Universidade de Brasília
Departamento de Estatística

Trabalho de Conclusão de Curso 2

Classificação de Alvos de Relatórios de Inteligência Financeira Utilizando Florestas Aleatórias e Medidas de Centralidade de Rede

Olivia Ziller e Silva

Orientador:
Prof **André Cançado**

Brasília
2/2019

Olivia Ziller e Silva
Estatística

Trabalho de Conclusão de Curso 2

Classificação de Alvos de Relatórios de Inteligência Financeira Utilizando Florestas Aleatórias e Medidas de Centralidade de Rede

Orientador:
Prof **André Cançado**

Relatório Final, apresentado como Trabalho de Conclusão de Curso da Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2/2019

Agradecimentos

Dentre as heranças e ensinamentos que a minha saudosa vida de atleta me deixou, está o dizer de que “eu seguro minha mão na sua e uno meu coração ao seu, para que juntos possamos conseguir aquilo que eu não posso conseguir sozinho”. A menção a esse clássico dos shows de fim de temporada me permite homenagear a imensa contribuição que o esporte, a patinação artística as pessoas que eles me trouxeram tiveram na construção de quem eu sou, e dizer que este trabalho não existiria sem muitas mãos e corações que me acompanharam nesta caminhada. Dito isto, não poderia deixar de agradecer àqueles foram essenciais para a conclusão deste longo ciclo que aqui se encerra.

À minha família, em especial meus pais a quem eu devo minhas vitórias. Muito obrigada por todo suporte, amor e carinho. Eu não estaria aqui e não seria nada se não fosse por vocês. Aos meus irmãos, Marta e Gabriel, meus cúmplices de vida, que são as duas pessoas no mundo com quem não tenho dúvidas de que posso contar.

Ao Rodrigo, meu amor, parceiro, companheiro de todas as horas, por todo apoio e amparo nos momentos de dificuldade, e torcida pelas vitórias ao longo do curso e da vida.

À Li, Ana, Mari, Sofs, Ju, Dora, Sólon, João, Marcola e Matheus pela amizade, por me manterem sã e me fazerem rir quando eu quis chorar.

Aos meus chefes e superiores Clesito, Rochelle, Mônica, Caetano e Ana Amélia, que confiaram em mim e me deram a oportunidade de trabalhar em uma instituição de tamanha importância e com uma função que respeito e acredito. Aos meus colegas do COAF. Em especial Alexandre e Virgínia, pela amizade, parceria, e por terem sido verdadeiros mentores. Sem vocês eu definitivamente não teria chegado até aqui. Espero que algum dia alguém possa ter por mim a gratidão que eu tenho por vocês. Muito obrigada.

Aos professores da UNB e aos funcionários do EST, em especial meu orientador Prof. André Cançado por toda ajuda, paciência, disponibilidade e flexibilidade com uma libriana indecisa com seu tema.

Aos meus colegas (futuros) estatísticos, em especial Bia, Gustavo’s, Luiz, Rodrigo, Fê, Carol, Gabi, Ana Vitória, Gongs e Gabrieis pelo companheirismo e por terem trazido um pouco de luz e humor aos momentos de terror e pânico que a vida acadêmica pode nos proporcionar.

A todos vocês, deixo aqui o meu muito obrigada.

Resumo

Reduzir esforço humano e fornecer auxílio e direcionamento na tomada de decisões utilizando modelagem estatística vem se tornando cada vez mais útil e necessário. Neste trabalho são utilizados conceitos de redes sociais, medidas de grafos e florestas aleatórias para buscar e classificar pessoas físicas e jurídicas como possíveis alvos de relatórios de inteligência financeira do Conselho de Controle de Atividades Financeiras (COAF), a Unidade de Inteligência Financeira do Brasil. Toda a extração das variáveis, criação da base de dados, análises e desenvolvimento do modelo foram feitas utilizando o SAS.

Palavras-chave: Modelagem Estatística, Redes Sociais, Grafos, Florestas Aleatórias, Alvos, COAF, SAS.

Abstract

The use of statistical models have become ever more useful and necessary to reduce human effort and to provide assistance and guidance in decision-making activities. This paper uses concepts regarding social network analysis, graph theory, and random forests to classify natural and legal persons as possible targets of a Financial Intelligence Report from the Brazilian Financial Intelligence Unit (FIU). The SAS software was used all through the project, from variables extraction and datasets creation, to analysis, development and selection of the final model.

Palavras-chave: Statistical Models, Social Networks, Graph Theory, Random Forests, target, Brazilian Financial Intelligence Unit - FIU, SAS.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Quadros

1 Introdução

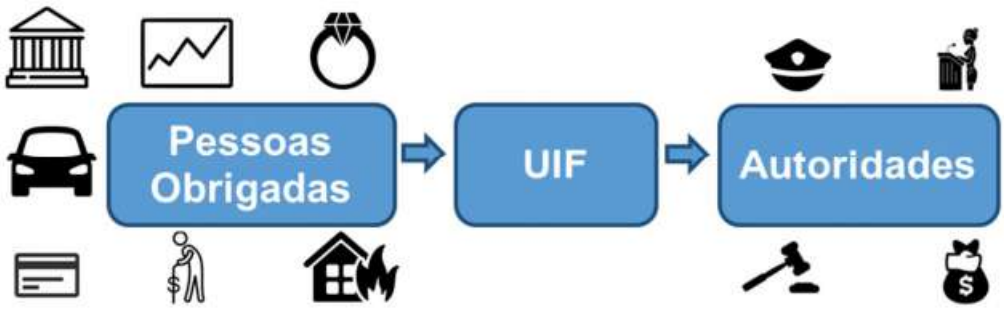
Solucionar problemas de classificação estatisticamente vem se tornando interesse nos mais diversos setores da sociedade. Considerando o aumento da complexidade dos problemas que surgem e a vasta quantidade de dados a serem avaliados, técnicas de caráter numérico são cada vez mais valorizadas. Seja visando automatizar um processo por completo ou reduzir o esforço associado ao seu desenvolvimento, ferramentas matemáticas e computacionais são auxílio relevante na tomada de decisões.

No processo de combate e prevenção à lavagem de dinheiro em especial, fornecer ferramentas que reduzam e acelerem o trabalho humano investido no reconhecimento de ilícitos pode ser crucial na identificação de atividades criminosas.

Nesse contexto, introduz-se o Conselho de Controle de Atividades Financeiras (COAF), Unidade de Inteligência Financeira (UIF) do Brasil, que tem como missão produzir inteligência financeira e proteger os setores econômicos contra a lavagem de dinheiro e o financiamento do terrorismo. [1]

Uma de suas atribuições é receber, examinar e identificar ocorrências suspeitas de atividade ilícita e (quando identificados fundados indícios) informá-las às autoridades competentes para instauração de procedimento. Neste encargo, surge um processo cujo fluxo se inicia no recebimento de informações de movimentações financeiras suspeitas enviadas pelos chamados setores obrigados (como instituições financeiras, empresas seguradoras e comércio de imóveis). Esta informação é denominada comunicação, e provém dela a análise que visa determinar se ali existem indícios de lavagem de dinheiro, de financiamento do terrorismo ou de outros crimes que levem à elaboração de um Relatório de Inteligência Financeira (RIF). Os RIFs elaborados pelo COAF são destinados às autoridades competentes para subsidiar eventuais procedimentos investigativos.

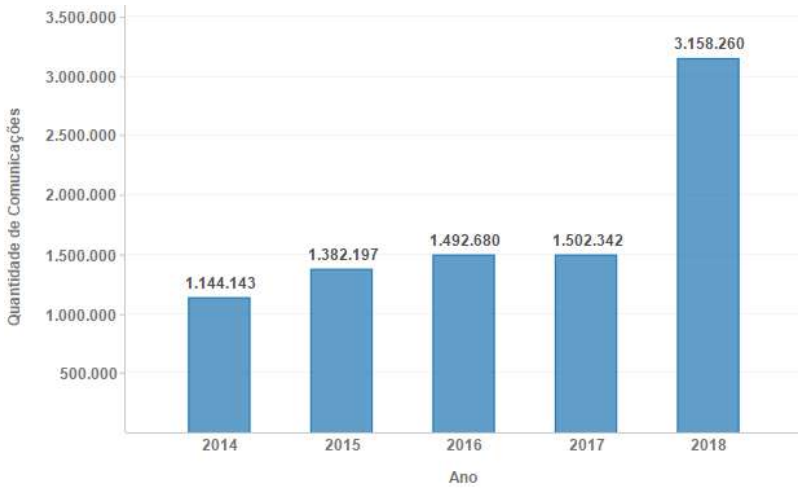
Figura 1: Ilustração do Fluxo do Processo de Produção de Inteligência Financeira



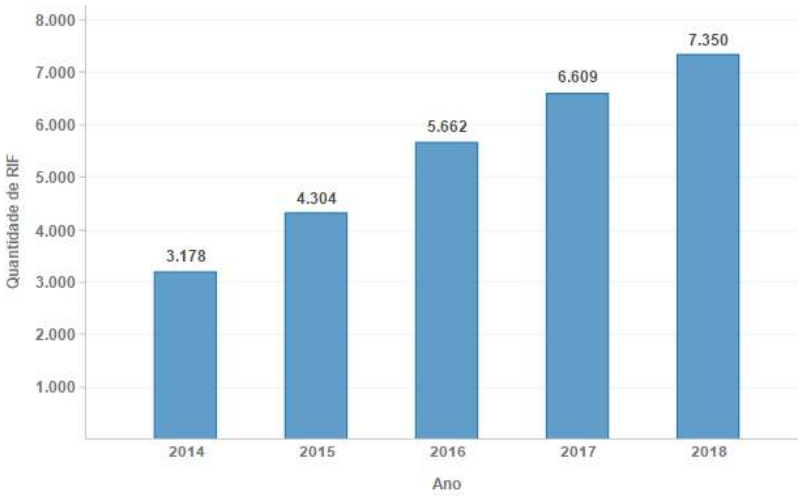
Fonte: Relatório “O que é o COAF?”[2]

Como pode ser observado nos gráficos 2a e 2b, é evidente que nos últimos anos a quantidade de comunicações recebidas pelo COAF e Relatórios de Inteligência produzidos tem crescido significativamente.

Figura 2: Quantidade de Comunicações Recebidas e RIFs Produzidos por Ano



(a) Comunicações



(b) RIFs

Fonte: "COAF em Números"[1]

Com isso, o esforço despendido no processo da produção de inteligência, descrito anteriormente, se manifesta de diversas formas. Dentre elas estão:

1. A necessidade de se avaliar e identificar se existe em toda essa extensa dimensão de informações suspeitas uma situação descrita que deva ser submetida a análise para a produção de um RIF.

2. A produção efetiva dos Relatórios de Inteligência Financeira, que se tornam mais frequentes, trabalhosos e complexos conforme se expande a quantidade de dados a serem considerados.

Em resposta ao primeiro item, foi desenvolvido um modelo estatístico que analisa o nível de risco de uma determinada comunicação. O modelo avalia e atribui à situação descrita uma probabilidade que determina se a comunicação será submetida à análise para a produção de um RIF.

Na intenção de proporcionar assistência ao produto descrito no segundo item, que trata da produção dos Relatórios de Inteligência Financeira, surgiu o objeto de estudo do trabalho em questão.

Uma das etapas da confecção de um RIF é determinar as pessoas físicas ou jurídicas que constarão como principais relacionadas do documento, que seriam as protagonistas do possível evento ilícito constatado.

Instigando-se nesta necessidade, enxerga-se propósito em desenvolver um modelo estatístico que seja capaz de classificar indivíduos como alvos ou não alvos de um Relatório de Inteligência Financeira produzido pelo COAF. Neste trabalho em especial, serão utilizados modelos de florestas aleatórias.

No trajeto de identificação dos alvos de um relatório de inteligência surge a necessidade de buscar os candidatos a serem avaliados pelo modelo. Visualizar e reconhecer estes candidatos e a relação entre eles pode ser uma tarefa difícil dada a quantidade de informações e pessoas a serem a consideradas. Justifica-se então, a procura por um método de seleção automática destes indivíduos. A utilização de conceitos e métricas relacionados a redes sociais criminosas neste trabalho parte da necessidade de cumprir essa tarefa. De maneira simplória, buscou-se utilizar os relacionamentos identificáveis pelas bases disponíveis ao COAF para atribuir aos citados em uma comunicação (em que tenha sido identificado possível cometimento de ilícito) os possíveis relacionadas ao evento a ser relatado em um RIF.

Toda a extração, análise, construção da base de dados, treino, e teste dos modelos foi feita utilizando o software SAS.

2 Objetivos

2.1 Geral

De maneira geral o objetivo do trabalho foi fornecer auxílio técnico na produção de Relatórios de Inteligência Financeira, de maneira que se indicasse com fundamentos estatísticos os possíveis principais relacionados daquele documento.

2.2 Específicos

Especificamente, buscou-se:

- Avaliar a possibilidade da utilização de conceitos de redes sociais e grafos para buscar os possíveis envolvidos dos relatórios.
- Investigar a viabilidade e eficiência de conceitos e medidas de teoria de grafos como variáveis explicativas para a resposta do modelo proposto.
- Construir a base de dados a ser utilizada para treino do modelo.
- Utilizar florestas aleatórias para construir um modelo de classificação de indivíduos como alvos ou não alvos de um relatório de inteligência financeira.

3 Revisão de Literatura

Será apresentada nesta seção uma breve revisão da literatura envolvendo os principais conceitos e técnicas utilizadas no desenvolvimento deste trabalho. A fundamentação exposta a respeito de redes e grafos foi utilizada nas etapas de seleção dos candidatos e extração de algumas das variáveis explicativas a serem avaliadas no momento da modelagem. Os conceitos que dizem respeito a problemas de classificação, árvores de decisão, florestas aleatórias e medidas de acurácia foram utilizados no desenvolvimento e escolha do modelo que tem como objetivo classificar as pessoas físicas e jurídicas como alvos ou não alvos de um relatório de inteligência financeira produzido pelo COAF.

3.1 Redes

Por sua capacidade de representar por meio da modelagem problemas de natureza real, o estudo de redes vem sendo utilizado em diversas áreas (Wasserman, 1994) [3]. Em sua forma mais simples, uma rede é dada por uma coleção de pontos ligados por linhas. Na área matemática, os pontos são chamados de vértices ou nós, e as linhas de arestas. Em "Networks: An Introduction", Mark Newmam dá exemplos de vértices e arestas para determinados tipos de rede.

Tabela 1: Vértices e Arestas em Redes

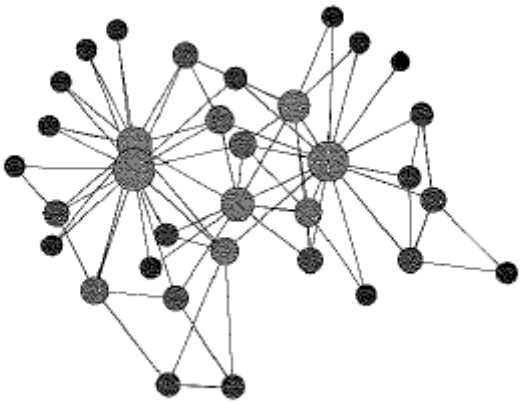
Rede	Vértice	Aresta
Internet	Computador	Cabo ou conexão
Rede de Amizades	Pessoas	Amizade
Rede Neural	Neurônio	Sinapses
Rede de Citação	Artigo, Livro, ou Patente	Citação

Fonte: "Networks: An Introduction"[4]

Dentre as redes mais estudadas podemos destacar as redes sociais, compostas por pessoas, ou grupo de pessoas que estão ligados por um ou mais tipos específicos de interdependência, (Quintilha, 2010) [5]. Redes sociais são amplamente estudadas na sociologia, área que a encara como uma estrutura de laços entre atores de um determinado sistema social. Esses atores (que apresentamos como vértices) podem ser papéis, indivíduos, organizações, setores ou nações. Os laços (que apresentamos como arestas) podem ser dados por amizade, afeto, conversação, parentesco, autoridade, troca econômica, troca de informação ou qualquer outro atributo que consista a base de uma relação. (Nohria & Eccles, 1992)[6]

Um exemplo famoso de rede social vem da literatura sociológica de Wayne Zachary [7], em que foram analisados os padrões de amizade entre membros de um clube de Karatê de uma universidade dos Estados Unidos. Na representação gráfica da rede de amizades do clube de karatê, apresentada a seguir, observam-se os 34 vértices identificados no estudo e a estrutura de seus relacionamentos, representados pelas arestas e baseados na amizade entre os atores.

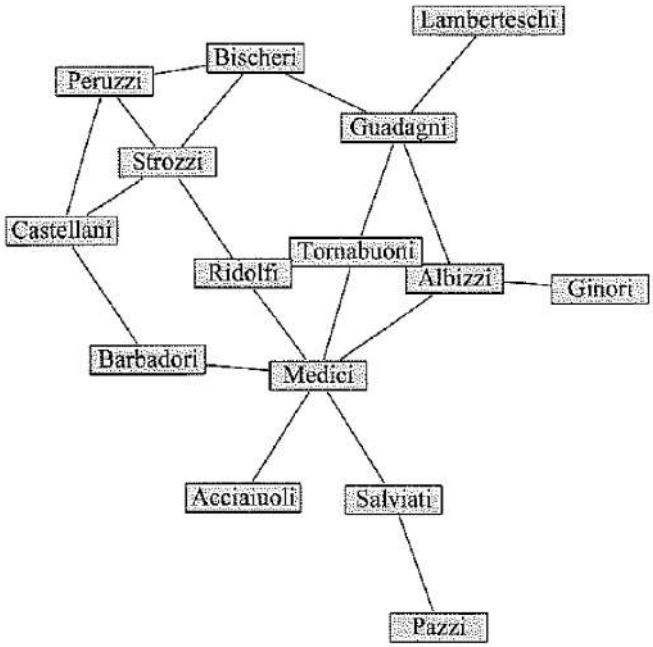
Figura 3: Rede de Amizade entre Membros do Clube de Karatê



Fonte: “Networks: An Introduction”[4]

Outro exemplo conhecido de redes sociais, exposto por Newman[4] é a rede de interações entre famílias influentes de Florença no século 15, estudada por Padgett e Ansell [8] com auxílio de registros históricos. Uma das redes resultantes desse estudo utiliza casamentos entre famílias como arestas. Sua representação gráfica é apresentada a seguir.

Figura 4: Rede de Casamentos entre Famílias Influentes de Florença no Século XV



Fonte: “Networks: An Introduction”[4]

É interessante notar que a família Medici encontra-se em uma posição central

da rede, tendo laços de casamento com 6 outras famílias influentes. Para Padgett e Ansell a manipulação destes tipos de relacionamento foi uma das maiores contribuições para colocar a família Medici em sua poderosa posição de dominância política e social em Florença.

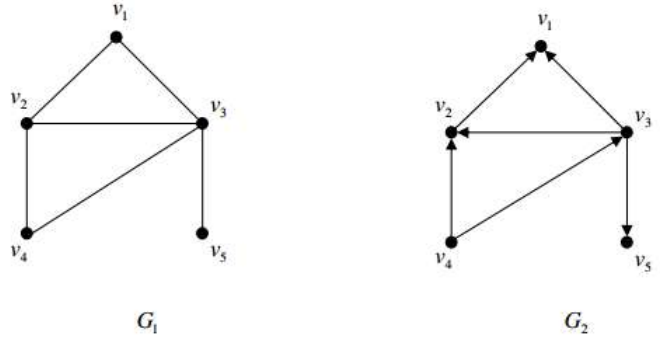
3.2 Teoria de Grafos

Uma rede pode ser analisada matematicamente de diferentes formas e com diferentes notações. Um dos modelos de análise de redes é dado por um objeto denominado grafo, utilizado para representar sua estrutura topológica (Quintilha, 2010) [5].

Um grafo $G = (V(G), E(G))$ consiste de um conjunto finito e não vazio de vértices $V = V(G)$ e um conjunto de arestas $E = E(G)$ formado por pares de elementos de V. O conjunto de arestas E induz uma conexão binária em elementos de V. Esta conexão é chamada de relação de adjacência dos vértices de G. O número de vértices n e o número de arestas m de G são respectivamente iguais a $|V(G)| = n$ e $|E(G)| = m$.(Quintilha, 2010)[5].

Grafos podem ser orientados ou não orientados. Um grafo orientado é definido quando um par de vértices v_i e $v_j \in V$ conectados por uma aresta tem sentido definido de v_i pra v_j ou vice-versa. Em um grafo não orientado, esse sentido não é definido. A imagem a seguir ilustra em $G1$ um grafo não direcionado, e em $G2$ um grafo direcionado.

Figura 5: Grafo Direcionado e Não Direcionado



Fonte: “Medidas de Centralidade em Grafos”(Quintilha, 2010) [5]

A notação teórica dos grafos é eficiente na utilização de métodos para investigação de centralidade e prestígio de um indivíduo na rede (Wasserman, 1994) [3]. Segundo Quintilha [5], a noção de centralidade em redes sociais foi introduzida por Bavelas em 1948, quando ele afirma que em um grupo de pessoas, um indivíduo que se encontra estrategicamente localizado no caminho mais curto de comunicação entre dois pares de indivíduos está numa posição mais central da rede.

As medidas de centralidade buscam medir a variação da importância dos

vértices, em função de alguns invariantes do grafo.

Seguem abaixo, definições de algumas medidas de centralidade e conceitos de grafos que serão utilizados no desenvolvimento do trabalho. Todas as medidas apresentadas podem ser encontradas em "Medidas de centralidade em grafos"[5], "Social Networks Analysis"[3] e "Data Mining: Conceitos,Técnicas, algoritmos, orientações e aplicações"[9].

3.2.1 Matriz de Adjacência

Seja G um grafo com n vértices. A matriz de adjacência $A(G)$ de G é uma matriz de ordem n cujas entradas são:

$$a_{ij} = \begin{cases} 1, & \text{se } (v_i, v_j) \in E \\ 0, & \text{Caso contrário} \end{cases}$$

3.2.2 Centralidade de Grau

Em um grafo não direcionado, a centralidade de grau de um vértice é determinada pela quantidade de arestas que o incidem (Quintilha, 2010) [5].

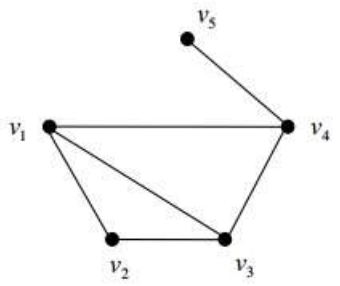
Seja G um grafo com n vértices e seja v_k um vértice de G. A centralidade de grau de v_k, denotada por d_k, é o número de arestas incidentes a v_k. Isto é:

$$d_k = \sum_{j=1}^{n} a_{kj}$$

Onde a_{kj} são elementos da matriz de adjacência $A(G)$.

Na figura abaixo os vértices v_1, v_3 e v_4 são aqueles com maior centralidade de grau (3).

Figura 6: Ilustração Centralidade de Grau



Fonte: "Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

3.2.3 Centralidade de Proximidade

A Centralidade de Proximidade (Closeness Centrality) é uma medida baseada na soma das menores distâncias entre um vértice e todos os outros da rede.

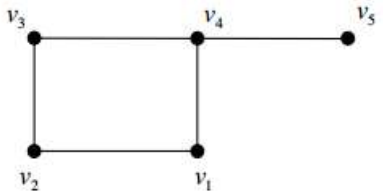
Seja G um grafo com n vértices e seja v_k um vértice de G. A centralidade de proximidade de v_k, denotada por c_{cc} é dada pelo inverso da soma das distâncias entre v_k e todos os demais vértices do grafo.

$$c_{cc} = \frac{1}{\sum_{j=1}^{n} dist(v_j, v_k)}$$

É interessante notar que v_k pode no mínimo estar a uma distância igual a 1 em relação a v_j num grafo conexo com n vértices, e no máximo estar ligado a n -1 outros vértices. (Quintilha, 2010) [5]

Na figura abaixo o vértices v_4 é o mais central segundo a centralidade de proximidade, considerando que à todas as arestas estão atríbuídas as mesmas distâncias.

Figura 7: Ilustração Centralidade de Proximidade



Fonte:"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

3.2.4 Centralidade de Intermediação

A Centralidade de Intermediação (Betweenness Centrality) mede quantos caminhos mais curtos, entre todos os pares de vértices da rede, passam por um determinado vértice (Quintilha, 2010) [5].

Seja G um grafo com n vértices e seja v_k um vértice de G. Considerando um par de vértices v_i e v_j em G tal que $i \neq j$, $i \neq k$ e $j \neq k$. A intermediação parcial de v_k com respeito a v_i e v_j é dada por:

$$b_{ij} = \begin{cases} 0, & \text{se não existir caminho entre } v_i \text{ e } v_j \\ \frac{g_{ij}(v_k)}{g_{ij}}, & \text{Caso contrário} \end{cases}$$

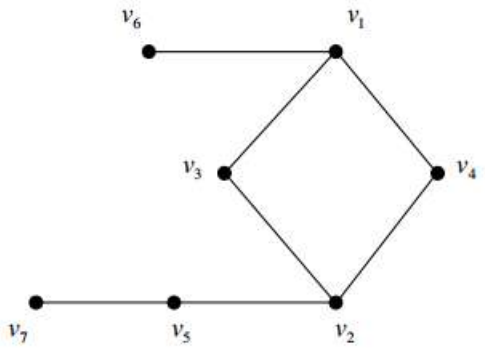
Onde g_{ij} é o número de menores distâncias entre v_i e v_j, e $g_{ij}(v_k)$ denota o número de menores distâncias que passam por v_k.

A centralidade de intermediação de um vértice v_k em um grau G com n vértices denotada por $c_B(v_k)$ é a soma das intermediações parciais de v_k em G. Ou seja,:

$$c_B(v_k) = \sum_{1 \leq i \leq j \leq n} b_{ij}(v_k)$$

Na figura abaixo o vértice v_2 é o mais central segundo a centralidade de intermediação.

Figura 8: Ilustração Centralidade de Intermediação



Fonte:"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

3.2.5 Centralidade de Autovetor

A Centralidade de Autovetor (Eigenvector Centrality), é baseada no conceito de autovalores e autovetores da matriz de adjacência do grafo G.

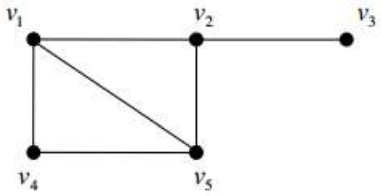
Seja G um grafo com n vértices e seja v_k um vértice de G. A centralidade de autovetor de v_k, denotada é dada por:

$$c_{eig}(v_k) = x_k$$

Onde x_k é a k-ésima coordenada do autovetor positivo unitário x associado ao índice do grafo.

Na ilustração abaixo os vértices v_1 e v_5 são aqueles com maior centralidade de autovetor.

Figura 9: Ilustração Centralidade de Autovetor



"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

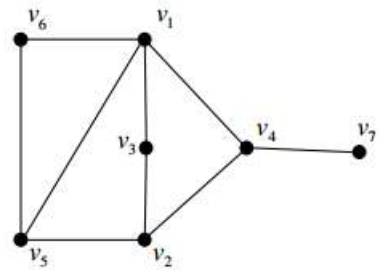
3.2.6 Ponto de Articulação

O grafo $G_1 = (V_1, E_1)$ é um subgrafo de G se $V_1 \subseteq V$ e $E_1 \subseteq E$. Além disso, se G_1 contém exatamente as arestas de G incidentes nos vértices de V_1 , G_1 é um subgrafo induzido de G. Um grafo G desconexo é formado por pelo menos dois subgrafos induzidos conexos, denominados componentes conexas de G.

Seja $G = (V, E)$ um grafo e v_k um vértice de G, se a remoção de v_k ocasionar um aumento na quantidade de componentes conexas, então v_k é denominado um ponto de articulação.

Na ilustração abaixo o vértice v_4 representa um ponto de articulação.

Figura 10: Ilustração Ponto de Articulação



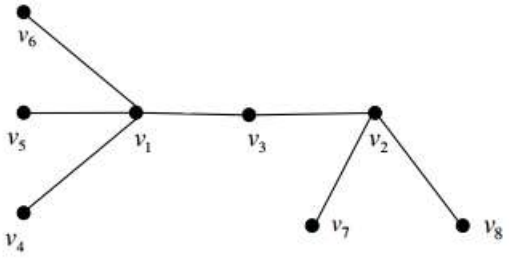
"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

3.2.7 Nó Terminal

Nós terminais são aqueles que encontram-se nas extremidades do grafo, e por consequência, tem 1 grau de centralidade (somente uma aresta incidente).

Na rede ilustrada abaixo, os vértices v_4, v_5, v_6, v_7 e v_8 representam nós terminais.

Figura 11: Ilustração Nó Terminal



Fonte:"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

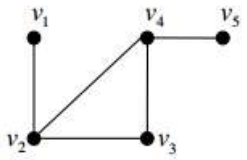
3.2.8 Maior Clique

Um clique de um grafo $G = (V(G), E(G))$ é um subgrafo em que todos os vértices são conectados entre si. Um clique máximo é aquele formado por vértices que não fazem parte de qualquer clique que seja maior do que ele mesmo.

O número de cliques máximos em um grafo particular pode crescer exponencialmente conforme se aumenta o número de nós (Newman, 2010) [4]. Por isso, o conceito de maior clique costuma ser utilizado. O maior clique do grafo é aquele que contém o maior número de vértices.

Na rede ilustrada abaixo, os vértices v_2, v_3 e v_4 formam um clique.

Figura 12: Ilustração Cliques



Fonte:"Medidas de Centralidade em Grafos"(Quintilha, 2010) [5]

3.2.9 Grafos Valorados

Há certas redes que quando modeladas por grafos, faz-se necessário avaliar não somente as possíveis ligações entre os pares de vértices como também a intensidade de tais ligações. Assim, precisamos atribuir valores às arestas. Grafos com essa composição, são chamados grafos valorados.

Um grafo é valorado se existe uma função w que relaciona $E(G)$ a valores numéricos, referindo-se cada número como o peso da aresta, denotado por ω. Estes valores podem ser custos, distâncias, tempo gasto no percurso, confiabilidade da transmissão e serão denominados nesse trabalho como peso das arestas.

Algumas das medidas de centralidade apresentadas podem ser utilizadas em grafos valorados. Neste caso, a centralidade de grau é dada pela soma dos pesos das arestas incidentes e a centralidade de proximidade é dada pelo inverso da soma dos pesos das arestas referentes a menor distância que liga pares de vértices. A centralidade de autovetor passa a considerar uma matriz onde as entradas correspondem aos valores de cada aresta, denominada matriz dos pesos. (Quintilha, 2010)[5].

3.3 Árvores de Decisão

Árvores de decisão são um tipo de modelo não paramétrico e supervisionado, que utiliza um conjunto de dados com respostas conhecidas para criar um método de previsão que pode ser utilizado para antecipar novos resultados. É uma formalização de um processo de decisão que utiliza uma sequência de perguntas atribuídas às variáveis explicativas para classificar uma variável resposta (Agresti, 2011)[10].

Árvores de decisão podem ser aplicadas a problemas de classificação (em que se busca avaliar uma resposta categórica) ou de regressão (em que se busca avaliar uma resposta numérica). Como o trabalho em questão trata de uma resposta categórica binária, será abordada a teoria que se refere a árvores de classificação.

Uma árvore de classificação prevê que cada observação avaliada pertence à categoria de treino do modelo que tenha maior ocorrência na região em que se encontra. Para determinar estas regiões, é feito um particionamento recursivo binário da variável resposta com base nas variáveis explicativas. Neste particionamento busca-se maximizar o ganho de informação. Informação seria quanto se ganha em termos de classificações corretas da variável resposta antes e depois de uma partição. No caso de árvores de classificação, o ganho de informação é calculado pelo índice de Gini, Entropia ou taxa de Erro de Classificação.

O índice de Gini, Entropia e Erro de Classificação (E) são dados por:

$$Gini = 1 - \sum_{j} p_{mj}^2$$

$$Entropia = \sum_{j} -p_{mj} * log_2(p_{mj})$$

$$E = 1 - max_2(p_{mj})$$

Onde p_{mj} é a proporção de observações do treino na m-ésima região que são da j-ésima classe.

As partições são realizadas até que se alcance um critério de parada pré-definido, como uma quantidade máxima de partições, nós ou observações que restam após uma partição.

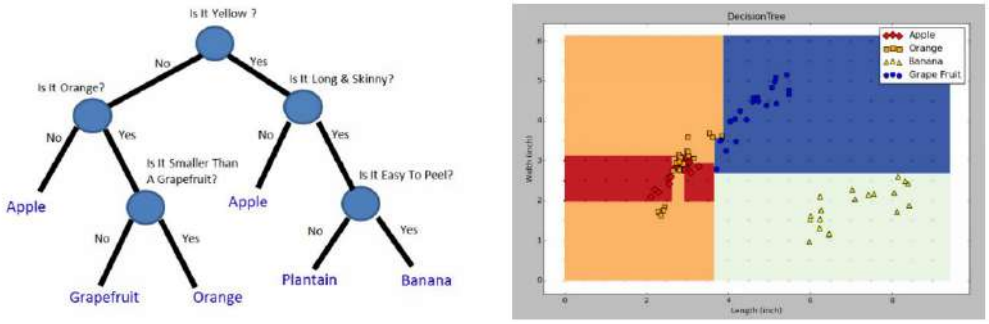
Um exemplo de árvore de decisão apresentado em “Machine Learning With Random Forests And Decision Trees” [11] trata da classificação de frutas. A figura

13(a) ilustra uma árvore de decisão que utiliza as perguntas (ou variáveis) “É amarelo?”, “É laranja?”, “É longa?”, “É menor do que uma toranja?” e “É fácil de descascar?” para classificar observações futuras como maças, toranjas, laranjas, bananas ou bananas da terra. Da forma como foi construída, respostas positivas para “É amarela?”, “É longa?” e “É fácil de descascar?” levariam a árvore a classificar uma fruta como uma banana.

Já a figura 13(b), ilustra uma árvore que utiliza largura e comprimento das frutas de treino para classificar novas observações avaliadas como laranjas, toranjas, maças, ou bananas. Na figura os círculos, triângulos, quadrados e losangos representam as frutas utilizadas para treino, e as diferentes cores de fundo representam a classificação que a árvore daria a uma nova fruta avaliada.

Segue abaixo uma ilustração da árvore tratada no exemplo.

Figura 13: Ilustrações de Árvores de Decisão



(a) Exemplo I (b) Exemplo II

Fonte: “Machine Learning With Random Forests And Decision Trees” [11]

3.4 Florestas Aleatórias

Árvores de decisão tem alta variância. Isto significa que se separarmos os dados aleatoriamente em duas partes, e ajustarmos um árvore para cada uma das partes, os resultados obtidos podem ser consideravelmente diferentes. Um método natural de reduzir a variância de um conjunto de dados é utilizar diversas bases de treino, construir um modelo de previsão para cada amostra e utilizar a classe de maior ocorrência como resposta. Outra opção é selecionar amostras aleatórias com reposição de uma mesma base de treino para obter os diferentes resultados, técnica conhecida como bootsrapping (Gareth, 2013) [12].

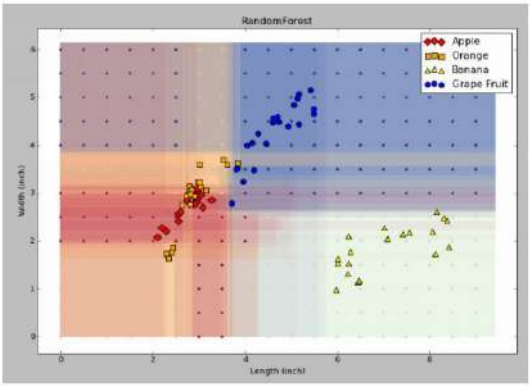
Florestas Aleatórias (Random Forests) utilizam amostras aleatórias com reposição de uma base de treino para gerar uma grande quantidade de árvores de decisão processadas em paralelo para construir o modelo de previsão da variável resposta.

Ao gerar essas árvores, em cada uma das partições é selecionada uma amostra aleatória de tamanho m dentre as p variáveis explicativas para serem utilizadas como candidatas àquela partição. Em outras palavras, florestas aleatórias forçam as parti-

ções a considerarem somente um subconjunto das variáveis explicativas disponíveis. Tipicamente se utiliza $m \approx \sqrt{p}$.

Um exemplo de floresta aleatória apresentado em "Machine Learning With Random Forests And Decision Trees" [11] em analogia ao à árvore aleatória da figura 13(b), ilustra a utilização de 16 árvores de decisão para classificar maças, laranjas, bananas e toranjas. A principal diferença entre as figuras, é que no caso da ilustração da floresta aleatória, as cores do fundo utilizadas para representar a classificação que o modelo daria a uma nova observação tem mais tons do que no caso da árvore. Isso representa os diferentes resultados de classificação das árvores utilizadas. A ideia seria utilizar a cor de fundo "mais predominante" como resultado da classificação para cada uma das regiões.

Figura 14: Ilustração de uma Árvore de Decisão



Fonte: "Machine Learning With Random Forests And Decision Trees" [11]

3.5 Medidas de Acurácia

Uma das maneiras de se avaliar o desempenho de modelos é dada pela análise do que chamamos de matriz de confusão.

Uma matriz de confusão é o cruzamento da classificação real dos dados avaliados com o a classificação predita pelo modelo, e pode ser representada da seguinte forma (Nguyen et al, 2009)[13].

Representação de uma Matriz de Confusão

		Predita	
		0	1
Real	0	**VN**	**FP**
	1	**FN**	**VP**

Onde:

- VP é a quantidade de verdadeiros positivos. Isto é, a quantidade de observações classificados corretamente como positivos pelo modelo.

- VN é a quantidade de verdadeiros negativos. Isto é, a quantidade de observações classificadas corretamente como negativas pelo modelo.
- FP é a quantidade de falsos positivos. Isto é, a quantidade de observações classificadas incorretamente como positivas pelo modelo.
- FN é a quantidade de falsos negativos. Isto é, a quantidade de observações classificadas incorretamente como negativas pelo modelo.

A partir da matriz de confusão e das informações por ela apresentadas podem ser calculadas diversas medidas para a avaliação da qualidade do modelo, como as apresentadas a seguir.

3.5.1 Acurácia (A)

A acurácia é uma medida que representa a proporção de acertos (sejam eles verdadeiros negativos ou positivos) em relação a quantidade total de registros avaliados. É definida da seguinte maneira:

$$A = \frac{VP + VN}{VN + FP + FN + VN} \tag{1}$$

3.5.2 Sensibilidade ou Recall (R)

Também conhecida como taxa de verdadeiros positivos, a sensibilidade mede quanto se acertou dentre o que de fato era positivo. É definida da seguinte maneira:

$$R = \frac{VP}{VP + FN} \tag{2}$$

3.5.3 Especificidade (E)

Também conhecido como taxa de verdadeiros negativos, a especificidade mede quanto se acertou dentre o que de fato era negativo. É definida da seguinte maneira:

$$E = \frac{VN}{VN + FP} \tag{3}$$

3.5.4 Precisão (P)

A precisão mede quanto se acertou dentre o que foi classificado como positivo pelo modelo. É definida da seguinte maneira:

$$P = \frac{VP}{FP + VP} \tag{4}$$

3.5.5 Medida F (F)

A medida F, ou F-Measure , é definida como a média harmônica de Precisão (4) e Recall (2)[14]. Isto é:

$$F = \frac{2PR}{P + R} \tag{5}$$

No entanto, em uma definição mais completa, pode ser adicionada à medida F um argumento β que controla o balanceamento entre Precisão e Recall. Neste caso, F é denominado F_β (Sasaki & Yutaka, 2007) [14]. Onde:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}, \qquad (0 \leq \beta \leq +\infty) \tag{6}$$

Sendo que:

- Se $\beta = 1$, F_1 é o equivalente à média harmônica definida em (5).
- Se $\beta > 1$, a medida F dá maior peso ao recall.
- Se $0 \leq \beta < 1$, a medida F dá maior peso à precisão.

3.5.6 Medidas para Dados Desbalanceados

A acurácia costuma ser a medida mais utilizada no propósito de avaliação de acertos e erros de um modelo de aprendizado de máquina. No entanto, para dados em que as classes da variável resposta são desbalanceadas (proporção desigual de positivos e negativos) utilizar a acurácia como medida de avaliação não é apropriado, pois a quantidade de registros da categoria influencia na sua contribuição no valor resultante. (Nguyen et al, 2009) [13]

Weiss e Provost [15] mostraram empiricamente que a utilização da acurácia como medida de avaliação dos modelos leva a uma performance ruim da classificação da categoria minoritária.

Algumas das medidas mais relevantes para medir o acerto de modelos com dados desbalanceados são precisão, recall, especificidade e medida F (apresentadas anteriormente). (Nguyen et al, 2009)[13]

3.6 Oversampling e Undersampling

A amostragem é uma das abordagens mais comuns quando tratamos de problemas com dados desbalanceados. A ideia é fazer alterações na base de treino do modelo, de maneira que se minimize a discrepância entre classes da variável resposta (Nguyen et al, 2009)[13].

De acordo com Weiss e Provost [15], dois métodos básicos de amostragem para reduzir o desbalanceamento de classes são denominadas undersampling (ou subamostragem) e oversampling (ou superamostragem). Undersampling consiste em extrair um subconjunto da classe majoritária e preservar a classe minoritária. Oversampling consiste em aumentar a quantidade referente a classe minoritária ao replicar seus registros. A amostragem associada ao aumento ou diminuição da classe minoritária e majoritária, respectivamente, pode ser feito de diversas maneiras.

4 Materiais, Métodos e Estudos Iniciais

Nesta seção serão descritos os procedimentos, métodos, materiais utilizados e estudos inicias realizados para que os objetivos propostos no trabalho fossem alcançados.

Primeiramente, será mencionada de maneira genérica a natureza dos dados utilizados no trabalho. Em seguida, será detalhado o processo de construção das redes de relacionamento utilizadas, e os resultados de dois estudos inicias. O objetivo do primeiro estudo foi determinar o nível das redes que seriam utilizadas para identificar os candidatos a serem avaliados pelo modelo. Isto é, a partir dos indivíduos citados em uma comunicação que daria origem a um relatório de inteligência, quantos graus de distância devem ser considerados para buscar os possíveis alvos do relatório que descreverá o evento ilícito suspeito?

O segundo estudo visava avaliar se as medidas de centralidade advindas da teoria de grafos seriam boas variáveis explicativas para a resposta de se um determinado indivíduo é ou não principal relacionado do RIF. Apesar dos estudos tratarem de resultados obtidos, suas conclusões foram determinantes para o direcionamento metodológico do trabalho. Por isso, optou-se por apresentá-los neste momento.

Por fim, será exposto como foram encontradas e o planejamento de extração das demais informações a serem consideradas como variáveis explicativas iniciais do modelo de florestas aleatórias para a classificação de indivíduos como alvos ou não alvos de um relatório de inteligência financeira.

4.1 Base de Dados

Os dados utilizados em todo o desenvolvimento do trabalho são de bases disponíveis ao COAF. No geral, tratam do histórico de informações de movimentações financeiras recebidas e dados cadastrais da Receita Federal, que contém informações como CPF, CNPJ, nome, razão social, participações societárias e dependentes, por exemplo (O que é o COAF, 2019) [2].

4.2 Construção das Redes para Identificação dos Possíveis Envolvidos

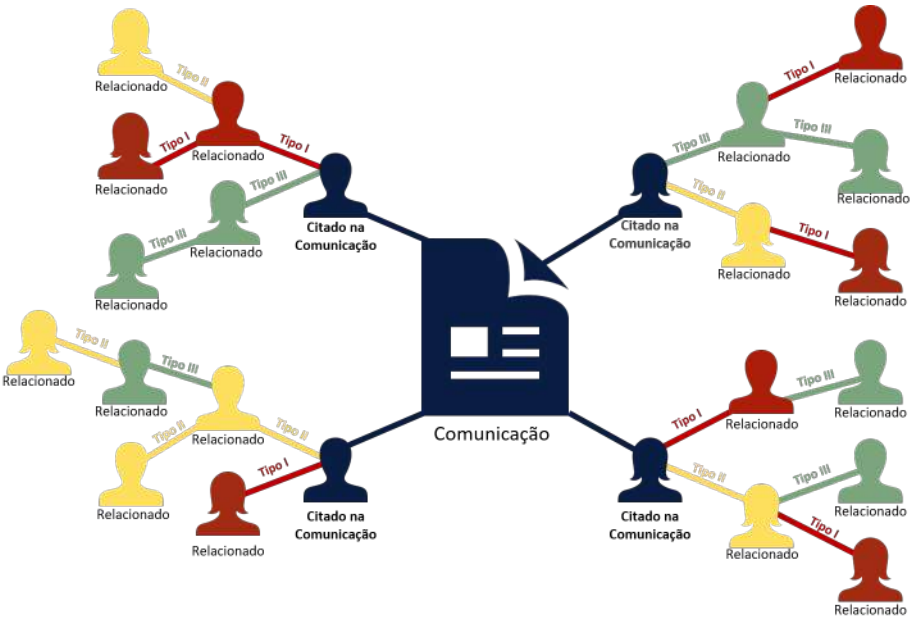
O desafio inicial do trabalho foi encontrar uma maneira automática de selecionar possíveis envolvidos do relatório de inteligência, que em um segundo momento seriam classificados como sendo ou não os principais relacionados. Partiu daqui, a inspiração de se utilizar os conceitos de redes sociais.

Como mencionado na seção teórica, uma rede é uma estrutura de laços entre atores de um determinado sistema social, e pode ser representada matematicamente por um objeto conhecido como grafo. Um grafo é composto por um conjunto de pontos denominados vértices, que representam os atores, e se unem ou não por linhas denominadas arestas, que representam os laços), e manifestam a associação

entre eles. (Quintilha, 2010) [5].

Os Relatórios de Inteligência Financeira (RIF) tem origem em comunicações de operações financeiras recebidas (COAF, Relatório de Atividades 2018) [16]. Por isso, a ideia foi partir dos indivíduos mencionados em determinada comunicação (em que tenha sido identificado possível cometimento de ilícitos) para associar por meio de relacionamentos identificáveis pelas bases disponíveis, uma rede de possíveis envolvidos no RIF a ser produzido. A imagem a seguir ilustra uma rede montada a partir dos indivíduos de uma comunicação com base nos relacionamentos "Tipo I", "Tipo II" e "Tipo III".

Figura 15: Ilustração montagem da Rede



Fonte: Autora

Assim, em termos de grafo, no problema em questão foram considerados vértices pessoas físicas ou jurídicas, e arestas qualquer tipo de relacionamento identificado que pudessem vincular os vértices entre si. Além disso, optou-se por utilizar grafos valorados (considerando a quantidade de repetições no relacionamento como peso), mas não orientados. A decisão por utilizar grafos não orientados se deu pela dificuldade de identificar a direção dos relacionamentos considerados.

Por questões de desempenho computacional, decidiu-se por montar as redes de cada um dos tipos de aresta separadamente, e consultá-las no momento de construir os grafos por relatório. Por exemplo, as informações de movimentações financeiras suspeitas enviadas pelos setores obrigados, denominadas doravante de comunicações, foram um tipo de relacionamento considerado como aresta. Então a partir de uma tabela inicial contendo identificador da comunicação e pessoas relacionadas, foi montada o que se chamou de rede de comunicações. Na tabela que representa a rede de comunicações constam todas as duplas de pessoas que se relacionam por meio de comunicações, o identificador do tipo de relacionamento utilizado

para buscar este vínculo (neste caso, a comunicação) e a quantidade de relacionamentos deste tipo encontrados entre determinada dupla (neste caso, quantidade de comunicações em comum). Para esclarecer este processo, segue abaixo uma representação das tabelas mencionadas.

Tabela 2: Representação da Tabela Inicial Utilizada para Construção da Rede de Comunicações

Nº da Comunicação	**Pessoa Relacionada**
1	PFPJ1
1	PFPJ2
1	PFPJ3
2	PFPJ1
2	PFPJ3
3	PFPJ2
3	PFPJ3
...	...
n	PFPJx
n	PFPJy

Tabela 3: Representação da Tabela da Rede de Comunicações

Pessoa 1	**Pessoa 2**	**Tipo de Relacionamento**	**Quantidade**
PFPJ1	PFPJ2	Comunicação	1
PFPJ1	PFPJ3	Comunicação	2
PFPJ2	PFPJ3	Comunicação	2
...	...	...	...
PFPJx	PFPJy	Comunicação	1

No exemplo das representações acima, consta na tabela inicial que as pessoas 1, 2 e 3 aparecem na comunicação 1. Enquanto as pessoas 1 e 3 aparecem na comunicação 2 e as pessoas 2 e 3 aparecem na comunicação 3. Por isso, na tabela da rede de comunicações do exemplo a observação que representa o vínculo entre pessoa 1 e 2 aparece com quantidade 1 (fruto das comunicação 1). A observação que representa o vínculo entre pessoa 1 e 3 aparece com quantidade 2 (fruto das comunicações 1 e 2). E a observação que representa o vínculo entre pessoa 2 e 3 aparece com quantidade 2 (fruto das comunicações 1 e 3).

Vale ressaltar que em termos de grafo as colunas “Pessoa 1” e “Pessoa 2” representam os vértices, “Tipo de Relacionamento” representa a aresta, e a “Quantidade” será utilizada como peso. O peso pode ser utilizado como ponderador para algumas medidas de centralidade, como apresentado na seção teórica.

Outros 6 tipos de relacionamento identificáveis pelas bases disponíveis foram considerados, no entanto não serão revelados para preservar o mecanismo de identificação de possíveis envolvidos. Para cada um dos 6 relacionamentos foi realizado o processo exemplificado na rede de comunicações, dando origem a 7 bases como as representadas a seguir.

Tabela 4: Representações das Tabelas das Redes de Relacionamento

Comunicação

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Comunicação	1
PFPJ1	PFPJ3	Comunicação	2
PFPJ2	PFPJ3	Comunicação	2
...	...	...	...
PFPJx	PFPJy	Comunicação	1

Relacionamento 2

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 2	1
PFPJ1	PFPJ3	Tipo 2	2
PFPJ2	PFPJ3	Tipo 2	2
...	...	...	...
PFPJx	PFPJy	Tipo 2	1

Relacionamento 3

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 3	1
PFPJ1	PFPJ3	Tipo 3	2
PFPJ2	PFPJ3	Tipo 3	2
...	...	...	...
PFPJx	PFPJy	Tipo 3	1

Relacionamento 4

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 4	1
PFPJ1	PFPJ3	Tipo 4	2
PFPJ2	PFPJ3	Tipo 4	2
...	...	...	...
PFPJx	PFPJy	Tipo 4	1

Relacionamento 5

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 5	1
PFPJ1	PFPJ3	Tipo 5	2
PFPJ2	PFPJ3	Tipo 5	2
...	...	...	...
PFPJx	PFPJy	Tipo 5	1

Relacionamento 6

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 6	1
PFPJ1	PFPJ3	Tipo 6	2
PFPJ2	PFPJ3	Tipo 6	2
...	...	...	...
PFPJx	PFPJy	Tipo 6	1

Relacionamento 7

V1	**V2**	**Aresta**	**Peso**
PFPJ1	PFPJ2	Tipo 7	1
PFPJ1	PFPJ3	Tipo 7	2
PFPJ2	PFPJ3	Tipo 7	2
...	...	...	...
PFPJx	PFPJy	Tipo 7	1

A construção da tabela inicial, contendo apenas identificador do relacionamento e pessoas envolvidas, é particular para cada rede, e de considerável complexidade dada a estrutura do armazenamento de dados da instituição. No entanto, uma vez construída a tabela inicial, o processo de estruturação de cada uma das redes é semelhante, e a codificação no SAS pôde ser generalizada, no entanto para fins de proteção das estruturas do COAF, não serão apresentadas no trabalho.

Por consequência da dimensão das bases que foram utilizadas, o esforço computacional associado a este processo foi considerável. O tempo de execução despendido na construção das redes para cada um dos relacionamentos é apresentado na tabela a seguir.

Quadro 1: Tempo de Processamento para a Construção das Redes

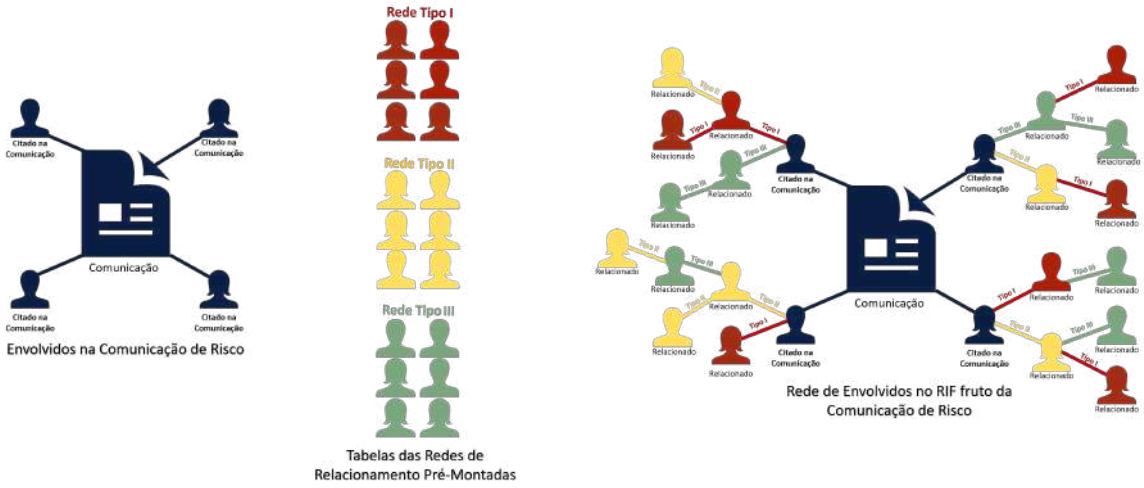
Rede	Tempo de Processamento
Comunicação	7 horas
Rede 2	6 horas e 15 minutos
Rede 3	1 hora e 50 minutos
Rede 4	27 minutos
Rede 5	16 minutos
Rede 6	22 minutos
Rede 7	1 hora e 10 minutos

Uma vez montadas as redes de cada um dos tipos de aresta, resta consultá-las para buscar os relacionamentos dos envolvidos em uma comunicação e construir a rede de envolvidos no RIF, que relatará o evento ilícito suspeito. Vale esclarecer que:

- Pode haver mais de um tipo de relacionamento entre dois indivíduos. Neste caso, foi utilizada a soma das quantidades de cada tipo como peso final para a aresta.
- Como "comunicações" foram um tipo de relacionamento utilizado, cria-se automaticamente uma aresta entre todos os citados na comunicação que daria origem a rede do RIF.

Na imagem a seguir ilustra-se o processo de montagem da rede de um relatório utilizando as redes pré-montadas e partindo dos citados em uma comunicação de risco.

Figura 16: Montagem da Rede do Relatório Utilizando as Bases de Redes de Relacionamentos



Fonte: Autora

Neste momento manifestou-se uma indagação quanto a distância a ser considerada entre os envolvidos nas comunicações de risco e os demais indivíduos da rede do relatório de inteligência financeira. Isto é, quantos graus de distância a partir dos citados nas comunicações deveriam ser considerados para buscar os candidatos a alvo do RIF? Os detalhes do porquê deste questionamento e os estudos feitos para respondê-lo serão apresentados na próxima seção.

4.3 Avaliação dos Graus de Separação das Redes

Em seu estudo que ficou conhecido como "Fenômeno do Mundo Pequeno", o psicólogo Stanley Milgram buscou responder a seguinte pergunta: Qual a probabilidade de que duas pessoas selecionadas aleatoriamente de uma grande população, como a dos Estados Unidos, se conheçam? No entanto, em sua formulação do problema, Milgram propôs que se levasse em conta o fato de que embora duas pessoas possam não se conhecer diretamente, pode existir uma pessoa (ou um conjunto de pessoas) que conheça tanto um indivíduo quanto o outro (Milgram, 1967) [17]. Isto é, dois indivíduos a e z, podem estar conectados por uma série de intermediários, $a - b - c - ... - y - z$.

Figura 17: Indivíduos Diretamente Conectados

Fonte: Autora

Figura 18: Indivíduos Conectados por Intermediários

Fonte: Autora

Com esta abordagem, a questão levantada por Milgram passou a ser: qual a probabilidade de que o número mínimo de intermediários entre dois indivíduos selecionados ao acaso de uma população seja 0,1,2,...,k? Ou alternativamente, qual a média de intermediários entre dois indivíduos?

Para responder a esta pergunta, ele utilizou o que ficou conhecido como "Método do Mundo Pequeno"[17] (Milgram, 1967). Milgram selecionou aleatoriamente 296 "indivíduos iniciais"das cidades de Boston e Omaha, nos Estados Unidos, e um "indivíduo alvo"nascido em Sharon (também nos Estados Unidos) que trabalhava em Boston. Aos indivíduos inicias foi enviado um documento que descrevia o estudo e pedia para que se tornassem participantes ao repassá-lo por correio, de maneira que se alcançasse o alvo pré-selecionado. A regra era que a carta só poderia ser

repassada para uma pessoa verdadeiramente conhecida. Ou seja, se as pessoas iniciais conhecessem o alvo, a carta poderia ser enviada diretamente. Caso contrário (situação que seria mais provável) o documento deveria ser repassado para alguém conhecido que fosse considerado por eles mais próximo do alvo informado. A corrente seria considerada finalizada quando a carta alcançasse o alvo, ou quando uma pessoa no caminho se negasse a participar.

Uma das conclusões do experimento foi que 64 cartas alcançaram o alvo. E dentre estas, em média 6 pessoas compuseram caminho percorrido para que as cartas chegassem ao seu destino. Isto é, existiu uma média de 6 intermediários entre os "indivíduos iniciais"selecionados aleatoriamente e o "indivíduo alvo". Este resultado teve tamanha relevância, e gerou tamanha surpresa, que deu origem à teoria dos "Seis Graus de Separação", tópico de estudos e experimentos de diversos cientistas, como Duncan J. Watts em "Six Degrees: The Science of a Connected Age"[18]. Em resumo os estudos estendem a população avaliada por Milgram, ao avaliar a teoria de que quaisquer duas pessoas no planeta podem estar conectadas por uma média de 6 intermediários, níveis ou graus de separação.

Com os estudos envolvendo os "Seis Graus de Separação" em mente, surgiram dois questionamentos em relação à estratégia de construção das redes:

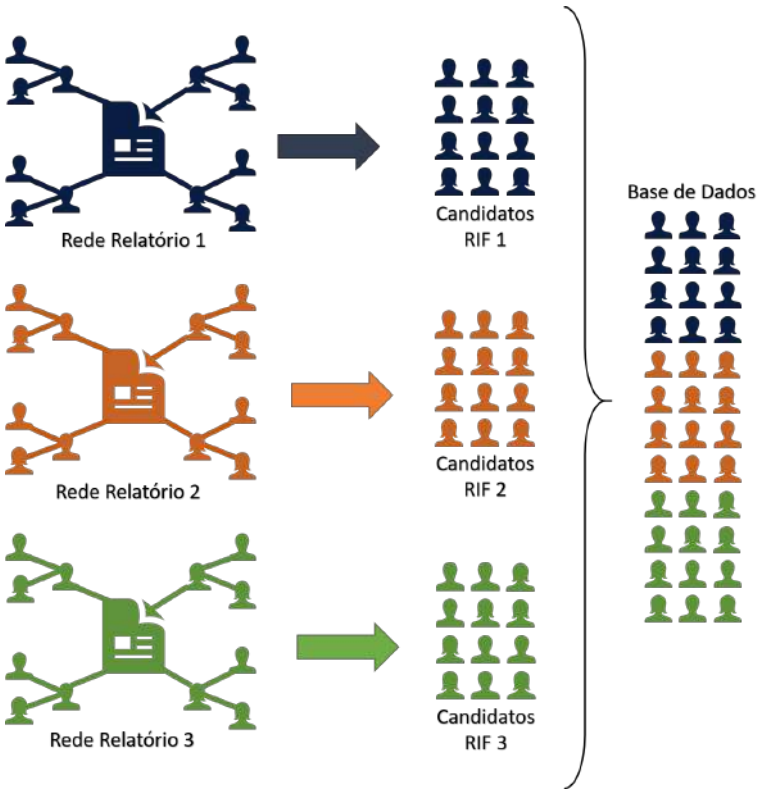
1. Em termos de redes de relacionamento, quantos níveis, ou quantos graus a partir dos indivíduos mencionados em determinada comunicação, devem ser contemplados para que todos os principais relacionados estejam na rede do RIF a ser analisado? Isto é, intermediários separam os indivíduos da comunicação que deu origem ao relatório e os demais alvos do possível evento ilícito ali presente?

2. No universo de pessoas acessíveis pelas bases disponíveis ao COAF, quantos graus de separação são necessários para que todos os indivíduos se conectem? A preocupação aqui foi avaliar se os graus de separação utilizados para buscar as pessoas a serem avaliadas fariam com que toda a população estivesse em todos os relatórios, o que faria com que o método fosse inutilizável.

Para responder a essas perguntas, foram construídas as redes (utilizando os relacionamentos mencionados na seção anterior) partindo dos indivíduos de 13.276 comunicações que deram origem a relatórios de inteligência, e considerando apenas 1 nível de distância dos "indivíduos inicias" (em analogia ao estudo de Milgram).

Ou seja, foram buscadas pessoas diretamente relacionadas (por alguma das bases de redes de relacionamento pré montadas e descritas anteriormente) aos indivíduos que constavam na comunicação onde se identificou um indício de cometimento de ilícito. Esta busca resultou em uma listagem de pessoas, em forma de base de dados, contendo os candidatos a alvo de cada RIF. Segue ilustração deste processo.

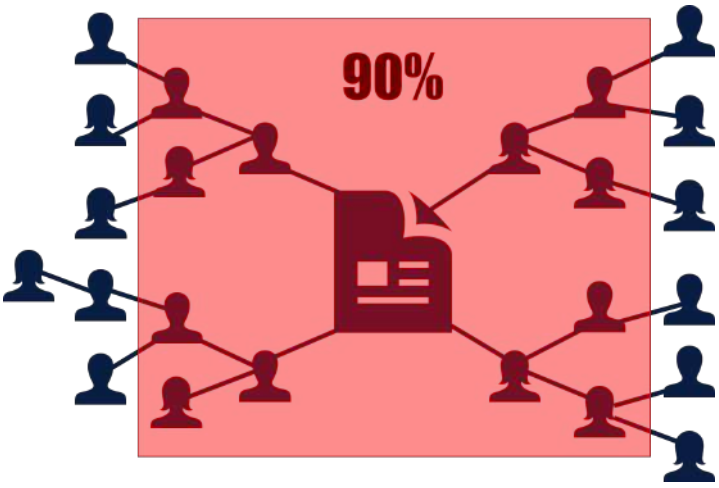
Figura 19: Busca por Candidatos Utilizando 1 Grau de Distância



Fonte: Autora

Ao comparar indivíduos identificados neste processo aos reais envolvidos nos relatórios de inteligência fruto das comunicações selecionadas, foi analisado que em média 90% dos alvos constavam naquela lista de pessoas apanhadas automaticamente.

Figura 20: Ilustração do Resultado do Estudo: 90% dos Alvos no 1º Grau



Fonte: Autora

Como o custo computacional e tempo de trabalho associado a busca e avaliação do resultado que se daria em maiores níveis era grande, decidiu-se por utilizar, nesta primeira versão do modelo, um grau de distância dos citados nas comunicações (em que se identifique fundados indícios de cometimentos de ilícitos) para selecionar os indivíduos a serem avaliados pelo modelo de classificação como alvos ou não alvos do relatório.

4.4 Avaliação das Medidas de Centralidade como Variáveis Explicativas

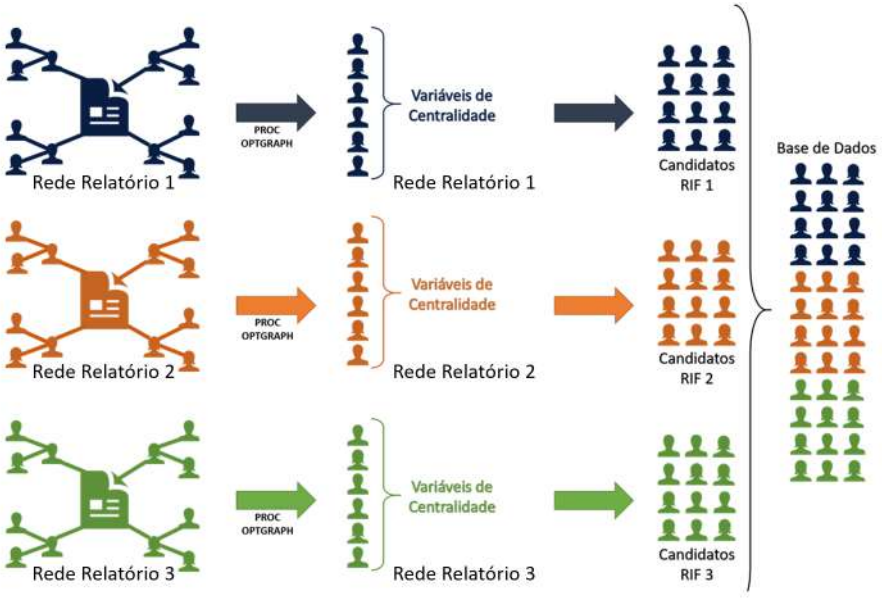
Existem na literatura diversos conceitos e métricas que visam classificar os vértices com base em seus papeis estruturais no grafo. Como o método de seleção de pessoas a serem avaliadas pelo modelo envolve conceitos de redes, surgiu ideia de se utilizar medidas de centralidade como variáveis explicativas.

Para investigar esta possibilidade, foi feito um estudo preliminar com uma análise exploratória de algumas métricas associadas aos atores da redes. Neste estudo foram utilizadas os mesmos 13.276 relatórios e indivíduos identificados na avaliação dos níveis das redes, detalhada na seção anterior.

Para calcular essas medidas, foi utilizada a PROC OPTGRAPH, disponível no SAS, que com auxílio de uma macro codificada fornece informações relacionadas a centralidade e grupos de cliques dos vértices de cada uma das redes de relatórios.

O desafio computacional e a necessidade da macro neste processo se deu porque as medidas deveriam ser relacionadas a centralidade de indivíduos por caso, e para tanto seria necessário que se rodasse a PROC para as redes de cada um dos relatórios separadamente. Ou seja, no caso deste estudo, seria necessário que se rodasse a PROC 13.276 vezes. A imagem a seguir, ilustra o fluxo do processo de utilização da PROC OPTGRAPH que foi automatizado pela macro. A diferença da base resultante no fluxo apresentado abaixo para a base resultante no fluxo ilustrado na figura 19, é que agora são atribuídos aos indivíduos as suas medidas de centralidade.

Figura 21: Busca pelas Medidas de Centralidade dos Indivíduos em suas Redes



Fonte: Autora

Com as saídas da PROC e algumas manipulações adicionais, obtiveram-se estatísticas gerais das redes e medidas dos vértices avaliados. A média de algumas das estatísticas por rede de relatório é apresentada a seguir.

Quadro 2: Estatísticas Estudo Inicial das Medidas de Centralidade por Relatório

Medida	Média por Relatório
Vértices	189
Arestas	383
Arestas por Vértices	1,73
Nós Terminais	147
Alvos	5
Proporção de Alvos no > Clique	57%

Como visto no quadro acima, a média de vértices e arestas por rede de relatório é de 189 indivíduos (sendo 147 nós terminais) e 383 relacionamentos, resultando 1,73 arestas por vértice.

Outro resultado interessante retirado do quadro apresentado, é que dos 189 vértices 5 (em média) são alvos. No entanto, mesmo com uma proporção minoritária em relação aos demais envolvidos, 57% dos vértices que compõem o maior clique da rede são representados por alvos.

Outra avaliação feita, foi a comparação entre as médias e medianas por rede de relatório referentes a “Ponto de articulação”(que indica se aquele vértice é um ponto de articulação da rede), “Centralidade de Grau”, “Centralidade de Intermediação”e “Centralidade de Auto Vetor”para vértices alvos e não alvos.

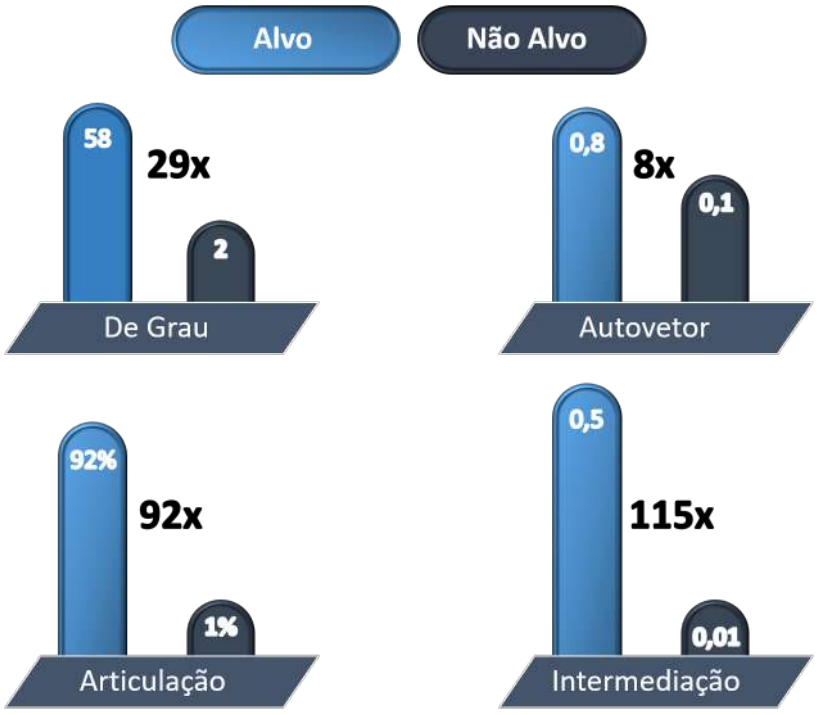
Vale comentar que no caso de “Ponto de Articulação”, que é uma variável binária indicando se aquele vértice é um ponto de articulação, foi calculado o percentual de casos positivos dentre alvos e não alvos.

Os resultados obtidos são resumidos pelo quadro e imagem a seguir.

Quadro 3: Medidas de Centralidade por Relatório para Alvos e Não Alvos

Medida	Média por Relatório		Mediana por Relatório	
	Alvos	Não Alvos	Alvos	Não Alvos
Centralidade de Grau	58	2	32	1,29
Centralidade de Autovetor	0,75	0,13	0,84	0,09
Centralidade de Intermediação	0,53	0,08	0,44	0,001
Ponto de Articulação	92%	1%	100%	0%

Figura 22: Comparação da Média das Medidas por Relatório para Alvos e Não Alvos



Fonte: Autora

Avaliando os resultados apresentados nessa seção, pode-se dizer que existiram indícios de que as variáveis de centralidade advindas da análise de grafos apresentavam diferenças significantes em seus resultados quando estratificados pela identificação de Principais Relacionados, podendo ser variáveis explicativas úteis para o modelo de classificação de alvos.

Assim, decidiu-se por utilizar 7 medidas associadas aos relacionamentos e centralidade de vértices dos grafos dos relatórios como variáveis explicativas iniciais do modelo:

1. **Presente no Maior Clique:** variável binária que indica se aquele vértice se encontra no maior dos cliques da rede do relatório.
2. **Nó terminal:** variável binária que indica se aquele vértice é um nó terminal da rede do relatório ao qual pertence.
3. **Ponto de Articulação:** variável binária que indica se aquele vértice é um ponto de articulação
4. **Centralidade de Autovetor Ponderada:** variável numérica que representa o valor atribuído a centralidade de autovetor ponderada do vértice.
5. **Centralidade de grau:** variável numérica que representa o valor atribuído a centralidade de grau ponderada do vértice.
6. **Centralidade de intermediação:** variável numérica que representa o valor atribuído a centralidade de intermediação ponderada do vértice.
7. **Centralidade de Proximidade:** variável numérica que representa o valor atribuído a centralidade de proximidade ponderada do vértice da rede do relatório ao qual pertence.

Para extrair a variável "Presente no Maior Clique" foi necessário desenvolver alguns códigos adicionais para manipulação da saída de "Cliques" da PROC OPTGRAPH. As variáveis "Nó terminal", "Ponto de articulação", "Centralidade de Grau", "Centralidade de Intermediação" e "Centralidade de Proximidade" são saídas diretas da PROC, atribuídas a cada um dos vértices avaliados.

4.5 Busca por Novas Variáveis

Afim de buscar novas variáveis, além das sete identificadas no tema de centralidade e prestígio, foram feitas entrevistas com os analistas para melhor entendimento das configurações da construção dos relatórios e identificação dos alvos.

A análise das entrevistas e dos estudos iniciais levou à identificação de 7 temas de análise e 38 variáveis iniciais. As variáveis foram inicialmente separadas em dois momentos:

- Momento 1: Variáveis a serem levantadas de forma imediata para serem analisadas e possivelmente utilizadas como variáveis explicativas no modelo de identificação de principais relacionados. Esse momento contou com 23 variáveis.
- Momento 2: Variáveis a serem levantadas e analisadas em momento posterior, possivelmente em caráter de revisão do modelo construído. Esse momento contou com 15 variáveis que, em sua maioria, contém informações de bases não disponíveis ao COAF, variáveis cujo levantamento foi considerado de extrema dificuldade ou com prazo longo de execução.

Dos 7 temas e 38 variáveis, 1 e 7 (respectivamente) se referem às variáveis de centralidade, descritas nas seções anteriores. Os demais 6 temas e 31 variáveis não serão revelados por questões relacionadas a segurança institucional do órgão e preservação dos critérios de identificação dos alvos do relatório. No entanto, os resultados serão apresentados de maneira descaracterizada, como na tabela abaixo, que resume as variáveis, seus temas e momentos de extração reservados para cada uma delas.

Tabela 5: Variáveis Explicativas

Nº	**Tema**	**Variável**	**Momento Extração**
1	Centralidade	Alvo do Maior Clique	1
2	Centralidade	Centralidade de auto vetor ponderada	1
3	Centralidade	Centralidade de grau	1
4	Centralidade	Centralidade de intermediação	1
5	Centralidade	Centralidade de proximidade	1
6	Centralidade	Nó terminal	1
7	Centralidade	Ponto de Articulação	1
8	Tema 1	Variável 1	1
9	Tema 1	Variável 2	1
10	Tema 1	Variável 3	1
11	Tema 2	Variável 4	1
12	Tema 2	Variável 5	1
13	Tema 3	Variável 6	1
14	Tema 3	Variável 7	1
15	Tema 3	Variável 8	1
16	Tema 3	Variável 9	1
17	Tema 7	Variável 10	1
18	Tema 4	Variável 11	1
19	Tema 3	Variável 12	1
20	Tema 3	Variável 13	1
21	Tema 3	Variável 14	1
22	Tema 3	Variável 15	1
23	Tema 5	Variável 16	1
24	Tema 3	Variável 17	2
25	Tema 3	Variável 18	2
26	Tema 5	Variável 19	2
27	Tema 1	Variável 20	2
28	Tema 1	Variável 21	2
29	Tema 1	Variável 22	2
30	Tema 1	Variável 23	2
31	Tema 1	Variável 24	2
32	Tema 1	Variável 25	2
33	Tema 1	Variável 26	2
34	Tema 6	Variável 27	2
35	Tema 6	Variável 28	2
36	Tema 3	Variável 29	2
37	Tema 3	Variável 30	2
38	Tema 3	Variável 31	2

5 Resultados

Serão apresentados nesta seção os principais resultados do trabalho. O primeiro resultado apresentado trata da base de dados obtida seguindo o que foi exposto na seção metodológica e os passos seguidos para a construção do que se chamou de "Base de Treino Final" e "Base de Validação Final". Dentre estes passos, estiveram:

- A partição da extração original em treino e teste.
- Um ajuste da base de treino inicial dada por uma percepção de que a base que estava sendo utilizada poderia conter observações cujas respostas não foram de fato validadas por um analista. Isto porque o método de seleção automática dos candidatos a alvo do relatório acaba buscando mais indivíduos "não alvos" do que aqueles que constam nos relatórios. Além disso, retirar tais indivíduos tornaria a base de treino mais balanceada, o que levou à decisão de não considerar tais observações no treino do modelo.
- A retirada de 6 variáveis explicativas cuja extração dependia de um processo de análise que pode vir a não existir no futuro. A decisão de retirar estas variáveis foi gerencial.

Em seguida, serão comentados os critérios para a escolha de alguns dos parâmetros definidos nos modelos de florestas aleatórias. Serão expostos também os 5 cenários de avaliação do modelo, levando em consideração diferentes bases de treino construídas utilizando undersampling e oversampling. O objetivo desta abordagem foi avaliar e prevenir que o desbalanceamento da classe resposta pudesse causar um mal desempenho do modelo.

Por fim, serão comparadas as matrizes de confusão e medidas de acurácia dos modelos ajustados em cada um dos cenários apresentados e o critério para a escolha do modelo final.

5.1 Base de Treino e Validação

Utilizando as redes de relacionamento pré-montadas e métodos descritos na seção anterior, foram selecionados os candidatos a alvos de 2.938 relatórios de inteligência produzidos pelo COAF, que trata de RIFs mais recentes do que os 13.276 utilizados nos estudos. Para cada um destes indivíduos em seus respectivos relatórios, foram extraídas as variáveis explicativas do "Momento 1" (apresentadas na tabela 5), e a marcação real de se aquela pessoa foi ou não uma principal relacionada do relatório, que é a variável resposta do modelo. O resultado desta extração foi uma base de dados com 23 variáveis:

- Identificador do Relatório
- Identificador do Envolvido
- Variável "Presente no Maior Clique"

- Variável “Centralidade de autovetor Ponderada”
- Variável “Centralidade de grau”
- Variável “Centralidade de Intermediação”
- Variável “Centralidade de Proximidade”
- Variável “Nó Terminal”
- Variável “Ponto de Articulação”
- Variáveis 1-16

Esta base, a qual nos referiremos como “Base de Dados Original” contou com 1.698.747 observações, das quais 8.080 e 1.690.667 representavam alvos e não alvos dos relatórios, respectivamente. Esta distribuição da classe da variável resposta é representada pela tabela a seguir.

Tabela 6: Distribuição da Classe da Variável Resposta na Base de Dados Original

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	8.080	0,4%
Não Alvos	1.690.667	99,6%%
Total	1.698.747	100%

Para que os indivíduos selecionados em um mesmo relatório não fossem divididos, foram selecionados 2.460 relatórios (com 1.280.338 indivíduos), e 478 relatórios (com 418.409 indivíduos), para compor as bases de treino e validação do modelo, respectivamente. As distribuições da classe da variável resposta em cada uma das bases, que tem um claro e grande desbalanceamento associado, são representados pelas tabelas a seguir.

Tabela 7: Distribuição da Classe da Variável Resposta

Base de Treino

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	6.815	0,5%
Não Alvos	1.273.523	99,5%
Total	1.280.338	100%

Base de Validação

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	1.265	0,3%
Não Alvos	417.144	99,7%
Total	418.409	100%

Já havia sido notado que o método proposto para a seleção automática dos candidatos a alvo trazia mais pessoas do que aquelas que constavam no relatório. Isto porque existem indivíduos que estão nos relatórios dos analistas, e que são divididos em duas classes: alvos e não alvos. Os indivíduos que não constaram do relatório podem ou não ter sido incluídos no grafo construído usando a metodologia descrita.

Tal aspecto não foi visto como um problema, pois o objetivo do modelo seria direcionar a determinação dos alvos do evento ilícito suspeito, e não todos os envolvidos. Portanto, aqueles que não constavam no relatório foram vistos simplesmente como "não alvos".

No entanto, surgiu a ideia de que utilizar os indivíduos que não constavam no relatório no treino do modelo poderia trazer uma imprecisão ao ajuste, pois estaríamos "assumindo" uma classificação que na verdade não existiu. Além disso, ao não considerar tais indivíduos, a base se tornaria mais equilibrada. Assim, na intenção de tornar a base de treino do modelo mais balanceada, e levando em conta a percepção de que o método de seleção automática dos candidatos a alvos dos relatórios traz mais "não alvos" do que aqueles que são de fato "marcados" pelo analista, optou-se por retirar do treino os indivíduos que não constavam nos relatórios. Por outro lado, do ponto de vista operacional, para uma nova observação avaliada não se saberia quais indivíduos seriam ou não incluídos no hipotético relatório, levando à decisão de não se alterar a base de validação (que avaliará o desempenho do modelo treinado).

Desta forma, a "Base de Treino Final" e a "Base de Validação Final" passaram a ter seguinte configuração:

Tabela 8: Distribuição da Classe da Variável Resposta nas Bases Finais

Base de Treino Final

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	6.815	6,8%
Não Alvos	94320	93,2%
Total	101.135	100%

Base de Validação Final

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	1.265	0,3%
Não Alvos	417.144	99,7%
Total	418.409	100%

Percebe-se que desta forma, a proporção de alvos em relação ao total aumenta de 0,5% para 6,8%.

Além disso, foi levantada uma preocupação do ponto de vista gerencial que manifestou a dependência das variáveis 11-16 em um processo interno do órgão que pode vir a não existir. Por isso, optou-se por não utilizá-las como variáveis explicativas do modelo. Desta forma, a "Base de Treino Final" e a "Base de Validação Final" acabaram permanecendo com 17 variáveis:

- Identificador do Relatório
- Identificador do Envolvido
- Variável "Presente no Maior Clique"
- Variável "Centralidade de autovetor Ponderada"
- Variável "Centralidade de grau"
- Variável "Centralidade de Intermediação"

- Variável “Centralidade de Proximidade”
- Variável “Nó Terminal”
- Variável “Ponto de Articulação”
- Variáveis 1-10

5.2 Escolha dos Parâmetros

Existem alguns parâmetros a serem determinados na modelagem de florestas aleatórias. Serão apresentados a seguir os parâmetros utilizados para o treino dos modelos ajustados neste trabalho. Especificamente, serão discutidos os valores de:

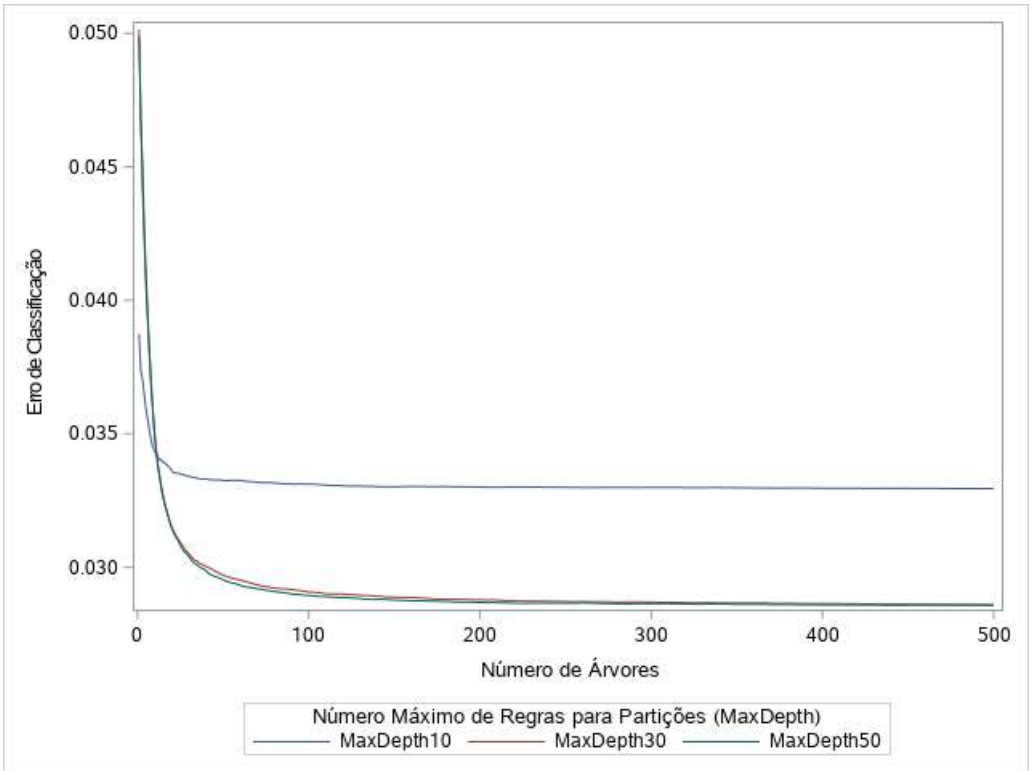
- Número Máximo de Árvores: especifica a quantidade máxima de árvores a serem utilizadas nas florestas aleatórias. No SAS é representado pelo argumento MAXTREES.
- Variáveis Avaliadas: especifica a quantidade de variáveis utilizadas em cada partição. No SAS é representado pelo argumento VARS_TO_TRY.
- Proporção “Inbag”: especifica a proporção de observações utilizadas para o bootstrapping de cada árvore. No SAS é representado pelo argumento TRAINFRACTION.
- Número Máximo de Regras: especifica o número máximo de regras de cada partição de cada árvore. No SAS é representado pelo argumento MAXDEPTH.

Como visto na revisão da literatura, o número de variáveis utilizadas em cada partição da árvore é tipicamente a raiz quadrada do total de variáveis explicativas disponíveis para o modelo. No caso do problema em questão, como temos 17 variáveis explicativas disponíveis, foram utilizadas $\sqrt{17} \approx 4$ variáveis em cada partição.

Outro valor proposto na literatura é a proporção dos dados que será avaliado em cada árvore, usualmente definida como 60%.

Para determinar o número máximo de regras, foi observado o erro de classificação médio em cada árvore em ajustes de florestas aleatórias na base de treino (com demais parâmetros padrão do SAS) considerando 3 valores diferentes: 10, 30 e 50. Como pode ser visto no gráfico a seguir, o erro diminui consideravelmente de 10 para 30 e 50. No entanto não há diferença entre os valores de 30 e 50, e como um menor número de regras exige menor esforço computacional, pelo princípio da parcimônia optou-se por utilizar o valor de 30 como parâmetro.

Figura 23: Gráfico da Redução do Erro de Classificação por Quantidade de Árvores e Valor de Maxdepth



Fonte: Autora

Apesar de não existirem indícios de que o aumento do número de árvores utilizadas cause problemas de sobreajuste, na prática se utiliza o valor suficientemente grande para que a taxa de erro se estabilize. Com base no gráfico da figura (23), percebe-se que para qualquer dos valores avaliados o erro se fixa em torno de 250 árvores.

Um resumo dos parâmetros utilizados é apresentado a seguir.

Quadro 4: Parâmetros do Modelo

Parâmetro	Valor
Variáveis Avaliadas	4
Proporção "Inbag"	0,6
Número Máximo de Regras	30
Número Máximo de Árvores	250

5.3 OverSampling e UnderSampling da Base de Treino

Alguns estudos sugerem que o procedimento de florestas aleatórias pode não classificar precisamente categorias menos expressivas em dados com grande desbalanceamento. Conforme visto na revisão bibliográfica, uma das maneiras de lidar

com dados desbalanceados é por meio da amostragem. A ideia é pré-processar a base de treino para minimizar as discrepâncias entre as classes. Como os dados em questão apresentaram uma desproporção significativa da classe da variável resposta, foram considerados 4 novas bases de treino, afim de avaliar cada caso e prevenir que o desbalanceamento pudesse causar um mal desempenho do modelo.

Para isto, foram utilizados os métodos de undersampling e oversampling expostos anteriormente. Para o caso do undersampling, foi selecionada uma amostra aleatória simples da classe majoritária. Para o caso do oversampling, feita com uma replicação aleatória dos registros da classe minotritária. Foram criadas 2 bases via undersampling, e 2 bases via oversampling, definidas da seguinte maneira:

1. Base Undersampling 50%: em que se partiu da "Base de Treino Final" para realizar uma subamostragem da classe majoritária, de forma que a classe de "alvos" representasse 50% do total.
2. Base Undersampling 25%: em que se partiu da "Base de Treino Final" para realizar uma subamostragem da classe majoritária, de forma que a classe de "alvos" representasse 25% do total.
3. Base Oversampling 50%: em que se partiu da "Base de Treino Final" para realizar uma sobreamostragem da classe minoritária, de forma que a classe de "alvos" representasse 50% do total.
4. Base Oversampling 25%: em que se partiu da "Base de Treino Final" para realizar uma sobreamostragem da classe minoritária, de forma que a classe de "alvos" representasse 25% do total.

Para esclarecimento, a distribuição de cada uma das bases é apresentada a seguir.

Tabela 9: Distribuição da Classe da Variável Resposta

Base Undersampling 50%

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	6.815	50%
Não Alvos	6.815	50%
Total	13.630	100%

Base Undersampling 25%

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	6.815	25%
Não Alvos	20.445	75%
Total	27.260	100%

Base Oversampling 50%

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	94.320	50%
Não Alvos	94.320	50%
Total	188.640	100%

Base Oversampling 25%

Classe	Quantidade Absoluta	Quantidade Relativa
Alvos	31.440	25%
Não Alvos	94.320	75%
Total	27.260	100%

5.4 Escolha do Modelo

Para a escolha do modelo, foram considerados 5 cenários que se diferenciam pela base de treino utilizada.

Quadro 5: Cenários de Ajuste

Cenário	Treino	Validação
I	Base de Treino Final	Base de Validação Final
II	Base Undesampling 50%	Base de Validação Final
III	Base Undesampling 25%	Base de Validação Final
IV	Base Oversampling 50%	Base de Validação Final
V	Base Oversampling 25%	Base de Validação Final

Utilizando os parâmetros definidos anteriormente, foram avaliadas as matrizes de confusão do ajuste da base de validação com os modelos treinados nos 5 cenários propostos. Como a saída do modelo de florestas aleatórias é a probabilidade de que aquela observação pertença a determinada classe, foram avaliadas as classificações pelo modelo para 4 pontos de corte diferentes:

- Corte de 10%: considerados alvos aqueles com probabilidade de ser alvo (atribuída pelo modelo) maior do que 10%.
- Corte de 15%: considerados alvos aqueles com probabilidade de ser alvo (atribuída pelo modelo) maior do que 15%.
- Corte de 25%: considerados alvos aqueles com probabilidade de ser alvo (atribuída pelo modelo) maior do que 25%.
- Corte de 50%: considerados alvos aqueles com probabilidade de ser alvo (atribuída pelo modelo) maior do que 50%.

As matrizes de confusão para cada corte e um resumo das medidas de acurácia para cada cenário são apresentados a seguir. Vale ressaltar que 1 representa a classe "alvo" e 0 representa a classe "não alvo". Além disso, foram calculadas as medidas Recall (R), Especificidade (E), Precisão (P), F_1, F_2, F_3, F_4 e Acurácia (A), que foram apresentadas na revisão bibliográfica e variam de 0% a 100%, sendo 100% uma classificação perfeita, e 0% uma classificação completamente errada.

Cenário I

Tabela 10: Matrizes de Confusão da Validação no Cenário I

Ponto de Corte: 10%

		Predita	
		0	1
Real	0	413.718	3.426
	1	66	1.199

Ponto de Corte: 15%

		Predita	
		0	1
Real	0	415.731	1.413
	1	99	1.166

Ponto de Corte: 25%

		Predita	
		0	1
Real	0	416.664	480
	1	132	1.133

Ponto de Corte: 50%

		Predita	
		0	1
Real	0	417.044	100
	1	372	893

Quadro 6: Medidas de Acurácia - Validação do Cenário I

Medida	Treino			
	10%	15%	25%	50%
R	94,78%	92,17%	89,57%	70,59%
E	99,18%	99,66%	99,88%	99,98%
P	25,92%	45,21%	70,24%	89,93%
F_1	40,71%	60,67%	78,74%	79,10%
F_2	61,90%	76,32%	84,89%	73,77%
F_3	74,89%	83,50%	87,17%	72,14%
F_4	81,97%	86,87%	88,14%	71,50%
A	99,17%	99,64%	99,85%	99,89%

Cenário II

Tabela 11: Matrizes de Confusão da Validação no Cenário II

Ponto de Corte: 10%

		Predita	
		0	1
Real	0	211.097	206.047
	2	66	1.263

Ponto de Corte: 15%

		Predita	
		0	1
Real	0	314.347	102.797
	1	5	1.260

Ponto de Corte: 25%

		Predita	
		0	1
Real	0	385.508	31.636
	1	16	1.249

Ponto de Corte: 50%

		Predita	
		0	1
Real	0	41.4005	3.139
	1	58	1.207

Quadro 7: Medidas de Acurácia - Validação do Cenário II

Medida	Undersampling 50%			
	10%	15%	25%	50%
R	99,84%	99,60%	98,74%	95,42%
E	50,61%	75,36%	92,42%	99,25%
P	0,61%	1,21%	3,80%	27,77%
F_1	1,21%	2,39%	7,31%	43,02%
F_2	2,97%	5,77%	16,46%	64,16%
F_3	5,78%	10,91%	28,21%	76,73%
F_4	9,44%	17,23%	39,97%	83,46%
A	50,75%	75,43%	92,44%	99,24%

Cenário III

Tabela 12: Matrizes de Confusão da Validação no Cenário III

Ponto de Corte: 10%

		Predita	
		0	1
Real	0	363.727	54.417
	1	18	1247

Ponto de Corte: 15%

		Predita	
		0	1
Real	0	392.319	24.825
	1	26	1.239

Ponto de Corte: 25%

		Predita	
		0	1
Real	0	408.479	8.665
	1	51	1.214

Ponto de Corte: 50%

		Predita	
		0	1
Real	0	416.428	716
	1	114	1.151

Quadro 8: Medidas de Acurácia - Validação Cenário III

Medida	Undersampling 25%			
	10%	15%	25%	50%
R	98,58%	97,94%	95,97%	90,99%
E	86,95%	94,05%	97,92%	99,83%
P	2,24%	4,75%	12,29%	61,65%
F_1	4,38%	9,07%	21,79%	73,50%
F_2	10,27%	19,90%	40,63%	83,08%
F_3	18,60%	33,08%	57,09%	86,85%
F_4	27,93%	45,49%	68,52%	88,51%
A	86,99%	94,06%	97,92%	99,80%

Cenário IV

Tabela 13: Matrizes de Confusão da Validação no Cenário IV

Ponto de Corte: 10%

		Predita	
		0	1
Real	0	312.834	104.310
	1	25	1.240

Ponto de Corte: 15%

		Predita	
		0	1
Real	0	388.842	28.662
	1	47	1.218

Ponto de Corte: 25%

		Predita	
		0	1
Real	0	412.861	4.283
	1	87	1.178

Ponto de Corte: 50%

		Predita	
		0	1
Real	0	416.876	268
	1	187	1.078

Quadro 9: Medidas de Acurácia - Validação Cenário IV

Medida	Oversampling 50%			
	10%	15%	25%	50%
R	98,02%	96,28%	93,12%	85,22%
E	74,99%	93,13%	98,97%	99,94%
P	1,17%	4,08%	21,57%	80,09%
F_1	2,32%	7,82%	35,03%	82,57%
F_2	5,61%	17,43%	55,98%	84,14%
F_3	10,60%	29,52%	69,93%	84,68%
F_4	16,76%	41,31%	77,92%	84,90%
A	75,06%	93,14%	98,96%	99,89%

Cenário V

Tabela 14: Matrizes de Confusão da Validação no Cenário V

Ponto de Corte: 10%

		Predita	
		0	1
Real	0	397.474	19.670
	1	34	1231

Ponto de Corte: 15%

		Predita	
		0	1
Real	0	408.697	8.447
	1	59	1.206

Ponto de Corte: 25%

		Predita	
		0	1
Real	0	415.443	1.701
	1	106	1.159

Ponto de Corte: 50%

		Predita	
		0	1
Real	0	416.951	193
	1	214	1.051

Quadro 10: Medidas de Acurácia - Validação Cenário V

Medida	Oversampling 25%			
	10%	15%	25%	50%
R	97,31%	95,34%	91,62%	83,08%
E	95,28%	97,98%	99,59%	99,95%
P	5,89%	12,49%	40,52%	84,49%
F_1	11,11%	22,09%	56,19%	83,78%
F_2	23,71%	40,98%	73,17%	83,36%
F_3	38,13%	57,32%	81,36%	83,22%
F_4	50,87%	68,58%	85,29%	83,16%
A	95,29%	97,97%	99,57%	99,90%

Vimos anteriormente que algumas das medidas mais relevantes para medir o acerto de modelos com dados desbalanceados são precisão, recall, especificidade e medida F. Como a base de treino utilizada é desbalanceada, e a medida F com $\beta = 1$ (F_1) é uma média harmônica de Precisão e Recall, optou-se por utilizá-la como critério de escolha dos modelo.

O resumo dos valores de F_1 em cada cenário e para cada ponto de corte é apresentado no quadro a seguir.

Quadro 11: Resultados F_1 Resumidos

Treino	Ponto de Corte			
	10%	15%	25%	50%
Cenário V	11,11%	22,09%	56,19%	83,78%
Cenário IV	2,32%	7,82%	35,03%	82,57%
Cenário I	40,71%	60,67%	78,74%	79,10%
Cenário III	4,38%	9,07%	21,79%	73,50%
Cenário II	1,21%	2,39%	7,31%	43,02%

Podemos observar que para todos os cenários, o maior F_1 se encontra no corte de 50%, e o melhor cenário (de acordo com a medida) é o Cenário V, que utiliza a base de treino com oversampling, fazendo com que "alvos" representassem 25% do total de observações.

Assim, optou-se por utilizar o modelo do Cenário V como modelo final.

A importância das variáveis pode ser avaliada de acordo com a redução do erro de classificação consequente de sua utilização. O quadro abaixo resume em ordem de importância (segundo redução do erro de classificação) as variáveis explicativas do modelo escolhido.

Quadro 12: Importância das Variáveis

#	Variável Explicativa	Redução do Erro de Classificação
1	Variável 5	0,037402
2	Presente no Maior Clique	0,017004
3	Centralidade de Autovetor	0,014272
4	Centralidade de Intermediação	0,013096
5	Variável 10	0,010103
6	Centralidade de Proximidade	0,008503
7	Variável 7	0,005954
8	Variável 6	0,005279
9	Centralidade de Grau	0,004693
10	Ponto de Articulação	0,003657
11	Variável 8	0,003209
12	Variável 9	0,002844
13	Nó Terminal	0,002843
14	Variável 3	0,002514
15	Variável 2	0,00211
16	Variável 1	0,000219
17	Variável 4	0,000022

6 Conclusões e Trabalhos Futuros

Considerando os resultados apresentados no decorrer deste trabalho, conclui-se que:

- É possível, dentro de suas limitações, utilizar redes sociais para buscar automaticamente os possíveis alvos de um relatório de inteligência produzido pelo COAF.
- Buscar indivíduos a um grau de distância dos citados nas comunicações que dariam origem a um relatório de inteligência foi suficiente para detectar em média 90% dos principais relacionados classificados pelos analistas.
- Medidas de centralidade e conceitos relacionados a importância de um vértice na estrutura de um grafo se mostraram úteis para diferenciar alvos e não alvos de um relatório do COAF. Especificamente, observou-se com base em uma amostra de 13.276 casos que:
    1. Alvos tem centralidade de grau 29 vezes maior, centralidade de autovetor 8 vezes maior e centralidade de intermediação 115 vezes maior do que não alvos.
    2. Em média, 93% dos pontos de articulação identificados nas redes dos relatórios se tratavam de alvos.
    3. A proporção de alvos dentro dos maiores cliques das redes dos relatórios é consideravelmente maior do que a proporção de alvos nos relatórios.
- O melhor modelo dentre os modelos de florestas aleatórias ajustados, com base na medida F_1, foi treinado com a base de dados com proporção de "alvos" representando 25% do total, em que se atingiu esta proporção por superamostragem da classe minoritária.
- O modelo final escolhido neste trabalho foi considerado útil e eficiente do ponto de vista técnico e gerencial, podendo fornecer aos analistas auxílio técnico e um direcionamento na identificação dos principais relacionados aos eventos ilícitos a serem relatados nos relatórios de inteligência. Tal assistência, pode ser capaz de reduzir o tempo despendido na construção e análise de um RIF, trazendo maior eficiência e melhor utilização dos recursos da instituição.
- O modelo não apenas identifica em alvo ou não alvo, mas apresenta uma probabilidade para cada caso. Dessa forma, o analista poderia conduzir sua investigação partindo-se dos indivíduos com maior probabilidade de serem alvos, não se restringindo a classificação consequente de um corte.
- As redes construídas para cada relatório têm valor em si mesmas. A simples visualização do grafo, acompanhada das probabilidades ajustadas pelo modelo, podem facilitar o processo de investigação, na medida em que avaliação dos riscos de diferentes indivíduos não é feita de maneira separada.
- A automação que se dá pelo uso do modelo proposto nesse trabalho, busca reproduzir um processo analítico, o qual pode ser deficiente. Nesse sentido, casos em que o modelo tenha classificado um indivíduo com alvo e que ele não

tenha sido incluído no relatório do COAF pode indicar que a técnica é capaz de detectar alvos que não seriam percebidos pelos analistas. Isso porque o método é capaz de avaliar uma rede maior do que seria factível para um humano.

Apesar do desempenho satisfatório desta primeira versão do modelo de detecção de principais relacionados de um relatório de inteligência financeira, foram propostas algumas melhorias e planos para trabalhos futuros, que não foram feitas até o momento. Dentre essas proposta está:

- Buscar por novas bases de dados que forneçam informações capazes de atribuir novos tipos de relacionamento para construir as arestas das redes de possíveis envolvidos do relatório.
- Considerar utilizar pesos distintos para os diferentes tipos de aresta considerados, valorizando os relacionamentos considerados mais importantes.
- Adicionar as variáveis cujas extrações haviam sido deixadas para um segundo momento.
- Avaliar a utilização de mais um nível de distância do grupo de indivíduos iniciais para buscar os candidatos a principais relacionados.
- Testar outros tipos de modelo de classificação, como regressão logística, análise de discriminantes e máquina de vetores de suporte.

Referências

[1] Ministério da Economia. *Conselho de Controle de Atividades Financeiras*, 2019. `http://www.fazenda.gov.br/orgaos/coaf`. Citado 2 vezes nas páginas 8 e 9.

[2] Conselho de Controle de Atividades Financeiras. *O que é a Unidade de Inteligência Financeira?*, 2019. Citado 2 vezes nas páginas 8 e 28.

[3] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. Citado 3 vezes nas páginas 14, 16 e 17.

[4] M.E.J. Newman. *Networks: An Introduction*, volume 1. Oxford University Press, 2010. Citado 3 vezes nas páginas 14, 15 e 21.

[5] Leandro Quintanilha de Freitas. *Medidas de centralidade em grafos*. PhD thesis, dissertação de mestrado, Universidade Federal do Rio de Janeiro, 2010. Citado 9 vezes nas páginas 14, 16, 17, 18, 19, 20, 21, 22 e 29.

[6] Nitin Nohria and R Eccles. Is a network perspective a useful way of studying organizations. *Leading Organizations: Perspectives for A New Era; Robinson, HG, Ed*, pages 287–301, 1992. Citado na página 14.

[7] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977. Citado na página 14.

[8] John F Padgett and Christopher K Ansell. Robust action and the rise of the medici, 1400-1434. *American journal of sociology*, 98(6):1259–1319, 1993. Citado na página 15.

[9] Ronaldo Goldschmidt and Emmanuel Passos. *Data mining: um guia prático*. Gulf Professional Publishing, 2005. Citado na página 17.

[10] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011. Citado na página 22.

[11] Scott Hartshorn. *Machine Learning With Random Forests And Decision Trees*. Citado 3 vezes nas páginas 22, 23 e 24.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: With Applications in R*, volume 112. Springer, 2013. Citado na página 23.

[13] Giang Hoang Nguyen, Abdesselam Bouzerdoum, and Son Lam Phung. Learning pattern classification tasks with imbalanced data sets. In *Pattern recognition*. IntechOpen, 2009. Citado 3 vezes nas páginas 24, 26 e 27.

[14] Yutaka Sasaki et al. The truth of the f-measure. *Teach Tutor mater*, 1(5):1–5, 2007. Citado na página 26.

[15] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research*, 19:315–354, 2003. Citado 2 vezes nas páginas 26 e 27.

[16] Conselho de Controle de Atividades Financeiras. *Relatório de Atividades 2018*, 2019. Citado na página 29.

[17] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977. Citado na página 33.

[18] Duncan J Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2004. Citado na página 34.