# APLICAÇÃO DE *MACHINE LEARNING* NA CLASSIFICAÇÃO DE TERMOS DE PESQUISA EM CAMPANHAS PUBLICITÁRIAS DIGITAIS.

*Yuri Lima Pereira – yurilp92@gmail.com*
*Marcelo Cortimiglia - cortimiglia@producao.ufrgs.br*

**RESUMO**

O presente estudo relaciona *machine learning* e marketing digital, através de um estudo da aplicação de um algoritmo de *machine learning* em termos de pesquisa gerados por uma campanha de mídia paga. O objetivo é identificar os termos de pior performance através de um algoritmo de clusterização, simplificando um processo manual e não assertivo. Utilizou-se um modelo de clusterização *K-Means* para identificar padrões dentro de um grande banco de dados, gerado diariamente por pesquisas feitas na plataforma de busca Google. Obteve-se como resultado um modelo, baseado em dados históricos, que segmenta a base em 7 diferentes clusters, onde analisou-se o cluster de pior performance para a negativação de termos de pesquisa em uma campanha. Com um algoritmo desenvolvido em Python e a partir de uma análise da performance dos estratos, foi possível, portanto, desenvolver um modelo que identificasse os termos de pior performance a serem negativados.

**Palavras-chave:** *machine learning, marketing digital*, clusterização, *K-Means*, campanhas de mídia paga, negativação de termos*, Python.*

## 1. INTRODUÇÃO

A massificação do acesso à Internet transformou a maneira como as empresas interagem com o público, tendo a possibilidade de alcançar um número muito maior de consumidores em potencial. Em janeiro de 2021, a taxa de penetração da internet na população mundial era aproximadamente de 59,5% (*DATAREPORTAL*, 2021) e, segundo um relatório realizado pela Global Web Index (2020), 81% dos usuários da internet na faixa etária de 16 a 64 anos pesquisaram online por um produto ou serviço em junho de 2020.

Os mecanismos de pesquisa têm um papel fundamental nesse processo, pois,

além de atender a busca por informação, facilitam o acesso a sites e produtos específicos. Conforme Ghose e Yang (2008), os mecanismos de pesquisa foram capazes de alavancar seu valor de mercado vendendo publicidade vinculada a consultas geradas pelo usuário que então são encaminhados às páginas da web dos anunciantes. Com o surgimento das campanhas de pesquisas pagas, os principais mecanismos de pesquisa, como Yahoo!, Google, entre outros, transformaram significativamente o comércio online (*BATTELLE*, 2005). Para Kaushik (2009), um dos principais diferenciais da criação e publicação de anúncios nessas plataformas é a ampla e imediata mensuração dos resultados, como a efetivação de uma transação online, quais páginas foram acessadas e quanto tempo foi despendido em cada uma antes da venda, além de dados que permitem esboçar o perfil do usuário comprador.

As agências de marketing digitais são empresas especializadas na prestação de serviços na área de mídias digitais, *Search Engine Optimization* (*SEO*) e *Search Engine Marketing (SEM)*, onde se inserem as pesquisas pagas, sendo responsáveis pela criação, configuração e análise das campanhas digitais, além da entrega dos resultados. Nos últimos anos houve um grande número de novos entrantes no ramo da publicidade em virtude do crescimento da comunicação através de ferramentas digitais de marketing e publicidade. Em consequência disso, as agências devem buscar inovar seus modelos de negócio (*PRUDÊNCIO*, *2018*).

Buscando auxiliar os profissionais de marketing e publicidade digital a tomarem decisões mais acuradas são necessárias diversas ferramentas e sistemas de apoio a decisões, considerando a enorme quantidade de dados que eles lidam diariamente (*KOTLER et al., 2016*). Seguindo a mesma linha de raciocínio, Júnior e Azevedo (2015), consideram que o futuro da tomada de decisões vai ser cada vez mais dominado pelos dados, inteligência na captação, estruturação e utilização dessas informações. Para que isso aconteça, torna-se cada vez mais essencial a integração entre as áreas de marketing e tecnologia para o desenvolvimento de projetos de marketing (*MCKENNA*, 2002). Percebe-se então, por parte das agências, a necessidade de gerir e analisar uma imensa quantidade de dados fornecidos pelas plataformas de pesquisas pagas e a necessidade de se manter competitivo utilizando novas tecnologias para otimização de tarefas operacionais.

A utilização de tecnologias exponenciais, como o *machine learning* (ML), cria oportunidades de vantagem competitiva ao aplicar abordagens orientadas a dados às práticas do marketing digital (*MIKOSLIK et al, 2019*). O aprendizado de máquina pode realizar previsões e apoiar as tomadas de decisões, extraindo insights de grandes quantidades de dados reduzindo o tempo despendido em tarefas operacionais e consequentemente dando mais espaço para a parte estratégica. Para Ma e Sun (2020), o ML aplicado ao marketing digital pode ajudar a ampliar o entendimento de seus consumidores-alvo e a promover e aprimorar a interação com eles.

Existem diversos veículos e formatos de campanhas de mídia paga, sendo a campanha de pesquisa uma das mais relevantes devido a busca ativa por parte do usuário. Visto isso, são necessárias diversas otimizações rotineiras e dentre eles, destaca-se a negativação de palavras-chave. As palavras-chaves negativadas são termos que são excluídos das campanhas para não serem ser associados a um anúncio quando algum usuário fizer uma busca. Essa negativação serve para impedir tráfego irrelevante, evitando cliques que geram custo e não geram receita, melhorando sua performance, tanto em receita quanto em engajamento. É um processo manual, com um grande volume de dados a serem analisados, onde o operador da ferramenta analisa esses termos de forma qualitativa, relacionando os termos pesquisados e os produtos vendidos pelo cliente, e quantitativa, analisando as métricas de cada termo e, dependendo do escopo de produtos oferecidos pelo cliente, é feita de forma arbitrária devido ao tempo que a tarefa.

Diante do exposto, este artigo busca responder o seguinte problema: negativar termos de pesquisa de forma assertiva em campanhas com grandes volumes de dados e produtos variados. Para isso, será utilizado uma abordagem de clusterização desses termos, considerando dados históricos e seus indicadores, com um algoritmo não-supervisionado em *machine learning,* Além da elaboração de um modelo que seja capaz de identificar padrões dentro de uma base de dados, pretende-se analisar qual o conjunto de dados que tem a pior performance em relação a termos com boa performance e realizar a negativação, melhorando a performance da campanha. Além disso, a aplicação pode fornecer aos analistas insights para criação de novas campanhas relacionadas aos clusters com ótimo desempenho. Conforme Lopes e Santos (2019), existe a necessidade da realização de estudos sobre aplicações de *machine learning* no marketing digital,

devido ao reduzido número de estudos presentes na literatura nacional relacionados ao tema.

O artigo está dividido em cinco seções, iniciando-se por esta introdução. A seção 2 consiste no referencial teórico consultado sobre marketing digital, *machine Learning* e possíveis aplicações, seguido pela seção 3, que traz a metodologia usada na implementação de um modelo de *machine Learning* com o intuito de otimizar campanhas de publicidade paga. A seção 4 apresenta os resultados obtidos, bem como uma discussão sobre os mesmos. Por fim, a seção 5 traz a conclusão do artigo e uma avaliação dos próximos passos.

## 2. REFERENCIAL TEÓRICO

Nesta seção são abordados conceitos e temáticas relevantes ao desenvolvimento da pesquisa, tais quais marketing digital, introduzindo conceitos de campanhas de publicidade digital paga, *machine learning* e algoritmos utilizados, além de aplicações de *machine learning* no marketing digital.

### 2.1. Marketing digital

Chaffey & Smith (2008), definem marketing digital como o tipo de marketing que busca aproximar-se dos clientes e compreendê-los melhor, agregando valor aos produtos, ampliando os canais de distribuição e impulsionando as vendas por meio da execução de campanhas usando canais de mídia digital, como mecanismos de pesquisa, publicidade online entre outros. Farias (2016) classifica marketing digital como um conjunto de ações feitas em diversos meios digitais com o objetivo de promover empresas, se diferenciando do marketing tradicional por envolver a utilização de diferentes plataformas, métodos e ferramentas que permitem a análise dos resultados em tempo real.

O marketing digital é facilitado por vários canais digitais, sendo esses, sistemas baseados na Internet que criam, aceleram e transmitem valor para toda cadeia de suprimentos, do produtor ao consumidor. Segundo Ken (2017), os canais de marketing digital mais comuns e amplamente usados são divididos em três domínios: e-mail, mídia

social e marketing de mecanismo de pesquisa (englobando pesquisas orgânicas e anúncios de pesquisas pagos). A Figura 1 destaca esse fluxo e suas principais atividades.

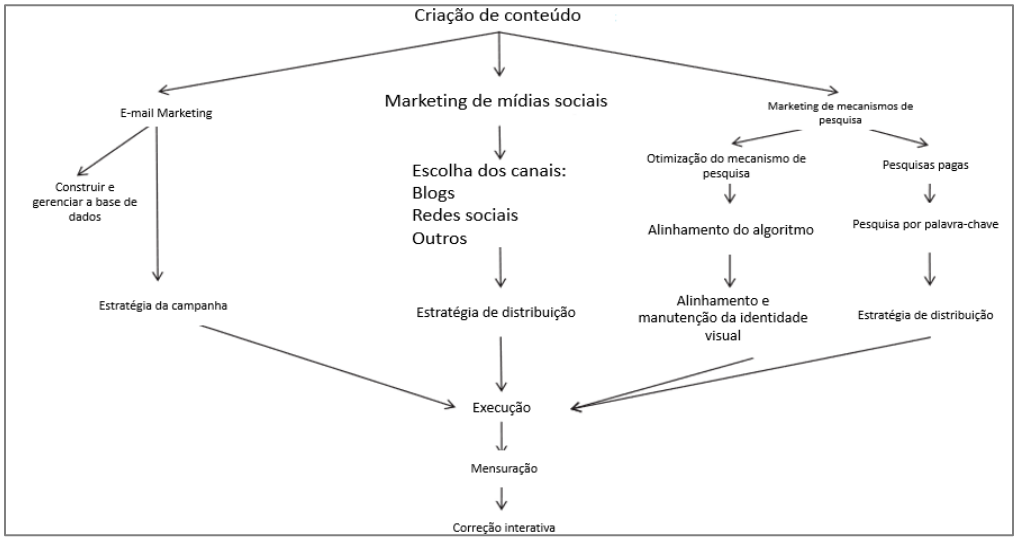



Figura 1: Fluxo de distribuição dos canais digitais. (Fonte: Adaptado de Ken, 2017).

### 2.1.1. Anúncios de pesquisas pagos

O Google Ads, que detinha em 2018 aproximadamente 73% do market-share no mercado de pesquisas pagas (Media Post, 2020), oferece sete tipos de campanhas diferentes em sua plataforma (GOOGLE, 2021a), conforme o Quadro 1.

| **Nome da Campanha** | **Descrição** |
|---|---|
| **Campanha de Rede de Pesquisa** | Campanha da rede de pesquisa: As campanhas da Rede de Display, configura-se anúncios gráficos que são exibidos em sites terceiros e aplicativos em forma de banners. Elas também permitem fazer o acompanhamento de clientes novos e atuais com os anúncios de *remarketing*. (GOOGLE, 2021b). |
| **Campanha de Vídeo** | Com as campanhas de vídeo, exibem-se anúncios em vídeo ou em outro conteúdo de streaming de vídeo no YouTube e em toda a Rede de *Display* do Google. (GOOGLE, 2021c). |

| Nome da Campanha | Descrição |
|---|---|
| **Campanha do Shopping** | Os anúncios de *Shopping* também são ativados por buscas, como na Rede de Pesquisa, porém apenas na parte superior da busca do Google, e na pesquisa do Google Shopping, como uma vitrine de produtos, onde aparece a imagem do produto, título e preço. (GOOGLE, 2021d). |
| **Campanha para App** | Este tipo de campanha tem como objetivo engajamento ou instalação de aplicativos, direcionando o usuário para a página para fazer download do aplicativo. (GOOGLE, 2021e). |
| **Campanha local:** | As campanhas locais ajudam você a atrair as pessoas para as lojas físicas, através de anúncios otimizados automaticamente para exibição na rede de pesquisa, na rede de display, no Google Maps e no YouTube. (GOOGLE, 2021f). |
| **Campanha inteligente** | As campanhas inteligentes foram projetadas para facilitar a experiência do anunciante economizando tempo na configuração e no gerenciamento das campanhas. Esse tipo de campanha permite que os anunciantes escolham as metas de negócios e onde anunciar, e o Google usa o aprendizado de máquina para oferecer resultados de acordo com esses objetivos. (GOOGLE, 2021g). |

Quadro 1: Tipos de campanhas de mídia (Fonte: Google).

Segundo uma pesquisa feita pelo IAB Brasil (2019), 33% dos investimentos em mídias no Brasil são no meio digital, totalizando mais de 16 bilhões de reais em 2018. Os anúncios de pesquisa pagos representam 18% desse investimento (*IAB*, 2019) e, segundo Fain (2006), são uma importante forma de publicidade online, exibindo anúncios patrocinados que correspondem às consultas dos usuários nos mecanismos de pesquisa. Esses mecanismos tiveram um crescimento exponencial e se tornaram o modelo de negócios central das principais empresas de mecanismos de busca (*JANSEN*, 2008).

Nas pesquisas pagas, um conjunto de anúncios é exibido, rotulados como “anúncios patrocinados”, juntamente com os resultados da pesquisa orgânica ao responder a uma consulta (*VARIAN*, 2007). Embora exibidos simultaneamente e em formas semelhantes, os resultados da pesquisa paga são gerados por um mecanismo bastante diferente daquele da pesquisa orgânica. Enquanto os resultados da pesquisa orgânica são produzidos de acordo com a relevância de cada página da web para a

consulta, os resultados da pesquisa pagam são mostrados de acordo com um processo de leilão, onde participam três atores: os anunciantes, a plataforma de pesquisa e os usuários que fazem a pesquisa (AGGARWAL et al, 2006), conforme a Figura 2. Aggarwal et al. (2006) explicam que, quando um usuário faz a pesquisa, o mecanismo de pesquisa primeiro aciona alguns anúncios cujas palavras-chave associadas correspondem à consulta, fornecidos pelos anunciantes antecipadamente. Em seguida, o mecanismo de pesquisa fará um leilão automatizado e em tempo real entre os anunciantes candidatos, considerando a qualidade do anúncio e o preço do lance, que varia de acordo com as faixas pré-determinadas de preço que os anunciantes estão dispostos a pagar.

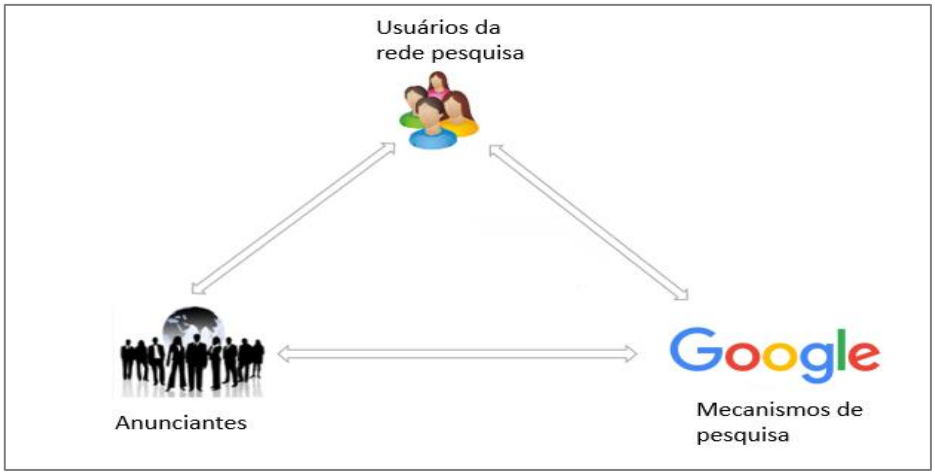



Figura 2: Relação entre os principais atores. (Fonte: Autoria própria).

## 2.2. Machine Learning

*Machine Learning* recebeu várias definições formais na literatura. Alpaydin (2006), definiu *machine learning* como o campo da programação de computadores para otimização de critérios de desempenho usando dados históricos e/ou experiências anteriores. Essas várias definições compartilham a noção de treinamento de algoritmos para execução de tarefas de forma inteligente, além da análise de números tradicionais, aprendendo a partir de exemplos repetidos.

Sterne (2017) conceitua o *machine learning* como uma maneira de o computador usar um determinado conjunto de dados para descobrir como executar uma função

específica por meio de tentativa e erro. A máquina analisa os resultados anteriores, formula uma conclusão enquanto aguarda os resultados dos testes de sua hipótese. Em seguida, consome esses resultados e atualiza seus fatores de ponderação, reconfigurando seu modelo. O aprendizado de máquina refere-se, também, a métodos baseados em computador para extrair padrões ou conhecimento de dados e realizar tarefas de otimização com o mínimo de intervenção humana (*CUI et al., 2006*). A maioria desses métodos tem suas raízes na inteligência artificial e na programação dinâmica. As técnicas de *machine learning* também podem ser divididas em três grandes grupos com base nos tipos de problemas que podem resolver, a saber, a aprendizagem supervisionada, semi-supervisionada e não supervisionada (*HUANG, 2006*).

A aprendizagem supervisionada se refere a qualquer processo de *machine learning* que aprende uma função de um input para um output a partir de exemplos com valores de entrada e saída. Dois exemplos típicos de aprendizagem supervisionada são: classificação e regressão, pontuam Sammut & Webb (2017). Segundo Brei (2020), a aprendizagem supervisionada é a forma mais comumente usada de *Machine Learning* e geralmente envolve classificação ou regressão. Seu objetivo é mapear entradas e saídas com base em um conjunto rotulado de pares de entrada-saída, que geralmente é realizado por humanos. O algoritmo generaliza para responder corretamente a todas as entradas possíveis. As entradas podem ser tão simples quanto um vetor de números, mas também podem incluir objetos mais complexos e estruturados, como imagens, frases, textos, séries temporais, etc. Cui (2006) também relata que os métodos de *machine learning* foram adotados em muitos campos como ferramentas eficazes de mineração de dados para descobrir padrões interessantes e não óbvios ou conhecimentos escondidos em uma base de dados. Esses métodos incluem regras de associação, árvores de decisão, redes neurais e algoritmos genéticos. A aprendizagem não supervisionada contrasta com a aprendizagem supervisionada; a primeira busca entender a estrutura dos dados sem definir um output, já a segunda busca entender a estrutura com os dados de entrada e saída definidos.

O objetivo do aprendizado não supervisionado é encontrar uma estrutura potencialmente útil nos dados. A aprendizagem não supervisionada refere-se a qualquer processo de machine learning que busca entender uma estrutura na ausência de um

output não identificado ou de um feedback. Três exemplos típicos de aprendizagem não supervisionada são clusterização, regras de associação e mapas auto-organizáveis (SAMMUT & WEBB, 2017). A aprendizagem não supervisionada é usada para encontrar categorizações implícitas das observações com base no comportamento dos dados, como por exemplo, agrupar uma base de clientes a partir de suas compras anteriores e, com a intervenção humana, rotular esses grupos com base em suas características gerais (AWAD & KHANNA, 2015)

As regras de associação (*AGARWAL* et al., 1993) podem ser extraídas de conjuntos de dados onde cada exemplo consiste em um conjunto de regras. A descoberta de regras de associação é uma das principais técnicas de mineração de dados e talvez seja a forma mais comum de descoberta de padrões locais em sistemas de aprendizagem não-supervisionados. Ela é capaz de identificar qualquer padrão dentro de um banco de dados, contudo gera um grande volume de informações e uma análise de sua usabilidade é difícil e demorada (*KANTARDZIC*, 2011).

Segundo *Kantardzic* (2011), o algoritmo de mapas auto-organizáveis é um método de análise que produz mapeamentos não lineares de dados para dimensões inferiores. Alternativamente, pode ser visto como um algoritmo de clusterização que produz um conjunto de agrupamentos organizados em uma grade regular. Suas raízes estão na computação neural (redes neurais) e tem sido usado como um modelo abstrato para a formação de mapas ordenados de funções cerebrais, como mapas de características sensoriais. Diversas variantes têm sido propostas, desde modelos dinâmicos a variantes bayesianas. Tem sido amplamente utilizado como uma ferramenta de engenharia para análise de dados, monitoramento de processos e visualização de informações, em vários aplicativos

Atual existem opções alternativas excelentes para muitas das tarefas específicas que os mapas foram utilizados ao longo dos anos, embora até mesmo o algoritmo básico ainda seja viável como uma ferramenta de engenharia versátil em tarefas de análise de dados. Contudo, há também algumas desvantagens, visto que o algoritmo demanda uma máquina robusta e é computacionalmente caro (*SAMMUT & WEBB, 2017*).

Outro tipo de aprendizagem não-supervisionada é a clusterização, cujo objetivo é particionar um conjunto de dados em grupos chamados clusters. Os dados possuem uma

familiaridade maior em questão de comportamento com os dentro do seu cluster do que dos outros clusters. Para medir a semelhança entre os dados, os algoritmos de agrupamento usam várias distorções ou medidas de distância.

Os dois algoritmos mais usados para clusterização são o método hierárquico, que é frequentemente retratado como a abordagem de clusterização de melhor qualidade, mas é limitado por causa de sua complexidade e o K-means, algoritmo desenvolvido por Steinhaus (1956). Embora vários outros algoritmos de agrupamento tenham sido desenvolvidos desde então, o K-means continua sendo um dos métodos mais usados devido à sua simplicidade, facilidade de implementação e eficiência (Jain, 2010). O algoritmo de clusterização K-means calcula a distância entre os pontos médios de uma base de dados, definindo “centros de gravidade” e realocando os dados em grupos de amostra próximos. No final deste processo, dados estarão todos nos seus respectivos centroides, que podem ser consideradas clusters ou categorias separadas de dados. Caso seja necessário a classificação de algum dado novo, o algoritmo realoca para o cluster mais próximo.

Existem diversas formas e fluxos de trabalho para a aplicação de algoritmos de machine learning. Awad e Khanna (2015) desenvolveram um método genérico, simples e eficiente, demonstrado na imagem 3, que é divido em 7 etapas:

Etapa 1 - Coleta e curadoria de dados: geração e seleção do subconjunto relevante e útil de dados disponíveis para a solução de problemas.

Etapa 2 - Pré-processamento de dados: consiste na formatação para um formato adequado, limpando os dados corrompidos e ausentes e transformando os dados conforme necessário, seja normalizando, discretizando, ou suavizando para otimizar a representatividade do conjunto.

Etapa 3 - Transformação de dados: transforma os dados de entrada (muitas vezes uma tabela), utilizando o redimensionamento, decomposição ou uma combinação.

Etapa 4 - Treinamento de algoritmo de aprendizagem: a partir da etapa anterior, divide-se o conjunto de dados em 3 conjuntos: treinamento, validação e conjuntos de dados de teste. O primeiro é utilizado no processo de aprendizagem, onde são obtidos os parâmetros do modelo. Essa etapa geralmente não é necessária para tarefas de aprendizagem não supervisionadas.

Etapa 5 - Teste e otimização do modelo: avalie a eficácia e o desempenho, por meio do conjunto de validação. Os parâmetros que não podem ser aprendidos (os chamados hiper parâmetros) devem ser otimizados usando este conjunto de dados. Uma vez que um conjunto ideal de parâmetros é obtido, o conjunto de teste é usado para avaliar o desempenho do modelo. Se o modelo obtido não for bem-sucedido, as etapas anteriores são repetidas com seleção de dados aprimorada, representação, transformação, amostragem e remoção de outliers, ou alterando o algoritmo completamente.

Etapa 6 – Aplicação de aprendizado de reforço: A maioria das aplicações teóricos de controle requerem um bom mecanismo de feedback para operações estáveis. Em muitos casos, os dados de feedback são esparsos, atrasados ou inespecíficos. Nesses casos, o aprendizado supervisionado pode não ser prático e pode ser substituído por aprendizado reforçado. Em contraste com aprendizagem supervisionada, emprega um desempenho dinâmico para aprender com as consequências das interações com o meio, sem treinamento explícito. Essa etapa não é necessária para tarefas de aprendizagem não supervisionadas.

Etapa 7 – Aplicação execução: usando o modelo validado para fazer previsões sobre dados desconhecidos. O modelo pode ser continuamente retreinado sempre que novos dados estiverem disponíveis.

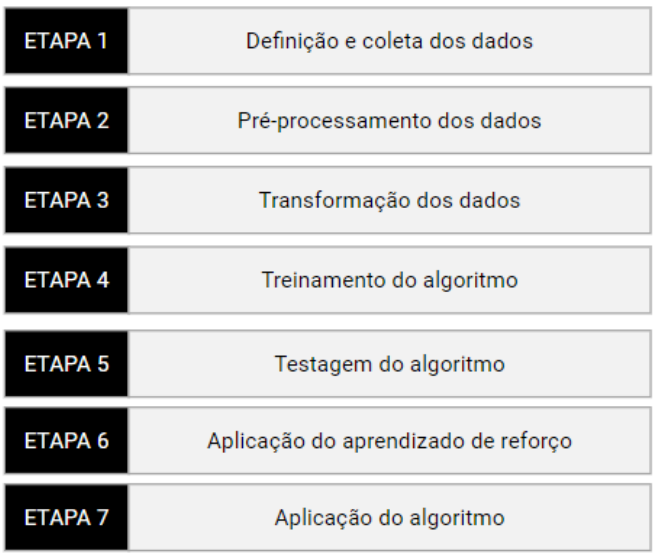



Figura 3: Relação entre os principais atores. (Fonte: AWAD e KHANNA, 2015).

## 2.3 *Machine learning aplicado ao Marketing Digital*

A maior vantagem do marketing digital sobre outras ferramentas e canais de marketing é sua mensurabilidade. A pegada digital de cada usuário da Internet contém uma quantidade significativa de dados que podem servir como inputs para variadas análises (*MIKLOSIK et al., 2019*). Considerando o imenso volume de dados disponíveis pelas plataformas virtuais, era questão de tempo para que a maioria das empresas utilizassem ferramentas automatizadas para a interpretação dos mesmos. Para Sterne (2017), as tecnologias que despontam na segunda década do século XXI, como o *machine learning*, serão popularizadas como serviços muito acessíveis e as empresas que não aderirem ficarão para trás. Miklosik et al. (2019), também afirmam que a expansão tecnológica exponencial cria oportunidades de vantagem competitiva ao aplicar novas abordagens *data-driven* às práticas de marketing digital. *machine learning*, extraindo insights de grandes quantidades de dados, pode gerar um grande impacto e agilizar o processo de tomada de decisões estratégicas das organizações.

A pesquisa quantitativa de marketing tradicionalmente usa modelos estatísticos ou econômicos (*PAUWELS, 2004*). Embora esses modelos muitas vezes tenham fundamentos matemáticos semelhantes aos métodos de aprendizagem de máquina, eles diferem na filosofia de orientação e no foco, e se usados apropriadamente, os métodos de aprendizado de máquina podem complementar abordagens econômicas para expandir a fronteira das pesquisas sobre marketing (*VARIAN, 2016*).

Atualmente, ferramentas analíticas são utilizadas na gestão de marketing para sistematizar processos, agilizar a tomada de decisões e automatizar o trabalho. Essas ferramentas usam o *machine learning* para aprender com os dados históricos e ajudar a planejar atividades futuras com mais eficácia (*HEIMBACH, 2015*).

Miklosik et al. (2019) argumentam que aplicações em *machine learning*, com base em um extenso processamento de dados fornecem informações necessárias para o processo de tomada de decisão dos especialistas em marketing. Segundo os autores, a aplicação de ferramentas baseadas em *machine learning* no marketing digital apresentam vários novos desafios e oportunidades, como a constância no desempenho, agilidade nas tomadas de decisão diminuindo a dependência de fatores subjetivos, além da automação de processos e redução nas taxas de erro.

## 3. METODOLOGIA

### 3.1. Descrição do cenário

A empresa utilizada para a realização do presente trabalho é uma agência de marketing digital, que atua nos setores de tecnologia e comunicação digital, prestando serviços para empresas de tamanhos e segmentos de atuação variados. Fundada em 2000, na cidade de Porto Alegre/RS, a organização iniciou suas operações como agência digital, especializada apenas em serviços de publicidade e marketing em mídias digitais, e posteriormente desenvolvendo soluções em tecnologia e *Customer Relationship Management* (CRM). Dentre esses serviços, a empresa foi pioneira em SEO e gestão de campanhas de links patrocinados. Atualmente, a agência tem atuação nacional e internacional com sedes em São Paulo e Londres.

A agência atua de forma híbrida, com foco em todos os aspectos de uma estratégia de marketing, oferecendo serviços de publicidade, consultoria e tecnologia. Sua proposta de valor busca oferecer estratégias que extraiam o máximo de performance na presença digital dos seus clientes, desde o planejamento à implementação e mensuração dos resultados, que variam de acordo com o objetivo dos mesmos. Atualmente, seu portfólio de serviços é dividido em quatro divisões de negócio: *Business & Strategy* (consultoria em marketing digital); *Digital Operations* (execução e gestão de campanhas de publicitárias digitais); *Solutions*, (venda de softwares próprios) e *Technology* (construção de sites para ecommerce). Possui, aproximadamente, 70% da receita advinda do setor de compra de mídia paga, o qual faz parte da divisão *Business & Strategy*.

A plataforma de anúncios digitais mais utilizada pela empresa é o Google Ads, onde são alocados, mensalmente, aproximadamente 65% dos investimentos em mídias pagas. Existem diversos tipos de campanhas oferecidas pelo Google (ver Quadro 1), contudo, a campanha para qual será desenvolvida a aplicação de *machine learning* é a campanha inteligente, denominada, *dynamic search ads* (DSA). A campanha DSA possui anúncios dinâmicos de pesquisa: o anunciante associa a campanha com uma URL específica e, quando um usuário faz uma pesquisa no Google com palavras relacionadas aos títulos e frases usados por essa URL, o Google Ads o redireciona para uma

ramificação da página escolhida, criando um título claro e relevante para o anúncio de forma autônoma. Ela tem como objetivo abranger os termos que não são cobertos pelas campanhas de pesquisa normais. Cada consulta direcionada pela campanha gera um conjunto de dados que ficam salvos e podem ser consultados no próprio Google Ads (indicadores pré-clique) ou no Google Analytics (indicadores pós-clique), uma plataforma da Google que permite monitorar o tráfego dos sites. Os principais indicadores atrelados às consultas no Google Analytics estão descritos no Quadro 2:

| **Variável** | **Variável numérica** | **Descrição** |
|---|---|---|
| Termo de pesquisa | Termo pesquisado | Termo que redireciona ao site. |
| Impressões | 0 a ∞ | É a exibição de um anúncio, ou seja, é o número de vezes que o anúncio foi visualizado. |
| Cliques | 0 a ∞ | Contabiliza o número de cliques no anúncio veiculado. |
| Custo | 0 a ∞ | É a quantia gasta pela veiculação do anúncio. |
| Sessões | 0 a ∞ | É uma visita ao site na qual o usuário realiza ao menos uma ação. |
| Páginas / sessão | 0 a ∞ | Número de páginas realizadas por sessão |
| Duração média da sessão | 0 a ∞ | Tempo médio de duração das sessões |
| Taxa de rejeição | 0 a 1 | A percentagem de sessões de página única nas quais não existiu interação com a página. |
| Transações | 0 a ∞ | É uma ação mensurável e valiosa para um anunciante, realizada por um usuário que clicou no seu anúncio. |
| Taxa de conversão do comércio eletrônico | 0 a 1 | A taxa de conversão corresponde ao número médio de conversões por interação. |
| Receita | 0 a ∞ | Total de receita gerada. |

Quadro 2: Descrição das variáveis (Fonte: Autoria própria).

Dentre as diversas otimizações dentro de uma campanha de pesquisa do Google, uma que se destaca é a negativação de termos de pesquisa. Segundo o Google (2021h) as palavras-chave negativas permitem excluir termos de pesquisa das suas campanhas e ajudam a segmentar melhor o usuário e a concentrar o tráfego somente nos mais relevantes, aumentando o retorno do investimento. É uma rotina bem importante, realizada ao menos uma vez na semana, onde o operador da plataforma Google Ads, entra na campanha escolhida, e dentro do grupo de anúncios específico, filtra os termos a serem negativados, a partir dos termos pesquisados na semana, relacionando os termos aos produtos oferecidos no e-commerce, e observando suas métricas de performance.

Será utilizado uma base de dados extraída de uma campanha DSA de um dos clientes, um *marketplace* focado em materiais para escritório, papelaria, e artigos diversos, mas que também atua como supermercado. Os dados a serem utilizados serão referentes ao período de 01/05/2021 à 30/09/2021. A campanha foi escolhida pois, além de ter um grande volume de dados históricos, os termos consultados são bem diversificados, principalmente, por existir uma grande variedade de produtos vendidos no site. Essa campanha tem como objetivo principal gerar receita para o cliente, provinda da venda de produtos na plataforma de *e-commerce do mesmo*.

### 3.2. Classificação da pesquisa

A aplicação de um modelo de *machine learning* desenvolvida neste artigo pode ser considerada, quanto à sua natureza, como uma pesquisa aplicada, pois busca gerar conhecimento, a partir de dados e informações já existentes, para solução de um problema específico. A pesquisa classifica-se como quantitativa, quanto à abordagem, pois utiliza uma base de dados numérica gerados pelos termos pesquisados por usuários que engajaram com os anúncios do Google Ads, além de utilizar diferentes análises estatísticas. Quanto ao objetivo de pesquisa, pode ser caracterizada como explicativa, considerando que busca analisar e classificar os termos de busca e determinar as variáveis mais influentes. Sobre os procedimentos, ela pode ser classificada como

experimental, uma vez que se definiu um objeto de estudo e se fez a avaliação das variáveis que o influenciam.

### 3.3. Etapas do procedimento metodológico

Foi utilizado como base para a metodologia deste artigo o framework de sete etapas para abordagem de problemas de *machine learning*, desenvolvido por Awad e Khanna (2015). Na primeira etapa ocorre a coleta dos dados, selecionando dentro dos atributos disponíveis os que mais podem ser úteis para a resolução do problema, visto que a seleção de todos os dados pode ser contraproducente. Logo após, na etapa dois, é feito o pré-processamento dos dados, reestruturando e combinando as bases em um único banco de dados em um formato a ser adaptado as necessidades da ferramenta a ser utilizada. Além disso, é feita uma limpeza dos removendo, substituindo ou corrigindo dados corrompidos e muito discrepantes, e normalizando os dados para uso eficiente.

Na terceira etapa é feita a transformação dos dados através do redimensionamento. Para esta etapa foi feito a utilização do *Principal Component Analysis*, ou Análise de Componentes Principais, método que tem como objetivo a redução da dimensionalidade da base, transformando um grande conjunto de variáveis em um menor que ainda contém a maior parte das informações do grande conjunto. A partir disso, o algoritmo é testado (na etapa 5) e aplicado na base de dados (etapa 7). Como foi utilizado um algoritmo de aprendizagem não-supervisionada, não foram necessárias as etapas 4 e 6: o treinamento do algoritmo, visto que não se tem um valor alvo definido e a aplicação de um algoritmo de reforço, respectivamente.

Por se tratar de um problema de clusterização, foi escolhido um método não-supervisionado denominado K-*Means.* A aplicação do algoritmo e análises serão realizadas utilizando a ferramenta Python, devido ao seu fácil acesso e sua extensa biblioteca de pacotes especializados em análise de dados e *machine learning*. Todos os códigos utilizados são apresentados em ANEXO II.

Por fim, após a separação da base em clusters e realizadas as análises dos resultados, identificou-se o cluster com o pior desempenho em relação a performance da campanha e negativou-se os termos presentes no cluster na campanha e,

posteriormente, utiliza-se o modelo gerado para a negativação dos termos de forma automatizada. Os resultados obtidos e conclusões constatadas serão descritos nos capítulos a seguir.

# 4. RESULTADOS

## 4.1. Coleta dos dados e escolha das variáveis

Para a seleção das variáveis relevantes, definiu-se junto aos analistas da empresa as métricas mais relevantes para uma campanha de mídia focada em performance, relacionadas ao desempenho dos termos que seriam o custo gerado pelo termo, o número de cliques que esse termo levou, o volume de impressões (número de visualizações do anúncio), as sessões geradas (uma sessão é gerada quando o usuário interage com a página dentro de um curto período de tempo), transações realizadas, taxa de conversão de cada termo (transações por sessão) e a receita. Além disso, definiu-se métricas relacionadas ao comportamento do usuário dentro do site a partir de um anúncio (duração média da sessão, número de páginas visualizadas por sessão e taxa de rejeição), totalizando 10 variáveis diferentes. Como cada termo pesquisado tem um custo e que o principal objetivo da campanha estudada é gerar receita para o e-commerce, essa clusterização tem como objetivo identificar padrões de comportamento dos termos e negativar os piores termos em relação a resultado, aumentando a performance da campanha. Contudo, não é interessante negativar todos os termos que não geram receita, visto que não necessariamente os usuários vão realizar transações sempre. Por esse motivo, outro ponto importante a ser observado é o comportamento do usuário dentro da plataforma, mensurados pela duração média da sessão, tempo que o usuário despende no site a partir de uma sessão (quanto maior melhor), taxa de rejeição, que mede a porcentagem de sessões que não geraram interações com a página (quanto menor melhor) e páginas por sessão, que mede o número de páginas visitadas em cada sessão (quanto maior melhor).

A partir disso, foram importados da plataforma Google Analytics todos os termos pesquisados entre o período de 1 de maio a 30 de setembro de 2021, totalizando 50.238 termos, juntamente com as métricas atreladas a cada termo utilizando a ferramenta de integração Supermetrics, que consolidou todos os valores em uma única base.

Posteriormente, foi realizada a limpeza dos dados inconsistentes da base referente ao período, eliminando os termos nos quais não haviam gerado nenhum custo, reduzindo a base para 46.421.

### 4.2. Pré-processamento dos dados

A seguir, foram realizadas a transformação do banco de dados em um formato apropriado para a utilização da linguagem Python e a importação para a ferramenta. Como última etapa do pré-processamento dos dados foi feito então uma segunda limpeza na base, normalizando todas as colunas com valores contínuos para evitar que a diferença de magnitude entre as variáveis inflasse a importância de algumas em detrimento de outras e retirando os outliers, valores que se diferenciam drasticamente de todos os outros e que acabam tendenciando ou causando anomalias nos resultados obtidos. Com isso foram excluídas 1.617 linhas que continham métricas com valores 4 desvios padrões acima da média, restando 44.750 termos. Sendo assim, chegou-se na base de dados final, apresentada na Figura , com a base de dados com todos os termos que não foram excluídos durante a etapa de pré-processamento, e com cada linha representando um termo e suas métricas, a qual será a aplicada o método PCA para reduzir sua dimensionalidade.

| Termo de Pesquisa | Custo | Clique | Impressões | Sessões | Duração Média da Sessão | Média de Páginas por Sessão | Taxa de Conversão | Transações | Receita | Taxa de Rejeicão |
|---|---|---|---|---|---|---|---|---|---|---|
| headset usb | 188.15 | 86 | 2786 | 107 | 158 | 3.31 | 0.0467 | 5 | 60783.09 | 0.3551 |
| prateleira de ferro | 35.59 | 22 | 443 | 83 | 447 | 7.04 | 0.3133 | 26 | 26575.25 | 0.2651 |
| celular | 2507.99 | 247 | 16369 | 254 | 200 | 3.55 | 0.0866 | 22 | 23067.68 | 0.4173 |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| microondas | 1148.25 | 231 | 13098 | 344 | 352 | 5.08 | 0.093 | 32 | 17143.71 | 0.4593 |
| papel cartao estampado | 5.01 | 1 | 5 | 20 | 392 | 9.65 | 0.65 | 13 | 16096.51 | 0.1 |

Figura 4: Base de dados processada (Fonte: Autoria Própria).

### 4.3. Transformação da base e aplicação do algoritmo

Como a base de dados possui 10 dimensões, buscando facilitar as análises e diminuir valores residuais, foi aplicado o PCA na base (Figura ). O PCA é um procedimento estatístico que permite resumir a informação em grandes tabelas de dados por meio de um conjunto menor de “índices de resumo” que podem ser mais facilmente visualizados e analisados, transformando uma base de dez dimensões em duas.

Depois da redução de dimensionalidade, foi utilizado o algoritmo *K-Means* na base, utilizando a ferramenta Python. Como um grande desafio durante a clusterização é a definição do número de clusters a serem estudados, em primeiro momento geramos 10 amostras, começando com um cluster de 2 estratos (C2), 3 estratos (C3), 4 estratos (C4), e assim por diante, até gerar um cluster com 10 estratos (C10). Podemos ver na tabela 2, a alocação dos termos de pesquisa dentro dos respectivos clusters. Os gráficos gerados pelas 10 amostras estão apresentados no ANEXO I.

| Cluster | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 42859 | 42632 | 40838 | 33489 | 31781 | 31626 | 836 | 436 | 10559 |
| 1 | 1891 | 1029 | 472 | 7539 | 956 | 718 | 25945 | 24921 | 445 |
| 2 | 0 | 1089 | 1004 | 470 | 870 | 838 | 320 | 317 | 310 |
| 3 | 0 | 0 | 2436 | 2279 | 7215 | 3617 | 3527 | 1114 | 24683 |
| 4 | 0 | 0 | 0 | 973 | 3632 | 344 | 280 | 188 | 1179 |
| 5 | 0 | 0 | 0 | 0 | 296 | 280 | 10845 | 10599 | 668 |
| 6 | 0 | 0 | 0 | 0 | 0 | 7327 | 2296 | 2298 | 131 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 701 | 705 | 4307 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4172 | 165 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2303 |

Figura 5: Distribuição dos termos dentro dos clusters (Fonte: Autoria Própria).

Geralmente, à medida que a quantidade de clusters aumenta no *K-Means*, as diferenças no comportamento entre os clusters se tornam muito pequenas, e as diferenças das observações intra-clusters vão aumentando. É preciso achar um equilíbrio em que as observações que formam cada agrupamento sejam o mais homogêneas possíveis e que os agrupamentos formados sejam os mais diferentes uns dos outros. Para determinar o número ideal de clusters para o modelo *K-Means*, foi feita uma medição da soma das distâncias quadradas dos pontos até o centro do cluster mais

próximo, conhecida como inércia. O Gráfico 1 mostra que, após 7 estratos, a mudança no valor da inércia não é mais significativa, definindo então que o número ideal de clusters para amostra é 7.

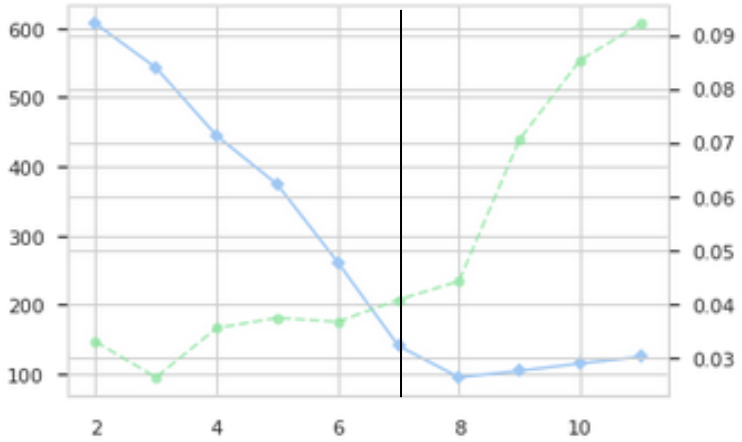



Gráfico 1: Definição do cluster ótimo (Fonte: Autoria Própria).

Após a definição do número ideal de clusters, foi gerado um gráfico representando a base projetada em duas dimensões (Gráfico 2), a partir da aplicação do PCA, e segmentada em 7 clusters distintos, cada um representado por uma cor e com comportamento distintos dos outros.

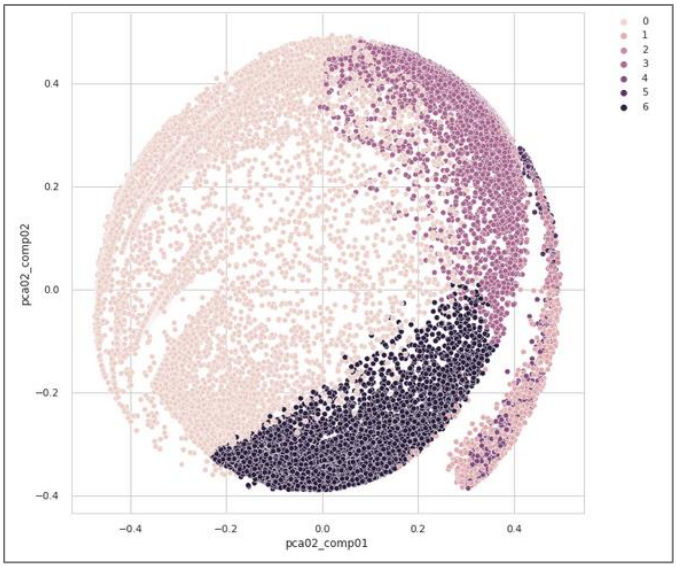



Gráfico 2: Base segmentada em 7 clusters (Fonte: Autoria Própria).

### 4.4. Análise da performance dos clusters

Após a aplicação do algoritmo, todos os termos foram identificados por clusters e, então, foi gerada uma tabela com as médias de cada cluster em relação às variáveis citadas anteriormente. A Figura  mostra as médias das métricas dos termos de pesquisa por cluster.

| Cluster | Custo | Cliques | Impressões | Sessões | Duração Média da Sessão | Média de Páginas por Sessão | Taxa de Conversão | Transações | Receita | Taxa de Rejeicão | N. de Termos |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | R$ 2.06 | 1.31 | 17.4 | 1.2 | 8.6 | 1.03 | 0% | 0.00 | R$ 0.00 | 69% | 31626 |
| 1 | R$ 12.29 | 4.33 | 88.6 | 7.86 | 421.66 | 5.41 | 16% | 0.98 | R$ 315.50 | 29% | 718 |
| 2 | R$ 29.95 | 13.89 | 388.98 | 14.59 | 115.52 | 2.26 | 1% | 0.12 | R$ 32.42 | 55% | 838 |
| 3 | R$ 11.76 | 5.19 | 60.4 | 4.2 | 71.59 | 1.8 | 0% | 0.00 | R$ 0.00 | 75% | 3617 |
| 4 | R$ 17.82 | 4.85 | 118.32 | 10.97 | 556.12 | 6.59 | 21% | 1.91 | R$ 1,073.06 | 23% | 344 |
| 5 | R$ 57.96 | 28.24 | 754.98 | 31.21 | 186.43 | 3.05 | 3% | 0.86 | R$ 371.77 | 49% | 280 |
| 6 | R$ 2.73 | 1.43 | 17.73 | 1.66 | 241.21 | 4.03 | 0% | 0.00 | R$ 0.00 | 7% | 7327 |

Figura 6: Média das variáveis por cluster (Fonte: Autoria Própria).

Os termos dentro dos clusters 1, 2, 4, e 5, resultaram em receita e vão ser desconsiderados da análise. Foi feita, então, uma análise comparativa entre os clusters que não resultaram em transações (os clusters 0, 3 e 6), em relação à média dos termos que geraram receita no mesmo período, buscando entender o cluster com o pior desempenho em relação a retorno financeiro e, também, ao comportamento do usuário dentro do site. A Figura  mostra a média das métricas de performance e comportamento dos 2441 termos que geraram receita no mesmo período, totalizando R$ 2.875.327,13.

| Custo | Clique | Impressões | Sessões | Duração Média da Sessão | Média de Páginas por Sessão | Taxa de Conversão | Transações | Receita | Taxa de Rejeicão |
|---|---|---|---|---|---|---|---|---|---|
| R$ 41.60 | 16.24 | 435.56 | 22.06 | 541.57 | 6.76 | 10% | 2.19 | R$ 1,177.93 | 28% |

Figura 7: Média das variáveis dos termos que geraram receita (Fonte: Autoria Própria).

Analisando o comportamento dos clusters de baixa performance em relação ao comportamento médio dos que geram receita, conforme a figura 7, é perceptível que o cluster 6 tem um bom desempenho nas métricas de comportamento comparado com os outros dois clusters, apresentando uma duração média das sessões apenas 55% abaixo da média, 40% abaixo da média de páginas por sessão e uma taxa de rejeição 76% menor. O cluster 3 possui um custo maior, com mais cliques e métricas de

comportamento semelhantes ao cluster 0, trazendo um impacto negativo pela maior veiculação das palavras, gerando, por consequência, um maior custo por clique, além de trazer usuários com qualificação inferior para o site devido ao baixa média de duração das sessões, número reduzido de páginas por sessão e alta taxa de rejeição. Além disso, pelo cluster 0 ser o conjunto com o maior número de entradas, possui termos que tiveram uma menor veiculação, com 92% menos cliques, 96% menos impressões, baixo engajamento do usuário no site, contudo, não é interessante excluir os mesmos devido ao pequeno volume de dados. Devido a isso, o cluster de pior performance relativa aos termos de bom desempenho é o cluster 3, com uma duração média de sessão 73% inferior, uma taxa de rejeição 172% maior do que a amostra comparada, e uma quantidade considerável de cliques, gerando custo e tráfego não qualificado para o site, mas não gerando nenhuma receita.

| Cluster | Custo | Clique | Impressões | Sessões | Duração Média da Sessão | Média de Páginas por Sessão | Taxa de Rejeicão |
|---|---|---|---|---|---|---|---|
| 0 | -95% | -92% | -96% | -95% | -98% | -85% | 149% |
| 3 | -72% | -68% | -86% | -81% | -87% | -73% | 172% |
| 6 | -93% | -91% | -96% | -92% | -55% | -40% | -76% |

Figura 3: Comparação dos clusters 0, 3 e 6 em relação aos termos que geraram receita. (Fonte: Autoria Própria).

Definido o cluster (3) e, consequentemente, os termos a serem negativados, é necessário adicionar os mesmos na ferramenta e criar uma rotina semanal de negativação, realizando a importação de uma base de mesmo formato e adicionando no algoritmo criado.

## 5. CONCLUSÕES

Ferramentas que conseguem processar grandes quantidades de dados, extrair informações e auxiliar na diminuição de tempo despendido em tarefas operacionais são uma das principais chaves para a sobrevivência das empresas no contexto atual. Tendo isso em vista, esse artigo buscou, com uma abordagem de *machine learning*, identificar padrões nos termos de pesquisa gerados por uma campanha de publicidade digital paga,

através da aplicação de um algoritmo de clusterização, com o objetivo de negativar os termos de pior performance, buscando aumentar a eficiência das campanhas.

Para isso, foi realizada uma pesquisa de natureza aplicada a partir do levantamento de dados de uma campanha de publicidade digital de uma agência de publicidade digital de Porto Alegre. Utilizou-se uma metodologia para aplicação de *machine learning* em uma base de dados de uma campanha de publicidade de um de seus clientes, com um algoritmo de clusterização *K-means.*

A base de dados deste estudo é baseada em métricas relacionadas ao desempenho de termos pesquisados no Google e que geraram tráfego ao site, através de anúncios de campanhas de mídia digitais pagas. Foram utilizadas variáveis numéricas que representam tanto a performance em questões financeiras como questões de comportamento do usuário dentro do site e, por isso, foram considerados apenas os termos que geraram algum custo para a empresa.

O algoritmo segmentou a base em 7 estratos diferentes, e a partir disso, foram feitas análises nos clusters que agruparam os termos que não geraram transações e receita, sendo realizada uma comparação com o comportamento dos termos que geraram receita no mesmo período. A partir da clusterização da base de dados gerada pelo modelo, juntamente com a análise dos clusters de pior performance, foi possível identificar padrões dentro da base de dados, entendendo qual cluster deveria ser negativado. Definiu-se então o cluster de pior performance para ser negativado na campanha, contendo os termos que não geravam receita, com um custo relativamente alto e com um baixo engajamento por parte dos usuários. O cluster selecionado teve um desempenho inferior em relação a duração média das sessões (-87%), média de páginas por sessão (-73%) e com uma taxa de rejeição 172% maior que a média dos termos que geraram receita. Além disso, foi o cluster que a teve a maior veiculação, gerando o maior custo médio, sem trazer nenhum resultado.

Com o modelo gerado, o responsável pela negativação das palavras precisa apenas importar os novos termos pesquisados na semana para o algoritmo, que vai realocar eles nos clusters definidos anteriormente e depois negativar essa lista de termos na campanha via Google Ads. Essa aplicação gerou dois ganhos significativos: a simplificação da rotina de negativação culminando em um aumento da produtividade do

analista de mídia e o aumento da assertividade da negativação, que era feita de forma arbitrária, visto a grande variedade de produtos do cliente e os termos pesquisados.

A principal limitação deste trabalho foi não ter avaliado o impacto da automatização em questão de tempo despendido pelo analista para a realização da tarefa. Em relação a trabalho futuros, podem ser consideradas aplicações de algoritmos de clusterização mais robustos, assim como uma comparação de desempenho entre os mesmos, além da inclusão de outras variáveis no banco de dados.

## BIBLIOGRAFIA


AGARWAL, R., IMIELINSKI, T., AND SWAMI, A. N. (1993). Mining association rules between sets of items in large databases. In proceedings of the 1993 ACM SIGMOD International Conference on Management of Data

AGGARWAL, G.; GOEL, A.; MOTWANI, R. (2006) Truthful auctions for pricing search keywords. In: Proceedings of the 7th ACM conference on electronic commerce. New York: ACM; p. 1–7.

ALPAYDIN E. (2014) Introduction to machine learning. 3rd ed. Cambridge, MA: The MIT Press.

AWAD, M.; KHANNA, R. (2015). Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress Open.

BATTELLE, J. (2005) The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture, New York, Penguin Group.

BREI, V. A. (2020). *Machine Learning* in Marketing: Overview, Learning Strategies, Applications, and Future Developments. Foundations and Trends® in Marketing, 14(3).

BOCK H.H. (2007). Clustering Methods: A History of K-Means Algorithms. In: Brito P., CUCUMEL G., BERTRAND P., DE CARVALHO F. (eds) Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg.

CHAFFEY, D.; SMITH, P. (2008). Emarketing Excellence: Planning and Optimizing your Digital Marketing. Ed. Routledge.

CHOLLET, F.; ALLAIRE, J.J. (2018). Deep Learning with R. Manning Publications Company.

CUI, G.; WONG, M. L.; LUI, H.K. (2006). *Machine Learning* for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. Management Science, 52(4), 597–612.

Datareportal – Global Digital Insights. Digital 2020: July Global Statshot. Disponível em: https://datareportal.com/reports/digital-2020-july-global-statshot. Acesso em: 16 de abril de 2021.

FARIAS, F. (2016). Série Épicos: O Guia Definitivo do Marketing Digital. Resultados Digitais.

FAIN D.C.; PEDERSEN J.O. (2006). Sponsored search: A brief history. Bull Am Soc Inf Sci Technol. 32(2):12–13.2.

FERREIRA J.; AZEVEDO, N. (2015). Marketing digital: uma análise do mercado 3.0. Curitiba: Ed. Intersaberes.

GOOGLE (2021a). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/2567043?hl=pt-BR>. Acessado em: 15 de abril 2021.

GOOGLE (2021b). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/2404190>. Acessado em: 15 de abril 2021.

GOOGLE (2021c). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/6340491>. Acessado em: 15 de abril 2021.

GOOGLE (2021d). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/7457632>. Acessado em: 15 de abril 2021.

GOOGLE (2021e). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/2454022>. Acessado em: 15 de abril 2021.

GOOGLE (2021f). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/9118422>. Acessado em: 15 de abril 2021.

GOOGLE (2021g). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/7457632>. Acessado em: 15 de abril 2021.

GOOGLE (2021h). Site para consulta. Disponível em < https://support.google.com/google-ads/answer/2453972>. Acessado em: 20 de novembro 2021.

GHOSE, A.; YANG, S. (2009). An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. Management Science,55(10), 1605-1622.

Global Web Index. (2020). Connecting the dots: Consumer trends that will shape 2020. Disponível em: https://www.globalwebindex.com/reports/trends-2020. Acesso em: 15 de abril de 2021.

HAGEN, L.; UETAKE, K.; YANG, N. et al. (2020). How can machine learning aid behavioral marketing research? Mark Lett 31, 361–370.

HEIMBACH, D.; S. KOSTYRA; O. HINZ. (2015). ‘‘Marketing automation,’’ Bus. Inf. Syst. Eng., vol. 57, no. 2, pp. 129–133.

HUANG, T.M.; KECMAN, V.; KOPRIVA, I. (2006). Kernel Based Algorithms for Mining Huge Data Sets, Studies in Computational Intelligence (SCI)17, 1–9.

IAB Brasil (*International Advertising Bureau* Brasil). Pesquisa Digital Adspend 2019. Disponível em < https://iabbrasil.com.br/pesquisa-digital-adspend-2019/ >. Acesso em: 15 de abr. 2021.

JAIN, A.K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters,

Volume 31, Issue 8.

JANSEN B.J.; MULLEN, T. (2008). Sponsored search: an overview of the concept, history, and technology. Int J Electron Bus. 6(2):114–30.

KOTLER, P; KARTAJAYA, H.; SETIAWAN, I. (2016) Marketing 4.0: Moving from traditional to digital. John Wiley & Sons, 2016.

KAUSHIK, A. (2010). Web Analytics 2.0: the art of online accountability & science of customer centricity. Indianapolis: Wiley.

KANTARDZIC, M. (2011). DATA MINING Concepts, Models, Methods, and Algorithms.

KEY, T. M. (2017). Domains of Digital Marketing Channels in the Sharing Economy. Journal of Marketing Channels, 24(1-2), 27–38.

LOPES, W.; SANTOS, J. (2019). A inteligência artificial aplicada ao marketing digital: um estudo prospectivo sobre tecnologias emergentes. International Symposium on Technological Innovation, Vol.10/n.1/ p.1121-1130.

Media Post. (2021). Site para consulta. Disponível em < https://www.mediapost.com/publications/article/341981/amazon-takes-us-search-ad-share-from-google.html>. Acessado em: 13 de abril 2021.

MA, L.; SUN, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. International Journal of Research in Marketing.

Marketing Week. (2017). “Rise of the machines: Are robots after your job?”. Disponível em: <https://www.marketingweek.com/rise-of-the-machines/>. Acesso em: 16 de abril de 2021.

MCKENNA, R. Total Access. (2002). Giving Customers What They Want in an Anytime, Anywhere World. Boston, Massachusetts: Harvard Business School, Print.

MIKLOSIK, A; KUCHTA, M.; EVANS, N.; ZAK, S. (2019). Towards the adoption of machine learning-based analytical tools in digital marketing. IEEE Access, 1–1.

PAUWELS, K.; CURRIM, I.; DEKIMPE, M.G. et al. (2004). Modeling Marketing Dynamics by Time Series Econometrics. Market Lett 15, 167–183.

PRUDÊNCIO, F. M. (2018). O modelo de negócio das agências de publicidade e propaganda na era digital: um estudo exploratório no contexto paulista.

SAMMUT, C.; WEBB, G. I. (2017). Encyclopedia of Machine Learning and Data Mining.

STEINHAUS, H. (1956). Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV:801

STERNE, J. (2017). Artificial Intelligence for Marketing.

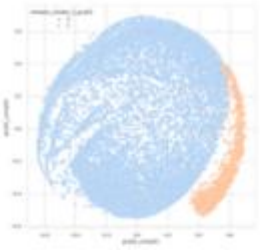
VARIAN, H. R. (2007) Position auctions. Machine Learning for Search Ranking and Ad Auctions

VARIAN, H. R. (2016). Causal inference in economics and marketing. Proceedings of the National Academy of Sciences.

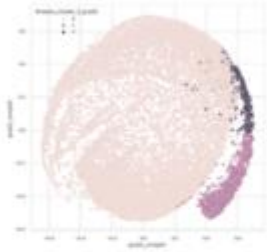
## ANEXOS

### ANEXO I – GRÁFICO DOS CLUSTERS GERADOS PELO ALGORÍTIMO

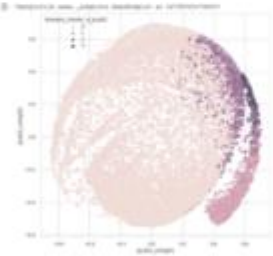
C2 – 2 Clusters



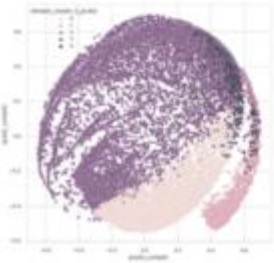
C3 – 3 Clusters



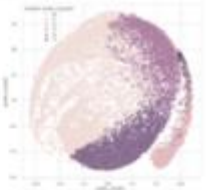
C4 – 4 Clusters



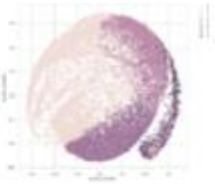
C5 – 5 Clusters



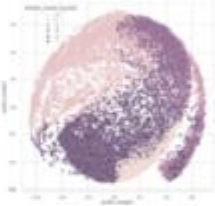
C6 – 6 Clusters



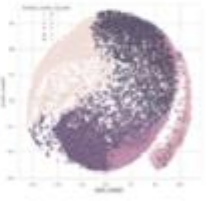
C7 – 7 Clusters



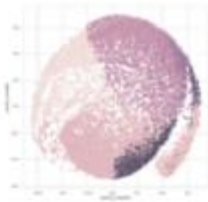
C8 – 8 Clusters



C9 – 9 Clusters



C10 – 10 Clusters



## ANEXO II – CÓDIGO UTILIZADO PARA A FORMATAÇÃO E TRANSFORMAÇÃO DA BASE DE DADOS, E APLICAÇÃO NA BASE DE DADOS

```
[ ] import pandas as pd
    from sklearn.preprocessing import StandardScaler
    from sklearn.preprocessing import Normalizer
    from sklearn.decomposition import PCA
    from sklearn.cluster import KMeans
    import matplotlib.pyplot as plt
    import seaborn as sns
    from mpl_toolkits.mplot3d import Axes3D
    from matplotlib.colors import ListedColormap
    import math
    from datetime import datetime
    from yellowbrick.cluster import KElbowVisualizer

    sns.set_theme(style="whitegrid", palette="pastel")
```

**IMPORTANDO E PREPROCESSANDO DADOS**

```
[ ] # Importando dados
    df = pd.read_csv('data.csv')

    # Renomeando colunas, dropando valores nulos e settando o nome da palavra como índice
    df.rename(columns={'Ad matched query': 'word'}, inplace=True)
    df.set_index('word', inplace=True)
    df.dropna(inplace=True, how='any')

    # Criando listas com todas as palavras e nomes de colunas (para ajustes de tabelas futuras)
    words_list = [word for word in df.index]
    column_names = [column for column in df.columns]

    # Printando primeiras cinco linhas dos dados
    df
```

```
    # Normalizando/Standardization dados (tranformando para distribuição normal)
    scaler = StandardScaler()
    scaler.fit(df)
    scaled_df = pd.DataFrame(scaler.transform(df))

    # Adicionando nomes de palavras, settando índices e nomes de colunas
    scaled_df['word'] = words_list
    scaled_df.set_index('word', inplace=True)
    scaled_df.rename(columns=dict(enumerate(column_names)), inplace=True)

    # Cortando outliebrs
    desvios_para_cortar = 4

    df_without_outliers = scaled_df.copy()
    for column in scaled_df.columns:
      df_without_outliers = df_without_outliers.loc[(df_without_outliers[column] <= desvios_para_cortar) & (df_without_outliers[column] >= - desvios_para_cortar)]

    # Printando primeiras cinco linhas dos dados
    df_without_outliers
```

```
[ ] selected_words = df_without_outliers.index.to_list()

    df.reset_index(inplace=True)
    df['to_keep'] = df.word.isin(selected_words)

    selected_df = df.copy()
    selected_df = selected_df.loc[selected_df.to_keep == True]
    selected_df.drop(columns='to_keep', inplace=True)
    selected_df.set_index('word', inplace=True)
    df.drop(columns='to_keep', inplace=True)

    selected_df.head()
```

```
[ ] # Normalizando/Standardization dados (tranformando para distribuição normal)
    scaler = StandardScaler()
    scaler.fit(selected_df)
    scaled_df = pd.DataFrame(scaler.transform(selected_df))

    # Adicionando nomes de palavras, settando índices e nomes de colunas
    scaled_df['word'] = selected_words
    scaled_df.set_index('word', inplace=True)
    scaled_df.rename(columns=dict(enumerate(column_names)), inplace=True)

    # Printando primeiras cinco linhas dos dados
    scaled_df.head()
```

**APLICANDO PCA PARA REDUÇÃO DE DIMENSIONALIDADE**

```
    # Criando e aplicando PCA para 2 componentes (2D)
    pca02 = PCA(n_components=2)
    pca02_results = pca02.fit_transform(scaled_df)
    print(f'PCA02 Explained Variance: {pca02.explained_variance_}')
    print(f'PCA02 Explained Variance Ratio: {pca02.explained_variance_ratio_}')
    scaled_df['pca02_comp01'] = pca02_results[:,0]
    scaled_df['pca02_comp02'] = pca02_results[:,1]
```

```
PCA02 Explained Variance: [4.28733655 2.40305526]
PCA02 Explained Variance Ratio: [0.42871096 0.24029286]
```

```
[ ] # Printando primeiras cinco linhas dos dados
    scaled_df.head()
```

**APLICANDO K-MEANS PARA CLUSTERIZAR DADOS**

```
[ ] clusters_columns_names = []
    max_clusters = 11

    # Aplicando Kmeans para PCA de 2 componentes (de 2 a 5 clusters)
    for id in range(2, max_clusters):
      kmeans_model = KMeans(n_clusters=id, random_state=420)
      scaled_df[f'kmeans_cluster_{id}_pca02'] = kmeans_model.fit_predict(scaled_df[['pca02_comp01', 'pca02_comp02']])
      clusters_columns_names.append(f'kmeans_cluster_{id}_pca02')

    scaled_df.head()
```

**NORMALIZANDO EIXOS DO PCA E GERANDO DATASET FINAL**

```
[ ] # Normalizando dados em MAX MIN(scaling de 0 até 1)
    normalizer = Normalizer()
    normalizer.fit(scaled_df.drop(columns=clusters_columns_names))
    final_df = pd.DataFrame(normalizer.transform(scaled_df.drop(columns=clusters_columns_names)))

    # Adicionando nomes de palavras, settando indices e nomes de colunas
    final_df['word'] = selected_words
    final_df.set_index('word', inplace=True)
    final_df.rename(columns=dict(enumerate([column for column in scaled_df.columns])), inplace=True)
    final_df.drop(columns=column_names, inplace=True)

    # Adicionando colunas com clusters
    for column in clusters_columns_names:
      final_df[column] = scaled_df[column]

    # Printando primeiras cinco linhas dos dados
    final_df = final_df.merge(df.set_index('word'), how='left', left_index=True, right_index=True)
    final_df.head()
```

```
[ ] data_frames_list = []
    for column in clusters_columns_names:
      to_group_columns_list = [og_column for og_column in df.columns if og_column != 'word']
      to_group_columns_list.append(column)
      grouped_df = final_df[to_group_columns_list].groupby(by=column).mean().reset_index()
      grouped_df['model_experiment'] = column
      grouped_df.drop(columns=[column], inplace=True)
      data_frames_list.append(grouped_df)

    final_grouped_df = pd.concat(data_frames_list)
    final_grouped_df.set_index('model_experiment', inplace=True)
    final_grouped_df
```

**ENCONTRANDO O NÚMERO DE CLUSTERS IDEALL**

```
[ ] # Instantiate the clustering model and visualizer
    model = KMeans()
    visualizer = KElbowVisualizer(model, k=(2,11), timing=False)

    visualizer.fit(7)        # Fit the data to the visualizer
    visualizer.show()        # Finalize and render the figure
```

**VISUALIZANDO DOS GRÁFICOS**

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_2_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_3_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_4_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_5_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_6_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_7_pca02', legend='full')
    plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_8_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_9_pca02')
```

```
[ ] fig = plt.figure(figsize=(10, 10))
    sns.scatterplot(data=final_df, x='pca02_comp01', y='pca02_comp02', hue='kmeans_cluster_10_pca02')
    plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
```