Aplicação de mineração de dados para identificação de possíveis inadimplentes em uma cooperativa do ramo agrícola

Jaisson Duarte[1], **Edimar Manica**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul - Câmpus Ibirubá
Rua Nelsi Ribas Fritsch, 1111 – CEP: 98200-000 – Ibirubá – RS – Brasil

{jaisson.duarte, edimar.manica}@ibiruba.ifrs.edu.br

Abstract. *Non-payment is a problem that affects both consumers and businesses, compromising credit and generating financial losses. Recent studies reveal Brazil's worrying situation in this context. In addition, agricultural cooperatives also face challenges in relation to non-payment, as they deal with financial transactions and have a large number of customers, which makes credit evaluation difficult. In this paper, we describe the development of a classification model to identify clients with the highest risk of non-payment. The model combines the RandomizableFilteredClassifier and NaiveBayes algorithms, achieving a recall of 67%. Real data, provided by a cooperative, was used, including financial information, sales history, and grain marketing. The resulting model of the work is intended to assist credit analysts in prioritizing the clients who will be evaluated and thus reduce future non-payment.*

Resumo. *A inadimplência é um problema que afeta tanto os consumidores quanto as empresas, comprometendo o crédito e gerando prejuízos financeiros. Estudos recentes revelam a preocupante situação do Brasil nesse contexto. Além disso, as cooperativas do ramo agropecuário também enfrentam desafios em relação à inadimplência, uma vez que lidam com transações financeiras e têm um grande número de clientes, o que dificulta a avaliação de crédito. Neste artigo, é descrito o desenvolvimento de um modelo de classificação capaz de identificar os clientes com maiores chances de inadimplência. O modelo combina os algoritmos RandomizableFilteredClassifier e NaiveBayes, alcançando uma revocação de 67%. Dados reais, cedidos por uma cooperativa, foram utilizados incluindo informações financeiras, histórico de vendas e comercialização de grãos. O modelo resultante do trabalho tem como objetivo auxiliar os analistas de crédito na priorização dos clientes que deverão ser avaliados e assim e reduzir a inadimplência futura.*

1. Introdução

A inadimplência é um problema que afeta tanto os consumidores quanto as empresas, comprometendo o crédito e gerando prejuízos financeiros. Um estudo divulgado por [Serasa 2023], informou que o Brasil conta com 71,44 milhões de pessoas em situação de inadimplência. São pessoas que não conseguiram honrar seus compromissos financeiros assumidos com empresas em geral, comprometendo o seu próprio crédito e gerando prejuízos às empresas, que, por sua vez, muitas vezes não conseguem repassar os valores financeiros aos seus fornecedores. As cooperativas do ramo agropecuário

estão inseridas nesse cenário, pois também realizam movimentações financeiras através da comercialização de produtos e serviços aos seus associados e clientes, enfrentando as mesmas dificuldades das demais empresas.

Um levantamento feito pela empresa Serasa [Experian 2023], considerando as 27 unidades federativas do país, revelou a situação de inadimplência do produtor rural brasileiro em novembro de 2022. De acordo com os dados, 27% desses trabalhadores estavam negativados nesse período. Neste sentido, avaliar os clientes é fundamental para reduzir a inadimplência futura.

A cooperativa objeto deste trabalho possui mais de 130 mil clientes, com um crescimento médio de 4% ao ano. Atualmente, o setor financeiro avalia em média apenas 5 mil clientes por ano para permitir ou negar a concessão de crédito. Isso acaba gerando uma sobrecarga de trabalho, podendo levar a possíveis falhas humanas. No período de estudo deste trabalho, apenas 12,46% das negociações a prazo foram avaliadas e autorizadas pelo setor de crédito. Os títulos que não passaram pela avaliação geraram uma inadimplência de mais de 88 milhões de reais. Isso demonstra a importância de melhorar os processos de avaliação de crédito.

Nesse contexto, este trabalho desenvolveu um modelo de classificação com o objetivo de identificar os clientes com maiores chances de inadimplência. Por meio da análise de dados históricos e utilizando os algoritmos disponíveis na ferramenta *Weka*, foi possível determinar se um novo cliente deve ser submetido à análise de crédito ou não.

Para a realização dos experimentos, foram utilizados dados reais de uma cooperativa do ramo agrícola da região. Esses dados foram extraídos do sistema *ERP*[1] da cooperativa entre os anos de 2015 a 2020. Neles, estão contidas informações pessoais dos clientes da cooperativa, históricos de movimentações financeiras, vendas e comercialização de grãos. Para garantir a privacidade dos clientes, todos os dados foram anonimizados e formatados estatisticamente, preparados para serem utilizados nos algoritmos de classificação da ferramenta *WEKA*.

Após a análise dos dados, o melhor modelo foi obtido pela combinação de dois algoritmos: *RandomizableFilteredClassifier* e o algoritmo *NaiveBayes*. Esse modelo alcançou uma revocação de 67%.

Este artigo está organizado da seguinte forma. A Seção 2 descreve a fundamentação teórica. A Seção 3 apresenta a metodologia aplicada neste trabalho. Os resultados obtidos são apresentados e discutidos na Seção 4. Por fim, a Seção 5 traz as considerações finais e aponta possíveis trabalhos futuros.

2. Fundamentação Teórica

O processo *KDD - Knowledge Discovery in Databases* (processo de descoberta de conhecimento em base de dados, em tradução livre) é um processo proposto por [Fayyad 1996], e consiste em uma sequência iterativa de etapas. Segundo [Ferreira 2018], O processo KDD tem sido utilizado pelos tomadores de decisão na busca de informação relevante de difícil detecção por métodos tradicionais de análise. Além disto, é um processo não trivial e estruturado de identificação de padrões, com a finalidade de extrair informações

[1] *Enterprise Resource Planning* - Sistema integrado de gestão empresarial.

e conhecimentos potencialmente úteis e implicitamente contidos em bases de dados. Este processo é composto pelas seguintes etapas:

1. Seleção: etapa de preparação e seleção dos dados utilizados;
2. Pré-processamento: etapa de remoção ou atenuação de possíveis ruídos presentes nos dados selecionados;
3. Transformação: etapa em que são aplicados tratamentos e transformações sobre os dados para melhor adequá-los à extração de padrões;
4. Mineração de dados: busca e extração de padrões nos dados por meio de algoritmos;
5. Interpretação e avaliação: análise da relevância e refinamento do conhecimento descoberto para o domínio em questão.

Vários trabalhos publicados demonstram a aplicação bem-sucedida de técnicas de mineração de dados para lidar com problemas de inadimplência e análise de crédito em diferentes setores. Nesse âmbito há o trabalho realizado por [Reis 2022] onde os algoritmos, *Gradient Boosting* e *Random Forest*, obtiveram destaque para substituir o modelo fundamentalista de avaliação de risco de crédito realizado por analistas humanos. O estudo realizado por [dos Santos 2022] tinha como o objetivo identificar um modelo preditivo de classificação de risco de crédito em operações comerciais de cheque especial para pessoas físicas, onde os resultados do estudo indicaram um desempenho para o algoritmo *Random Forest*. No estudo realizado por [da Silva Ribeiro 2020], foi aplicado um algoritmo de classificação com o objetivo de classificar os clientes e assim permitir uma análise comportamental de crédito dos clientes. O propósito deste trabalho é o desenvolvimento de um modelo de classificação capaz de identificar os clientes de uma cooperativa do ramo agrícola com maiores chances de inadimplência.

3. Metodologia

Esta seção descreve a metodologia utilizada neste trabalho, que seguiu o processo de descoberta de conhecimento em base de dados proposto por [Fayyad 1996]. As próximas subseções descrevem como foi realizada cada etapa deste processo.

3.1. Seleção

A fase na qual é definida a fonte dos dados a serem analisados é a de seleção. Para este trabalho, juntamente com os analistas financeiros da cooperativa, foi construída uma base de dados com as informações relevantes à análise de crédito. Estas informações foram divididas em dois tipos: individuais e coletivas. As informações individuais referem-se exclusivamente a uma pessoa, tais como, informações pessoais, histórico de movimentação financeira, histórico da produção de grãos e histórico de comercialização. Enquanto que as informações coletivas referem-se aos dados da região do produtor como, por exemplo, se houve aumento na produção de soja, milho, trigo e outros grãos no ano corrente. Essas informações servem para identificar se houve seca ou uma queda na produção em determinado ano daquela região, fato que influencia no valor de diversos atributos individuais.

Neste trabalho, foram extraídos os dados a partir do banco de dados relacional do *ERP* da cooperativa. Foram obtidos os dados entre os anos de 2015 a 2020, por serem anos com dados já consolidados desde a troca do *ERP*, esses dados foram extraídos das tabelas de movimentação financeira, notas fiscais, movimentação de produção de grãos, cadastro de pessoas e tabelas adjacentes, resultando um total de 23 tabelas envolvidas.

Após selecionados os dados que deveriam ser analisados, estes foram exportados para um arquivo em formato *CSV* (*Comma-separated values* - dados separados por vírgula em tradução livre), num total de mais de 908 mil registros, representando mais de 118 mil clientes. A base criada é apresentada na Tabela 1, contendo 33 atributos individuais e 4 atributos coletivos.

Cada registro retrata as informações de um produtor ao final de cada ano de forma anonimizada, tendo as informações da produção entregue de grãos, as comercializações realizadas em cada negócio, as características do produtor, tais como, estado civil, associação, faixa etária e principalmente as movimentações financeiras.

Os dados foram rotulados de forma automática por meio de uma consulta *SQL* (*Structured Query Language* - Linguagem de Consulta Estruturada) que definiu o valor do atributo **inadimplente**, o qual representa a classe alvo da classificação. Foi atribuído o valor **SIM** caso o produtor tenha fechado o ano com alguma dívida vencida, e o valor **NÃO**, caso contrário.

Tabela 1. Base de dados

Tipo	Atributo	Descrição	Tipo de Dado
Individual	Inadimplente	Ficou inadimplente no ano	Boolean
Individual	Renegociou	Caso tenha Renegociado alguma dívida	Boolean
Individual	Tempo Relacionamento	Tempo em Anos	Inteiro
Individual	Média de Pagamento Dia	Tempo em Dias	Inteiro
Individual	Idade	Idade	Inteiro
Individual	Natureza Pessoa	Física / Jurídica	Lista (F/J)
Individual	Associado	Associado da cooperativa	Boolean
Individual	Dapiano	Faz parte do programa DAP	Boolean
Individual	Estado Civil	Estado Civil	Lista (S/C/V/D/Q/A/J/U)
Individual	Hectares	Quantidade de Hectares	Double
Individual	Mesoregiao	Localização do Produtor no Estado	Texto
Individual	Ramo	Predominio da Atividade do Produtor	Texto
Individual	Valor em Reais em Aberto	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ Varejo	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ Ração	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ Peças	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ Sementes	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ Insumos	Valor em Reais	Double
Individual	Total de Vendas em Reais p/ outras negócios	Valor em Reais	Double
Individual	Total de Vendas em Reais	Valor em Reais	Double
Individual	Cultivo de Soja (ha)	Quantidade de Hectares	Double
Individual	Cultivo de Milho (ha)	Quantidade de Hectares	Double
Individual	Cultivo de Trigo (ha)	Quantidade de Hectares	Double
Individual	Cultivo de Outros Grãos (ha)	Quantidade de Hectares	Double
Individual	Total de Faturamento em Reais de Grãos	Valor em Reais	Double
Individual	Total de Faturamento p/ Soja	Quantidade em Sacas	Double
Individual	Total de Faturamento p/ Milho	Quantidade em Sacas	Double
Individual	Total de Faturamento p/ Trigo	Quantidade em Sacas	Double
Individual	Total de Faturamento p/ De Outros Grãos	Quantidade em Sacas	Double
Individual	Total de Produção Entregue de Soja	Quantidade em Sacas	Double
Individual	Total de Produção Entregue de Milho	Quantidade em Sacas	Double
Individual	Total de Produção Entregue de Trigo	Quantidade em Sacas	Double
Individual	Total de Produção Entregue de Outros Grãos	Quantidade em Sacas	Double
Coletivo	Aumento da Produção de Soja	Aumento em Relação ao Ano Anterior	Boolean
Coletivo	Aumento da Produção de Trigo	Aumento em Relação ao Ano Anterior	Boolean
Coletivo	Aumento da Produção de Milho	Aumento em Relação ao Ano Anterior	Boolean
Coletivo	Aumento da Produção de Outros Grãos	Aumento em Relação ao Ano Anterior	Boolean

3.2. Pré-processamento

Nesta fase, foi realizada uma limpeza e algumas correções dos dados obtidos na etapa anterior, visando garantir a qualidade da informação extraída, bem como eliminar a in-

consistência, a incompletude e também os ruídos. Por exemplo, os registros com campos nulos foram removidos.

A limpeza dos dados se deu através da remoção dos registros onde algumas informações não foram localizadas, como tempo de relacionamento, bem como, clientes que não tiveram nenhuma movimentação ou que tiveram movimentações fragmentadas no período da análise. Também foram removidos os clientes com os dados básicos faltantes, como data de nascimento, estado civil. Além disso, diversos registros tiveram alguns campos ajustados, como a data de nascimento, que não estava formatada adequadamente. Esta limpeza reduziu de 908 mil registros para um pouco mais de 19 mil registros.

Após, verificou-se que os dados estavam desbalanceados, ou seja, existiam mais valores de uma classe do que de outra. Evitar esse desbalanceamento de classe é importante antes de aplicar um algoritmo de aprendizado de máquina pois o objetivo final é treinar um modelo de aprendizado de máquina que generalize bem para todas as classes possíveis [Kharwal 2021]. A base de dados deste trabalho estava desbalanceada com 86% dos registros classificados como inadimplente e 14% como adimplente. Para solucionar esse problema, foi aplicado o balanceamento por meio do método *ClassBalancer* [Frank 2023] presente na ferramenta *Weka*. Este método ajusta o peso das instâncias nos dados para que cada classe tenha o mesmo peso total.

3.3. Transformação

Após a etapa do pré-processamento, foram efetuadas transformações em todos os atributos do tipo *Double* e *Integer* de forma a discretiza-los em faixas de valores. A seguir é descrito como foi realizado cada transformação.

Para que as técnicas de discretização aplicadas a seguir sejam claras, é necessário conhecer o programa DAP[2], pois ele fornece informações sobre os pequenos produtores, sendo esses, pessoas que também compõem os clientes das cooperativas do ramo agrícola. Esse programa é um instrumento utilizado para identificar e qualificar as Unidades Familiares de Produção Agrária (UFPA) da agricultura familiar e suas formas associativas. É um programa de incentivo à produção e geração de renda [Brasil 2023].

Os atributos cuja unidade de medida é sacas, como a produção entregue em soja, milho e trigo, e como a comercialização desses grãos, foram discretizados em faixas de 70 sacas para facilitar a análise. Esse agrupamento foi feito com base na média da produtividade por hectare DAP das cidades da região do Alto Jacuí na cultura de soja.

Os atributos cuja unidade de medida são hectares, foram discretizados em faixas de 100 hectares. Esse agrupamento também se deve à média DAP, onde tem direito à emissão da DAP o produtor com área rural de até quatro módulos fiscais. Dessa forma, a média arredondada de quatro módulos fiscais é de 100 hectares na região.

Os dados do atributo faixa etária foram inicialmente agrupados de acordo com o modelo de agrupamento aplicado pelo Instituto Brasileiro de Geografia e Estatística [IBGE 2022], que gera faixas de 5 anos. Como muitas faixas foram criadas, decidiu-se faixas de 10 anos, para reduzir a quantidade de opções.

O atributo média de pagamento dia foi agrupado a cada 30 dias, em razão das

[2]Declaração de Aptidão ao Programa Nacional de Fortalecimento da Agricultura Familiar.

conferências mensais realizadas pelo setor de crédito. Enquanto que o atributo tempo de relacionamento foi agrupado a cada cinco anos devido às políticas internas da cooperativa. O cliente que não realiza movimentações nesse período é desassociado.

Os atributos de valores monetários (vendas varejo, vendas ração, vendas peças, vendas sementes, vendas insumos, demais vendas e total de vendas em todos os negócios), foram agrupados em faixas de 1.500 reais. Esse agrupamento foi definido a partir do arredondamento do salário mínimo.

3.4. Mineração de Dados

Concluindo a etapa da transformação dos dados, a próxima fase consiste na mineração de dados, onde os dados são submetidos aos algoritmos que buscam extrair padrões e assim informações valiosas para a tomada de decisão. Existem diversas técnicas de mineração, neste trabalho foi utilizada a técnica de classificação, pois o objetivo final é detectar possíveis inadimplentes.

Para este trabalho, foram utilizados os algoritmos de classificação com seus parâmetros já pré-definidos pela ferramenta *Weka*, onde foi catalogado como execução de combinação heterogênea e de execução individual.

As execuções de combinação heterogênea são constituídas pela junção dos algoritmos que aplicam um pré-processamento ou filtros juntamente com algum outro algoritmo de classificação, como por exemplo, o *AdaBoost* que impulsiona um classificador de classe nominal melhorando o desempenho deste [Weka 2023]. As execuções individuais foram realizadas através dos algoritmos executados sem o auxílio de uma combinação, como por exemplo, os algoritmos da família *Bayes*.

Foram executados 12 algoritmos de execução individual e 12 algoritmos de combinação heterogênea, cada algoritmo de combinação utilizou-se de cada um dos algoritmos de execução individual, tento um total de 106 execuções [3].

Este trabalho utilizou-se da técnica de validação cruzada (*K-fold*), que consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os $k-1$ restantes são utilizados para estimação dos parâmetros [dos Santos 2022]. O presente trabalho utilizou-se do parâmetro $k = 10$ e a reamostragem foi realizada dentro de cada *fold*.

Com a etapa de mineração concluída, a próxima etapa consiste da interpretação e avaliação do desempenho dos algoritmos, a fim de identificar padrões e tendências que podem ser usados para apoiar a tomada de decisão.

3.5. Interpretação e Avaliação

Interpretar e avaliar os resultados é uma etapa crítica no processo de descoberta de conhecimento em base de dados, pois é nessa etapa onde o modelo é validado se atende o objetivo estipulado. Neste trabalho, é esperado que o modelo resultante seja capaz de determinar se o cliente será um possível inadimplente e assim encaminhar ao analista financeiro para uma avaliação mais criteriosa antes da liberação de crédito.

[3] A lista completa das execuções está disponível no endereço eletrônico: https://github.com/jaisson/mineracao/blob/main/experimentos.pdf.

Avaliar a qualidade dos resultados obtidos pode ser feito através de diversas métricas. Neste trabalho foram utilizadas as métricas de revocação, precisão e *F1-Score*. Essas são métricas tradicionais para avaliar o desempenho dos algoritmos classificadores, segundo [Cássio Oliveira Camilo 2009].

A principal métrica utilizada para determinar a eficácia dos algoritmos foi a revocação. Neste trabalho, a revocação avalia a proporção das instâncias classificadas corretamente como inadimplentes em relação ao total de instâncias que eram realmente inadimplentes. Uma alta revocação indica que o modelo é capaz de identificar a maioria dos casos de inadimplência. Portanto, no momento em que se tem uma dúvida no resultado o modelo encaminha ao analista financeiro, para que esse faça a interpretação mais adequada e assim liberar ou negar a concessão do crédito.

Neste trabalho, a precisão avalia a proporção de instâncias classificadas corretamente como inadimplentes em relação ao total de instâncias classificadas como inadimplentes.

Por fim. a métrica F-1, de acordo com [dos Santos 2022], é uma maneira de medir as métricas de precisão e revocação juntas. É um calculo que usa a média harmônica entre as duas métricas. Um modelo que apresenta um bom F1 é um modelo capaz tanto de acertar suas predições (precisão alta) quanto de recuperar os exemplos da classe de interesse (revocação alta).

4. Resultados e Discussão

Esta seção descreve os experimentos realizados com o objetivo de analisar os atributos, os algoritmos e os parâmetros mais eficazes para identificar possíveis inadimplências na base de dados da cooperativa alvo deste trabalho, bem como discutir os resultados obtidos. A Subseção 5.1 apresenta os atributos mais relevantes, enquanto que a Subseção 5.2 descreve os algoritmos mais eficazes.

4.1. Identificação dos Atributos mais relevantes

Este experimento teve como objetivo encontrar os atributos mais relevantes para identificar possíveis inadimplências na base de dados utilizada. Os atributos mais relevantes no contexto deste trabalho são aqueles com maior poder discriminatório entre inadimplentes e adimplentes. O conhecimento sobre tais atributos pode auxiliar os analistas de crédito da cooperativa a elaborar estratégias visando a diminuição da inadimplência.

Para entender como funciona a relação entre os atributos inicialmente é utilizado a entropia, que segundo segundo [Castanheira 2008], a entropia é uma medida de informação calculada pelas probabilidades de ocorrência de eventos individuais ou combinados, ou seja, é a medida da quantidade de desordem. De acordo com [Almeida 2004] quanto maior o grau da entropia maior é a desordem e, quanto menor o grau da entropia melhor a organização. Desta forma, para [Castanheira 2008], a medida de ganho da informação representa a redução esperada da entropia de um atributo preditivo.

Para este trabalho, utilizou-se do avaliador de desempenho *InfoGainAttributeEval*, ele avalia o valor de um atributo medindo o ganho de informação em relação à classe [Weka 2023]. O resultado dessa avaliação está descrito na tabela 2, onde está ranqueado os 25 atributos com maior relevância a classificação.

Tabela 2. Ranking dos Atributos

Ranking	Atributo	Ganho de Informação
01°	RAMO	0.12224
02°	VENDAS VLR INSUMOS	0.08442
03°	MEDIA PGTO DIA	0.07535
04°	VENDAS VLR RACAO	0.06786
05°	DEP SOJA SACAS	0.04421
06°	VENDAS VLR SEMENTES	0.03348
07°	DEP TRIGO SACAS	0.01973
08°	FAT SOJA SACAS	0.01490
09°	HE CULT SOJA	0.01448
10°	VENDAS VLR PECAS	0.01371
11°	HECTARES	0.01225
12°	DEP MILHO SACAS	0.01172
13°	ESTCIVIL	0.01021
14°	DAPIANO	0.00948
15°	VENDAS VLR VAREJO	0.00846
16°	DEP DE MAIS	0.00461
17°	NATUREZA PESSOA	0.00432
18°	TEMPO RELACIONAMENTO	0.00418
19°	FAT TRIGO SACAS	0.00313
20°	MESOREGIAO	0.00292
21°	ASSOCIADO	0.00289
22°	RENEGOCIOU	0.00270
23°	VENDAS VLR DEMAIS	0.00267
24°	AUMENTOU SOJA	0.00260
25°	AUMENTOU TRIGO	0.00213

4.2. Algoritmo mais eficaz

O objetivo deste experimento foi encontrar o algoritmo mais eficaz para identificar possíveis inadimplentes em uma cooperativa do ramo agrícola. A eficácia do algoritmo foi analisada por meio das métricas de revocação, precisão e F1, priorizando a revocação, pois ela mede a capacidade do algoritmo de encontrar todos os exemplos positivos em um conjunto de dados. Esse é um aspecto importante para os analistas de crédito, pois eles precisam identificar corretamente os inadimplentes para avaliar a concessão de crédito.

Foram efetuados um total de 106 execuções de algoritmos, incluindo algoritmos isolados e de combinação heterogênea. A tabela 3 apresenta o ranking dos 15 algoritmos mais eficazes.

Tabela 3. 15 Algoritmos mais eficazes

Algoritmo	Revocação	Precisão	F1
RandomizableFilteredClassifier.NaiveBayes	0,668	0,591	0,586
RandomizableFilteredClassifier.NaiveBayesUpdateable	0,668	0,591	0,586
InputMappedClassifier.RandomizableFilteredClassifier.NaiveBayes	0,668	0,591	0,586
InputMappedClassifier.RandomizableFilteredClassifier.NaiveBayesUpdateable	0,668	0,591	0,586
AdaBoostM1.NaiveBayes	0,649	0,674	0,673
AdaBoostM1.NaiveBayesUpdateable	0,649	0,674	0,673
InputMappedClassifier.AdaBoostM1.NaiveBayes	0,649	0,674	0,673
InputMappedClassifier.AdaBoostM1.NaiveBayesUpdateable	0,649	0,674	0,673
AdaBoostM1.BayesNet	0,648	0,672	0,671
InputMappedClassifier.AdaBoostM1.BayesNet	0,648	0,672	0,671
NaiveBayes	0,638	0,679	0,677
NaiveBayesUpdateable	0,638	0,679	0,677
FilteredClassifier.NaiveBayes	0,638	0,679	0,677
FilteredClassifier.NaiveBayesUpdateable	0,638	0,679	0,677
WeightedInstancesHandlerWrapper.NaiveBayes	0,638	0,679	0,677

Os algoritmos mais eficazes foram : *RandomizableFilteredClassifier* em combinação com os algoritmos *NaiveBayes* e *NaiveBayesUpdateable*, bem como as

combinações de *InputMappedClassifier* com *RandomizableFilteredClassifier* com os algoritmos *NaiveBayes* e *NaiveBayesUpdateable*.

O *RandomizableFilteredClassifier* é uma abordagem de classificação que combina filtragem de dados e classificação em um único modelo, por meio de técnicas de pré-processamento para filtrar ou transformar os dados antes de aplicar um algoritmo de classificação, assim impulsionando outro algoritmo tornando-o mais eficiente [Weka 2023]. Essa abordagem de combinar filtragem aleatória com classificação permite lidar com problemas de dados desbalanceados, redução de dimensionalidade, normalização e outras tarefas de pré-processamento.

Continuando a análise da tabela 3, percebe-se a família *bayesiana* entre os algoritmos mais eficientes, eles são métodos de classificação probabilísticos baseados no teorema de Bayes, que calcula a probabilidade de um evento acontecer, com base em um conhecimento que pode estar relacionado ao evento [Weka 2023].

5. Conclusão

A inadimplência é um problema significativo tanto para os consumidores quanto para as empresas, afetando negativamente o crédito e gerando prejuízos financeiros. No contexto das cooperativas do ramo agropecuário, a inadimplência também é uma realidade preocupante. Essas cooperativas lidam com transações financeiras e têm um grande número de clientes, o que dificulta a avaliação de crédito e aumenta os desafios relacionados à inadimplência.

Nesse contexto, o presente trabalho teve como objetivo desenvolver um modelo de classificação capaz de identificar os clientes com maiores chances de inadimplência e assim reduzir a inadimplência futura na cooperativa em questão. Para alcançar esse objetivo, foram utilizados dados reais da cooperativa, incluindo informações financeiras, histórico de vendas e comercialização de grãos entre os anos de 2015 a 2020, obtendo um total final de 19 mil registros. Esses dados foram analisados pelos algoritmos de classificação disponíveis na ferramenta *Weka*. Foram executados 12 algoritmos de execução individual e 12 algoritmos de combinação heterogênea, tendo um total de 106 execuções.

O resultado final deste trabalho mostrou que o melhor modelo para identificar inadimplentes foi a combinação heterogênea de dois algoritmos: *RandomizableFilteredClassifier* e *NaiveBayes*, atingindo uma revocação de 67%. Isso permitirá aos analistas de crédito priorizarem esses clientes para uma avaliação mais criteriosa, contribuindo para a redução da inadimplência e mitigação dos prejuízos financeiros. Ainda assim, buscou-se o refinamento desta combinação, bem como da combinação *RandomizableFilteredClassifier* e a família *bayesiana*, na tentativa de melhorar a revocação, mas não teve melhoras significativas em relação ao modelo encontrado.

Em conclusão, o modelo de classificação desenvolvido neste trabalho apresenta uma abordagem promissora para lidar com o desafio da inadimplência, oferecendo suporte aos analistas de crédito na tomada de decisões mais informadas e eficientes. A aplicação desse modelo pode ajudar a cooperativa a identificar e gerenciar melhor os riscos de inadimplência, aumentando a eficácia de suas estratégias de avaliação de crédito e reduzindo os impactos negativos causados pela inadimplência.

Ainda assim, outros estudos deverão ser realizados, como a busca de mais atribu-

tos relacionados ao cliente a fim de melhorar o modelo, além de conseguir dados de outras cooperativas e também obter mais dados históricos devido às sazonalidades e mudanças climáticas, que interferem diretamente na produção de grãos impactando consideravelmente a vida financeira dos produtores. Além disso, o modelo desenvolvido será validado com novos dados e ajustado conforme necessidades futuras. Também será desenvolvida uma ferramenta *WEB* que utilizará o modelo criado neste trabalho para priorizar os clientes a serem avaliados pelos analistas para a concessão de crédito.

Referências

Almeida, L. M. (2004). Uma ferramenta para extração de padrões. *Centro Universitário Luterano de Palmas ULBRA*.

Brasil (2023). Declaração de aptidão ao pronaf (dap). Disponível em: `https://www.gov.br/agricultura/pt-br/assuntos/agricultura-familiar/dap`. Acesso em: 10/06/2023.

Castanheira, L. G. (2008). Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. Master's thesis, Universidade Federal de Minas Gerais.

Cássio Oliveira Camilo, J. C. d. S. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Instituto de Informática Universidade Federal de Goiás*.

da Silva Ribeiro, H. (2020). Classificação de clientes utilizando mineração de dados. Master's thesis, Pontifícia Universidade Católica de Goiás, Goiânia.

dos Santos, P. F. (2022). Uso de técnicas de machine learning para análise de risco de crédito. Master's thesis, Universidade de Brasília, Brasília.

Experian (2023). Inadimplência atinge 27% dos produtores rurais brasileiros, revela serasa experian. Disponível em: `https://www.serasaexperian.com.br`. Acesso em: 30/05/2023.

Fayyad, U. M. (1996). *Advances in Knowledge Discovery Data Mining*. MIT Press, Massachusetts, Estados Unidos, 1º edição edition.

Ferreira, J. C. (2018). knowledge discovery in database e data mining: uma contribuição bibliométrica. *encontro nacional de engenharia de producao*, XXXVIII(18).

Frank, E. (2023). Class classbalancer. Disponível em: `https://weka.sourceforge.io/doc.dev/weka/filters/supervised/instance/ClassBalancer.html`. Acesso em: 09/07/2023.

IBGE (2022). Instituto brasileiro de geografia e estatística. Disponível em: `https://www.ibge.gov.br/`. Acesso em: 10/06/2022.

Kharwal, A. (2021). Class balancing in machine learning. Acesso em: 09/07/2023.

Reis, M. A. (2022). Modelo preditivo de risco de crédito para cooperativas de agronegócio. Master's thesis, Universidade de Brasília, Brasília.

Serasa (2023). Número de inadimplentes cai pelo segundo mês seguido, diz serasa. Disponível em: `https://www.serasa.com.br`. Último acesso em: 30/05/2023.

Weka (2023). Weka wiki. Disponível em: `https://waikato.github.io/weka-wiki/documentation/`. Acesso em: 09/07/2023.