

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA MECATRÔNICA

ARTHUR FRAGA KRELLING

PREDIÇÃO DE INADIMPLÊNCIA EM FATURAS DA COMPANHIA DE ÁGUAS DE
JOINVILLE COM O USO DE APRENDIZADO DE MÁQUINA

Joinville
2025

ARTHUR FRAGA KRELLING

PREDIÇÃO DE INADIMPLÊNCIA EM FATURAS DA COMPANHIA DE ÁGUAS DE
JOINVILLE COM O USO DE APRENDIZADO DE MÁQUINA

Trabalho apresentado como requisito parcial para obtenção do título de bacharel em Engenharia Mecatrônica, no Curso de Engenharia Mecatrônica, do Centro Tecnológico de Joinville, da Universidade Federal de Santa Catarina.

Orientador: Dr. Benjamin Grando Moreira

Joinville
2025

AGRADECIMENTOS

Primeiramente, estendo meus agradecimentos à Companhia de Águas de Joinville (CAJ), cuja colaboração institucional foi fundamental ao fornecer os dados históricos que viabilizaram a execução e a análise central deste trabalho.

Minha profunda gratidão ao meu orientador, Dr. Benjamin Grando Moreira. Sou grato não apenas por todo o conhecimento técnico e científico compartilhado durante as disciplinas e a realização deste projeto, mas, principalmente, por acreditar na minha conclusão do curso quando nem eu mesmo acreditava. Sua paciência, seu esforço contínuo e sua bondade foram essenciais e possibilitaram que eu concluísse minha formação.

Agradeço de coração a toda minha família, especialmente à minha mãe, Lucélia, ao meu pai, Adalberto, e às minhas irmãs, Amanda, Camila e Barbara. Sem o apoio incondicional e o carinho de vocês, eu jamais teria chegado até aqui.

RESUMO

A inadimplência no pagamento de contas de água representa um desafio significativo para as concessionárias de saneamento, afetando diretamente o planejamento financeiro e a qualidade dos serviços prestados à população. Neste contexto, técnicas de machine learning têm se mostrado promissoras para prever a probabilidade de inadimplência, possibilitando a adoção de estratégias proativas de gestão e contribuindo para a redução de perdas financeiras. Apresenta-se neste estudo uma análise dos dados históricos de faturamento referentes ao ano de 2024, provenientes de unidades consumidoras da Companhia Águas de Joinville. O principal objetivo é desenvolver e treinar um modelo preditivo capaz de identificar, com acurácia satisfatória, consumidores com maior propensão à inadimplência, auxiliando na tomada de decisões estratégicas por parte da empresa. A síntese dos achados empíricos revelou a progressão da performance mediante a aplicação de técnicas de pré-processamento. A abordagem baseline, estabelecida com dados brutos, fixou o ponto de partida com baixa performance; o maior score F1 atingido foi de 0.241, obtido pelo modelo Gradient Boosting. A inclusão de atributos derivados resultou em uma melhora marginal no desempenho geral, sendo o maior score F1 alcançado de 0.255, com o modelo Random Forest, confirmando que a Engenharia de Atributos sozinha não foi suficiente para superar a assimetria dos dados. A abordagem mais transformadora foi o Downsampling, que alterou o foco da otimização para o Recall. Com esta técnica, o modelo alcançou um alto Recall, com o modelo Gradient Boosting atingindo o maior valor de 0.862, porém, incorreu em uma baixa Precisão.

Palavras-chave: Machine Learning. Inadimplência. Saneamento Básico. Modelo Preditivo.

ABSTRACT

Delinquency in water bill payments represents a significant challenge for sanitation utility companies, directly impacting financial planning and the quality of services provided to the population. In this context, machine learning techniques have proven promising for predicting the probability of default, enabling the adoption of proactive management strategies and contributing to the reduction of financial losses. This study presents an analysis of historical billing data spanning 24 months, originating from consumer units of Companhia Águas de Joinville. The main objective is to develop and train a predictive model capable of identifying, with satisfactory accuracy, consumers with a higher propensity for delinquency, thereby assisting in the company's strategic decision-making. The methodology employed utilized classification algorithms, including Random Forest, Gradient Boosting, and Neural Network, with the process being challenged by the extreme class imbalance present in the billing database. Three distinct approaches were compared: the Baseline, the inclusion of engineered features, and balancing via Downsampling. Initial approaches, focused on the F1-score, yielded limited results; the highest F1-score attained in the Baseline was 0.241 (Gradient Boosting), and with Derived Features, it was 0.255 (Random Forest). The most transformative approach was Downsampling, which shifted the focus to Recall. With this technique, the model achieved a high Recall, with the Gradient Boosting model reaching the highest score of 0.862. However, this occurred at the expense of Precision, demonstrating the crucial operational trade-off for the utility company: high risk detection versus high operational cost caused by False Positives. The study concludes that these models are viable tools for proactive risk management, but suggests the future integration of external socio-economic data to enhance predictive reliability.

Keywords: Machine Learning. Delinquency. Basic Sanitation. Predictive Model.

LISTA DE FIGURAS

Figura 1 – Exemplo de um fragmento da base de dados após o tratamento . . .	24
Figura 2 – Base de dados pivotada	25
Figura 3 – Exemplo de fluxo de trabalho na Orange	26
Figura 4 – Parâmetros de configuração dos modelos preditivos	27
Figura 5 – Distribuição temporal de inadimplentes por data de referência	29
Figura 6 – Fluxo de trabalho implementado na ferramenta Orange para Análise Baseline	30
Figura 7 – Classificação dos atributos para abordagem Baseline	31
Figura 8 – Classificação dos atributos para abordagem Atributos Derivados . .	32
Figura 9 – Fluxo de trabalho implementado na ferramenta Orange para Análise com Downsampling	34

LISTA DE TABELAS

Tabela 1 – Distribuição de Atrasos	28
Tabela 2 – Comparativo de scores dos modelos na abordagem Baseline	31
Tabela 3 – Matriz de Confusão do modelo GB na abordagem Baseline	32
Tabela 4 – Comparativo de scores dos modelos na abordagem Atributos Derivados	32
Tabela 5 – Variação do score F1 em função do número de atributos selecionados	33
Tabela 6 – Variação do score Recall em função do número de atributos selecionados	33
Tabela 7 – Comparativo de scores dos modelos na abordagem com Downsampling	35
Tabela 8 – Matriz de Confusão do modelo GB na abordagem com Downsampling	35
Tabela 9 – Comparativo do Score F1 entre as diferentes abordagens	37
Tabela 10 – Comparativo do Score Precision entre as diferentes abordagens . .	37
Tabela 11 – Comparativo do Score Recall entre as diferentes abordagens	37

SUMÁRIO

1	INTRODUÇÃO	9
1.1	Objetivo	10
1.1.1	Objetivo Geral	10
1.1.2	Objetivos Específicos	10
2	FUNDAMENTAÇÃO TEÓRICA	11
2.1	Saneamento Básico	11
2.2	Concessionárias de Água e Modelos de Negócio	12
2.2.1	Geração de Receita	12
2.2.2	Impacto da Inadimplência no Fluxo de Caixa	13
2.2.3	Principais Causas da Inadimplência	13
2.3	Companhia Águas de Joinville	14
2.4	Machine Learning para Previsão de Inadimplência	14
2.4.1	Métricas de Avaliação	15
2.4.2	Implicações Estratégicas das Métricas de Avaliação	16
2.5	Técnicas de Análise de Dados	17
2.5.1	Pivotagem e Reorganização dos Dados	17
2.5.2	Atributos Derivados e Engenharia de Atributos	18
2.5.3	Downsampling e Desbalanceamento de Classes	18
2.5.4	Trabalhos Similares	19
3	MATERIAIS E MÉTODOS	21
3.1	Base de dados	21
3.2	Tratamento dos Dados	22
3.2.1	Formatação	22
3.2.2	Filtragem	22
3.2.3	Atributos Derivados	23
3.2.4	Base de Dados Transformada	25
3.2.5	Software Utilizado Para a Análise	26
3.2.6	Validação dos Modelos	26
4	ANÁLISES E RESULTADOS	28
4.1	Análise Exploratória dos Dados	28
4.2	Resultados	29
4.2.1	Resultados da Abordagem Baseline	29
4.2.2	Resultados com Atributos Derivados	31
4.2.3	Análise com Downsampling	33

4.3	COMPARAÇÃO GERAL DAS ABORDAGENS	35
5	CONCLUSÕES	39
	REFERÊNCIAS	41

1 INTRODUÇÃO

O setor de saneamento básico desempenha um papel fundamental na promoção da saúde pública e no desenvolvimento socioeconômico das cidades. A gestão eficiente dos recursos hídricos e a prestação de serviços de abastecimento de água tratada são essenciais para garantir qualidade de vida à população. Dentro desse contexto, a sustentabilidade financeira das concessionárias de água torna-se crucial para a manutenção e a expansão dos serviços prestados.

Dentre os diversos desafios enfrentados pelas concessionárias de saneamento, a inadimplência no pagamento das faturas de consumo se destaca como um dos mais relevantes. Este trabalho se insere no campo da aplicação de técnicas de machine learning para a gestão de riscos financeiros em serviços públicos, mais especificamente, na previsão de inadimplência no setor de abastecimento de água. Estudos recentes têm mostrado que a utilização de algoritmos preditivos pode contribuir significativamente para antecipar comportamentos de pagamento e melhorar as estratégias de cobrança.

Dentre os diversos problemas existentes na gestão do faturamento das concessionárias, este trabalho foca especificamente na seguinte questão: é possível prever com acurácia satisfatória as unidades consumidoras com maior probabilidade de inadimplência utilizando modelos de machine learning treinados com dados históricos de faturamento? Pesquisas anteriores sugerem que modelos preditivos podem atingir níveis consideráveis de precisão em diferentes setores, mas ainda há a necessidade de adaptar essas soluções ao contexto específico de concessionárias de água, considerando as particularidades de consumo e comportamento de pagamento.

A literatura analisada demonstra que a previsão de atrasos e defaults por meio de Machine Learning é um campo robusto, porém sem uma solução algorítmica universal. Os estudos de DONG, LIU e THAM (2024) e BASHAR et al. (2023) reforçam a premissa do Teorema do "No Free Lunch" PENG et al. (2011), indicando que a performance ideal varia significativamente entre modelos (Random Forest, Decision Tree, Gradient Boosting) e é altamente sensível ao volume, tipo e tratamento dos dados. Adicionalmente, observa-se que os trabalhos similares em predição financeira comumente empregam uma gama maior de features, frequentemente incorporando dados detalhados sobre a vida financeira e socioeconômica dos consumidores.

Diante desse contexto, o objetivo geral desta pesquisa é desenvolver e avaliar um modelo preditivo, utilizando técnicas de machine learning, para identificar unidades consumidoras com alto risco de inadimplência, contribuindo para a melhoria da gestão de faturamento da Companhia Águas de Joinville.

Para atingir esse objetivo, será realizada uma análise exploratória dos dados

históricos de faturamento do ano de 2024, seguida do pré-processamento das informações, seleção de atributos relevantes e treinamento de modelos de machine learning.

O conjunto de dados utilizado para a modelagem preditiva é composto por 2.478.768 registros, os quais representam as faturas emitidas, correspondentes a 103.282 usuários únicos no sistema, abrangendo um período de 24 meses. A base de dados contém 8 atributos, que englobam variáveis como a Matrícula, a Data Referência, a Data Vencimento e o Valor Fatura. Dentro do período total analisado, o volume consolidado de inadimplência corresponde a 26.511 faturas em default. Os atributos de data foram cruciais para a construção de features históricas, permitindo a análise da evolução do comportamento de pagamento das unidades consumidoras ao longo do tempo.

1.1 OBJETIVO

Para prever possíveis usuários inadimplentes da Companhia de Águas de Joinville, propõe-se neste trabalho os seguintes objetivos.

1.1.1 Objetivo Geral

Desenvolver e avaliar um modelo preditivo baseado em aprendizado de máquina para identificar os potenciais usuários inadimplentes, utilizando o histórico de faturas das unidades consumidoras atendidas pela Companhia de Águas de Joinville.

1.1.2 Objetivos Específicos

- Conduzir uma análise exploratória dos dados para compreender a distribuição, a natureza e as características do histórico de faturas;
- Discernir e validar as variáveis cruciais que demonstrem maior correlação com o evento de inadimplência, estabelecendo a base para a predição;
- Estruturar e refinar o conjunto de dados, realizando tarefas de limpeza, transformação de variáveis e engenharia de atributos para otimizar o processamento;
- Construir e parametrizar modelos preditivos, aplicando uma seleção diversificada de algoritmos de aprendizado de máquina;
- Avaliar o desempenho e comparar a eficácia dos modelos preditivos.

2 FUNDAMENTAÇÃO TEÓRICA

A construção e a avaliação rigorosa de um modelo preditivo demandam o conhecimento aprofundado dos fundamentos básicos da área de estudo, bem como a compreensão do contexto de negócio ao qual a solução se destina. Neste capítulo, são abordadas as questões gerais relativas ao funcionamento das concessionárias de água e saneamento e a estrutura de sua receita. Em seguida, é apresentada uma breve retrospectiva da história e do papel da Companhia de Águas de Joinville no município. Posteriormente, são discutidos os conceitos teóricos elementares empregados para o desenvolvimento e a análise dos modelos de predição utilizados neste trabalho, focando no Aprendizado de Máquina e nas técnicas de preparação de dados. Finalmente, o capítulo conclui com uma análise de trabalhos similares que aplicaram metodologias análogas em contextos de previsão de risco de pagamento.

2.1 SANEAMENTO BÁSICO

O saneamento básico, que abrange o conjunto de serviços, infraestruturas e instalações de abastecimento de água, esgotamento sanitário, limpeza urbana e manejo de resíduos sólidos e de drenagem de águas pluviais, é um pilar fundamental para o desenvolvimento de uma sociedade. A garantia do acesso à água potável é um dos principais desafios globais e um direito humano fundamental, conforme reconhecido pela Organização das Nações Unidas (ONU) em 2010 (ORGANIZAÇÃO DAS NAÇÕES UNIDAS, 2010).

A ausência de saneamento adequado é um fator de risco significativo para a saúde. A falta de acesso a serviços de saneamento básico e água potável de qualidade está associada ao aumento de diversas doenças. A Organização Mundial da Saúde (OMS) estima que a água insegura, o saneamento e a higiene precários são responsáveis pela morte de cerca de 1.000 crianças menores de 5 anos todos os dias (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2019b). Além disso, a literatura demonstra que o saneamento e a higiene são cruciais para a saúde materna e infantil, pois as infecções, muitas vezes ligadas a nascimentos em ambientes insalubres, representam 26% das mortes neonatais e 11% da mortalidade materna, totalizando 1 milhão de óbitos por ano (ORGANIZAÇÃO MUNDIAL DA SAÚDE, 2019a).

Além dos benefícios diretos para a saúde, o saneamento básico tem um impacto socioeconômico significativo. A universalização desses serviços promove a dignidade humana, a equidade e a inclusão social. A ausência de uma infraestrutura adequada afeta desproporcionalmente as populações mais vulneráveis, perpetuando ciclos de pobreza e marginalização. A Organização Mundial da Saúde ORGANIZAÇÃO

MUNDIAL DA SAÚDE (2012) estima que as perdas econômicas globais anuais associadas à falta de água e saneamento adequados chegam a US\$ 260 bilhões, o que representa 1,5% do Produto Interno Bruto (PIB) dos países analisados no estudo.

Ainda de acordo com a mesma fonte, o investimento em saneamento gera um retorno econômico considerável. Os benefícios incluem a redução de doenças diarreicas e da desnutrição, além da economia de tempo para as comunidades. A melhoria da saúde da população leva a menos gastos com assistência médica, à redução da perda de tempo produtivo devido a doenças e à diminuição da mortalidade prematura. Um exemplo notável desse retorno é a relação custo-benefício do investimento em saneamento, que chega a US\$ 5,5 de retorno para cada US\$ 1,00 investido globalmente.

2.2 CONCESSIONÁRIAS DE ÁGUA E MODELOS DE NEGÓCIO

As concessionárias de água e saneamento são entidades, sejam elas de natureza pública (empresas estaduais, autarquias) ou privada, responsáveis pela prestação de serviços públicos essenciais de infraestrutura. Estas empresas operam sob o regime de delegação ou concessão do Poder Público (titular dos serviços), sendo encarregadas do abastecimento de água potável, esgotamento sanitário, drenagem e manejo de águas pluviais, e gestão de resíduos sólidos (INSTITUTO RUI BARBOSA, 2025). O setor é caracterizado por ser um monopólio natural regulamentado, exigindo um arcabouço legal extenso e um intenso volume de capital para investimentos de longo prazo em redes de distribuição e estações de tratamento (ALBUQUERQUE, 2011).

2.2.1 Geração de Receita

O principal pilar financeiro de uma concessionária de água e saneamento está na cobrança das tarifas pelo serviço essencial de distribuição de água tratada e coleta/tratamento de esgoto. A estrutura tarifária é definida por uma agência reguladora e se baseia fundamentalmente no volume de água consumido (medido por hidrômetros), complementada por uma tarifa mínima e a cobrança específica pelo serviço de esgoto, geralmente um percentual da tarifa de água (BATISTA; CAMPOLI, 2023). As receitas principais são, portanto, as Receitas de Água e Receita de Esgoto (ALBUQUERQUE, 2011).

Este caráter regulado assegura uma previsibilidade de receita, uma vez que os reajustes e as revisões tarifárias são periodicamente ajustados para cobrir os custos operacionais e de investimento, proporcionando o retorno sobre o capital exigido pelo setor. No entanto, para a sustentabilidade do modelo, que exige investimentos de longo prazo em infraestrutura, a gestão eficiente dos custos e a garantia do retorno financeiro

são críticas, sendo a tarifa o instrumento chave para cobrir o custeio e viabilizar o CAPEX (investimento em capital).

2.2.2 Impacto da Inadimplência no Fluxo de Caixa

A inadimplência dos usuários constitui um dos maiores riscos operacionais e financeiros para as concessionárias, impactando diretamente o fluxo de caixa operacional. A não materialização da receita esperada reduz a liquidez da empresa, dificultando o cumprimento de obrigações de curto prazo, como o pagamento de fornecedores (energia e insumos químicos) e a manutenção essencial da infraestrutura (AESBE, 2020). Essa situação não apenas afeta o resultado contábil, mas também exige que a empresa provisione recursos para devedores duvidosos.

Em um contexto de necessidade constante de investimento em universalização, a perda de receita por inadimplência compromete seriamente a capacidade de investimento da concessionária em projetos de expansão e melhoria da qualidade do saneamento. Taxas de inadimplência elevadas, que já atingiram cerca de 25% em algumas operadoras AESBE (2020), criam um desequilíbrio financeiro que pode, indiretamente, onerar os usuários adimplentes, demandando políticas de cobrança rigorosas e, em casos extremos, o corte do fornecimento para mitigar o dano financeiro.

2.2.3 Principais Causas da Inadimplência

As causas da inadimplência no setor de saneamento são complexas, sendo as principais: a vulnerabilidade socioeconômica, o consumo variável e a sazonalidade.

- **Vulnerabilidade Socioeconômica:** É o fator dominante, onde famílias de baixa renda enfrentam instabilidade financeira e são forçadas a priorizar outras necessidades básicas, resultando na dificuldade ou incapacidade de pagar as faturas de água e esgoto (ONDAS, 2020). Essa situação está intimamente ligada à distribuição da pobreza e à falta de acesso em áreas mais carentes EBC (2024), sendo mitigada, em parte, pela implementação de programas de Tarifa Social.
- **Consumo Variável:** A elevação inesperada do valor da fatura, frequentemente causada por vazamentos internos não detectados a tempo ou por um uso incomum da água, gera contestações e, conseqüentemente, o não pagamento por parte do usuário até que a situação seja revista. Mudanças abruptas nas regras de medição ou na inclusão de novos serviços (como a cobrança de esgoto) também podem gerar resistência e inadimplência.
- **Sazonalidade:** Em menor grau, a variação do consumo em períodos de seca ou calor, que pode levar a um aumento significativo no volume faturado, exerce pressão adicional sobre o orçamento familiar, elevando pontualmente as taxas de inadimplência em períodos específicos do ano.

2.3 COMPANHIA ÁGUAS DE JOINVILLE

A Companhia Águas de Joinville (CAJ) é uma empresa pública do município de Joinville, Santa Catarina, com a responsabilidade crucial de gerenciar os serviços de saneamento básico da cidade. Sua atuação é abrangente, cobrindo todo o ciclo da água e do esgoto para os cerca de 600 mil habitantes do município (COMPANHIA ÁGUAS DE JOINVILLE, 2025).

A CAJ é encarregada de explorar diretamente os serviços públicos de abastecimento de água e de esgotamento sanitário. Isso engloba desde a captação de água bruta, o tratamento, a adução, a reservação e a distribuição para consumo público, até a coleta de esgotos sanitários, seu transporte, tratamento, reaproveitamento e disposição final, além de buscar soluções alternativas para esses processos.

Desde o início de suas atividades em 2005, a Companhia Águas de Joinville tem expandido significativamente sua infraestrutura e capacidade operacional. Atualmente, a CAJ faz a gestão de duas estações de tratamento de água (ETAs), 13 estações de tratamento de esgoto (ETEs), 13 reservatórios e cinco unidades de atendimento.

O crescimento da Companhia também se reflete em sua força de trabalho. Inicialmente com 40 funcionários contratados, a CAJ hoje conta com mais de 500 colaboradores próprios, além de uma equipe de funcionários terceirizados que executam serviços essenciais como obras, consertos de vazamentos, leitura de hidrômetros e teleatendimento.

2.4 MACHINE LEARNING PARA PREVISÃO DE INADIMPLÊNCIA

O Machine Learning (Aprendizado de Máquina) representa um subcampo da Inteligência Artificial que se dedica ao desenvolvimento de algoritmos capazes de aprender padrões e fazer previsões a partir de dados, dispensando a programação explícita para cada tarefa. O objetivo central é que o desempenho do sistema em uma tarefa específica melhore com o aumento da experiência, conforme o modelo clássico de (MITCHELL, 1997). Esta capacidade de generalização torna o Machine Learning uma ferramenta estratégica para concessionárias de água, permitindo a previsão de risco de crédito por meio da análise do histórico de consumo e pagamento dos usuários. Ao identificar consumidores com alta probabilidade de inadimplência, a concessionária pode transicionar de uma gestão reativa para uma abordagem proativa de mitigação de perdas financeiras.

A modelagem da inadimplência é formalizada como um problema de Aprendizado Supervisionado (Supervised Learning). Neste paradigma, o algoritmo é treinado com um conjunto de dados onde cada cliente possui um rótulo de saída que indica seu status de pagamento histórico. O modelo aprende a função de mapeamento que conecta as variáveis de entrada (e.g., média de consumo, atrasos) ao rótulo de

saída, ajustando seus parâmetros para minimizar o erro de previsão. Especificamente, o problema se configura como uma Classificação Binária (Binary Classification), na qual o modelo deve alocar cada observação em uma das duas categorias mutuamente exclusivas: Classe Positiva (inadimplente, o evento de interesse) ou Classe Negativa (adimplente). O resultado é uma pontuação de probabilidade que, após a aplicação de um limiar, resulta na decisão de risco (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Para que o modelo de classificação binária atinja alta acurácia preditiva, a Engenharia de Atributos (Feature Engineering) é indispensável. Esta etapa crítica envolve a transformação e a criação de variáveis preditivas (features) a partir dos dados brutos, visando aumentar o poder discriminatório do algoritmo (TAYLOR & FRANCIS GROUP, 2019). Em vez de utilizar apenas valores absolutos, a extração de atributos derivados, como a taxa de variação percentual do consumo ou a frequência de atrasos no último semestre, permite que o modelo capture o comportamento dinâmico do consumidor. Uma Engenharia de Atributos robusta e semanticamente rica é crucial, devendo ser conduzida com cautela para mitigar o risco de Data Leakage (vazamento de dados futuros no treinamento), garantindo que os atributos sejam calculados apenas com informações disponíveis no momento da previsão.

Apesar da sua relevância, a modelagem preditiva no setor de serviços enfrenta o severo Problema de Classes Desbalanceadas (Imbalanced Classes). Isto ocorre porque, no universo total de clientes analisados, os consumidores inadimplentes representam uma parcela numérica significativamente reduzida (classe minoritária) em comparação com os clientes adimplentes (classe majoritária). Este desequilíbrio leva a um viés algorítmico, onde o modelo tende a favorecer a classe majoritária, resultando em modelos com alta acurácia superficial, mas ineficazes na identificação real dos casos de inadimplência (HE; GARCIA, 2009). Para contornar esse desafio metodológico, são aplicadas técnicas de reamostragem, como undersampling (subamostragem) da classe majoritária e oversampling (sobreamostragem) da classe minoritária, sendo o SMOTE (Synthetic Minority Over-sampling Technique) uma alternativa popular para a criação de instâncias sintéticas, buscando equilibrar o conjunto de treinamento e aumentar a capacidade de sensibilidade (Recall) do modelo na identificação da classe de interesse.

2.4.1 Métricas de Avaliação

Em contextos de classes desbalanceadas, a avaliação do desempenho do modelo exige cautela. A Acurácia (Accuracy), definida como a proporção de previsões corretas sobre o total de observações, é uma métrica insuficiente e frequentemente enganosa, visto que um alto valor pode ser alcançado apenas pela correta classificação da classe majoritária. Torna-se, portanto, essencial utilizar um conjunto de métricas mais detalhadas, todas derivadas da Matriz de Confusão (HE; GARCIA, 2009).

A Matriz de Confusão é uma tabela de contingência que resume os resultados de um modelo de classificação, cruzando os valores de classe reais com os valores de classe preditos. Ela permite a quantificação precisa dos Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN) (ALPAYDIN, 2020). É a partir desses quatro componentes que se podem calcular métricas mais informativas, como Precision e Recall, que detalham a performance do modelo na identificação da classe de interesse. HAN, KAMBER e PEI (2012) definem um conjunto de métricas de desempenho relevantes para cenários de classificação, incluindo as seguintes:

- **Precision (Precisão):** Esta métrica avalia a pureza das previsões positivas feitas pelo modelo. Ela é calculada como a razão entre os Verdadeiros Positivos (TP) e o total de observações que o modelo classificou como positivas ($TP + FP$). A Precision indica, portanto, a probabilidade de um resultado positivo predito ser, de fato, correto.
- **Recall:** O Recall, ou sensibilidade, avalia a completude da identificação da classe positiva. É a razão entre os Verdadeiros Positivos (TP) e o total de observações que pertencem à Classe Positiva na realidade ($TP + FN$). Um alto Recall significa que o modelo conseguiu identificar a maioria dos casos positivos reais.
- **F1-Score:** O F1-Score representa a média harmônica entre Precision e Recall. Ao combinar essas duas medidas, ele oferece um indicador único de desempenho que penaliza modelos que exibem um desequilíbrio significativo entre Precision e Recall. Por esta razão, o F1-Score é frequentemente considerado uma métrica mais robusta do que a acurácia para avaliação de classificadores em bases desbalanceadas.

2.4.2 Implicações Estratégicas das Métricas de Avaliação

A escolha e a interpretação das métricas em um problema de classificação desbalanceada, como a previsão de inadimplência, exigem uma análise cuidadosa dos custos assimétricos associados aos erros. A meta não é apenas obter o melhor resultado estatístico, mas sim selecionar um modelo cujos trade-offs minimizem o prejuízo decorrente da falha de classificação (HE; GARCIA, 2009).

A métrica Recall (Sensibilidade) mede a capacidade do modelo de identificar corretamente os casos positivos reais (Verdadeiros Positivos - TP), o que está ligado à minimização de Falsos Negativos (FN). Por outro lado, a métrica Precision (Precisão) é usada para mitigar os Falsos Positivos (FP), que ocorrem quando um cliente adimplente é erroneamente sinalizado como de risco. Em problemas de risco, onde os custos de um FN (falha em detectar a inadimplência) e de um FP (erro ao sinalizar um cliente adimplente) são desiguais, a escolha da métrica prioritária é guiada pela minimização do custo total de misclassificação.

A determinação de qual métrica (Recall alto ou Precision alta) traria maior benefício para a concessionária depende da Matriz de Custo, que quantifica o prejuízo financeiro e operacional de cada tipo de erro (HE; GARCIA, 2009). No entanto, a ausência de uma Matriz de Custo previamente definida e validada para o caso da Companhia de Águas de Joinville impede a quantificação e a comparação do custo real de um Falso Negativo versus um Falso Positivo. Consequentemente, a priorização da otimização entre Recall e Precision não pode ser estabelecida com base em um cálculo financeiro neste estudo.

Para contornar esta limitação metodológica e preparar o terreno para futuras implementações, seria necessária a utilização de técnicas mais avançadas, como as propostas por (BOCK; COUSSEMENT; LESSMANN, 2020). O estudo em questão propõe o uso de otimização multicritério para gerar um conjunto de modelos ensemble que mapeiam diferentes trade-offs de erros no espaço de custo. Essa abordagem é notável por permitir que os modelos sejam treinados na ausência da Matriz de Custo, de modo que o analista possa utilizar o conjunto de modelos gerado para selecionar o candidato ideal com base no custo conhecido no momento da predição. No entanto, a exploração dessa metodologia avançada de otimização de matrizes de custo foge do escopo definido para esta pesquisa.

2.5 TÉCNICAS DE ANÁLISE DE DADOS

A qualidade e a organização dos dados são fatores determinantes para o sucesso de qualquer projeto de Machine Learning, uma vez que algoritmos preditivos são altamente sensíveis à forma como a informação lhes é apresentada. A etapa de Análise e Preparação de Dados (Data Preprocessing) é, portanto, crucial, pois envolve a limpeza, a transformação e a estruturação dos dados brutos em um formato adequado para o consumo do modelo (TAYLOR & FRANCIS GROUP, 2019). A aplicação rigorosa de técnicas como a pivotagem para reorganização de dados, a engenharia de atributos para extração de valor preditivo e o tratamento do desbalanceamento de classes não apenas corrige inconsistências e vieses, mas também maximiza a capacidade de generalização do modelo, garantindo a robustez e a confiabilidade dos resultados obtidos.

2.5.1 Pivotagem e Reorganização dos Dados

A organização dos dados é crucial para a modelagem preditiva, sendo comumente categorizada em dois formatos: long e wide. O formato long (longo) é caracterizado por ter múltiplas linhas para a mesma unidade de observação (e.g., um cliente), com uma coluna indicando o tipo de variável e outra contendo seu respectivo valor. Já o formato wide (largo) representa cada variável em uma coluna separada, de

modo que cada unidade de observação se encontra em uma única linha (WICKHAM, 2014).

A pivotagem (pivoting) é o processo de transformação que converte dados do formato long para o formato wide. Esta conversão é fundamental, pois a maioria dos modelos tabulares de Machine Learning (como Gradient Boosting ou Random Forests) requer que cada instância de treinamento (cliente) esteja em uma única linha, com todas as suas características dispostas em colunas. Pivotar permite consolidar uma série temporal de dados transacionais, como o histórico de consumo mensal, em variáveis estáticas e preditivas, como o consumo médio ou o desvio padrão de atraso, que são essenciais para alimentar o modelo.

2.5.2 Atributos Derivados e Engenharia de Atributos

A Engenharia de Atributos (Feature Engineering) é o processo técnico e criativo de transformar dados brutos em variáveis preditivas (atributos derivados) com maior poder discriminatório. A eficácia do modelo preditivo está diretamente ligada à capacidade dessa etapa de sintetizar o histórico e as relações entre as variáveis (TAYLOR & FRANCIS GROUP, 2019). Em vez de utilizar apenas o valor absoluto de uma transação, atributos derivados, como a taxa de variação de um pagamento ou a frequência de um evento, permitem ao algoritmo capturar a semântica do risco que os dados brutos não expressam diretamente.

Um risco crítico na criação de atributos é o Data Leakage (Vazamento de Dados), que ocorre quando informações do conjunto de teste (dados futuros) são inadvertidamente utilizadas no cálculo de features do conjunto de treinamento. Isso resulta em uma performance de treinamento excessivamente otimista (KAUFMAN et al., 2012). Para evitar o leakage, a Engenharia de Atributos deve ser aplicada seguindo o princípio temporal: os atributos de uma observação devem ser calculados somente com dados que estariam disponíveis no momento da previsão. Uma boa prática é isolar o cálculo de atributos dentro do ciclo de validação cruzada para garantir a estrita separação entre os conjuntos de dados.

2.5.3 Downsampling e Desbalanceamento de Classes

O Imbalanced Learning (Aprendizado com Dados Desbalanceados) é uma situação comum em tarefas de classificação, como a previsão de inadimplência, onde a classe minoritária (inadimplentes) é significativamente sub-representada. Este cenário é problemático porque algoritmos de Machine Learning, otimizados para a acurácia geral, tendem a ser enviesados para a classe majoritária. Consequentemente, a Acurácia torna-se uma métrica enganosa, pois um valor alto pode ser alcançado pelo simples fato de o modelo classificar corretamente a maioria dos casos negativos, falhando na

identificação da classe de interesse (HE; GARCIA, 2009).

Para contornar o viés e aprimorar a capacidade de identificação da classe minoritária, são aplicadas técnicas de reamostragem. O Undersampling (Subamostragem) consiste na redução do número de instâncias da classe majoritária. O Oversampling (Sobreamostragem) envolve a replicação de instâncias da classe minoritária. O SMOTE (Synthetic Minority Over-sampling Technique) é uma técnica avançada de oversampling que cria instâncias sintéticas para a classe minoritária, interpolando entre instâncias vizinhas. O uso dessas técnicas é essencial para que o modelo aprenda a fronteira de decisão da classe de interesse, aumentando sua sensibilidade (Recall) (HE; GARCIA, 2009).

2.5.4 Trabalhos Similares

A produção acadêmica sobre a aplicação de métodos preditivos para a gestão do risco financeiro tem se expandido significativamente na última década (DONG; LIU; THAM, 2024). Este campo de estudo concentra-se na utilização de algoritmos avançados para antecipar atrasos ou defaults no pagamento de contas e empréstimos em diversos setores, com o intuito de aprimorar a capacidade de decisão e mitigar perdas. O consenso acadêmico estabelece que a eficácia de um modelo é altamente dependente do contexto específico e da natureza dos dados. Nesse sentido, o Teorema do "No Free Lunch" (No Free Lunch Theorem) sugere que não existe um único algoritmo de classificação que seja o melhor para todas as situações, exigindo-se a comparação e a adaptação metodológica para cada problema em particular (PENG et al., 2011).

O estudo conduzido por DONG, LIU e THAM (2024) exemplifica essa necessidade de avaliação comparativa, ao analisar o desempenho de cinco distintos modelos de classificação: k-nearest neighbour, Naïve-Bayes, Decision Tree, Regressão Logística e Random Forest. Os autores utilizaram uma base de dados de cartões de crédito para prever o default de clientes em pagamentos de empréstimo. O trabalho conclui que os diferentes modelos apresentam desempenhos variados em função do volume, tipo e tratamento dos dados. Por exemplo, o modelo Decision Tree demonstrou uma performance superior com volumes de dados menores, mas resultou em um desempenho insatisfatório ao ser aplicado a um conjunto de dados de grande escala. Os resultados mais robustos foram atingidos pelo algoritmo Random Forest, que obteve um F1-score de 0.8424 e uma acurácia de 0.9802.

De forma complementar, BASHAR et al. (2023) realizaram uma pesquisa com o objetivo de prever a propensão a pagar de consumidores de uma companhia de energia elétrica australiana. Os autores empregaram uma variedade de modelos de Machine Learning, incluindo Decision Tree, Random Forest, Neural Network, Gradient Boosting e Regressão Logística. A análise detalhada das métricas revelou que o melhor score de Precision (0.76) foi obtido com a Neural Network, enquanto o Recall mais

elevado (0.82) foi alcançado pelo Random Forest. O melhor desempenho geral, medido pelo F1-score (0.68), foi atribuído ao algoritmo Gradient Boosting, reforçando o fato de que a escolha do melhor algoritmo é uma função da métrica de avaliação priorizada.

Em comparação com os trabalhos supracitados, o presente estudo enfrenta uma limitação metodológica notável, que se manifesta na restrição das features iniciais disponíveis para a modelagem. Enquanto outros trabalhos similares, notadamente aqueles focados em risco de crédito, contam com informações mais detalhadas e abrangentes sobre a vida financeira dos consumidores, como score de crédito, renda mensal ou histórico de empréstimos, esta pesquisa se restringiu primariamente aos dados transacionais internos da concessionária (consumo e histórico de pagamento). As métricas de desempenho alcançadas — F1-Score e Recall — não foram tão elevadas quanto as observadas nos estudos de referência, o que sugere que a falta de features preditivas externas sobre a capacidade socioeconômica e a propensão de pagamento do cliente é um fator preponderante na limitação da performance dos modelos de previsão de inadimplência.

Por fim, torna-se relevante salientar que a Regressão Logística é uma abordagem frequentemente associada à previsão de atraso ou inadimplência, como evidenciado nos estudos de ZAIED et al. (2020) e (VÁSQUEZ; ALICEA-PLANAS, 2017). Adicionalmente, ela está frequentemente entre os modelos de Machine Learning empregados para essa finalidade, conforme demonstrado nos trabalhos de BASHAR et al. (2023) e (NASCIMENTO; MAIA; ALMADA, 2021).

3 MATERIAIS E MÉTODOS

O procedimento metodológico deste estudo foi executado através de quatro fases distintas e sucessivas. Inicialmente, ocorreu a obtenção e o entendimento do conjunto de dados, que envolveu a inspeção da composição das informações de faturamento e a seleção das variáveis cruciais para a análise. Subsequentemente, implementou-se a fase de preparo e saneamento dos dados, abrangendo a padronização de formatos, a exclusão de elementos inconsistentes e a geração de features que condensassem o histórico de comportamento financeiro das unidades consumidoras.

Na terceira fase, os dados foram reestruturados em múltiplas configurações para a modelagem preditiva, com a criação de uma base de dados transformada (pivotada) por período mensal e de um dataset consolidado com indicadores estatísticos relacionados a atrasos e montantes de faturas. Finalmente, na quarta fase, procedeu-se ao treinamento e à aferição de múltiplos modelos de classificação, comparando diversas estratégias (baseline, atributos derivados e downsampling) com base em indicadores de performance como F1-Score, Recall e a matriz de confusão. Essa sequência de atividades permitiu a avaliação sistemática da capacidade preditiva dos algoritmos de machine learning na identificação de unidades com risco elevado de não cumprimento de obrigações financeiras.

Os detalhes de cada etapa são apresentados nas seções a seguir.

3.1 BASE DE DADOS

Os dados analisados, apresentados em formato tabular, compreendem um recorte histórico das faturas das unidades consumidoras atendidas pela Companhia de Águas de Joinville. A estrutura da tabela é organizada de modo que cada linha representa uma fatura específica, associada a uma determinada unidade consumidora e data de referência. A seguir, detalha-se a lista das colunas, juntamente com uma breve descrição do significado de cada atributo.

- **Matrícula com Máscara:** Identificador único da unidade consumidora. A aplicação de uma máscara criptográfica garante a preservação da privacidade do consumidor.
- **Categoria:** Classificação da unidade consumidora, categorizada como "Residencial", "Comercial", "Industrial" ou "Público".
- **Grupo de Faturamento:** Classificação numérica que agrupa as unidades consumidoras com base em regiões geográficas específicas de Joinville.

- **Referência:** Data (mês/ano) à qual a fatura se refere.
- **Data Vencimento:** Data limite para o pagamento da fatura.
- **Situação Fatura:** Indica o status da fatura no momento da extração dos dados, podendo ser "Quitada" ou "Pendente".
- **Data Situação:** Para faturas "Quitadas", representa a data de pagamento. Para faturas "Pendente", corresponde à data de extração dos dados.
- **Valor Fatura:** Montante monetário total da fatura a ser pago pela unidade consumidora.

Estes representam os atributos originais do conjunto de dados. Antes de sua utilização para a derivação de novas variáveis ou para o treinamento de qualquer modelo preditivo, faz-se necessário um processo de tratamento dos dados que abrange a formatação para facilitar a manipulação, a eliminação de inconsistências e erros humanos, e a filtragem de outliers e casos atípicos, como o parcelamento de faturas.

3.2 TRATAMENTO DOS DADOS

A análise inicial foi realizada diretamente nos dados por meio de consultas SQL executadas no software Superintendent.app¹. Dado o grande volume de informações, a escolha dessa ferramenta permitiu uma manipulação eficiente e otimizada dos dados.

O Superintendent.app é um aplicativo de desktop que permite executar consultas SQL diretamente em arquivos .csv, eliminando a necessidade de criar e configurar previamente um banco de dados.

SQL (Structured Query Language) é uma linguagem de programação padrão projetada para gerenciar e manipular dados em bancos de dados relacionais. Conforme descrito por BEAULIEU (2020), ela é amplamente utilizada para consultar, atualizar, inserir e excluir dados, sendo essencial para a interação com sistemas de gerenciamento de banco de dados.

3.2.1 Formatação

A formatação inicial necessária envolveu apenas a conversão das colunas "Referência", "Data Vencimento" e "Data Situação" do formato string para datetime, o que foi feito com um simples comando SQL.

3.2.2 Filtragem

Para garantir a qualidade do treinamento do modelo e a representatividade dos dados nas análises, é crucial evitar o uso de informações errôneas ou que não

¹ <https://superintendent.app>

correspondam aos casos de interesse. A seguir, a lista de filtros aplicados que acarretam na remoção de uma linha da tabela:

- **Linha com Dado Nulo:** A ausência de valor em qualquer atributo impede a análise da linha, resultando em sua remoção.
- **Matrícula com Múltiplas Categorias e/ou Grupos de Faturamento:** A ambiguidade na classificação correta da matrícula leva à remoção de todas as suas linhas, pois não é possível determinar a informação precisa.
- **Linha com Data Situação Posterior à Referência:** Indica um cenário atípico de pagamento antecipado ou uma inconsistência nos dados, justificando a remoção.
- **Linha com Data de Vencimento Anterior à Referência:** Caracteriza um erro lógico, pois a fatura já seria emitida com o pagamento em atraso, tornando a linha inválida para análise.
- **Linhas com Valor de Fatura Idêntico em Múltiplas Referências para a Mesma Matrícula:** Sugere um parcelamento de fatura ou um erro de registro, necessitando de remoção.

Para assegurar que as análises se refiram a um período mínimo de 24 meses, foi aplicado um filtro final. Após as etapas de remoção mencionadas, este filtro seleciona apenas as matrículas que contêm 24 ou mais faturas. As matrículas que não atingiram esse número mínimo de faturas válidas após a filtragem inicial foram descartadas.

Após esta filtragem temporal, o dataset contava com 2.674.200 faturas. Destas, 111.425 correspondiam à data de referência que seria utilizada para prever a situação de inadimplência. Observou-se que a soma das faturas inadimplentes nesta data, pertencentes às categorias Comercial, Industrial e Público, totalizava apenas 126 registros. Essa baixa representatividade, dividida entre três categorias distintas, inviabilizou a identificação de padrões preditivos robustos. Diante disso, os registros destas categorias foram filtrados, optando-se por focar a análise exclusivamente na categoria Residencial, que constitui a parcela majoritária dos dados e oferece maior volume de observações para a modelagem.

Após os processos iniciais de filtragem, restaram um número 2.478.768 de faturas, correspondentes a 103.282 matrículas. A Figura 1 apresenta, a título de exemplo, um pequeno trecho da base de dados após a conclusão do processo de tratamento, sendo visualizado por meio da ferramenta Superintendent.app.

3.2.3 Atributos Derivados

A partir dos atributos originais, foram derivados novos dados, resultando na inclusão de novas colunas à tabela existente. Esta etapa é crucial, pois facilita o processamento computacional por parte do modelo preditivo e pode melhorar

Figura 1 – Exemplo de um fragmento da base de dados após o tratamento

2,478,768 rows (Only 1,000 are shown)

	Matricula_com_Mascara	Categoria	Grupo_de_Faturamento	Referencia	Data_Vencimento	Data_Situacao	Valor_Fatura	Situacao_Fatura
1	10011913272749-4	Residencial	67	2022-06-01T00:00:00.000Z	2022-07-19T00:00:00.000Z	2022-07-19T00:00:00.000Z	32.57	Quitada
2	10011913272749-4	Residencial	67	2022-07-01T00:00:00.000Z	2022-08-19T00:00:00.000Z	2022-08-19T00:00:00.000Z	35.03	Quitada
3	10011913272749-4	Residencial	67	2022-08-01T00:00:00.000Z	2022-09-19T00:00:00.000Z	2022-09-19T00:00:00.000Z	33.8	Quitada
4	10011913272749-4	Residencial	67	2022-09-01T00:00:00.000Z	2022-10-19T00:00:00.000Z	2022-10-19T00:00:00.000Z	32.57	Quitada
5	10011913272749-4	Residencial	67	2022-10-01T00:00:00.000Z	2022-11-19T00:00:00.000Z	2022-11-19T00:00:00.000Z	36.26	Quitada
6	10011913272749-4	Residencial	67	2022-11-01T00:00:00.000Z	2022-12-19T00:00:00.000Z	2022-12-19T00:00:00.000Z	33.8	Quitada
7	10011913272749-4	Residencial	67	2022-12-01T00:00:00.000Z	2023-01-19T00:00:00.000Z	2023-01-19T00:00:00.000Z	35.03	Quitada
8	10011913272749-4	Residencial	67	2023-01-01T00:00:00.000Z	2023-02-19T00:00:00.000Z	2023-02-22T00:00:00.000Z	35.03	Quitada
9	10011913272749-4	Residencial	67	2023-02-01T00:00:00.000Z	2023-03-19T00:00:00.000Z	2023-03-20T00:00:00.000Z	35.03	Quitada
10	10011913272749-4	Residencial	67	2023-03-01T00:00:00.000Z	2023-04-19T00:00:00.000Z	2023-04-19T00:00:00.000Z	38.35	Quitada

Fonte: Elaborado pelo Autor (2025)

significativamente seus resultados. Para a presente análise, os atributos mais básicos a serem derivados são:

- **Atraso_Pagamento:** Um número inteiro correspondente ao intervalo, em dias, entre a Data de Vencimento e a Data Situação. Números positivos representam um atraso na fatura, enquanto os negativos representam pagamentos antecipados.
- **Default:** Um booleano que assume o valor de 1 se a fatura tem Atraso maior que 50 dias e, caso contrário, assume valor 0.

Estes novos atributos representam mais diretamente o objetivo de nossa análise. O campo "Atraso_Pagamento" sintetiza a diferença entre duas datas em um único número inteiro, enquanto "Default" simplifica essa informação de forma binária, indicando a inadimplência. Isso é particularmente útil, visto que nosso objetivo não é prever a quantidade exata de dias de atraso no pagamento da última fatura de uma unidade consumidora, mas sim determinar se esse atraso ultrapassará o limite de 50 dias.

Além disso, novos atributos foram calculados com o objetivo de otimizar o desempenho do modelo preditivo. Esses atributos são de valor agregado, como o valor médio de fatura por matrícula. Para evitar data leakage durante o treinamento do modelo, a fatura referente ao último mês de referência de cada matrícula foi excluída dos cálculos. A seguir, uma breve descrição de cada novo atributo:

- **Atraso_Max:** O valor máximo de Atraso_Pagamento registrado para a matrícula durante o período analisado.
- **Atraso_Min:** O valor mínimo de Atraso_Pagamento para cada matrícula.
- **Atraso_Medio:** O valor médio de Atraso_Pagamento para cada matrícula.
- **Atraso_Mediana:** A mediana dos valores de Atraso_Pagamento para cada matrícula.
- **Atraso_Variance_Medio:** O valor de variância média dos valores de Atraso_Pagamento para cada matrícula.
- **Fatura_Max:** O valor máximo de Valor_Fatura registrado para a matrícula durante o período analisado.

- **Fatura_Min:** O valor mínimo de Valor_Fatura para cada matrícula.
- **Fatura_Medio:** O valor médio de Valor_Fatura para cada matrícula.
- **Fatura_Mediana:** A mediana dos valores de Valor_Fatura para cada matrícula.
- **Fatura_Variance_Medio:** O valor de variância média dos valores de Valor_Fatura para cada matrícula.
- **Default_Somatorio:** O número de ocorrências em que a matrícula apresentou a situação de Default ao longo dos período de 23 meses analisados.
- **Ultimo_Default:** O número de dias transcorridos desde a última ocorrência de Default da matrícula, calculado em relação à data de referência alvo da análise.

3.2.4 Base de Dados Transformada

Para tentar replicar os resultados obtidos por GARAY e OLIVEIRA (2022), foi construída uma nova tabela com uma estrutura pivotada. Em vez de múltiplas linhas por matrícula, essa nova organização apresenta apenas uma linha por matrícula, mas com 24 novas colunas. Cada uma dessas colunas é nomeada com a data de referência no formato mês\ano e armazena o valor de Atraso_Pagamento correspondente ao respectivo mês.

A tabela foi consolidada, reduzindo o número de registros de 2.478.768 para 103.282. Em seguida, os valores calculados na subseção anterior foram integrados como novas colunas. Por fim, foi adicionada a coluna Is_Default_Target, que armazena o status de Default de cada matrícula para o mês de maio de 2024. O valor dessa nova coluna é o alvo da predição dos modelos.

Figura 2 – Base de dados pivotada

103,282 rows (Only 1,000 are shown)									
	Matricula_com_Mascara	06/2022	07/2022	08/2022	...	02/2024	03/2024	04/2024	Is_Default_Target
1	12675451042726-0	0	2	0		0	2	0	0
2	12675788446091-7	31	38	7		0	-10	0	0
3	12676601985508-3	3	2	-4		2	0	0	0
4	12680427450712-4	-6	25	55		-10	-5	1	0
5	12681812534697-4	0	-4	4		-3	0	-2	0
6	12682504447273-1	15	18	17		-14	-11	-12	0
7	12683895272830-3	1	0	2		-3	12	31	0
8	1268422239171-6	-20	-26	-25		30	28	-2	0
9	12686528551862-4	2	0	0		2	0	0	0
10	12688038488308-9	-12	-19	-9		-20	-19	-18	0

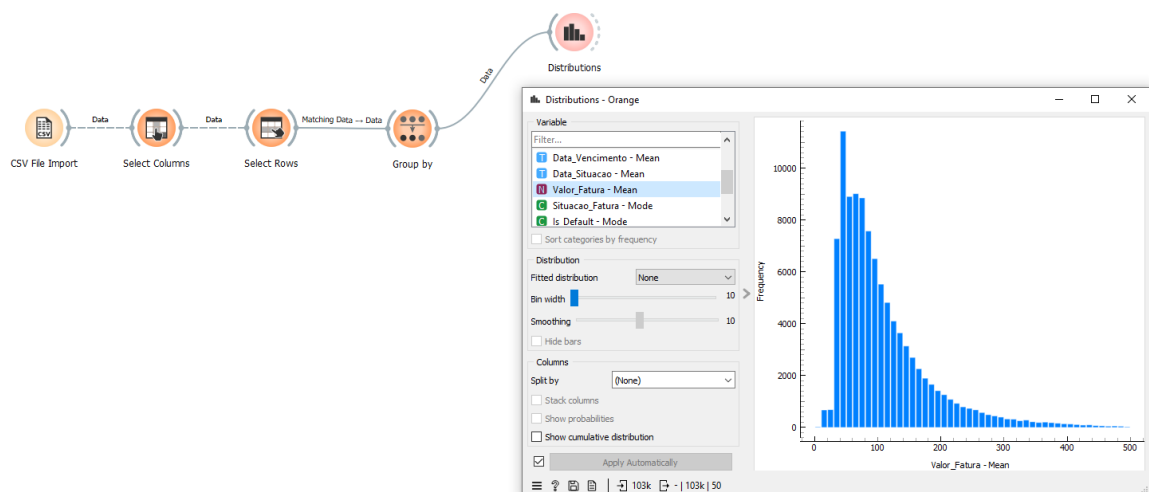
Fonte: Elaborado pelo Autor (2025)

3.2.5 Software Utilizado Para a Análise

Para a etapa de análise e aplicação de técnicas de machine learning, optou-se pela utilização do software Orange Datamining². Trata-se de uma plataforma de código aberto que oferece um ambiente visual e interativo para análise de dados, aprendizado de máquina e visualização, permitindo a construção de workflows de análise por meio de uma interface intuitiva.

O software Orange organiza suas funcionalidades por meio de widgets, que são conectados visualmente para construir o workflow de análise desejado. Para este trabalho, foram destacados dois widgets de particular relevância: o Rank, que avalia cada atributo utilizando diversos scores para determinar sua importância para a análise, e o Test and Score, que testa os modelos com os dados fornecidos e exibe os scores dos resultados obtidos, incluindo acurácia, F1, Precision e Recall.

Figura 3 – Exemplo de fluxo de trabalho na Orange



Fonte: Elaborado pelo Autor (2025)

3.2.6 Validação dos Modelos

A validação do desempenho dos modelos preditivos neste trabalho foi conduzida utilizando a estratégia Holdout (conjunto de retenção), implementada por meio da configuração "Test on Test Data" no widget de avaliação. O método Holdout representa a forma mais direta de avaliação, consistindo na divisão do conjunto de dados original em duas partições independentes: uma utilizada para o treinamento do algoritmo e a outra reservada para o teste e a aferição de performance. Para garantir a robustez da avaliação, 80% dos dados foram alocados para o treinamento, enquanto os

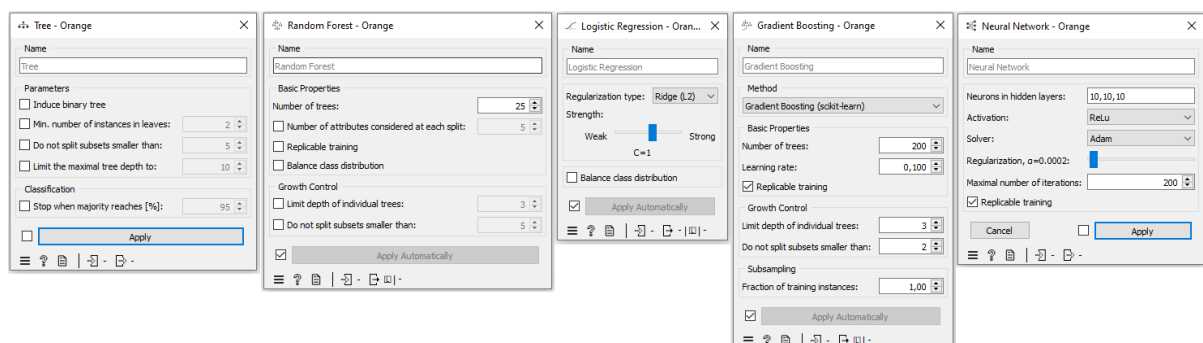
² <https://orangedatamining.com>

20% restantes foram destinados à validação. Esta proporção é amplamente aceita na literatura e representa um equilíbrio adequado para o aprendizado e a testabilidade do modelo, conforme recomendado por trabalhos como o de (BICHRI; CHERGUI; HAIN, 2024).

Os algoritmos de classificação binária empregados e testados neste estudo foram Decision Tree, Random Forest, Logistic Regression, Gradient Boosting e Neural Network. A seleção desses modelos foi baseada nos resultados e nas comparações de desempenho encontrados em trabalhos similares na área de previsão de risco de crédito e atraso em pagamentos. A literatura demonstra que a eficácia varia conforme o tipo e o volume de dados, sendo estes modelos escolhidos por apresentarem alta performance em diferentes métricas e contextos, justificando sua inclusão para a busca do melhor trade-off preditivo para o presente conjunto de dados da concessionária.

Os parâmetros empregados nos modelos foram definidos de forma experimental sendo, por fim, utilizados os parâmetros que podem ser visualizados na Figura 4. A Figura 4 mostra a tela de configuração da Orange para, da esquerda para a direita, modelo de Árvore de decisão, Random Forest, Regressão Logística, Gradient Boosting e da Rede Neural.

Figura 4 – Parâmetros de configuração dos modelos preditivos



Fonte: Elaborado pelo Autor (2025)

4 ANÁLISES E RESULTADOS

O presente capítulo detalha os achados empíricos e a performance dos modelos preditivos desenvolvidos. A seção inicia-se com a análise exploratória dos dados, visando a compreensão da estrutura da base de faturas. Em seguida, as análises são estruturadas em três abordagens sequenciais e progressivas, visando isolar o impacto de técnicas específicas de pré-processamento e engenharia de atributos no poder preditivo do modelo de classificação. Inicialmente, estabelece-se uma abordagem baseline para fixar um ponto de comparação de desempenho. Em seguida, avalia-se a contribuição dos atributos derivados na melhoria da performance. Por fim, investiga-se o efeito das técnicas de tratamento do desbalanceamento de classes (downsampling) sobre as métricas de interesse. O capítulo é concluído com uma comparação geral dos resultados obtidos em cada uma das abordagens.

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A análise exploratória inicial do conjunto de dados revelou a magnitude e a distribuição do comportamento de pagamento das unidades consumidoras da concessionária. O volume total de faturas analisadas ao longo dos 24 meses do período de referência (entre junho de 2022 e maio de 2024) foi de 2.478.768 registros. Desse total, 26.511 faturas tiveram atraso maior que 50 dias, o que corresponde a apenas 1,07% do volume total.

Tabela 1 – Distribuição de Atrasos

Dias de atraso	Ocorrências	% total
<=0	1.486.840	59,98
>=1	565.082	22,80
>=15	228.487	9,22
>=31	171.848	6,93
>50	26.511	1,07
Total	2.478.768	100,00

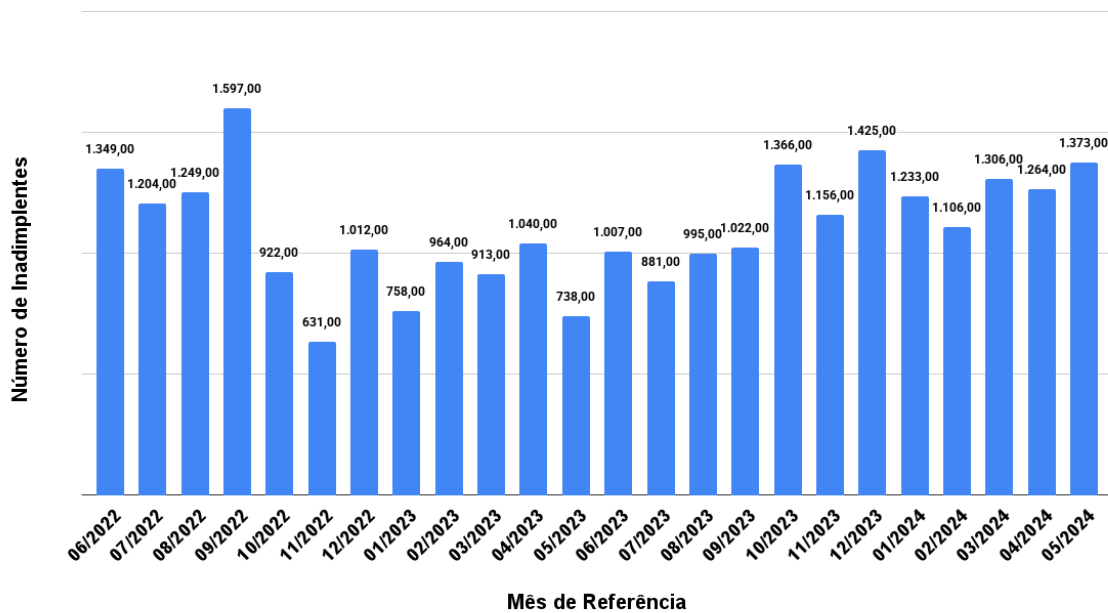
Fonte: Elaborado pelo Autor (2025)

A Tabela 1 apresenta a distribuição dos atrasos de pagamento, indicando que a grande maioria dos consumidores honra seus compromissos de forma antecipada ou pontual, totalizando 59,98% das faturas. A análise temporal dos defaults mensais mostrou variações ao longo do período: o ponto de pico de inadimplência ocorreu em setembro de 2022, quando foram registrados 1.597 consumidores inadimplentes, conforme ilustrado na Figura 1. A média de inadimplentes por mês no período estudado foi de 1.104,63, com uma mediana de 1.073, sugerindo que a distribuição mensal de

defaults é relativamente simétrica.

Em relação aos valores faturados, a análise descritiva mostrou que a média do valor das faturas foi de 124,8 reais, enquanto a mediana se situou em 79,9 reais. Essa diferença entre a média e a mediana é um indicativo de que a distribuição do valor das faturas é assimétrica positiva, sugerindo a presença de um número menor de faturas com valores significativamente mais altos, que elevam a média.

Figura 5 – Distribuição temporal de inadimplentes por data de referência



Fonte: Elaborado pelo Autor (2025)

4.2 RESULTADOS

Esta seção marca o início do treinamento e da avaliação dos modelos preditivos, apresentando a performance de cada algoritmo em diferentes abordagens metodológicas.

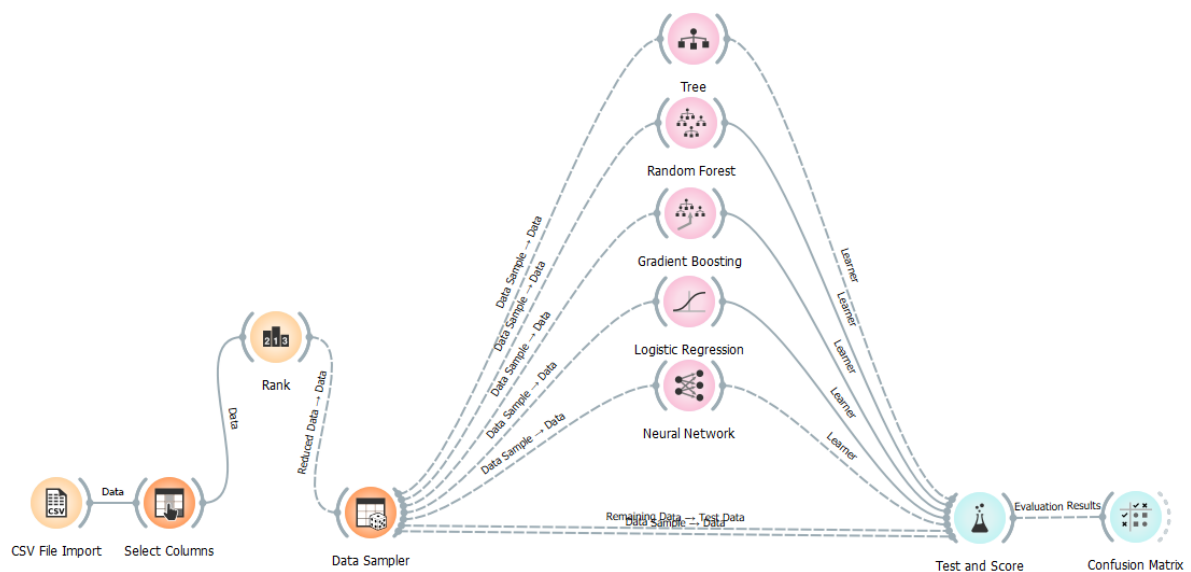
4.2.1 Resultados da Abordagem Baseline

Para estabelecer uma linha de base para os resultados, iniciou-se a análise empregando os mesmos métodos de GARAY e OLIVEIRA (2022). Os dados foram organizados conforme detalhado na Subseção 3.2.4. Para esta fase, foram utilizados como atributos as 23 colunas contendo o Atraso_Pagamento de cada mês, além da coluna Default_Any, totalizando 24 atributos.

A Figura 6 ilustra o fluxo de trabalho empregado no software Orange para o treinamento e a avaliação dos modelos preditivos. O processo inicia-se com o widget

"Select Columns", que desempenha a função de selecionar os atributos que serão utilizados como variáveis de entrada, bem como o atributo designado como alvo da predição. Imediatamente após, o widget "Rank" é utilizado para classificar os atributos com base em sua contribuição individual para o treinamento dos modelos, sendo uma ferramenta capaz de filtrar aqueles de menor relevância. No entanto, para esta etapa, todos os atributos disponíveis foram utilizados. Em seguida, o widget "Data Sampler" é o responsável por particionar a base de dados nas proporções requeridas, que, neste estudo, foram definidas como 80% para o conjunto de treinamento e 20% para o conjunto de teste, garantindo a amostragem estratificada. O fluxo de trabalho prossegue com a inclusão dos widgets correspondentes aos diferentes modelos de Machine Learning testados. Finalmente, o widget "Test and Score" é utilizado para a avaliação quantitativa do desempenho dos modelos, e o widget "Confusion Matrix" permite a visualização gráfica e detalhada dos resultados das predições.

Figura 6 – Fluxo de trabalho implementado na ferramenta Orange para Análise Baseline



Fonte: Elaborado pelo Autor (2025)

A Figura 7 ilustra a classificação dos atributos conforme avaliado pelo widget Rank. O atributo Default_Any se destaca como o mais relevante, exibindo scores significativamente superiores de Information Gain e Gain Ratio em comparação com os demais. Em segundo lugar na classificação, encontra-se o atributo 04/2024, seguido por 03/2024, 02/2024 e, em quinto lugar, 01/2024.

Esse padrão de relevância é observado apenas nas quatro datas mais recentes. Os atributos subsequentes seguem uma ordem semi-aleatória, embora a tendência geral de relevância continue sendo de, quanto mais antigo, menos relevante.

Os modelos foram treinados utilizando 80% dos registros, o que corresponde a

Figura 7 – Classificação dos atributos para abordagem Baseline

		#	Info. gain	Gain ratio	Gini
1	 Default_Any	2	0.021	0.042	0.001
2	 04/2024		0.013	0.007	0.001
3	 03/2024		0.012	0.006	0.001
4	 02/2024		0.012	0.006	0.000
5	 01/2024		0.011	0.006	0.000
6	 11/2023		0.011	0.006	0.000
7	 12/2023		0.011	0.005	0.000
8	 10/2023		0.010	0.005	0.000
9	 08/2023		0.009	0.005	0.000
10	 09/2023		0.009	0.005	0.000

Fonte: Elaborado pelo Autor (2025)

82.626 linhas, e testados com os 20% restantes, que correspondem a 20.656 linhas.

Tabela 2 – Comparativo de scores dos modelos na abordagem Baseline

	CA	F1	Prec	Recall
Tree	0.977	0.204	0.191	0.218
Random Forest	0.987	0.208	0.574	0.127
Logistic Regression	0.987	0.127	0.760	0.069
Gradient Boosting	0.987	0.241	0.568	0.153
Neural Network	0.987	0.161	0.694	0.091

Fonte: Elaborado pelo Autor (2025)

Conforme observado na Tabela 2 e na Tabela 3, que exibe a matriz de confusão do modelo com o melhor score F1, nenhum dos modelos apresentou desempenho satisfatório. Todos se mostraram incapazes de classificar adequadamente o status de Default das unidades consumidoras para o mês 05/2024. No conjunto de dados de teste, o modelo alcançou 42 Verdadeiros Positivos (TP) em um total de 275 casos de inadimplência, indicando um Recall muito limitado. O modelo também gerou 32 Falsos Positivos (FP). Dessa forma, das 74 classificações positivas totais (TP + FP), quase metade (43,24%) constituiu um erro de classificação, demonstrando a baixa Precisão (Precision) da abordagem.

4.2.2 Resultados com Atributos Derivados

Estabelecida uma linha de base de resultados, foi possível prosseguir com a exploração de outras abordagens para buscar uma melhora no desempenho dos

Tabela 3 – Matriz de Confusão do modelo GB na abordagem Baseline

	0	1	Σ
0	20349	32	20381
1	233	42	275
Σ	20582	74	20656

Fonte: Elaborado pelo Autor (2025)

modelos. Esta etapa teve início com a adição dos atributos derivados apresentados na Seção 3.2.3. O atributo Default_Any foi substituído por Default_Somatorio, uma vez que a informação fornecida por Default_Any seria redundante. No total, foram utilizados 35 atributos.

A estrutura do fluxo de trabalho implementado no software Orange para esta etapa manteve-se inalterada em relação à abordagem anterior. A única modificação reside no conjunto de atributos empregados e na utilização da funcionalidade do widget Rank para realizar a filtragem e a seleção dos atributos com menor poder discriminatório.

Tabela 4 – Comparativo de scores dos modelos na abordagem Atributos Derivados

	CA	F1	Prec	Recall
Tree	0.978	0.203	0.195	0.109
Random Forest	0.988	0.255	0.778	0.153
Logistic Regression	0.987	0.185	0.600	0.109
Gradient Boosting	0.987	0.242	0.531	0.156
Neural Network	0.987	0.203	0.195	0.211

Fonte: Elaborado pelo Autor (2025)

Figura 8 – Classificação dos atributos para abordagem Atributos Derivados

		#	Info. gain	Gain ratio	Gini
1	 Default_Somatorio		0.025	0.039	0.002
2	 Ultimo_Default		0.021	0.042	0.001
3	 Atraso_Max		0.018	0.009	0.001
4	 Atraso_Medio		0.015	0.008	0.001
5	 Atraso_Variance_Medio		0.014	0.007	0.001
6	 04/2024		0.013	0.007	0.001
7	 03/2024		0.012	0.006	0.001
8	 Atraso_Mediana		0.012	0.006	0.000
9	 02/2024		0.012	0.006	0.000
10	 01/2024		0.011	0.006	0.000

Fonte: Elaborado pelo Autor (2025)

Na Tabela 4, pode-se observar uma melhora bastante discreta nos

resultados. Além disso, a análise do Rank revelou que, entre os atributos derivados, apenas Default_Somatorio, Ultimo_Default, Atraso_Max, Atraso_Medio, Atraso_Variance_Medio e Atraso_Mediana apresentaram alguma relevância para a predição.

A etapa subsequente envolveu o descarte dos atributos menos relevantes. Esta ação é fundamental, visto que um excesso de features pode comprometer a performance e a capacidade de generalização do modelo preditivo. Para realizar a seleção progressiva de variáveis, os modelos foram testados inicialmente com o atributo que demonstrou a melhor classificação isoladamente. Em seguida, os atributos subsequentes foram adicionados um a um, sendo os resultados obtidos comparados após cada inclusão, de modo a identificar o ponto de saturação e a combinação ótima de features.

Tabela 5 – Variação do score F1 em função do número de atributos selecionados

	1	2	3	4	5	6	7	8	9	10	11	12	13
Tree	0.083	0.155	0.206	0.166	0.188	0.178	0.221	0.212	0.203	0.224	0.209	0.211	0.209
Random Forest	0.083	0.155	0.200	0.188	0.191	0.257	0.257	0.269	0.251	0.260	0.262	0.247	0.234
Logistic Regression	0.166	0.137	0.136	0.143	0.142	0.205	0.227	0.217	0.217	0.216	0.215	0.215	0.215
Gradient Boosting	0.083	0.187	0.186	0.202	0.226	0.245	0.229	0.219	0.216	0.223	0.224	0.234	0.229
Neural Network	0.126	0.126	0.108	0.130	0.131	0.255	0.201	0.198	0.276	0.255	0.229	0.269	0.213

Fonte: Elaborado pelo Autor (2025)

Tabela 6 – Variação do score Recall em função do número de atributos selecionados

	1	2	3	4	5	6	7	8	9	10	11	12	13
Tree	0.044	0.087	0.145	0.164	0.196	0.185	0.229	0.229	0.211	0.236	0.233	0.229	0.222
Random Forest	0.044	0.087	0.138	0.131	0.116	0.160	0.160	0.167	0.156	0.164	0.164	0.149	0.142
Logistic Regression	0.098	0.076	0.076	0.080	0.080	0.120	0.135	0.127	0.127	0.127	0.127	0.127	0.127
Gradient Boosting	0.044	0.109	0.113	0.120	0.138	0.153	0.138	0.135	0.131	0.138	0.138	0.142	0.142
Neural Network	0.069	0.069	0.058	0.073	0.073	0.160	0.116	0.116	0.178	0.127	0.138	0.167	0.127

Fonte: Elaborado pelo Autor (2025)

Na Tabela 5, observa-se uma melhora progressiva nos resultados para a maioria dos modelos até o ponto em que 12 atributos são utilizados, momento em que a performance começa a decair. A Neural Network, embora tenha alcançado o maior score F1, demonstrou uma performance bastante irregular. Já a Random Forest, com um desempenho mais consistente, obteve o segundo melhor score F1 ao utilizar 8 atributos.

Como pode-se observar na Tabela 6, o modelo Decision Tree apresentou os melhores scores de Recall, apesar de uma performance mais baixa em seu score F1.

4.2.3 Análise com Downsampling

Até o momento, o objetivo da modelagem concentrou-se em prever com alta precisão quais unidades consumidoras entrariam em default no mês analisado, buscando um baixo número de Falsos Positivos e Falsos Negativos, o que é refletido pelo score F1. No entanto, em face dos resultados insatisfatórios obtidos nessa

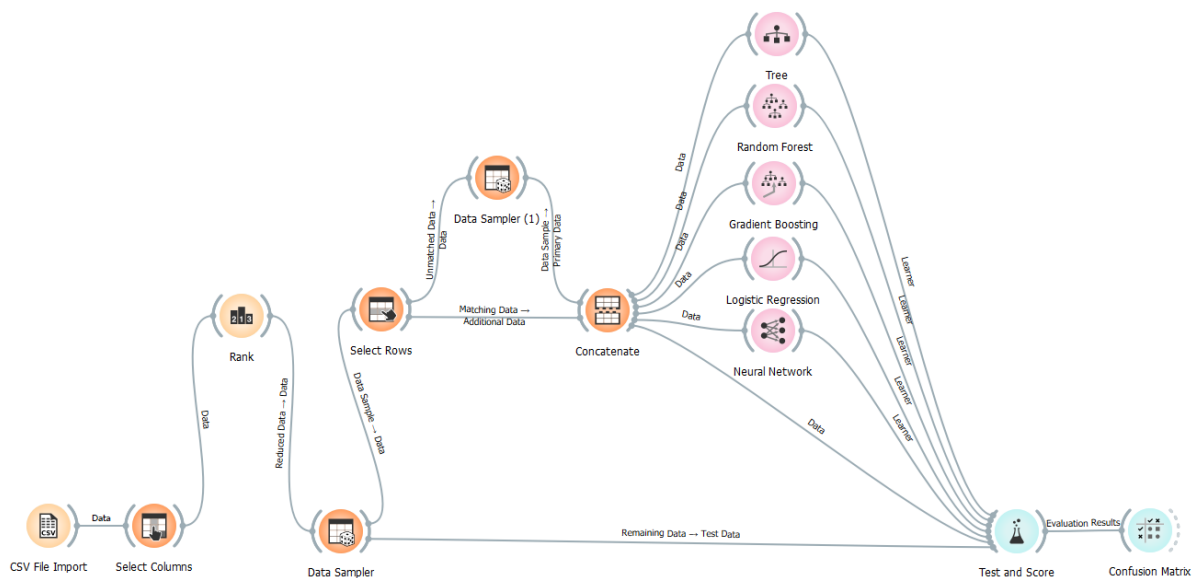
abordagem, decidiu-se alterar o foco para a minimização dos Falsos Negativos, mesmo que isso acarrete um aumento nos Falsos Positivos. A partir deste ponto, a análise priorizará a otimização do score Recall.

O conjunto de dados analisado apresenta uma distribuição de classes significativamente desbalanceada, com apenas 1,33% dos 103.282 registros indicando situação de Default. Essa disparidade leva os modelos a tenderem à classificação de todos os registros como não inadimplentes, já que essa estratégia eleva a probabilidade de acerto.

Para incentivar os modelos a classificarem um maior número de matrículas como inadimplentes, empregou-se o balanceamento das classes por meio de Downsampling nos dados de treinamento. Nesse método, o volume de dados da classe majoritária é reduzido até que se alcance a proporção de classes desejada.

O fluxo de trabalho da Orange utilizado nesta etapa de Downsampling difere dos anteriores pela inserção de três widgets após o primeiro "Data Sampler" (que realiza a divisão inicial entre treino e teste). Essa alteração, que pode ser vista na Figura 9, incluiu o uso dos widgets "Select Rows", "Data Sampler" (aplicado novamente) e "Concatenate", com o objetivo de rebalancear e reconstituir o conjunto de treinamento na proporção de classes desejada.

Figura 9 – Fluxo de trabalho implementado na ferramenta Orange para Análise com Downsampling



Fonte: Elaborado pelo Autor (2025)

Para esta análise, foram utilizados apenas 2.196 registros para o treinamento dos modelos, estabelecendo uma proporção de 1:1 entre matrículas em default e não em default. Os mesmos 8 atributos previamente classificados como os de melhor desempenho foram mantidos. Os resultados obtidos são apresentados na Figura 7.

Com essa abordagem, foram alcançados valores consideravelmente altos para o score Recall, com 86,2% dos casos positivos sendo classificados corretamente. Entretanto, o score Precisão sofreu uma queda significativa, resultando em um score F1 muito mais baixo. A matriz de confusão, apresentada na Tabela 7, ilustra o problema inerente a esse tipo de análise: o modelo com o melhor Recall classificou erroneamente 4.033 matrículas como inadimplentes.

Tabela 7 – Comparativo de scores dos modelos na abordagem com Downsampling

	CA	F1	Prec	Recall
Tree	0.769	0.085	0.045	0.804
Random Forest	0.811	0.108	0.057	0.858
Logistic Regression	0.836	0.114	0.062	0.796
Gradient Boosting	0.803	0.104	0.056	0.862
Neural Network	0.809	0.106	0.057	0.855

Fonte: Elaborado pelo Autor (2025)

Tabela 8 – Matriz de Confusão do modelo GB na abordagem com Downsampling

	0	1	Σ
0	16348	4033	20381
1	38	237	275
Σ	16386	4270	20656

Fonte: Elaborado pelo Autor (2025)

Com essa abordagem, são alcançados valores consideravelmente altos para o score Recall, com 86,2% dos casos positivos sendo classificados corretamente. No entanto, o score Precisão sofreu uma queda significativa, resultando em um score F1 muito mais baixo. A matriz de confusão, apresentada na Tabela 8, ilustra claramente o problema inerente a esse tipo de análise: o modelo com o melhor Recall classificou erroneamente 4.033 matrículas como inadimplentes.

4.3 COMPARAÇÃO GERAL DAS ABORDAGENS

A avaliação dos resultados demonstrou a importância das etapas de pré-processamento e balanceamento de classes na otimização do modelo de classificação de inadimplência. A comparação do desempenho das três abordagens testadas – Baseline, Atributos Derivados e Downsampling – revelou diferentes trade-offs entre a capacidade de identificar corretamente os inadimplentes (Recall) e a precisão das previsões (Precision), conforme sintetizado a seguir:

- **Abordagem Baseline:** O modelo inicial, treinado apenas com variáveis brutas e sem tratamento de desbalanceamento, apresentou um desempenho intermediário em termos de F1-Score. Sua pontuação, que foi inferior à obtida com a inclusão de atributos derivados, refletiu a dificuldade em lidar com a assimetria das

classes, resultando em um modelo incapaz de diferenciar efetivamente os casos minoritários de inadimplência.

- **Abordagem com Atributos Derivados:** A inclusão de features criadas proporcionou uma melhoria percentual pequena no desempenho geral. Essa melhora sugere que a Engenharia de Atributos adicionou algum valor preditivo ao modelo, mas foi insuficiente para superar o problema fundamental do desbalanceamento inerente ao conjunto de dados.
- **Abordagem com Downsampling:** A aplicação da técnica de reamostragem (Downsampling) da classe majoritária resultou no pior F1-Score geral. Contudo, essa técnica cumpriu seu objetivo principal ao gerar um alto Recall, indicando que o modelo se tornou muito eficaz em capturar a maioria dos casos reais de inadimplência. Essa alta sensibilidade foi obtida ao custo de uma baixa Precision. O modelo passou a gerar um volume elevado de Falsos Positivos, demonstrando o trade-off inerente ao balanceamento na fronteira de decisão.

Tabela 9 – Comparativo do Score F1 entre as diferentes abordagens

	Baseline	Derivados	Downsampling
Tree	0.204	0.203	0.085
Random Forest	0.208	0.255	0.108
Logistic Regression	0.127	0.185	0.114
Gradient Boosting	0.241	0.242	0.104
Neural Network	0.161	0.203	0.106

Fonte: Elaborado pelo Autor (2025)

Tabela 10 – Comparativo do Score Precision entre as diferentes abordagens

	Baseline	Derivados	Downsampling
Tree	0.191	0.195	0.045
Random Forest	0.574	0.778	0.057
Logistic Regression	0.760	0.600	0.062
Gradient Boosting	0.568	0.531	0.056
Neural Network	0.694	0.195	0.057

Fonte: Elaborado pelo Autor (2025)

Tabela 11 – Comparativo do Score Recall entre as diferentes abordagens

	Baseline	Derivados	Downsampling
Tree	0.218	0.109	0.804
Random Forest	0.127	0.153	0.858
Logistic Regression	0.069	0.109	0.796
Gradient Boosting	0.153	0.156	0.862
Neural Network	0.091	0.211	0.855

Fonte: Elaborado pelo Autor (2025)

A tabulação comparativa dos resultados entre as abordagens, encontradas nas Tabelas 9, 10 e 11, demonstrou que a inclusão dos atributos derivados proporcionou uma leve melhoria no score F1 para a maioria dos modelos, sendo a Decision Tree a única exceção quando comparada ao método Baseline. Em contrapartida, para o score Precision, a melhoria foi observada exclusivamente nos modelos Decision Tree e Random Forest. Os demais modelos apresentaram uma significativa piora no score, notavelmente a Neural Network, que sofreu uma queda de 0.694 para 0.195. O método de Downsampling apresentou os valores mais baixos de F1 e Precision, comportamento já esperado em função da sua estratégia de balanceamento.

Já em relação ao score Recall, observou-se uma leve melhoria em quase todos os modelos ao se transitar do Baseline para a abordagem com Atributos Derivados, com a Decision Tree novamente fugindo a essa tendência. Confirmando o objetivo da técnica, o Downsampling apresentou, de longe, os melhores resultados de Recall, com o modelo Gradient Boosting atingindo um score de 0.862.

A análise dos resultados evidencia as limitações do estudo e aponta para a complexidade inerente à previsão de eventos raros no setor de serviços.

A primeira limitação reside na extensão da base temporal utilizada. A análise de um período relativamente curto (23 meses) restringe a capacidade do modelo de capturar tendências de longo prazo ou padrões cíclicos de inadimplência que poderiam estar ligados a fatores macroeconômicos. Além disso, a inadimplência é um evento

raro e é influenciada por uma série de fatores não observados no conjunto de dados da concessionária. A decisão de não pagar uma conta pode ser precipitada por eventos estritamente pessoais (desemprego, doenças) ou conjunturais (mudanças súbitas na política tarifária) que não estão registrados nas variáveis de consumo ou pagamento, limitando o teto de Recall que o modelo pode atingir.

Por fim, os modelos apresentaram uma dependência significativa das variáveis transacionais internas da concessionária. Para aumentar drasticamente o poder preditivo, seriam necessários atributos financeiros e socioeconômicos mais significativos. Exemplos desses atributos incluem dados do consumidor como score de crédito (informação externa de bureaus), a renda mensal da unidade familiar ou o nível de endividamento registrado em outras fontes (MOSCATELLI et al., 2020). A inclusão dessas informações externas ao faturamento de água forneceria ao modelo uma visão mais completa da capacidade e da propensão de pagamento do cliente, o que é inviável apenas com os dados internos da concessionária.

5 CONCLUSÕES

O presente trabalho teve como objetivo o desenvolvimento e a avaliação de um modelo preditivo baseado em Machine Learning capaz de identificar potenciais usuários inadimplentes da Companhia de Águas de Joinville. A questão central da pesquisa foi determinar a viabilidade de antecipar o risco de default utilizando-se apenas o histórico transacional interno da concessionária. Os resultados demonstram que, embora a modelagem tenha alcançado níveis de performance superiores aos de uma classificação aleatória, ela ressaltou a complexidade inerente à previsão de eventos raros e a importância crítica do pré-processamento de dados.

A síntese dos achados empíricos revelou a progressão da performance mediante a aplicação de técnicas de pré-processamento. A abordagem baseline, estabelecida com dados brutos, fixou o ponto de partida com baixa performance, o que ressaltou o desafio imposto pelo extremo desbalanceamento de classes. A inclusão de atributos derivados, embora tenha conferido maior inteligência ao modelo, resultou em apenas uma melhora marginal no score F1, confirmando que a Engenharia de Atributos sozinha não foi suficiente para superar a assimetria dos dados. A abordagem mais transformadora foi o Downsampling, que alterou o foco da otimização para o Recall. Com esta técnica, o modelo alcançou um alto Recall (maior capacidade de detecção de inadimplentes), porém, incorreu em uma baixa Precision, demonstrando o trade-off crucial para a concessionária: alta detecção de risco versus alto custo operacional causado por Falsos Positivos.

A principal limitação identificada no estudo reside na restrição aos dados transacionais internos (histórico de consumo e pagamento). A ausência de features socioeconômicas (como renda familiar, score de crédito ou taxa de desemprego) impediu o modelo de capturar os fatores exógenos que realmente impulsionam a inadimplência, limitando o teto de performance alcançável. Adicionalmente, o ambiente de desenvolvimento utilizado (Orange) impôs restrições técnicas, como a indisponibilidade da arquitetura Long Short-Term Memory (LSTM) para Redes Neurais de forma nativa, exigindo a adição de scripts externos de Python para sua implementação completa. A LSTM é particularmente indicada para a modelagem de dados sequenciais e temporais, como é o caso das séries de faturamento, o que poderia ter beneficiado a captura de padrões de longo prazo e aprimorado a performance preditiva.

Para dar continuidade a esta pesquisa e superar as limitações encontradas, sugerem-se estudos futuros focados em três eixos principais: (i) Inclusão de Dados Exógenos: Obter e integrar informações de fontes externas (como bureaus de crédito ou

dados censitários) para aumentar o poder discriminatório das features. (ii) Modelagem Temporal Avançada: Replicar o estudo em ambientes de programação flexíveis (como Python) para permitir a aplicação de arquiteturas avançadas de Deep Learning, como as redes LSTM, visando otimizar a previsão em séries temporais. (iii) Otimização de Custos: Explorar modelos de Cost-Sensitive Learning para otimizar a classificação não apenas com base na precisão estatística, mas sim no custo financeiro real de cada Falso Negativo e Falso Positivo.

REFERÊNCIAS

- AESBE. **Saneamento já tem quase 25% de inadimplência**. Brasília, 2020. Disponível em: <https://aesbe.org.br/saneamento-ja-tem-quase-25-de-inadimplencia>. Acesso em: 21 nov. 2025.
- ALBUQUERQUE, G. D. R. Estruturas de financiamento aplicáveis ao setor de saneamento básico. **Banco Nacional de Desenvolvimento Econômico e Social**, n. 34, p. 45–94, 2011.
- ALPAYDIN, E. **Introduction to Machine Learning**. 4. ed. Cambridge: The MIT Press, 2020.
- BASHAR, A. et al. Machine learning for predicting propensity-to-pay energy bills. **Intelligent Systems with Applications**, v. 17, 2023.
- BATISTA, D. M. A.; CAMPOLI, J. S. V-1374 - business model canvas aplicado para universalização dos serviços de saneamento básico em um município brasileiro. In: **32 CONGRESSO BRASILEIRO DE ENGENHARIA SANITÁRIA E AMBIENTAL**. Porto Alegre: ABES, 2023. Disponível em: https://abes-dn.org.br/anaisletronicos/32cbesa/1374_tema_v.pdf. Acesso em: 21 nov. 2025.
- BEAULIEU, A. **Learning SQL: Generate, Manipulate, and Retrieve Data**. 3. ed. Sebastopol: O'Reilly Media, 2020.
- BICHRI, H.; CHERGUI, A.; HAIN, M. Investigating the impact of train / test split ratio on the performance of pre-trained models with custom datasets. **International Journal of Advanced Computer Science and Applications**, v. 15, n. 2, jan. 2024.
- DE BOCK, K. W.; COUSSEMENT, K.; LESSMANN, S. Cost-sensitive business failureprediction when misclassification costs are uncertain: A heterogeneous ensemble selectionapproach. **European Journal of Operational Research**, v. 285, n. 2, p. 612–630, jan. 2020.
- COMPANHIA ÁGUAS DE JOINVILLE. **História**. Joinville, 2025. Disponível em: <https://www.aguasdejoinville.com.br/?servico=historia>. Acesso em: 21 nov. 2025.
- DONG, H.; LIU, R.; THAM, A. W. Accuracy comparison between five machine learning algorithms for financial risk evaluation. **Journal of Risk and Financial Management**, v. 17, n. 50, p. 1–21, jan. 2024.
- EBC. **Doze milhões de crianças e adolescentes vivem sem acesso a esgoto**. Brasília, 2024. Disponível em: <https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2024-10/doze-milhoes-de-criancas-e-adolescentes-vivem-sem-acesso-adequado-esgoto>. Acesso em: 21 nov. 2025.
- GARAY, H. S.; OLIVEIRA, R. M. **Aplicando técnicas de Machine Learning e Artificial Intelligence no setor de infraestrutura pública**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Joinville, Santa Catarina, 2022.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3. ed. Waltham: Morgan Kaufmann Publishers, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2. ed. Heidelberg: Springer, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **Institute of Electrical and Electronics Engineers**, v. 21, n. 9, p. 1263–1284, 2009.

INSTITUTO RUI BARBOSA. **Gestão, Controle e Regulação do Saneamento Básico**. Brasília, 2025. Disponível em: <https://irbcontas.org.br/wp-content/uploads/2025/05/gestao-controle-e-regulacao-do-saneamento-basico.pdf>. Acesso em: 21 nov. 2025.

KAUFMAN, S. et al. Leakage in data mining: Formulation, detection, and avoidance. **ACM Journals**, v. 6, n. 4, p. 1–21, 2012.

MITCHELL, T. M. **Machine Learning**. 1. ed. Columbus: McGraw-Hill Science/Engineering/Math, 1997.

MOSCATELLI, M. et al. Corporate default forecasting with machine learning. **Expert Systems With Applications**, v. 161, n. 113567, 2020.

NASCIMENTO, B.; MAIA, D.; ALMADA, L. Applying machine learning to improve collection and to reduce write-offs in utilities. In: **International Conference on Artificial Intelligence and Blockchain Technology**. Beijing: Institute of Electrical and Electronics Engineers, 2021. p. 1–5. Disponível em: <https://doi.org/10.1109/aibt53261.2021.00017>. Acesso em: 21 nov. 2025.

ONDAS. **É urgente garantir água à população em situação de vulnerabilidade social e econômica**. 2020. Disponível em: <https://ondasbrasil.org/e-urgente-garantir-agua-a-populacao-em-situacao-de-vulnerabilidade-social-e-economica-artigo>. Acesso em: 21 nov. 2025.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. **The human right to water and sanitation. Resolution 64/292**. Nova Iorque, 2010. Disponível em: <https://docs.un.org/en/A/RES/64/292>. Acesso em: 21 nov. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Global costs and benefits of drinking-water supply and sanitation interventions to reach the MDG target and universal coverage - WHO/HSE/WSH/12.01**. Geneva, 2012. Disponível em: <https://www.who.int/publications/i/item/WHO-HSE-WSH-12.01>. Acesso em: 21 nov. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **1 in 4 health care facilities lacks basic water services – UNICEF, WHO**. Geneva, 2019. Disponível em: <https://www.who.int/news/item/03-04-2019-1-in-4-health-care-facilities-lacks-basic-water-services-unicef-who>. Acesso em: 21 nov. 2025.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Burden of disease attributable to unsafe drinking-water, sanitation and hygiene: 2019 Update**. Geneva, 2019. Disponível em: <https://www.who.int/publications/i/item/9789240075610>. Acesso em: 21 nov. 2025.

PENG, Y. et al. An empirical study of classification algorithm evaluation for financial risk prediction. **Official Journal of the World Federation on Soft Computing**, v. 11, n. 2, p. 2906–2915, mar. 2011.

TAYLOR & FRANCIS GROUP. **Feature Engineering and Selection: A Practical Approach for Predictive Models**. 2019. Disponível em: <http://www.feat.engineering>. Acesso em: 21 nov. 2025.

VÁSQUEZ, W.; ALICEA-PLANAS, J. Factors associated with nonpayment behavior in the water sector of nicaragua. **Utilities Policy**, v. 47, p. 50–57, 2017.

WICKHAM, H. Tidy data. **Journal of Statistical Software**, v. 59, n. 10, p. 1–23, 2014.

ZAIED, Y. et al. Delayed payment of residential water invoice and sustainability of water demand management. **Journal of Cleaner Production**, v. 276, n. 123517, 2020.