UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Prevendo a quantidade de postos de atendimento cooperativo no Brasil

Amanda Pereira Ferraz

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Amanda Pereira Ferraz

Prevendo a quantidade de postos de atendimento cooperativo no Brasil

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Pereira Ferraz, Amanda
P436p Prevendo a quantidade de postos de atendimento cooperativo no Brasil / Amanda Pereira Ferraz; orientador Ricardo Rodrigues Ciferri. -- São Carlos, 2025.
38 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. . I. Rodrigues Ciferri, Ricardo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Amanda Pereira Ferraz

Predicting the number of cooperative service stations in Brazil

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Ricardo Rodrigues Ciferri

Original version

São Carlos

2025

RESUMO

Este trabalho tem como objetivo predizer a quantidade de Postos de Atendimento Cooperativo (PACs) em municípios brasileiros, utilizando modelos de aprendizado de máquina e variáveis socioeconômicas e demográficas. A pesquisa busca apoiar o planejamento de expansão geográfica das cooperativas de crédito. A contextualização apresenta a evolução do cooperativismo de crédito no Brasil e sua relevância para a inclusão financeira e o desenvolvimento regional, destacando o papel do Sistema Nacional de Crédito Cooperativo (SNCC) e o crescimento das cooperativas frente à retração das agências bancárias tradicionais. Foram utilizados os microdados do IBGE e do Banco Central do Brasil (BCB) referentes a 2021. As variáveis analisadas foram: região, estado, população estimada, área, densidade demográfica, PIB per capita e faixa do PIB per capita. Dos 5.570 municípios brasileiros, 2.857 possuíam pelo menos um PAC. Foram construídos e comparados modelos supervisionados: Árvore de Decisão (AD), *Random Forest* (RF), Gradiente *Boosting* (GB) e *XGBoost* (XGB). As métricas utilizadas foram R^2, MAE, RMSE e MAPE. O modelo Gradient Boosting apresentou o melhor desempenho, com $R^2 = 0{,}73$, RMSE = 1,74, MAE = 0,94 e MAPE = 0,43%, indicando boa capacidade preditiva considerando o volume de dados disponível. Os resultados mostram que a população estimada foi a variável com maior correlação positiva com a quantidade de PACs, seguida da densidade demográfica, enquanto o PIB per capita apresentou pouca influência direta. O estudo demonstra a viabilidade do uso de modelos de machine learning para prever a rede de atendimento das cooperativas, contribuindo com o planejamento estratégico do setor financeiro cooperativo e com políticas de inclusão bancária. Recomenda-se incluir, em pesquisas futuras, variáveis adicionais como PIB do governo, PEA, IDH, renda média, número de empresas e indicadores de crédito, além de aplicar técnicas de clusterização para refinar as análises regionais.

Palavras-chave: Árvore de Decisão, Gradiente *Boosting*, *XGBoost*, Cooperativismo, Aprendizado de Máquina, *Random Forest*

ABSTRACT

This study aims to predict the number of Cooperative Service Points (PACs) in Brazilian municipalities using machine learning models and socioeconomic and demographic variables. The research seeks to support the geographic expansion planning of credit cooperatives. The contextualization presents the evolution of credit cooperativism in Brazil and its relevance for financial inclusion and regional development, highlighting the role of the National Credit Cooperative System (SNCC) and the growth of cooperatives compared to the reduction of traditional bank branches. Microdata from the Brazilian Institute of Geography and Statistics (IBGE) and the Central Bank of Brazil (BCB) for the year 2021 were used. The variables analyzed were: region, state, estimated population, area, population density, GDP per capita, and GDP per capita range. Among the 5,570 Brazilian municipalities, 2,857 had at least one PAC. Supervised models were built and compared: Decision Tree (AD), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB). The metrics used were R^2, MAE, RMSE, and MAPE. The Gradient Boosting model showed the best performance, with $R^2 = 0.73$, RMSE = 1.74, MAE = 0.94, and MAPE = 0.43%, indicating good predictive capacity given the available data volume. The results show that the estimated population was the variable with the highest positive correlation with the number of PACs, followed by population density, while GDP per capita had little direct influence. The study demonstrates the feasibility of using machine learning models to predict the service network of cooperatives, contributing to the strategic planning of the cooperative financial sector and to banking inclusion policies. It is recommended that future research include additional variables such as government GDP, economically active population (PEA), Human Development Index (HDI), average income, number of companies, and credit indicators, as well as apply clustering techniques to refine regional analyses.

Keywords: Decision Tree, Gradient Boosting, XGBoost, Cooperativism, Machine Learning, Random Forest

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCB	Banco Central do Brasil
BPC	Benefício de Prestação Continuada
CCS	Centro Cooperativo Sicoob
FIPE	Fundação Instituto de Pesquisas Econômicas
IBGE	Instituto Brasileiro de Geografia e Estatística
ICA	*International Cooperative Alliance*
IDH	Índice de Desenvolvimento Humano
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
MSE	Erro Quadrático Médio
PA	Posto de Atendimento
PAB	Posto de Atendimento Bancário
PAC	Posto de Atendimento Cooperativo
PCA	Análise de Componentes Principais
PEA	População Economicamente Ativa
R^2	Coeficiente de Determinação
RMSE	Raiz do Erro Quadrático Médio
SFN	Sistema Financeiro Nacional
SICOOB	Sistema de Cooperativas de Crédito do Brasil
SISBR	Sistema de Informática do Sicoob
SNCC	Sistema Nacional de Crédito Cooperativo
SVM	Máquina de Vetores de Suporte

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

As primeiras experiências com o cooperativismo de crédito foram empreendidas na Alemanha no século XIX. Em 1852 foi criada a primeira cooperativa de crédito urbana, já as primeiras cooperativas de crédito rural foram criadas na década de 1860. (FIPE, 2019)

No Brasil, as cooperativas de crédito podem se organizar em sistemas de segundo ou terceiro nível, central e confederação, respectivamente. Atualmente são 4 Sistemas de Cooperativas de Crédito de Terceiro Nível (Confederação): SICOOB, Sicredi, UNICRED e CRESOL; 4 Sistemas de Cooperativas de Crédito de Segundo Nível (Centrais): CECOOP, AILOS, CrediSIS e Uniprime; além de cooperativas singulares de crédito não filiadas a centrais.

Esses sistemas propiciam economia de escala sob uma estrutura piramidal, as cooperativas singulares (primeiro nível) ocupam a base, têm o objetivo de prestar serviços direto ao associado e são constituídas por um mínimo de vinte cooperados; as cooperativas centrais (segundo nível) ocupam a zona intermediária, têm o objetivo de organizar, em maior escala, os serviços das filiadas, facilitando a utilização recíproca de serviços e são constituídas por, no mínimo, três cooperativas singulares; e por fim, a confederação (terceiro nível) fica no topo, possui personalidade jurídica própria e reúne no mínimo três centrais, com o objetivo de defender seus interesses, promover a padronização, supervisão e integração operacional, financeira, normativa e tecnológica.

O Sistema Nacional de Crédito Cooperativo (SNCC) possui, ainda, um quarto elemento: os bancos cooperativos, que devem ter controle acionário de cooperativas centrais de crédito, e fornecem produtos e serviços financeiros especialmente para os membros do sistema, tais como poupança e fundo de investimento. No Brasil, existem dois bancos cooperativos: o Banco Cooperativo Sicredi S/A fundado em 16 de outubro de 1995 e o Banco Cooperativo do Brasil S/A fundado em 21 de julho de 1997.

O Sistema Financeiro Nacional (SFN, 1988) é formado por um conjunto de instituições que promovem a intermediação financeira conectando credores e tomadores de recursos. O sistema financeiro possibilita que pessoas, empresas e governo circulem a maior parte dos ativos, paguem as dívidas e realizem os investimentos.

De acordo com pesquisa do Instituto Locomotiva (MEIRELLES, 2021), realizada em Janeiro/2021, ainda existem 34 milhões de brasileiros com acesso precário ao sistema bancário - o equivalente a 21% da população que movimenta aproximadamente R$ 347 bilhões ao ano. A prtir desses dados é possíve perceber que o Brasil é um país que apresenta uma grande quantidade de Instituições Financeiras em funcionamento e apesar

desse número ser bastante expressivo, ainda existe uma grande parcela desse mercado a ser explorada, o que contribui para a expansão geográfica das Cooperativas de Crédito.

Atualmente, muitas das transações bancárias podem ser realizadas por meio de aplicativos móveis, sendo possível resolver praticamente tudo que envolve os bancos por meio do Internet Banking, entretanto ainda existe uma parcela significativa da população brasileira que é resistente ao uso de aplicativos para movimentação da conta, como é demonstrado na Global Mobile Consumer Survey 2016 (DELOITTE, 2016). Pensando nesse público, é de interesse do Sistema Cooperativo ter unidades físicas próximas a esses potenciais clientes. Assim, seria útil verificar a cobertura existente no país em termos de agências físicas para atendimento e, a partir disso, verificar se variáveis sociais, demográficas e econômicas seria possível prever a quantidade de postos de atendimento cooperativo em cada município do país, ou seja, prever a rede de atendimento da região considerada.

Segundo dados do Banco Central (BCB, 2022a), as unidades das cooperativas de crédito praticamente dobraram em quantidade. As cooperativas de crédito têm expandido o seu atendimento presencial e por não serem bancos, não há agência, o que existem são os postos de atendimento cooperativo (PAC), em 2022 existiam 8.593 PACs, um aumento de 99,3% no número postos de atendimento cooperativo em relação a Março/2015 quando existiam 4.281 unidades. A maior rede de PACs é do Sicoob, 4.018 unidades. Desde 2020, o Sicoob abriu 690 postos de atendimento cooperativo.

O Banco Central iniciou o projeto do *Open Banking* em 2020, ampliando-o posteriormente para Open Finance, para mostrar a sua maior abrangência, que inclui não somente informações sobre produtos e serviços financeiros mais tradicionais (como contas e operações de crédito), mas também dados de produtos e serviços relacionados a investimentos, e, posteriormente, os referentes a câmbio, credenciamento, seguros e previdência (BCB, 2022b). O *Open Banking* é um conjunto de regras e tecnologias que vai permitir o compartilhamento de dados e serviços de clientes entre instituições financeiras por meio da integração de seus respectivos sistemas. O *Open Banking* trará uma maior liberdade para os usuários do sistema bancário, pois os clientes terão a possibilidade de permitir o compartilhamento de suas informações de produtos e serviços contratados entre diferentes instituições autorizadas pelo Banco Central e de movimentar as contas bancárias por meio de diferentes plataformas de forma segura, ágil e conveniente. O *Open Banking* trará também mais competição, pois as instituições participantes poderão fazer ofertas de produtos e serviços para os clientes de seus concorrentes, com benefícios para o consumidor, que poderá obter tarifas mais baixas e condições mais vantajosas e também melhorará a experiência quanto ao uso de produtos e serviços financeiros, pois possibilitará que as instituições participantes ofereçam soluções que facilitam o controle financeiro dos clientes, visto que consolidam as informações em um único local.

Segundo o presidente do Banco Central (CAMPOS, 2022), a expansão das coopera-

tivas de crédito possibilita o desenvolvimento das comunidades em que elas estão inseridas, pois impacta positivamente diversas variáveis, como renda, emprego e empreendedorismo, e também impulsiona a inclusão e a educação financeira, além de alcançar localidades remotas com menor ou nenhuma presença do Sistema Financeiro Nacional, buscando diminuir a desigualdade geográfica existente, visto que muitas vezes por inviabilidade econômica na avaliação das instituições financeiras alguns municípios brasileiros continuam desprovidos de atendimento bancário presencial e, nesse cenário, as cooperativas de crédito surgem como uma excelente alternativa, pois assumem os riscos em prol da comunidade, promovendo o desenvolvimento local por meio da poupança e do microcrédito.

Em Janeiro/2023, o Sicoob alcançou a marca de 7 milhões de cooperados, um crescimento de 18% na comparação com Janeiro/2022 (VALOR, 2023). Dos clientes conquistados ao longo dos 12 últimos meses, 839 mil são pessoas físicas e 209 mil pessoas jurídicas. O crescimento da base de cooperados reforça o papel do Sicoob como propulsor de um movimento importante que tem como premissa possibilitar o acesso de todo brasileiro a uma instituição financeira.

Pensando no projeto de expansão e abertura de novos PACs pelo Sicoob, este trabalho se mostra relevante, pois se variáveis sociais, demográficas e econômicas permitirem predizer a quantidade de postos de atendimento cooperativo do municípios, em municípios onde não há PACs seria possível saber a quantidade de postos que seria necessário considerando aspectos sociais, demográficos e econômicos.

1.2 Objetivos

O objetivo geral é predizer, a partir de modelos de aprendizado de máquina e variáveis sociais, demográficas e econômicas a quantidade de postos de atendimento cooperativo de cada município.

1.2.1 Objetivos Específicos

Os objetivos específicos deste estudo englobam a realização das seguintes tarefas:

- Escolher os modelos de aprendizado de máquina mais adequados para a atividade de prever a quantidade de postos de atendimento cooperativo em cada município;

- Comparar a performance dos modelos, utilizando a linguagem de programação Python, com o auxílio das medidas de avaliação da capacidade preditiva, como erro percentual absoluto médio - MAPE, erro quadrático médio - MSE, raiz do erro quadrático médio - RMSE, erro absoluto médio - MAE e coeficiente de determinação - R^2;

1.3 Contribuições

Este trabalho contribui ao revelar as possíveis localidades para abertura de novos PACs, considerando variáveis sociais, demográficas e econômicas. A possibilidade de predizer a quantidade de PACs em municípios em que eles ainda não existem é interessante do ponto de vista do sétimo princípio do cooperativismo mundial (ICA, 1995), que demonstra o interesse pela comunidade, esse princípio cooperativista se expressa por meio da atuação sem fins lucrativos e, mais do que isso, pela atuação orientada à geração de benefícios sociais e econômicos para toda a sociedade.

1.4 Organização do trabalho

O capítulo 2 apresenta os Fundamentos Teóricos deste trabalho, incluindo a descrição dos modelos de Aprendizado de Máquina: Árvore de Decisão, *Random Forest*, Gradiente *Boosting* e *XGBoost*, são discutidas também diversas métricas de avaliação dos modelos de AM.

No capítulo 3, é apresentada a metodologia usada no desenvolvimento.

Os resultados são apresentados e discutidos no capítulo 4.

Finalmente, destacam-se as principais contribuições, conclusões, limitações e sugestões para trabalhos futuros no capítulo 5.

Os códigos-fonte desenvolvidos ao longo deste trabalho são apresentados no Anexo.

2 FUNDAMENTAÇÃO TEÓRICA

O referencial teórico deste trabalho abordará os seguintes tópicos: Aprendizado de Máquina, Aprendizado Supervisionado, Árvore de Decisão, *Random Forest* e Métricas para avaliação de modelos. (Faceli *et al.*, 2022)

O Aprendizado de Máquina é um método de análise de dados que automatiza a construção de modelos analíticos, sendo um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

O aprendizado supervisionado utiliza conjuntos de dados rotulados para treinar algoritmos que classificam dados ou preveem resultados. Os dados de entrada são apresentados ao modelo e os pesos são ajustados até que o esse modelo apresente bons valores para as métricas de avaliação do erro, com o objetivo de garantir que não haja *overfitting* nem *underfitting*.

2.1 Técnicas de Aprendizado de Máquina

A árvore de decisão estabelece nós (decision nodes) que se relacionam entre si por meio de uma hierarquia. Existe o nó-raiz (root node), que é o mais importante, e os nós-folha (leaf nodes), que são os resultados finais. No contexto de Aprendizado de Máquina, o nó-raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que será gerado como resposta. Assim, um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia na medida em se avança em novos níveis de profundidade. As soluções dos subproblemas podem ser combinadas na forma de uma árvore para produzir uma solução do problema complexo. A força dessa proposta vem da capacidade de dividir o espaço de instâncias em subespaços e cada subespaço é ajustado usando diferentes modelos. Para definir os posicionamentos, é preciso calcular a entropia das classes de saída e o ganho de informação dos atributos da base de dados, assim, aquele que tiver maior ganho de informação entre os atributos será o nó-raiz.

A árvore de decisão é um modelo não-paramétrico (Géron, 2021), visto que o número de parâmetros do modelo não é determinado antes do treinamento, ou seja, não assume qualquer distribuição para os dados.

Para evitar o sobreajuste dos dados de treinamento é necessário limitar a liberdade da árvore de decisão durante esse processo, normalmente restringindo a profundidade máxima da árvore de decisão (max_depth), pois sintetizar esse parâmetro diminui o risco de sobreajuste, uma vez que regulariza o modelo.

Outros parâmetros também conseguem restringir de maneira semelhante a forma da árvore de decisão, o min_samples_split é o número mínimo de amostras que um nó deve ter, antes de poder ser dividido; o min_samples_leaf é o número mínimo que um nó da folha deve ter e só pode ser expresso como uma fração do número total de instâncias ponderadas; o max_features é o número máximo de características avaliadas para cada divisão em um nó.

A busca deve ser por um modelo que generalize melhor, visto que modelos sobreajustados, de modo geral, funcionam bem para conjuntos de teste específicos.

O valor predito para cada região é sempre o valor-alvo médio das instâncias nessa região. O algoritmo divide cada região de uma maneira que faz com que boa parte das instâncias de treinamento fique o mais próximo possível desse valor.

As árvores de decisão apresentam diversas vantagens, versatilidade, flexibilidade, robustez, interpretabilidade e eficiência, pois a árvore de decisão é um método invariante a transformações estritamente monótonas das variáveis de entrada, apresentando pouca sensibilidade a *outliers* e as decisões são baseadas nos valores dos atributos utilizados para descrever o problema (nó-raiz e nós-folha). No entanto, elas utilizam muito fronteiras de decisão ortogonais, ou seja, todas as divisões são perpendiculares a um eixo, o que as tornam sensíveis à rotatividade (pequenas variações) do conjunto de treinamento.

Uma forma de determinar esse problema das fronteiras nas árvores de decisão é usar a análise de componentes principais (PCA), que resulta em uma melhor orientação dos conjuntos de treinamento, além disso, outra possibilidade seria fazer uso dos modelos *Random Forest*, que por calcular a média de predição em relação a muitas árvores, consegue limitar essa instabilidade das fronteiras de decisão ortogonais.

O *Random Forest* é um algoritmo criado a partir da combinação de árvores de decisão e adiciona aleatoriedade extra ao modelo com o objetivo de reduzir ainda mais a correlação entre os modelos de base. Ao invés de procurar pela melhor característica ao fazer a partição de nós, ele busca a melhor variável em um subconjunto aleatório dos atributos. Este processo cria uma grande diversidade, o que normalmente leva a geração de modelos melhores. Assim, quando uma árvore é criada no modelo floresta aleatória, apenas um subconjunto aleatório das características é considerado na partição de um nó. É possível fazer as árvores ficarem mais aleatórias utilizando limiares (thresholds) aleatórios para cada atributo, ao invés de procurar pelo melhor limiar (como uma árvore de decisão geralmente faz).

O modelo *Random Forest Regressor* surge da agregação das predições de um grupo de preditores regressores (árvores de decisão) e, de maneira geral, um método *ensemble* obterá melhores predições que do que o melhor com o preditor individual. A despeito de sua simplicidade, o *Random Forest* é um dos algoritmos de aprendizado de máquina mais

poderosos disponíveis atualmente. Por ser um modelo criado a partir da combinação de árvores de decisão, o *Random Forest* tem a vantagem de dificilmente apresentar *overfitting*.

O modelo *Random Forest* introduz aleatoriedade extra ao cultivar árvores, em vez de procurar a melhor característica quando divide um nó, ele busca o melhor aspecto entre um subconjunto de predicados, com isso o algoritmo resulta em uma maior diversidade de árvores, que pode trocar um viés maior por uma variância menor, produzindo um modelo geral melhor.

O *Random Forest* também possibilita que seja medida a importância relativa de cada uma das características, a importância de cada um desses atributos é calculada analisando o quanto os nós das árvores que utilizam esses aspectos reduzem em média a impureza em todas as árvores do modelo, é uma média ponderada em que o peso de cada nó é igual ao número de amostras de treinamento associadas a ele e a soma de todas as importâncias é igual a 1.

Os métodos *ensemble* funcionam melhor quando os preditores são tão independentes dos outros quanto possível.

2.2 Métricas de Avaliação de Modelos de Aprendizado de Máquina

Em relação às métricas de avaliação do modelo, tem-se que o erro absoluto médio (MAE) não é afetado por *outliers* e como a métrica é apresentada na mesma escala dos dados utilizados na previsão, a interpretação do seu valor é mais fácil.

Uma característica da raiz do erro quadrático médio (RMSE) é que como os erros (diferença entre os valores reais e as predições do modelo) são elevados ao quadrado antes de a média ser calculada, os pesos atribuídos à soma são diferentes e, conforme os valores dos erros das instâncias aumentam, o valor do RMSE aumenta consideravelmente, ou seja, se houver um *outlier* no conjunto de dados, seu valor terá maior peso para o cálculo do RMSE e isso deixará a métrica superestimada, o que evidencia o fato de que o erro quadrático médio (MSE) aumenta na medida em que as previsões feitas pelo modelo pioram. Apesar disso, o RMSE tem a vantagem de penalizar os erros de maior magnitude e como a métrica é apresentada na mesma escala do dado original, isso resulta em uma melhor interpretabilidade do seu resultado, o que não ocorre com a métrica MSE que, por ser calculada elevada ao quadrado, fica distorcida, pois não é apresentada na mesma escala que o dado original. Para pequenas variâncias, tem-se uma pequena diferença entre as métricas com MAE $<$ RMSE, uma vez que, no mundo real, o RMSE sempre resultará em um valor maior que o MAE.

O erro percentual absoluto médio (MAPE) exprime uma porcentagem obtida por meio da divisão da diferença entre o valor predito e o valor real pelo valor real. Como o MAPE divide o erro absoluto pelos dados reais, os valores próximos a zero acabam por

superestimá-lo. Assim como o MSE e o MAE, quanto menor o seu valor, mais preciso é o modelo de regressão. Os *outliers* têm menos efeito sobre o MAPE do que sobre o MSE.

2.3 Trabalhos Correlatos

O artigo (Mohd; Masrom; Johari, 2019) revela que, para prever o preço de moradias na Malásia, modelos mais simples como *Random Forest Regressor* e *Decision Tree Regressor* foram os que apresentaram os melhores resultados considerando o RMSE.

O artigo (Ho; Tang; Wong, 2021) mostra que para prever o preço de moradias em Hong Kong os algoritmos de Aprendizado de Máquina, SVM, *Random Forest* e Gradiente *Boosting*, são adequados e, considerando o poder preditivo e as métricas de avaliação dos modelos, o *Random Forest* e o Gradiente *Boosting* obtiveram um desempenho melhor, entretanto o SVM também se revela um algoritmo útil no ajuste de dados porque pode produzir previsões razoavelmente precisas mesmo havendo uma restrição de tempo.

O artigo (Erraissi; Banane, 2021) elenca o MSE, o RMSE, o MAE e o MAPE como as principais medidas para avaliar a qualidade do modelo de regressão.

Esses trabalhos mostram que pode não ser viável desconsiderar os modelos mais tradicionais porque eles podem apresentar um bom resultado para o caso em questão e, além disso, possibilitam que seja compreendido o passo-a-passo na construção dos modelos.

O artigo (Fernandes; Dantas; Porfírio, 2019) revela que a existência de municípios brasileiros sem atendimento cooperativo é uma realidade e que apesar de ter sido encontrada correlação positiva entre o fato de o possuir cooperativa e o tamanho do PIB per capita, bem como o tamanho da população, não é possível afirmar que o surgimento de novos postos de atendimento seja diretamente influenciado por essas variáveis.

3 METODOLOGIA

Foram utilizados os microdados disponibilizados nos sites do Instituto do Intituto Brasileiro de Geografia e Estatística (IBGE, 1936) e do Banco Central do Brasil (BCB, 1965).

As informações do Banco Central disponíveis para serem utilizadas na construção dos modelos de Aprendizado de Máquina são as seguintes:

Tabela 2 – Variáveis disponíveis para elaboração dos modelos de predição de quantidade de PACs

Variável Explicativa	**Definição da Variável Explicativa**
Região	Região a que pertence o município
Estado	Estado a que pertence o município
População Estimada	Quantidade de habitantes do município
Tamanho do município (área)	Superfície do território do município em km^2
Densidade Demográfica	Relação entre a população e o território
PIB per capita	Renda nacional dividida pela população
Faixa do PIB per capita	7 faixas publicadas anualmente pelo IBGE

Tabela 3 – Estratificação do PIB per capita para o ano de 2021 (valores em reais)

Faixa	**PIB per capita (R$)**
1	5.408 a 14.000
2	14.000 a 20.000
3	20.000 a 30.000
4	30.000 a 42.247
5	42.247 a 55.000
6	55.000 a 100.000
7	100.000 a 920.834

As faixas do PIB per capita são definidas de forma que 3 faixas fiquem abaixo da média nacional e 3 faixas fiquem acima da média nacional, de maneira que a faixa intermediária contenha a média nacional.

O conjunto de dados foi criado tendo o município como chave primária para todos os cruzamentos que foram realizados.

Esse conjunto de variáveis será utilizado para criar os modelos de aprendizado de máquina para verificar se eles possuem a capacidade de predizer, de forma adequada, a quantidade de postos de atendimento cooperativo.

3.1 Preparação dos Dados

O conjunto de dados foi construído no SAS Enterprise Guide da seguinte forma:

Foram consideradas as informações dos municípios: região, estado, PIB per capita, faixa do PIB per capita, população, área e densidade demográfica.

Foram considerados somente os municípios que tinham pelo menos um PAC em Dezembro/2021.

A base de dados apresenta a quantidade de PACs como última variável (variável resposta).

Dos 5.570 municípios existentes em 2021, 2.713 não tinham nenhum posto de atendimento cooperativo.

O conjunto de dados apresenta 2.857 registros e como foi utilizada a validação cruzada, o conjunto de dados foi dividido em grupos aleatórios, mantendo um grupo como teste e treinando o modelo nos grupos restantes. Este processo foi repetido para cada grupo considerado como grupo de teste, então a média dos modelos é utilizada para o modelo resultante.

Os registros apresentam a seguinte divisão:

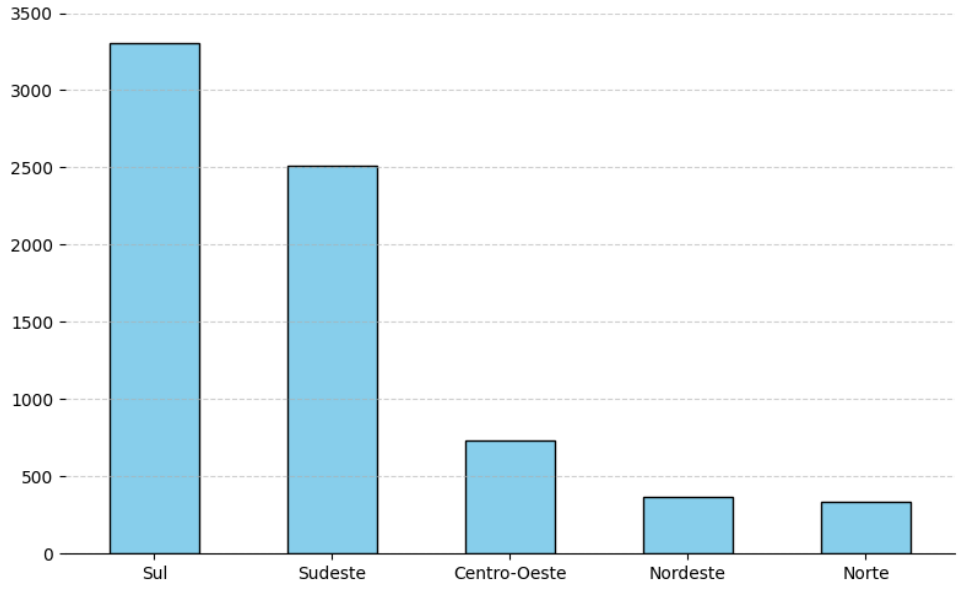

Figura 1 – Quantidade de Postos de Atendimento Cooperativo por Região do Município em que estão localizados

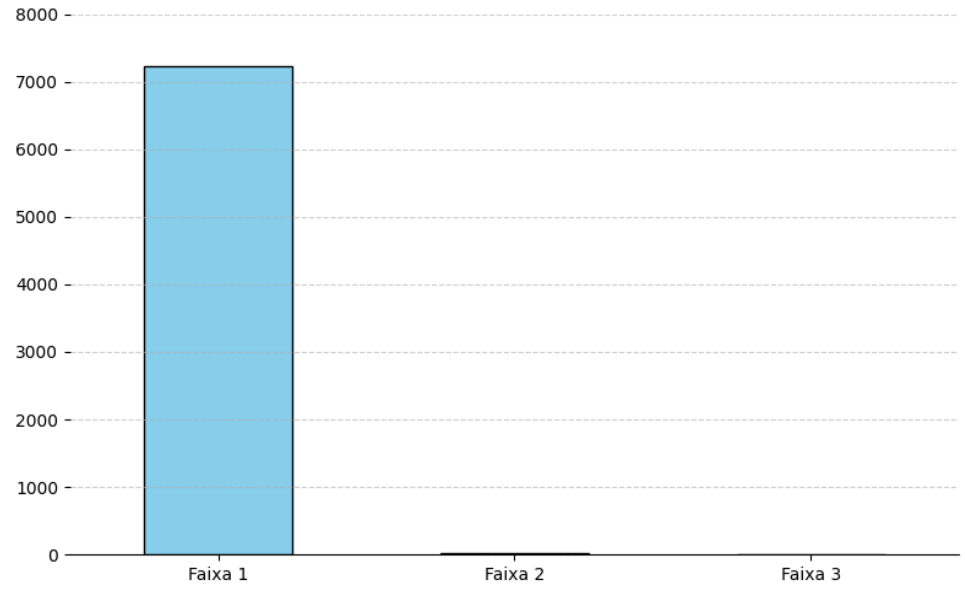

Figura 2 – Quantidade de Postos de Atendimento Cooperativo por faixa do PIB per capita do município em que estão localizados

A partir dessas informações é possível perceber que os postos de atendimento cooperativo, existentes em Dezembro/2021, em sua maioria, estão localizados nas regiões Sul e Sudeste do país, em municípios com um menor PIB per capita, faixa 1.

3.2 Análise Descritiva dos Dados

Considerando a quantidade de postos de atendimento cooperativo, tem-se as seguintes correlações das variáveis explicativas em relação à variável resposta:

Tabela 4 – Correlação das Variáveis Explicativas em relação à quantidade de PACs

Variável Explicativa	**Correlação com a quantidade de PACs**
Região	0,04
População Estimada	0,64
Densidade Demográfica	0,33
PIB per capita	0,06
Faixa do PIB per capita	-0,01

Analisando a tabela 4, é possível perceber que a variável população estimada tem uma correlação moderada positiva com a quantidade de PACs existentes no município, ou seja, municípios mais populosos tendem a ter mais postos de atendimento cooperativo, o que faz muito sentido e que a variável densidade demográfica tem uma correlação fraca positiva com a quantidade de PACs existentes no município, ou seja, locais mais densos também tendem a ter mais postos de atendimento cooperativo, mas essa relação é bem menor. Já as demais variáveis explicativas têm uma correlação praticamente irrelevante

com a variável resposta, visto que o Coeficiente de Correlação de Pearson foi inferior a 0,10 para esses casos. Sabe-se, no entanto, que o coeficiente de Correlação de Pearson pode não ser muito bom em identificar relações não lineares entre variáveis e não pode isoladamente ser usado para descartar possível relação.(Prado, 2020)

Segundo dados do Banco Central do Brasil (BCB, 2007), os municípios que são o foco da análise, os 2.857 municípios, entre os 5.570 municípios brasileiros, estão distribuídos da seguinte forma:

Tabela 5 – Distribuição dos Municípios em que existe pelo menos um PAC por Região

Região	**Quantidade de Municípios com pelo menos um PAC**
Sul	1.128
Sudeste	1.062
Centro-Oeste	323
Nordeste	210
Norte	134
Total	**2.857**

A tabela 5 mostra que o Sul e Sudeste são as regiões com a maior quantidade de municípios com pelo menos um PAC.

Tabela 6 – Distribuição dos municípios em que existe pelo menos um PAC por PIB per capita

PIB per capita do município	**Quantidade**
Até R$ 14 mil	2.842
Entre R$ 14 mil e R$ 20 mil	13
Entre R$ 20 mil e R$ 30 mil	2
Entre R$ 30 mil e R$ 42.247 mil	0
Entre R$ 42.247 mil e R$ 55 mil	0
Entre R$ 55 mil e R$ 100 mil	0
Acima de R$ 100 mil	0
Total	**2.857**

A tabela 6 revela que, quase 99,5% dos municípios que têm pelo menos 1 PAC apresenta um PIB per capita de até R$ 14 mill reais, ou seja, são em sua maioria municípios que estão na fixa mais baixa do PIB per capita.

A população está distribuída da seguinte forma:

Tabela 7 – Distribuição dos Municípios em que existe pelo menos um PAC por Região

Região	**Quantidade de Municípios com pelo menos um PAC**
Sul	1.128
Sudeste	1.062
Centro-Oeste	323
Nordeste	210
Norte	134
Total	**2.857**

A tabela 7 mostra que quase a população dos municípios com pelo menos um PAC é quase de 170 milhões (IBGE, 2021).

3.3 Análise de Desempenho e Seleção dos Modelos de Aprendizado de Máquina

A análise dos dados, construção e avaliação dos modelos de Aprendizado de Máquina foram feitas utilizando a linguagem Python.

A atividade de construção e seleção dos modelos foi feita com o conjunto de dados descrito neste capítulo. Foram construídos os seguintes modelos de predição para a quantidade de PACs nos municípios:

Quatro modelos, sendo um de árvore de decisão, um de *Random Forest*, um de *XGBoost* e outro de Gradiente *Boosting*, que consideram como variáveis independentes a região, a faixa do PIB per capita, a população e a densidade demográfica.

As métricas de avaliação MAE, MSE, RMSE, MAPE e R^2 foram utilizadas para avaliar os modelos, assim, tem-se que:

- A métrica R^2, conhecida coeficiente de determinação, representa o percentual da variância dos dados que é explicado pelo modelo. Quanto maior é o valor de R^2, mais explicativo é o modelo em relação aos dados previstos.

- O erro médio absoluto (MAE) mede a média da diferença entre o valor real com o predito, considerando o módulo da diferença.

- O erro percentual absoluto médio (MAPE) é uma métrica que mostra a porcentagem de erro em relação aos valores reais.

- O erro quadrático médio (MSE) é uma métrica que calcula a média de diferença entre o valor predito e o valor real, considerando a diferença elevada ao quadrado. Quanto maior é o valor de MSE, pior é a performance do modelo em relação às previsões.

- A raiz do erro quadrático médio (RMSE) é a raiz quadrática do erro quadrático médio (MSE), como a unidade fica na mesma escala que o dado original, isso permite que o resultado da métrica seja interpretado facilmente.

Os modelos construídos foram testados em diversos cenários e seus resultados analisados. Estes cenários e resultados são discutidos no capítulo seguinte.

4 AVALIAÇÃO EXPERIMENTAL

Os modelos descritos na definição da metodologia foram testados nas estimativas da quantidade de PACs. Esses modelos estão detalhados na tabela abaixo.

Tabela 8 – Descrição dos Modelos de Aprendizado de Máquina Testados

Nome do Modelo	**Técnica de Aprendizado de Máquina**	**Variáveis Independentes**
RF	Random Forest	região, faixa do PIB per capita, população e densidade demográfica
AD	Árvore de Decisão	região, faixa do PIB per capita, população e densidade demográfica
XGB	*XGBoost*	região, faixa do PIB per capita, população e densidade demográfica
GB	Gradiente *Boosting*	região, faixa do PIB per capita, população e densidade demográfica

Apresentam-se na tabela 9 as métricas do conjunto de teste para os modelos GB, RF, XGB e AD.

Tabela 9 – Desempenho dos modelos GB, RF, XGB e AD na predição da quantidade de PACs

Modelo	R^2	**RMSE**	**MAPE**	**MAE**	**MSE**
Gradiente Boosting	0,73	1,737385	0,434064	0,941208	3,018507
RF	0,68	1,871971	0,440599	0,991738	3,504276
XGBoost	0,62	2,050164	0,564555	1,147580	4,203174
Árvore de Decisão	0,25	2,889274	0,488352	1,239510	8,347902

Os modelos criados foram testados no conjunto de teste tal como descrito no capítulo 3. Na tabela 9, apresenta-se o desempenho dos modelos GB, RF, XGB e AD respectivamente.

Em relação ao modelo GB para a predição da quantidade de PACs, que considera as variáveis explicativas região, faixa do PIB per capita, população e densidade demográfica, tem-se que ele apresentou um erro percentual absoluto médio de aproximadamente 0,43%, com um erro absoluto médio de quase 1 PAC e com a raiz do erro quadrático médio de quase 2 PACs.

Em relação ao modelo RF para a predição da quantidade de PACs, que considera as variáveis explicativas região, faixa do PIB per capita, população e densidade demográfica,

tem-se que ele apresentou um erro percentual absoluto médio de aproximadamente 0,44%, com um erro absoluto médio de quase 1 PAC e com a raiz do erro quadrático médio de quase 2 PACs.

Em relação ao modelo XGB para a predição da quantidade de PACs, que considera as variáveis explicativas região, faixa do PIB per capita, população e densidade demográfica, tem-se que ele apresentou um erro percentual absoluto médio de aproximadamente 0,56%, com um erro absoluto médio de pouco mais de 1 PAC e com a raiz do erro quadrático médio de pouco mais de 2 PACs.

Em relação ao modelo AD para a predição da quantidade de PACs, que considera as variáveis explicativas região, faixa do PIB per capita, população e densidade demográfica, tem-se que ele apresentou um erro percentual absoluto médio de aproximadamente 0,49%, com um erro absoluto médio de pouco mais de 1 PAC e com a raiz do erro quadrático médio de quase 3 PACs.

Considerando as métricas de avaliação tem-se que o Gradiente Boosting foi o modelo que apresentou o melhor desempenho entre todos.

5 CONCLUSÕES

Com os dados utilizados foi possível construir modelos que tiveram ter um bom desempenho para o conjunto de dados em questão. Os modelos testados tiverem desempenho dentro da espectativa, mas o baixo volume de dados provavelmente impediu que os modelos apresentassem um resultado superior. Apesar disso, o problema de prever a quantidade de PAC para os municípios brasileiros é extremamente importante para todo o sistema cooperativo financeiro, visto que o modelo econômico cooperativo representa competitividade e inclusão, com comprovado retorno para a comunidade.

Um aspecto que poderia tornar os modelos mais precisos seria o uso de outras informações demográficas, sociais e populacionais, tais como: PIB da Administração Pública (PIB do Governo), Bolsa Família, Benefício de Prestação Continuada (BPC), população economicamente ativa (PEA), pessoas jurídicas por município, índice de desenvolvimento humano (IDH) do município (total, de longevidade, de educação), renda média do trabalhador, quantidade de leitos hospitalares, classificação de trabalhadores formais com ensino superior, volumes financeiros de operações de crédito e depósitos por município e classificação total da frota de automóveis.

Tendo em vista a disparidade existente entre as regiões do Brasil, entre os municípios e dentro do próprio município, seria importante estratificar os dados, clusterizando com base em determinados perfis, para assim estudar as pequenas áreas e partes de um mesmo município.

REFERÊNCIAS

BCB: Estabilidade Financeira. BANCO CENTRAL DO BRASIL, 1965. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/relacao_instituicoes_funcionamento. Acesso em: 10/01/2023.

BCB: Relação de Agências e Postos de Atendimento das Instituições Financeiras. BANCO CENTRAL DO BRASIL, 2007. Disponível em: https://www.bcb.gov.br/fis/info/agencias.asp?frame=1. Acesso em: 11/10/2022.

BCB: Brasil perde 5,8 mil Agências Bancárias em 7 anos; Cooperativas Dobram. VALOR ECONÔMICO, 2022. Disponível em: https://valor.globo.com/financas/noticia/2022/09/21/brasil-perde-58-mil-agncias-bancrias-em-7-anos-cooperativas-dobram.ghtml. Acesso em: 06/12/2022.

BCB: Open Banking. BANCO CENTRAL DO BRASIL, 2022. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/openbanking. Acesso em: 07/12/2022.

CAMPOS: Cooperativas de Crédito crescem mais que o Sistema Financeiro como um todo, diz Presidente do BC, Roberto Campos Neto. VALOR ECONÔMICO, 2022. Disponível em: https://valorinveste.globo.com/mercados/brasil-e-politica/noticia/2022/08/30/cooperativas-de-credito-crescem-mais-que-o-sistema-financeiro-como-um-todo-diz-presidente-do-bc.ghtml. Acesso em: 08/12/2022.

DELOITTE: 2016 Global Mobile Consumer Survey: US Edition. DELOITTE, 2016. Disponível em: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-global-mobile-consumer-survey-2016-executive-summary.pdf. Acesso em: 05/12/2022.

ERRAISSI, A.; BANANE, M. Machine learning model to predict the number of cases contaminated by covid-19. **International Journal of Computing and Digital Systems**, v. 10, n. 1, p. 991–1001, 2021. Disponível em: http://dx.doi.org/10.12785/ijcds/100189. Acesso em: 21 jan. 2023.

FACELI, K. *et al.* **Inteligência Artificial: uma abordagem de Aprendizado de Máquina**. 2nd. ed. Rio de Janeiro: LTC, 2022.

FERNANDES, B. V. R.; DANTAS josé A.; PORFíRIO, L. V. Um retrato do cooperativismo de crédito no brasil: Perfil dos municípios brasileiros em dezembro de 2017. **Revista de Gestão e Organizações Cooperativas**, v. 6, n. 12, p. 201–218, jul. 2019. Disponível em: https://periodicos.ufsm.br/rgc/article/download/35743/pdf. Acesso em: 22 sep. 2024.

FIPE: Benefícios econômicos do cooperativismo de crédito na economia brasileira - relatório técnico. FUNDAÇÃO INSTITUTO DE PESQUISAS ECONÔMICAS, 2019. Disponível em: https://www.sicredi.com.br/media/produtos/sicredi-beneficios-do-cooperativismo-de-credito.pdf. Acesso em: 02/12/2022.

GéRON, A. **Mãos a Obra: Aprendizado de Máquina com Scikit-Learn, Keras e TensorFlow - Atualizada com a TensorFlow 2: Conceitos, Ferramentas e**

Técnicas para a Construção de Sistemas Inteligentes. 2nd. ed. Rio de Janeiro: Alta Books, 2021.

HO, W. K.; TANG, B.-S.; WONG, S. W. Predicting property prices with machine learning algorithms. **Journal of Property Research**, v. 38, n. 1, p. 48–70, 2021. Disponível em: https://doi.org/10.1080/09599916.2020.1832558. Acesso em: 23 jan. 2023.

IBGE: Estatísticas Socias de População. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 1936. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao.html. Acesso em: 10/01/2023.

IBGE: Estimativas da População Residente no Brasil. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2021. Disponível em: https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2021/estimativa_dou_2021.pdf. Acesso em: 10/10/2022.

ICA: Cooperative Identify, Values & Principles. INTERNACIONAL COOPERATIVE ALLIANCE, 1995. Disponível em: https://www.ica.coop/en/cooperatives/cooperative-identity. Acesso em: 30/11/2022.

MEIRELLES: 34 Milhões de Brasileiros não têm acesso a Serviços Bancários. INSTITUTO LOCOMOTIVA, 2021. Disponível em: https://ilocomotiva.com.br/clipping/labs-news-34-milhoes-de-brasileiros-nao-tem-acesso-a-servicos-bancarios/. Acesso em: 04/12/2022.

MOHD, T.; MASROM, S.; JOHARI, N. Machine learning housing price prediction in petaling jaya, selangor, malaysia. **International Journal of Recent Technology and Engineering**, v. 8, n. 2, p. 542–546, set. 2019. Disponível em: https://www.ijrte.org/wp-content/uploads/papers/v8i2S11/B10840982S1119.pdf. Acesso em: 22 jan. 2023.

PRADO, M. L. de. Statistical association (presentation slides). **SSRN**, 2020. Disponível em: https://ssrn.com/abstract=3512994. Acesso em: 03 may. 2024.

SFN: Sistema Financeiro Nacional. BANCO CENTRAL DO BRASIL, 1988. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/sfn. Acesso em: 03/12/2022.

VALOR: Sicoob alcança 7 milhões de cooperados, com expansão anual de 18%. VALOR ECONÔMICO, 2023. Disponível em: https://valor.globo.com/google/amp/financas/noticia/2023/01/26/sicoob-alcana-7-milhes-de-cooperados-com-expanso-anual-de-18-pontos-percentuais.ghtml. Acesso em: 26/01/2023.