Determination of Dropout Student Profile Based on Correspondence Analysis Technique

T. Barros, I. Silva, *Member, IEEE*, and L. Affonso

***Abstract*—The purpose of this paper is to analize how the Correspondence Analysis Techinique can be used to enhance the Exploratory Data Analysis on the drop-out for a set of educational data with a majority of the categorical type. In order to do these analisys, the independence calculation was implemented from the chi-square test, and the heatmap and perceptual map were generated with the indexes on the socioeconomic data of students and academic performance in Portuguese and Mathematics subjects. As a result of this paper is presented the relations of attraction and repulsion between socioeconomic attributes and drop-out, and the profile of the student most vulnerable to evasion is drawn. Some characteristics are: students who have already failed at least once, who do not live with their parents, elementary school in public school, low level of education of the financial responsible. The studies were carry out using a real academic database of students of the secondary education with training in professional education through technical courses with duration of four years, in the face-to-face modality of the Federal Institute of Rio Grande do Norte (IFRN) in Brazil.**

***Index Terms*—Correspondence analysis, Education data mining, Data visualization.**

I. Introdução

MINERAÇÃO de Dados Educacionais, ou em inglês *Education Data Mining* (EDM), é definida como a interseção entre as grandes áreas de estatística, mineração de dados e educação [1]. Essa área está se tornando uma grande aliada dos professores e da gestão de institutos educacionais no auxílio de descoberta de novos conhecimentos e de novos padrões sobre dados de estudantes, subsidiando a tomada de decisão para os novos desafios da educação na era digital. Dentro do processo de descoberta do conhecimento, uma das fases que vem ganhando cada vez mais destaque é a análise exploratória de dados, em inglês *Exploratory data analysis* (EDA), a qual tem como objetivo principal entender os dados, a partir da sua distribuição, padrões e tendências, auxiliando na tomada de decisão de quais testes estatísticos serão apropriados para se usar [2]. A EDA une técnicas avançadas de visualização de dados e modelos estatísticos, as quais, explorando o sentido mais desenvolvido ao homem, a visão, utilizam-se de gráficos bem elaborados, permitindo realizar inferências e análise complexas, mesmo quando os gráficos estão baseados em estatísticas básicas [3].

Dentre as diversas aplicações da EDM, a predição de evasão vem ganhando destaque devido ao fato que cada aluno ou aluna que se evade da escola representa oportunidade de mudança de vida desperdiçada, menos mão-de-obra qualificada no mercado, menor chance de mobilidade social [4], principalmente em um continente com índices de desigualdades sociais elevados, como o caso da América Latina. Para efeito de ilustração e mensuração da relevância do problema, a evasão escolar no Brasil em 2010 foi de 11,4% (considerando-se os alunos que abandonaram o curso para o qual foram admitidos). Já em 2014, esse número chegou a 49% [5], em um país que apresenta o percentual de 75% de jovens de 20 a 24 anos de idade que não estudam, sendo o maior índice no mundo entre os países pesquisados pelo relatório Education at a Glance [6]. Em termos econômicos, estima-se que 7 bilhões de reais por ano é o valor investido em 1,9 milhão de jovens de 15 a 17 que abandonam o ensino médio antes do final do ano ou são reprovados ao final dele [4], valor equivalente ao custeio de todos institutos e universidades federais do país no ano de 2017 [7].

Diante desse cenário, ferramentas de mineração e visualização de dados podem auxiliar na descoberta de relações entre variáveis disponíveis para gestão (geralmente extraídas dos sistemas de controle acadêmico) e a evasão escolar, dando subsídios para melhores tomadas de decisões sobre o fenômeno da evasão. Entretanto, é importante enfatizar que, além das notas das disciplinas, a maioria das variáveis geralmente disponíveis nos sistemas de controle acadêmico são de natureza categóricas e nominais (como os dados demográficos e socioeconômicos). Esse tipo de variável não possui uma relação de ordem (maior ou menor) entre seus valores, por exemplo o atributo raça pode assumir os valores ”branco”, ”negro”, ”pardo”, ”amarelo”, em que não é possível definir uma relação de ordem entre eles. Para esses casos devem ser utilizadas técnicas de análise de independência, como a análise de correspondência, que medem se os valores de duas variáveis ocorrem juntos ou não. Situação diferente a de correlações clássicas, como a de *Pearson*, em que são geralmente utilizadas para dados numéricos (como notas de disciplina) para medir a variação de crescimento entre variáveis [8].

Diante do exposto acima, o objetivo principal do presente artigo é investigar como o uso de técnicas de visualização junto a técnica estatística de análise de correspondência podem identificar as principais características socioeconômicas e demográficas de alunos evadidos. Para propósito de validação do estudo, serão analisados dados educacionais de alunos dos cursos do Integrado atualizados em janeiro de 2018, na modalidade presencial, do Instituto Federal do Rio Grande do Norte (IFRN), Brasil.

Com o intuito de alcançar o objetivo acima, este artigo está dividido em mais cinco seções. Na Seção II, denominada de

T. M. Barros, Campus EaD, Instituto Federal do Rio Grande do Norte (IFRN), Natal, RN, Brasil, thiago.medeiros@ifrn.edu.br.

I. Silva, Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brasil, ivan@imd.ufrn.br.

L. A. Guedes, Departamento de Engenharia de Computação e Automação, Universidade Federal do Rio Grande do Norte (UFRN), Natal, RN, Brasil, affonso@dca.ufrn.br.

"Análise de Correspondência", é apresentada essa técnica estatística, o seu cálculo e os trabalhos relacionados, destacando-se as variáveis e as técnicas utilizadas. Na Seção III, denominada "Metodologia de Análise de Dados Utilizada", é apresentado o ambiente de desenvolvimento e a base utilizada. Na Seção IV, denominada "Estudo de Caso", são apresentados os principais gráficos e suas interpretações. Por fim, na Seção 6, denominada "Conclusão", é descrito o potencial dos gráficos utilizados e indicação de trabalhos futuros.

II. Técnica de Análise de Correspondência

A análise de correspondência (AC) é uma técnica exploratória em dados multivariados frequentemente utilizada para redução de dimensionalidade e mapeamento perceptual em base de dados composta por dados categóricos [8]. O objetivo é estabelecer a relação entre os valores nominais de duas variáveis categóricas disposta em uma tabela de contingência, a fim de descobrir uma explicação de baixa dimensão para possíveis desvios da independência dessas variáveis [9]. A análise de correspondência se destaca pela construção do mapa perceptual a partir da associação de objetos descritos pelos atributos selecionados. Sua aplicação principal é exibir a correspondência entre categorias em escalas nominais e permitir representar duas variáveis categóricas em um mesmo diagrama. É importante destacar que essa técnica também pode ser utilizada em variáveis com valores contínuos, desde que sejam discretizados. Para a realização da análise de correspondência é necessário o cálculo da tabela de contingência entre duas variáveis de escala nominal. A tabela de contingência representa a frequência conjunta entre os valores nominais de cada uma das variáveis. Após a criação da tabela de contingência, é realizado o cálculo do teste estatístico do qui-quadrado, a fim de padronizar os valores e gerar um índice de associação ou similaridade utilizado para criação do diagrama (mapa perceptual) [8].

A partir desse procedimento é gerado o mapa percentual de tal forma que quanto mais próximos estiverem os valores de dois atributos mais similares eles são.

1) Trabalhos Relacionados: Para o levantamento dos trabalhos relacionados foi realizado uma seleção de artigos do portal *Web of Science* utilizando a *String: ((("correspondence analysis") AND (dropout OR drop out OR drop-out)) OR (("análise de correspondência") AND evasão))* . Na primeira pesquisa foram retornados 14 trabalhos dos quais apenas 5 estão relacionados ao tema de educação.

No trabalho [10] a técnica de análise de correspondência é utilizada para relacionar os motivos de desistências do ensino médio extraídos de entrevistas dos alunos desistentes ou seus responsáveis, com característica demográficas extraídas da base de dados, em que é destacado a influência da região do aluno, indicando que a evasão é um fenômeno contextualizado. No trabalho [11] a técnica de análise de correspondência foi utilizada para obter a relação entre a participação em cursos *on-line* e as seguintes variáveis: nível de escolaridade, conhecimento da língua de sinais e número de membros no domicílio, entretanto os principais resultados obtidos no estudo estão dispostos em forma de tabelas, o que dificulta a sua interpretação para gestores e professores. Nos artigos [12] e [13], diferente do presente trabalho, foi utilizado a técnica análise de correspondência múltipla (uma derivação em que geralmente se analisa a distância de atributos de todas as variáveis de uma única vez, a partir da redução de dimensionalidade com perdas para apenas 2 componentes, gerando um mapa 2D), em que no primeiro foi utilizado para analisar padrões de matrícula e estudar a eficiência e conclusão entre os estudantes em programas com qualificações profissionais da Suécia, e no segundo dados socioeconômicos e demográficos da Universidade de Santander na Colômbia. Por último, no trabalho [14], é utilizado outra derivação da técnica de análise de correspondência, denominada de GCA (em inglês, *grade correspondence analysis*), em que pode ser aplicada a qualquer matriz de variáveis não negativas, levando a resultados bem interpretáveis, apesar das limitações teóricas, nesse caso sobre uma base de dados simulada de uma universidade.

Ao utilizar o portal *IEEE Explorer* com a *String* de pesquisa anteriormente definida, não foram encontrados trabalhos. Entretanto, foram encontrados artigos relacionados ao tema predição de evasão e desempenho na *IEEE Latin American* com destaque os artigos: [15] em que analisa o desempenho de alunos na área de matemática e ciências a partir da técnica de regressão linear múltipla; e no trabalho [16] a utilização da técnica de árvore de decisão para predição de evasão de alunos na Universidade Simón Bolívar.

III. Metodologia de Análise de Dados Utilizada

Os dados utilizados neste trabalho são de 8908 alunos do ensino Integrado (ensino médio com formação em educação profissional através de cursos técnicos com duração de quatro anos, na modalidade presencial) do Instituto Federal do Rio Grande do Norte (IFRN), localizado no nordeste do Brasil e distribuídos por 20 *Campi* (nas cidades de Apodi, Caicó, Canguaretama, Ceará-Mirim, Currais Novos, Ipanguaçu, João Câmara, Lajes, Macau, Mossoró, Natal-Central, Natal-Cidade Alta, Natal-Zona Norte, Nova Cruz, Parelhas, Parnamirim, Pau dos Ferros, Santa Cruz, São Gonçalo, São Paulo do Potengi). A base disponibilizada foi extraída do Sistema Unificado de Administração Pública, SUAP (suap.ifrn.edu.br), desenvolvido pelo próprio IFRN, e possui informações demográficas, caracterização socioeconômica e média final dos alunos nas disciplinas. A última atualização dos dados foi em janeiro de 2018. Os dados selecionados para o trabalho são todos os atributos demográficos e socioeconômicos disponíveis no SUAP e as notas das disciplinas de Português e de Matemática, uma vez que essas duas são disciplinas em comum de todos os cursos no 1º ano de ingresso dos alunos. Todas as variáveis utilizadas estão descritas na Tabela I.

O ambiente de desenvolvimento utilizado foi a linguagem de programação *Python* e os pacotes: *Pandas* [17], para manipulação dos dados; *Prince* [18], para criação do mapa perceptual da análise de correspondência; *Searborn* [19] e *Matplot* [20], para os gráficos.

Após a criação da base de dados e implementado o ambiente de desenvolvimento, foi realizado o cálculo de análise de

TABELA I
DESCRIÇÃO DAS VARIÁVEIS SELECIONADAS

Atributo	Descrição
LnguaPortuguesae LiteraturaI90H	Média de 0 a 10 da disciplina língua portuguesa
LnguaPortuguesae LiteraturaI90H Dependencia	Quantidade de dependências (repetição da disciplina devido a reprovação) do aluno na disciplina de língua portuguesa
LnguaPortuguesae LiteraturaI90Hfreq	Porcentagem de 0 a 100 da frequência na disciplina de língua portuguesa
MatemticaI120H	Média de 0 a 10 da disciplina de matemática
MatemticaI120H_ dependencia	Quantidade de dependência (repetição da disciplina devido a reprovação) do aluno na disciplina de matemática
MatemticaI120H_freq	Porcentagem de 0 a 100 da frequência na disciplina de matemática
aluno_exclusivo_ rede_publica	Verdadeiro se o aluno é exclusivo da rede pública durante todo o ensino fundamental
descricao_area_residencial	Área residencial do aluno: Urbana, Rural, Indígena, Quilombola, não informada
descricao_companhia_ domiciliar	Companhia domiciliar: cônjuge, mãe, pai, pais, outros, não informado, parente(s) ou amigo(s), sozinho(a)
descricao_estado_civil	Descrição do estado civil do aluno: casado(a), divorciado(a), não declarado, solteiro(a), união estável
descricao_historico	Qual curso técnico o aluno faz
descricao_imovel	Qual situação financeira do imóvel em que o aluno mora: alugado, cedido ou emprestado, financiado, não informado, outro, pensionato ou alojamento, próprio
descricao_mae_ escolaridade	Escolaridade da mãe do aluno: alfabetizado, fundamental completo, fundamental incompleto, médio completo, médio incompleto, superior completo, superior incompleto, não conhece, não estudou, pós graduação completo, pós graduação incompleto
descricao_pai_ escolaridade	Escolaridade do pai do aluno (mesmos valores do item anterior)
descricao_raca	Raça autodeclarada do aluno: amarela, branca, indígena, não declarado, parda, preta
descricao_responsavel_ escolaridade	Escolaridade do responsável legal do aluno (mesmos valores do item 'descricao_mae_ escolaridade'
descricao_responsavel_ financeiro	Quem é o responsável financeiro do aluno: avô(ó), cônjuge, irmão(a), mãe, o próprio aluno, outros, pai, parentes, tio(a)
descricao_trabalho	Descrição do trabalho do aluno: aposentado, autônomo, beneficiário ou pensionista do INSS, empresa privada, estágio ou bolsa, nunca trabalhou, não está trabalhando, não informado, pescador, serviço público, trabalha com vínculo empregatício, trabalhador rural/agricultor
pessoa_fisica_sexo	Sexo do aluno: M, F
possui_necessidade_ especial	Verdadeiro para alunos com necessidades especiais
qtd_pessoas_domicilio	Quantidade de pessoas que moram com o aluno
Sigla	Qual o campus do aluno
qnt_pc	Resultado da soma de computadores, notebooks e netbooks com valor máximo de 4
qnt_salarios	Renda bruta familiar representada em quantidade de salários mínimos com valor máximo de 10
tempo_entre_conclusao_ ingresso	Tempo entre a conclusão do ensino fundamental e entrada no IFRN com valor máximo de 3

correspondência definido na Seção II e então gerados três gráficos para cada um dos 25 atributos da base de dados, relacionando cada um deles à classe (classe 0 se o aluno evadiu, classe 1 o aluno regular). Os gráficos gerados foram: o mapa perceptual, a partir da análise de correspondência do pacote *Prince*; o *Heatmap* da análise de correspondência calculada de acordo com o descrito na Seção II; e o *bar-plot* do tipo *stacked*. O mapa perceptual apresenta a noção de distância entre cada valor nominal do atributo e as classes 0 e 1. O *Heatmap* apresenta, a partir das cores, a atração ou repulsão entre cada valor nominal do atributo e as classes 0 e 1, sendo o vermelho com significado de atração e o azul de repulsão. O *bar-plot* representa em valores absolutos a quantidade de instâncias da classe 0 e classe 1 para cada valor nominal do atributo.

A Figura 1 apresenta a metodologia proposta nesse trabalho, que pode ser descrita como:

1) **Calcular AC**: Etapa para calcular os índices de AC a partir do qui-quadrado sobre a tabela de contingência entre os atributos e a classe (0 aluno evadido, 1 aluno regular).
   a) Gerar tabela de contingência entre cada um dos atributo e a classe do aluno;
   b) Gerar índice de independência a partir do cálculo do qui-quadrado sobre a tabela de contingência.
2) **Gerar Gráficos**: Etapa para gerar os gráficos sobre os índices calculados da etapa 01.
   a) Gerar Heatmap sobre os valores do qui-quadrado;
   b) Gerar Mapa Perceptual sobre os valores do qui-quadrado;
   c) Gerar Bar-plot sobre as quantidades absolutas.
3) **Interpretar**: Etapa para interpretar os gráficos gerados na etapa 02.
   a) Interpretação visual sobre o Heatmap em que o vermelho escuro representa atração entre o valor de um atributo e a classe do aluno, já o azul escuro a repulsão;
   b) Interpretação visual sobre o Mapa Perceptual em que menores distância representa uma maior dependência entre o valor do atributo e a classe do aluno;
   c) Interpretação visual sobre o Bar-plot em que apresenta em números absolutos a quantidade de instâncias da classe do aluno para cada valor do atributo.

Fig. 1. Sistematização da metodologia adotada para identificação do perfil do aluno evadido a partir da AC.

Todo o código do trabalho e os dados utilizados estão disponíveis em [21].

IV. Resultados

Para este trabalho foram gerados 75 gráficos (para cada um dos 25 atributos foi feito o gráfico mapa perceptual, *heatmap* e *stacked*). Nas Figuras 2, 3 e 4 são colocados alguns exemplos dos gráficos gerados a partir da relação entre a evasão e os atributos ”escolaridade do responsável financeiro” e ”raça”.

Como visto na Figura 2, o mapa perceptual dá uma indicação de distância entre os valores nominais dos atributos e das classes. No ”Caso I”, há uma maior aproximação entre a classe 0 e os valores relacionados à baixa escolaridade do responsável financeiro do estudante (representado pelos valores ”Alfabetizado”, ”Ensino fundamental incompleto”, ”Ensino médio incompleto”, ”Não estudou”). Já no ”Caso II”, destaca-se a proximidade entre a classe 1 e as raças ”Branca” e ”Parda”, informação contrastante com o ”Caso II” da Figura 4, uma vez que neste gráfico é mostrado um maior quantitativo em números absolutos de alunos da classe 0 para essas raças, o que poderia levar a uma interpretação incorreta, já que as raças mais relacionada com a evasão, de acordo com o cálculo de independência e visualizado nas Figuras 2 e 3, são a ”Amarela” e a ”Preta”. Ou seja, um maior quantitativo absoluto entre valores nominais de duas variáveis, não significa necessariamente que há uma alta dependência entre eles.

A partir da análise visual da Figura 3 é possível realizar uma análise de independência entre os atributos e a classe que representa a evasão, como, por exemplo, no ”Caso I” a forte repulsão entre alunos evadidos e a escolaridade com de ”Ensino médio completo” do responsável financeiro e mais moderada com ”Ensino superior completo”, ”Pós-graduação completo” e ”Pós-graduação incompleto”. No ”Caso II” confirma-se a análise de distância da Figura 2, uma relação de atração entre a classe 0 e a raça ”Amarela” e ”Preta”, e uma relação de afastamento entre ”Branca” e ”Parda”. Nos dois casos, que compõem a Figura 3, é verificada que a força de atração ou repulsão não está bem representada no lado da classe dos alunos não evadidos (classe 1). Acredita-se que o motivo seja que a base de dados utilizada neste trabalho é fortemente desbalanceada (uma razão de 1:10) entre as classes de alunos evadidos e alunos regulares e a informação acaba se tornando diluída para o último. Também é importante destacar que a análise de correspondência gera um índice para cada relação entre cada um dos 25 atributos e a classe de evasão, logo, a força de atração ou repulsão não pode ser comparada entre os 25 atributos, mas sim apenas entre os valores assumidos por cada atributo, por exemplo, não faz sentido dizer que o índice de independência da raça ”Amarela” com a classe de evasão é maior ou menor do que o índice de independência do valor ”Alfabetizado” do atributo escolaridade do responsável financeiro.

A Figura 4 mostra o gráfico *bar-plot* do tipo *stacked*, apresentando o quantitativo absoluto de alunos evadidos ou regulares com os atributo escolaridade do responsável financeiro e raça. No ”Caso I” é destacado que o valor ”Ensino médio completo” e ”Ensino fundamental incompleto” tem um grande quantitativo de evadidos, entretanto, ao analisar os gráficos mapa perceptual e *heatmap*, é verificado que para o primeiro atributo não há uma força de atração com a classe 0, diferente do valor ”Ensino fundamental incompleto”, no qual é verificado que há uma atração forte com a classe 0. Portanto, o uso do gráfico do tipo *stacked* oculta relações evidenciadas pelos gráficos mapa perceptual e *heatmap*.

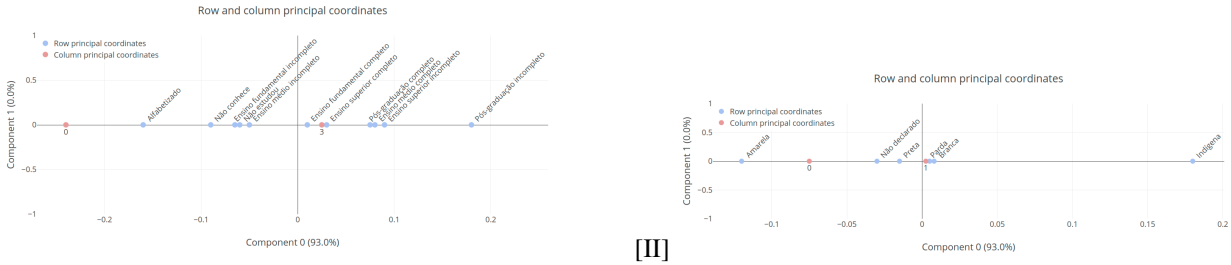
A partir da análise visual do gráfico de *heatmap* e do mapa perceptual dos atributos, na Tabela II são listadas as relações de atração entre os valores nominais para cada atributo com a classe evasão (classe 0).

Após a análise da Tabela II, podemos traçar de forma preliminar as características do perfil mais vulnerável para evasão, dado o contexto educacional do estudo de caso, quais sejam: alunos com baixo desempenho e pouca assiduidade, que já reprovaram ao menos uma vez, que não moram com os pais, com ensino fundamental em escola pública, que a escolaridade do responsável financeiro seja baixa, raça preta ou amarela, do sexo masculino, que convivam com mais de 6 pessoas no mesmo domicílio, que não tenha acesso a computador em casa, com renda bruta familiar de apenas 1 salário mínimo e que passou mais de 2 anos entre a conclusão do ensino fundamental e o ingresso no ensino integrado do IFRN. Portanto, de forma preliminar, há um forte indício que o problema da evasão do integrado do IFRN esteja ligado principalmente a vulnerabilidade social do aluno.

Para traçar de forma preliminar as características do aluno que tende a se manter no curso, foi utilizado principalmente os gráficos de mapa perceptual, uma vez que a informação no gráfico *Heatmap* não está clara devido ao desbalanceamento de dados, são elas: alunos com notas acima da média (6) e assíduos, que não reprovaram, que estudaram em escola particular, que moram com os pais ou mãe em imóvel alugado ou próprio, localizado na zona urbana ou rural, solteiros, com o responsável financeiro com ensino superior completo, com 3 a 4 computadores em casa, com renda bruta familiar a partir de 2 salários mínimos, convivendo com menos de 6 pessoas, que terminou o ensino fundamental no ano anterior, de raça parda ou branca. Essas características são semelhantes às famílias de classe média brasileira. Logo, os resultados indicam forte indício da relação da evasão com a vulnerabilidade social com que o aluno se encontra.

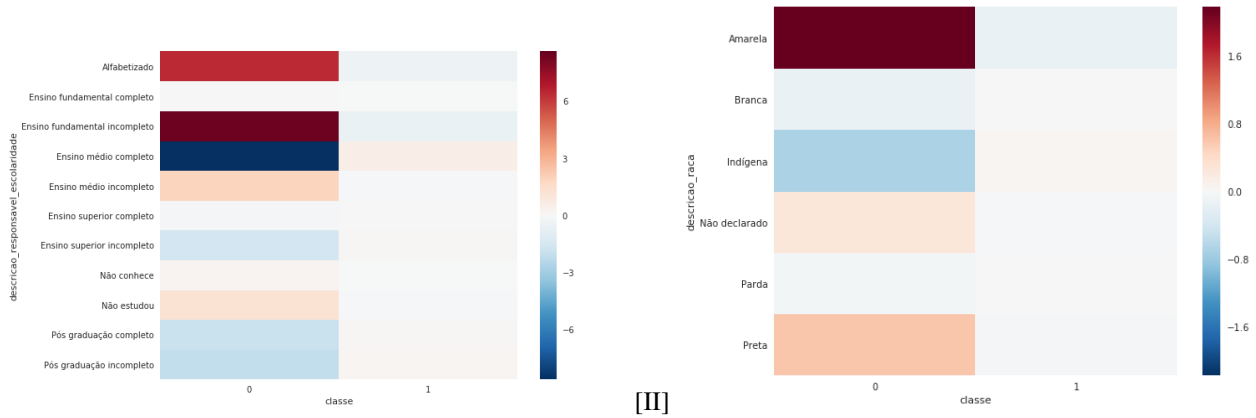
V. Conclusão

Devido à relevância do problema de evasão escolar em países da América Latina em geral e no Brasil em particular, neste trabalho foi avaliado o uso da técnica de análise de correspondência associado com técnicas de visualização de dados para análise de caracterização do perfil de alunos evadidos. A partir da análise, foram destacados as seguintes características do perfil do aluno evadido: alunos com baixo desempenho e pouca assiduidade, que já reprovaram ao menos uma vez, que não moram com os pais, com ensino fundamental em escola pública, que a escolaridade do responsável financeiro seja baixa, raça preta ou amarela, do sexo masculino, que convivam com mais de 6 pessoas no mesmo domicílio, que não tenha acesso a computador em casa, com renda bruta familiar



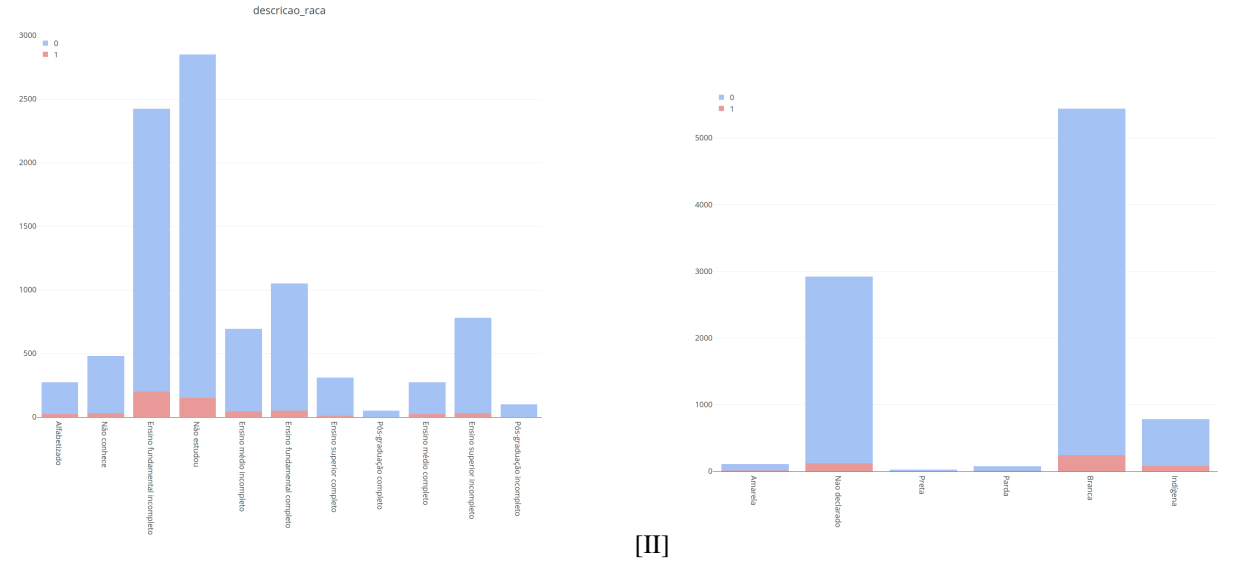
[I] [II]

Fig. 2. Mapa perceptual. Os pontos vermelhos representam as classes (”0” aluno evadido, ”1”: aluno persistente) e os pontos azuis representam os valores do atributo. Quanto mais próximos os pontos, maior a relação de dependência entre eles. Caso I: escolaridade do responsável financeiro x classe. Caso II: raça x classe.



[I] [II]

Fig. 3. *Heatmap*. O eixo x representa a classe (”0” aluno evadido, ”1”: aluno persistente) e o eixo y representa os valores do atributo. Quanto mais escuro for o vermelho, maior a relação de dependência. Caso I: escolaridade do responsável financeiro x classe. Caso II: raça x classe.



[I] [II]

Fig. 4. *Bar-plot*. Gráfico em barras quantificando e relacionando em números absolutos as instâncias de cada classe (”0” aluno evadido, ”1”: aluno persistente) por atributo. O verde representa a quantidade de alunos persistentes e o azul a quantidade de alunos evadidos. Caso I: escolaridade do responsável financeiro x classe. Caso II: raça x classe.

de apenas 1 salário mínimo e que passou mais de 2 anos entre a conclusão do ensino fundamental e o ingresso no ensino integrado do IFRN. Ou seja, um forte indício que o problema da evasão do integrado do IFRN esteja ligado principalmente a vulnerabilidade social do aluno. Devido à natureza categórica de boa parte dos dados sobre os estudantes, a determinação do grau de independência entre variáveis via uso de técnica de análise de correspondência se mostrou mais adequada do que a tradicional análise de correlação de Pearson. Os resultados obtidos indicam que essa abordagem se mostrou eficiente para traçar as características do perfil do aluno mais vulneráveis ao fenômeno de evasão e a facilidade de interpretação visual

TABELA II
RELAÇÃO ENTRE ATRIBUTOS E EVASÃO

Atributo	Descrição
LnguaPortuguesae LiteraturaI90H	Relação de atração com notas abaixo de 50
LnguaPortuguesae LiteraturaI90H Dependencia	Atração forte com 1 dependência
LnguaPortuguesae LiteraturaI90Hfreq	Começa a aparecer atração a partir de 85%
MatemticaI120H	Não é tão perceptível como de português, mas notasse a relação de atração com notas baixas
MatemticaI120H_ dependencia	Fortemente ligado a 1 dependência
MatemticaI120H_freq	Começa a aparecer atração a partir de 85%
aluno_exclusivo_ rede_publica	Atração forte com o valor Verdadeiro
descricao_area_residencial	Atração forte com o valor "não informado", e repulsão leve com "urbana"
descricao_companhia_ domiciliar	Atração forte com o valor "Cônjuge" e moderada com "Outros"
descricao_estado_civil	Atração forte com o valor "Divorciado"
descricao_historico	Atração forte com os cursos de "Informática" e "Têxtil" e moderada com "Meio Ambiente"
descricao_imovel	Atração forte com "Não informado" e de repulsão com "Financiado"
descricao_mae_ escolaridade	Atração forte com "Fundamental Incompleto", moderado com "Alfabetizado", "Não estudou", "Médio incompleto" e repulsão forte com "Médio Completo"
descricao_pai_ escolaridade	Atração forte com "Alfabetizado" e "Não estudou", moderado com "Fundamental Incompleto" e repulsão forte com "Médio Completo"
descricao_raca	Atração forte com "Amarelo" e moderada com "Preta", e de repulsão com "Indígena"
descricao_responsavel_ escolaridade	Atração forte com "Fundamental Incompleto" e "Alfabetizado", e repulsão forte com "Médio Completo" e moderada com "Superior incompleto", "Pós graduação completo", "Pós graduação incompleto"
descricao_responsavel_ financeiro	Atração forte com "O próprio aluno" e moderada com "Cônjuge" e "Avô(ó)". Repulsão moderada com "Pai"
descricao_trabalho	Atração forte "Não informado". Repulsão leve "Nunca trabalhou" (alunos que apenas estudam)
pessoa_fisica_sexo	Atração forte masculino. Repulsão forte feminino
possui necessidade especial	Atração forte "True", repulsão moderada "False"
qtd_pessoas_domicilio	Atração acima de 6 e com o valor 0. Repulsão moderada com o valor 4
Sigla	Atração forte MC, moderada JC e leve CA, LAJ, NC, SC, SPP. Repulsão forte PAR, moderada para CN, e leve em CANG, CNAT, MO, PF, SGA
qnt_pc	Atração forte com 0, repulsão a partir de 1
qnt_salarios	Atração forte com renda bruta familiar de 1 salário mínimo. Repulsão a partir de 2.
tempo_entre_conclusao_ ingresso	Atração forte com 3 anos, repulsão com 1 ano (significa que não parou os estudos)

dos resultados potencializa o seu emprego por professores e gestores. Deste modo, a abordagem adotada aqui pode ser bastante útil para se analisar conjuntamente dados acadêmicos e socio-econômico de estudantes visando o entendimento do motivo pelo qual essas variáveis estão relacionadas ao grupo de alunos evadidos. Como trabalhos futuros, sugere-se a criação de métrica para seleção dos atributos de forma mais sistemática para a caracterização do grupo de risco de evadidos e uso de técnicas para tratar com o efeito do desbalanceamento de dados entre classes.

REFERÊNCIAS

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, Nov 2010.

[2] V. Cox, *Translating Statistics to Make Decisions: A Guide for the Non-Statistician*. Apress, 2017.

[3] D. McCandless, *Knowledge Is Beautiful*, 1st ed. Harper Design, 2014.

[4] R. P. Barros, "Políticas públicas para a redução do abandono e da evasão escolar de jovens," Fundação Brava, Insper, Instituto Unibanco e Instituto Ayrton Senna, Página na Internet, 2017. [Online]. Available: http://gesta.org.br/wp-content/uploads/2017/09/Politicas-Publicas-para-reducao-do-abandono-e-evasao-escolar-de-jovens.pdf

[5] BRASIL, "Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro," http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior, 2016. [Online]. Available: http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior

[6] ——, "Panorama da educação destaques do education at a glance 2016," DEED/MEC, Tech. Rep., 2016. [Online]. Available: http://download.inep.gov.br/acoes_internacionais/eag/documentos/2016/panorama_da_educacao_2016_eag.PDF

[7] ——, "Mec libera 100% do orçamento de custeio para universidades e institutos federais," http://redefederal.mec.gov.br/links/1204-mec-libera-100-do-orcamento-de-custeio-para-universidades-e-institutos-federais, 2018. [Online]. Available: http://redefederal.mec.gov.br/links/1204-mec-libera-100-do-orcamento-de-custeio-para-universidades-e-institutos-federais

[8] J. F. Hair, B. Black, B. Babin, R. E. Anderson, and R. L. Tatham, *Análise Multivariada de Dados*, 6th ed. bookman, 2009.

[9] A. J. Izenman, *Modern Multivariate Statistical Techniques*. Springer, 2008.

[10] K. S. Ziemer, B. Pires, V. Lancaster, S. Keller, M. Orr, and S. Shipp, "A New Lens on High School Dropout: Use of Correspondence Analysis and the Statewide Longitudinal Data System," *American Statistician*, 2018.

[11] E. Ferreiro-Lago and S. Osuna-Acedo, "Factors affecting the participation of the deaf and hard of hearing in e-learning and their satisfaction: A quantitative study," *International Review of Research in Open and Distance Learning*, 2017.

[12] C. Carlhed, "The Social Space of Educational Strategies: Exploring Patterns of Enrolment, Efficiency and Completion among Swedish Students in Undergraduate Programmes with Professional Qualifications," *Scandinavian Journal of Educational Research*, 2017.

[13] M. O. Pérez-Pulido, F. Aguilar-Galvis, G. Orlandoni-Merli, and J. Ramoni-Perazzi, "Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en la Universidad de Santander, Colombia - Statistical analysis of the results of state tests for admission to higher education at the University of Sa," *Revista Científica*, 2016.

[14] O. Matyja and W. Szczesny, "Visualization in prediction based on grade correspondence analysis," in *Advances in Soft Computing*, 2000.

[15] L. G. F. Silva, M. E. P. S. Rocha, and R. A. A. Fagundes, "Enade: Math and science students' performance analysis," *IEEE Latin America Transactions*, vol. 15, no. 09, 2017.

[16] E. B. Y. Amaya and D. Heredia, "Student dropout predictive model using data mining techniques," *IEEE Latin America Transactions*, vol. 13, no. 09, 2015.

[17] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51 – 56.

[18] M. Halford, "Prince: library for doing factor analysis," 2016. [Online]. Available: https://pypi.org/project/prince/

[19] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, J. Ostblom, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Brunner, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, and A. Qalieh, "mwaskom/seaborn: v0.9.0 (july 2018)," Jul. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1313201

[20] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science & Engineering*, 2007.
[21] T. M. Barros, “Modelo ifrn,” 2018. [Online]. Available: http://bit.ly/321Mr8l

Thiago Medeiros Barros é graduado em Engenharia da computação pela Universidade Federal do Rio Grande do Norte (UFRN), 2009. Obteve o título de mestre em engenharia elétrica e computação pela UFRN, 2013, em que atualmente também faz o doutorado. É professor de tecnologias educacionais e Diretor de Produção de Material Didático no Instituto Federal do Rio Grande do Norte. Suas pesquisas incluem mineração e visualização de dados educacionais, tecnologias educacionais, produção de material didático e recursos educacionais abertos.

Ivanovitch Silva Possui graduação em Engenharia de Computação (2006), mestrado (2008) e doutorado (2013) em Engenharia Elétrica e Computação pela Universidade Federal do Rio Grande do Norte e coparticipação da Universidade do Porto (Sanduíche), e curso de aperfeiçoamento tecnológico em Big Data & Social Analytics pelo Massachusetts Institute of Technology (MIT, 2016). Desde 2013 é docente da Universidade Federal do Rio Grande do Norte sendo lotado no Instituto Metrópole Digital (IMD). Seus interesses em pesquisa incluem: modelagem e análise científica de dados, internet das coisas, indústria 4.0 e cidades inteligentes.

Luiz Affonso Guedes Luiz Affonso Guedes, graduado em Engenharia Elétrica na UFPA em 1998. Mestre em Engenharia Eletrônica e Computação pelo ITA em 1991 e Doutor pela Unicamp na área de Automação e Engenharia de Computação em 1999. Atualmente é professor titular do Departamento de Automação e Engenharia de Computação da UFRN. Tem interesse em pesquisa análise científica de dados nas áreas de automação industrial e educacional.