Uso da Técnica de Análise de Correspondência para Análise Exploratória de Dados no Contexto Educacional

Thiago Medeiros Barros[1], **Ivanovitch Silva**[2], **Luiz Affonso Guedes**[3]

[1]Campus EaD – Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN) – Natal – RN – Brasil

[2]Instituto Metrópole Digital
Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil

[3]Departamento de Engenharia de Computação e Automação
Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil

thiago.medeiros@ifrn.edu.br, ivan@imd.ufrn.br, affonso@dca.ufrn.br

Abstract. *The purpose of this paper is to show how the Correspondence Analysis 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. To this, the independence calculation was implemented from the chi-square, and the heatmap and perceptual map were generated with the indexes on the socioeconomic data of students of courses of IFRN. 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.*

Resumo. *A proposta deste trabalho é mostrar como a técnica de Análise de Correspondência pode ser utilizada para potencializar a Análise Exploratória de Dados sobre o fenômeno da evasão para um conjunto de dados educacionais com maioria do tipo categórico. Para tal, foi implementado o cálculo de independência a partir do qui-quadrado, e produzidos os gráficos heatmap e mapa perceptual com os índices gerados sobre os dados socioeconômicos de alunos de cursos integrados do Instituto Federal do Rio Grande do Norte (IFRN). Como resultados são apresentadas as relações de atração e repulsão entre os atributos socioeconômicos e a evasão, e traçado o perfil de aluno mais vulnerável para evasão, sendo algumas características: alunos que já reprovaram ao menos uma vez, que não moram com os pais, com ensino fundamental em escola pública, a escolaridade do responsável financeiro baixa.*

1. 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 [Romero and Ventura 2010]. A área está se tornando uma grande aliada aos professores e à gestão de institutos educacionais, a fim de auxiliar na descoberta de novos

DOI: 10.5753/cbie.wcbie.2018.389

conhecimentos e de novos padrões, com o objetivo de subsidiar na 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 quais dados existem, quais as tendências possíveis, e, portanto, quais testes estatísticos serão apropriados para usar [Cox 2017]. 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 [McCandless 2014].

Dentre as diversas aplicações da EDM, a predição de evasão vem ganhando destaque nacionalmente. Essa preocupação se dá pelo fato de que a evasão de um aluno representa oportunidades de mudança de vida desperdiçadas, menos mão de obra qualificada no mercado, menor chance de mobilidade social [Barros 2017], principalmente em um país com índices de desigualdades sociais como o Brasil. Para ilustrar a evasão escolar no país, em 2010, 11,4% dos alunos abandonaram o curso para o qual foram admitidos. Já em 2014, esse número chegou a 49% [BRASIL 2016a], 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 [BRASIL 2016b]. Além disso, estima-se que 7 bilhões 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 [Barros 2017].

Diante desse cenário, ferramentas de mineração e visualização de dados podem auxiliar na descoberta de relações entre variáveis disponíveis para gestão através dos sistemas de controle acadêmico e a evasão, permitindo melhores tomadas de decisões para diminuir 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). Para esse tipo de variável, devem ser utilizadas técnicas de análise de independência, como a análise de correspondência, ao invés de medidas de correlações clássicas, por exemplo a de *Pearson*, geralmente utilizadas para dados numéricos como notas de disciplina [HAIR et al. 2009].

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 pode identificar as principais características socioeconômicas e demográficas de alunos evadidos no contexto do IFRN. Serão analisados, neste artigo, dados educacionais de alunos dos cursos do Integrado atualizados em janeiro de 2018, na modalidade presencial.

Com o intuito de alcançar o objetivo acima, este estudo está dividido em mais seis seções. Na Seção 2, denominada de "Fundamentação Teórica", é apresentada a área de Mineração de Dados Educacionais e a técnica estatística de análise de correspondência. Na Seção 3, nomeada de "Trabalhos Relacionados", são apresentadas outras referências, destacando as variáveis e os algoritmos utilizados para abordar a evasão. Na Seção 4, denominada "Metodologia", é apresentado o ambiente de desenvolvimento e a base utilizada. Na Seção 5, denominada "Resultados e Discussão", 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.

2. Fundamentação Teórica

2.1. *Education Data Mining (EDM)*

EDM é uma área de pesquisa emergente que vem crescendo principalmente a partir da explosão de dados educacionais gerados com o desenvolvimento de sistemas de informação para controle acadêmico. Sendo considerada uma área interdisciplinar, a EDM utiliza métodos de *Data Mining* e estatística para explorar os dados que têm origem em contextos educacionais. As principais subáreas são: Análise e Visualização de Dados, *Feedback* para Instrutores, Recomendações para Estudantes, Predição do Desempenho do Aluno, Modelagem do Aluno, Detecção de Comportamentos de Estudantes, Agrupamento de Alunos, Análise de Redes Sociais, Desenvolvimento de Mapas Conceituais, Construção de Cursos e Planejamento e Agendamento [Romero and Ventura 2010]. De forma mais detalhada, a predição de desempenho do aluno é uma das aplicações mais populares e tradicionais da EDM, que tem como objetivo prever o valor da variável que representa o desempenho do aluno, sendo essa variável do tipo numérica ou categórica. Quando a previsão envolve variáveis numéricas, geralmente são utilizadas técnicas de regressão analítica, a qual encontra a relação entre uma variável dependente e uma ou mais variáveis independentes. Já quando a variável é representada por valores categóricos, geralmente são usadas técnicas de agrupamento em que itens individuais são colocados em grupos com base em cálculos de similaridade entre as instâncias.

2.2. Análise de Correspondência

A análise de correspondência é uma técnica exploratória em dados multivariados frequentemente utilizada para redução de dimensionalidade e mapeamento perceptual em base de dados composta por dados categóricos [HAIR et al. 2009]. O objetivo é esclarecer 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 [Izenman 2008]. 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 discretizadas. Para a análise de correspondência, é calculada a 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 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). Para o cálculo do qui-quadrado, devem ser realizados os passos a seguir [HAIR et al. 2009]:

1. Valor esperado: representa o valor esperado para uma célula da tabela de contingência. O cálculo é feito a partir da probabilidade conjunta da combinação da coluna com a linha, através da probabilidade marginal para a coluna (total da coluna / total geral) vezes a probabilidade marginal para a linha (total da linha / total geral). Esse cálculo é descrito na equação abaixo:

$$FrequenciaEsperada = \frac{TotalColuna * TotalLinha}{TotalGeral} \quad (1)$$

2. Diferença entre as frequências esperadas e reais: representa o quão distante o valor real está da frequência esperada naquela célula da tabela de contingência. A diferença é calculada a partir do passo anterior e os valores reais observados na tabela de contingência, sendo obtida via a equação:

$$diferenca = FrequenciaEsperada - FrequenciaObservada \quad (2)$$

3. Valor do teste qui-quadrado: relaciona-se esse valor é relacionado à intensidade de associação entre os valores nominais das variáveis. O cálculo é realizado a partir da razão entre a diferença ao quadrado calculada no passo anterior e a Frequência esperada calculada no primeiro passo e descrita pela equação a seguir:

$$QuiQuadrado = \frac{Diferenca^2}{FrequenciaEsperada} \quad (3)$$

4. Sinal da medida de similaridade: o valor definido no teste qui-quadrado deve retomar a direção removida ao elevar a diferença ao quadrado no passo 03. Para tornar a medida mais intuitiva, deve ser atribuído o sinal inverso ao gerado no cálculo de diferença no passo 02. Portanto, caso o sinal tenha valor negativo, ele representa a repulsão entre os valores nominais de cada atributo. Por outro lado, caso o sinal seja positivo, o valor representa a força de atração entre os valores nominais de cada atributo.

A partir desse valor, é gerado o mapa percentual de tal forma que quanto mais próximos estiverem dois atributos mais similares o são.

3. Trabalhos Relacionados

A Tabela 1 apresenta trabalhos relacionados com o estudo de predição de desempenho e/ou evasão escolar, destacando os objetivos de cada trabalho, as variáveis independentes utilizadas no modelo e os algoritmos usados para predição. A tabela foi produzida a partir da seleção de artigos do portal *Web of Science* utilizando as palavras-chaves *predict, perfomance, student*.

Tabela 1: Comparação com Trabalhos Relacionados.

Referência	**Objetivo do Trabalho**	**Variáveis do Modelo**	**Algoritmo**
[Huang and Fang 2013]	Predição de Desempenho	Notas	Regressão Linear, *Multilayer Perceptron*, Função de Base Radial, Máquina de vetores de suporte
[Rovira et al. 2017]	Predição de Desempenho e Evasão	Notas	Regressão Logística, Naive Bayes, Máquina de vetores de suporte, *Random Forest*, *Adaptive Boosting*, Filtro colaborativo, Regressão Linear
[Burgos et al. 2017]	Predição de Evasão	Notas	Regressão

[Li et al. 2013]	Predição de Desempenho	Notas	Análise de Componentes Principais
[Xu et al. 2017]	Predição de Desempenho	Notas	Fatoração de matriz probabilística
[Meier et al. 2015]	Predição de Desempenho	Notas	Algoritmo próprio utilizando cálculo de semelhança e Regressão

Os trabalhos apresentados constroem modelos preditivos de evasão ou desempenho utilizando apenas variáveis que indicam a performance escolar de disciplinas já consolidadas pelo aluno e/ou de atividades avaliativas que compõem a nota final de uma disciplina, negligenciando informações demográficas e socioeconômicas que revelam situações de vulnerabilidade as quais podem influenciar no desempenho escolar do aluno e na sua evasão.

Nesses trabalhos não foram utilizados dados socioeconômicos e demográficos dos alunos, que poderiam auxiliar em modelos preditivos mais robustos e precisos, além de uma melhor compreensão da influência desses atributos sobre o fenômeno de evasão, tornando possível traçar um perfil de vulnerabilidade sobre tal fenômeno em um dado contexto educacional.

4. Metodologia

Os dados, atualizados em janeiro de 2018, 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 IFRN distribuídos por 20 *Campi* (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.ifrn.edu.br) desenvolvido pelo IFRN e possui informações demográficas, caracterização socioeconômica e média final dos alunos nas disciplinas. 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 2.

Tabela 2: Descrição das variáveis selecionadas

Atributo	**Descrição**
LnguaPortuguesaeLiteraturaI90H	Média de 0 a 10 da disciplina língua portuguesa
LnguaPortuguesaeLiteraturaI90HDependencia	Quantidade de dependências (repetição da disciplina devido a reprovação no ano anterior) do aluno na disciplina de língua portuguesa
LnguaPortuguesaeLiteraturaI90Hfreq	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 no ano anterior) do aluno na disciplina de matemática
MatemticaI120H_freq	Porcentagem de 0 a 100 da frequência na disciplina de matemática
aluno_exclusivo_rede_publica	Se o aluno é exclusivo da rede público 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
descricao_estado_civil	Descrição do estado civil do aluno
descricao_historico	Qual curso técnico o aluno faz
descricao_imovel	Qual situação financeira do imóvel em que o aluno mora
descricao_mae_escolaridade	Escolaridade da mãe do aluno
descricao_pai_escolaridade	Escolaridade do pai do aluno
descricao_raca	Raça autodeclarada do aluno
descricao_responsavel_escolaridade	Escolaridade do responsável legal do aluno
descricao_responsavel_financeiro	Quem é o responsável financeiro do aluno
descricao_trabalho	Descrição do trabalho do aluno
pessoa_fisica_sexo	Sexo do aluno
possui_necessidade_especial	Se o aluno possuí necessidades especiais
qtd_pessoas_domicilio	Quantidade de pessoas que moram com o aluno
Sigla	Qual o campus do aluno

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

Após a criação da base de dados e implementado o ambiente de desenvolvimento, foi realizado o cálculo de análise de correspondência definido na Seção 2.2 e então gerados três gráficos para cada um dos 22 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 2; 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.

Todo o código do trabalho e os dados utilizados estão disponíveis em

[Barros 2018].

5. Resultados e Discussão

Para o trabalho, foram gerados 66 gráficos (para cada um dos 22 atributos foi feito o gráfico mapa perceptual, *heatmap* e *stacked*). Abaixo são colocados alguns exemplos dos gráficos gerados em relação ao atributo escolaridade do responsável financeiro.

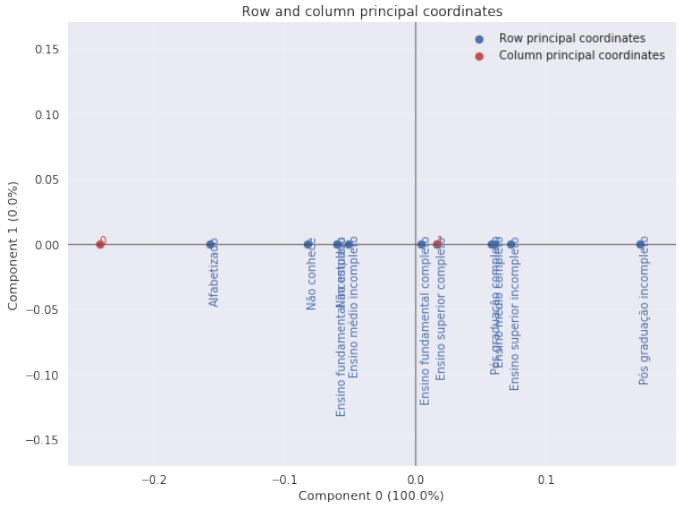

Figura 1. Mapa perceptual escolaridade do responsável financeiro x classe.

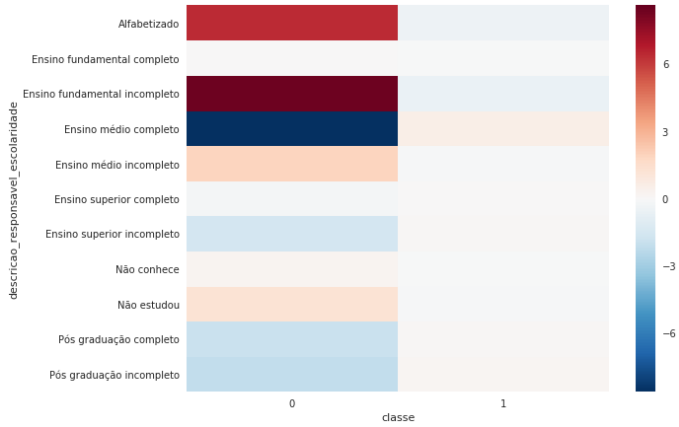

Figura 2. Gráfico *Heatmap* escolaridade do responsável financeiro x classe.

Como visto na Figura 1, o mapa perceptual dá uma indicação de distância entre os valores nominais dos atributos e das classes. Nota-se uma relação clara entre a baixa escolaridade do responsável financeiro (representado pelos valores "Alfabetizado", "Ensino fundamental incompleto"e "Ensino médio incompleto") e a atração com a classe 0. Na Figura 2, é mais fácil realizar de forma visual uma análise de quais valores para cada atributo estão mais relacionados com a evasão, como, por exemplo, 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". Também é verificada que a força de atração ou repulsão

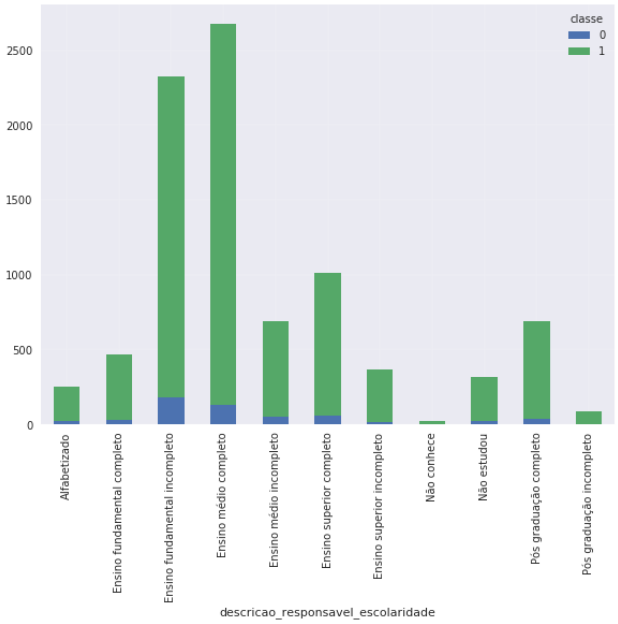

Figura 3. Gáfico *Bar-plot* escolaridade do responsável financeiro x classe.

não está bem representada no lado da classe dos alunos não evadidos. Acredita-se que o motivo seja que a base de dados 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 a classe 1. Também é importante notar que a análise de correspondência gera um índice para cada par de atributos, logo, a força de atração ou repulsão não pode ser comparada entre os 22 atributos. A Figura 3 mostra o gráfico *bar-plot* do tipo *stacked*, apresentando o atributo escolaridade do responsável financeiro. 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 *heatmap* e do mapa perceptual de todos os atributos, na Tabela 3 estão listadas as relações de atração entre os valores nominais para cada atributo com a classe evasão (classe 0).

Tabela 3: Relação entre atributos e evasão

Atributo	**Descrição**
LnguaPortuguesaeLiteraturaI90H	Notas abaixo de 60
LnguaPortuguesaeLiteraturaI90H_dependencia	Atração forte com 1 dependência
LnguaPortuguesaeLiteraturaI90H_freq	Começa a aparecer atração a partir de 85%
MatemticaI120H	Não é tão perceptível como de português, mas notasse a relação 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"
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 o valor "Informática" e "Têxtil" e moderada com "Meio Ambiente"
descricao_imovel	Atração forte com "Não informado"
descricao_mae_escolaridade	Atração forte com "Ens. Fund. Incompleto", moderado com "Alfabetizado" e repulsão forte com Ens. Med. Completo
descricao_pai_escolaridade	Atração forte com "Ens. Fund. Incompleto" e "Não estudou", moderado com "Alfabetizado" e repulsão forte com Ens. Med. Completo
descricao_raca	Atração forte com "Amarelo" e moderada com "Preta"
descricao_responsavel_escolaridade	Atração forte com "Ens. Fund. Incompleto" e "Alfabetizado" e repulsão forte com Ens. Med. Completo
descricao_responsavel_financeiro	Atração forte com "O próprio aluno" e moderada "Cônjuge". Repulsão moderada com "Pai"
descricao_trabalho	Atração forte "Não informado". Repulsão leve "Nunca trabalhou"
pessoa_fisica_sexo	Atração forte masculino. Repulsão feminino
possui_necessidade_especial	Atração forte "True", repulsão moderada "False"
qtd_pessoas_domicilio	Atração forte acima de 10 e 0. Repulsão moderada 4
Sigla	Atração forte MC, moderada JC e leve CA, LAJ, NC, SC, SPP. Repulsão moderada PAR, CN

Após a análise acima, podemos traçar de forma preliminar os perfis mais vulneráveis para evasão, dado contexto educacional descrito, qais sejam: alunos que já reprovaram ao menos uma vez, que não moram com os pais, com ensino fundamental em escola pública, escolaridade do responsável financeiro baixa, negro ou amarelo, do sexo masculino.

6. Conclusão

O uso de análise de correspondência associado ao uso de técnicas de visualização de dados se mostrou um método eficiente para traçar o perfil de risco de evasão. A partir desses resultados, podemos realizar trabalhos para o entendimento mais profundo do motivo pelo qual essas variáveis estão mais relacionadas ao grupo de alunos evadidos e quais ações

a instituição de educação pode realizar para evitar a evasão. Como trabalho futuro, é sugerida a criação de uma métrica para seleção dos atributos que caracterizem o grupo de risco de evadidos, e como minimizar efeito do desbalanceamento de dados.

Referências

Barros, R. P. (2017). Políticas públicas para a redução do abandono e da evasão escolar de jovens. Página na internet, Fundação Brava, Insper, Instituto Unibanco e Instituto Ayrton Senna.

Barros, T. M. (2018). Modelo ifrn. `https://github.com/tmedeirosb/modelo_ifrn_integrado/blob/master/versao_2/workflow_CA.ipynb`.

BRASIL (2016a). 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`.

BRASIL (2016b). Panorama da educação destaques do education at a glance 2016. Technical report, DEED/MEC.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers Electrical Engineering*, pages –.

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

HAIR, J. F., BLACK, B., BABIN, B., ANDERSON, R. E., and TATHAM, R. L. (2009). *Análise Multivariada de Dados*. bookman, 6th edition.

Huang, S. and Fang, N. (2013). Predicting student academic performance in an engineering dynamics course : A comparison of four types of predictive mathematical models. *Computers & Education*, 61:133–145.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques*. Springer.

Li, K. F., Rusk, D., and Song, F. (2013). Predicting student academic performance. In *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 27–33.

McCandless, D. (2014). *Knowledge Is Beautiful*. Harper Design, 1th edition.

Meier, Y., Xu, J., Atan, O., and v. d. Schaar, M. (2015). Personalized grade prediction: A data mining approach. In *2015 IEEE International Conference on Data Mining*, pages 907–912.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Rovira, S., Puertas, E., and Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS ONE*, pages 1–21.

Xu, J., Moon, K. H., and van der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, PP(99):1–1.