Área temática: AI – Administração da Informação

USO DE *MACHINE LEARNING* PARA ENTENDIMENTO DAS RELAÇÕES DE CONSUMO NO MODELO DE NEGÓCIO (*SaaS enabled Marketplace*) DO OLIST

RESUMO
A adoção do comércio eletrônico e dos serviços de nuvem podem potencializar as oportunidades e, consequentemente, aumentar o desempenho das empresas no contexto da economia digital. Porém, há ainda uma dificuldade de inserção ou de alcance para empreendedores de menor porte, no que a *startup* brasileira Olist assume o papel de intermediar o lojista e os grandes *marketplaces* aumentando o alcance e automatizando processos. Desta forma, o objetivo geral desta pesquisa é identificar os fatores que impactam no *review score* dado pelo consumidor após a realização da compra via Olist Store, em algum *marketplace*, por meio da análise de dados e aplicação do modelo de *Machine Learning* para prever a nota baseada em CSAT. A pesquisa tem natureza preditiva, uma vez que consiste na aplicação de algoritmos para compreender a estrutura dos dados existentes e gerar regras de predição, e utilizou algoritmos de *Machine Learning* para analisar dados secundários obtidos de fontes abertas e de acesso público. As técnicas de *Machine Learning, Random Forest* e *Feature Importance* foram empregadas e tem um importante papel na interpretação de padrões em um grande volume de dados, assim como no auxílio a solução de problemas complexos. Como resultado, verificou-se que a variável *expected_diff*, que representa a diferença de dias entre a data que estava prevista para a chegada do produto e a data que realmente chegou, é a de maior importância para todas as categorias, com destaque para beleza/saúde e brinquedos. Por fim, concluiu-se que é de suma importância que empresas que desejam obter vantagem competitiva no ramo do comércio eletrônico invistam em uma boa logística.

Palavras-chave: Comércio eletrônico. Técnicas de *Machine Learning*. Categorias de produtos. Previsão do *review score*.

ABSTRACT
The adoption of e-commerce and cloud services can enhance opportunities and, consequently, increase the performance of companies in the context of the digital economy. However, there is still a difficulty of insertion or reach for smaller entrepreneurs, in which the Brazilian startup Olist assumes the role of intermediating the shopkeeper and the large marketplaces, increasing reach and automating processes. In this way, the general objective of this research is to identify the factors that impact the review score given by the consumer after purchasing the Olist Store, in some marketplace, through data analysis and application of the Machine Learning model to predict the score based on CSAT. The research is predictive since it applies algorithms to understand the structure of existing data and generate prediction rules. It used Machine Learning algorithms to analyze secondary data obtained from open and publicly accessible sources. Machine Learning, Random Forest, and Feature Importance techniques were used and had an essential role in interpreting patterns in a large volume of data and helping solve complex problems. As a result, it was found that the expected_diff variable, which represents the difference in days between the expected date of arrival of the product and the date that it arrived, is the most important for all categories, especially beauty/ health and toys. Finally, it was concluded that companies that want to gain a competitive advantage in the field of e-commerce invest in good logistics.

Keywords: e-commerce. Machine Learning Techniques. Product categories. Review score prediction.

INTRODUÇÃO

O empreendedorismo tem crescido em termos de importância em nações desenvolvidas e em desenvolvimento, tendo em vista seu impacto no desenvolvimento nacional e econômico (KUMAR; SYED; PANDEY, 2021). Nesse contexto, de acordo com Pobee (2021), o comércio eletrônico ganha destaque, pois se constitui em uma ferramenta importante para os empreendedores por permitir, entre outros, 24 horas de atendimento, 7 dias por semana, promovendo a troca informações em tempo real. Shehata e Montash (2020) destacam que pequenas empresas, incluindo empreendedores individuais, podem obter vantagem competitiva e estratégica ao adotar o comércio eletrônico. Isso ficou ainda mais evidente quando da eclosão da pandemia da Covid-19 (SCUTARIU et al., 2022).

De acordo com o relatório 2021-2022 disponibilizado pelo Global Entrepreneurship Monitor, os empreendedores brasileiros que estão iniciando ou administrando um novo negócio, ou ainda administrando um negócio já estabelecido, 84% esperam usar mais tecnologias digitais para vender seus produtos ou serviços (GEM, 2022).

Nesse contexto, a computação em nuvem se apresenta com grande potencial para transformar a maneira como as empresas de comércio eletrônico fazem negócios, pois o sucesso desse tipo de iniciativa depende de oferecer ao cliente uma loja *online* segura e confiável (SOHAIB et al., 2019). De acordo com o National Institute of Standards and Technology (NIST), os modelos atuais de serviços em nuvem são Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) e Software-as-a-Service (SaaS). Cada modelo representa um tipo diferente de serviço de nuvem (PINOCHET et al., 2021).

Inserindo a lógica destes modelos a uma empresa de comércio eletrônico é possível observar que o IaaS é usado para comprar infraestrutura de computador como um serviço sob demanda, como servidores, armazenamento, redes e sistemas operacionais. O PaaS fornece aos clientes uma plataforma de aplicativos pré-criada que pode ser usada, conforme necessário, para desenvolvimento em vez de investir em sua própria infraestrutura subjacente. Por fim, o SaaS pode ser usado para iniciar rapidamente um site de comércio eletrônico sem se preocupar com configurações de servidor, investimentos em equipamentos ou atualizações de *software* (SOHAIB et al., 2019).

A adoção do comércio eletrônico e dos serviços de nuvem podem potencializar as oportunidades e, consequentemente, aumentar o desempenho das empresas no contexto da economia digital (SCUTARIU et al., 2022). Além disso, é importante que as empresas levem em consideração a mudança de comportamento do consumidor *online*, sendo necessário estar atento à sua satisfação e à necessidade de implementar estratégias centradas nele (ROSÁRIO; RAIMUNDO, 2021).

Para acompanhar o nível de satisfação, existem diferentes caminhos, mas o método *CSAT (Customer Satisfaction Score)*, traduzido como a nota atribuída a determinado serviço/produto indicando a satisfação do consumidor com determinada experiência (BIRKETT, 2021), se faz necessário para o sucesso de qualquer estratégia. Neste trabalho, o método CSAT é representado pela nota da experiência no Olist Store, maior loja de departamentos dentro de grandes *marketplaces* brasileiros, intermediando o consumo no meio digital.

Assim, o objetivo geral desta pesquisa é identificar os fatores que impactam no *review score* dado pelo consumidor após a realização da compra via Olist Store, em algum *marketplace*, por meio da análise de dados e aplicação do modelo de *Machine Learning* para prever a nota baseada em CSAT.

Com este estudo espera-se promover avanços nas técnicas analíticas de *Machine Learning* de *Random Forest* e *Feature Importantece*, como forma de ranquear as

principais variáveis que impactam na satisfação do cliente via CSAT (nota dada pelo usuário), e apresentar uma estratégia de comparação do desempenho de diferentes categorias de produtos da venda no Olist Store na previsão do *review* dado pelo usuário por meio de técnicas de *Machine Learning*.

A Olist ocupa um importante papel de impulsionar o empreendedor de menor porte, com destaque para o contexto brasileiro no qual empresas como Amazon não foram aceitas, facilitando o acesso às primeiras páginas dos *marketplaces* (NIWATE, 2021). De acordo com a versão brasileira da Technology Review, plataforma de conteúdo do Instituto de Tecnologia de Massachusetts (MIT), a Olist está entre as 20 empresas mais inovadoras do Brasil (EXAME, 2022).

A escolha pela aplicação de técnicas de *Machine Learning* ocorreu tendo em vista que os algoritmos de aprendizado de máquina têm obtido sucesso ao serem capazes de igualar e superar o desempenho humano em diversos campos, como tradução de idiomas, jogos de tabuleiro e carros inteligentes (MCCOY; AURET, 2019) ao aprenderem com os próprios dados, o que coloca os dados como a nova versão do petróleo (PORTELA, 2022).

Em suma, a aplicação dessas técnicas podem trazer inúmeros benefícios às organizações, como aumentar a eficiência e a eficácia dos processos de auditoria (CHAN; VASARHELYI, 2011), ou prever, objetivo deste estudo, o comportamento de consumidores (PAÇO et al., 2018).

REFERENCIAL TEÓRICO

A influência do intermediário no comércio eletrônico

Um dos maiores avanços proporcionados pela internet é o comércio eletrônico, e as transações comerciais na rede podem ser definidos como qualquer transação comercial efetuada em algum ambiente *online*, entre diferentes partes, sendo as mais famosas: entre empresas, popularmente conhecido como B2B (*business-to-business*) e entre empresas e consumidores finais B2C (*business-to-consumer*) (RODRIGUES, 2020).

Com a convergência desses e de outros elementos, as trocas comerciais por meio do comércio eletrônico se tornaram viáveis, fazendo com que, o que inicialmente era uma alternativa dentro do varejo tradicional, extrapolasse as expectativas com um alto nível de faturamento (ALBERTIN, 2020).

Nesse contexto, tendo em vista a complexidade crescente do comércio eletrônico, surgem várias empresas, com destaque para a *startups*, que, segundo a Associação Brasileira de *Startups* (2019), estão alavancando seus negócios com base em 3 modelos: *Saas*, *Marketplace* e *SaaS enabled Marketplace.* O primeiro, SaaS (*Software as a Service*), automação usando *softwares*, baseado em algoritmos, para determinado processo que era feito manualmente. Como exemplo de *startup* com este modelo de negócio no cenário brasileiro tem-se o Nubank que é responsável por uma revolução no setor bancário e a mais valiosa dentre todas as *startups*.

Na sequência tem-se o *Marketplace,* cujo modelo consiste em conectar oferta e demanda, cobrando uma parcela por tal intermediação, como é o caso do iFood. Por fim, tem-se o *SaaS enabled Marketplace* que consiste na automatização de algum processo do usuário, além de conectar de alguma forma, a oferta e demanda, monetizando essa intermediação. Como exemplo temos a Olist que hoje é considerada uma *startup* unicórnio, valendo mais de US$ 1 bilhão (PADRÃO, 2021).

A Olist é considerada a maior loja de departamentos dentro dos maiores *marketplaces* brasileiros e ganhou mercado ajudando lojas físicas a venderem nesses *marketplaces* como Mercado Livre, Amazon e Submarino (Olist [s/d]). O modelo de negócio da Olist consiste em automatizar um processo que seria feito manualmente por meio de

softwares, no caso o cadastro e a gestão do lojista para cada produto em cada *marketplace* distinto que deseja ofertar, visto que o Olist centraliza todos os cadastros e unifica a gestão na plataforma única da empresa, com acesso exclusivo ao lojista (NIWATE, 2021).
Além disso, conecta a oferta com a demanda, pois se posiciona como uma loja única perante os consumidores no menu de busca dos *marketplaces*, mas é composta por todos os lojistas credenciados, monetizando também este intermédio. A Figura 1 ilustra o ecossistema ao qual a Olist Store faz parte.

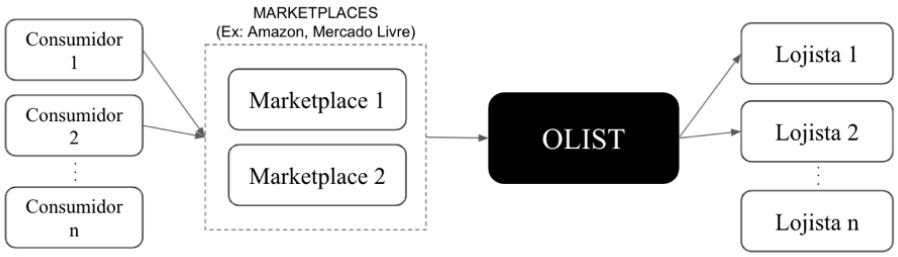

Figura 1: Fluxo de compras via Olist Store
Fonte: elaboração própria

Podemos diferenciar a Olist dos demais *marketplaces,* tendo em vista que pode ser definida como um *shopping* virtual, ou seja, um *site* de *e-commerce* que reúne ofertas de produtos e serviços de diferentes vendedores (EUROMONITOR, 2018). Ou seja, a Olist anuncia os produtos nos demais *sites*, portanto tendo seu negócio intermediado pelo *marketplace*, mas também sendo o intermediário entre a plataforma do *marketplace* e o lojista. Portanto, conforme observado na Figura 1, nas transações via Olist Store, são dois intermediários entre o consumidor e o lojista.
O modelo de negócios da empresa, auxilia, principalmente, pequenos e médios lojistas que enfrentam muitas vezes problemas por incompatibilidade tecnológica ao ingressarem nos aplicativos de comércio, visto que, por meios digitais, a gestão se torna muito mais complexa, pois aumentam os canais de fontes de informação, elevando também a dificuldade em manter uma correta gestão de inventário (MATOS, 2020). A Olist propõe o fim deste problema, por meio de automatização e monitoramento integrado.

Machine Learning e a compreensão de padrões

Nas últimas décadas, a tecnologia vem ganhando cada vez mais espaço na vida das pessoas, consequentemente vem sendo cada vez mais discutida, com *Machine Learning* não é diferente. O *Machine Learning* é um ramo da Inteligência Artificial, mas, em contraste ao modelo tradicional, não possui como seu objetivo principal a necessidade de automatizar uma tarefa que o ser humano teria que fazer repetidamente ou que demande um longo tempo, mas sim de explorar algoritmos que podem aprender e fazer predições com base nos dados de entrada e no passado (HUTTER; KOTTHOFF; VANSCHOREN, 2019).
Nas últimas décadas ocorreu uma evolução na complexidade dos problemas a serem tratados pelos computadores, somado ao aumento exponencial do número de dados gerados e armazenados por diversos setores e à sofisticação das técnicas computacionais para reduzir cada vez mais a necessidade de intervenção e manutenção humana. Ou seja, que as tecnologias fossem capazes de criar hipóteses baseadas nas experiências passadas fornecidas para resolver o problema (FACELI et al., 2011). Os algoritmos de *Machine Learning* podem ser categorizados, conforme o

tipo de aprendizado, em 3 tipos, sendo eles: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço (CASTLE, 2018; COSTI, 2020).
A principal diferença entre o aprendizado de máquina supervisionado e não supervisionado é a questão dos algoritmos de aprendizado de máquina supervisionados terem o treino em conjuntos de dados rotulados (CASTLE, 2018), ou seja, você possui o que pretende prever na base, que orientam o algoritmo a entender quais recursos são relevantes para o problema em questão. Para Silva (2021), o aprendizado supervisionado se refere aos algoritmos de *Machine Learning* que aprendem a "resposta correta" mapeando os *inputs* e *outputs* por meio de amostras inseridas no sistema, ou seja, o algoritmo entra em contato com a base de treino que possui a variável resposta, desta forma pode aprender, encontrar os padrões, criar uma regra interna e prever o resultado de uma nova amostra.
Assim, normalmente utilizado para a predição de eventos, o objetivo do aprendizado supervisionado é aprender uma função que, dada uma amostra de dados e resultados desejados, se aproxima melhor da relação entre entrada e saída observável nos dados. Ao conduzir o aprendizado supervisionado, as principais considerações a serem feitas são em relação a complexidade do modelo e o *trade-off* de viés e variância (SONI, 2018).
Já os não supervisionados, são treinados em dados não rotulados, ou seja, chega em respostas não presentes no banco de dados original, e devem determinar a importância do recurso independentemente, com base nos padrões inerentes à amostra (COSTI, 2020). Ou seja, são utilizados para a descrição de eventos ainda não conhecidos e recebem um conjunto de dados de treinamento não rotulados. Seu objetivo é procurar alguma estrutura em dados de amostra, agrupá-los em grupos de regras semelhantes e descobrir padrões ocultos. Métodos de agrupamento são usados para encontrar uma partição dos dados e classificar novos dados de entrada com uma regra de previsão (JORDAN; MITCHELL, 2015).
Por fim, a terceira categoria diz respeito ao aprendizado por reforço, cujo dados de treinamento indicam uma recompensa ou uma punição ao algoritmo com base nas metas estabelecidas. Na aprendizagem por tentativa e erro, usando recompensas e punições como *feedback*, os algoritmos encontram uma solução adequada maximizando as recompensas totais (JORDAN; MITCHELL, 2015; SILVA, 2021).
O presente estudo está alinhado na categoria do aprendizado de máquina supervisionado, visto que ele visa prever o campo de pontuação dada pelo cliente após a compra.

METODOLOGIA

A pesquisa que se apresenta tem natureza preditiva, uma vez que consiste na aplicação de algoritmos para compreender a estrutura dos dados existentes e gerar regras de predição (SANTOS et al., 2019), e utilizou algoritmos de *Machine Learning* para analisar dados secundários obtidos de fontes abertas e de acesso público.
Técnicas de *Machine Learning* estão cada vez sendo mais exploradas porque, quando os modelos são expostos a novos dados, eles são capazes de se adaptar independentemente, significando que é possível produzir, de forma rápida e automaticamente, modelos capazes de analisar dados maiores e mais complexos, e entregar resultados mais rápidos e precisos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

A escolha da base de dados

A base de dados que utilizada é composta por dados públicos da Olist, disponível no Kaggle, plataforma *online* gratuita fundada em 2010 e considerada uma grande comunidade de cientistas de dados com fóruns de discussão, desafios, competições e *datasets* públicos com grande quantidade de informações para exploração (BANACHEWICZ; MASSARON, 2022).

A base de dados do estudo é composta por dados temporais de pedidos da Olist Store, de 2016 a 2018 (dados mais atuais não foram disponibilizados pelo Kaggle ou empresa por serem considerados dados estratégicos e sensíveis no mercado), e informações de aproximadamente 100 mil transações. Composto por 9 bancos de dados, Figura 2, possui uma grande diversidade de tipos de informação com visibilidade de informações de itens do pedido, preço, frete, pagamento, localização do cliente, atributos do produto de diferentes setores, suas características do anúncio digital e algumas mais. A estrutura do banco de dados permite que o relacionamento entre elas seja facilitado, visto que é composta por identificações únicas entre as 9 bases que compõem o banco de dados.

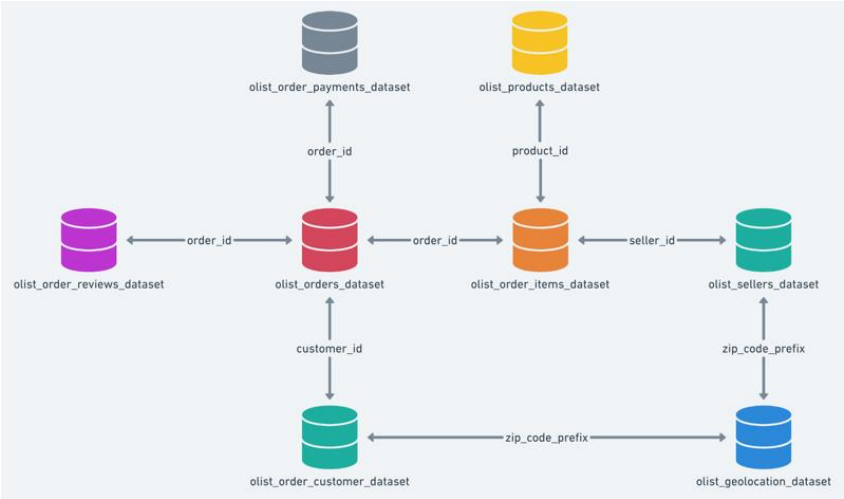

Figura 2: Esquema dos conjuntos de dados do estudo
Fonte: Kaggle

É importante ressaltar que os bancos de dados são compostos por algumas chaves primárias e únicas entre as bases, o que permite a união de todas, mas sem permitir qualquer tipo de identificação dos envolvidos, para respeitar a Lei Geral de Proteção de Dados Pessoais (LGPD).

Técnicas de pesquisa e Análise de dados

O presente estudo utilizou o Google Colaboratory, também conhecido como Google Colab, como principal plataforma na nuvem para o desenvolvimento de toda a análise, desde o momento exploratório dos dados até o algoritmo de *Machine Learning* (GUNAWAN et al., 2020).

Por ser um ambiente totalmente *online* e diretamente no seu navegador, a plataforma nos proporciona conexão direta com a plataforma Kaggle, local onde o banco de dados foi disponibilizado, por meio de uma chave API única por usuário. As principais bibliotecas utilizadas no estudo são: NumPy, Pandas, Math, Matplotlib, Seaborn, Datetime e Sickit-Learn.

As principais funções utilizadas no trabalho foram com base nas bibliotecas explicadas acima. Após conexão e *download* dos bancos de dados no Google Colab, iniciou-se

o processo de tratamento da base. Um dos estágios previstos na literatura é o *Data Cleaning*, que ajuda na manutenção dos dados de entrada para o modelo, visto que o modelo de *Machine Learning* não terá resultados positivos caso a entrada de dados não seja consistente (Tembusai et al., 2021). A primeira parte desta etapa se refere aos dados duplicados. Neste caso optou-se pela aplicação da função de *drop_duplicates()*, responsável por identificar todas as linhas que tivessem os mesmos dados em todas as colunas, evitando assim que houvesse um peso duplicado onde não há necessidade.

A segunda parte dessa etapa diz respeito aos dados faltantes. O mais comum na literatura é entender primeiro se a origem desses dados faltantes é por ser um dado que não existe, ou um dado que não foi gravado. Caso se trate de um dado que não existe, não faz sentido tentar prever. Exemplo: tentar preencher a coluna de *review_score* de um pedido que está no *status* de *aguardando aprovação*. Como o dado não está preenchido por não existir, ele deverá permanecer nulo, enquanto a coluna *data de entrega* para um pedido com o *status* de *entregue* não estar preenchida representa não ter sido gravado apesar do dado já existir, pode ser inferida por meio de algumas das ferramentas utilizadas na comunidade.

Neste trabalho, em busca de um melhor resultado, houve um corte de todas as linhas que, por algum motivo, não tivessem todos os dados de todas as colunas preenchidos, permanecendo a frequência da coluna de menor valor por meio da função *dropna()*. Essa decisão se deu principalmente visto que mesmo nas colunas com tais dados faltantes, ainda há muitos dados de entrada, o que, portanto, não afetaria o modelo.

Um primeiro filtro feito na base de pedidos (*df_order)* é a continuidade apenas de compras que chegaram ao *status* de *delivered* (entregue, em português), para que sejam comparadas apenas compras que possuem já uma nota atribuída, pois tal avaliação apenas é enviada ao cliente após cumprimento de todas as etapas de compra até o recebimento.

O próximo passo no pré-processamento será a substituição de colunas antes com datas, pela diferença entre os eventos, visto que a grande maioria dessas datas são de fenômenos subsequentes. As datas que não se enquadram nisso serão descartadas. As datas que foram consideradas são listadas na Tabela 1 com seus respectivos *dataframes* de origem:

Tabela 1 – colunas de datas consideradas

Coluna	***Dataframe* de Origem**
review_creation_date	df_reviews
review_answer_timestamp	df_reviews
order_purchase_timestamp	df_orders
order_approved_at	df_orders
order_delivered_carrier_date	df_orders
order_delivered_customer_date	df_orders
order_estimated_delivery_date	df_orders

Fonte: elaboração própria

Para um melhor entendimento da seleção das datas, a Figura 3 traz o processo de compra com as respectivas etapas.

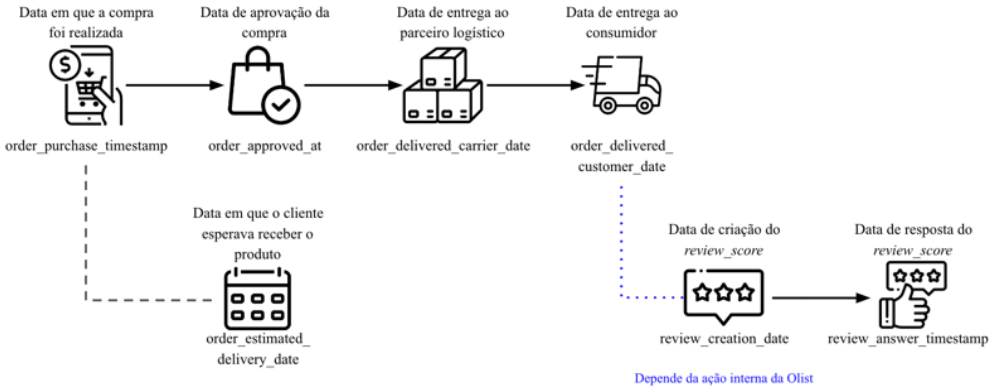

Figura 3: Fluxo das colunas de datas dentro do processo de compras do Olist
Fonte: Elaboração própria

A primeira coluna foi descartada e as próximas substituídas por um número inteiro, como o número de dias que o pedido atrasou ou adiantou frente ao esperado pelo consumidor, tendo uma maior relevância do que ter duas colunas contendo datas distintas. A Figura 4 ilustra o processo de compras após as alterações de colunas.

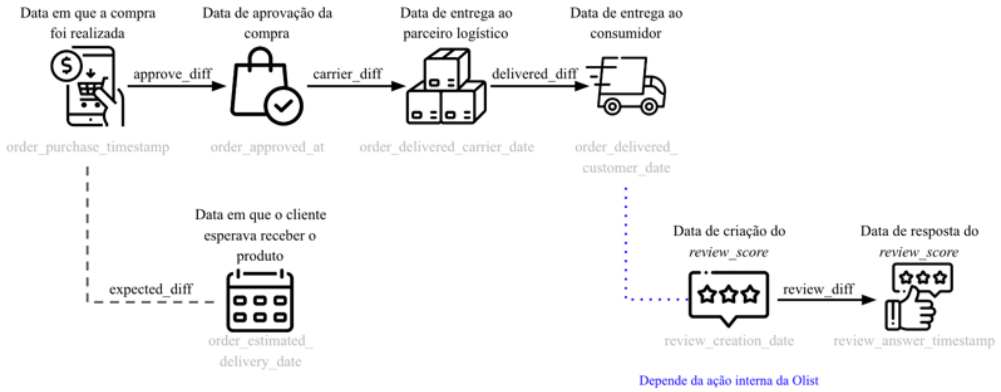

Figura 4: Fluxo das colunas de datas após alterações
Fonte: elaboração própria

Por conta da formatação presente no banco de dados do estudo, foi necessário primeiro converter as respectivas colunas para o formato *date time* e então feita a criação da nova coluna, contendo a operação mencionada. Após todas as análises iniciais com esse formato mais visual, houve uma nova conversão dos formatos resultantes da operação para *time delta*, números inteiros, para assim o modelo ser capaz de interpretá-los.
Outra etapa cumprida foi a criação de uma coluna chamada '*total_product_qtd*' que é responsável por apresentar quantos produtos são por compra, o que auxiliará em etapas seguintes.
A função *merge* também foi utilizada neste trabalho com a finalidade de unir *data frames* pelas suas chaves primárias (únicas). Antes de unir as bases, é preciso identificar quais dados realmente são interessantes para o objetivo do estudo, para assim escolher o melhor formato da aplicação (ESTRELLA, 2020), por meio do requerimento *how* e da identificação por meio da Teoria dos Conjuntos, Figura 5.

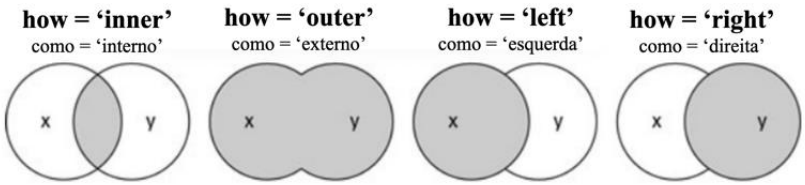

Figura 5: Teoria dos Conjuntos

Na aplicação da Teoria dos Conjuntos a junção de *data frames*, como mostra a Figura 5, inicialmente sempre será determinado pelo df que ocupará a posição da esquerda (exemplificado pela letra x), e qual df ocupará a posição da direita (exemplificado pela letra y). Portanto, o *how* definirá o limite que o padrão deverá respeitar ao fazer o encontro dos dados, e no presente estudo utilizou-se o *left*. A função será utilizada em diferentes momentos da pesquisa, como no agrupamento do *df_orders* com *df_items*, mas tendo sua principal função em unir todos os bancos, que inicialmente eram separados, em uma única base que será utilizada para a análise final.

Uma das possíveis quebras presentes no banco de dados são as categorias, que vieram por padrão da base original, e são classificadas em 71 categorias, desde itens de casa até informática. Para tentar criar um recorte um pouco mais coerente, visto que as categorias podem ter comportamentos muito diferentes entre cada uma, buscou-se entender quais eram as 10 principais categorias, no quesito volume de pedidos, mais significativas, como mostra a Tabela 2.

Tabela 2 – 10 principais categorias

Categoria	Número de pedidos
beleza_saude	954
cama_mesa_banho	922
relogios_presentes	843
utilidades_domesticas	693
esporte_lazer	629
informatica_acessorios	596
moveis_decoracao	513
automotivo	492
telefonia	385
ferramentas_jardim	284

Fonte: elaboração própria

O segundo filtro foi focado em apenas compras que tivessem 1 (um) item por pedido, visto que a variável dependente do estudo e algumas outras variáveis seguem o padrão de ser uma resposta por pedido, e não por item do pedido. À essa base final, deu-se o nome de df_categorias.

Para uma melhor compreensão dos resultados, optou-se pela conversão das 5 classes (nota de 1 a 5) para 2, utilizando a média das notas para definir o limite, sendo a nova variável chamada de *score_class*. Sendo a classe 0 o *score* baixo (notas de 1 a 3) e a classe 1 o score alto (notas 4 e 5).

Por fim, as colunas finais presentes no banco de dados e que, portanto, são utilizadas para a aplicação dos modelos de *Machine Learning* estão descritas na Tabela 3, todas com o preenchimento de 5384 linhas:

Tabela 3 – colunas do banco de dados *df_categorias*

Colunas	Descrição
review_diff	Diferença de dias entre a criação do *review* da compra e o cliente respondê-la.
approve_diff	Diferença de dias entre a realização da compra e a aprovação dela.
carrier_diff	Diferença de dias entre a aprovação da compra e a entrega ao parceiro logístico.
delivered_diff	Diferença de dias entre a entrega ao parceiro logístico e a entrega ao consumidor.
expected_diff	Diferença de dias entre a data que o cliente esperava receber e a data em que recebeu.
price	Preço do pedido.
total_freight_value	Preço do frete.
product_name_lenght	Número de caracteres extraídos do nome do produto.
product_description_lenght	Número de caracteres extraídos da descrição do produto.
product_photos_qty	Número de fotos publicadas do produto.
product_weight_g	Peso do produto medido em gramas.
product_length_cm	Comprimento do produto medido em centímetros.
product_height_cm	Altura do produto medida em centímetros.
product_width_cm	Largura do produto medida em centímetros.
score_class	Conversão do *review score* em *Score* Alto e *Score* Baixo

Fonte: elaboração própria

Modelo de Machine Learning

No presente estudo, optou-se pela utilização do algoritmo de *Random Forest*, que é uma popular e poderosa técnica de Machine Learning (BREIMAN, 2001). Ele é feito pelo conjunto de muitas árvores de decisão individuais, que calculam entre si o resultado predito com partes randômicas da amostragem total, o que auxilia na redução do efeito de *overfitting* e melhora a generalização, e então escolhem pela maioria a classificação final (LOOSVELT et al., 2012). Os modelos de Árvore de Decisão são baseados em probabilidade em que cada *Tree* seria uma Árvore de Decisão individual, enquanto o *Random Forest* é o conjunto de muitas árvores.

Para iniciar a modelagem de *Machine Learning*, separou-se a base em duas sendo o *dataframe y_cat* contendo apenas a variável dependente (*score_class*) e o *dataframe x_cat* com todas as demais variáveis, que servirão para treino e teste do modelo. Em seguida, foi realizada a separação de ambos os *dataframes*, mas ainda sem haver dispersão da correspondência entre as duas tabelas, em treino e teste de forma aleatória e randômica.

Essa etapa tem a relevância de permitir que o modelo treine em uma base, o que permite ao pesquisador fazer ajustes, para então validar realmente sua performance rodando na base de teste. Assim, a base de teste se torna uma base desconhecida para o modelo, simulando a prática na vida real. A separação das compras foi feita em 80% para treino com uma amostragem de 4308, e 20% para teste com uma amostragem de 1078. O algoritmo *Random Forest* tende a ter uma melhor performance com a divisão selecionada (AVUÇLU; ELEN, 2020).

A etapa seguinte baseou-se em reduzir a variância das variáveis que seriam analisadas, por meio do método chamado *Z score* que é ilustrada pela função: $z\ score = \frac{(x-\mu)}{\sigma}$. Na expressão, x representa o valor que será normalizado em questão, como o valor da compra por exemplo, μ representa a média das variáveis e σ representa o desvio padrão (AHMED et al., 2019). A essa etapa dá-se o nome de *Feature Scaling*, e deve ser feita após a separação em treino e teste para que uma não tenha influência na normalização da outra, afinal o intuito é principalmente que a

base de teste nunca tenha sido vista pela máquina. Em seguida, iniciou-se o processo de aplicação do método do *Random Forest*.

Matriz de Confusão

A apresentação dos resultados é feita por meio da Matriz de Confusão, Figura 6, que é comumente utilizada para entender o comportamento de classificação de cada categoria em modelos supervisionados de classificação (HASNAIN et al., 2020). A matriz é composta por linhas e colunas, e apresenta duas possibilidades de resposta, *score* alto ou *score* baixo, a matriz terá dimensão 2x2. Para a análise dos resultados no estudo, foi criada uma Matriz de Confusão referente ao percentual de acerto de determinada categoria.

		Valor real		
		Score Baixo	Score Alto	
Valor previsto	Score Baixo	Verdadeiro Positivo (VP %)	Falso Negativo (FN %)	= 100%
	Score Alto	Falso Positivo (FP %)	Verdadeiro Negativo (VN %)	= 100%

Figura 6: Modelo prático da Matriz de Confusão no estudo

Cada valor pode ser descrito por:

- Verdadeiro Positivo (VP %): mostra quantos registros foram classificados como *score* alto e o *score* realmente era alto, dividido pelo total de vezes que o classificador previu *score* baixo, independentemente do valor real.
- Verdadeiro Negativo (VN %): mostra quantos registros foram classificados como *score* baixo e o *score* realmente era baixo, dividido pelo total de vezes que o classificador previu *score* alto, independentemente do valor real.
- Falso Positivo (FP %): mostra quantos registros foram classificados como *score* alto e o *score* era baixo, dividido pelo total de vezes que o classificador previu *score* alto, independentemente do valor real.
- Falso Negativo (FN %): mostra quantos registros foram classificados como negativos incorretamente, ou seja, a resposta do classificador foi que o *score* era baixo e o *score* era alto.

E por fim, utilizou-se o *Feature Importance,* que é um valor usado para ordenar as variáveis por importância de impacto na variável dependente (KANG; RYU, 2019), *score_class*, em uma escala de 0 a 100%. Esse *ranking* que pode auxiliar na seleção de variáveis para testes posteriores, auxiliando na performance e reduzindo os recursos necessários, se tornando mais rápido. Estas preocupações com performance computacional tendem a ser necessárias proporcionalmente ao tamanho da amostragem.

O processo todo descrito a partir da separação do df_categorias em treino e teste até a obtenção do *Feature Importance* foi repetido 10 vezes e então retirada as médias simples. x é representado pela saída do *Feature Importance*: $média = \frac{\Sigma x}{n}$. Após isso, repetiu-se o processo com foco nas 5 variáveis de maior relevância e bases separadas por categoria, focada nas 10 principais categorias em volume de compras.

RESULTADOS E DISCUSSÃO

A primeira aplicação foi com base no *df_categorias*, data frame contendo toda a base das 10 categorias juntas, com foco em entender por meio do *Feature Importance* quais variáveis tinham uma maior relevância, para então utilizá-las na análise por categoria. Com 4308 amostras no treino, sendo desses 73,9% (3184) da classe 1, ou seja,

definidos como *score* alto, e 1078 no teste, sendo desses coincidentemente 73,9% (797) também da classe 1. Com essas amostragens, obteve-se uma acurácia média das 10 aplicações de 79,1%. Também tirou se uma média simples da matriz de confusão de cada *loop*, tendo o resultado mostrado na Figura 7.

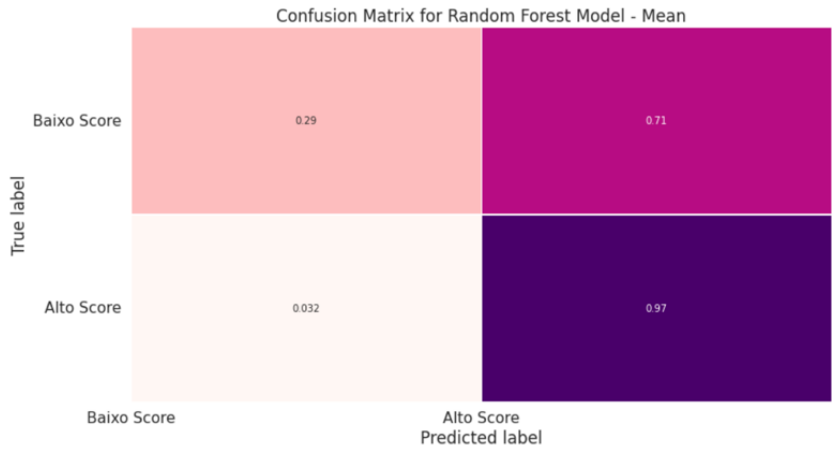

Figura 7: Matriz de Confusão com resultados médios do *Random Forest*

Outros resultados de performance são observados na Tabela 4.

Tabela 4 – Indicador de performance do algoritmo de *Machine Learning*

Classe	Precisão
Baixo Score	0,76
Alto Score	0,79

Por fim, ao analisar o resultado de cada *looping* da aplicação do *Random Forest* para o *df_categorias*, concluiu-se que as 5 variáveis mais relevantes para a decisão da nota atribuída após todo o serviço são *expected_diff, delivered_diff, review_diff, carrier_diff* e *approve_diff*. Os resultados de cada *loop*, sendo 0 o primeiro e 9 o 10º, e a média final são mostrados na Tabela 5.

Tabela 5 – *Feature Importance* por *loop*

Variável	0	1	2	3	4	5	6	7	8	9	Média
expected_diff	0,13	0,13	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14
delivered_diff	0,11	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
review_diff	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09	0,09
carrier_diff	0,09	0,09	0,09	0,08	0,09	0,08	0,08	0,08	0,08	0,08	0,08
approve_diff	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07	0,07

Por concluir que essas são as mais relevantes, então repetiu-se o processo em bases separadas por categoria, para que não haja influência do padrão de uma categoria no resultado da outra, apenas com as variáveis tidas como mais relevantes. Após efetuada a análise por categoria, foi considerada a média dos *Feature Importances* médios de cada variável por categoria, assim como a acurácia média, para então ter uma melhor análise dos comportamentos e predições. O resultado médio da relevância de cada variável e a acurácia para cada categoria é mostrado na Tabela 6, e a acurácia média por categoria.

Tabela 6 – *Feature Importanc*e médio de cada variável e Acurácia por categoria

Categoria	expected_diff	delivered_diff	review_diff	carrier_diff	approve_diff	Acurácia
cama_mesa_banho	22,75	21,77	19,74	18,87	16,87	78,61
beleza_saude	28,22	19,78	19,11	17,94	14,92	83,17
esporte_lazer	26,32	22,44	18,32	17,90	14,99	80,45
informatica_acessorios	22,47	20,23	19,74	19,25	18,29	78,03
moveis_decoracao	23,21	20,26	19,36	18,93	18,21	72,81
utilidades_domesticas	26,85	20,71	18,63	17,51	16,28	84,07
relogios_presentes	24,07	23,39	19,72	17,69	15,15	68,37
telefonia	22,30	21,29	20,42	19,42	16,54	69,43
automotivo	24,88	21,35	19,09	18,05	16,60	79,33
brinquedos	27,59	19,99	19,38	18,26	14,75	83,00

Com resultado unânime entre as categorias tem-se a variável *expected_diff* que é a de maior relevância para a predição do modelo e ilustra que o fator entrega têm um grande peso ao se tratar de compras feitas pela internet, mesmo antes da pandemia da COVID-19. O desafio das redes varejistas é conseguir acompanhar a evolução deste novo tipo de compra, e consumidores cada vez mais empoderados, que querem um prazo cada vez menor e pontualidade de entrega, tendo a urgência com grande relevância (SHETTY et al., 2018).

Para suprir a necessidade de uma entrega rápida e pontual ao cliente, é necessário que cada vez mais as empresas foquem em desenvolver uma boa logística, com processos claros e eficientes. O presente estudo mostrou que as categorias em que essa necessidade se torna mais presente são de beleza/saúde ($\bar{x}_{FI} = 28{,}73\%$) e brinquedos ($\bar{x}_{FI} = 27{,}60\%$), também é importante ressaltar que ambas obtiveram bons resultados de acurácia, reforçando o diagnóstico.

Segundo pesquisa feita pela McKinsey & Company, focada no investimento de alguns países para classificações de bem-estar, é possível observar uma grande importância das áreas de saúde e aparência, se refere ao investimento em produtos de beleza para o Brasil, o que traz um reflexo na necessidade crescente de prazos mais pontuais para os consumidores da categoria (CALLAGHAN, 2021).

Na visão contrária, pode-se concluir que as categorias de informática/acessórios ($\bar{x}_{FI} = 22{,}47\%$) e móveis/decoração ($\bar{x}_{FI} = 23{,}21\%$) são as categorias com a melhor distribuição entre as variáveis, mas com acurácias não tão altas.

Entretanto, segundo pesquisa feita pela Opinion Box (2021), consumidores do mercado de móveis e decoração *online* são mais apegados a marca em si, seguidos por outros fatores como material, durabilidade, preço e, por último, o frete. Diante disso, estas comparações permitem confirmar que os comportamentos podem ser particulares a cada categoria.

CONCLUSÕES

O estudo atingiu o objetivo de identificar os fatores que impactam no *review score* dado pelo consumidor após a experiência via Olist Store, o que ilustra a satisfação com o produto e serviço prestados por meio da técnica de *Machine Learning* do *Random Forest*, e após isso aplicado o *Feature Importance* para ranqueamento da relevância das variáveis perante a predição.

Com a forte relevância das *startups* no cenário econômico brasileiro e mundial, e com a importância da Olist dentro desse cenário, associado a aplicação de técnicas de Inteligência Artificial para a resolução de problemas complexos e trabalhosos para um ser humano, deu-se a presente pesquisa. Por meio da aplicação da técnica do *Random Forest,* modelo de *Machine Learning*, conseguiu-se ensinar a máquina a tentar prever

qual *score_class* seria atribuído a cada compra, e com a aplicação do *Feature Importance* descobriu-se quais variáveis tiveram uma maior influência nessa nota.
Após realizar o processo em uma base geral, houve a quebra e repetição do processo dentro de cada categoria, dentre as mais relevantes em volume de compras, permitindo assim a conclusão de que as categorias de beleza/saúde e brinquedos são as mais afetadas pela diferença de dias entre o esperado para a entrega do produto, e a data realmente entregue, reforçando a necessidade que as empresas de *e-commerce* estejam muito alinhadas com os processos de logística. Além disso, observa-se que a pandemia ampliou a expansão do mercado de compras *online*, em que o consumidor está cada vez mais exigente quanto a prazo de entrega, e o cumprimento desses, sendo este um item que merece atenção dos empreendedores.
Por fim, concluiu-se que as empresas de *e-commerce,* possuem vantagens competitivas ao explorar e investir em tecnologia, para entender melhor seu mercado e padrões de consumo, assim como buscar oportunidades e melhorar a eficiência, sendo positivo em todos os campos, desde que haja uma boa coleta e armazenamento de dados.
Como limitação e sugestão para futuras pesquisas, tem-se a possibilidade de olhar para a classificação do *review_score* e, por exemplo, entender se a distância entre comprador e vendedor pode ter uma influência positiva no *expected_diff*, consequentemente melhorando a satisfação com a integração de comércios regionais.

REFERÊNCIAS

ACCENTURE. *What Is Artificial Intelligence* | Accenture, 2021. Disponível em: https://www.accenture.com/us-en/insights/artificial-intelligence-summary-index. Acesso em: 15/07.

ALBERTIN, A. L. Pesquisa FGVcia de Comércio Eletrônico no Mercado Brasileiro. *FGV EAESP*, 2020. Disponível em: https://eaesp.fgv.br/producao-intelectual/pesquisa-fgvcia-comercio-eletronico-mercado-brasileiro Acesso em: 03/03/2022.

AHMED, U.; MUMTAZ, R.; ANWAR, H.; SHAH, A. A.; IRFAN, R.; GARCÍA-NIETO, J. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water*, vol. 11, n. 11, p. 2210, 2019.

ASSOCIAÇÃO BRASILEIRA DE *STARTUPS*. (2019). Modelo de Negócio para Startups: É melhor criar um SaaS ou uma Plataforma Marketplace? *Abstartups.* Disponível em: https://abstartups.com.br/modelo-de-negocio-para-startups-e-melhor-criar-um-saas-ou-uma-plataforma-marketplace/ Acesso em: 03/03/2022.

AVUÇLU, E.; ELEN, A. Evaluation of train and test performance of machine learning algorithms and Parkinson diagnosis with statistical measurements. *Medical & Biological Engineering & Computing*, vol. 58 n.11, p. 2775–2788, 2020.

BANACHEWICZ, K.; MASSARON, L. *The Kaggle Book*: Data analytics and machine learning for competitive data science. Packt Publishing; 1ª Edition, 2022.

BIRKETT, A. *What Is Customer Satisfaction Score (CSAT)?,* 2021. Disponível em: https://blog.hubspot.com/service/customer-satisfaction-score Acesso em: 03/03/2022.

BREIMAN, L. Random Forests. *Machine Learning*, vol. 45 n. 1, p. 5–32, 2001.

CASTLE, N. What is Semi-Supervised Learning? *Oracle AI & Data Science Blog*, 2018. Disponível em: https://blogs.oracle.com/ai-and-datascience/post/what-is-semi-supervised-learning Acesso em: 03/03/2022

CALLAGHAN, S.; LOSCH, M., PIONE, A.; TEICHNER, W. Sentir-se bem: O futuro do mercado de bem-estar de $ 1,5 trilhão | *McKinsey*, 2021, maio 17. https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/feeling-good-the-future-of-the-1-5-trillion-wellness-market/pt-BR Acesso em: 03/03/2022

COSTI, G. Aprendizagem Não Supervisionada. *Lambda3*, 2020. Disponível em:

https://lambda3.com.br Acesso em: 03/03/2022.
CHAN, D. Y.; VASARHELYI, M. A. Innovation and practice of continuous auditing. *International Journal of Accounting Information Systems*, vol. 12 n.2, p. 152–160, 2011.
ESTRELLA, C. Pandas: Combinando data frames com merge() e concat(). *Data Hackers*, 2020. Disponível em: https://medium.com/data-hackers/pandas-combinando-data-frames-com-merge-e-concat-10e7d07ca5ec Acesso em: 03/03/2022.
EUROMONITOR. Understanding Global Marketplace Trends. *Market Research Report*, 2018. Disponível em: https://www.euromonitor.com/understanding-global-marketplace-trends/report Acesso em: 03/03/2022.
EXAME. *Conheça as 20 empresas mais inovadoras do Brasil, segundo a MIT Tech Review*, 2022. Disponível em: https://exame.com/tecnologia/as-20-empresas-mais-inovadoras-do-brasil/ Acesso em: 03/03/2022.
FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. DE L. F. DE. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC, 2011.
GEM (Global Entrepreneurship Monitor). *Global Entrepreneurship Monitor 2021/2022*. Global Report: Opportunity Amid Disruption. London: GEM, 2022.
GUNAWAN, T. S.; ASHRAF, A.; RIZA, B. S.; HARYANTO, E. V.; ROSNELLY, R.; KARTIWI, M.; JANIN, Z. Development of video-based emotion recognition using deep learning with Google Colab. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18 n. 5, p. 2463–2471, 2020.
HASNAIN, M.; PASHA, M. F.; GHANI, I.; IMRAN, M.; ALZAHRANI, M. Y.; BUDIARTO, R. Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, vol. 8, p. 90847–90861, 2020.
HUTTER, F.; KOTTHOFF, L.; & VANSCHOREN, J. *Automated Machine Learning: Methods, Systems, Challenges*. The Springer Series on Challenges in Machine Learning, 2019.
JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, vol. 349 n. 6245, p. 255–260, 2015.
KANG, K.; RYU, H. Predicting types of occupational accidents at construction sites in Korea using random forest model. *Safety Science*, vol. 120, p. 226–236, 2019.
KUMAR, A.; SYED, A. A.; PANDEY, A. Adoption of online resources to improve the marketing performance of SMES. *Asia-Pacific Journal of Health Management*, vol. 16 n. 3, 2021.
LOOSVELT, L.; PETERS, J.; SKRIVER, H.; DE BAETS, B.; VERHOEST, N. Impact of reducing polarimetric SAR input on the uncertainty of crop classifications based on the random forests' algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50 n.10, p. 4185–4200, 2012.
MATOS, M. P. C. de. Barreiras à adoção dos marketplaces B2C: A relutância das micro e pequenas empresas Portuguesas (Região Norte) em vender na Amazon. Dissertação de Mestrado. Universidade Católica Portuguesa, 2020. Disponível em: https://repositorio.ucp.pt/handle/10400.14/30484 Acesso em: 03/03/2022.
MCCOY, J. T.; AURET, L. Machine learning applications in minerals processing: A review. *Minerals Engineering*, vol. 132, p. 95–109, 2019.
MOHRI, M., ROSTAMIZADEH, A., & TALWALKAR, A. *Foundations of Machine Learning* (2° ed). MIT Press, 2018.
NIWATE, T. S. *Impact on revenue generation of Olist ecommerce company on the basis of various product parameters*. Electronic Theses, Projects, and Dissertations. 1315, 2021. Disponível em: https://scholarworks.lib.csusb.edu/etd/1315 Acesso: 03/03/2022.
OLIST ([s/d]). Olist store. Disponível em: https://olist.com/pt-br/solucoes-para-

comercio/vender-em-marketplaces/ Acesso em: 03/03/2022.
PAÇO, A. DO.; SHIEL, C.; ALVES, H. A New Model for Testing Green Consumer Behaviour. *Journal of Cleaner Production*, vol. 207, 2018.
Padrão, M. Olist unicórnio: startup de soluções de e-commerce vale mais de US$ 1 bilhão. *Canaltech*, 2021. Disponível em: https://canaltech.com.br/startup/olist-unicornio-startup-de-solucoes-de-e-commerce-vale-mais-de-us-1-bilhao-204649/ Acesso em: 03/03/2022.
PINOCHET, L. H. C.; ALVES, F. R. R.; LOPES, E. L.; HERRERO, E.; BRELAZ, G. 'From cloud to the board' - identification of triggers adoption in Brazilian companies. *International Journal of Business Information Systems*, vol. 38, p. 343-366, 2021.
POBEE, F. Modeling the Factors that Influence Ghanaian Entrepreneurs to Adopt e-Commerce. *International Journal of Innovation and Technology Management*, vol. 18 n. 6, 2021.
PORTELA, M. Dados sintéticos: A chave para a inovação sustentável - MIT Technology Review. *MIT Technology Review – Brasil*, 2022. Disponívem em: https://mittechreview.com.br/dados-sinteticos-a-chave-para-a-inovacao-sustentavel/ Acesso em: 03/03/2022.
RODRIGUES, G. A. B.; PLENS, M.; PIANTINO, N. P. DE A. Gestão em vendas *online*: Estudo de caso de empresa calçadista com modelo de negócio em *marketplace*. *Revista Empreenda UniToledo Gestão, Tecnologia e Gastronomia*, vol. 4 n.1, 2020.
SALGADO, D. Papo de Pesquisa: Mercado de Móveis e Decoração. *Opinion Box*, 2021. Disponível em: https://blog.opinionbox.com/moveis-e-decoracao/. Acesso em: 03/03/2022.
SANTOS, H. G. DOS.; NASCIMENTO, C. F. DO.; IZBICKI, R.; DUARTE, Y. A. DE O.; PORTO CHIAVEGATTO, A. D. Machine learning para análises preditivas em saúde: Exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. *Cadernos de Saúde Pública*, vol. 35, 2019.
SCUTARIU, A-L.; SUSU, S., HUIDUMAC-PETRESCU, A-E.; GOGONEA, R-M. A Cluster Analysis Concerning the Behavior of Enterprises with E-Commerce Activity in the Context of the COVID-19 Pandemic. *Journal Theoretical and Applied Electronic Commerce Research*, vol. 17 n. 1, p. 47-68, 2022.
SHEHATA, G. & MONTASH, M. Driving the internet and e-business technologies to generate a competitive advantage in emerging markets Evidence from Egypt. *Information Technology & People*, vol. 33 n. 2, p. 389–423, 2020.
SHETTY, A. S.; JEEVANANDA, S. How to win back the disgruntled consumer? The omni-channel way. *Journal of Business & Retail Management Research (JBRMR),* vol. 12 n. 4, p. 200- 207, 2018.
SILVA, V. M. da. Estudos de futuro e *foresight* para ciência, tecnologia e inovação: tendências do uso de *big data* e *machine learning*. Tese de Doutorado. Universidade Estadual de Campinas, 2021. Disponível em: http://repositorio.unicamp.br/Acervo/Detalhe/1164751 Acesso em: 03/03/2022.
SONI, D. Supervised vs. Unsupervised Learning. *Medium*, 2018. Disponível em: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d Acesso em: 03/03/2022.
SOHAIB, O.; NADERPOUR, M.; HUSSAIN, W.; MARTINEZ, L. Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method. *Computers & Industrial Engineering*, vol. 132, p. 47-58, 2019.
TEMBUSAI, Z. R.; MAWENGKANG, H.; ZARLIS, M. K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification. *International Journal of Advances in Data and Information Systems*, vol. 2 n. 1, p. 1–8, 2021.