

Após um ano repleto de desafios e momentos altos e baixos, quero agradecer a todos aqueles que estiveram constantemente ao meu lado e contribuíram para tornar esta jornada mais fácil. Compartilharam comigo experiências, histórias, risos e, acima de tudo, possibilitaram o meu crescimento tanto pessoal quanto profissional. Portanto, este espaço é reservado a todos vocês.

Em primeiro lugar, começo por agradecer ao meu orientador, Professor Doutor Davide Carneiro, por me ter dado a oportunidade de realizar este projeto sob a sua orientação, por todo o apoio prestado durante o processo de pesquisa e desenvolvimento deste projeto e por todo o conhecimento e ensinamentos partilhados não só ao longo deste ano, como nos 3 anos de Licenciatura e 2 de Mestrado.

Um especial agradecimento a todos os colegas de trabalho, principalmente ao Bruno Ribeiro e ao André Gouveia. Bruno, agradeço por me teres dado a oportunidade de realizar este projeto, pelas orientações que me deste e por toda a ajuda. André, obrigado pela paciência e disponibilidade que demonstraste comigo. A tua experiência e conhecimento foram muito importantes para o sucesso alcançado neste projeto.

Não posso deixar de agradecer a todos os meus colegas universitários, em especial ao Carlos Sousa e ao Pedro Lopes pelas discussões e troca de ideias ao longo destes 5 anos académicos. Além de colegas de turma, tornaram-se bons amigos e o vosso apoio e as experiências partilhadas tornaram esta caminhada acadêmica mais fácil e divertida.

Para os meus pais, obrigado por estarem presentes em todos os momentos, por serem os meus pilares, por me ajudarem a tomar as melhores decisões e, principalmente, por me ajudarem a tornar na pessoa que sou hoje. O fim desta etapa também se deve a vocês.

Ao meu irmão Guilherme, agradeço por partilhar a vida contigo. Sei que a vida ainda te vai sorrir muito, espero estar ao teu lado a acompanhar-te e a ver-te alcançar coisas muito boas. O teu sucesso será sempre o meu sucesso também!

Por último, agradeço à minha namorada Beatriz. A tua ajuda e palavras certas no momento certo, foram a chave principal para terminar este projeto. Obrigado por todo o apoio, por me incentivares a não desistir nunca, por seres o meu "braço direito" e, acima de tudo, obrigado por estares sempre comigo. Nada disto seria possível sem ti ao meu lado!

A todos os mencionados e àqueles que de alguma forma contribuíram para este trabalho, o meu profundo agradecimento. As vossas contribuições foram inestimáveis e sou imensamente grato pela vossa ajuda.

Resumo

Nos últimos anos, temos presenciado um aumento significativo na quantidade e qualidade dos dados nas organizações. Isso tem levado as empresas a adaptarem-se e a aproveitarem ao máximo esses dados. Neste projeto, o objetivo é focar no processo de compra *online* de uma empresa nacional, onde os clientes podem comprar produtos por meio de uma loja virtual. Além disso, aborda-se o problema das sessões abandonadas, procurando entender o motivo de ocorrer e encontrar formas de converter essas sessões em vendas lucrativas para a empresa.

O principal objetivo é analisar os dados do processo de compra e das sessões abandonadas, a fim de melhorar a experiência do cliente, otimizar os fluxos de compra e aumentar o lucro. Ao compreender as razões por trás das sessões abandonadas e desenvolver estratégias eficazes para reverter essa situação, espera-se aumentar as taxas de conversão e fortalecer o relacionamento com os clientes, tornando a empresa mais competitiva no mercado.

Em resumo, este projeto visa aproveitar os dados disponíveis no processo de compra *online* de uma empresa nacional, com o objetivo de entender e converter as sessões abandonadas em vendas concretas. Isso impulsionará o sucesso financeiro da empresa e fortalecerá a sua posição num mercado que é competitivo.

Palavras-chave: processo de compra *online;* sessões abandonadas; vendas; *machine learning*

Abstract

In recent years, we have witnessed a significant increase in the quantity and quality of data within organizations. This has led companies to adapt and make the most of these data. In this project, the goal is to focus on the online purchasing process of a national company, where customers can purchase products through a virtual store. Additionally, we address the issue of abandoned sessions, seeking to understand the reasons for their occurrence and find ways to convert these sessions into profitable sales for the company.

The main objective, is to analyze the data from the purchasing process and abandoned sessions to enhance the customer experience, optimize purchasing flows, and increase profitability. By understanding the reasons behind abandoned sessions and developing effective strategies to reverse this situation, we aim to boost conversion rates and strengthen customer relationships, making the company more competitive in the market.

In summary, this project aims to leverage the data available in the online purchasing process of a national company, with the goal of understanding and converting abandoned sessions into concrete sales. This will drive the financial success of the company and strengthen its position in the competitive market.

Keywords: online purchasing process; abandoned sessions; sales; machine learning

Índice

Lista de quadros

Lista de figuras

Lista de abreviaturas

IA – Inteligência Artificial

ML – *Machine Learning*

DL – *Deep Learning*

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

RF – *Random Forest*

RLI – Regressão Linear

RLO – Regressão Logística

XGB – *Extreme Gradient Boosting*

LGBM –*Light Gradient-Boosting Machine*

NB – Naïve Bayes

CB – CatBoost

SMOTE - *Synthetic Minority Over-sampling Technique*

TPR – *True Positive Rate*

FPR – *False Positive Rate*

FNR – *False Negative Rate*

TNR – *True Negative Rat*

1. Introdução

Nos últimos anos, temos testemunhado um crescimento exponencial do comércio eletrónico, impulsionado pela evolução da tecnologia e pelas mudanças nos hábitos de consumo das pessoas. Cada vez mais, indivíduos optam por realizar as suas compras por meio de lojas virtuais, em vez de se deslocarem até estabelecimentos físicos. Essa transformação no comportamento de compra é resultado de diversos fatores.

Em primeiro lugar, a facilidade e conveniência proporcionadas pelo *e-commerce* são atrativos irresistíveis. Com apenas alguns cliques, os consumidores podem explorar uma vasta gama de produtos e serviços, comparar preços, ler avaliações e realizar transações financeiras seguras, tudo isso sem sair de casa. Essa comodidade permite que as pessoas economizem tempo e esforço, especialmente num meio de um estilo de vida agitado e cheio de obrigações.

Além disso, a variedade de opções disponíveis no comércio eletrónico é outro fator que contribui para a sua crescente popularidade. Os consumidores têm acesso a um amplo mercado global, podendo escolher entre produtos de diferentes marcas, tamanhos, modelos e preços. Isso amplia as suas possibilidades de encontrar exatamente o que desejam, atendendo às suas necessidades e preferências específicas.

Outro aspeto importante é a evolução das plataformas de *e-commerce*, que têm investido em melhorias significativas para garantir uma experiência de compra mais agradável e segura. O desenvolvimento de *interfaces* intuitivas, recursos de procura aprimorados, sistemas de recomendação personalizados e opções de pagamento diversificadas são apenas algumas das melhorias que tornam a compra *online* mais atraente para os consumidores.

Além disso, a pandemia de COVID-19 teve um impacto significativo na aceleração do comércio eletrónico. As restrições de distanciamento social e os bloqueios impostos em muitas regiões fizeram com que as pessoas recorressem ainda mais ao *e-commerce* para suprir as suas necessidades básicas e satisfazer os seus desejos. Esse aumento,

impulsionou ainda mais o setor e incentivou as empresas a investirem em infraestrutura digital e logística para atender às necessidades dos consumidores.

No entanto, apesar de todos esses benefícios, é importante destacar que o comércio eletrónico também apresenta desafios. Uma das principais desvantagens do comércio eletrónico em comparação às lojas físicas é a falta de interação física com os produtos antes da compra.

Nas lojas *online*, os consumidores não têm a oportunidade de ver, tocar ou experimentar os produtos antes de efetuar a compra. Essa limitação pode levar a uma menor confiança por parte do utilizador. A experiência presencial numa loja física permite que os clientes examinem os produtos de perto, sintam a sua textura, testem a sua funcionalidade e avaliem a sua qualidade antes de tomar uma decisão de compra. Essa interação direta com os produtos proporciona uma sensação de segurança e confiança, pois os consumidores têm a certeza de que estão adquirindo algo que atende às suas expectativas.

Por outro lado, nas lojas *online*, os consumidores dependem de informações fornecidas pelas descrições dos produtos, imagens, vídeos e avaliações de outros clientes. Embora esses recursos possam ajudar a fornecer uma ideia geral sobre o produto, eles nem sempre são suficientes para substituir a experiência física de inspecioná-lo pessoalmente. A falta de comunicação cara a cara também pode levar a um nível mais alto de incerteza para os consumidores. Não ter a oportunidade de conversar diretamente com um vendedor ou receber recomendações personalizadas pode gerar dúvidas e inseguranças. Os consumidores podem questionar-se sobre o tamanho correto de uma peça de roupa, a adequação de um produto às suas necessidades específicas ou até mesmo sobre a confiabilidade da loja virtual em si.

Outro fator que influencia o comércio eletrónico é a segurança dos dados pessoais dos consumidores. É essencial que as empresas adotem medidas robustas de proteção da privacidade dos clientes. Isso inclui a implementação de políticas claras de privacidade, a utilização de criptografia de dados durante as transações, a adoção de práticas de cibersegurança e o cumprimento de regulamentações específicas como o Regulamento Geral de Proteção de Dados (GDPR) na União Europeia. Ao mesmo tempo, os consumidores também devem estar cientes dos riscos e adotar medidas de proteção, como

o uso de senhas seguras, a atualização de *software* e a verificação dos certificados de segurança das plataformas de *e-commerce*.

A confiabilidade nas entregas é outra preocupação que precisa de ser estudada. As empresas devem investir numa logística eficiente, estabelecer parcerias confiáveis com serviços de transporte e fornecer informações claras sobre prazos de entrega e rastreamento de encomendas. Os consumidores, por sua vez, podem pesquisar a reputação da empresa e ler avaliações de outros clientes para tomar decisões mais informadas. Em casos de problemas com entregas, é importante que as empresas sejam responsivas e ofereçam suporte ao cliente para resolver quaisquer questões de forma adequada.

A possibilidade de fraudes também é uma preocupação significativa que as empresas devem ter em conta. Estas, devem investir em medidas de segurança adicionais, como autenticação de dois fatores e sistemas antifraude, para proteger as transações e evitar atividades fraudulentas. Por sua vez, os consumidores devem estar atentos a sinais de possíveis fraudes, como ofertas muito boas para serem verdadeiras, sites com *design* duvidoso e solicitações de informações sensíveis não justificadas.

Em relação a este tópico, existe um comportamento nas compras *online* que requer uma melhor compreensão: o abandono do carrinho de compras virtual. No âmbito deste assunto, existe uma fase fundamental do processo de compra *online*, que é a fase de finalização da compra, conhecida como *checkout*.

Esta etapa marca o encerramento do processo de compra *online*, e apenas após a sua conclusão se pode afirmar que a compra *online* foi bem-sucedida. Nessa fase, é crucial compreender o fenómeno de não-compra, que é frequentemente observado no contexto do comércio eletrónico. Esse comportamento ocorre quando os consumidores selecionam produtos nos seus carrinhos de compras virtuais, mas não finalizam a compra, resultando no abandono do carrinho de compras *online*.

O abandono do carrinho de compras *online* é um desafio significativo para os retalhistas *online*, pois representa uma perda de potenciais vendas. Por isso, é importante para as empresas entenderem as razões por trás desse comportamento e adotarem estratégias para reduzir o abandono, como a simplificação do processo de *checkout*, oferecer descontos

ou incentivos, melhorar a confiança do consumidor e fornecer um suporte ao cliente eficiente.

O ponto principal desta dissertação é prever em que situações as sessões dos clientes não terminam em compra. A principal contribuição é fornecer informações valiosas para os retalhistas *online*, a fim de aumentar as vendas nas suas lojas virtuais.

O intuito deste trabalho passa por recolher dados reais de uma loja virtual real sobre o seu processo de compra/*checkout* e os carrinhos virtuais dos seus clientes. O objetivo é recolher esses dados para, posteriormente, serem utilizados em processos de previsão com recurso a *machine learning* de modo a determinar se um determinado carrinho se transformará em abandonado ou se, por outro lado, o cliente efetivará a compra.

1.1. Relevância do tema

Para salientar a importância desta área de conhecimento, é necessário examinar dados do mercado na prática. Um estudo conduzido pela Barilliance (Serrano, 2020) revelou que a taxa média de abandono do carrinho de compras *online* foi de 78,65% em 2017 e 77,13% em 2019. Em termos simples, esse estudo estimou que cerca de três quartos dos compradores desistem das suas compras antes de concluí-las. Além disso, de acordo com uma pesquisa realizada pelo Baymard Institute (2020), as perdas decorrentes do abandono de carrinhos no comércio eletrónico dos Estados Unidos e da União Europeia chegam a aproximadamente US$ 260 biliões.

Com base nisso, o objetivo desta dissertação é analisar o abandono do carrinho de compras *online* e as razões que levam os consumidores a agirem dessa forma. Em termos simples, o abandono do carrinho de compras virtual acontece quando o consumidor adiciona itens ao seu carrinho de compras *online*, mas não efetua a compra de nenhum item durante aquela sessão de compras em específico (Kukar-Kinney & Close, 2010).

Sendo assim, a importância prática para os empreendedores que operam no mercado *online*, e o facto de existir um número elevado de carrinhos abandonados que levam a um crescente aumento de perdas económicas, sustentam a importância de explorar as razões que levam ao abandono do carrinho de compras *online*.

Como resultado disso, o objetivo deste estudo é compreender as razões por detrás do comportamento de não compra *online* e fornecer orientações para que as lojas virtuais possam melhorar as suas taxas de conversão em vendas. Por outras palavras, pretende-se compreender as motivações que levam os consumidores a desistirem da compra *online* e fornecer informações úteis para ajudar as lojas *online* a aumentarem as suas taxas de sucesso nas vendas, desenvolvendo para isso, um modelo de previsão de *machine learning* que apresentará previamente se um determinado carrinho será convertido em abandonado, podendo as empresas tomar ações para prevenir esse acontecimento.

1.2. Objetivos

Com o objetivo de tentar diminuir o elevado número de carrinhos abandonados, a finalidade deste projeto é obter resposta para o seguinte problema:

"Que utilizadores não concluirão o processo de *checkout*?"

Com base nos resultados obtidos, o objetivo é gerar conhecimento científico sobre o assunto e fornecer contribuições que permitam aos gestores de lojas *online* reduzir as interrupções no processo de finalização da compra. A intenção é garantir que os consumidores que selecionam produtos consigam concluir com êxito a etapa final da compra. Por outras palavras, procura-se oferecer *insights* e estratégias que ajudem os gestores de lojas *online* a minimizar obstáculos durante o processo de *checkout*, permitindo que os consumidores que adicionam itens ao carrinho de compras alcancem com sucesso a fase de conclusão da compra.

Por consequência, a fim de definir um caminho para responder ao problema exposto, também faz parte dos objetivos desenvolver um modelo de previsão de *machine learning* que seja capaz de prever se um carrinho se transformará em abandonado ou numa venda, para que os gestores das lojas *online* consigam ter essa informação e tomar decisões para reduzir o número de carrinhos abandonados.

1.3. Metodologia de trabalho

A metodologia de trabalho utilizada neste projeto foi o CRISP-DM (*Cross Industry Standard Process for Data Mining)* que consiste num modelo de processamento de mineração de dados e análise de dados.

Esta metodologia consiste em descrever abordagens comuns e utilizadas diversas vezes por especialistas em problemas de mineração de dados.

O CRISP-DM é uma abordagem iterativa que divide o processo de mineração de dados em várias etapas distintas, permitindo que as equipas de projeto organizem e gerem da melhor forma os seus esforços.

As etapas principais do CRISP-DM são as seguintes:

1- ***Business Understanding:*** Nesta fase inicial, os objetivos e requisitos do projeto são definidos. A equipa de projeto deve entender o contexto de negócios e as metas que a análise de dados visa alcançar. Isso envolve interações com os *stakeholders* para definir claramente o problema.

2- ***Data Understanding:*** Nesta etapa, a equipa recolhe, explora e avalia os dados disponíveis para o projeto. Isso inclui a identificação de fontes de dados, a obtenção dos dados necessários e a avaliação da qualidade e relevância dos dados.

3- ***Data Preparation****:* Os dados recolhidos são processados e preparados para a análise. Isso pode incluir limpeza de dados, transformação e seleção de características. O objetivo é criar um conjunto de dados pronto para análise.

4- ***Modeling:*** Nesta fase, a equipa constrói modelos de mineração de dados utilizando técnicas apropriadas, como árvores de decisão, redes neurais, regressão, etc. Esses modelos são ajustados e avaliados para encontrar o melhor desempenho possível.

5- ***Evaluation:*** Os modelos construídos são avaliados com base em métricas relevantes, como precisão, *recall*, F1-score, etc. A equipa determina se os modelos atendem aos objetivos de negócio e faz ajustes conforme necessário.

6- ***Deployment:*** Os modelos validados são implementados num ambiente de produção. Isso pode envolver a criação de sistemas ou aplicações que utilizam os resultados da mineração de dados para tomar decisões.

7- ***Monitoring:*** Após a implementação, é importante monitorizar o desempenho contínuo dos modelos e fazer ajustes à medida que os dados mudam ou evoluem.

Figura 1 - CRISP-DM

O CRISP-DM é um modelo flexível que reconhece a natureza iterativa do processo de mineração de dados. As etapas podem ser repetidas conforme necessário para refinar os modelos ou incorporar novos dados. Esta abordagem fornece uma estrutura sólida para o desenvolvimento de projetos de mineração de dados bem-sucedidos, ajudando as organizações a extrair *insights* valiosos dos seus dados para tomar decisões informadas.

Cada fase do CRISP-DM é cuidadosamente delineada, detalhando as atividades realizadas, os métodos empregues e os resultados obtidos. Isto proporciona uma visão completa do processo de análise de dados aplicado durante o desenvolvimento do projeto, permitindo uma compreensão profunda de como cada etapa contribuiu para alcançar os objetivos estabelecidos e gerar *insights* valiosos.

1.4. Estrutura da dissertação

Este documento está organizado em cinco partes. O primeiro capítulo diz respeito à introdução, onde são apresentados o tema a ser desenvolvido, a sua justificação e relevância, além da definição do problema, dos objetivos e da metodologia seguida. O segundo capítulo aborda o enquadramento teórico, apresentando uma síntese e definição de pontos importantes para este projeto, como *grandes volumes de dados*, *e-commerce analytics*, comportamento do consumidor, inteligência artificial, *machine learning* e terminologia fundamental para o projeto. O terceiro capítulo detalha e explica como o projeto será conduzido e executado, além de apresentar os resultados e as discussões do projeto, incluindo a avaliação do modelo de *machine learning* desenvolvido, a interpretação dos resultados obtidos e as conclusões extraídas desses resultados. O último capítulo apresenta as conclusões do projeto, discute as limitações encontradas e recomendações para trabalhos futuros.

2. Enquadramento teórico

No contexto académico, o enquadramento teórico desempenha um papel fundamental na construção e orientação de qualquer investigação, fornecendo a base conceitual sobre a qual a pesquisa é edificada. Este capítulo visa explorar e delinear as principais teorias e modelos que sustentam o estudo em questão, oferecendo uma visão abrangente dos conceitos chave e das abordagens que moldam a compreensão do fenómeno investigado. Através da análise crítica da literatura existente e da integração dos diferentes paradigmas teóricos, será possível estabelecer um referencial sólido que não só contextualiza a pesquisa, mas também identifica lacunas e oportunidades para novas contribuições no campo. O objetivo é fornecer uma fundamentação teórica robusta que ajude a orientar as questões de investigação e as metodologias a serem empregues, assegurando a coerência e a relevância dos resultados obtidos.

2.1. A era dos grandes volumes de dados

Desde o início do século XXI, ocorreram muitas mudanças tecnológicas significativas na indústria das tecnologias da informação, como a computação em nuvem, a Internet das Coisas e as redes sociais. O desenvolvimento dessas tecnologias tem feito a quantidade de dados aumentar continuamente e acumular a uma velocidade sem precedentes. Todas as tecnologias mencionadas anunciam a chegada dos grandes volumes de dados. Atualmente, a quantidade de dados globais está a crescer exponencialmente (Cai et al., 2015).

Na era da grande explosão de informações, a velocidade de geração de informações está a aumentar de dia para dia, e o volume de informações produzidas em todo o mundo é imenso. Nos últimos anos, o termo "Grandes volumes de dados" tornou-se uma das palavras mais utilizadas nos setores industriais, financeiros e de saúde (Yang et al., 2020).

Estamos perante um aumento exponencial na quantidade de dados recolhidos e armazenados digitalmente. Esse crescimento vertiginoso reflete a expansão significativa do universo de informações digitais que estão a ser acumuladas em diferentes fontes e plataformas. A cada instante, uma enorme quantidade de dados é gerada e capturada por meio de diversos dispositivos, sensores, transações *online* e interações digitais. Essa

avalanche de informação está a redefinir os limites do que é possível armazenar e analisar, impulsionando a necessidade de soluções avançadas de gestão, armazenamento e processamento de dados. A magnitude dessa expansão de dados desafia não apenas as capacidades técnicas, mas também oferece uma oportunidade única de explorar e extrair conhecimentos valiosos que podem impulsionar a inovação, melhorar a tomada de decisão e avançar a sociedade como um todo (Yang et al., 2020).

A análise de grandes volumes de dados está no centro da ciência e dos negócios modernos. Estes dados são produzidos a partir de transações *online*, *e-mails*, vídeos, áudios, imagens, registos de cliques, *logs*, pesquisas. Os dados são armazenados em bases de dados que crescem massivamente e tornam-se difíceis de capturar, organizar, armazenar, gerir, compartilhar, analisar e visualizar por meio das ferramentas típicas de *software* de base de dados (Smari, W. W, 2013).

Antes da revolução deste tema, as empresas não tinham a capacidade de armazenar todos os seus arquivos por longos períodos, nem gerir eficientemente conjuntos de dados enormes. As tecnologias tradicionais possuíam capacidade de armazenamento limitada, ferramentas de gestão rígidas e eram caras. Estas ferramentas careciam de escalabilidade, flexibilidade e desempenho necessários no contexto de grandes volumes de dados. Na verdade, a sua gestão requer recursos significativos, novos métodos e tecnologias poderosas. Mais precisamente, exige limpeza, processamento, análise, segurança e acesso granular a conjuntos de dados em constante evolução. As empresas e indústrias estão cada vez mais conscientes de que a análise de dados se tornou um fator principal para serem competitivas, descobrir novos *insights* e personalizar serviços (Oussous et al., 2018).

Para extrair conhecimento de grandes volumes de dados, diversos modelos, programas, *softwares*, *hardwares* e tecnologias foram desenvolvidos e propostos. Eles visam garantir resultados mais precisos e confiáveis para as aplicações de grandes volumes de dados. No entanto, nesse ambiente, pode ser demorado e desafiador escolher entre inúmeras tecnologias disponíveis. De facto, muitos parâmetros devem ser considerados: compatibilidade tecnológica, complexidade de implementação, custo, eficiência, desempenho, confiabilidade, suporte e riscos de segurança (Oussous et al., 2018).

A área de grandes volumes de dados possui uma natureza complexa que requer tecnologias poderosas e algoritmos avançados. Portanto, as ferramentas tradicionais de *Business Intelligence* estáticas já não são eficientes no caso de aplicações de grandes volumes de dados. Por outras palavras, as abordagens tradicionais de análise de dados não são adequadas para lidar com a vasta variedade e velocidade de dados, tornando necessário o uso de tecnologias mais avançadas e flexíveis (Oussous et al., 2018).

O conceito de grandes volumes de dados pode ser descrito com base nas seguintes características, conhecidas como os cinco V's (Parashar, 2013)*:*

1. Variedade - A variedade de dados gerados não se limita a uma única categoria, pois inclui não apenas os dados tradicionais, mas também os dados semiestruturados provenientes de diversas fontes, como páginas da *web*, *e-mails*, documentos e dados de dispositivos de sensores, tanto de dispositivos ativos quanto passivos. Todos estes dados são totalmente diferentes, consistindo em dados brutos, estruturados, semiestruturados e até mesmo não estruturados, o que dificulta o seu tratamento pelos sistemas analíticos tradicionais existentes.

2. Volume - Atualmente, os dados existentes estão na ordem de *petabytes* e espera-se que aumentem para *zettabytes* num futuro próximo. Os *sites* de redes sociais em si estão a gerar dados na ordem de *terabytes* todos os dias, e essa quantidade de dados é definitivamente difícil de ser tratada utilizando os sistemas tradicionais existentes.

3. Velocidade - A velocidade é um conceito que lida com a rapidez dos dados provenientes de várias fontes. Essa característica não se limita apenas à velocidade dos dados recebidos, mas também à velocidade com que os dados fluem. Por exemplo, os dados provenientes de dispositivos de sensores estarão constantemente em movimento para o armazenamento da base de dados, e essa quantidade não será pequena. Portanto, os sistemas tradicionais não são capazes de realizar análises nos dados que estão constantemente em movimento.

4. Veracidade - Refere-se à qualidade e confiabilidade dos dados. Muitas vezes com a grande quantidade de dados enfrenta-se o desafio de lidar com dados incertos,

inconsistentes e ruidosos. A veracidade dos dados é essencial para garantir resultados precisos e confiáveis nas análises.

5. Valor - Os utilizadores podem executar determinadas consultas nos dados armazenados e, assim, podem deduzir resultados importantes a partir dos dados filtrados obtidos, além de poder classificá-los de acordo com as dimensões que necessitam. Estes relatórios ajudam essas pessoas a identificar as tendências de negócio, com base nas quais podem alterar as suas estratégias. À medida que os dados armazenados por diferentes organizações são utilizados para análise de dados, surge uma lacuna entre os líderes de negócios e os profissionais de TI. A principal preocupação dos líderes de negócios seria agregar valor aos seus negócios e obter cada vez mais lucro, ao contrário dos líderes de TI, que precisariam de se preocupar com as questões técnicas do armazenamento e processamento. Portanto, os principais desafios enfrentados pelos profissionais de TI ao lidar com grandes volumes de dados são:

    - Sistemas capazes de lidar de forma eficiente e eficaz com uma grande quantidade de dados;
    - Filtrar os dados mais importantes de todos os dados recolhidos pela organização. Por outras palavras, podemos dizer que é adicionar valor ao negócio.

Em resumo, os desafios enfrentados pelos profissionais de TI incluem o desenvolvimento de sistemas robustos para gerir grandes volumes de dados e a capacidade de identificar e extrair informações valiosas desses dados para agregar valor aos negócios.

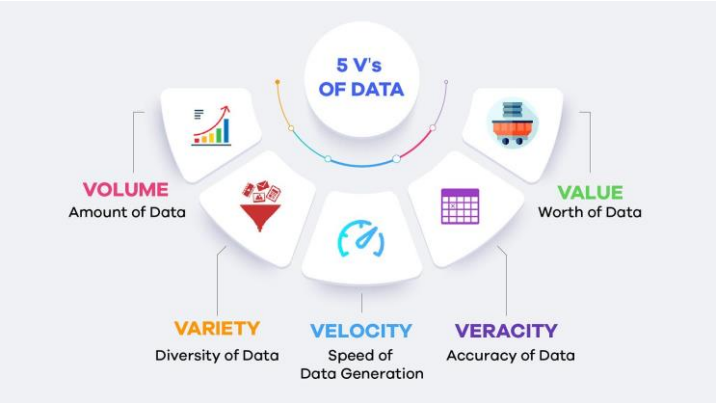


Figura 2 - Os 5 V's dos dados

Ao contrário dos dados tradicionais, o termo "Grandes volumes de dados" refere-se a conjuntos de dados em constante crescimento que incluem formatos heterogéneos (Eberendu, 2016):

1. Dados estruturados - referem-se a dados que possuem um formato e tamanho definidos, sendo fáceis de armazenar e analisar, com alto grau de organização. Isso significa que os dados são organizados numa estrutura identificável, permitindo que eles respondam a consultas para recuperar informações para uso organizacional. Um exemplo típico de dados estruturados é uma base de dados relacional, como o *SQL*, que contém números, datas e grupos de palavras e números organizados, chamados de *strings*/texto. Devido à estrutura contínua da base de dados, esta pode ser pesquisada com algoritmos de pesquisa simples e diretos, que podem ser baseados no tipo de dados ou no conteúdo real. A análise tradicional tem se concentrado em dados estruturados, negligenciando uma grande quantidade de outros tipos de dados.

2. Dados semiestruturados - são dados irregulares que podem ser incompletos e possuem uma estrutura que muda rapidamente ou de forma imprevisível, mas não segue um esquema fixo ou explícito. Isto significa que eles não são orientados a tabelas, como num modelo de base de dados relacional, nem em formato de grafo, como em bases de dados de objetos. Por outras palavras, dados semiestruturados são aqueles que possuem uma estrutura menos rígida do que os dados estruturados, permitindo que diferentes tipos de informações sejam integrados, mesmo que não sigam um esquema específico.

3. Dados não estruturados - O termo "dados não estruturados" tem sido utilizado com frequência nos últimos anos, mas não foi definido de forma exaustiva e significativa. Os dados não estruturados (ou informações não estruturadas) referem-se a informações que não possuem um modelo de dados pré-definido ou não se encaixam em tabelas relacionais (Sint et al., 2009). Informações não estruturadas geralmente são texto, mas também podem incluir dados como datas. Isso introduz irregularidades e ambiguidades que são difíceis de serem compreendidas por programas de computador tradicionais em comparação com dados armazenados em bases de dados ou anotados em documentos (Sint et al., 2009).

   Os dados não estruturados geralmente consistem em arquivos como documentos de processamento de texto, folhas de cálculo, *PDFs*, redes sociais e conteúdo de mensagens, vídeos, multimédia e gráficos. Ao contrário dos dados relacionais, as informações não estruturadas não têm nenhum tipo de dados definido e nenhuma regra imposta que indique onde os dados são armazenados (Rusu et al.).

2.2. *E-commerce analytics*

Nos últimos anos, tem-se dado ênfase crescente à análise de grandes volumes de dados no comércio eletrónico (Safara, 2020).

Na maioria dos casos, as empresas de comércio eletrónico lidam com dados estruturados e não estruturados. Enquanto os dados estruturados se centram em dados demográficos, incluindo nome, idade género, data de nascimento, morada e preferências, os dados não estruturados incluem cliques, gostos, ligações, *tweets*, etc. No ambiente de grandes volumes de dados, o desafio é lidar com ambos os tipos de dados, a fim de gerar conhecimento significativo para aumentar as vendas (Safara, 2020).

Atualmente, o panorama do comércio eletrónico está repleto de grandes volumes de dados que estão a ser utilizados para resolver problemas comerciais. A utilização de grandes volumes de dados disparou no comércio eletrónico "devido [às] redes sociais, à Internet, à telefonia móvel e a todos os tipos de novas tecnologias que criam e captam dados" (Kauffman et al., 2012). Com a ajuda de uma capacidade de armazenamento e

processamento económicas, e de ferramentas analíticas de ponta, os grandes volumes de dados permitem agora às empresas de comércio eletrónico reduzir custos e gerar benefícios sem qualquer dificuldade (Kauffman et al., 2012).
No entanto, a análise que capta os grandes dados é diferente dos dados tradicionais em muitos aspetos. Especificamente, devido aos elementos da sua natureza única (ou seja, volume, variedade, velocidade e veracidade), os grandes volumes de dados podem ser facilmente distinguidos da forma tradicional.

2.2.1. Tipos de grandes volumes de dados utilizados em *e-commerce*

O comércio eletrónico refere-se às transações *online*: venda de bens e serviços na *Internet*, quer numa única transação (por exemplo Amazon, Zappos, eBay, Expedia) ou através de uma transação contínua (por exemplo, Netflix, Match.com, LinkedIn, etc.) (Strauss et al., 2016).

As empresas de comércio eletrónico, desde a Amazon à Netflix, capturam vários tipos de dados (por exemplo, encomendas, carrinhos de compras, visitas, utilizadores, palavras-chave, navegação em catálogos, dados sociais, etc), que podem ser classificados em quatro categorias: (a) dados sobre transações ou atividades comerciais; (b) dados sobre o *click-stream*; (c) dados de vídeo; e (d) dados de voz.

No comércio eletrónico, os dados são a chave para seguir o comportamento de compra dos consumidores e personalizar as ofertas que são recolhidas ao longo do tempo utilizando a navegação do consumidor e pontos transacionais.

Esta secção analisa os diferentes tipos de grandes volumes de dados e as suas implicações para o comércio eletrónico.

2.2.1.1. Dados sobre transações ou atividades comerciais

Os dados relativos a transações ou atividades comerciais evoluem em resultado de trocas entre o cliente e a empresa ao longo do tempo. Estes dados são estruturados por natureza e têm origem em muitas fontes de dados que vão desde programas de relacionamento com

clientes (por exemplo perfis de clientes mantidos pela empresa, a ocorrência de reclamações de clientes) até às transações de vendas.
Por exemplo, a Amazon utiliza um tipo de técnica de modelação preditiva denominada filtragem colaborativa, utilizando os dados dos clientes para gerar avisos "talvez também queira" para cada produto comprado ou visitado. A Amazon revelou, numa determinada altura, que 30 % das vendas eram geradas através do seu motor de recomendação (Manyika et al., 2011).

Em geral, é evidente que os retalhistas eletrónicos podem obter numerosos benefícios em toda a cadeia de valor utilizando dados de transações.

2.2.1.2. Dados sobre *click-stream*

Os dados do fluxo de cliques têm origem na *Web* e em anúncios *online*, bem como em conteúdos das redes sociais, como *tweets*, *blogues*, publicações do Facebook, etc, de empresas de comércio eletrónico.

No ambiente conectado de hoje, as redes sociais e os anúncios *online* desempenham um papel fundamental na estratégia promocional das empresas, como a utilização de dados sobre o fluxo de cliques que são muito importantes para a gestão na tomada de decisões informadas, estratégicas e táticas. Estudos anteriores concluíram que muitas empresas de comércio eletrónico em todo o mundo (por exemplo, Amazon, eBay, Zappos, Alibaba, etc.) recorrem a dados sobre o fluxo de cliques nos seus esforços para captar dados. Os dados sobre o fluxo de cliques podem ser aplicados para prever preferências e gostos dos clientes. A Netflix capta e analisa mais de mil milhões de dados da *Web* relacionados com críticas de filmes que são apreciados, amados, odiados, etc, para compreender os gostos dos clientes (Davenport et al., 2005).

As empresas de cartão de crédito, por meio de dados do *site* e do *call center* mantêm bases de dados (designadas por *ready-to-market*) para oferecer produtos personalizados ao cliente em milissegundos e otimizar as ofertas através do acompanhamento das respostas dos clientes (Davenport, 2012). Algumas empresas utilizam essas bases de dados não só para abordar os clientes, mas também para oferecer serviços *online*. Ao analisar dados da *Web*, os retalhistas eletrónicos recebem um alerta vermelho quando os preços dos

produtos dos seus concorrentes estão abaixo do seu próprio nível de preços (Biesdorf et al., 2013). Por conseguinte, os retalhistas podem ajustar os seus preços para se manterem competitivos.

2.2.1.3. Dados de vídeo

Os dados de vídeo são dados que resultam da captação de imagens em direto. Atualmente, as empresas de comércio eletrónico estão interessadas em utilizar não só dados de fluxo de cliques ou dados de transação, mas, em associação com *software* de análise de imagem, tendem também a captar dados de vídeo.

As empresas de comércio eletrónico têm as competências necessárias para analisar dados extremamente não estruturados, como dados de vídeo ou de voz. Estes dados têm o potencial de acrescentar valor para as empresas de comércio eletrónico (Schroeck et al., 2012). Por exemplo, a Netflix utiliza dados de vídeo para prever hábitos de visualização e avaliar a qualidade das experiências (Ramaswamy, 2012). Além disso, a ferramenta de visualização e de análise da procura baseada no tipo de consumo de filmes ajudam a Netflix a compreender as preferências dos seus clientes.

Além deste exemplo, existem alguns retalhistas que utilizam *software* sofisticado de análise de imagens ligado às suas câmaras de videovigilância para detetarem padrões de tráfego e comportamento do consumidor (Manyika et al., 2011). Assim, a utilização de dados de vídeo é essencial para as empresas tomarem melhores decisões do que os seus concorrentes.

2.2.1.4. Dados de voz

Outro tipo de dados ligado à família dos grandes volumes de dados são os dados de voz, ou seja, dados tipicamente provenientes de chamadas telefónicas, *call centers* ou atendimento ao cliente. Conforme evidenciado em pesquisas recentes, os dados de voz são vantajosos para analisar o comportamento de compra do consumidor ou consumidores ou para direcionar novos clientes. As empresas de cartões de crédito, por exemplo a American Express, utilizam e rastreiam dados relacionados com as atividades dos centros de atendimento telefónico para que possam ser feitas ofertas personalizadas

em milissegundos (Davenport et al., 2012). As empresas de comércio eletrónico utilizam avançadas técnicas para analisar texto e transcrições convertidas de conversas de *call center* (Schroeck et al., 2012). Para além disso, numerosas nuances da linguagem, como o sentimento, o calão e as intenções, podem ser lidas e reconhecidas no contexto do comércio eletrónico.

Uma vez que a natureza e o tipo de grandes volumes de dados são únicos e provenientes de várias redes de plataformas digitais, existe a possibilidade de uma nova teoria para a resolução de novos problemas. A economia dos dados indica também que os grandes dados são "relacionais" e "em rede", que exigem novos desenvolvimentos em termos de capacidades e algoritmos de TI, qualidade dos sistemas e dos dados, privacidade e implicações éticas, alinhamento estratégico e cultura empresarial.

2.2.2. Valor comercial da análise de grandes volumes de dados para as empresas de comércio eletrónico

O desafio final da análise de grandes volumes de dados é gerar valor comercial para a organização (Beath et al., 2012). O termo "valor" no contexto dos grandes volumes de dados implica a geração de conhecimentos e/ou benefícios economicamente válidos através da análise, extração e transformação de dados. O valor comercial da análise de grandes volumes de dados consiste nos benefícios transacionais, informativos e estratégicos para as empresas de comércio eletrónico (Wixom et al., 2013). Enquanto o valor transacional se centra na melhoria da eficiência e na redução dos custos, o valor informativo permite a tomada de decisões em tempo real e o valor estratégico diz respeito à obtenção de vantagens competitivas.

Por exemplo, ao realizar análises no comércio eletrónico, os gestores podem obter valor comercial global ao satisfazerem as necessidades dos clientes, ao criarem produtos e serviços, ao expandirem-se para novos mercados, e aumentando as vendas e as receitas (Columbus, 2014).

Existem seis mecanismos que têm como objetivo melhorar os valores práticos do negócio na economia dos dados: personalização, preços dinâmicos, serviço ao cliente, visibilidade da cadeia de valores, segurança e deteção de fraudes e análise preditiva (Safara, 2020).

2.2.2.1. Personalização

A primeira aplicação dos grandes volumes de dados para as empresas de comércio eletrónico é o fornecimento de serviços personalizados ou produtos (Koutsabasis et al., 2008). Os estudos argumentam que os consumidores gostam de fazer compras no mesmo retalhista utilizando diversos canais, e que os grandes volumes de dados destes diversos canais podem ser personalizados em tempo real (Kopp, 2013). A análise de dados em tempo real permite que as empresas ofereçam serviços personalizados, incluindo conteúdos e promoções especiais aos clientes. Além disso, esses serviços personalizados ajudam as empresas a separar clientes fiéis de novos clientes e a fazer ofertas promocionais em conformidade (Mehra, 2013).

A personalização pode aumentar as vendas em 10% ou mais (Liebowitz, 2013). A Bloomspot, neste contexto, explorou os dados dos cartões de crédito dos clientes para seguir os registos de despesas dos clientes mais fiéis e oferecer-lhes prémios através de ofertas e benefícios que ajudaram a aumentar a fidelidade dos clientes (Miller, 2013). A Wine.com conseguiu um aumento maciço das suas vendas utilizando o *marketing* personalizado por correio eletrónico (Zhao, 2013).

2.2.2.2. Preços dinâmicos

No atual ambiente de mercado extremamente competitivo, os clientes são considerados "reis". Por conseguinte, para atrair novos clientes, as empresas de comércio eletrónico devem ser ativas e vibrantes e, ao mesmo tempo, estabelecer um preço competitivo (Kung, 2013). O sistema de preços dinâmicos da Amazon monitoriza os preços da concorrência e alerta a Amazon de 15 em 15 segundos, o que resultou num aumento de 35 % em todas as vendas. Para oferecer preços competitivos aos clientes em vésperas de possíveis aumentos de vendas (como no Natal ou noutras épocas festivas), a Amazon processa os dados tendo em conta os preços dos concorrentes, as vendas de produtos, as ações dos clientes e quaisquer preferências regionais ou geográficas (Kopp, 2013). O acesso a estas informações através da utilização de grandes volumes de dados é suscetível de permitir às empresas de comércio eletrónico estabelecer preços dinâmicos (Leloup et al., 2001).

2.2.2.3. Serviço ao cliente

Outra área fundamental em que as empresas de comércio eletrónico podem utilizar os resultados da análise de grandes volumes de dados é o serviço ao cliente. As queixas dos clientes comunicadas através de formulários de contacto nas lojas *online*, juntamente com *tweets*, permitem às empresas de comércio eletrónico fazer com que os clientes se sintam valorizados quando ligam para o centro de assistência resultando numa pronta prestação de serviços (Mehra, 2013).

2.2.2.4. Visibilidade da cadeia de valor

Quando os clientes fazem uma encomenda numa plataforma *online*, é lógico que eles esperam que as empresas forneçam o serviço de seguimento da encomenda enquanto as mercadorias ainda estão em trânsito. Os clientes esperam informações importantes, como a disponibilidade exata, o estado atual e a localização das suas encomendas (Kopp, 2013).

As empresas de comércio eletrónico têm frequentemente dificuldade em responder a estas expectativas dos clientes, uma vez que vários terceiros, como o armazenamento e o transporte, estão envolvidos no processo da cadeia de valor (Kopp, 2013).

A análise de grandes volumes de dados desempenha um papel fundamental neste contexto ao recolher múltiplas informações de múltiplas partes sobre vários produtos (Mehra, 2013), e subsequentemente e, posteriormente, informa com precisão a data de entrega prevista aos clientes.

2.2.2.5. Segurança e deteção de fraude

As perdas relacionadas com a fraude ascendem, em média, a 9000 USD por cada 1 milhão de USD de receitas (Mehra, 2013). Este montante significativo de perdas pode ser evitado através da identificação relevante e através da utilização de grandes volumes de dados. Com a ajuda da infraestrutura certa, como o Hadoop, as empresas de comércio podem analisar os dados a um nível agregado para identificar fraudes relacionadas com cartões de crédito, devoluções de produtos e roubo de identidade (Mehra, 2013).

Além disso, as empresas de comércio eletrónico podem identificar fraudes em tempo real, combinando dados de transação com o histórico de compras dos clientes, registos da *Web*, *feed* social e dados de localização geoespacial de aplicações para *smartphones*. Por exemplo, a Visa instalou um sistema de gestão de fraudes baseado em megadados que permite a inspeção de 500 aspetos diferentes de uma transação. Com este sistema permite poupar 2 mil milhões de dólares em potenciais perdas anualmente.

2.2.2.6. Análise preditiva

A análise preditiva refere-se à identificação de eventos antes de estes ocorrerem, através da utilização de grandes volumes de dados (Kopp, 2013). Neste contexto, Loveman (2003), diretor executivo e presidente da Caesar's Entertainment, afirmou que: "[a] melhor maneira de se envolver em ... *marketing* orientado por dados é recolher informações cada vez mais específicas sobre as preferências dos clientes, efetuar experiências e análises com os novos dados e determinar formas de apelar aos interesses dos clientes".

Por conseguinte, a análise preditiva ajuda as empresas a preparar os seus orçamentos de receitas. A preparação destes orçamentos ajuda as empresas de comércio eletrónico a reconhecer futuros padrões de vendas a partir de dados de vendas anteriores (por exemplo, anuais ou trimestrais). Isto, por sua vez, ajuda as empresas a prever melhor e a determinar as necessidades de inventário, o que permite evitar ruturas de *stock* de produtos e a perda de clientes (Kopp, 2013).

As empresas de comércio eletrónico extraem cada vez mais valor comercial dos conhecimentos sobre grandes volumes de dados, quer para resolver problemas comerciais quer para tomar decisões. Este novo desenvolvimento no domínio do comércio eletrónico baseado em dados desencadeia o desenvolvimento de novas teorias no contexto do valor comercial tangível (por exemplo, melhoria da produtividade) e intangível (por exemplo, compreensão estratégica da empresa) utilizando pessoas, processos e tecnologia.

2.3. Comportamento do Consumidor

O comportamento do consumidor refere-se à ação de adquirir um produto ou serviço, seja para uso pessoal ou para terceiros, com o intuito de satisfazer uma necessidade imediata ou de modificar ou trocar algo que seja necessário (Gomes et al.).

O comportamento do consumidor está diretamente relacionado com a satisfação das necessidades e desejos dos consumidores (Solomon et al., 2008). Portanto, é crucial realizar uma análise minuciosa do comportamento das pessoas antes, durante e após a aquisição para compreender completamente o processo de compra.

O comportamento do consumidor abrange as ações relacionadas com a obtenção, consumo e descarte de produtos e serviços (Vieira, 2000). Isso inclui os processos de tomada de decisão que precedem e seguem essas atividades.

2.3.1. Novas estratégias *online* de vendas

O crescimento dos canais de venda e comunicação *online*, como lojas virtuais e redes sociais, tornou-se uma realidade para muitos vendedores como uma forma de agregar valor aos seus negócios. Essa abordagem tem-se mostrado eficaz na expansão do alcance para mais clientes, resultando no aumento da receita para aqueles que adotam esse modelo. No entanto, lidar com a diversidade de canais apresenta o desafio da integração eficiente entre eles. Nesse contexto, para uma gestão adequada e eficaz, é essencial implementar uma estratégia conhecida como "*e-commerce omnichannel*", que envolve a utilização de vários canais de comunicação diferentes dentro de uma única plataforma (Cavalcanti et al., 2021).

Além disso, tornou-se cada vez mais frequente que as empresas estejam ativas nas redes sociais e forneçam aplicações aos consumidores, utilizando essas plataformas como facilitadoras e mantedoras de uma conexão direta com o público. Nesse contexto, um exemplo ilustrativo dessa estratégia de vendas é a adoção do *Instagram*, que permite a exposição visual de produtos a potenciais clientes, proporcionando uma experiência completa de compra. Esta plataforma, inicialmente centrada em relações interpessoais, expandiu as suas funcionalidades para incluir ferramentas abrangentes que possibilitam

aos utilizadores realizar transações de compra e venda dentro do próprio ambiente da aplicação, atraindo assim, marcas de renome internacional (Cavalcanti et al., 2021).

Vinculado às estratégias de vendas através de aplicações e plataformas, surge o fenómeno do comportamento de compra *online* influenciado pelos "*digital influencers*". Estes indivíduos destacam-se como formadores de opinião cujo principal canal de comunicação com o público é constituído pelas redes sociais, incluindo *Facebook*, *Youtube* e *Twitter*. Os influenciadores digitais caracterizam-se por partilhar as suas opiniões nas redes sociais, atraindo assim um público que se identifica com as suas visões. Consequentemente, os utilizadores que consomem esse conteúdo manifestam interesse nas marcas, produtos e serviços recomendados pelos influenciadores, adotando, assim, o comportamento de compra, especialmente nas plataformas *online*. Diante desse padrão de comportamento *online*, muitas empresas identificaram uma oportunidade única para atrair mais clientes e concretizar vendas por meio dessa estratégia (Cavalcanti et al., 2021).

Estas estratégias inovadoras estão a tornar-se cada vez mais comuns e representam uma parte significativa das vendas de produtos e serviços através dos meios digitais para muitas marcas. Mesmo que os métodos de venda mais tradicionais continuem a existir e a atrair consumidores, fica evidente que incorporar abordagens novas para alcançar os clientes é crucial. Essa necessidade não se limita apenas a superar desafios momentâneos, mas também se estende a adaptar o negócio às tendências em evolução do mercado. Portanto, compreender os fatores-chave levados em consideração pelos consumidores no momento da decisão de compra *online* é essencial.

2.3.2. Fatores determinantes na decisão de compra *online*

Durante todo o processo de decisão de compra *online*, o consumidor geralmente analisa vários elementos que podem influenciar a sua decisão de concluir ou não a compra. Esses fatores desempenham um papel crucial no sucesso das empresas que operam no ambiente de vendas *online*. Portanto, para compreender as complexidades envolvidas na tomada de decisões de compra no comércio eletrónico, foram identificados na literatura os fatores mais recorrentes que exercem influência nesse processo, os quais serão detalhados a seguir.

O estudo realizado por Antunes (2011) que investigou como a confiança afeta a decisão de compra *online*, considerando que todo cliente precisa de sentir confiança ao realizar transações eletrônicas (Cavalcanti et al., 2021). O fator de confiança foi examinado em relação a elementos como a segurança dos dados pessoais, a certeza de que o pagamento será processado corretamente e a confiança na entrega do produto em perfeitas condições, acreditando que o processo de entrega não comprometerá a integridade do que foi adquirido.

A pesquisa conduzida por Rita, Oliveira e Farisa (2019), identificou a qualidade do serviço *online* como um fator significativo (Cavalcanti et al., 2021). Este elemento é uma compilação de vários aspetos, desde a experiência do cliente ao aceder à loja *online* até concluir o processo de compra. Por outras palavras, a qualidade abrange desde o interesse inicial em comprar até o recebimento efetivo do produto. O cliente percebe todo esse ciclo de compra de forma única, influenciada pela imagem que cada loja virtual projeta. Portanto, a qualidade do serviço *online* reflete todos os sentimentos do consumidor ao longo desse período.

No mesmo estudo, a satisfação do cliente foi destacada como um aspeto crucial nas compras *online*. A satisfação do cliente difere da qualidade do serviço, pois está relacionada com tudo o que o cliente percebe após concluir a compra *online* e receber o produto ou serviço. Após passar por todo o ciclo para avaliar a qualidade, o cliente precisa de determinar se o que experimentou atende ou não às suas expectativas. Isso pode envolver a comparação entre o que foi adquirido e o que foi recebido, ou a avaliação de qualquer aspeto que o cliente não tenha gostado durante o processo de compra. No final, após ponderar todas essas perceções sobre a qualidade, o cliente decide a sua satisfação com a experiência na loja virtual. Esse fator tem o potencial de influenciar decisões futuras de compra, tanto para o próprio cliente quanto para outros potenciais compradores (Cavalcanti et al., 2021).

O fator relacionado à loja virtual refere-se ao *design* do *site* ou aplicação destinado ao cliente final, garantindo que ele compreende de forma clara e simples todas as características do produto, seja em termos físicos, como dimensões, ou em aspetos situacionais, como preço. É relevante destacar que, para lojas ainda não conhecidas pelos clientes, a ênfase recai na facilidade de efetuar uma compra e na navegabilidade geral do

site. Essa facilidade está diretamente relacionada a um componente anterior à realização da compra do produto ou serviço, ou seja, a decisão de utilizar ou não o *site* para uma potencial compra. Se o *site* não oferecer uma experiência de utilização atraente ao utilizador, é provável que ele não considere utilizar essa plataforma (Cavalcanti et al., 2021).

A conveniência refere-se à perceção dos benefícios da compra *online* por parte do cliente. Isso pode envolver a ideia de que é mais fácil receber o produto em casa, em comparação com ir até uma loja física. Também pode surgir da falta de disponibilidade que o cliente precisa localmente, levando-o a ver o meio *online* como a única alternativa para atender às suas necessidades. Além disso, a conveniência está estreitamente relacionada com a perceção do cliente sobre como será o processo de compra dentro de uma determinada plataforma *online*, ou seja, como ele espera que o *site* o auxilie nas suas necessidades (Cavalcanti et al., 2021).

Outro elemento significativo na decisão de compra *online* é a ansiedade (Cavalcanti et al., 2021). Frequentemente, as pessoas procuram na compra *online* uma forma de satisfazer as suas necessidades, mas a facilidade de acesso a uma variedade de produtos e serviços também as torna entusiastas, desejando consumir tudo disponível nesse meio.

As influências ambientais também podem desempenhar um papel na tomada de decisão de compra *online*. Por exemplo, em países mais desenvolvidos, onde o acesso à *Internet* é generalizado em toda a população, é de se esperar que as pessoas estejam mais familiarizadas com a presença de grandes marcas no ambiente *online*. No entanto, essa cultura pode não ser tão difundida em países menos desenvolvidos. Portanto, os fatores ambientais de cada cliente tornam-se um ponto importante para a marca avaliar a viabilidade desse canal (Cavalcanti et al., 2021).

Como a experiência de compra *online* difere da compra física, os clientes necessitam de garantias quanto à integridade do produto e à resolução de possíveis problemas associados a ele (Cavalcanti et al., 2021). Portanto a garantia e serviço ao cliente tornam-se fatores importantes que levam o cliente a comprar numa loja *online*.

2.3.3. Prever o comportamento do consumidor

A previsão do comportamento do consumidor é uma tarefa complexa e vital para as estratégias de *marketing* e negócios. Diversas técnicas são empregues para entender as preferências, decisões de compra e interações dos consumidores. A análise de dados e os grandes volumes de dados desempenham um papel crucial, utilizando técnicas como mineração de dados e análise preditiva para identificar padrões e tendências. A aplicação de ML permite a construção de modelos que podem prever comportamentos futuros com base em dados históricos. A inteligência artificial, com as suas capacidades de processamento de linguagem natural e sistemas de recomendação, oferece *insights* valiosos. Além disso, a análise de sentimentos, por meio da monitorização de redes sociais e avaliações *online*, ajuda a compreender as opiniões dos consumidores. Técnicas tradicionais, como pesquisas e inquéritos *online*, continuam a ser ferramentas essenciais. Em conjunto, estas técnicas formam um arsenal diversificado para antecipar e compreender as complexidades das decisões de compra e interações dos consumidores.

Os sistemas de sugestões de produtos para adicionar ao carrinho representam uma estratégia perspicaz na experiência de compra *online*, destacando-se por oferecer recomendações personalizadas com base no comportamento do consumidor. Um exemplo notável é a abordagem da Amazon com o seu sistema de recomendação. Utilizando algoritmos sofisticados que analisam o histórico de compras, visualizações de produtos e padrões de navegação, a Amazon sugere produtos relevantes que se alinham aos interesses e preferências do utilizador. Ao fornecer sugestões contextualmente precisas durante a navegação, no carrinho ou durante o *checkout*, a Amazon cria uma experiência de compra altamente personalizada. Este sistema não facilita apenas a descoberta de novos produtos, mas também incentiva compras adicionais, maximizando a satisfação do cliente e contribuindo para o sucesso duradouro da plataforma.

2.4. *Machine Learning*

Para simplificar a estrutura de IA, a Figura 3 ilustra como esses componentes-chave se relacionam entre si de uma perspetiva de alto nível. A IA é a família mais ampla que tem como componentes o *Machine Learning* (ML) e *Deep Learning* (DL). O ML é um subconjunto da IA, enquanto o DL é um subconjunto do ML. (Miguel, 2022).

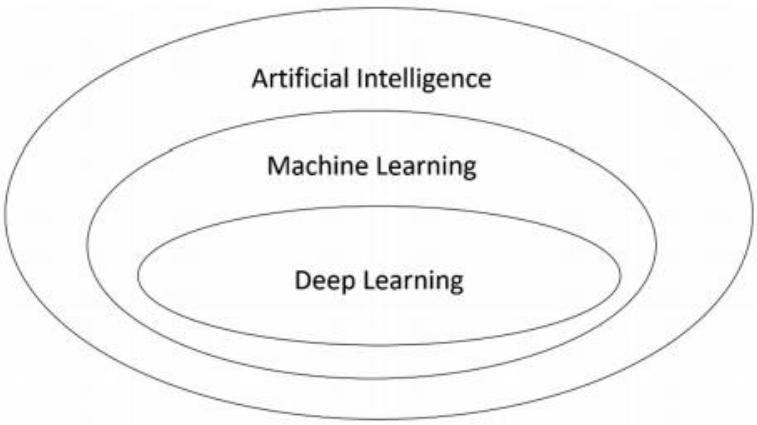


Figura 3- Inteligência Artificial vs Machine Learning vs Deep Learning

O termo "inteligência artificial" nasceu da ambição de permitir que os computadores executem as tarefas do cérebro humano. Um conjunto de dados de entrada é crucial para formar um parâmetro ou número para obter dados de resposta, que são basicamente expressos em função de probabilidades.

A inteligência artificial consiste em fazer com que as máquinas repitam as tarefas que os humanos desempenham. Trata-se da capacidade das máquinas de aprender, reconhecer, perceber e decidir. Isso requer uma grande quantidade de dados de alta qualidade com consistência, integridade, precisão e conformidade (Monteiro et al., 2022).

A IA é um conceito amplo que possui com diversas subáreas, entre elas o ML, que é uma parte da IA, e o DL, uma parte do ML. Este último é o processo pelo qual os computadores desenvolvem o reconhecimento de padrões ou a capacidade de aprender continuamente ou fazer previsões com base em dados e fazer alterações sem serem especificamente programados para isso. Este processo automatiza efetivamente o processo de criação de modelos analíticos e permite que as máquinas se adaptem a novos cenários de forma independente (Monteiro et al., 2022).

A IA é basicamente o estudo de treino de máquinas para imitar um cérebro humano e as suas capacidades de pensamento. O ML, é um subconjunto da IA e permite ao sistema aprender automaticamente por conta própria através das experiências que teve e melhorar de acordo com isso, sem que seja explicitamente programado para tal, e sem a intervenção

humana. Por sua vez, o DL é um subconjunto da família mais ampla do ML que faz uso de redes neuronais para imitar o comportamento do cérebro humano.

Muitos investigadores e desenvolvedores da área acreditam que dentro de poucos anos a IA possa exceder a capacidade dos humanos de aprender ou raciocinar sobre qualquer assunto. No entanto, outros permanecem céticos devido ao facto de toda a atividade cognitiva estar ligada a julgamentos de valor que estão sujeitos à experiência humana. Atualmente, as aplicações da IA são infinitas, desde a área da saúde, financeira, jogos, entre outros.

ML é definido como "o estudo de algoritmos de computador que podem aprender a melhorar o seu desempenho de tarefas com base na experiência passada", ou seja, ML é um campo da IA que se concentra no desenvolvimento de algoritmos e técnicas que permitem que os computadores aprendam a realizar tarefas específicas sem serem explicitamente programados para tal (Miguel, 2022).

As técnicas de ML são baseadas em reconhecimento de padrões, ciência da computação e inferência estatística. Estes métodos funcionam melhor quando visam diretamente campos com uso intensivo de dados, como finanças, economia, medicina entre outros. O objetivo principal geralmente é obter *insights* e previsões valiosas com base em evidências e acontecimentos passados.

Nas abordagens tradicionais de programação, os desenvolvedores escrevem algoritmos que dizem ao computador exatamente como realizar uma determinada tarefa. Por exemplo, se quisermos que um programa reconheça imagens de gatos, precisaríamos de escrever um código que descrevesse as características específicas de um gato e, em seguida, instruísse o computador a procurar essas características nas imagens. No entanto, esta abordagem é limitada, pois requer que os programadores antecipem e codifiquem todas as possíveis situações.

Por outro lado, em ML, em vez de escrever regras específicas, os algoritmos de ML são projetados para aprender com dados de entrada. Estes são treinados num conjunto de exemplos, chamados dados de treino, que contém os dados de entrada e as respostas

correspondentes para cada entrada. O algoritmo analisa esses exemplos e procura padrões e relações entre as entradas e as respostas.

Uma vez que o algoritmo tenha sido treinado, ele pode fazer previsões ou tomar decisões com base em novos dados que não foram usados durante o treino. Em essência, o algoritmo aprende a melhorar o seu desempenho numa determinada tarefa à medida que é exposto a mais dados e ganha experiência.

Atualmente, existem exemplos de aplicações de ML ao nosso redor. Algumas delas são, por exemplo: os *sites* que recomendam produtos, filmes e músicas com base no que compramos, assistimos ou ouvimos anteriormente. Os detetores de *spam* impedem que *emails* indesejados cheguem à nossa caixa de correio. Os sistemas de análise de imagens médicas ajudam os profissionais de saúde a detetar tumores que eles possam ter perdido. Sem esquecer que já existem carros que conduzem sozinhos e estão equipados com tecnologia que evita os acidentes.

2.4.1. Tipos de modelos de *Machine Learning*

Existem quatro tipos de categorias de ML, ilustrados na figura 4 (Miguel, 2022):

1. ***Supervised learning*** **–** nesta categoria utilizam-se dados etiquetados para o treino do modelo, isto é, os dados são rotulados com informações acerca do modelo que se está a construir. Por exemplo, um modelo de visão para computador projetado para identificar cães pastor alemão de raça pura pode ser treinado com um conjunto de dados de várias imagens de cães classificadas com a sua raça. Este tipo de ML não necessita de tantos dados para treinar o modelo, tornando o treino mais fácil. Contudo dados etiquetados corretamente são caros e existe o risco de *overfiting* ou a criação de um modelo tão adaptado aos dados de treino que não consegue lidar com variações de novos dados.

1.1. ***Classification*** **–** Esta técnica tem como objetivo atribuir uma classe ou categoria a uma determinada instância de dados com base nas suas características. Qualquer que seja o tipo de ML, geralmente envolve duas etapas principais: treino e teste.

Durante a etapa de treino, um algoritmo de classificação é alimentado com um conjunto de exemplos etiquetados, chamado de conjunto de treino. Cada exemplo consiste num conjunto de características (também chamadas de atributos) e a classe ou categoria correta associada a essas características. O algoritmo de classificação analisa os exemplos de treino e procura padrões ou relações entre as características e as classes associadas. Com base nesses padrões, o algoritmo constrói um modelo que representa o relacionamento entre as características e as classes. Existem vários algoritmos de classificação populares, como Árvores de Decisão, *Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM)* e Redes Neuronais. Uma vez que o modelo de classificação tenha sido treinado, ele pode ser usado para fazer previsões com novos dados. Durante a etapa de teste, o modelo é avaliado usando um conjunto separado de exemplos, chamado de conjunto de teste, que também contém características, mas as classes corretas não são fornecidas ao modelo. O modelo faz previsões com base nas características dos exemplos de teste e é avaliado comparando as previsões com as classes corretas. A precisão do modelo é medida em termos de quantos exemplos foram classificados corretamente.

1.2. *Regression* - é um método usado para prever valores contínuos com base num conjunto de características. Enquanto a classificação procura atribuir uma classe ou categoria a uma instância de dados, a regressão visa estimar um valor numérico. O processo de regressão também envolve as etapas de treino e teste. Durante a etapa de treino, um algoritmo de regressão é alimentado com um conjunto de exemplos rotulados, semelhante ao processo de classificação. O algoritmo de regressão analisa as características dos exemplos de treino e procura padrões ou relações entre essas características e os valores das variáveis de saída. Com base nessas relações, o algoritmo constrói um modelo matemático que pode fazer previsões para novos dados. Existem vários algoritmos de regressão comumente usados, como regressão linear, regressão logística, regressão polinomial, regressão de árvore de decisão, regressão de vetores de suporte (SVR), entre outros. Cada algoritmo tem as suas próprias suposições e propriedades específicas. Uma vez que o modelo de regressão tenha sido treinado, ele pode ser usado para fazer previsões com novos dados. Durante a etapa de teste, o modelo é avaliado usando um conjunto separado de exemplos, semelhante ao

processo de classificação. O modelo faz previsões com base nas características dos exemplos de teste e é avaliado comparando as previsões com os valores reais das variáveis de saída. Diversas métricas de avaliação são usadas para medir a qualidade do modelo de regressão, como o erro médio quadrático (*Mean Squared Error* - MSE) e o coeficiente de determinação (R^2).

2. ***Unsupervised*** **learning –** nesta categoria são utilizados dados que não são classificados e utiliza algoritmos para classificar os dados em tempo real sem a intervenção humana. Esta categoria está relacionada com a descoberta de padrões nos dados, que possam não ser percetíveis ao olho humano.

2.1. ***Clustering*** **-** é um método utilizado para agrupar dados com base na sua similaridade, sem ter rótulos ou categorias pré-definidas. O objetivo do *clustering* é encontrar estruturas ou padrões intrínsecos nos dados, agrupando-os de forma que instâncias semelhantes sejam colocadas no mesmo grupo e instâncias diferentes sejam colocadas em grupos distintos. No processo de *clustering*, um algoritmo de *clustering* analisa os dados de entrada e identifica padrões ou similaridades entre as instâncias. O algoritmo agrupa as instâncias de acordo com suas características, procurando maximizar a similaridade intra-grupo e minimizar a similaridade inter-grupo. Existem vários algoritmos de *clustering* populares, cada um com suas próprias características e abordagens. Alguns exemplos comuns incluem o algoritmo *K-means*, o algoritmo de agrupamento hierárquico, o algoritmo *DBSCAN (Density-Based Spatial Clustering of Applications with Noise*), entre outros. Estes algoritmos podem diferir em termos de como definem a similaridade entre as instâncias, como atribuem instâncias a grupos e como determinam o número de grupos. Um aspeto importante no *clustering* é a escolha da métrica de similaridade ou dissimilaridade utilizada para comparar as instâncias. Essa métrica pode ser baseada em distâncias euclidianas, distâncias de Manhattan, coeficientes de correlação, entre outros. Após a conclusão do processo de *clustering*, cada grupo formado é considerado um *cluster*. Os *clusters* podem ser visualizados através de gráficos ou representações espaciais, permitindo a análise e interpretação dos resultados obtidos.

2.2. ***Association*** **-** é um método utilizado para descobrir relações ou padrões de coocorrência entre itens num conjunto de dados. Especificamente, o método de associação concentra-se na identificação de associações frequentes entre itens, ou seja, em encontrar itens que são frequentemente comprados, usados ou ocorrem juntos. O objetivo da análise de associação é revelar regras de associação, que são declarações do tipo "se X, então Y". Estas regras descrevem as relações entre diferentes itens ou conjuntos de itens num conjunto de dados. As regras de associação são compostas por duas partes principais: o antecedente (ou conjunto de itens antecedentes) e o consequente (ou conjunto de itens consequentes). A regra é aplicada quando o antecedente é satisfeito e, em seguida, o consequente é previsto ou recomendado. Um algoritmo de associação comumente usado é o *Apriori*. O algoritmo *Apriori* segue uma abordagem iterativa para descobrir regras de associação. Ele explora a propriedade de que um subconjunto de itens frequentes num conjunto de dados também é frequente. O algoritmo começa por encontrar os itens frequentes individualmente (itens que ocorrem acima de um limiar de suporte pré-definido) e, em seguida, gera combinações de itens frequentes para encontrar conjuntos maiores de itens frequentes. A partir desses conjuntos, são extraídas as regras de associação. As métricas comumente usadas para avaliar as regras de associação incluem suporte, confiança e *lift*. O suporte mede a frequência com que uma regra ocorre no conjunto de dados. A confiança mede a probabilidade condicional de que o consequente seja verdadeiro quando o antecedente é verdadeiro. O *lift* é uma medida de associação que compara a frequência observada de uma regra com a frequência esperada se os itens fossem independentes. Uma métrica de *lift* maior que 1 indica que a regra tem uma associação positiva.

3. ***Semi-Supervised learning*** **–** esta categoria oferece um meio termo entre as duas categorias descritas anteriormente, ou seja, aprende com um número pequeno de instâncias etiquetadas e tenta classificar muitos dados não etiquetados. Esta categoria utiliza dois métodos que foram explicados anteriormente: *classification* e *clustering*.

4. ***Reinforcement learning*** **–** lida com a interação de um agente com um ambiente, em que o agente aprende a tomar ações para maximizar uma recompensa

cumulativa ao longo do tempo. É um paradigma inspirado pela forma como os seres humanos e outros organismos aprendem através da tentativa e erro, recebendo *feedback* positivo ou negativo. No *Reinforcement Learning*, o agente não é treinado com exemplos rotulados como no *supervised learning*, nem procura identificar padrões como no *unsupervised learning*. Em vez disso, o agente aprende através de um processo de tentativa e erro, explorando o ambiente, tomando ações e observando as recompensas resultantes. O agente toma decisões num ambiente através de uma política, que é uma estratégia para selecionar ações com base nas informações disponíveis. O objetivo do agente é encontrar uma política que maximize a recompensa cumulativa ao longo do tempo. Para fazer isso, ele precisa de aprender a tomar ações que levem a resultados desejáveis e evitar ações que levem a resultados indesejáveis. O agente recebe *feedback* sobre o desempenho das suas ações na forma de recompensas ou penalidades. A recompensa é uma medida numérica que reflete a qualidade da ação tomada em determinado estado. O agente ajusta a sua política com base nas recompensas recebidas, usando algoritmos de otimização para aprender a escolher as melhores ações em diferentes estados. Ou seja, este método é uma abordagem de ML em que um agente aprende a tomar ações para maximizar uma recompensa cumulativa ao interagir com um ambiente. Ele é útil em situações em que não há exemplos rotulados disponíveis, mas o agente pode explorar o ambiente e aprender através de tentativa e erro.

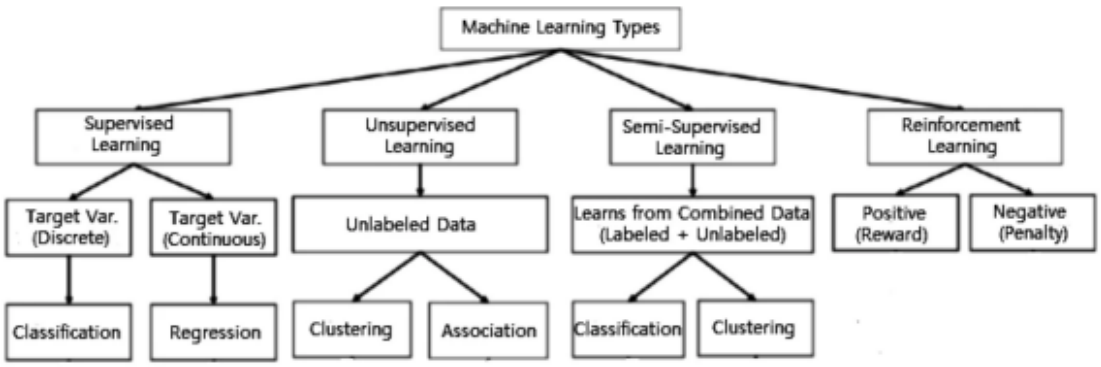


Figura 4 - Tipos de Machine Learning

2.4.2. Terminologia fundamental

Esta secção apresenta e explica algumas terminologias concetuais multidisciplinares relevantes que se aplicam ao ML para melhorar a compreensão.

***Dataset*:** Um conjunto de dados que está em conformidade com um esquema sem requisitos de ordenação. Num conjunto de dados típico, cada coluna representa um atributo e cada linha representa um membro do conjunto de dados.

Accuracy: É uma métrica para avaliar os modelos de classificação. É calculada utilizando uma fração das previsões do modelo, como se segue: $\frac{TP+TN}{TP+TN+FP+FN}$ [1]

***Precision*:** A precisão é, intuitivamente, a capacidade do classificador de não rotular como positiva uma amostra que é negativa. É calculada com base na seguinte expressão: $\frac{TP}{TP+FP}$

Recall: O *recall* determina a capacidade para encontrar amostras positivas. É calculado com base na seguinte expressão: $\frac{TP}{TP+FN}$

F1-score: Utilizado em problemas de classificação que combina *precision* e *recall* numa única pontuação, fornecendo uma forma mais abrangente de avaliar o desempenho do modelo. O *F1* é calculado através da seguinte fórmula: $\frac{2*(Precision*Recall)}{(Precision+Recall)}$

***Root-mean-square error* (RMSE):** É definido como o erro quadrático médio (erros de previsão). Os resíduos medem a distância a que os pontos de dados se encontram da linha de regressão. O RMSE é uma medida do grau de dispersão destes resíduos. Por outras palavras, indica a robustez dos dados.

R-Squared (R^2): É definido como o coeficiente de determinação e utilizado em análise de regressão para avaliar o quão bem o modelo se ajusta aos dados observados. Esta

[1] TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

métrica fornece uma medida da proporção da variação na variável dependente que é explicada pelas variáveis independentes. O R^2 varia entre 0 e 1, em que 0 indica que o modelo não consegue explicar nenhuma variação nos dados, enquanto 1 indica que o modelo explica toda a variação nos dados e ajusta-se perfeitamente aos pontos.

ROC Area Under the Curve (ROC AUC): É uma métrica utilizada para avaliar a qualidade do desempenho de modelos de classificação binária. A área sob a curva *ROC* mede a capacidade global do modelo distinguir entre classes positivas e negativas. Quanto maior for a área sob a curva *ROC*, melhor o modelo está em separar as duas classes.

Confusion matrix: A matriz de confusão é uma medida abrangente para resolver problemas de classificação. Pode ser utilizada tanto para problemas de classificação binária como de classificação multiclasse. As matrizes de confusão representam a soma das contagens previstas e reais. O resultado "TN" significa *True Negative* (Verdadeiro Negativo) e representa o número de exemplos negativos corretamente classificados. Do mesmo modo, "TP" significa *True Positive* (Verdadeiro Positivo) e representa o número de exemplos positivos corretamente classificados. A abreviatura "FP" refere-se ao número de exemplos negativos reais classificados como positivos, enquanto "FN" refere-se ao número de exemplos positivos reais classificados como negativos. Na Figura 5 está representada esta matriz.

	Predicted Class	
True Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Figura 5 - Confusion Matrix

3. Preparação e manipulação dos dados

Este capítulo tem como intuito descrever e explicar todos os passos e etapas que foram seguidas ao longo de todo o projeto. Primeiramente, serão abordados quais os dados adquiridos para a realização do projeto, a construção do(s) *dataset(s)* utilizado(s) e a apresentação e explicação de alguns gráficos desenvolvidos. De seguida, serão apresentadas as técnicas de análise de dados aplicadas e quais os modelos de ML que foram utilizados e treinados, e os seus respetivos resultados. Por fim, será selecionado o melhor modelo e realizada a sua avaliação com a utilização de diversas métricas.

Ao contrário dos sistemas de recomendação de produtos que já existem atualmente e são altamente eficazes em lojas virtuais, em que o objetivo é sugerir ao cliente produtos para além daqueles já adicionados ao carrinho, com base nas suas visualizações, histórico de compras e padrões de navegação, o intuito da realização deste projeto é ligeiramente diferente.

O objetivo primordial deste projeto consiste em utilizar dados reais fornecidos por uma organização para desenvolver um sistema que seja capaz de antecipar se uma sessão de compras vai ser terminada com sucesso ou não, ou seja, se a sessão vai ser abandonada pelo cliente antes de efetivar a compra, ou se o cliente vai efetivamente realizar a compra. Este estudo visa empregar informações concretas e históricas para criar modelos analíticos ou algoritmos que possam prever com precisão o comportamento dos consumidores, permitindo à organização tomar medidas proativas para reduzir o abandono de sessões e melhorar a eficiência do processo de compra *online*.

3.1. *Business Understanding*

Na fase de *Business Understanding* da *framework* CRISP-DM, é crucial compreender profundamente os objetivos e requisitos do negócio antes de iniciar qualquer análise de dados. Neste projeto, o objetivo principal é prever se uma sessão será abandonada ou não, utilizando dados históricos de produtos e sessões de clientes fornecidos pela empresa.

Para isso, inicia-se esta etapa por definir claramente o problema de negócio: o abandono de sessões, que pode resultar em perda de potenciais vendas e insatisfação do cliente. A

previsão precisa deste comportamento pode permitir a implementação de estratégias proativas para reter clientes e aumentar a conversão.

3.2. *Data Understanding* – 1º *dataset*

Os dados adquiridos para a realização deste projeto pertencem a uma organização de venda *online* de produtos, cujas sessões dos utilizadores não são, muitas vezes, terminadas com sucesso. Desta forma, os dados adquiridos inicialmente são dados relacionados com os produtos ativos da loja e as sessões dos utilizadores. Todos estes dados foram analisados e recolhidos de forma a serem úteis e prestáveis para construir um *dataset* que seja possível de utilizar aquando do desenvolvimento e teste dos modelos produzidos.

Após a recolha inicial dos dados, criou-se um conjunto de dados que incluía as seguintes colunas: identificação da sessão, loja, plataforma de acesso, sucesso e identificação do produto. Cada linha deste *dataset* representa a interação do cliente com o respetivo produto e se a sessão foi terminada com sucesso ou não.

A tabela 1 representa a análise descritiva realizada a este conjunto de dados:

Tabela 1 - Análise descritiva dos dados

Coluna	**Tipo de dados**	**Descrição**
Id_sessão	*String*	Código identificador da sessão
Loja	*String*	Identificador do país da sessão
Plataforma	*String*	Identificador da plataforma da sessão
Produto	*Int*	Representa o produto
Sucesso	*Int*	Identifica se a compra foi efetuada ou não

Com o objetivo de realizar uma análise mais profunda e detalhada aos dados, foram construídos alguns gráficos que pretendem demonstrar como os dados estão distribuídos e organizados no conjunto de dados.

O gráfico circular (Figura 6) representa visualmente a distribuição dos dados na coluna 'plataforma'. Ao analisar este gráfico, torna-se evidente que a maioria das sessões é iniciada através da plataforma *mobile-web*, correspondendo a 64,9% do total. Em segundo lugar, a plataforma *web* contribui com 22,7% dos dados, enquanto a plataforma *mobile-app* representa apenas 12,4% das sessões iniciadas.

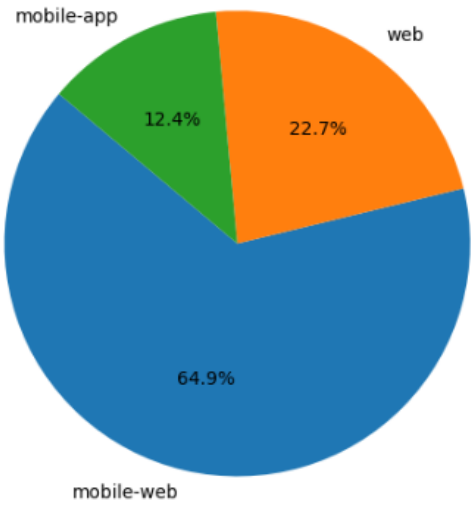


Figura 6 - Distribuição dos dados na coluna 'plataforma'

O próximo gráfico circular (Figura 7) ilustra como os dados estão distribuídos na coluna 'loja'. Este gráfico revela que a maior parte das sessões tem origem em Itália, representando 27,1% do total, seguido por Portugal, que contribui com 23,7%. Espanha é responsável por 19% das sessões, enquanto a Alemanha representa 11,7%, e França fica com 18,5% das sessões iniciadas.

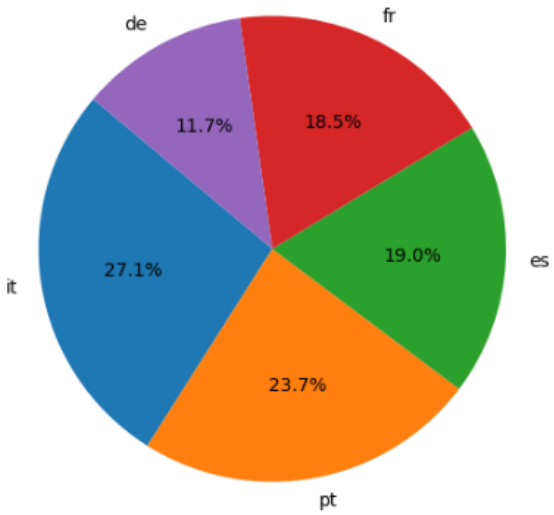


Figura 7 - Distribuição dos dados na coluna 'loja'

O terceiro gráfico elaborado (Figura 8) é um gráfico de barras que exibe a contagem total de ocorrências dos valores 0 e 1 na coluna 'sucesso'. Há um total de 118.239 registos com o valor 0 (aproximadamente 92%) e 9.984 registos com o valor 1 (aproximadamente 8%). Através destas evidências, é possível afirmar que este *dataset* é desbalanceado, ou seja, uma grande parte dos valores que existem nesta coluna correspondem ao valor 0, que significa que as sessões dos clientes não são terminadas com sucesso.

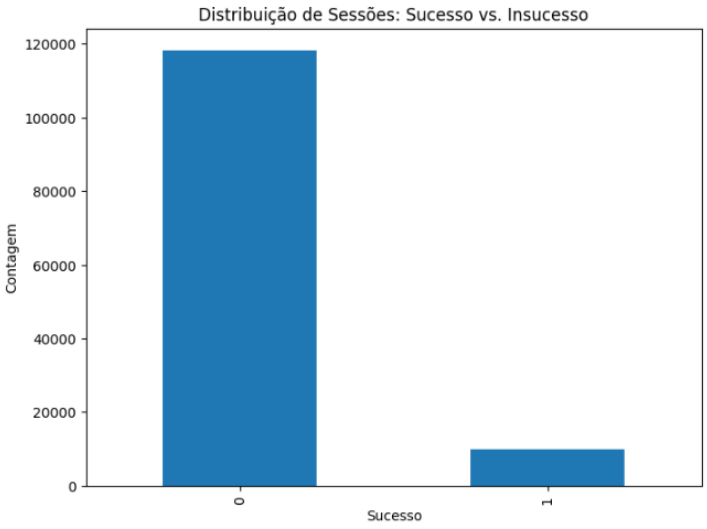


Figura 8 - Distribuição dos dados na coluna 'sucesso'

3.3. *Data Preparation* – 1º *dataset*

No entanto, a principal intenção ao estruturar o conjunto de dados a ser utilizado nos modelos de ML era reinterpretar as colunas como se fossem os próprios produtos. Assim, efetuaram-se algumas transformações no conjunto de dados original para atingir este objetivo.

Desta forma, foi construído um *dataset* composto por 6504 colunas e 128915 linhas. As colunas deste *dataset* representam todos os produtos ativos da loja enquanto as linhas representam as interações da sessão do utilizador com determinado produto. Caso a linha esteja preenchida com o valor 1, significa que o utilizador interagiu com aquele produto, ou seja, pode ter visualizado o produto, adicionado ou removido o produto do carrinho. Caso a linha esteja preenchida com o valor 0, significa que o utilizador não teve nenhuma interação com aquele produto.

No processo de treino dos modelos de ML, foi decidido que, se alguma das colunas estivesse completamente preenchida com valores iguais a zero, essa coluna seria excluída e não seria considerada no treino do respetivo modelo. Desta forma, o conjunto de dados passa a ser composto por um total de 3797 colunas.

Devido ao facto de existir uma grande diferença de registos com valor 0 e 1, utilizou-se a técnica de *underslamping* que tem como objetivo balancear o *dataset*, mantendo o número de registos da classe minoritária (classe 1) e diminuindo ao número de registos da classe maioritária (classe 0). Assim, o conjunto de dados passou a ser constituído por 9670 registos correspondentes à classe 0 e 9601 correspondentes à classe 1.

3.4. Modeling – Modelos de classificação - 1º dataset

Os primeiros modelos treinados consistem num problema de regressão, em que o objetivo é indicar a probabilidade de uma sessão terminar com sucesso, ou seja, finalizar a compra. Para tal, foram utilizados dois modelos de ML: *Random Forest* e Regressão Linear e foram treinados utilizando as configurações *default.* Os resultados obtidos com o treino dos modelos, encontram-se identificados na tabela 2:

Tabela 2 - Resultados obtidos para os modelos de RF e RLI

Modelo	**RMSE**	**R^2**
RF	0.12	0.08
RLI	0.26	0.07

Os resultados obtidos com estes dois modelos não foram os esperados. Apesar de o valor 0.12 de RMSE ser relativamente bom, o valor de 0.08 de R^2 torna o modelo um pouco limitado visto que este valor indica que aproximadamente 8% da variabilidade dos dados é explicada pelo modelo, significando que o mesmo tem uma capacidade reduzida de explicar ou prever a variação dos dados. Em termos práticos, um R^2 de 0.08 indica que o modelo não está ajustado aos seus dados e que os mesmos podem ser altamente dispersos e difíceis de prever com base nas variáveis independentes. Devido ao facto de os resultados apresentados não serem os melhores, as configurações utilizadas inicialmente como *default* para treinar os modelos bastaram para perceber que não seria o caminho ideal.

3.5. Modeling – Modelos de regressão – 1º dataset

Desta forma, decidiu-se treinar modelos de classificação ao invés de regressão, como descrito anteriormente. Assim sendo, o objetivo é prever se a sessão vai ser terminada com sucesso (1) ou não (0).

Para o treino destes modelos foi utilizada a técnica de *k-fold cross-validation* com 5 *splits*. Neste caso específico, o conjunto de dados é dividido em 5 partes de aproximadamente tamanhos iguais. O processo de *cross-validation* com 5 *splits* ocorre da seguinte forma:

1. **Divisão do conjunto de dados:** O conjunto de dados é dividido em 5 partes iguais. Cada uma dessas partes é designada de *fold*;
2. **Treino e avaliação:** O modelo é treinado e avaliado 5 vezes, em que cada vez se utiliza uma combinação diferente de 4 *folds* para treino e o *fold* restante para teste. Isto significa que em cada iteração, existe um conjunto de treino e de teste diferentes, como exemplificado na Figura 9;

3. **Métricas de avaliação:** Após cada treino, o desempenho do modelo é avaliado utilizando métricas relevantes;
4. **Média das métricas:** No final das 5 iterações, as métricas de desempenho obtidas em cada *fold* de teste são geralmente médias para fornecer uma estimativa mais robusta do desempenho geral do modelo.

Figura 9 - Processo de cross-validation com 5 splits

Os modelos treinados e os respetivos resultados encontram-se expressos na tabela 3.

Tabela 3 - Resultados obtidos para os modelos RF, RLO, XGB, LGBM, NB e CB com utilização da técnica cross-validation

Modelo	***Accuracy***	***Precision***	***Recall***	***F1-score***	***ROC AUC***
RF	0.854	0.451	0.158	0.234	0.767
RLO	0.858	0.494	0.155	0.236	0.777
XGB	0.859	0.504	0.156	0.238	0.787
LGBM	0.861	0.537	0.106	0.178	0.792
NB	0.783	0.281	0.343	0.309	0.712
CB	0.860	0.515	0.160	0.244	0.791

Pela análise dos resultados expostos na tabela acima, é possível verificar que em todos os modelos os valores da *accuracy* são bastante bons, todos eles acima de 80%, com exceção do modelo NB. No entanto, estes resultados não são os expectáveis, visto que, os modelos

não conseguem prever as situações em que a sessão vai ser terminada com sucesso (1), como se verifica pelos baixos valores do *recall* existentes em todos os modelos treinados.

De forma a analisar melhor os valores obtidos em cada uma das métricas dos modelos treinados, foram criados 5 *boxplots* em que cada um tem a distribuição de uma métrica para cada *fold* do *cross*-validation. Para o desenvolvimento destes gráficos, foi utilizada a configuração *default* de cada modelo treinado.

O primeiro gráfico, representado na Figura 10, mostra os valores de *accuracy* obtidos para os modelos treinados. Estes resultados indicam que o modelo LGBM obteve a maior precisão média (0.861), seguido de perto pelos modelos XGB e CB. O modelo NB teve a menor precisão média (0.783) entre os modelos testados. É importante observar que, além da precisão média, o desvio padrão é uma medida importante, pois indica a variabilidade dos resultados. Neste caso, todos os modelos têm desvios padrão bastante baixos (0.002 a 0.004), o que sugere que os resultados são consistentes.

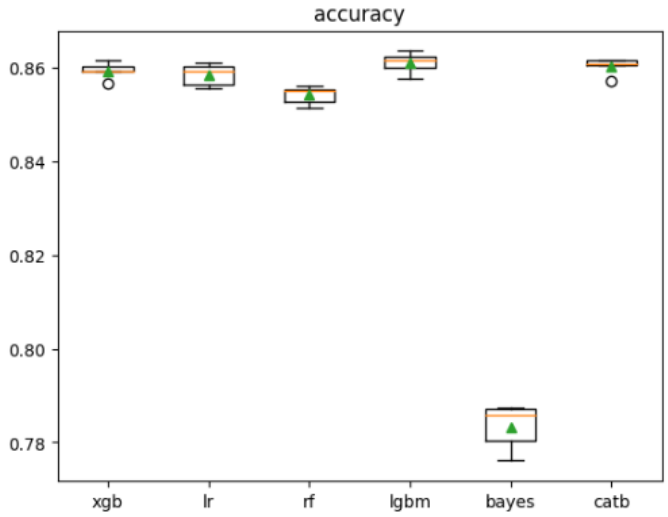


Figura 10 - Boxplot com os valores de accuracy obtidos em cada modelo

A análise da Figura 11 indica que o modelo LGBM obteve a maior precisão média (0.537), seguido pelo modelo XGB. O modelo NB teve a menor precisão média (0.281) entre os modelos testados. Os desvios padrão indicam a variabilidade dos resultados, e o modelo LGBM também apresentou o maior desvio padrão (0.034), sugerindo maior variabilidade nas medidas de precisão. Por outro lado, o modelo NB teve um desvio padrão bastante baixo (0.010), indicando menor variabilidade nas medidas de precisão.

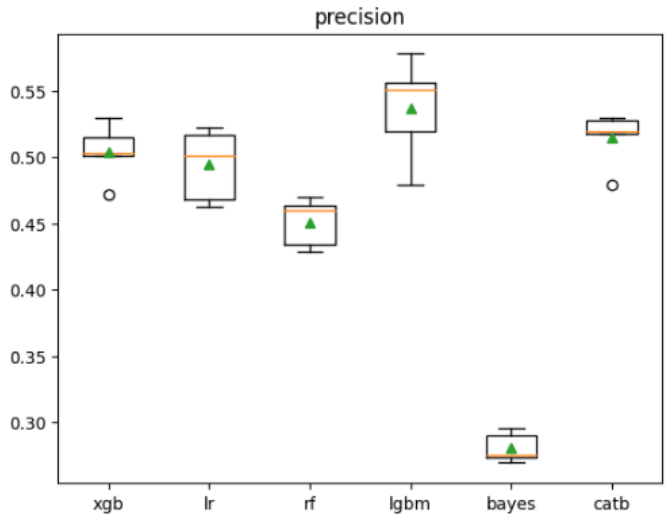


Figura 11 - Boxplot com os valores de precision obtidos em cada modelo

Com base nos resultados representados na Figura 12, o modelo NB obteve o *recall* mais alto (0.343), indicando uma boa capacidade de identificar positivos verdadeiros. O modelo LGBM teve o *recall* mais baixo (0.106) entre os modelos testados. Os desvios padrão, neste caso, são relativamente baixos no geral, o que sugere consistência nos resultados de *recall.*

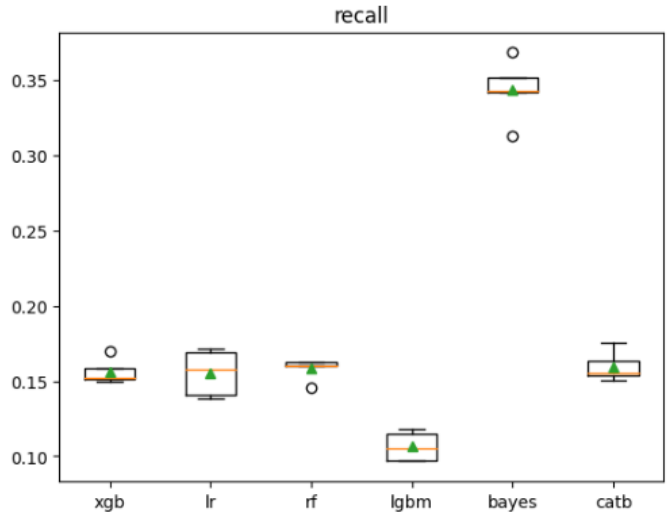


Figura 12 - Boxplot com os valores de recall obtidos em cada modelo

Com base nos resultados da Figura 13, o modelo NB obteve o *F1-score* mais alto (0.309), indicando um bom equilíbrio entre *precision* e *recall*. O modelo LGBM teve o *F1-score*

mais baixo (0.178) entre os modelos testados. Em geral, os desvios padrão são relativamente baixos, sugerindo consistência nos resultados *de F1-score.*

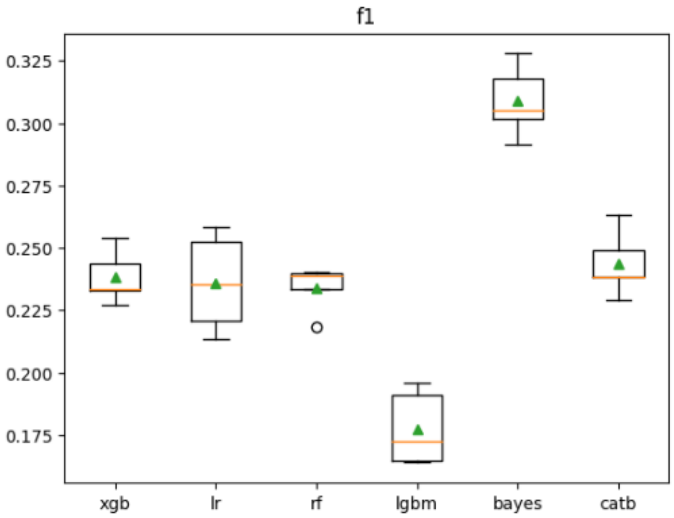


Figura 13 - Boxplot com os valores de f1-score obtidos em cada modelo

Com base nos resultados apresentados na Figura 14, o modelo LGBM obteve o ROC AUC médio mais alto (0.792), indicando um bom desempenho na separação das classes. O modelo NB teve o ROC AUC médio mais baixo (0.712) entre os modelos testados. Em geral, os desvios padrão são relativamente baixos, sugerindo consistência nos resultados de ROC AUC.

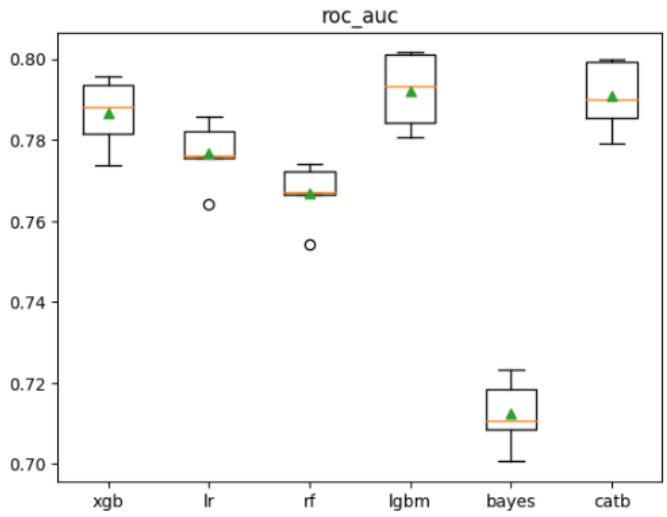


Figura 14 - Boxplot com os valores de ROC-AUC obtidos em cada modelo

Como referido e exposto anteriormente, os resultados obtidos no treino dos modelos anteriores não são os resultados pretendidos, visto que o *recall*, a métrica mais importante

para este projeto pois é o que indica a previsão das situações em que a sessão termina com sucesso, estão muito abaixo do expectável e pretendido para este projeto.

3.6. *Data Understanding – 2º dataset*

Desta forma, foi seguida uma nova abordagem com um *dataset* diferente do utilizado até aqui, com o intuito de alcançar melhores resultados nos modelos treinados.

Para além deste problema dos baixos resultados obtidos no treino dos modelos anteriores, o conjunto de dados utilizado é bastante grande, visto que as colunas são compostas por todos os produtos ativos da loja. Isto faz com que o treino dos modelos seja muito demorado, sendo que foi um dos pontos principais que levou à alteração do *dataset*. Assim sendo, foi desenvolvido um conjunto de dados um pouco diferente daquele utilizado até ao momento. O *dataset* utilizado anteriormente era constituído por um elevado número de colunas (que correspondiam aos produtos ativos na loja) e um pequeno número de linhas. Este último era relativamente pequeno devido ao facto de existirem recursos limitados no treino dos modelos, ou seja, sempre que se tentou treinar os modelos com mais linhas, mantendo o número de colunas, o treino não terminava com sucesso devido ao facto de o *cluster* utilizado ter poucos recursos.
Deste modo, no novo *dataset* desenvolvido, passou-se a ter categorias de produtos ao invés dos produtos, diminuindo significativamente o número de colunas deste novo conjunto de dados (devido ao facto de o número de categorias ser bastante inferior ao número de produtos existentes). Como se conseguiu diminuir o número de colunas, foi possível aumentar o número de linhas, de forma a fornecer mais informação no treino do modelo, para este conseguir obter bons e melhores resultados.

Por esse motivo, foi criado um conjunto de dados que incluía as seguintes colunas: identificação da sessão, loja, plataforma de acesso, sucesso e identificação da categoria do produto.

A tabela 4 representa a análise descritiva realizada a este conjunto de dados.

Tabela 4 - Análise descritiva dos dados

Coluna	**Tipo de dados**	**Descrição**
Id sessão	*String*	Código identificador da sessão
Loja	*String*	Identificador do país da sessão
Plataforma	*String*	Identificador da plataforma da sessão
Categoria	*Int*	Representa as categorias dos produtos
Sucesso	*Int*	Identifica se a compra foi efetuada ou não

Com o objetivo de realizar uma análise mais profunda e detalhada aos dados, foram construídos alguns gráficos que pretendem demonstrar como os dados estão distribuídos e organizados no conjunto de dados.

O gráfico circular apresentado na Figura 15 oferece uma representação visual da forma como os dados estão distribuídos na categoria 'plataforma'. Ao examinar esta representação gráfica, fica claro que a grande maioria das sessões é originária da plataforma *mobile-web*, constituindo 66,9% do total. Em segundo lugar, a plataforma *web* contribui com uma fração de 20,8% dos dados, enquanto a plataforma *mobile-app* representa apenas 12,3% das sessões iniciadas.

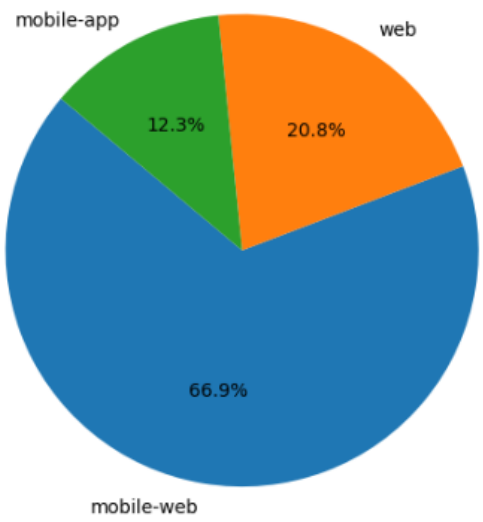


Figura 15 - Distribuição dos dados na coluna 'plataforma'

No gráfico circular seguinte (Figura 16), é possível visualizar a distribuição dos dados na categoria 'loja'. Os dados revelam que a maioria das sessões provém de Itália, com uma representação de 27,3% do total, seguida de Portugal, que contribui com 23,5%. França ocupa a terceira posição, sendo responsável por 20,6% das sessões, enquanto a Espanha representa 19% e a Alemanha corresponde a 9,6% das sessões iniciadas.

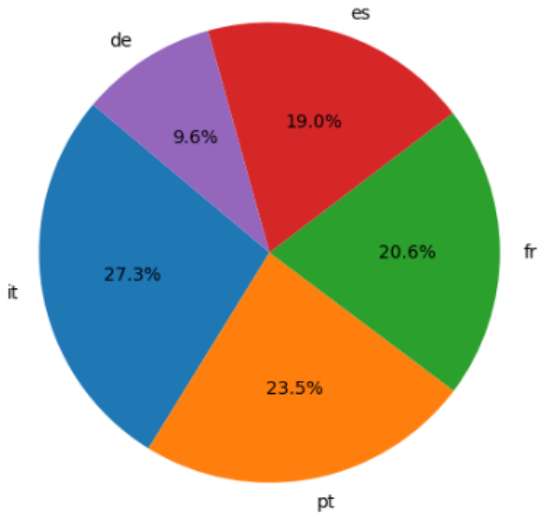


Figura 16 - Distribuição dos dados na coluna 'loja'

No gráfico apresentado na Figura 17, é exibida uma representação em formato de gráfico de barras que mostra o número total de ocorrências dos valores 0 e 1 na coluna 'sucesso'. Observa-se que existem 1.259.453 registos com o valor 0, o que equivale a aproximadamente 92,5% do total, e 101.049 registos com o valor 1, correspondendo a

aproximadamente 7,5% do total. Com base nestas constatações, pode-se concluir que este conjunto de dados possui um desequilíbrio significativo, denominando-se um *dataset* desbalanceado.

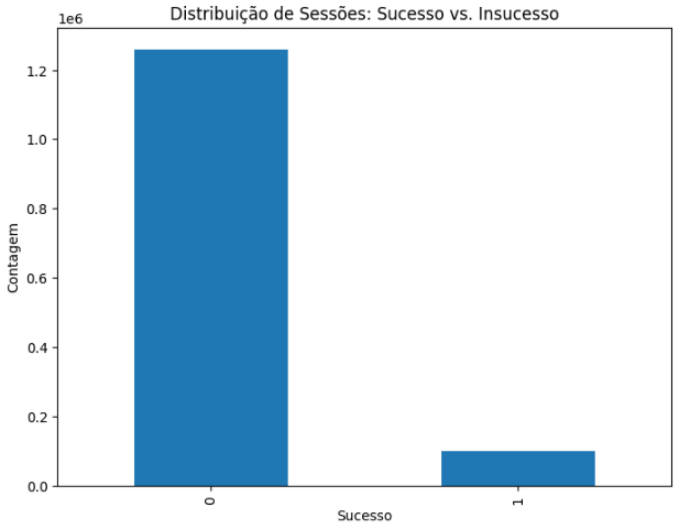


Figura 17 - Distribuição dos dados na coluna 'sucesso'

3.7. *Data Preparation – 2º dataset*

No entanto, tal como no primeiro *dataset* construído, o objetivo era reinterpretar as colunas como se fossem as próprias categorias dos produtos. Assim, efetuaram-se algumas alterações neste conjunto de dados para atingir este objetivo.

E, assim sendo, o segundo *dataset* é constituído por 529 colunas e 1 360 502 linhas. As diferenças existentes entre os dois conjuntos de dados é que, enquanto as colunas do primeiro se referem aos produtos, neste segundo *dataset* as suas colunas correspondem às categorias dos produtos. Ao invés de representar as linhas como sendo as interações dos clientes com os produtos, neste *dataset*, as linhas apresentam a pontuação da sessão do cliente por categoria dos produtos.

No treino dos modelos de ML, assim como no caso dos modelos iniciais, foi determinado que se alguma coluna estivesse preenchida exclusivamente com zeros, essa coluna seria

removida e não seria levada em consideração durante o treino do modelo correspondente. Desta forma, o conjunto de dados passa a ser composto por um total de 529 colunas.

Tal como o *dataset* anteriormente utilizado, também este se encontra desbalanceado. Com o objetivo de resolver este problema, foi utilizada uma técnica de *underslamping* denominada *OneSideSelection.* A principal ideia desta técnica é equilibrar o conjunto de dados removendo amostras da classe maioritária (0) de forma seletiva, de modo a preservar a informação essencial para o problema. A principal vantagem do *OneSideSelection* é que ele tenta preservar informações valiosas da classe maioritária, reduzindo ao mesmo tempo o desequilíbrio de classe.

O processo de *undersampling* utilizando a técnica *OneSideSelection* ocorre da seguinte forma:

1. **Identificação de amostras ruidosas:** Primeiro, o algoritmo avalia o conjunto de dados e identifica quais exemplos da classe maioritária podem ser considerados "ruidosos" ou potencialmente mal rotulados. Esses exemplos são geralmente aqueles que estão próximos da fronteira de decisão entre as duas classes;
2. **Subamostragem seletiva:** Após a identificação das amostras ruidosas, o algoritmo remove algumas dessas amostras, tentando manter um equilíbrio melhor entre as duas classes. A ideia é remover seletivamente as amostras que têm menos probabilidade de serem representativas da classe maioritária;
3. **Treino do modelo:** Com o conjunto de dados resultante após a subamostragem seletiva, um modelo de ML é treinado.

É importante referir que antes de se utilizar a técnica de *underslamping* foram retirados 20% do número de linhas verdadeiras do *dataset* para depois serem utilizadas no teste do modelo, com o objetivo de utilizar algumas linhas verdadeiras na avaliação do modelo, para se conseguir obter um resultado mais próximo da realidade. O número de linhas verdadeiras retiradas antes de realizar a técnica *underslamping* foram 272100.

Após a utilização da técnica de *undersampling* mencionada acima, o conjunto de dados passou a ser constituído por 341916 registos correspondentes à classe 0 e 101876 correspondentes à classe 1.

De forma a obter um *dataset* ainda mais balanceado, utilizou-se uma técnica de *oversampling* denominada SMOTE que consiste numa abordagem que lida especificamente com o problema em que a classe minoritária (1) é sub-representada em comparação com a classe maioritária (0).

A técnica SMOTE funciona da seguinte forma:

1. **Seleção de exemplos:** Primeiro, o SMOTE, seleciona aleatoriamente um exemplo da classe minoritária do conjunto de dados;
2. **Geração de exemplos:** Em seguida, escolhe aleatoriamente um ou mais vizinhos próximos a esse exemplo (os vizinhos são escolhidos com base na distância euclidiana ou outra medida de similaridade). A partir desses vizinhos, o SMOTE cria exemplos interpolando características entre o exemplo original e os vizinhos selecionados;
3. **Incorporação dos exemplos:** Os exemplos criados são adicionados ao conjunto de dados, aumentando assim a representação da classe minoritária.

O principal objetivo ao utilizar esta técnica de *oversampling* é aumentar o número total de registos que correspondem à classe 1 e, assim, obter um *dataset* balanceado e com um *ratio* de 1:1 entre registos correspondentes às classes 0 e 1. Desta forma, após a utilização da técnica SMOTE, o conjunto de dados é composto por 341916 registos correspondentes à classe 0 e à classe 1.

3.8. *Modeling* – Modelos de Regressão – 2º *dataset*

Após as alterações ao conjunto de dados referidas anteriormente, iniciou-se o treino dos modelos para problemas de classificação. Para o treino destes modelos, também foi utilizada a técnica de *cross-validation* referida anteriormente. Neste caso foram treinados

os mesmos 6 modelos, utilizados anteriormente: RF, RLO, XGB, LGBM, NB e CB. A tabela 5 apresenta os resultados que foram obtidos no treino destes modelos:

Tabela 5 - Resultados obtidos nos modelos RF, RLO, XGB, LGBM, NB e CB

Modelo	***Accuracy***	***Precision***	***Recall***	***F1-score***	***ROC AUC***
RF	0.860	0.315	0.432	0.364	0.816
RLO	0.861	0.347	0.567	0.431	0.821
XGB	0.894	0.422	0.375	0.397	0.827
LGBM	0.894	0.426	0.402	0.414	0.847
NB	0.858	0.285	0.349	0.313	0.790
CB	0.899	0.451	0.391	0.419	0.842

Pela análise dos resultados obtidos e, comparando com os resultados obtidos anteriormente com um conjunto de dados distinto, verifica-se que os valores obtidos na métrica do *recall* foram bastante superiores em todos os modelos treinados mostrando que, com este *dataset* os modelos conseguem prever com mais exatidão as situações em que a sessão vai ser terminada com sucesso (1).

Tal como efetuado para o primeiro *dataset*, foram desenvolvidos 6 *boxplots* em que cada um deles mostra os valores obtidos em cada métrica do treino do modelo com o objetivo de perceber o valor médio obtido e a variabilidade dos dados existentes. Com base nos resultados representados na Figura 18, modelo RLO teve o *recall* mais alto (0.567) entre os modelos testados indicando uma boa capacidade de identificar positivos verdadeiros. O modelo NB obteve o *recall* mais baixo (0.349) entre os modelos testados, por outro lado, verifica-se que foi o modelo que apresentou valores mais altos de desvio padrão (0.012). O valor de desvio padrão mais baixo (0.003) apresentado corresponde ao modelo RLO.

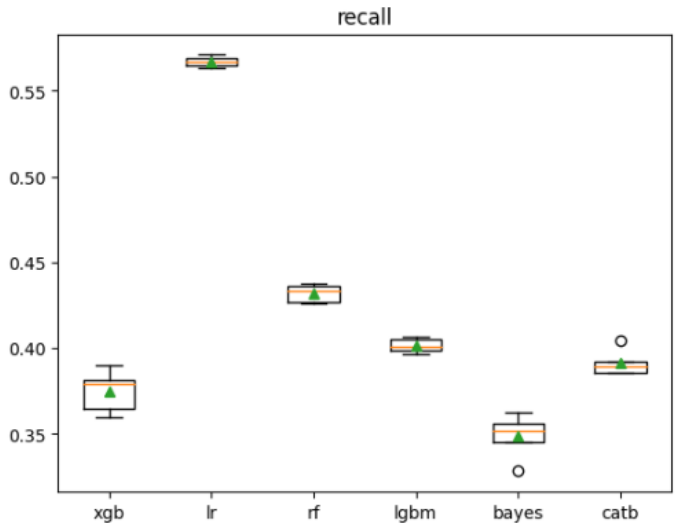


Figura 18 - Boxplot com os valores de recall obtidos em cada modelo

A análise da Figura 19 indica que o modelo CB obteve a maior precisão média (0.451), seguido pelos modelos LGBM e XGB. O modelo NB teve a menor precisão média (0.285) entre os modelos testados. Os desvios padrão indicam a variabilidade dos resultados, e o modelo RF apresentou o maior desvio padrão (0.009), sugerindo maior variabilidade nas medidas de precisão.

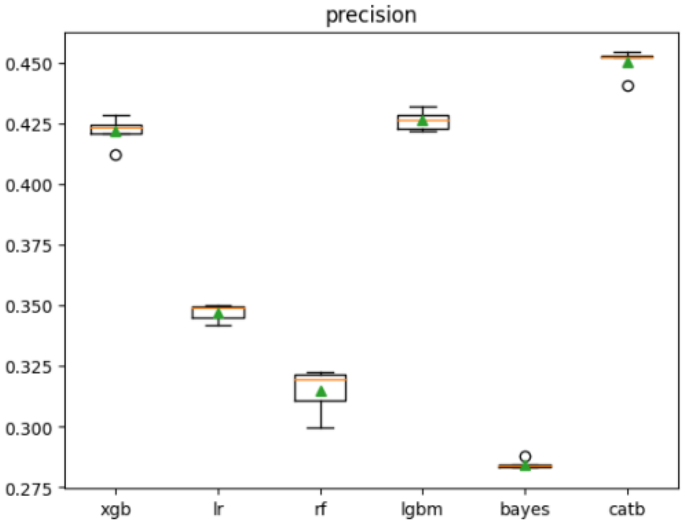


Figura 19 - Boxplot com os valores de precision obtidos em cada modelo

O gráfico representado na Figura 20, mostra os valores de *accuracy* obtidos para os modelos treinados. Estes resultados indicam que o modelo CB obteve a maior *accuracy* média (0.899), seguido pelos modelos XGB e LGBM. O modelo NB teve a menor

accuracy média (0.858) entre os modelos testados. Neste caso, o modelo com maior desvio padrão foi o RF (0.003).

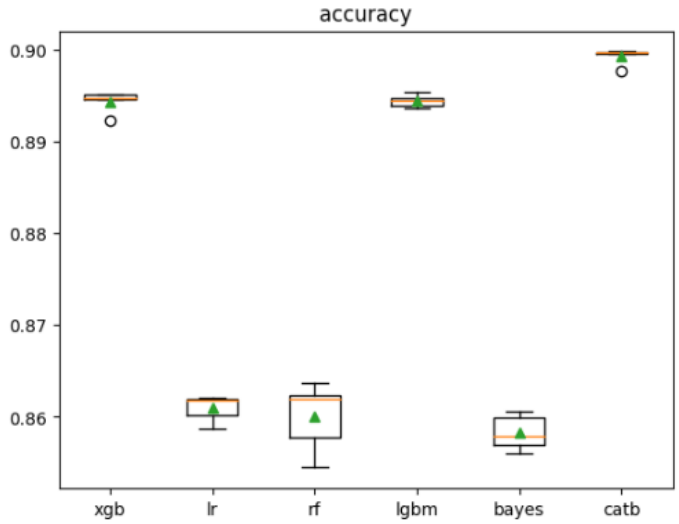


Figura 20 - Boxplot com os valores de accuracy obtidos em cada modelo

Com base nos resultados da Figura 21, o modelo RLO obteve o *F1-score* mais alto (0.431), indicando um bom equilíbrio entre *precision* e *recall*. O modelo NB teve o *F1-score* mais baixo (0.313) entre os modelos testados. Os valores dos desvios padrão variam entre 0.003 (modelos RLO) e 0.007 (modelo XGB).

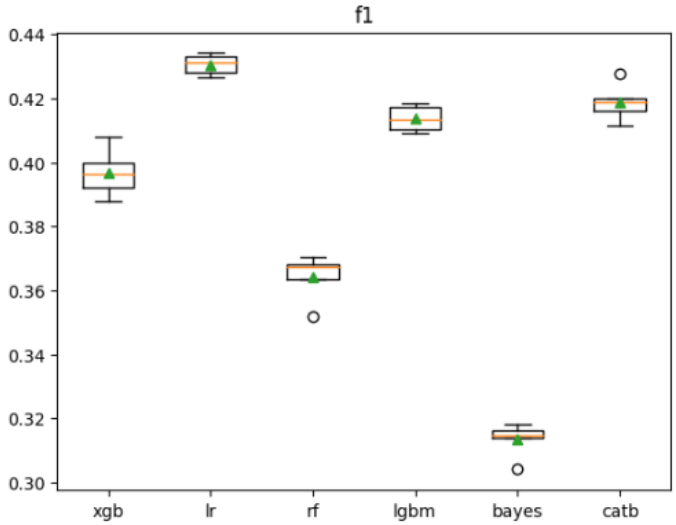


Figura 21 - Boxplot com os valores de F1-score obtidos em cada modelo

A análise da Figura 22 indica que o modelo LGBM obteve o ROC AUC médio mais alto (0.847), indicando um bom desempenho na separação das classes. O modelo NB teve o

ROC AUC médio mais baixo (0.790) entre os modelos testados. Em geral, os desvios padrão são relativamente baixos, sugerindo consistência nos resultados de ROC AUC.

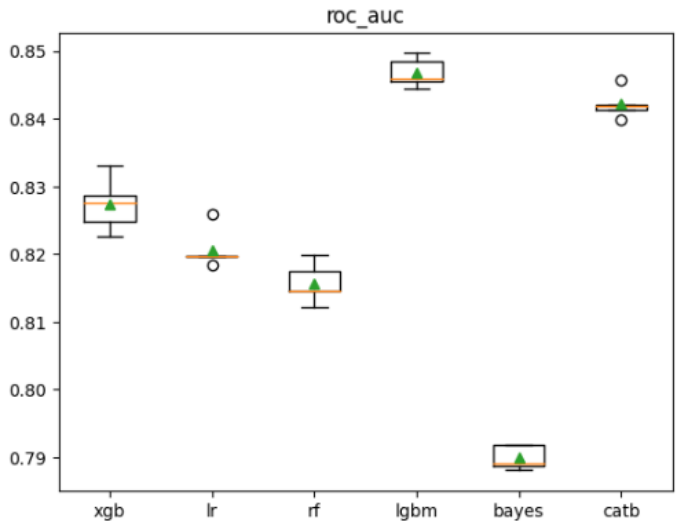


Figura 22 - Boxplot com os valores de ROC-AUC obtidos em cada modelo

Após a análise dos resultados obtidos, verifica-se que o modelo RLO é o modelo que apresenta melhores resultados. Assim, com o treino dos modelos terminado e com a seleção do melhor, o próximo passo é o teste do modelo.

3.9. *Evaluation*

De modo a avaliar a *performance* do modelo escolhido, desenvolveram-se alguns gráficos que mostram o comportamento do modelo RLO com dados de teste que foram anteriormente retirados do *dataset* antes de realizar a técnica *underslamping.* O primeiro gráfico representa a *confusion matrix* que é uma ferramenta fundamental na avaliação de algoritmos de classificação. A *confusion matrix* permite a visualização do desempenho de um modelo de classificação, comparando as previsões feitas pelo modelo com as classes reais conhecidas. Neste caso, pela análise da Figura 23 é possível constatar que a *confusion matrix* está organizada numa tabela 2x2, onde existem duas classes possíveis: 0 e 1. Com base nos valores da *confusion matrix* obtidos, é possível concluir o seguinte:

- TP: Há 230844 casos em que o modelo previu corretamente a classe 0 e a amostra realmente pertencia a essa classe. Isso significa que o modelo identificou com sucesso 230844 instâncias.

- FP: Existem 21021 casos em que o modelo previu incorretamente a classe 1 quando a amostra na verdade pertencia à classe 0. Isto representa erros em que o modelo classificou erradamente amostras negativas como positivas.
- FN: Há 8904 casos em que o modelo previu incorretamente a classe 0 quando a amostra na verdade pertencia à classe 1. Estes são erros em que o modelo não conseguiu identificar amostras positivas.
- TN: Existem 11325 casos em que o modelo previu corretamente a classe 1 e a amostra realmente pertencia a essa classe. Isto indica que o modelo identificou com sucesso 11325 instâncias.

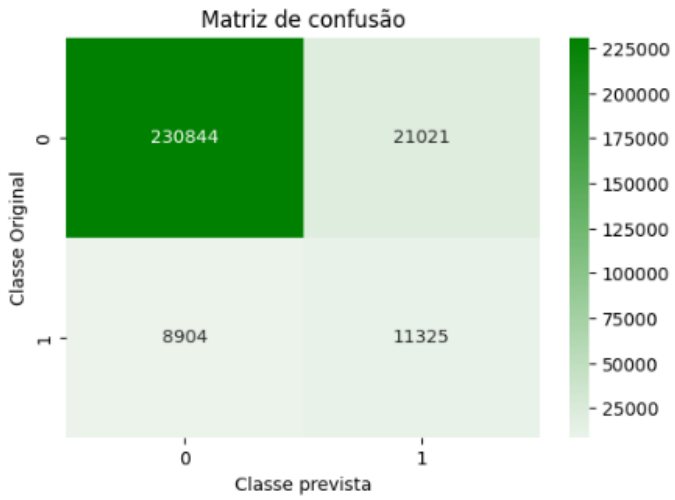


Figura 23 - Confusion Matrix

A Figura 24 apresenta os resultados obtidos da *precision* do teste do modelo em forma de matriz. Através da análise dos resultados obtidos na *precision matrix* é possível concluir que:

- TPR: O valor de 0.963 sugere que o modelo obteve uma taxa de verdadeiros positivos de 96.3%, o que significa que ele identificou corretamente 96.3% das instâncias da classe 0 (positiva).
- FPR: O valor de 0.650 indica que o modelo teve uma taxa de falsos positivos de 65.0%, o que significa que ele classificou incorretamente 65.0% das instâncias da classe 0 (negativa) como pertencentes à classe 1.
- FNR: O valor de 0.037 sugere que o modelo teve uma taxa de falsos negativos de 3.7%, o que significa que ele não conseguiu identificar corretamente 3.7% das instâncias da classe 1.

- TNR: O valor de 0.350 indica que o modelo teve uma taxa de verdadeiros negativos de 35.0%, o que significa que ele classificou corretamente 35.0% das instâncias da classe 1.

Em resumo, com base nas taxas de verdadeiros positivos, falsos positivos, falsos negativos e verdadeiros negativos obtidos, é possível avaliar o desempenho do modelo em relação às classes 0 e 1. O modelo parece ter uma boa capacidade de identificar verdadeiros positivos (TPR alto) e verdadeiros negativos (TNR alto), mas também comete alguns erros de falsos positivos (FPR) e falsos negativos (FNR).

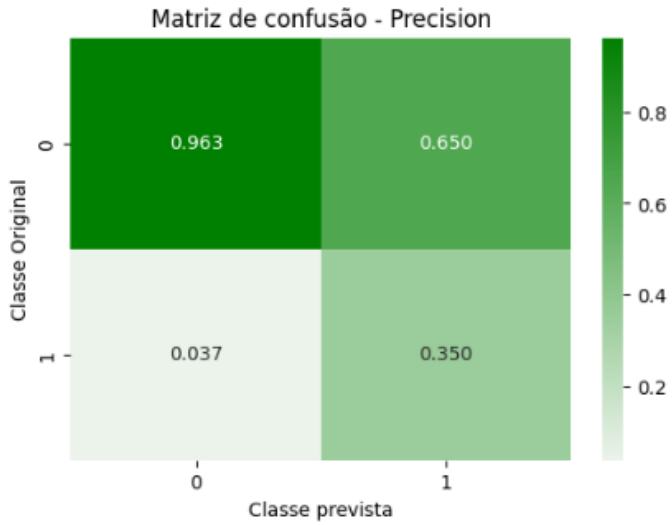


Figura 24 - Precision Matrix

Na Figura 25 está representada uma *recall matriz* cujo objetivo é mostrar os resultados obtidos na métrica do *recall* para os TP, FP, FN e FP para o modelo de RF selecionado. Desta forma, é possível concluir que:

- TPR: O valor de 0.917 sugere que o modelo obteve uma taxa de verdadeiros positivos de 91.7%, o que significa que ele identificou corretamente 91.7% das instâncias da classe 0 (positiva).
- FPR: O valor de 0.083 indica que o modelo teve uma taxa de falsos positivos de 8.3%, o que significa que ele classificou incorretamente 8.3% das instâncias da classe 0 (positiva) como pertencentes à classe 1.

- FNR: O valor de 0.440 sugere que o modelo teve uma taxa de falsos negativos de 44.0%, o que significa que ele não conseguiu identificar corretamente 44.0% das instâncias da classe 1.
- TNR: O valor de 0.560 indica que o modelo teve uma taxa de verdadeiros negativos de 56.0%, o que significa que ele classificou corretamente 56.0% das instâncias da classe 1.

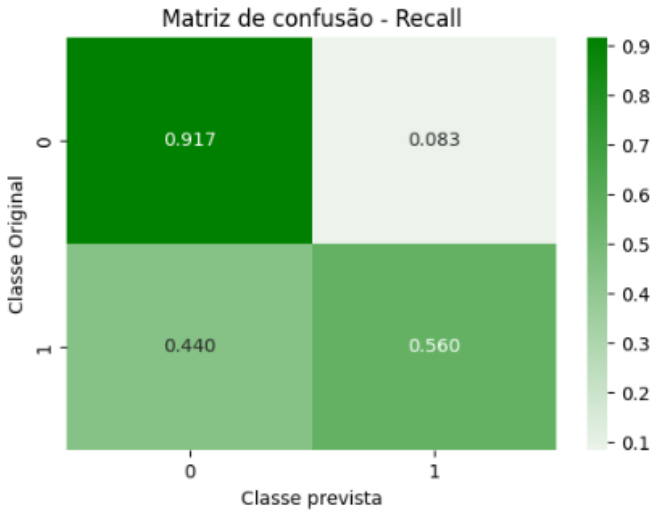


Figura 25 - Recall Matrix

A Figura 26 apresenta o gráfico ROC que é uma ferramenta importante para avaliar o desempenho de modelos de classificação. O eixo do X representa o FPR, enquanto o eixo do Y representa o TPR. O valor do ROC AUC é frequentemente utilizado como uma métrica para resumir o desempenho do modelo e, neste caso, o valor obtido foi de 0.81. A partir do gráfico ROC e do valor AUC é possível concluir que:

- Um AUC de 0.81 é considerado muito bom. Quanto mais próximo o valor de AUC estiver de 1, melhor é o desempenho do modelo em discriminar entre as classes. Um AUC de 0.81 indica que o modelo é capaz de distinguir eficazmente entre as duas classes, com uma alta taxa de verdadeiros positivos (TPR) em comparação com a taxa de falsos positivos (FPR).
- O facto de a curva ROC estar localizada significativamente acima da linha diagonal (que representaria um desempenho aleatório) sugere que o modelo tem uma baixa taxa de falsos positivos em relação à taxa de verdadeiros positivos.

Isto é uma boa indicação de que o modelo não está a classificar incorretamente muitas amostras da classe negativa como pertencentes à classe positiva.

- A curva ROC próxima ao canto superior esquerdo do gráfico indica que o modelo tem uma alta sensibilidade, ou seja, é capaz de identificar eficazmente as amostras da classe positiva.

Em resumo, um valor de AUC de 0.81 no gráfico ROC é um indicador positivo de que o modelo de classificação está a funcionar bem, com uma boa capacidade de distinção entre as classes.

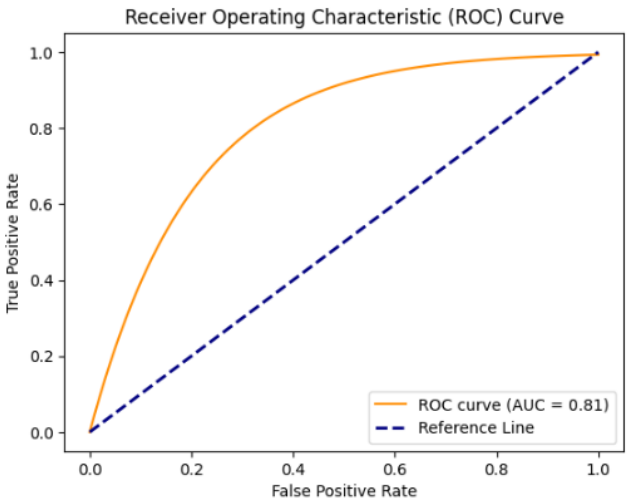


Figura 26 - ROC

De forma a avaliar o desempenho do modelo RLO, foi utilizada a técnica de *hyperparameter tuning*. Este processo consiste em encontrar a combinação ideal dos *hyperparameter* que maximiza o desempenho do modelo. O objetivo é encontrar a configuração que resulta no melhor desempenho do modelo num conjunto de dados. Assim, ao utilizar esta técnica para o modelo de RLO, verifica-se que os melhores parâmetros são os seguintes (Figura 27):

Parameter	Value
C	78.47599703514607
class_weight	None
createdate	2024-03-19
dual	False
fit_intercept	True
intercept_scaling	1
l1_ratio	None
max_iter	150
multi_class	auto
n_jobs	None
penalty	l2
random_state	42
solver	lbfgs
tol	0.0001
verbose	0
warm_start	False

Figura 27 - Resultados obtidos utilizando a técnica de hyperparameter tuning

Os resultados obtidos com a utilização destes parâmetros são os seguintes:

- *Precision:* 0.35
- *Recall:* 0.56
- *F1-score:* 0.44
- *ROC AUC:* 0.81
- *Accuracy:* 0.89

Os resultados da utilização do processo de *hyperparameter tunning* são muito idênticos aos resultados obtidos com o primeiro modelo de RLO treinado.

3.10. *Deployment*

Considerando que este projeto foi desenvolvido dentro do contexto real de uma organização portuguesa cujo mercado de atuação é *online*, o presente documento tem como objetivo apresentar e detalhar a implementação desse projeto específico nessa organização. Ao longo deste capítulo, serão explicados os diversos passos envolvidos na implementação do projeto na respetiva organização.

Para a realização desta implementação, foram concebidos e desenvolvidos dois serviços (*API's*) distintos alojados num *cluster* K8S, de forma a garantir a escalabilidade dos serviços:

- **Serviço com método *POST*** – Este primeiro serviço foi cuidadosamente desenvolvido com o objetivo específico de receber dados através de um método *POST*. Estes dados incluem as categorias dos produtos que foram adicionados ao carrinho de compras pelos utilizadores, o número total de vezes que cada produto foi adicionado ao carrinho, bem como as visualizações que cada categoria de produtos recebeu em cada sessão de utilizador. Após a recolha destes dados detalhados, o serviço realiza cálculos para determinar a pontuação correspondente a cada categoria de produto. Estas pontuações são essenciais e são, subsequentemente, encaminhadas para o modelo de ML para análises mais avançadas e tomadas de decisão informadas. Este serviço foi desenvolvido em Flask, um *microframework* para desenvolvimento *web* em Python, projetado para ser simples e flexível, ou seja, fornece apenas o essencial para a construção de aplicações *web*. Esta ferramenta é conhecida pela sua simplicidade, modularidade e pela facilidade de utilização, sendo ideal tanto para desenvolvedores iniciantes como para projetos mais complexos. Este serviço foi desenvolvido utilizando o Flask porque, dentro da organização, já está estabelecido que qualquer projeto que envolva a leitura de dados provenientes de modelos de ML deve ser desenvolvido com essa ferramenta específica. Esta decisão foi tomada para garantir a consistência, a padronização e a eficiência no desenvolvimento de tais projetos, aproveitando a flexibilidade que o Flask oferece para lidar com esse tipo de aplicações. Além disso, o uso contínuo do Flask dentro da organização promove

uma maior familiaridade entre os membros da equipa, resultando num desenvolvimento mais ágil e na manutenção simplificada dos serviços.

- **Serviço com método *GET*** – Este serviço foi projetado para receber a identificação da sessão do cliente, juntamente com a lista completa dos produtos que o cliente visualizou e adicionou ao carrinho durante essa sessão. Após a recolha destes dados, o serviço transmite todos os dados para o serviço anterior. Esta transmissão é crucial para garantir que o serviço anterior tenha todas as informações necessárias para realizar os cálculos precisos das pontuações das categorias dos produtos, permitindo assim uma análise completa e integrada do comportamento do cliente durante a sessão. Este serviço foi desenvolvido em Scalatra que consiste num *microframework* para desenvolvimento de aplicações *web* em Scala. Esta ferramenta oferece uma forma simples e eficaz de criar serviços *web* com Scala, sendo ideal tanto para pequenos projetos quanto para aplicações mais complexas, proporcionando uma base leve e extensível. Este serviço foi implementado utilizando o Scalatra devido à política interna da organização, que determina que todos os projetos envolvendo a leitura de dados de sessões de clientes (em tempo real) devem ser desenvolvidos com esta ferramenta específica. Esta decisão foi tomada para assegurar a consistência e a padronização no desenvolvimento desses projetos, beneficiando-se das capacidades robustas que o Scalatra proporciona. Além disso, a utilização contínua do Scalatra dentro da organização promove uma maior especialização e familiaridade entre os membros da equipa, o que facilita o desenvolvimento rápido e a manutenção dos serviços. Esta escolha estratégica permite otimizar os recursos e garantir que os sistemas sejam escaláveis e de fácil gestão.

Além dos dois serviços mencionados, foi desenvolvido um *dashboard* no Databricks com o objetivo de facilitar a análise dos dados e a avaliação do modelo de ML. Este *dashboard* foi projetado para ser uma ferramenta abrangente e intuitiva, proporcionando uma visão detalhada e precisa do desempenho do modelo e do comportamento dos clientes.

Uma das funcionalidades principais do *dashboard* é a capacidade de receber como entrada diversos parâmetros, incluindo datas específicas, tipos de clientes (como ativos, *logados*, etc.), e filtros para visualizar diferentes resultados preditivos do modelo. Por exemplo, é

possível filtrar os casos em que o modelo previu corretamente um valor de 1 ou 0, bem como os casos de falso positivo (modelo previu 1 e deveria ser 0) e falso negativo (modelo previu 0 e deveria ser 1).

Principais Dados e Funcionalidades do Dashboard:

1. Métricas de Avaliação do Modelo:
    - O *dashboard* apresenta uma série de métricas essenciais para a avaliação do desempenho do modelo, como *recall*, *accuracy* e precisão (Fig.28). Estas métricas oferecem uma visão clara da *performance* do modelo em diferentes cenários, ajudando a identificar áreas de melhoria e a validar a eficácia do modelo.

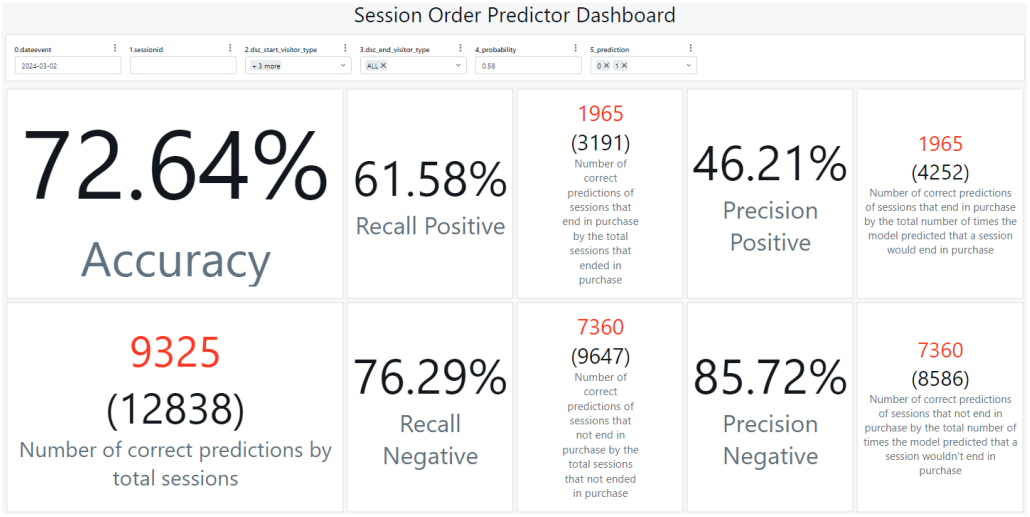
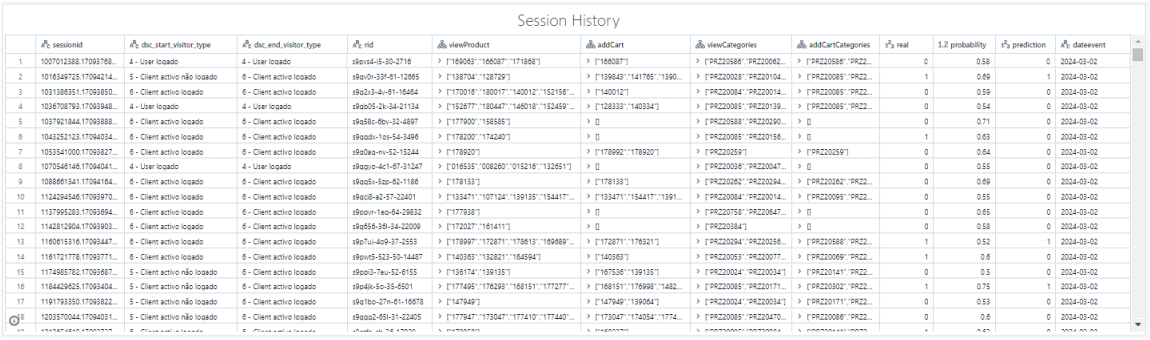


Figura 28 - Métricas de avaliação do modelo

2. Tabela Detalhada de Sessões:
    - Incluí uma tabela detalhada que lista cada sessão de cliente, os produtos visualizados e adicionados ao carrinho, o resultado real (o que efetivamente aconteceu) e a previsão do modelo para cada caso (Fig.29). Esta tabela permite um acompanhamento minucioso das interações dos clientes com os produtos e a precisão das previsões do modelo em cada situação.



	sessionid	dsc_start_visitor_type	dsc_end_visitor_type	rid	viewProduct	addCart	viewCategories	addCartCategories	real	1.2 probability	prediction	dateevent
1	1007012388.17093768...	4 - User logado	4 - User logado	s9ovs4-i5-30-2716	["169063","166087","171868"]	["166087"]	["PRZ20586","PRZ20062...	["PRZ20586","PRZ2...	0	0.58	0	2024-03-02
2	1016349725.17094214...	5 - Client activo não logado	6 - Client activo logado	s9ov0r-33f-61-12665	["138704","128729"]	["139843","141765","1390...	["PRZ20028","PRZ20104...	["PRZ20085","PRZ2...	1	0.69	1	2024-03-02
3	1031386351.17093850...	6 - Client activo logado	6 - Client activo logado	s9o2r3-4r-61-16464	["170016","160017","140012","152156"...	["140012"]	["PRZ20084","PRZ20014...	["PRZ20085","PRZ2...	0	0.59	0	2024-03-02
4	1036708793.17093948...	4 - User logado	4 - User logado	s9qb05-2k-34-21134	["152677","160447","146018","152459"...	["128333","140334"]	["PRZ20085","PRZ20139...	["PRZ20085","PRZ2...	0	0.54	0	2024-03-02
5	1037921844.17093888...	6 - Client activo logado	6 - Client activo logado	s9o58c-6bv-32-4897	["177900","158585"]	[]	["PRZ20588","PRZ20290...	[]	0	0.71	0	2024-03-02
6	1043252123.17094034...	6 - Client activo logado	6 - Client activo logado	s9qadx-1ox-54-3496	["178200","174240"]	[]	["PRZ20085","PRZ20156...	[]	1	0.63	0	2024-03-02
7	1053541000.17093827...	6 - Client activo logado	6 - Client activo logado	s9o0aq-nv-52-15244	["178920"]	["178992","178920"]	["PRZ20259"]	["PRZ20259"]	0	0.64	0	2024-03-02
8	1070546146.17094041...	4 - User logado	4 - User logado	s9qqvo-4c1-67-31247	["016535","008260","015216","132651"]	[]	["PRZ20036","PRZ20047...	[]	0	0.55	0	2024-03-02
9	1088661341.17094164...	6 - Client activo logado	6 - Client activo logado	s9qa5x-5zp-62-1186	["178133"]	["178133"]	["PRZ20262","PRZ20294...	["PRZ20262","PRZ2...	0	0.69	0	2024-03-02
10	1124294546.17093970...	6 - Client activo logado	6 - Client activo logado	s9qc8-a2-57-22401	["133471","107124","139135","154417"...	["133471","154417","1391...	["PRZ20084","PRZ20014...	["PRZ20095","PRZ2...	0	0.55	0	2024-03-02
11	1137995283.17093694...	6 - Client activo logado	6 - Client activo logado	s9qovr-1eq-64-29832	["177938"]	[]	["PRZ20758","PRZ20647...	[]	0	0.65	0	2024-03-02
12	1142812904.17093903...	6 - Client activo logado	6 - Client activo logado	s9q656-36l-34-22009	["172027","161411"]	[]	["PRZ20384"]	[]	0	0.58	0	2024-03-02
13	1160615316.17093447...	6 - Client activo logado	6 - Client activo logado	s9q7ul-4o9-37-2553	["178997","172871","178613","169689"...	["172871","176321"]	["PRZ20294","PRZ20256...	["PRZ20588","PRZ2...	1	0.52	1	2024-03-02
14	1161721778.17093771...	6 - Client activo logado	6 - Client activo logado	s9qwt5-523-50-14487	["140363","132821","164594"]	["140363"]	["PRZ20053","PRZ20077...	["PRZ20069","PRZ2...	1	0.6	0	2024-03-02
15	1174985782.17093687...	5 - Client activo não logado	6 - Client activo logado	s9qoi3-7eu-52-6155	["136174","139135"]	["167536","139135"]	["PRZ20024","PRZ20034"]	["PRZ20141","PRZ2...	0	0.5	0	2024-03-02
16	1184429625.17093404...	5 - Client activo não logado	6 - Client activo logado	s9qlik-5o-35-6501	["177495","176293","168151","177277"...	["168151","176998","1482...	["PRZ20085","PRZ20171...	["PRZ20302","PRZ2...	1	0.75	1	2024-03-02
17	1191793350.17093822...	5 - Client activo não logado	6 - Client activo logado	s9q1bo-27n-61-16678	["147949"]	["147949","139064"]	["PRZ20024","PRZ20034"]	["PRZ20171","PRZ2...	0	0.53	0	2024-03-02
18	1203570044.17094031...	5 - Client activo não logado	6 - Client activo logado	s9qaa2-65l-31-22405	["177947","173047","177410","177440"...	["173047","174054","1774...	["PRZ20085","PRZ20470...	["PRZ20086","PRZ2...	0	0.6	0	2024-03-02

Figura 29 - Tabela detalhada de sessões

3. Gráfico de Linhas para Evolução da Sessão:
    - Um gráfico de linhas ilustra a evolução de cada sessão, mostrando como a probabilidade de compra, indicada pelo modelo, varia ao longo do tempo. Este gráfico é crucial para entender as mudanças no comportamento do cliente durante a sessão e como essas mudanças influenciam as previsões do modelo.

4. Gráficos de Barras para Acontecimentos Reais e Previsões do Modelo:
    - O *dashboard* inclui gráficos de barras que indicam o número total de acontecimentos reais de 1 e 0, fornecendo uma visão geral do comportamento observado nos dados. Além disso, há gráficos de barras que mostram o número total de vezes que o modelo previu 1 e 0, permitindo uma comparação direta entre as previsões do modelo e os resultados reais (Fig.30). Estes gráficos são úteis para avaliar a distribuição e a frequência das previsões, bem como para identificar possíveis padrões ou discrepâncias.

Figura 30 - Gráficos de barras para acontecimentos reais e previsões do modelo

Benefícios do *Dashboard*:

Este *dashboard* não facilita apenas a análise dos dados e do modelo de ML, mas também proporciona *insights* valiosos para a tomada de decisões estratégicas. A capacidade de filtrar e visualizar diferentes aspetos do desempenho do modelo e do comportamento dos clientes permite uma análise mais granular e informada, auxiliando os *stakeholders* na identificação de oportunidades de melhoria e na otimização das estratégias de negócio.

Em suma, este *dashboard* é uma ferramenta poderosa e versátil, que centraliza e simplifica a análise de dados complexos, fornecendo uma base sólida para a avaliação contínua e a melhoria do modelo de ML utilizado.

4. Conclusões

Neste capítulo são apresentadas as conclusões do projeto realizado. Está dividido em duas secções, sendo que no primeiro ponto é apresentada uma síntese/discussão dos resultados obtidos. Na segunda e última secção é apresentado o trabalho futuro a ser realizado.

4.1. Discussão

Ao concluir um projeto, é apropriado avaliar os resultados à luz dos objetivos iniciais delineados. É fundamental adotar uma abordagem crítica em relação às restrições e limitações que surgiram ao longo do projeto.

Em contraste com os objetivos inicialmente traçados para o projeto, constata-se que estes foram atingidos de forma satisfatória. Isto evidencia-se no desempenho positivo do modelo final escolhido, que demonstra índices robustos de previsão. Esses indicadores sugerem com confiança a capacidade do modelo em antecipar com precisão situações em que as sessões tendem a ser concluídas com sucesso ou não. O resultado reflete a eficácia do modelo preditivo, fornecendo uma base sólida para a tomada de decisões informadas no contexto das interações do utilizador. Este alinhamento bem-sucedido com os objetivos iniciais ressalta não apenas a adequação do método adotado, mas também a relevância das previsões geradas, consolidando assim, a contribuição significativa do projeto para a compreensão e aperfeiçoamento das dinâmicas relacionadas ao encerramento de sessões.

Os resultados alcançados revelam-se altamente promissores, conferindo ao projeto uma aplicabilidade abrangente não apenas dentro da organização que partilhou os dados, mas também em qualquer empresa que faça uso de plataformas de comércio virtual. A solidez dos resultados obtidos sugere que a implementação bem-sucedida deste projeto não está restrita a um contexto específico, mas pode ser extrapolada para diversas organizações que operam no espaço do comércio *online*. A generalização destes resultados destaca a robustez e a adaptabilidade do modelo desenvolvido, indicando o seu potencial valor em diversos cenários. Essa versatilidade amplia consideravelmente o alcance e o impacto do projeto, posicionando-o como uma ferramenta valiosa para aprimorar as estratégias de retenção e conversão em ambientes de compras virtuais.

Uma das grandes limitações que se fizeram sentir na realização deste projeto, foi o facto de os conjuntos de dados serem extremamente grandes, principalmente o primeiro *dataset* desenvolvido, o que levava a um custo temporal altíssimo para o treino dos modelos, sendo que esta limitação, além dos fracos resultados obtidos no treino dos modelos com a utilização deste *dataset*, foi uma das causas que levou à diminuição do conjunto de dados, caso contrário seria extremamente difícil realizar o treino dos modelos de ML.

4.2. Trabalho Futuro

Ao concluir este projeto, o trabalho futuro pensado consiste em conceber e desenvolver uma aplicação com a finalidade de fornecer informações em tempo real sobre a probabilidade de êxito de uma determinada sessão. Especificamente, esta aplicação seria capaz de oferecer *insights* imediatos sobre a possibilidade de uma transação ser concluída com sucesso ou não, antecipando se a compra será efetivada. Em essência, esta proposta sugere a criação de uma ferramenta que não apenas monitoriza as interações do utilizador, mas também emprega algoritmos preditivos avançados para avaliar os indicadores relevantes durante uma sessão, permitindo uma previsão informada sobre os resultados. Esta abordagem procura aprimorar a eficiência e a experiência do utilizador, promovendo tomadas de decisões mais rápidas e informadas no contexto de transações *online*.

Além disso, em adição à funcionalidade de antecipação de sessões que possam ser encerradas sem sucesso, a proposta contempla a implementação de um mecanismo de estímulo personalizado aos clientes. Em cenários nos quais o sistema prevê que uma sessão está propensa a ser abandonada antes da conclusão bem-sucedida da compra, estratégias de incentivo seriam automaticamente acionadas. Essas estratégias podem variar desde ofertas exclusivas, descontos personalizados até sugestões de produtos complementares, criando uma abordagem proativa para reter o interesse e motivar os clientes a finalizarem as suas transações. Esta abordagem visa não apenas aprimorar a taxa de conversão, mas também fortalecer a lealdade do cliente ao oferecer uma experiência de compra mais envolvente e personalizada. O objetivo é estabelecer uma dinâmica onde o cliente se sinta valorizado e incentivado a seguir com a compra, transformando potenciais abandonos em transações bem-sucedidas.

5. Referências bibliográficas

Afonso Vieira, V. (2000). Comportamento do Consumidor. https://doi.org/10.1590/S1415-65552002000300015

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the grandes volumes de dados era. Data Science Journal, 14. https://doi.org/10.5334/dsj-2015-002

CAVALCANTI, Leonardo José Elias; DONEUX, Nicolas Franco. Análise de fatores determinantes na decisão de compra online: reflexões sobre o impacto da pandemia no comportamento do consumidor brasileiro. 2021. Trabalho de Conclusão de Curso (Graduação em Administração) – Universidade Federal de São Carlos, Sorocaba, 2021. Disponível em: https://repositorio.ufscar.br/handle/ufscar/14042.

Columbus, L. (2014). Making analytics accountable: 56 % Of executives expect analytics to contribute To 10 % Or more growth in 2014. Forbes. Available at: http://www.forbes.com/sites/louiscolumbus/2014/12/10/making-analytics-accountable-56-of-executives-expect-analytics-to-contribute-to-10-or-more-growth-in-2014/#761c65a95b56 (Accessed on the 2nd of February, 2023).

Davenport, T. H. (2005). Competing on Analytics. www.hbr.orgorcall800-988-0886.www.hbrreprints.org

Mikalef, Patrick & Giannakos, Michail & Pappas, Ilias & Krogstie, John. (2018). The Human Side of Big Data: Understanding the skills of the data scientist in education and industry. https://doi.org/10.1109/EDUCON.2018.8363273

Davenport, T. H., Barth, P., & Bean, R. (2012). How "Big Data" is Different. MIT Sloan Management Review, 54 (1), 22–24.

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Grandes volumes de dados Encouraging Female Students to Program View project Unstructured Data: an

overview of the data of Grandes volumes de dados. Article in International Journal of Emerging Trends & Technology in Computer Science, 38(1). https://doi.org/10.14445/22312803/IJCTT-V38P109

Ferreira Ribeiro, P. M. (2022). Universidade do Minho School of Engineering Machine Learning Applied to Companies Management. Disponível em: https://hdl.handle.net/1822/84499

Gomes, E. G. da S., Domingues, D. A. dos S. D., & Biazon, V. V. . (2020). Comportamento do consumidor: fatores que influenciam o poder de compra. Scientific Electronic Archives, 14(4). https://doi.org/10.36560/14420211252

Grandes volumes de dados, Explained: The 5V s of Data. (2022). https://medium.com/@get_excelsior/big-data-explained-the-5v-s-of-data-ae80cbe8ded1 (Accessed 2nd of March, 2023)

International Conference on Collaboration Technologies and Systems San Diego, Calif.) (2013 :, Smari, W. W., Fox, G. C., Agostini, A., & Institute of Electrical and Electronics Engineers. (2013). Proceedings of the 2013 International Conference on Collaboration Technologies and Systems : May 20-24, 2013, the Sheraton San Diego Hotel & Marina, San Diego, California, USA. IEEE.

Kauffman, R. J., Srivastava, J., & Vayghan, J. (2012). Business and data analytics: New innovations for the management of e-commerce. In Electronic Commerce Research and Applications (Vol. 11, Issue 2, pp. 85–88). https://doi.org/10.1016/j.elerap.2012.01.001

Kopp, M., (2013). Seizing the grandes volumes de dados opportunity, Ecommerce Times. Available at: http://www.ecommercetimes.com/story/78390.html (Accessed 2nd of March, 2023).

Koutsabasis, P., Stavrakis, M., Viorres, N., Darzentas, J. S., Spyrou, T., & Darzentas, J. (2008). A descriptive reference framework for the personalisation of e-business applications. Electronic Commerce Research, 8(3), 173–192. https://doi.org/10.1007/s10660-008-9021-1

Kung, D. S., Gordon, L. C., Lin, F., Shayo, C., & Dyck, H. (2013). 2013. IT-based System with Dynamic Pricing Algorithm. Business Journal for Entrepreneurs: Business Analytics.

Leloup, B., & Deveaux, L. Dynamic Pricing on the Internet: Theory and Simulations. Electronic Commerce Research 1, 265–276 (2001). https://doi.org/10.1023/A:1011546021787

Liebowitz, J. (Ed.). (2013). Big Data and Business Analytics (1st ed.). Auerbach Publications. https://doi.org/10.1201/b14700

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Grandes volumes de dados: The next frontier for innovation, competition, and productivity. www.mckinsey.com/mgi.

Mehra, G., (2013). 6 uses of grandes volumes de dados for online retailers, Practical Ecommerce. Available at: http://www.practicalecommerce.com/articles/3960-6-Uses-of-Big-Data-for-Online-Retailers (Accessed 2nd of March, 2023).

Miguel Ferreira Ribeiro, P. (2022). Universidade do Minho School of Engineering Machine Learning Applied to Companies Management. Disponível em: https://hdl.handle.net/1822/84499

Miller, G., (2013). 6 ways To use “grandes volumes de dados” To increase operating margins By 60 %. Available at: http://upstreamcommerce.com/blog/2012/04/11/6-ways-big-data-increase-operating-margins-60-part-2 (Accessed 2nd of March, 2023).

Monteiro, R., de Castro Machado Rabello, G., Vidal de Arruda Junior, F., & Biscegli Jatene, F. (2022). Inteligência Artifical, Deep Learning, Machine Learning, Redes Neurais na Medicina e Biomarcadores Vocais: Conceitos, Onde Estamos e para Onde Vamos. Revista Da Sociedade de Cardiologia Do Estado de São Paulo, 32(1), 11–17. https://doi.org/10.29381/0103-8559/2022320111-7

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Grandes volumes de dados technologies: A survey. In Journal of King Saud University - Computer and Information Sciences (Vol. 30, Issue 4, pp. 431–448). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2017.06.001

Parashar, M., Jaypee Institute of Information Technology University, University of Florida. College of Engineering, Institute of Electrical and Electronics Engineers. Delhi Section, & Institute of Electrical and Electronics Engineers. (n.d.). 2013 sixth International Conference on Contemporary Computing (IC3-2013) : 8-10 August 2013, Jaypee Institute of Information Technology, Noida, India.

Ramaswamy, S., (2013). What the Companies Winning at Grandes volumes de dados Do Differently. Bloomberg, June: http://www.bloomberg.com/news/2013–06-25/what-the-companies-winning-at-big-data-do-differently.html (Accessed 16th of March, 2023).

Rusu, O. & Halcu, Ionela & Grigoriu, O. & Neculoiu, Giorgian & Sandulescu, V. & Marinescu, M. & Marinescu, Viorel. (2013). Converting unstructured and semi-structured data into knowledge. Proceedings - RoEduNet IEEE International Conference. 1-4. https://doi.org/10.1109/RoEduNet.2013.6511736.

Safara, F. (2020). A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic. Computational Economics, 59(4), 1525–1538. https://doi.org/https://doi.org/10.1007/s10614-020-10069-3

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of grandes volumes de dados. www.sbs.ox.ac.uk

Sint, R., Schaffert, S., Stroka, S., Ferstl, R., Haringer Str, J., & von Siemens-Platz, W. (2009). Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis. Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis. https://www.researchgate.net/publication/220706578

Solomon M. R., Ribeiro L. B., & Farias Salomão Alencar de. (2008). O comportamento do consumidor comprando possuindo e sendo (7. ed.). Bookman.

Stefan Biesdorf, David Court, & Paul Willmott. (2013). big_data_web.

Strauss, J., & Frost, R. (2016). E-marketing. https://doi.org/10.4324/9781315506531

Structured Data vs. Unstructured Data: what are they and why care? (2019). https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/ (Accessed 2nd of March, 2023)

Wixom, Barbara & Yen, Bruce & Relich, Michael. (2013). Maximizing Value from Business Analytics. MIS Quarterly Executive. 12. 111-123.

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A., & Lyu, J. (2020). Brief introduction of medical database and data mining technology in grandes volumes de dados era. In *Journal of Evidence-Based Medicine* (Vol. 13, Issue 1, pp. 57–69). Blackwell Publishing. https://doi.org/10.1111/jebm.12373

Zhao, D. (2013). Frontiers of Grandes volumes de dados Business Analytics. In *Grandes volumes de dados and Business Analytics* (pp. 43–68). Auerbach Publications. https://doi.org/10.1201/b14700-4

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

M MESTRADO
ENGENHARIA INFORMÁTICA

Análise do processo de *checkout* e carrinhos abandonados

Pedro Silva
Davide Carneiro