Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Alessandro Roberto Luz

Aplicação de técnicas de mineração de dados no auxílio à investigações criminais

Curitiba
2022

Alessandro Roberto Luz

Aplicação de técnicas de mineração de dados no auxílio à investigações criminais

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: André Ricardo Abed Grégio

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Aplicação de técnicas de mineração de dados no auxílio à investigações criminais

Alessandro Roberto Luz[1]
André Ricardo Abed Grégio[2]

Resumo

Técnicas de aprendizado de máquina têm sido cada vez mais adotadas para auxiliar na análise de grandes volumes de dados. O campo de aplicação é amplo e abrange diversos ramos do conhecimento. Considerando os benefícios potenciais, o presente trabalho propõe a aplicação de técnicas de aprendizado de máquina não supervisionado no auxílio à investigação criminal. Foram utilizados dados de movimentações financeiras decorrentes do afastamento do sigilo bancário e a análise foi realizada em três perspectivas com a aplicação de algoritmos distintos: detecção de anomalias com o algoritmo Isolation Forest, análise de agrupamento de dados mistos com k-Prototype e mineração de regras de associação com Apriori. Ao final, o trabalho realizado mostrou-se viável como estratégia de auxílio ao exame preliminar dos dados revelando aspectos que poderão direcionar outras análises promovidas pelo analista.
Palavras-chave: Isolation Forest, k-Prototype, Apriori, aprendizado de máquina não supervisionado.

Abstract

Machine learning techniques have been increasingly adopted to aid in the analysis of large volumes of data. The field of application is broad and encompasses several branches of knowledge. Considering the potential benefits, the present work proposes the application of unsupervised machine learning techniques to aid criminal investigation. Data from financial transactions resulting from the removal of bank secrecy were used and the analysis was performed in three perspectives with the application of different algorithms: anomaly detection with the Isolation Forest algorithm, analysis of mixed data clusters with k-Prototype and rule mining of association with Apriori. In the end, the work carried out proved to be viable as a strategy to aid the preliminary examination of the data, revealing aspects that may direct other analyzes promoted by the analyst.
***Keywrods**: Isolation Forest, k-Prototype, Apriori, unsupervised machine learning*

[1]Aluno do programa de Especialização em Data Science & Big Data, alessandroluz@ufpr.br.
[2]Professor do Departamento de Informática - DInf/UFPR, gregio@inf.ufpr.br.

1 Introdução

O volume crescente de dados produzidos por organizações diversas popularizou a adoção de técnicas de aprendizado de máquina para melhor compreensão da realidade, possibilitando assim a descoberta de conhecimento útil para auxílio à tomada de decisões.

O campo de aplicação dos algoritmos de aprendizado de máquina é bastante amplo, abrangendo os mais diversos ramos do conhecimento. Havendo um conjunto de dados estruturado, há potencial para a extração de conhecimento por meio das técnicas apropriadas. Nesse sentido, dados pertinentes à segurança pública também podem ser objetos de análise com a aplicação de métodos de inteligência artificial, viabilizando novas maneiras de se lidar com o combate à criminalidade.

Em todo o mundo, o crime organizado movimenta cifras astronômicas. O caráter empresarial do crime é cada vez mais evidente, pois as organizações criminosas são estruturadas com a finalidade precípua de auferir lucros com atividades ilícitas.

Muitos delitos geram ganhos financeiros. Por exemplo, um traficante obtém lucro indevido com a venda de substâncias entorpecentes, assim como um estelionatário igualmente percebe vantagem ilícita ludibriando vítimas inocentes. Por outro lado, para gozar dos lucros ilícitos os criminosos utilizam de artifícios ou técnicas com o objetivo de burlar os mecanismos de fiscalização e controle das instituições competentes e da repressão estatal. O cenário é propício para a lavagem de dinheiro, crime previsto em lei (artigo 1o, da Lei nº 9.613/98) [1] e que consiste em dissimular ou ocultar a natureza, origem, de bens, valores ou direitos provenientes de infração penal.

Em breves palavras, a lavagem de capitais é a aplicação de técnicas ou artifícios no intuito de conferir aparência lícita a ganhos obtidos ilicitamente, justamente para que os criminosos possam usufrir do lucro ilegal sem despertar atenção dos mecanismos de controle, fiscalização e repressão estatal.

Há inúmeros métodos ou tipologias utilizados para lavar dinheiro [2], desde práticas simples até esquemas altamente sofisticados e complexos em que diversas operações financeiras são realizadas, em várias camadas e envolvendo diversos atores.

Didaticamente, a lavagem de dinheiro é fundamentada em três etapas:

1. Colocação;
2. Ocultação;
3. Integração;

Na etapa da **colocação**, o criminoso insere os valores ilícitos no sistema financeiro. É a fase mais vulnerável da lavagem do ponto de vista delituoso, mas também a mais relevante da perspectiva repressiva, pois neste momento há atuação dos mecanismos de controle e compliance das instituições financeiras, respaldado por normas legais que auxiliam na investigação e combate ao crime.

Na etapa de **ocultação** são realizadas diversas manobras e operações financeiras objetivando dificultar o rastreamento dos valores ou promover seu distanciamento da origem criminosa.

Por fim, na etapa de **integração**, os valores com aparência lícita são reintegrados ao patrimônio do indivíduo. Nessa etapa é comum a aquisição de bens.

Diversos trabalhos de pesquisa já foram realizados objetivando aplicar técnicas de aprendizado de máquina para a detecção de fraudes. São conhecidos os exemplos de modelos preditivos para classificar operações com cartões de créditos em fraudulentas ou não. Nesse tipo de tarefa, os dados são rotulados, isto é, foram previamente classificados por interferência humana e o aprendizado supervisionado é aplicado para prever classificações futuras.

Uma das principais medidas adotadas em investigações policiais de crimes que apresentam reflexos financeiros é o afastamento judicial do sigilo bancário. É uma diligência salutar para averiguar anormalidades financeiras, investigar ganhos ilícitos ou detectar tipologias de lavagem de dinheiro em transações bancárias. A obtenção desses dados, efetivada somente com prévia autorização judicial, possibilita que as instituições financeiras enviem à polícia investigativa informações sobre movimentações bancárias realizadas por aqueles que foram alvo da quebra do sigilo de dados e referentes a um período delimitado.

Em geral, o trabalho de análise de dados bancários na busca de evidências não é uma tarefa fácil, demandando conhecimento analítico do profissional de investigação no tocante à identificação de ilícitos, na detecção de padrões ou de transações suspeitas que possam contribuir de alguma forma para a elucidação dos crimes.

Abordagens de prevenção ou detecção de lavagem de dinheiro baseadas em mineração de dados têm sido propostas pela literatura. Algumas com o propósito de detectar padrões, outras baseadas na classificação de operações fraudulentas em registros rotulados, no agrupamento de dados mediante aprendizado não supervisionado, dentre outras. Salehi et al. [3] catalogaram, de forma sistemática, diversas técnicas de mineração de dados atualmente existentes que podem auxiliar na detecção de lavagem de dinheiro. Kute et al. [4] efetuaram a revisão de técnicas de *deep learning* para a detecção de transações suspeitas com indícios de lavagem de dinheiro. Já Paula et al. [5] apresentaram o resultado de pesquisas na construção de um modelo de *deep learning* de aprendizado não supervisionado de suporte à investigação de fraudes e lavagem de dinheiro em operações de exportações de mercadorias no território brasileiro. Enfim, a gama de ferramentas disponíveis é diversificada, abrangendo métodos de agrupamento, de classificação, de predição, de análise de redes sociais, etc.

Vale ressaltar que a detecção de indícios de atividades ilícitas em dados bancários nem sempre é uma tarefa trivial. Na prática, diversos fatores devem ser levados em consideração como o volume de dados disponível, a complexidade do caso, o contexto fático em apuração, o perfil econômico-financeiro dos investigados e as evidências previamente coletadas na investigação. Como naturalmente as transações bancárias não são classificadas em legítimas ou ilegítimas, a tarefa se torna ainda mais complexa e a intervenção humana na avaliação ainda é relevante e não deve ser dispensada.

Em investigações de elevada complexidade envolvendo dezenas de suspeitos e abrangendo vários anos de afastamento do sigilo bancário, o conjunto de dados pode facilmente chegar a milhares de transações bancárias.

Sendo assim, o presente trabalho propõe a aplicação de técnicas de mineração de dados, essencialmente métodos de aprendizado não supervisionado, no auxílio à investigação financeira criminal. Contudo, a proposta é orientada ao conhecimento (*knowledge-based*), pois acredita-se que, devido às nuances presentes em dados bancários, à variabilidade das operações financeiras e às peculiaridades do caso concreto subjacente, a forma mais viável de se extrair insights relevantes seria a aplicação de métodos de descoberta do conhecimento aliados à experiência do analista. A adoção das técnicas adequadas aliada à expertise do profissional potencializa a capacidade de identificação de transações suspeitas, sobretudo tipologias de lavagem dinheiro.

Sendo assim, a proposta idealizada é fundamentada tanto na aplicação de algoritmos de *data mining* com o propósito de retratar o "comportamento" dos dados e detectar eventuais anomalias ou padrões, quanto no conhecimento humano e na análise criteriosa de operações que possam indicar anormalidades ou o cometimento de ilícitos.

O trabalho baseou-se em experimentos com três algoritmos de aprendizagem não-supervisionada: *Isolation Forest* para detecção de anomalias, *k-Prototype* para análise de agrupamento de dados mistos (categóricos e numéricos) e mineração de regras de associação com *Apriori*.

A escolha do *Isolation Forest* foi devido ao bom desempenho do algoritmo em comparação a outros métodos de detecção de *outliers*. O *Isolation Forest* isola as anomalias que estiverem mais próximas da raiz da árvore. Essa característica única possibilita ao algoritmo construir modelos parciais, proporcionalmente menores que um modelo completo. Segundo Liu et al. [6], essa abordagem tem se mostrado altamente eficaz na detecção de anomalias. Na avaliação realizada pelos citados autores, o algoritmo apresenta performance significa-

tivamente superior, sobretudo em tempo de execução. Ademais, a complexidade do algoritmo é linear, apresentando bom desempenho em grandes conjuntos de dados e com baixo custo de memória.

No tocante ao algoritmo *k-Prototype* para a análise de agrupamentos, Huang [7] assevera que o método preserva a eficiência do algoritmo *k-Means*, porém sem a limitação de agrupar apenas dados numéricos. Conforme o mencionado autor, o algoritmo é eficiente para a clusterização de grandes conjuntos de dados com valores numéricos e categóricos (mistos), comuns em aplicações de mineração de dados.

Já o algoritmo *Apriori* é amplamente conhecido para a mineração de regras de associação e detecção de padrões, razão pela qual a escolha foi natural.

2 Conjunto de Dados

Para a realização dos experimentos foi utilizado conjunto de dados de transações bancárias reais legitimamente obtido mediante autorização judicial para instruir investigação criminal. Os dados foram enviados por diversas instituições financeiras que figuraram como destinatárias da ordem de afastamento do sigilo bancário e recepcionados por intermédio do Sistema de Investigação de Movimentação Bancária (SIMBA).

2.1 Características dos dados

Os dados são referentes à quebra do sigilo bancário de 09 pessoas físicas e 01 pessoa jurídica, abrangendo o período aproximado de cinco anos e se referem a movimentações financeiras dos investigados. Ao todo o *dataset* possui 24819 linhas e 16 colunas consistentes em:

Tabela 1: Dicionário do conjunto de dados

Variável	Descrição
Chave extrato E.g.	Indexador 1,2,3...
Banco E.g.	(discreta/categórica) Número do banco 001, 260
Agência E.g.	(discreta/categórica) Número da agência bancária 1, 237
Conta	(discreta/categórica) Número da conta
Tipo de conta	(discreta/categórica) 1- conta corrente, 2 - conta de poupança, 3 - conta investimento e 4 - outros casos
Data do lançamento	Data da operação bancária
Descrição do lançamento E.g.	(texto) Descrição do lançamento efetuada pelo banco Transferência entre contas, saque, etc.
CNAB	(discreta/categórica) Código para o tipo da operação bancária
Descrição CNAB	(texto) descrição correspondente ao código CNAB
Valor E.g.	(contínua) Valor da operação 20,00, 150,00
Natureza	(categórica) C – Crédito, D - Débito
Banco O/D	(discreta/categórica) Número do banco de origem ou destino
Agência O/D	(discreta/categórica) Número da agência de origem ou destino
Conta O/D	(discreta/categórica) Número da conta de origem ou destino
Tipo de conta O/D	(discreta/categórica) 1- conta corrente, 2 - conta de poupança, 3 - conta investimento e 4 - outros casos
Tipo de pessoa O/D	(discreta/categórica)

2.2 Limpeza e preparo dos dados

O tratamento geral dos dados consistiu em selecionar variáveis de interesse e formatar apropriadamente alguns campos, quando necessário. Atributos com dados ausentes de preenchimento facultativo ou sujeitos a inconsistências foram preservados. Na análise de dados bancários, transações só devem ser descartadas se for necessário para o objetivo pretendido, pois a premissa inicial é que qualquer operação pode ser relevante a depender do contexto da investigação.

Eventuais preparações específicas serão implementadas conforme a necessidade dos experimentos retratados neste trabalho.

3 Modelos Empregados

O trabalho propõe a análise de dados bancários sob três perspectivas: detecção de anomalias, análise por agrupamento e mineração de regras de associações ou padrões por intermédio da aplicação de algoritmos de aprendizado não supervisionado. Para tanto foram estabelecidos três modelos gerais, um com base no algoritmo *Isolation Forest* para detecção de anomalias, o segundo com aplicação do algoritmo *k-Prototypes* para o agrupamento de dados mistos e o terceiro utilizando o algoritmo *Apriori* para formulação de regras de associação.

3.1 Detecção de anomalias com *Isolation Forest*

O *Isolation Forest* é um algoritmo de aprendizado de máquina não supervisionado para detecção de anomalias, baseado em árvores de decisão e aplicável em dados não rotulados. Ele trabalha isolando valores discrepantes (*outliers*) presentes no conjunto de dados. As anomalias

são pontos incomuns que desviam da normalidade e não seguem o padrão esperado. Geralmente a detecção de anomalias é uma tarefa relevante na análise de dados, pois outliers podem impactar negativamente o desempenho de modelos de machine learning. Em contrapartida, eles também podem indicar pontos de interesse que merecem a atenção do analista, conforme o contexto da análise.

A detecção de anomalias em dados bancários pode revelar transações atípicas capazes de direcionar a apuração dos fatos. Exemplificando, indivíduo assalariado cujo comportamento financeiro regular é o recebimento mensal de crédito salarial pago em parcela única de valor médio, mas inesperadamente e sem motivo justificável recebe crédito de elevada monta em sua conta bancária.

Em *datasets* com expressiva quantidade de registros e com padrões complexos a identificação de anormalidades pela simples observação dos dados não é eficaz. Portanto, a aplicação de algoritmos voltados para esse fim é de grande valia.

A maioria dos algoritmos para detecção de anomalias baseia-se na elaboração de um perfil de instâncias normais para então identificar ocorrências que, a princípio, não estão de acordo com esse perfil, ou seja, eles são otimizados para construir um perfil de objetos comuns, mas não são otimizados para a detecção de objetos incomuns. Consequentemente, nem sempre o resultado é tão bom quanto o esperado [6].

Em razão disso, o algoritmo *Isolation Forest* propõe a construção de um modelo que isola explicitamente as anomalias ao invés de estabelecer o padrão de registros normais. O método trabalha com duas premissas intrínsecas às anomalias: a) as anomalias são a minoria composta por menos instâncias; b) os registros têm valores de atributos discrepantes em relação às demais instâncias do conjunto de dados. Sendo assim, as anormalidades são escassas e "diferentes", o que as tornam mais suscetíveis de isolamento, ou seja, na separação de uma instância discrepante do restante das demais instâncias.

O algoritmo também apresenta vantagem em termos de desempenho, uma vez que não utiliza medidas de distância ou densidade para detectar anomalias, exigindo assim menor custo computacional no trabalho.

Para identificar anomalias, o Isolation Forest constrói uma árvore aleatória e efetua o particionamento recursivo de instâncias até que todas as instâncias sejam isoladas. Os grupos diferentes são separados da raiz da árvore, situando-se nos ramos. O particionamento aleatório produz caminhos mais curtos para as anomalias. Em contrapartida, amostras que penetram mais fundo na árvore e precisam de mais cortes para separá-las possuem menor probabilidade de serem observações anormais.

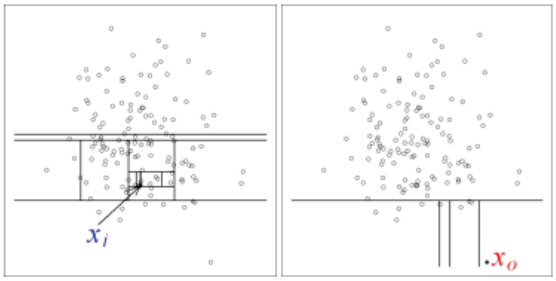

Figura 1: Particionamento dos dados com o isolamento dos dados anômalos. [6]

3.1.1 Experimento 1

Inicialmente, foi realizada a inspeção dos dados e a representação visual das operações de créditos e débitos. Na figura 2, os créditos estão representados na cor azul e débitos em vermelho:

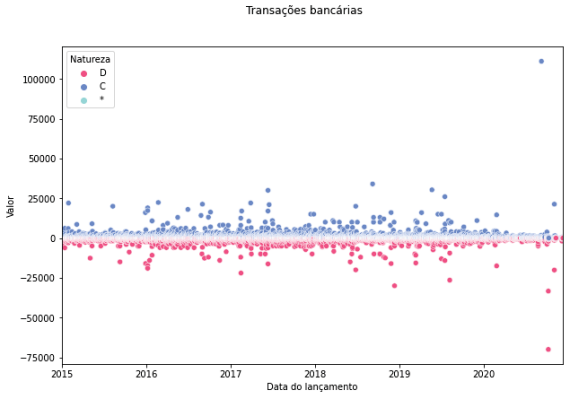

Figura 2: Créditos e débitos no conjunto de dados.

O experimento foi realizado com o auxílio da biblioteca *scikit-learn*, amplamente utilizada em ciência de dados. O pacote disponibiliza ao analista vários algoritmos de *machine learning* e mineração de dados para diversos propósitos.

A principal tarefa foi criar um modelo instanciando a classe *"IsolationForest"*. O construtor da classe apresenta quatro parâmetros principais:

- **Número de estimadores ("n_stimators")**: o número de árvores que serão construídas. É um parâmetro opcional. Para o estudo foi arbitrado o valor de 50 árvores;
- **Número máximo de amostras ("max_samples")**: representa o número de amostras utilizadas para o treinamento do modelo;
- **Contaminação ("contamination")**: consiste na proporção esperada de outliers no conjunto de dados. Após alguns testes, foi arbitrado para o experimento o valor de 0.0020, isto é, estipula-se que até 0,20% dos dados sejam considerados anomalias;
- **Número máximo de características ("max_features")**: é a quantidade máxima de

características utilizadas no modelo. Será considerada apenas uma característica, o atributo "Valor" do dataset.

Definido o modelo, é efetuado o treinamento com base na característica estipulada. O resultado pode ser observado conforme o índice de pontuação de precisão do modelo ("scores") e a categorização das instâncias em anômala ou não. Para tanto foi adicionada no dataset uma coluna "Anomalia" com valores "1" e "-1" resultantes do algoritmo. Valores negativos para o "score" e "-1" para a coluna "Anomalia" indicam registros anômalos.

Após o processamento foram detectadas 50 instâncias consideradas anormais para o conjunto de dados.

3.1.2 Resultado e constatações

Ao final do experimento o algoritmo mostrou-se computacionalmente eficiente realizando a tarefa de detecção de anomalias com relativa rapidez.

Para o conjunto de dados de 24.819 registros, foram detectadas 50 observações anômalas (figura 3). Conforme esperado, a quantidade de anomalias detectadas depende diretamente do valor do parâmetro de contaminação estipulado para o modelo. Foram testados diversos valores e, mesmo ao mínimo incremento, o modelo resultou em quantidade expressiva de pseudo-anomalias. Por outro lado, com um valor demasiadamente baixo o modelo não foi capaz de detectar registros discrepantes que eram claramente evidentes. Conclui-se que, quanto maior o conjunto de dados, mais sensível é o ajuste da contaminação. Por essa razão, é aconselhável que o analista já tenha uma ideia pré-definida acerca da quantidade de dados anômalos aceitáveis para o conjunto de dados.

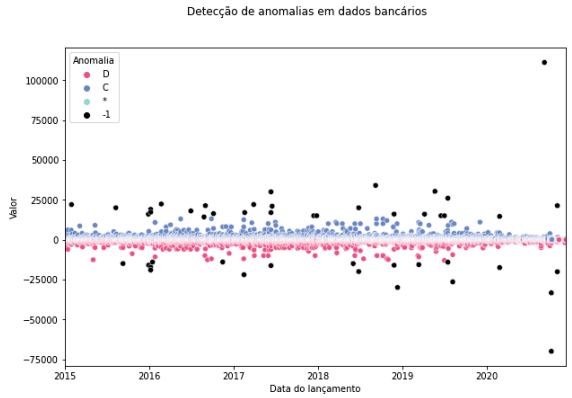

Figura 3: Anomalias detectadas pelo algoritmo *Isolation Forest*.

Do ponto de vista da análise orientada ao conhecimento, o experimento foi satisfatório para confirmar que o conjunto de dados, como um todo, apresenta uma faixa de valores estratificada relativamente uniforme tanto em operações a crédito quanto a débito. As instâncias mais discrepantes foram notadas no ano de 2020. Dependendo do contexto da investigação, tal informação poderia ser útil para direcionar a concentração de esforços de trabalho nesse período. Outliers em operações bancárias podem indicar mudança brusca no comportamento financeiro. Não há qualquer problema se as transações forem plenamente justificáveis. (Ex.: recebimento de herança, alienação de bens, etc.). Caso contrário, se forem devidamente justificadas e houver ilícitos em apuração, podem indicar operações suspeitas ou até mesmo indicativo de lavagem de dinheiro.

3.2 Agrupamento com *k-Prototypes*

Uma das técnicas conhecidas para análise de dados não rotulados é o agrupamento ou clusterização. O aprendizado de máquina por agrupamento, também denominado segmentação de dados, permite separar os objetos em grupos (*clusters*) com características similares. O método é útil para representar o comportamento de grandes conjuntos de dados. Também é uma das técnicas de redução de dados (*data reduction*) que pode ser empregada para se obter uma representação reduzida do conjunto de dados, mas mantendo sua integridade original.

Para definir a similaridade, geralmente os algoritmos de agrupamento realizam cálculos de distância para determinar o quão próximos os objetos estão entre si no espaço. Entretanto, a eficácia desta técnica depende da natureza dos dados. Serão obtidos resultados mais satisfatórios para os dados que puderem ser organizados em grupos distintos do que para dados difusos [8].

Existem diversos algoritmos de agrupamento baseados em abordagens distintas, e por isso os métodos podem gerar grupos diferentes para o mesmo conjunto de dados. A clusterização também pode ser utilizada para a detecção de *outliers* quando possível identificar grupos discrepantes do restante dos dados.

Enfim, o campo de aplicação da clusterização é vasto, abrangendo pesquisas em mineração de dados, estatística, aprendizado de máquina, biologia, marketing, etc. Acrescente-se que, com o volume crescente de dados que são produzidos atualmente, a análise por agrupamento é uma técnica cuja aplicação deve ser considerada pelo analista.

Comumente, para a efetivação dos cálculos de distância entre os objetos, os algoritmos de clusterização utilizam atributos numéricos. Entretanto, o conjunto de dados objeto de estudo neste trabalho (transações bancárias) é essencialmente categórico. De todos os atributos existentes no conjunto de dados, somente a variável "Valor" (valor da transação) é genuinamente. Embora os campos "Banco", "Agência" e "Conta" sejam expressos em números, eles atuam apenas como elementos identificadores e, portanto, são categóricos.

Uma alternativa teoricamente possível e frequentemente utilizada seria transformar os dados categóricos textuais em variáveis numéricas para então serem submetidas à tarefa de agrupamento (*on-hot encoding*, por exemplo). A abordagem é válida, contudo o resultado final poderia ser de difícil interpretação ou contextualização em relação ao propósito investigativo. Segundo lecionam Han et al. [8], usuários desejam que o resultado

do agrupamento seja interpretável, compreensível e útil, além de ser necessário que o agrupamento seja contextualizado com interpretações ou semânticas específicas.

Para fins de pesquisa e experimentação, optou-se em adotar um método de agrupamento especialmente aplicável a dados heterogêneos composto por atributos categóricos e numéricos.

Foi escolhido o algoritmo *k-Prototypes* proposto por Huang [7]. O algoritmo é o resultado da integração dos algoritmos *k-Means*, amplamente conhecido e utilizado em tarefas de clusterização, e o algoritmo *k-Modes*. Apesar de bastante utilizado em análise de agrupamentos, o *k-Means* é aplicável somente em dados numéricos. Já o método *k-Modes* proposto por Huang [7] permite a clusterização de dados categóricos. Ao invés de calcular a distância euclidiana para os objetos do cluster e centróides, o *k-Modes* adota a medida de dissimilaridade entre os objetos, ou seja, o quanto as observações são diferentes entre si. Dessa forma, a distância é representada pelo número de discrepâncias entre cada objeto e o centro do cluster. O *k-Prototype* idealizado pelo mesmo autor nada mais é do que uma modificação do *k-Modes* combinada com o algoritmo *k-Means* possibilitando o agrupamento de dados mistos ou heterogêneos (variáveis numéricas e categóricas). Assim como o *k-Means*, o *k-Prototype* adota a estratégia de particionamento dos dados para realizar os agrupamentos. A aplicação prática do algoritmo é evidente, haja vista que geralmente os bancos de dados do mundo real armazenam dados mistos. O algoritmo também é capaz de lidar com grandes conjuntos de dados.

O *k-Prototypes* considera a distância dos atributos numéricos efetuando o cálculo da distância euclidiana (assim como o *k-Means*), mas também mede a distância entre os recursos categóricos por meio do número de categorias correspondentes.

Sendo assim, a dissimilaridade entre dois objetos de tipo misto X e Y descritos por atributos A^r_1, A^r_2, ... A^r_p, A^c_{p+1},..., A^c_m pode ser medida pela equação:

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

O primeiro termo é a distância euclidiana para os atributos numéricos e o segundo termo é a medida de dissimilaridade de correspondência simples das variáveis categóricas [7].

3.2.1 Preparação dos dados

Foi necessário implementar algumas medidas de pré-processamento a fim de preparar os dados para a aplicação do algoritmo. Primeiramente, foram selecionadas apenas as transações com valores superiores a R$2,00, excluindo-se assim operações com valores irrisórios, irrelevantes do ponto de vista investigativo e que poderiam gerar "ruídos" no resultado. Também foram desconsideradas as variáveis que apresentavam grande variabilidade de valores e que poderiam impactar negativamente a tarefa de agrupamento, como "Data do lançamento" e dados de origem/destino dos recursos transacionados. Acrescente-se também que os campos 'Banco O/D', 'Agência O/D', 'Conta O/D', 'CPF/CNPJ O/D' nem sempre são integralmente preenchidos pelas instituições financeiras e também não foram considerados em razão da presença de valores ausentes.

Ao final, as variáveis de interesse selecionadas resumiram-se em: 'Banco', 'Agência', 'Conta', 'CNAB', 'Descrição CNAB', 'Valor'.

Por fim, os dados dos atributos banco, agência e conta bancária foram devidamente anonimizados para preservar o sigilo das informações.

Definidas as colunas categóricas, o próximo passo foi transformar o dataset de trabalho em um vetor, estrutura adequada para a aplicação do algoritmo.

O experimento foi realizado em duas etapas idênticas, uma para as operações a crédito e outra para os débitos.

3.2.2 Experimento 2

A primeira fase do experimento consistiu no agrupamento de operações a crédito. O modelo foi construído testando arbitrariamente uma quantidade máxima de 10 *clusters* (grupos), sendo que o algoritmo identificou 8 grupos válidos. Vale ressaltar que quanto maior o número de grupos, maior o tempo de processamento, fato que deverá ser levado em consideração ao trabalhar com conjunto de dados extenso.

Para escolha do número ideal de *clusters* foi adotada uma adaptação do Método de Elbow (ou "método do cotovelo"). Ao invés da soma dos quadrados das distâncias euclidianas, o algoritmo *k-Prototype* fornece uma função de custo (ou função de perda) que retrata o "custo" computacional associado ao evento da combinação de variáveis numéricas e categóricas. Traçando um gráfico relacionando o número de *clusters* com o custo exigido para a tarefa é possível identificar visualmente o número ideal de grupos quando o "cotovelo" da linha se tornar menos acentuado, ou seja, quando a inclinação da curva for menos íngreme. A partir desse ponto o incremento de grupos não mais impactará significativamente para o resultado final (figura 4).

Para os dados de trabalho, 4 *clusters* aparentam ser suficientes para um agrupamento satisfatório.

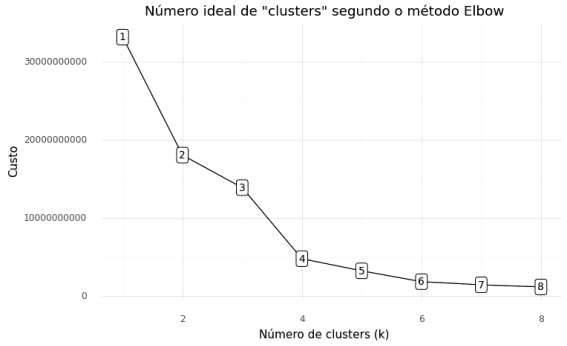

Figura 4: Número ideal de 4 grupos para operações a crédito.

3.2.3 Resultado e constatações

Após o processamento, o resultado com os quatro grupos pode ser visualizado na Tabela 2. Particularmente para este caso, o algoritmo identificou um grupo com uma única observação, justamente o registro com valor discrepante detectado no experimento 1 deste trabalho. Tal constatação corrobora o fato que, sob certas condições, técnicas de agrupamento de dados também podem ser úteis para realçar registros anômalos em um conjunto de dados.

Tabela 2: Grupos resultantes - operações a crédito

Grupo	Elementos
3	8512
4	1303
2	72
1	1

Os centróides de cada cluster estão retratados na figura 5.

	Banco	Agência	Conta	CNAB	Descrição	Valor	Natureza
0	Banco 3	Agência 18	Conta 38	205	LANÇAMENTO AVISADO	111117.1000000005	C
1	Banco 4	Agência 5	Conta 19	201	DEPÓSITOS	13625.737638888893	C
2	Banco 4	Agência 9	Conta 20	205	LANÇAMENTO AVISADO	263.58263862781996	C
3	Banco 4	Agência 5	Conta 16	201	DEPÓSITOS	2122.4441135840348	C

Figura 5: Centroides para o agrupamento de operações a crédito.

A representação gráfica do resultado pode ser visualizada na figura 6. Note-se que o grupo 3 apresenta a maior quantidade de operações a crédito (8512) e englobando todos os tipos de lançamentos, porém os valores movimentados são relativamente baixos. Denota-se que, em linhas gerais, o padrão financeiro global é mediano:

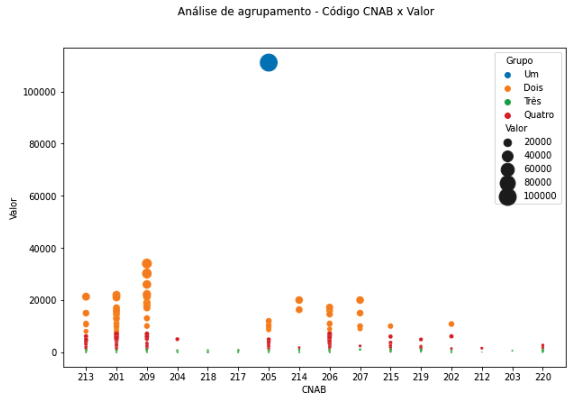

Figura 6: Gráfico Código CNAB x Valor (créditos).

Todavia, o grupo 2 contém significativa quantidade de lançamentos do tipo 209 – "transferência interbancária (DOC, TED)"e do tipo 201 – "depósitos", em valores que ultrapassam R$ 20 mil. Tal constatação poderia ser melhor investigada a fim de se averiguar se houve motivos justificáveis para tais operações.

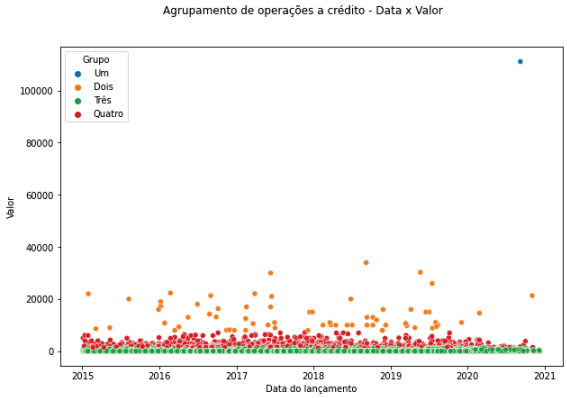

Figura 7: Gráfico Data do lançamento x Valor (créditos).

Note-se também que os valores mais expressivos (grupo 2) foram movimentados nos anos de 2018 e 2019 (figura 7), atingindo o patamar mais alto em 2020 (*outlier* do grupo 1). Conforme o contexto da investigação, a atenção poderia ser direcionada para esses anos.

A segunda etapa do experimento implementou basicamente o mesmo procedimento no tocante à preparação dos dados, porém foram selecionadas as transações a débito.

Optou-se, segundo o método de Elbow, pelo número ideal de 5 grupos (figura 8).

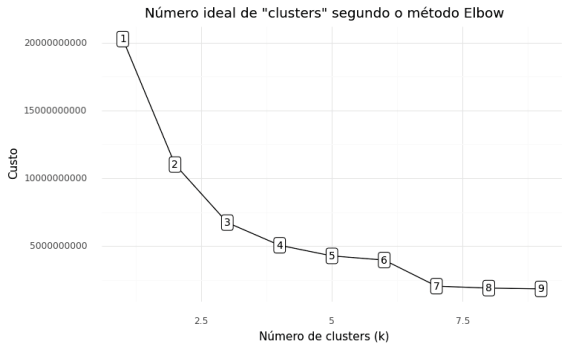

Figura 8: Número ideal de grupos para operações a débito.

Os grupos resultantes estão consignados na tabela 3. O grupo 5 apresentou o maior número de observações, com 9159 instâncias.

Tabela 3: Grupos resultantes - operações a débito

Grupo	Elementos
5	9159
2	3174
3	607
1	109
4	17

Observa-se na figura 9 os centroides dos grupos resultantes:

	Banco	Agência	Conta	CNAB	Descrição	Valor	Natureza
0	Banco 4	Agência 5	Conta 19	101	CHEQUES	6825.942110091755	D
1	Banco 4	Agência 5	Conta 16	101	CHEQUES	761.6549023314425	D
2	Banco 4	Agência 5	Conta 16	102	ENCARGOS	2169.939192751235	D
3	Banco 4	Agência 5	Conta 16	101	CHEQUES	22854.52470588247	D
4	Banco 4	Agência 5	Conta 16	104	LANÇAMENTO AVISADO	133.25718637405836	D

Figura 9: Centroides para o agrupamento de operações a débito.

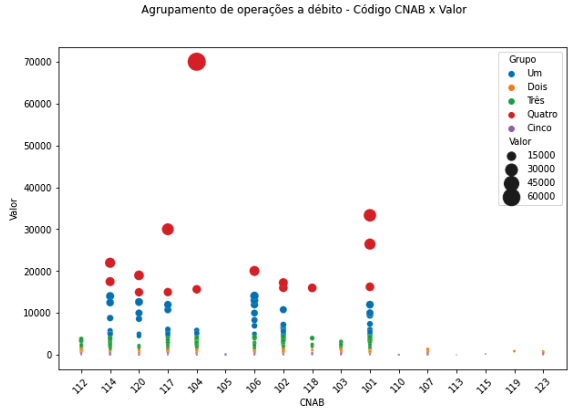

Figura 10: Gráfico Código CNAB x Valor (débitos).

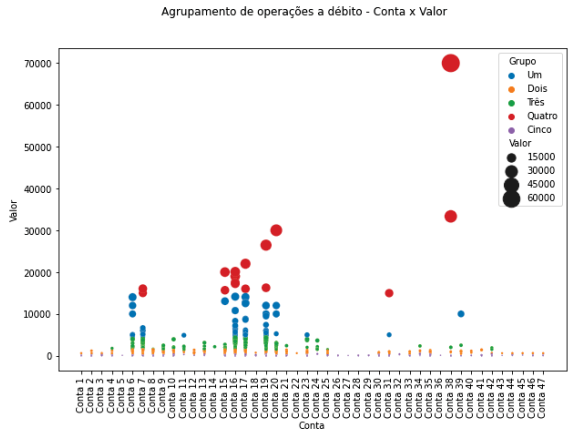

Figura 11: Gráfico Conta x Valor (débitos).

3.2.4 Resultado e constatações

É possível notar na figura 8 que a "Conta 16" exerce papel relevante do conjunto de dados, uma vez que apresentou quatro operações representativas dos centroides dos cinco grupos resultantes. Tal fato é corroborado no gráfico da figura 11, sendo possível visualizar que a conta apresentou movimentação significativa em relação às demais. Em um contexto investigativo, seria prudente concentrar as atenções nesta conta bancária.

Observa-se também que algumas contas apresentaram movimentação a débito com valores superiores a cerca de R$ 18 mil. São contas com transações pertencentes ao grupo 4, o menor grupo.

3.3 Regras de associação com *Apriori*

O descobrimento de regras de associação consiste em localizar um conjunto de itens que ocorram simultaneamente e de forma frequente em um banco de dados [9]. A aplicação clássica de regras de associação é a descoberta de produtos que sejam frequentemente vendidos de forma conjunta, como em vendas realizadas em mercados (carrinho de compras), por exemplo. Desse modo, a ocorrência de um conjunto de itens presentes no antecedente da regra leva a propensão de compra dos itens no consequente da regra. Esta técnica também é especialmente útil para detectar padrões em grandes conjuntos de dados.

Formalmente, uma regra de associação é uma implicação na forma $X \rightarrow Y$, onde X (antecedente) e Y (consequente) são conjuntos de itens tais que $X \cap Y = \oslash$. A intersecção vazia entre antecedente e consequente assegura que não sejam extraídas regras em que um item esteja associado a ele mesmo. [9].

Segundo Goldschmidt and Passos [9], "uma associação é considerada frequente se o número de vezes em que a união de conjuntos de itens ($X \cup Y$) ocorrer em relação ao número total de transações do banco de dados for superior a uma frequência mínima (denominada suporte mínimo) que é estabelecida em cada aplicação".

A força de uma associação pode ser medida pelo suporte e a confiança. O suporte determina a frequência na qual um regra de associação é aplicável ao conjunto de dados. [10]. Seja N o total de transações, o suporte é representado pela seguinte fórmula [10]:

$$Suporte, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

Em contrapartida, a confiança é a frequência na qual os elementos de Y aparecem em transações que contenham X. Para uma determinada regra $X \rightarrow Y$, "quanto maior a confiança, maior a probabilidade de que Y esteja presente em transações que contenham X" [10]. É expressa pela fórmula:

$$Confiança, c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Por fim, o *lift* indica a força de associação entre os elementos da regra, isto é, representa a chance de ocorrer Y se X estiver presente. É a razão entre a confiança (c) da regra e o suporte (s) do conjunto de itens no consequente da regra [10]:

$$Lift = \frac{c(X \rightarrow Y)}{s(Y)}$$

- Se lift = 1, então não há correlação no conjunto de dados, ou seja, X e Y são independentes. Indica que X e Y sempre aparecem juntos;
- Se lift > 1, então há uma correlação positiva no conjunto de dados, isto é, X tem efeito positivo sobre a ocorrência de Y. Indica que X e Y aparecem mais frequentemente juntos do que o esperado;

- Se lift < 1, então há uma correlação negativa no conjunto de dados, significando que X tem efeito negativo sobre a ocorrência de Y. Indica que X e Y aparecem no conjunto com menos frequência.

Sendo assim, a formulação do problema da mineração de regras de associação pode ser enunciado como: dado um conjunto de transações T, elabore todas as regras de associação a partir da especificação de um suporte e confiança mínimos.

Existem diversos algoritmos desenvolvidos para a descoberta de associações. Para a realização do experimento foi adotado o algoritmo *Apriori* dada sua sua popularidade e pelo fato de que os demais algoritmos possuem uma estrutura comum nele inspirada. Goldschmidt e Passos [10] asseveram que os algoritmos baseiam-se na propriedade de anti-monotonicidade do suporte, na qual "Um *k-itemset* somente pode ser frequente se todos os seus (k-1)-subconjuntos forem frequentes".

O experimento foi baseado em duas estratégias para a geração das regras de associação: regras envolvendo a conta bancária do titular em relação à conta bancária de origem ou destino dos recursos movimentados nas operações financeiras e regras envolvendo o tipo de lançamento bancário em relação à origem/destino.

Para evitar "ruídos" desnecessários e regras irrelevantes, foram consideradas apenas as operações com valores acima de R$2,00. Além disso, os dados de banco, agência e conta foram concatenados resultando em duas colunas, uma para as contas dos titulares investigados e outra para as contas origem/destino. Lembrando que só foram consideradas as operações cuja origem/destino do recurso foram identificadas com valores do banco, agência e conta O/D. Por fim, todos os valores de banco, agência e conta foram anonimizados.

Os valores de suporte mínimo, confiança mínima e *lift* mínimo foram parametrizados em 0,003, 0,2 e 3, respectivamente.

3.3.1 Resultado e constatações

A primeira parte do experimento resultou em 21 regras de associação. A título de exemplo, foi selecionada a regra com o maior índice de suporte:

- Regra: Banco O/D 8-Agência O/D 102-Conta O/D 157 → Banco 4-Agência 6-Conta 10
  Suporte: 0.09336455893832943
  Confiança: 0.8978978978978979
  Lift: 9.50584468766287

A interpretação da regra revela que em 9,3% de todas as transações com origem/destino identificadas no conjunto de dados, a conta bancária O/D nº 157 transacionou de alguma forma com a conta bancária nº 10 titularizada por algum investigado. 89,8% das operações da conta nº 10 com origem/destino identificadas foram efetuadas com a conta nº 157. Ademais, a conta nº 10 tem 9,5 vezes mais probabilidade de figurar no conjunto de transações quando a conta bancária O/D for a de nº 157. A regra revelou e existência de vínculo entre contas bancárias que merece atenção por parte do analista, haja vista a frequência com que as contas transacionaram entre si.

Ordenando o conjunto de regras pelo índice de *lift* mais alto, é possível observar a seguinte associação:

- Regra: Banco O/D 11-Agência O/D 122-Conta O/D 1163 -> Banco 3-Agência 15-Conta 31
  Suporte: 0.003434816549570648
  Confiança: 1.0
  Lift: 291.1363636363636

Nota-se a forte correlação positiva entre as contas nº 31 e nº 122 (*lift* ≅ 291), indicando que todas as transações da conta nº 31 com origem/destino identificados ocorreram com a conta bancária nº 1163 (confiança de 100%).

A segunda parte do experimento levou em consideração os tipos de lançamentos bancários realizados em relação à conta bancária de origem/destino dos valores movimentados. Foram mantidos os mesmos valores para os parâmetros de suporte mínimo, confiança mínima e *lift*.

Após o processamento, o algoritmo produziu 36 regras de associação. Mais uma vez selecionando a regra com maior índice de suporte temos:

- Regra: Banco O/D 11-Agência O/D 19-Conta O/D 48 → SAQUE ELETRÔNICO
  Suporte: 0.028727556596409055
  Confiança: 0.9945945945945945
  Lift: 33.705705705705704

A regra revela que em 2,8% das operações realizadas houve saque eletrônico envolvendo a conta nº 48, do Banco 11. A probabilidade de ocorrer esse lançamento com a citada conta foi de 99,4%. Há 33 vezes mais possibilidade do saque eletrônico ocorrer quando a conta nº 48 estiver envolvida.

Outra regra com o segundo maior índice de suporte consistiu em:

- Regra: TRANSFERÊNCIA ENTRE CONTAS → Banco O/D 8-Agência O/D 109-Conta O/D 631
  Suporte: 0.02529274004683841
  Confiança: 1.0
  Lift: 6.784957627118644

É possível notar que em 2,5% das operações bancárias houve transferência de valores oriundos ou destinados à conta bancária nº 631. Há 100% de certeza que tal conta figurou nos dados bancários quando ocorreu esse tipo de lançamento.

4 Conclusão

Por todo o exposto, conclui-se que o resultado do trabalho foi positivo, pois demonstrou a viabilidade da aplicação de técnicas de aprendizado de máquina não

supervisionado na análise de dados bancários como instrumento de auxílio à investigação. Vale ressaltar que a intenção das análises foi extrair conhecimento dos dados de forma geral, despertando a atenção do analista para pontos de interesse que poderão ser melhor averiguados com maior profundidade em etapa posterior da investigação, possivelmente adotando-se outras ferramentas e métodos de análise. Nesse contexto é de fundamental importância a experiência do profissional e seu conhecimento sobre os fatos subjacentes à investigação.

Referências

[1] Brasil. Lei nº 9.613, de 03 de março de 1998. *Diário Oficial [da] República Federativa do Brasil*, 1998. ISSN 1677-7042. URL http://www.planalto.gov.br/ccivil_03/leis/L9613compilado.htm.

[2] Carla Veríssimo De Carli. *Lavagem de Dinheiro: Prevenção e Controle Penal*. Verbo Jurídico, 2013. ISBN 9788576994459.

[3] Ahmad Salehi, Mehdi Ghazanfari, and Mohammad Fathian. Data mining techniques for anti money laundering. *International Journal of Applied Engineering Research*, 12:10084–10094, 01 2017.

[4] Dattatray Vishnu Kute, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE Access*, 9:82300–82317, 2021.

[5] Ebberth Paula, Marcelo Ladeira, Rommel Carvalho, and Thiago Marzagao. Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. pages 954–960, 12 2016. doi: 10.1109/ICMLA.2016.0172.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

[7] Zhexue Huang. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998. doi: 10.1023/a:1009769707641. URL https://doi.org/10.1023%2Fa%3A1009769707641.

[8] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2012. ISBN 0123814790.

[9] Ronaldo Goldschmidt and Emmanuel Passos. *Data Mining: um guia prático*. Elsevier, Rio de Janeiro, 2005.

[10] Pang-Hing Tan, Michael Steinbach, and Vipin Kumar. *Introdução ao Data Mining - Mineração de Dados*. Editora Ciência Moderna, Rio de Janeiro, 2009. ISBN 9788573937619.