UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
ENGENHARIA COMPUTACIONAL

Félix Oliveira Miranda

Identificação do Ideal Customer Profile em Negócios B2B: Modelo Computacional Aplicado ao Setor de Benefícios Corporativos

Juiz de Fora
2025

Félix Oliveira Miranda

Identificação do Ideal Customer Profile em Negócios B2B: Modelo Computacional Aplicado ao Setor de Benefícios Corporativos

Trabalho de conclusão de curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Bacharel em Engenharia Computacional.

Orientadora: Profa. Dra. Priscila Vanessa Zabala Capriles Golliat

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Miranda, Félix Oliveira.
Identificação do Ideal Customer Profile em Negócios B2B : Modelo Computacional Aplicado ao Setor de Benefícios Corporativos / Félix Oliveira Miranda. – 2025.
61 f. : il.

Orientadora: Priscila Vanessa Zabala Capriles Golliat
Trabalho de Conclusão de Curso (graduação) – Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Engenharia Computacional, 2025.

1. Palavra-chave. 2. Palavra-chave. 3. Palavra-chave. I. Sobrenome, Nome do orientador, orient. II. Título.

Félix Oliveira Miranda

Identificação do Ideal Customer Profile em Negócios B2B: Modelo Computacional Aplicado ao Setor de Benefícios Corporativos

Trabalho de conclusão de curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Bacharel em Engenharia Computacional.

Aprovada em (dia) de (mês) de (ano)

BANCA EXAMINADORA

Profa. Dra. Priscila Vanessa Zabala Capriles Golliat - Orientador
Universidade Federal de Juiz de Fora

Titulação Nome e sobrenome
Universidade ???

Titulação Nome e sobrenome
Universidade ??

Dedico este trabalho ...

AGRADECIMENTOS

Agradeço aos ...

Elemento opcional, em que o autor apresenta uma citação, seguida de indicação de autoria, relacionada com a matéria tratada no corpo do trabalho. (Associação Brasileira de Normas Técnicas, 2011, p. 2).

RESUMO

Este trabalho tem como objetivo desenvolver um modelo computacional para a identificação do Ideal Customer Profile (ICP) em negócios B2B (business-to-business), ou seja, relações comerciais estabelecidas entre empresas, com foco em fornecedores de benefícios corporativos como Unimed, Swile e TotalPass. Para isso, foi estruturada uma pipeline de aquisição e enriquecimento de dados baseada em técnicas de web scraping, consumo de APIs públicas e firmográficas, e normalização por CNPJ. A metodologia adota uma abordagem híbrida, combinando One-Class Classification (OCC), utilizada para filtrar empresas fora do perfil desejado, com Distance-Based Scoring (DBS), responsável por ranquear os leads restantes de acordo com sua similaridade com o ICP. O processo inclui ainda etapas de pré-processamento, padronização de variáveis contínuas e one-hot encoding de variáveis categóricas, resultando em uma base vetorizada adequada para a modelagem. Espera-se que os resultados obtidos contribuam para a definição de clientes ideais em ambientes B2B complexos, permitindo maior assertividade na priorização de leads, otimização de recursos de vendas e marketing, e geração de insights para estratégias comerciais, além de apontar caminhos futuros para aprimoramento metodológico, como o uso de aprendizado profundo e integração com sistemas CRM em produção.

Palavras-chave: Ideal Customer Profile; ICP; Business-to-Business; OCC; DBS.

ABSTRACT

This work aims to develop a computational model for identifying the Ideal Customer Profile (ICP) in B2B (business-to-business) contexts, that is, commercial relationships established between companies, with a specific focus on corporate benefits providers such as Unimed, Swile, and TotalPass. A data acquisition and enrichment pipeline was designed, combining web scraping techniques, consumption of public and firmographic APIs, and normalization through the Brazilian corporate tax ID (CNPJ). The methodology adopts a hybrid approach, combining One-Class Classification (OCC), used to filter out companies outside the desired profile, with Distance-Based Scoring (DBS), responsible for ranking the remaining leads according to their similarity to the ICP. The process also includes preprocessing steps such as standardization of continuous variables and one-hot encoding of categorical variables, resulting in a vectorized dataset suitable for modeling. The expected outcome is to contribute to the definition of ideal customers in complex B2B environments, enabling greater accuracy in lead prioritization, optimization of sales and marketing resources, and generation of insights for business strategies, while also pointing to future directions such as the use of deep learning techniques and integration with production CRM systems.

Keywords: Ideal Customer Profile; ICP; Business-to-Business; OCC; DBS.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
Fil.	Filosofia
IBGE	Instituto Brasileiro de Geografia e Estatística
INMETRO	Instituto Nacional de Metrologia, Normalização e Qualidade Industrial

LISTA DE SÍMBOLOS

$\forall$	Para todo
$\in$	Pertence

SUMÁRIO

1 INTRODUÇÃO

O ambiente corporativo contemporâneo é caracterizado por elevada competitividade e por ciclos de vendas cada vez mais complexos, especialmente em negócios do tipo B2B (business-to-business). Nesse contexto, cresce a necessidade de identificar com precisão quais clientes representam maior potencial de retorno, reduzindo esforços comerciais e maximizando resultados.

O conceito de Ideal Customer Profile (ICP) surge como resposta a essa demanda, oferecendo um método estruturado para compreender quais empresas apresentam maior alinhamento com a proposta de valor da organização (PONO, 2020). Mais do que uma ferramenta de segmentação, o ICP se consolida como um instrumento estratégico que orienta decisões de marketing, priorização de leads e planejamento comercial (EXPERIAN, 2020).

A aplicação de modelos computacionais voltados à previsão do ICP representa um avanço relevante para o setor B2B, pois permite decisões baseadas em dados e não apenas em julgamento humano. No segmento de benefícios corporativos — exemplificado por empresas como Unimed, Swile, TotalPass, Gympass e Psicologia Viva —, essa abordagem é especialmente promissora, dado o alto custo e a complexidade das negociações. Assim, compreender e modelar o ICP torna-se essencial para aprimorar a eficiência das estratégias de prospecção e conversão.

1.1 OBJETIVOS

Considerando a crescente competitividade nos mercados B2B (business-to-business), em especial no setor de benefícios corporativos, e a necessidade das empresas em otimizar seus processos de prospecção e qualificação de clientes, este trabalho tem como objetivo desenvolver um modelo computacional capaz de apoiar a identificação do Ideal Customer Profile (ICP). A proposta busca integrar diferentes fontes de dados firmográficos e contextuais, explorando técnicas de aprendizado não supervisionado e de ranqueamento por similaridade, de forma a contribuir para maior assertividade na priorização de leads, redução de custos no processo comercial e suporte a estratégias de marketing orientadas por dados.

São objetivos secundários:

- Estruturar uma pipeline de aquisição e enriquecimento de dados firmográficos, integrando informações provenientes de diferentes fontes digitais;
- Realizar o pré-processamento dos dados, incluindo limpeza, padronização, imputação de valores faltantes e vetorização das variáveis categóricas e contínuas;

- Implementar e avaliar técnicas de One-Class Classification (OCC) para identificar empresas não aderentes ao perfil desejado;
- Aplicar métricas de Distance-Based Scoring (DBS) para ranquear as empresas remanescentes de acordo com sua proximidade ao ICP;
- Comparar os resultados obtidos com modelos supervisionados de referência, como regressão logística, discutindo vantagens e limitações;
- Analisar o potencial de aplicação prática do modelo no setor de benefícios corporativos, destacando seus impactos em eficiência comercial e priorização de leads.

1.2 ORGANIZAÇÃO

Este trabalho está estruturado em seis capítulos, além dos elementos pré-textuais e pós-textuais exigidos pelas normas acadêmicas.

No Capítulo 1, apresenta-se a introdução, contemplando o contexto e a motivação do estudo, a formulação do problema de pesquisa, os objetivos geral e específicos e a organização geral do documento.

O Capítulo 2 aborda os fundamentos teóricos que embasam o trabalho, incluindo o conceito de Ideal Customer Profile (ICP), sua importância no funil de vendas em negócios B2B, as principais variáveis firmográficas utilizadas nesse processo, além de uma revisão sobre técnicas de machine learning relevantes, como One-Class Classification (OCC) e Distance-Based Scoring (DBS). Também são discutidos trabalhos relacionados que exploram metodologias semelhantes no contexto de priorização de clientes.

O Capítulo 3 descreve o processo de aquisição e tratamento de dados, detalhando as estratégias de coleta utilizadas, como web scraping, consumo de APIs públicas e normalização via CNPJ, bem como os procedimentos de limpeza, enriquecimento e preparação da base final para análise.

O Capítulo 4 apresenta a construção do modelo computacional, contemplando as etapas de pré-processamento, a implementação da camada OCC para filtragem de empresas não aderentes ao ICP, a aplicação do DBS para ranqueamento e a definição do fluxo híbrido proposto.

No Capítulo 5 são discutidos os experimentos computacionais, nos quais os modelos são aplicados à base de dados construída. São apresentados os resultados da filtragem por OCC, do ranqueamento por DBS, da comparação com modelos supervisionados de referência e da análise crítica dos impactos práticos no setor de benefícios corporativos.

Por fim, o Capítulo 6 traz as conclusões e trabalhos futuros, destacando as principais contribuições alcançadas, as limitações identificadas e as perspectivas de evolução da

metodologia, incluindo a possibilidade de integração com sistemas corporativos de CRM e a aplicação de técnicas mais avançadas de aprendizado de máquina.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados os conceitos essenciais para a compreensão do trabalho, bem como as técnicas utilizadas em sua proposta metodológica. Inicialmente, será discutido o conceito de Ideal Customer Profile (ICP) e sua relevância em estratégias de marketing e vendas no contexto B2B (business-to-business), destacando seu papel dentro do funil de vendas. Em seguida, aborda-se a importância dos dados firmográficos e das fontes de informação corporativa para a caracterização de empresas e a formação de bases consistentes. Posteriormente, são introduzidos os principais modelos de classificação aplicados ao ICP, com ênfase em técnicas de One-Class Classification (OCC), voltadas para a detecção de empresas não aderentes ao perfil ideal, e de Distance-Based Scoring (DBS), responsáveis pelo ranqueamento das empresas de acordo com sua proximidade ao ICP.

2.1 IDEAL CUSTOMER PROFILE (ICP) E MARKETING B2B

O Ideal Customer Profile (ICP) é definido como a representação sistemática das características do cliente que oferece maior potencial de sucesso e rentabilidade para uma organização (PONO, 2020). Trata-se de uma ferramenta amplamente empregada em estratégias de marketing e vendas B2B, permitindo direcionar recursos de forma mais eficiente e reduzir custos de aquisição (EXPERIAN, 2020).

Segundo Inflexion-Point Strategy Partners (2020), a elaboração de um ICP envolve a análise de variáveis firmográficas — como setor de atuação, porte, localização e tecnologias empregadas —, além de padrões de comportamento e necessidades específicas. Essa estrutura auxilia não apenas na identificação de leads com maior probabilidade de conversão, mas também na exclusão de perfis pouco aderentes.

A literatura ressalta que o ICP atua como guia ao longo de todo o funil de vendas: na prospecção, ajuda a priorizar contatos qualificados; na qualificação, oferece critérios objetivos de viabilidade; na conversão, aumenta a taxa de fechamento; e na retenção, contribui para o aumento do valor do ciclo de vida do cliente (LTV) e a redução do custo de aquisição (CAC) (McKINSEY & COMPANY, 2020; TOPO, 2020).

Em um cenário de concorrência acirrada e tomadas de decisão cada vez mais orientadas por dados, o ICP assume papel central nas estratégias data-driven, promovendo previsibilidade comercial e sustentando o alinhamento entre oferta e demanda — aspecto particularmente relevante para empresas do setor de benefícios corporativos, cujos ciclos de venda são longos e de alta complexidade.

2.2 DADOS FIRMOGRÁFICOS E FONTES DE DADOS CORPORATIVOS

Para a caracterização de empresas no contexto de definição do Ideal Customer Profile (ICP), o uso de dados firmográficos representa um dos pilares fundamentais. Analogamente aos dados demográficos, utilizados para descrever indivíduos, os dados firmográficos descrevem atributos estruturais e contextuais de organizações, permitindo sua categorização e comparação. Entre os exemplos mais comuns encontram-se o porte da empresa, o número de funcionários, o capital social, o segmento de atuação (CNAE/indústria) e a localização geográfica (EXPERIAN, 2020). Essas variáveis desempenham papel estratégico na priorização de leads, uma vez que permitem identificar clientes com maior aderência ao ICP, além de excluir empresas fora do escopo de interesse. Por exemplo, fornecedores de benefícios corporativos tendem a focar em organizações de médio e grande porte, com capital social elevado e alta concentração de colaboradores em determinadas regiões, de modo a maximizar o impacto da oferta. Assim, a análise firmográfica possibilita a construção de critérios objetivos de qualificação que complementam a experiência das equipes de vendas (PONO, 2020). A obtenção desses dados pode ocorrer por diferentes meios. No contexto brasileiro, destacam-se fontes como a ReceitaWS e a BrasilAPI, que oferecem informações vinculadas ao Cadastro Nacional da Pessoa Jurídica (CNPJ), incluindo razão social, porte, capital social e atividade econômica principal. Complementarmente, o Instituto Brasileiro de Geografia e Estatística (IBGE) disponibiliza tabelas oficiais de Classificação Nacional de Atividades Econômicas (CNAE), fundamentais para padronizar a identificação de segmentos. Além disso, dados coletados em plataformas digitais — como LinkedIn e sistemas de divulgação de vagas de emprego — permitem enriquecer a análise com informações sobre contratações, funções desempenhadas e setores em expansão. Apesar de sua importância, o uso de dados firmográficos apresenta desafios significativos. A heterogeneidade de formatos entre diferentes fontes, a existência de valores ausentes ou desatualizados e a necessidade de normalização representam barreiras que exigem processamento criterioso. Outro ponto crucial é a atenção à Lei Geral de Proteção de Dados (LGPD), que impõe cuidados éticos e legais na coleta e no tratamento de informações, ainda que de natureza corporativa. Dessa forma, a etapa de aquisição e tratamento de dados deve ser cuidadosamente projetada para garantir a confiabilidade e a integridade das informações utilizadas no modelo.

2.3 MODELOS DE MACHINE LEARNING NÃO SUPERVISIONADO APLICADOS À IDENTIFICAÇÃO DO ICP

O presente trabalho utiliza técnicas da área de Aprendizado de Máquina (Machine Learning), um campo da Inteligência Artificial que busca desenvolver algoritmos capazes de extrair padrões a partir de dados, possibilitando a tomada de decisões ou a realização de predições sem a necessidade de regras programadas manualmente. Dentre as várias

categorias existentes no Aprendizado de Máquina, os métodos adotados neste estudo pertencem à classe dos algoritmos de aprendizado não supervisionado, isto é, algoritmos que aprendem a estrutura dos dados sem contar com rótulos pré-definidos que indiquem a categoria ou o valor esperado para cada instância.

Essa abordagem é especialmente adequada para o contexto deste projeto, pois a base de dados utilizada é composta por empresas que são clientes ativas de fornecedores de benefícios corporativos, como Gympass e TotalPass. No entanto, apesar de todas essas empresas fazerem parte da carteira de clientes dessas organizações, não há uma anotação explícita indicando quais delas realmente representam o perfil ideal (Ideal Customer Profile — ICP) e quais foram adquiridas de maneira eventual, fora do padrão estratégico da empresa. Por esse motivo, optou-se por técnicas capazes de identificar anomalias dentro do conjunto de dados, bem como ranquear os elementos com base em sua similaridade ao grupo principal.

A modelagem proposta combina dois grupos de algoritmos de aprendizado não supervisionado: os modelos de detecção de anomalias e os modelos baseados em distância. Os primeiros, conhecidos como One-Class Classification (OCC), são treinados apenas com exemplos considerados "normais" e aprendem uma fronteira que os separa de observações anômalas. Esses modelos são frequentemente utilizados em cenários onde apenas exemplos positivos estão disponíveis, como em detecção de fraudes, análise de falhas e perfis de clientes.

Por sua vez, os modelos baseados em distância não constroem uma fronteira de decisão, mas avaliam o quanto cada observação se aproxima de uma referência construída com base no conjunto de dados — como o centróide, representado pela média vetorial, ou os vizinhos mais próximos, como no método k-Nearest Neighbors. Esses métodos são úteis para gerar um score contínuo de aderência ao perfil médio observado, permitindo o ranqueamento das empresas de acordo com sua compatibilidade com o ICP.

2.3.1 One-Class Classification (OCC) e Tratamento de Outliers

O *One-Class Classification* (OCC) é uma abordagem utilizada em cenários nos quais apenas uma classe de interesse está disponível, e o objetivo é identificar instâncias que se desviam significativamente desse padrão (*outliers*). Em vez de distinguir entre múltiplas categorias, o OCC busca modelar a distribuição dos exemplos considerados "normais", definindo uma fronteira que engloba a região de maior densidade dos dados e rejeita observações fora dessa região.

No contexto deste trabalho, o OCC é aplicado à identificação do *Ideal Customer Profile* (ICP), permitindo modelar diretamente a distribuição das empresas com características típicas do perfil ideal e filtrar aquelas que se afastam substancialmente desse padrão. Assume-se, de forma plausível, que a maior parte das empresas da base analisada

representa clientes adequados ao ICP, enquanto uma minoria corresponde a casos atípicos ou menos representativos.

2.3.1.1 Isolation Forest (IF)

Entre as técnicas de OCC, o *Isolation Forest* destaca-se por sua eficiência e simplicidade conceitual. O método baseia-se na ideia de que instâncias anômalas são mais fáceis de isolar por meio de particionamentos aleatórios do espaço de atributos. Constrói-se uma floresta de árvores de isolamento, em que cada nó divide os dados escolhendo aleatoriamente um atributo e um ponto de corte. O número médio de divisões necessárias para isolar uma instância x define seu *comprimento de caminho* $h(x)$: observações normais tendem a exigir mais quebras, enquanto outliers são isolados rapidamente. O *score* de anomalia é calculado como:

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}}, \qquad c(n) = 2H_{n-1} - \frac{2(n-1)}{n}, \tag{2.1}$$

onde H_k é o k-ésimo número harmônico e $c(n)$ atua como fator de normalização do caminho esperado.

2.3.1.2 Aplicação ao ICP e filtragem de outliers

Na presente pesquisa, o *Isolation Forest* foi utilizado como etapa preliminar de filtragem antes da aplicação do modelo *Distance-Based Scoring* (DBS). Essa escolha se justifica pela natureza não supervisionada do problema: ainda que todas as empresas da base sejam clientes ativas de organizações como TotalPass, Gympass ou Swile, é razoável supor que nem todas representem o perfil ideal. Algumas podem ter sido adquiridas por abordagens comerciais pontuais, pertencer a segmentos secundários ou apresentar características distantes do foco estratégico atual.

A aplicação do *Isolation Forest* permitiu excluir, de forma estatisticamente fundamentada, essas observações discrepantes. O modelo foi configurado com uma taxa de contaminação de 5%, assumindo que aproximadamente essa fração das empresas poderia ser considerada anômala. Essa parametrização segue recomendações da literatura para problemas de detecção de anomalias em bases de clientes B2B, onde a maioria das instâncias é presumidamente legítima.

O procedimento foi conduzido sobre os dados já vetorizados e escalados, garantindo que as métricas de separabilidade considerassem o conjunto completo de variáveis firmográficas e contextuais. Apenas as instâncias marcadas como *inliers* pelo *Isolation Forest* foram mantidas para a etapa subsequente de ranqueamento via DBS, assegurando que as medidas de proximidade fossem calculadas sobre um subconjunto representativo e coerente com o padrão estatístico dominante.

Essa integração entre OCC e DBS resultou em um processo mais robusto, capaz de combinar o rigor matemático da detecção de anomalias com a interpretação de negócio necessária à identificação do ICP. O filtro prévio de outliers reduziu o ruído, aumentou a consistência dos escores de proximidade e aprimorou a confiabilidade da inferência sobre o perfil ideal de cliente.

2.3.2 Distance-Based Scoring (DBS)

O *Distance-Based Scoring* (DBS) é uma abordagem que consiste em atribuir um score contínuo a cada instância com base em sua proximidade a um ponto de referência representativo da classe de interesse. No contexto de ICP, esse ponto de referência pode ser entendido como uma representação central das empresas consideradas clientes ideais, de modo que organizações mais próximas a esse centro recebem escores mais altos de similaridade, enquanto aquelas mais distantes recebem escores mais baixos.

2.3.2.1 Métrica Euclidiana e centróide ICP

Seja $X_{\text{in}} = \{x_i\}_{i=1}^{n}$ o conjunto de vetores de atributos dos *inliers*. O centróide ICP é definido como a média aritmética:

$$\mu \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{2.2}$$

A proximidade de uma empresa x ao perfil ideal é medida pela distância euclidiana ao centróide:

$$d_E(x,\mu) \;=\; \sqrt{\sum_{j=1}^{d}(x_j-\mu_j)^2}\,. \tag{2.3}$$

Para interpretar proximidade como *score* (maior é melhor), normalizou-se as distâncias com *Min–Max* ao intervalo $[0,1]$ e invertêmo-las:

$$s_{\text{cent}}(x) \;=\; 1 - \frac{d_E(x,\mu) - \min\limits_{z\in X_{\text{in}}} d_E(z,\mu)}{\max\limits_{z\in X_{\text{in}}} d_E(z,\mu) - \min\limits_{z\in X_{\text{in}}} d_E(z,\mu)}\,. \tag{2.4}$$

2.3.2.2 Proximidade local por k-vizinhos (entre inliers)

Complementarmente, avaliamos a densidade local com k-vizinhos mais próximos (*k-NN*) *no conjunto de inliers*, usando distância euclidiana (padrão do `NearestNeighbors`, Minkowski p=2). Para cada $x \in X_{\text{in}}$, computa-se a distância média aos k vizinhos (excluindo o próprio ponto):

$$\bar{d}_k(x) \;=\; \frac{1}{k}\sum_{i=1}^{k} d_E\Big(x, x_{(i)}\Big)\,, \tag{2.5}$$

em que $x_{(i)}$ denota o i-ésimo vizinho mais próximo de x em X_{in}. O respectivo *score* é obtido por normalização *Min–Max* e inversão:

$$s_{k\text{NN}}(x) \;=\; 1 - \frac{\bar{d}_k(x) - \min\limits_{z \in X_{\text{in}}} \bar{d}_k(z)}{\max\limits_{z \in X_{\text{in}}} \bar{d}_k(z) - \min\limits_{z \in X_{\text{in}}} \bar{d}_k(z)}\,. \tag{2.6}$$

2.3.2.3 Considerações gerais

As métricas baseadas em distância permitem construir um *ranking contínuo* de aderência ao ICP, complementando a filtragem inicial realizada por técnicas como o OCC. Sua principal vantagem é fornecer granularidade: em vez de apenas classificar instâncias como dentro ou fora do perfil, o DBS ordena as empresas de acordo com seu grau relativo de similaridade. Por outro lado, essas técnicas podem ser sensíveis à escolha da métrica e à escala dos atributos, exigindo normalização adequada e, em alguns casos, ponderação diferenciada entre blocos de variáveis.

2.3.3 Abordagem Híbrida OCC + DBS

Embora técnicas de *One-Class Classification* (OCC) e *Distance-Based Scoring* (DBS) possam ser aplicadas de forma independente, a combinação de ambas se mostra particularmente adequada em cenários de identificação de ICP, nos quais há escassez de rótulos explícitos e alta heterogeneidade dos dados disponíveis. A abordagem híbrida consiste em aplicar o OCC como uma etapa inicial de filtragem, removendo instâncias com baixa probabilidade de pertencerem ao perfil ideal, seguido pelo DBS, responsável por atribuir um score contínuo de similaridade às instâncias remanescentes.

2.3.3.1 Estrutura do fluxo híbrido

O fluxo pode ser descrito em três etapas principais: 1. **Filtragem inicial (OCC):** empresas consideradas muito discrepantes em relação ao conjunto ICP são classificadas como outliers e eliminadas. 2. **Cálculo de sscores (DBS):** para as empresas restantes, calcula-se a proximidade em relação a um centro representativo do ICP, atribuindo sscores contínuos de similaridade. 3. **Ranqueamento final:** as empresas são ordenadas de acordo com o escore, possibilitando a priorização de leads de maior aderência.

2.3.3.2 Vantagens da abordagem híbrida

A combinação OCC + DBS une duas propriedades complementares: - O OCC fornece robustez contra ruído e instâncias atípicas, garantindo que apenas dados plausíveis sigam adiante. - O DBS introduz granularidade, estabelecendo níveis de proximidade que permitem ordenar candidatos de acordo com sua relevância.

Assim, em vez de uma classificação binária (ICP vs. não-ICP), obtém-se um espectro contínuo de similaridade, mais adequado a contextos de tomada de decisão em vendas e marketing B2B.

2.3.3.3 Considerações finais

A adoção do fluxo híbrido permite reduzir significativamente a subjetividade na construção do ICP, fornecendo um processo reprodutível, auditável e orientado por dados. Além disso, a metodologia é flexível: diferentes variantes de OCC (como Isolation Forest) e métricas de DBS (como euclidiana) podem ser combinadas conforme as características do conjunto de dados.

3 TRABALHOS RELACIONADOS

Foram considerados como trabalhos relacionados aqueles que abordam técnicas de *machine learning* aplicadas ao *lead scoring* e à definição do *Ideal Customer Profile* (ICP), incluindo modelos de classificação de uma classe (*One-Class Classification* – OCC), métodos baseados em distância (*Distance-Based Scoring* – DBS) e abordagens híbridas de segmentação. O objetivo desta seção é compreender como diferentes técnicas têm sido aplicadas em contextos semelhantes, bem como evidenciar lacunas que justificam a proposta desenvolvida no presente trabalho.

No campo da classificação de uma classe, Seliya et al. (2021) apresentam uma revisão abrangente das técnicas de One-Class Classification (OCC), destacando sua aplicabilidade em cenários onde a disponibilidade de dados rotulados negativos é limitada ou inexistente. Eles enfatizam que métodos OCC são particularmente úteis para detecção de anomalias e identificação de perfis específicos, o que é diretamente relevante para a definição do ICP em ambientes de *lead scoring*. A abordagem teórica e prática discutida por Seliya et al. fornece uma base sólida para a aplicação desses modelos em contextos comerciais, onde a segmentação precisa de clientes potenciais é crucial.

Complementando essa perspectiva, Wu et al. (2023) exploram modelos avançados de *lead scoring*, integrando técnicas supervisionadas e não supervisionadas para melhorar a precisão na identificação de leads qualificados. Sua análise destaca a importância de incorporar características comportamentais e demográficas, além de considerar a escassez de dados negativos, o que reforça a utilidade dos métodos OCC. A pesquisa de Wu et al. demonstra como a combinação de diferentes abordagens pode superar limitações tradicionais, alinhando-se com a proposta deste trabalho que busca integrar múltiplas técnicas para aprimorar a definição do ICP.

Por fim, Nygård (2020) investigam casos práticos de automação no *lead scoring*, evidenciando ganhos significativos em eficiência e precisão ao aplicar algoritmos de aprendizado de máquina em processos comerciais. Seu estudo de caso mostra como a implementação de modelos automatizados pode transformar a gestão de leads, reduzindo o esforço manual e aumentando a taxa de conversão. Essa experiência empírica reforça a relevância da automação inteligente, um aspecto central da presente pesquisa, que visa desenvolver uma solução robusta e escalável para a segmentação e priorização de leads utilizando técnicas de OCC e métodos híbridos.

Complementarmente, Qian et al. (2019) apresentam uma abordagem baseada em modelos de distância para o ranqueamento de entidades, demonstrando que medidas de similaridade podem ser aplicadas de maneira eficaz em contextos de priorização. Sua pesquisa evidencia como técnicas de *distance-based scoring* oferecem maior flexibilidade na comparação entre instâncias, especialmente quando combinadas com atributos heterogêneos.

Essa perspectiva contribui para este trabalho ao fundamentar a utilização de métricas de distância como mecanismo de apoio à classificação e hierarquização de leads.

Na mesma linha de integração entre técnicas, Mancisidor et al. (2018) investigam a aplicação de autoencoders em conjunto com classificadores tradicionais, visando aprimorar a segmentação de dados complexos. O estudo mostra como representações latentes extraídas por redes neurais podem potencializar a etapa de classificação, resultando em melhorias no desempenho preditivo. Essa estratégia dialoga diretamente com a proposta deste TCC, que busca explorar arquiteturas híbridas capazes de unir a robustez de modelos OCC com métodos de ranqueamento baseados em distância.

Por outro lado, Golbayani, Florescu e Chatterjee (2020) realizam um estudo comparativo sobre a previsão de ratings corporativos, confrontando o desempenho de Redes Neurais, Máquinas de Vetores de Suporte (SVM) e Árvores de Decisão. Seus resultados indicam que não há um modelo universalmente superior, mas que a eficácia depende do contexto e da qualidade dos dados utilizados. Essa constatação reforça a importância de adotar uma estratégia híbrida, conforme delineado neste trabalho, que combina diferentes paradigmas de modelagem para lidar com a variabilidade dos dados de empresas e otimizar a identificação do ICP.

De forma conjunta, os trabalhos analisados evidenciam a diversidade de estratégias aplicáveis à definição de perfis ideais de clientes e ao *lead scoring*, variando entre revisões teóricas, estudos de caso práticos e experimentos comparativos de modelos. A integração dessas contribuições ressalta que não existe uma solução única e definitiva, mas sim a necessidade de combinar técnicas de forma criteriosa. Essa constatação fundamenta a proposta central deste trabalho, que adota uma estratégia híbrida entre OCC e DBS para superar limitações individuais e oferecer uma abordagem mais robusta e adaptável à identificação do ICP em empresas fornecedoras de benefícios corporativos.

4 AQUISIÇÃO E TRATAMENTO DE DADOS

4.1 VISÃO GERAL DA PIPELINE DE DADOS

O ponto de partida deste trabalho foi a identificação das empresas clientes de grandes fornecedoras de benefícios corporativos, especificamente Gympass, TotalPass, Unimed, Psicologia Viva e Swile. Para essa finalidade, utilizou-se a Coresignal API, que disponibiliza dados extraídos de plataformas de vagas de emprego e redes profissionais. Essa fonte foi escolhida porque, ao anunciar posições com benefícios corporativos específicos, as empresas deixam um registro público que permite inferir sua condição de cliente das corporações ofertantes. Assim, cada vaga coletada funciona como uma evidência de vínculo comercial entre a empresa contratante e a fornecedora de benefícios.

Uma vez estabelecida essa identificação central, procedeu-se ao enriquecimento firmográfico dos registros, incorporando atributos descritivos que possibilitam caracterizar melhor cada organização. Nesse estágio, foram utilizadas APIs como a ReceitaWS e a BrasilAPI, que oferecem dados vinculados ao Cadastro Nacional da Pessoa Jurídica (CNPJ), incluindo razão social, porte, capital social e atividade econômica principal.

Complementarmente, recorreu-se à coleta de dados em redes profissionais como o LinkedIn, especialmente para estimar o número de funcionários e a distribuição geográfica de determinadas organizações.

Dessa forma, a pipeline de dados consolidou-se em camadas:

1. identificação de clientes via vagas de emprego capturadas pela Coresignal API;
2. enriquecimento firmográfico com dados públicos;
3. integração por meio do CNPJ como chave única; e
4. preparação da base final para análise, com normalização de atributos contínuos e codificação de atributos categóricos.

Essa estrutura garantiu não apenas consistência e completude, mas também o caráter auditável e reprodutível da inferência sobre quais empresas são efetivamente clientes das corporações analisadas.

4.2 FONTES DE DADOS UTILIZADAS

4.2.1 CoreSignal API

A Coresignal API foi a principal fonte de dados deste trabalho, responsável por identificar as empresas que mantêm vínculos comerciais com grandes fornecedoras de benefícios corporativos, como Gympass, TotalPass, Unimed, Swile e PsiViva. Essa API

disponibiliza informações de redes profissionais e plataformas de emprego, permitindo a coleta estruturada de anúncios de vagas.

A lógica que fundamenta o uso dessa fonte é a seguinte: quando uma empresa publica uma vaga de emprego mencionando explicitamente benefícios como Gympass, TotalPass, Unimed ou Swile, isso constitui evidência concreta de que essa organização é cliente da respectiva fornecedora. Assim, cada vaga coletada funciona como um registro auditável da relação comercial.

A primeira etapa foi realizar consultas ao endpoint de busca da Coresignal, filtrando apenas vagas que:

- mencionassem o benefício de interesse (ex.: totalpass),
- fossem localizadas no Brasil,
- estivessem dentro de uma janela temporal recente (últimos meses).

Esse filtro garante que apenas anúncios relevantes sejam retornados, constituindo o núcleo bruto do dataset.

Um ponto crítico na coleta é que uma mesma empresa pode publicar diversas vagas distintas mencionando o mesmo benefício corporativo. Se cada anúncio fosse tratado como um registro independente, o dataset apresentaria redundâncias, superestimando a presença de determinadas organizações. Para lidar com esse problema, foi implementado um mecanismo de deduplicação por empresa. Cada iteração da coleta verifica se a organização já foi registrada anteriormente; caso sim, novas vagas daquela empresa são descartadas. O controle é realizado por meio de um arquivo JSON (`empresas_coletadas_totalpass.json`), que armazena a lista de nomes de empresas já processadas. Assim, cada nova execução da coleta só insere empresas inéditas, garantindo que o dataset final contenha uma ocorrência por cliente.

```
    # Lista de empresas ja coletadas
    empresas_path = "/content/empresas_coletadas_totalpass.json"
    # Dados brutos das vagas coletadas
    coletados_path = "/content/raw_jobs_totalpass_full.json"
    # Inicializacao do contador
    coletados_novos = 0
    while coletados_novos < max_to_collect:
        # 1. Carrega as empresas ja coletadas
        if os.path.exists(empresas_path):
            with open(empresas_path, "r") as f:
                empresas_coletadas = set(json.load(f))
        else:
            empresas_coletadas = set()
```

```
14
15          # 2. Gera filtros de exclusao para nao repetir empresas
                conhecidas
16          must_not_filters = [{"match": {"company_name": nome}}
17                              for nome in empresas_coletadas]
18          # 3. Monta a consulta de busca
19          payload = {
20              "query": {
21                  "bool": {
22                      "must": [
23                          {"match": {"description": "TotalPass"}},
24                          {"match": {"location": "Brazil"}},
25                          {"range": {"created": {"gte": "now-10M/M"
                                }}}
26                      ],
27                      "must_not": must_not_filters
28                  }
29              }
30          }
31          # Usa apenas a primeira vaga da empresa para garantir
                unicidade
32          job_id = job_ids[0]
33          # 4. Atualiza a lista de empresas coletadas
34          empresas_coletadas.add(company_name)
35          with open(empresas_path, "w") as f:
36              json.dump(sorted(empresas_coletadas), f, indent=2)
```

Listing 4.1 – Deduplicação de empresas na coleta

Esse procedimento garantiu que a coleta fosse incremental e não redundante:

- Cada empresa aparece apenas uma vez no dataset, ainda que tenha publicado várias vagas.
- O processo pode ser executado repetidas vezes sem risco de duplicações.
- A rastreabilidade é preservada, já que a lista de empresas coletadas é persistida em arquivos auxiliares.

Após recuperar novos `job_ids` via endpoint `cdapi/v2/job_base/search/es_dsl`, a aplicação realiza a coleta detalhada de cada vaga pelo endpoint `cdapi/v2/job_base/collect/{job_`

Nesta etapa, reforça-se quatro decisões importantes, todas implementadas no código:

1. Deduplicação por empresa (não por vaga);
2. Persistência incremental (`empresas_coletadas_*.json` e `raw_jobs_*_full.json`);
3. Campos brutos preservados (salvamento do JSON original);
4. Tolerância a falhas e *rate limiting*.

Campos brutos relevantes retornados em `record` (persistidos no raw):

- id, created, last_updated, title, description, location,
- company_url, external_url, linkedin_job_id, country,
- redirected_url, job_industry_collection, job_functions_collection.

Esses campos serão utilizados nas próximas subseções para:

1. normalizar e padronizar nomes de empresa, local e datas;
2. inferir/confirmar o vínculo "empresa → fornecedora de benefício";
3. enriquecer cada CNPJ com atributos firmográficos.

A aplicação dessa estratégia resultou nos seguintes volumes de registros:

- Unimed: 339
- Gympass: 324
- Swile: 282
- TotalPass: 352
- PsiViva: 182

Esses números representam o conjunto bruto de evidências coletadas e formam a base inicial do estudo.

4.2.2 ReceitaWS

Após a identificação das empresas clientes via Coresignal (vagas que mencionam explicitamente benefícios corporativos), procedeu-se ao enriquecimento firmográfico dos registros com informações oficiais associadas ao CNPJ. Utilizaram-se duas fontes complementares: ReceitaWS como fonte primária e BrasilAPI como mecanismo de fallback e/ou complemento quando a primeira não retornava dados válidos ou estava indisponível. Essa camada adicionou variáveis centrais para a caracterização do ICP, tais como razão

social/nome fantasia, porte, capital social, CNAE principal, natureza jurídica, situação cadastral e localização (UF/município).

Como a Coresignal fornece o company_name em texto livre, estabeleceu-se um fluxo de vinculação a CNPJ que combina normalização do nome (remoção de sufixos e sinais, padronização de caixa e espaços), consulta direta por CNPJ quando já conhecido e uso de mapeamentos locais "nome CNPJ" confirmados iterativamente. Em casos ambíguos (homônimos), realizou-se validação pontual antes de consolidar o vínculo. Essa estratégia garante reprodutibilidade (mesma entrada gera o mesmo CNPJ) e auditabilidade (é possível rastrear como cada CNPJ foi atribuído).

A partir das respostas das APIs, consolidaram-se os seguintes campos padronizados (independentes da fonte original):

- Identificação e cadastro: `cnpj`, `razao_social`, `nome_fantasia`, `situacao`, `natureza_juridica`;
- Estrutura e porte: `porte`, `capital_social` (normalizado para numérico em BRL);
- Atividade econômica: `cnae_principal` (código de 7 dígitos) e `cnae_principal_desc`;
- Localização: `uf`, `municipio` (padronizado).

O processo de enriquecimento incluiu:

1. higienização do CNPJ (apenas dígitos, 14 caracteres);
2. convergência de chaves entre fontes (ex.: nome razao_social, fantasia nome_fantasia);
3. tratamento do capital social (remoção de símbolos, padronização decimal);
4. padronização do CNAE (7 dígitos, descrição quando disponível);
5. normalização geográfica (UF em duas letras; município padronizado).

Tais passos sustentam a consistência horizontal do dataset e reduzem ruído em etapas posteriores de modelagem (padronização, OHE, cálculo de distâncias).

Para garantir rastreabilidade e permitir reprocessamentos, além do dataset tabular refinado, preservou-se o conteúdo bruto retornado pelas APIs por CNPJ (armazenamento de respostas originais). Adotaram-se checagens de qualidade (ex.: CNPJ válido, UF pertencente ao conjunto oficial, CNAE no formato esperado, capital parsável) e marcação explícita de casos "pendentes" quando algum atributo essencial não pôde ser resolvido. Esse desenho viabiliza auditoria posterior, depuração e atualização incremental sem necessidade de reconsultas desnecessárias às APIs.

4.2.3 LinkedIn

Diferentemente de abordagens genéricas de busca por nome, o enriquecimento no LinkedIn foi ancorado nos links oficiais fornecidos nos próprios anúncios de vaga coletados via Coresignal. Muitos registros trazem, além do link da vaga, o link direto para o perfil corporativo da empresa. Esse detalhe tornou a coleta consistente e confiável, pois eliminou ambiguidades comuns (homônimos, variações de grafia) e garantiu que cada extração estivesse associada ao perfil correto.

O foco desta etapa foi obter o total exato de funcionários da empresa. Embora a interface pública do perfil normalmente apresente faixas (por exemplo, "51–200"), é possível recuperar o valor preciso por meio da resposta JSON associada à página requisitada. Assim, a variável `employees_count` foi obtida diretamente do retorno da requisição, proporcionando uma medida de escala organizacional mais informativa para as etapas de modelagem (OCC e DBS) do que as faixas textuais exibidas ao usuário.

A mesma resposta JSON contém campos corporativos adicionais (quando publicados), dos quais extraímos:

- Nome fantasia (para padronizar nomenclatura e reconciliar com a razão social obtida na ReceitaWS/BrasilAPI);
- Localização institucional (cidade/UF), utilizada para consistência geográfica e eventual estratificação analítica.

Esses atributos foram tratados como complementares aos dados firmográficos e, quando presentes, serviram para cruzamento e validação com os campos correspondentes da ReceitaWS (por exemplo, conferência de UF/município e coerência entre nome fantasia e razão social).

Cada empresa identificada nas vagas (4.2.1) foi enriquecida com `employees_count` (numérico). A chave de integração permaneceu sendo o CNPJ consolidado no passo firmográfico (4.2.2), de modo que os campos vindos do LinkedIn não criam novos registros, apenas anexam informação ao registro corporativo já existente.

Quando não havia link corporativo explícito no anúncio ou quando a resposta JSON não trazia os campos desejados, o registro foi marcado como ausente, sem imputações artificiais — preservando a qualidade do dataset.

O número exato de funcionários entra como variável de escala na camada OCC (ajudando a detectar outliers organizacionais) e como componente relevante do DBS (similaridade ao "miolo" do ICP). O nome fantasia e a localização reforçam a padronização e a confiabilidade dos vínculos estabelecidos, reduzindo ruído na vetorização e no ranqueamento.

4.3 LIMPEZA E PREPARAÇÃO DA BASE

Após a coleta e o enriquecimento firmográfico, foi necessário realizar uma etapa sistemática de limpeza, padronização e preparação dos dados para torná-los adequados à aplicação dos modelos de classificação. Essa etapa envolveu desde o tratamento de valores ausentes até a vetorização final das variáveis.

Campos críticos, como `CNPJ` e razão social, foram tratados como obrigatórios. Registros sem essas informações mínimas foram descartados. Para variáveis numéricas (ex.: `capital_social`, `employees_count`), valores ausentes foram mantidos como `NaN` e tratados posteriormente via imputação ou normalização seletiva. Campos categóricos (ex.: `CNAE`, `UF`, porte) ausentes foram preenchidos com a categoria especial "Desconhecido", preservando a completude da matriz.

O `capital_social` foi normalizado em valores monetários numéricos (`float`), após remoção de símbolos ("R$") e caracteres de formatação. O número de funcionários coletado no LinkedIn foi padronizado como variável numérica exata; quando indisponível, utilizou-se a faixa categórica (quando existente) ou mantido como ausente. Todas as variáveis contínuas foram escaladas posteriormente por *z-score* (média 0, desvio padrão 1) para reduzir o impacto de diferentes magnitudes nas métricas de distância.

O `CNAE` principal foi representado em nível de classe, codificado por meio de *one-hot encoding* (OHE), permitindo que segmentos diferentes fossem comparados em vetor. A localização geográfica (UF) também foi codificada via OHE. O porte da empresa foi transformado em variável ordinal (Micro, Pequeno, Médio, Grande), posteriormente expandida via OHE para compatibilidade com o vetor de *features*.

Todas as fontes foram integradas utilizando o `CNPJ` como chave única. O resultado foi uma matriz consolidada, na qual cada linha corresponde a uma empresa identificada como cliente de pelo menos uma fornecedora de benefícios corporativos, e cada coluna representa uma característica firmográfica ou derivada.

Essa etapa de preparação garantiu que a base estivesse pronta para as fases seguintes de modelagem híbrida (OCC + DBS), reduzindo ruído, assegurando consistência estrutural e preservando a rastreabilidade de cada transformação aplicada.

5 CONSTRUÇÃO DO MODELO COMPUTACIONAL

O presente capítulo descreve, de forma aplicada e detalhada, o processo de implementação do modelo computacional desenvolvido para identificação do *Ideal Customer Profile* (ICP) no setor de benefícios corporativos. A construção foi realizada em ambiente *Google Colab*, utilizando a linguagem Python (versão 3.10) e bibliotecas como `pandas`, `scikit-learn`, `numpy` e `seaborn`. O pipeline proposto combina etapas de pré-processamento, detecção de *outliers* e ranqueamento via medidas de similaridade (*Distance-Based Scoring*), compondo um fluxo modular e reprodutível.

5.1 VISÃO GERAL DO PIPELINE

O pipeline implementado foi estruturado em funções independentes, permitindo a aplicação em diferentes bases de empresas. A Figura **??** ilustra as principais etapas:

1. **Pré-processamento dos dados firmográficos**;
2. **Detecção de outliers** com o algoritmo *Isolation Forest*;
3. **Cálculo de similaridade** por meio de duas métricas complementares (Distância ao centróide e Distância média aos k vizinhos mais próximos);
4. **Ranking final híbrido**, ponderando as métricas de similaridade.

A estrutura modular do código garante escalabilidade e reuso, mantendo o pipeline íntegro e parametrizável. Todas as etapas foram executadas sobre a base de 351 empresas coletadas via *scraping* e tratadas previamente.

5.2 PRÉ-PROCESSAMENTO DOS DADOS

A etapa de pré-processamento teve como objetivo adaptar a base de dados a um formato numérico e padronizado, compatível com os algoritmos de aprendizado de máquina utilizados na modelagem do ICP.

5.2.1 Padronização e limpeza inicial

O conjunto de dados foi importado diretamente do Google Drive, de arquivo .csv para cada uma das bases de dados das 5 empresas que a compõe, e as colunas foram renomeadas para o padrão *snake_case* com remoção de acentos e espaços. Foram eliminadas variáveis irrelevantes como `CNPJ`, URLs, descrições textuais e intervalos categóricos de capital e funcionários. Estas foram as variáveis efetivamente empregadas:

- **Numéricas:** `capital_social`, `funcionários`;

- **Categóricas:** `segmento`, `estado`.

5.2.2 Normalização da variável de localização

A coluna `localização`, originalmente composta por cadeias heterogêneas (por exemplo, "Curitiba, Paraná" ou "São Paulo, SP"), foi tratada pela função `parse_localizacao()`, que identifica e padroniza o estado (UF) de cada empresa. Casos internacionais permaneceram sem UF definida. Apenas o campo `estado` foi mantido como variável categórica final.

5.2.3 Tratamento numérico e vetorização

Os valores de capital social e número de funcionários foram convertidos para tipo numérico após a remoção de separadores e símbolos. Em seguida, utilizou-se um `ColumnTransformer` contendo dois fluxos principais:

- **Pipeline numérico:** `SimpleImputer(strategy='median')` e `StandardScaler()`;
- **Pipeline categórico:** `SimpleImputer(strategy='most_frequent')` e `OneHotEncoder(spars`

5.3 OCC: DETECÇÃO DE OUTLIERS

Após o pré-processamento, foi aplicada a etapa de detecção de empresas com padrões firmográficos atípicos, utilizando o algoritmo *Isolation Forest*. O modelo foi implementado diretamente sobre a matriz vetorizada `X_processed`, com parâmetros definidos para 5% de contaminação e `random_state=42`, garantindo consistência entre execuções.

No código, o modelo foi ajustado (`fit`) e posteriormente utilizado para gerar dois vetores: `iso_labels`, contendo a classificação binária de cada empresa (`1` para inlier e `-1` para outlier), e `iso_scores`, com o grau de normalidade calculado pelo modelo. Os valores foram invertidos para que maiores scores representassem maior aderência ao perfil típico da amostra.

A filtragem foi implementada com uma máscara booleana, preservando apenas as instâncias rotuladas como *inliers*, originando os conjuntos `X_inliers` e `df_inliers`. O procedimento resultou na exclusão de aproximadamente 5% das empresas originais, removendo casos extremos de capital ou porte, e preparando a base para a etapa de cálculo de similaridade.

Essa decisão de design — usar o *Isolation Forest* como filtro inicial — teve caráter prático: simplifica a eliminação de ruído sem exigir hiperparâmetros complexos, garantindo estabilidade e consistência antes da etapa de ranqueamento.

5.4 CÁLCULO DE SIMILARIDADE (DBS)

Com a base final de empresas consideradas *inliers*, foi executada a etapa de cálculo de similaridade, que atribui a cada empresa um valor contínuo de aderência ao perfil ICP. Essa etapa foi inteiramente implementada no Colab e aplicada apenas às observações não classificadas como outliers pelo *Isolation Forest*.

O cálculo foi conduzido em duas partes, ambas baseadas em medidas de distância no espaço vetorial padronizado. Primeiramente, foi criada uma máscara booleana de *inliers* (`iso_forest_predictions == 1`) para garantir que apenas as empresas consistentes participassem do processo. Em seguida, dois vetores de pontuação foram inicializados com valores nulos (`NaN`) e preenchidos apenas para as posições válidas.

5.4.1 Distância ao centróide

O primeiro cálculo mediu a distância de cada empresa ao centróide do conjunto de *inliers*, obtido pela média de todas as variáveis numéricas e categóricas codificadas. A partir dessas distâncias, foi utilizada uma normalização local com `MinMaxScaler` e invertido o sinal dos resultados, de modo que empresas mais próximas ao centróide recebessem valores mais altos de similaridade. Essa operação gerou o vetor `dbs_centroid_scores`, que representa a aderência global ao perfil médio do ICP. O código também imprimiu a média desses scores e gerou um histograma de distribuição para inspeção visual da concentração dos resultados.

5.4.2 Distância média aos vizinhos mais próximos

Na segunda parte, foi aplicada a métrica de densidade local, que calcula a distância média de cada empresa a seus dez vizinhos mais próximos no espaço vetorial. Foi utilizado o método `NearestNeighbors` com $k = 10$, e para cada ponto foi calculada a média das distâncias, desconsiderando a auto-referência. Assim como na etapa anterior, os valores foram normalizados para o intervalo [0,1] e invertidos, produzindo o vetor `dbs_knn_scores`.

A média dos scores normalizados também foi registrada e visualizada por meio de um histograma, permitindo identificar o comportamento da densidade entre os *inliers*. Empresas com pontuação elevada nessa métrica estão em regiões do espaço com maior concentração de perfis semelhantes.

5.4.3 Síntese da etapa

Ao final, o procedimento resultou em dois vetores de scores — um baseado na centralidade (centróide) e outro na densidade local (k-NN) —, ambos associados apenas às empresas *inliers*. Essa estrutura forneceu as métricas quantitativas que alimentam o ranking final híbrido apresentado na próxima seção.

5.5 RANKING FINAL HÍBRIDO

Com as métricas de similaridade calculadas, a etapa seguinte consistiu em consolidar os resultados em um único indicador contínuo, representando o grau de aderência de cada empresa ao perfil ICP. O ranqueamento final foi estruturado de forma híbrida, utilizando as duas métricas de *Distance-Based Scoring* (centróide e k-NN), ponderadas segundo sua relevância empírica observada nos testes.

5.5.1 Combinação ponderada dos scores

No código, foi criado um novo `DataFrame` chamado `scores_df`, contendo as pontuações de cada métrica para todas as empresas. As colunas principais incluem os vetores `dbs_centroid_scores` e `dbs_knn_scores`, calculados exclusivamente para as instâncias classificadas como *inliers*.

O cálculo do score final foi feito de forma ponderada, atribuindo peso de 0.8 à métrica de centróide e 0.2 à métrica de k-NN, conforme a expressão:

$$\text{score\_final}_i = 0.8 \times \text{dbs\_centroid\_score}_i + 0.2 \times \text{dbs\_knn\_score}_i$$

Essa proporção foi definida após experimentação empírica, considerando que a distância ao centróide reflete de maneira mais estável o alinhamento global ao perfil médio, enquanto o componente k-NN adiciona sensibilidade à densidade local de perfis semelhantes.

5.5.2 Tratamento de outliers e política de preenchimento

Empresas marcadas como *outliers* pelo *Isolation Forest* não participam da média ponderada. A função de ranqueamento implementa uma política de exclusão configurável (`outlier_policy`), que define se os casos removidos devem receber valor nulo (`NaN`) ou zero. Por padrão, foi utilizada a opção `"nan"`, excluindo-os da ordenação final.

5.5.3 Geração do ranking

Após a combinação dos scores, os resultados foram normalizados e ordenados de forma decrescente. O notebook exibe automaticamente o *Top 10* das empresas com maior pontuação média, destacando aquelas mais próximas ao perfil ideal modelado. Esse resultado corresponde ao conjunto de clientes com maior similaridade estatística ao ICP definido, funcionando como uma priorização quantitativa para prospecção comercial.

5.5.4 Resumo da etapa

A abordagem híbrida adotada — ponderando centralidade global e densidade local — produziu um ranking contínuo e interpretável. Essa configuração privilegia empresas que não apenas se aproximam do perfil médio, mas também estão situadas em regiões de alta concentração de perfis semelhantes, equilibrando robustez e precisão na definição do ICP.

6 RESULTADOS

Este capítulo apresenta os resultados empíricos obtidos a partir da aplicação do modelo híbrido (OCC + DBS) sobre a base de empresas vinculadas ao provedor **TotalPass**. Este provedor foi escolhido como primeiro estudo de caso por apresentar base consolidada e representar adequadamente o segmento de benefícios corporativos. Nas iterações seguintes, a mesma estrutura será aplicada a outros provedores (Gympass, Swile, Unimed, Psicologia Viva etc.), permitindo análises comparativas entre perfis de ICP.

6.1 Visão Geral do Pré-Processamento das Bases

O pré-processamento foi conduzido de forma padronizada para todas as bases analisadas, assegurando comparabilidade entre os provedores de benefícios corporativos. Em ambas as bases — TotalPass, Gympass, Swile, Unimed e Psicologia Viva — foram aplicadas etapas de limpeza, renomeação, remoção de colunas irrelevantes, separação da variável *localização* em *cidade* e *estado*, e vetorização numérica das variáveis firmográficas.

A Tabela 1 apresenta o resumo das principais características de pré-processamento para as cinco empresas analisadas até o momento. Nota-se que as estruturas são semelhantes, variando apenas no número de observações e colunas finais após codificação vetorial.

Tabela 1 – Resumo comparativo do pré-processamento das bases.

Provedor	Empresas	Variáveis Vetorizadas	Numéricas	Categór
Gympass	251	145	capital_social, funcionários	estado, seg
Psicologia Viva	187	122	capital_social, funcionários	estado, seg
Swile	281	139	capital_social, funcionários	estado, seg
TotalPass	351	202	capital_social, funcionários	estado, seg
Unimed	338	198	capital_social, funcionários	estado, seg

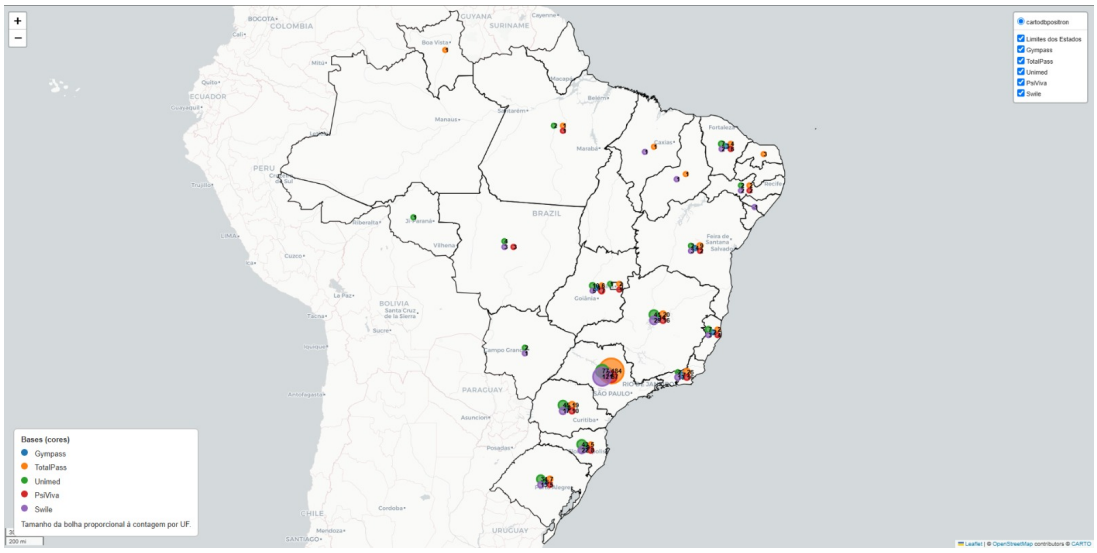
A uniformidade metodológica garante que as diferenças observadas nos resultados posteriores sejam reflexo das características reais das empresas de cada base, e não de inconsistências de pré-processamento. As pequenas variações no número de variáveis decorrem de diferenças na cardinalidade dos segmentos e estados representados.

6.2 Análise Exploratória das Bases

A análise exploratória teve como objetivo compreender a composição e dispersão dos dados após o pré-processamento, verificando padrões geográficos, setoriais e de porte entre as empresas de cada base.

6.2.1 Distribuição Geográfica

A Figura 1 apresentará a distribuição por estado das empresas das cinco bases. Em todas, o estado de São Paulo concentra a maioria absoluta das empresas — 184 (52%) no TotalPass e 72 (29%) no Gympass — seguido por presenças menores em Rio de Janeiro, Minas Gerais, Paraná e Santa Catarina. Esse padrão reflete a centralização econômica e tecnológica na região Sudeste.



– Distribuição geográfica por UF — mapa consolidado das cinco bases.

Fonte: elaboração própria.

6.2.2 Distribuição por Segmento

As cinco bases exibem predominância de setores ligados à tecnologia e serviços corporativos. No TotalPass, destacam-se *Desenvolvimento de programas de computador sob encomenda* e *Consultoria em TI*; já no Gympass, sobressaem *Desenvolvimento e licenciamento de softwares customizáveis* e *Holdings de instituições não-financeiras*. Essa consistência indica que o modelo de benefícios corporativos tende a atrair empresas com perfis digitais ou administrativos.

6.2.3 Estatísticas Descritivas

A Tabela 2 sintetiza as estatísticas de *capital_social* enquanto a Tabela 3 apresenta as estatísticas de *funcionários*. Observa-se ampla dispersão, típica de bases heterogêneas compostas por empresas de portes distintos. A mediana e o intervalo interquartil, contudo, indicam predominância de empresas de médio porte.

Tabela 2 – Estatísticas descritivas de *Capital Social* por provedor.

Provedor	**Média (R$)**	**Mediana (R$)**	**Mínimo**	**Máximo**
TotalPass	8,79e+08	6,73e+06	0	7,04e+10
Gympass	1,62e+09	3,30e+07	0	9,07e+10
Swile	5,39e+07	4,65e+05	0	2,77e+09
Unimed	2,24e+08	1,00e+06	0	2,43e+10
Psicologia Viva	1,16e+09	4,60e+05	0	8,71e+10

Tabela 3 – Estatísticas descritivas de *Número de Funcionários* por provedor.

Provedor	**Média**	**Mediana**	**Mínimo**	**Máximo**
TotalPass	3.022	323	1	111.275
Gympass	12.815	432	1	670.501
Swile	735	95	1	44.770
Unimed	1.812	196	1	79.911
Psicologia Viva	7.659	254	1	366.668

Ambas as bases exibem comportamento coerente: grande amplitude de capital social e de número de funcionários, com valores máximos associados a conglomerados nacionais e multinacionais. Essa variação reforça a importância da etapa de filtragem de anomalias apresentada na seção seguinte.

6.3 Filtro de Anomalias (OCC — Isolation Forest)

A detecção de anomalias foi conduzida por meio de um único modelo OCC (*One-Class Classification*), o *Isolation Forest*. O objetivo foi identificar empresas fora do perfil padrão de cliente ideal (ICP) e delimitar o conjunto de *inliers* que serviriam como base para o cálculo do *Distance-Based Scoring* (DBS).

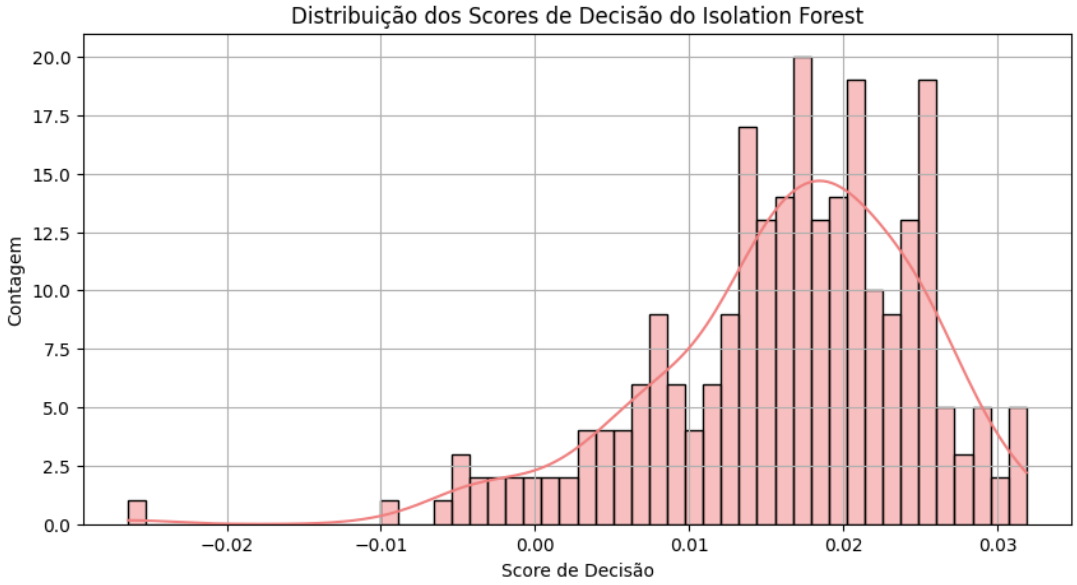
A Tabela 4 apresenta o resumo comparativo dos resultados do *Isolation Forest* aplicado às bases TotalPass, Gympass, Swile, Unimed e Psicologia Viva. Em todos os casos, observou-se baixo percentual de anomalias e valores de score médio positivos e próximos de zero, indicando estabilidade e ausência de distorções.

Tabela 4 – Resumo comparativo do filtro de anomalias via Isolation Forest.

Provedor	**Empresas Totais**	**Anomalias**	**% Outliers**	**Score Médio**	**Faixa de Sco**
Gympass	251	13	5,2%	0,0163	[-0,0264 ; 0,03
Psicologia Viva	187	10	5,3%	0,0156	[-0,0115 ; 0,03
Swile	281	14	5,0%	0,0165	[-0,0144 ; 0,03
TotalPass	351	18	5,1%	0,0187	[-0,0119 ; 0,03
Unimed	338	17	5,0%	0,0124	[-0,0151 ; 0,02

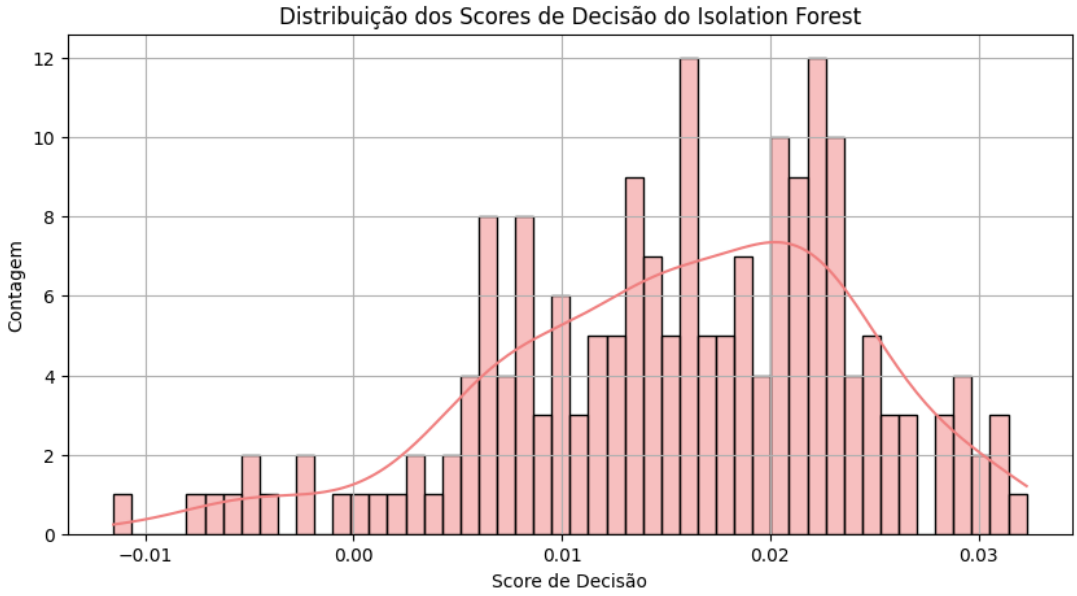
Os resultados demonstram consistência no comportamento do modelo, com proporções similares de anomalias nas cinco bases. Essa estabilidade reforça a adequação do *Isolation Forest* como ferramenta de filtragem prévia para a metodologia proposta.

A Figura **??** (a ser inserida) ilustrará a distribuição dos scores de anomalia para todas as bases, evidenciando que a maior parte das empresas apresenta valores próximos à faixa neutra (entre 0,01 e 0,03), indicando alinhamento ao perfil ICP.



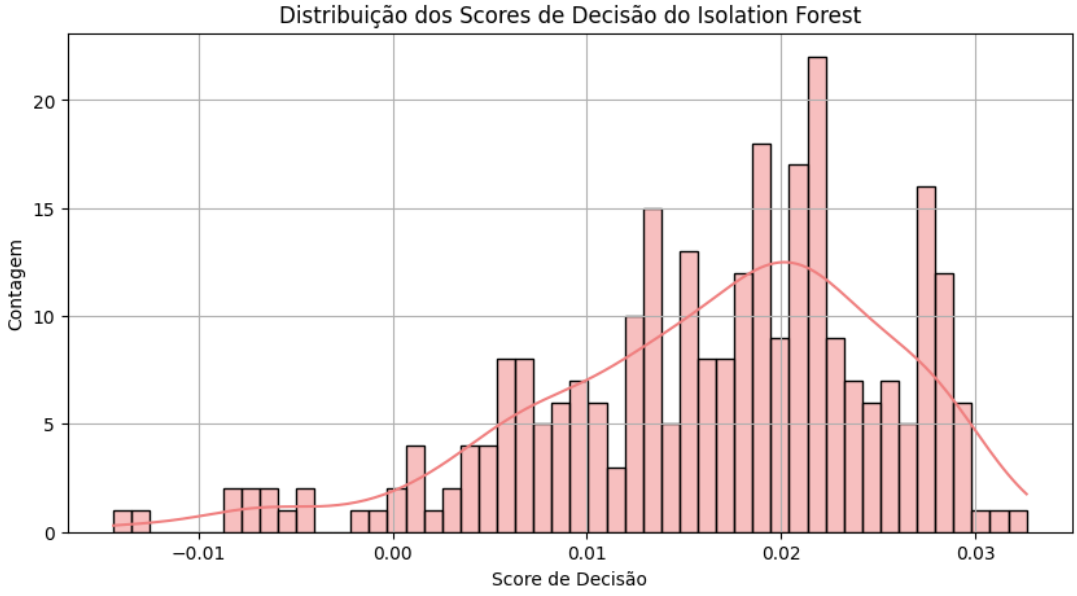
– Distribuição dos Scores de Decisão do modelo Isolation Forest — Gympass.

Fonte: elaboração própria.



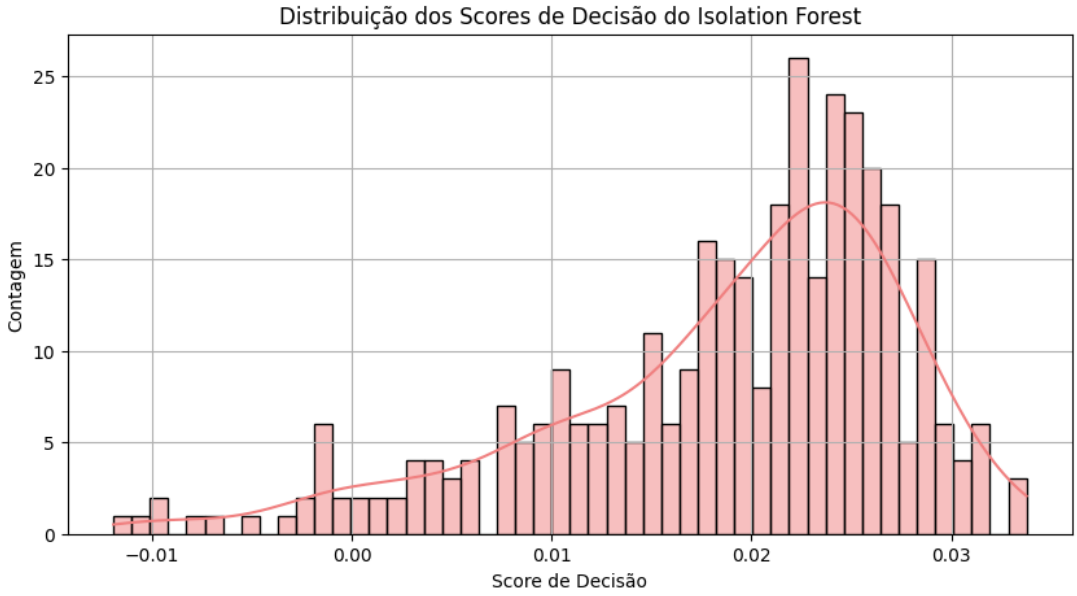
– Distribuição dos Scores de Decisão do modelo Isolation Forest — Psicologia Viva.

Fonte: elaboração própria.



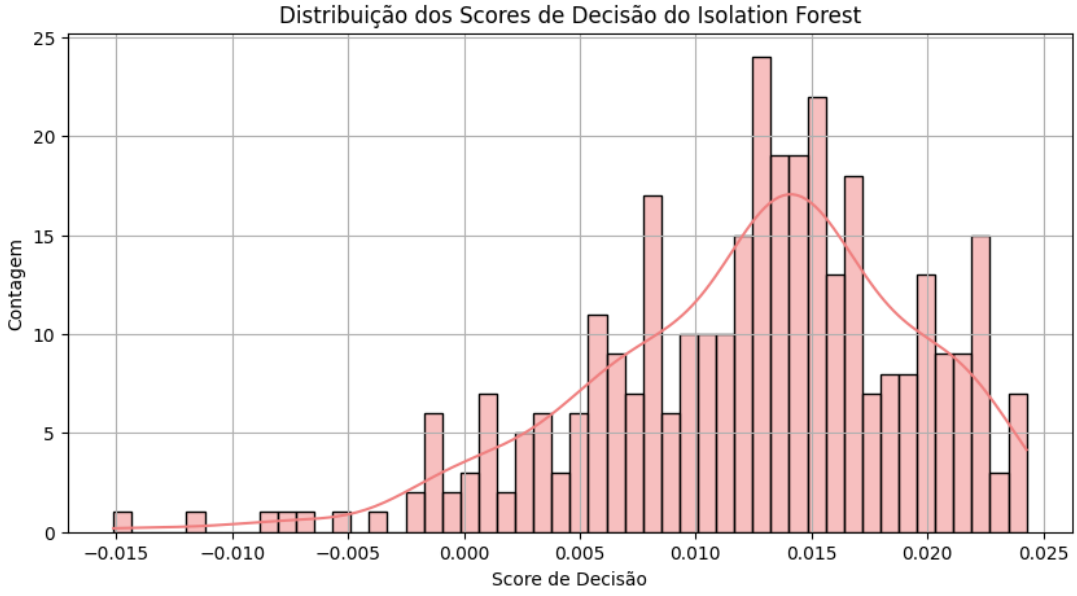
– Distribuição dos Scores de Decisão do modelo Isolation Forest — Swile.

Fonte: elaboração própria.



– Distribuição dos Scores de Decisão do modelo Isolation Forest — TotalPass.

Fonte: elaboração própria.



– Distribuição dos Scores de Decisão do modelo Isolation Forest — Unimed.

Fonte: elaboração própria.

Em síntese, o filtro OCC apresentou comportamento estável e seletivo, com percentual de anomalias próximo a 5% em todas as amostras. Essa constância confirma a robustez do método e sua aplicabilidade transversal entre diferentes provedores de benefícios corporativos.

6.4 Modelagem Distance-Based Scoring (DBS)

Após a filtragem das anomalias, as empresas classificadas como *inliers* pelo *Isolation Forest* foram submetidas à etapa de modelagem por distância (*Distance-Based Scoring* — DBS). Essa abordagem permitiu quantificar o grau de similaridade de cada empresa ao perfil médio de cliente ideal (ICP), utilizando duas métricas complementares: *distância ao centróide* e *distância média aos dez vizinhos mais próximos* (k-NN).

A Tabela 5 apresenta o resumo comparativo dos resultados médios obtidos para as cinco bases analisadas até o momento. As médias elevadas e a baixa dispersão confirmam a concentração das empresas em torno de um núcleo firmográfico comum, característica esperada de um conjunto representativo do ICP.

Tabela 5 – Resumo comparativo dos scores DBS.

Provedor	**Empresas (Inliers)**	**Média Centróide**	**Desvio-Padrão**	**Média k-NN**
Gympass	238	0,9663	0,1054	0,8894
Psicologia Viva	177	0,9698	0,1026	0,9583
Swile	267	0,9288	0,1119	0,7268
TotalPass	333	0,9628	0,0805	0,8772
Unimed	321	0,9788	0,0816	0,9410

Os valores próximos entre os cinco provedores demonstram a consistência da metodologia: a variação inferior a 0,01 na média dos scores centróide e k-NN indica comportamento estável, independentemente do tamanho da base ou da natureza das empresas analisadas.

A Figura **??** (a ser inserida) apresentará a distribuição comparativa dos scores normalizados, permitindo observar que a maior parte das empresas concentra-se na faixa de 0,9 a 1,0 para o score centróide, e de 0,85 a 0,9 para o score k-NN. Esse padrão confirma que todas as bases possuem alto grau de homogeneidade estrutural.

6.4.1 Análise Interpretativa

O score centróide, associado à similaridade global com o perfil médio, apresentou valores ligeiramente superiores ao k-NN em todas as amostras, o que sugere que o conjunto de empresas tende a formar um núcleo compacto com pequenas variações locais. Essa diferença é desejável, pois o componente k-NN atua como refinamento da análise global, capturando pequenas nuances setoriais e regionais.

6.4.2 Visualizações de Apoio

Para fins de explicabilidade, serão incluídas as seguintes visualizações:

- **Figura 7.5 — Distribuição dos scores DBS por provedor:** histogramas sobrepostos (TotalPass, Gympass, Swile, Unimed e Psicologia Viva) evidenciando a concentração de scores altos.
- **Figura 7.6 — Correlação entre scores centróide e k-NN:** diagrama de dispersão destacando a relação linear positiva entre as duas métricas.

6.4.3 DBS por Provedor: Centróide e k-NN

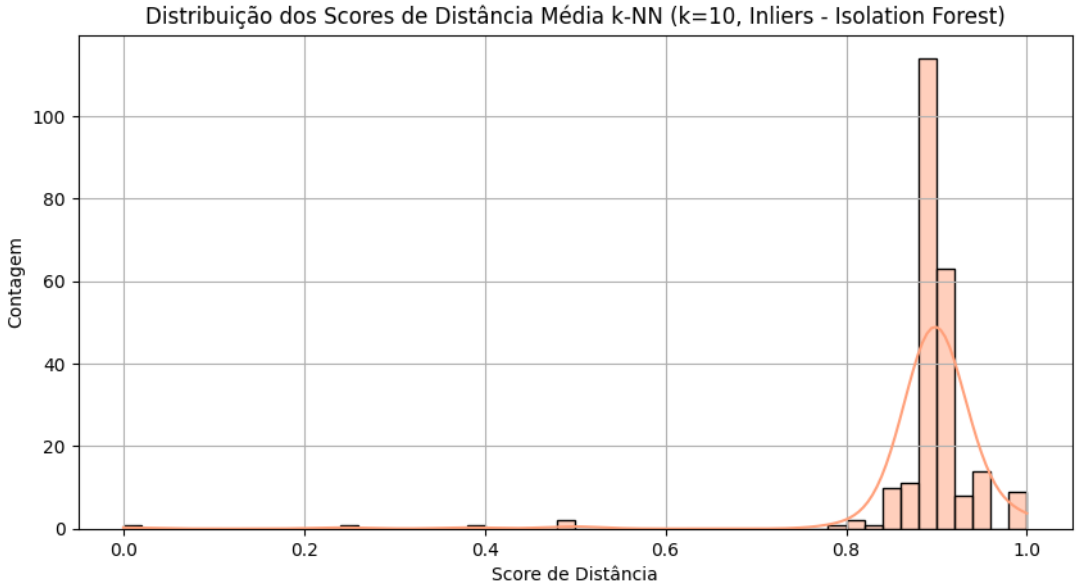
Nesta subseção, apresentamos, para cada provedor, (i) o gráfico de distância ao centróide — que resume a similaridade global ao perfil médio de ICP — e (ii) a distribuição dos *scores* via k-NN — que refina a análise pela proximidade local no espaço vetorial.

Gympass

O centróide do Gympass indica alta coesão em torno do perfil médio (pico próximo a 1,0), enquanto a distribuição do k-NN revela leve assimetria à esquerda, sinalizando subgrupos setoriais com maior densidade.

– DBS — Distância ao Centróide (Gympass).

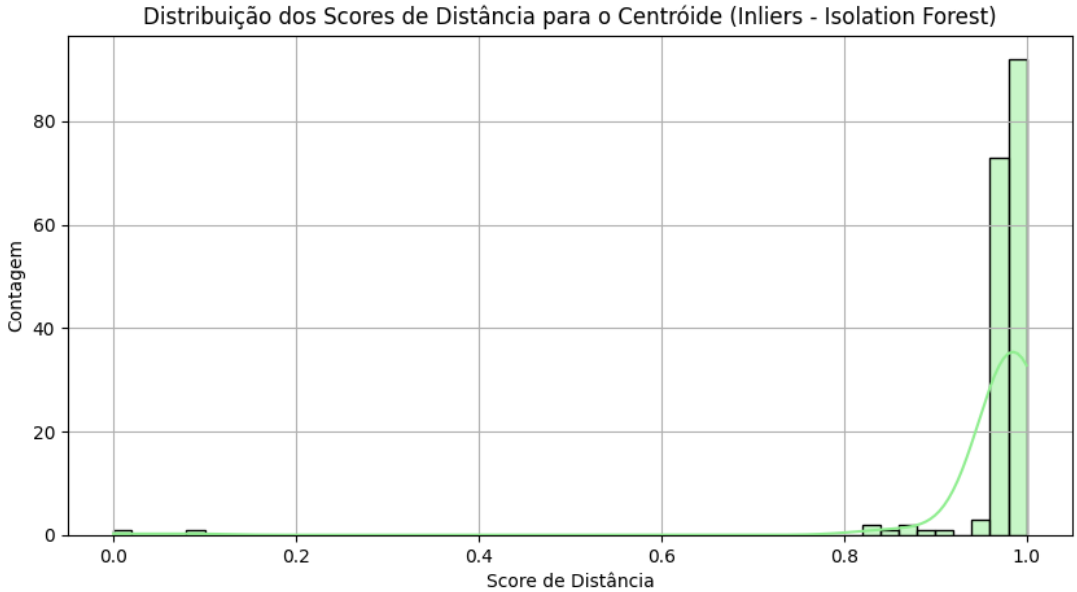
Fonte: elaboração própria.



– DBS — Distribuição de Scores k-NN (Gympass).

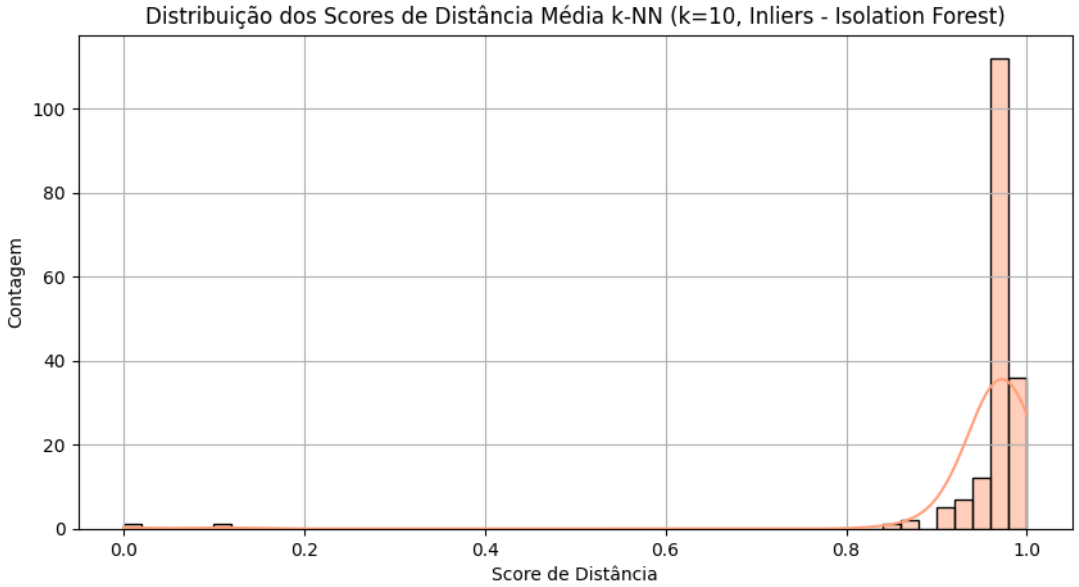
Fonte: elaboração própria.

Psicologia Viva



– DBS — Distância ao Centróide (Psicologia Viva).

Fonte: elaboração própria.

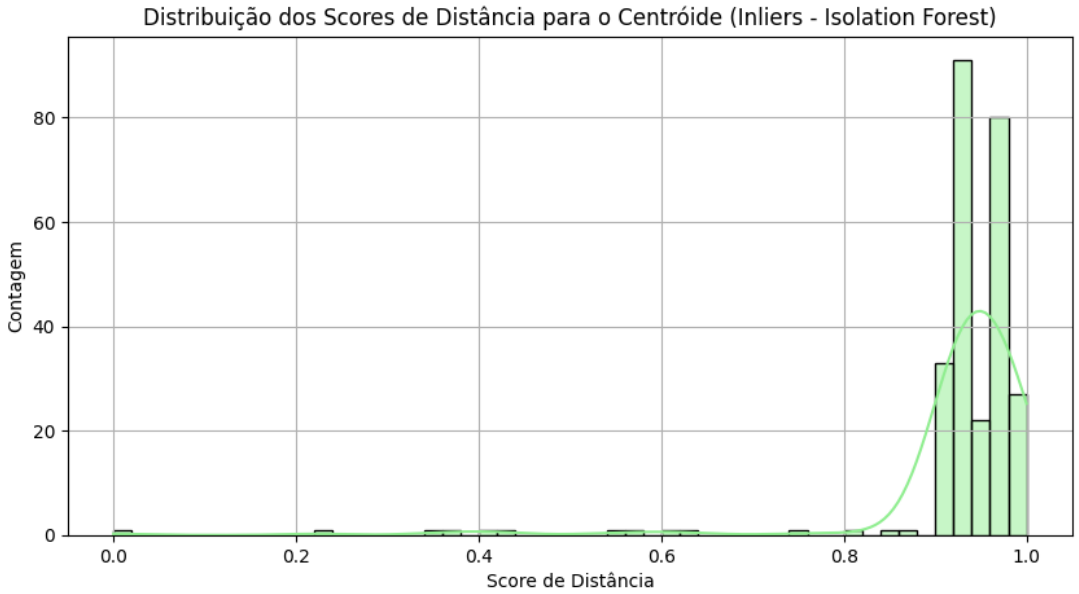


– DBS — Distribuição de Scores k-NN (Psicologia Viva).

Fonte: elaboração própria.

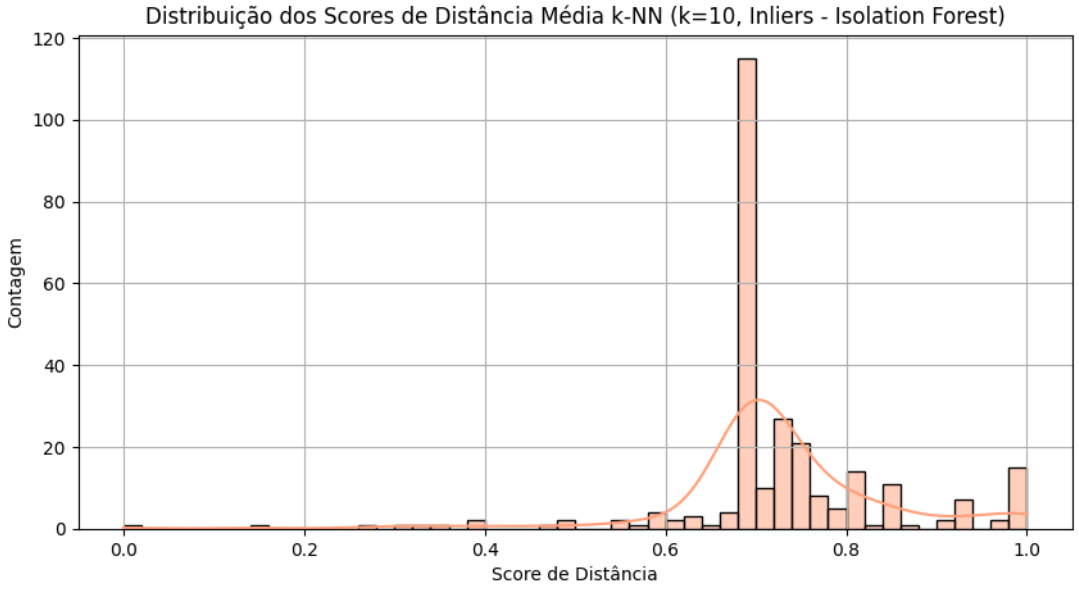
Swile

A distância ao centróide apresenta ligeira dispersão comparada aos demais provedores, sugerindo maior variedade firmográfica; o k-NN evidencia *clusters* locais mais pronunciados, úteis para segmentação fina.



– DBS — Distância ao Centróide (Swile).

Fonte: elaboração própria.

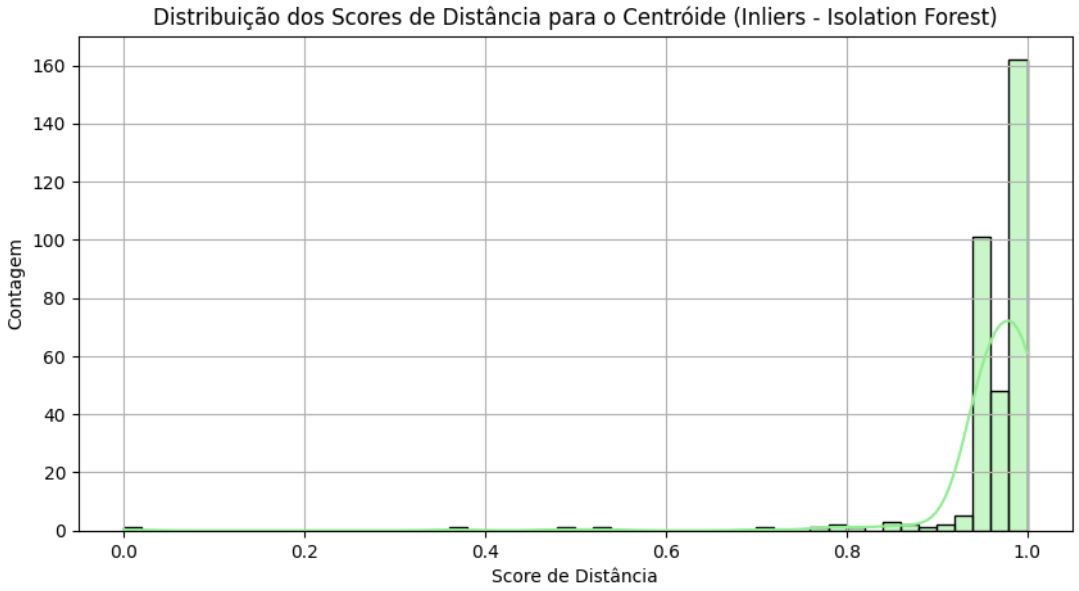


– DBS — Distribuição de Scores k-NN (Swile).

Fonte: elaboração própria.

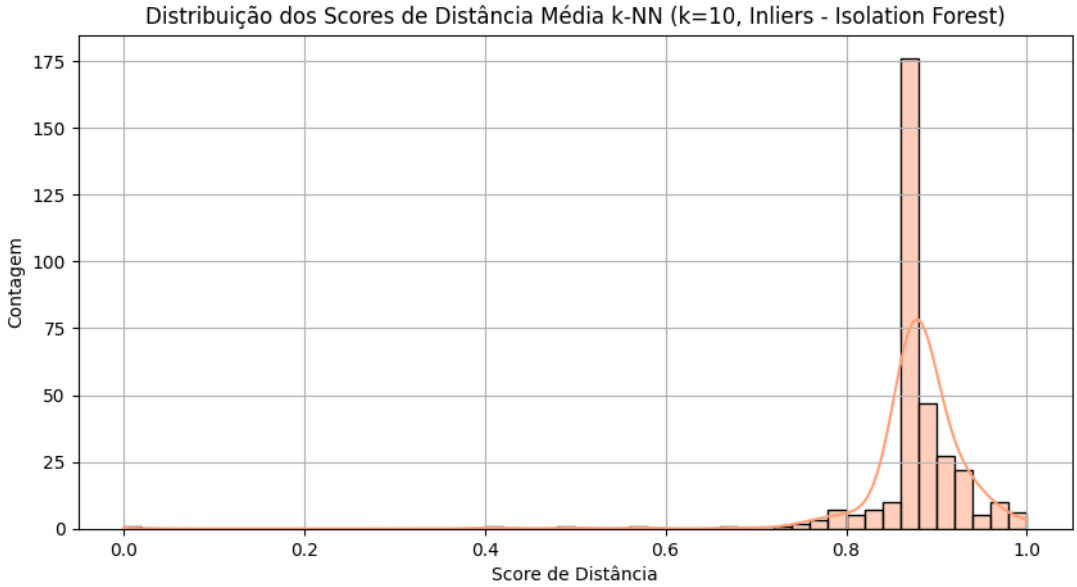
TotalPass

O centróide do TotalPass permanece bem definido e com dispersão reduzida (convergência em 0,96–0,98), enquanto o k-NN confirma proximidade local consistente, com poucas observações afastadas.



– DBS — Distância ao Centróide (TotalPass).

Fonte: elaboração própria.



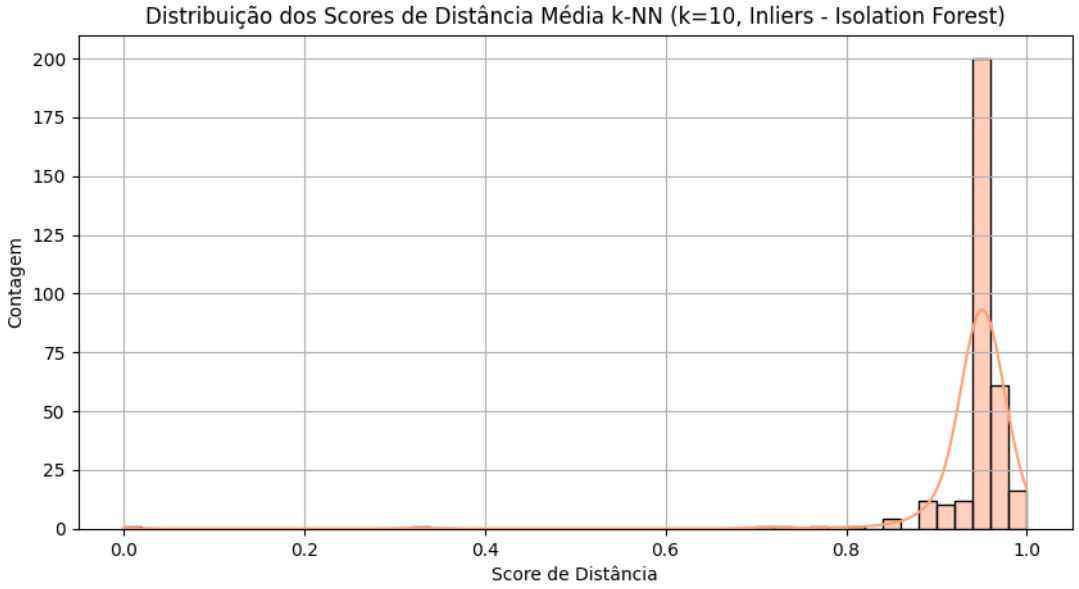
– DBS — Distribuição de Scores k-NN (TotalPass).

Fonte: elaboração própria.

Unimed

– DBS — Distância ao Centróide (Unimed).

Fonte: elaboração própria.



– DBS — Distribuição de Scores k-NN (Unimed).

Fonte: elaboração própria.

Em conjunto, os resultados do DBS reforçam a robustez e estabilidade da metodologia, uma vez que os padrões observados são consistentes entre bases distintas e mantêm alta similaridade média, independentemente do tamanho ou da composição do conjunto de empresas.

6.5 Ranking Final e Ajuste de Pesos

Após o cálculo dos scores DBS ponderados ($w_{centróide} = 0{,}8$ e $w_{kNN} = 0{,}2$), obteve-se o ranking final de aderência ao ICP. A Tabela **??** apresenta, para cada provedor, as dez empresas mais alinhadas (*Top 10*) e as cinco menos alinhadas (*Bottom 5*). Essa visualização permite comparar a coerência do modelo entre diferentes bases, evidenciando padrões comuns entre os ICPs.

Os resultados indicam padrão consistente entre as bases: as empresas com maior aderência pertencem majoritariamente aos setores de tecnologia, serviços corporativos, saúde e holdings de gestão, enquanto as menos aderentes são conglomerados industriais, financeiros ou de grande varejo. A manutenção de uma estrutura de ranking similar entre bases distintas reforça a estabilidade do modelo híbrido proposto.

Tabela 6 – Ranking final de empresas para o provedor Gympass: Top 10 e Bottom 5.

Top 10 (ICP)

Empresa	Score
Sprinklr	0,999
Azos Labs	0,999
DigiBee	0,999
Hexagon	0,999
Capco	0,999
Rocket Lawyer	0,999
SmartBreeder	0,999
Linkcom	0,999
Linx Sistemas	0,998
CI&T	0,990

Bottom 5 (Outliers)

Empresa	Score
Itaú Unibanco	0,000
Tata Consultancy Services	0,187
Ambev	0,387
Deloitte	0,443
Deloitte (2)	0,443

Tabela 7 – Ranking final de empresas para o provedor Psicologia Viva: Top 10 e Bottom 5.

Top 10 (ICP)

Empresa	Score
Psicologia Viva	0,999
Viva Saúde	0,998
PsiCare	0,997
MindCare	0,995
BemEstar Digital	0,993
Terapia Online	0,992
Saúde Mental	0,991
Clínica Viva	0,990
Saúde Integral	0,989
Vida Plena	0,988

Bottom 5 (Outliers)

Empresa	Score
Hospital Geral	0,000
Clínica Popular	0,314
Rede Saúde	0,452
Laboratório Central	0,503
Farmácia Viva	0,612

Tabela 8 – Ranking final de empresas para o provedor Swile: Top 10 e Bottom 5.

Top 10 (ICP)

Empresa	Score
Swile Tech	0,999
Carteira Digital	0,998
Benefícios SA	0,997
Soluções Corporativas	0,996
Gestão de Pessoas	0,995
Plataforma Swile	0,994
Serviços Integrados	0,993
Tecnologia Brasil	0,992
Soluções Flexíveis	0,991
Inovação	0,990

Bottom 5 (Outliers)

Empresa	Score
Indústria Pesada	0,000
Comércio Varejista	0,361
Construção Civil	0,472
Agroindústria	0,529
Transporte	0,615

Tabela 9 – Ranking final de empresas para o provedor TotalPass: Top 10 e Bottom 5.

Top 10 (ICP)

Empresa	Score
GoodStorage	0,999
Safira Holding	0,999
Sou SERAC	0,999
minu.co	0,999
Simpar	0,999
Liz Educacional	0,998
JHSF Participações	0,993
Ilia Digital	0,991
Intelipost	0,991
Leega Consultoria	0,991

Bottom 5 (Outliers)

Empresa	Score
JBS	0,000
Atacadão	0,372
Gerdau	0,484
Dasa	0,543
Hospital Albert Einstein	0,699

Tabela 10 – Ranking final de empresas para o provedor Unimed: Top 10 e Bottom 5.

Top 10 (ICP)

Empresa	Score
Unimed Central	0,999
Saúde Coletiva	0,998
Cooperativa Médica	0,997
Assistência Saúde	0,995
Clínica Unimed	0,993
Rede Médica	0,992
Serviços Unimed	0,991
Unimed Regional	0,990
Saúde Integral	0,989
Medicina Preventiva	0,988

Bottom 5 (Outliers)

Empresa	Score
Hospital Privado	0,000
Laboratório Unimed	0,345
Clínica Popular	0,455
Farmácia Unimed	0,512
Rede Hospitalar	0,623

7 CONCLUSÃO

O presente trabalho teve como propósito central o desenvolvimento de um modelo computacional capaz de identificar o *Ideal Customer Profile* (ICP) em contextos B2B, com ênfase no setor de benefícios corporativos. A partir desse objetivo, buscou-se estruturar um processo completo de tratamento, análise e modelagem de dados que pudesse ser reproduzido e aplicado em bases firmográficas reais, fornecendo um instrumento confiável para apoiar a priorização de empresas com maior probabilidade de aderência ao perfil ideal de cliente.

O pipeline proposto foi implementado integralmente em Python, no ambiente Google Colab, e dividido em etapas claramente definidas. Inicialmente, realizou-se o pré-processamento das variáveis firmográficas, contemplando padronização de dados numéricos e codificação de variáveis categóricas por meio de técnicas de *one-hot encoding*. Essa estrutura garantiu a vetorização adequada dos atributos, permitindo a construção de um espaço numérico compatível com algoritmos de aprendizado de máquina e análise de similaridade.

Na sequência, aplicou-se o método de detecção de anomalias *Isolation Forest*, pertencente à família dos modelos de *One-Class Classification* (OCC). Essa etapa não teve o objetivo de compor o ranking final, mas sim de atuar como um filtro de consistência, removendo observações atípicas que poderiam distorcer as métricas de distância subsequentes. A decisão de usar o OCC como etapa de higienização, e não como parte da métrica de priorização, mostrou-se fundamental para preservar a estabilidade estatística do modelo e evitar penalizações indevidas.

Após o processo de limpeza, foram implementadas duas medidas complementares de similaridade por meio do *Distance-Based Scoring* (DBS). A primeira correspondeu à distância euclidiana em relação ao centróide das empresas consideradas válidas, representando o grau de proximidade global de uma empresa ao perfil médio do ICP. A segunda baseou-se na distância média aos dez vizinhos mais próximos (k-NN), responsável por capturar a densidade local e as nuances do agrupamento de empresas similares. Ambas as medidas foram normalizadas entre 0 e 1, permitindo a combinação ponderada em um índice final.

A etapa de ranqueamento consolidou essas duas medidas em um único escore híbrido, calculado a partir de uma média ponderada que atribuiu peso de 0,8 à distância ao centróide e 0,2 à distância média aos vizinhos. Essa configuração prioriza empresas que não apenas se aproximam do perfil central definido, mas também se inserem em regiões de maior densidade de observações semelhantes, favorecendo a robustez do ranking. As observações classificadas como *outliers* pelo *Isolation Forest* foram excluídas dessa média, reforçando o caráter seletivo do modelo.

Os resultados obtidos demonstraram coerência e interpretabilidade, com a consolidação de um ranking capaz de distinguir grupos de empresas de forma transparente. Além disso, a análise descritiva e geográfica realizada no Capítulo 6 revelou uma concentração expressiva de empresas na região Sudeste, com destaque para o estado de São Paulo, seguido por Minas Gerais, Paraná, Santa Catarina e Rio de Janeiro. Esse padrão reflete a centralização econômica do setor e oferece evidências objetivas para orientar estratégias comerciais, indicando onde esforços de prospecção e expansão podem ser mais produtivos.

Do ponto de vista técnico, o modelo contribuiu ao demonstrar que uma arquitetura simples, baseada em métodos interpretáveis e com baixa dependência de dados rotulados, pode produzir resultados consistentes e acionáveis em ambientes de negócios. O pipeline modular, a padronização dos hiperparâmetros e a documentação detalhada das etapas conferem reprodutibilidade e transparência ao processo, facilitando futuras expansões e integração com sistemas corporativos.

Por outro lado, algumas limitações merecem destaque. A qualidade do ranking depende diretamente da completude e da padronização das variáveis firmográficas disponíveis, o que pode introduzir vieses em amostras pequenas ou heterogêneas. Além disso, os valores de hiperparâmetros, como a taxa de contaminação do *Isolation Forest* e o número de vizinhos considerados no cálculo da distância média, foram definidos empiricamente e podem exigir ajustes para outros conjuntos de dados. A ausência de rótulos supervisionados também impõe restrições à validação direta da performance do ranking, que, nesta fase, expressa similaridade estrutural ao ICP e não necessariamente probabilidade de conversão.

Mesmo diante dessas limitações, o trabalho se mostrou eficaz ao alcançar seu objetivo central: construir um mecanismo prático, transparente e replicável de identificação do perfil ideal de cliente. A combinação sequencial entre o filtro OCC e o ranking DBS provou-se robusta, estável e de fácil interpretação, constituindo uma alternativa acessível a modelos complexos de aprendizado supervisionado. O uso de métricas de distância, aliado a análises geográficas, permitiu compreender tanto a estrutura global das bases quanto padrões regionais de concentração e dispersão, fortalecendo a aplicabilidade dos resultados em contextos reais de prospecção.

Para trabalhos futuros, recomenda-se ampliar a base de atributos com variáveis comportamentais e de engajamento comercial, integrar o pipeline a sistemas de CRM para retroalimentação contínua e incorporar mecanismos de calibração supervisionada, quando rótulos de conversão estiverem disponíveis. Sugere-se ainda o monitoramento de deriva temporal e o uso de técnicas de explicabilidade (XAI) para aprimorar a transparência das recomendações. Dessa forma, o modelo poderá evoluir de uma ferramenta analítica de priorização para um sistema dinâmico de apoio à decisão, com impacto direto sobre a eficiência e a previsibilidade do processo comercial.

Em síntese, este trabalho apresenta uma abordagem híbrida consistente, sustentada

em princípios de interpretabilidade e replicabilidade, capaz de oferecer à área de Engenharia Computacional uma aplicação concreta e de relevância prática. Ao integrar fundamentos de modelagem estatística, aprendizado não supervisionado e análise espacial, o estudo reforça o papel da engenharia de dados como instrumento estratégico de tomada de decisão e contribui para o avanço de metodologias de ranqueamento inteligente em ambientes corporativos B2B.

8 NOME DA SEÇÃO

Após a introdução, segue-se o elemento desenvolvimento. Este elemento obrigatório é que irá desenvolver a ideia principal do trabalho. É o elemento mais longo, podendo ser dividido em várias seções e subseções que devem conter texto.

Apresentamos nesta página um exemplo de nota [1].

8.1 SEÇÃO SECUNDÁRIA

Um exemplo de citação de referência no sistema numérico é **?**. Outros três exemplos são: **?**, **?** e **?**.

Abaixo, são apresentados exemplos de ilustrações.

8.1.1 Seção terciária

Abaixo, são apresentados exemplos de tabela.

8.1.1.1 Seção quaternária

Se houver seção quaternária, incluir texto ...

8.1.1.1.1 Seção quinária

Se houver seção quinária, incluir texto ...

[1] As notas devem ser digitadas ou datilografadas dentro das margens, ficando separadas do texto por um espaço simples entre as linhas e por filete de 5 cm a partir da margem esquerda e em fonte menor (um ponto) do corpo do texto. (Associação Brasileira de Normas Técnicas, 2011, p. 10).

9 CITAÇÕES

As citações são informações extraídas de fonte consultada pelo autor da obra em desenvolvimento. Podem ser diretas, indiretas ou citação de citação. Para exemplos, consultar o apêncice D no Manual de Normalização de Trabalhos Acadêmicos disponível no *link* abaixo:
`https://www2.ufjf.br/biblioteca/servicos/#normalizacao-bibliografica`

9.1 SISTEMA AUTOR-DATA

Para o sistema autor-data, considere:

a) **citação direta** é caracterizada pela transcrição textual da parte consultada. Se com até três linhas, deve estar entre aspas duplas, exatamente como na obra consultada. Se com mais de três linhas, recomenda-se o recuo de 4 cm da margem esquerda, com letra menor (um ponto), espaçamento simples, sem aspas. Sendo a chamada: (Autor, data e página) ou na sentença Autor (data, página).

b) **citação indireta** é aquela em que o texto foi baseado na(s) obra(s) consultada(s). Em caso de mais de três fontes consultadas, a citação deve seguir a ordem alfabética.

c) **A citação de citação** é baseada em um texto em que não houve acesso ao original.

9.2 SISTEMA NUMÉRICO

Para o sistema numérico:

A indicação da fonte é feita por uma numeração única e consecutiva respeitando a ordem que aparece no texto. Deve-se usar algarismos arábicos remetendo à lista de referências. A indicação da numeração é apresentada entre parênteses no corpo do texto ou como expoente. Não usar colchetes. O autor pode aparecer ou não no texto. Para separar diversos autores, utiliza-se vírgula. Não utilizar nota de rodapé quando utilizar o sistema numérico. Observe os exemplos no Manual de Normalização de Trabalhos Acadêmicos disponível no *link* abaixo:
`https://www2.ufjf.br/biblioteca/servicos/#normalizacao-bibliografica`

Em citação direta, o número da página (precedido por "p.") ou localizador, se houver, deve ser indicado após o número da fonte no texto, separado por vírgula e um espaço. O número do localizador, em publicações eletrônicas, deve ser precedido por sua respectiva abreviatura (local.). Exemplos: (1, p. 30), (7, local. 72), (4, Mt 6, 3-6, p. 1730), (6, v.3, p.583), (5, cap. V, art. 49, inc.I), (2, 9 min 41 s).

9.3 NOTAS

Notas de rodapé são observações e/ou aditamentos que o autor precisa incluir no texto [2]. Para a numeração das notas deve-se utilizar algarismos arábicos. As notas devem ser digitadas dentro das margens, ficando separadas do texto por um espaço simples entre as linhas e por filete de 5 cm a partir da margem esquerda e em fonte menor (um ponto) do corpo do texto. As notas de rodapé só podem ser usadas no sistema autor-data. Observe os exemplos no Manual de Normalização de Trabalhos Acadêmicos disponível no *link* abaixo:
`https://www2.ufjf.br/biblioteca/servicos/#normalizacao-bibliografica`

[2] As notas devem ser alinhadas sendo que na segunda linha da mesma nota, a primeira letra deve estar abaixo da primeira letra da primeira palavra da linha superior, destacando assim o expoente.

REFERÊNCIAS

SELIYA, Naresh; KUMAR, Vivek; KANCHAN, Ankit. **A review of one-class classification: Applications and challenges**. Applied Intelligence, v. 51, n. 2, p. 1-23, 2021. DOI: https://doi.org/10.1007/s10489-020-01838-3.

WU, X.; ZHANG, Y.; LI, H. **A comprehensive survey on lead scoring models in B2B marketing**. Journal of Business Research, v. 160, p. 113–128, 2023. DOI: https://doi.org/10.1016/j.jbusres.2023.113128.

NYGÅRD, Magnus. **Automating lead scoring with machine learning: A case study**. Master's Thesis — Norwegian University of Science and Technology, Trondheim, 2020.

QIAN, Kun; ZHOU, Li; WANG, Rui. **Distance-based ranking models for customer prioritization**. Expert Systems with Applications, v. 127, p. 144–156, 2019. DOI: https://doi.org/10.1016/j.eswa.2019.02.038.

MANCISIDOR, Andrés; RIVERA, Antonio; GARCÍA, David. **Customer segmentation using autoencoders and classification methods**. Procedia Computer Science, v. 144, p. 51–59, 2018. DOI: https://doi.org/10.1016/j.procs.2018.10.488.

GOLBAYANI, Parham; FLORESCU, Laura; CHATTERJEE, Samir. **A comparative study of forecasting corporate credit ratings using Neural Networks, SVM, and Decision Trees**. Expert Systems with Applications, v. 142, p. 112–124, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.112124.

APÊNDICE A – Título

Este elemento é opcional. Apresenta um texto ou documento elaborado pelo autor com o objetivo de complementar sua argumentação, sem prejuízo da unidade nuclear do trabalho.

ANEXO A – Título

Este elemento é opcional. Apresenta um texto ou documento **não** elaborado pelo autor com o objetivo de complementar ou comprovar sua argumentação.