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.
45 f.

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 14
1.1 Objetivos 15
1.2 Organização 15
2 FUNDAMENTAÇÃO TEÓRICA 17
2.1 IDEAL CUSTOMER PROFILE (ICP) E MARKETING B2B 17
2.2 DADOS FIRMOGRÁFICOS E FONTES DE DADOS CORPORATIVOS 18
2.3 MODELOS DE MACHINE LEARNING NÃO SUPERVISIONADO APLICADOS À IDENTIFICAÇÃO DO ICP 19
2.3.1 One-Class Classification (OCC) 20
2.3.1.1 One-Class Support Vector Machine (OC-SVM) 20
2.3.1.2 Isolation Forest (IF) 20
2.3.1.3 Outras variantes 21
2.3.1.4 Considerações gerais 21
2.3.2 Distance-Based Scoring (DBS) 21
2.3.2.1 Métrica Euclidiana 21
2.3.2.2 Similaridade do Cosseno 22
2.3.2.3 Método dos k-vizinhos mais próximos (k-NN) 22
2.3.2.4 Considerações gerais 22
2.3.3 Abordagem Híbrida OCC + DBS 22
2.3.3.1 Estrutura do fluxo híbrido 23
2.3.3.2 Vantagens da abordagem híbrida 23
2.3.3.3 Considerações finais 23
2.4 TRATAMENTO DE OUTLIERS 23
3 TRABALHOS RELACIONADOS 25
4 AQUISIÇÃO E TRATAMENTO DE DADOS 27
4.1 VISÃO GERAL DA PIPELINE DE DADOS 27
4.2 FONTES DE DADOS UTILIZADAS 28
4.2.1 CoreSignal API 28
4.2.2 ReceitaWS 31
4.2.3 LinkedIn 32
4.3 LIMPEZA E PREPARAÇÃO DA BASE 33
5 CONSTRUÇÃO DO MODELO COMPUTACIONAL 35
5.1 PRÉ-PROCESSAMENTO DOS DADOS 35
5.2 DETECÇÃO DE OUTLIERS COM MODELOS ONE-CLASS CLASSIFICATION 36
5.3 CÁLCULO DE SIMILARIDADE VIA ESTRATÉGIAS DISTANCE-BASED (DBS) 37

5.3.1 Distância ao Centróide . . . 37
5.3.2 Distância Média aos k-Vizinhos Mais Próximos (k-NN) . . . 37
5.3.3 Considerações . . . 37
5.4 RANKING FINAL . . . 38
5.4.1 Cálculo da Média Final . . . 38
6 NOME DA SEÇÃO . . . 39
6.1 SEÇÃO SECUNDÁRIA . . . 39
6.1.1 Seção terciária . . . 39
6.1.1.1 Seção quaternária . . . 39
6.1.1.1.1 Seção quinária . . . 39
7 CITAÇÕES . . . 40
7.1 SISTEMA AUTOR-DATA . . . 40
7.2 SISTEMA NUMÉRICO . . . 40
7.3 NOTAS . . . 41
8 CONCLUSÃO . . . 42
REFERÊNCIAS . . . 43
APÊNDICE A – Título . . . 44
ANEXO A – Título . . . 45

1 INTRODUÇÃO

O ambiente corporativo contemporâneo é marcado por elevada competitividade e por ciclos de vendas cada vez mais complexos, sobretudo em negócios do tipo B2B (business-to-business), nos quais a relação comercial se estabelece entre empresas. Nesse cenário, o conceito de Ideal Customer Profile (ICP) tem se tornado uma ferramenta fundamental para empresas que buscam otimizar seus processos de vendas e marketing (PONO, 2020). O ICP oferece uma definição clara das características ideais dos clientes-alvo, permitindo que as empresas direcionem suas estratégias de forma mais eficiente (EXPERIAN, 2020). Ao identificar padrões de comportamento, necessidades específicas e características firmográficas, como setor de atuação, tamanho da empresa e tecnologias utilizadas, as organizações podem alinhar seus produtos ou serviços com os clientes que realmente se beneficiariam de suas ofertas (INFLEXION-POINT STRATEGY PARTNERS, 2020).

A relevância do ICP se evidencia ao longo de todo o funil de vendas: na etapa de prospecção, permite priorizar leads com maior aderência ao perfil desejado, reduzindo o esforço desperdiçado em contatos pouco promissores; na etapa de qualificação, orienta a análise de viabilidade comercial e acelera o processo de tomada de decisão; já na etapa de conversão, aumenta a taxa de sucesso ao alinhar a proposta de valor da empresa com as necessidades reais do cliente. Além disso, um ICP bem definido impacta desdobramentos estratégicos, como o planejamento de marketing direcionado, a redução do custo de aquisição de clientes (CAC) e o aumento do valor do ciclo de vida do cliente (LTV).

A relevância dessa abordagem é ainda mais evidente no contexto atual, onde a concorrência está cada vez mais acirrada e o ciclo de vendas se torna mais complexo (McKINSEY & COMPANY, 2020). Empresas que conseguem identificar e priorizar seus potenciais clientes com maior precisão obtêm uma vantagem significativa, maximizando a taxa de conversão e minimizando os recursos desperdiçados em leads que não geram valor real (TOPO, 2020). Dessa forma, a criação de um modelo computacional que seja capaz de prever o ICP com base em dados firmográficos e contextuais pode representar uma inovação poderosa para as áreas de vendas e marketing de diversas organizações.

No setor de benefícios corporativos, empresas como Unimed, Swile e TotalPass exemplificam um mercado em expansão e fortemente dependente da capacidade de selecionar clientes estratégicos. Nesse contexto, a correta definição do ICP pode representar a diferença entre ciclos comerciais longos e custosos e um processo de vendas mais eficiente, baseado em decisões orientadas por dados.

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 conceito de Ideal Customer Profile (ICP) tem se consolidado como uma ferramenta fundamental para empresas que buscam aumentar a eficiência de suas estratégias de vendas e marketing, sobretudo em ambientes de negócios B2B (business-to-business) (PONO, 2020). O ICP pode ser entendido como a representação estruturada das características que definem o cliente ideal, ou seja, aquele que apresenta maior probabilidade de gerar valor para a organização no longo prazo. Essa definição permite que empresas direcionem seus esforços comerciais de forma mais precisa, minimizando desperdícios de recursos e ampliando a assertividade na geração de oportunidades (EXPERIAN, 2020). A aplicação prática do ICP se manifesta ao longo de todas as etapas do funil de vendas. Na fase de prospecção, auxilia na priorização de leads com maior aderência, reduzindo o esforço dedicado a contatos pouco promissores. Durante a qualificação, fornece critérios objetivos para avaliação da viabilidade comercial e acelera a tomada de decisão por parte da equipe de vendas. Na etapa de conversão, contribui para o aumento da taxa de fechamento ao alinhar a proposta de valor com as necessidades específicas do cliente. Finalmente, no estágio de retenção, favorece a manutenção de clientes estratégicos, ampliando o valor do ciclo de vida (LTV) e reduzindo o custo de aquisição de clientes (CAC). De acordo com Inflexion-Point Strategy Partners (2020), a definição de ICP envolve a análise de padrões de comportamento, necessidades do mercado e variáveis firmográficas, como setor de atuação, porte da empresa, localização geográfica e tecnologias utilizadas. Esses fatores permitem não apenas identificar os clientes mais propensos a gerar retorno, mas também excluir aqueles que, embora possam parecer atrativos em um primeiro momento, demandariam alto custo de manutenção ou não se beneficiariam plenamente da solução oferecida. McKinsey & Company (2020) reforça que, em um cenário de competição acirrada e ciclos de vendas cada vez mais longos, a adoção de ICPs bem definidos oferece uma vantagem competitiva

significativa, elevando a eficiência comercial e maximizando o retorno sobre investimento. Nesse sentido, TOPO (2020) argumenta que organizações que priorizam a identificação criteriosa de seus clientes-alvo conseguem não apenas aumentar suas taxas de conversão, mas também alinhar melhor suas estratégias de marketing ao perfil do mercado em que atuam. Assim, o ICP se configura como um elemento central em estratégias data-driven, sustentando decisões mais racionais e reduzindo a subjetividade no processo de vendas. No setor de benefícios corporativos, esse papel se torna ainda mais crítico, uma vez que a complexidade das negociações exige elevado alinhamento entre a proposta de valor da empresa fornecedora e as características de seus potenciais clientes.

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, TotalPass e Swile. 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. Entre os métodos utilizados nesta categoria estão o One-Class SVM (Support Vector Machine) e o Isolation Forest, que será detalhado posteriormente.

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 escore 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)

O *One-Class Classification* (OCC) é uma abordagem utilizada em problemas nos quais existe apenas uma classe de interesse, denominada "normal", e o objetivo é identificar instâncias que se desviam significativamente desse padrão, classificando-as como anomalias ou outliers. No contexto do ICP, o OCC é relevante por permitir modelar diretamente a distribuição das empresas com características típicas do perfil ideal, rejeitando observações distantes dessa distribuição. De forma intuitiva, o OCC busca construir uma **fronteira de decisão** que envolva a região de maior densidade dos dados, marcando como anômalos os pontos que ficam fora dela.

2.3.1.1 One-Class Support Vector Machine (OC-SVM)

O OC-SVM **?** é uma das formulações mais utilizadas de OCC. A ideia central é separar a origem dos dados no espaço de características com máxima margem. Formalmente, resolve-se o seguinte problema de otimização:

$$\min_{w,\rho,\xi} \ \frac{1}{2}\|w\|^2 \ + \ \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \ - \ \rho \tag{2.1}$$

sujeito a:

$$(w^\top \phi(x_i)) \ \geq \ \rho \ - \ \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n, \tag{2.2}$$

onde $\phi(\cdot)$ é o mapeamento dos dados para o espaço de características induzido por um kernel. O parâmetro $\nu \in (0, 1]$ controla a fração máxima de outliers admitidos e a fração mínima de vetores de suporte. A função de decisão é:

$$f(x) = \text{sign}\Big(\sum_{i=1}^{n}\alpha_i K(x_i, x) - \rho\Big), \tag{2.3}$$

em que α_i são os multiplicadores de Lagrange e $K(\cdot, \cdot)$ é a função kernel, como o RBF ou polinomial.

2.3.1.2 Isolation Forest (IF)

O Isolation Forest **?** baseia-se na ideia de que outliers são mais fáceis de isolar por particionamentos aleatórios. Constrói-se uma floresta de árvores de isolamento, nas quais cada nó divide os dados selecionando aleatoriamente um atributo e um ponto de corte. O número esperado de quebras necessárias para isolar uma instância x define o seu *comprimento de caminho* $h(x)$: instâncias normais tendem a exigir mais quebras, enquanto outliers são isolados rapidamente. O score de anomalia é dado por:

$$s(x, n) = 2^{-\frac{\mathbb{E}[h(x)]}{c(n)}}, \qquad c(n) = 2H_{n-1} - \frac{2(n-1)}{n}, \tag{2.4}$$

onde H_k é o k-ésimo número harmônico e $c(n)$ normaliza o caminho esperado.

2.3.1.3 Outras variantes

Além do OC-SVM e do Isolation Forest, outras técnicas incluem o *Elliptic Envelope*, que assume distribuições aproximadamente gaussianas e utiliza estimadores robustos de covariância, e o *Local Outlier Factor (LOF)*, que avalia a densidade local em relação à vizinhança **?**.

2.3.1.4 Considerações gerais

O OCC é particularmente útil em contextos nos quais não há rótulos confiáveis para todas as instâncias, mas presume-se que a maior parte dos dados pertença a uma classe "normal". No caso de ICP, isso significa assumir que a base de dados contém, em sua maioria, empresas plausivelmente dentro do perfil ideal, de modo que as técnicas OCC podem aprender suas características comuns e rejeitar as instâncias mais discrepantes.

2.3.2 Distance-Based Scoring (DBS)

O *Distance-Based Scoring* (DBS) é uma abordagem que consiste em atribuir um escore 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

A distância euclidiana é a forma mais comum de mensurar proximidade em espaços vetoriais. Dado um vetor de atributos $x \in \mathbb{R}^d$ e um centro de referência $\mu \in \mathbb{R}^d$, a distância euclidiana é definida como:

$$d_E(x, \mu) = \sqrt{\sum_{j=1}^{d} (x_j - \mu_j)^2}. \tag{2.5}$$

Escores de proximidade podem ser calculados de forma inversa à distância, permitindo interpretar empresas mais próximas ao centro como mais aderentes ao ICP.

2.3.2.2 Similaridade do Cosseno

Outra medida amplamente utilizada é a similaridade do cosseno, especialmente adequada para dados de alta dimensionalidade e representações esparsas. Para dois vetores x e μ, define-se:

$$\text{sim}_{\cos}(x, \mu) = \frac{x \cdot \mu}{\|x\| \, \|\mu\|}. \tag{2.6}$$

Essa métrica avalia o ângulo entre os vetores, retornando valores próximos de 1 quando os vetores estão fortemente alinhados, mesmo que suas magnitudes sejam diferentes. No caso de ICP, empresas com perfis de atributos similares em direção, ainda que em escalas distintas, podem ser consideradas próximas.

2.3.2.3 Método dos k-vizinhos mais próximos (k-NN)

O ranqueamento também pode ser construído a partir do cálculo das distâncias de cada empresa para seus k vizinhos mais próximos dentro do conjunto ICP. Define-se o escore médio como:

$$s_{kNN}(x) = \frac{1}{k} \sum_{i=1}^{k} d(x, x_i), \tag{2.7}$$

em que x_i são os vizinhos mais próximos de x. Quanto menor o escore, maior a proximidade do ponto ao conjunto ICP.

2.3.2.4 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 escore 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 escores (DBS):** para as empresas restantes, calcula-se a proximidade em relação a um centro representativo do ICP, atribuindo escores 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 OC-SVM ou Isolation Forest) e métricas de DBS (como euclidiana ou cosseno) podem ser combinadas conforme as características do conjunto de dados.

2.4 TRATAMENTO DE OUTLIERS

A presença de observações discrepantes, conhecidas como *outliers*, é um desafio recorrente em projetos de análise de dados e modelagem preditiva. Em contextos de aprendizado não supervisionado, onde não há rótulos disponíveis para indicar quais instâncias são desejáveis ou não, os *outliers* representam um risco ainda maior, pois podem distorcer significativamente o espaço de representação dos dados. Isso é particularmente relevante quando se busca identificar perfis ideais, como no caso deste trabalho, em que se pretende caracterizar o *Ideal Customer Profile* (ICP) a partir de dados de empresas que já são clientes de fornecedoras de benefícios corporativos.

Apesar de todas as empresas da base analisada serem clientes ativas de organizações como Gympass, TotalPass ou Swile, é razoável assumir que nem todas representam o ICP genuíno. Algumas podem ter sido adquiridas por estratégias pontuais, por abordagens comerciais não direcionadas, ou ainda podem pertencer a segmentos fora do foco estratégico atual. A presença dessas observações pode comprometer a definição do que é o perfil ideal de cliente, especialmente em algoritmos sensíveis à densidade ou à distribuição das variáveis.

Diante desse cenário, foi adotada uma etapa explícita de filtragem de *outliers* antes da aplicação dos modelos de ranqueamento. A técnica escolhida para essa tarefa foi o *Isolation Forest*, um algoritmo de detecção de anomalias baseado no princípio da separabilidade de instâncias. Ao contrário de métodos que calculam distâncias ou densidades, o *Isolation Forest* funciona construindo árvores binárias aleatórias que particionam o espaço dos dados. A intuição por trás do algoritmo é que observações anômalas são mais fáceis de isolar — ou seja, requerem um menor número de divisões para serem separadas do restante da base — do que observações normais, que tendem a estar embutidas em regiões mais densas e complexas.

O algoritmo foi configurado com uma taxa de contaminação de 5%, isto é, assumiu-se que aproximadamente 5% das empresas presentes na base poderiam ser consideradas discrepantes em relação ao padrão médio observado. Essa escolha foi embasada tanto em critérios empíricos quanto na literatura, que sugere faixas similares em aplicações de perfis de clientes. O uso do *Isolation Forest* como etapa preliminar permitiu ao modelo excluir, com base estatística, aquelas empresas cujas características destoavam significativamente do conjunto analisado, reduzindo o ruído e aprimorando a qualidade da inferência posterior.

Essa filtragem foi aplicada diretamente sobre os dados já vetorizados e escalados, garantindo que os critérios de anormalidade considerassem o conjunto completo de variáveis utilizadas no modelo. Somente após essa etapa é que os métodos de *Distance-Based Scoring* (DBS) foram aplicados, assegurando que o ranqueamento fosse calculado apenas sobre empresas cuja estrutura firmográfica estivesse alinhada com o padrão estatístico geral da base de clientes. Esse procedimento combinou, portanto, rigor matemático com coerência de negócio, e contribuiu para a robustez e a confiabilidade da abordagem adotada.

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, como Gympass, TotalPass, Unimed 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. Fontes auxiliares, como o Instituto Brasileiro de Geografia e Estatística (IBGE) e a tabela CONCLA, também foram integradas, permitindo padronizar os segmentos de atuação e a localização geográfica das empresas.

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).

```
    payload = {
        "query": {
            "bool": {
                "must": [
                    {"match": {"description": "TotalPass"}},
                    {"match": {"location": "Brazil"}},
                    {"range": {"created": {"gte": "now-10M/M"}}}
                ],
                "must_not": must_not_filters
            }
        }
    }

    # Envia a requisicao POST para buscar novos job_ids
    response = requests.post(
        "https://api.coresignal.com/cdapi/v2/job_base/search/
           es_dsl",
        headers=headers, data=json.dumps(payload)
    )
```

Listing 4.1 – Consulta à API da Coresignal para coletar vagas

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()

        # 2. Gera filtros de exclusao para nao repetir empresas
           conhecidas
        must_not_filters = [{"match": {"company_name": nome}}
                            for nome in empresas_coletadas]
        # 3. Monta a consulta de busca
        payload = {
            "query": {
                "bool": {
                    "must": [
                        {"match": {"description": "TotalPass"}},
                        {"match": {"location": "Brazil"}},
```

```
                        {"range": {"created": {"gte": "now-10M/M"
                            }}}
                    ],
                    "must_not": must_not_filters
                }
            }
        }
        # Usa apenas a primeira vaga da empresa para garantir
            unicidade
        job_id = job_ids[0]
        # 4. Atualiza a lista de empresas coletadas
        empresas_coletadas.add(company_name)
        with open(empresas_path, "w") as f:
            json.dump(sorted(empresas_coletadas), f, indent=2)
```

Listing 4.2 – 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 prática e detalhada, o processo de construção do modelo computacional proposto para identificação do *Ideal Customer Profile* (ICP) no setor de benefícios corporativos. A abordagem combina técnicas de pré-processamento, detecção de outliers, medição de similaridade e ranqueamento final, estruturadas em um pipeline reproduzível e modular. A implementação foi realizada no ambiente Google Colab, utilizando a linguagem Python e bibliotecas como `scikit-learn`, `pandas` e `seaborn`. Cada etapa do fluxo foi encapsulada em funções que permitem aplicação em diferentes bases de dados, garantindo escalabilidade e adaptabilidade do método.

5.1 PRÉ-PROCESSAMENTO DOS DADOS

A etapa de pré-processamento foi responsável por transformar os dados brutos obtidos via *scraping* em uma estrutura adequada para aplicação de algoritmos de aprendizado de máquina. Inicialmente, os nomes das colunas foram padronizados para letras minúsculas e formato *snake_case*, além da remoção de espaços, acentos e colunas irrelevantes (como CNPJ, URLs, textos descritivos e campos auxiliares).

Em seguida, os dados da variável *localização*, originalmente armazenados como uma string heterogênea (e.g., “Curitiba, Paraná”, “São Paulo, SP”, “New York”), foram processados por meio de uma função de *parsing*. Essa função normalizou acentos e capitalização, identificou tanto siglas (UFs) quanto nomes de estados brasileiros por extenso, e criou dois novos campos: cidade e estado. Após essa etapa, apenas o campo estado foi mantido como variável categórica, enquanto valores internacionais permaneceram sem UF identificada, garantindo consistência nos casos nacionais.

O campo `capital_social` passou por uma limpeza que removeu separadores de milhar e casas decimais irrelevantes, sendo convertido para tipo numérico. O mesmo procedimento foi adotado para a variável funcionários, cujos valores foram convertidos em inteiros representando a força de trabalho de cada empresa.

Após essa limpeza, foram definidas duas classes principais de atributos:

- Atributos numéricos: `capital_social` e `funcionários`;
- Atributos categóricos: `segmento` e `estado`.

Para tornar os dados compatíveis com os modelos de *machine learning* utilizados, aplicou-se um `ColumnTransformer` contendo:

- um pipeline de atributos numéricos com `SimpleImputer` (mediana) e `StandardScaler`;

- um pipeline de atributos categóricos com `SimpleImputer` (moda) e `OneHotEncoder` (`sparse=False`).

A saída dessa etapa foi uma matriz vetorizada (`X_processed`) e um `DataFrame` equivalente (`X_df_processed`) com todas as colunas expandidas e normalizadas, permitindo total transparência e reprodutibilidade nos passos subsequentes.

5.2 DETECÇÃO DE OUTLIERS COM MODELOS ONE-CLASS CLASSIFICATION

Após a vetorização dos dados, aplicaram-se técnicas de *One-Class Classification* (OCC) com o objetivo de identificar e remover observações que destoam significativamente do perfil geral da base — isto é, empresas que não apresentam similaridade estrutural suficiente para serem consideradas candidatas ao ICP. Essa filtragem é fundamental para evitar que *outliers* distorçam as métricas de similaridade nas etapas subsequentes do modelo.

Foram empregados dois algoritmos não supervisionados clássicos para essa tarefa: o *One-Class Support Vector Machine* (OC-SVM) e a *Isolation Forest*. Ambos foram treinados diretamente sobre os vetores `X_processed`, obtidos após o pré-processamento.

O One-Class SVM utiliza um kernel RBF para identificar a fronteira que envolve a maioria dos dados considerados "normais", classificando como anomalias os pontos que se encontram fora dessa superfície. Para essa aplicação, utilizou-se um valor de $\nu = 0.05$, assumindo que até 5% da base poderiam ser *outliers*. O método foi mantido na pipeline apenas para fins comparativos e diagnósticos.

Já a *Isolation Forest*, modelo baseado em árvores aleatórias de partição, demonstrou maior robustez e interpretabilidade ao identificar anomalias por meio da facilidade com que uma instância pode ser isolada. Esse método também foi configurado com uma taxa de contaminação de 5%, e sua saída foi utilizada para filtrar os dados efetivamente: apenas as empresas classificadas como *inliers* (isto é, não-anômalas) foram mantidas para os próximos estágios.

Ao final dessa etapa, a base vetorizada foi reduzida de acordo com as predições da *Isolation Forest*. Os dados excluídos representam organizações com perfis atípicos, cuja presença poderia comprometer a construção de uma métrica de similaridade confiável. A Tabela 5.1 ilustra as estatísticas dos scores produzidos por ambos os modelos, e a Figura 5.2 mostra a distribuição dos valores.

A decisão de utilizar exclusivamente a *Isolation Forest* para filtragem foi motivada por sua simplicidade interpretativa, maior estabilidade nas iterações e melhor alinhamento com as propriedades estatísticas da base. O resultado foi uma matriz `X_filtrado` contendo apenas os vetores de empresas compatíveis com o perfil estrutural dominante, fornecendo uma base sólida para a análise de similaridade descrita na próxima seção.

5.3 CÁLCULO DE SIMILARIDADE VIA ESTRATÉGIAS DISTANCE-BASED (DBS)

Com a base de empresas "não anômalas" previamente filtrada pela *Isolation Forest*, o modelo passa a atribuir um grau de aderência de cada empresa ao perfil ideal (ICP) por meio de técnicas baseadas em distância. Essa abordagem, denominada *Distance-Based Scoring* (DBS), considera que empresas mais próximas ao "centro" da nuvem de dados são mais representativas do ICP, enquanto aquelas mais distantes tendem a se afastar desse padrão.

Duas estratégias distintas foram adotadas para esse cálculo de similaridade:

5.3.1 Distância ao Centróide

A primeira métrica parte do pressuposto de que o perfil médio das empresas ICP pode ser representado pelo centróide vetorial — isto é, a média de todas as variáveis numéricas e categóricas vetorizadas. A distância euclidiana de cada ponto em relação a esse centróide é então interpretada como uma medida inversa de similaridade: quanto menor a distância, maior a aderência ao ICP.

Matematicamente, essa distância é dada por:

$$d_i = \|x_i - \bar{x}\|_2 \tag{5.1}$$

Onde x_i é o vetor da empresa i e $\bar{x}$ é o centróide global. Para tornar o score mais interpretável, os valores foram normalizados via `MinMaxScaler` e invertidos, de forma que o score final assume valores entre 0 e 1, sendo 1 a empresa mais próxima ao centróide.

5.3.2 Distância Média aos k-Vizinhos Mais Próximos (k-NN)

A segunda métrica de similaridade foi baseada em *k-Nearest Neighbors* (k-NN). Para cada empresa, calcularam-se as distâncias para seus $k = 10$ vizinhos mais próximos no espaço vetorial, desconsiderando a si mesma. A média dessas distâncias foi utilizada como score: quanto menor a média, mais densa e típica é a vizinhança da empresa no espaço ICP.

Tal como no método anterior, os scores foram normalizados e invertidos para seguir a mesma lógica interpretativa: quanto maior o score, maior a probabilidade da empresa pertencer ao ICP.

5.3.3 Considerações

As duas abordagens se complementam. Enquanto a primeira captura a proximidade ao perfil médio, a segunda mede a densidade local de similaridade. Ao combiná-las, o

modelo oferece uma visão mais robusta da aderência de cada organização ao padrão desejado.

O uso dessas estratégias permite a criação de um ranking contínuo de empresas, de forma granular e interpretável. A próxima seção descreve como os diferentes scores — tanto de OCC quanto de DBS — foram combinados para formar o escore final de ICP.

5.4 RANKING FINAL

A etapa final do pipeline consiste em combinar os diferentes scores produzidos pelos modelos OCC (*One-Class Classification*) e DBS (*Distance-Based Scoring*) em uma métrica unificada de priorização. O objetivo é construir um ranking contínuo de empresas, ordenado pela sua similaridade ao perfil ideal de cliente (ICP).

Para cada empresa mantida após a filtragem da *Isolation Forest*, foram obtidos os seguintes indicadores:

- `oc_svm_score` — Score de decisão do One-Class SVM;
- `iso_forest_score` — Score de decisão da Isolation Forest;
- `dbs_centroid_score` — Score de similaridade baseado na distância ao centróide (já normalizado e invertido);
- `dbs_knn_score` — Score baseado na densidade local (média dos 10 vizinhos mais próximos), também normalizado e invertido.

Embora todos os quatro indicadores contenham informações relevantes, a maior confiabilidade foi atribuída aos dois últimos, por refletirem diretamente a similaridade vetorial entre empresas. Ainda assim, decidiu-se manter os scores OCC na combinação para capturar possíveis nuances de estrutura latente não linear nos dados.

5.4.1 Cálculo da Média Final

A composição do score final se deu pela média aritmética:

$$\text{score\_final}_i = \frac{1}{4}\left(\text{oc\_svm}_i + \text{iso\_forest}_i + \text{dbs\_centroid}_i + \text{dbs\_knn}_i\right) \tag{5.2}$$

Este valor representa a probabilidade relativa de que uma empresa pertença ao perfil ICP, sendo que quanto mais próximo de 1, maior a similaridade com as empresas originalmente identificadas como clientes.

6 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].

6.1 SEÇÃO SECUNDÁRIA

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

Abaixo, são apresentados exemplos de ilustrações.

6.1.1 Seção terciária

Abaixo, são apresentados exemplos de tabela.

6.1.1.1 Seção quaternária

Se houver seção quaternária, incluir texto ...

6.1.1.1.1 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).

7 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`

7.1 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.

7.2 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).

7.3 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.

8 CONCLUSÃO

Este elemento é obrigatório e é a parte final do texto. Nele, são apresentadas as conclusões identificadas a partir do desenvolvimento da pesquisa.

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.