Sheilla Pereira de Barros

Fronteira entre Risco Operacional e Risco de Crédito: classificação de eventos de perdas operacionais

Brasília-DF

2022

Sheilla Pereira de Barros

Fronteira entre Risco Operacional e Risco de Crédito: classificação de eventos de perdas operacionais

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

Universidade de Brasília - UnB

Faculdade de Administração Contabilidade e Economia - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Orientador: Prof. Dr. Herbert Kimura

Brasília-DF

2022

Sheilla Pereira de Barros
Fronteira entre Risco Operacional e Risco de Crédito: classificação de eventos de perdas operacionais/ Sheilla Pereira de Barros. – Brasília-DF, 2022-
57p. : il. (algumas color).

Orientador: Prof. Dr. Herbert Kimura

Dissertação (Mestrado) – Universidade de Brasília - UnB
Faculdade de Administração Contabilidade e Economia - FACE
Departamento de Economia - ECO
Programa de Pós-Graduação, 2022.

1. Aprendizagem de máquina. 2. Classificação supervisionada. 3. Perda operacional. I. Universidade de Brasília. II. Faculdade de Administração, Contabilidade e Economia - FACE. III. Departamento de Economia IV. Fronteira entre Risco Operacional e Risco de Crédito: classificação de eventos de perdas operacionais

Sheilla Pereira de Barros

Fronteira entre Risco Operacional e Risco de Crédito: classificação de eventos de perdas operacionais

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

Trabalho aprovado. Brasília-DF, 15 de junho de 2022:

Prof. Dr. Herbert Kimura
Orientador

Prof. Dr. João Gabriel de Moraes

Prof. Dr. Leonardo Fernando Cruz Basso

Brasília-DF
2022

Este trabalho é dedicado a todos aqueles que sonham mais alto que drones, para que amanhã não seja só um ontem, com um novo nome...

Agradecimentos

Primeiramente agradeço a Deus e aos meus antepassados que, com muita generosidade e amor, me conduziram até esse momento - em todos os aspectos. Agradeço às minhas meninas por todo o apoio recebido em todos os dias da minha vida, não só àqueles dedicados ao mestrado, mas em especial quero que saibam o quanto cada palavra de incentivo (e foram muitas) me motivou a seguir em frente e não desistir no meio do caminho. Muito obrigada ao meu pai. Eu sinto o orgulho que você está sentindo por esta conquista! Aonde quer que você esteja, receba a minha eterna gratidão... Muito obrigada à minha mãe, por me mostrar que não há nada que eu não possa fazer e por estar sempre ao meu lado. Muito obrigada à minha irmã, por nunca soltar a minha mão, independente do terreno. Muito obrigada à minha namorada, por acreditar no meu potencial até nos dias em que eu não acredito. Agradeço ao meu empregador pela maravilhosa oportunidade de integrar o grupo discente da tão almejada Universidade de Brasília (e por tantas outras coisas). Agradeço a todo o corpo docente que, com grande dedicação e generosidade, passaram fins de semana inteiros em sala de aula conosco, em especial ao professor Dr. Daniel Cajueiro por todo o seu ensinamento e paciência. Agradeço ao meu orientador professor Dr. Herbert Kimura que me auxiliou até este momento. Agradeço aos meus colegas de classe, que de uma forma ou de outra foram relevantes para esse trabalho, em especial à Tainá e ao Victor que tanto me ajudaram em todas as etapas deste mestrado.

Agradeço, agradeço, agradeço, muito obrigada!

"Foi o tempo que dedicaste à tua rosa
que a fez tão importante".
(Antoine de Saint-Exupéry)

Resumo

Neste estudo avaliamos a aplicação de técnicas de aprendizagem de máquina para classificação de eventos de perda operacional com provável impacto no risco de crédito. Para tanto, utilizamos base de dados de perdas de uma instituição financeira brasileira. Na sequência selecionamos uma combinação de modelos probabilísticos (näive bayes e regressão logística) e de aprendizagem de máquina (árvore de decisão e random forest) para avaliar a precisão da classificação. Os resultados se mostraram satisfatórios para a maior parte dos algoritmos testados em especial àqueles com métrica em Árvore de Decisão. O presente estudo pretende contribuir não só academicamente (frente à escassez de referencial bibliográfico diretamente relacionado a Risco de Fronteira entre o Risco Operacional e o Risco de Crédito), mas também corporativamente mediante a otimização de classificação de informações pelas instituições financeiras, o que pode acarretar no aprimoramento da eficiência operacional, melhoria no gerenciamento de riscos (de forma integrada) e qualificação dos dados de perdas operacionais.

Palavras-chave: aprendizagem de máquina, classificação supervisionada, perda operacional.

Abstract

In this study, we evaluated the application of machine learning techniques to classify operational loss events with a probable impact on credit risk. For this, we used a database of losses from a Brazilian financial institution. Next, we selected a combination of probabilistic models (näive bayes and logistic regression) and machine learning (decision tree and random forest) to evaluate the classification accuracy. The results were satisfactory for most of the algorithms tested, especially those with Decision Tree metrics. The present study intends to contribute not only academically (in view of the scarcity of bibliographic reference directly related to Border Risk between Operational Risk and Credit Risk), but also corporately by optimizing the classification of information by financial institutions, which can lead to in improving operational efficiency, improving risk management (in an integrated manner) and qualifying operational loss data.

Keywords: machine learning, supervised classification, operational loss.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AD	Árvore de Decisão
BACEN	Banco Central do Brasil
BIS	Bank for International Settlements (Banco de Compensações Internacionais)
CART	Classification and Regression Trees
CMN	Conselho Monetário Nacional
ICAAP	Internal Capital Adequacy Assessment Process (Processo de Avaliação de Adequação de Capital Interno)
NB	Näive Bayes
RF	Random Forest
RL	Regressão Logística
SMOTE	Synthetic Minority Over-sampling Technique (Técnica de Sobreamostragem de Minoria Sintética)
SUSEP	Superintendência de Seguros Privados

Sumário

1 Introdução

Delimitar a fronteira entre Risco Operacional e os demais tipos de riscos existentes no mercado financeiro é uma expectativa do Banco Central do Brasil e um grande desafio para as instituições financeiras. Primeiramente esclarecemos que os limites entre os tipos de riscos não são explicitamente definidos. Há uma dificuldade na identificação de fatores operacionais relacionados a falhas e fraudes que resultem numa possível inadimplência do tomador – por exemplo. Além disso, o crime organizado e a atuação de fraudadores estão cada vez mais aprimorados, o que gera inúmeros gastos das instituições financeiras com a adoção de medidas mitigadoras de riscos na tentativa de prevenir perdas. O direcionamento de esforços na mitigação e regularização de causas que não correspondam à real origem das consequências obtidas com a materialização de riscos representa prejuízo aos bancos. Por esse motivo, faz-se necessário o desenvolvimento de metodologias que possibilitem aos bancos a identificação e a avaliação dos riscos de fronteira, pois a correta classificação auxilia as instituições na gestão de riscos e tomada de decisão, além de possibilitar o direcionamento adequado de capital e atendimento à legislação vigente.

Cabem às instituições financeiras definir, reconhecer e prever os eventos de risco operacional que deem abertura para a materialização de outros tipos de riscos. Mas para tanto, é necessária a aplicação de técnicas que auxiliem na classificação das perdas operacionais e que sinalizem eventual fronteira entre os tipos de riscos.

Neste estudo, temos o intuito de sinalizar a fronteira entre o Risco Operacional e o Risco de Crédito. Para tanto, avaliamos a base de perdas operacionais de uma instituição financeira brasileira (banco múltiplo), definimos indícios de risco de crédito nos lançamentos de eventos operacionais vinculados a produtos da carteira comercial, aplicamos diferentes técnicas de classificação supervisionada para as situações de fronteira identificadas na base de dados avaliada e, na sequência, comparamos os resultados obtidos.

Chernobai, Jorion e Yu (2011) demonstrou uma estreita relação entre risco operacional e risco de crédito ao verificar a frequência anual de eventos de perda de uma amostra entre os anos 1980 e 2005, focada em instituições financeiras dos Estados Unidos. Na análise realizada, foi verificado que a frequência dos eventos de risco operacional foi aumentando desde 1980, mas apresentou um declínio acentuado depois do ano 2001. O padrão é semelhante ao do número de inadimplências do setor financeiro no mesmo período. Essa conjectura direciona as instituições financeiras a estimar uma distribuição de perdas considerando não somente o risco operacional, assim como o risco de crédito gerado indiretamente.

Com o mesmo enfoque, o estudo conduzido por Ko, Lee e Anandarajan (2019)

evidenciou que o risco de crédito é positivamente associado a incidentes de risco operacional. Quanto maiores os incidentes de risco operacional, maior o risco de crédito.

A intenção deste estudo é facilitar a classificação de eventos de perdas operacionais com possível reflexo em risco de crédito, o que será benéfico academicamente, tanto quanto servirá para o aprimoramento das informações de perdas operacionais de instituições financeiras. O uso de ferramentas que auxiliem na classificação de eventos dessa natureza, além de garantir ganho em eficiência operacional, aprimora o gerenciamento de riscos da instituição, o que impacta diretamente na tomada de decisões estratégicas, mitigação de riscos e diminuição de perdas. Como afirmado por Shah (2019):

> Nosso conhecimento de risco bancário aumenta à medida que nossa capacidade de quantificar o risco aumenta. Nosso conhecimento de risco bancário aumenta à medida que nossa capacidade de desagregar o risco para níveis mais granulares aumenta.

Para abordar o tema proposto, dentro das limitações deste estudo, o objetivo principal da pesquisa realizada é responder à questão: "como classificar os lançamentos de perdas operacionais com possibilidade de impacto em risco de crédito?".

Dentre os objetivos acessórios constam: a) compreender a eventual relação entre risco operacional e risco de crédito; b) avaliar algoritmos de classificação supervisionada de aprendizagem de máquina e sua aplicabilidade na base de perdas operacionais de um banco brasileiro; c) apresentar o resultado da pesquisa.

Embora alguns autores confirmem a relação entre risco operacional e risco de crédito, tais como Chernobai, Jorion e Yu (2011) e Ko, Lee e Anandarajan (2019), em nossas pesquisas, não identificamos artigos científicos acadêmicos (ou mesmo material bibliográfico em formato distinto) que abordasse especificamente o tema de fronteira entre os riscos. Isso demonstra a relevância do presente estudo, assim como a necessidade de aprofundamento do assunto em trabalhos futuros.

Organizamos este trabalho em 5 capítulos: os capítulos 1 e 2 são abordados a problemática, a regulamentação vigente e a contextualização necessária à compreensão do tema; enquanto no capítulo 3 apresentamos a metodologia utilizada, os dados e o seu tratamento, e resumimos as funcionalidades e objetivos de cada um dos modelos de aprendizagem de máquina aplicado; no capítulo 4, demonstramos os resultados obtidos e o diagnóstico referente aos modelos aplicados; por fim, no capítulo 5 de encerramento, finalizamos o presente estudo contemplando a aplicabilidade dos algoritmos na base de perdas da instituição avaliada, assim como sugerimos novos estudos para ampliar as análises sobre o tema.

2 Relação entre os Riscos Operacional e o de Crédito

O risco operacional é definido como a possibilidade de ocorrência de perdas resultantes de falha, deficiência ou inadequação de processos internos, pessoas e sistemas, ou de eventos externos (BANCO CENTRAL DO BRASIL, 2006). Para Feuerverger (2016), o risco operacional é uma fonte de risco (...) que surge em todas as linhas de negócios de um banco (por exemplo, banco de varejo, banco comercial e operações de corretagem) e envolve uma variedade de tipos de perdas que podem ocorrer em qualquer linha de negócios (por exemplo, fraude externa, fraudes internas e falhas do sistema).

Por sua vez, risco de crédito é a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação (BANCO CENTRAL DO BRASIL, 2009). Para Psillaki, Tsolas e Margaritis (2010), o risco de crédito é o risco mais significativo para as instituições financeiras, e por meio de uma gestão eficaz da exposição ao risco de crédito, essas instituições não só apoiam a viabilidade e rentabilidade do seu próprio negócio, como também contribuem para a estabilidade sistêmica e para uma alocação eficiente de capital na economia.

A possível relação entre os dois tipos de riscos é sinalizada pelo Banco Central do Brasil na constituição da base de dados de risco operacional na qual devem constar as perdas operacionais associadas a risco de mercado e a risco de crédito (BANCO CENTRAL DO BRASIL, 2017a).

A fronteira entre ambos riscos está consignada nos eventos de risco operacional (como falhas e fraudes) que impactam no não cumprimento das obrigações pactuadas pelo tomador de crédito, gerando assim a materialização do risco de crédito. Para auxiliar no entendimento, a SUSEP (2014) exemplifica:

> suponha uma aplicação financeira aprovada por um gerente de investimentos e cuja contraparte envolvida tenha declarado falência, não honrando com a restituição dos valores investidos. A princípio, a perda verificada estaria relacionada a um evento de risco de crédito. Contudo, foi constatado que o gerente de investimentos extrapolou a alçada para ele definida na política de investimentos da empresa no que tange ao limite de aplicação que poderia ser por ele aprovada. Ou seja, os controles de segurança da empresa não impediram

a violação da citada alçada.

O Banco Central do Brasil (2013) estabelece que as perdas operacionais relacionadas a risco de crédito cuja causa seja claramente identificada como risco operacional devam compor a base de cálculo de risco operacional. São vários os exemplos possíveis para essa situação como, eventos em que, por falha operacional, um cliente é avaliado erroneamente e disso resulta uma futura inadimplência do contrato, ou seja, o risco de crédito foi originado pelo risco operacional.

Para determinados casos o risco operacional não representa o fim em si, sendo mais aderente a atribuição de outro tipo de risco, como de crédito ou de mercado ou liquidez (dentre outros). E é neste contexto que está a relevância de investigar uma possível fronteira que determine os riscos operacionais com impacto em outros tipos de riscos: para que mediante a classificação de tais casos, a instituição financeira aprimore os mitigadores adotados para minimizar as perdas incorridas.

Para o Bank for International Settlements (2006), algumas perdas são claramente resultado do risco operacional, entretanto, em outros casos, não resta claro se as perdas devem ser classificadas como risco operacional ou de crédito, por exemplo. Nestes casos, poderia ser apropriado alocar parte da perda ao risco operacional e parte para o risco de crédito. Tais problemas de classificação são descritos como questões de "limite". O BIS afirmou, em 2006, que o Acordo de Basileia II era relativamente claro em relação ao limite de risco operacional e risco de mercado, mas deixava margem para interpretação em relação ao limite entre risco operacional e risco de crédito.

Decerto que o assunto é pouco explorado academicamente, restando às instituições financeiras o aprofundamento do tema e a elaboração interna de metodologias para identificação e classificação da fronteira entre os riscos operacional e de crédito.

3 Metodologia

O intuito deste estudo é classificar, dentro da base de perdas operacionais da instituição financeira, aqueles lançamentos que sinalizam provável impacto em risco de crédito mediante a utilização de técnicas de aprendizagem de máquina para classificação supervisionada.

O início deste trabalho deu-se por meio de pesquisa bibliográfica com ênfase no tema objeto, na regulamentação brasileira relacionada a risco de fronteira e em assuntos correlatos ao nosso objetivo. O segundo passo foi coletar e analisar o conjunto de dados aos quais tivemos acesso. Na sequência, realizamos o tratamento dos dados descrito no item 3.1.1, aplicamos os modelos de aprendizagem de máquina apresentados no item 3.2 e consolidamos os resultados a partir da aplicação dos modelos no conjunto de dados (4).

3.1 Dados

Neste estudo, utilizamos dados da base de perdas operacionais de instituição financeira brasileira (banco múltiplo), lançados manualmente, referente ao período de Julho de 2020 a Dezembro de 2021.

De acordo com o Banco Central do Brasil (2017b), é definido como perda operacional o valor quantificável associado aos eventos de risco operacional. Assim, o regulador direciona à composição da base de perdas operacionais, também, aqueles eventos de risco operacional que gerem risco de crédito, como falhas humanas ou sistêmicas.

A base é composta, originalmente, por 1.979.836 lançamentos, segregados nos 8 tipos de eventos de nível 1 estabelecidos pelo Banco Central do Brasil (2020):

Evento Risco Operacional N1	**Qtde Lançamentos**
Danos a Ativos Físicos Próprios ou em Uso	275
Demandas Trabalhistas e Segurança Deficiente de Trabalho	1.958
Situações que Acarretem a Interrupção das Atividades	46
Falhas Sistemas de TI	38.019
Falhas Execução, Prazos e Gerenciamento Atividades	81.283
Fraudes Externas	485.687
Fraudes Internas	4.124
Práticas Inadequadas Relativas a clientes, Produtos e Serviços	1.368.444

Tabela 1 – Lançamentos por Tipo de Evento de Risco Operacional

3.1.1 Tratamento dos Dados

Realizamos o tratamento prévio dos dados para melhor direcioná-los aos objetivos deste estudo, conforme os seguintes parâmetros:

- utilização apenas de lançamentos com valores superiores a R$ 1.000,00 (um mil reais);
- exclusão de lançamentos com preenchimento NULL (dados faltantes); dos lançamentos em situação de estorno (SL2); e, de lançamento *outlier* [1];
- exclusão de atributos não necessários ao presente estudo (como, por exemplo, informações de estado/UF e agência de vinculação);
- criação de atributo classe denominado fronteira;
- seleção apenas de lançamentos cuja data de origem do evento corresponda ao período de referência da base.

Na base de dados utilizada, consta a data de origem do evento assim como, a data de lançamento; ou seja, a data em que tal valor passou a integrar o balanço da instituição. Durante o tratamento dos dados, desconsideramos todos os lançamentos com origem em data anterior ao período de referência - quer seja, Julho de 2020 a Dezembro de 2021. Para Chernobai, Ozdagli e Wang (2021), esse recurso evita problemas de identificação associados a informações desatualizadas, como perdas no balanço patrimonial decorrentes de riscos assumidos anos antes.

Após o tratamento, a base foi reduzida significativamente, se consolidando em 261.858 lançamentos, ou seja, 13,22% da base primária, divididos conforme demonstrado na tabela 2:

Evento Risco Operacional N1	**Qtde Lançamentos**
Danos a Ativos Físicos Próprios ou em Uso	91
Demandas Trabalhistas e Segurança Deficiente de Trabalho	0
Situações que Acarretem a Interrupção das Atividades	4
Falhas Sistemas de TI	5.514
Falhas Execução, Prazos e Gerenciamento Atividades	206
Fraudes Externas	244.716
Fraudes Internas	3.196
Práticas Inadequadas Relativas a clientes, Produtos e Serviços	8.131

Tabela 2 – Lançamentos por Tipo de Evento de Risco Operacional Pós Tratamento

[1] lançamento único em valor significativamente superior aos demais.

Para este estudo, dividimos as variáveis entre 6 atributos previsores e 1 atributo classe. Os atributos previsores correspondem às variáveis que são utilizadas para prever o valor da classe. E definimos como atributo classe a variável denominada fronteira:

Atributos Previsores	**Atributo Classe**
Nome do Evento Contábil	
Produto	
Valor	Fronteira
Evento de Risco Operacional N1	
Evento de Risco Operacional N2	
Evento de Risco Operacional N3	

Tabela 3 – Divisão de atributos previsores e classe

As categorias de evento de risco operacional nível 1 (N1) e nível 2 (N2) são estabelecidas pelo Banco Central do Brasil (2020) (Vide Anexo A), enquanto que o nível 3 (N3) é definido pelas instituições financeiras.

Como a base é composta por variáveis numéricas e categóricas, durante o tratamento, convertemos os atributos de *string* para números (função Label Encoder) e realizamos a transformação dos dados em variáveis binárias (função One Hot Encoder).

A dificuldade no tratamento dos dados relaciona-se, principalmente, ao caráter propriamente contábil da base de perdas operacionais da instituição. Os bancos de dados contábeis não precisam ser tão abrangentes e detalhados quanto os bancos de dados de perdas operacionais: os últimos exigem maiores quantidades e melhores qualidades de dados de perdas (JONGH et al., 2013).

3.1.1.1 Atributo Classe (Fronteira)

Criamos o atributo fronteira na base de dados como variável classe para aplicação das técnicas de aprendizagem de máquina analisadas, sendo que para tanto, foram sinalizados os eventos de possível fronteira, da seguinte forma:

- Classe 0: Não há indícios de fronteira
- Classe 1: Há indícios de fronteira

A base de perdas operacionais utilizada não dispõe de vínculo com contratos da carteira ativa, ou seja, embora os lançamentos sejam vinculados a produtos/operações de crédito, não há relação com um contrato ou cliente específico - impossibilitando a análise real da adimplência ou inadimplência (*default*) no crédito em si. Dessa forma, avaliamos como necessária a sinalização de lançamentos passíveis de risco de fronteira.

Para fins deste estudo, classificamos como possibilidade de risco de fronteira aqueles lançamentos vinculados a operações de crédito cujos eventos de risco operacional estejam relacionados a fraude interna e fraude externa.

Selecionamos os lançamentos vinculados a eventos de fraude externa e fraude interna devido à alta probabilidade de *default* (não pagamento) de operações de crédito nessas situações. Inclusive, fraude é utilizado como exemplo pelo Banco Central do Brasil (2017a) no texto da Carta Circular nº 3.841/2017 onde define que devam constar no relatório ICAAP as informações relacionadas a riscos relevantes, dentre eles os "riscos de fronteira entre operacional e crédito, tais como fraude em crédito"(grifo nosso).

Jongh et al. (2013) define fraude interna como as perdas decorrentes de atos destinados a fraudar, apropriar-se indevidamente de bens ou burlar regulamentos, a lei ou a política da empresa, que envolvam pelo menos uma parte interna (empregado da instituição). Por sua vez, fraude externa é definida como perdas decorrentes de atos destinados a fraudar, apropriar-se indevidamente de bens ou burlar a lei por um terceiro à instituição.

Ainda que não tenha sido abordado neste estudo, reconhecemos que outros tipos de eventos de perdas operacionais, como por exemplo, eventos vinculados às categorias de falha em sistemas de tecnologia da informação (TI) e de práticas inadequadas relativas a clientes, produtos e serviços também possam gerar risco de crédito futuro advindo de uma inadimplência do cliente tomador dado a eventual equívoco/erro na negociação e liberação da transação - assim como outros.

3.1.1.2 Balanceamento da Base

A base apresenta desbalanceamento em vista da diminuta ocorrência de lançamentos classe 1, ou seja, eventos com indícios de fronteira:

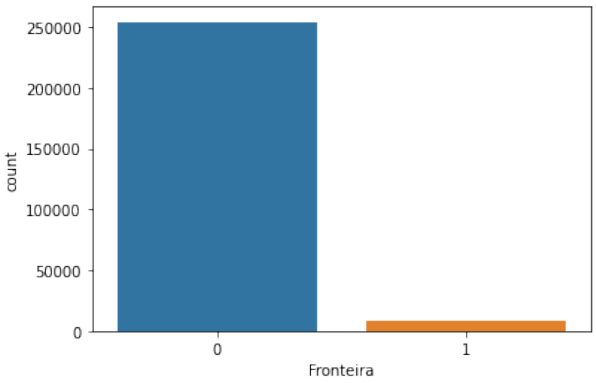

Figura 1 – Divisão Fronteira de Riscos Original

De acordo com Chawla et al. (2002), um conjunto de dados é desequilibrado se as categorias de classificação não forem representadas de forma aproximadamente igual. A

figura 1 demonstra o desbalanceamento da variável classe fronteira. Como pode ser visto, a maior parte dos lançamentos correspondem a 0, ou seja, não há indício de fronteira na maioria dos lançamentos avaliados (aplicados os critérios de seleção mencionados anteriormente).

Para balanceamento da base, aplicamos o algoritmo SMOTE (da biblioteca imblearn.over_sampling), (como descrito nos itens 4.1 e 4.2). Dessa forma, a classe minoritária é super amostrada introduzindo exemplos sintéticos ao longo dos segmentos de linha que unem qualquer/todos os k vizinhos mais próximos da classe minoritária (CHAWLA et al., 2002).

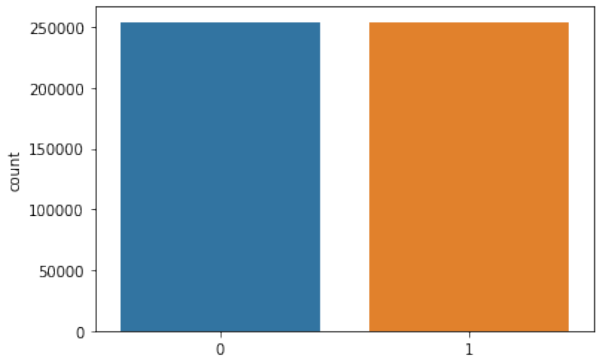

Figura 2 – Divisão Fronteira de Riscos - SMOTE

3.2 Aprendizagem de Máquina

Diferentes técnicas de aprendizagem de máquina são estudadas academicamente para detecção de transações suspeitas de lavagem de dinheiro (CHEN et al., 2018), para identificação de risco de fraude em demonstrações financeiras (SONG et al., 2014), para detecção de fraude em cartão de crédito (WANG; CHEN; CHEN, 2019), (CHEN; CHEN; LIN, 2006), (SHEN; TONG; DENG, 2007), (SÁ; PEREIRA; PAPPA, 2018), previsão de falência de empresas como ferramenta de gerenciamento de risco de crédito (BARBOZA; KIMURA; ALTMAN, 2017), dentre outras. Considerando a premissa de elementos de risco operacional com impacto em risco de crédito nos exemplos citados, depreendemos a viabilidade de utilização de determinados algoritmos de aprendizagem de máquina para classificar os eventos com provável risco de fronteira.

A aprendizagem de máquina supervisionada é a busca por algoritmos que analisam a partir de instâncias fornecidas externamente para produzir hipóteses gerais, e então fazem previsões sobre instâncias futuras. Em outras palavras, o objetivo do aprendizado supervisionado é construir um modelo conciso da distribuição de rótulos de classe em termos de recursos previsores. O classificador resultante é então usado para atribuir rótulos

de classe às instâncias de teste em que os valores dos recursos do previsor são conhecidos, mas o valor do rótulo de classe é desconhecido. (KOTSIANTIS, 2007).

Para distinguir os eventos de risco operacional daqueles em que há a possível consequência de um risco de crédito adjunto, selecionamos técnicas de aprendizagem de máquina que permitem fazer o reconhecimento de padrões para, assim, classificar novos casos. Seguimos o raciocínio de Sá, Pereira e Pappa (2018), por também estarmos interessados em algoritmos que gerem modelos facilmente interpretáveis (classificadores), que é o caso das árvores de decisão e classificadores de redes bayesianas.

Na classificação, o objetivo de um algoritmo de aprendizagem é construir um classificador dado um conjunto de exemplos de treinamento com rótulos de classe (ZHANG, 2004). Tendo em vista o objetivo de classificação e as variáveis dispostas na base de dados utilizada, optamos pela aplicação de duas técnicas estatísticas tradicionais (Näive Bayes e Regressão Logística) e duas técnicas de aprendizagem de máquina (Árvore de Decisão e Random Forest) para comparação dos resultados.

Para a aplicação dos algoritmos apresentados, utilizamos, principalmente, as bibliotecas scikit-learn e numpy, em linguagem python, no ambiente do Google Colab [2]. Scikit-learn é um módulo Python que integra uma ampla gama de algoritmos de aprendizagem de máquina de última geração para problemas supervisionados e não supervisionados de média escala. Este pacote se concentra em levar a aprendizagem de máquina para não especialistas usando uma linguagem de alto nível de uso geral (PEDREGOSA et al., 2011).

3.2.1 Naïve Bayes

É um classificador probabilístico baseado no teorema de Bayes que calcula a probabilidade de hipóteses e estima a classificação de novos objetos.

A probabilidade de ocorrência de um evento B pode ser estimada pela frequência com que ele ocorre, assim como é possível estimar a probabilidade de que um evento B ocorra, para cada classe ou evento A, $P(B|A)$. Em complemento, o teorema de Bayes fornece uma maneira de calcular a probabilidade de um evento ou objeto pertencer a uma classe $P(A|B)$ utilizando a probabilidade *a priori* da classe $P(A)$, a probabilidade de observar vários objetos que pertencem à classe $P(B|A)$ e a probabilidade de ocorrência desses objetos $P(B)$ (CARVALHO et al., 2011).

Para Zhang (2004) Näive Bayes é a forma mais simples de rede bayesiana, em que todos os atributos são independentes, dado o valor da variável de classe. E é um classificador útil para procedimentos que precisam ser repetidos muitas vezes, como testes de permutação, por ser muito mais rápido de treinar do que os outros (PEREIRA; MITCHELL; BOTVINICK, 2009).

[2] Google Colaboratory: ferramenta em nuvem que permite criar e executar códigos na linguagem Python.

Neste estudo, utilizamos o algoritmo Gaussiano Näive Bayes (GaussianNB na biblioteca sklearn.naive_bayes do scikit-learn, em linguagem python).

3.2.2 Regressão Logística

A Regressão Logística está entre os algoritmos de regressão que também podem ser utilizados para classificação. É comumente utilizada para estimar a probabilidade de uma instância pertencer a determinada classe (GÉRON, 2019), ou seja, é utilizada para a análise de dados multivariados envolvendo respostas binárias - assim como o objetivo deste estudo.

Utilizamos o algoritmo classificador Regressão Logística (LogisticRegression na biblioteca sklearn.linear_model do scikit-learn.

3.2.3 Árvore de Decisão

Uma árvore de decisão usa a estratégia de dividir para conquistar para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. As soluções dos subproblemas podem ser combinadas, na forma de uma árvore, para produzir uma solução do problema complexo (CARVALHO et al., 2011).

Na árvore de decisão, o atributo mais importante é apresentado como o primeiro nó e os atributos menos relevantes são mostrados nos nós subseqüentes (LEMOS; STEINER; NIEVOLA, 2005).

Nos nossos estudos utilizamos o algoritmo classificador de Árvore de Decisão (DecisionTreeClassifier na biblioteca sklearn.tree do scikit-learn). O scikit-Learn utiliza o algoritmo de treinamento CART para treinar árvores de decisão.

O funcionamento do algoritmo é relativamente simples: ele primeiro divide o conjunto de treinamento em dois subconjuntos (busca pelo par mais puro, ponderados pelo tamanho). Em seguida, ele divide os subconjuntos utilizando a mesma lógica, depois os subconjuntos e assim por diante, recursivamente (GÉRON, 2019).

Para Lemos, Steiner e Nievola (2005), a vantagem principal das árvores de decisão é a tomada de decisões levando em consideração os atributos mais relevantes, além de compreensíveis para a maioria das pessoas. Ao escolher e apresentar os atributos em ordem de importância, as árvores de decisão permitem aos usuários conhecer quais fatores mais influenciam os seus trabalhos.

3.2.4 Random Forest

Trata-se de uma técnica de aprendizagem formada por um conjunto de previsores (*Ensemble Learning*), especificamente, um classificador que consiste em uma coleção de árvores de decisão, estruturadas em k vetores aleatórios, independentes e identicamente distribuídos, onde cada árvore lança um voto unitário para a classe mais popular na entrada x (BREIMAN, 2001).

Para Géron (2019), o random forest (ou Floresta Aleatória, em português) é um dos mais poderosos algoritmos de aprendizagem de máquina disponível atualmente, apesar da sua simplicidade.

Utilizamos o algoritmo classificador Floresta Aleatória (RandomForestClassifier na biblioteca sklearn.ensemble do scikit-learn).

4 Análises e Resultados

4.1 Execução dos Algoritmos

Para a execução dos algoritmos, dividimos a base de dados em duas partes: um conjunto de dados para treinamento (85%) e um conjunto de dados para teste (15%).

Nesse cenário, dada a divisão[1] dos dados entre treinamento e teste aplicamos o balanceamento da base (SMOTE) apenas no conjunto de treinamento. O conjunto de dados de teste permaneceu original para melhor simular uma aplicação real, que de maneira geral, diversifica a quantidade de lançamentos em cada classe.

A acurácia obtida em cada um dos algoritmos na base de teste está descrita na tabela 4

Algoritmo	**Acurácia**
Näives Bayes	96,62%
Regressão Logística	98,29%
Árvore de Decisão	99,99%
Random Forest	99,99%

Tabela 4 – Acurácia Algoritmos

Embora os resultados aparentem positivos, visto que todos ultrapassaram os 96% de acurácia, a matriz de confusão gerada para cada execução demonstra a dificuldade do algoritmo Näive Bayes na identificação da classe 1, assim como um pequeno erro do algoritmo de Regressão Logística na atribuição de classe 1 a lançamentos da classe 0 (3).

Além disso, podemos verificar no relatório de classificação o diminuto *F1-Score* atribuído ao algoritmo Näive Bayes. A métrica *F1-Score* trata-se de uma média harmônica entre a precisão e a sensibilidade (*recall*[2]).

Classificar um lançamento de perda operacional como um evento com possível fronteira direcionará o dado à avaliação de um especialista em risco operacional, o que consequentemente o levaria à desconsiderá-lo da classe 1. Em contraponto, um lançamento de perda operacional com possibilidade de impacto no risco de crédito se classificado como 0, não seria reavaliado. Nesse contexto, a classificação de um evento classe 1 como classe 0 tem um custo muito maior à instituição e à qualificação da base de dados do que um evento classe 0 ser classificado como classe 1.

[1] para divisão da base em treinamento e teste utilizamos o algoritmo train_test_split da biblioteca sklearn.model_selection.

[2] *recall* é a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas.

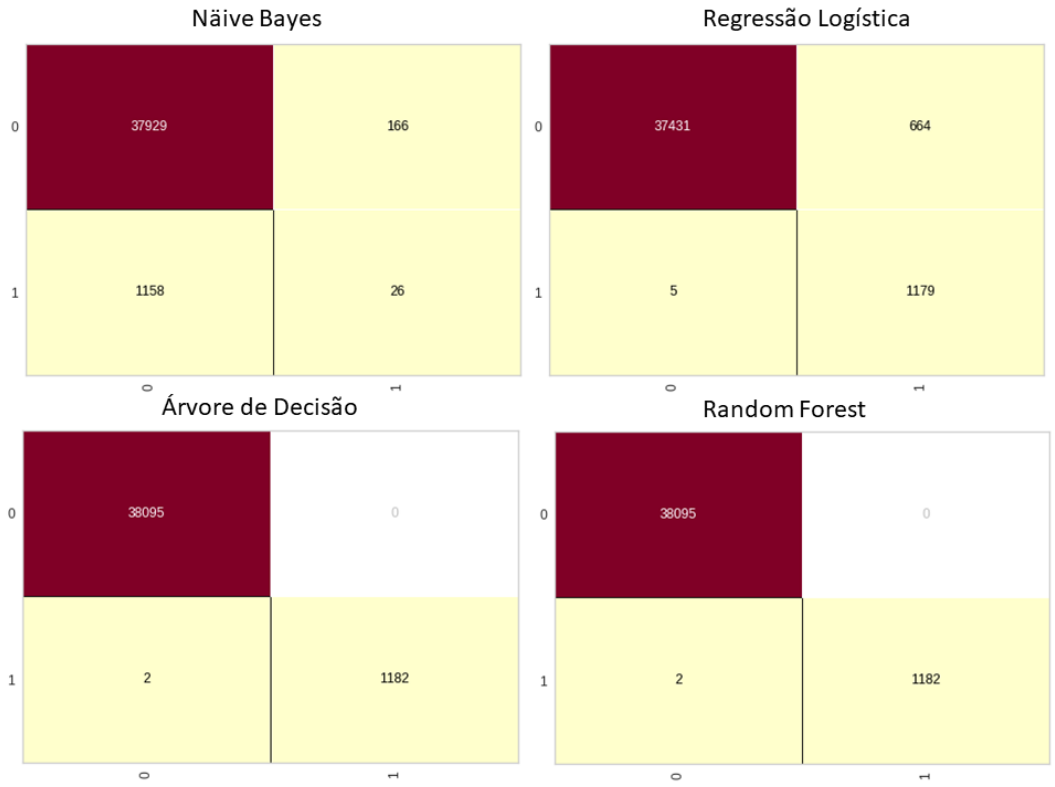

Figura 3 – Matriz de Confusão - Algoritmos

Näive Bayes			
Class	Precision	Recall	F1-Score
0	0,97	1	0,98
1	0,14	0,02	0,04

Regressão Logística			
Class	Precision	Recall	F1-Score
0	1	0,98	0,99
1	0,64	1	0,78

Árvore de Decisão			
Class	Precision	Recall	F1-Score
0	1	1	1
1	1	1	1

Random Forest			
Class	Precision	Recall	F1-Score
0	1	1	1
1	1	1	1

Figura 4 – Relatório de Classificação

Os algoritmos Árvore de Decisão e Random Forest se comportaram de maneira semelhante, obtendo 99,99% de acurácia e 100% em precisão e sensibilidade. Nota-se na matriz de confusão desses algoritmos (figura 3), que foram classificados apenas 2 lançamentos como classe 1, embora pertencentes à classe 0. No Random Forest utilizamos 5 árvores (*n estimators*) e seus resultados foram praticamente idênticos àqueles obtidos com uma única árvore de decisão. Nesse contexto, inclusive por conta de eventual limitação computacional e eficiência operacional, o algoritmo Árvore de Decisão pode ser melhor aplicado devido às características dos dados.

4.2 Validação Cruzada

Após a execução dos algoritmos (descrita na seção anterior), utilizamos a validação cruzada em K-*folds* (onde, K=10) para ratificar os resultados obtidos.

Na validação cruzada, a base de dados é separada continuamente em dois conjuntos distintos onde $\frac{9}{10}$ dos dados são utilizados para treino e $\frac{1}{10}$ dos dados para teste. Em teoria, a cada vez que a validação cruzada é realizada o algoritmo é testado 10 vezes, visto que o resultado de cada execução é a média dos resultados parciais obtidos em cada uma das 10 partes. Considerando que a validação cruzada divide a base de dados entre treinamento e teste de forma aleatória, para este cenário, aplicamos o balanceamento da classe fronteira (SMOTE) no conjunto de dados (diferente do executado anteriormente, onde somente a base de treinamento fora balanceada).

Como executamos a validação cruzada 10 vezes, podemos afirmar que cada algoritmo foi executado 100 vezes (cada execução em 10-*folds*). Os resultados de cada uma das execuções estão dispostos na tabela 5.

Seq	**Acur.**	**Alg.**	**Seq**	**Acur.**	**Alg.**	**Seq**	**Acur.**	**Alg.**	**Seq**	**Acur.**	**Alg.**
0	0.51026	NB	10	0.99997	AD	20	0.99999	RF	30	0.99153	RL
1	0.51025	NB	11	0.99998	AD	21	0.99999	RF	31	0.99068	RL
2	0.51027	NB	12	0.99998	AD	22	0.99998	RF	32	0.99050	RL
3	0.51022	NB	13	0.99997	AD	23	0.99999	RF	33	0.99233	RL
4	0.51030	NB	14	0.99998	AD	24	0.99999	RF	34	0.99081	RL
5	0.51030	NB	15	0.99997	AD	25	0.99998	RF	35	0.99328	RL
6	0.51032	NB	16	0.99998	AD	26	0.99999	RF	36	0.99118	RL
7	0.51027	NB	17	0.99997	AD	27	0.99999	RF	37	0.99319	RL
8	0.51031	NB	18	0.99998	AD	28	0.99999	RF	38	0.99059	RL
9	0.51026	NB	19	0.99997	AD	29	0.99999	RF	39	0.99066	RL

Tabela 5 – Resultados

Em termos de distribuição, verificamos que os resultados obtidos com cada um dos algoritmos apresenta um curva normal de distribuição (figura 5).

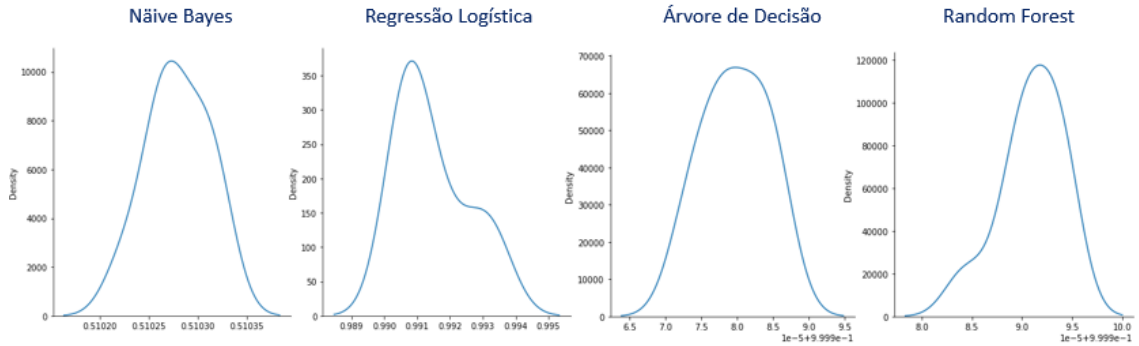

Figura 5 – Distribuição - Resultados Algoritmos

A figura 6 apresenta a correlação dos resultados obtidos com os algoritmos:

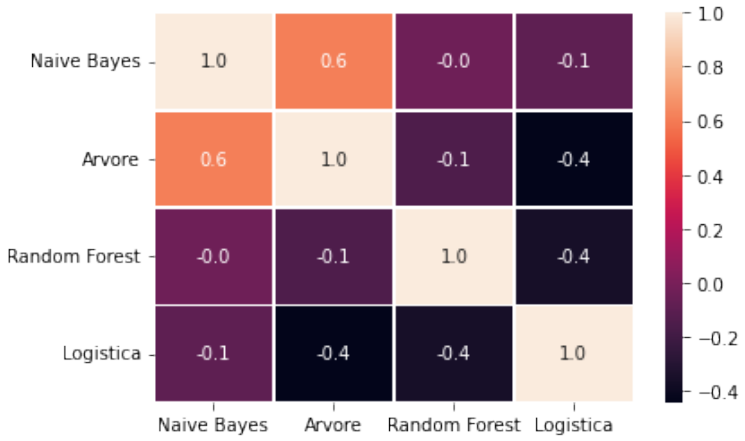

Figura 6 – Matriz de Correlação

A figura 7 demonstra a comparação entre os algoritmos considerando o resultado de cada um na validação cruzada. Notamos que os resultados se assemelham àqueles obtidos na seção anterior (4.1) inclusive em termos de semelhança entre o desempenho dos algoritmos Árvore de Decisão e Random Forest. Assim como no caso anterior, o algoritmo Regressão Logística apresentou um resultado satisfatório, embora inferior àqueles obtidos nos algoritmos que utilizam a métrica de Árvores de Decisão. E novamente o algoritmo Näive Bayes apresentou um resultado inferior aos demais.

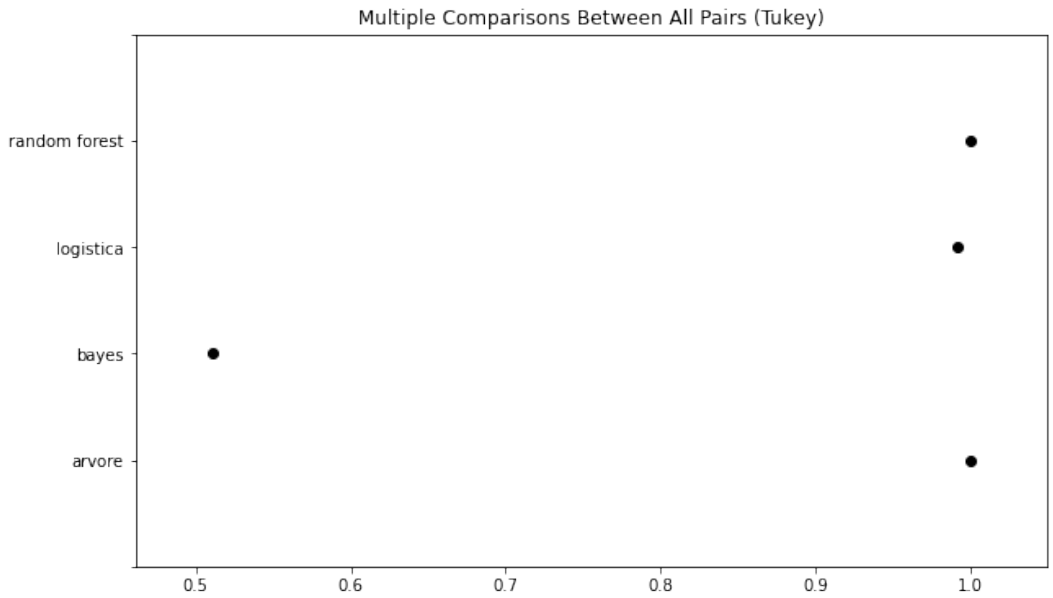

Figura 7 – Múltipla Comparação entre Pares

Os resultados obtidos com os algoritmos de aprendizagem de máquina de Regressão Logística, Árvore de Decisão e Random Forest demonstraram-se satisfatórios para a classificação de eventos propensos à fronteira entre os riscos operacional e de crédito.

5 Conclusão

A aplicação dos modelos aplicados neste estudo para classificação de eventos de risco operacional com provável impacto no risco de crédito representa melhoria na rotina atualmente executada pela instituição, ainda que necessite de tratamento posterior por analistas de risco para verificação dos eventos sinalizados com o atributo fronteira. Mesmo envolvendo a atuação humana de um especialista, o uso das técnicas aplicadas neste estudo reduzirá consideravelmente os eventos para análise. Na base utilizada, por exemplo, é a diferença entre avaliar 261.858 lançamentos ou apenas os 7.829 lançamentos sinalizados com fronteira (classe 1).

Entendemos que uma base de perdas operacionais que contemple dados mais específicos dos contratos ou mesmo dados de clientes para uma vinculação direta ao risco de crédito apresentará ganho nos resultados das técnicas aplicadas neste estudo. O cruzamento de dados de perdas de risco operacional e base de inadimplência de operações de crédito (contratos em *default*) tende a possibilitar uma classificação efetiva do risco de fronteira. Inclusive, para a identificação do percentual de lançamento que deva ser vinculado a cada tipo de risco - a real perda operacional e a consecutiva perda com a inadimplência da contraparte. Um modelo de classificação contemplando ambos os dados reduzirá a necessidade de avaliação humana dos resultados obtidos neste estudo.

Os resultados apresentados no capítulo 4 confirmam que a implementação de métodos de aprendizagem de máquina para classificação da base de perdas operacionais pode representar ganho e eficiência operacional na escala das avaliações especialistas (humana).

Se consolidaram entre os algoritmos com melhores resultados as técnicas de Árvore de Decisão e Random Forest. Mediante resultados tão similares e observando as relações de custo X benefício, a aplicação de modelo com a execução de Árvore de Decisão pode vir a representar ganho em eficiência operacional e agilidade de resposta, ainda que o Random Forest seja considerado, geralmente, um algoritmo mais evoluído.

Referências

BANCO CENTRAL DO BRASIL. *Resolução nº 3.380, de 29 de junho de 2006*. [S.l.], 2006. Disponível em: <https://www.bcb.gov.br/pre/normativos/res/2006/pdf/res_3380_v2_L.pdf>. Citado na página 25.

BANCO CENTRAL DO BRASIL. *Resolução nº 3.721, de 30 de abril de 2009*. [S.l.], 2009. Disponível em: <https://www.bcb.gov.br/pre/normativos/res/2006/pdf/res_3380_v2_L.pdf>. Citado na página 25.

BANCO CENTRAL DO BRASIL. *Circular nº 3.647, de 4 de março de 2013*. [S.l.], 2013. Disponível em: <https://ww.bcb.gov.br/htms/Normativ/CIRCULAR3647.pdf>. Citado na página 26.

BANCO CENTRAL DO BRASIL. *CARTA CIRCULAR Nº 3.841, DE 14 DE SETEMBRO DE 2017*. [S.l.], 2017. Disponível em: <https://www.bcb.gov.br/content/estabilidadefinanceira/especialnor/CartaCircular3841.pdf>. Citado 2 vezes nas páginas 25 e 30.

BANCO CENTRAL DO BRASIL. *Resolução nº 4.557, de 23 de fevereiro de 2017*. [S.l.], 2017. Disponível em: <https://www.bcb.gov.br/pre/normativos/busca/downloadNormativo.asp?arquivo=/Lists/Normativos/Attachments/50344/Res_4557_v2_P.pdf>. Citado na página 27.

BANCO CENTRAL DO BRASIL. *Circular nº 3.979, de 30 de janeiro de 2020*. [S.l.], 2020. Disponível em: <https://www.bcb.gov.br/pre/normativos/busca/downloadNormativo.asp?arquivo=%2FLists%2FNormativos%2FAttachments%2F50913%2FCirc_3979_v1_O.pdf>. Citado 2 vezes nas páginas 27 e 29.

BANK FOR INTERNATIONAL SETTLEMENTS. *Basel Committee on Banking Supervision - Observed range of practice in key elements of Advanced Measurement Approaches (AMA)*. [S.l.], 2006. Disponível em: <https://www.bis.org/publ/bcbs131.pdf>. Citado na página 26.

BARBOZA, F.; KIMURA, H.; ALTMAN, E. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, Elsevier, v. 83, p. 405–417, 2017. Citado na página 31.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 34.

CARVALHO, A. et al. Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, p. 45, 2011. Citado 2 vezes nas páginas 32 e 33.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *The Journal of artificial intelligence research*, AI Access Foundation, San Francisco, v. 16, p. 321–357, 2002. ISSN 1076-9757. Citado 2 vezes nas páginas 30 e 31.

CHEN, R.-C.; CHEN, T.-S.; LIN, C.-C. A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and*

Artificial Intelligence, World Scientific, v. 20, n. 02, p. 227–239, 2006. Citado na página 31.

CHEN, Z. et al. Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, Springer, v. 57, n. 2, p. 245–285, 2018. Citado na página 31.

CHERNOBAI, A.; JORION, P.; YU, F. The determinants of operational risk in us financial institutions. *Journal of Financial and Quantitative Analysis*, Cambridge University Press, v. 46, n. 6, p. 1683–1725, 2011. Citado 2 vezes nas páginas 23 e 24.

CHERNOBAI, A.; OZDAGLI, A.; WANG, J. Business complexity and risk management: Evidence from operational risk events in u.s. bank holding companies. *Journal of monetary economics*, Elsevier B.V, v. 117, p. 418–440, 2021. ISSN 0304-3932. Citado na página 28.

FEUERVERGER, A. On goodness of fit for operational risk. *International Statistical Review*, Wiley Online Library, v. 84, n. 3, p. 434–455, 2016. Citado na página 25.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. [S.l.]: Alta Books, 2019. Citado 2 vezes nas páginas 33 e 34.

JONGH, E. D. et al. A review of operational risk in banks and its role in the financial crisis. *South African Journal of Economic and Management Sciences*, AOSIS, v. 16, n. 4, p. 364–382, 2013. Citado 2 vezes nas páginas 29 e 30.

KO, C.; LEE, P.; ANANDARAJAN, A. The impact of operational risk incidents and moderating influence of corporate governance on credit risk and firm performance. *International Journal of Accounting & Information Management*, Emerald Publishing Limited, 2019. Citado 2 vezes nas páginas 23 e 24.

KOTSIANTIS, S. Supervised machine learning: a review of classification techniques. *Informatica (Ljubljana)*, Slovenian Society Informatika, v. 31, n. 3, p. 249, 2007. ISSN 0350-5596. Citado na página 32.

LEMOS, E. P.; STEINER, M. T. A.; NIEVOLA, J. C. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. *Revista de Administração-RAUSP*, Universidade de São Paulo, v. 40, n. 3, p. 225–234, 2005. Citado na página 33.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 32.

PEREIRA, F.; MITCHELL, T.; BOTVINICK, M. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage (Orlando, Fla.)*, Elsevier Inc, United States, v. 45, n. 1, p. S199–S209, 2009. ISSN 1053-8119. Citado na página 32.

PSILLAKI, M.; TSOLAS, I. E.; MARGARITIS, D. Evaluation of credit risk based on firm performance. *European journal of operational research*, Elsevier, v. 201, n. 3, p. 873–881, 2010. Citado na página 25.

SÁ, A. G. de; PEREIRA, A. C.; PAPPA, G. L. A customized classification algorithm for credit card fraud detection. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 72, p. 21–29, 2018. Citado 2 vezes nas páginas 31 e 32.

SHAH, S. A. A. Integration of financial risks with non financial risks: an exploratory study from pakistani context. *Copernican Journal of Finance & Accounting*, v. 8, n. 2, p. 49–65, 2019. Citado na página 24.

SHEN, A.; TONG, R.; DENG, Y. Application of classification models on credit card fraud detection. In: IEEE. *2007 International conference on service systems and service management*. [S.l.], 2007. p. 1–4. Citado na página 31.

SONG, X.-P. et al. Application of machine learning methods to risk assessment of financial statement fraud: evidence from china. *Journal of Forecasting*, Wiley Online Library, v. 33, n. 8, p. 611–626, 2014. Citado na página 31.

SUSEP. *Padrões para o Reporte de Perdas Operacionais no BDPO - Orientações da Susep ao Mercado, de 6 de agosto de 2014*. [S.l.], 2014. Disponível em: <http://www.susep.gov.br/setores-susep/cgsoa/coris/requerimentos-de-capital/arquivos/Padroes%20para%20o%20Reporte%20de%20Perdas%20Operacionais.pdf>. Citado na página 25.

WANG, D.; CHEN, B.; CHEN, J. Credit card fraud detection strategies with consumer incentives. *Omega*, Elsevier, v. 88, p. 179–195, 2019. Citado na página 31.

ZHANG, H. The optimality of naive bayes. *Aa*, v. 1, n. 2, p. 3, 2004. Citado na página 32.

Apêndices

APÊNDICE A – Script Python

A.1 Tratamento dos Dados - Execução Algoritmos

```
import pandas as pd
import numpy as np
import seaborn as sns

base_perda = pd.read_excel('/content/base_RO_py.xlsx')

#definindo atributos previsores
x_base = base_perda.iloc[:, 1:7].values

#definindo atributo classe
y_base = base_perda.iloc[:, 7].values

#converter string para números
from sklearn.preprocessing import LabelEncoder

label_encoder_Nome_do_Evento = LabelEncoder ()
label_encoder_Nome_Produto = LabelEncoder ()
label_encoder_N1_RISCO = LabelEncoder ()
label_encoder_N2_RISCO = LabelEncoder ()
label_encoder_N3_RISCO = LabelEncoder ()

x_base[:,0] = label_encoder_Nome_do_Evento.fit_transform(x_base [:,0])
x_base[:,1] = label_encoder_Nome_Produto.fit_transform(x_base [:,1])
x_base[:,3] = label_encoder_N1_RISCO.fit_transform(x_base [:,3])
x_base[:,4] = label_encoder_N2_RISCO.fit_transform(x_base [:,4])
x_base[:,5] = label_encoder_N3_RISCO.fit_transform(x_base [:,5])

#converter valores numéricos em variáveis binárias
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer

OneHotEncoder_base = ColumnTransformer(transformers=[("OneHot",
    OneHotEncoder(),[0,1,3,4,5])],remainder='passthrough')
```

```
x_base = OneHotEncoder_base.fit_transform(x_base).toarray()

#dividir conjunto de dados em treino e teste
from sklearn.model_selection import train_test_split

x_base_treinamento_over, x_base_teste_over, y_base_treinamento_over,
    y_base_teste_over =   train_test_split(x_base, y_base, test_size = 0.15,
    random_state = 0)
x_base_treinamento_over.shape, x_base_teste_over.shape

#balancear dados de treino
from imblearn.over_sampling import SMOTE

smote = SMOTE(sampling_strategy='minority')
x_base_treinamento_over, y_base_treinamento_over = smote.fit_resample
    (x_base_treinamento_over, y_base_treinamento_over)
```

A.2 Näive Bayes

```
from sklearn.naive_bayes import GaussianNB

naive_base_data = GaussianNB()
naive_base_data.fit (x_base_treinamento_over, y_base_treinamento_over)

previsoes = naive_base_data.predict (x_base_teste_over)

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

#resultado acurácia
accuracy_score(y_base_teste_over, previsoes)

#matriz de confusão
from yellowbrick.classifier import ConfusionMatrix
cm = ConfusionMatrix (naive_base_data)
cm.fit(x_base_treinamento_over, y_base_treinamento_over)
cm.score(x_base_teste_over, y_base_teste_over)
```

```
#relatório de classificação
print(classification_report(y_base_teste_over, previsoes))
```

A.3 Regressão Logística

```
from sklearn.linear_model import LogisticRegression

logistic_base = LogisticRegression(random_state = 1)
logistic_base.fit(x_base_treinamento_over, y_base_treinamento_over)

previsoes = logistic_base.predict(x_base_teste_over)

from sklearn.metrics import accuracy_score, classification_report
accuracy_score(y_base_teste_over, previsoes)

from yellowbrick.classifier import ConfusionMatrix
cm = ConfusionMatrix(logistic_base)
cm.fit(x_base_treinamento_over, y_base_treinamento_over)
cm.score(x_base_teste_over, y_base_teste_over)

print(classification_report(y_base_teste_over, previsoes))
```

A.4 Árvore de Decisão

```
from sklearn.tree import DecisionTreeClassifier

arvore_base = DecisionTreeClassifier(criterion='entropy', random_state=0)
arvore_base.fit(x_base_treinamento_over, y_base_treinamento_over)

previsoes = arvore_base.predict (x_base_teste_over)

from sklearn.metrics import accuracy_score, classification_report
accuracy_score(y_base_teste_over, previsoes)

from yellowbrick.classifier import ConfusionMatrix
```

```
cm = ConfusionMatrix(arvore_base)
cm.fit(x_base_treinamento_over, y_base_treinamento_over)
cm.score(x_base_teste_over, y_base_teste_over)

print(classification_report(y_base_teste_over, previsoes))
```

A.5 Random Forest

```
from sklearn.ensemble import RandomForestClassifier

random_forest_base = RandomForestClassifier(n_estimators=5,
    criterion='entropy', random_state=0)
random_forest_base.fit(x_base_treinamento_over, y_base_treinamento_over)

previsoes = random_forest_base.predict (x_base_teste_over)

from sklearn.metrics import accuracy_score, classification_report
previsoes = random_forest_base.predict (x_base_teste_over)
accuracy_score(y_base_teste_over, previsoes)

from yellowbrick.classifier import ConfusionMatrix
cm = ConfusionMatrix(random_forest_base)
cm.fit(x_base_treinamento_over, y_base_treinamento_over)
cm.score(x_base_teste_over, y_base_teste_over)

print(classification_report(y_base_teste_over, previsoes))
```

A.6 Tuning de Parâmetros

```
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
```

```
x_base = np.concatenate((x_base_treinamento_over,
    x_base_teste_over), axis = 0)

y_base = np.concatenate((y_base_treinamento_over,
    y_base_teste_over), axis = 0)

#Árvore de Decisão
parametros = {'criterion': ['gini', 'entropy'],
              'splitter': ['best', 'random'],
              'min_samples_split': [2, 5, 10],
              'min_samples_leaf': [1, 5, 10]}

grid_search = GridSearchCV(estimator=DecisionTreeClassifier(),
    param_grid=parametros)
grid_search.fit(x_base, y_base)
melhores_parametros = grid_search.best_params_
melhor_resultado = grid_search.best_score_
print(melhores_parametros)
print(melhor_resultado)

#Random Forest
parametros = {'criterion': ['gini', 'entropy'],
              'n_estimators': [10, 40, 100, 150],
              'min_samples_split': [2, 5, 10],
              'min_samples_leaf': [1, 5, 10]}

grid_search = GridSearchCV(estimator=RandomForestClassifier(),
    param_grid=parametros)
grid_search.fit(x_base, y_base)
melhores_parametros = grid_search.best_params_
melhor_resultado = grid_search.best_score_
print(melhores_parametros)
print(melhor_resultado)

#Regressão Logística
parametros = {'tol': [0.0001, 0.00001, 0.000001],
              'C': [1.0, 1.5, 2.0],
              'solver': ['lbfgs', 'sag', 'saga']}
```

```
grid_search = GridSearchCV(estimator=LogisticRegression(),
    param_grid=parametros)
grid_search.fit(x_over, y_over)
melhores_parametros = grid_search.best_params_
melhor_resultado = grid_search.best_score_
print(melhores_parametros)
print(melhor_resultado)
```

A.7 Validação Cruzada

```
from sklearn.model_selection import cross_val_score, KFold

resultados_arvore = []
resultados_random_forest = []
resultados_logistica = []
resultados_bayes = []

for i in range(10):
  #print(i)
  kfold = KFold(n_splits=10, shuffle=True, random_state=i)

  bayes = GaussianNB( )
  scores = cross_val_score(bayes, x_base, y_base, cv = kfold)
  print(scores)
  print(scores.mean())
  resultados_bayes.append(scores.mean())

  arvore = DecisionTreeClassifier(criterion='gini',
    min_samples_leaf=1, min_samples_split=2, splitter='random')
  scores = cross_val_score(arvore, x_base, y_base, cv = kfold)
  print(scores)
  print(scores.mean())
  resultados_arvore.append(scores.mean())

  random_forest = RandomForestClassifier(criterion = 'entropy',
    min_samples_leaf = 1, min_samples_split=5, n_estimators = 100)
  scores = cross_val_score(random_forest, x_base, y_base, cv = kfold)
```

```
  resultados_random_forest.append(scores.mean())

  logistica = LogisticRegression(C = 1.0, solver = 'lbfgs', tol = 0.0001)
  scores = cross_val_score(logistica, x_base, y_base, cv = kfold)
  resultados_logistica.append(scores.mean())
```

A.8 Resultados Validação Cruzada

```
resultados = pd.DataFrame({'Naive Bayes': resultados_bayes,
    'Arvore': resultados_arvore, 'Random Forest': resultados_random_forest,
    'Logistica': resultados_logistica})
resultados

#correlação
resultados.corr()

#aplicação teste de shapiro
from scipy.stats import shapiro

shapiro(resultados_bayes), shapiro(resultados_arvore)
    , shapiro(resultados_random_forest), shapiro(resultados_logistica)

#curvas de distribuição
sns.displot(resultados_bayes, kind = 'kde');
sns.displot(resultados_arvore, kind = 'kde');
sns.displot(resultados_random_forest, kind = 'kde');
sns.displot(resultados_logistica, kind = 'kde');

#teste de hipótese
from scipy.stats import f_oneway

_, p = f_oneway(resultados_bayes, resultados_arvore, resultados_random_forest,
    resultados_logistica)
p

alpha = 0.05
if p <= alpha:
```

```
  print('Hipótese nula rejeitada. Dados são diferentes')
else:
  print('Hipótese alternativa rejeitada. Resultados são iguais')

resultados_algoritmos = {'accuracy': np.concatenate([resultados_bayes,
    resultados_arvore, resultados_random_forest, resultados_logistica]),
    'algoritmo': ['bayes', 'bayes', 'bayes', 'bayes', 'bayes', 'bayes',
    'bayes', 'bayes', 'bayes', 'bayes', 'arvore', 'arvore', 'arvore',
    'arvore', 'arvore', 'arvore', 'arvore', 'arvore', 'arvore', 'arvore',
    'random forest', 'random forest', 'random forest', 'random forest',
    'random forest', 'random forest', 'random forest', 'random forest',
    'random forest', 'random forest', 'logistica', 'logistica', 'logistica',
    'logistica', 'logistica', 'logistica', 'logistica', 'logistica',
    'logistica', 'logistica']}

resultados_df = pd.DataFrame(resultados_algoritmos)
resultados_df

#multicomparação
from statsmodels.stats.multicomp import MultiComparison

compara_algoritmos = MultiComparison(resultados_df['accuracy'],
    resultados_df['algoritmo'])

teste_estatistico = compara_algoritmos.tukeyhsd()
print(teste_estatistico)

#gráfico
teste_estatistico.plot_simultaneous();
```

Anexos

ANEXO A – Categorias de Eventos de Risco Operacional

Categorias de Eventos de Risco Operacional	
Categoria Nível 1	**Categoria Nível 2**
Fraudes internas	Atividade não autorizada
	Roubo e fraude (origem interna)
Fraudes externas	Roubo e fraude (origem externa)
	Segurança de sistemas
Demandas trabalhistas e segurança deficiente do local de trabalho	Relações de trabalho
	Segurança do local de trabalho
	Diversidade e discriminação
Práticas inadequadas relativas a clientes, produtos e serviços	Adequação de produto a cliente, divulgação de informações sobre produtos e serviços, desrespeito ao dever fiduciário
	Práticas impróprias de negócios e em mercados
	Falhas no produto
	Seleção, patrocínio e exposição
	Atividades de assessoramento
Danos a ativos físicos próprios ou em uso pela instituição	Desastres e outros eventos
Situações que acarretem a interrupção das atividades da instituição	Interrupção de atividades
Falhas em sistemas, processos ou infraestrutura de tecnologia da informação (TI)	Falhas em sistemas, processos ou infraestrutura de TI
Falhas na execução, no cumprimento de prazos ou no gerenciamento das atividades da instituição	Captura, execução e manutenção de transações
	Monitoramento e reporte
	Aquisição de clientes e documentação
	Gestão de contas correntes e de não correntistas
	Contrapartes em transações
	Representantes e fornecedores

Figura 8 – Categorias de Eventos de Risco Operacional N1 e N2

Fonte: Circular BACEN nº 3.979, de 30 de janeiro de 2020