Método Baseado em Estatística e Aprendizado de Máquina para Detecção de Fraude em Concursos Públicos

Roberto Paulo Moreira Nunes

Janeiro / 2022

Dissertação de Mestrado em Ciência da Computação

Método Baseado em Estatística e Aprendizado de Máquina para Detecção de Fraude em Concursos Públicos

Esse documento corresponde à dissertação apresentada à Banca Examinadora no curso de Mestrado em Ciência da Computação da UNIFACCAMP – Centro Universitário Campo Limpo Paulista.

Campo Limpo Paulista, 27 de janeiro de 2022.

Roberto Paulo Moreira Nunes

Orientador: Prof. Dr. Ferrucio de Franco Rosa

O presente trabalho foi realizado com apoio da
Coordenação de Aperfeiçoamento de Pessoal de Nível
Superior - Brasil (CAPES) - Código de Financiamento 001.
Número do Processo: 88887.625555/2021-00.

FICHA CATALOGRÁFICA

Ficha catalográfica elaborada pela

Biblioteca Central da Unifaccamp

N928m

Nunes, Roberto Paulo Moreira

Método baseado em estatística e aprendizado de máquina para detecção de fraude em concursos públicos / Roberto Paulo Moreira Nunes. Campo Limpo Paulista, SP: Unifaccamp, 2022.

Orientador: Profº. Dr. Ferrucio de Franco Rosa

Dissertação (Programa de Mestrado Profissional em Ciência da Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Método de detecção de fraude. 2. Concurso público. 3. Mineração de dados. 4. Estatística. I. Rosa, Ferrucio de Franco. II. Centro Universitário Campo Limpo Paulista. III. Título.

CDD- 005.8

Dedico este trabalho à minha família.

AGRADECIMENTOS

Ao meu orientador Prof. Dr. Ferrucio de Franco Rosa, pelas valiosas e incontáveis horas dedicadas a este orientado, pela maneira otimista, sistematizada e clara com a qual conduziu sua orientação, sempre disponível e assertivo, coluna mestra na produção de dois artigos publicados e na conclusão dessa dissertação.

A todos os professores do Programa de Mestrado em Ciência da Computação da UNIFACCAMP, em especial ao Prof. Dr Osvaldo, coordenador do curso; ao Prof. Dr. Rodrigo Bonacin, também partícipe em um dos artigos publicados e membro da banca de qualificação; à Prof. Dra. Ana Monteiro, pela excelente condução da disciplina de Mineração de Dados e pela participação nas bancas de qualificação e mestrado.

Ao Exército Brasileiro, instituição a qual pertenço há 28 anos, contumaz incentivadora do autoaperfeiçoamento dos seus quadros, pela autorização e liberação parcial de expediente para a realização do Curso de Mestrado.

Ao incentivo e apoio irrestrito de minha esposa Nereida, pelas diversas trocas de ideias e pela tolerância aos meus frequentes desvios de humor. Ainda no âmbito familiar, agradeço ao meu filho Fabrício, jovem prodígio da matemática, por pontuais e profícuas sugestões. E, ainda, à minha filha Fabiana pelos carinhosos bilhetes de incentivo colados no meu monitor.

Ao Major Alex Sandro Faria Manuel, mestre e professor de Cálculo, pelo incentivo constante, pelas orientações matemáticas e o apoio na obtenção de artigos.

RESUMO

O concurso público é uma forma eficaz de selecionar servidores, civis e militares, para admissão em diversos setores do serviço público. Esses exames geralmente atraem pessoal bem treinado, apresentando uma seleção altamente competitiva e meritocrática. No entanto, também atraem criminosos que oferecem aos candidatos a possibilidade de admissão fácil e ilegítima. Com o objetivo de proporcionar maior segurança e confiança aos processos seletivos, propõe-se o Método de Detecção de Fraude em Concursos Públicos (MDFCP), que provê um índice de suspeição, utilizando Machine Learning e Estatística. O método engloba quatro análises: (i) perfil do candidato, com base em dados geográficos, biográficos e de desempenho; (ii) respostas às questões de múltipla escolha, avaliadas isoladamente e em comparação aos outros candidatos; (iii) verificação de notas outliers agrupadas por afinidade e (iv) registros anteriores de suspeição de origens diversas. Destas análises resultam índices parciais que, aplicados pesos específicos a cada um, resultarão em um índice de suspeição final para cada candidato de uma determinada ocorrência de concurso. Um candidato avaliado pelo método, quando apresenta esse índice de suspeição acima de um determinado limite, é colocado em uma lista de suspeição, e exigirá uma investigação mais aprofundada. Cada uma das quatro análises que compõem o método MDFCP foi validada, com a aplicação de três anos de bases de dados parcialmente simuladas. Especificamente para o teste das respostas em questões de múltipla escolha foi utilizada uma técnica de inserção simulada de fraude, que possibilitou mensurar a eficiência de detecção de fraude desse componente do método. Como resultado da aplicação do método foram encontrados valores significativos nas análises parciais e obtenção de valores de suspeição final, gerando percentual médio de 3,04% de candidatos suspeitos nas três ocorrências de concurso analisadas. Ao final, o método se mostrou mais abrangente e parametrizável do que métodos similares, podendo determinar pesos que desconsiderem totalmente ou aumentem significativamente a importância das quatro análises realizadas.

Palavras-chave: Método de Detecção de Fraude, Concurso Público, Mineração de Dados, Estatística.

ABSTRACT

A public service examination is an effective manner of selecting civil and military servants for admission to various sectors of public service. These examinations usually attract well-trained personnel by presenting highly competitive and meritocratic selection. Nevertheless, they also attract criminals that offer candidates the possibility of easy and illegitimate admission. We propose the Method for Detecting Fraud in Public Service Examinations (MDFPSE), providing a suspicion index by using Machine Learning and Statistics. We are aiming to provide security and trust in the selection processes. MDFPSE comprises four analyses: (i) candidate's profile, based on geographic, biographical, and performance data; (ii) multiple-choice questions answers, evaluated separately and compared to other candidates; (iii) verification of outlier's grades grouped by affinity; (iv) previous suspicion records from different sources. These analyzes result in partial indices that, applying specific weights to each one, will result in a final suspicion index for each candidate in each public examination. A candidate evaluated by the method with its suspicion index above a certain threshold is placed on a suspicion list and could require further investigation. Each of the four analyzes of MDFPSE was validated, by the application of three years of partially simulated databases. Specifically for the test of answers in multiple-choice questions, we used a simulated fraud insertion technique to measure the fraud detection efficiency of this method component. As a result of applying the method, significant values were found in the partial analyzes and obtaining of final suspicion values, generating an average percentage of 3.04% of suspect candidates in the three instances of examination analyzed. MDFPSE proved to be more comprehensive and parameterizable than similar methods, being able to determine weights that totally disregard or significantly increase the importance of the four analyzes performed.

***Keywords**:* Method for Fraud Detection, Data Mining, Statistics, Public Service Examination.

SUMÁRIO

GLOSSÁRIO

Abreviatura/Sigla	Significado
API	*Application Programming Interface*
BAC	*Baccalauréat /Bacharelado (França)*
BD	*Banco de Dados*
CAp	*Candidato Aproveitado*
ECG	*Eletrocardiograma*
ED	*Especialista do Domínio*
ENEM	*Exame Nacional do Ensino Médio*
Gab	*Gabarito*
Gaokao	*Exame Nacional Para o Ingresso no Ensino Superior (China)*
HCA	*Hierarchical Cluster Analysis*
HMM	*Hidden Markov Method*
IDQ	*Índice de Dificuldade da Questão*
IS	*Índice de Suspeição*
ISF	*Inserção Simulada de Fraude*
IVL	*Incremental Virtual Learning*
LOF	*Local Outlier Factor*
LS	*Limiar de Suspeição*
LSi	*Limiar de Similaridade*
KNN	*K-Nearest Neighbors*
KNNW	*K-Nearest Neighbors-Weighted*
LSi	*Limiar de Similaridade*
MDFCP	*Método para Detecção de Fraude em Concurso Público*
ML	*Machine Learning*
PCA	*Probabilidade do Candidato ser Aproveitado*
PCA	*Análise do Componente Principal*
RC	*Respostas Corretas*
RL	*Regressão Logística*
RQME	*Respostas às Questões de Múltipla Escolha*
SAT	*Scholastic Assessment Test (EUA)*
TF-IDF	*Term Frequency — Inverse Document Frequency*

LISTA DE TABELAS

LISTA DE FIGURAS

1 INTRODUÇÃO

Os governos cumprem suas missões com o apoio de servidores públicos, em sua maioria, de carreira. Normalmente, os cargos oferecidos visam perfis específicos e são preenchidos por meio de concursos públicos, selecionando os candidatos com base no mérito profissional (L. Barbosa, 2014). Dentre esses cargos, encontram-se funcionários administrativos, juízes, procuradores, assistentes legislativos, médicos, professores, militares, policiais, entre outras carreiras diversas, que totalizam, nas diversas esferas de poder no Brasil, uma massa de aproximadamente 11,4 milhões de servidores públicos e militares, sendo que o poder executivo municipal é o grande empregador no setor público brasileiro, com 6,4 milhões (Guedes *et al*., 2021).

Os concursos públicos apresentam grande concorrência, em grande parte devido à estabilidade e aos benefícios proporcionados por essas vagas. Um destes benefícios é o salário, normalmente maior do que os de cargos similares oferecidos pela iniciativa privada (A. L. N. de H. Barbosa & Souza, 2012). Vagas importantes são disputadas por profissionais altamente qualificados, que passam por complexos exames e testes de avaliação. Com efeito, muitas vezes existe um ecossistema em torno de concursos públicos envolvendo diversos atores, tais como organizadores, candidatos, inspetores, escolas e cursos preparatórios, sites, editoras, entre outros. Os concursos se norteiam por meritocracia, e os envolvidos se dedicam, dispensam tempo e ambicionam uma carreira pública, depositando a confiança de que participam de processos lícitos.

Os concursos são, com alguma variação, compostos por uma primeira etapa com provas objetivas (múltipla-escolha) e discursivas/dissertativas. Etapas posteriores podem contemplar exames médicos, psicológicos, físicos, provas de títulos e provas orais. Requisitos, métodos de seleção e outros detalhes variam de acordo com o país e as idiossincrasias dos cargos oferecidos. Além dos concursos públicos, existem exames para admissão ao Ensino Superior, também com provas de múltiplas escolha e dissertativas, tais como o ENEM, no Brasil, Gaokao, na China, SAT, nos Estados Unidos, Bac, na França, dentre outros. Embora não sejam meios de admissão em carreiras públicas, o são para cursos de nível superior com alta concorrência.

Neste cenário, surgem oportunidades para que criminosos, especialistas em fraudes em concursos públicos, ofereçam aos candidatos a possibilidade de admissão

facilitada e ilegal. As fraudes podem ser realizadas de diversas formas, tais como manipulação de resultados, substituição por terceiros por meio de documentação falsa de identidade, obtenção prévia de gabaritos ou provas, ou ainda a utilização de dispositivos eletrônicos para transmissão de respostas aos candidatos. O preço cobrado pelas quadrilhas é relativo à importância da posição desejada e à técnica de fraude.

Para evitar fraudes, procedimentos e dispositivos de segurança devem ser previamente implementados, tais como: i) seleção criteriosa do pessoal envolvido na organização do concurso; ii) manter um alto nível de confidencialidade na preparação das provas; iii) adoção de uma logística segura; iv) uso de fiscais de sala; v) uso de detectores de ondas eletromagnéticas; vi) coleta de dados biométricos (por exemplo, impressões digitais, fotografia e amostra caligráfica). Os concursos públicos, normalmente, compreendem algumas etapas, tais como testes intelectuais (e.g., testes de múltipla escolha, redações), verificação de qualificação ou título (e.g., mestrado, doutorado ou qualificações específicas), testes físicos (tipicamente para carreiras policiais e militares), exames médicos e avaliação psicológica, como pode ser observado em exemplos apresentados na Tabela 1.1.

Tabela 1.1 – Exemplos de Concursos para Carreiras Públicas.

Esfera/Carreira	Nível	Tipo Prova	Outras etapas	Efetivo(*)
Federal/Sargento do Exército	Médio	Objetiva, discursiva	Física, Médica e Psicológica	126.000
Federal/Agente da Polícia Federal	Superior	Objetiva, discursiva	Física, Médica e Psicológica	220.000
Federal/Policial Rodoviário Federal	Superior	Objetiva, discursiva	Física, Médica, Psicológica e Títulos	304.000
Federal/Auditor-Fiscal da Receita Federal	Superior	Objetiva, discursiva	Sindicância da vida pregressa	68.500
Estadual/Escrivão da Polícia Civil de Minas Gerais	Superior	Objetiva, discursiva e prática	Médica, Física, Títulos e Sindicância	35.000
Estadual/Defensoria Pública do Mato Grosso	Superior	Objetiva	Médica	21.500
Municipal/Prefeitura de São Paulo	Médio/ Superior	Objetiva	Prova de Títulos, médica	118.000
Municipal/Prefeitura de Salvador	Médio/ Superior	Objetiva	Médica, física e Psicológica	99.000

(*) Efetivo aproximado de candidatos. Fonte: Comissões Organizadoras.

As instituições responsáveis pelo gerenciamento dos exames mantêm os dados de inscrição do candidato (e.g., biográficos, geográficos), bem como informações sobre o tempo de execução do teste, as respostas aos testes de múltipla escolha e as pontuações obtidas por assunto. Além disso, eles podem armazenar dados de exames anteriores. Este conjunto de dados, como será visto posteriormente, é fundamental para a aplicação de métodos de detecção de fraude.

1.1 Questão de Pesquisa, Contribuição Principal e Objetivos

A partir da problemática exposta, aborda-se a seguinte **questão de pesquisa**: "*Como identificar e atribuir computacionalmente, com base em dados de um concurso público, um grau de suspeição de fraude a candidatos?*"

Uma solução para a questão de pesquisa passa pela superação de alguns óbices: i) o custo computacional de verificação entre candidatos em concursos com um grande número de inscritos; ii) identificar e testar as soluções de aprendizado de máquina/mineração de dados e estatística mais apropriados para o desenvolvimento do método; e iii) contornar a inserção de viés em bases de dados incompletas.

Como contribuição principal propõe-se um método para detecção de fraude em concursos públicos, cujas avaliações contenham provas de múltipla escolha (combinadas ou não com questões dissertativas), utilizando um conjunto de dados mantido pelos gestores dos exames. Com base em técnicas estatísticas e de mineração de dados, é gerado um índice de suspeição que, filtrado por parâmetros, irá compor uma lista final de suspeição. O método proposto poderá ser utilizado em diferentes concursos públicos, vestibulares e exames e deve ser aplicado ao conjunto de candidatos aprovados para identificar potenciais fraudadores.

Como objetivos da pesquisa, destacamos: i) conceber e desenvolver o método de detecção de fraudes; ii) definir indicadores de suspeição de candidatos; e iii) propor um processo de aplicação do método.

1.2 Estrutura da Dissertação

Esta dissertação está organizada da seguinte forma: o Capítulo 2 apresenta uma síntese de uma revisão sistemática da literatura sobre métodos de detecção de fraude em diversos domínios, além de métodos com objetivos similares ao proposto nesse projeto.

O Capítulo 3 apresenta um referencial teórico, contendo as técnicas de mineração de dados e de estatística que serviram de base para o desenvolvimento do Método de Detecção de Fraude em Concursos Públicos (MDFCP). No Capítulo 4 é apresentado o MDFCP. O Capítulo 5 apresenta a aplicação do método. O Capítulo 6 apresenta uma discussão. O Capítulo 7 apresenta as conclusões e trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta uma síntese da revisão sistemática de literatura (Nunes, Bonacin, *et al*., 2021), que foi conduzida para identificar métodos voltados à detecção de fraude em diferentes domínios. Na Seção 2.1 é descrito o protocolo de revisão; na Seção 2.2 é apresentado o resultado da revisão e na Seção 2.3 apresentam-se os trabalhos relacionados.

2.1 Protocolo de Revisão

O mapeamento sistemático da literatura realizado foi baseado, com adaptações, no método de Kitchenham (2004). O processo de revisão, incluindo as atividades e os documentos produzidos, é apresentado resumidamente na Figura 2.1.

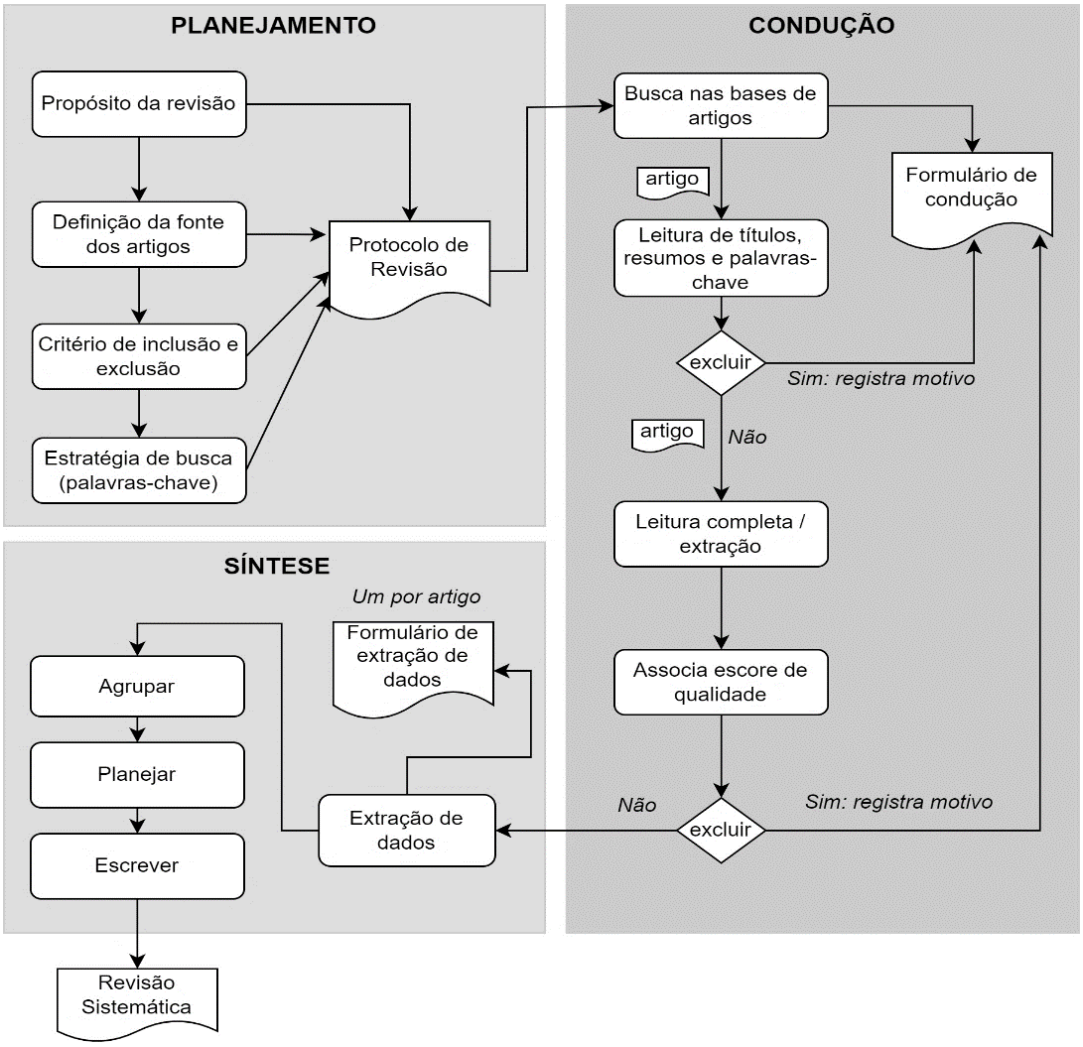

Figura 2.1 - Fases do Processo de Revisão literatura (Nunes, Bonacin, *et al*., 2021)

O processo proposto consiste em 3 fases: planejamento, condução e síntese. O planejamento foi realizado com base no protocolo de revisão, contendo o tema e os objetivos da revisão. Foram selecionadas as seguintes bases de dados científicas, uma vez que abrangem pesquisas relevantes sobre o tema: ACM *Digital Library*[1], IEEE *Xplore*[2], *Google Scholar*[3] e *Springer Link*[4].

Os seguintes critérios de inclusão foram definidos:

1	Artigos oriundos de periódicos ou anais de eventos científicos, cujo texto completo esteja disponível nas bases de dados científicas
2	Trabalhos publicados após 2010
3	Trabalhos que apresentam métodos e técnicas para detecção de fraude
4	Trabalhos publicados na Língua Inglesa

Os seguintes critérios de exclusão foram definidos:

1	Área de pesquisa diferente de Ciência da Computação
2	Trabalhos que não enfocam tópicos relacionados à revisão
3	Artigos curtos e resumos

A revisão buscou identificar trabalhos que expusessem abordagens voltadas à detecção de fraude em exames ou concursos admissionais. Assim, a questão motivadora para a *string* de busca é: Quais são os métodos, arquiteturas ou processos disponíveis voltados a detectar ou avaliar fraudes em concursos ou exames de seleção pública?

Foram definidas as palavras-chave e uma *string* de busca norteadora: *"(method OR procedure OR process OR architecture) AND (detection OR evaluation OR assessment) AND (contest OR competition OR examination) AND fraud*".

Durante a fase de condução, a *string* de busca foi adaptada, de acordo com as particularidades dos buscadores (Tabela 2.1); especificamente, foram consideradas a forma de inserção de palavras-chave, campos selecionados e refinamento nas buscas. No

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org
[3] https://scholar.google.com.br/
[4] https://link.springer.com/

caso de retorno nulo de artigos, as palavras-chave foram ajustadas para uma restrição menor.

Tabela 2.1 – Strings de Busca e Quantidade de Artigos Coletados

Bases de dados científicas e *string* de busca específica	**Coletados (incluídos)**
IEEE Xplore (2010-2020): (("Document Title":"method" OR "Document Title":"procedure" OR "Document Title": architecture") AND ("Document Title": "detection" OR "Document Title": "evaluation" OR "Document Title": "assessment") AND ("Document Title": "fraud")).	16 (9)
IEEE Xplore (2018-2020): ((("Document Title":detection) AND " Document Title ":fraud) AND " Document Title ":survey).	1 (1)
Springer Link (2010-2020): field "where the title contains": "detection fraud".	1 (1)
Google Scholar (2010-2020): allintitle: method OR process "detection fraud".	6 (2)
ACM Digital Library (2010-2020): [[[Publication Title: "method"] OR [Publication Title: "procedure"] OR [Publication Title: "architecture"]] AND [[Publication Title: "detection"] OR [Publication Title: "evaluation"] OR [Publication Title: "assessment"]] AND [Publication Title: "fraud"]	7 (2)

Restrições de período: Artigos publicados entre 2010 e 2020; e as revisões de literatura mais recentes, publicadas entre 2018 e 2020. Outros 5 estudos foram citados pelos estudos analisados (*Snowballing*), neste caso, mesmo fora dos períodos, foram incluídos na revisão. Uma planilha auxiliar e o software *Mendeley* foram utilizados nas fases de condução e síntese.

Foi realizada uma leitura completa das obras selecionadas. Embora preconizado por Kitchenham (2004), nenhum artigo foi adicionado ou excluído com base nos critérios de qualidade da pesquisa. Na fase de síntese, os trabalhos coletados nas fases anteriores foram classificados, comparados e resumidos.

2.2 Resultados da Revisão

Nesta seção é apresentada uma análise dos estudos selecionados. A Subseção 2.2.1 apresenta soluções baseadas em Aprendizado de Máquina / *Data Minning*; a Subseção 2.2.2 apresenta soluções baseadas em técnicas estatísticas; a Subseção 2.2.3

apresenta pesquisas relacionadas à detecção de fraudes em exames de múltipla escolha; a Subseção 2.2.4 apresenta pesquisas sobre fraude em licitações públicas e a Subseção 2.2.5 apresenta uma análise sintética dos artigos selecionados.

2.2.1 Soluções baseadas em Aprendizado de Máquina / *Data Mining*

Uma abordagem para detecção de fraude em leilões *online* é proposta por Chang & Chang (2012). O objetivo é reduzir a quantidade dos atributos avaliados para descrever as características dos participantes desta modalidade de vendas. O comportamento dos vendedores é avaliado, considerando apenas a quinta extremidade do histórico disponível, reduzindo assim os esforços computacionais. Ao reduzir o histórico de transações, o custo total de detecção pode ser bastante reduzido, ao mesmo tempo em que mantém uma precisão de detecção razoável. O modelo construído para validar o estudo considerou diversas técnicas de mineração de dados, a saber: C.45, Cart, *Naive Bayes Trees*, Regressão Logística e *Ada Boost* (Wu *et al*., 2008). A partir da aplicação das técnicas, uma precisão de 91% a 95% foi obtida na detecção dos modelos de perfil tardio construídos.

Uma abordagem para detecção de fraude em compras com cartão de crédito é proposta por Xie *et al*. (2019). Os recursos são extraídos, com base na frequência das transações do usuário, e nas regras de comportamento individual e de grupo, para classificar as transações como legítimas ou fraudulentas. Para validar a proposta, a técnica *Randon Forest* (Breiman, 2001) foi utilizada como classificador binário.

Zhang *et al*. (2020) apresentam uma solução para aumentar a precisão na detecção de fraudes de usuários com baixa frequência de transações, ou seja, usuários que não permitem a criação de um perfil individual mais preciso. Um algoritmo de *clustering* DBSCAN é usado para atribuir o usuário a um grupo de usuários semelhantes. Em seguida, por meio de um algoritmo Naive Bayes, verifica-se se uma transação é fraudulenta, considerando o ambiente privado e de grupo.

Um método que considera análises em subespaços definidos por duas ou três variáveis registradas nas transações é apresentado em Salazar *et al*. (2019). Destes subespaços, a velocidade e a aceleração da transação são estimadas como vetores de entrada para um processo de classificação. A análise discriminante linear e quadrática e

a floresta aleatória são implementadas como classificadores. Os resultados da classificação obtidos para cada subespaço são então mesclados para obter um resultado geral usando o algoritmo de integração alfa. As redes neurais são usadas para construir modelos de detecção de fraudes financeiras. No entanto, a precisão dos modelos diminui com o tempo, quando eles são implantados em sistemas de detecção *online*.

Ma *et al*. (2019) abordam a aplicação de redes neurais para construir modelos de detecção de fraudes financeiras. Os autores ressaltam que o desempenho dos modelos diminui com o tempo, quando eles são implantados em sistemas de detecção *online*. Para manter o desempenho do modelo em sistemas *online*, quando os rótulos de novas transações não estão disponíveis, é proposto um método de aprendizado virtual incremental (IVL), para atualizar as redes neurais continuamente.

Um método baseado em federação para detectar fraudes de cartão de crédito é proposto por Yang *et al*. (2019). É proposta a utilização descentralizada de ML nos dados de cada instituição da federação. O modelo de aprendizagem é então compartilhado com as outras entidades federadas, sem expor seus dados de transação.

Zhao *et al*. (2019) apresentam um método baseado no algoritmo de propagação de rótulos para extrair recursos de uma rede conectada. A partir da rede conectada que contém usuários fraudulentos conhecidos e a relação entre eles, um algoritmo de propagação de rótulo personalizado é usado para inferir a probabilidade de fraude por parte do usuário desconhecido.

Y. J. Chen & Wu (2017) abordam o uso de *big data* em conjunto com ML para identificar fraudes nos balanços das empresas.

Um método de agrupamento de ambientes de usuário baseado em dados não balanceados para detecção de fraude de cartão de crédito é proposto por Li & Xie (2019). Os autores propõem dividir o ambiente do usuário em vários grupos de ambientes usando k-means, remover o ruído e classificar a amostra. O cluster k-means é usado e, em seguida, discrepâncias são encontradas nos clusters resultantes usando HMM (Método de

Markov Oculto). O algoritmo divide efetivamente os números em *clusters* e, em seguida, detecta *outlier*s; o número do cartão de crédito é validado usando o algoritmo de *Luhn*[5].

Um *framework* para detecção de fraude em transações com cartão de crédito é proposto por Eshghi & Kargari (2019). O *framework* consiste nos seguintes componentes: i) um componente baseado em regras, que usa uma árvore de decisão; ii) um componente que realiza uma análise de tendência, com cálculo de dissimilaridades (semi-supervisionado); e iii) um componente baseado em cenário, onde é calculada a extensão das semelhanças na sequência de transações com os cenários de fraude conhecidos.

2.2.2 Soluções baseadas em técnicas estatísticas

Alnajem & Zhang (2013) propõem um método de detecção de fraude, que mede um valor de risco de fraude para uma determinada transação de *mobile-banking*. Ao contrário dos métodos existentes, que geralmente assumem que os diferentes fatores de risco para fraude marginal são independentes uns dos outros, o método proposto poderia capturar padrões de fraude evasivos causados por fatores de risco de fraude que são dependentes ou independentes uns dos outros.

Um método de detecção de fraudes em transações *online* baseado em ambientes individuais é proposto por L. Chen *et al*. (2019). Ao considerar várias dimensões de registros de transações históricas, um ambiente de transação do usuário é gerado. Em seguida, um algoritmo é proposto para determinar o limite de risco ótimo para cada usuário. Por fim, combinando o ambiente de transação e o limite de risco ideal, forma-se um *benchmark* do ambiente do usuário, que será utilizado para construir o modelo da hiperesfera multidimensional. As transações são ajustadas a esta hiperesfera, mostrando se as novas transações são normais ou fraudulentas.

Um método baseado em assinatura para detecção de fraudes no comércio eletrônico é proposto por Belo *et al*. (2016). A assinatura proposta é definida por um conjunto de atributos que recebem um conjunto de variáveis relacionadas ao comportamento de um usuário em um ambiente de *e-commerce*. A assinatura aponta

[5] O algoritmo de Luhn é uma fórmula de soma de verificação frequentemente usada para validar números de cartão de crédito

desvios de comportamento em relação à atividade recente do usuário, permitindo detectar potenciais situações de fraude.

2.2.3 Detecção de fraude em exames de múltipla escolha

Um método para detecção e avaliação de trapaça em exames de admissão à universidade usando classificação supervisionada é proposto por Cavalcanti *et al*. (2012). Wesolowsky (2000) usa mineração de dados e correção de Bonferroni para detectar similaridade excessiva em respostas de múltipla escolha entre pares de candidatos.

Algoritmos de mineração de dados e ferramentas estatísticas (agrupamento hierárquico, árvore de dendrogramas e pesos de dificuldade da pergunta) são aplicados por M. Chen (2017) para determinar semelhanças entre respostas. Uma análise visual é realizada, usando um mapa de calor para identificar padrões nas pontuações do exame.

2.2.4 Pesquisas sobre fraude em licitações públicas

Carvalho *et al*. (2009) propõem o uso de uma ontologia probabilística, por meio de Probabilistic OWL (PR-OWL), para automatizar a detecção de fraudes em compras públicas. Ralha & Silva (2012) propõem um modelo que visa extrair informações de big data relacionadas às compras governamentais. Com o objetivo de identificar a formação de cartéis (associação fraudulenta entre empresas) em contratações públicas, foi criada uma ferramenta informática utilizando técnicas de *data mining* (regras de agrupamento e associação) em conjunto com uma abordagem multiagente.

2.2.5 Análise sintética dos artigos selecionados

A Figura 2.2 apresenta um resumo dos trabalhos analisados, categorizados por técnicas e a Figura 2.3 apresenta os domínios de aplicação tratados nos artigos. Em geral, os domínios de aplicação predominantes estavam relacionados a transações financeiras, especialmente aquelas realizadas *online* com o uso de cartões de crédito.

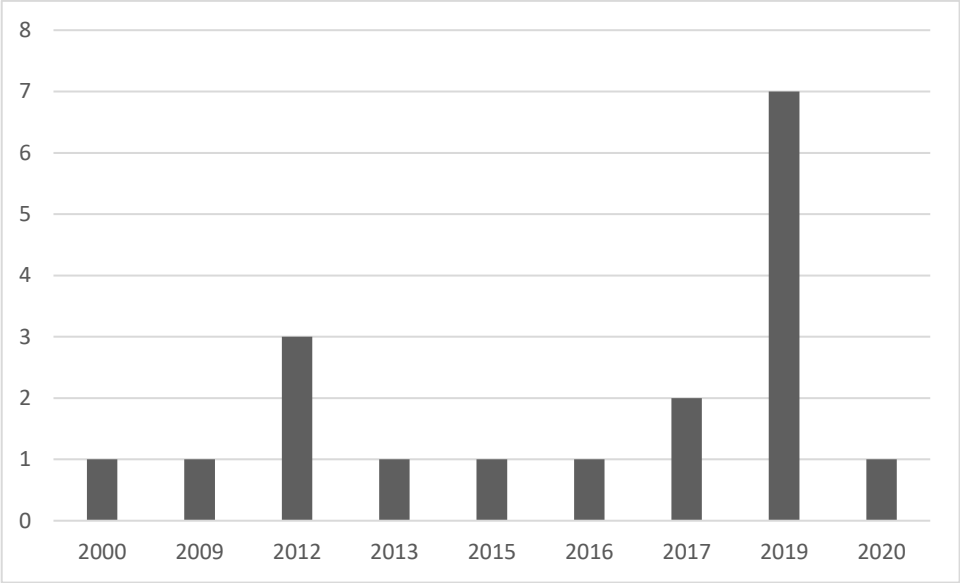

Figura 2.2 – Artigos classificados por ano de publicação

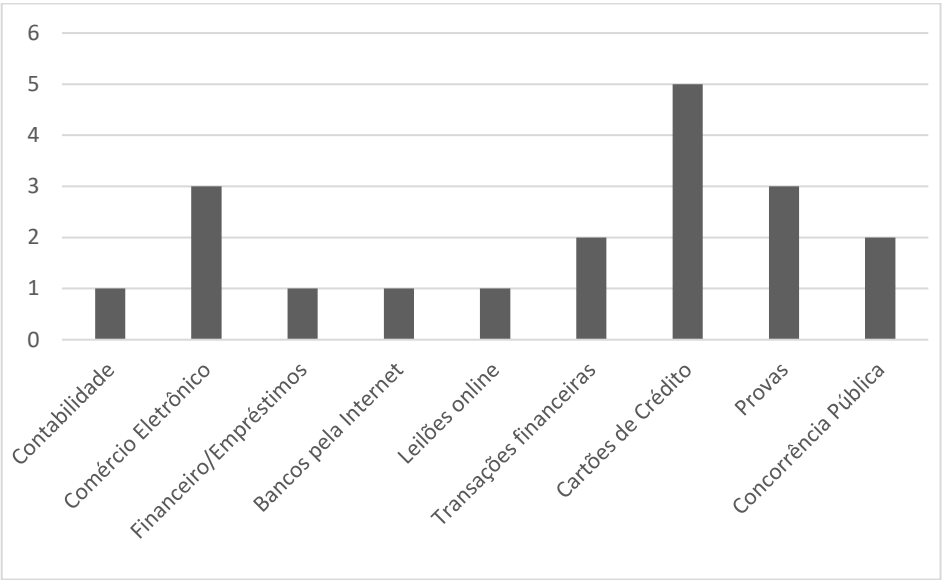

Figura 2.3 – Artigos classificados por domínio de aplicação

A Tabela 2.2 apresenta uma síntese dos trabalhos analisados, categorizados por técnica e domínio de aplicação, conforme segue: *Técnica*: (ML) *Machine Learning*/(DM) *Data Mining*; e (ET) Estatística. *Domínio de Aplicação*: (1) Contabilidade; (2) Comércio Eletrônico; (3) Financeiro/Empréstimos; (4) Banco pela Internet; (5) Leilões on-line; (6) Transações Financeiras; (7) Transações com Cartões de Crédito; (8) Provas; e (9) Concorrência Pública.

Tabela 2.2 – Síntese dos trabalhos analisados

Referência	Técnica		Domínio de aplicação								
	ML/DM	ST	1	2	3	4	5	6	7	8	9
(Chang & Chang, 2012)	X						X				
(Xie *et al*., 2019)	X								X		
(Zhang *et al*., 2020)	X			X							
(Salazar *et al*., 2019)	X								X		
(Ma *et al*., 2019)	X							X			
(Yang *et al*., 2019)	X								X		
(Zhao *et al*., 2019)	X				X						
(Y. J. Chen & Wu, 2017)	X		X								
(Li & Xie, 2019)	X								X		
(Bhati & Sharma, 2015)	X								X		
(Eshghi & Kargari, 2019)	X			X							
(Alnajem & Zhang, 2013)		X				X					
(L. Chen *et al*., 2019)		X						X			
(Belo *et al*., 2016)		X		X							
(Cavalcanti *et al*., 2012)	X									X	
(Wesolowsky, 2000)	X									X	
(M. Chen, 2017)	X	X								X	
(Carvalho *et al*., 2009)		X									X
(Ralha & Silva, 2012)	X	X									X

2.3 Trabalhos Relacionados

No decorrer da condução da revisão sistemática, observou-se uma predominância de artigos referentes à área financeira, comércio eletrônico e uso de cartão de crédito em ambientes *online*. Por isso, como trabalhos relacionados, foram selecionados três artigos sobre fraude em provas e detalhados nesta seção, a saber: 1) Cavalcanti *et al*. (2012) propõem um método de classificação supervisionada para detecção e avaliação de fraudes em vestibulares (Subseção 2.3.1); 2) Wesolowsky (2000) usa *Data Mining* e Correção de Bonferroni para detectar similaridade excessiva em respostas de múltipla escolha entre pares de candidatos (Subseção 2.3.2); 3) M. Chen

(2017) usa algoritmos de mineração de dados e ferramentas estatísticas para determinar semelhanças entre respostas (Subseção 2.3.3).

2.3.1 Detection and evaluation of cheating on college exams using supervised classification (Cavalcanti *et al*., 2012)

Cavalcanti *et al*. (2012) contextualiza o problema de trapaças ("cola") como sendo comum, constatado em relatos acadêmicos referentes ao Brasil e a outros países do mundo, em vários segmentos da educação, sem uma solução para o problema. A começar por faltar uma definição concreta do que se enquadra como fraude. O trabalho aborda uma aplicação prática de *text mining* aplicado no domínio da educação: a avaliação e detecção de plágio ou cola em exames escolares. Diferentemente de outros trabalhos, que focam em verificação de fraude em testes de múltipla escolha, este trabalho abrange textos abertos ("dissertativos").

No trabalho são elencadas uma série de outras ferramentas comerciais e livres, que tem como objetivo a detecção de plágio (e.g., Ephorus[6], Plagium[7], Sherlock[8] e Urkund[9]). A originalidade do artigo é baseada no fato de que possui um contexto local, avaliando questões abertas de provas; e as demais ferramentas comparam textos locais com conteúdo disponível na Internet, normalmente sobre trabalhos ou artigos científicos.

Assim como as demais ferramentas, técnicas de mineração de dados são utilizadas, mais especificamente *text mining*. Uma das técnicas levantadas é a que mede a frequência de termos combinada com o inverso de frequência de termos, esse último para se evitar o sobrepeso de certos tipos de termos (e.g., artigos, proposições) na medida de similaridade (método TF-IDF). A similaridade entre dois documentos D1 e D2 pode ser medida pela distância, por técnicas como cosseno ou coeficiente *Jaccard*[10]. O método também classifica a fraude em um exame como parcial ou total.

A aplicação do método envolveu a seleção dos dados e pré-processamento: um conjunto de 30 provas de Língua Portuguesa foram selecionadas, cada uma com quatro

[6] https://www.ephorus.com/pt/home
[7] https://www.plagium.com/
[8] https://warwick.ac.uk/fac/sci/dcs/research/ias/software/sherlock/
[9] https://www.urkund.com/pt-br/
[10] O índice de Jaccard, também conhecido como coeficiente de similaridade de *Jaccard*, é uma estatística usada para medir a similaridade e diversidade de conjuntos de amostras

questões abertas. Os textos submetidos pelos estudantes foram armazenados em formato digital. A acentuação e pontuação, além de palavras com menos de 3 letras, foram retiradas dos textos avaliados. Os textos também passaram por uma regularização morfológica, onde palavras com o mesmo radical (verbos com diferentes conjugações, plural, aumentativo, diminutivo, gênero, se tornavam equivalentes). Para cada questão aberta foram definidas palavras que poderiam ocorrer na solução apresentada. Quando questões comparadas de dois alunos apresentaram um alto número de palavras idênticas, um indício de fraude ("cola") foi evidenciado. As questões de cada aluno foram transformadas em vetores de palavras, como resumido na Tabela 2.3.

Tabela 2.3 – Documento representado como vetores. Adapt. de Cavalcanti *et al*. (2012)

	t_1	t_2	...	t_j	...	t_M	***Classe***
d_1	a_{11}	a_{12}	...	a_{1j}	...	a_{1M}	y_1
d_2	a_{21}	a_{22}	...	A_{2J}	...	a_{2M}	y_1
⋮	⋮	⋮	⋱	⋮	⋱	⋮	y_3
d_i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{iM}	y_2
⋮	⋮	⋮	⋱	⋮	⋱	⋮	y_2
d_N	a_{N1}	a_{N2}	...	a_{Nj}	...	a_{NM}	y_3

Dois indicadores foram extraídos de cada par de documentos avaliados: similaridade de cosseno e coeficiente *overlap*. Os valores apurados foram avaliados por um especialista do domínio, que indicava se o nível de fraude era alto, intermediário, baixo ou nenhum. O resultado, além da informação numérica, era também apresentado de forma gráfica, com a relação de suspeição entre alunos (Figura 2.4).

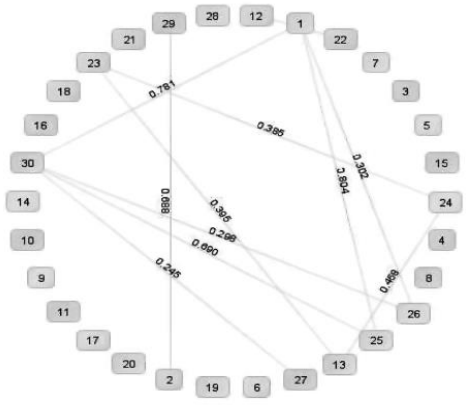

Figura 2.4 – Grafos de similaridade de uma questão, entre alunos (Cavalcanti *et al*., 2012)

No trabalho não fica claro como se obtém os textos, mas presume-se que o método seja viável somente com provas ou trabalhos realizados em ambiente digital.

2.3.2 Detecting excessive similarity in answers on multiple choice exams (Wesolowsky, 2000)

Wesolowsky (2000) apresenta um método para detecção de fraude (“cola”) em provas. Segundo o autor, os exames de múltipla escolha são muito presentes na vida acadêmica, assim como as suspeitas de fraude, em especial pela cópia de respostas entre alunos. Contra esse tipo de fraude, na maioria das vezes usam-se provas com múltiplas versões, normalmente com a alteração da ordem de questões e alternativas. Mesmo os que têm conhecimento das ferramentas de detecção de fraude, temem usá-los, por desgastes legais, para evitar confronto ou pelo receio de falsos positivos. No método proposto busca-se a simplicidade, procurando avaliar as respostas coincidentes, manter o método compreensível e aplicável, além de evitar falsos positivos.

A maioria dos métodos similares, busca a probabilidade de uma resposta correta ou incorreta, a partir de respostas reais de uma classe. A maioria não incorpora a capacidade do aluno no modelo, sendo que para isso seria útil a estratificação dos alunos por desempenho, assumindo que alunos em cada estrato têm capacidade igual. O método proposto é uma modificação do método apresentado em Frary *et al*. (1977), com a adição da possibilidade de suspeição partir do fiscal de prova. Utiliza-se a desigualdade de Bonferroni e uma simulação é conduzida para demonstrar pontos de corte de suspeição com margem de segurança para se evitar erros do tipo falso positivo.

No trabalho se supõe que: i) os alunos que têm uma pontuação geral mais alta são mais propensos a responder a uma pergunta corretamente do que aqueles que têm uma pontuação geral inferior; ii) a probabilidade de um aluno ter uma resposta certa depende da dificuldade da pergunta; isto é, a dificuldade se reflete na proporção da turma que responde incorretamente à questão.

O método considera uma série de informações, calculadas por técnicas estatísticas: n representa o número de estudantes em uma classe; q representa o número de questões do exame; m_{jki} representa a probabilidade de coincidência de resposta entre os estudantes j e k na questão i; M_{jk} representa o número de coincidências de resposta

observadas entre os estudantes *j* e *k*; p_{ji} representa a probabilidade do estudante *j* acertar a questão *i*; r_i representa a proporção da classe que respondeu corretamente a questão *i*; c_j representa a proporção de questões respondidas corretamente pelo estudante *j*; v_i representa o número de respostas erradas na questão *i*; w_{ti} representa a probabilidade de que, dada uma resposta incorreta, a escolha errada *t* é escolhida na questão *i*. Sobre estas informações é calculada, utilizando q-Bernoulli, a média de semelhanças esperadas e o valor computado, fazendo a checagem de suspeição. Como dito, se há uma suspeição prévia, oriunda do fiscal de prova, um acréscimo de suspeição é adicionado à comparação.

Embora trate do assunto, o método não faz uso do plano de assentos, para verificar a proximidade de dois alunos sendo comparados e, assim, incrementar o índice de suspeição. Analisando o trabalho, a adoção de suspeição prévia pode trazer um viés indesejado, uma vez que o comportamento do aluno, avaliado pelo fiscal de setor, é, muitas vezes, subjetivo.

2.3.3 Detect multiple choice exam cheating pattern by applying multivariate statistics (M. Chen, 2017)

M. Chen (2017) aplica uma série de algoritmos de *data mining* (e.g., correlação multivariada) e um mapa de calor para identificar padrões de suspeição. Aliado a isso, utiliza-se agrupamento hierárquico e árvores de dendrogramas (Figura 2.5) para verificar possíveis relacionamentos entre alunos. No artigo, foram consideradas duas premissas: i) a probabilidade de escolher as mesmas respostas erradas nas perguntas difíceis é ainda mais improvável, apenas por acaso, em comparação com a escolha das respostas certas para as perguntas fáceis; ii) é estatisticamente ainda mais improvável que os alunos selecionem, sem querer, as mesmas respostas erradas em questões difíceis e, com isso, forneçam evidências de trapaça.

No estudo, foram coletadas as respostas de 75 alunos, aos quais foram submetidas 3 diferentes versões de prova (diferenciação somente na ordem, não no conteúdo apresentado). Três alunos estavam alocados dentre 25 mesas disponíveis. Para aumentar a eficiência da detecção, a avaliação se concentrou nas questões mais difíceis (primeiro quintil). Além disso, descartou as 25 provas de múltipla escolha com notas mais

baixas, alegando baixo desempenho, para aumentar a eficiência na detecção. Um mapa de calor também foi utilizado para identificar padrões.

As seguintes análises foram realizadas: i) Análise multivariada de correlação entre notas de alunos; ii) Análise da localização do aluno x nota obtida; iii) Análise de Clusterização Hierárquica (HCA): com o uso da abordagem aglomerativa, o objetivo desta análise foi pesquisar o grau de semelhança entre as respostas dos exames e pesquisar padrões de semelhança entre os alunos; iv) Análise por mapa de calor; e v) Análise do Componente Principal (PCA), para avaliar o resultado de análises anteriores (Figura 2.6).

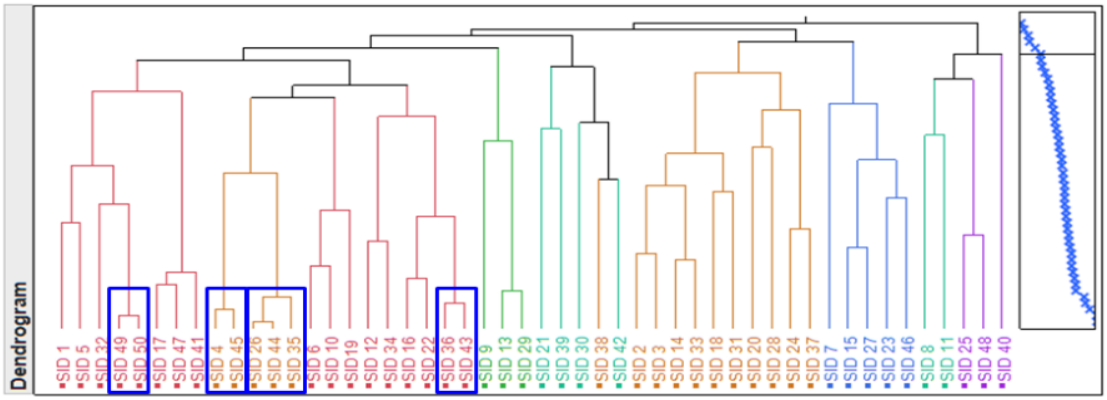

Figura 2.5 – Árvore Hierárquica / Dendrograma. (M. Chen, 2017)

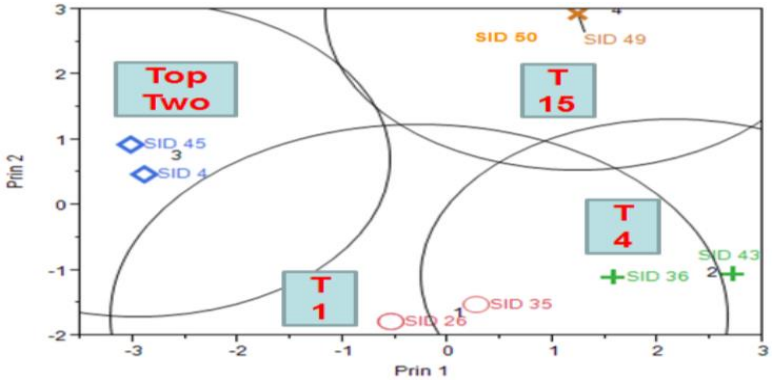

Figura 2.6 – Resultado apresentado pela análise PCA. (M. Chen, 2017)

Em uma análise do método, consideramos que o descarte das 25 provas de notas mais baixas pode mascarar aqueles alunos que fizeram cópia parcial de questões, o que, em uma análise empírica, parece ser o mais comum em provas. Além desse fator, a indicação precisa de local de prova, que permita majorar alunos próximos é mais exequível em um contexto local, de prova realizada em um só local.

2.3.4 Análise comparativa dos trabalhos relacionados

A Tabela 2.4 apresenta uma Análise comparativa dos trabalhos relacionados. Além dos pontos descritos nas subseções anteriores e na tabela, destaca-se que o MDFCP se adapta às particularidades dos exames (por exemplo, notas de dissertações e exames anteriores), além de apresentar um indicador de suspeita abrangente que considera diferentes elementos de suspeição.

Tabela 2.4 – Análise comparativa dos trabalhos relacionados.

Artigo / **Características**	**(Cavalcanti *et al.*, 2012)**	**(Wesolowsky, 2000)**	**(M. Chen, 2017)**	**MDFCP**
Contexto	*Plágio em provas locais*	*Coincidências em uma classe de alunos*	*Coincidências em questões de múltipla escolha*	*Levantamento de suspeição em concursos públicos*
Técnica	*Data Mining*	*Estatística*	*Data Mining Estatística*	*Data Mining Estatística*
Detalhamento da técnica	*Text Mining (TF-IDF)*	*Desigualdade de Bonferroni; q-Bernoulli*	*Clusterização Hierárquica / Árvore de Dendrogramas*	*Regressão Logística; Detecção de Outlier; Modelo Estatístico*
Tipo de prova	*Questões abertas*	*Múltipla Escolha*	*Múltipla Escolha*	*Múltipla Escolha*
Parametrizável	*Não*	*Não*	*Não*	*Limites de suspeição e similaridade*
Estrutura de dados	*Não estruturada*	*Fixa*	*Fixa*	*Ajustável*
Agentes externos	*ED participa do método*	*Fiscais de prova subsidiam suspeitas*	---	*ED parametriza e valida*
Retroalimenta-ção de dados	*Não*	*Não*	*Não*	*Sim*
Apresentação	*Numérica Visual*	*Numérica*	*Mapa de calor Numérica*	*Numérica*
Tipo de informação	*Índice*	*Índice*	*Índice*	*Índice*
Problemas ou limitações observados	*Converter informação para o meio digital*	*Viés na indicação de suspeição por fiscais de setor*	*Descarte parcial de alunos com pior desempenho Informação do assento em prova*	*Detalhes no capítulo de discussão*

A primeira diferença entre o MDFCP e os três trabalhos destacados como relacionados (Tabela 2.4) está no foco de cada um, seja em plágio em questões dissertativas ou coincidência em questões de múltipla escolha.

Os trabalhos relacionados exigem i) conversão de informação escrita para digital (Cavalcanti *et al*., 2012), ii) apresentam viés na indicação de suspeição por fiscais de setor de prova (Wesolowsky, 2000), e iii) descartam parcialmente alunos com pior desempenho e exigem informações do assento em sala (M. Chen, 2017). Embora exija tratamento prévio dos dados e estudo pelo especialista de domínio, MDFCP não apresenta tais características.

No MDFCP a abordagem de identificação de suspeitos é mais abrangente, indo além de similaridades em questões. De forma geral, o MDFCP se mostrou mais flexível e parametrizável do que os métodos relacionados.

3 REFERENCIAL TEÓRICO

Este capítulo tem por finalidade descrever as abordagens teóricas e metodológicas que fundamentam o método proposto, englobando conceitos de estatística, mineração de dados (regressão logística e detecção de *outliers* em dados multidimensionais) e preparação dos dados. A Subseção 3.1 apresenta o método de detecção de *outlier* adotado; a Subseção 3.2 introduz a Regressão Logística como método de predição de perfil de candidato aprovado; a Subseção 3.3 trata da origem e dos tipos de dados esperados; a Subseção 3.4 trata da preparação dos dados.

3.1 Detecção de *Outlier*

Um *outlier* pode ser definido como uma observação que aparenta ser inconsistente com o restante de um conjunto de dados. Valores *outliers* podem ser considerados como ruídos (valores incorretos, transformação de dados incoerente etc.), que devem ser eliminados para não afetar processos posteriores de agrupamento, por exemplo (Bouguessa, 2012). No contexto deste trabalho, *outlier*s são evidenciados e não suprimidos, sendo usados para indicar um certo grau de suspeição expresso em um índice parcial.

Algoritmos de detecção de *outliers* são comumente aplicados na detecção de fraudes financeiras, surtos de pandemia na área de saúde pública, entre outros domínios (Belo *et al*., 2016). A maneira pela qual uma observação é julgada inconsistente ou suspeita não é definida de maneira geral, ao contrário, depende do cenário de aplicação e do método de detecção a ser utilizado e, sob esta ótica, nota-se então que esta área envolve um tipo de subjetividade que também está presente, por exemplo, em agrupamento não supervisionado de dados (Campos, 2015).

Na identificação do método mais adequado deve-se considerar os seguintes aspectos: i) a natureza dos dados de entrada: binários, categóricos ou contínuos (Singh & Upadhyaya, 2012); ii) o tipo de instância de dados: univariáveis, quando são compostas por somente um atributo, ou multivariáveis, quando são compostas por vários atributos do mesmo tipo ou de diferentes tipos (Tan *et al*., 2013); iii) a disponibilidade de rótulos nos dados de classificação de instâncias de dados; iv) a classificação de saída: rótulo ou pontuação (*score*), sendo a última mais flexível para se definir, pelo analista, de um limiar

para o que se classifica como *outlier* (Prasad *et al*., 2009; Varun Chandola *et al*., 2009); e v) o tipo de *outlier*: 1) Global, onde um objeto se desvia consideravelmente dos demais (exemplo na Figura 3.1); 2) Contextual, onde objetos são *outliers* em comparação ao contexto em que está inserido (exemplo na Figura 3.2); e 3) Coletivos, sendo um subconjunto de dados considerado anormal quando formam um grupo, exemplificado na Figura 3.3 (Freitas, 2019).

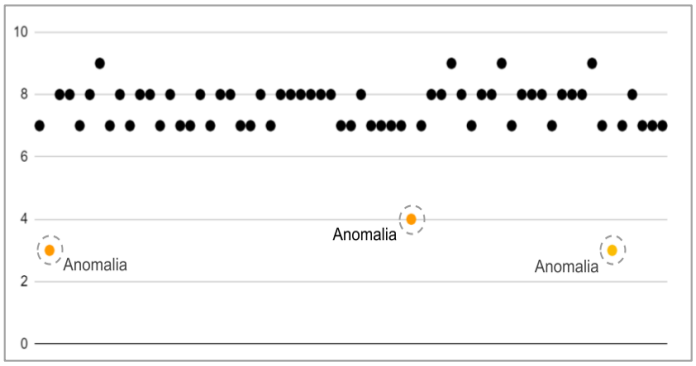

Figura 3.1 – Exemplo de Outlier Global

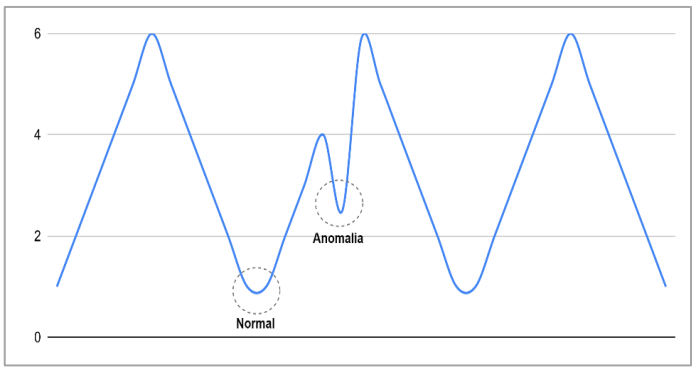

Figura 3.2 – Exemplo de Outlier Contextual – temperatura no tempo. Adaptado de (Singh & Upadhyaya, 2012)

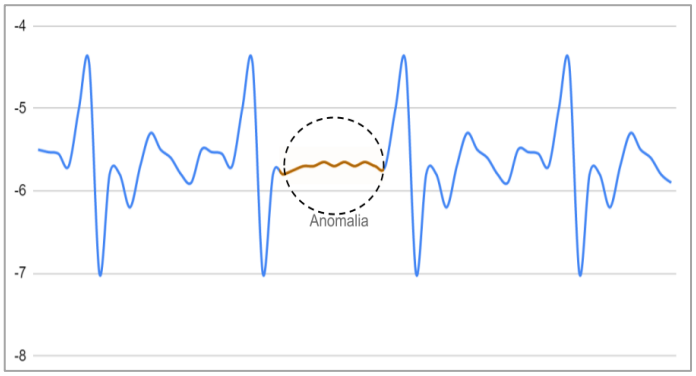

Figura 3.3 – Exemplo de Outlier Coletivo em um ECG. Adaptado de Singh & Upadhyaya (2012)

No caso dos dados submetidos ao algoritmo de detecção de *outliers*, notas de um candidato são dados contínuos e multivariáveis; o dado *outlier*, caso exista, é coletivo. Não há rótulos que distingam grupos, pois todos são candidatos aproveitados (convocados) e a classificação de saída é dependente do método utilizado.

A maior parte das técnicas de detecção de *outliers* foram desenvolvidas para domínios de aplicação específicos, enquanto outras são mais gerais, mas não tão genéricas de forma a ser eficiente para qualquer tipo de aplicação (Niu *et al*., 2011). A detecção de *outlier* pode ser realizada por técnicas supervisionadas ou não-supervisionadas. Um *outlier*, considerando a detecção de *outliers* não supervisionada, pode ser descrito como "uma observação (ou subconjunto de observações) as quais podem parecer inconsistentes com o restante do conjunto de dados (Campos, 2015). Este tipo de técnica é usado para identificar fraudes desconhecidas (Tan *et al*., 2013).

A detecção de *outliers* não-supervisionada é a de utilização mais comum e engloba algoritmos como *K-Nearest Neighbors* (KNN), *K-Nearest Neighbors-Weighted* (KNNW), *Local Outlier Factor* (LOF), dentre vários outros (Campos, 2015).

A implementação e os testes dos algoritmos de detecção de *outlier* (esquema básico na Figura 3.4) se valem da biblioteca Scikit-learn (Pedregosa *et al*., 2011), uma API de aprendizado de máquina de código aberto em Python, implementada de forma gráfica e parametrizável através do software Orange Data Mining[11].

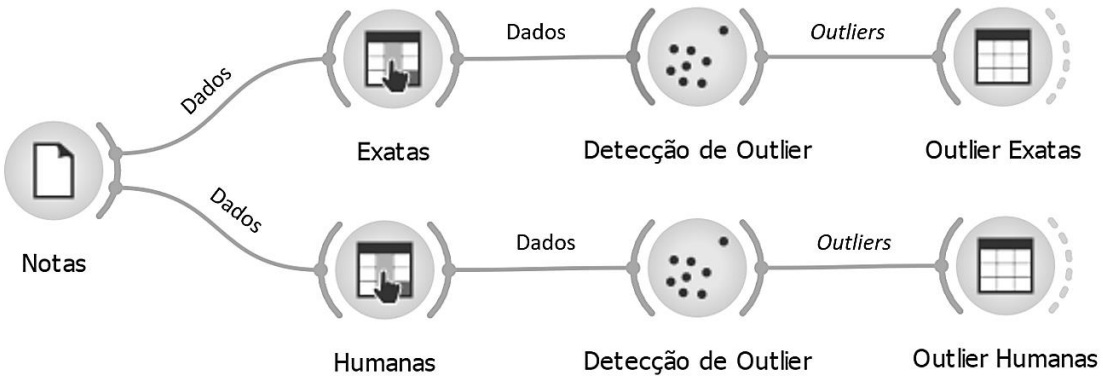

Figura 3.4 – Esquema de detecção de outlier no software Orange

[11] https://orangedatamining.com/

3.2 Regressão Logística (RL)

A regressão linear é usada para indicar uma relação entre uma variável de resposta e um conjunto de variáveis previsoras. No entanto, para algumas aplicações de dados, a variável de resposta é categórica em vez de contínua. Nestes casos, é aconselhável a utilização do método de Regressão Logística (RL) para descrever a relação entre uma variável de resposta categórica e um conjunto de variáveis previsoras (Larose & Larose, 2015). Recomenda-se RL para situações em que a variável dependente (categórica) é de natureza dicotômica ou binária. As demais variáveis, independentes, podem ser categóricas ou não.

Na Figura 3.6 é mostrado um exemplo de esquema geral do funcionamento da RL. Uma tabela com dados de treinamento é preparada, composta por uma série de dados (variáveis previsoras), acrescida de uma variável de resposta categórica. Com isso, é possível "treinar" um modelo preditivo, que será capaz de estimar, submetido um conjunto de novas variáveis (tabela para validação), a probabilidade de que estes dados pertençam a um dentre dois estados possíveis. Um resultado, relativo à probabilidade de sucesso de uma situação indicada pela variável de resposta categórica, é gerado para cada conjunto de dados submetida, e fica contido no intervalo [0..1]. Para o cálculo probabilístico utilizado em regressão logística, é utilizada a seguinte fórmula:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}$$, onde $g(x) = B_0 + B_1X_1 + ... + B_pX_p$

Os coeficientes B0 , B1, ..., Bp são estimados a partir do conjunto de dados, pelo método da máxima verossimilhança. A função descrita é representada graficamente como na Figura 3.5.

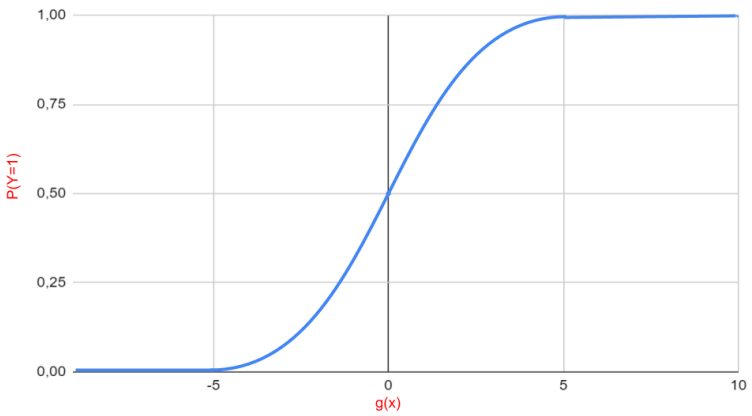

Figura 3.5 – Gráfico da função Sigmoide ou Logística

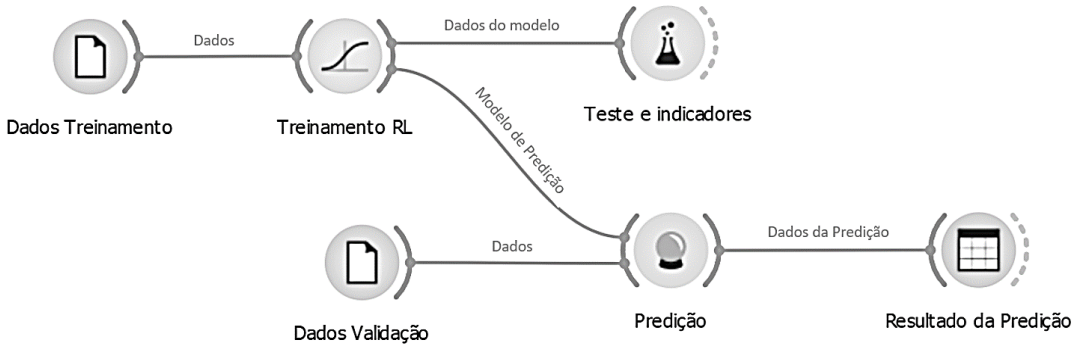

Figura 3.6 – Exemplo de utilização de algoritmo de RL no software Orange

RL pode ser aplicada, por exemplo, para calcular: i) a probabilidade de um tomador de empréstimo pagar regularmente ou não as parcelas devidas; ii) a probabilidade de que um aluno, com base em desempenho anterior, seja aprovado ou reprovado no curso atual; iii) a probabilidade de que um candidato seja aprovado em um concurso, com base em seus dados biográficos, geográficos e históricos, caso específico que será tratado na descrição de MDFCP.

Existem medidas de qualidade da classificação de um modelo de RL, a saber:

i) Acurácia: percentual de casos que são corretamente classificados.

$$Acurácia = \frac{positivos\ verdadeiros + negativos\ verdadeiros}{n}$$

ii) Precisão: percentual de observações em uma classe (0 ou 1, sucesso ou insucesso), que foram corretamente classificadas.

$$Precisão_{sucesso} = \frac{positivos\ verdadeiros}{positivos\ verdadeiros + falsos\ negativos}$$

$$Precisão_{insucesso} = \frac{negativos\ verdadeiros}{negativos\ verdadeiros + falsos\ positivos}$$

iii) Recall: para uma classe, corresponde ao percentual de previsões da classe que foram corretamente classificadas.

$$Recall_{sucesso} = \frac{positivos\ verdadeiros}{positivos\ verdadeiros + falsos\ positivos}$$

$$Recall_{insucesso} = \frac{negativos\ verdadeiros}{negativos\ verdadeiros + falsos\ negativos}$$

iv) Score F-1: para uma determinada classe (0 ou 1, sucesso ou insucesso), corresponde a uma combinação entre precisão e recall

$$ScoreF1_{sucesso} = \frac{2\ x\ Precisão_{sucesso}\ x\ \ Recall_{sucesso}}{Precisão_{sucesso}\ x\ \ Recall_{sucesso}}$$

$$ScoreF1_{insucesso} = \frac{2\ x\ Precisão_{insucesso}\ x\ \ Recall_{insucesso}}{Precisão_{insucesso}\ x\ \ Recall_{insucesso}}$$

Esses valores podem ser mais precisos dependendo das variáveis previsoras disponíveis e da qualidade da informação atribuída a cada variável. Testes possibilitam avaliar quais variáveis previsoras tornarão a RL mais precisa. A quantidade e a qualidade dos dados contidos nas variáveis submetidas ao modelo impactarão na precisão de suas predições. Um estudo, com base nos índices elencados, pode determinar as variáveis que poderão representar um maior ganho de precisão.

Técnicas de validação cruzada podem ser usadas para garantir que resultados descobertos em uma análise são generalizáveis para outros conjuntos de dados. *K-fold* é uma técnica de validação cruzada, no qual um conjunto de dados é separado em k partições independentes e de tamanho similar. Os resultados obtidos nas k simulações podem ser avaliados, por exemplo, utilizando de média. A Figura 3.7 apresenta um exemplo de partições de um conjunto de dados pela técnica de *K-fold.*

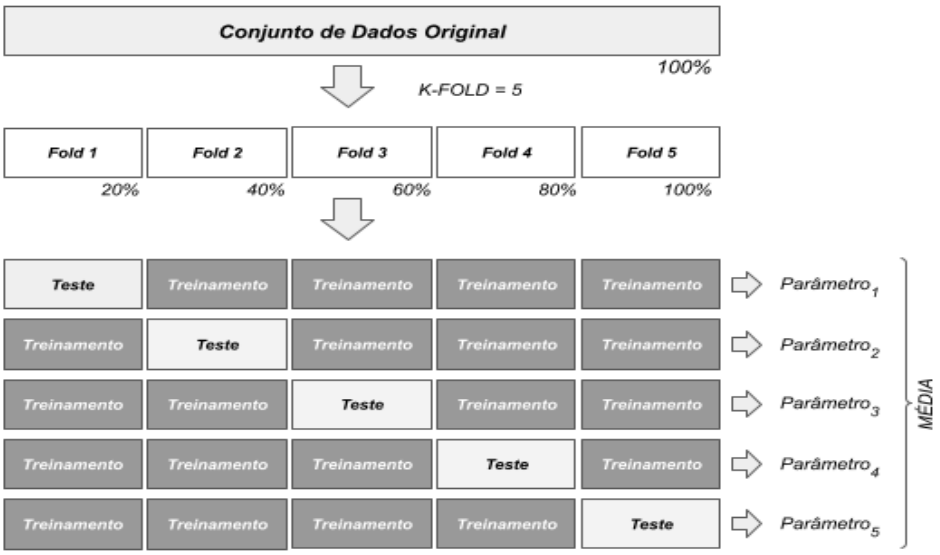

Figura 3.7 – Exemplo de partições de um conjunto de dados pela técnica de K-fold

Neste trabalho, os testes e a execução do método de RL são realizados com o uso do *software Orange DataMining*[12], que implementa a biblioteca *Scikit-learn*.

3.3 Origem e tipo dos dados

Os dados disponíveis divergem entre diferentes concursos, oferecidos em diferentes periodicidades, geridos por diferentes instituições e empresas. Estes dados são, em sua maioria, obtidos de forma eletrônica, durante o processo de inscrição. A depender das particularidades de cada concurso, serão complementados com outros dados, obtidos em fases posteriores. Os dados mais comuns, observados nos principais concursos, estão discriminados na Tabela 3.1.

Tabela 3.1 – Principais informações coletadas pelas gestoras de concursos.

Tipo	Rótulo
Biográfico	idade, sexo, origem racial, nome da mãe, origem escolar, nível escolar, profissão dos pais, foto, preparação para o concurso, investigação social
Geográfico	local de nascimento, residência e realização da prova
Social	renda familiar, dimensão familiar, indicadores de classe social
Desempenho	gabaritos, respostas de questões de múltipla escolha, notas em provas dissertativas e de múltipla escolha, índice de acerto global por questão, tempo de resolução de prova, desempenho em provas físicas, desempenho em provas específicas de aptidão
Contato	*e-mail*, telefone, redes sociais
Documento	número de documento de identidade, CPF
Médico	exames e laudos médicos, testes psicológicos
Histórico	número de vezes que realizou o concurso, anormalidades registradas, desempenho anterior

Durante a inscrição, no formulário eletrônico, é impossível se testar a veracidade dos dados submetidos, sendo comumente realizada uma consistência básica de dados informados, como limites de idade, validação de CPF, dentre outras. A checagem dos dados, normalmente, é realizada em fases posteriores do concurso, como entrega de documentação e/ou comprovação biográfica.

[12] https://orangedatamining.com/

3.4 Preparação dos dados

Apesar de existirem dados básicos, em que a maioria dos concursos se assemelha, há uma série de dados específicos, em função de diferenças de etapas, tipos de prova aplicadas, exigências em particular e detalhamento do perfil do candidato. O MDFCP tem como premissa avaliar concursos que tenham, dentre outras provas e etapas, uma prova de múltipla escolha, por se tratar de análise central dentre as várias avaliadas pelo método, além de dados básicos relativos ao perfil dos candidatos.

A preparação dos dados é uma tarefa importante prévia à execução do método proposto. Esse pré-processamento visa adequar os dados à modelagem e aos requisitos de dados previstos no MDFCP. Por isso, esta atividade é específica em função dos dados disponíveis e os necessários à execução do modelo, em que pese as várias possibilidades de ajustes previstos no método.

De maneira geral, como apontado em Brownlee (2020), as tarefas normalmente atinentes à preparação de dados são i) Limpeza dos dados: identificar e corrigir erros nos dados; ii) Seleção de características: identificar as variáveis que serão relevantes para a tarefa; iii) Transformação dos dados: mudar a escala ou a distribuição de variáveis; iv) Engenharia de características: derivar novas variáveis dos dados disponíveis; v) Redução da dimensionalidade: criar projeções compactas dos dados.

As especificidades de preparação dos dados, voltadas para o MDFCP, serão expostas na seção de aplicação do método.

4 MÉTODO PARA DETECÇÃO DE FRAUDE EM CONCURSOS PÚBLICOS (MDFCP)

Neste capítulo, apresenta-se o Método para Detecção de Fraude em Concursos Públicos (MDFCP). O método visa atribuir um índice de suspeita a cada candidato aproveitado (CAp). Candidatos aproveitados são aqueles, dentre o universo de candidatos de um determinado concurso, convocados para a realização de fases posteriores do concurso ou para tomar posse do cargo a que concorre.

Conforme apresentado na Figura 4.1, o método se vale de quatro análises, que avaliam diferentes aspectos de suspeição, gerando índices parciais. As análises se valem de técnicas estatísticas, aprendizado de máquina e um modelo próprio. Esses índices parciais, aplicados pesos determinados, compõem um índice final de suspeição. Ao final do processo, os candidatos com índice de suspeita acima de um limite pré-determinado são incluídos em uma lista de suspeição.

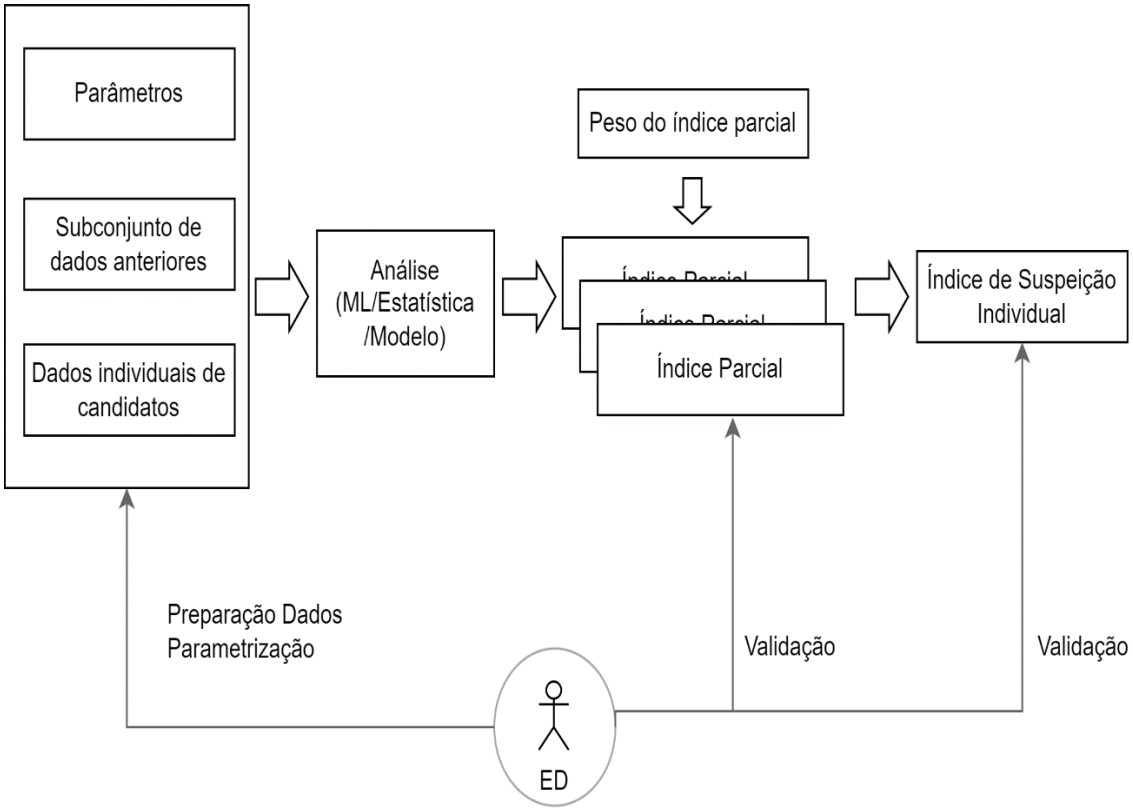

Figura 4.1 – Processo Geral de MDFCP para Geração do Índice de Suspeição (IS)

Como entrada do método são utilizados dados do concurso avaliado, tais como os do candidato (perfil, desempenho, histórico), o conjunto de respostas corretas (gabarito) e um valor geral de acerto das questões. O MDFCP também usa como entradas uma série de parâmetros, como um limite de suspeita, pesos específicos de cada índice parcial, um limite de similaridade de respostas entre dois candidatos, além de bancos de dados de exames anteriores e registros de suspeição diversos.

Atendidas as premissas básicas – possuir questões de múltipla escolha e dados sobre o candidato que permitam definir um perfil –, o método se mostra bastante flexível, valendo-se de ajustes dos dados recebidos e a utilização de diversos parâmetros. O Especialista de Domínio (ED) exerce papel fundamental no método, pois atua na preparação dos dados e parametrização, além de validar os índices gerados pelo método.

A Figura 4.2 apresenta a arquitetura conceitual do método, com os respectivos parâmetros e o fluxo de dados pelos componentes. As subseções deste capítulo detalham os componentes da arquitetura conceitual: análise de perfil, análise de notas, análise de respostas em questões de múltipla escolha, análise de registros anteriores.

O MDFCP contempla um mecanismo de *feedback*, se alimentando de dados de processamentos anteriores com o objetivo de adicionar uma maior precisão em usos posteriores do método. Esse *feedback* pode ser o registro de suspeição anterior detectada (item “r3” da Figura 4.2) ou os registros de similaridades entre candidatos (item “r2” da Figura 4.2).

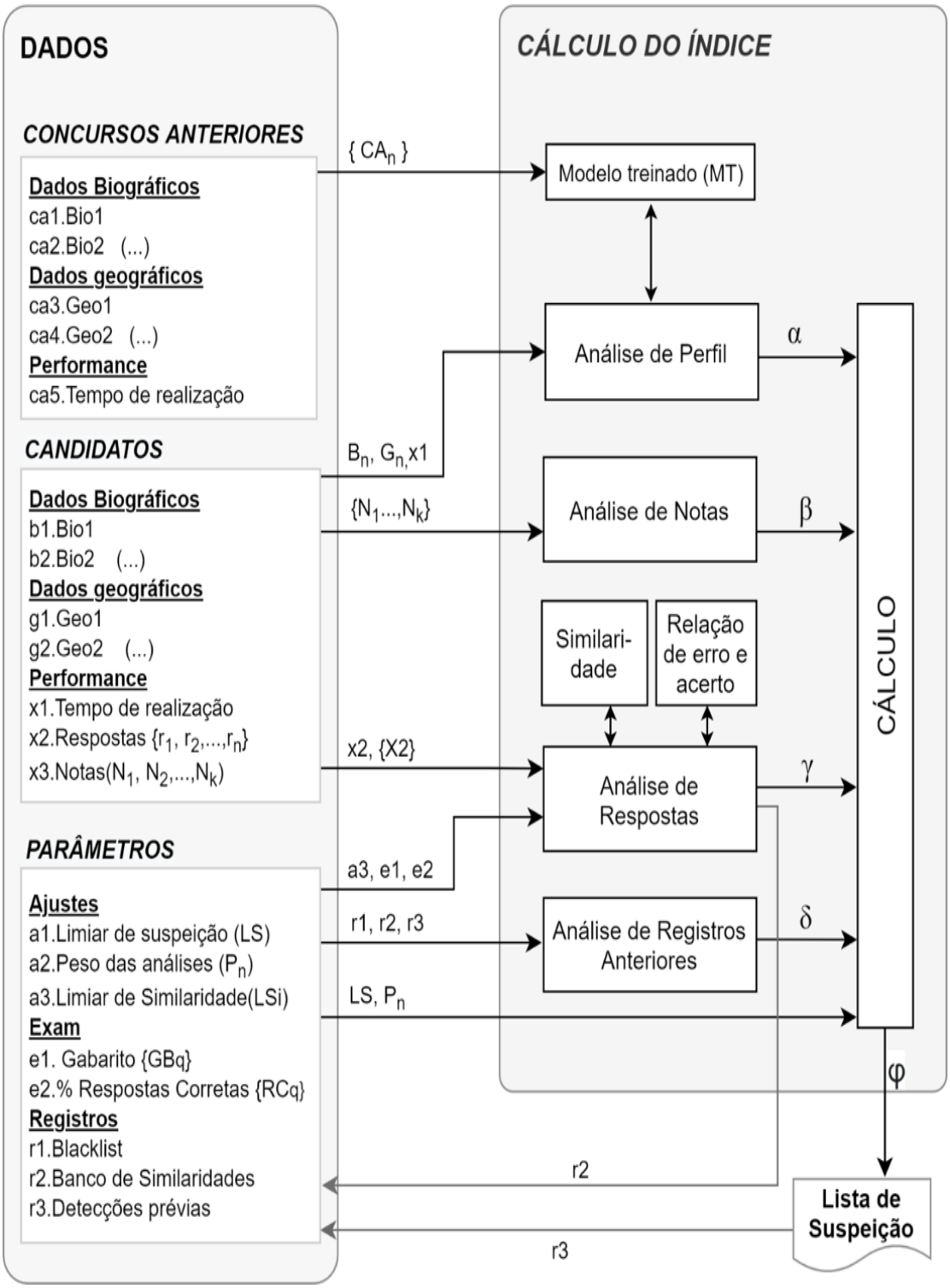

Figura 4.2 - Arquitetura Conceitual do MDFCP

4.1 Análise de Perfil

A primeira das análises realizadas no método é a da adequação do perfil do candidato. Esta ocorre em duas etapas. Na primeira, utiliza-se uma amostra de dados

biográficos, geográficos e de tempo de realização do teste, de candidatos aproveitados (CAp) e de não aproveitados em um concurso anterior (Tabela 4.1). Esses dados são usados para treinar um modelo de RL (Figura 4.3), utilizando o software Orange Data Mining.

Tabela 4.1 – Exemplo de conjunto de dados para treinamento do modelo de RL

Dados Biográficos
Dado Biográfico 1
Dado Biográfico 2
(...)
Dado Biográfico n
Dados Geográficos
Dado Geográfico 1
Dado Geográfico 2
(...)
Dado Geográfico n
Outros
Outros 1
(...)
Outros n
Rótulo
Situação (candidato aproveitado[13]/não aproveitado)

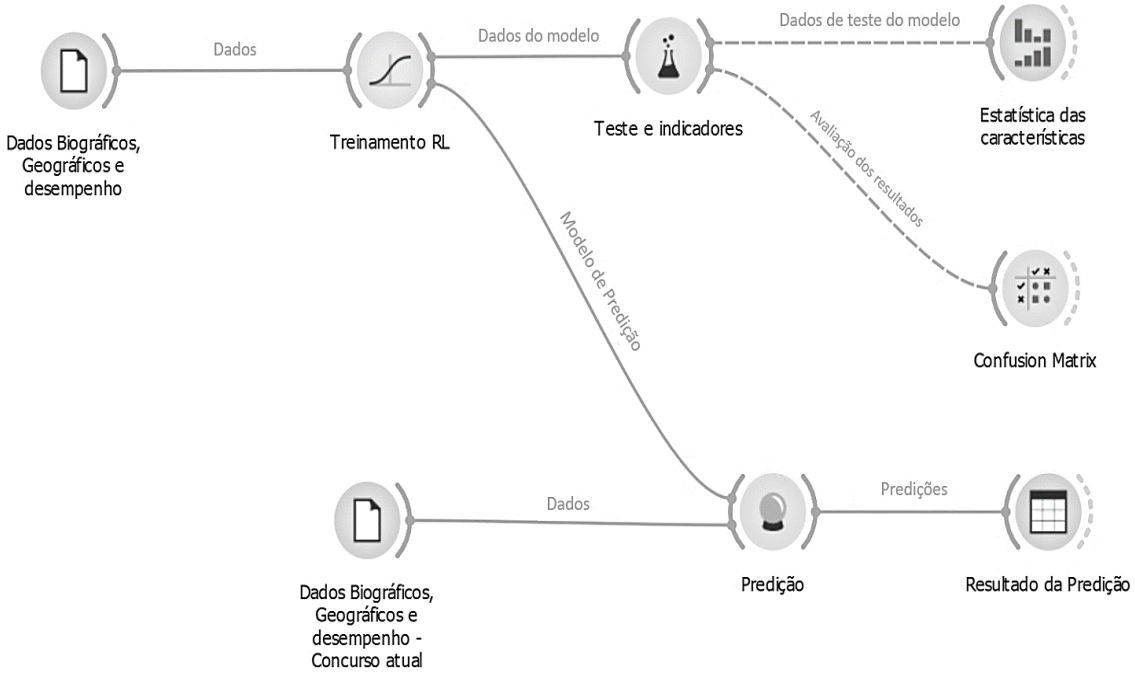

Figura 4.3 – Modelo de Regressão Logística

[13] Candidato aproveitado é uma nomenclatura adotada nesta dissertação para generalizar termos como candidato aprovado, candidato convocado, candidato chamado que aprecem em diferentes concursos.

Em um segundo momento, utiliza-se um conjunto de dados do candidato, do concurso sendo avaliado, com a mesma estrutura que foi utilizada para treinar o modelo, com a exceção do rótulo (candidato aproveitado/não aproveitado). Na verdade, nesse momento da execução da análise, todos pertencem à classe de candidatos aproveitados. Estes dados são submetidos ao modelo, que retorna um valor no intervalo [0..1], para cada um dos candidatos avaliados, referente à probabilidade de que um determinado candidato esteja no universo de aproveitados (doravante, Probabilidade de Candidato ser Aproveitado, PCA). O índice parcial α é o complemento desse valor, calculado como segue:

$$\alpha = 1 - PCA$$

Portanto, candidatos com baixa probabilidade de serem aproveitados, isto é, com um perfil divergente daqueles normalmente aproveitados, terão um índice parcial de suspeição α maior, como exemplificado na Tabela 4.2.

Tabela 4.2 – Exemplo de cálculo do índice parcial α

Identificação do Candidato	**Probabilidade RL (PCA)**	**α = 1 – PCA**
4079E66D9	0,221273773	0,778726227
10A753A15	0,44754627	0,55245373
2E753D20E	0,490501532	0,509498468
479EDF39C	0,851275999	0,148724001
9102230F	0,756058847	0,243941153

4.2 Análise de Notas

Este componente do método busca identificar valores discrepantes dentre as notas de disciplinas de candidatos aproveitados. As notas de todos os candidatos devem ser agrupadas por afinidade, como notas em disciplinas de humanas, notas em disciplinas de exatas etc., conforme exemplificado na Tabela 4.3. A avaliação agrupada, com uma maior dimensionalidade, busca avaliar, em um contexto mais amplo, disciplinas de alguma forma correlacionadas. A avaliação de *outlier* sobre cada uma das disciplinas, de forma individual, apresentaria menos contexto. Por exemplo, notas baixas, dentro de um grupo de disciplinas correlatas, avaliadas de forma individual, poderiam gerar 3 *outlier*s individuais que, em grupo, seriam coerentes e não discrepantes.

Tabela 4.3 – Notas agrupadas por afinidade

Grupo de notas 1
Grupo1. Nota 1
Grupo1. Nota 2
(...)
Grupo1. Nota k
(...)
Grupo de notas n
Grupo n. Nota 1
Grupo n. Nota 2
(...)
Grupo n. Nota z

Nessa análise, devem ser utilizadas, exclusivamente, as notas do concurso avaliado, sem considerar outros concursos. Notas de concursos anteriores referem-se a provas distintas, com graus de dificuldade distintos, portanto, a análise é feita de forma isolada ao evento sendo analisado.

Cada um destes grupos de notas relacionadas, de todos os candidatos avaliados, deve ser submetido a um algoritmo de detecção de *outlier*s, que detecta os valores discrepantes, conforme esquema apresentado na Figura 4.4. A detecção de *outlier*s se mostrou semelhante, mas o algoritmo *Local Outlier Factor* (Breunig *et al*., 2000), com a função de distância cosseno, foi selecionado dentre os disponíveis, através de experimentações, por apresentar valores mais coerentes com o que se espera por *outlier* nesse contexto, em uma avaliação do ED. A Tabela 4.4 apresenta uma comparação dos métodos de detecção de *outlier* testados.

Tabela 4.4 – Comparação dos métodos de detecção de outlier testados

Grupo de Disciplinas	SVM	Isolation Forest	LOF
Grupo 1	19 (2,6%)	22 (3,0%)	26 (3,5%)
Grupo 2	24 (3,3%)	20 (2,7%)	22 (2,7%)
Grupo 3	24 (3,3%)	22 (3,0%)	18 (2,5%)

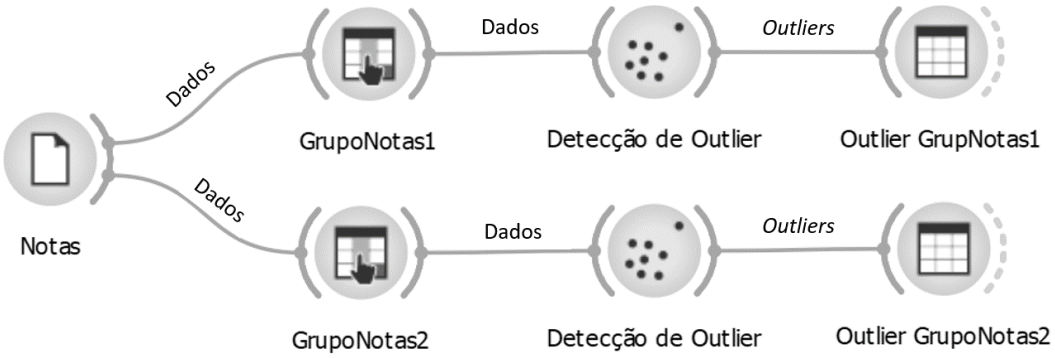

Figura 4.4 – Esquema de detecção de outlier em conjuntos distintos de notas

Para cada um dos candidatos, é calculado índice β:

$$\textbf{Se } Q_o = 0 \textbf{ então } \beta = 0\textbf{; senão } \beta = \frac{2^{QNO}}{2^{MaxNO}}$$

Onde:

QNO representa a quantidade de grupos de notas com valores *outlier*, de um candidato, considerando todos os agrupamentos de disciplinas avaliados.

MaxNO representa a quantidade máxima de agrupamentos detectados como *outlier* para um candidato, apurado de forma global. O intuito é dar maior significância ao valor aferido para cada um dos candidatos. Assim, quanto mais próximo dessa quantidade máxima, maior será o valor de *β*, no intervalo [0..1].

Supondo um concurso com 4 diferentes grupos de disciplinas (ou áreas específicas de conhecimento) de forma global e o máximo que um determinado candidato tenha figurado como *outlier* tenha sido em 3 destes grupos (MaxNO=3), a Tabela 4.5 demonstra os valores do índice parcial *β* conforme a quantidade de grupos de notas *outlier* de um candidato.

Tabela 4.5 – Exemplo de valores de β para MaxNO =3

QNO	β
0	0
1	0,25
2	0,50
3	1

4.3 Análise de Respostas de Questões de Múltipla Escolha (RQME)

A análise de respostas em questões de múltipla escolha tem como objetivo identificar suspeição no conjunto de respostas em provas de múltipla escolha, relativas aos candidatos aproveitados em um determinado concurso.

Como será detalhado, há a combinação de duas análises, que dá uma cobertura mais ampla a diferentes tipos de fraude: uma que compara um candidato a todos os outros do universo avaliado (1:n); e outra que avalia isoladamente as respostas de um candidato (1:0).

As duas análises visam ser as mais abrangentes, sendo eficientes nos casos em que fraudadores evitem notas muito elevadas, introduzindo erros propositais, introduzidos de forma idêntica ou aleatório.

Além das respostas dos candidatos e do gabarito com as respostas corretas, a análise também tem como entrada o Índice de Dificuldade da Questão (IDQ). Cada uma das questões de uma prova de múltipla escolha, de cada concurso, tem um nível de dificuldade associado. Este índice, individualmente associado à cada questão, é inversamente correlacionado ao índice de acertos global desta questão. O valor é calculado anteriormente à análise das RQME, com base no gabarito informado e no rol de respostas dos candidatos. O índice adiciona um componente à análise, que é a menor probabilidade de coincidência de respostas entre dois candidatos quando uma determinada questão apresenta um maior valor associado. Esse índice é calculado da seguinte forma:

$$\mathbf{IDQq = 1 - PAG_q}$$

Sendo, PAG_q o percentual de acerto geral, considerando todas as respostas em uma determinada questão *q*, em formato decimal, no intervalo [0..1]. O IDQ é amplamente utilizado na geração do índice parcial de suspeição, nas duas abordagens, que serão demonstradas a seguir.

Ao final, essa análise produz um índice de suspeição parcial γ, obtido pela análise e seleção de um dos índices gerados pelas duas diferentes abordagens. O maior desses índices é considerado como o índice parcial de suspeição.

4.3.1 Abordagem 1 (*γ1)*: Grau de similaridade entre RQME de dois candidatos

Nesta abordagem, analisa-se o maior grau de similaridade entre respostas de dois candidatos ($Cand_A$, $Cand_B$). Para calcular esse valor, confrontam-se as respostas de cada candidato avaliado com os demais candidatos aproveitados, o que se denominou avaliação 1:n. Para cada uma das avaliações 1:n, os seguintes passos são necessários:

Passo 1 – Obter o maior valor de coincidências possível (MaxCC) entre dois candidatos sendo comparados; valor dado pela menor quantidade de acertos dentre ambos (Figura 4.5).

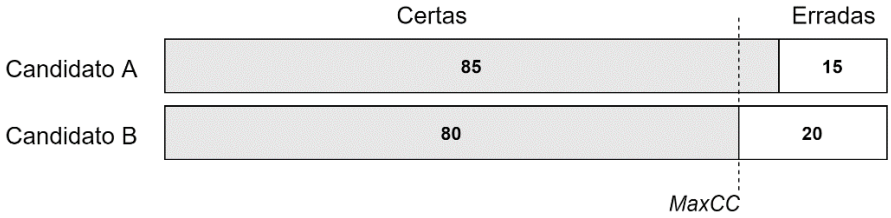

Figura 4.5 – Obtenção do máximo de questões coincidentes entre dois candidatos

Passo 2 – Obter a média dos Índices de Dificuldade de Questão (IDQ), das respostas corretas do Candidato A (MedIDQCandA).

Passo 3 – Obter a expectativa do valor máximo da soma de IDQ de respostas coincidentes e corretas (MaxIDQCandAB), entre os candidatos A e B, dado pela multiplicação do maior valor de coincidências de respostas corretas (MaxCC) pela média dos IDQ das respostas corretas do candidato A, conforme segue:

$$\mathrm{MaxIDQCandAB} = \mathrm{MaxCC} * \mathrm{MedIDQCandA}.$$

Passo 4 – Obter o redutor (Red) relativo à discrepância de desempenho entre os candidatos A e B. Para esse cálculo, são necessários: i) o valor máximo que pode ser atribuído à nota de um candidato (maxNota); ii) nota do candidato A (notaA); e iii) nota do candidato B (notaB). O redutor tem a função de evitar que comparações entre candidatos com desempenhos divergentes ocasionem um falso positivo (Wesolowsky, 2000). O redutor pode ser obtido pela fórmula:

$$Red = 1 - \frac{|\ NotaA - \ NotaB\ |}{MaxNota}$$

Passo 5 – Obter a soma dos IDQ de todas as questões coincidentes (resposta do Candidato A igual ao B) e corretas (igual ao Gabarito), como demonstrado na Figura 4.6, e dividir por MaxIDQCandAB (passo 3) e, por fim, aplicar o redutor (Passo 4), conforme a fórmula:

$$\gamma_1 = \frac{\sum_{n=1}^{qcc} IDQn}{\text{MaxIDQCandAB}} \ .Red$$

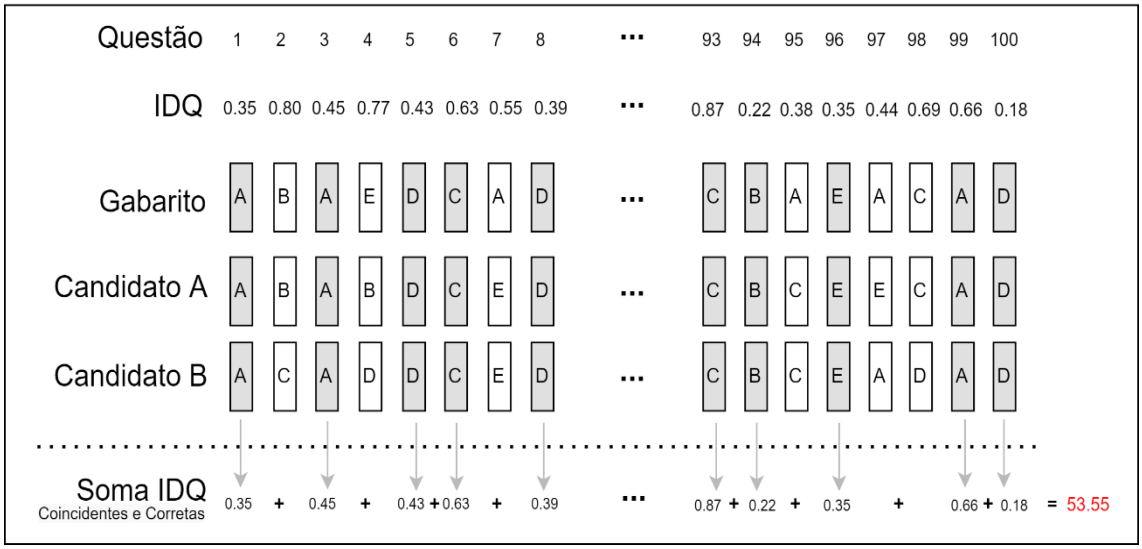

Figura 4.6 – soma dos IDQ das questões coincidentes e corretas

Como dito anteriormente, γ_1, representa o cálculo relativo a uma única comparação das respostas de dois candidatos (CandA, CandB). Como se trata de uma comparação 1:n, são realizados comparações e cálculos do candidato com todos os outros candidatos aproveitados. Ao final, o maior valor obtido de todas as comparações é assumido por γ_1.

4.3.2 Abordagem 2 (*γ2)*: Relação IDQ questões acertadas x IDQ questões erradas

Nesta abordagem, faz-se uma análise somente das respostas do candidato avaliado (Candidato A), sem comparações com outros candidatos, o que se convencionou de avaliação 1:0. O objetivo é obter a razão de dois valores: i) a média dos IDQ das

respostas corretamente respondidas; ii) a média dos IDQ das respostas incorretamente respondidas, conforme segue:

$$\gamma_2 = \frac{medIDQcorretas_{CandA}}{medIDQincorretas_{CandA}}$$

O racional por trás dessa análise está no fato de que, em uma situação normal, um candidato aproveitado, bem-preparado, tende a acertar muitas questões, perdendo poucas questões fáceis e errando somente as mais difíceis. Com isso, quando um candidato induz erro intencional ou recebe as questões erradas já definidas, esta lógica pode ser contrariada e o indicador gerado tende a ser próximo ou ligeiramente maior que 1. O IDQ é usado para pesar a dificuldade das questões respondidas pelo candidato.

4.3.3 Seleção do índice parcial *γ* e desdobramentos

Com respeito à seleção de um valor para o índice parcial *γ*, apurados os dois índices (*γ1* e *γ2*), que representam as duas abordagens computadas, têm-se alguns desdobramentos, a saber:

1) O maior dos dois índices apurado (*γ1* e *γ2*), limitado ao valor 1, será o índice de suspeição *γ'*. Após todas as comparações entre candidatos (1:n) serem realizadas, apurado o maior valor de *γ'*, este será atribuído ao índice parcial *γ* (exemplo apresentado na Tabela 4.6).

2) Caso o índice *γ1* supere o limite de similaridade (LSi)[14], é adicionada a tupla [CandA, CandB, *γ1*] ao banco de similaridades. Esse banco não tem aplicação na avaliação em curso, pois não compõe o índice de suspeição final do candidato, mas serve de base para retroalimentar o método como um dos bancos de suspeição. Também há um uso, não incluído nesse estudo, que é o de estabelecer relações entre candidatos, com índices de similaridade alto, o que poderá servir para uma futura extensão do método ou para subsidiar eventuais investigações. Na Figura 5.9 há um exemplo desse relacionamento obtido com a aplicação do MDFCP.

[14] Índice calculado pelo algoritmo *Covariance Estimator*, da biblioteca *Scikit-learn*.

Tabela 4.6 – Exemplo de valores de γ1, γ2 e γ testados

Id	Classe	Nota	*γ1*	γ2	γ
4079E66D9	1	94,2590	0,9423	0,9306	0,9423
10A753A15	2	94,0740	0,9373	0,9600	0,9600
9A62BA534	3	94,0740	0,9470	0,8301	0,9470
1BC7F9AA7	5	90,7410	0,9475	0,9202	0,9475
1EE1AF513	10	89,0740	0,9454	0,8676	0,9454
479EDF39C	11	88,7040	0,9285	0,8762	0,9285

4.4 Análise de registro

A última análise realizada é a de eventuais registros anteriores de suspeição. As fontes de conhecimento destas informações (Figura 4.7) podem ser:

1) *Registros de Suspeição* obtidos em execuções do MDFCP, sobre ocorrências anteriores de um mesmo concurso. Devido à alta concorrência apresentada na disputa pelos cargos oferecidos, é comum que candidatos participem de um determinado concurso em mais de uma oportunidade.

2) *Registros de Anormalidades* (uso de meio ilícito, por exemplo) de candidatos (*blacklists*), de diversos concursos – incluindo os de organizadoras externas ou autoridades públicas.

3) *Registros da Tabela de Similaridades*, gerados pela Análise de RQME. Todas as similaridades acima de um limite (LSi) são incluídas nesta tabela e, independentemente de serem ou não aproveitadas no índice de suspeição de um candidato, podem ser aproveitados nesta análise.

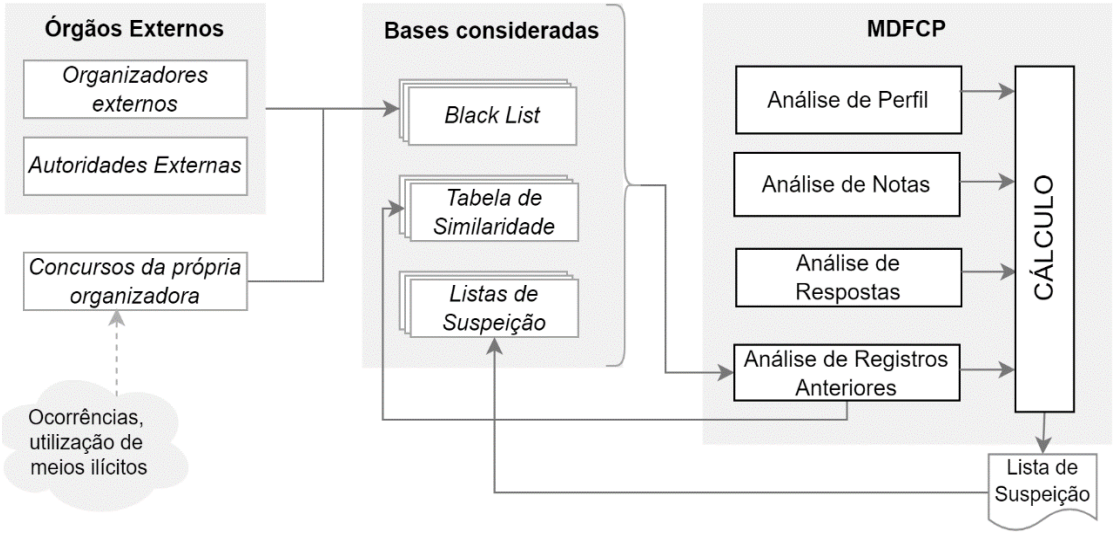

Figura 4.7 – Fontes de conhecimento utilizadas na análise de registros

O índice parcial δ é calculado com base nesses dados anteriores – internos e externos ao concurso. O índice δ é um valor no intervalo [0..1], sendo *n* o número de bases onde foram encontrados registros suspeitos para um determinado candidato e *b* é o número de bases selecionadas, conforme segue:

$$\delta = \frac{n}{b}$$

O ED define as bases que devem entrar nesta análise. Por exemplo, definidas que serão utilizadas 3 bases, e um determinado candidato figura em uma destas bases, gerará o seguinte índice parcial de suspeição: $\delta = \frac{1}{3} = 0{,}333$.

4.5 Cálculo do Índice de Suspeição (IS)

Obtidos os quatro índices, oriundos das análises de perfil, notas, RQME e registros anteriores, é possível calcular o índice de suspeição do candidato, aplicando os pesos atribuídos a cada análise, conforme a fórmula:

$$IS = \frac{\alpha.w_\alpha + \beta.w_\beta + \gamma.w_\gamma + \delta.w_\delta}{w_\alpha + w_\beta + w_\gamma + w_\delta}$$

Onde, *α, β, γ* e *δ* são os índices parciais já descritos. *Wα, wβ, wγ* e *wδ*, são pesos associados a cada índice parcial. Esses pesos são atribuídos pelo especialista do domínio, com base no perfil dos dados, da avaliação da importância que deve ser dada a cada uma das análises consideradas, incluindo desconsiderar totalmente uma ou mais de uma das análises, se julgado necessário.

A soma dos índices parciais ponderados é dividida pela soma dos pesos definidos, resultando no valor do índice de suspeição entre [0..1]. Se o valor do IS for maior que o Limiar de Suspeição (LS) determinado, o candidato é incluído em uma lista de suspeitos. O LS é calculado com a aplicação do algoritmo de detecção de *outlier Covariance Estimator*, da biblioteca *Scikit-learn*, aplicado sobre os índices de suspeição de todos os candidatos avaliados.

O Algoritmo 1 (Apêndice 1 – Algoritmo 1) formaliza como o SI é calculado e como a lista de suspeição é gerada.

Considerando os pesos 4, 2, 8 e 1; e os valores de 0.5142, 0.0000, 0.8689 e 0.2500, respectivamente para as análises de perfil, Notas *Outlier*, RQME e Registros anterior, o Índice de Suspeição de um candidato seria calculado como:

$$IS = \frac{0{,}5142 \,.\, 4 + \ 0{,}0000 \,.\, 2 + \ 0{,}8689 \,.\, 8 + \ 0{,}2500 \,.\, 1}{4 + 2 + 8 + 1} = \frac{9{,}258}{15} = 0{,}6172$$

4.6 O papel do Especialista do Domínio

O método exige a assistência de um especialista do domínio. Esse profissional deve ter uma visão aprofundada dos dados e das diversas características de um determinado concurso. O método exige supervisão em determinados pontos de sua execução e o ED é responsável por avaliar e apontar os ajustes necessários para melhorar a eficiência na detecção de eventuais fraudes. Ao final, o ED realiza a validação dos resultados apresentados.

Dentre as ações que podem ser atribuídas ao ED estão: i) preparação dos dados para uso do método: dados anteriores, dados do concurso avaliado e bases de registros anteriores; ii) determinar os pesos de cada análise componente do MDFCP, podendo sobrepesar ou eliminar da composição do índice de suspeição; iii) validação dos valores parciais e finais gerados pelo método.

O ED pode tanto ser o próprio usuário do método como ser o alimentador dos dados e informações para que um terceiro faça o papel de auditor do processo de seleção.

5 APLICAÇÃO DO MDFCP

Um estudo de caso foi conduzido, visando a validação de hipóteses e testes de eficiência de diferentes análises que compõem o MDFCP. A aplicação do MDFCP foi realizada com a utilização de bases de dados de três ocorrências de um concurso, ajustadas ao método. Um maior detalhamento sobre as características das bases é apresentado na Subseção 5.1.3.

O perfil destas bases de dados é apresentado, juntamente com os parâmetros e ajustes previstos e utilizados na aplicação. Ao final, os dados são consolidados e um resultado geral é apresentado. Cada uma das análises exigiu a utilização de um suporte computacional específico, quer seja a preparação de planilhas, modelos no software Orange DataMining, ou o desenvolvimento de um programa específico para testes, validação e aplicação. Esses suportes e os dados utilizados na execução do método, em cada uma das análises, são apresentadas na Tabela 5.1. O fluxo dos dados entre as bases e as diversas análises e o cálculo do IS, é apresentado na Figura 5.1.

Tabela 5.1 – Componentes, Dados e Plataformas

Componente	Tipo de Dado	Plataforma
Análise de Perfil	Biográficos, geográficos, sociais, número de vezes que realizou o concurso	Orange Data Mining
Análise de RQME	Respostas em questões de múltipla escolha, gabarito, índice de acerto global por questão, parâmetros	Protótipo de Software Desenvolvido
Notas *outlier*	Notas nas disciplinas, agrupadas por área/afinidade (humanas, exatas, biológicas, profissionais etc.)	Orange Data Mining
Análise de registros anteriores	Detecções ou suspeitas registradas anteriormente	Planilha Desenvolvida
Cálculo do IS	Índices parciais oriundos das análises anteriores; parâmetros	Planilha Desenvolvida

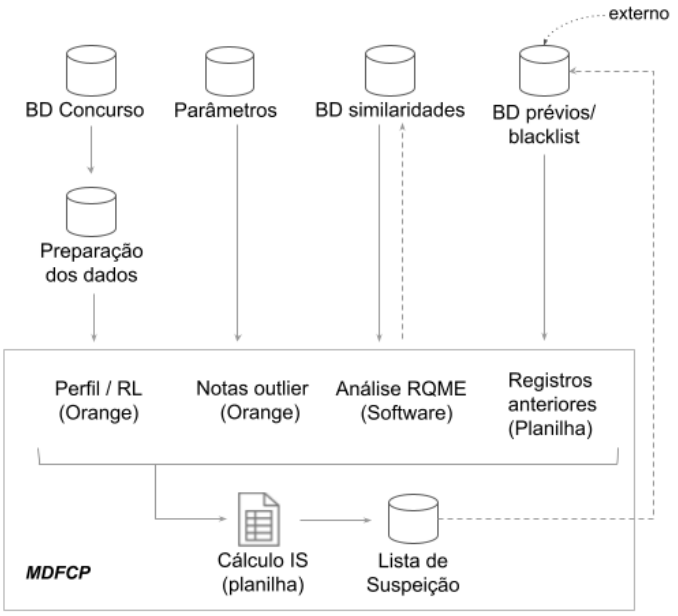

Figura 5.1 – Fluxo de Dados na execução do MDFCP

Nas seções deste capítulo serão apresentadas: i) as bases de dados utilizadas (perfil e preparação); ii) a preparação dos dados necessários à aplicação do método; ii) a aplicação da análise de perfil; iii) a aplicação da análise de notas outlier; iv) a análise de RQME e a utilização da fraude induzida para validação desse índice parcial de suspeição; v) análise de registros anteriores; e vi) consolidação dos índices e a produção dos índices de suspeição.

5.1 Bases de dados utilizadas

As bases utilizadas contemplam um perfil do candidato (dados biográficos, geográficos e realização anterior do concurso), respostas às questões de múltipla escolha, gabaritos, índice de dificuldade das questões, notas e informações de concursos anteriores. No Apêndice 6 há uma descrição sucinta dos dados utilizados nesta execução do MDFCP.

Outro aspecto considerado é o da separação dos candidatos, seja por divisões como introdução de cotas sociais ou raciais, diferentes carreiras selecionadas por um mesmo concurso ou outras situações. Quando este fator cria grupos muito diferenciados, torna-se aconselhável que a avaliação seja realizada de forma separada, para cada um dos grupos.

5.1.1 Perfil da base de dados

Foram preparadas bases de dados relativas a três ocorrências subsequentes de um concurso. As bases foram denominadas “A-2”, “A-1” e “A”, sendo a última a mais atual. Com a utilização das três bases, foi possível conduzir a aplicação das análises, conforme resumido na Tabela 5.2. Também, aproveitou-se de uma das características do método, que é se retroalimentar com dados de aplicações em concursos anteriores.

O método poderia ter sido aplicado a uma única ocorrência, aliás, como foi feito para a validação individualizada de cada um dos seus componentes. Especificamente para a ocorrência A-2, devido a não existir concurso anterior, não foram realizadas as análises de perfil (método requer treinar um modelo) e de registros anteriores.

Tabela 5.2 – Quadro resumo: análises do método por ocorrência do concurso

Método / **Ocorrência**	**A-2**	**A-1**	**A**
Análise de Perfil	Não	Sim	Sim
Análise de Notas *Outlier*	Sim	Sim	Sim
Análise de RQME	Sim	Sim	Sim
Análise de Registros Anteriores	Não	Sim	Sim

Cada uma das ocorrências apresentou os dados de candidatos, com os seguintes atributos: identificador, quantidade de vezes que realizou o concurso, local de realização da prova, idade, tipo de escola, cidade e estado de residência. Cada candidato tem associado, também, um conjunto de 100 respostas de múltipla escolha, com o domínio de respostas sendo {A, B, C, D, E, “em branco”}. Estas respostas pertencem a 3 disciplinas da área de exatas, 2 disciplinas em linguagens e a 2 disciplinas da área de humanas.

Um gabarito com as respostas corretas e o percentual de acerto de cada questão também foram utilizados. Também compõem a base de dados as notas de cada candidato em cada uma das sete disciplinas. Cada base de dados tem entre 585 e 725 candidatos aproveitados e outras características compiladas na Tabela 5.3.

Tabela 5.3 – Resumo dos dados por ocorrência do concurso

Ocorrência	Total Candidatos	Total Aproveitados	Questões	Locais de Exame
A-2	22.063	725	100	38
A-1	31.979	685	100	36
A	29.516	585	100	36

Por se tratar de candidatos aproveitados e com desempenho superior, as notas finais se concentram em valores altos, como se observa na Figura 5.2.

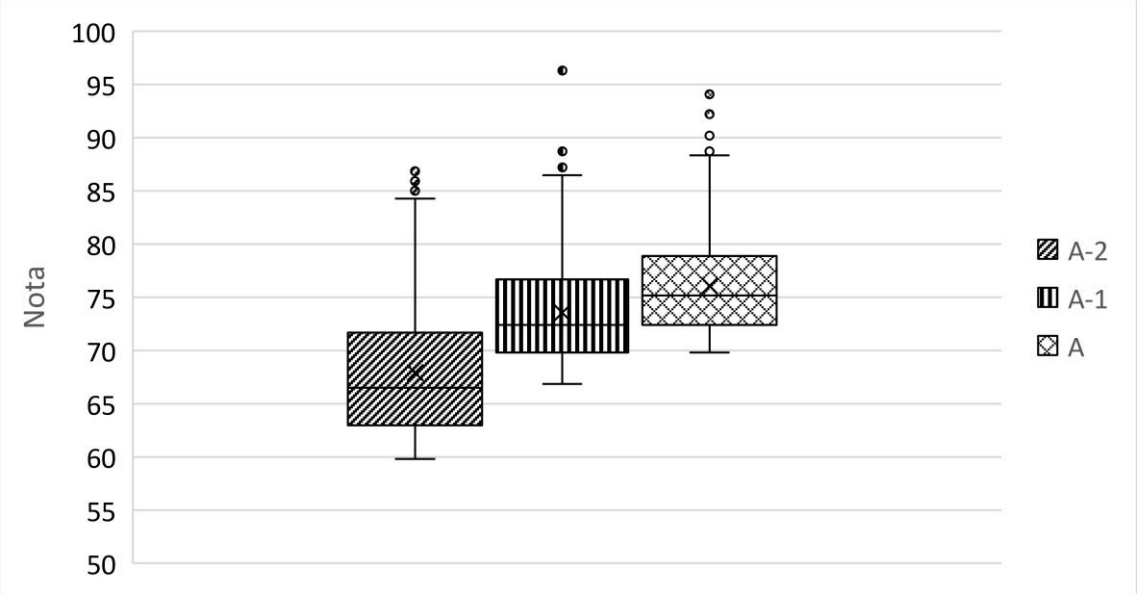

Figura 5.2 – Distribuição de notas por ocorrência do concurso

5.1.2 Preparação das bases de dados

Como abordado no referencial teórico, as bases de dados obtidas devem sofrer uma série de procedimentos, visando a sua preparação para utilização no método. Na Tabela 5.4, está relacionado o preconizado em Brownlee (2020) às tarefas realizadas sobre as bases de dados utilizadas no MDFCP.

Tabela 5.4 – Preparação de dados para uso no MDFCP

Tarefas	Utilização na preparação ao MDFCP
Limpeza dos Dados	Pouco esforço foi empregado devido às várias consistências impostas pelas bancas dos concursos e pelas características dos dados.
Seleção de características	Identificar os dados biográficos, geográficos e de desempenho que servirão para treinar um modelo de perfil. Identificar as notas e agrupá-las com base em critério de semelhança (exatas, humanas, técnicas etc.).
Transformação dos dados	Dentre milhares de cidades, considerar as 90 mais frequentes e um valor "Outras Cidades" para as demais (limitação do algoritmo de RL). Reduzir a dimensionalidade de "Tipo de Escola" para uma das três seguintes opções: particular; pública; pública com restrição (com seleção ou critério restritivo para admissão).
Engenharia de características	Código de local de prova → cidade/UF Data de nascimento → idade Dados de vários concursos → quantidade de vezes que realizou o concurso Provas com questões em ordem diferenciada → respostas em uma mesma ordem

Pelas características dos dados envolvidos, não há de se esperar ruídos ou erros significativos no banco de dados. Os dados do candidato são consistidos na sua ficha eletrônica de inscrição, parametrizando valores como locais de prova, data de nascimento dentro da faixa etária prevista, dentre outros. Pode-se argumentar que os dados podem ser informados de forma imprecisa, entretanto, o conjunto de dados se referem a candidatos aproveitados, isto é, convocados para as demais fases do concurso. Sobre estes candidatos, normalmente, é realizada uma checagem mais detalhada dos dados, confrontados com a apresentação de documentação comprobatória.

Além das tarefas de preparação dos dados, outras específicas para o domínio podem ser aplicadas em função das características do BD:

1) Anonimizar dados pessoais de candidatos, omitindo informações tais como nome, números de documentos, telefones, endereços, nome de mãe, dentre outros.

2) Gerar uma chave de identificação aleatória para cada candidato, associando os seus dados, as respostas em prova, notas apuradas e aos demais dados relacionados. Com isso, garante-se a segurança dos dados repassados ao método, assegurando que, mesmo que acessados ou repassados indevidamente, não infringirão as leis vigentes de

privacidade e guarda de dados. A associação “candidato – chave aleatória” fica em poder da gestora do concurso. Ressalta-se que, um candidato que participe de diversos concursos avaliados deve receber um mesmo número de identificação, visando o aproveitamento de informações de concursos anteriores, que é utilizada no método.

3) Ajustar as respostas de múltipla escolha a uma mesma ordem de alternativas e questões, caso existam provas em ordens diferenciadas.

4) Excluir registros de candidatos que tenham feito a prova de forma parcial ou tenham sido eliminados.

5) Filtrar os dados, devendo ser utilizados na execução do método somente os dados de candidatos aproveitados, *i.e.*, aqueles aprovados e convocados para fases posteriores ou para tomar posse do cargo, por uma questão de racionalização. Assim, o processamento fica restrito a uma menor quantidade de candidatos, o que impacta significativamente em processos com complexidade polinomial [$\Theta(n^2)$], como o que ocorre na análise de respostas de um candidato em relação a todos os outros [1:n]. A exceção fica por conta de dados de treinamento de perfil, que devem considerar candidatos aproveitados e não aproveitados, para aumentar a precisão do modelo de perfil gerado pelo método de regressão logística.

6) Quando o concurso possuir segmentação de candidatos (*e.g.*, sexo, cota racial, carreiras distintas), com a formação de grupos heterogêneos entre os segmentos, é conveniente executar o método sobre cada universo separadamente.

5.1.3 Simulação parcial das bases

Em outros domínios de aplicação é comum se obter bases de dados para validação de métodos de detecção de fraude, como de algumas instituições financeiras, por exemplo. No domínio de concursos públicos o cenário é mais complexo e tais dados não são ofertados, pois há a preocupação das instituições em que eventuais fraudes detectadas as exponham e atentem contra um dos seus principais pilares, que é a segurança e a lisura dos processos que conduzem. E, mesmo que existissem tais bases de dados, estas seriam extremamente desbalanceadas. Pela imensa quantidade de transações financeiras que são realizadas, é razoável presumir que existam muito mais transações

financeiras fraudulentas e que acabam expostas e registradas com o tempo, do que fraudes em concursos públicos.

Assim, para a aplicação do MDFCP, não houve acesso a bases de dados completas, nem interesse das instituições em fornecer tais dados. Em face a essa dificuldade, usou-se no estudo um banco de dados parcialmente simulado. Parte dos dados que o compõem, como dados do perfil do candidato (biográficos, geográficos e notas de disciplina), são públicos. Entretanto, dados que poderiam identificar um candidato são eliminados ou anonimizados, como tratado no item 5.1.2 (preparação dos dados).

As respostas de um candidato são simuladas, com base nas notas reais. O viés introduzido por esse procedimento pode ser contornado na validação do método, como será demonstrado no item 5.4 (Análises de Respostas de Questões de Múltipla Escolha).

5.2 Análise de Perfil

Nos testes realizados, a Regressão Logística atingiu uma boa precisão, se mostrando uma técnica adequada para verificar a adequação do perfil de um candidato. É favorável a um cenário com variáveis independentes e uma variável categórica dependente (“Situação”: aproveitado/não aproveitado). Os dados dos candidatos seguem a estrutura constante da Tabela 5.5.

Tabela 5.5 – Conjunto de dados utilizados no treinamento do modelo de RL

Dados Biográficos
Idade
Tipo de Escola
Recorrência
Dados Geográficos
Cidade de Residência
UF de Residência
Local de realização da prova
Rótulo
Situação (aproveitado/não aproveitado[15])

[15] Candidato aproveitado é uma nomenclatura adotada nesta dissertação para generalizar termos como candidato aprovado, candidato convocado, candidato chamado que aprecem em diferentes concursos.

A análise do perfil foi realizada sobre as ocorrências “A-1” e “A”. A ocorrência “A-2” não possuía concurso anterior que permitisse treinar um modelo, portanto, não foi realizada.

O conjunto de dados de treinamento (amostra apresentada na Tabela A2.1, no Apêndice 2) baseou-se em uma amostra de candidatos de um concurso anterior ao em avaliação (“A-2” e “A-1”), divididos em dois grupos: (1) aproveitados, chamados para as próximas etapas do concurso, e (2) não aproveitados, nas quantidades explicitadas na Tabela 5.6.

Tabela 5.6 – Quantidade de registros da tabela de treinamento, por ocorrência

	Treinamento	
Ocorrência	**Aproveitados**	**Não aproveitados**
A-2	700	700
A-1	696	696

Os dados de validação de cada ocorrência do concurso foram submetidos ao seu respectivo modelo de RL (amostra destes dados na Tabela A2.2, no Apêndice 2). Esse conjunto de dados continha dados de todos os candidatos aproveitados, ou seja 685 para a ocorrência “A-1” e 585 para a ocorrência “A”. O algoritmo de predição (RL) retornou os mesmos valores de entrada mais a probabilidade de aprovação dos índices e o valor previsto para o rótulo “Aproveitado” (Apêndice 2 – Tabela A2.3). O valor de complemento para 1 dessa probabilidade é o índice parcial de suspeição atribuído a cada candidato. O modelo treinado apresentou os indicadores e valores de validação explicitados na Tabela 5.7.

Tabela 5.7 – Indicadores dos modelos de RL

	Indicadores				
Ocorrência	Validação	K-Fold	Precisão	Recall	Score F1
A-1	72.3%	5	0.771	0.740	0.755
A	75.9%	5	0.731	0.728	0.727

5.3 Análise de Notas *Outlier*

A análise de notas *outlier* foi realizada com dados de três ocorrências, “A-2”, “A-1” e “A”. Para cada uma delas foi aplicado o algoritmo de detecção de *outlier* sobre cada um dos grupos de disciplinas: 1) Exatas; 2) Humanas; e 3) Linguagens. A Figura 5.3

apresenta o esquema utilizado para detecção de notas *outlier* da ocorrência A-2, desenvolvido no software Orange Data Mining.

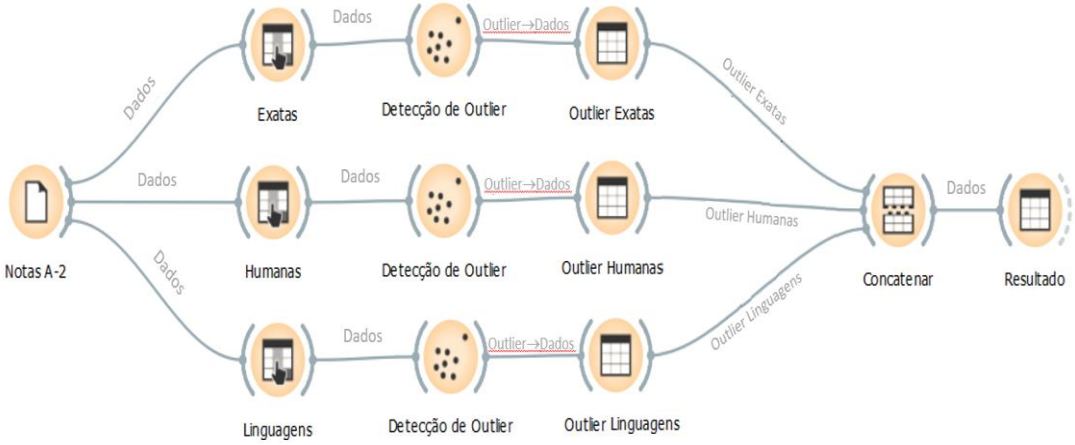

Figura 5.3 – Esquema para detecção de notas outlier

Foram detectados valores *outliers* em todos os grupos de disciplinas (extrato no Apêndice 3, Tabela A3.1), mas não houve candidato que tenha figurado em mais de um grupo. De forma resumida, a Tabela 5.8 apresenta as quantidades de *outlier* por contexto, por grupo de disciplinas, o total e o percentual de *outliers*.

Tabela 5.8 – Valores outliers detectados

Contexto	Exatas	Humanas	Linguagens	Soma	*Outlier*
A-2	26	22	18	66	9,1%
A-1	9	18	28	55	8,0%
A	8	18	14	40	6,8%

Observou-se que uma maior concorrência trouxe como consequência a concentração das notas em valores mais altos, com valores menos diferenciados e menor detecção de valores *outliers*.

5.4 Análise de Respostas em Questões de Múltipla Escolha

A Análise de Respostas em Questões de Múltipla Escolha (RQME), como requer somente os dados do próprio concurso avaliado, foi realizada sobre as três bases disponíveis. Cada uma das bases, para cada candidato, apresentou as respostas para 100 questões de múltipla escolha. Além disso, a análise requereu os gabaritos de cada ocorrência e o IDQ de cada questão. A Figura 5.4 apresenta a distribuição de IDQ de candidatos aproveitados e geral, em 3 ocorrências.

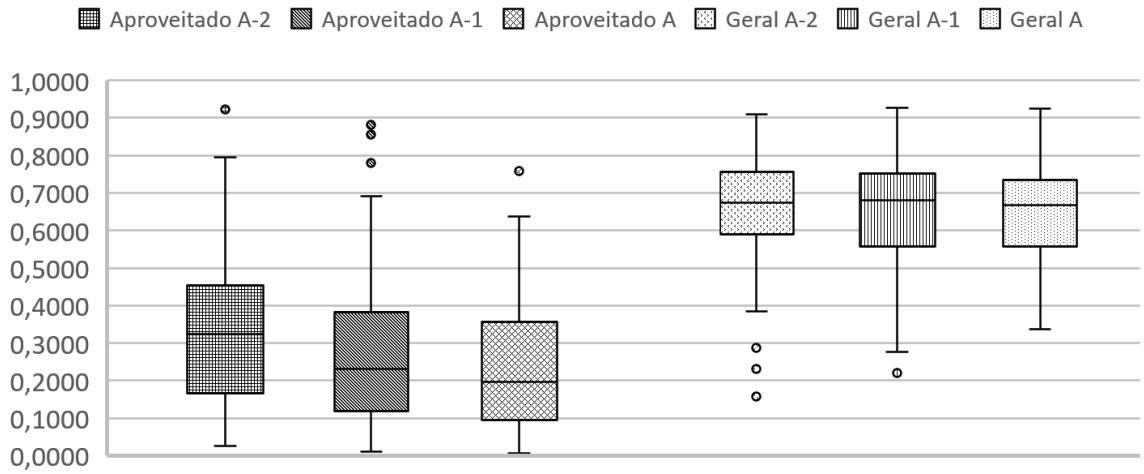

Figura 5.4 – Distribuição de IDQ de candidatos aproveitados e geral

Uma vez que não estão disponíveis bases de testes que contemplem fraudes em concursos públicos, foi usada uma abordagem, chamada neste trabalho de Inserção Simulada de Fraude (ISF), como uma alternativa para verificar a precisão do método, especificamente na análise de Respostas em Questões de Múltipla Escolha (RQME).

Existe uma linha lógica que direcionou a indução de fraude, que começa pela identificação de dois métodos de fraude. Um é a comunicação de um elemento externo com o candidato realizando a prova, seja por ponto eletrônico, *smartwatch* ou dispositivo semelhante. Nesse caso, é situação comum que existam vários candidatos fraudadores, até em diferentes locais de prova. Outro tipo de fraude comum é a inscrição de candidatos homônimos, com documentação falsa, preparados para o concurso e que, em algum momento prévio ao início ou por comunicação em sala durante as provas, repassam as questões ou preenchem cartões de respostas trocados.

Nesse cenário, imagina-se que exista a precaução para que o fraudador tenha um desempenho mínimo para aprovação e, por outro lado, não acerte todas as questões da prova (“Gabaritar”), não atraindo atenção para si. Como os dados de notas de candidatos aprovados são públicos e estão ao alcance dos criminosos, isso é plausível. Para se adequar a esse cenário, foram cogitadas duas maneiras de indução de erro. Uma em que os erros já seriam repassados oriundos da origem externa, o que induziria a conjuntos de respostas semelhantes. Outra, em que os candidatos fraudadores, uma vez recebidas as questões, seriam orientados a induzirem erros aleatórios, evitando, assim, semelhança exacerbada nas respostas.

Na Figura 5.5 estão evidenciados os dois tipos de indução de erro, cada um com quatro “fraudadores” e suas respectivas respostas. A primeira linha evidencia o gabarito. Observa-se que as respostas determinam uma nota, e esse conjunto de respostas é inserido no banco de respostas na posição adequada (Figura 5.6). A eficiência da Análise de RQME é medida pela quantidade de respostas de fraudadores detectadas através dos índices de suspeição gerados.

	Gabarito	A	E	B	B	C	A	D	C	A	E	C	D	...	B	A	C	D	B	D	B	E	C	A	E	C	
																											Nota
Tipo 1	Fraude A1	E	E	B	B	C	A	D	C	C	E	C	D		A	A	C	D	E	D	B	A	C	A	D	C	78,59
	Fraude A2	A	E	A	B	C	A	D	C	E	E	C	D	...	B	A	C	C	B	D	C	E	C	A	E	C	80,73
	Fraude A3	A	A	B	B	E	A	D	C	C	E	C	B		C	A	C	D	E	D	B	B	C	A	A	C	75,73
	Fraude A4	E	E	B	B	A	A	D	C	C	E	C	D		B	E	C	D	B	D	C	E	E	B	E	C	80,02
Tipo 2	Fraude B1	A	A	B	B	C	E	D	C	A	E	C	E	...	B	B	C	D	B	C	B	E	E	A	C	C	77,88
	Fraude B2	A	B	B	B	C	D	D	C	A	E	C	A		B	E	C	D	B	A	B	E	A	A	C	C	77,88
	Fraude B3	A	C	B	B	C	C	D	C	A	E	C	E		B	B	C	D	B	E	B	E	E	A	A	C	77,88
	Fraude B4	A	A	B	B	C	C	D	C	A	E	C	A	...	B	B	C	D	B	C	B	E	D	A	D	C	77,88

Figura 5.5 – Amostra de dois tipos de fraude induzida

A50CDF91	A	D	B	B	C	E	C	C	A	E	C	D	B	D		C	D	B	D	B	E	C	A	E	C	64	80,741
327ADB64F	A	E	A	B	C	A	D	C	E	E	C	D	B	B		C	C	B	D	C	E	C	A	E	C	65	80,730
AD00F70F2	A	C	B	B	C	E	D	C	A	E	C	D	B	B		C	D	B	C	E	E	A	A	E	C	66	80,370
159E9E65C	C	B	A	C	C	E	A	C	A	E	C	D	B	B	...	C	D	B	D	B	E	C	A	E	C	67	80,370

Figura 5.6 – Fraude introduzida na base de dados em posição adequada à nota

O MDFCP considera duas abordagens para a geração do índice de suspeição, gerado na Análise de RQME. Isso, e a adoção da ISF, foram extensões do método proposto em Nunes, Jino, *et al*. (2021). Os testes para validação da Análise de RQME, com o uso do ISF, utilizaram duas bases de dados (“A-2 Simulação” e “A-1 Simulação”), realizado de forma isolada dos demais testes de validação.

As fraudes induzidas foram evidenciadas pela Análise de RQME, como se observa parcialmente na Figura 5.7, por uma ou outra das abordagens utilizadas (“Similaridade” ou “Relação IDQ_Certas e Erradas”). A eficiência das abordagens foi validada com a identificação de 15 das 16 fraudes, inseridas nas duas bases testadas, como apontado na Tabela 5.9.

idCandidato	RespostasCndt		Class	Média	Similaridade	Rel_IDQ_CxE
3E6B9CEF1	EACBABACBCDDBDCBDBDCABDEBCEDC		99	75,7	0,878029935	1,143369176
257712902	EEEBEBECBEEDBECBDEDCCBBEACEDA		145	73,6	1,021371113	1,044735513
36A62B542	EDEBCBECBDEDBBCBDDDCCBAEACEDA		146	73,6	1,021371113	1,044735513
1A31DB4EF	ECEBBBECBAEDBACBDADCCBAEACEDA		147	73,6	1,021371113	1,044735513
37FE122A5	EBEBCBECBAEDBBCBDCDCCBEEACEDA	...	148	73,6	1,021371113	1,044735513
3E230FEF7	EAEBCAECDAEDBACAEAACCBBEDCEBA		680	60,5	0,897443996	0,987897968
3BD9DFFE	ABBBABECBCEAADBCBBDACBDEACEDA		105	75,1	0,953207929	0,983223487
E383A5A0C	AAEBABECBCEDBDEBDBDACBDEACEDA		12	82,7	0,927956152	0,968520966
DD2D4045	EAEBBBECBCBDBDEBBBDCCBDEACEDA	...	43	78,8	0,932653411	0,964952106
3E6B51C08	EBEBAEECBCECBDABDBDCEBDEACECA		53	77,9	0,891936113	0,963001758
976C6779A	EBEDAAECACEAADBBDBDACBDEACEBA		80	76,4	0,918898153	0,961080506
92110A73	BCEDEAECDCEDBABDDBDACBDEACEDA		98	75,7	0,915868916	0,958971219
2A0734C98	ABABABECBCEECDDDDBDACBDEACEDA	...	29	80,9	0,953207929	0,956704522
DB360770	EAECABECBCEABACBEBDACBDBACEDA		106	75,	0,925921674	0,9557074
94FD48AD2	AAECCBECBCEACACBEBDECBDEACEBA		10	83,5	0,917903244	0,950995671
1054655B80	AAEBCBACBCEDBDCCDDDACBDEACADA		25	81,4	0,872674671	0,950158203
1E035DFDC	DAACABECACEDBDDBCBDCCBDEACEDA		167	72,2	0,905443709	0,948761641
3E51A4849	BAEBABECACEDBABBEBDACBDEACEDA		31	80,3	0,929446678	0,947674419

Figura 5.7 – Fraude evidenciada na Análise de RQME

Tabela 5.9 – Identificação de fraude via ISF

Ocorrência	Detectada	Não detectada	% Acerto
A-1 Simulação	8	0	100%
A-2 Simulação	7	1	87,5%
Geral	15	1	93,75%

5.4.1 Resultados da Análise de RQME

As duas abordagens previstas na Análise de RQME foram aplicadas nas bases disponíveis, gerando dois índices parciais (*γ1 e γ2)*, sendo considerado o maior dentre os dois como o índice parcial de suspeição (totais considerados por abordagem na Tabela 5.10).

O valor de LSi (Tabela 5.10) foi utilizado para determinar os valores que serão armazenadas na tabela de similaridades e foi determinado pela aplicação de um algoritmo de *outlier*[16] sobre os valores dos índices parciais *γ1'* (independente se o maior encontrado é que assumirá como valor de *γ)*, calculados para todos os candidatos de uma determinada ocorrência de um concurso.

[16] *Covariance Estimator*, da biblioteca *Scikit-learn* usando *Orange DataMining.*

Tabela 5.10 – Quantidade de índices de similaridade considerados

Ocorrência	γ1	γ2	LSi	Intervalo γ' [17]
A-2	633	92	0.9075	[0.6341 .. 0.9448]
A-1	417	279	0.7263	[0.4544 .. 0.8319]
A	401	184	0.7018	[0.4636 .. 0.8357]

Na Figura 5.8 estão distribuídos os valores dos índices relativos à Análise de RQME, calculados para todos os candidatos, nesse exemplo, para a ocorrência "A-2". Nos Índices de Similaridade (γ2), marcadores em laranja no gráfico, há uma discreta correlação entre a classificação e o valor desse índice, situação que foi observada e previamente atenuada com a adoção do redutor relativo à disparidade de desempenho entre dois candidatos com as respostas comparadas. A linha laranja é a linha de tendência.

Ainda no gráfico da Figura 5.8, os marcadores azuis apresentam a distribuição dos Índices (γ1) relativos à relação da soma de IDQ de questões certas por questões erradas, também calculados para cada candidato da ocorrência "A-1". Nessa ocorrência não se visualiza nenhuma correlação classificação x índice, mas, de forma geral, os valores *outlier*s apontam para situações suspeitas, tanto que a maior parte das "capturas" da ISF foram devidas a essa abordagem. A linha azul é a linha de tendência.

Observa-se no gráfico da Figura 5.8 uma quantidade maior de índices γ1 com valores mais elevados. De fato, como também consta na Tabela 5.10, os índices γ1 são mais considerados na detecção de anomalia nas RQME, mas ambos os índices trabalham em conjunto para obter maior precisão nessa abordagem do MDFCP.

[17] Maior índice dentre γ1 e γ2, limitado ao valor 1.

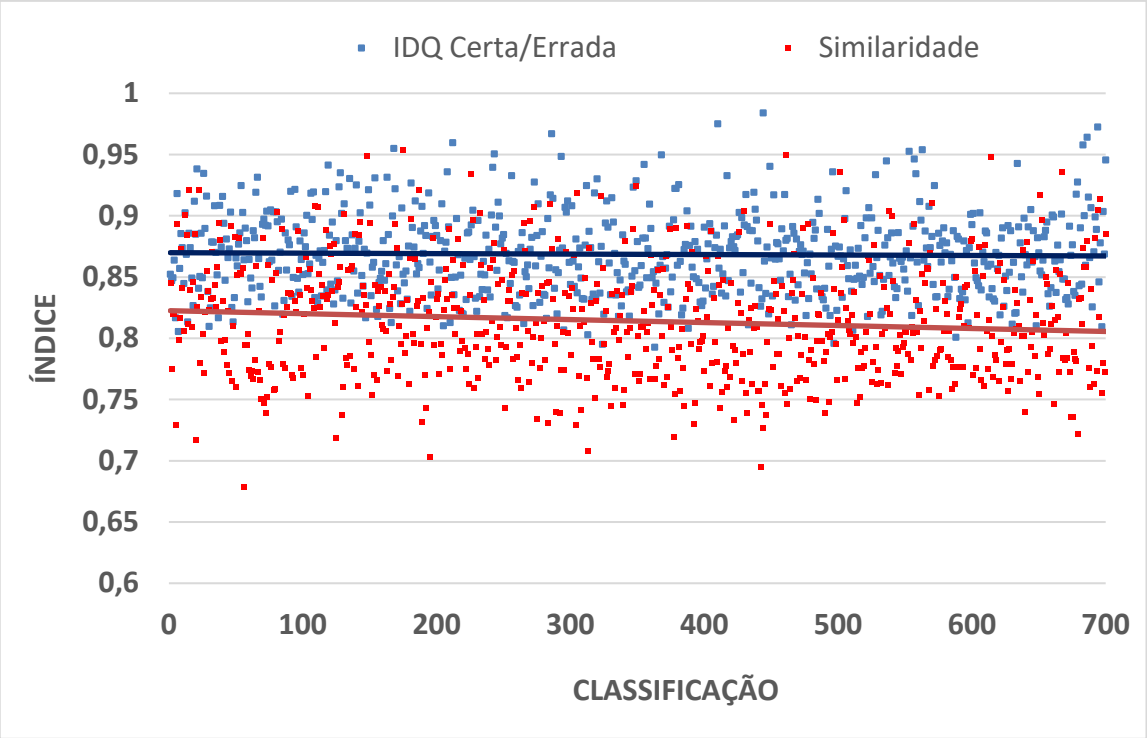

Figura 5.8 – índice de similaridade γ1 e γ2 x Classificação do Candidato

Como resultado adicional, não considerado para o cálculo do índice de suspeição, há o registro de todas as similaridades entre dois candidatos, além do limite delimitado por LSi, na tabela de similaridades. Esses dados ajudarão a identificar relacionamentos entre candidatos com suspeição. Como exemplo, um candidato com esse tipo de registro foi selecionado e relacionado aos respectivos candidatos com respostas similares e ao índice de similaridade entre eles, como apresentado na Figura 5.9.

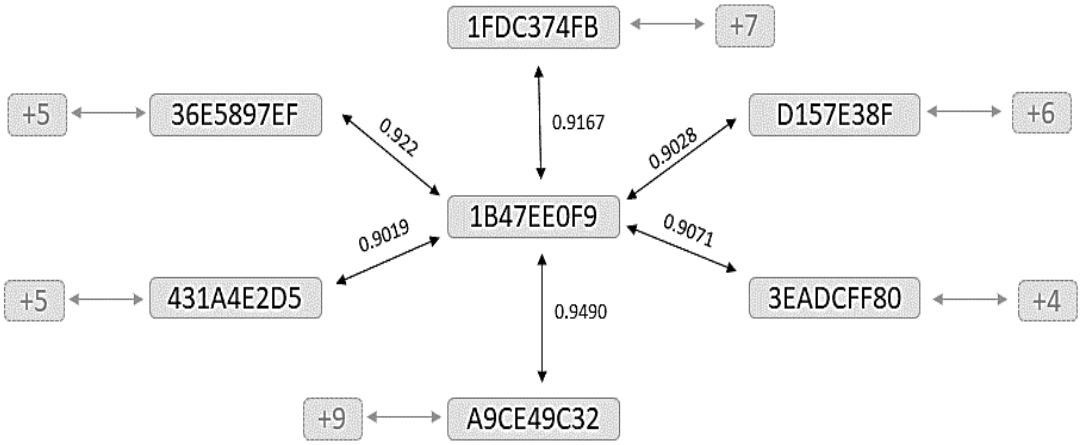

Figura 5.9 – Exemplo de relacionamento entre candidatos com similaridade alta em RQME

5.5 Análise de Registros Anteriores

A análise de dados de aplicações anteriores do método teve um menor peso, em certa medida porque candidatos de um concurso não constarão em listas de suspeitas de concursos anteriores, uma vez que foram aproveitados e não há lógica de repetirem o concurso. Por isso, foram considerados os dados de similaridade da tabela de similaridades de anos anteriores (Tabela 5.11), determinado por todas as comparações entre candidatos que excederam LSi. Não foram consideradas bases de terceiros ou registros internos sobre candidatos.

Tabela 5.11 – Percentual de registros anteriores encontrados

Ocorrência	Cand. com registros	% Total
A-1	2	0,2873%
A	10	0,1709%

5.6 Geração do Índice de Suspeição (IS)

Para o cálculo do Índice de Suspeição (IS) de cada candidato, de um concurso em particular, com valor no intervalo [0..1], são necessários: i) cada um dos índices parciais calculados (obtidos nos passos anteriores) e ii) os pesos atribuídos, pelo ED, a cada índice parcial (Tabela 5.12).

Tabela 5.12 – Pesos / parâmetro utilizados e percentual de suspeitos

	Peso nas Análises do MDFCP[18]						
Ocorrência	Perfil	Nota	RQME	Anterior	LS	Suspeitos	% Suspeitos
A-2	0	2	8	0	0,907526	22	3,03%
A-1	4	2	8	1	0,726337	21	3,02%
A	4	2	8	1	0,701773	18	3,08%

Após o cálculo, o valor é confrontado com o Limiar de Suspeição (LS) específico de um concurso, com base em um algoritmo de detecção de *outlier*[19] aplicado sobre os IS de todos os candidatos. Os IS que superarem os LS, serão adicionados à lista de suspeição. Os LS obtidos e os suspeitos prospectados estão parcialmente listados na Tabela 5.13, que apresenta um extrato do cálculo do IS, com base em todos os índices

[18] Valor arbitrado pelo ED.
[19] *Covariance Estimator*, da biblioteca *Scikit-learn* usando *Orange DataMining*.

parciais calculados e respectivos pesos atribuídos, relativos à ocorrência “A”. Na 8ª linha da tabela está destacado um IS superior ao LS definido.

Tabela 5.13 – Extrato de tabela com cálculo do IS (ocorrência “A”)

Id. Cand.	Análise RQME			Perfil (Peso 4)	RQME (Peso 8)	*Outlier* (Peso 2)	Anterior (Peso 1)	IS
	Rel IDQ C/E	Simila-ridade	Índice					
1042BEA20	0,8851	0,8893	0,8893	0,1340	0,8893	0,0000	0,0000	0,5100
1056BC76A9	0,8788	0,8972	0,8972	0,1351	0,8972	0,0000	0,0000	0,5146
105AE8008A	0,8626	0,8689	0,8689	0,5142	0,8689	0,0000	0,0000	0,6005
105DE56911	0,8879	0,8260	0,8879	0,0699	0,8879	0,0000	0,0000	0,4922
105FF25697	0,9163	0,8943	0,9163	0,1940	0,9163	0,0000	1,0000	0,6071
10697C4C47	0,8921	0,8284	0,8921	0,2197	0,8921	0,0000	0,0000	0,5344
106A10D7B6	0,9106	0,8561	0,9106	0,6550	0,9106	0,0000	0,0000	0,6603
10BF0DDA3	0,9565	0,8259	0,9565	0,2203	0,9565	1,0000	0,0000	**0,7022**
10D926985	0,9274	0,8716	0,9274	0,3700	0,9274	0,0000	0,0000	0,5933
10FB0B594	0,9400	0,8798	0,9400	0,1340	0,9400	0,0000	0,0000	0,5371
111135776	0,8905	0,9220	0,9220	0,3700	0,9220	0,0000	0,0000	0,5904
111987B8D	0,8733	0,9091	0,9091	0,7722	0,9091	0,0000	0,0000	0,6908

Com respeito à parametrização de MDFCP, existe a possibilidade de se sobrepesar determinada análise em relação a outra, em função da avaliação dos dados disponíveis e suas características. É possível, inclusive, eliminar determinada análise do método, atribuindo peso zero, o que foi feito na ocorrência “A-2”, na Análise de Perfil e Registros Anteriores, uma vez que não existia concurso anterior disponível para utilização. Não só por isso, mas pela avaliação de que determinada análise, em confronto com os dados disponíveis, não seja adequada e deva ser suprimida.

6 DISCUSSÃO

Este Capítulo apresenta uma discussão sobre o método (Seção 6.1), as bases de dados utilizadas (Seção 6.2) e uma breve conclusão sobre os trabalhos relacionados (Seção 6.3), além de dificuldades e limitações da proposta (Seção 6.4).

6.1 MDFCP

MDFCP é proposto com o objetivo de apoiar a detecção de fraudes em uma grande variedade de concursos, desde que inclua questões de múltipla escolha e tenha um banco de dados com um mínimo de atributos sobre os candidatos.

Por ser parametrizável, MDFCP se adequou bem às três ocorrências de concurso utilizadas para validação do método. Este se mostrou viável e apontou indícios de suspeição. Embora atenuado pelo fato de o método avaliar diversos aspectos, como o método se baseia em estatística e técnicas de aprendizado de máquina, há uma possibilidade não desprezível de se obter falsos positivos. Por uma questão ética e para que se tenha valor legal, com base nos resultados obtidos na aplicação do MDFCP, é necessária uma investigação complementar, que consubstancie a suspeita apontada pelo método.

Na Análise de Perfil, pelos indicadores (Tabela 5.7) e pela validação do modelo, a Regressão Logística se mostrou uma ferramenta adequada na predição de que um determinado candidato teria o perfil de "candidato aproveitado". O retorno do valor probabilístico (intervalo [0..1]) se mostrou coerente com o adotado no método, sem a necessidade de cálculos ou procedimentos complexos, necessitando somente do cálculo do valor complementar para 1 para indicar a inadequação do perfil.

Na Análise de Notas *Outlier*, as validações prévias do algoritmo mais apropriado, *Local Outlier Factor*, se confirmaram nos testes, com a obtenção de valores *outlier*s condizentes, na avaliação do ED. Cabe ressaltar que a participação do ED foi em parte reduzida quando adotamos algoritmos de *outlier* para determinar LS e LSi. O ED contribuiu na construção do método. Na execução, o ED é mais exigido na distribuição dos pesos de cada análise e na validação dos resultados.

A Análise de RQME exigiu uma série de testes prévios até se chegar à sua forma final, alguns detalhados no Apêndice 5. Ao final, as dúvidas se resumiam a duas questões

principais. A primeira no uso de um IDQ mais amplo (“IDQ Geral”), que considerasse todos os candidatos no seu cálculo; ou um mais restritivo (“IDQ Top”), que consideraria somente os candidatos aproveitados. Nesse caso, nos testes, o IDQ Geral se mostrou mais adequado, uma vez que diferenciava melhor as questões pelo nível de dificuldade (Figura 5.4). A segunda questão dizia respeito ao uso ou não de um redutor de disparidade de performance entre candidatos avaliados e qual seria o redutor mais adequado. Os índices calculados se mostraram mais adequados com a utilização de um redutor que considerasse uma depreciação maior quanto mais distante fosse o desempenho de dois candidatos, considerando a diferença de estratos baseados na nota final. A quantidade de estratos foi definida pela avaliação do intervalo de valores de notas finais apresentados no banco de dados.

O uso da técnica ISF contribuiu sobremaneira para a validação da Análise de RQME, em suas duas abordagens complementares, pois apresentou uma excelente taxa de detecção da fraude simulada de 93,75% nas suas ocorrências simuladas (Tabela 5.9).

Tanto a aplicação da Análise de RQME quanto os testes de detecção de fraude inseridas pela ISF foram realizadas por meio do software de apoio desenvolvido especificamente para essa análise. Esse software se mostrou uma ferramenta versátil para simulações das diversas variações testadas, gerando todas as versões de índices testados simultaneamente, facilitando as comparações de resultado. O Apêndice 4 detalha as características do software desenvolvido.

Observou-se que a concorrência aumentou de forma crescente entre concursos, impactando no perfil das notas finais (Figura 5.2) e na geração de IDQ com valores menores (Figura 5.3), indicando que, no geral, o índice de acerto de questões estava crescendo. Como o IDQ é a base para o cálculo dos índices da Análise de RQME, o aumento da concorrência impactou também na geração destes índices. Com as alterações do perfil dos dados, se fizeram necessários ajustes nos parâmetros utilizados no método, como nos limites de suspeição e similaridade, assim como nos pesos atribuídos a cada índice parcial.

A adoção de um algoritmo de detecção de *outlier*, usado para determinar o LSi na Análise de RQME, substituiu um valor arbitrado, provavelmente pelo ED, e tornou mais homogênea a execução da análise sobre diversas ocorrências do concurso.

A complementariedade das duas abordagens da Análise de RQME foi avaliada na detecção da ISF e na adoção compartilhada do índice de uma delas para associar ao índice γ (Tabela 5.10).

Como procedimento secundário, pois não compõem diretamente o cálculo do índice de suspeição, o registro de todas as similaridades entre dois candidatos ($\gamma 2$) se mostrou eficiente em traçar relações com alto grau de similaridade em RQME, conforme exemplificado na Figura 5.9.

Por fim, a Análise de Registros Anteriores tinha o objetivo de dar um valor parcial, quando um determinado candidato estivesse registrado em bancos anteriores de suspeição, similaridades ou *blacklists*. Na execução do método, não foram utilizadas bases de dados externas e praticamente não houve registros de listas de suspeição anteriores, uma vez que candidatos aproveitados não deveriam participar de concursos posteriores. Por isso, aproveitou-se os dados constantes da tabela de similaridade para compor esse índice parcial. Não foram muitas ocorrências, como se pode notar na Tabela 5.11.

O cálculo do Índice de Suspeição (IS) se vale de todos os índices parciais de suspeição, anteriormente calculados, sendo atribuídos pesos específicos. Esses pesos foram configurados, de acordo com o perfil do BD, dos resultados parciais obtidos e da avaliação do ED. O LS foi determinado pelo algoritmo de detecção previsto no método, sobre todos os índices de todos os candidatos de cada ocorrência do concurso.

De forma geral, o método se mostrou aplicável e viável. Com base no tempo mensurado na aplicação sobre as três bases disponíveis, se mostrou adequado até para concursos com efetivos maiores, pois são considerados somente os candidatos aproveitados, número que tende a não crescer mesmo com uma concorrência maior.

6.2 Bases de dados utilizadas

Todas as bases de dados a serem utilizadas devem ser preparadas como detalhado na Subseção 5.1.2. A anonimização é uma tarefa importante, não interfere na execução do método e garante a necessária proteção dos dados. A relação código x candidato real fica sob a guarda de quem tutela e tem responsabilidade sobre os dados, não necessariamente sendo do conhecimento de quem executa o MDFCP.

Quanto às bases utilizadas na Análise de Perfil, supõe-se que a adição de mais variáveis previsoras, obtidas na inscrição ou em fases posteriores do concurso, podem delimitar melhor um perfil de candidato aproveitado, aumentando a precisão desta análise.

6.3 Trabalhos relacionados

Na revisão sistemática de literatura foram identificados três trabalhos relacionados (Tabela 2.4). A primeira diferença entre eles está no foco de cada um, seja em plágio em questões dissertativas ou coincidência em questões de múltipla escolha.

Os trabalhos relacionados exigem i) conversão de informação escrita para digital (Cavalcanti *et al*., 2012) , ii) apresentam viés na indicação de suspeição por fiscais de setor de prova (Wesolowsky, 2000), e iii) descartam parcialmente alunos com pior desempenho e exigem informações do assento em sala (M. Chen, 2017). Embora exija tratamento prévio dos dados e estudo pelo ED, o MDFCP não apresenta tais características negativas.

No MDFCP a abordagem de identificação de suspeitos é mais abrangente, indo além de similaridades em questões. De forma geral, o MDFCP se mostrou mais flexível e parametrizável do que os métodos relacionados.

6.4 Dificuldades e limitações

A maior dificuldade na condução desse estudo foi a utilização de base de dados parcialmente simuladas. A obtenção de bases de dados reais de concursos é uma tarefa difícil, devido à legislação de privacidade de dados, normas de segurança interna das instituições organizadoras de concursos e do receio destas que eventuais detecções ou divulgações de fraudes causem dano à sua imagem. Para contornar essa limitação, usamos um banco de dados parcialmente simulado. Porém, cabe ressaltar que esta base de dados utilizou como referência a estrutura de dados de uma base de dados real, bem como mantém uma coerência com a base de dados real. Os dados públicos (perfil e notas) foram anonimizados, preparados e complementados com dados simulados (respostas dos candidatos).

Uma primeira limitação existe em função do foco do método, limitado a provas que tenham questões de múltipla escolha e uma base de dados mínima. Outra limitação é

a apresentação dos resultados, que carece de construção de uma interface gráfica mais funcional.

Durante a aplicação do método identificou-se que algumas alterações no perfil das provas, especificamente na distribuição do nível de dificuldade das questões, podem ajudar a diferenciar melhor os IDQ, aumentando a eficiência do método.

Desde o princípio do desenvolvimento do método foi pensado no ED como um ator importante, pois possui uma visão aprofundada de seu banco de dados, da sistemática do concurso e das limitações do processo. Por isso, a presença do ED contribuiu na construção do método, quer seja no levantamento e validação de padrões que sugerissem suspeição, quer seja na avaliação dos resultados apresentados e nos ajustes necessários para melhorar o desempenho do método. Isso continua válido, mas sua participação foi em parte reduzida quando da adoção de algoritmos de *outlier* para determinar o LS e o LSi. Na execução o ED é mais exigido na distribuição dos pesos de cada análise e na validação dos resultados parciais e finais obtidos.

7 CONCLUSÃO

Com a existência de inúmeros concursos que apresentam grande concorrência como única forma de admissão no serviço público e o surgimento de quadrilhas de fraudadores, foi levantada a seguinte questão de pesquisa: "*Como identificar e atribuir computacionalmente, com base em dados de um concurso público, um grau de suspeição de fraude a candidatos*?". Em resposta a isso, foi proposto o Método para Detecção de Fraude em Concursos Públicos (MDFCP), com o objetivo de atribuir um índice de suspeição aos candidatos participantes de um concurso, desde que tenham provas com questões de múltipla escolha e um banco de dados com informações dos candidatos.

Com base na execução completa do método usando três ocorrências de um concurso público, foi possível exercitar MDFCP e validar a proposta com dados simulados.

Espera-se que, com a aplicação de MDFCP em bases de dados de concursos reais, e em outros concursos com perfis diferenciados, o método seja aprimorado e se torne ainda mais flexível e parametrizável, de modo a ser usado em um amplo espectro de provas e concursos.

Os resultados e contribuições do projeto de pesquisa e trabalhos futuros são apresentados a seguir.

7.1 Resultados da Pesquisa

Esta dissertação apresenta contribuição para a área de Ciência da Computação, apoiando a seleção de recursos humanos nas esferas pública e privada. Mais especificamente, MDFCP visa ser aplicável e útil aos gestores de concursos públicos e vestibulares. Além do MDFCP, esta Dissertação apresentou as seguintes contribuições:

1) Artigo Científico de Revisão Sistemática de Literatura, intitulado "*Methods for Detecting Fraud in Civil and Military Service Examinations: a systematic mapping*" (Nunes, Bonacin, *et al*., 2021), apresentado e publicado na *International Conference on Information Technology-New Generations* (ITNG 2021) [QUALIS A4].

> *Abstract: Civil and military service examinations are carried out in several countries for the recruitment and admission of public servants in various*

spheres/levels of government. This is considered an effective and rational method for selection based on merit. Due to the constant economic variations and the stability provided by public offices, the interest in some offered positions can be huge. Criminals specialized in defrauding public examinations offer candidates the possibility of facilitated and illegal admission. Various types of information could be submitted to methods and techniques (e.g., application data, test performance, geodata, etc.) to detect fraud. We present a systematic mapping of the literature on fraud detection methods in several domains, which can be adapted and improved to detect fraud in public examinations. 31 articles were identified, and after analysis, 19 selected works were analyzed and classified. The usages of machine learning and data mining techniques were uppermost methods adopted in the analyzed papers. This work is aimed at researchers who seek to develop fraud detection techniques in admission exams.

2) Artigo Científico da Proposta Conceitual, intitulado "*A Conceptual Proposal of a Hybrid Method for Detecting Fraud in Civil and Military Service Entrance Examinations"* (Nunes, Jino, *et al*., 2021), apresentado e publicado na 18th ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2021 [QUALIS B1].

Abstract: A public service examination is an effective manner of selecting civil and military servants for admission to various sectors of public service. These examinations usually attract well trained personnel by presenting highly competitive and meritocratic selection. Nevertheless, they also attract criminals that offer candidates the possibility of easy and illegitimate admission. Aiming to provide better security and trust in the selection process, we propose a conceptual method that uses data mining and statistical techniques to detect fraud in public service examinations. The method analyzes geographic and biographical data of candidates, scores, and similarity of answers of one candidate to those of other candidates, as well as past information, to compose a comprehensive index of suspicion. When the value of the index is above a certain limit, it indicates a degree of suspicion of an approved candidate, which may call for further investigation.

3) Artigo da Proposta de Projeto de Pesquisa, intitulado "*Research Project: A Method based on Profile Signature and Statistical Parameters for Detecting Fraud in Civil and Military Service Examinations*", apresentado e publicado no XVII Workshop de Computação do Unifaccamp (WCF 2021).

> *Abstract: We present a research project aiming at detecting fraud in civil and military service examinations. A method based on profile signature and statistical parameters is proposed. Our conceptual proposal considers statistical, historical, geographical, social, and behavioral parameters. A suspicion signature, generated for each candidate, is converted into a suspicion index to provide a profile ranking of candidates.*

4) Registro de Programa de Computador, intitulado "*Programa para análise de respostas em questões de múltipla escolha*", registrado no INPI sob o código *512022000015-6.*

7.2 Trabalhos Futuros

Como trabalhos futuros, propõe-se aprimorar e ampliar a abrangência do método nos seguintes aspectos:

(1) Comparação de desempenho.

O MDFCP já engloba análises variadas. A adição de uma análise, sobre a comparação do desempenho, por exemplo entre disciplinas similares do concurso avaliado e um concurso externo, pode tornar o método ainda mais abrangente e mais efetivo.

(2) Aprimorar Análise de RQME.

Na busca por aprimorar a Análise de RQME, espera-se levar em consideração os grupos delimitados por disciplinas, ou o conjunto de questões relacionadas, na geração do índice de similaridade das respostas (γ).

(3) Analisar registros anteriores.

Como a funcionalidade de retroalimentação de dados no MDFCP não foi totalmente explorada nos testes, a utilização de bancos de dados externos de suspeição, de uma maior de quantidade de listas de suspeição e de tabelas de similaridade geradas

com base em várias ocorrências anteriores de um concurso deve possibilitar um maior peso nessa análise.

(4) Apresentação dos resultados

Como as informações produzidas pelo método são numéricas e relacionais, há espaço para se trabalhar os dados e produzir informações visuais, através do desenvolvimento de uma interface gráfica. Dentro dessa perspectiva, pode-se oferecer uma visão georreferenciada, apontando conexões entre candidatos com altos níveis de suspeita.

BIBLIOGRAFIA

Alnajem, A. A. I. & Zhang, N. (2013). A copula-based fraud detection (CFD) method for detecting evasive fraud patterns in a corporate mobile banking context. *2013 International Conference on IT Convergence and Security, ICITCS 2013*, 0–3. https://doi.org/10.1109/ICITCS.2013.6717772

Barbosa, A. L. N. de H. & Souza, P. H. G. F. de. (2012). *Diferencial salarial Público-Privado e desigualdade dos rendimentos do trabalho no Brasil*.

Barbosa, L. (2014). Meritocracia à brasileira: o que é desempenho no Brasil? *Revista Do Serviço Público*, *47*(3), 58–102. https://doi.org/10.21874/rsp.v47i3.396

Belo, O., Mota, G. & Fernandes, J. (2016). *A Signature Based Method for Fraud Detection on E-Commerce Scenarios* (Issue January, pp. 497–506). https://doi.org/10.1007/978-3-319-25226-1_45

Bhati, P. & Sharma, M. (2015). *Credit Card Number Fraud Detection Using K-Means with Hidden Markov Method Transition probabilities*. *2*(3), 104–108.

Bouguessa, M. (2012). Modeling *Outlier* Score Distributions. In *Advanced Data Mining and Applications* (pp. 713–737). Springer.

Breiman, L. R. (2001). Random Forests. *Machine Learning 45*, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. (2000). LOF: Identifying Density-Based Local *Outlier*s. *ACM SIGMOD Record*, *29*(2), 93–104. https://doi.org/10.1145/335191.335388

Campos, G. O. (2015). *Estudo, avaliação e comparação de técnicas de detecção não supervisionada de outliers*. 67.

Carvalho, R. N., Laskey, K. B., Costa, P. C. G., Ladeira, M., Santos, L. L. & Matsumoto, S. (2009). Probabilistic ontology and knowledge fusion for procurement fraud detection in Brazil. *CEUR Workshop Proceedings*, *527*, 3–14. https://doi.org/10.1007/978-3-642-35975-0_2

Cavalcanti, E. R., Pires, C. E., Cavalcanti, E. P. & Pires, V. F. (2012). Detection and evaluation of cheating on college exams using supervised classification.

Informatics in Education, *11*(2), 169–190. https://doi.org/10.15388/infedu.2012.09

Chang, J. S. & Chang, W. H. (2012). A cost-effective method for early fraud detection in *online* auctions. *International Conference on ICT and Knowledge Engineering*, 182–188. https://doi.org/10.1109/ICTKE.2012.6408551

Chen, L., Zhang, Z., Liu, Q., Yang, L., Meng, Y. & Wang, P. (2019). A method for *online* transaction fraud detection based on individual behavior. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3321408.3326647

Chen, M. (2017). Detect multiple choice exam cheating pattern by applying multivariate statistics. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, *2017*(OCT), 173–181.

Chen, Y. J. & Wu, C. H. (2017). On Big Data-Based Fraud Detection Method for Financial Statements of Business Groups. *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, 986–987. https://doi.org/10.1109/IIAI-AAI.2017.13

Eshghi, A. & Kargari, M. (2019). Introducing a Method for Combining Supervised and Semi-Supervised Methods in Fraud Detection. *Proceedings of 2019 15th Iran International Industrial Engineering Conference, IIIEC 2019*, 23–30. https://doi.org/10.1109/IIIEC.2019.8720642

Frary, R. B., Tideman, T. N. & Watts, T. M. (1977). Indices of Cheating on Multiple-Choice Tests. *Journal of Educational Statistics*, *2*(4), 235–256. https://doi.org/10.3102/10769986002004235

Freitas, I. W. S. de. (2019). *Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados*.

Guedes, A., Portes, B., Teles, J., Silveira, L., Ferreira, U., dos Santos, M., Elgaly, P. & Fonseca, R. (2021). *Funcionalismo público no Brasil: Análise dos dados nas últimas três décadas*. www.ipea.gov.br

Kitchenham, B. (2004). Procedures for Performing Systematic Literature Reviews. *Joint Technical Report, Keele University TR/SE-0401 and NICTA TR-0400011T.1*, *33*(TR/SE-0401), 33.

Larose, D. T. & Larose, C. D. (2015). Logistic Regression. In *Data Mining and Predictive Analytics* (2nd ed., pp. 359–413). John Wiley & Sons.

Li, Q. & Xie, Y. (2019). A behavior-cluster based imbalanced classification method for credit card fraud detection. *ACM International Conference Proceeding Series*, 134–139. https://doi.org/10.1145/3352411.3352433

Ma, T., Qian, S., Cao, J., Xue, G., Yu, J., Zhu, Y. & Li, M. (2019). An unsupervised incremental virtual learning method for financial fraud detection. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2019-Novem*. https://doi.org/10.1109/AICCSA47632.2019.9035259

Niu, Z., Shi, S., Sun, J. & He, X. (2011). A Survey of *Outlier* Detection Methodologies and Their Applications. In *Artificial Intelligence and Computational Intelligence* (pp. 380–387). Springer-Verlag.

Nunes, R. P. M., Bonacin, R. & de Franco Rosa, F. (2021). Methods for Detecting Fraud in Civil and Military Service Examinations: A Systematic Mapping. *ITNG 2021 18th International Conference on Information Technology-New Generations*, 203–208. https://doi.org/10.1007/978-3-030-70416-2_26

Nunes, R. P. M., Jino, M., Bonacin, R. & Rosa, F. de F. (2021). A Conceptual Proposal of a Hybrid Method for Detecting Fraud in Civil and Military Service Entrance Examinations. *18th ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2021*, *1-8*. https://doi.org/10.1109/AICCSA53542.2021.9686932.

Nunes, R. P. M. & Rosa, F. de F. (2021). *Programa de Análise de Respostas às Questões de Múltipla Escolha / MDFCP* (1.7). https://github.com/rpmnunes/MDFCP-Answer-Analyze

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V;Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ralha, C. G. & Silva, C. V. S. (2012). A multi-agent data mining system for cartel detection in Brazilian government procurement. *Expert Systems with Applications*, *39*(14), 11642–11656. https://doi.org/10.1016/j.eswa.2012.04.037

Salazar, A., Safont, G. & Vergara, L. (2019). A new method for fraud detection in credit cards based on transaction dynamics in subspaces. *Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019*, 722–725. https://doi.org/10.1109/CSCI49370.2019.00137

Singh, K. & Upadhyaya, S. (2012). *Outlier Detection: Applications And Techniques*. www.IJCSI.org

Tan, P.-N., Steinbach, M. & Kumar, V. (2013). Anomaly Detection. In *Introduction to Data Mining* (1ª Ed, pp. 651–683). Pearson Education.

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909–921. https://doi.org/10.1080/02664760050120588

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J. & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Xie, Y., Liu, G., Cao, R., Li, Z., Yan, C. & Jiang, C. (2019). A Feature Extraction Method for Credit Card Fraud Detection. *Proceedings - 2019 2nd International Conference on Intelligent Autonomous Systems, ICoIAS 2019*, 70–75. https://doi.org/10.1109/ICoIAS.2019.00019

Yang, W., Zhang, Y., Ye, K. & Li, L. (2019). FFD: A Federated Learning Based Method for Credit Card Fraud Detection. In *Chen K., Seshadri S., Zhang LJ. (eds) Big Data – BigData 2019. Lecture Notes in Computer Science* (Vol. 11514, pp. 18–32). https://doi.org/10.1007/978-3-030-23551-2

Zhang, Z., Chen, L., Liu, Q. & Wang, P. (2020). A Fraud Detection Method for Low-Frequency Transaction. *IEEE Access*, *8*, 25210–25220. https://doi.org/10.1109/ACCESS.2020.2970614

Zhao, P., Fu, X., Wu, W., Li, D. & Li, J. (2019). Network-Based Feature Extraction Method for Fraud Detection via Label Propagation. *2019 IEEE International Conference on Big Data and Smart Computing, BigComp 2019 - Proceedings*, *1*. https://doi.org/10.1109/BIGCOMP.2019.8679414

APÊNDICE 1 – ALGORITMO DE CÁLCULO DO IS

Algoritmo 1: Cálculo do Índice de Suspeição (CIS)

Requer: '{} indica várias ocorrências
CA: conjunto de dados de exames anteriores
B_n: candidato.dados biográficos
G_n: candidato.dados geográficos
Tt: candidato.tempo de teste
Perfil_RL: modelo treinado para candidato aproveitado (CAp)
$\{N_1...N_k\},\{N_J...N_L\}$: notas de candidatos aproveitados (CAp), agrupadas por afinidade
$\{RC_q\}$: índice de acerto das questões
$\{r_1, ..., r_n\}$: respostas dos candidatos
LSi: Limiar de similaridade
LS: Limiar de Suspeição
{Pn}: Pesos das análises
{RCq}: respostas corretas das questões (gabarito)
r1, r2, r3: Registros em BD de detecções anteriores/black list

Para cada CandAv em ListaCandidatos

// (1) Análise de perfil: enviando dados para treinar modelo.
α ← (1 – PerfilRL[Bn, Gn, tt])

// (2) Análise de notas outliers
QNo ← Σ (*Outlier* {gruposnotas}
Se QNo = 0 então β ← 0; senão β ← (2QSO + Ad) / 2QD

// (3) Análise de Respostas
// 3a. Calcule a relação IDQ Certas/Erradas. Valor ajustado [0..1]
γ_2 = RelaçãoIDQCertasxErradas(respostasCandAv, RCq)
// 3b. Calcule a maior similaridade comparada com CandAv
Para cada CandB no conjunto de candidatos aproveitados < > CandAv
γ_1← Analisar similaridade (Respostas CandAv, Respostas CandB, RCq)
Se $\gamma_1' > \gamma_1$ **então** $\gamma_1 \leftarrow \gamma_1'$
// Registra semelhanças maiores LSi em BD de similaridades
Se γ_1' >= SiT, **então** adicione (γ_1',CandAv,CandB) no BD Similaridades
FimPara
γ ← maior(γ_1, γ_2)

// (4) Análise de Registros. Para cada base com reg. CandAv, adicionar 1.
Para cada candAv em (base1,...,baseb)
n+=1
FimPara
δ ← n / b
// Cálculo do Índice de Suspeição (ϕ)
ϕ = (α.wα + β.wβ + γ.wγ + δ.wδ) / (wα + wβ + wγ + wδ)
// Lista de Suspeição: ϕ maior que o limite de suspeição (LS), adicionar à lista
Se ϕ >= ST **então**
Add CandAv, ϕ to Suspicious List
FimSe
FimPara

APÊNDICE 2 – DADOS SUBMETIDOS NA ANÁLISE DE PERFIL

Tabela A2.1 – Extrato da Tabela de Treinamento (A-1)

APROVEI-TADO	REALI-ZAÇÕES	LOCAL PROVA	IDADE	TIPO ESCOLA	CIDADE
SIM	0	RIO DE JANEIRO-RJ	19	PRIVADO	RIO DE JANEIRO-RJ
SIM	0	RIO DE JANEIRO-RJ	17	PRIVADO	OUTRAS
SIM	0	SAO PAULO-SP	17	PRIVADO	SAO PAULO-SP
SIM	1	GOIANIA-GO	18	PUBLICO COM RESTRICAO	GOIANIA-GO
SIM	1	FORTALEZA-CE	18	PUBLICO COM RESTRICAO	OUTRAS
SIM	0	BELO HORIZONTE-MG	18	PUBLICO COM RESTRICAO	OUTRAS
SIM	1	BRASILIA-DF	18	PUBLICO COM RESTRICAO	BRASILIA-DF
SIM	0	TAUBATE-SP	17	PRIVADO	TAUBATE-SP
SIM	1	TERESINA-PI	17	PRIVADO	TERESINA-PI
SIM	0	JUIZ DE FORA-MG	18	PUBLICO COM RESTRICAO	OUTRAS
SIM	1	RIO DE JANEIRO-RJ	18	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ
SIM	1	CASCAVEL-PR	17	PUBLICO	OUTRAS
SIM	1	RESENDE-RJ	19	PUBLICO	RESENDE-RJ
SIM	1	RIO DE JANEIRO-RJ	18	PRIVADO	RIO DE JANEIRO-RJ
NAO	0	SALVADOR-BA	18	PUBLICO COM RESTRICAO	SALVADOR-BA
NAO	0	CAMPINAS-SP	21	PRIVADO	PIRASSUNUNGA-SP
NAO	0	RIO DE JANEIRO-RJ	17	PRIVADO	RIO DE JANEIRO-RJ
NAO	0	RIO DE JANEIRO-RJ	18	PRIVADO	DUQUE DE CAXIAS-RJ
NAO	1	SAO PAULO-SP	18	PUBLICO	CAMPINAS-SP
NAO	0	TAUBATE-SP	17	PRIVADO	LORENA-SP
NAO	0	PORTO VELHO-RO	16	PRIVADO	PORTO VELHO-RO
NAO	0	SAO PAULO-SP	18	PUBLICO	SAO PAULO-SP
NAO	0	SAO PAULO-SP	19	PUBLICO	OSASCO-SP
NAO	0	ARACAJU-SE	20	PUBLICO	ARACAJU-SE
NAO	0	CAMPINAS-SP	19	PUBLICO	OUTRAS
NAO	0	RESENDE-RJ	17	PUBLICO	OUTRAS
NAO	0	MANAUS-AM	17	PUBLICO	MANAUS-AM
NAO	0	RIO DE JANEIRO-RJ	19	PUBLICO COM RESTRICAO	SAO GONCALO-RJ
NAO	0	SAO PAULO-SP	17	PUBLICO	SANTOS-SP
NAO	1	BOA VISTA-RR	20	PUBLICO	BOA VISTA-RR
NAO	0	CASCAVEL-PR	18	PUBLICO	OUTRAS

Tabela A2.2 – Amostra dos dados de teste (A)

IDCAND	REALI-ZAÇÕES	LOCAL PROVA	IDADE	TIPO ESCOLA	CIDADE
4079E66D9	0	RIO DE JANEIRO-RJ	19	PUBLICO	RIO DE JANEIRO-RJ
10A753A15	1	BELO HORIZONTE-MG	17	PUBLICO	OUTRAS
9A62BA534	1	SAO PAULO-SP	18	PUBLICO	SAO PAULO-SP
14211AB1E7	1	SALVADOR-BA	18	PUBLICO COM RESTRICAO	SALVADOR-BA
1BC7F9AA7	0	SAO PAULO-SP	21	PRIVADO	SAO PAULO-SP
2E753D20E	0	BRASILIA-DF	18	PUBLICO COM RESTRICAO	BRASILIA-DF
D26EC9E5	0	PORTO ALEGRE-RS	18	PRIVADO	PORTO ALEGRE-RS
74F61E1E	0	BELO HORIZONTE-MG	17	PRIVADO	BELO HORIZONTE-MG
12D09EB46	1	PORTO ALEGRE-RS	19	PUBLICO COM RESTRICAO	PORTO ALEGRE-RS
1EE1AF513	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE
479EDF39C	1	RIO DE JANEIRO-RJ	17	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ
37D768AA9	0	TRES CORACOES-MG	18	PRIVADO	TRES CORACOES-MG
9102230F	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE
AAF00BBB5	0	LINS-SP	16	PRIVADO	OUTRAS
31956EFAF	1	RIO DE JANEIRO-RJ	21	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ
1B31AFBB4	0	FORTALEZA-CE	17	PRIVADO	FORTALEZA-CE
1C75B822C	2	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE
3AD46E645	0	RIO DE JANEIRO-RJ	17	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ
134E50497	1	RECIFE-PE	17	PUBLICO COM RESTRICAO	RECIFE-PE
2AD604B07	1	JUIZ DE FORA-MG	19	PUBLICO	BARBACENA-MG
36FDEF1B1	0	RIO DE JANEIRO-RJ	18	PRIVADO	RIO DE JANEIRO-RJ
2C92B76D6	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE
B979E050E	1	TAUBATE-SP	17	PRIVADO	TAUBATE-SP
1A2C309C4	0	TERESINA-PI	18	PRIVADO	TERESINA-PI
C8AD95B3	2	FORTALEZA-CE	21	PRIVADO	FORTALEZA-CE
4667921BF	1	RIO DE JANEIRO-RJ	17	PUBLICO	SAO JOAO DE MERITI-RJ
13D777895	0	CAMPO GRANDE-MS	19	PRIVADO	OUTRAS
A18B5EBC	0	FORTALEZA-CE	19	PRIVADO	ARACAJU-SE
E7F5EB92	1	ARACAJU-SE	18	PRIVADO	ARACAJU-SE
261E1AA79	0	RESENDE-RJ	18	PRIVADO	RESENDE-RJ
111135776	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE
9B4BE4A9	0	SAO PAULO-SP	21	PRIVADO	RIO DE JANEIRO-RJ
AF51CDAA9	1	RESENDE-RJ	17	PRIVADO	LORENA-SP
3C1BDE5E4	2	RESENDE-RJ	20	PRIVADO	RESENDE-RJ
20901C6EE	2	BELEM-PA	21	PRIVADO	RIO DE JANEIRO-RJ
14DCC845D	1	CAMPO GRANDE-MS	18	PUBLICO COM RESTRICAO	CAMPO GRANDE-MS

Tabela A2.3 – Resultados da Regressão Logística ("A")

IDCAND	REALI-ZAÇÕES	LOCAL PROVA	IDADE	TIPO ESCOLA	CIDADE	APROVEI-TADO	LR [20](Rótulo)	LR (SIM)
4079E66D9	0	RIO DE JANEIRO-RJ	19	PUBLICO	RIO DE JANEIRO-RJ	SIM	NAO	0,221273773
10A753A15	1	BELO HORIZONTE-MG	17	PUBLICO	OUTRAS	SIM	NAO	0,44754627
9A62BA534	1	SAO PAULO-SP	18	PUBLICO	SAO PAULO-SP	SIM	SIM	0,531788533
14211AB1E7	1	SALVADOR-BA	18	PUBLICO COM RESTRICAO	SALVADOR-BA	SIM	SIM	0,822737199
1BC7F9AA7	0	SAO PAULO-SP	21	PRIVADO	SAO PAULO-SP	SIM	NAO	0,441926576
2E753D20E	0	BRASILIA-DF	18	PUBLICO COM RESTRICAO	BRASILIA-DF	SIM	NAO	0,490501532
D26EC9E5	0	PORTO ALEGRE-RS	18	PRIVADO	PORTO ALEGRE-RS	SIM	NAO	0,450163149
74F61E1E	0	BELO HORIZONTE-MG	17	PRIVADO	BELO HORIZONTE-MG	SIM	NAO	0,452914984
12D09EB46	1	PORTO ALEGRE-RS	19	PUBLICO COM RESTRICAO	PORTO ALEGRE-RS	SIM	SIM	0,821110838
1EE1AF513	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE	SIM	SIM	0,756058847
479EDF39C	1	RIO DE JANEIRO-RJ	17	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ	SIM	SIM	0,851275999
37D768AA9	0	TRES CORACOES-MG	18	PRIVADO	TRES CORACOES-MG	SIM	NAO	0,450163149
9102230F	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE	SIM	SIM	0,756058847
AAF00BBB5	0	LINS-SP	16	PRIVADO	OUTRAS	SIM	NAO	0,371259951
31956EFAF	1	RIO DE JANEIRO-RJ	21	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ	SIM	SIM	0,845560481
1B31AFBB4	0	FORTALEZA-CE	17	PRIVADO	FORTALEZA-CE	SIM	SIM	0,758102395
1C75B822C	2	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE	SIM	SIM	0,986308527
3AD46E645	0	RIO DE JANEIRO-RJ	17	PUBLICO COM RESTRICAO	RIO DE JANEIRO-RJ	SIM	SIM	0,542805526
134E50497	1	RECIFE-PE	17	PUBLICO COM RESTRICAO	RECIFE-PE	SIM	SIM	0,824351937
2AD604B07	1	JUIZ DE FORA-MG	19	PUBLICO	BARBACENA-MG	SIM	SIM	0,529020874
36FDEF1B1	0	RIO DE JANEIRO-RJ	18	PRIVADO	RIO DE JANEIRO-RJ	SIM	NAO	0,499629356
2C92B76D6	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE	SIM	SIM	0,756058847
B979E050E	1	TAUBATE-SP	17	PRIVADO	TAUBATE-SP	SIM	SIM	0,822525945
1A2C309C4	0	TERESINA-PI	18	PRIVADO	TERESINA-PI	SIM	NAO	0,450163149
C8AD95B3	2	FORTALEZA-CE	21	PRIVADO	FORTALEZA-CE	SIM	SIM	0,985850992
4667921BF	1	RIO DE JANEIRO-RJ	17	PUBLICO	SAO JOAO DE MERITI-RJ	SIM	SIM	0,534554237
13D777895	0	CAMPO GRANDE-MS	19	PRIVADO	OUTRAS	SIM	NAO	0,363512602
A18B5EBC	0	FORTALEZA-CE	19	PRIVADO	ARACAJU-SE	SIM	SIM	0,751965445
E7F5EB92	1	ARACAJU-SE	18	PRIVADO	ARACAJU-SE	SIM	SIM	0,797862861
261E1AA79	0	RESENDE-RJ	18	PRIVADO	RESENDE-RJ	SIM	NAO	0,450163149
111135776	0	FORTALEZA-CE	18	PRIVADO	FORTALEZA-CE	SIM	SIM	0,756058847

[20] Coluna "Aproveitado" original da tabela de validação e coluna LR (Rótulo) com o resultado da predição baseada em RL.

APÊNDICE 3 – DADOS OBTIDOS DA ANÁLISE DE NOTAS

Tabela A3.1 – Amostra de valores outliers (“A-2”)

IdCand	Exatas 1	Exatas 2	Exatas 3	Humanas 1	Humanas 2	Linguagem 1	Linguagem 2
3FE645374	90.000	100.000	91.667	---	---	---	---
400161689	90.000	91.667	91.667	---	---	---	---
3CD81AD8C	90.000	100.000	91.667	---	---	---	---
3A9785DB5	95.000	83.333	83.333	---	---	---	---
30B78B02B	90.000	91.667	91.667	---	---	---	---
3E81B10C9	90.000	100.000	91.667	---	---	---	---
1C5439AAB	90.000	91.667	91.667	---	---	---	---
E4FABB3F	75.000	33.333	58.333	---	---	---	---
311FC71CC	60.000	25.000	91.667	---	---	---	---
E691BE297	60.000	66.667	58.333	---	---	---	---
346A50CC5	60.000	66.667	58.333	---	---	---	---
3E246D910	30.000	83.333	83.333	---	---	---	---
4273FB15D	60.000	66.667	58.333	---	---	---	---
212B1ECA	---	---	---	83.333	91.667	---	---
6EBF612B	---	---	---	91.667	75.000	---	---
249997F4E	---	---	---	91.667	75.000	---	---
198A47E26	---	---	---	83.333	91.667	---	---
457647D7F	---	---	---	58.333	91.667	---	---
137F6D75E	---	---	---	58.333	91.667	---	---
16EBEC14E	---	---	---	75.000	41.667	---	---
2B5C255E4	---	---	---	33.333	91.667	---	---
254C89074	---	---	---	33.333	83.333	---	---
A8ED50CF	---	---	---	58.333	91.667	---	---
3790BB32E	---	---	---	33.333	83.333	---	---
2B7412131	---	---	---	58.333	91.667	---	---
328A715D7	---	---	---	---	---	80.000	58.333
4A1B56D1	---	---	---	---	---	80.000	83.333
186624322	---	---	---	---	---	55.000	41.667
2E6BE7B6C	---	---	---	---	---	55.000	41.667
72242378	---	---	---	---	---	80.000	50.000
80939D6BC	---	---	---	---	---	55.000	41.667
39744128D	---	---	---	---	---	80.000	58.333
D991FC7C	---	---	---	---	---	40.000	83.333
195FD1477	---	---	---	---	---	40.000	41.667
3BE24C145	---	---	---	---	---	40.000	41.667
329BD77F6	---	---	---	---	---	55.000	41.667
38B45032D	---	---	---	---	---	55.000	41.667

APÊNDICE 4 – PROGRAMA DE APOIO À ANÁLISE DE RQME

Foi desenvolvido um *software* para o apoio aos testes prévios e a validação da Análise de RQME. Este programa serviu para os testes iniciais, contemplando as diferentes abordagens testadas, como a comparação entre respostas de dois candidatos para obtenção de um índice de similaridade (1:n) e a observação isolada das respostas de um candidato (1:0), em busca de disparidade na relação da soma do IDQ das respostas corretas e respostas erradas. Nessas abordagens foram testadas as variações de cálculo do IDQ, considerando todo o universo de candidatos ou somente aqueles aproveitados. Também foi testado, especificamente na abordagem 1:n, a utilização ou não de um redutor de disparidade de desempenho. Esse redutor de desempenho poderia ser calculado com base em diferença de estratos de desempenho ou em um peso relativo à diferença entre notas de dois candidatos. Essas variações testadas e as que, ao final, foram utilizadas no método, estão explicitadas no Apêndice 5.

Na aplicação final do método, sobre as três ocorrências do concurso, o programa foi utilizado para o cálculo dos índices correspondentes às duas abordagens previstas na Análise de RQME.

A solução foi implementada em .Net e banco de dados Sql Server. O programa se mostrou bastante eficiente, realizando quase meio bilhão de relações por execução e oferecendo a possibilidade de comparação dos diversos índices gerados (Tabela A4.1), relativos a cada uma das variações, no intuito de identificar os mais eficientes.

O código-fonte do protótipo de software que implementa partes do algoritmo e uma interface gráfica (Figura A4.1) e a estrutura de dados (Figura A4.2), estão disponíveis no repositório Github (Nunes & Rosa, 2021).

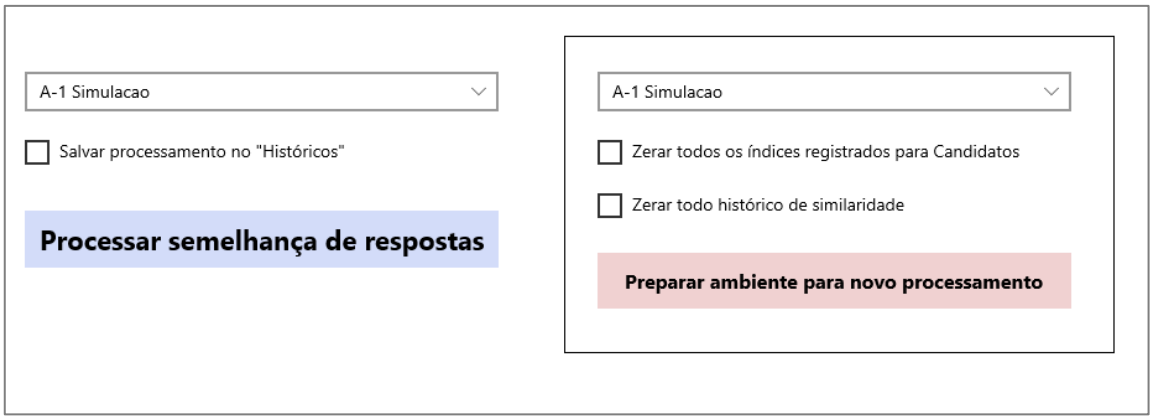

Figura A4.1 – Tela de exemplo da Interface Gráfica do Protótipo

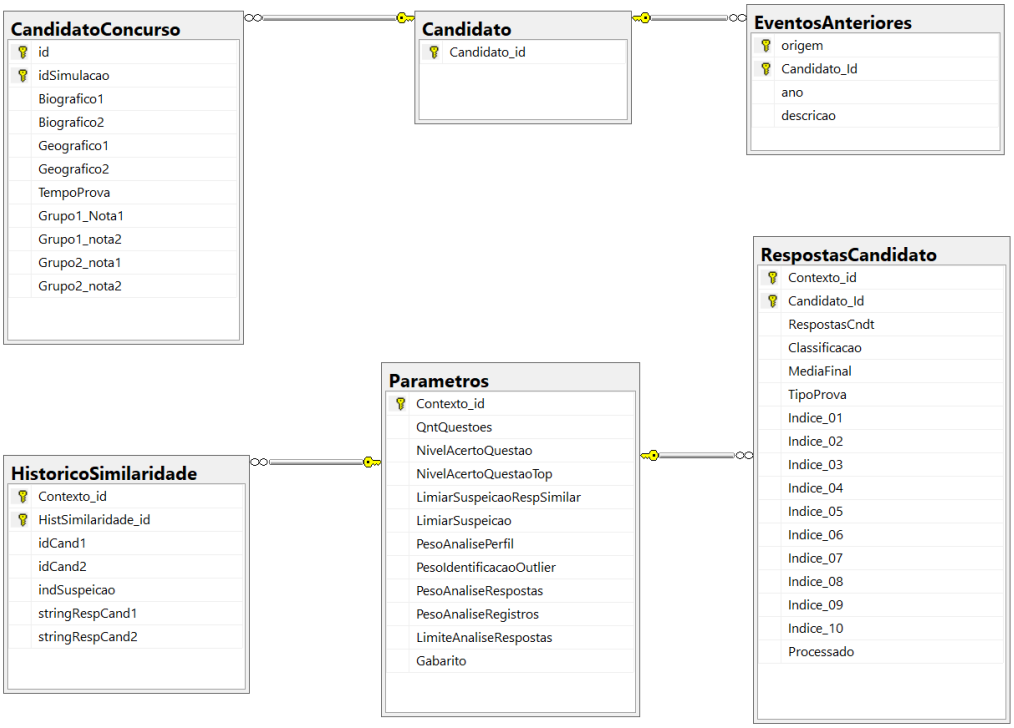

Figura A4.2 – Modelagem dos dados utilizados no MDFCP

Tabela A4.1 – Amostra de resultado da execução do programa de apoio ao cálculo do índice relacionado à Análise de RQME

Id	Class	Nota	Ind01[21]	Ind02	Ind03	Ind04	Ind05	Ind06	Ind07	Ind08
4079E66D9	1	94,2590	0,9306	0,4303	0,9828	0,9468	0,9423	0,8890	0,9353	0,8872
10A753A15	2	94,0740	0,9600	0,5965	0,9859	0,9585	0,9373	0,8748	0,9373	0,8748
9A62BA534	3	94,0740	0,8301	0,4050	0,9955	0,9455	0,9470	0,9148	0,9464	0,9148
14211AB1E7	4	92,2220	0,8524	0,4525	0,9786	0,9164	0,9527	0,9103	0,9321	0,8924
1BC7F9AA7	5	90,7410	0,9202	0,4330	0,9876	0,9420	0,9475	0,8870	0,9419	0,8983
2E753D20E	6	90,1850	0,8640	0,4227	0,9779	0,9383	0,9430	0,8969	0,9374	0,9019
D26EC9E5	7	90,1850	0,8335	0,3695	1,0031	0,9679	0,9589	0,9288	0,9533	0,9280
74F61E1E	8	89,0740	0,8867	0,5705	0,9575	0,9490	0,9155	0,8221	0,8993	0,8591
12D09EB46	9	89,0740	0,8701	0,4632	0,9613	0,9163	0,9269	0,8755	0,9120	0,8651
1EE1AF513	10	89,0740	0,8676	0,4370	0,9668	0,9367	0,9454	0,8956	0,9379	0,8885
479EDF39C	11	88,7040	0,8762	0,5130	0,9732	0,9597	0,9285	0,8683	0,9195	0,9050

[21] Dos índices testados, foram aproveitados o índice 01 (abordagem de certas/erradas) e o índice 05 (abordagem da similaridade entre respostas de candidatos)

APÊNDICE 5 – ÍNDICES TESTADOS NA ANÁLISE DE RQME

Tabela A5.1 – Variações dos índices testados no software de apoio à Análise de RQME

Indicador	Compara[22]	Contempla[23]	Observações
índice_01	1:0	IDQ Geral	Eficiente. **Abordagem 1, selecionada na análise das RQME.**
índice_02	1:0	IDQ Aproveitados	Menor diferenciação do IDQ tornou a identificação de anormalidades pouco eficiente
índice_03	1:n	IDQ Geral, sem redutor	Sem redutor, a comparação entre candidatos de desempenhos diferenciados gerou uma quantidade excessiva de índices altos de similaridade (falso positivo)
índice_04	1:n	IDQ Aproveitados, sem redutor	
índice_05	1:n	IDQ Geral, redutor / estratificado	Eficiente. **Abordagem 2, selecionada na análise das RQME.**
índice_06	1:n	IDQ Aproveitados, redutor /estratificado	Menor diferenciação do IDQ tornou a identificação de anormalidades pouco eficiente
índice_07	1:n	IDQ Geral, redutor / diferença nota	Eficiente.
índice_08	1:n	IDQ Aproveitados, redutor/diferença nota	Menor diferenciação do IDQ tornou a identificação de anormalidades pouco eficiente

[22] 1:0 representa a comparação das respostas de um candidato isoladamente; 1:n representa a comparação das respostas de um candidato com as respostas de todos os outros candidatos em avaliação pelo método.

[23] IDQ Geral e aproveitados considera para a geração do IDQ, respectivamente, as respostas de todos os candidatos em uma determinada questão ou somente as respostas de candidatos aproveitados.

APÊNDICE 6 – DESCRIÇÃO DOS DADOS UTILIZADOS PELO MDFCP

Grupo	Dado	Tipo	Observação
Biográfico	Idade	Numérico discreto	---
	Tipo de Escola	Categórico nominal	Pública; Pública com restrição; Privada
	Recorrência	Numérico discreto	Quantidade de vezes que realizou o concurso
Geográficos	Cidade de Residência	Categórico	90 principais e "outras cidades"
	UF de residência	Categórico	---
	Local de realização da prova	Categórico	---
RQME	Respostas	Categórico nominal	{A,B,C,D,E,branco}
	Gabarito	Categórico nominal	{A,B,C,D,E}
	Nível de acerto de questões	Numérico contínuo	[0..1]
Notas	Exatas 1	Numérico contínuo	[0..100]
	Exatas 2	Numérico contínuo	[0..100]
	Exatas 3	Numérico contínuo	[0..100]
	Humanas 1	Numérico contínuo	[0..100]
	Humanas 2	Numérico contínuo	[0..100]
	Linguagem 1	Numérico contínuo	[0..100]
	Linguagem 2	Numérico contínuo	[0..100]
Parâmetros	Limiar de Suspeição	Numérico contínuo	[0..1]
	Peso das Análises	Numérico discreto	[0..1]
	Limiar de Similaridade	Numérico contínuo	[0..1]