Aplicação de Inteligência Artificial para Análise de Oportunidades Comerciais no Agronegócio

Victor Ramires da Silva

Bruno Magalhães Nogueira (Orientador)

[1]Universidade Federal de Mato Grosso do Sul

victor.ramires157@gmail.com, bruno.nogueira@facom.ufms.br

Abstract. Este artigo apresenta a aplicação de técnicas de inteligência artificial na análise de oportunidades comerciais no setor agropecuário brasileiro com o desafio de identificar os municípios mais promissores para expansão de uma rede varejista do setor, buscando maior chance de sucesso e retorno financeiro. Utilizando dados de faturamento fornecidos por uma empresa parceira, junto a dados populacionais, geográficos e agropecuários de fontes públicas, foram empregados algoritmos de aprendizado de máquina supervisionado, como o *Random Forest Regressor*, e de aprendizado de máquina não supervisionado, como a *clusterização hierárquica*. A técnica *k-Nearest Neighbors (KNN)* também foi utilizada para análise de similaridade entre municípios. O modelo, implementado em Python, busca apoiar estratégias de expansão com base em dados.

1. INTRODUÇÃO

A adoção de técnicas de inteligência artificial no agronegócio tem crescido significativamente em todo o mundo, impulsionando ganhos de produtividade, automação de processos e melhor tomada de decisão em toda a cadeia produtiva. Soluções baseadas em IA vêm sendo utilizadas para prever safras, otimizar uso de insumos, monitorar a saúde de rebanhos e automatizar operações logísticas. No entanto, apesar do avanço global, ainda são escassas as aplicações voltadas à expansão estratégica e direcionada na busca por melhores oportunidades comerciais, sobretudo no contexto brasileiro, caracterizado por uma grande diversidade territorial e produtiva.

Segundo dados do Instituto de Pesquisa Econômica Aplicada (IPEA), o Produto Interno Bruto (PIB) do agronegócio brasileiro cresceu cerca de 40% entre 2009 e 2022, passando de R$1,8 trilhões para R$2,5 trilhões. No mesmo intervalo, as exportações do

setor aumentaram 146%, atingindo US$159 bilhões, enquanto a produção de grãos mais que dobrou, saltando de 135 milhões para mais de 320 milhões de toneladas (CEPEA, 2023; CONAB, 2023).

A crescente disponibilidade de dados estruturados tem viabilizado o desenvolvimento de soluções de inteligência de mercado voltadas à expansão no varejo, permitindo a identificação de áreas com maior potencial de lucratividade e a redução de riscos associados à abertura de novas filiais. No setor varejista, a inteligência artificial tem se destacado em aplicações como personalização e planejamento de demanda. Além disso, seu impacto vai além de casos específicos, trazendo valor para toda a organização ao transformar os sistemas internos de gestão do conhecimento, reduzindo o tempo de busca por informações e elevando a produtividade dos colaboradores e acelerando a tomada de decisão (McKINSEY & COMPANY, 2023).

Reduzir ao máximo os erros na escolha de novos locais para expansão comercial é um fator estratégico no varejo, uma vez que decisões equivocadas podem resultar em investimentos com baixo retorno. Considerando a vasta diversidade territorial e produtiva do Brasil, torna-se crucial identificar municípios que apresentem características favoráveis ao sucesso de unidades comerciais. A análise criteriosa de indicadores regionais permite reduzir incertezas e direcionar os recursos de forma mais eficiente.

Sob essa ótica, este trabalho tem como objetivo principal desenvolver um modelo preditivo para apoiar a expansão comercial de uma rede do varejo agropecuário, combinando técnicas de inteligência artificial para identificar padrões territoriais, estimar o faturamento potencial de novos pontos de venda e traçar um panorama preditivo da viabilidade comercial em regiões onde a empresa não está presente. Para isso, propõe-se a integração de dados populacionais e agropecuários com algoritmos de clusterização, regressão supervisionada e K-Nearest Neighbors (KNN), oferecendo uma ferramenta completa para orientar decisões estratégicas baseada em dados. A escolha dessas abordagens permite uma análise direcionada, já que agrupa cidades com perfis produtivos semelhantes, possibilitando análises dos perfis de cidades e, em seguida, utiliza modelos robustos de regressão para estimar o faturamento de cada mês do primeiro ano operacional de uma loja nova.

Este trabalho está organizado em quatro seções principais: a primeira apresenta

a fundamentação teórica, com os principais conceitos e algoritmos aplicados à análise de dados e à previsão de faturamento; a segunda aborda a metodologia, detalhando as etapas de preparação e modelagem dos dados, com ênfase na aplicação das técnicas de clusterização hierárquica e regressão por meio do algoritmo Random Forest; a terceira seção apresenta os resultados obtidos, com destaque para o agrupamento de municípios com base em suas características produtivas, também detalha a análise da base de vendas e os resultados obtidos nos modelos preditivos; e, por fim, a quarta seção reúne as referências bibliográficas utilizadas no desenvolvimento e pesquisa.

2. FUNDAMENTAÇÃO TEÓRICA

Esta seção tem como objetivo apresentar os conceitos e métodos fundamentais que sustentam todo o trabalho. Por se tratar da aplicação de técnicas de inteligência artificial voltadas à análise de oportunidades comerciais no agronegócio, torna-se indispensável entender os alicerces teóricos que justificam a aplicação de algoritmos de clusterização, modelos de regressão e estratégias de tratamento de dados, como pré-processamento e pós-processamento.

2.1 Base de Dados

A base de dados representa o conjunto estruturado de informações utilizadas para alimentar os algoritmos de clusterização e regressão empregados neste trabalho. Ela é composta por duas partes principais: os registros reais de faturamento de unidades comerciais já em operação, fornecidos por uma rede varejista do setor agropecuário; e os indicadores populacionais, geográficos e produtivos dos municípios brasileiros (IBGE, 2022), obtidos a partir de fontes públicas e oficiais, como o Instituto Brasileiro de Geografia e Estatística (IBGE), o sistema SIDRA e o projeto MapBiomas. A construção da base de dados se deu por dois processos principais: extração de dados e pré-processamento.

2.1.1 Extração de dados

A extração é a etapa inicial, onde os dados brutos são obtidos a partir de fontes diversas. Essa fase é essencial para reunir as informações que servirão de base para os algoritmos de clusterização e regressão. No contexto do trabalho foram utilizados abordagens de extração distintas, a primeira consistiu em coletar os dados à partir dos dados públicos do IBGE, SIDRA, MAPBIOMAS e Produção agrícola municipal (PAM). Os dados são

disponibilizados pelas instituições e estão disponíveis em arquivos CSV ou XLSX, por município, estado ou nível nacional. Os dados coletados das fontes públicas nessa etapa abrangem três principais características: Demográficas, Geográficas e Agropecuárias.

Demográficos: População, Taxa de Alfabetização (%), Idade Mediana, Índice de Envelhecimento (razão entre idosos e jovens), Domicílios, Razão de Sexo (razão entre homens e mulheres) e População Indígena

Geográficos: Área (km^2), Latitude, Longitude, Município (UF), Código IBGE (identificador único)

Agropecuários: Rebanho, Ordenha, Imóveis Rurais, Área de Pastagem (ha), Soja (ha), Milho (ha), Cana-de-Açúcar (ha), Sorgo (ha), Café (ha), Algodão (ha) e Eucalipto/Pinus (ha). E, por fim, a extração dos dados de vendas, que foi realizada por meio de consulta direta ao banco de dados da empresa parceira. A base de dados disponibilizada pela empresa continha: histórico de vendas de 2008 a 2019, data da venda, cidade da loja, tipo de produto e quantidade vendida. Todos os dados extraídos nessa etapa foram armazenados em arquivos CSV e tratados utilizando Python na etapa de pré processamento de dados.

2.1.2 Pré-processamento de dados

O pré-processamento é uma etapa crucial, na qual os dados brutos são preparados para análise por meio de transformações que garantem sua consistência, removendo valores ausentes, ruídos e informações irrelevantes. Esse processo envolve tarefas como limpeza, integração, transformação e redução de dados, sendo essencial para garantir a qualidade e a eficácia dos algoritmos de aprendizagem de máquina (Jiawei Han, 2011). Nesta fase, a base de faturamento da empresa parceira, foi anonimizada a pedido da fornecedora, substituindo os identificadores reais dos municípios por códigos, assim como os tipos de produto, de modo a preservar a confidencialidade sem comprometer as análises. [1]

2.2 Técnicas de inteligência artificial aplicadas

Das técnicas de inteligência artificial utilizadas, podemos destacar duas abordagens principais: algoritmos de clusterização e modelos de regressão. A combinação dessas técnicas permite explorar padrões territoriais e realizar projeções fundamentadas, contribuindo para a definição de regiões prioritárias na expansão comercial.

[1] A anonimização dos dados de faturamento foi realizada para garantir a privacidade e a confidencialidade, em conformidade com princípios éticos, assegurando que nenhuma cidade, loja ou tipo específico de produto pudesse ser identificado, visando o uso responsável da informação.

2.2.1 Algoritmo de clusterização

Algoritmos de clusterização são técnicas de aprendizado não supervisionado utilizadas para agrupar dados com base em similaridades, sem a necessidade de rótulos previamente definidos. Esses modelos analisam as características dos dados e formam grupos, permitindo a identificação de padrões ocultos no conjunto de dados. Neste trabalho, foi aplicada a técnica de clusterização hierárquica aglomerativa (WARD, 1963), que constrói uma árvore de agrupamentos (dendrograma) ao juntar iterativamente os dados mais semelhantes até atingir um número predeterminado de grupos. Essa abordagem foi empregada para segmentar os municípios brasileiros com base em indicadores populacionais, geográficos e produtivos, possibilitando a análise de diferentes perfis de cidades e apoiando a identificação de padrões que orientam decisões estratégicas. Por se tratar de um algoritmo de aprendizado não supervisionado, a clusterização não requer um conjunto de treino e teste. Em vez disso, o modelo avalia as relações entre os dados como um todo e constrói os agrupamentos diretamente com base nas distâncias euclidianas entre cada uma das features e na estrutura da variância, que representa o quanto os dados se afastam do centróide de cada cluster.

O resultado final da clusterização foi incorporado ao conjunto de dados por meio da criação de uma nova coluna denominada Cluster na base de dados, permitindo análises comparativas entre os diferentes grupos de municípios formados.

2.2.2 Algoritmo de regressão

Modelos de regressão são técnicas estatísticas e computacionais voltadas à estimativa de variáveis numéricas contínuas com base em variáveis explicativas. Eles buscam aprender a relação entre variáveis independentes e uma variável alvo, ajustando funções capazes de realizar previsões com base em dados observados.

Neste trabalho, foram testados diferentes modelos de regressão para prever o faturamento potencial de municípios brasileiros utilizando informações sobre população, localização e produção agropecuária, retiradas de fontes públicas. Entre os algoritmos avaliados estão o Support Vector Regressor (SVR), a rede neural MLP (Multi-Layer Perceptron), a regressão linear tradicional e, por fim, um método mais simples baseado em média móvel. O modelo que apresentou melhor desempenho foi o Random Forest Regressor (BREIMAN, 2001), que se destaca por combinar múltiplas árvores de decisão construídas

a partir de subconjuntos aleatórios dos dados, oferecendo maior robustez contra overfitting e melhor capacidade de generalização.

Para validar o modelo, os dados de municípios com lojas foram divididos em dois conjuntos: um para treinamento e outro para teste. Em todos os modelos citados, 80% da base foi utilizada como teste e 20% como treino. O conjunto de treino serve para ajustar os parâmetros internos do modelo com base em padrões históricos, enquanto o conjunto de teste permite avaliar como o modelo se comporta com dados desconhecidos, como se fossem previsões reais. Após o treinamento e validação, o modelo foi utilizado para prever o faturamento de cada mês do primeiro ano de operação das cidades-alvo. A avaliação foi realizada por meio do Root Mean Squared Error (RMSE) e do R-squared (R^2), métricas de erro que serão explicadas com mais detalhes nas seções 2.3.1 e 2.3.2 respectivamente. O resultado final fornece uma visão preditiva detalhada do desempenho previsto por município, apoiando decisões estratégicas de expansão comercial no setor agropecuário.

2.3 Pós-processamento de dados

O pós-processamento é a etapa dedicada à análise e interpretação dos resultados obtidos. No caso da clusterização, envolve a validação dos grupos formados, a verificação da coerência entre os municípios agrupados e a adequação do número de clusters. Para os modelos de regressão, a qualidade das previsões é avaliada por meio de métricas estatísticas que indicam a proximidade entre os valores previstos e os valores reais. Neste trabalho, foram utilizadas duas métricas principais: o Root Mean Squared Error (RMSE), e o R-squared (R^2). Quando bem conduzido, o pós-processamento assegura que os resultados sejam consistentes, relevantes e aplicáveis à tomada de decisões estratégicas.

2.3.1 Root Mean Squared Error (RMSE)

O RMSE (Root Mean Squared Error) mede a média das diferenças ao quadrado entre os valores previstos e os reais, sendo expresso na mesma unidade da variável target e é definida pela equação (1):

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \tag{1}$$

Quanto menor o RMSE, melhor o desempenho do modelo, indicando menor dispersão dos erros (CHAI; DRAXLER, 2014).

2.3.2 R-squared (R^2)

R^2 (R-squared) representa a proporção da variabilidade dos dados que é explicada pelo modelo. Seus valores próximos de 1 indicam que o modelo possui boa capacidade explicativa, valores próximos de zero indicam baixa capacidade explicativa (CHICCO; WARRENS; JURMAN, 2021). Apesar de ser muito utilizada,é importante destacar que existem autores que criticam essa métrica pois ela é sensível a outliers, podendo não representar exatamente como o modelo se adaptou aos dados, tornando-a uma métrica menos confiável (LEWIS-BECK; SKALABAN, 1990) e, por esse motivo, deve ser utilizada junto com outras métricas. A fórmula utilizada para o cálculo da métrica R^2 é mostrada na equação 2.

$$\mathrm{R2}(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2}$$

3. CLUSTERIZAÇÃO E ANÁLISE DOS CLUSTERS

Nesta etapa do trabalho, foi realizada o agrupamento dos municípios brasileiros com base em indicadores agropecuários, populacionais e de produção agrícola. O objetivo foi identificar grupos de cidades com características semelhantes, permitindo uma análise mais estratégica do território nacional.

Os dados utilizados nessa etapa foram obtidos de diferentes fontes públicas, como o IBGE, o SIDRA e a Produção Agrícola Municipal (PAM). A base reuniu todas as informações agropecuárias, demográficas e geográficas extraídas na etapa 2.1.1 e, antes a aplicação da clusterização, foi realizado o cálculo da distância entre cada uma das cidades do Brasil, utilizando a fórmula de Haversine, que calcula a distância em linha reta entre dois pontos(C. C. Robusto, 1957).

$$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{3}$$

As cidades consideradas para esta etapa foram classificadas como “Oportunidade”, com base em critérios definidos em conjunto com a área de negócios da empresa parceira do trabalho. São localidades que se destacam por sua densidade agropecuária, escala produtiva ou influência sobre outras cidades no entorno, e que, por isso, merecem uma

análise mais aprofundada. Para essa classificação, foi calculado o somatório regional das variáveis: população, área de pastagem (ha), milho (ha), rebanho, imóveis rurais, ordenha, cana (ha) e soja (ha), considerando todas as cidades localizadas em um raio de 75 km de cada município do Brasil. Um município foi considerado uma oportunidade se atendesse a pelo menos uma das condições a seguir:

- **Condição 1:** Menos de 30 cidades no entorno, com somatório mínimo de 250 mil cabeças de rebanho, 250 mil hectares de pastagem e 2.000 imóveis rurais. Essa condição identifica áreas produtivas cujo potencial de consumo é sustentado pela própria base agropecuária local.
- **Condição 2:** Pelo menos 30 cidades no entorno, com somatório de 500 mil cabeças de rebanho, 300 mil hectares de pastagem e 4.500 imóveis rurais. Neste caso, o que prevalece é a força da região como um todo — cidades com maior influência territorial, capazes de atrair produtores de toda a área ao redor.

A Tabela 1 apresenta um exemplo prático do somatório das variáveis para uma cidade que atende aos critérios de "Oportunidade". Nela, são listadas todas as cidades localizadas em um raio de 75 km de Campo Grande – MS, com os respectivos valores das variáveis consideradas, incluindo também os dados da própria cidade. Cidades que não atenderam aos critérios de oportunidade foram removidas da base.

Tabela 1 – Cidades dentro do raio de 75 km de Campo Grande - MS e total da soma

Distância (km)	Municípios	População	Rebanho 2023	Ordenha 2023	Imóveis Rurais	Pastagem	Soja	Milho	Cana
65.38	Bandeirantes	7940	180121	2306	1348	128059	93044	23487	0
70.58	Corguinho	4783	214525	5400	6840	150892	2870	500	150
72.62	Dois Irmãos do Buriti	11100	190490	3127	8380	109809	20806	1384	0
42.24	Jaraguari	7139	158230	1857	2267	137544	41564	10411	100
28.00	Rochedo	5199	139724	1213	9480	73625	9478	3895	250
63.48	Sidrolândia	47118	192460	3272	1861	87598	2510	2008	1177
22.30	Terenos	17652	246446	3344	2192	125154	400	175	50
0.00	Campo Grande	898100	728363	32332	2883	422849	11013	5038	18
Total	**8**	**999031**	**1757399**	**19841**	**35251**	**1235530**	**181685**	**46898**	**1745**

Esses dois critérios permitiram contemplar diferentes perfis de oportunidade — tanto cidades altamente produtivas de forma isolada quanto aquelas com papel estratégico no contexto regional. A partir dessa filtragem inicial, foi possível avançar para a etapa de análise e clusterização com um total de 3118 cidades. Para o agrupamento dos municípios, foi adotado o algoritmo de clusterização hierárquica aglomerativa com método 'Ward'. Essa abordagem permite a visualização de dendrogramas, que mostram graficamente os

agrupamentos e a distância entre cada um dos clusters, facilitando a identificação da quantidade ideal de clusters.

A quantidade de clusters foi feita com base na análise visual do dendrograma, observando os maiores saltos nas alturas dos ramos e definindo pontos de corte buscando diversidade entre os perfis e o equilíbrio entre a quantidade de cidades em cada cluster, evitando clusters extremamente pequenos ou dominantes e a definição do valor da quantidade de clusters foi fundamentada na interpretação dos agrupamentos resultantes. A distribuição dos clusters pode ser visualizada na Figura 1.

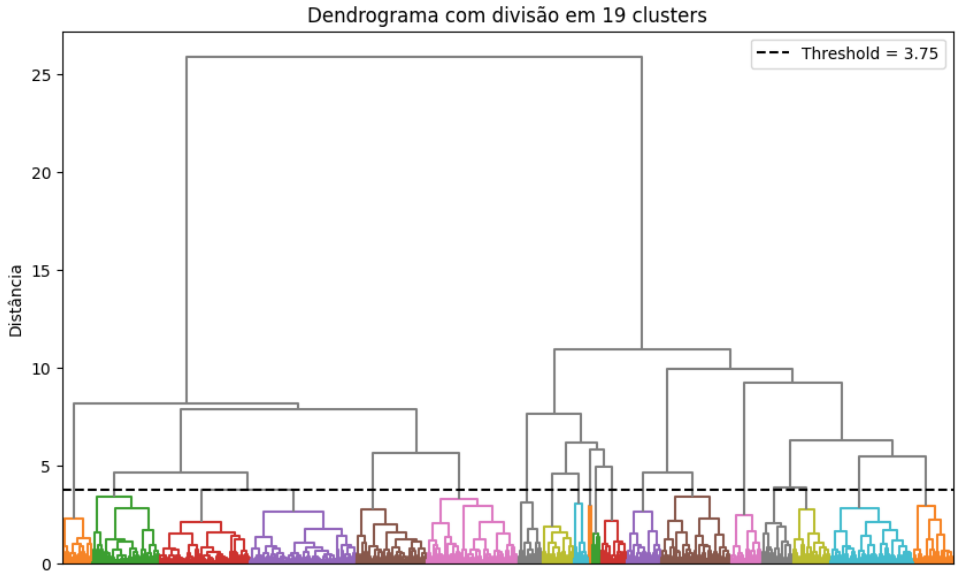

Figura 1 – Dendrograma das cidades classificadas como Oportunidade

Após a análise, o agrupamento foi concluído com 19 clusters e a quantidade de cidades por cluster pode ser visualizada na Tabela 2, que mostra, em ordem decrescente, a quantidade de cidades em cada cluster.

A partir da definição dos clusters, foi possível representar graficamente como esses agrupamentos se distribuem no território nacional. Enquanto o dendrograma fornece uma visão da proximidade entre os municípios com base nas variáveis, a Figura 2 permite observar a distribuição espacial dos grupos, revelando padrões regionais e a concentração de determinados perfis agropecuários em diferentes partes do país. Essa representação ajuda a entender como os clusters se organizam geograficamente, o que é fundamental para orientar estratégias de expansão territorial de forma mais direcionada.

Tabela 2 – Quantidade de cidades por cluster

Código do cluster	**Quantidade de cidades**
6	373
3	319
18	315
5	288
12	247
0	245
1	236
2	141
8	132
14	119
16	111
17	108
9	108
11	106
10	91
7	85
15	49
13	30
4	15

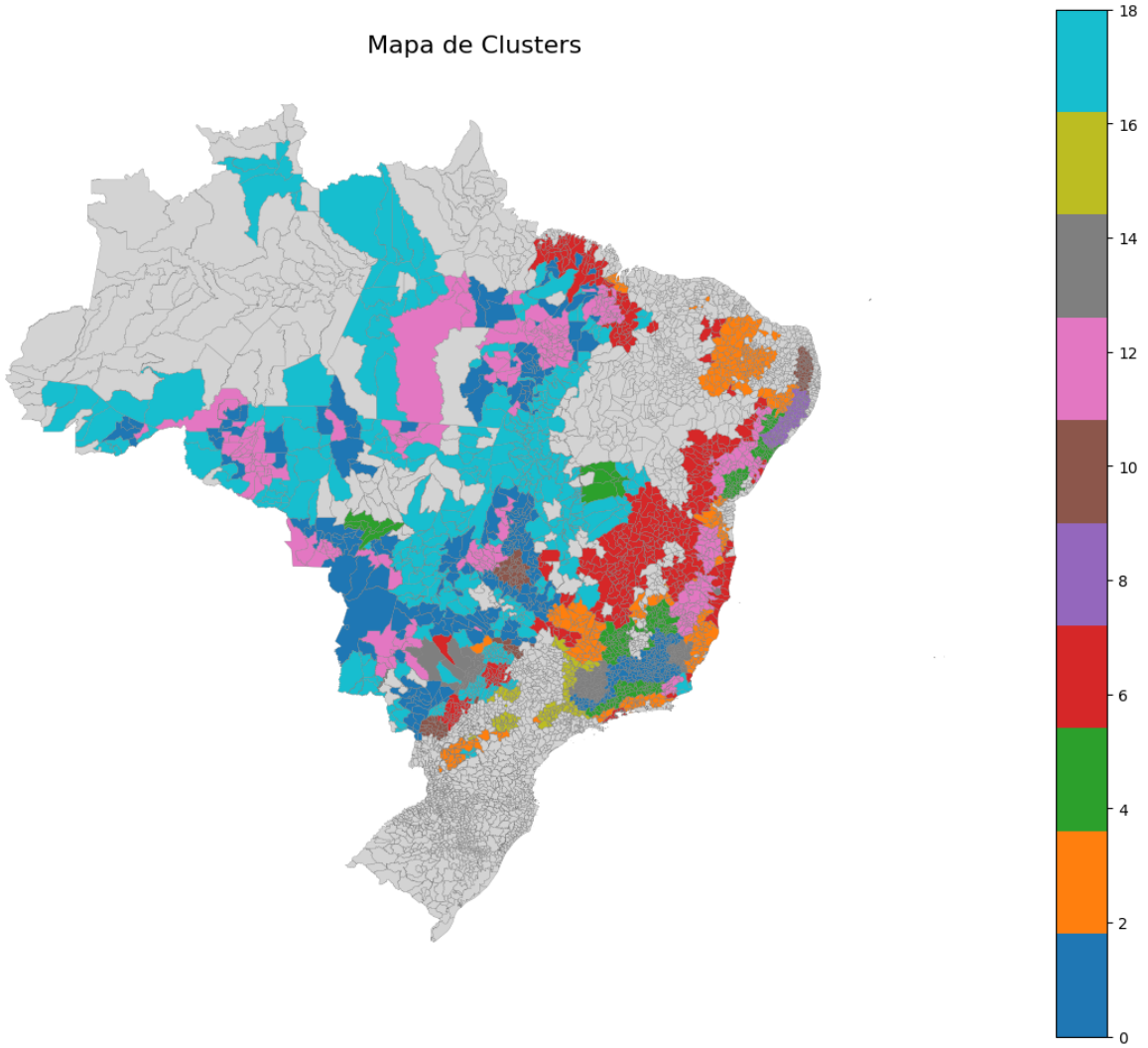

Figura 2 – Distribuição espacial dos clusters

Para entender melhor o perfil de cada cluster, foi inicialmente calculada a média de cada uma das variáveis agropecuárias e populacionais, com os valores agregados por grupo. O resultado foi apresentado na figura 3, um gráfico comparativo, que permite identificar com clareza quais clusters concentram os maiores valores médios em cada uma das variáveis.

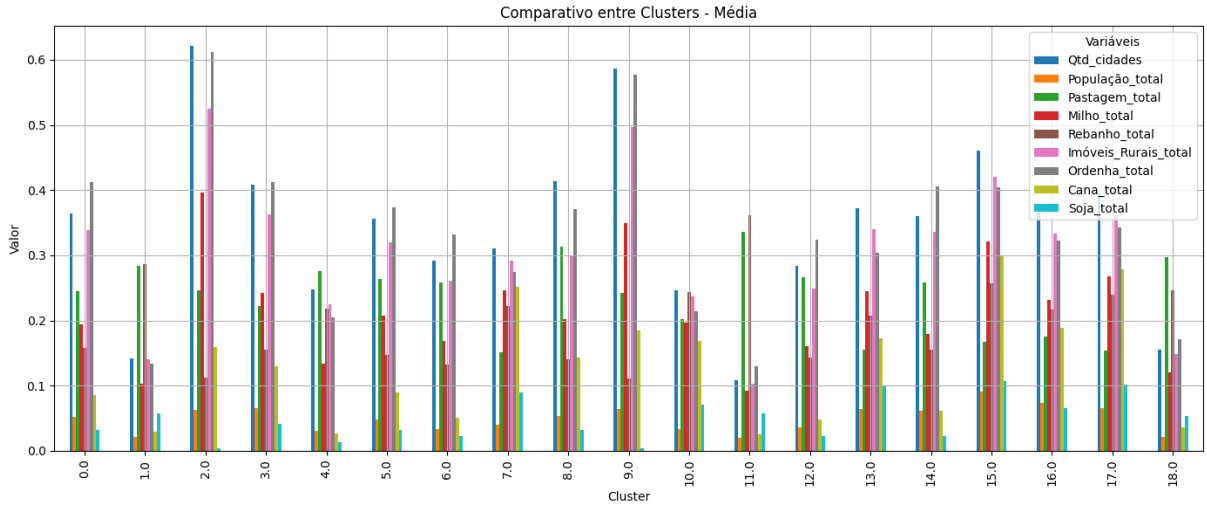

Figura 3 – Média normalizada das variáveis de cada cluster

Esse tipo de análise ajuda a entender melhor cada grupo. Por exemplo, os clusters que apresentam valores médios mais altos em imóveis rurais e rebanho costumam indicar regiões com uma estrutura produtiva focada na pecuária extensiva. Por outro lado, os clusters que mostram médias elevadas em produção agrícola, como milho e soja, apontam para áreas com uma forte presença da agricultura. Complementando essa análise, foi construído um heatmap com os mesmos indicadores, com o objetivo de facilitar a comparação relativa entre os clusters.

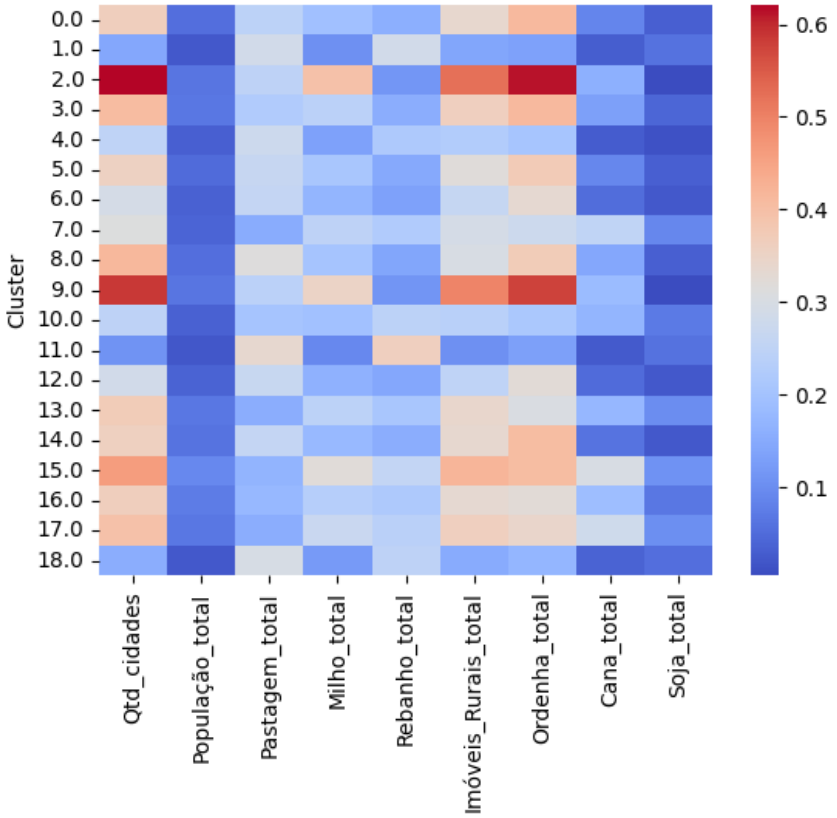

Figura 4 – Heatmap de cada cluster

Com a etapa de clusterização concluída, foi possível identificar diferentes perfis de municípios com base em suas características agropecuárias e populacionais. Essa segmentação contribui para uma compreensão mais estratégica do território, permitindo reconhecer padrões regionais, potencial produtivo e níveis distintos de influência entre os municípios classificados como oportunidade. Essa abordagem também será fundamental na próxima fase do projeto, que busca prever o desempenho comercial das futuras unidades, assim como ajuda a compreender como os diferentes perfis territoriais podem influenciar diretamente nos resultados de faturamento.

4. Análise e previsão de vendas

Com a base de vendas consolidada, iniciou-se a análise das lojas com relação ao seu cluster e a construção de modelos preditivos com o objetivo de estimar o faturamento mensal durante os doze primeiros meses de operação em cada cidade. Para isso, foram utilizadas apenas lojas inauguradas entre 2009 e 2019, excluindo-se unidades mais antigas que já estavam maduras comercialmente e poderiam distorcer o padrão de comportamento inicial.

4.1 Análise exploratória e comportamental das vendas

A primeira análise realizada foi do ticket médio por tipo de produto, que consiste na divisão entre o faturamento total por tipo de produto e a quantidade total de vendas para cada tipo de produto. Esse cálculo foi feito de forma geral, considerando todas as vendas registradas no conjunto de lojas da base de dados, sem segmentação por unidade federativa (UF) ou por cluster. O objetivo foi identificar quais categorias de produtos, independentemente da localização, apresentaram maior ticket médio.

A tabela mostra os resultados agrupados por código de tipo de produto e revela uma variação considerável entre os itens. Alguns apresentam tickets médios mais altos, sugerindo compras de maior valor por transação, enquanto outros têm valores mais baixos, o que pode estar ligado a produtos mais acessíveis ou comprados com maior frequência.

Tabela 3 – Ticket médio por tipo de produto

Tipo de produto	**Ticket Médio**
110	R$ 313,11
106	R$ 186,06
103	R$ 127,07
104	R$ 87,10
119	R$ 69,48
114	R$ 47,18
111	R$ 47,07
109	R$ 46,73
117	R$ 27,10
105	R$ 26,05
100	R$ 12,19
116	R$ 11,58
115	R$ 6,56
102	R$ 5,50
120	R$ 4,35
113	R$ 3,67
118	R$ 0,85

Para entender melhor a relação entre o perfil produtivo dos municípios e o faturamento por tipo de produto, foi realizada uma análise de correlação entre as variáveis agropecuárias e o faturamento de cada código de produto, utilizando como exemplo o Cluster 11. Esse grupo foi escolhido para exemplificar a análise principalmente, por contar com quatro lojas em operação. A matriz de correlação mostra valores entre -1 e 1, onde, cada célula da matriz representa o grau de correlação entre uma variável territorial e o

faturamento de um determinado tipo de produto. Valores positivos indicam que, à medida que determinada característica do município aumenta, também cresce o faturamento daquele produto. Já valores negativos sugerem uma relação inversa. Quanto mais próximo de 1 ou -1, mais forte é essa associação linear.

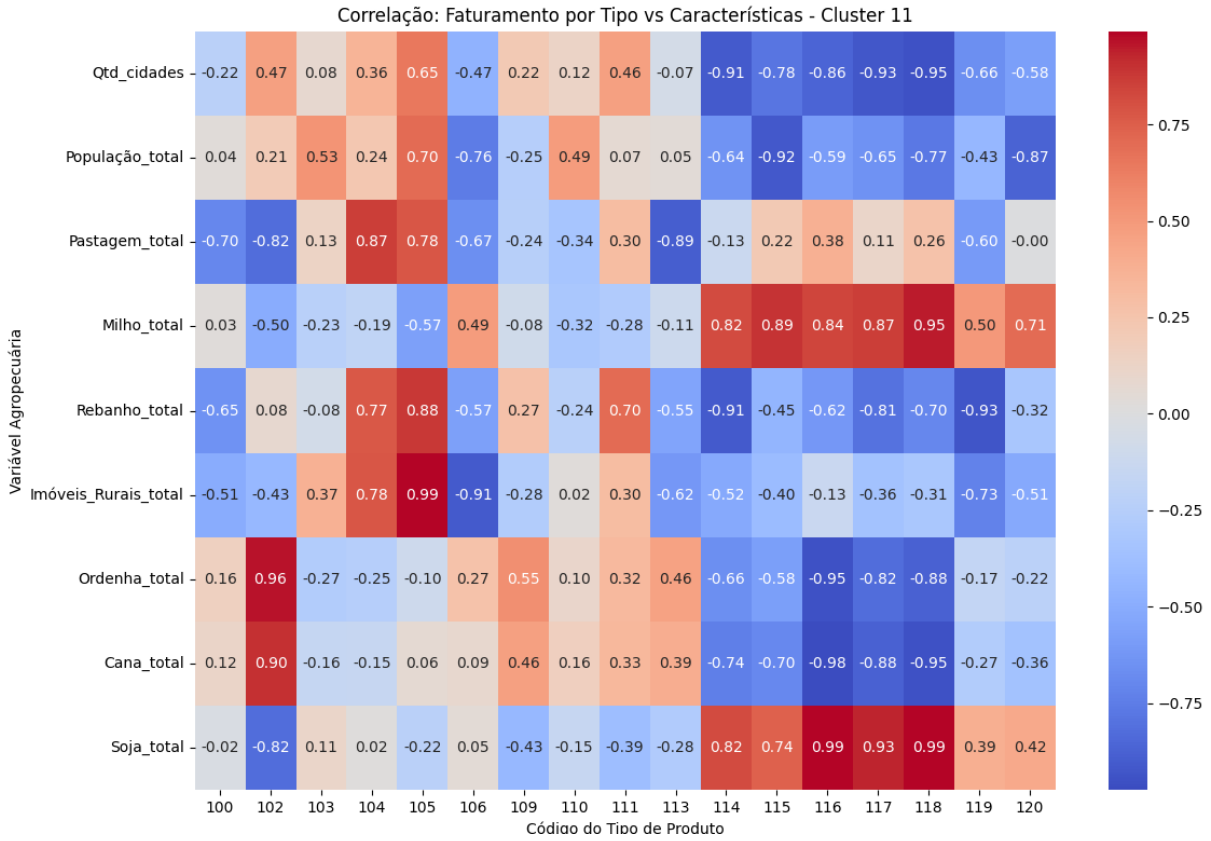

Figura 5 – Correlação entre faturamento e características - Cluster 11

Ao analisar os resultados, é possível notar dois grupos distintos de produtos, cada um com comportamento associado a diferentes características do cluster. O primeiro grupo é formado pelos produtos de código 114, 115, 116, 117 e 118, que apresentaram alta correlação com a produção agrícola, especialmente milho e soja. Isso indica que esses produtos tendem a ter um desempenho melhor em cidades onde essas culturas estão mais presentes. Já o segundo grupo, representado pelos produtos 104 e 105, mostra correlações mais elevadas com variáveis relacionadas à estrutura da pecuária extensiva, como o número de imóveis rurais, o tamanho do rebanho bovino e a área de pastagem. Essas correlações ajudam a entender quais produtos são mais relevantes em cada contexto, o que é essencial para decisões sobre provisão de mercadorias, posicionamento de lojas e estratégias de previsão de faturamento com base nas características locais.

O gráfico a seguir apresenta o faturamento total de cada tipo de produto nos clusters 1, 11, 13 e 18, permitindo visualizar de forma clara quais produtos são mais representativos em cada grupo de municípios. Essa visualização complementa a análise anterior ao evidenciar, na prática, como o contexto produtivo de cada cluster influencia a composição das vendas. É possível observar que certos tipos de produto aparecem com destaque em clusters mais ligados à pecuária, enquanto outros se sobressaem em regiões com perfil agrícola.

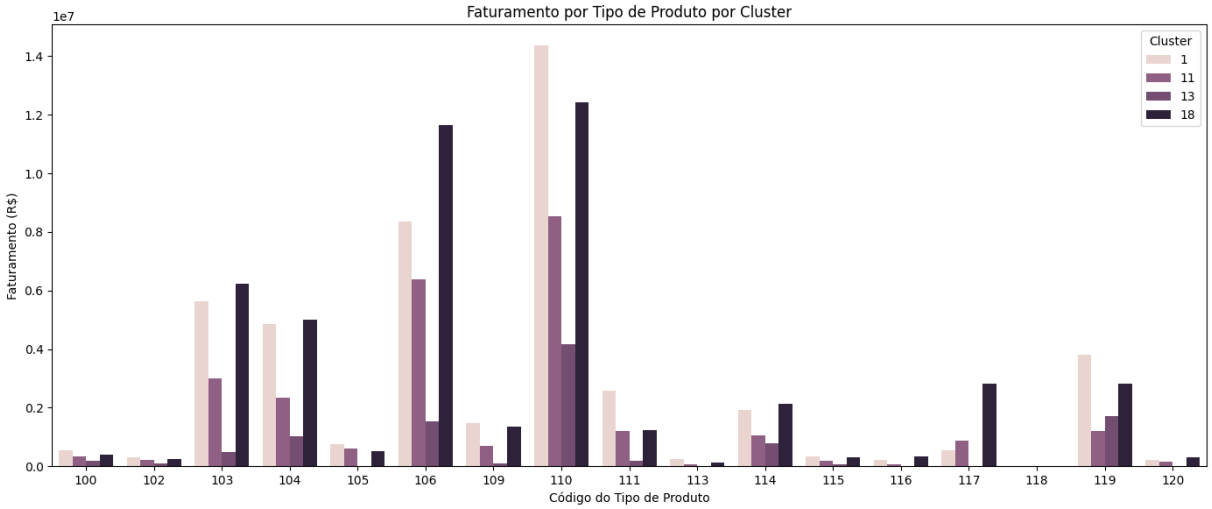

Figura 6 – Faturamento por tipo de produto x cluster

4.2 Modelagem preditiva do faturamento mensal

A partir das análises anteriores, foi desenvolvida uma modelagem preditiva com o objetivo de estimar o faturamento mensal durante o primeiro ano de operação de novas lojas. A proposta é oferecer uma ferramenta que ajude a projetar o desempenho comercial em cidades ainda não atendidas, com base em características semelhantes às daquelas onde a rede já atua. Para isso, o modelo utiliza variáveis estruturais do município e informações temporais relacionadas às vendas, mostrando a evolução da loja ao longo dos meses.

A base de dados utilizada no treinamento dos modelos foi construída a partir de vendas registradas entre 2009 e 2019. Foram consideradas apenas as lojas que possuíam registros completos desde o primeiro mês de funcionamento. Unidades mais antigas, foram deixadas de fora para evitar distorções, já que seu padrão de faturamento tende a ser diferente do padrão encontrado no início da operação por estarem em um estágio de maior

maturidade.

Com esse filtro, foram selecionadas 16 lojas, com registros mensais de faturamento, correspondentes ao seu primeiro ano de atividade. Para facilitar a análise, os dados foram organizados com base no tempo desde a abertura — ou seja, do mês 0 ao mês 11 — criando uma estrutura padronizada para todas as lojas, independente de qual mês ou ano a loja foi aberta.

Além do faturamento, cada linha da base foi enriquecida com as variáveis de população, número de imóveis rurais, área de pastagem, rebanho bovino e produção de milho, soja, cana e ordenha. No total, a base contou com 192 registros.

4.2.1 Modelos testados e erro

Para estimar o faturamento mensal durante o primeiro ano de operação das lojas, foram testados diferentes modelos de regressão. O objetivo era identificar qual algoritmo apresentaria melhor desempenho preditivo, ou seja, menor erro e maior capacidade de explicar a variação nos dados. Foram avaliados cinco métodos distintos: média móvel simples, regressão linear, Random Forest Regressor, Support Vector Regression (SVR) e rede neural do tipo MLP (Multi-layer Perceptron). Cada um deles apresentou resultados diferentes para R^2 e RMSE, como pode ser observado na Tabela 4.

Tabela 4 – Desempenho dos modelos preditivos testados

Modelo	**RMSE médio (R$)**	**R^2 médio**
Média móvel	212.971,42	0,3708
Random Forest Regressor	280.277,21	0,1989
Regressão Linear	290.807,37	–0,0543
SVR (Support Vector Regression)	322.494,50	–0,0500
MLP (Rede Neural)	701.629,19	–3,9744

Como mostra a Tabela 4, a média móvel se destacou como o modelo com melhor desempenho geral entre os testados, apesar de sua simplicidade, esse método interpreta bem as tendências em contextos com dados limitados ou com alta variabilidade (HYNDMAN; ATHANASOPOULOS, 2021). Ela apresentou o menor erro médio (RMSE) e o maior coeficiente de determinação (R^2), mostrando que, para os dados atuais, basear-se apenas no histórico de crescimento das lojas existentes foi mais eficaz. Isso sugere que, mesmo com a base de dados enriquecida, talvez ela ainda não capture todos os fatores que explicam a variação de faturamento entre os municípios, ou que o número de observações ainda é

limitado para modelos mais complexos. Entre os algoritmos supervisionados, o Random Forest Regressor foi o que mais se sobressaiu. Ele teve o menor RMSE e foi o único modelo desse grupo com um R^2 positivo, indicando que conseguiu explicar parte da variação do faturamento com base nas variáveis independentes. Por isso, os dois algoritmos foram utilizados para uma análise mais detalhada. Esses resultados podem ser visualizados nas Figuras 7 e 8, que comparam os valores reais e previstos por ambos os modelos em uma das lojas, selecionada de maneira aleatória na base de dados. Na Figura 7, é possível observar o comportamento da média móvel, que suaviza as variações mês a mês e consegue seguir razoavelmente bem a tendência geral do faturamento real. A previsibilidade mais estável desse modelo contribui para seu bom desempenho médio, conforme indicado pelas métricas de erro.

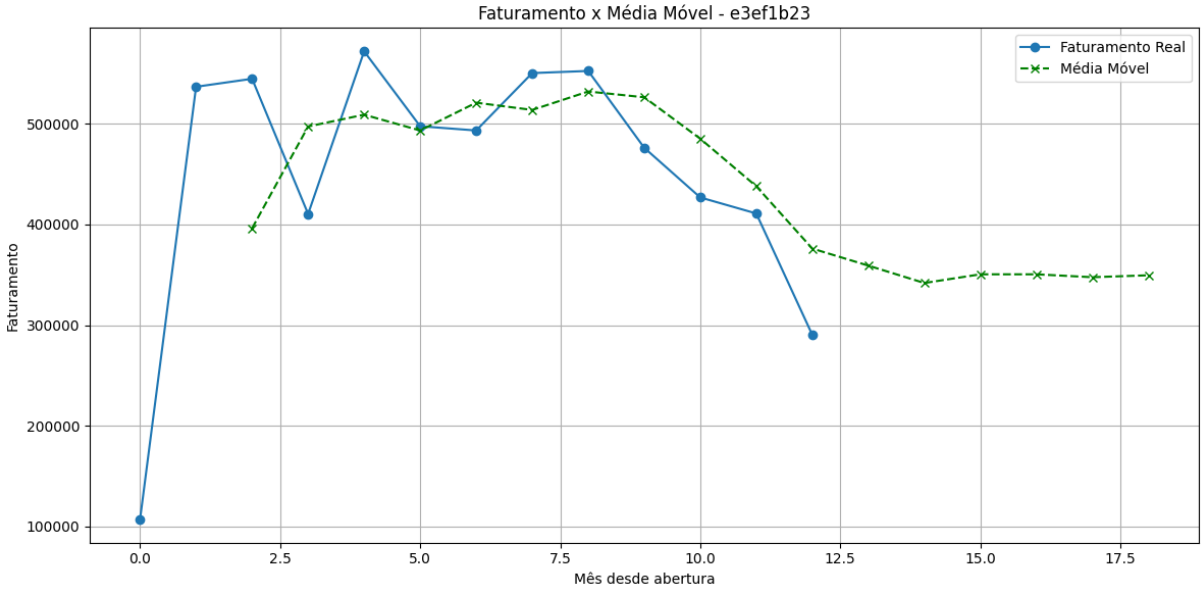

Figura 7 – Faturamento real x Média móvel

Já na Figura 8, temos o desempenho da Random Forest Regressor. Esse modelo funciona por meio da combinação de várias árvores de decisão e, neste caso, foram utilizadas 100 árvores. Embora o modelo capture bem a trajetória de crescimento nos primeiros meses e consiga antecipar certos picos de faturamento, ele tende a superestimar os valores nos meses finais. Esse comportamento ajuda a entender por que o modelo, mesmo sendo o melhor entre os supervisionados, ainda apresentou desempenho inferior à média móvel no conjunto geral.

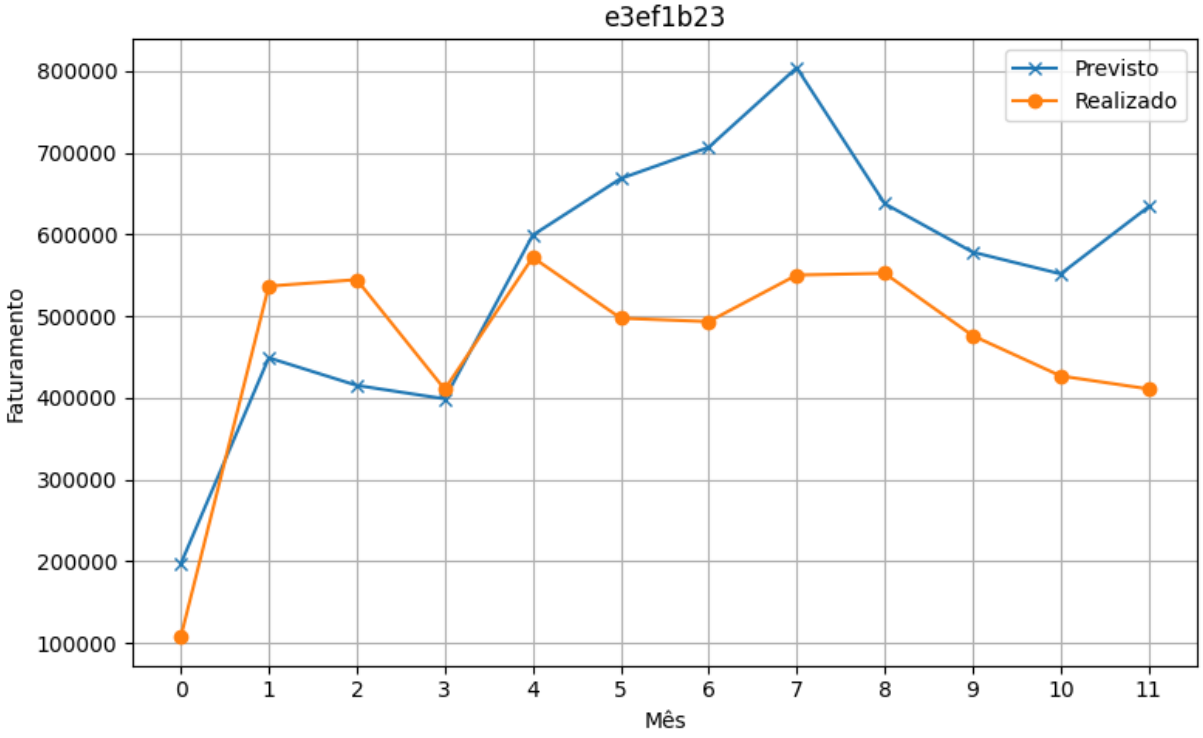

Figura 8 – Faturamento real x RandomForestRegressor

4.2.2 Análise de similaridade entre municípios para previsão indireta

Além dos modelos de regressão, foi aplicado ao final do processo um algoritmo do tipo K-Nearest Neighbors (KNN), com o objetivo de identificar quais cidades ainda não atendidas são mais semelhantes, em termos de perfil agropecuário e populacional, às cidades onde a empresa já possui lojas. O algoritmo calcula a distância euclidiana entre as variáveis independentes e, para cada cidade nova, aponta qual loja existente possui o perfil mais próximo. Para essa comparação, foram utilizadas as seguintes variáveis: população, número de imóveis rurais, área de pastagem (ha), rebanho bovino, e a produção de milho (ha), soja (ha), cana (ha) e leite (ordenha).

Esse tipo de comparação é útil principalmente para projeções rápidas de faturamento em municípios sem histórico de vendas, permitindo associar cada nova cidade à loja mais parecida e herdar seu desempenho como referência inicial. Embora o KNN não seja um modelo preditivo, ele funciona como um suporte estratégico à tomada de decisão, oferecendo um ponto de partida confiável para estimativas quando não se deseja ou não se pode aplicar modelos mais complexos. Para análise do resultado, foi observado os valores

das distribuição das distâncias e se organiza da seguinte maneira:

- **0 a 3** — Muito semelhantes (alta similaridade)
- **3 a 7** — Semelhantes (perfil próximo)
- **7 a 12.3** — Moderadamente semelhantes (diferenças perceptíveis)
- **12.3 a 37.5** — Pouco semelhantes (perfil distante)
- **Acima de 37.5** — Muito diferentes (sem similaridade relevante)

Essa escala foi definida utilizando os quartis, média e análise dos outliers. Os quartis dividem a distribuição em quatro partes iguais e ajudam a identificar os limites onde se concentram a maioria dos dados, já a média é usada como referência de tendência central, validando que distâncias até 15 ainda representam certa proximidade e por fim, os outliers que possuem valor muito acima do restante.

A distribuição da distância entre as cidades pelo KNN mostra que, em média, as cidades sem loja apresentam uma distância de 12.3 em relação à cidade mais parecida com loja. A mediana é um pouco menor, cerca de 7.0, o que indica que a maior parte das cidades analisadas possui uma correspondência razoavelmente próxima com algum município atendido. No entanto, os percentis mais altos chamam atenção: 5% das cidades apresentam distância superior a 37,5, e em casos extremos, esse valor chega a 248,1.

Esses valores elevados sugerem a presença de outliers — ou seja, cidades com características muito diferentes de todas as demais da base de lojas existentes. Uma análise adicional foi realizada especificamente sobre essas cidades mais distantes, e de fato foram identificados casos com perfis pouco similares ao perfil das cidades com loja.

Por essa razão, o tratamento desses outliers será considerado em trabalhos futuros, em conjunto com a área de negócios da empresa, que, com sua expertise e conhecimento do mercado, poderá ajudar a refinar os critérios que definem as características das cidades consideradas como oportunidade no início do estudo. Esse cuidado é essencial para garantir que as projeções feitas com base em cidades comparáveis sejam realmente confiáveis e representem de forma mais precisa o potencial comercial de cada novo município analisado.

5. DISCUSSÃO

Apesar dos resultados alcançados, o estudo apresenta limitações importantes que impactam nos modelos preditivos. A base de dados utilizada conta com apenas 16 lojas, o que restringe a diversidade de cenários observados e compromete a capacidade dos modelos de generalizar para contextos distintos. Essa limitação aumenta o risco de overfitting, onde o modelo se ajusta excessivamente aos dados disponíveis, aprendendo padrões específicos demais e com baixo poder de generalização para novos casos. Com uma base mais ampla, incluindo lojas de diferentes portes e regiões, é provável que os algoritmos testados apresentem desempenho superior, com previsões mais confiáveis ao longo do tempo.

Além disso, embora a base contemple o número de imóveis rurais por município, não foi possível estimar a taxa de penetração da marca em cada região, ou seja, a proporção de propriedades rurais efetivamente atendidas pelas unidades comerciais existentes. Essa métrica seria fundamental para avaliar o quanto do mercado potencial está sendo alcançado e em que medida há espaço real para novas lojas em determinadas áreas. Da mesma forma, o estudo não considerou possíveis sobreposições territoriais entre unidades já em operação, o que pode gerar efeitos de concorrência interna e prejudicar o desempenho de novos pontos de venda — uma dinâmica conhecida como canibalização comercial, em que o crescimento de uma loja ocorre às custas da queda de desempenho de outra localizada na mesma região de influência. Tais aspectos são relevantes para definição do plano de expansão, pois impactam diretamente na eficiência do capital investido, no retorno sobre novas unidades e no crescimento da rede no médio prazo.

Os resultados obtidos ao longo deste estudo demonstram que a aplicação de técnicas de inteligência artificial, especialmente a combinação de clusterização hierárquica, regressão supervisionada e análise de similaridade via K-Nearest Neighbors, é promissora como suporte à tomada de decisão na expansão de lojas agropecuárias. O agrupamento dos municípios brasileiros revelou perfis produtivos e populacionais distintos, permitindo uma leitura mais estratégica do território nacional. Essa abordagem contribui diretamente para o reconhecimento de regiões com maior potencial de retorno, mesmo na ausência de histórico comercial. Na prática, isso oferece à gestão uma base concreta para priorização territorial, definição de metas e mitigação de incertezas em mercados ainda inexplorados.

6. CONCLUSÃO

Este trabalho teve como objetivo apoiar a expansão estratégica de uma rede de lojas voltadas ao setor agropecuário, por meio da análise de dados territoriais, produtivos e comerciais. A abordagem adotada combinou técnicas de agrupamento, análise comportamental de vendas e modelagem preditiva para estimar o faturamento mensal esperado no primeiro ano de operação em novos municípios.

A clusterização dos municípios brasileiros permitiu identificar grupos com características agropecuárias e populacionais semelhantes, o que ajudou a direcionar a análise comercial para diferentes perfis territoriais. A análise exploratória de vendas mostrou forte relação entre as variáveis do território e o tipo de produto comercializado, reforçando a importância do contexto local no desempenho das lojas.

Na etapa preditiva, cinco modelos foram testados para estimar o faturamento mensal: média móvel, regressão linear, SVR, MLP e Random Forest. A média móvel apresentou o melhor desempenho geral. Entre os modelos supervisionados, o Random Forest foi o que mais se destacou, os modelos de SVR, MLP e regressão linear apresentaram resultados ruins para a base de dados. Além disso, o algoritmo KNN foi utilizado para identificar cidades sem loja com perfil semelhante às já atendidas, funcionando como uma ferramenta complementar para estimativas rápidas. A análise também revelou a existência de outliers entre essas cidades, que serão tratados em estudos futuros com apoio da área de negócios. Os resultados reforçam o potencial do uso de dados estruturados na tomada de decisão comercial. Com ajustes e expansão da base, a metodologia proposta pode ser aprimorada para apoiar a escolha de novos pontos de venda de forma mais precisa e eficiente.

Como trabalhos futuros, propõe-se o retreinamento do modelo com uma base de dados mais ampla, incluindo um número maior de lojas com histórico completo de vendas, o que poderá aumentar sua robustez e viabilizar a aplicação de algoritmos mais robustos. Também se destaca a importância de considerar fatores ainda não abordados, como a taxa de penetração da marca em cada região e os efeitos de sobreposição territorial entre unidades, que podem gerar concorrência interna e afetar negativamente o desempenho das lojas — fenômeno conhecido como canibalização comercial. Além disso, sugere-se o aprofundamento na análise das condições do solo e do calendário agrícola dos municípios, incorporando informações sobre sazonalidade, safras e entressafras, que influenciam diretamente o comportamento de compra do público agropecuário. A integração desses elementos pode

contribuir para a construção de modelos mais precisos e alinhados à realidade do setor.

Referências

[1] ASSIS, Ricardo; GEOUÊG, L. *As categorias territoriais e o campo brasileiro*. In: SILVEIRA, M. L. (Org.). **O Campo no Século XXI: território e territorialidades**. São Paulo: Contexto, 2006. p. 128–148.

[2] BRASIL. Ministério da Agricultura e Pecuária. **Projeções do agronegócio: Brasil 2022/23 a 2032/33 – Projeções de longo prazo**. Brasília: Mapa, 2023. Disponível em: `https://www.gov.br/agricultura`. Acesso em: 20 jun. 2025.

[3] CALIL, Leonardo Aparecido de Almeida et al. **Mineração de dados e pós-processamento em padrões descobertos**. *Exatas Terra e Ciências Agrárias*, Ponta Grossa, 2008 . Disponível em: `https://revistas.uepg.br/index.php/exatas/article/view/946`. Acesso em: 20 jun. 2025.

[4] HAN, Jiawei; Kamber, Micheline; Pei, Jian. **Data Mining: Concepts and Techniques**. 3. ed. Amsterdam: Elsevier, 2011.

[5] HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

[6] IPEA – Instituto de Pesquisa Econômica Aplicada. VIEIRA FILHO, J. E. R. **O agronegócio brasileiro: a contribuição do IPEA nos debates**. *Boletim Regional, Urbano e Ambiental*, n. 30, jul.–dez. 2023. Disponível em: `https://www.ipea.gov.br`. Acesso em: 20 jun. 2025.

[7] MAPBIOMAS. **Coleção 9 da Série Anual de Mapas de Uso e Cobertura da Terra do Brasil**. Disponível em: `https://mapbiomas.org/`. Acesso em: 20 jun. 2025.

[8] IBGE – Instituto Brasileiro de Geografia e Estatística. **SIDRA – Sistema IBGE de Recuperação Automática**. Disponível em: `https://sidra.ibge.gov.br/`. Acesso em: 20 jun. 2025.

[9] CEPEA. **Indicadores do Agronegócio Brasileiro 2023**. Piracicaba: ESALQ/USP, 2023.

[10] CONAB. **Séries Históricas da Produção Agrícola Brasileira: Safra 2023/24**. Brasília: Companhia Nacional de Abastecimento, 2023.

[11] McKINSEY & COMPANY. **Global AI Survey: State of AI in 2023**. New York, 2023. Disponível em: `https://www.mckinsey.com`. Acesso em: 20 jun. 2025.

[12] PwC. **Global Retail and Consumer Insights Survey 2024**. London, 2024.

[13] WARD, J. H. **Hierarchical Grouping to Optimize an Objective Function**. *Journal of the American Statistical Association*, v. 58, n. 301, p. 236–244, 1963.

[14] BREIMAN, L. **Random Forests**. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

[15] HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. 3. ed. Melbourne: OTexts, 2021. Disponível em: `https://otexts.com/fpp3/`. Acesso em: 20 jun. 2025.

[16] IBGE. **Sistema de Contas Regionais 2024**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2024. Disponível em: `https://www.ibge.gov.br`. Acesso em: 20 jun. 2025.

[17] CHAI, T.; DRAXLER, R. R. *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments*. Environmental Modelling and Software, v. 35, p. 92–95, 2014.

[18] CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, v. 7, p. e623, 2021. DOI: `https://doi.org/10.7717/peerj-cs.623`.

[19] LEWIS-BECK, M. S.; SKALABAN, T. J. *Applied Regression: An Introduction*. 2. ed. Thousand Oaks: Sage Publications, 1990.

[20] ROBUSTO, C. C. The Cosine-Haversine Formula. *American Mathematical Monthly*, v. 64, n. 1, p. 38–40, 1957.