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

Abstract. *Non-payment is a problem that affects both consumers and businesses, compromising credit and generating financial losses. Recent studies by Serasa and Serasa-Experian 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 graduation project, a classification model was developed 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 realizados pelas empresas Serasa e Serasa-Experian 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 trabalho de conclusão de curso, foi desenvolvido 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 heterogênea de dois algoritmos: *RandomizableFilteredClassifier* e o algoritmo *NaiveBayes*. Esse modelo alcançou uma revocação de 67%.

O escopo deste artigo está organizado da seguinte maneira. A Seção 2 descreve a fundamentação teórica, abordando os principais termos, técnicas e tecnologias relevantes para a compreensão deste trabalho. A Seção 3 discute os principais trabalhos relacionados à temática abordada neste artigo. A Seção 4 apresenta a metodologia aplicada neste trabalho. Os resultados obtidos são apresentados e discutidos na Seção 5. Por fim, a Seção 6 traz as considerações finais e aponta possíveis trabalhos futuros.

2. Fundamentação Teórica

A Mineração de Dados é uma área da ciência da computação que se preocupa com a descoberta de padrões e relacionamentos em grandes conjuntos de dados, tendo como objetivo identificar informações ocultas nos dados que podem ser úteis para a tomada de decisões. A mineração de dados corresponde a uma etapa do processo de descoberta de conhecimento em base de dados, que foi adotado neste trabalho.

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

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 e conhecimentos potencialmente úteis e implicitamente contidos em bases de dados. As etapas deste processo são representadas na figura 1.

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.

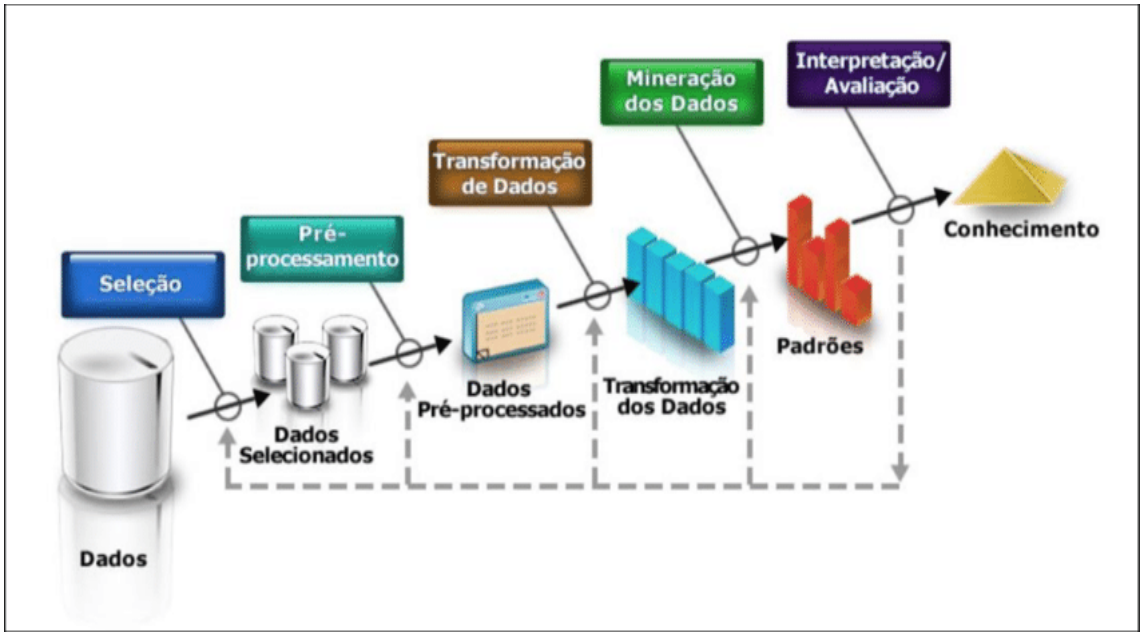

Figura 1. Etapas do processo de KDD. Fonte: Adaptado de Fayyad (1996)

A etapa de seleção dos dados de acordo com Ferreira (2018), é também conhecida como "Redução de Dados". É a primeira etapa no processo de descoberta de informação e possui papel fundamental no resultado final, uma vez que nesta etapa é definido o conjunto de dados contendo todas as possíveis variáveis (atributos) e registros (instâncias, casos, observações ou padrões) que se pretende analisar. Em sua grande maioria, esta seleção é realizada por um especialista da área, ou seja, alguém que realmente entende do assunto em questão.

Conforme Camilo (2009), devido às diversas origens possíveis, é comum que os dados não estejam preparados para que os métodos de Mineração de Dados sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser

necessárias. Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher ou retirar valores vazios, Essas ações devem ser realizadas na etapa do pré-processamento.

A etapa de transformação dos dados ou codificação dos dados tem como objetivo principal converter o conjunto bruto de dados em uma forma padrão de uso, segundo Ferreira (2018). Esta etapa é implementada através de um processamento dos dados, visando organizar os dados para auxiliar o trabalho sucedido pelas fases posteriores do processo *KDD*.

Na fase da mineração de dados, são aplicados os algoritmos. Para Souza (2013), nesta etapa, uma das tarefas comumente resolvidas com técnicas de mineração de dados é a classificação. Ela pode ser definida como um processo de aprendizagem de máquina que visa compreender a estrutura subjacente de semelhanças entre as instâncias de uma mesma classe.

De acordo com Camilo (2009), a Mineração de Dados é comumente classificada pela sua capacidade em realizar determinadas tarefas, as tarefas mais comuns são classificação, regressão, agrupamento e regras de associação. Essas tarefas são descritas a seguir.

- Classificação: visa identificar a qual classe um determinado registro pertence. Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado);
- Regressão: similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não categórico. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais;
- Agrupamento: tem como finalidade identificar e aproximar os registros similares. Um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares;
- Associação: esta tarefa consiste em identificar quais atributos estão relacionados. Apresentam a forma: **SE** *atributo X* **ENTÃO** *atributo Y*.

Os métodos de mineração de dados, de acordo com Camilo (2009), são divididos em aprendizado não-supervisionado (descritivo) e supervisionado (preditivo). Os métodos não-supervisionados não precisam de uma pré-categorização para os registros. Um dos métodos mais comuns para o aprendizado descritivo é o agrupamento. Já no aprendizado supervisionado, os métodos são providos com um conjunto de dados que possuem uma variável alvo pré-definida e os registros são categorizados em relação a ela. As técnicas mais comuns de aprendizado preditivo são a classificação e a regressão.

Visando melhorar o desempenho geral da modelagem ou otimizar o processo de aprendizado, existem os meta-algoritmos, são técnicas que usam outros algoritmos de aprendizado podendo serem nas tarefas citadas anteriormente. Existem diferentes tipos de meta-algoritmos. A seguir são citados alguns mais utilizados.

- *Ensemble Learning*: combinam as previsões de múltiplos modelos de aprendizado de máquina, os mais comuns são os algoritmos *Random Forests*, *AdaBoost* e *Bagging*, conforme Dietterich (2000);
- Aprendizado ativo: interagem com o especialista humano para obter *feedback* e selecionar amostras de treinamento mais informativas, conforme Settles (2010).

A etapa de interpretação e avaliação, segundo Ferreira (2018), é uma fase que envolve todos os participantes que avaliam de forma criteriosa os resultados, proporcionando uma interpretação para o modelo desenvolvido. Caso o resultado não seja satisfatório, o processo pode retornar a qualquer uma das etapas anteriores.

Como este trabalho visa a identificação de possíveis inadimplentes, seguindo o processo de descoberta de conhecimento em uma base de dados. Utilizou-se do aprendizado supervisionado, tendo como método de mineração de dados a classificação, juntamente com a técnica de *Ensemble Learning*, que combina várias técnicas para melhorar a precisão do modelo de classificação. Dados de movimentações financeiras, comercializações entre outras características, foram utilizadas para o desenvolvimento deste modelo.

3. Trabalhos Relacionados

Esta seção discute os estudos relacionados ao contexto em que este trabalho está inserido. Esses estudos foram identificados por meio de pesquisas realizada na plataforma *Google Google (2023)* utilizando os termos "mineração de dados", "crédito", "finanças"e "inadimplência". A partir dos resultados obtidos, foram selecionados os cinco trabalhos mais recentes e relevantes para a proposta deste trabalho.

No trabalho realizado por Reis (2022), o objetivo era verificar qual método de aprendizado de máquina, dentre os principais, apresentaria os melhores resultados ao substituir o modelo fundamentalista de avaliação de risco de crédito realizado por analistas humanos. Foram testados vários modelos amplamente utilizados no mercado, utilizando as métricas MAE[2], MSE[3], MAPE[4] e RMSE[5] como base para o modelo proposto pelo pesquisado, sendo observado que dois modelos se destacaram para a proposta do estudo: *Gradient Boosting* e *Random Forest*. Além disso, o estudo utilizou as demonstrações financeiras dos anos de 2017 a 2022, abrangendo diferentes ciclos produtivos, visando compreender a realidade de forma abrangente. Foram avaliadas um total de 249 cooperativas do agronegócio, localizadas em diferentes regiões do Brasil.

No estudo realizado por Santos (2022), o objetivo era identificar um modelo preditivo de classificação de risco de crédito em operações comerciais de cheque especial para pessoas físicas. Para isso, foram utilizados algoritmos de *Machine Learning*, incluindo Regressão Logística, *k-Nearest Neighbors (k-NN), Support Vector Machines (SVM)*, Árvore de Decisão, *Random Forest* e *Gradient Boosting*. Os resultados do estudo indicaram um desempenho superior para o modelo *Random Forest*. A base de dados utilizada no trabalho consistiu em uma amostra anonimizada de operações de crédito de um banco brasileiro, abrangendo o período de maio a outubro de 2020.

[2]Mean absolute error regression loss - Perda média de regressão de erro absoluto.
[3]Mean squared error regression loss - Perda de regressão de erro quadrático médio.
[4]Mean absolute percentage error regression loss - Perda média de regressão de erro percentual absoluto.
[5]Root mean squared error regression loss - Perda de regressão de erro quadrático médio raiz.

No estudo conduzido por Silva (2022), foram utilizadas técnicas de Regressão Logística (*Logit*) e modelos de *Machine Learning*, como *Random Forest* e *Gradient Boosting*, com o objetivo de prever quando um contrato habitacional poderia se tornar um Ativo Problemático (AP). Essa previsão visa subsidiar decisões para lidar com cenários extremos, em conformidade com as exigências do órgão regulador. Através das métricas de Curva ROC, precisão e revocação, o algoritmo *Gradient Boosting* demonstrou maior capacidade de previsão desejada, em termos de acurácia, eficiência e sensibilidade. Os dados utilizados no estudo foram extraídos de uma instituição financeira brasileira, no contexto do segmento imobiliário, envolvendo pessoas físicas que obtiveram anteriormente aprovação e contratação de crédito relacionado a financiamentos no Sistema Brasileiro de Poupança e Empréstimo (SBPE), abrangendo o segundo semestre de 2021.

No trabalho realizado por Beserra (2021), foi utilizada a técnica de mineração de dados conhecida como regressão logística, sendo utilizada uma base de dados chamada German Credit Data, disponibilizada pela Universidade da Califórnia Irvine UCI no repositório Machine Learning Repository's UCI (2023). Essa base de dados contém informações financeiras e pessoais de 1000 clientes de um cartão de crédito na Alemanha. O objetivo do estudo foi aplicar a técnica de regressão logística para identificar clientes adimplentes e inadimplentes. Os resultados do trabalho demonstraram uma acurácia de 72% e uma sensibilidade de 87%. Dos 140 clientes adimplentes, o modelo proposto pelo autor identificou corretamente 122 deles, apresentando também uma especificidade de 38% para os 60 clientes inadimplentes, onde o modelo acertou 23. Além disso, a probabilidade de um cliente ser adimplente foi estimada em 76%.

No estudo realizado por Ribeiro (2020), foi aplicado um algoritmo de classificação de mineração de dados em uma empresa atacadista. O objetivo do trabalho foi utilizar atributos comerciais e financeiros para classificar os clientes em quatro perfis: bronze, prata, ouro e diamante. Essa classificação permitiria uma análise de comportamento e crédito dos clientes. Os dados utilizados no estudo foram anonimizados pelo autor e abrangeram um período de seis meses em um ano. O algoritmo escolhido para realizar a classificação foi o J48, que obteve uma acurácia de 89% no trabalho. Isso indica que o modelo desenvolvido pelo autor foi capaz de classificar corretamente 89% dos clientes nos perfis determinados.

Estes estudos demonstram a aplicação bem-sucedida de técnicas de mineração de dados e algoritmos de classificação para lidar com problemas de inadimplência e análise de crédito em diferentes setores, como cooperativas do agronegócio, operações de cheque especial, contratos habitacionais e análise de clientes em uma empresa atacadista. Essas abordagens têm o potencial de auxiliar na tomada de decisões estratégicas e na redução da inadimplência, fornecendo informações valiosas para os analistas de crédito, bem como para os gestores financeiros.

Na tabela 1, é apresentado um resumo dos trabalhos relacionados, as métricas utilizadas e os algoritmos que alcançaram o melhor desempenho nos seus respectivos trabalhos. Observa-se que os trabalhos tem como foco contextos diferentes desde contratos habitacionais até cartão de crédito. Apenas o trabalho de Reis (2022) é voltado para cooperativas do agronegócio, assim como este trabalho de conclusão de curso. As métricas utilizadas também são diversas, sendo que a acurácia foi utilizada pela maioria dos trabalhos. Por fim, os algoritmos com melhor desempenho também foram diferentes em

cada trabalho demonstrando a necessidade de avaliação de diferentes classificadores para a base de dados alvo do trabalho.

Tabela 1. Trabalhos Relacionados

Autor	Contexto	Métricas	Melhor Algoritmo
(REIS, 2022)	Cooperativas do Agronegócio	MAE, MSE, MAPE, RMSE .	*Gradient Boosting* e *Random Forest*
(SANTOS, 2022)	Cheques especiais para PF	Curva ROC, precisão e revocação	*Random Forest*
(SILVA, 2022)	Contratos habitacionais	acurácia, eficiência e sensibilidade	*Gradient Boosting*
(BESERRA, 2021)	Cartão de crédito	acurácia e sensibilidade	Regressão Logística
(RIBEIRO, 2020)	Empresa atacadista	acurácia	J48

4. 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.

4.1. 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 2, 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 rotulado 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 2. 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

4.2. 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 inconsistência, 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 solu-

cionar 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.

4.3. 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[6], pois ele fornece informações sobre os pequenos produtores que são o foco das técnicas de discretização, 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 conferências mensais realizadas pelo setor de crédito. Enquanto que o atributo tempo de relacionamento foi agrupado a cada cinco anos devido as políticas internas da cooperativa. Onde 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.

O resultado final dessas transformações é apresentado na Tabela 3. Observa-se nessa tabela as faixas de discretização e a referência que determina cada faixa.

4.4. 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

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

Tabela 3. Atributos Transformados

Atributo	Faixa	Referência
Tempo Relacionamento	5	Politica Interna
Média de Pagamento Dia	30	Mensal
Idade	10	Adaptação IBGE
Hectares	100	4 módulos fiscais / DAP
Valor em Reais em Aberto	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ Varejo	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ Ração	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ Peças	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ Sementes	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ Insumos	1500	Salário mínimo arredondado
Total de Vendas em Reais p/ outras negócios	1500	Salário mínimo arredondado
Total de Vendas em Reais	1500	Salário mínimo arredondado
Cultivo de Soja (ha)	100	4 módulos fiscais / DAP
Cultivo de Milho (ha)	100	4 módulos fiscais / DAP
Cultivo de Trigo (ha)	100	4 módulos fiscais / DAP
Cultivo de Outros Grãos (ha)	100	4 módulos fiscais / DAP
Total de Faturamento em Reais de Grãos	70	Produtividade média de 1 hectare
Total de Faturamento p/ Soja	70	Produtividade média de 1 hectare
Total de Faturamento p/ Milho	70	Produtividade média de 1 hectare
Total de Faturamento p/ Trigo	70	Produtividade média de 1 hectare
Total de Faturamento p/ De Outros Grãos	70	Produtividade média de 1 hectare
Total de Produção Entregue de Soja	70	Produtividade média de 1 hectare
Total de Produção Entregue de Milho	70	Produtividade média de 1 hectare
Total de Produção Entregue de Trigo	70	Produtividade média de 1 hectare
Total de Produção Entregue de Outros Grãos	70	Produtividade média de 1 hectare

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 isolados, tento um total de 106 execuções, a lista completa das execuções está disponível no endereço eletrônico [7] .

Os experimentos foram realizados em um computador pessoal, cedido pela cooperativa, com as seguintes características:

- Sistema Operacional - Windows 10 Pro;
- Processador - Intel(R) Core(TM) i7-9700 CPU 3.00 GHz;
- Memória RAM - 16,0 GB;
- Versão do Java - 11.0.12+8-LTS-237;
- Versão do Weka - 3.8.3;
- Acesso ao banco de dados Oracle, somente visualização.

[7] https://drive.google.com/file/d/1AcdgwyYyqNYDAubnKXJtfXn6lSlZRdmU/view

Segundo Bouckaert (2004), uma maneira natural de medir o desempenho de um algoritmo em um determinado conjunto de dados é prever seu desempenho futuro. Uma das técnicas para medir o desempenho é chamada de validação cruzada. Esta técnica fornece um método de avaliação onde os dados são divididos em conjuntos de treinamento e de validação.

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 (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.

4.5. 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.

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 Camilo (2009).

Para entender como cada métrica funciona, antes é necessário o entendimento do conceito de matriz de confusão. Essa é uma matriz de ordem 2, que contém os valores dos testes aplicados indicando os erros e os acertos do algoritmo comparando com o resultado esperado. A Figura 2 exemplifica uma matriz de confusão que inclui os seguintes conceitos conforme Camilo (2009):

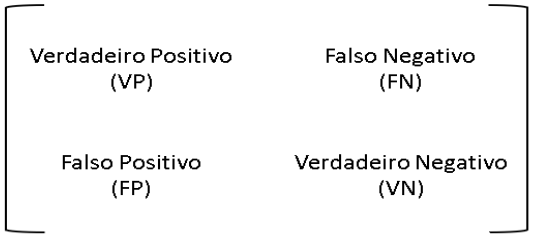

Figura 2. Exemplo de matriz de confusão

- Verdadeiro Positivo (VP): Total dos registros onde o algoritmo classificou corretamente, ou seja, classificou como Positivo e o valor realmente era Positivo. No caso deste trabalho, se o cliente previsto como inadimplente, verificou-se de fato como inadimplente.

- Falso Negativo (FN): Total dos registros onde o algoritmo classificou como Negativo a classe, porém o valor real era Positivo. Para este trabalho, foi avaliado se os registros previstos como adimplente eram na verdade inadimplentes.
- Falso Positivo (FP): Total dos registros onde o algoritmo classificou como Positivo, entretanto o valor real era Negativo. Neste caso, foi verificado se o cliente previsto como inadimplente foi classificado incorretamente, ou seja, o cliente era adimplente.
- Verdadeiro Negativo (VN): Total dos registros onde o algoritmo classificou como Negativo e o valor realmente era Negativo. No presente trabalho, se o cliente previsto como adimplente foi classificado como adimplente.

Neste trabalho, a principal métrica utilizada para determinar a eficácia dos algoritmos é a revocação. Essa métrica é utilizada para avaliar o desempenho de modelos de classificação, através da proporção de instâncias positivas corretamente identificadas em relação ao total de instâncias positivas presentes nos dados. Seu cálculo está na razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos + falsos negativos, como demonstrado na Equação 1

$$Rv = \frac{VP}{(VP + FN)} \tag{1}$$

O valor da revocação varia de 0 a 1, onde 1 representa uma revocação perfeita, ou seja, todos os casos positivos foram corretamente identificados. Um valor baixo de revocação indica que o modelo está falhando em identificar uma quantidade significativa de casos positivos.

A métrica de revocação é muito útil quando o foco está na detecção de casos positivos relevantes, como identificar clientes inadimplentes. Uma alta revocação indica que o modelo é capaz de identificar a maioria dos casos positivos, minimizando assim os falsos negativos, de acordo com Camilo (2009). 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 negando a concessão do crédito.

Segundo Santos (2022), a precisão é uma das métricas mais comuns para avaliar modelos de classificação, ela é utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas), conforme a Equação 2. No caso deste trabalho, se todos os clientes classificados como inadimplentes quantos foram corretamente classificados.

$$Pr = \frac{VP}{(VP + FP)} \tag{2}$$

Enquanto que a métrica F-1, de acordo com 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. Conforme é mostrado na Equação 3. 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).

$$F1 = \frac{2 * Pr * Rv}{Pr + Rv} \tag{3}$$

5. 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, a Subseção 5.2 descreve os algoritmos mais eficazes e a Subseção 5.3 analisa os parâmetros mais eficazes.

5.1. 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. Desda 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 descrita na tabela 4, onde está ranqueado os 25 atributos com maior relevância a classificação.

Tabela 4. Ranking dos Atributos

Ranking	Attribute
01°	RAMO
02°	VENDAS VLR INSUMOS
03°	MEDIA PGTO DIA
04°	VENDAS VLR RACAO
05°	DEP SOJA SACAS
06°	VENDAS VLR SEMENTES
07°	DEP TRIGO SACAS
08°	FAT SOJA SACAS
09°	HE CULT SOJA
10°	VENDAS VLR PECAS
11°	HECTARES
12°	DEP MILHO SACAS
13°	ESTCIVIL
14°	DAPIANO
15°	VENDAS VLR VAREJO
16°	DEP DE MAIS
17°	NATUREZA PESSOA
18°	TEMPO RELACIONAMENTO
19°	FAT TRIGO SACAS
20°	MESOREGIAO
21°	ASSOCIADO
22°	RENEGOCIOU
23°	VENDAS VLR DEMAIS
24°	AUMENTOU SOJA
25°	AUMENTOU TRIGO

5.2. 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 5 apresenta o ranking dos 15 algoritmos mais eficazes.

Tabela 5. 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 5, percebe-se a família *bayesiana* entre os algoritmos mais eficientes, eles são métodos de classificação probabilístico 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).

5.3. Refinamento dos parâmetros mais eficazes do algoritmo *RandomizableFilteredClassifier*

O objetivo desse experimento foi refinar os parâmetros do algoritmo mais eficaz para identificar inadimplentes. Para atingir esse objetivo foram analisados os seguintes parâmetros da combinação dos algoritmos *RandomizableFilteredClassifier* e *NaiveBayes*: filtro, dimensão e distribuição.

O parâmetro filtro tem como objetivo definir o algoritmo que será utilizado para filtrar os dados pelo classificador *RandomizableFilteredClassifier*. Foram testados os seguintes filtros:

- *Randomize*: é um filtro de instância não supervisionado que embaralha aleatoriamente a ordem das instâncias passadas por ele.
- *ReservoirSample* é um filtro de instância não supervisionado que produz uma subamostra aleatória de um conjunto de dados por meio do algoritmo "R"de Vitter.
- *Resample* é um filtro de instância não supervisionado que produz uma subamostra aleatória de um conjunto de dados, usando amostragem com ou sem substituição.
- *SpreadSubsample* é um filtro de instância supervisionado que produz uma subamostra aleatória de um conjunto de dados. Ele permite especificar o "*spread*"máximo entre a classe mais rara e a mais comum, para evitar que o conjunto de dados fique desequilibrado.
- *RandomProjection* é um filtro de atributo não supervisionado que reduz a dimensionalidade dos dados projetando-os em um subespaço dimensional inferior usando uma matriz aleatória com colunas de comprimento unitário.

No experimento de análise do parâmetro filtro, foram utilizadas a dimensão e a distribuição padrões da ferramenta *Weka* (respectivamente, 42 e *Sparse1*) e foram testadas as seguintes opções de filtro: *RandomProjection*, *ReservoirSample*, *Resample*, *Randomize* e *SpreadSubsample*. A Figura 3 apresenta os resultados, onde observa-se que o filtro mais eficaz foi *RandomProjection*, pois apresentou o maior valor de revocação, logo o parâmetro com a maior capacidade de identificar a inadimplência para uma análise futura pelos analistas de crédito da cooperativa.

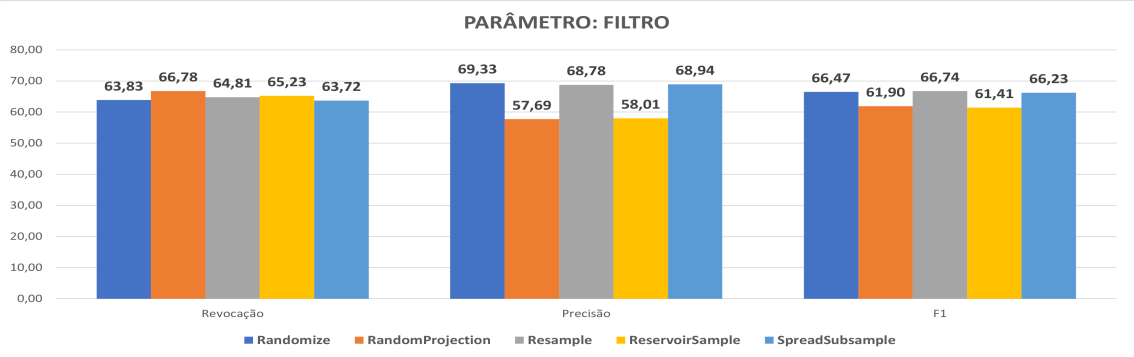

Figura 3. Testes de Refinamento - Parâmetro: Filtro

A segunda análise teve como objetivo identificar a distribuição mais eficaz para o objetivo desse trabalho. Esse parâmetro determina a distribuição utilizada para calcular a matriz de colunas que é aplicada pelo filtro de *RandomProjection* para reduzir a dimensionalidade dos dados. Nessa análise, foi utilizada a dimensão padrão da ferramenta Weka (42) e o filtro *RandomProjection*, identificado como o mais eficaz no analise anterior. Foram testadas as seguintes opções de distribuição: *Sparse1*, *Gaussian* e *Sparse2*. A Figura 4 apresenta os resultados, onde observa-se que a distribuição mais eficaz foi *Sparse1*, pois apresentou o maior valor de revocação. O próximo experimento usou este valor de distribuição.

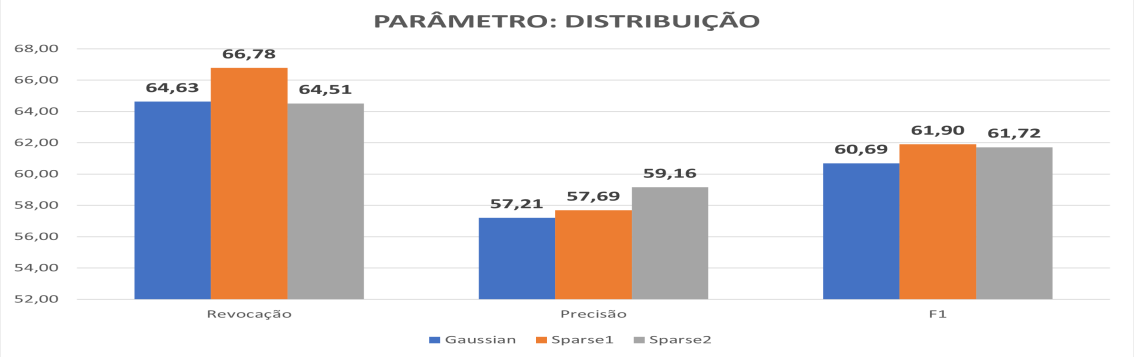

Figura 4. Testes de Refinamento - Parâmetro: Distribuição

A terceira análise teve como objetivo identificar a dimensão mais eficaz para a tarefa de identificar inadimplentes. Nessa análise, foi utilizado o filtro mais eficaz, o *RandomProjection*, e a distribuição mais eficaz, o *Sparse1*. Foram testadas as dimensões com os seguintes parâmetros: padrão de 0% mas com *seed* de 42, 50%, 75% e 100%. A Figura 5 apresenta os resultados, onde observa-se que a dimensão padrão foi a mais eficaz, pois obteve o maior valor de revocação.

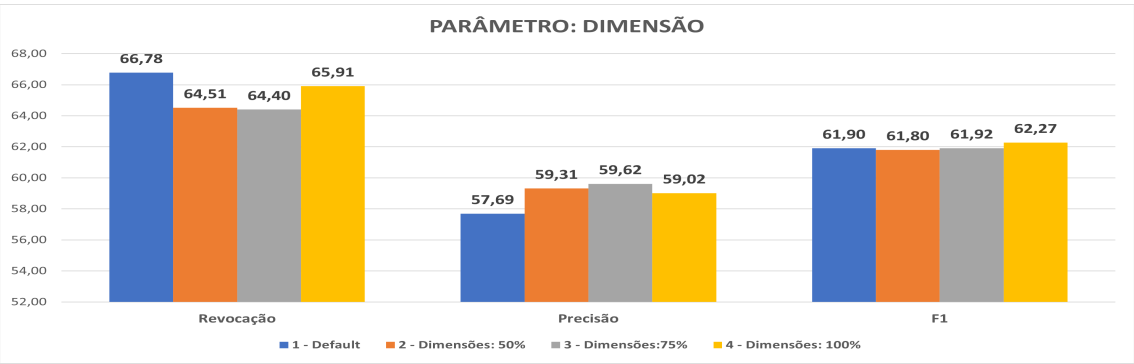

Figura 5. Testes de Refinamento - Parâmetro: Dimensão

Ao avaliar os resultados, percebe-se que o melhor algoritmo é a combinação dos algoritmos *RandomizableFilteredClassifier* e *NaiveBayes* sem a necessidade de alterações nos parâmetros pré definidos. Concluindo então que o modelo que melhor auxiliará os analistas financeiros na concessão de crédito é formado pelos atributos definidos e a combinação dos algoritmos antes mencionados.

6. 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, tento um total de 106 execuções.

O resultado final deste trabalho tem como modelo resultante da 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 atributos 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. Uma ferramenta para extração de padrões. *Centro Universitário Luterano de Palmas ULBRA*, 2004. Disponível em: <https://www.cin.ufpe.br/~lma3/UmaFerramentaParaExtracaoDePadroes.pdf>.

BESERRA, R. S. *Modelagem Com Regressão Logística Para Análise De Concessão De Crédito*. Dissertação (Mestrado) — Universidade Estadual Da Paraíba, Paraíba, 2021.

BOUCKAERT, R. R. Bayesian network classifiers in weka. *University of Waikato*, 2004. Disponível em: <https://weka.sourceforge.io/manuals/weka.bn.pdf>.

BRASIL. *Declaração de Aptidão ao Pronaf (DAP)*. 2023. Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/agricultura-familiar/dap>. Acesso em: 10/06/2023.

CAMILO, J. C. d. S. C. O. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Instituto de Informática Universidade Federal de Goiás*, 2009. Disponível em: <https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf>.

CASTANHEIRA, L. G. *Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2008.

DIETTERICH, T. G. Ensemble methods in machine learning. *Oregon State University*, 2000. Disponível em: <https://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf>.

EXPERIAN. *Inadimplência atinge 27% dos produtores rurais brasileiros, revela Serasa Experian*. 2023. Disponível em: <https://www.serasaexperian.com.br/sala-de-imprensa/agronegocios/inadimplencia-atinge-27-dos-produtores-rurais-brasileiros-revela-serasa-experian/>. Acesso em: 30/05/2023.

FAYYAD, U. M. *Advances in Knowledge Discovery Data Mining*. 1º edição. ed. Massachusetts, Estados Unidos: MIT Press, 1996.

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

FRANK, E. *Class ClassBalancer*. 2023. Disponível em: <https://weka.sourceforge.io/doc.dev/weka/filters/supervised/instance/ClassBalancer.html>. Acesso em: 09/07/2023.

GOOGLE. *Google Acadêmico*. 2023. Disponível em: <https://scholar.google.com.br>. Acesso em: 13/08/2023.

IBGE. *Instituto Brasileiro de Geografia e Estatística*. 2022. Disponível em: <https://www.ibge.gov.br/>. Acesso em: 10/06/2022.

KHARWAL, A. *Class Balancing in Machine Learning*. 2021. Acesso em: 09/07/2023.

REIS, M. A. dos. *Modelo Preditivo de Risco de Crédito para Cooperativas de Agronegócio*. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2022.

RIBEIRO, H. da S. *Classificação de clientes utilizando mineração de dados*. Dissertação (Mestrado) — Pontifícia Universidade Católica de Goiás, Goiânia, 2020.

SANTOS, P. F. dos. *Uso de técnicas de machine learning para análise de risco de crédito*. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2022.

SERASA. *Número de inadimplentes cai pelo segundo mês seguido, diz Serasa*. 2023. Disponível em: <https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/>. Último acesso em: 30/05/2023.

SETTLES, B. Active learning literature survey. *University of Wisconsin Madison*, 2010. Disponível em: <https://burrsettles.com/pub/settles.activelearning.pdf>.

SILVA, J. S. *Gerenciamento Integrado de Riscos: Modelos de Predição de Risco de Crédito em Machine Learning para a dentificação de Ativos Problemáticos em uma Instituição Financeira – Segmento Habitacional PF*. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2022.

SOUZA, R. C. T. de. *Uma metodologia para classificação de dados nominais baseada no processo kdd: ênfase aos algoritmos culturais, estimação de distribuição e análise*

de correspondência múltipla. Tese (Doutorado) — Universidade Federal do Paraná, Programa de PósGraduação em Métodos Numéricos em Engenharia, Curitiba, 2013.

UCI. *UCI Machine Learning Repository*. 2023. Disponível em: <https://archive.ics.uci.edu/>. Acesso em: 20/08/2023.

WEKA. *Weka Wiki*. 2023. Disponível em: <https://waikato.github.io/weka-wiki/documentation/>. Acesso em: 09/07/2023.