UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

RODRIGO DOS SANTOS DE MIRANDA

DESENVOLVIMENTO DE UM MODELO PREDITIVO PARA AUXILIAR NA IDENTIFICAÇÃO DE LAVAGEM DE DINHEIRO

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

RODRIGO DOS SANTOS DE MIRANDA

DESENVOLVIMENTO DE UM MODELO PREDITIVO PARA AUXILIAR NA IDENTIFICAÇÃO DE LAVAGEM DE DINHEIRO

DEVELOPMENT OF A PREDICTIVE MODEL TO ASSIST IN THE IDENTIFICATION OF MONEY LAUNDERING

Trabalho de conclusão de curso de Especialização em Ciência de Dados apresentado como requisito para obtenção do título de Especialista em Ciência de Dados da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Prof. Me. Francisco Pereira Junior.

DOIS VIZINHOS

2022

RODRIGO DOS SANTOS DE MIRANDA

DESENVOLVIMENTO DE UM MODELO PREDITIVO PARA AUXILIAR NA IDENTIFICAÇÃO DE LAVAGEM DE DINHEIRO

Trabalho de conclusão de curso de Especialização em Ciência de Dados apresentado como requisito para obtenção do título de Especialista em Ciência de Dados da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 11/novembro/2022

Francisco Pereira Junior
Mestre
Universidade Tecnológica Federal do Paraná

Rosangela de Fátima Pereira Marquesone
Doutora
Universidade Tecnológica Federal do Paraná

Henrique Yoshikazu Shishido
Doutor
Universidade Tecnológica Federal do Paraná

DOIS VIZINHOS

2022

RESUMO

Esse trabalho apresenta o uso de técnicas de *machine learning* para detecção de indícios de lavagem de dinheiro. A partir de estudos de aprendizagem de máquina com algoritmos de aprendizado supervisionado, foi desenvolvido um modelo capaz de identificar padrões de movimentações suspeitas de crime de lavagem de dinheiro em conta corrente de clientes de uma instituição financeira que precisam ser comunicadas aos órgãos competentes para uma maior investigação. A metodologia desenvolvida, seus resultados e sua aplicabilidade foram testados e validados no presente estudo, assim com as sugestões de melhorias futuras do modelo permitindo que tais situações sejam identificadas para tomada de decisão dos responsáveis.

Palavras-chaves: Economia, lavagem de dinheiro, financiamento ao Terrorismo, sistema financeiro nacional, cooperativas de crédito.

ABSTRACT

This work presents the use of machine learning techniques to detect evidence of money laundering. From machine learning studies with supervised learning algorithms, a model was developed capable of identifying patterns of suspicious movements of money laundering crime in the current account of customers of a financial institution that need to be communicated to Organs competent bodies for greater investigation. The methodology developed, its results and its applicability were tested and validated in the present study, as well as suggestions for future improvements to the model, allowing such situations to be identified for decision-making by those responsible.

Keywords: Economy, money laundering, terrorist financing, national financial system, credit unions.

SUMÁRIO

1 INTRODUÇÃO

A ciência de dados é uma área em constante crescimento. Ao longo dos anos, várias áreas de conhecimento passaram a integrar a análise dos dados às suas rotinas com o objetivo de melhorar os resultados de seus objetivos estratégicos.

Entre as diversas áreas do conhecimento que podem prover diferentes aplicações está o sistema financeiro nacional. Nesse contexto, a cada ano que passa, cresce as áreas de Analytics, seja nas equipes do Banco Central do Brasil (BACEN), Conselho de Controle de Atividades Financeiras (COAF) e nas próprias instituições financeiras, buscando soluções que aumentem o diferencial competitivo e promovam controles capazes de acompanhar o crescimento do negócio.

Nesse sentido e com o desenvolvimento da economia a modo global, os sistemas financeiros dos países estão cada vez mais modernos e aumentando o número de transações de recursos. Contudo, todo crescimento traz consigo situações de riscos que precisam ser analisadas e respondidas.

Uma das principais fraudes financeiras que tem preocupado os governantes é a lavagem de dinheiro, que pode causar danos à segurança nacional, ao sistema financeiro e ao desenvolvimento da economia global. Além disso, esse tipo de prática está ligado a crimes como terrorismo, corrupção e tráfico de drogas.

Dessa forma, esse estudo busca unir as áreas da economia e aprendizado de máquina, para obter uma avaliação satisfatória para auxiliar na identificação de casos de lavagem de dinheiro. Para buscar esse resultado é necessário analisar movimentações bancárias e fatores de riscos que elevam a probabilidade de lavagem de dinheiro por criminosos.

A principal motivação deste trabalho é definir um modelo preditivo com o objetivo de detectar indícios de lavagem de dinheiro, com base na legislação vigente no Brasil.

1.1 Objetivos

Esse trabalho tem como objetivo geral utilizar aprendizado de máquina, com base de dados de resultados para prever indícios de lavagem de dinheiro no sistema financeiro nacional.

1.1.1 Objetivos específicos

- Selecionar e pré-processar uma base de dados com movimentações e fatores que indicam lavagem de dinheiro;
- Treinar algoritmos de aprendizado de máquina para predizer os indícios de lavagem de dinheiro;
- Propor e avaliar uma estrutura de dados para suportar o algoritmo de detecção de indícios de lavagem de dinheiro;
- Verificar se o resultado do modelo desenvolvido é capaz de solucionar o problema ou se será necessário adicionar mais variáveis para ter um resultado satisfatório.

1.1.2 Justificativa

O presente estudo propôs a investigar o tema em razão de sua relação intrínseca com crimes de maior potencial ofensivo, tais como: tráfico de drogas e armas, terrorismo e a corrupção do sistema econômico nacional nas instituições controladas pelo governo. Diante de tal panorama, de potenciais crimes que abalam de forma sistêmica a sociedade, é necessário dentro do sistema financeiro nacional criar formas de impedir a movimentação de recursos provenientes de origem ilícita. Dito isto, o conhecimento adquirido nesse estudo poderá ser aplicado e compartilhado para várias instituições financeiras nacionais para identificação de possíveis crimes contra o sistema financeiro nacional, uma vez que o comportamento das operações financeiras desses usuários tem um padrão a ser seguido e pode ser prevenido.

1.1.3 Estrutura deste documento

A seção 2 traz uma fundamentação teórica buscando contextualizar como a lavagem de dinheiro ocorre dentro do sistema financeiro nacional. Além de apresentar o desenvolvimento do referencial metodológico *Cross Industry Standart Process for Data Mining* (CRISP-DM) aplicado neste trabalho. E, por fim, um comparativo com outros trabalhos com a mesma temática.

A seção 3 apresenta o desenvolvimento do estudo e a preparação dos dados. A seção 4 demonstra como foi realizada a modelagem e metodologia utilizada para a busca do atingimento dos objetivos propostos. A seção 5 apresenta as conclusões, os resultados alcançados e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A lavagem de dinheiro é o meio pelo qual os criminosos procuram introduzir um bem, direito ou valor de origem ilícita na atividade econômica legal com aparência de lícito. Essa prática consiste em utilizar o sistema financeiro do país através de transações bancárias ou transações comerciais para ocultar a origem do bem. Para regulamentar o crime de lavagem de dinheiro no Brasil foi criada a Lei n. 9.613, de 03/03/1998[1]. Ela é utilizada para definir o crime, condenações e responsabilidades, e específica quem deve implantar mecanismos de controle para identificar os indícios dessa fraude.

Dessa forma, mais especificamente o Banco Central do Brasil (BACEN), aplicou uma regulamentação sobre as instituições financeiras brasileiras através da circular n. 3.978, de 23/01/2020[2] e a carta circular n. 4.001, de 29/01/2020[3], a fim de estabelecer procedimentos a serem realizados por essas instituições, com o objetivo de combater a lavagem de dinheiro, com a definição de regras para identificação de indícios de crimes que possam ocorrer no ambiente de instituição financeira. Quando esses indícios são identificados, eles devem ser comunicados ao Conselho de Controle de Atividades Financeiras (COAF).

Segundo o órgão regulador (BACEN) existem alguns fatores que aumentam o risco de lavagem de dinheiro, eles estão elencados na carta circular 4.001/2020.

O crime de lavagem de dinheiro é um tema que ganhou bastante notoriedade ultimamente, apesar de já fazer parte da história do sistema financeiro nacional. Com os recentes casos de corrupção e diversas operações fraudulentas, esse tema vem sendo noticiado cada vez mais pela mídia, o que tornou o tema mais conhecido. Com esse recente ganho midiático, muito se fala em aumento dos controles e mecanismo de prevenção de tais operações por meio das instituições financeiras. Visto isso, há uma necessidade muito grande de usar os dados do

---

[1] Para as regras gerais de apresentação das citações consultar: BRASIL. Presidência da República. Lei n° 9.613, de 3 de março de 1998. Disponível em: L9613 (planalto.gov.br)

[2] Para as regras gerais de apresentação das citações consultar: BANCO CENTRAL DO BRASIL. Circular nº 3.978 de 23 de janeiro de 2020. Disponível em: Circular Nº 3.978 (bcb.gov.br)

[3] Para as regras gerais de apresentação das citações consultar: BANCO CENTRAL DO BRASIL. Carta Circular nº 4.001 de 29 de janeiro de 2020. Disponível em: Minuta (bcb.gov.br)

sistema financeiro nacional para criar modelos capazes de prevenir ou descobrir esse tipo de movimentação atípica por parte dos clientes.

O processo de lavagem de dinheiro geralmente possui três fases independentes, conforme Figura 1.

Figura 1 - Fases da lavagem de dinheiro

FONTE: Adaptado de Conselho de Controle de Atividades Financeiras (2014)

A primeira fase, denominada **colocação**, busca encobrir a origem ilícita do dinheiro. Nessa fase, o criminoso busca inserir o dinheiro ilícito na economia formal. (PITOMBO, 2003).

Na sequência, ocorre a fase de **ocultação**, cujo objetivo é dar legalidade ao bem provido de crimes, dificultando, assim, o rastreamento contábil desses recursos (VILARDI, 2004). Para dar essa legalidade, o criminoso movimenta o dinheiro, geralmente de forma eletrônica, para contas de pessoas físicas que fornecem seu nome para ocultar o destinatário do dinheiro ou por empresas de fachada.

Por fim, a **integração** é a fase em que o capital, já com características legais, é utilizado para aquisição de ativos em geral, tais como: imóveis, ações, veículos, embarcações etc. Esses ativos geralmente são utilizados em suas atividades criminosas a fim de facilitar e ampliar a sua prática (CALLEGARI, 2000).

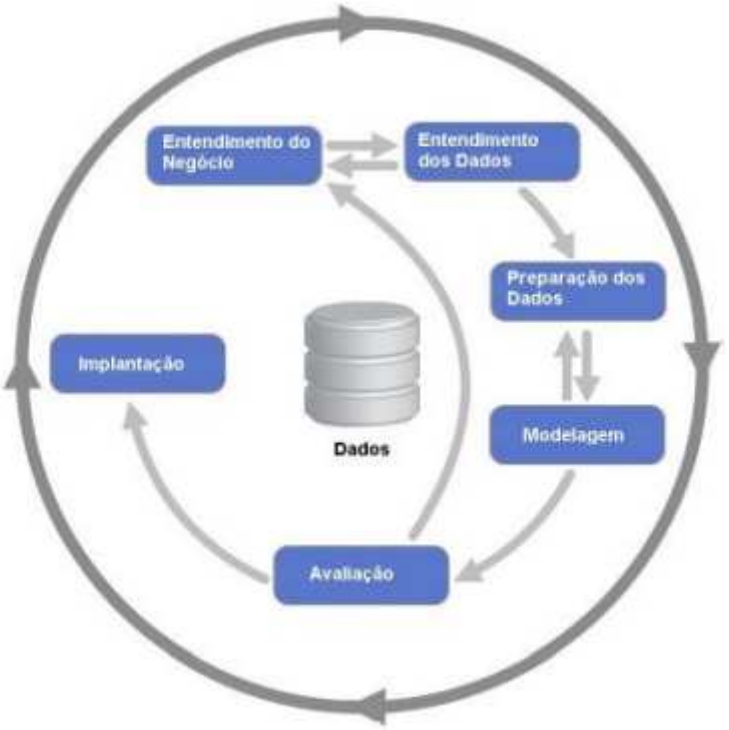
Portanto, ao final dessas três fases, pode-se concluir que o dinheiro, adquirido de forma ilícita, está com a aparência de lícito e distante de sua origem, dificultando, assim, a detecção de seus autores.

2.1 Modelo de referência CRISP-DM

Uma das metodologias utilizadas por cientistas de dados para demonstrar a mineração de dados é o *Cross Industry Standard Process for Data Mining* (CRISP-DM). Esse processo define uma hierarquia que começa com o entendimento do negócio, passando ao entendimento dos dados, preparação dos dados e

modelagem, uma fase de avaliação e, por fim, a implantação do processo. A exemplificação dessa metodologia está representada através da Figura 2.

Figura 2 - Fases do CRISP-DM



***Fonte: CRISP-DM* (2000)**

O detalhamento das fases foi baseado no guia de mineração de dados do CRISP-DM [4].

Entendimento do negócio – esta fase objetiva ter um claro entendimento do que se pretende a partir da mineração de dados e como os resultados alcançados se parecerão em termos dos processos de negócios que serão beneficiados.

Entendimento dos dados – esta fase parte de uma coleta inicial dos dados, seguida de atividades que possibilitem a familiarização com seu conjunto, a identificação de problemas de qualidade e a descoberta de *insights* dentro dos dados, que permitam a formulação de hipóteses para informações que não estejam aparentes.

[4] Para as regras gerais de apresentação das citações consultar: ***CRISP-DM 1.0: Step-by-step data mining guide*.** Disponivel em: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Preparação dos dados – a fase de preparação de dados é constituída de atividades que visam a construção, a partir dos dados brutos iniciais, do conjunto de dados final. Tarefas de preparação de dados não possuem uma ordem prescrita e são susceptíveis de serem realizadas várias vezes. Essas tarefas incluem a seleção de tabela, registro e atributo, bem como transformação e limpeza de dados para ferramentas de modelagem.

Modelagem – nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos específicos sobre a forma dos dados. Portanto, voltar à fase de preparação de dados é muitas vezes necessário.

Avaliação – esta fase do projeto se dá após a construção de um modelo (ou modelos) que, a partir de uma perspectiva de análise de dados, aparenta ter alta qualidade. Para se ter certeza de que o modelo atinge adequadamente os objetivos de negócios, antes de proceder à implantação final do modelo, é importante avaliá-lo cuidadosamente e rever as etapas executadas para criá-lo. Um dos principais objetivos aqui é o de determinar se existe alguma questão de negócio importante que não tenha sido suficientemente considerada. No final desta fase uma decisão sobre o uso dos resultados de mineração de dados deve ser alcançada.

Implantação – esta fase geralmente envolve a aplicação de modelos ‘ao vivo’ dentro dos processos de tomada de decisão de uma organização. A complexidade desta fase depende dos requisitos, podendo ser tão simples quanto gerar um relatório ou tão complexa como implementar um processo de mineração de dados repetível em toda a empresa. Em muitos casos esta fase não é executada pelo analista de dados. No entanto, mesmo se o analista realizar o esforço de implantação, é importante que o cliente compreenda as ações que precisam ser realizadas para realmente usar os modelos criados.

2.2 Trabalhos Correlatos

Trabalhos utilizando aprendizado de máquina e modelos preditivos têm sido amplamente utilizados na detecção de transação com indícios de lavagem de dinheiro, especialmente quando utilizam reconhecimento de padrões. Dessa forma,

é possível notar que alguns trabalhos similares já foram desenvolvidos nos últimos anos.

O trabalho apresentado por Luo (2014) também propõe um algoritmo de classificação, com a finalidade de detectar transações suspeitas. Essa pesquisa possui duas fases: na primeira acontece a extração das regras mais frequentes por meio da base de treinamento; no segundo estágio, ocorre a classificação das transações, onde como resultado o algoritmo apresentou 317 contas destacadas como suspeitas em um universo de mais de 100 milhões de transações, com uma taxa de precisão de mais de 80%.

O trabalho apresentado por Socreppa (2016) propõe a utilização do algoritmo baseado em árvore de decisão e o processo de KDD para classificar transações com indícios de lavagem de dinheiro se baseando na lei em vigor na época do estudo. O algoritmo teve resultados satisfatórios na detecção de lavagem de dinheiro uma vez que a eficácia dos algoritmos foi superior a 0,99. Todavia, esse modelo apresenta fragilidades pois é baseado na legislação vigente e com a mudança da legislação o modelo necessita de ajustes. Assim como esse estudo, o método proposto aqui foi realizado utilizando uma base de dados real.

O estudo realizado por Suresh, Reddy e Sweta (2016) propõem a utilização de um método que faz uso de um algoritmo *Hash* para compactar os dados e diminuir o caminho para chegar no resultado. Aliado a esse algoritmo os autores utilizam o algoritmo de *Apriori* para realizar as associações e detectar contas suspeitas de lavagem de dinheiro. Os algoritmos tiveram sucesso em encontrar as contas com maior probabilidade de transações suspeitas.

O trabalho desenvolvido por Borba (2017) propõe realizar um escore de risco para classificar transações suspeitas de lavagem de dinheiro através de um algoritmo de regressão logística ordinal, com uma base de dados amostral com mais de 80.000 observações, a utilização de qualquer técnica teria uma alta significância, portanto foi realizada um balanceamento dos dados para treinar o modelo que tinha uma taxa de acerto geral de 79,71%. Com isso, o resultado foi de 9,22% dos casos considerados como risco de lavagem de dinheiro.

Por fim, o trabalho apresentado por Pacheco Junior (2019) propõe a utilização de algoritmos baseados em classificação, sendo que os resultados foram muito semelhantes entre eles. Para verificar qual algoritmo teve o melhor desempenho foi utilizado o indicador de curva ROC, e o algoritmo que apresentou

maior destaque foi o *Random Forest*. Esse trabalho utilizou uma base de dados real e teve resultados satisfatórios, conseguiu apresentar um modelo capaz de prever fraudes com a utilização de aprendizado de máquina.

O modelo proposto por este estudo será usar o modelo de árvore de decisão, utilizando uma base de dados real, baseando se na legislação em vigor para identificar fatores de risco que podem influenciar a decisão de comunicar um fato suspeito de lavagem de dinheiro, além de utilizar fatores ligados aos padrões de movimentações realizadas pelos clientes.

A Tabela 1 apresenta um resumo comparativo entre os trabalhos correlatos com a presente pesquisa, de acordo com os critérios a seguir:

- Detecção: este critério tem o objetivo de demonstrar o alvo de detecção, seja por movimentações financeiras ou contas correntes.
- Tarefa: apresenta se o modelo utilizou ou não aprendizado supervisionado.
- Algoritmo: objetiva qual é o algoritmo utilizado pelo modelo.
- Regras do País: demonstra se as formas de detecção de lavagem de dinheiro levaram em conta as regras do órgão fiscalizador.
- Base de dados: indica a utilização de uma base de dados real ou simulada.

Tabela 1 - Comparação entre os trabalhos correlatos

	Luo (2014)	Socreppa (2016)	Suresh, Reddy e Sweta (2016)	Borba, Maria Clara Vieira (2017)	Pacheco Junior (2019)	Método Proposto
FORMA DE DETECÇÃO	Transações	Transações	Contas	Contas	Transações	Transações
TIPO DE APRENDIZADO DE MÁQUINA	Classificação	Classificação	Associação	Classificação	Classificação	Classificação
ALGORITMO	Classificação baseada no FP-TREE	Árvore de Decisão	Hash Based association	Regressão Logísitica Ordinal	Random Forest	Árvore de Decisão
REGRAS DO PAÍS	Não	Sim	Não	Sim	Não	Sim
BASE DE DADOS	Simulada	Real	Simulada	Real	Real	Real

FONTE: O autor (2022)

3 DESENVOLVIMENTO

Para prevenir o crime de lavagem de dinheiro as instituições financeiras precisam analisar o perfil das movimentações financeiras realizadas por seus clientes, com o intuito de verificar atipicidades na movimentação, para tal é necessário criar métricas capazes de evidenciar quais movimentações têm maior risco e necessitam uma análise mais aprofundada.

Este estudo propõe-se uma abordagem fundamentada em 25 fatores, onde 20 fatores são baseados na movimentação dos clientes, tendo como base a carta circular 4.001/2020 do BACEN. E, outros 5 fatores, baseados em informações cadastrais do cliente que ajudarão para a tomada de decisão do modelo.

- **Tipo de Pessoa** – este fator informa se o cliente é pessoa física ou pessoa jurídica.
- **Idade / tempo de constituição** – este fator apresenta a idade das pessoas físicas ou o tempo de constituição das pessoas jurídicas.
- **Tempo de relacionamento** – este fator apresenta o tempo de relacionamento dos clientes com a instituição financeira.
- **Pessoa exposta politicamente (PEP)** – este fator apresenta se os clientes são ou não pessoas expostas politicamente.
- **Mídias negativas** – este fator apresenta se os clientes têm algum registro de mídias com informações atrelados a algum crime.

A Tabela 2, elenca quais foram os fatores utilizados baseados na movimentação dos clientes, com objetivo de melhorar a classificação por parte do algoritmo.

Tabela 2 - Listagem dos Fatores

Fatores	Título dos Fatores	Objetivo dos Fatores
1	Depósitos fragmentados	Verificar se há depósitos fragmentados com soma superior a R$ 50.000,00 no mesmo dia.
2	Atipicidade nas distâncias de transações com TEDs	Verificar se há indícios de atipicidades em transações com TED's para clientes classificados com mesmo padrão.
3	Análise das distâncias de transações com TEDs - PEPs	Verificar se há indícios de atipicidades em distâncias das transações com TED's para clientes classificados como PEP.
4	Saques fragmentados	Verificar se há saques fragmentados com soma superior a R$ 50.000,00 no mesmo dia.
5	Atipicidades nas distâncias de depósitos	Verificar se há indicios de atipicidades em distâncias das transações com depósitos para clientes classificados com mesmo padrão.
6	Movimentação atípica X renda	Verificar os clientes que têm movimentação acima de 300% da renda/faturamento cadastrado.
7	Aumento substancial na proporção de depósitos em espécie	Verificar clientes que tiveram um aumento substancial de depósitos em relação ao total de créditos recebidos.
8	Burla em comunicações de saques e depósitos	Verificar se há tentativa de burla em depósitos/saques fragmentados com soma superior a R$ 50.000,00 nos últimos 5 dias úteis.
9	Movimentação em espécie após recebimento de entes públicos	Verificar se os clientes receberam recursos de entes públicos e realizaram saques em espécie no período.
10	Pagamentos fragmentados de boletos	Verificar se há pagamentos de boletos fragmentados com soma superior a R$ 10.000,00 no mesmo dia.
11	Recebimentos PIX em horários atípicos	Verificar se houver recebimentos de PIX em horários atípicos.
12	Dispositivos com diversas contas	Verificar se há clientes com muitos dispositivos cadastrados para mesma conta.
13	Relação comercial de alto valor	Vericar se houve transação comercial com clientes que comercializem bens de alto valor comercial (bens de luxo).
14	Recursos para servidores públicos	Verificar se houve recebimento de recursos por um terceiro de um ente público e subsequente repasse para servidor público que trabalha para o mesmo ente público.
15	Saques e depósitos fragmentados em dias consecutivos	Verificar se há depósitos/saques fragmentados com soma superior a R$ 50.000,00 nos últimos 5 dias úteis.
16	Depósitos fragmentados no mesmo momento	Verificar depósitos em espécie fragmentados realizados no mesmo momento, na mesma agência e com o mesmo depositante.
17	Saques fragmentados no mesmo momento	Verificar saques em espécie fragmentados com indícios de burla do provisionamento.
18	Envio de recursos para pessoas ligadas à terrorismo	Verificar se houve transações através de PIX, TED e transferência enviada para clientes que tenham mídia negativa relacionada ao terrorismo.
19	Recebimento de recursos de pessoas ligadas à terrorismo	Verificar se houve transações através de PIX, TED e transferência recebida de clientes que tenham mídia negativa relacionada ao terrorismo.
20	Atipicidade nas distâncias de transações com PIX e TEDs - geral	Verificar se há indicios de atipicidades em distâncias das transações com PIX para clientes classificados com o mesmo padrão.

FONTE: O autor (2022)

A base de dados utilizada foi disponibilizada por uma instituição financeira cooperativa do sistema financeiro nacional, o Banco Cooperativo Sicredi, referente a todas as movimentações financeiras realizadas por meio das contas correntes de pessoas físicas e jurídicas, com domicílio bancário nos estados do Paraná, São Paulo e Rio de Janeiro, incluindo regiões de fronteira, principalmente com o Paraguai e Argentina entre dezembro/2021 à junho/2022. Ao todo a amostragem

consistia em 32.599 associados, cuja consolidação das movimentações foi acima da capacidade financeira comprovada.

Para manter o sigilo das informações foi utilizada uma chave para mascarar os dados de identificação do cliente e códigos para não demonstrar o tipo de transação que está sendo efetuada.

Na fase de preparação dos dados, foi necessário incluir um fator chamado comunicação, que determina quais os casos realmente têm indícios de lavagem de dinheiro e devem ser comunicados ao COAF para uma análise mais assertiva.

A construção de cada fator foi baseada nas movimentações em conta corrente dos clientes que apresentaram ao menos uma ocorrência daquele padrão de movimentação com risco de lavagem de dinheiro, durante todo o período analisado. Todos esses dados foram consolidados em uma única tabela com classificação binária dos 20 fatores que levavam em consideração a movimentação financeira dos clientes. Na Tabela 3, tem-se a descrição dos dados e o que cada coluna representa nesta tabela de consolidação dos dados.

Tabela 3 – Colunas de um relatório detalhadas

Nome da Coluna	Representa	Tipo de Dado	Necessário	Binário
IDENTIFICACAO	Identificação do cliente.	Numérico	Sim	Não
TIPO_PESSOA	Pessoa fisica ou Juridica.	Numérico	Sim	Sim
IDADE	Idade das pessoas fisicas ou tempo de constituição das pessoas juridicas.	Numérico	Sim	Não
TEMPO_RELAC	Tempo de relacionamento com a instituição financeira.	Numérico	Sim	Não
PEP	Classificação se o cliente é pessoa politicamente exposta.	Numérico	Sim	Sim
MIDIAS_NEGATIVA_ASSOC	Classificação se o cliente tem midias negativas.	Numérico	Sim	Sim
FATOR 1	Depósitos fragmentados	Numérico	Sim	Sim
FATOR 2	Atipicidade nas distâncias de transações com TEDs	Numérico	Sim	Sim
FATOR 3	Análise das distâncias de transações com TEDs - PEPs	Numérico	Sim	Sim
FATOR 4	Saques fragmentados	Numérico	Sim	Sim
FATOR 5	Atipicidades nas distâncias de depósitos	Numérico	Sim	Sim
FATOR 6	Movimentação atípica X renda	Numérico	Sim	Sim
FATOR 7	Aumento substancial na proporção de depósitos em espécie	Numérico	Sim	Sim
FATOR 8	Burla em comunicações de saques e depósitos	Numérico	Sim	Sim
FATOR 9	Movimentação em espécie após recebimento de entes públicos	Numérico	Sim	Sim
FATOR 10	Pagamentos fragmentados de boletos	Numérico	Sim	Sim
FATOR 11	Recebimentos PIX em horários atípicos	Numérico	Sim	Sim
FATOR 12	Dispositivos com diversas contas	Numérico	Sim	Sim
FATOR 13	Relação comercial de alto valor	Numérico	Sim	Sim
FATOR 14	Recursos para servidores públicos	Numérico	Sim	Sim
FATOR 15	Saques e depósitos fragmentados em dias consecutivos	Numérico	Sim	Sim
FATOR 16	Depósitos fragmentados no mesmo momento	Numérico	Sim	Sim
FATOR 17	Saques fragmentados no mesmo momento	Numérico	Sim	Sim
FATOR 18	Envio de recursos para pessoas ligadas à terrorismo	Numérico	Sim	Sim
FATOR 19	Recebimento de recursos de pessoas ligadas à terrorismo	Numérico	Sim	Sim
FATOR 20	Atipicidade nas distâncias de transações com PIX e TEDs - geral	Numérico	Sim	Sim
COMUNICAÇÃO	Classificação se o cliente foi comunicado ao COAF.	Numérico	Sim	Sim

FONTE: O autor (2022)

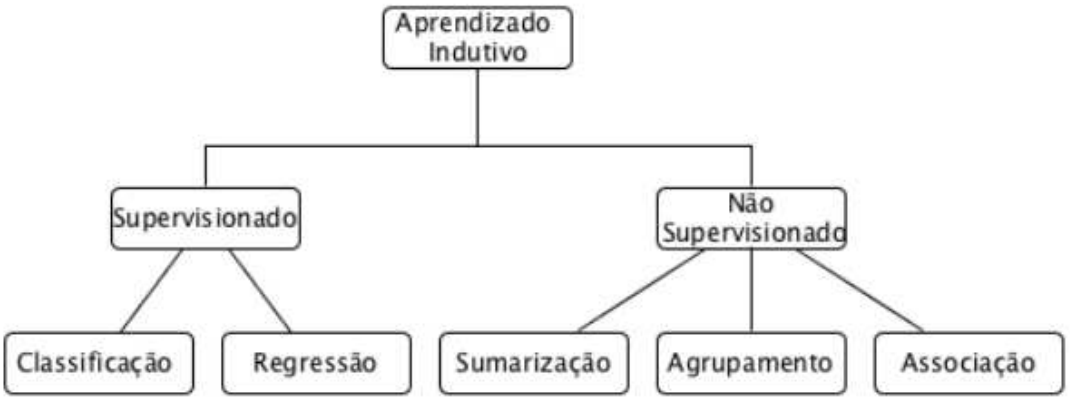
Este estudo utilizou o *SAS Guide* e o *SAS Miner* como softwares de computação estatística e de mineração de dados para grandes volumes de dados,

para tanto as linguagens utilizadas foram *SQL* e *SAS base*. Basicamente os 25 fatores foram construídos através do *SAS Guide*, juntamente com a consolidação dos dados para que o aprendizado de máquina fosse realizado no *SAS Miner.*

4 MODELAGEM (ÁRVORE DE DECISÃO)

Os algoritmos de aprendizado de máquina (AM) podem ser classificados em duas categorias: Aprendizado supervisionado e Aprendizado não supervisionado, conforme ilustrado na Figura 3.

Figura 3 - Aprendizado de Máquina



FONTE: Adaptado de FACELI (2011)

O aprendizado supervisionado ou modelo preditivo, trabalha com base em uma hipótese, obtida por meio de treinamento, ou seja, exemplos. Essa hipótese é utilizada para rotular novos registros. Os algoritmos utilizados neste tipo de aprendizado são de:

- Classificação – tem o objetivo de reconhecer, em um conjunto de dados, os exemplos que tenham a mesma caraterística e associar a uma classe previamente definida. Para isso é necessária uma base com dados já classificados para o algoritmo realizar o treinamento.
- Regressão – este modelo tem similaridades com o modelo de classificação, mas seu objetivo é encontrar uma função por meio de uma base de treinamento. Essa função tem por objetivo estimar um valor numérico contínuo para o atributo classe de um registro desconhecido.

Já o aprendizado não supervisionado, denominado modelo descritivo, identifica informações relevantes nos dados sem a utilização de uma base de treinamento para fazer o aprendizado. Neste caso os algoritmos usados são de:

- Sumarização – fornece uma descrição resumida dos dados, onde são apresentadas as principais características deles.
- Agrupamento – esse modelo tem o objetivo de promover um conjunto de dados a um subconjunto chamado de *clusters*. Esses *clusters* armazenam os objetos que são semelhantes através de medidas.
- Associação – ele busca o relacionamento entre os atributos por meio das interações e conexões entre os dados da base.

Os algoritmos de árvore de decisão constroem uma árvore a partir de uma base de treinamento. Seu objetivo é encontrar uma árvore de decisão mínima (em termos de números de nós) por meio da minimização dos erros da generalização (ROKACH; MAIMOM, 2007).

A árvore de decisão é considerada um modelo preditivo e supervisionado que pode ser usado para a tarefa de classificação. Esse modelo apresenta algumas vantagens:

- Flexibilidade – como é um modelo que não assume nenhuma distribuição dos dados, os espaços dos objetos são divididos em subespaços que são ajustados com diferentes modelos.
- Robustez – por mais que ocorram transformações nas variáveis de entrada, no final produz árvores com a mesma estrutura.
- Interpretabilidade – todas as decisões são baseadas nos valores definidos na descrição do problema.

Além das vantagens descritas, esse algoritmo é considerado um classificador robusto, pois utiliza a estrutura de uma árvore para modelar o relacionamento entre as características e as classes.

5 RESULTADOS

O algoritmo de árvore de decisão utilizado nesse estudo baseou-se em um universo de 32.599 clientes, avaliando 25 fatores que determinariam quais teriam maior risco de lavagem de dinheiro e deveriam ser comunicados aos órgãos competentes.

Foi utilizado um modelo de aprendizado de máquina baseado em classificação, onde dos 32.599 clientes, um total de 436 foram classificados com risco de lavagem de dinheiro e de acordo com a base histórica disponibilizada para esse estudo, foram comunicados ao órgão competente. Estatisticamente o número de clientes comunicados ao COAF é de 1,33% durante o período analisado.

A base de dados foi separada 50% para treinamento do algoritmo e 50% para validação dos resultados.

Em uma base de treinamento, o algoritmo apresentou um percentual de acerto de 99,95% para não comunicar, ou seja, dos 16.298 clientes analisados pelo algoritmo, ele decidiu por não comunicar em 16.073. Todavia, em apenas 3 ocasiões o algoritmo acertou a decisão por comunicar os casos que realmente deveriam ser comunicados. Para 214 clientes que deveriam ser comunicados, o modelo decidiu por não comunicar. E, por fim, para 8 clientes que o modelo deveria decidir por não comunicar, a decisão do algoritmo foi por comunicar.

A Figura 4 apresenta o resumo dos resultados obtidos pelo modelo na base que foi separada para treinamento.

Figura 4 - Tabela de decisões do Modelo - Treinamento do Algoritmo (Captura de tela de algoritmo)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage	Adjusted Percent of Predict/Decision Variable
1	1	27.2727	1.3825	3	0.0184	0.0691
1	2	1.3139	98.6175	214	1.3130	4.9309
2	1	72.7273	0.0497	8	0.0491	0.0473
2	2	98.6861	99.9503	16073	98.6195	94.9527

Fonte: O autor (2022)

Na base de validação dos resultados, o algoritmo apresentou um percentual de acerto de 99,98%, ainda maior do que na base de treinamento, para os casos em que não deveria comunicar, ou seja, dos 16.301 clientes analisados pelo modelo, ele decidiu por não comunicar em 16.079. Mas, semelhante ao acontecido na base de treinamento, o percentual de acerto por comunicar quando realmente deveria, o algoritmo acertou em apenas 2 ocasiões. Dos 219 clientes que deveriam ser comunicados, 217 não foram, portanto o percentual de acerto ficou menor que 1%. E, por fim, para 3 clientes que o modelo deveria decidir por não comunicar, a decisão do algoritmo foi por comunicar.

Na Figura 5 tem um resumo dos resultados obtidos na base que foi separada para validação do modelo.

Figura 5 - Tabela de decisões do Modelo - Validação do Algoritmo (Captura de tela de algoritmo)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage	Adjusted Percent of Predict/Decision Variable
1	1	40.0000	0.9132	2	0.0123	0.0461
1	2	1.3316	99.0868	217	1.3312	4.9991
2	1	60.0000	0.0187	3	0.0184	0.0177
2	2	98.6684	99.9813	16079	98.6381	94.9707

Fonte: O autor (2022)

O modelo proposto se mostrou capaz de tomar decisões com alto nível de acerto, de um ponto de vista estatístico, mas do ponto de vista do risco do negócio, o risco de não comunicar uma movimentação atípica e com indícios de lavagem de dinheiro é maior do que o de comunicar ao (COAF). Portanto, a proposta é definir uma base de treinamento mais balanceada para que o modelo possa ter uma maior probabilidade de identificar casos de lavagem de dinheiro e tomada de decisão.

Diante do exposto, foi realizado um balanceamento na base de dados, diminuindo a base para uma amostra de 800 clientes, onde o único parâmetro para seleção foi que a base apresentasse 50% dos casos classificados para comunicar ao órgão competente e os outros 50% por não comunicar.

A nova base de dados amostral foi separada em aproximadamente 50% para treinamento do algoritmo e 50% para validação dos resultados.

Realizado o balanceamento dos dados para treinamento do algoritmo, dos 398 clientes selecionados, notou-se que em 60,05% dos casos o modelo acertou na decisão (ver Figura 6), seja para os 137 clientes que deveriam ser comunicados ou para os 102 clientes que não deveriam ser comunicados. Para 97 clientes que estavam classificados previamente para não comunicar o algoritmo decidiu por comunicar e para 62 clientes que deveriam ser comunicados o modelo errou e decidiu por não comunicar.

Figura 6 - Decisões do Modelo - Treinamento do Algoritmo após balanceamento dos dados (Captura de tela de algoritmo)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
1	1	58.5470	68.8442	137	34.4221
2	1	41.4530	48.7437	97	24.3719
1	2	37.8049	31.1558	62	15.5779
2	2	62.1951	51.2563	102	25.6281

Fonte: O autor (2022)

A Figura 7 apresenta os dados da validação do modelo após o balanceamento dos dados, onde o percentual de assertividade foi de 60,44%, que representa a soma dos 132 clientes que foram comunicados de forma assertiva e os 111 clientes que foi decidido por não comunicar, também de forma assertiva. O resultado extraído na validação foi muito próximo do resultado extraído na base de treinamento do algoritmo.

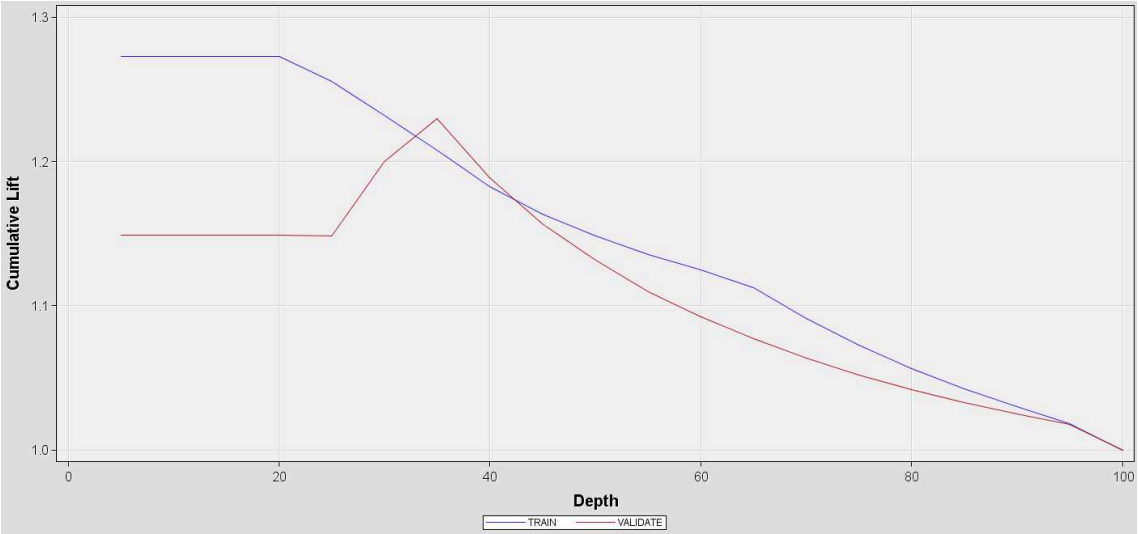
Figura 7 - Decisões do Modelo - Validação do algoritmo após balanceamento dos dados (Captura de tela de algoritmo)

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
1	1	59.4595	65.6716	132	32.8358
2	1	40.5405	44.7761	90	22.3881
1	2	38.3333	34.3284	69	17.1642
2	2	61.6667	55.2239	111	27.6119

Fonte: O autor (2022)

A Figura 8 é a demonstração gráfica da taxa de aprendizado do algoritmo ao decorrer da amostra utilizada. Portanto o eixo Y é o percentual de alavancagem acumulada obtido durante o aprendizado de máquina ao classificar cada observação da base de dados. É possível notar que a maior taxa de aprendizado do algoritmo foi realizada nos primeiros 20% dos dados da base separada para treinamento, quanto mais clientes o modelo analisou, a taxa de aprendizado ou de ganho foi diminuindo proporcionalmente e gradativamente.

Figura 8 - Gráfico de *Gain / Lift* da classificação do modelo



Fonte: O autor (2022)

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Os resultados obtidos sugerem que o modelo baseado em árvore de decisão trouxe resultados significativos e que podem apresentar clientes que através da sua movimentação tragam alertas de indícios de lavagem de dinheiro, através dos fatores que foram selecionados de acordo com a legislação vigente. Vale ressaltar que o balanceamento dos dados realizado como uma proposta de trabalhar melhor com os dados e conseguir resultados diferentes, apresentando uma alternativa de resultado para treinamento do modelo que diminuísse o risco de não comunicar uma atividade suspeita ao órgão competente. Mas o balanceamento dos dados não se encaixa na métrica proposta por esse estudo que tem premissa de usar dados reais.

O modelo preditivo apresentado se mostrou bastante útil para elencar alertas de lavagem de dinheiro que podem ser comunicados ao órgão competente, através da movimentação com indícios.

A partir do presente trabalho é possível determinar diversos caminhos a serem percorridos no futuro. Esse modelo ficou limitado aos 25 fatores escolhidos como premissas, fatores bem específicos que foram elencados pela legislação vigente, mas há vários outros fatores que poderiam ser desenvolvidos e que agregariam bastante valor ao modelo existente, como, por exemplo o desenvolvimento de um fator que utilizasse uma base de dados que fosse capaz de verificar o padrão de movimentação financeira de pessoas jurídicas por ramo de atividade e posição geográfica, da mesma forma para pessoas físicas com a mesma ocupação de uma determinada localização.

Além de calibrar melhor os fatores, identificando quais seriam mais úteis de manter ou excluir da base, além da criação de novos que sejam mais assertivos para uma melhor classificação das características de crimes de lavagem de dinheiro. A classificação pela comunicação ou não de cada cliente na base real utilizada nesse estudo foi realizada de forma subjetiva pelo analista que tomou a decisão pela comunicação, portanto dos 436 clientes comunicados ao COAF, podem existir situações que não são riscos de lavagem de dinheiro ou que não foram elencados pelos fatores dispostos no presente trabalho. A criação de fatores com critérios mais bem definidos poderia trazer melhorias para o modelo, além da melhor classificação

dos casos comunicados na base histórica, poderia trazer ganhos significativos para o estudo.

O combate à lavagem de dinheiro e o financiamento ao terrorismo tem caráter subjetivo de decisão do que deve ou não ser comunicado ao (COAF), portanto, quanto mais assertivos forem os fatores melhor serão para contribuir para filtrar casos.

Quando leva-se em consideração o alto volume de transações em comparação a taxa de comunicação, encontra-se um percentual muito baixo, nesta pesquisa essa taxa foi em torno de 1,33%. Portanto, há necessidade de uma melhoria do modelo para ser mais assertivo naqueles casos em que realmente precisa ser comunicado, mas diante das considerações e resultados obtidos nesse estudo há possibilidades de retorno positivos para futuros trabalhos.

REFERÊNCIAS

BRASIL. Presidência da República. **Lei n° 9.613, de 3 de março de 1998:** Dispõe sobre os crimes de "lavagem" ou ocultação de bens, direitos e valores; a prevenção da utilização do sistema financeiro para os ilícitos previstos nesta Lei; cria o Conselho de Controle de Atividades Financeiras - COAF, e dá outras providências. Diário Oficial [da] Presidência da República Federativa do Brasil. Brasília, DF. 03 de mar. 1998. Não paginado.

BANCO CENTRAL DO BRASIL. **Circular nº 3.978 de 23 de janeiro de 2020:** Dispõe sobre a política, os procedimentos e os controles internos a serem adotados pelas instituições autorizadas a funcionar pelo Banco Central do Brasil visando à prevenção da utilização do sistema financeiro para a prática dos crimes de "lavagem" ou ocultação de bens, direitos e valores, de que trata a Lei nº 9.613, de 3 de março de 1998, e de financiamento do terrorismo, previsto na Lei nº 13.260, de 16 de março de 2016. Não paginado.

BANCO CENTRAL DO BRASIL. **Carta Circular nº 4.001 de 29 de janeiro de 2020:** Divulga relação de operações e situações que podem configurar indícios de ocorrência dos crimes de "lavagem" ou ocultação de bens, direitos e valores, de que trata a Lei nº 9.613, de 3 de março de 1998, e de financiamento ao terrorismo, previstos na Lei nº 13.260, de 16 de março de 2016, passíveis de comunicação ao Conselho de Controle de Atividades Financeiras (Coaf).

BORBA, M. C. V. **Um escore de risco para classificação de transações suspeitas de lavagem de dinheiro via regressão ordinal.** 2017. Disponível em: https://repositorio.unb.br/handle/10482/32349. Acesso em 02 nov. 2022.

BORGES, H. C. O. C.; MALTA, B. P. **Crime de lavagem de dinheiro no brasil: Atuação do COAF na prevenção e combate.** 2018. Disponível em: https://www.unirv.edu.br/conteudos/fckfiles/files/CRIME%20DE%20LAVAGEM%20DE%20DINHEIRO%20NO%20BRASIL_%20ATUA%C3%87%C3%83O%20DO%20%20COAF%20NA%20PREVEN%C3%87%C3%83O%20E%20COMBATE.pdf. Acesso em 30 out. 2022.

CALLEGARI, A. L. Problemas pontuais da lei de lavagem de dinheiro. Revista Brasileira de Ciências Criminais, **Revista dos Tribunais**, São Paulo, n. 31, p. 183-200. 2000.

CHAPMAN. P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0 Step-bystep data mining guide.** IBM, Aug. 2000. [Online]. Disponível em: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf. Acesso em 02 nov. 2022.

CUNHA, I. R.; MACEDO, D. L.; ESTENDER, A. C. **Prevenção à lavagem de dinheiro no sistema financeiro.** 2016. Disponível em: https://assets.unitpac.com.br/arquivos/Revista/77/Artigo_5.pdf. Acesso em 30 out. 2022.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, C. P. L. F. A. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina.** Rio de Janeiro: LTC, 2011.

JUNIOR, J. C. P. **Modelos de detecção de fraudes utilizando técnicas de aprendizado de máquina.** 2019. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27166/Dissertacao_Joao_Carlos_Pacheco_VFinal_2.pdf. Acesso em 27 out. 2022.

LUO, X. **Suspicious Transaction Detection for Anti-Money Laundering.** 2014 International Journal of Security and Its Applications, Vol.8, p.157-166. 2014.

PAULA, E. L. **Mineração de dados como suporte à detecção de lavagem de dinheiro.** 2016. Disponível em: https://repositorio.unb.br/bitstream/10482/22598/1/2016_EbberthLopesdePaula.pdf. Acesso em 30 out. 2022.

PITOMBO, A. S. de M. Lavagem de dinheiro: A tipicidade do crime antecedente. São Paulo: **Revista dos Tribunais**, 2003.

ROKACH, L.; MAIMOM, O. **Data mining with decision trees: theory and applications.** World Scientific Publishing Co., Inc., River Edge, NJ, 2007

SOCREPPA, R. B. **Um modelo de inferência utilizado na detecção de indícios de lavagem de dinheiro**. 2016. Disponível em: https://lactec.org.br/wp-content/uploads/2019/11/DissertacaoFinal-VF.pdf. Acesso em 10 out. 2022.

SURESH, C.; REDDY, T. K.; SWETA N. **A Hybrid Approach for Detecting Suspicious Accounts in Money Laundering Using Data Mining Techniques.** International Journal of Information Technology and Computer Science (IJITCS), Vol.8, p.37- 43. 2016.