Programa de Pós-Graduação em Pesquisa Operacional ITA/UNIFESP Disciplina:
Resolução de Problemas via Modelagem Matemática

Gabriel Adiron Ribeiro
Guilherme de Almeida Ferracini
Gustavo Guardia dos Santos Prado
Jardel Ferreira dos Santos
Kadem Gabriel Aídar

Clusterização de Portfólio Comercial para o Sistema de Ensino Poliedro Baseado em Dados

São José dos Campos, SP - Brasil
2025

1. SISTEMA DE ENSINO POLIEDRO

O Poliedro Educação é um grupo educacional com mais de 32 anos de história, possuindo colégios e cursos pré-vestibulares em São José dos Campos, São Paulo e em Campinas, além de possuir um sistema de ensino próprio. Atualmente, o Sistema de Ensino Poliedro está presente em todos os estados brasileiros e no Distrito Federal, com mais de 700 escolas associadas e aproximadamente 220 mil alunos.

1.1. Estrutura e Funcionamento do Sistema de Ensino Poliedro

O Sistema de Ensino Poliedro (SEP) atua como uma solução educacional completa, oferecendo às escolas parceiras um conjunto integrado de recursos pedagógicos, tecnológicos e de gestão. Os Colégios Poliedro estão entre as 10 primeiras posições nacionais do ENEM e são referência em aprovações nos vestibulares mais concorridos.

1.2. Justificativa do Projeto

A expansão no setor de educação básica, especificamente na venda de sistemas de ensino, exige estratégias devido à complexidade do ciclo de vendas e à diversidade do território e das escolas nacionais. Atualmente, a comercialização do Sistema de Ensino Poliedro (SEP) é realizada por uma equipe de consultores distribuídos pelo país, responsáveis pela prospecção e conversão de escolas privadas que ainda não integram a rede parceira.

No modelo vigente, a distribuição das carteiras de clientes e a definição de rotas de visita baseiam-se predominantemente no conhecimento empírico da gestão comercial. Embora funcional, essa abordagem apresenta limitações: a heterogeneidade das regiões brasileiras gera desequilíbrios na carga de trabalho e no potencial de vendas de cada consultor.

Diante desse cenário, justifica-se a necessidade de uma abordagem quantitativa para otimizar a força de vendas. A proposta deste trabalho é, através de um modelo matemático de clusterização, agrupar escolas com base em um potencial de venda, e não apenas por proximidade física, o projeto visa fornecer

uma base analítica para decisões estratégicas. Isso permitirá equilibrar as carteiras dos consultores (**equidade**), priorizar leads com maior probabilidade de conversão (**afinidade**) e assim maximizar o retorno sobre o investimento nas atividades de prospecção.

2. DEFINIÇÃO DO PROBLEMA E OBJETIVOS

2.1. Definição do Problema

O problema central deste trabalho reside em clusterizar cerca de 40 mil escolas privadas brasileiras não parceiras do Poliedro com objetivo de dividir de forma justa entre os consultores, ao minimizar a distância ponderando a restrição de equidade. Além disso, classificamos uma afinidade para separar potenciais clientes para o SEP com uma prioridade inferida a partir de atributos geográficos, demográficos, socioeconômicos e educacionais como nota do ENEM, infraestrutura, quantidade alunos etc.

Portanto, o problema configura-se como um desafio de otimização de clusters geográficos com restrições de carteira e classificação de afinidade com o SEP.

2.2. Objetivos

2.2.1. Objetivo Geral

Desenvolver um modelo de clusterização de escolas privadas brasileiras que auxilie o planejamento estratégico das visitas comerciais do Poliedro Sistema de Ensino, otimizando recursos, equilibrando carteiras e aumentando a efetividade da equipe de vendas.

2.2.2. Objetivos Específicos

1. Identificar e organizar as escolas brasileiras em clusters com igual oportunidade entre consultores comerciais, considerando o potencial de venda;
2. Desenvolver uma forma de quantificar a afinidade;

3. Desenvolver uma ferramenta que crie e otimize os clusters, a partir de inputs padrão;
4. Oferecer subsídios para a tomada de decisão do time comercial e permitir priorizar regiões estratégicas.

3. FUNDAMENTAÇÃO TEÓRICA

Para atingir o objetivo geral de desenvolver um modelo de clusterização que auxilie o planejamento estratégico das visitas comerciais do Poliedro, este trabalho recorre a um conjunto de técnicas de Ciência de Dados e Pesquisa Operacional. Assim, esta seção apresenta os conceitos essenciais para transformar dados brutos em conhecimento aplicável, cobrindo desde a limpeza das bases até a alocação dos consultores.

3.1. Descoberta de Conhecimento

A capacidade de extrair informações estratégicas de grandes bases de dados é um diferencial competitivo. Esse processo é formalmente definido na literatura clássica por **Fayyad, Piatetsky-Shapiro e Smyth (1996)** como *Knowledge Discovery in Databases* (KDD). Segundo os autores, o KDD é o "processo não trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados". Ele não se resume a uma única etapa, mas constitui um fluxo contínuo que engloba desde a seleção e tratamento dos dados até a interpretação dos resultados.

3.2. Processamento e Transformação de Dados

O trabalho com grandes volumes de informações públicas do governo impõe, inicialmente, desafios relacionados à dimensionalidade e à qualidade dos dados.

Embora seja intuitivo imaginar que quanto mais atributos e registros uma base possui, maior será a qualidade das análises, essa relação não é tão direta. Como explicam **Castro e Ferrari (2016)**, quando a dimensionalidade cresce demais, os dados tendem a se tornar esparsos, o que prejudica a estabilidade das medidas matemáticas empregadas nos algoritmos de mineração. Além disso, bases muito

amplas aumentam o custo computacional do processamento e tornam os modelos mais complexos e difíceis de interpretar.

Nesse contexto, a melhoria da qualidade dos dados exige procedimentos de **Transformação** e **Engenharia de Atributos**. Segundo **Castro e Ferrari (2016)**, essa etapa do processo de descoberta de conhecimento (KDD) envolve a conversão ou consolidação dos dados em formatos mais apropriados, como a agregação de valores e a construção de novos atributos que representem melhor o fenômeno estudado. Complementarmente, **Morettin e Singer (2021)** destacam que a criação de indicadores sintéticos atua como uma estratégia para condensar a informação, preservando a variabilidade essencial dos dados originais em um número menor de dimensões.

Para operacionalizar essa redução e evitar a instabilidade numérica, aplica-se a **Análise de Componentes Principais (PCA)**. Segundo **Morettin e Singer (2021)**, essa técnica de aprendizado não supervisionado permite reduzir a dimensão do conjunto de dados preservando a maior parte da sua variabilidade original, transformando variáveis correlacionadas em um novo conjunto de componentes principais não correlacionados.

Ainda no âmbito da qualidade dos dados, a presença de observações discrepantes pode comprometer a análise. **Morettin e Singer (2021)** definem *outliers* como observações que se desviam consideravelmente do padrão. Em técnicas baseadas em média e variância, como o PCA, esses valores podem distorcer as estimativas. Por isso, a identificação e o tratamento desses dados — seja via inspeção de boxplots ou cortes baseados em quantis — são etapas fundamentais para garantir a robustez do modelo.

3.3. Geocodificação

Superada a etapa de tratamento tabular, a dimensão espacial desempenha um papel determinante na estratégia comercial. A manipulação de dados espaciais exige a conversão de informações descritivas, como endereços postais, em referências geométricas precisas.

Segundo **Câmara et al. (2001)**, a característica fundamental da geoinformação é a capacidade de associar atributos a uma localização específica na superfície terrestre. Nesse sentido, a **geocodificação** atua como o processo

técnico de tradução de endereços textuais em coordenadas geográficas (latitude e longitude). Essa transformação é pré-requisito obrigatório para a aplicação de métricas espaciais, pois permite o cálculo de distâncias geodésicas e a identificação de relações de vizinhança, viabilizando análises quantitativas sobre a distribuição territorial das escolas.

3.4. Modelagem de Semelhança e Detecção de Perfil

Com os dados tratados e georreferenciados, o passo seguinte consiste na identificação das escolas com maior potencial de conversão. Para isso, adota-se a estratégia de *Lookalike Modeling* (Modelagem de Semelhança).

Segundo **Chacko, Pranav e Poornima (2016)**, essa abordagem consiste em utilizar aprendizado de máquina para encontrar novos prospectos que compartilhem características estruturais e comportamentais com uma base de "clientes semente" (*seed set*) já existente. A premissa central é que a similaridade nos atributos observáveis correlaciona-se positivamente com a probabilidade de aquisição do serviço.

Dada a natureza do problema — onde se dispõe apenas de dados da classe positiva (clientes atuais) —, a classificação tradicional torna-se inviável. Para contornar essa limitação, utiliza-se o algoritmo **One-Class Support Vector Machine (OC-SVM)**. Conforme definido por **Schölkopf et al. (2001)**, o OC-SVM é uma técnica que busca estimar o "suporte" de uma distribuição de alta dimensão. O algoritmo aprende uma fronteira de decisão que engloba a região onde se concentram os casos "normais" (o perfil do cliente típico), permitindo classificar novas observações como pertencentes a esse grupo ou como anomalias.

3.5. Otimização Combinatória e Designação de Recursos

Por fim, a alocação eficiente de recursos limitados (consultores) para atender à demanda identificada enquadra-se no escopo da Otimização Combinatória.

O problema clássico de Designação (*Assignment Problem*) consiste em encontrar o emparelhamento ideal entre tarefas e agentes. No entanto, quando os agentes possuem limitações reais — como tempo disponível, deslocamento

geográfico ou limite de contas — o modelo evolui para o **Problema de Designação Capacitado (CAP – Capacitated Assignment Problem)**.

Segundo **Hillier e Lieberman (2013)**, o CAP busca determinar a atribuição de n tarefas a m agentes de modo a otimizar a função objetivo, respeitando a restrição fundamental de que a soma das demandas atribuídas a um agente não pode exceder sua capacidade individual. Por ser um problema classificado como **NP-difícil**, ele exige modelagem matemática rigorosa para garantir que as carteiras geradas sejam operacionalmente viáveis.

No contexto específico de gestão comercial, esse desafio é tratado na literatura como o problema de **Alinhamento de Territórios de Vendas** (*Sales Territory Alignment*). Conforme estabelecido no trabalho seminal de **Zoltners e Sinha (1983)**, o alinhamento de territórios não é apenas uma questão logística, mas um modelo estratégico que busca maximizar o lucro ou a cobertura de mercado através da alocação ótima de contas e áreas geográficas aos vendedores, balanceando o potencial de vendas e a carga de trabalho.

Sob a ótica da pesquisa operacional, essa alocação modelada com restrições de capacidade é formalmente descrita como uma variação do **Problema de Designação Generalizada**. De acordo com **Cattrysse e Van Wassenhove (1992)**, o diferencial deste modelo em relação à designação simples é a permissão de que múltiplos itens (neste caso, escolas ou clientes) sejam atribuídos a um único agente (consultor), desde que o consumo acumulado de recursos não viole a capacidade disponível.

Os autores reiteram que, devido à natureza combinatória explosiva dessas permutações, o CAP exige algoritmos robustos e, muitas vezes, o uso de heurísticas para encontrar soluções ótimas em tempos computacionais aceitáveis para o negócio.

3.6. Resolução de Problemas de Grande Escala

Dada a magnitude dos dados envolvidos — com dezenas de consultores e dezenas de milhares de escolas gerando milhões de variáveis de decisão —, a resolução exata do modelo matemático requer o uso de solvers de alto desempenho baseados em Programação Linear Inteira Mista (MILP).

Neste cenário, a eficiência do método de solução é crítica. **Huangfu e Hall (2018)**, desenvolvedores do solver HiGHS, demonstram que a modernização das implementações do método Simplex Dual, combinada com técnicas de paralelismo, é fundamental para garantir a estabilidade numérica e a velocidade de convergência em problemas de otimização de grande escala. A utilização de tais ferramentas permite transpor a barreira teórica dos problemas NP-difícil, viabilizando a aplicação prática de modelos complexos de alocação em cenários corporativos reais.

4. METODOLOGIA

4.1. Ingestão, Tratamento e Processamento de Dados

O projeto seguiu um pipeline linear; iniciamos com conhecimento de negócio específico, esclarecemos as dúvidas iniciais com a equipe do Poliedro e definimos o conjunto de dados que seria utilizado.

4.1.1. Coleta e Integração

O processo iniciou-se com a integração de cinco fontes primárias de dados:

1. Censo Escolar (INEP 2024): Principal levantamento estatístico educacional brasileiro, utilizado para obter o cadastro das escolas, dados de infraestrutura e dependência administrativa;
2. Microdados do ENEM (INEP 2024): Base contendo o desempenho individual dos estudantes, agregada para gerar indicadores de desempenho por escola;
3. Dados Internos (Poliedro): Histórico comercial contendo a base de clientes ativos;
4. A localização dos consultores de vendas;
5. O preço médio do kit de materiais de cada aluno.

4.1.2. Seleção e Filtragem de Variáveis

O dataset inicial, composto por 215.545 instituições de educação básica, passou por uma filtragem inicial para isolar o público-alvo comercial do Poliedro, resultando em um universo de 37.178 escolas. Os critérios de seleção foram:

1. Dependência administrativa: inclusão apenas de escolas privadas;
2. Perfil institucional: exclusão de escolas confessionais, militares e instituições que não utilizam mediação presencial;
3. Tamanho da escola: remoção de escolas muito pequenas (menos de 20 alunos), por estarem fora do escopo estratégico.

4.1.3. Tratamento de Dados Geográficos (Geocodificação)

Para viabilizar a análise espacial, realizou-se a conversão dos CEPs das escolas e dos endereços dos consultores em coordenadas geográficas (latitude e longitude).

Foi desenvolvido um algoritmo em Python para consultas assíncronas à AwesomeAPI. Para otimizar a execução, implementou-se uma estratégia de cache local, onde o sistema consulta primeiramente um banco de dados proprietário, recorrendo à API apenas para novos registros.

4.1.4. Construção de Indicadores de Potencial e Engenharia de Atributos

Com o objetivo de caracterizar o perfil socioeconômico e estrutural das escolas, foram selecionadas variáveis do Censo Escolar e transformadas em indicadores sintéticos. Nesse processo, diversas colunas binárias foram combinadas e reescaladas, resultando em scores que sintetizam características relevantes das instituições, tais como infraestrutura, recursos pedagógicos e condições de funcionamento. Esses indicadores funcionam como dimensões de similaridade a serem exploradas nos algoritmos de clusterização. A tabela 4.1 apresenta a seleção das variáveis e a composição de cada score.

Paralelamente, foi desenvolvido um indicador complementar associado ao potencial econômico das escolas. A métrica valor_venda foi definida como uma estimativa de receita potencial anual por instituição. Para isso, utilizou-se o preço médio do kit de materiais de cada nível de ensino por aluno matriculado, somando os resultados para compor o total por escola.

Esse procedimento permite capturar heterogeneidades de porte e capacidade econômica das unidades, gerando uma dimensão adicional para a análise de segmentação.

Além disso, foram criadas métricas de densidade pedagógica (alunos por professor) e densidade ocupacional (alunos por sala), além de uma variável categórica relacionada ao tipo de ocupação do prédio (Próprio, Alugado ou Cedido). Também foi realizada a agregação de dados quantitativos de alunos e professores, com o objetivo de identificar o porte das instituições.

Tabela 4.1: Indicadores de potencial

Indicador	**Exemplos**	**Qtd**
Infraestrutura	Água, energia, esgoto, lixo	4
Estrut. Básica	Almox., pátio, refeit., cozinha	10
Estrut. Padrão	Audit., biblio., labs	8
Estrut. Premium	Quadra, piscina, música, acess.	12
Ativos Básicos	TV, DVD, retro	3
Ativos Premium	Lousa, port., tablets	3
Pessoal Básico	Adm., secr., serv.	5
Pessoal Premium	Psic., nutr., fono	5

Após a aplicação de filtros e combinação de colunas, o conjunto de dados foi reduzido de 426 variáveis para apenas 21.

4.1.5. Processamento dos Microdados do ENEM

O objetivo desta etapa foi calcular a Média da Nota do ENEM por escola, a ser utilizada como proxy de desempenho educacional no processo analítico. O tratamento aplicado aos Microdados do ENEM seguiu as etapas descritas abaixo:

1. Filtros de qualidade e elegibilidade: o dataset inicial continha aproximadamente 4,3 milhões de registros. Para assegurar a consistência

estatística das médias por escola, foram aplicados os seguintes critérios de filtragem:

a. Manutenção apenas dos participantes com código de escola válido, isto é, códigos presentes na tabela do Censo Escolar 2024;
b. Exclusão dos alunos que apresentaram ausência em qualquer um dos dias de prova;
c. Exclusão das escolas com menos de cinco participantes válidos, de modo a evitar distorções decorrentes de amostras muito reduzidas.

2. Cálculo das métricas: após a aplicação dos filtros (resultando em 1,18 milhão de observações), as notas foram agregadas por código de escola. Para cada instituição, foi calculada a média aritmética das notas das cinco áreas avaliadas: Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação;
3. Construção do indicador final: a variável nota_enem foi definida como a soma entre a média das notas objetivas e a média da nota de Redação por escola. Essa etapa resultou em um dataset final contendo 26.302 escolas e seus respectivos indicadores de desempenho no ENEM.

4.1.6. Limpeza e Unificação

Para garantir a robustez estatística das novas variáveis, aplicou-se um tratamento de outliers com base no percentil 98. Valores acima desse limite foram tratados como outliers e, dependendo do caso, substituídos pela mediana da variável ou limitados ao próprio valor do percentil, de modo a reduzir o impacto de observações discrepantes no modelo de clusterização.

Por fim, todos os datasets processados — Censo INEP, Microdados do Enem e dados dos consultores, a identificação de clientes ativos e o ticket médio — foram integrados em um arquivo mestre, consolidando a base analítica para as etapas subsequentes.

4.2. Modelagem de Afinidade

A afinidade expressa a semelhança que uma escola possui de escolas que adotam o Sistema de Ensino Poliedro.

4.2.1. Pré-processamento

Nesta fase, os dados foram preparados para a modelagem através das seguintes subetapas:

1. Separação dos Dados: Os dados foram separados em um conjunto de treinamento (escolas que são clientes conhecidos, onde cliente = 1) e um conjunto de teste (escolas prospecto, onde cliente = null);
2. Pipeline de Pré-Processamento: Um *pipeline* de pré-processamento foi definido e aplicado exclusivamente ao conjunto de treinamento:

    a. Imputação: Valores ausentes nas variáveis numéricas (como as notas do ENEM) foram preenchidos com a mediana (SimpleImputer(strategy="median"));
    b. Escalonamento: As *features* foram padronizadas (StandardScaler) para garantir que nenhuma variável dominasse o cálculo de distância do modelo;
    c. PCA (Análise de Componentes Principais): Foi aplicada a redução de dimensionalidade para eliminar redundâncias e ruídos, retendo 95% da variância original dos dados (n_components=0.95).

4.2.2. Modelagem One-Class SVM (OCSVM)

Para a detecção de perfil, foi aplicado o algoritmo One-Class SVM, treinado exclusivamente com os dados das escolas que já são clientes ("classe positiva"). O processo incluiu uma etapa de otimização de hiperparâmetros via validação cruzada (*cross-validation*) para definir os melhores valores de *nu* e *gamma*, procurando a melhor generalização da fronteira de decisão.

O modelo final foi aplicado a toda a base de dados, gerando dois indicadores para cada escola:

1. ocsvm_score: Variável contínua que quantifica a similaridade com o perfil de cliente (quanto maior o valor, maior a aderência);
2. ocsvm_binary: Classificador binário que rotula a escola como pertencente ao perfil (1) ou anômala (-1).

4.2.3. Resultado

O produto desta etapa foi a atribuição de um score de afinidade para cada escola do Brasil, representando a probabilidade estimada de conversão em cliente.

4.3. Preparação para Otimização

4.3.1. Cálculo da matriz de Distâncias

Foi calculada uma matriz de distâncias N x M entre todas as escolas e todos os consultores. Para garantir precisão, utilizou-se a distância geodésica (geopy.distance.geodesic), que considera a curvatura da Terra. Isso é uma matriz com 32 consultores do input do poliedro com as 37 mil escolas filtradas resultando em aproximadamente 1 milhão e 200 mil variáveis.

4.3.2. Exportação

Os dados processados foram consolidados contendo ID de cada escola, coordenadas, valor ponderado e a matriz de distâncias. Este artefato serve como input estruturado para o solver de otimização HIGHS na etapa subsequente.

4.4. Otimização das carteiras

Para a alocação final das escolas aos consultores, adotou-se uma abordagem de otimização exata utilizando **Programação Linear Inteira Mista (MILP)**. O problema abordado classifica-se formalmente como o **Problema de Designação Capacitado (CAP – Capacitated Assignment Problem).**

Diferente da aplicação simplificada do problema de designação, onde a relação entre tarefa e agente é de um para um, o CAP permite que um único consultor gerencie um conjunto de escolas, desde que a soma das demandas respeite os limites operacionais estabelecidos. O modelo foi implementado em Python utilizando a biblioteca PuLP e resolvido através do solver HiGHS. A escolha deste solver justifica-se pela sua eficiência em problemas de grande escala através do método Simplex Dual paralelo, garantindo a otimalidade global da solução, além de ser a melhor opção entre os solvers gratuitos. A formulação matemática é detalhada a seguir:

4.4.1. Definição de Conjuntos e Parâmetros

Para a modelagem, definem-se os seguintes conjuntos:

- **I**: Conjunto de **escolas** (*leads* qualificados) a serem alocadas, indexadas por *i*.
- **J**: Conjunto de **consultores** comerciais ativos, indexados por *j*.

Os parâmetros utilizados são:

- $\mathbf{d_{i}}$: Distância geodésica entre a escola *i* e o consultor *j*.
- $\mathbf{p_{i}}$: Potencial de Venda da escola *i*, calculado como o produto do *ticket* médio pela afinidade(score aproximado calculado anteriormente).
- $M_{med} = \frac{1}{\#J}\sum_{i}^{\#I} p_{i}$ Média global do potencial por consultor, utilizada como referência de equilíbrio.
- **α**: Parâmetro de cobertura mínima, que define quantos % das escolas serão distribuídas entre os consultores, esse parâmetro é definido pelo usuário na hora de rodar o solver.

4.4.2. Variáveis de Decisão

A alocação é controlada por variáveis binárias $\mathbf{x_{i}}$:

$$x_{ij} = \{1,\ se\ a\ escola\ i\ for\ atribuída\ ao\ consultor\ j\ e\ 0,\ caso\ contrário\}$$

4.4.3. Função Objetivo

O objetivo do modelo é **minimizar a distância total percorrida** pela equipe para atender a base de escolas, otimizando a logística de visitas

$$min \sum_{i} \sum_{j} d_{ij} x_{ij}$$

(Onde o primeiro somatório percorre todas as escolas $i \in I$ e o segundo todos os consultores $j \in J$).

4.4.4. Restrições

O modelo está sujeito às seguintes restrições operacionais:

1. **Restrição de Atribuição Única:** Garante que cada escola seja atribuída a exatamente um consultor, evitando duplicidade de atendimento.

$$\sum x_{ij} \leq 1 \; \forall i \in I$$

(Somatório sobre todas as escolas i deve ser menor ou igual a um consultor);

2. **Restrição de Carga Mínima (Cobertura):** Para garantir a equidade econômica da carteira, a soma do potencial das escolas alocadas a cada consultor deve ser, no mínimo, α% da média global.

$$\sum p_i x_{ij} \geq \alpha \, M_{med}, \; \forall \, j \in J$$

(Somatório sobre todos os consultores j).

4.4.5. Considerações sobre a Solução

Nesta formulação, optou-se por não impor um limite superior rígido (teto) à carteira. Essa estratégia permite que o *solver* aloque volume extra a consultores que estejam geograficamente muito próximos de aglomerados de escolas, priorizando a minimização do deslocamento global (Função Objetivo), desde que o piso mínimo de todos os demais consultores seja respeitado.

5. RESULTADOS

Esta seção apresenta os resultados obtidos a partir do tratamento, consolidação e análise das bases de dados utilizadas no estudo. Diferentemente da seção metodológica, aqui são destacados os efeitos práticos das etapas de processamento, evidenciando como os dados foram reduzidos, organizados e caracterizados, bem como as propriedades observadas na base final que subsidiam as análises subsequentes.

5.1. Consolidação e caracterização da base educacional

A base de dados educacionais utilizada neste estudo foi construída a partir dos microdados do Censo Escolar da Educação Básica de 2024, originalmente composta por 215.545 registros e 426 variáveis. Após o processo de consolidação e filtragem, obteve-se uma base final formada por 37.178 instituições de ensino, descritas por 23 variáveis, conforme sintetizado na Tabela 5.1.

O conjunto resultante representa exclusivamente escolas privadas em funcionamento, com oferta presencial e porte mínimo compatível com análises estruturais, constituindo uma amostra consistente para as análises subsequentes.

Tabela 5.1: Evolução da base de dados educacionais

Etapa do processamento	Número de escolas	Número de variáveis
Base original (Censo Escolar 2024)	215.545	426
Base consolidada final	37.178	23

5.1.1. Porte e escala das instituições

A análise do porte das instituições revela uma predominância de escolas de pequeno e médio porte. O número de alunos por escola variou entre 21 e 882 estudantes, com maior concentração nas faixas inferiores de matrícula. Esse comportamento evidencia um perfil institucional majoritariamente composto por unidades educacionais de menor escala de atendimento.

A distribuição do número de alunos por escola apresenta assimetria à direita, com cauda longa associada a instituições de grande porte, porém pouco frequentes na amostra.

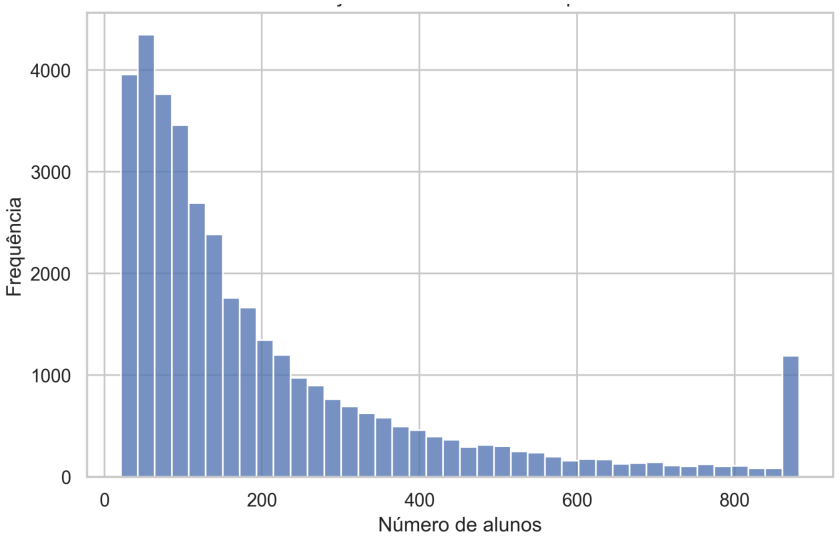

Figura 5.1: Distribuição de alunos

5.1.2. Indicadores operacionais e estruturais

A base consolidada permite a análise de indicadores operacionais relevantes, relacionados à disponibilidade de recursos humanos e físicos das instituições. A tabela 5.2 apresenta estatísticas descritivas das principais variáveis operacionais, incluindo quantidade de alunos, quantidade de professores, relação aluno-professor e relação aluno-sala.

É importante destacar que, por se tratar de uma base administrativa de abrangência nacional, podem ocorrer valores pontuais incoerentes ou inconsistentes, como registros nulos ou extremamente reduzidos para determinadas variáveis. A presença de valores mínimos iguais a zero, por exemplo, não representa necessariamente uma condição real de funcionamento, mas pode estar associada a falhas de preenchimento, inconsistências cadastrais ou defasagens no registro das informações.

Apesar dessas limitações, os indicadores centrais apresentados oferecem uma visão consistente do perfil operacional predominante das escolas analisadas. Observa-se que a relação média entre alunos e professores foi de 12,14 alunos por docente, enquanto a relação média entre alunos e salas utilizadas foi de 18,07

alunos por sala, sugerindo níveis moderados de ocupação dos recursos pedagógicos.

Tabela 5.2: Estatísticas descritivas das principais variáveis operacionais

Variável	Mínimo	Mediana	Média	Máximo
Quantidade de alunos	21	132	208.69	882
Quantidade de professores	1	12	19.22	91
Alunos por professor	0.78	10.73	12.14	41
Alunos por sala	0.63	15.8	18.07	63

5.1.3. Síntese da base final

A base educacional resultante apresenta-se como um conjunto de dados robusto e heterogêneo, com variações significativas entre instituições quanto ao porte e à estrutura operacional. Essa diversidade fornece suporte adequado para análises posteriores de afinidade institucional, proximidade geográfica e alocação ótima de recursos, desenvolvidas nas seções seguintes.

5.2. Consolidação e enriquecimento da base de escolas e consultores

Após a consolidação da base de escolas particulares apresentada na Seção 5.1, foi realizada uma etapa adicional de enriquecimento dos dados, com o objetivo de incorporar variáveis estratégicas para análises posteriores de afinidade, priorização e alocação territorial.

Ao final desse processo, obteve-se uma base composta por 37.178 escolas, agora descritas por 27 variáveis, preservando integralmente o recorte anterior e ampliando o potencial analítico do conjunto de dados.

Tabela 5.3: Resumo do enriquecimento e cobertura das variáveis adicionadas

Variável	Descrição	Total de escolas	Com informação	Cobertura (%)
Valor de venda	Estimativa de potencial econômico	37178	37178	100
Nota do enem	Proxy de desempenho acadêmico	37178	6564	17.66
Lat / Lon	Coordenadas geográficas por consultor	37178	37178	100
Cliente	Indica se a escola já pertence ao sistema Poliedro	37178	533	1.43

5.2.1. Cálculo do valor de venda estimado por escola

O potencial econômico de cada escola foi estimado por meio da variável valor de venda, construída a partir do número de matrículas por etapa de ensino, ponderadas por um ticket médio específico. Essa estimativa foi calculada para 100% das 37.178 escolas da base consolidada.

Observou-se elevada heterogeneidade nos valores estimados, refletindo a diversidade de portes e composições de oferta educacional. Essa característica reforça a necessidade de análises baseadas em estatísticas robustas e segmentações posteriores.

Tabela 5.4: Estatísticas descritivas do valor de venda estimado

Variável	Mínimo	P25	Mediana	Média	P75	Máximo
valor de venda (kR$)	11.812	48.051	104.568	251.006	311.871	2221.688

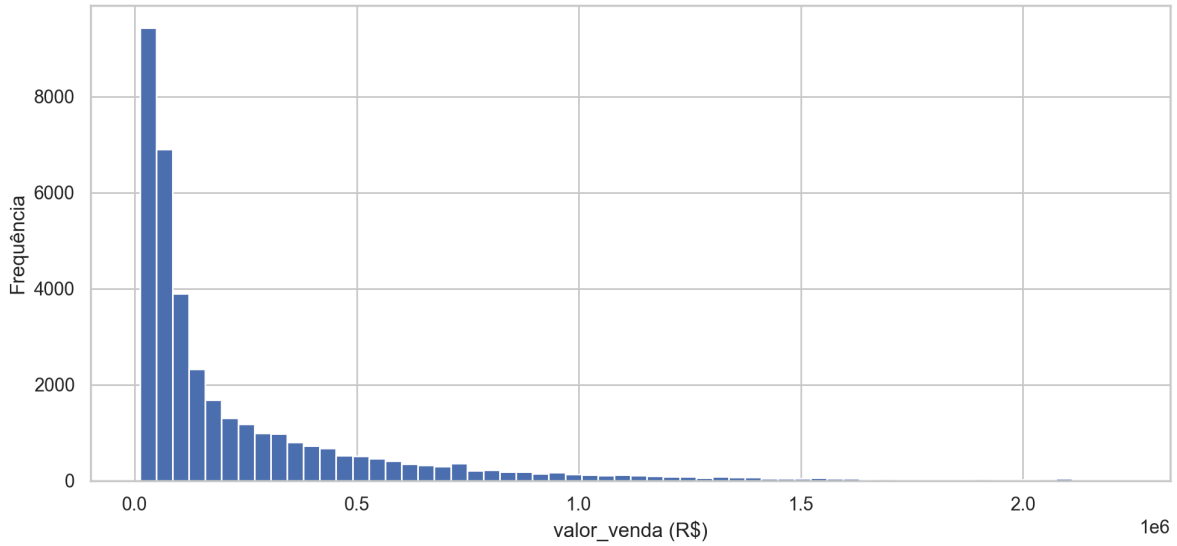

Figura 5.2: Distribuição do valor de venda estimado

5.2.2. Agregação das notas do ENEM por escola

Como proxy de desempenho acadêmico, foram incorporadas à base as notas do ENEM agregadas por escola. Após o processo de filtragem e agregação, 6.564 escolas passaram a contar com essa informação, correspondendo a 17,66% do total de instituições analisadas.

A cobertura parcial é esperada, dado que nem todas as escolas possuem participação representativa no exame ou correspondência válida de código no recorte considerado. Ainda assim, o subconjunto obtido é suficientemente amplo para análises comparativas e segmentações baseadas em desempenho acadêmico.

Tabela 5.5: Cobertura de escolas com nota ENEM

Total de escolas	Escolas com ENEM	Escolas sem ENEM	Cobertura ENEM (%)
37178	6564	30614	17.66

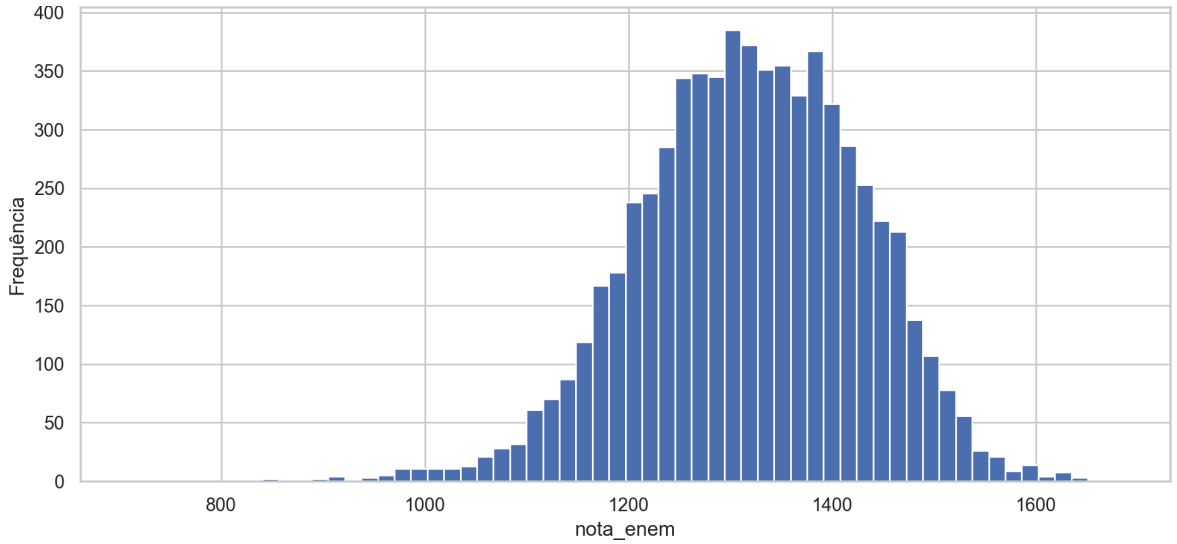

Figura 5.3: Distribuição da nota ENEM por escola

5.2.3. Georreferenciamento das escolas e consultores

A base consolidada foi integralmente georreferenciada, resultando em 100% das 37.178 escolas com coordenadas geográficas válidas. Adicionalmente, foram incorporadas informações de localização referentes a 24 consultores, também representados por latitude e longitude.

Essa estrutura espacial permite análises de proximidade, cálculo de distâncias e validação territorial dos dados.

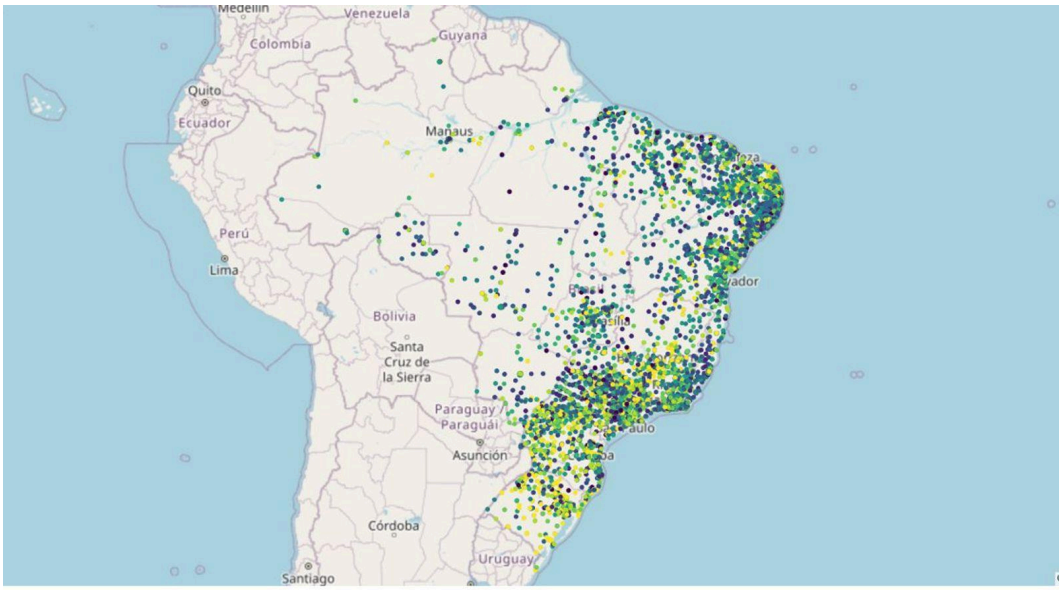

Figura 5.4: Dispersão geográfica de escolas

Figura 5.5: Dispersão geográfica de consultores

5.2.4. Identificação de escolas já pertencentes ao grupo

Por fim, foi incorporada à base a variável indicadora cliente, utilizada para identificar escolas já pertencentes ao grupo. Foram identificadas 533 escolas com correspondência válida entre a base consolidada e o cadastro interno.

Diferenças entre o número de escolas originalmente informadas e aquelas efetivamente encontradas decorrem, em geral, de inconsistências cadastrais ou ausência no recorte filtrado. Ainda assim, a informação obtida é suficiente para distinguir escolas já atendidas daquelas elegíveis para análises de expansão.

Tabela 5.6: Consolidação das escolas já pertencentes ao grupo

Códigos informados	Encontrados na base	Não encontrados	Taxa de match (%)
555	520	25	94.5

5.3. Resultados da alocação otimizada de escolas por consultor

Esta seção apresenta os resultados do modelo de alocação, com foco na distribuição das escolas entre consultores sob diferentes cenários de cobertura. Os resultados evidenciam o comportamento do modelo em termos de equilíbrio econômico, quantidade de escolas atribuídas e impacto da variação do nível de cobertura sobre o perfil das escolas selecionadas.

Ressalta-se que os consultores analisados nesta seção são apresentados apenas como exemplos ilustrativos, não representando a totalidade dos profissionais envolvidos na operação.

5.3.1. Comparação entre consultores no cenário de 30% de cobertura

A tabela 5.7 apresenta os resultados da alocação considerando um cenário de 30% de cobertura da base de escolas, para três consultores selecionados aleatoriamente. São reportadas, para cada consultor, a média do valor de venda das escolas atribuídas, a quantidade de escolas e a soma total do valor de venda.

Tabela 5.7: Comparação de indicadores econômicos entre consultores (30%)

Consultor:	**Média (R$)**	**Contagem**	**Soma (R$)**
Adriana Mota	411.819	357	147.019.608
Flavio Paim	469.715	313	147.020.894
Ricardo Melo	266.290	552	146.992.491

A análise dos resultados revela um comportamento consistente com o objetivo de distribuição econômica equilibrada entre os consultores. Observa-se que, embora a média do valor de venda por escola e a quantidade de escolas atribuídas variem significativamente entre os consultores, o valor total de venda alocado permanece muito próximo entre eles.

Esse resultado indica que o modelo de alocação cumpre adequadamente a restrição de balanceamento econômico, compensando diferenças no perfil das

escolas atribuídas por meio do ajuste no número de instituições. Consultores que recebem escolas com menor valor médio tendem a receber um número maior de escolas, enquanto aqueles com escolas de maior valor médio recebem menos unidades, mantendo o somatório final praticamente equivalente.

Tal comportamento confirma, do ponto de vista empírico, que a estratégia de alocação proposta consegue equilibrar a carga econômica entre consultores, mesmo diante da heterogeneidade natural das instituições educacionais.

5.3.2. Comparação entre níveis de cobertura: 10%, 30% e 95%

A tabela 5.8 amplia a análise ao comparar os resultados obtidos para três níveis distintos de cobertura da base: 10%, 30% e 95%. Para cada consultor, são apresentados conjuntamente a média do valor de venda por escola e a quantidade de escolas atribuídas em cada cenário.

Tabela 5.8: Comparação dos indicadores econômicos por nível de cobertura

Consultor:	**10%** **Média / Cont**	**30%** **Média / Cont**	**95%** **Média / Cont**
Adriana Mota	1.400.026 / 35	411.819 / 357	272.852 / 1706
Flavio Paim	664.364 / 74	469.715 / 313	462.234 / 1007
Ricardo Melo	701.091 / 70	266.290 / 552	372.996 / 1248

Os resultados evidenciam um padrão claro e consistente. À medida que o nível de cobertura aumenta, observa-se uma redução significativa no valor médio de venda por escola, acompanhada por um aumento expressivo no número de escolas atribuídas a cada consultor.

No cenário de 10% de cobertura, o modelo seleciona predominantemente as escolas com maior potencial econômico, resultando em médias elevadas de valor de venda e um número reduzido de instituições por consultor. Esse comportamento é esperado, uma vez que a seleção é fortemente concentrada nas melhores oportunidades disponíveis na base.

No cenário intermediário de 30% de cobertura, ocorre um equilíbrio entre valor médio e quantidade de escolas, refletindo uma expansão controlada do conjunto selecionado. Já no cenário de 95% de cobertura, praticamente toda a base elegível é considerada, incluindo escolas de menor porte e menor potencial econômico, o que explica a queda acentuada do valor médio e o aumento substancial da quantidade de escolas atribuídas.

Esses resultados reforçam que o modelo não realiza uma seleção aleatória, mas sim prioriza sistematicamente as escolas de maior valor quando a cobertura é restrita, incorporando gradualmente instituições de menor valor à medida que a cobertura se expande.

5.3.3. Síntese dos resultados da alocação

Em conjunto, os resultados apresentados nesta seção demonstram que o modelo de alocação atende simultaneamente a dois objetivos fundamentais:

1. Equilíbrio econômico entre consultores, mesmo diante de perfis distintos de escolas;
2. Controle explícito do perfil das instituições selecionadas por meio do parâmetro de cobertura.

A comparação entre diferentes níveis de cobertura evidencia como o modelo permite ajustar estrategicamente o foco da operação, priorizando qualidade econômica em cenários restritivos ou ampliando alcance territorial e institucional em cenários mais abrangentes, sem comprometer a coerência da distribuição entre consultores.

6. OPORTUNIDADES FUTURAS

Para garantir a escalabilidade e a melhoria contínua da solução desenvolvida, identificam-se duas frentes estratégicas de evolução:

- **Enriquecimento dos dados de entrada:** Incorporar aos dados públicos informações adicionais coletadas pelos consultores em campo, como a

mensalidade das escolas e a estrutura do bairro. Além disso, verificar a veracidade dos dados públicos, de modo a aumentar a assertividade da *afinidade*.

- **Aprimoramento do Modelo de Afinidade:** Aperfeiçoar o algoritmo OC-SVM adicionando novos dados e calibrar os pesos das variáveis preditivas, visando aumentar a precisão do score de afinidade.

7. CONCLUSÕES

O presente trabalho demonstrou a eficácia da integração entre técnicas de Pesquisa Operacional e Ciência de Dados para a resolução de problemas complexos de gestão de força de vendas. Ao transpor o modelo de decisão empírico para uma abordagem puramente baseada em dados, foi possível mitigar as distorções causadas pela heterogeneidade territorial brasileira e pela subjetividade na alocação de territórios.

A formulação do problema como um Problema de Designação Capacitado (CAP) garantiu que a minimização do deslocamento geográfico não ocorresse em detrimento da equidade econômica das carteiras. Os resultados obtidos nos diferentes cenários de cobertura (10%, 30% e 95%) confirmam que o modelo prioriza sistematicamente as melhores oportunidades de mercado, mantendo um equilíbrio robusto no potencial de venda total entre os consultores. Aliado a isso, a aplicação do algoritmo One-Class SVM permitiu uma modelagem de semelhança identificando padrões em clientes atuais para priorizar prospecções com maior probabilidade de conversão.

Do ponto de vista estratégico, a solução entrega mais do que uma simples lista de visitas; ela fornece uma ferramenta de planejamento de longo prazo. O uso do solver HiGHS provou ser fundamental para lidar com a magnitude dos dados — atingindo mais de 1,2 milhão de variáveis de decisão — e garantir a otimalidade global em tempo computacional viável para o negócio.

8. REFERÊNCIAS

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Censo Escolar da Educação Básica 2023: Notas Estatísticas**. Brasília: INEP, 2024. Disponível em: https://www.gov.br/inep. Acesso em: 14 set. 2024.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Microdados do Enem 2023**. Brasília: INEP, 2024. Disponível em: https://www.gov.br/inep. Acesso em: 14 set. 2024.

CÂMARA, G. et al. **Anatomia de sistemas de informação geográfica**. Campinas: Instituto de Computação, UNICAMP, 2001.

CASTRO, L. N. de; FERRARI, D. G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016

CATTRYSSE, D. G.; VAN WASSENHOVE, L. N. A survey of algorithms for the generalized assignment problem. **European Journal of Operational Research**, v. 60, n. 3, p. 260-272, 1992.

CHACKO, A.; PRANAV, B. A.; POORNIMA, A. A Study of Machine Learning Techniques for Customer Lookalike Modeling. In: IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW), 16., 2016, Barcelona. **Proceedings...** Barcelona: IEEE, 2016. p. 239-243.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

HILLIER, F. S.; LIEBERMAN, G. J. **Introdução à pesquisa operacional**. 9. ed. Porto Alegre: AMGH, 2013.

HUANGFU, Q.; HALL, J. A. J. Parallelizing the dual simplex method. **Mathematical Programming Computation**, v. 10, n. 1, p. 119-142, 2018.

MORETTIN, P. A.; SINGER, J. M. **Estatística e Ciência de Dados**. Rio de Janeiro: LTC, 2021.

SCHÖLKOPF, B. et al. Estimating the support of a high-dimensional distribution. **Neural Computation**, v. 13, n. 7, p. 1443-1471, 2001.

ZOLTNERS, A. A.; SINHA, P. Sales Territory Alignment: A Review and Model. **Management Science**, v. 29, n. 11, p. 1237-1256, 1983.