Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

HELTON SOUZA LIMA

R3D: UMA ABORDAGEM UTILIZANDO CLASSIFICAÇÃO TRADICIONAL E SENSÍVEL AO CUSTO PARA PREDIÇÃO DE PEDIDOS DE REVISÃO DE DÍVIDA ATIVA

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2023

Helton Souza Lima

R3D: uma abordagem utilizando classificação tradicional e sensível ao custo para predição de Pedidos de Revisão de Dívida Ativa

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dra. Damires Yluska Souza Fernandes
Coorientador: Prof. Dr. Thiago José Marques Moura

João Pessoa

2023

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha do IFPB, *campus* João Pessoa

L732r Lima, Helton Souza.

R3D : uma abordagem utilizando classificação tradicional e sensível ao custo para predição de pedidos de revisão de dívida ativa / Helton Souza Lima. - 2023.

61 f. : il.

Dissertação (Mestrado -Tecnologia da Informação) - Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2023.

Orientação : Profº. D.ra Damires Yluska Souza Fernandes.

Coorientação : Profº D.r Thiago José Marques Moura.

1.Análise preditiva. 2. Ciência de dados – metodologia CRISP - DM. 3. Algoritmos. 4. Administração tributária – dívida ativa. 5. Métodos para resolução de problemas. I. Título.

CDU 004.02:336.22(043)

Lucrecia Camilo de Lima
Bibliotecária – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

HELTON DE SOUZA LIMA

R3D: UMA ABORDAGEM UTILIZANDO CLASSIFICAÇÃO TRADICIONAL E SENSÍVEL AO CUSTO PARA PREDIÇÃO DE PEDIDOS DE REVISÃO DE DÍVIDA ATIVA

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 28 de fevereiro de 2023

Membros da Banca Examinadora:

Dr(a) Damires Yluska Souza Fernandes

IFPB - PPGTI

Dr(a) Thiago José Marques Moura

IFPB - PPGTI

Dr(a) Alex Sandro da Cunha Rêgo

IFPB (Membro Interno)

Dr(a) Ticiana Linhares Coelho da Silva

UFC (Membro Externo)

João Pessoa/2023

Documento assinado eletronicamente por:

- **Damires Yluska de Souza Fernandes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/03/2023 13:50:15.
- **Thiago Jose Marques Moura**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/03/2023 20:26:46.
- **Alex Sandro da Cunha Rego**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 03/03/2023 11:36:47.
- **Ticiana Linhares Coelho da Silva**, **PROFESSOR DE ENSINO SUPERIOR NA ÁREA DE ORIENTAÇÃO EDUCACIONAL**, em 23/03/2023 14:22:24.

Este documento foi emitido pelo SUAP em 14/02/2023. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código 388883
Verificador: 15e554ea98
Código de Autenticação:

AGRADECIMENTOS

Após 3 anos de mestrado, percebo um aumento substancial na minha bagagem de conhecimentos. Posso relacionar esse aumento em relação a matérias específicas como ciência de dados, engenharia de sotware e inovação. Entretanto, eu elencaria o aprendizado a respeito do que é ciência e de como colaborar com a ciência como o que mais agregou valor à minha formação. Sinto que valeu a pena todo o esforço e dedicação empreendidos neste ciclo, que está chegando ao fim. O sentimento de gratidão me acompanhou durante toda essa jornada e fico muito feliz em ter a oportunidade de registrar uma parte desse sentimento aqui.

Primeiramente gostaria de agradecer à minha esposa Bruna, que me apoiou desde o primeiro dia que demonstrei interesse em fazer o curso de mestrado. Minha maior incentivadora, adaptou os horários do seu próprio trabalho para assumir algumas das minhas funções domésticas. Sempre paciente e amorosa, compreendeu meus momentos de indisponibilidade, incluindo os que não haviam sido combinados previamente.

Gostaria de agradecer aos meus filhos Sophia e Henrique e ao meu enteado Igor, pois souberam compreender e me ajudar no que estava ao alcance deles. Só pelo motivo de existirem, eles me fazem querer ser uma pessoa cada vez melhor.

Também gostaria de agradecer à professora Damires pela intensa dedicação a nosso trabalho. Sua brilhante orientação soube dosar os momentos de maior aproximação, onde há mais colaboração mútua, com aqueles de maior afastamento, onde há o ganho de autonomia para quem está aprendendo, mas que demanda grande paciência para quem está ensinando. Seu exemplo é fonte de motivação para nós.

Ao meu co-orientador, professor Thiago, agradeço pela sua disposição em estar nesta jornada, mesmo com seus ajustes de rota, acelerações e desacelerações. Nas inúmeras reuniões, está sempre pronto a contribuir com seu conhecimento, a participar do trabalho em equipe e nos motivar ainda mais.

Agradeço ao nosso consultor de negócio, Dr. Daniel Sabóia, que nos apresentou o desafio e esteve sempre presente e confiou em nosso grupo de pesquisa para ser um braço auxiliar para os seus objetivos.

Gostaria de agradecer ao nosso coordenador, professor Francisco Petrônio, que foi muito além de suas obrigações. Por fim, ao Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal da Paraíba pela oportunidade de aprendizado concedida. Tive excelentes professores e ótimas instalações e ferramentas à minha disposição. Em especial, incentivos financeiros para participação em eventos, me proporcionando uma formação ainda mais completa. Em especial, pelo sucesso na adaptação das metodologias de ensino para enfrentamento da situação da pandemia do covid-19.

RESUMO

A administração tributária é uma área complexa existente em governos de todo o mundo. As melhorias realizadas em seus processos operacionais aumentam a arrecadação e recuperação tributária. No âmbito da administração tributária brasileira, encontra-se disponível na internet um serviço para que os contribuintes possam registrar pedidos de ajustes nos processos de cobrança de dívidas. O referido serviço recebe um grande volume de pedidos e atualmente apresenta um alto tempo de resposta. Este trabalho tem como objetivo propor uma abordagem utilizando classificação tradicional e sensível ao custo capaz de predizer a classificação do resultado final para novos registros do serviço de Pedido de Revisão da Dívida Inscrita como apoio à decisão dos procuradores responsáveis pela análise dos pedidos. A abordagem, chamada de R3D, utiliza modelos de classificação tradicionais e sensíveis ao custo de forma híbrida para predizer novos pedidos do referido serviço. Os classificadores sensíveis ao custo têm demonstrado serem mais efetivos quanto ao custo do que os classificadores tradicionais em muitos problemas de classificação do mundo real, evitando grandes perdas financeiras. Isso geralmente é medido por uma métrica chamada de *savings*. Por outro lado, os classificadores sensíveis ao custo têm apresentado, em alguns casos, uma maior quantidade de erros de classificação, reduzindo valores de métricas de desempenho convencionais como acurácia, *f-score*, precisão e sensibilidade. Considerando a administração tributária e outros problemas de classificação no âmbito de órgãos públicos, é importante evitar perdas financeiras, porém, a procura na obtenção do menor custo não deve ser uma prioridade exclusiva. A abordagem R3D desenvolvida alcançou resultados mais equilibrados ao considerar, simultaneamente as métricas de *savings*, acurácia, *f-score*, precisão e sensibilidade de acordo com as regras de negócio e dos dados disponíveis.

Palavras-chaves: Análise preditiva, classificação sensível ao custo, *Savings* score, Administração tributária.

ABSTRACT

Tax administration is a complex area that exists in governments around the world. The improvements made to its operational processes increase tax collection and recovery. In Brazil, a service is available on the internet so that taxpayers can register requests for adjustments in debt collection processes. This service receives a large volume of requests and currently has a high response time. In this light, this work proposes an approach involving classification models applied to the problem of predicting new requests for the referred service. The approach, called R3D, allows the application of traditional and cost-sensitive classification models in a hybrid way to classify new requests for that service. Cost-sensitive classifiers have been shown to be more cost-effective than traditional classifiers in many real-world classification problems, avoiding large financial losses. This is usually measured by a metric called *savings*. On the other hand, cost-sensitive classifiers have shown, in some cases, a greater amount of classification errors, reducing traditional performance metrics such as accuracy, *f-score*, precision and sensitivity. Considering tax administration and other problems of classification within public domain, it is important to avoid financial losses, however, the pursuit for the lowest cost should not be an exclusive priority. The proposed R3D approach achieves more balanced results when simultaneously considering *savings*, accuracy, *f-score*, precision and sensitivity according to business rules and available data.

Key-words: Predictive analysis, Cost-sensitive classification, *Savings* score, Tax administration.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AB	Adaboost
AB	Adaboost Sensível ao Custo Dependente de Exemplo
AD	Árvore de Decisão
ADSCDE	Árvore de Decisão Sensível ao Custo Dependente de Exemplo
AM	Aprendizado de Máquina
Bag	*Bagging*
BagSCDE	*Bagging* Sensível ao Custo Dependente de Exemplo
BMR	*Bayes Minimum Risk*
CRISP-DM	*Cross Industry Standard Process for Data-Mining*
DAU	Dívida Ativa da União
IBPT	Instituto Brasileiro de Planejamento e Tributação
KDD	Knowledge Discovery in Databases
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PGFN	Procuradoria-Geral da Fazenda Nacional
PRDI	Pedido de Revisão de Dívida Inscrita
RF	*Random Forest*
RFSCDE	*Random Forest* Sensível ao Custo Dependente de Exemplo
RL	Regressão Logística
RLSCDE	Regressão Logística Sensível ao Custo Dependente de Exemplo
RFB	Receita Federal do Brasil

SUMÁRIO

1 INTRODUÇÃO

Este capítulo introduz o contexto de negócio no qual esta pesquisa está inserida, seguida da apresentação da motivação e das questões de pesquisa e contribuições esperadas. Ao final, é indicada a estrutura deste documento.

1.1 Contexto de Negócio

A evasão fiscal acarreta impacto negativo na qualidade de vida dos cidadãos de qualquer território pois, sem uma arrecadação de impostos adequada, não é possível manter serviços públicos essenciais como os de saúde, saneamento básico, mobilidade urbana, segurança, educação, entre outros (MATHEWS et al., 2018). Além do problema da ilegalidade presente na evasão fiscal, existe uma consequência que é a de concorrência desleal, quando a empresa que sonega está no mesmo ambiente de negócios de outras empresas que pagam os impostos regularmente.

A atribuição de combater a evasão fiscal é do setor responsável pela administração tributária em cada governo, seja em nível federal, estadual ou municipal. Conforme apresentado na Figura 1, a administração tributária geralmente é responsável por um conjunto de processos sequenciais que dizem respeito ao contato/relação com o contribuinte (OECD, 2020), podendo variar um pouco dependendo da natureza da taxação. Esses processos geralmente compreendem algumas etapas que se iniciam a partir de um Cadastro, que consiste em realizar a identificação do contribuinte (pessoa física ou jurídica). Na etapa de Taxação aplicam-se regras tributárias e calcula-se o valor do imposto devido. Para alguns tipos de impostos, existe uma etapa de Verificação, na qual realiza-se a declaração de transações e rendimentos e suas respectivas comprovações. Na etapa de Coleta, acontece o pagamento de impostos.

Figura 1 – Macro-processos geralmente encontrados na administração tributária.

Fonte: autoria própria, adaptado de OECD (2020)

Quando um determinado tributo não é pago até a sua data de vencimento, geralmente o valor devido passa a ser considerado uma dívida ou débito. Esse valor passa a ser cobrado através de processos que podem ser administrativos ou judiciais, na chamada etapa de Disputa. É na etapa de Disputa que o escopo desta proposta está inserido, por lidar com dados de dívidas

registradas em processo administrativo. É importante também destacar que existem serviços ofertados aos contribuintes em cada uma das etapas.

No Brasil, considerando a esfera federal, a Procuradoria-Geral da Fazenda Nacional (PGFN) é o órgão responsável pelos processos de cobrança de dívidas tributárias junto aos contribuintes pessoas físicas ou jurídicas (PGFN, 2021a). A PGFN, atualmente subordinada à Advocacia-Geral da União (AGU[1]), é composta por um corpo de procuradores cujas atribuições residem, principalmente, na representação da União em causas fiscais, na cobrança judicial e administrativa de qualquer débito (tributário e não-tributário) em que a União é a parte credora e, por fim, no assessoramento e consultoria jurídica junto aos outros órgãos para assuntos fiscais.

O Pedido de Revisão de Dívida Inscrita (PRDI) é um dos serviços oferecidos pela PGFN aos contribuintes, disponível desde 2018 (PGFN, 2021b). Consiste em permitir que o contribuinte realize uma solicitação para reanálise da situação da dívida tributária.

O PRDI é oferecido através do site[2] de serviços da PGFN, onde o contribuinte pode solicitar que a situação do débito seja reanalisada em casos de: informar que já houve pagamento do débito, informar que o débito foi parcelado e está sendo pago, solicitar suspensão do débito por decisão judicial, informar que o débito prescreveu, solicitar compensação do débito entre outros. Após a escolha do tipo do pedido, o contribuinte pode encaminhar diversas informações específicas do pedido, submeter o pedido para análise das equipes da PGFN e ficar no aguardo do seu resultado.

Após o registro do PRDI, o pedido atravessa um fluxo, ilustrado na Figura 2, onde, após o cadastro do registro na internet, o pedido passa inicialmente por uma fase de triagem, etapa esta na qual decide-se a equipe de especialistas que realizará a análise final do pedido. Na etapa da análise final, o pedido pode ser DEFERIDO ou INDEFERIDO, sendo o resultado disponibilizado ao contribuinte. Se o pedido for deferido, o débito pode ser cancelado, retificado ou suspenso. No caso de um indeferimento, o débito mantém sua validade e passa a seguir às demais etapas dentro da PGFN. Todo o processo do PRDI é monitorado pelas equipes de gestão da PGFN.

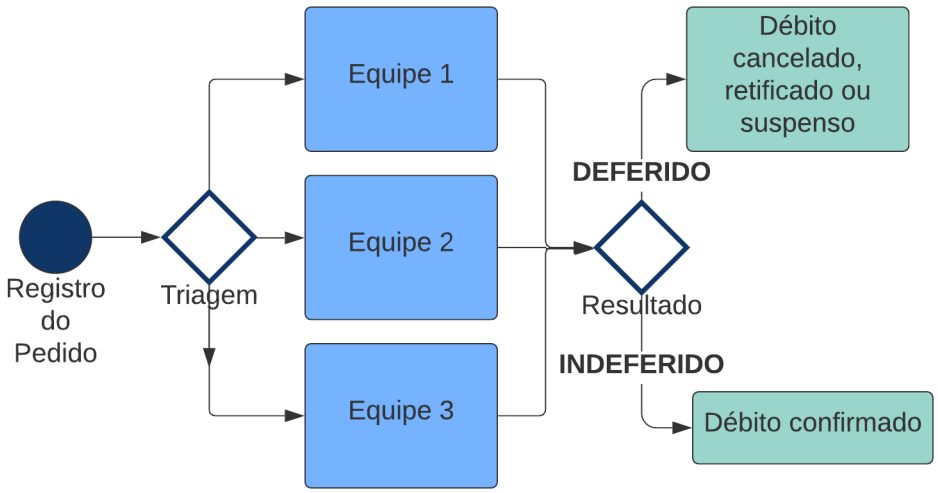
1.2 Motivação e Definição do Problema

O PRDI é um dos processos existentes dentro do contexto organizacional da PGFN. O serviço vem sendo cada vez mais procurado pelos contribuintes, de maneira que, em 2019, foram registrados 44 mil pedidos e, no ano de 2020, foram registrados 116,9 mil pedidos.

Para que um procurador da PGFN possa analisar um pedido e concluir sobre o seu deferimento ou indeferimento, o mesmo realiza diversos levantamentos de informação para subsidiar a análise, acessando vários sistemas diferentes, com o objetivo de obter o máximo de informações possíveis (atuais e históricas) do contribuinte e da dívida em si. A etapa de

[1] www.gov.br/agu/pt-br/acesso-a-informacao/institucional/sobre
[2] https://www.regularize.pgfn.gov.br/

Figura 2 – Fluxo de etapas de análise do PRDI.



O PRDI se inicia com seu registro pelo contribuinte, é tratado por uma das equipes da PGFN e, por fim, o pedido pode ser deferido ou indeferido. Fonte: autoria própria.

análise apresenta alto nível de complexidade e, portanto, consome uma quantidade de tempo considerável. O tempo médio de resposta a um PRDI é de 30 dias desde o seu registro até a emissão do parecer final.

O tempo de resposta médio do PRDI tem preocupado a administração da PGFN. Considerando que vem sendo registrado um aumento na utilização do serviço, identificado em mais de 150% de aumento no volume de pedidos entre 2019 e 2020, a tendência é de também aumentar o tempo médio de resposta.

Alguns dos mecanismos considerados para ajudar na aceleração das análises envolvem a utilização de soluções de software que possam servir como suporte à decisão dos procuradores. Este trabalho encontra-se inserido nesse contexto e busca desenvolver um mecanismo computacional de apoio à decisão para ser utilizado durante a etapa de análise de Pedidos de Revisão de Dívida Inscrita da União.

Neste sentido, evidências científicas apontam que a produtividade de uma organização aumenta quando há um maior uso de mecanismos de apoio à decisão baseado em dados (PROVOST; FAWCETT, 2013). Especificamente para a administração tributária, o relatório publicado pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE) evidencia que a utilização de técnicas de aprendizado de máquina está relacionada ao aumento da eficiência em diversos processos deste domínio de negócio (OECD, 2016). Neste sentido, alguns exemplos que se apresentam são a detecção de possíveis contribuintes inadimplentes e a melhoria de serviços junto aos contribuintes (OECD, 2020).

À medida que os processos da administração tributária, que antes eram baseados na utilização de papel, vão migrando para a era digital, mais dados vão sendo captados. Esse

grande volume de dados tem aumentado as oportunidades de utilização dos mesmos em diversos processos de cobrança, com o objetivo de aumentar a arrecadação tributária (NAZAROV; MIKHALEVA; CHERNOUSOVA, 2019). Trabalhos acadêmicos, conforme discutidos na Seção 2.4, também têm demonstrado que o uso de técnicas de aprendizado de máquina pode aumentar a eficiência e eficácia na administração tributária.

A partir da disponibilização de um conjunto de dados com informações históricas do PRDI, foram construídos modelos de classificação para predizer o resultado da análise de novos registros de PRDI, em "deferido"ou "indeferido"(LIMA et al., 2021). Além da predição de cada pedido, os modelos utilizados também emitem a informação da probabilidade estimada para determinada predição. Ou seja, além de predizer se determinado PRDI será provavelmente deferido ou indeferido, os modelos informam qual a probabilidade de confiança desta predição.

Os resultados do referido trabalho demonstraram que o uso de modelos de classificação podem ser úteis dentro do processo de análise do PRDI. Em um primeiro momento, a informação da predição de um pedido pode auxiliar o procurador responsável, de forma complementar às informações normalmente levantadas pelo procurador. Essa complementação pode levar a um aumento na confiança na decisão a ser tomada e, consequentemente, na diminuição do tempo de análise. Em um segundo momento, avaliar situações de análise automática dos pedidos, fornecendo respostas imediatas, com o objetivo de dar uma maior vazão na fila dos pedidos.

Para que seja definido o melhor modelo que possa ser aplicado para o problema de classificação do PRDI, é necessário que seja utilizada uma métrica de avaliação que melhor se adeque ao contexto do negócio e que possa representar os melhores resultados esperados, atendendo aos objetivos do domínio em questão. No caso do serviço do PRDI, é importante que o classificador apresente uma maior quantidade de acertos em suas predições, ou seja, apontar corretamente o deferimento ou indeferimento dos pedidos. Em relação aos erros de classificação, é preciso considerar que há uma diferença na gravidade de um falso positivo e de um falso negativo, a ser detalhada posteriormente. Além disso, existem pedidos que envolvem dívidas com valor baixo e algumas dívidas que envolvem valores muito altos em relação à maioria. Dessa forma, é importante que o classificador não incorra em erros de classificação, especialmente se o pedido envolver dívidas com alto valor.

Algoritmos de classificação tradicionais assumem que todos os erros de classificação carregam o mesmo custo (HARRINGTON, 2012). No âmbito desta dissertação, estes algoritmos também são chamados de insensíveis ao custo. No caso do PRDI, modelos de classificação tradicionais (insensíveis ao custo) podem correr o risco de deferir um pedido envolvendo milhões de reais em dívidas que não deveriam ser perdoadas.

Em diversas outras aplicações do mundo real, o custo relacionado a diferentes erros de classificação também podem ser diferentes. Por exemplo, falhar na aprovação de um empréstimo a um fraudador leva a perdas maiores do que negá-lo a um mutuário de boa fé. Métodos de classificação que levam em consideração diferentes custos de erros de classificação são

conhecidos como classificadores sensíveis ao custo (apresentados na Seção 2.3) (ELKAN, 2001; ZADROZNY; ELKAN, 2001).

Sendo assim, a questão de pesquisa que primeiramente norteia este trabalho é:

QP1 O uso de modelos sensíveis ao custo obterá resultados melhores do que os modelos tradicionais no que diz respeito ao custo envolvido nos erros de classificação quando aplicados ao problema de classificação do PRDI?

Alguns resultados dos experimentos utilizando modelos de classificação sensíveis ao custo aplicados ao problema de classificação do PRDI demonstraram que, em comparação aos modelos tradicionais, embora os modelos sensíveis ao custo tenham obtido melhores resultados em termos de custo/economia (medidos por uma métrica que contabiliza os custos envolvidos na classificação), eles apresentaram uma maior quantidade de erros de classificação, impactando métricas tradicionais, como a acurácia (fundamentada no Capítulo 2). Este comportamento também foi observado em outras referências na literatura (WU et al., 2012). No caso do problema de classificação do PRDI, modelos sensíveis ao custo podem evitar, por exemplo, o perdão equivocado de uma dívida com valor muito alto. Porém, ao priorizar o objetivo de minimizar os custos, esses modelos apresentam mais erros do que os modelos tradicionais. No âmbito do PRDI, identificou-se que esses erros estavam associados, em grande parte, a dívidas de menor valor.

Entretanto, para a administração tributária, é importante ser justo na análise dos pedidos, mesmo que a dívida possua um valor baixo, porque um dos objetivos nesse contexto de negócio é incentivar uma cultura de conformidade tributária. Ao mesmo tempo, ao lidar com dívidas contendo valores muito altos, é imprescindível não haver deferimentos equivocados, porque um outro objetivo desse contexto de negócio é aumentar as taxas de recuperação tributária [3].

De forma a buscar um maior equilíbrio entre uma menor quantidade de erros (obtida por modelos tradicionais) e um menor custo (obtido por modelos sensíveis ao custo), este trabalho propõe uma abordagem de classificação híbrida, chamada de R3D, que permite a integração dos dois tipos de modelos como um único modelo. Essa proposta busca responder, neste sentido, a uma segunda questão de pesquisa:

QP2 Como integrar modelos de aprendizado de máquina sensíveis ao custo e modelos tradicionais de forma a manter os níveis de custo de classificação atingidos pelos primeiros e mantendo uma menor quantidade de erros de classificação atingidos pelos segundos no âmbito do serviço do PRDI?

[3] https://www.gov.br/pgfn/pt-br/acesso-a-informacao/institucional/competencia

1.3 Objetivos

Com a finalidade de atender às questões de pesquisa apresentadas neste capítulo, estão descritos a seguir o objetivo geral da dissertação e os objetivos específicos para que as contribuições esperadas sejam alcançadas.

1.3.1 Objetivo geral

- Propor uma abordagem utilizando classificação tradicional e sensível ao custo capaz de predizer a classificação do resultado final para novos registros do serviço de Pedido de Revisão da Dívida Inscrita como apoio à decisão dos procuradores responsáveis pela análise dos pedidos.

1.3.2 Objetivos específicos

- Construir modelos de classificação utilizando algoritmos tradicionais com o objetivo de realizar a predição de resultados de novos registros de PRDI;
- Construir modelos de classificação utilizando algoritmos sensíveis ao custo com o objetivo de realizar a predição de resultados de novos registros de PRDI;
- Avaliar resultados dos modelos de classificação construídos de acordo com métricas de desempenho tradicionais e específicas escolhidas;
- Propor e desenvolver uma abordagem específica para o problema de classificação do PRDI de forma a se obter um maior equilíbrio entre os resultados dos modelos tradicionais e os modelos sensíveis ao custo;
- Avaliar resultados da abordagem desenvolvida de acordo com as métricas mais importantes para a classificação de novos registros do PRDI;

1.4 Estrutura do Documento

Os capítulos subsequentes estão organizados da seguinte forma: o Capítulo 2 introduz conceitos utilizados nesta dissertação e descreve os trabalhos relacionados, comparativamente. A abordagem R3D é apresentada no Capítulo 3. O Capítulo 4 descreve os experimentos realizados e seus respectivos resultados. Por fim, as conclusões e considerações finais estão expostas no Capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Este capítulo descreve os principais conceitos envolvidos na pesquisa realizada por este trabalho, iniciando com uma introdução à área de ciência de dados e modelos de processos de extração de conhecimento. Logo após, apresenta conceitos de aprendizado de máquina, particularmente o aprendizado supervisionado, seguido do detalhamento de alguns dos algoritmos de aprendizado supervisionado mais utilizados e algumas métricas de avaliação. Em seguida, são abordados algoritmos de aprendizado de máquina sensíveis ao custo. Por fim, descreve os trabalhos relacionados a esta proposta e pontua o diferencial deste trabalho em relação aos existentes na literatura.

2.1 Ciência de Dados e Metodologia CRISP-DM

Uma das formas de se produzir significado a partir de um conjunto de dados é utilizar mecanismos computacionais para embasar decisões (PROVOST; FAWCETT, 2013). Por exemplo, um publicitário pode selecionar alguns anúncios baseado na sua longa experiência em um determinado negócio. Ou então, pode-se basear a decisão em dados que indicam como são as diferentes reações dos consumidores quando apresentados a diferentes tipos de anúncios. Na verdade, pode-se utilizar um pouco das duas opções, pois o processo de decisão baseado em dados possui diferentes níveis de influência nos diversos pontos de tomada de decisão dentro de uma organização.

Termos como ciência de dados, mineração de dados e aprendizado de máquina têm sido usados de formas indistintas ou complementares. Apesar de existirem muitos elementos em comum com as áreas de aprendizado de máquina e mineração de dados, a ciência de dados é considerada mais ampla em seu escopo (PROVOST; FAWCETT, 2013). A Ciência de Dados compreende uma área multidisciplinar, envolvendo conhecimentos da estatística, da informática, da computação, de gerenciamento de dados e da sociologia, que estuda e faz uso de princípios e ferramentas com o objetivo de obter e gerar conhecimento a partir de dados (CAO, 2017). Em sistemas computacionais, a experiência existe na forma de dados, e o objetivo principal do aprendizado de máquina é o desenvolvimento e utilização de algoritmos que constróem modelos (mecanismos computacionais que realizam tarefas específicas) a partir dos dados (ZHOU, 2021b).

Para o desenvolvimento de projetos na área de ciência de dados, utiliza-se algum modelo de processo que possa nortear o cumprimento de suas etapas. A metodologia CRISP-DM (*Cross Industry Standard Process for Data-Mining*) (CHAPMAN et al., 1999) é uma das mais utilizadas. Ela provê diretrizes para as etapas mais comuns no contexto de projetos de ciência de

dados (MARTÍNEZ-PLUMED et al., 2019). O CRISP-DM é uma extensão das etapas originais propostas no modelo de processo *KDD* (*Knowledge Discovery in Databases (KDD)*) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), conhecido como um processo de descoberta de conhecimento a partir de dados.

No CRISP-DM todas as etapas do *KDD* estão presentes, porém, o CRISP-DM se diferencia ao dar um foco complementar ao domínio do negócio e por acrescentar uma etapa de implantação do modelo em ambiente de produção. O processo, que é iterativo e incremental, pode ser utilizado em projetos de diversos domínios de negócio no contexto da Ciência de Dados. Uma breve descrição das fases do processo, que não são necessariamente sequenciais, é apresentada a seguir:

- **Entendimento do negócio**: inclui a definição de objetivos do negócio e os objetivos da(s) tarefa(s) de mineração de dados, geralmente produzindo um plano de projeto. Para o problema de classificação do PRDI, abrange o conhecimento dos objetivos envolvendo o serviço de Pedidos de Revisão de Dívida Inscrita e a definição do objetivo da etapa de mineração de dados: classificação (Subseção 2.1.1).
- **Entendimento dos dados**: consiste em coletar os dados, descrevê-los e explorá-los, verificando sua qualidade. Em outras palavras, é necessário o conhecimento de todas as variáveis que compõem o conjunto de dados e como elas estão relacionadas com o domínio do problema.
- **Preparação dos dados**: compreende tarefas como a seleção, limpeza, integração e formatação dos dados.
- **Modelagem**: esta etapa é focada na construção de modelos de aprendizado de máquina, que inclui a seleção e aplicação do algoritmo de aprendizado a ser usado e o planejamento dos testes.
- **Avaliação**: promove a avaliação dos resultados dos testes a partir dos modelos criados. Os resultados são avaliados a partir de métricas como, por exemplo, a acurácia e a sensibilidade.
- **Implantação**: é a etapa responsável por planejar a implantação do modelo de aprendizado de máquina em ambiente de produção.

2.1.1 Aprendizado de Máquina

O aprendizado de máquina é uma área dedicada a construir programas de computador com a capacidade de "aprender", ou seja, de forma análoga ao aprendizado humano que aprende com base em experiências, os sistemas construídos com o uso de aprendizado de máquina têm a

capacidade de melhorar sua performance a partir de experiências (dados) anteriores (MITCHELL; MCGRAW-HILL, 1997).

Ainda que os sistemas atualmente não sejam capazes de aprender da mesma forma que os humanos aprendem, alguns algoritmos foram criados para lidar com certas tarefas de aprendizado. Por exemplo, resolver tarefas como detectar e-mails como sendo de *spam* observando a ocorrência de uma única palavra pode não ser muito eficaz. Mas observar ocorrências de algumas palavras sendo usadas ao mesmo tempo, combinado com a informação do tamanho do e-mail e outros fatores, pode trazer uma ideia mais clara se o e-mail é um *spam* ou não. Outros exemplos de utilização com sucesso do aprendizado de máquina são o reconhecimento de fala, carros autônomos, reconhecimento de imagens e recomendação de produtos em lojas de varejo virtuais (HARRINGTON, 2012).

As tarefas do aprendizado de máquina estão divididas em dois tipos mais comumente referenciados (ALPAYDIN, 2010): supervisionado e não-supervisionado. No primeiro tipo, que será explorado no escopo desta dissertação, o objetivo é aprender uma função de mapeamento através de exemplos de entrada e saída com valores reais ou simulados, este último fornecido por um supervisor. No segundo tipo, não há supervisor e, portanto, não há mapeamento entre entradas e saídas, existindo apenas os dados de entrada. O objetivo, no caso de aprendizado não-supervisionado, é encontrar normalmente similaridades e/ou relacionamentos entre os dados observados.

Um exemplo de aprendizado não-supervisionado é a segmentação de clientes em que, a partir de dados como endereço, idade e lista de produtos comprados por cada cliente, é possível formar grupos com perfis semelhantes com o objetivo de direcionar ações específicas de relacionamento para diferentes grupos de clientes. A identificação de diferentes agrupamentos entre os registros de um conjunto de dados é também chamada de *clusterização* (HARRINGTON, 2012).

Outro exemplo de aprendizado não-supervisionado consiste nas análises de regras de associação. Trata-se de encontrar relacionamentos entre itens em um conjunto de dados, podendo ser um grupo de itens com alta frequência de ocorrerem juntos ou a identificação de regras de associação entre dois itens (HARRINGTON, 2012). Por exemplo, em um sistema de um supermercado, um grupo de alta frequência pode ser o conjunto dos seguintes itens: vinho, fraldas, soja e leite. Uma regra de associação que pode ser encontrada é que, se alguém compra fraldas, existem grandes chances dessa pessoa também comprar vinho.

O aprendizado supervisionado se propõe a resolver problemas cujo objetivo é determinar uma função que, para uma dada entrada X (que consiste em um conjunto de dados com alguma quantidade de atributos), resulte em uma saída Y (que consiste em um conjunto de rótulos para as classes do problema). É dito supervisionado pela existência de exemplos de instâncias da entrada X terem a saída Y previamente conhecida. As variáveis de saída podem ser categóricas (qualitativas) ou numéricas (quantitativas). As variáveis de saída categóricas são definidas por um

conjunto finito de valores, podendo ser chamados de rótulos. Por outro lado, as variáveis de saída numéricas ou quantitativas são definidas por um conjunto contínuo ou discreto de dados, que podem ser números reais ou o conjunto (ou um subconjunto) dos números inteiros (ALPAYDIN, 2010).

Para os casos em que a variável pertence a um intervalo contínuo de números, a tarefa é chamada de regressão (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). Por exemplo, para o caso de concessão de crédito, além de classificar o cliente em baixo ou alto risco, a instituição bancária precisa também, antes de oferecer um valor de empréstimo, saber qual será o valor oferecido. Neste exemplo, o ideal é que o sistema não precise encaixar o cliente em valores pré-determinados, porém possa definir um valor customizado para cada cliente, sendo uma variável definida em um intervalo contínuo de números e, portanto, um problema de regressão.

Quando a tarefa de aprendizado supervisionado resulta em uma variável categórica ou em um número que corresponda a uma determinada classe, essa tarefa é chamada de classificação (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). Por exemplo, em instituições bancárias, é muito útil poder, antes de oferecer um empréstimo a um cliente físico ou jurídico, ter um indicativo de que esse cliente possui uma alta ou baixa probabilidade de pagar o empréstimo. Neste caso, é possível criar um problema de classificação em que, a partir dos dados anteriores de diversos empréstimos realizados pela instituição no passado, relacionados a informações inerentes à capacidade financeira dos clientes (e.g., renda, gastos, profissão, idade, histórico financeiro, etc), seja possível classificar um determinado cliente em duas classes distintas: **alto risco** ou **baixo risco**. Este exemplo diz respeito a um problema de classificação binária, pois envolve apenas duas classes. De forma similar, a classificação de risco também se aplica aos contribuintes dentro de órgãos da administração tributária.

Para que um modelo classificador seja criado, uma parte do conjunto de dados rotulados disponíveis é utilizado para alimentação do modelo, em uma etapa denominada de treinamento. O subconjunto dos dados utilizados para execução desta fase é chamado de dados de treinamento. Uma outra parte dos dados, ainda não conhecida pelo modelo, é utilizada para verificar se as predições realizadas pelo classificador estão corretas ou não. Este último subconjunto é denominado de dados de teste (HARRINGTON, 2012).

Quando o modelo (algoritmo já treinado) passa a classificar novas instâncias cuja classificação ainda não é definida, pode-se utilizar os verbos predizer/estimar/prever, no sentido de que o classificador está se antecipando na identificação da classe para a qual determinado registro em um conjunto de dados será classificado.

Como o problema de classificar PRDI envolve apenas duas classes (deferido e indeferido), a escolha e uso dos algoritmos será focado em classificação binária. As seções seguintes apresentam brevemente os algoritmos de classificação utilizados neste trabalho, divididos em algoritmos de classificação tradicionais (insensíveis ao custo) e algoritmos de classificação sensíveis ao custo.

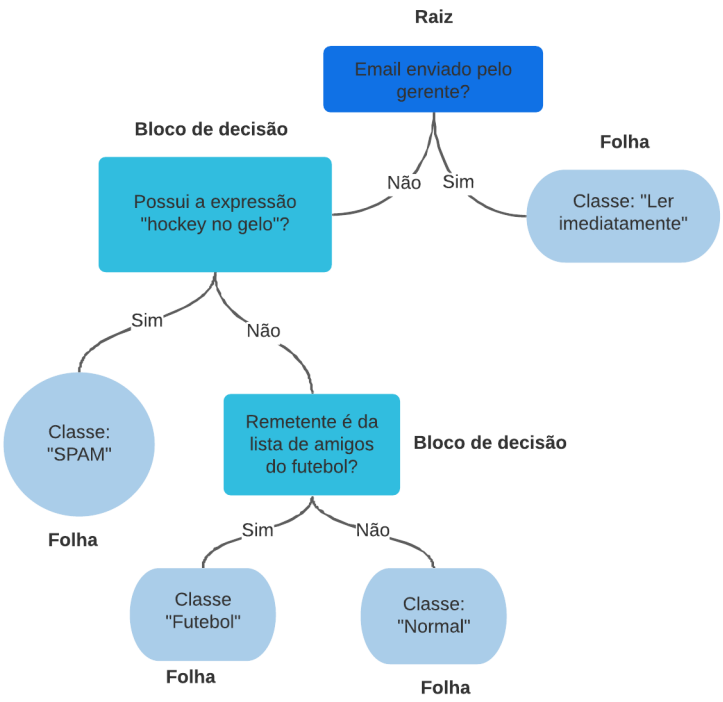
2.2 Classificação tradicional

Nesta seção são apresentados brevemente alguns dos principais algoritmos de aprendizado supervisionado para a tarefa de classificação e que são utilizados nesta dissertação.

2.2.1 Árvore de decisão

O algoritmo de árvore de decisão possui estruturas denominadas de nós, nos quais realiza testes baseados no valor das variáveis dos dados recebidos como entrada produzindo uma decisão de caminho a ser percorrido até um novo nó intermediário ou um nó final, na qual uma decisão final é processada. Esse encadeamento de nós é realizado de forma hierárquica em que o primeiro nó é chamado de raiz e os nós finais são chamados de folhas (HARRINGTON, 2012).

Figura 3 – Exemplo de aplicação de árvore de decisão.



Fonte: autoria própria, adaptado de Harrington (2012)

A Figura 3 ilustra um exemplo de árvore de decisão em uma aplicação de gerenciamento de e-mails, na qual realiza a classificação do e-mail recebido de acordo com as classes representadas pelos blocos finais: se o e-mail tiver sido encaminhado pelo gerente, então o e-mail será classificado como "ler imediatamente"; caso contrário, se possuir a expressão "hockey no gelo", então deverá ser classificado como "SPAM"; caso contrário, se o remetente está incluído na lista de amigos, então será classificado como "Futebol", senão, por fim, será classificado como "Normal".

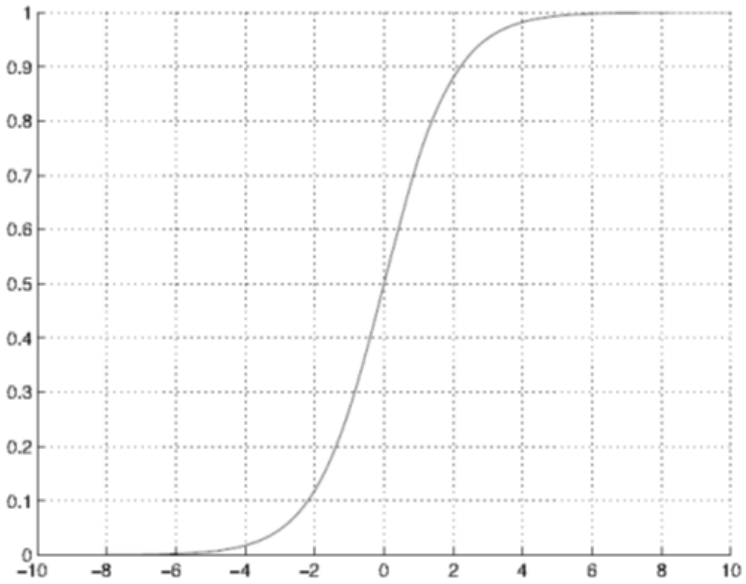
O processo de construção de uma árvore de decisão envolve a criação de novos nós e uma etapa final chamada de poda. A árvore de decisão é construída na fase de treinamento do modelo, em que se utiliza de um cálculo que busca minimizar o fator de "impureza"definido. Métricas de impureza padrão podem ser a quantidade de erros de classificação ou entropia ou o cálculo do coeficiente de Gini (ALPAYDIN, 2010). Esse cálculo pode ser utilizado para definir a regra de criação de novos nós ou na fase de poda da árvore.

Árvores de decisão são classificadores bastante populares que têm as vantagens de exigirem pouco poder de processamento e, além disso, seus resultados são fáceis para a compreensão humana podendo, dessa forma, ser validados por especialistas de domínio (ALPAYDIN, 2010).

2.2.2 Regressão Logística

O algoritmo de Regressão Logística busca encontrar os parâmetros para uma função logística, chamada de sigmóide, que melhor se adeque aos dados de treinamento. A função sigmóide tem formato semelhante ao apresentado na Figura 4, em que há uma variação de 0 até 1 no eixo Y que indica uma probabilidade de pertencimento a determinada classe a partir da definição de um valor em X. A busca pelos parâmetros a serem utilizados na função sigmóide se dá através de algoritmos de otimização (HARRINGTON, 2012).

Figura 4 – Imagem de uma função sigmóide.



Fonte: Alpaydin (2010)

A predição realizada pela regressão logística se dá através do valor do eixo *y*, que possui valor entre 0 e 1. Por exemplo, caso o valor seja igual ou maior que 0.5, então será considerado *true*. Caso contrário, será considerado *false*. O valor exato do eixo *y* é considerado a probabilidade estimada da predição.

2.2.3 *Ensembles*

Algoritmos de classificação que geram diversos modelos que trabalham juntos para resolver um problema são chamados de modelos baseados em comitês, sistemas de múltiplos classificadores ou *ensembles*. A presente dissertação utiliza esta última nomenclatura.

Em geral, os *ensembles* são classificadores compostos por diversos modelos gerados por um ou mais algoritmos utilizados como base. Os modelos de base se diferenciam por terem sido criados por algoritmos diferentes ou apenas por terem sido criados a partir de diferentes dados utilizados no treinamento ou, finalmente, apenas pela mudança de alguns hiper-parâmetros. A classe final resultante do *ensemble* consiste em uma combinação das classes resultantes de cada modelo de base. A forma mais utilizada na combinação é o uso da regra do voto majoritário (simples ou ponderado) (ZHOU, 2021a) (POLIKAR, 2012). O interesse a respeito dos *ensembles* se tornou crescente nas últimas décadas principalmente após o trabalho de Hansen e Salamon (1990) concluir que as predições realizadas por uma combinação de redes neurais geralmente alcançam uma acurácia maior do que uma única rede neural. A atenção dispensada aos *ensembles* foi gradativa, pois verificou-se que os mesmos provaram ser bastante efetivos e extremamente versáteis em uma larga gama de problemas e aplicações no mundo real (ZHOU, 2009; SAGI; ROKACH, 2018).

Diversos *ensembles* foram desenvolvidos e publicados nos últimos anos, porém, muitos são variações de alguns algoritmos iniciais que se mostraram bem sucedidos após extensivos testes e publicações. Dois dos mais conhecidos desses métodos são chamados de *Bagging* e *Boosting* e são descritos a seguir.

O algoritmo de *Bagging* (*Bootstrap Aggregation*) (BREIMAN, 1996) é um dos mais antigos e simples entre os métodos *ensembles*. O algoritmo consiste em gerar subconjuntos dos dados de treinamento, cujas instâncias são escolhidas aleatoriamente. Para cada subconjunto são geradas cópias de algumas instâncias. A geração de cópias pode ser realizada com repetição de algumas instâncias e com ausência de outras instâncias, de forma que o subconjunto possa ter a mesma quantidade de instâncias que o conjunto de treinamento original. Esta técnica de geração de cópias é chamada de *bootstrap*.

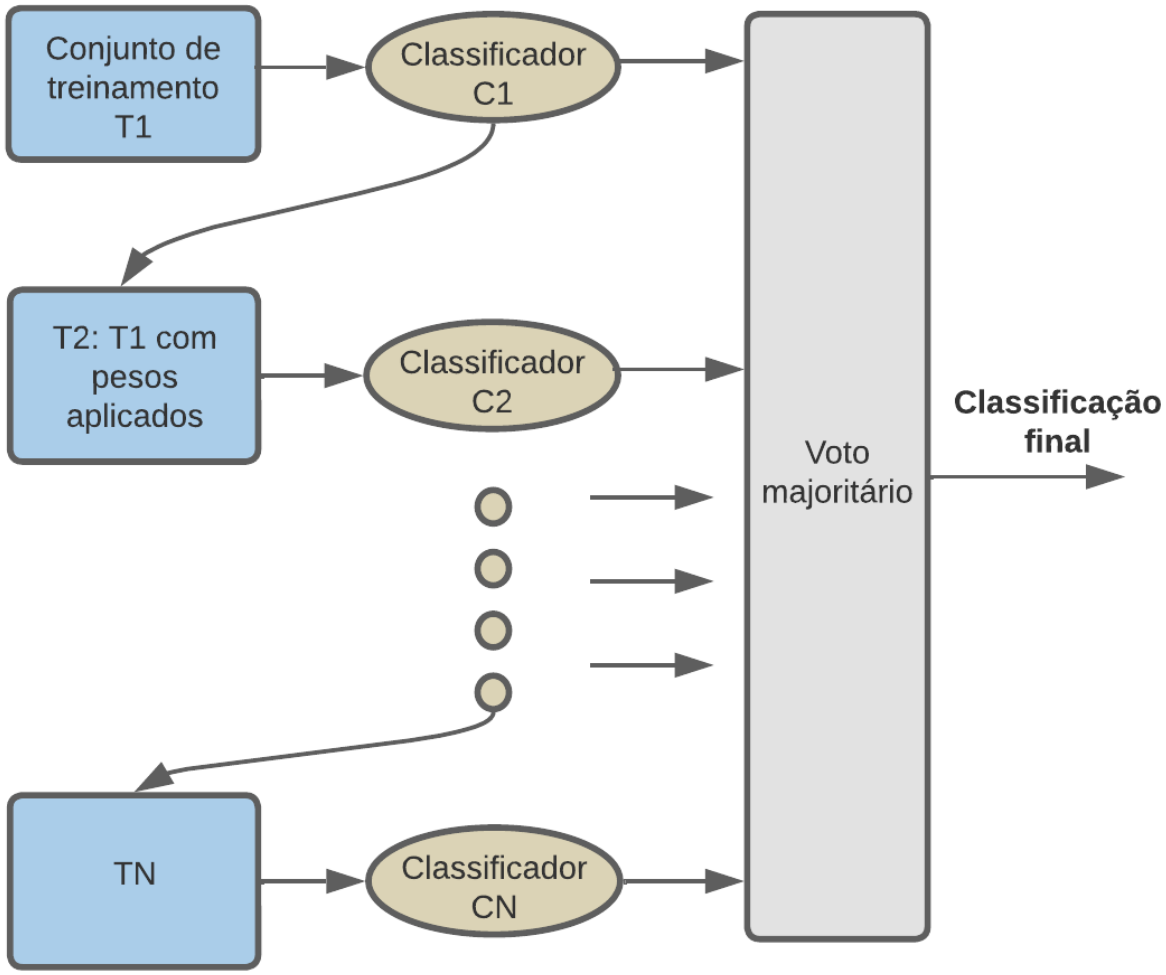
Cada modelo que faz parte do *ensemble* é treinado por um subconjunto diferente de todos os outros. Ao classificar uma nova instância, o *ensemble* utiliza, por exemplo, uma combinação por voto majoritário de todos os classificadores que o compõem (POLIKAR, 2012).

O *Random Forest* consiste em uma combinação de árvores de decisão (BREIMAN, 2001). Primeiramente, como no algoritmo de *bagging*, a fase de treinamento de cada árvore é alimentada por um subconjunto de treinamento diferente dos subconjuntos utilizados para as outras árvores de decisão da floresta. Além de utilizar a técnica de *bootstrap*, o *Random Forest* também utiliza apenas um subconjunto das variáveis para realizar o treinamento de cada árvore de decisão, técnica chamada de *random subspace*. Para cada novo exemplo a ser classificado,

todas as árvores classificam de forma independente, e a classe final escolhida consiste na classe com maior representatividade apresentada, seja por voto majoritário simples ou ponderado (BREIMAN, 2001).

O algoritmo de *Boosting* (FREUND; SCHAPIRE et al., 1996) também utiliza diferentes subconjuntos no treinamento dos classificadores, porém com dois diferenciais: i) não há utilização da técnica de reposição; ii) os subconjuntos com padrões em que são identificados erros de classificação por classificadores iniciais passam a ter uma maior probabilidade de serem selecionados para o treinamento dos classificadores subsequentes, dentro das demais iterações realizadas.

Figura 5 – Ilustração gráfica do algoritmo *AdaBoost*.



Fonte: adaptado de Hastie (2001)

Uma das primeiras e mais conhecidas implementações do algoritmo de *boosting* chama-se *AdaBoost* (*Adaptive Boosting*) (FREUND; SCHAPIRE et al., 1996), cuja ideia central é a utilização de pesos para as instâncias do conjunto de treinamento conforme ilustrado na Figura 5. O conjunto de treinamento original, chamado de T1, é utilizado no treinamento do primeiro classificador, quando os pesos são iguais para todas as instâncias. Em seguida, os pesos são

atualizados, sendo maior para as instâncias em que foram identificados erros de classificação do primeiro classificador. O conjunto de treinamento T2, juntamente com os pesos atualizados, são utilizados pelo novo classificador. Isso permite que cada novo classificador possa dar maior ênfase para as instâncias que não foram classificadas corretamente nas iterações anteriores. A classificação final é decidida por voto majoritário de todos os classificadores.

2.2.4 Avaliação de modelos tradicionais

Para que seja possível a realização de uma comparação entre os diferentes algoritmos de classificação é necessária a utilização de métricas que indicam a quantidade de classificações acertadas e erradas na fase de teste. Essas métricas possibilitam que as observações sejam padronizadas e haja uma comparação justa entre os modelos gerados e as diferentes experiências.

Para entender as métricas de avaliação, é fundamental, primeiramente, entender como funciona a matriz de confusão, que é uma representação sumária dos resultados das predições bastante conhecida e utilizada (HARRINGTON, 2012). Trata-se de uma tabela construída com as quantidades de erros e acertos realizados pelo classificador. Em uma aplicação para classificação binária (que é o caso do problema de origem desta dissertação), constrói-se uma tabela com duas linhas e duas colunas: a primeira linha contendo a quantidade de acertos da primeira classe seguida da quantidade de erros da primeira classe; a segunda linha contendo mais dois números informando a quantidade de erros e acertos da segunda classe. A Tabela 1 apresenta uma matriz de confusão para uma aplicação utilizando classificação de pacientes com probabilidade de estarem com doença cardíaca. A classe "Doente"é considerada como o valor positivo (ou seja, da classe de interesse) da classificação, enquanto a classe "Saudável"é considerada o valor negativo. Quando o classificador realiza uma predição corretamente de um caso (ou instância) real de doença, diz-se que se trata de um verdadeiro positivo (VP). Quando o classificador realiza uma predição incorreta de um caso como doença quando se trata de um caso saudável, diz-se que se trata de um falso positivo (FP). O mesmo raciocínio se aplica às predições de casos como sendo saudáveis.

Real / Predição	Doente	Saudável
Doente	30 (VP)	2 (FN)
Saudável	5 (FP)	95 (VN)

Tabela 1 – Matriz de Confusão para uma aplicação de classificação de casos de doenças cardíacas.

Neste exemplo, a matriz de confusão mostra que 30 casos foram classificados corretamente como doentes e 95 casos foram classificados corretamente como saudáveis. Por outro lado, 5 casos foram classificados como doentes, mas eram de pessoas saudáveis, enquanto 2 casos foram classificados como saudáveis mas eram pessoas doentes. A soma dos números da diagonal verdadeira (30 + 95 = 125) indica quantos casos foram classificados corretamente pelo algoritmo, enquanto que a soma dos números da diagonal falsa (2 + 5 = 7) indica quantos casos foram classificados incorretamente.

A partir desses valores é possível calcular diferentes métricas de avaliação, quais sejam:

- **Acurácia:** A acurácia mede a taxa de predições corretas em relação ao total de exemplos avaliados e é definida pela seguinte fórmula (HOSSIN; SULAIMAN, 2015):

$$Acuracia = \frac{(VP + VN)}{(VP + VN + FP + FN)} \tag{1}$$

  No exemplo de doenças cardíacas, a acurácia é aproximadamente 0,95. A acurácia é a métrica de avaliação de modelos mais utilizada para classificações. É uma métrica com baixa complexidade em seu cálculo e mais facilmente entendida por humanos. Porém, por não fazer distinção em relação ao tipo do erro (falso positivo ou falso negativo), apenas com a acurácia não é possível avaliar se há algum tipo de erro de classificação ocorrendo com maior frequência.

  Para o problema de classificação do PRDI, é importante avaliar o percentual de classificações incorretas em relação ao volume total de pedidos. Neste contexto, os dois tipos de erros de classificação são indesejados, mesmo considerando que cada um dos tipos possui impactos ou gravidades diferentes para o referido contexto de negócio.

- **Precisão:** A precisão consiste no índice de classificações de uma determinada classe que foram corretas, definida pela seguinte fórmula:

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

  No exemplo de doenças cardíacas, a precisão é aproximadamente 0,86. A precisão indica o percentual de acertos considerando apenas a classe positiva. Para o problema de classificação do PRDI, a classe positiva significa o indeferimento do pedido, ou sua rejeição. O falso positivo significa que um pedido tem sua predição como indeferido, porém sua classe real é deferido. Ou seja, significa que a dívida será erroneamente confirmada, o que leva à continuidade de cobrança indevida por parte da administração tributária. Essa situação é indesejada considerando tanto a visão da administração tributária quanto a visão do contribuinte. Para o primeiro, significa que utilizará seus recursos para uma cobrança indevida de uma dívida. Para o segundo, significa optar por utilizar seus recursos para continuar pleiteando a extinção/correção da dívida ou pagará por uma dívida cobrada indevidamente.

  Portanto, a precisão foi incluída neste trabalho e, consequentemente, nos experimentos para que se possa avaliar se há uma grande quantidade de falsos positivos conforme o domínio de dados do PRDI.

- **Sensibilidade:** A sensibilidade (também chamada de cobertura ou *recall*) é uma métrica que avalia se há muitos falsos negativos em relação à quantidade de verdadeiros positivos,

através da seguinte fórmula:

$$Sensibilidade = \frac{VP}{(VP + FN)} \tag{3}$$

No exemplo de doenças cardíacas, a sensibilidade é aproximadamente 0,94. Para o problema de classificação do PRDI, a classe negativa significa o deferimento do pedido, ou seja, o seu aceite. O falso negativo significa que um pedido tem sua predição como deferido, porém sua classe real é indeferido. Para o contexto do negócio, significa que a dívida será erroneamente perdoada e o processo de cobrança encerrado. Essa situação não é desejável pois supostamente resulta na perda da arrecadação do valor da dívida, caso o processo de cobrança seja exitoso.

Portanto, a sensibilidade é uma métrica com grande importância para este trabalho, pois classificadores com alta sensibilidade não possuem um grande número de falsos negativos (HARRINGTON, 2012).

- **F-score:** Representa a média harmônica entre a sensibilidade (cobertura ou *recall*) e a precisão sendo definida pela seguinte fórmula:

$$F - score = \frac{2 * Precisao * Sensibilidade}{(Precisao + Sensibilidade)} \tag{4}$$

No exemplo de doenças cardíacas, o *f-score* é aproximadamente 0,90. Dessa forma, esta métrica é bastante útil, dentre outras, em que a acurácia pode estar com altos valores, porém o classificador não está com acertos significativos para a classe minoritária (HOSSIN; SULAIMAN, 2015).

Para o problema de classificação do PRDI, o *f-score* será utilizado para dar uma visão geral do classificador em termos de falsos positivos e falsos negativos, de forma complementar à acurácia e por ser utilizado em vários trabalhos relacionados, para fins de comparação.

2.3 Classificação sensível ao custo

Algoritmos de aprendizado tradicionais (insensíveis ao custo) buscam, em geral, maximizar a acurácia. Esses algoritmos assumem que o custo de uma classificação errada é o mesmo para qualquer caso ou instância (MITCHELL; MCGRAW-HILL, 1997). Algoritmos de aprendizado sensíveis ao custo buscam reduzir o custo dos erros de classificação mais do que a quantidade de erros de classificação (KIM et al., 2012).

Os algoritmos de aprendizado sensíveis ao custo podem ser divididos em dois tipos: dependentes de classe ou dependentes de exemplo.

2.3.1 Classificação sensível ao custo dependente de classe

Em muitas aplicações do mundo real, o custo dos falsos positivos e falsos negativos são diferentes. Por exemplo, em problemas de classificação na área médica, predizer que um paciente doente está saudável geralmente é um erro mais sério do que predizer que um paciente saudável está doente.

Um problema de classificação sensível ao custo é chamado de *dependente de classe* quando os custos são diferentes entre as classes, porém são constantes entre os exemplos/instâncias (ELKAN, 2001). Para um problema de classificação binária, os custos de classificação podem ser representados por uma matriz de custo, representada na Tabela 2, onde C_{FP} = custo de um falso positivo; C_{FN} = custo de um falso negativo; C_{VN} = custo de um verdadeiro negativo e C_{VP} = custo de um verdadeiro positivo.

Predição / Real	**Classe positiva**	**Classe Negativa**
Classe positiva	C_{VP}	C_{FP}
Classe negativa	C_{FN}	C_{VN}

Tabela 2 – Matriz de custo dependente de classe.

Abordagens de classificação sensíveis ao custo geralmente assumem um custo constante para cada tipo de erro de classificação (falso positivo e falso negativo). Em geral, também assumem que $C_{FP} \neq C_{FN}$, porque em um problema de classificação sensível ao custo dependente de classe, os custos entre as classes geralmente são diferentes. É também convencionado que C_{FP} deve ser maior que C_{VP} e C_{FN} deve ser maior que C_{VN}, porque o custo associado a classificações incorretas devem ser maiores do que o custo de classificações corretas quando é considerada a mesma classe.

Importante destacar que podem existir alguns problemas de classificação que envolvam custos nas classificações corretas. Por exemplo, em aplicações de detecção de fraudes em cartões de crédito, existe um custo associado à classificação correta de uma fraude. Provavelmente, será necessário bloquear a transação, realizar ligações para o cliente e demais tarefas de um processo de confirmação de uma fraude. Essas tarefas têm um custo associado, em termos de infraestrutura de comunicação e recursos humanos.

2.3.2 Classificação sensível ao custo dependente de exemplo

Abordagens de classificação sensíveis ao custo dependentes de classe podem não ser adequadas a diversos problemas do mundo real. Por exemplo, para o problema de detecção de fraudes em transações de cartões de crédito, uma transação fraudulenta pode estar associada a uma compra com pequeno valor monetário ou um enorme valor monetário. Para o problema de classificação do PRDI, as dívidas envolvidas nos pedidos vão desde dívidas de alguns milhares de reais até dívidas de mais de um bilhão de reais. Um erro ao deferir um pedido envolvendo uma dívida pequena deve ter um menor custo do que um erro ao deferir uma dívida bilionária.

Abordagens de classificação sensíveis ao custo *dependentes de exemplo* ocorrem quando os custos não são constantes entre os exemplos. Portanto, existe uma diferença na definição da matriz de custo, que representa diferentes custos para cada exemplo *i*, como ilustrado na Tabela 3.

Predição / Real	**Classe positiva**	**Classe Negativa**
Classe positiva	C_{VPi}	C_{FPi}
Classe negativa	C_{FNi}	C_{VNi}

Tabela 3 – Matriz de custo dependente de exemplo, onde C_{FPi} = custo do falso positivo para uma instância *i*; C_{FNi} = custo de um falso negativo para uma instância *i*; C_{VNi} = custo de um verdadeiro negativo para uma instância *i* and C_{VPi} = custo de um verdadeiro positivo para uma instância *i*.

Abordagens de aprendizado sensíveis ao custo dependentes de exemplo utilizam a matriz de custo tanto para a construção de modelos quanto para a realização de predições. A matriz de custo deve estar associada ao conjunto de dados em um formato de *array* com 4 posições unido a cada exemplo do conjunto de dados.

Os métodos de classificação sensíveis ao custo dependentes de exemplo podem ser agrupados em relação à fase em que os custos são inseridos na solução (BAHNSEN; AOUADA; OTTERSTEN, 2015b): (i) durante o treinamento ou (ii) após o treinamento. No primeiro caso, são realizadas modificações que fazem o algoritmo original de aprendizado levar em consideração o custo durante a fase de treinamento para produzir classificadores sensíveis ao custo. A seguir, são apresentados os algoritmos modificados de árvore de decisão, regressão logística, *bagging*, *Random Forest* e *Adaboost*. Por fim, no segundo caso, o método sensível ao custo é aplicado após o treinamento de um outro classificador tradicional.

Uma **árvore de decisão sensível ao custo dependente de exemplo (ADSCDE)** leva em consideração o custo de cada exemplo durante a criação de novos nós e ao realizar a poda da árvore (BAHNSEN; AOUADA; OTTERSTEN, 2015b). Ao invés de cálculos tradicionais como a entropia, o custo de cada nó é calculado para avaliar o ganho em termos de custos, de forma a minimizar o custo total do modelo. A fase de poda da árvore também é modificada para levar em consideração a minimização dos custos.

Métricas de impureza padrão como quantidade de erros de classificação, entropia ou coeficiente de Gini, consideram a distribuição das classes em cada folha/nó para avaliar o poder preditivo de uma regra de divisão de nós. Essas métricas visam minimizar uma menor quantidade de erros de classificação.

A métrica de impureza sensível ao custo considera os custos envolvidos quando todos os exemplos de um nó são classificados como positivos ou negativos, avaliando o menor custo esperado em cada regra de divisão. Essa métrica está definida na Equação 5.

$$I_c(S) = min(C(f_0(S)), C(f_1(S)) \tag{5}$$

onde $I_c(S)$ é a métrica de impureza para um conjunto de exemplo S em um nó; $C(f_0(S))$ e $C(f_1(S))$ são os custos quando todos os exemplos de um conjunto em um nó são classificados como positivos ou negativos, respectivamente. A classificação de cada conjunto de um nó é calculado como a predição que gera o menor custo (Equação 6):

$$f(S) = 0 \; se \; C(f_0(S)) \leq C(f_1(S)); \; 1 \; senao \tag{6}$$

No que diz respeito ao critério de poda baseado nos custos, segue o mesmo raciocínio da métrica de impureza apresentada. Nós que não contribuem para a minimização dos custos devem ser podados, independentemente do impacto desses nós para a acurácia do algoritmo.

A **regressão logística sensível ao custo dependente de exemplo (RLSCDE)** introduz os custos alterando a função objetivo do modelo para uma função sensível ao custo (BAHNSEN; AOUADA; OTTERSTEN, 2014). A modificação da função objetivo utiliza algoritmos de otimização para encontrar os melhores parâmetros para a função logística (sigmóide) que minimiza o custo total do modelo.

***Ensembles* que utilizam árvores de decisão sensíveis ao custo** utilizam uma combinação de várias árvores de decisão sensíveis ao custo dependentes de exemplo (ADSCDE) (BAHNSEN; AOUADA; OTTERSTEN, 2015a). Nestes *ensembles*, o conjunto de ADSCDE podem ser treinadas a partir das técnicas de *Bagging* e *Random Forest*. Na etapa de combinação, é utilizada a técnica de votação ponderada sensível ao custo. Essa técnica é uma extensão da técnica de votação ponderada (ZHOU, 2012) e consiste em considerar que os modelos de base que mais contribuem para a minimização de custos têm maior peso na etapa de votação.

O método ***Adaboost* sensível ao custo dependente de exemplo (ABSCDE)** considera o custo de cada exemplo na função que determina o erro dos classificadores treinados na iteração anterior (ZELENKOV, 2019), chamada função de perda. Essa função de perda define o peso a serem utilizados na próxima iteração da fase de treinamento. Essa função de perda também define a importância do classificador de base na etapa de combinação: o classificador que mais contribui para a minimização de custos também têm maior peso na etapa de votação.

O método ***Bayes Minimum Risk* (BMR)** é um método aplicado para tornar classificadores tradicionais em classificadores sensíveis ao custo (BAHNSEN et al., 2013). Após o treinamento de um classificador tradicional, ele utiliza a probabilidade estimada em cada predição para calcular o risco da predição para cada uma das classes, considerando os custos dos erros de classificação. Por fim, escolhe a classe que tem o menor risco estimado. Por exemplo, para o problema de classificação do PRDI, o método BMR pode ser definido através das Equações 7 and 8.

$$R(p_a|x) = C(p_a, y_a)P(p_a|x) + C(p_a, y_r)P(p_r|x) \tag{7}$$

$$R(p_r|x) = C(p_r, y_r)P(p_r|x) + C(p_r, y_a)P(p_a|x) \tag{8}$$

onde p_a, p_r são os resultados das predições para aceitar ou rejeitar um pedido, respectivamente; y_a, y_r são as classes reais de um determinado pedido, aceito ou rejeitado; $C(a,b)$ é o custo quando um determinado pedido tem sua predição como a e o resultado real é b; $P(p_a|x)$, $P(p_r|x)$ são as probabilidades estimadas pelo classificador para aceitar ou rejeitar um pedido, respectivamente. Cada PRDI vai ter sua predição final como aceito se $R(p_a|x) \leq R(p_r|x)$. Caso contrário, sua predição final será rejeitado.

2.3.3 Avaliação de modelos sensíveis ao custo

Métricas de performance tradicionais como acurácia, precisão, f-score ou sensibilidade assumem que os custos para os diferentes tipos de erros de classificação são iguais (HARRINGTON, 2012). Em relação aos problemas de classificação sensíveis ao custo dependentes de exemplo, os custos das predições entre dois classificadores com mesma quantidade de erros totais mas diferentes quantidades de falsos positivos e de falsos negativos não são os mesmos. É necessário contabilizar o erro para cada predição, de forma individualizada.

Uma das métricas utilizadas para essa problemática é chamada de *savings*. Essa métrica considera o custo de cada exemplo para comparar a performance entre dois classificadores, no que diz respeito ao custo (BAHNSEN et al., 2013). Essa métrica é definida da seguinte forma:

Seja Z um conjunto de N exemplos, em que cada exemplo i do conjunto de dados X é associado a seus respectivos custos, que podem ser representados da seguinte forma:

$$Z_i = [X_i,\ C_{VPi},\ C_{FPi},\ C_{FNi},\ C_{VNi}] \tag{9}$$

e um classificador f que cuja predição é o resultado da função $f(Z_i)$ para cada elemento i, então o valor absoluto do custo total ao usar f classificando Z, considerando y como a classe real, é definido da seguinte forma (BAHNSEN et al., 2013):

$$C(y, f(Z)) = \sum_{i=1}^{N} C(y_i, f(Z_i)) \tag{10}$$

O *savings* é definido como o custo total de usar um determinado classificador versus não utilizar um classificador, chamado de custo de base. O custo de base é o menor custo entre classificar todos os exemplos como positivos $(f(Z) = 1)$ ou negativos $(f(Z) = 0)$ e é definido na Equação 11:

$$C_{base} = min(C(y,1), C(y,0)) \tag{11}$$

O *savings* pode ser interpretado como a economia fornecida pelo uso do classificador que está sendo avaliado e é expressado na Equação 12. Ao final da avaliação, o melhor classificador consiste naquele que obtiver o valor de *savings* mais próximo de 1.

$$Savings(y, f(Z)) = \frac{C_{base} - C(y, f(Z))}{C_{base}} \tag{12}$$

Utilizando o problema de classificação do PRDI como exemplo, C_{base} é o custo de rejeitar todos os pedidos. Se a soma do custo de todos os erros de classificação for 0, então o *savings* será 1. Se a soma do custo de todos os erros de classificação for maior do que 0, então o *savings* será o percentual equivalente em relação ao custo de C_{base}.

2.4 Trabalhos Relacionados

Apresentam-se nesta seção alguns trabalhos relacionados, que foram identificados através de buscas na literatura que fazem uso de técnicas de classificação supervisionada para o domínio da administração tributária e os trabalhos que utilizam técnicas de classificação sensíveis ao custo.

O objetivo do trabalho exposto por González e Velásquez (2013) consiste na detecção de evasão fiscal empresarial do Chile. Para isso, os autores utilizaram um conjunto de dados com informações de 582.161 empresas. Dentro do conjunto de dados, havia 1.692 registros com informações sobre auditorias realizadas que resultaram, para alguns casos, na identificação do uso de notas fiscais falsas por determinada empresa, ou seja, aproximadamente 0,29% das empresas. Sendo assim, foi possível treinar algoritmos de aprendizado supervisionado para detectar fraudes no uso de notas fiscais por determinada empresa. Os resultados encontrados pelo trabalho foram utilizados pela receita federal chilena com o objetivo de priorizar quais empresas devem ser alvo de auditorias.

De forma semelhante ao trabalho anterior, o objetivo do trabalho apresentado por Rahimikia et al. (2017) consiste em detectar evasão fiscal por parte de empresas do Irã. Entretanto, ao invés de utilizar dados de notas fiscais, utilizou dados do sistema de imposto de renda anual. O conjunto de dados foi montado com informações de 7.477 empresas dos setores têxtil e de alimentos. Dentro do conjunto de dados, em torno de 5% das empresas possuíam um registro de não-conformidade no pagamento de suas obrigações tributárias. Este trabalho utilizou o algoritmo *harmony search* (GEEM; KIM; LOGANATHAN, 2001) para selecionar uma combinação dos melhores parâmetros dos algoritmos e quais as variáveis mais significativas para aumentar a performance dos algoritmos, concluindo que, para setores distintos, diferentes variáveis foram selecionadas, sugerindo que os setores não agem uniformemente no que diz respeito à forma de operar a evasão fiscal.

O objetivo do trabalho descrito por López, Rodríguez e Santos (2019) busca identificar

a probabilidade de que um determinado indivíduo possa cometer fraude no imposto de renda da Espanha. O trabalho utilizou um conjunto de dados com informações do preenchimento do formulário de imposto de renda de 2 milhões de contribuintes do ano de 2014 que foi disponibilizado pela Receita Federal da Espanha.

O objetivo do trabalho desenvolvido por Battiston, Gamba e Santoro (2020) consiste em identificar os contribuintes do tipo microempreendedores com maior risco de evasão fiscal na Itália. O trabalho utilizou um conjunto de dados disponibilizado pela Receita Federal Italiana, com informação de 5 anos referente a mais de 600 mil microempreendedores. Além da contribuição da construção do modelo preditivo como ferramenta de subsidiar as decisões da administração tributária italiana, o trabalho apresentou uma função para calcular os diferentes custos nas duas situações de erros de classificação: é mais custoso predizer que um contribuinte efetivamente fraudulento não cometeu fraude do que predizer que um contribuinte efetivamente honesto cometeu fraude. A função desenvolvida foi utilizada para priorizar os casos em que há chance de um maior retorno financeiro em prováveis ações de combate a fraudes realizadas pelo governo italiano.

O trabalho apresentado por Ippolito e Lozano (2020) avaliou a utilização de algoritmos de aprendizado supervisionado com a finalidade de predizer crimes fiscais no âmbito da cidade de São Paulo, Brasil. O trabalho utilizou um conjunto de dados com informações de apenas 217 ocorrências de fiscalizações realizadas em que algumas delas resultaram na identificação de crime fiscal, enquanto outras resultaram na identificação que não houve crime fiscal.

O trabalho exposto por Soares e Cunha (2020) buscou criar um mecanismo para auxiliar na identificação de contribuintes com maior risco de se tornar um devedor contumaz (que comete inadimplementos fiscais de foma recorrente) em uma capital brasileira. Para tal, utilizou um conjunto de dados contendo informações de 20.737 empresas, fornecidas pela secretaria municipal de finanças.

O trabalho demonstrado por Silva, Carvalho e Souza (2015), diferentemente dos trabalhos anteriores, não tem como objetivo a detecção de fraudes. O referido trabalho desenvolveu modelos de classificação de pedidos de compensação de crédito, solicitados em um serviço virtual disponibilizado pela Receita Federal do Brasil. Os pedidos podem ser classificados como "Deferido"ou "Indeferido". O objetivo do trabalho é auxiliar na construção de uma ferramenta que possa assistir as decisões dos pedidos, de forma a acelerar a análise dos mesmos. Para treinar os algoritmos, utilizou 18.000 instâncias de pedidos realizados anteriormente (representando apenas uma região do país), com seu respectivo resultado de deferimento ou indeferimento (aproximadamente 50% de representação para cada classe).

Lima et al. (2021) foi a primeira publicação a lidar com o problema de classificação do PRDI. Utilizou uma primeira versão do conjunto de dados do PRDI e apresentou a utilização de modelos tradicionais. Os algoritmos utilizados foram: redes neurais, *Random Forest*, *SVM* e *Naive Bayes*. O modelo com *Random Forest* alcançou os maiores níveis de acurácia (88%) e

sensibilidade (92%).

Comparando esse primeiro grupo de trabalhos relacionados com este, o aspecto mais importante de diferenciação em relação a esta dissertação diz respeito à falta de experimentação de métodos de classificação sensíveis ao custo.

No que diz respeito a trabalhos envolvendo modelos de classificação sensíveis ao custo, diferentes novos métodos foram propostos em alguns trabalhos nos últimos dez anos. No primeiro trabalho do pesquisador Alejandro Correa Bahnsen, ao utilizar inferência Bayesiana e cálculo do valor esperado (GHOSH; DELAMPADY; SAMANTA, 2006), propôs o método *Bayes Minimum Risk* (BMR) e demonstrou, além da importância de realizar o cálculo do ganho ou da perda financeira com os métodos de classificação, que o BMR apresenta maior efetividade financeira em aplicações de detecção de fraude em cartão de crédito (BAHNSEN et al., 2013).

Outros métodos de classificação sensíveis ao custo dependentes de exemplo foram propostos por Bahnsen, ao alterar algoritmos tradicionais de aprendizado supervisionado. As propostas foram a modificação dos algoritmos de regressão logística (BAHNSEN; AOUADA; OTTERSTEN, 2014), árvore de decisão (BAHNSEN; AOUADA; OTTERSTEN, 2015b) e os algoritmos de *bagging* e *Random Forest* (BAHNSEN; AOUADA; OTTERSTEN, 2015a). Por fim, outros autores propuseram alterações em métodos de *boosting*, no sentido de transformá-los em sensíveis ao custo dependentes de exemplo (HÖPPNER et al., 2022; ZELENKOV, 2019).

De forma geral, os referidos trabalhos realizam comparações entre os novos métodos propostos e respectivas versões insensíveis ao custo. Os mais recentes incluem comparações com alguns métodos sensíveis ao custo já publicados anteriormente. Todos os trabalhos referenciados utilizam conjuntos de dados da área financeira: bancos e cartões de crédito. As aplicações são, em sua maioria, para detecção de fraudes, definição do *score* de crédito e análise de marketing.

O trabalho de Wang e Tiong (2022) apresenta uma proposta de um novo método sensível ao custo dependente de exemplo para ser aplicado na predição de encerramentos precoces em contratos de parcerias entre as áreas pública e privada. Definiu-se uma nova função de perda para 3 diferentes algoritmos: redes neurais, light gradient boosting (KE et al., 2017) e extreme gradient boosting (CHEN et al., 2015). Este trabalho comparou os novos métodos de classificação com métodos tradicionais e concluiu que os métodos sensíveis ao custo eram mais adequados apenas em contextos de negócio onde se prioriza a minimização dos custos. Para outros casos, os métodos tradicionais eram considerados mais adequados devido à menor quantidade de erros de classificação ocasionados.

O trabalho de Mehta et al. (2020), ao buscar classificar casos de evasão fiscal, demonstrou que o uso de métodos de classificação sensíveis ao custo obtém melhores resultados em termos de *savings* em comparação aos métodos tradicionais, para um problema de classificação da administração tributária. O referido trabalho usou um conjunto de dados do departamento da receita federal do governo de Telangana, na Índia. Em geral, o trabalho mostrou que, ao atingir

maiores índices em relação ao *savings*, maior era a probabilidade em diminuir os índices de acurácia e *f-score*.

A Tabela 4 mostra um resumo comparativo dos trabalhos relacionados, focando na diferenciação em relação ao uso de métodos sensíveis ao custo dependentes de exemplo, qual o domínio de aplicação dos dados, se houve análise comparativa entre diversos algoritmos de classificação e, por fim, se fez uso de uma abordagem de classificação híbrida (que permite a integração de modelos de classificação tradicionais e sensíveis ao custo).

Trabalhos relacionados	**Utilização de métodos sensíveis ao custo?**	**Domínio dos dados**	**Comparação entre algoritmos?**	**Utiliza abordagem híbrida?**
Rahimikia et al. (2017); González e Velásquez (2013); Soares e Cunha (2020); Ippolito e Lozano (2020); López, Rodríguez e Santos (2019); Battiston, Gamba e Santoro (2020); Silva, Carvalho e Souza (2015); Lima et al. (2021);	**Não**	Adm. tributária	Sim	**Não**
Bahnsen et al. (2013) Bahnsen, Aouada e Ottersten (2014); Bahnsen, Aouada e Ottersten (2015b); Bahnsen, Aouada e Ottersten (2015a); Höppner et al. (2022); Zelenkov (2019);	Sim	Cartões de crédito	Sim	**Não**
Wang e Tiong (2022)	Sim	Contratos Públicos	**Não**	**Não**
Mehta et al. (2020)	Sim	Adm. tributária	**Não**	**Não**
Nossa proposta	Sim	Adm. tributária	Sim	Sim

Tabela 4 – Comparativo entre os trabalhos relacionados.

O primeiro grupo refere-se aos trabalhos que lidam com problemas de classificação para a administração tributária (SOARES; CUNHA, 2020; IPPOLITO; LOZANO, 2020; LÓPEZ; RODRÍGUEZ; SANTOS, 2019; BATTISTON; GAMBA; SANTORO, 2020). O mais importante diferencial em relação a esta dissertação é que nenhum deles utilizou métodos de classificação sensíveis ao custo.

O segundo grupo refere-se aos trabalhos que propõem novos métodos sensíveis ao custo (BAHNSEN et al., 2013; BAHNSEN; AOUADA; OTTERSTEN, 2014; BAHNSEN; AOUADA; OTTERSTEN, 2015b; BAHNSEN; AOUADA; OTTERSTEN, 2015a; HÖPPNER et al., 2022; ZELENKOV, 2019). O principal diferencial em relação a esta dissertação é que os referidos trabalhos não lidam com problemas de classificação no âmbito da administração tributária. Ao lidar com problemas de classificação de empresas privadas, há uma maior possibilidade da priorização de custos. Por exemplo, um método de classificação sensível ao custo pode predizer que determinada transação em um cartão de crédito, com valor muito pequeno, não deve ser considerada fraudulenta, ao considerar que existe um custo operacional para verificar uma fraude. Esse custo é relacionado ao tempo utilizado por analistas no sentido de avaliar os dados da transação, entrar em contato com o cliente até concluir todo o ciclo de tratamento de fraude. Portanto, não há nenhuma discussão levantada a respeito de uma maior quantidade de erros de classificação e, portanto, não utilizam nenhuma abordagem híbrida.

O trabalho apresentado por Wang e Tiong (2022) realizou observações em relação ao *trade-off* entre *savings* e métricas de avaliação tradicionais, e trabalhou com problemas de classificação que envolvia órgãos públicos. Porém, a cada experimento, ou utilizava modelos tradicionais ou sensíveis ao custo e não utilizou uma abordagem híbrida.

Por fim, o trabalho de Mehta et al. (2020) lidou com um problema de classificação na visão da administração tributária, utilizando um método sensível ao custo. Porém, não realizou comparativos em relação às versões tradicionais dos modelos e não utilizou uma abordagem híbrida.

Assim, não foram identificados trabalhos que buscassem um maior equilíbrio entre bons resultados quando se considera tanto a métrica de *savings* quanto as de acurácia e *f-score*. Esta dissertação apresenta uma proposta de abordagem híbrida de classificação. Nesse panorama, ela tece uma análise sobre o trade-off entre as métricas de avaliação comumente usadas e o savings, em busca de ajudar a identificar a opção ideal ao lidar com o problema de classificação do PRDI. A abordagem proposta é apresentada no próximo capítulo, seguida da apresentação dos experimentos com modelos tradicionais, modelos sensíveis ao custo e da abordagem híbrida.

3 A ABORDAGEM R3D

Com o objetivo de lidar com problemas de classificação em que há diferentes custos associados aos diversos tipos de erros de classificação (e também, eventualmente, dos acertos), métodos de classificação sensíveis ao custo dependentes de exemplo vem sendo propostos, tendo alguns deles sido detalhados no Capítulo 2 (SCOTT, 2011; BAHNSEN et al., 2013; BAHNSEN; AOUADA; OTTERSTEN, 2014; BAHNSEN; AOUADA; OTTERSTEN, 2015b; BAHNSEN; AOUADA; OTTERSTEN, 2015a; ZELENKOV, 2019). Estes trabalhos mostram que, em relação à métrica de avaliação *savings*, os métodos propostos apresentam melhores resultados do que métodos de classificação tradicionais e, inclusive, em relação aos métodos sensíveis ao custos dependentes apenas da classe (ZELENKOV, 2019). Entretanto, a maioria dos referidos trabalhos mostram também que a quantidade de erros de classificação aumenta quando se compara aos métodos de classificação tradicionais (WU et al., 2012).

No que diz respeito a órgãos públicos, com sua característica de não ter o lucro como uma prioridade, alguns cortes de custos, que podem impactar a execução da atividade-fim ou esbarram em restrições legais, nem sempre são adequados (WANG; TIONG, 2022). Considerando o uso de métodos sensíveis ao custo neste tipo de contexto de negócio, um alto número de falsos positivos e/ou negativos não são uma consequência aceitável.

Para a administração tributária, ter processos eficientes e de baixo custo é muito importante (WANG; TIONG, 2022). Porém, a administração tributária desempenha um papel de regulação do mercado, através da busca pela conformidade tributária de empresas e pessoas. Dentro desse contexto, é também importante ter baixos índices de falsos positivos e/ou negativos. Especificamente para o problema de classificação do PRDI, deve-se buscar ser justo na análise dos pedidos, independentemente do valor da dívida.

A proposta da presente abordagem, chamada de R3D (Request for Revision of Registered Debts), busca possibilitar um maior equilíbrio entre uma menor quantidade de erros de classificação e um menor impacto em relação a custos. Esse equilíbrio é medido através da métrica baseada em custos (*savings*) e das métricas tradicionais (*f-score*, acurácia, sensibilidade e precisão). A abordagem R3D visa propor uma solução para a segunda questão de pesquisa replicada a seguir:

QP2 Como integrar modelos de aprendizado de máquina sensíveis ao custo e modelos tradicionais de forma a manter os níveis de economia (considerando a métrica sensível ao custo *savings*) atingidos pelos primeiros e mantendo uma menor quantidade de erros atingidos pelos segundos no âmbito do serviço do PRDI?

3.1 O problema de classificação do PRDI

A abordagem R3D foi concebida para ser aplicada ao problema de classificação do PRDI. Os modelos de classificação devem predizer o resultado da análise de cada pedido, podendo ser "deferido/aceito"ou "indeferido/rejeitado". A classe positiva é o resultado "indeferido", enquanto a classe negativa é o resultado "deferido".

O conjunto de dados utilizado durante esta pesquisa foi disponibilizado pela PGFN e foi criado por uma equipe composta por especialistas de domínio e cientistas de dados. Foram coletados dados de diversas fontes da PGFN, incluindo sistemas analíticos e transacionais. Os registros incluídos no conjunto de dados representam os registros de PRDI realizados entre novembro de 2018 e março de 2022. Os dados disponibilizados foram previamente tratados, de forma que não foram identificados dados faltantes nem outliers.

O conjunto de dados possui 23 variáveis independentes e um total de 173.709 instâncias, abrangendo dados de todas as regiões brasileiras. Todas as informações de identificação pessoal ou corporativa (e qualquer outra variável considerada sensível) foram descartadas. Existe uma variável que indica o resultado do PRDI, com dois valores possíveis: deferido ou indeferido. Em torno de 40% dos registros são de deferimento e 60% de indeferimento.

Por questões de segurança e confidencialidade, a PGFN não autoriza a publicação de detalhes das variáveis utilizadas. Entretanto, é possível citar algumas informações que estão contidas nos atributos, conforme a seguir:

- Informações sobre o registro do PRDI, como por exemplo sua data e hora de abertura;
- Informações sobre o contribuinte;
- Informações sobre o débito: valor, idade, tipo e situação;
- Informações sobre o histórico de relacionamento do contribuinte com a PGFN;
- Informações sobre alguns relacionamentos do contribuinte com outras entidades, pessoa física ou jurídica;
- Uma variável que indica se o PRDI foi deferido ou indeferido.

A abordagem R3D também utiliza modelos de classificação sensíveis ao custo. Para o PRDI, o custo dos erros de classificação tem uma relação com o valor da dívida. Portanto, a Tabela 5 apresenta mais detalhes quanto às características da distribuição dos valores das dívidas no conjunto de dados.

Identifica-se que os valores das dívidas apresentam distribuição semelhante a uma distribuição exponencial. Muitos valores abaixo da média e poucos valores que acabam impactando no cálculo da média. Por exemplo, 69% das instâncias têm valor abaixo de R$ 50.000,00. Porém, se

Quantidade	173.709
Mediana	R$ 12.778,58
Desvio padrão	R$ 16.325.820,00
Primeiro quartil até	R$ 3.488,18
Segundo quartil até	R$ 12.778,58
Terceiro quartil até	R$ 64.158,57
Valor máximo	R$ 2.690.038.000,00
Soma de todos os valores	R$ 127.756.051.982,00

Tabela 5 – Principais estatísticas a respeito da distribuição dos valores das dívidas envolvidas nos pedidos do PRDI.

somadas todas essas instâncias, elas resultariam em apenas 1,08% da soma das dívidas envolvidas em todos os pedidos do conjunto de dados. Por outro lado, a soma dos outros 31% maiores valores resultam em 98,92% dos valores de todo o conjunto de dados.

3.1.1 A matriz de custo do PRDI

Para o problema do PRDI, a matriz de custo, apresentada na Tabela 6, foi definida em conjunto com um especialista de domínio e de acordo com as regras do negócio.

Real / Predição	**Rejeitado**	**Aceito**
Rejeitado	$C_{VPi} = 0$	$C_{FNi} =$ valor da dívida
Aceito	$C_{FPi} = R\$50.000,00$	$C_{VNi} = 0$

Tabela 6 – Matriz de custo para o problema de classificação do PRDI.

A classe positiva significa o indeferimento do pedido, ou sua rejeição. A situação de verdadeiro positivo não tem custos associados, sendo definido como zero. O falso positivo significa que um pedido tem sua predição como rejeitado, porém sua classe real é aceito. Para o contexto de negócio, significa que a dívida será erroneamente ratificada, o que leva à continuidade de cobrança indevida por parte da administração tributária. Para definir o custo do falso positivo, fez-se necessário a definição do custo de um processo de cobrança de dívidas tributárias.

Os processos de cobrança, dentro da PGFN, incluem custos relacionados à utilização de infraestrutura e recursos humanos para cobrança e gerenciamento de cada dívida. No trabalho de Cunha, Klin e Pessoa (2011), reuniu-se dados de forma a rastrear e estimar o custo médio de um processo de cobrança tributária no âmbito da PGFN. Por ser tratar de um trabalho realizado em 2011, o valor médio resultante deste estudo foi atualizado de acordo com a inflação, sendo definido em valores atuais por R$ 50.000,00. Este valor também foi validado pelo especialista de negócio. É importante salientar que esse valor é o mesmo para todo processo, independente do valor da dívida. Portanto, passa a ser um valor constante na matriz de custo, na situação do falso positivo.

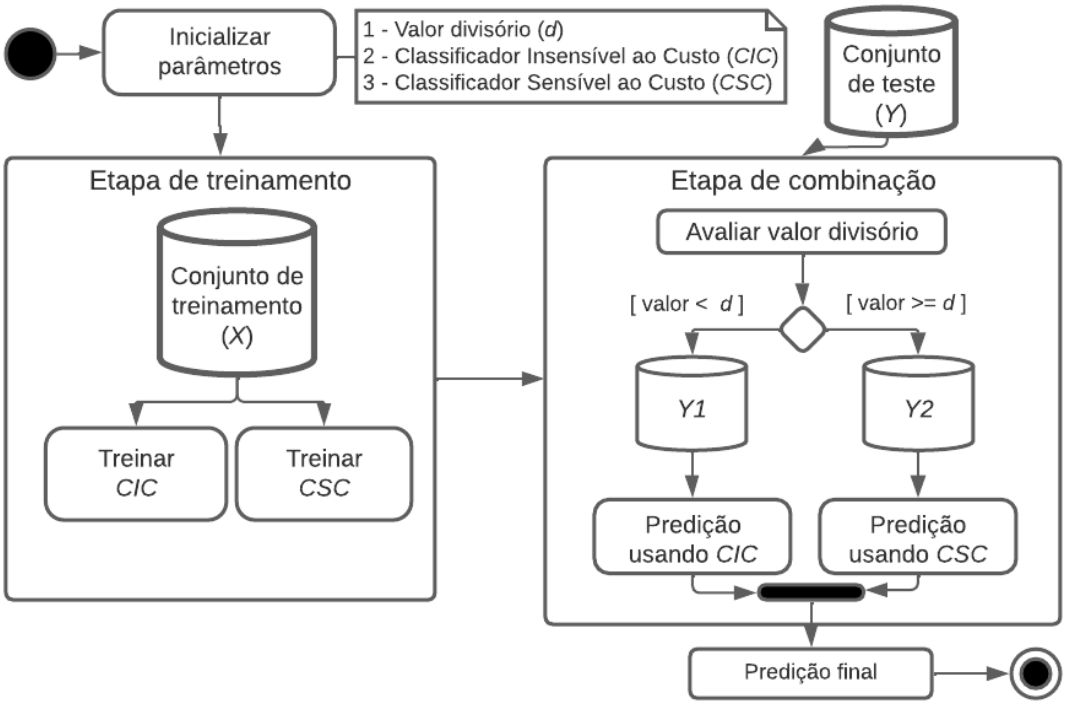
A classe negativa significa o deferimento do pedido, ou o seu aceite. A situação de verdadeiro negativo também não tem custos associados, sendo definido como zero. O falso

negativo significa que um pedido tem sua predição como aceito, porém sua classe real é rejeitado. Para o contexto do negócio, significa que a dívida será erroneamente perdoada e o processo de cobrança encerrado. Essa situação supostamente resulta na perda da arrecadação do valor da dívida, caso o processo de cobrança seja exitoso. Portanto, o falso negativo é totalmente dependente do valor da dívida e foi definido como seu valor exato.

3.2 Definições da abordagem R3D

A abordagem R3D consiste em 3 etapas, conforme mostradas na Figura 6 e explicadas a seguir: (i) Inicialização de Parâmetros; (ii) Etapa de treinamento; e (iii) Etapa de combinação.

Figura 6 – Visão geral da abordagem R3D.



Os parâmetros são definidos na primeira etapa: o valor divisório *d*, o classificador insensível ao custo *CIC* e o classificador sensível ao custo *CSC*. Ambos são treinados na segunda etapa. Na etapa de combinação, o conjunto de teste é dividido de acordo com o valor divisório, gerando subconjuntos *Y1* e *Y2* que terão suas predições realizadas por *CIC* e *CSC* respectivamente.

Na primeira etapa, a abordagem R3D foi construída de forma flexível por permitir a definição de 3 parâmetros por usuários da abordagem: um valor divisório (*d*), um classificador insensível ao custo (*CIC*) e um classificador sensível ao custo (*CSC*).

O valor divisório (*d*) desempenha o papel de separar, dentro do conjunto de dados de teste, os pedidos com valores menores de dívidas dos pedidos com valores maiores de dívidas de forma a criar dois subconjuntos de teste. Essa separação permite que dois classificadores

diferentes possam ser responsáveis pelas predições em dois subconjuntos de teste diferentes. Conforme apresentado na Tabela 5, o valor das dívidas apresenta uma distribuição similar a uma distribuição exponencial. Isso significa que o subconjunto *Y1* possui mais instâncias, mas com dívidas de valores menores. Da mesma forma, o subconjunto *Y2* possui menos instâncias, porém com dívidas de valores maiores, podendo em alguns casos ser bem grandes (bastante acima da média).

Os outros parâmetros a serem definidos dentro da abordagem são os classificadores sensíveis (*CSC*) e insensíveis (*CIC*) ao custo. A abordagem é considerado híbrida porque os dois classificadores são usados intercambiavelmente na etapa de combinação. O *CIC* é usado para as predições de pedidos com valor da dívida menor que *d* e o *CSC* é usado para as predições de pedidos com valor da dívida maior que *d*.

Na etapa de treinamento, todo o conjunto de treino é usado duas vezes: para treinar o *CIC* e o *CSC*. Por fim, na etapa de combinação, cada instância do conjunto de dados de teste é avaliada em relação ao valor da dívida que o pedido em questão possui. Se o valor da dívida for menor que *d*, então a instância é inserida em *Y1* e o *CIC* realizará as predições deste subconjunto. Se o valor da dívida for maior ou igual a *d*, então a instância é inserida em *Y2* e o *CSC* realizará as predições deste subconjunto. Todas as predições são inseridas no conjunto final das predições resultantes da abordagem R3D.

O objetivo principal da abordagem R3D é minimizar a quantidade de erros de classificação que os modelos sensíveis ao custo apresentam mas, ao mesmo tempo, se beneficiar de sua caracterísitica sensível ao custo. Neste sentido, a intenção é manter os pedidos com valores maiores sendo tratados pelo classificador sensível ao custo, porque os valores envolvidos em *Y2* causam um maior impacto em *savings*, caso algum erro de classificação ocorra. Por conter menos instâncias, os erros de classificação causados pelo *CSC* podem causar menor impacto nas métricas tradicionais como *f-score*, acurácia, sensibilidade e precisão.

Por outro lado, ao utilizar um classificador tradicional para o subconjunto *Y1*, haveria menos impacto em relação ao *savings*, pois os pedidos deste subconjunto envolvem valores de dívidas que não são tão significativas. Porém, como *Y1* possui muitos valores, é importante ter um índice de acertos nas predições, o que pode aumentar *f-score* e acurácia.

No sentido de verificar se essas hipóteses se confirmam, alguns experimentos foram realizados, conforme descritos no próximo capítulo.

4 EXPERIMENTOS E RESULTADOS

Neste capítulo são apresentados aspectos referentes aos resultados obtidos neste trabalho. Para isso, foram realizados experimentos visando avaliar as soluções implementadas e responder às principais questões de pesquisa (apresentadas no Capítulo 1). O presente capítulo inclui então os objetivos esperados para cada cenário experimental, a configuração utilizada para a execução dos experimentos (modelos selecionados e métricas avaliadas), resultados e respectivas análises e considerações dos experimentos de cada cenário.

4.1 Modelos sensíveis ao custo vs Modelos tradicionais

Objetivo

O objetivo deste experimento foi avaliar os resultados obtidos por modelos de classificação tradicionais em comparação com os modelos de classificação sensíveis ao custo quando aplicados ao problema de classificação do PRDI. Esse objetivo busca responder à primeira questão de pesquisa, replicada a seguir.

QP1 O uso de modelos sensíveis ao custo obterá resultados melhores do que os modelos tradicionais no que diz respeito ao custo envolvido nos erros de classificação quando aplicados ao problema de classificação do PRDI?

Configuração

O problema de classificação do PRDI consiste em um problema com custos de classificação dependentes de cada exemplo/instância e não apenas da classe. Essa característica se dá pela definição do custo do falso negativo, que é totalmente dependente da dívida envolvida no pedido em questão. Portanto, foram selecionados para o cenário deste experimento alguns métodos de classificação sensíveis ao custo dependentes de exemplo.

Por terem sido identificados na literatura e estarem disponíveis para utilização, os seguintes métodos de classificação sensíveis ao custo dependentes de exemplo foram selecionados: árvore de decisão (ADSCDE), regressão logística (RLSCDE), *bagging* (BagSCDE), *Random Forest* (RFSCDE) e Adaboost (ABSCDE). Os respectivos modelos de classificação tradicionais correspondentes a cada modelo de classificação sensível ao custo também foram incluídos, instanciados com os parâmetros padrão da biblioteca *SciKit-Learn*[1]. Por fim, o método BMR foi aplicado para cada método de classificação tradicional, em um total de 15 modelos avaliados. As implementações dos métodos ADSCDE, RLSCED, BagSCDE, RFSCDE e BMR foram obtidos

[1] https://scikit-learn.org/

na página da biblioteca COSTCLA [2]. A implementação do ABSCDE foi obtida do repositório do autor da proposta deste método [3].

O cenário de execução foi realizado utilizando 10 execuções de validação cruzada estratificada com 10 partições. Para a avaliação dos experimentos, *savings* é a métrica prioritária. São incluídas as métricas de *f-score*, acurácia, sensibilidade e precisão no sentido de observar o comportamento em relação à quantidade de erros de classificação em ambas as classes. A Tabela 7 apresenta os resultados obtidos nos experimentos.

Classifier	*Savings*	*F-score*	*Acurácia*	*Sensibilidade*	*Precisão*
AD	-113,05% (±140,29%)	78,41% (±0,36%)	82,74% (±0,28%)	78,48% (±0,54%)	78,35% (±0,42%)
AD-BMR	72,29% (±0,59%)	40,46% (±0,66%)	62,16% (±0,38%)	32,18% (±0,61%)	54,46% (±0,80%)
ADSCDE	61,81% (±0,52%)	19,77% (±0,50%)	53,09% (±0,29%)	14,47% (±0,41%)	31,19% (±0,70%)
RL	-548,52% (±174,96%)	41,09% (±16,91%)	65,84% (±2,72%)	32,88% (±14,12%)	59,89% (±12,84%)
RL-BMR	61,60% (±0,68%)	23,43% (±0,91%)	53,07% (±0,67%)	17,97% (±0,70%)	33,646% (±1,41%)
RLSCDE	60,56% (±0,48%)	19,09% (±0,61%)	52,27% (±0,37%)	14,11% (±0,57%)	29,56% (±0,66%)
Bag	-121,23% (±136,31%)	**81,41%** (±0,35%)	85,75% (±0,25%)	78,16% (±0,56%)	84,95% (±0,43%)
Bag-BMR	80,53% (±0,60%)	58,46% (±0,80%)	72,44% (±0,44%)	48,56% (±0,88%)	73,44% (±0,76%)
BagSCDE	61,82% (±0,52%)	19,47% (±0,53%)	53,23% (±0,34%)	14,16% (±0,45%)	31,18% (±0,74%)
AB	-277,89% (±175,87%)	66,67% (±0,41%)	75,04% (±0,30%)	62,52% (±0,52%)	71,42% (±0,49%)
AB-BMR	70,38% (±0,70%)	36,39% (±0,90%)	61,24% (±0,44%)	27,76% (±0,81%)	52,81% (±1,04%)
ABSCDE	-26,70% (±33,60%)	65,79% (±0,19%)	61,15% (±0,27%)	**94,26%** (±0,24%)	50,74% (±0,18%)
RF	-132,09% (±147,85%)	81,36% (±0,40%)	**85,82%** (±0,28%)	77,51% (±0,58%)	**85,62%** (±0,43%)
RF-BMR	**80,75%** (±0,59%)	58,80% (±0,80%)	72,69% (±0,44%)	48,80% (±0,88%)	73,96% (±0,75%)
RFSCDE	19,17% (±24,05%)	47,77% (±15,23%)	44,54% (±4,68%)	74,45% (±32,99%)	37,98% (±4,39%)

Tabela 7 – Média e desvio-padrão de validação cruzada 10 x 10-fatias.

[2] http://albahnsen.github.io/CostSensitiveClassification/index.html
[3] https://github.com/yzelenkov/Cost-sensitive-AdaBoost

Análise dos resultados

Todos os classificadores tradicionais apresentam resultado negativo de *savings*. Um valor negativo significa que o classificador avaliado obtém índices de *savings* piores do que um classificador que rejeita todos os pedidos.

Os classificadores tradicionais também apresentam um alto desvio-padrão em *savings*, demonstrando instabilidade quando esta métrica de custo é considerada. Este resultado é um índicio de que ocorre um ou mais falsos negativos envolvendo valores altos de dívidas. Entretanto, como os altos valores são raros no conjunto de dados, nem sempre podem ser selecionados para o conjunto de teste. Se não forem selecionados ou não houver falso negativo para um valor alto, o *savings* alcança um bom resultado positivo. Caso contrário, um falso negativo com alto valor de dívida causa um grande impacto no *savings*. Sendo assim, falso negativos para altos valores não são situações aceitáveis do ponto de vista deste contexto de negócio.

Em relação às versões sensíveis ao custo dos classificadores utilizados, quase todos apresentam um valor positivo para *savings*. Fica evidente que os classificadores sensíveis ao custo, de fato, apresentam um desempenho superior em relação aos classificadores tradicionais no que diz respeito ao custo. Isso significa que os classificadores sensíveis ao custo podem contribuir significativamente ao lidar com decisões de problemas reais envolvendo custos diferentes para os diferentes tipos de erros de classificação.

Os classificadores ADSCDE, RLSCDE e BagSCDE apresentam resultados muito próximos em todas as cinco métricas. Como BagSCDE é um *ensemble* composto de várias ADSCDE, então a baixa variância apresentada pela ADSCDE provavelmente causou votos similares na etapa de combinação do *ensemble*. As performances similares apresentadas por ADSCDE e RLSCDE também foram reportadas em outro trabalho relacionado (BAHNSEN; AOUADA; OTTERSTEN, 2015b).

O classificador ABSCDE, apesar de não atingir um resultado positivo em *savings*, obteve um resultado melhor que a versão tradicional AB. ABSCDE é um *ensemble* que usa a versão tradicional de árvore de decisão como classificador base, enquanto RFSCDE e BagSCDE utilizam a ADSCDE. São propostas diferentes, realizadas por autores diferentes. Essa diferença pode explicar o pior resultado do ABSCDE quando comparado aos demais métodos sensíveis ao custo.

O comportamento do RFSCDE precisa de um estudo mais aprofundado. É possível que a seleção randômica das variáveis de treinamento produzam árvores de decisão significativamente diferentes e, por isso, apresenta um desvio-padrão em *savings* e *f-score* mais altos que os demais modelos baseados em árvore de decisão sensível ao custo.

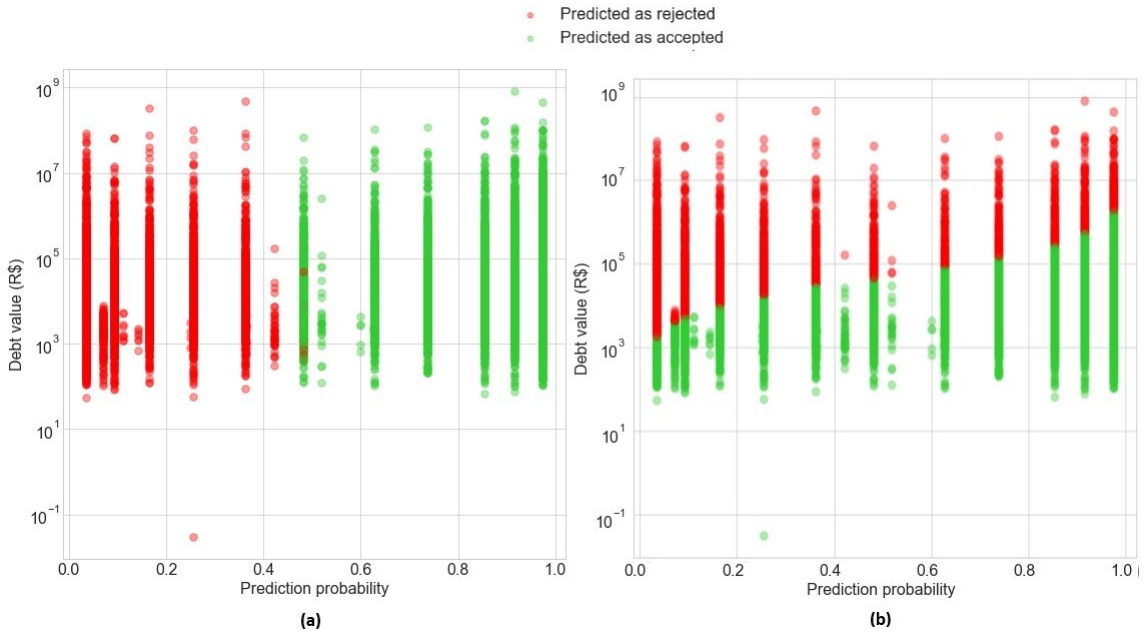
Em relação ao BMR, percebe-se que ele atinge os melhores índices de *savings* tanto em relação ao modelo tradicional quanto aos modelos sensíveis ao custo respectivos. Também observa-se um baixo desvio-padrão, o que evidencia estabilidade quanto aos custos. O classificador *Random Forest* após a aplicação do BMR apresentou o melhor resultado em relação ao

savings, atingindo 80,75%.

Verifica-se no método BMR que, em alguns casos, há uma mudança na predição de aceito para rejeitado ou vice-versa. A Figura 7(a) apresenta um gráfico do *Random Forest* antes e depois da aplicação do BMR. As probabilidades de cada predição são representandas pelo eixo X e os valores das dívidas pelo eixo Y. A Figura 7(a) mostra que, se a probabilidade de aceitação é menor do que 50%, então a predição resulta em rejeição do pedido (pontos vermelhos). Caso contrário, a predição aponta aceitação (pontos verdes).

A Figura 7(b) apresenta o resultado das predições quando o BMR é aplicado. Percebe-se que, à medida que o valor da dívida vai aumentando, o cálculo do risco tende a rejeitar alguns pedidos com valores altos, mesmo se a probabilidade de aceitação é maior ou igual a 50%. Por outro lado, o cálculo do risco tende a apontar uma aceitação para pedidos com dívidas de baixo valor, mesmo se a probabilidade de aceitação menor do que 50%.

Figura 7 – Probabilidade da predição vs Valor da dívida antes (a) e depois (b) da utilização do método BMR associado ao classificador Random Forest.



Fonte: autoria própria

Conclui-se, portanto, que os modelos de classificação tradicionais, quando aplicados ao problema de classificação do PRDI, apesar de apresentarem uma menor quantidade de erros de classificação, correm o risco de cometerem alguns erros com alto custo, considerados inaceitáveis para este contexto de negócio na qual foram aplicados. Estes erros são, principalmente, devido à possibilidade da perda de arrecadação de milhões e até bilhões de reais em dívidas registradas por alguns pedidos.

Os modelos de classificação sensíveis ao custo demonstraram que não cometem erros de classificação com alto custo e podem ser considerados uma alternativa para o problema de classificação do PRDI. Porém, apesar de alcançar um melhor resultado em relação ao *savings*,

os modelos sensíveis ao custo apresentam piores resultados em relação a *f-score*, acurácia, sensibilidade e precisão. Esse comportamento (*trade-off*) também foi reportado em outros trabalhos relacionados (BAHNSEN; AOUADA; OTTERSTEN, 2015a; ZELENKOV, 2019).

Apresentar uma quantidade alta de erros de classificação, mesmo obtendo melhores resultados no que diz respeito à métrica *savings* também não é um comportamento desejado do ponto de vista do contexto de negócio na qual está inserido o PRDI.

De forma geral, os classificadores sensíveis ao custo que obtiveram os melhores resultados em *savings* (acima de 50%), apresentaram um impacto de redução maior na sensibilidade do que na precisão quando comparados com a respectiva versão tradicional. Em outras palavras, nesta comparação, o impacto é maior no aumento de falsos negativos em relação ao aumento de falsos positivos.

4.2 Abordagem R3D

Dentro do contexto do PRDI, o falso negativo significa um pedido erroneamente aceito. Por priorizar a minimização de custos, os modelos sensíveis ao custo estão realizando as predições com a preferência pela aceitação de pedidos. Conforme análise específica do método BMR, esses falsos negativos se dão em grande parte para pedidos com baixo valor (abaixo de 100 mil reais). De acordo com o cálculo realizado pelo método BMR, conclui-se que esse comportamento acontece porque o cálculo do risco do falso negativo é menor do que o risco do falso positivo (que tem valor fixo de R$ 50.000,00). Entretanto, não há embasamento jurídico que possa validar uma maior propensão na aceitação de pedidos apenas por envolver valores próximos ou abaixo do custo médio de um processo de cobrança tributária.

A abordagem R3D proposta foi concebida com o objetivo de possibilitar um maior equilíbrio entre uma menor quantidade de erros de classificação e um menor impacto em relação a custos para o problema de classificação do PRDI. Os experimentos realizados para avaliar a abordagem R3D estão descritos nesta seção.

Os experimentos foram divididos em três cenários, cujos objetivos são mostrados a seguir:

1. Cenário 1: Avaliar o comportamento da abordagem R3D quando alguns valores são definidos para o valor divisório *d*, com objetivo de identificar o valor mais apropriado para *d* de acordo com as necessidades de negócio.

2. Cenário 2: Avaliar as diferenças de comportamento entre os classificadores sensíveis e insensíveis ao custo em relação aos diferentes subconjuntos de teste *Y1* e *Y2* no contexto da R3D.

3. Cenário 3: Avaliar se a abordagem R3D, definida como híbrida, apresenta um maior equilíbrio entre os classificadores *baseline* definidos, com o objetivo de verificar sua aplicabilidade no contexto real de negócio.

4.2.1 *Baselines* considerados

Como primeiro *baseline* foram considerados os resultados do primeiro trabalho abordando o problema de classificação do PRDI, conforme resumidamente descrito na Seção 2.4 (LIMA et al., 2021). No referido trabalho, apenas métodos de classificação tradicionais foram utilizados (Redes Neurais, *Naive Bayes*, *Random Forest* e SVM), e os melhores índices foram alcançados pelo *Random Forest*, com os seguintes resultados: 88% de acurácia e 92% de sensibilidade.

O bom desempenho alcançando pelo *Random Forest* foi novamente observado na seção anterior, na qual os seguintes métodos de classificação tradicionais foram utilizados: Árvore de decisão, Regressão Logística, *Bagging*, *Random Forest* e *Adaboost*. Considerando a acurácia, o *Random Forest* atingiu novamente o melhor índice: 85,82%.

Como segundo *baseline*, foram considerados os resultados dos experimentos que utilizaram métodos sensíveis ao custo aplicados ao problema de classificação do PRDI. Considerando o *savings*, o método BMR aplicado ao *Random Forest* atingiu o melhor índice: 80,75%.

Portanto, no âmbito da avaliação da abordagem R3D, o *Random Forest* (RF) é considerado como primeiro *baseline*, em relação à acurácia, *f-score*, precisão e sensibilidade. O método BMR aplicado ao *Random Forest* (RF-BMR) é considerado o segundo *baseline*, em relação a *savings*.

4.2.2 Cenário 01

A abordagem R3D é baseada na utilização de um valor divisório d. Neste sentido, com a finalidade de avaliar o comportamento da abordagem em relação à definição de diversos valores divisórios diferentes, foram utilizados oito valores para d, de R$ 2.500,00 até R$ 400.000,00, representando todos os quatro quartis.

Configuração

Para completar a definição dos parâmetros exigidos pela abordagem R3D, definiu-se o classificador *Random Forest* como *CIC* e o método BMR aplicado ao *Random Forest* como *CSC* pois, quando aplicados ao problema de classificação do PRDI, foram considerados os melhores classificadores tradicional e sensível ao custo, respectivamente. Os resultados estão apresentados na Tabela 8.

Valor divisório/Medida	***Savings***	***Acurácia***	***F-score***	***Sensibilidade***	***Precisão***
R$ 2.500,00	78,84% (±0,63%)	78,60% (±0,38%)	71,45% (±0,56%)	67,03% (±0,68%)	76,48% (±0,53%)
R$ 5.000,00	77,84% (±0,54%)	80,66% (±0,40%)	75,38% (±0,53%)	74,14% (±0,64%)	76,67% (±0,53%)
R$ 10.000,00	77,30% (±0,62%)	81,81% (±0,37%)	77,51% (±0,47%)	78,48% (±0,59%)	76,57% (±0,47%)
R$ 25.000,00	77,09% (±0,66%)	82,27% (±0,37%)	78,42% (±0,44%)	80,66% (±0,50%)	76,31% (±0,50%)
R$ 50.000,00	77,05% (±0,69%)	82,30% (±0,37%)	78,48% (±0,45%)	80,79% (±0,57%)	76,30% (±0,48%)
R$ 100.000,00	76,92% (±0,58%)	82,32% (±0,35%)	78,39% (±0,44%)	80,26% (±0,55%)	76,60% (±0,45%)
R$ 200.000,00	76,35% (±0,63%)	82,44% (±0,36%)	78,37% (±0,45%)	79,62% (±0,56%)	77,16% (±0,48%)
R$ 400.000,00	74,85% (±0,66%)	82,73% (±0,35%)	78,53% (±0,43%)	79,09% (±0,55%)	77,98% (±0,48%)

Tabela 8 – Utilizando diferentes valores divisórios na abordagem R3D.

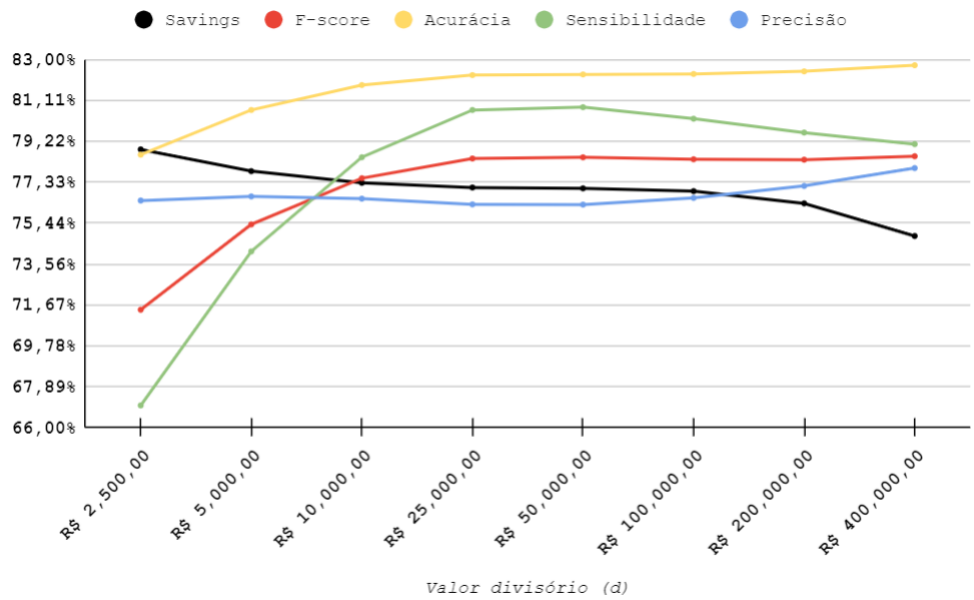
Análise dos resultados

A Figura 8 apresenta um gráfico de linhas que facilita a observação dos resultados de *savings* em uma tendência oposta quando comparada com as demais métricas. É importante ressaltar que a abordagem R3D, conforme apresentada no Capítulo 3, apresenta um comportamento de que, quanto maior for o valor divisório definido, mais predições serão realizadas pelo classificador insensível ao custo (*CIC*) e menos predições serão realizadas pelo classificador sensível ao custo (*CSC*). Observa-se também, à medida que o valor divisório aumenta, o *savings* diminui lentamente, enquanto que acurácia, *f-score*, sensibilidade e precisão aumentam lentamente.

Um efeito mais evidente se trata de um maior crescimento da sensibilidade quando o valor divisório aumenta, principalmente nos três primeiros valores. Isso mostra que, quanto mais pedidos têm a sua predição realizada pelo CSC (RF-BMR), mais casos de falsos negativos ocorrem. Considerando que o *f-score* é uma medida harmônica entre precisão e sensibilidade, percebe-se um aumento quando o valor divisório vai aumentando, mesmo com a precisão apresentando certa estabilidade. Por fim, a acurácia também é afetada pelo maior número de falsos negativos quando o valor divisório é menor.

O valor de R$ 50.000,00 aparenta ser o valor que traz um maior equilíbrio entre as métricas, inclusive por ter resultados similares ao número imediatamente menor (R$ 25.000,00) e o número imediatamente maior (R$ 100.000,00.). É importante ressaltar que esse valor é o mesmo valor definido para o falso positivo na matriz de custo (Tabela 6). Trata-se de um comportamento esperado, considerando que este valor muda o comportamento do RF-BMR, afetando as predições realizadas por este classificador.

Figura 8 – R3D: Efeitos observados em mudanças do valor divisório.



Resultados observados em *savings*, *f-score*, *acurácia*, *sensibilidade* e *precisão* quando diferentes valores divisórios são utilizados na abordagem R3D. Fonte: autoria própria.

4.2.3 Cenário 02

Com o objetivo de avaliar o comportamento dos classificadores RF e RF-BMR nos diferentes subconjuntos de teste, este cenário foi executado com a configuração definida a seguir.

Configuração

O treinamento foi realizado com todo o conjunto de treino (X) em ambos os classificadores. Porém, cada classificador foi testado separadamente com os subconjuntos de teste *Y1* e *Y2*, considerando o valor de d igual a R$ 50.000,00. Os resultados do RF-BMR são apresentados na Tabela 9 e os resultados do RF são apresentados na Tabela 10.

Classificador/Métrica	***Savings***	***Acurácia***	***F-score***	***Sensibilidade***	***Precisão***
RF-BMR em *Y1*	43,76% (±1,50%)	71,84% (±0,51%)	55,05% (±1,14%)	39,31% (±1,10%)	**91,89%** (±0,56%)
RF-BMR em *Y2*	**59,13%** (±1,68%)	**76,91%** (±0,80%)	**70,03%** (±0,95%)	**90,37%** (±0,758%)	57,17% (±1,15%)

Tabela 9 – Resultados do RF-BMR nos dois subconjuntos de teste.

Análise dos resultados

Os resultados do RF-BMR mostram uma diferença significativa em relação à sensibilidade e precisão, quando comparados os dois subconjuntos. Em relação a *Y1*, o RF-BMR alcança uma precisão de mais de 90% e uma sensibilidade de apenas 39%. Em relação a *Y2*,

Classificador/Métrica	***Savings***	***Acurácia***	***F-score***	***Sensibilidade***	***Precisão***
Random Forest em *Y1*	**4,03%** ($\pm 2,84\%$)	84,72% ($\pm 0,32\%$)	**81,96%** ($\pm 0,40\%$)	**79,04%** ($\pm 0,59\%$)	85,10% ($\pm 0,46\%$)
Random Forest em *Y2*	-580,73% ($\pm 490,82\%$)	**89,44%** ($\pm 0,44\%$)	80,91% ($\pm 0,85\%$)	75,04% ($\pm 1,09\%$)	**87,79%** ($\pm 0,91\%$)

Tabela 10 – Resultados do *Random Forest* nos dois subconjuntos de teste.

observa-se praticamente uma inversão: o RF-BMR alcança uma sensibilidade de mais de 90% e uma precisão de apenas 57%. Este resultado demonstra que o RF-BMR tende a rejeitar pedidos com alto valor (causando mais falsos positivos no subconjunto *Y2*) e aceitar pedidos com baixo valor (causando mais falsos negativos no subconjunto *Y1*). A acurácia, porém, alcança níveis similares em ambos os subconjuntos, com uma diferença de menos de 5%.

Os resultados do *Random Forest* apresentam *savings* negativo (com o subconjunto *Y2*) ou próximo de zero (com o subconjunto *Y1*). No que diz respeito a todas as outras métricas tradicionais (*f-score*, acurácia, sensibilidade e precisão), o RF apresenta valores acima de 75% em todas elas, inclusive apresentando estabilidade no número de falsos positivos e falsos negativos, independentemente do subconjunto de teste ao qual foi submetido.

Como visto, o subconjunto *Y2* contém os pedidos com altos valores de dívidas (valores acima do valor divisório de R$ 50.000,00). Particularmente, esse subconjunto contém alguns pedidos com valores bem maiores do que o valor divisório, em alguns casos até dez vezes maiores. Estes casos podem ser considerados raros sob o ponto de vista do conjunto de dados. Considerando as características da validação cruzada, quando acontece alguma validação em que nenhum desses casos raros são selecionados para fazer parte do conjunto de teste, o RF não erra sua predição, e o *savings* não é afetado significativamente. Entretanto, quando alguns desses casos são selecionados dentro do conjunto de testes e o RF erra sua predição, o *savings* é afetado significativamente, tornando negativa a média final. O alto desvio-padrão demonstra essa variação nas situações entre as iterações de validação cruzada: nem sempre ocorrem.

Estes resultados confirmam que RF-BMR tende a predizer aceitação para pedidos com valores menores (*Y1*), gerando um alto número de falsos negativos que afetam a sensibilidade. Também confirmam que o RF-BMR tende a predizer rejeição para pedidos com altos valores (*Y2*), gerando um alto número de falsos positivos que afetam a precisão.

Os resultados do *Random Forest* (RF) confirmam que este classificador não alcança índices aceitáveis em relação a *savings* em nenhum dos dois subconjuntos. Principalmente no subconjunto *Y2*, onde apresenta um resultado negativo e um alto desvio-padrão. Para o subconjunto *Y1*, consegue um valor próximo de zero e um desvio-padrão igualmente pequeno, demonstrando um melhor resultado e uma maior estabilidade.

4.2.4 Cenário 03

O objetivo deste cenário é avaliar se a abordagem R3D apresenta um maior equilíbrio em comparação aos classificadores *baseline* definidos (RF e RF-BMR), com o objetivo de verificar sua aplicabilidade no contexto real de negócio.

Configuração

A abordagem R3D foi configurada com o valor divisório d igual a R$ 50.000,00 por apresentar um maior equilíbrio para a abordagem R3D. Dessa forma, a Tabela 11 apresenta os resultados obtidos pela abordagem R3D ao lado dos resultados obtidos com as *baselines* definidas: RF como modelo tradicional e referência para as métricas de *f-score*, acurácia, sensibilidade e precisão; RF-BMR como modelo sensível ao custo e referência para a métrica de *savings*.

Classificador/Métrica	***Savings***	***Acurácia***	***F-score***	***Sensibilidade***	***Precisão***
R3D	77,05% (±0,69%)	82,30% (±0,37%)	78,48% (±0,45%)	**80,79%** (±0,57%)	76,30% (±0,48%)
RF-BMR	**79,01%** (±1,07%)	68,75% (±1,05%)	51,76% (±1,79%)	41,99% (±1,86%)	67,54% (±2,33%)
Random Forest	-132,09% (±147,85%)	**85,82%** (±0,28%)	**81,36%** (±0,40%)	77,51% (±0,58%)	**85,62%** (±0,43%)

Tabela 11 – Comparação entre a abordagem R3D, *Random Forest* e RF-BMR.

Análise dos resultados

Em comparação com o *Random Forest* (RF), a abordagem R3D demonstra uma maior efetividade quanto aos custos, alcançando um índice consideravelmente mais alto em relação a *savings*, e uma maior estabilidade, devido ao menor desvio-padrão. Apesar do RF apresentar melhores índices em *f-score*, acurácia e precisão, obtém valor negativo e um alto desvio-padrão para *savings*. Para o contexto de negócio do PRDI, o classificador RF não demonstra evitar perdas financeiras em relação à arrecadação tributária.

Em comparação com o RF-BMR, a abordagem R3D atinge níveis semelhantes quando considera-se a métrica *savings*, sendo apenas 1,96% menor. O RF-BMR alcança um bom nível em *savings*, ao predizer o aceite para muitos pedidos apenas por terem um baixo valor em sua dívida. Existem muitas instâncias com valores menores que R$ 50.000,00 (valor definido para d): a Tabela 5 mostra que mais da metade das dívidas envolvidas nos pedidos são menores que R$ 13.000,00. Como o custo de um falso negativo é o valor da dívida, então o RF-BMR calcula que um eventual falso negativo tem menor custo do que um eventual falso positivo. Por exemplo, a probabilidade da predição de um pedido com dívida de R$ 13.000,00 precisa ser maior do que 74% de rejeição (para o classificador-base) para que a predição do RF-BMR também seja de rejeição.

Concluindo a comparação com o RF-BMR, a abordagem R3D apresenta um aumento na sensibilidade, *f-score*, acurácia e precisão em quase 39%, aproximadamente 27%, 16% e 9% respectivamente. A melhora dos índices se dá principalmente pelo menor número de falsos negativos: menos pedidos estão sendo erroneamente aceitos, mesmo se a dívida tiver um valor pequeno.

A abordagem R3D alcança um melhor equilíbrio entre as métricas tradicionais e a métrica sensível ao custo. Em relação a *savings*, uma diminuição de apenas 1,96% não é desejável, porém não é significativa. Por outro lado, um aumento de 39% na sensibilidade é muito importante.

Em relação ao contexto de negócio no qual está inserido o PRDI, é importante mencionar que a sensibilidade pode ser considerada uma métrica mais importante do que a precisão. A sensibilidade expõe um alto número de falsos negativos, enquanto que a precisão expõe os falsos positivos. Para o problema de classificação do PRDI, um falso negativo significa perdoar indevidamente uma dívida e encerrar o processo, sem chance para correção futura. O falso positivo é um problema que pode ser considerado menos grave, pois mesmo se um pedido for erroneamente rejeitado, o contribuinte terá outros meios de pleitear o ajuste do processo de cobrança da dívida.

O maior equilíbrio alcançado entre as métricas faz sentido neste ambiente de negócio, pois não é justo aceitar o pedido apenas porque ele possui um valor baixo de dívida. Se a administração tributária assume esse comportamento, os contribuintes podem começar a registrar pedidos com dívidas de baixo valor mesmo se não existir justificativa para tal. A administração tributária exerce um papel importante na direção de uma economia saudável e justa, ao regular a conformidade tributária.

A análise dos pedidos deve se basear em diversos critérios e não apenas no custo. Ao mesmo tempo, é importante utilizar um classificador sensível ao custo para proteger erros de classificação que envolvem pedidos com altos valores. Este equilíbrio pode ser de interesse de outros órgãos públicos, onde a busca pelo menor custo nem sempre é aceitável, apesar de ser um fator importante a ser considerado.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O Pedido de Revisão de Dívida Inscrita (PRDI) consiste em um serviço oferecido pela Procuradoria-Geral da União (PGFN) aos contribuintes que possuem alguma dívida junto ao governo federal brasileiro. Esta dissertação de mestrado apresentou uma abordagem utilizando métodos de classificação tradicional e sensível ao custo, de forma híbrida, para a predição de deferimento/indeferimento de pedidos de revisão de dívida ativa.

A abordagem desenvolvida, denominada R3D, é baseada na utilização de modelos de classificação convencionais e sensíveis ao custo. A concepção dessa abordagem híbrida nasceu das observações de pontos de melhoria identificados ao utilizar apenas modelos tradicionais ou apenas modelos sensíveis ao custo no panorama do domínio do PRDI.

No caso dos modelos tradicionais, apesar de atingirem resultados promissores quando consideradas métricas tradicionais como a acurácia e *f-score*, eles demonstravam alguns erros de classificação considerados graves para o contexto de negócio como, por exemplo, a predição incorreta de deferimento de um pedido de milhões de reais em dívidas. Para a medição dos custos gerados através dos erros de classificação dos modelos, foi utilizada uma métrica baseada em custos, denominada de *savings*.

No caso dos modelos sensíveis ao custo, apesar de atingirem resultados muito melhores quanto ao *savings*, observa-se uma maior quantidade de erros de classificação quando comparados aos modelos tradicionais, por priorizar a minimização do risco de perdas financeiras.

As seguintes contribuições do trabalho de pesquisa apresentado por esta dissertação são identificadas:

- A construção e análise de diversos modelos de classificação utilizando algoritmos tradicionais com o objetivo de realizar a predição de registros do PRDI. Os resultados obtidos demonstraram resultados promissores na produção de um mecanismo de apoio à decisão a ser disponibilizado para a PGFN;
- A definição de uma matriz de custo específica para o problema de classificação do PRDI;
- A construção e análise de diversos modelos de classificação utilizando algoritmos sensíveis ao custo dependentes de exemplo e a avaliação dos modelos utilizando uma métrica baseada em custo. Para o problema de classificação do PRDI, demonstrou-se que esses modelos são mais eficientes do que os modelos tradicionais quando considerados os custos dos erros de classificação;
- A proposta de uma abordagem híbrida, usando modelos de classificação tradicionais e sensíveis ao custo de forma intercambiável, de acordo com um valor divisório;

- A instanciação da referida abordagem através de uma composição do melhor classificador tradicional junto com o melhor classificador sensível ao custo
- A definição de um valor divisório a ser utilizado na abordagem proposta, de forma a atingir um melhor equilíbrio em relação às métricas consideradas;
- Uma análise do *trade-off* demonstrado pela abordagem proposta em relação ao melhor classificador tradicional e o melhor classificador sensível ao custo.

O trabalho de pesquisa apresentado nesta dissertação resultou na publicação de 2 artigos e submissão de um terceiro artigo: Publicação de artigo no *23rd International Conference on Enterprise Information Systems* (LIMA et al., 2021); Publicação de artigo no Encontro Nacional de Inteligência Artificial e Computacional, premiado como o segundo melhor artigo do referido evento (LIMA; FERNANDES; MOURA, 2022); Submissão de um terceiro artigo ao Journal of Information and Data Management.

Ao comparar com o melhor classificador sensível ao custo (RF-BMR), a abordagem proposta apresentou um aumento de mais de 39% na sensibilidade, mais de 16% na acurácia, mais de 27% no *f-score* e mais de 9% na precisão. Esses resultados são atingidos com uma diminuição de apenas menos de 2% no *savings*. Dessa forma, a diminuição não significativa observada na métrica *savings* é compensada com um aumento significativo nas métricas tradicionais, demonstrando que a abordagem R3D pode ser considerada mais adequada quando aplicada ao problema de classificação do PRDI.

O uso bem-sucedido de técnicas de aprendizado na resposta aos PRDI também pode estimular o uso de tais técnicas em outras áreas e processos da PGFN ou outros órgãos federais, como Receita Federal do Brasil. Também pode estimular o uso de tais técnicas nas diversas secretarias de finanças municipais e estaduais e órgãos de auditoria como Controladoria-Geral da União e Tribunais de Conta em nível federal e estadual.

As limitações deste trabalho incluem a aplicação da abordagem proposta apenas em um problema de classificação proveniente da administração tributária, restando ser aplicada em problemas de classificação no âmbito das aplicações de cartões de crédito, conforme outros trabalhos relacionados. O caráter muito específico do problema de classificação do PRDI dificulta a comparação dos resultados obtidos com outros trabalhos relacionados. Por fim, a impossibilidade de divulgação de detalhes das variáveis utilizadas no conjunto de dados também consiste em uma limitação intrínseca a esta dissertação.

Em relação a trabalhos futuros, identificam-se:

- A implantação do modelo híbrido R3D dentro do fluxo de trabalho da PGFN. A implantação poderá permitir a avaliação da efetividade da abordagem do ponto de vista da aceleração das análises do PRDI;

- Aprofundamento no estudo da utilização de métodos de classificação sensíveis ao custo e suas aplicações;
- A utilização de abordagem de aprendizado de máquina tradicional, sensível ao custo e híbrida em outros problemas de classificação da administração tributária. A partir da disponibilização de conjuntos de dados dentro da própria PGFN ou de outros órgãos federais, estaduais ou municipais da adminsitração tributária, entende-se que poderá ser oportuno a avaliação de modelos de classificação sensíveis ao custo (além dos tradicionais) e a possibilidade do desenvolvimento/aplicação de abordagens híbridas, como a apresentada por esta dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

ALPAYDIN, E. *Introduction to machine learning.[Sl]*. [S.l.]: The MIT Press, 2010. Citado 3 vezes nas páginas 21, 22 e 24.

BAHNSEN, A. C.; AOUADA, D.; OTTERSTEN, B. Example-dependent cost-sensitive logistic regression for credit scoring. In: IEEE. *2014 13th International conference on machine learning and applications*. [S.l.], 2014. p. 263–269. Citado 5 vezes nas páginas 32, 36, 38, 39 e 40.

BAHNSEN, A. C.; AOUADA, D.; OTTERSTEN, B. Ensemble of example-dependent cost-sensitive decision trees. *arXiv e-prints*, p. arXiv–1505, 2015. Citado 6 vezes nas páginas 32, 36, 38, 39, 40 e 49.

BAHNSEN, A. C.; AOUADA, D.; OTTERSTEN, B. Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, Elsevier, v. 42, n. 19, p. 6609–6619, 2015. Citado 6 vezes nas páginas 31, 36, 38, 39, 40 e 47.

BAHNSEN, A. C. et al. Cost sensitive credit card fraud detection using bayes minimum risk. In: IEEE. *2013 12th international conference on machine learning and applications*. [S.l.], 2013. v. 1, p. 333–338. Citado 6 vezes nas páginas 32, 33, 36, 38, 39 e 40.

BATTISTON, P.; GAMBA, S.; SANTORO, A. Optimizing tax administration policies with machine learning. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper*, n. 436, 2020. Citado 3 vezes nas páginas 35, 38 e 39.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 25.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 25 e 26.

CAO, L. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 50, n. 3, p. 1–42, 2017. Citado na página 19.

CHAPMAN, P. et al. The crisp-dm user guide. In: *4th CRISP-DM SIG Workshop in Brussels in March*. [S.l.: s.n.], 1999. v. 1999. Citado na página 19.

CHEN, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, v. 1, n. 4, p. 1–4, 2015. Citado na página 36.

CUNHA, A. d. S.; KLIN, I. d. V.; PESSOA, O. A. G. Custo e tempo do processo de execução fiscal promovido pela procuradoria geral da fazenda nacional. Instituto de Pesquisa Econômica Aplicada (Ipea), 2011. Citado na página 42.

ELKAN, C. The foundations of cost-sensitive learning. In: LAWRENCE ERLBAUM ASSOCIATES LTD. *International joint conference on artificial intelligence*. [S.l.], 2001. v. 17, n. 1, p. 973–978. Citado 2 vezes nas páginas 17 e 30.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Citado na página 20.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156. Citado na página 26.

GEEM, Z. W.; KIM, J. H.; LOGANATHAN, G. V. A new heuristic optimization algorithm: harmony search. *simulation*, Sage Publications Sage CA: Thousand Oaks, CA, v. 76, n. 2, p. 60–68, 2001. Citado na página 34.

GHOSH, J. K.; DELAMPADY, M.; SAMANTA, T. Bayesian inference and decision theory. *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, p. 29–63, 2006. Citado na página 36.

GONZÁLEZ, P. C.; VELÁSQUEZ, J. D. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, Elsevier, v. 40, n. 5, p. 1427–1436, 2013. Citado 2 vezes nas páginas 34 e 38.

HANSEN, L. K.; SALAMON, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 12, n. 10, p. 993–1001, 1990. Citado na página 25.

HARRINGTON, P. *Machine learning in action*. [S.l.]: Manning Publications Co., 2012. Citado 8 vezes nas páginas 16, 21, 22, 23, 24, 27, 29 e 33.

HASTIE, T. *Tibshirani R. Friedman J.: The Elements of Statistical Learning*. [S.l.]: Springer-Verlang (es), 2001. Citado na página 26.

HÖPPNER, S. et al. Instance-dependent cost-sensitive learning for detecting transfer fraud. *European Journal of Operational Research*, Elsevier, v. 297, n. 1, p. 291–300, 2022. Citado 3 vezes nas páginas 36, 38 e 39.

HOSSIN, M.; SULAIMAN, M. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015. Citado 2 vezes nas páginas 28 e 29.

IPPOLITO, A.; LOZANO, A. C. G. Tax crime prediction with machine learning: A case study in the municipality of são paulo. In: *ICEIS (1)*. [S.l.: s.n.], 2020. p. 452–459. Citado 3 vezes nas páginas 35, 38 e 39.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 36.

KIM, J. et al. Classification cost: An empirical comparison among traditional classifier, cost-sensitive classifier, and metacost. *Expert Systems with Applications*, Elsevier, v. 39, n. 4, p. 4013–4019, 2012. Citado na página 29.

LIMA, H. S.; FERNANDES, D. Y. de S.; MOURA, T. J. On the evaluation of example-dependent cost-sensitive models for tax debts classification. In: SBC. *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2022. p. 425–436. Citado na página 57.

LIMA, H. S. et al. On the evaluation of classification methods applied to requests for revision of registered debts. 2021. Citado 5 vezes nas páginas 16, 35, 38, 50 e 57.

LÓPEZ, C. P.; RODRÍGUEZ, M. J. D.; SANTOS, S. de L. Tax fraud detection through neural networks: an application using a sample of personal income taxpayers. *Future Internet*, Multidisciplinary Digital Publishing Institute, v. 11, n. 4, p. 86, 2019. Citado 3 vezes nas páginas 34, 38 e 39.

MARTÍNEZ-PLUMED, F. et al. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, 2019. Citado na página 20.

MATHEWS, J. et al. Regression analysis towards estimating tax evasion in goods and services tax. In: IEEE. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. [S.l.], 2018. p. 758–761. Citado na página 13.

MEHTA, P. et al. Deepcatch: Predicting return defaulters in taxation system using example-dependent cost-sensitive deep neural networks. In: IEEE. *2020 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2020. p. 4412–4419. Citado 3 vezes nas páginas 36, 38 e 39.

MITCHELL, T.; MCGRAW-HILL, M. L. *Edition*. [S.l.]: New York: McGraw-Hill, Inc, 1997. Citado 2 vezes nas páginas 21 e 29.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of machine learning*. [S.l.]: MIT press, 2018. Citado na página 22.

NAZAROV, M.; MIKHALEVA, O.; CHERNOUSOVA, K. Digital transformation of tax administration. In: SPRINGER. *International Scientific Conference Digital Transformation of the Economy: Challenges, Trends, New Opportunities*. [S.l.], 2019. p. 144–149. Citado na página 16.

OECD. *Advanced Analytics for Better Tax Administration*. [s.n.], 2016. 60 p. Disponível em: <https://www.oecd-ilibrary.org/content/publication/9789264256453-en>. Citado na página 15.

OECD. *Tax Administration 3.0: The Digital Transformation of Tax Administration*. [s.n.], 2020. 77 p. Disponível em: <http://www.oecd.org/tax/forum-on-tax-administration/publications-and-products/tax-administration-3-0-the-digital-transformation-of-tax-administration.htm>. Citado 2 vezes nas páginas 13 e 15.

PGFN. 2021. Sobre a PGFN. Disponível em: <https://www.gov.br/pgfn/pt-br/acesso-a-informacao/institucional/sobre-a-pgfn>. Acesso em: 15 jun 2021. Citado na página 14.

PGFN. 2021. Revisão de Dívida Inscrita (PRDI). Disponível em: <https://www.gov.br/pgfn/pt-br/servicos/orientacoes-contribuintes/revisao-de-divida-inscrita-prdi>. Acesso em: 15 jun 2021. Citado na página 14.

POLIKAR, R. Ensemble learning. In: *Ensemble machine learning*. [S.l.]: Springer, 2012. p. 1–34. Citado na página 25.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013. Citado 2 vezes nas páginas 15 e 19.

RAHIMIKIA, E. et al. Detecting corporate tax evasion using a hybrid intelligent system: A case study of iran. *International Journal of Accounting Information Systems*, Elsevier, v. 25, p. 1–17, 2017. Citado 2 vezes nas páginas 34 e 38.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. Citado na página 25.

SCOTT, C. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. In: *ICML*. [S.l.: s.n.], 2011. Citado na página 40.

SILVA, L. S. da; CARVALHO, R. N.; SOUZA, J. C. F. Predictive models on tax refund claims-essays of data mining in brazilian tax administration. In: SPRINGER. *International Conference on Electronic Government and the Information Systems Perspective*. [S.l.], 2015. p. 220–228. Citado 2 vezes nas páginas 35 e 38.

SOARES, G. de V.; CUNHA, R. Predição de irregularidade fiscal dos contribuintes do tributo iss. In: SBC. *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*. [S.l.], 2020. p. 223–228. Citado 3 vezes nas páginas 35, 38 e 39.

WANG, Y.; TIONG, R. L. Public–private partnership contract failure prediction using example-dependent cost-sensitive models. *Journal of Management in Engineering*, American Society of Civil Engineers, v. 38, n. 1, p. 04021079, 2022. Citado 4 vezes nas páginas 36, 38, 39 e 40.

WU, R.-S. et al. Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, Elsevier, v. 39, n. 10, p. 8769–8777, 2012. Citado 2 vezes nas páginas 17 e 40.

ZADROZNY, B.; ELKAN, C. Learning and making decisions when costs and probabilities are both unknown. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2001. p. 204–213. Citado na página 17.

ZELENKOV, Y. Example-dependent cost-sensitive adaptive boosting. *Expert Systems with Applications*, Elsevier, v. 135, p. 71–82, 2019. Citado 6 vezes nas páginas 32, 36, 38, 39, 40 e 49.

ZHOU, Z.-H. Ensemble learning. *Encyclopedia of biometrics*, v. 1, p. 270–273, 2009. Citado na página 25.

ZHOU, Z.-H. *Ensemble methods: foundations and algorithms*. [S.l.]: CRC press, 2012. Citado na página 32.

ZHOU, Z.-H. Ensemble learning. In: *Machine Learning*. [S.l.]: Springer, 2021. p. 181–210. Citado na página 25.

ZHOU, Z.-H. *Machine learning*. [S.l.]: Springer Nature, 2021. Citado na página 19.