Modelagem e Visualização Científica de Dados Educacionais: Estudo de Caso sobre o Desempenho em Componentes Curriculares

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 main objective of this paper is to investigate how scientific modeling and visualization techniques can enhance the interpretation of educational data. For data modeling, k-means algorithm and correspondence analysis were used for clustering and dimensionality reduction, respectively. For data visualization, boxplot, violin chart and perceptual map techniques were analyzed. The analyses were performed using academic data from two courses of Instituto Federal do Rio Grande do Norte in the period from 2008 to 2016. One of the obtained results shows that students' grades follow multimodal distributions, indicating that the use of classic boxplots is not adequate.*

Resumo. *A proposta deste trabalho é mostrar técnicas de visualização de dados em conjunto com técnicas estatísticas, a fim de potencializar a interpretação de dados educacionais. Para tal, são analisados as notas dos alunos dos cursos de Licenciatura em Espanhol e Segurança do Trabalho do IFRN no período de 2008 a 2016, na modalidade presencial e a distância. Para modelagem, foram utilizados o algoritmo de k-means para agrupamentos de dados e a análise de correspondência para redução de dimensionalidade. Para visualização de dados, são utilizados o boxplot, o gráfico de violino e mapa perceptual. Um dos resultados obtidos indica que as notas dos alunos seguem distribuições multimodais, as quais o emprego dos clássicos boxplots não é indicado.*

1. Introdução

Dentre diversos desafios no contexto educacional, os índices de evasão nos cursos superiores das Universidades Brasileiras trazem números preocupantes. 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 estão estudando, sendo o maior índice no mundo entre os países pesquisados pelo relatório *Education at a Glance* [BRASIL 2016b]. Portanto,

DOI: 10.5753/cbie.wcbie.2017.654

identificar alunos com potencial de evasão permite à Instituição de Ensino realizar ações para retê-los de forma eficaz e eficiente [Rovira et al. 2017], por exemplo.

Devido a isto, abordagens de mineração de dados voltadas ao contexto educacional vêm ganhando atenção da comunidade acadêmica, ao ponto de atualmente já ser caracterizada como uma área de pesquisa em si e denominada de *Educational Data Mining* (EDM) [Romero and Ventura 2010]. EDM não se limita à análise estatística sobre dados educacionais, mas visa principalmente inferir relacionamentos a partir de análise de dados para melhor embasar ações e políticas educacionais.

Atualmente, atividades de mineração de dados estão incluídas numa área maior, denominada de Análise Científica de Dados (*Data Science Analysis*, em inglês). Essa área se utiliza primordialmente de técnicas e ferramentas advindas da estatística e da aprendizagem de máquina para análise dos dados. Mais recentemente, técnicas avançadas de visualização de dados deixaram de ser apenas assessórias, para se tornarem essenciais para a área de *Data Science* [McCandless 2014]. Isto se deve ao fato de que a forma como os dados são apresentados aos usuários é fundamental para a sua interpretação cognitiva. Em muitos casos, o emprego de técnicas de visualização de dados adequadas possibilita inferências complexas por parte dos usuários, mesmo utilizando-se apenas estatísticas básicas. É importante ressaltar que técnicas de visualização de dados não são substitutas ou concorrentes das técnicas estatísticas e de aprendizagem de máquina, mas sim uma aliada que visa potencializar as análises sobre os dados.

O objetivo principal do presente artigo é investigar como técnicas de visualização de dados quando empregadas em conjunto com técnicas estatísticas e de aprendizagem de máquina podem potencializar a interpretação dos resultados de dados educacionais por parte dos usuários dos mesmos. Para tal, neste artigo serão analisados dados educacionais de alunos dos cursos de Licenciatura em Espanhol e Técnico de Segurança do Trabalho do Instituto Federal do Rio Grande do Norte no período de 2008 a 2016, na modalidade presencial e a distância respectivamente. Serão analisados dados socio-econômicos e de média final dos alunos nos componentes curriculares. Para analisar esses dados serão utilizados o algoritmo de *k-means* para determinação dos agrupamentos de dados e análise de correspondência para redução de dimensionalidade. Em relação a visualização de dados, serão utilizados os clássicos *boxplot*, além dos gráficos de violino e mapa perceptual. Ressalta-se que os gráficos *boxplot* não são adequados para distribuições multimodais.

O restante do artigo está dividido em mais quatro seções. Na Seção Fundamentação Teórica são apresentadas as principais técnicas de visualização de dados utilizadas no trabalho. Na seção Metodologia de Pesquisa é apresentado a sequência em que as técnicas são utilizadas. Em Resultados é apresentado os principais gráficos e suas interpretações. Na Conclusão é descrito o potencial dos gráficos utilizados e indicação de trabalhos futuros.

2. Fundamentação Teórica

2.1. *Educational Data Mining* (EDM)

EDM é uma área de pesquisa interdisciplinar e emergente que trabalha com o uso de 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, Planejamento e Agendamento [Romero and Ventura 2010].

Predição do desempenho do aluno é uma das aplicações mais antigas e mais populares de *Data Mining* na educação, e diferentes técnicas e modelos foram aplicados (Redes neurais, redes bayesianas, sistemas baseados em regras, regressão e análise de correlação) [Romero and Ventura 2010]. O objetivo da predição de desempenho do aluno é estimar o valor de uma variável que descreve o desempenho do aluno. Na educação o valor das variáveis que representam o desempenho podem ser do tipo numérico ou categórico. Para os valores numéricos são utilizadas técnicas de regressão analítica, a qual encontra a relação entre uma variável dependente e uma ou mais variáveis independentes. Para valores categóricos geralmente são utilizadas técnicas de agrupamento em que itens individuais são colocados em grupos com base em cálculos de similaridade entre as instâncias.

A Tabela 1 apresenta a comparação das técnicas utilizadas para predição de desempenho do aluno, destacando o objetivo do trabalho, as variáveis independentes utilizadas no modelo, e os algoritmos utilizados para predição. A tabela foi produzida a partir da seleção de artigos do portal *Web of Science* utilizando as palavras chaves *predict, perfomance, student*. A partir da Tabela 1 fica evidente que a principal variável utilizada para predição de desempenho são os históricos das notas. Logo, entender o comportamento estatístico dessa variável é fundamental para melhorar o modelo preditivo sobre comportamento de alunos.

2.2. Algoritmo *K-means* para Agrupamento de Dados

O *k-means* é um dos algoritmos mais simples e popular para agrupamento de dados. Ele busca encontrar centros dos agrupamentos que sejam representativos para grupo de dados, assumindo o conhecimento *a priori* do número de grupos. O algoritmo alterna entre duas etapas: (i) atribuição de pontos de dados aos centros do modelo a partir do cálculo de distância e, em seguida, (ii) configuração de cada centro do grupo como a média dos dados que lhe são atribuídos [Müller and Guido 2016].

2.3. Gráfico Violino

O gráfico de violino é a combinação do *boxplot* e a linha de densidade no mesmo diagrama, permitindo visualizar detalhes que as duas técnicas sozinhas não seriam capaz de detectar [Hintze and Nelson 1998]. O *boxplot* consegue capturar características importantes dos dados em sua visualização: o centro, o espalhamento, assimetria e *outliers*. Essas características são identificadas a partir da indicação no gráfico do limite inferior, primeiro quartil, mediana, terceiro quartil, limite superior.

Entretanto, quando se trata de distribuições multimodal o *boxplot* não consegue capturar essa informação, uma vez que oculta a forma da distribuição dos dados, como pode ser observado entre os gráficos 1 e 2 que representam os mesmos dados. A linha de densidade complementa as informações fornecida pelo *boxplot*, a partir da visualização gráfica da forma da distribuição dos dados. Para isso a linha de densidade é colocada simetricamente para esquerda e direita sobre o *boxplot*.

Tabela 1. Comparação com Trabalhos Relacionados

Referência	Objetivo do Trabalho	Variáveis do Modelo	Algoritmo
[Huang and Fang 2013]	Predição de Desempenho	Notas	Multiple Linear Regression, Multilayer Perception Network, Radial Basis Function Network, and Support Vector Machine
[Rovira et al. 2017]	Predição de Desempenho e Evasão	Notas	Logistic Regression, Gaussian Naive Bayes, SVM, Random Forest, Adaptive Boosting, Predict performance Collaborative Filtering Recommendation System, Linear Regression, Support Vector Regression
[Burgos et al. 2017]	Predição de Evasão	Notas	Regression Model
[Li et al. 2013]	Predição de Desempenho	Notas	Principal Component Analysis
[Xu et al. 2017]	Predição de Desempenho	Notas, Dados demográficos	Probabilistic Matrix Factorization Algorithm
[Meier et al. 2015]	Predição de Desempenho	Notas	Algoritmo próprio utilizando cálculo de semelhança e Regressão

2.4. Gráfico Q-Q

Se x e y são variáveis distribuídas de forma idêntica, então o gráfico de x-quantis versus y-quantis será uma linha reta com inclinação 1, apontada para a origem. Uma propriedade elementar dos gráficos de Q-Q é que se y for uma função linear de x, então o gráfico Q-Q correspondente ainda será linear, mas com possível mudança de localização e inclinação. Esta propriedade de invariância linear torna o uso do gráfico Q-Q uma abordagem valiosa, uma vez que a linearidade é uma configuração geométrica que os seres humanos são capazes de perceber com mais facilidade [WILK and GNANADESIKAN 1968].

Para o caso em que as variáveis têm caudas longas, o gráfico Q-Q tende a enfatizar a estrutura comparativa nas caudas não dando tanta importâncias as distinções no "meio", local em que as densidades são altas. A razão para isso é que o quantil é uma função de mudança rápida de probabilidade p quando a densidade é esparsa (nas caudas), e uma função de mudança lenta de probabilidade p a qual a densidade é alta (no meio).

O Q-Q fornece um indicador muito sensível de discrepâncias de distribuição e fornece uma base útil para examinar a adequação de hipóteses compostas, em que os parâmetros não são especificados.

Além do uso do gráfico Q-Q com fins de comparação de distribuição entre dois conjuntos de dados observados, ele também é utilizado para comparações entre uma distribuição observada e teórica, como a normal, exponencial e qui-quadrado.

2.5. Análise de Correspondência

A análise de correspondência [HAIR et al. 2009] é uma técnica de interdependência geralmente utilizada para redução de dimensionalidade e mapeamento perceptual para base de dados composta por dados categóricos. Também conhecida como escalonamento ótimo, ela se destaca pela construção do mapa perceptual a partir da associação de objetos e um conjunto de características descritivas ou atributos especificados pelo autor. Sua aplicação principal é exibir a correspondência entre categorias em escalas nominais, e permite representar duas variáveis, com escalas nominais, em um mesmo diagrama.

Para o cálculo da análise de correspondência é necessário 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. A partir da tabela de contingência é utilizado a técnica estatística do qui-quadrado para padronizar os valores e criar um índice de similaridade que será utilizado no diagrama.

Para o cálculo do qui-quadrado deve ser realizado:

1. Cálculo do valor esperado: a probabilidade conjunta da combinação da coluna com a linha, a partir da probabilidade marginal para a coluna (total da coluna / total geral) vezes a probabilidade marginal para a linha (total da linha / total geral).
$$FrequnciaEsperada = \frac{TotalColuna * TotalLinha}{TotalGeral} \quad (1)$$
2. Cálculo da diferença entre as frequências esperadas calculada no passo anterior e a observada $diferena = FrequnciaEsperada - FrequnciaObservada$
3. Cálculo do valor do qui-quadrado é realizado a partir da razão entre a diferença calculada no passo anterior e a Frequência esperada calculada no primeiro passo.
$$Valorqui - quadrado = \frac{Diferena^2}{FrequnciaEsperada} \quad (2)$$

O valor do teste qui-quadrado é relacionado à intensidade de associação entre os atributos das variáveis. Caso o sinal desse valor seja negativo, ele representa a repulsão entre os atributos. Por outro lado, caso o sinal seja positivo, o valor representa a força de atração entre os atributos. A partir desse valor é gerado o mapa percentual de tal forma que quanto mais próximos estiverem dois atributos, mais similares o são.

3. Metodologia de Pesquisa

Como visto na tabela 1, a variável mais utilizada para predição de desempenho são as notas dos alunos, logo, para um melhor entendimento dos dados, técnicas de visualização fundamentadas em estatística permite ao humano uma melhor interpretação do fenômeno modelado.

Os dados utilizados são dos cursos de Licenciatura em Espanhol e Técnico de Segurança do Trabalho do Instituto Federal do Rio Grande do Norte no período de 2008 a

2016, nas modalidades presencial e EaD (Ensino a Distância) com um total de 34799 instâncias. Todo código e dados do trabalho estão disponíveis publicamente [Barros 2017].

Inicialmente foram utilizados os gráficos violino e Q-Q, a fim de ter uma visão geral da distribuição dos dados na disciplina. Após a primeira análise visual, foi utilizado o algoritmo *k-means* para a clusterização dos dados e, com as classes atribuídas, foi feita uma nova análise a partir do gráfico Q-Q e do Mapa Perceptual gerado pela Análise de Correspondência.

4. Resultados

Inicialmente foi utilizado o gráfico de violino por disciplina, a fim de visualizar o comportamento da variável aleatória média final por componente curricular.

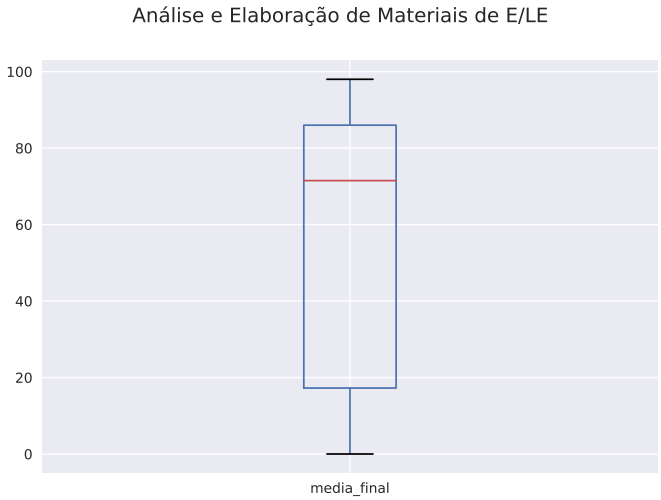

Figura 1. Gráfico Boxplot

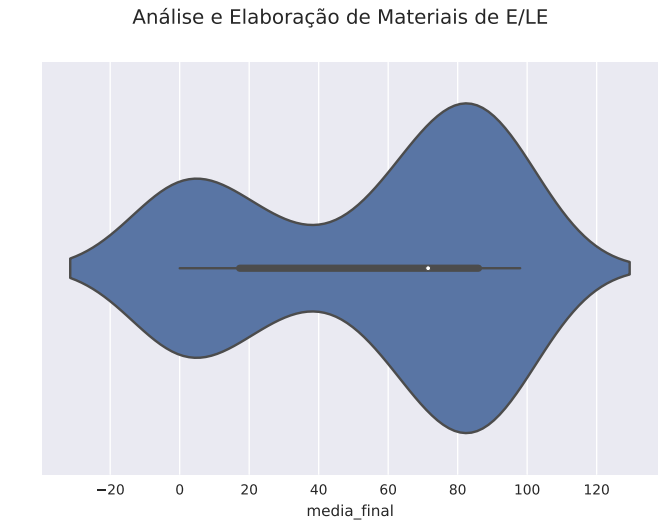

Figura 2. Gráfico Violino

Como visto na figura 2 nota-se que, diferente do senso comum, a distribuição dos valores das notas não segue uma distribuição normal, mas sim uma distribuição bimodal. Interessante notar que se fosse apenas utilizado apenas o boxplot como na figura 1, haveria uma interpretações errônea em relação aos dados e não seria verificado a existência dos possíveis dois grupos na distribuição. Na figura 3 é exibido o gráfico Q-Q entre os dados observados e a distribuição teórica Normal. Uma vez que os pontos não estão dispostos sobre uma reta, significa que a distribuição dos dados observados não segue uma Normal.

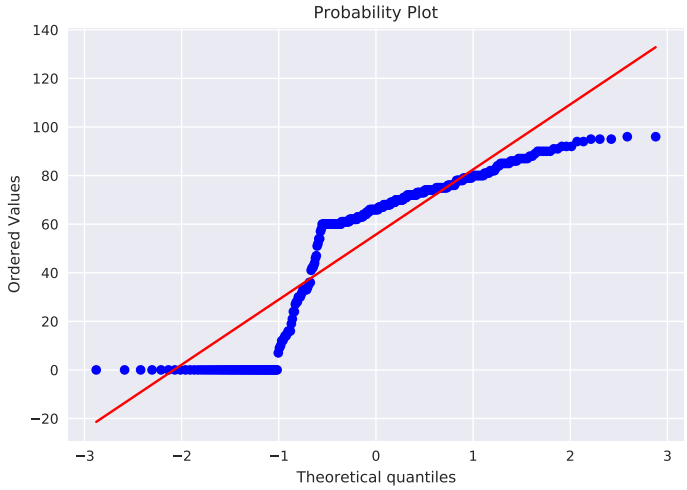

Figura 3. Gráfico Q-Q entre notas e a distribuição Normal

Uma vez verificado esse padrão bimodal da distribuição dos valores de notas, foi utilizado o algoritmo *k-means*, a fim de confirmar se era possível a clusterização dos dados em dois grupos como visto na figura 4

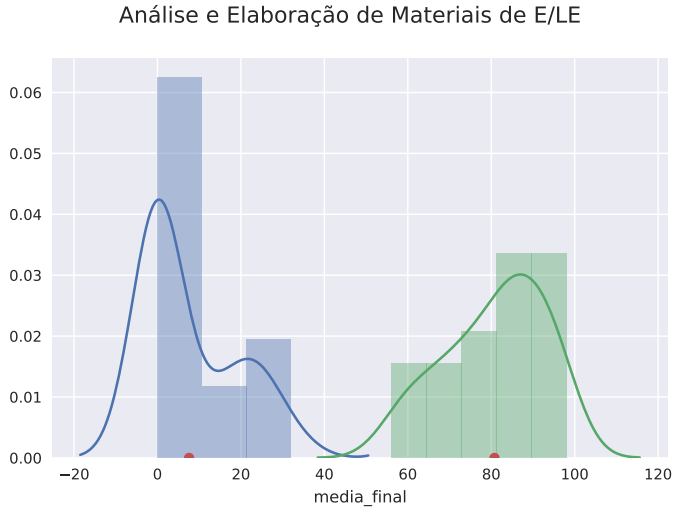

Figura 4. Clusterização em 2 classes das notas de uma disciplina

A partir dos dois grupos gerados é utilizado o gráfico Q-Q (figura 5) entre os dados observados e a distribuição teórica Normal.

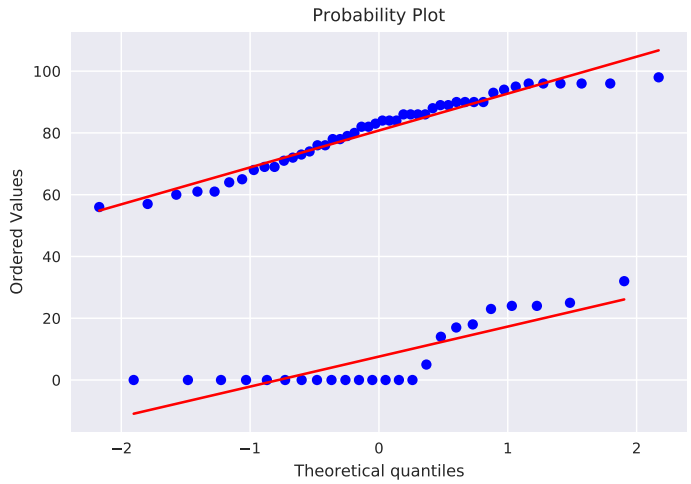

Figura 5. Gráfico Q-Q para cada grupo gerado pelo k-means e a distribuição Normal

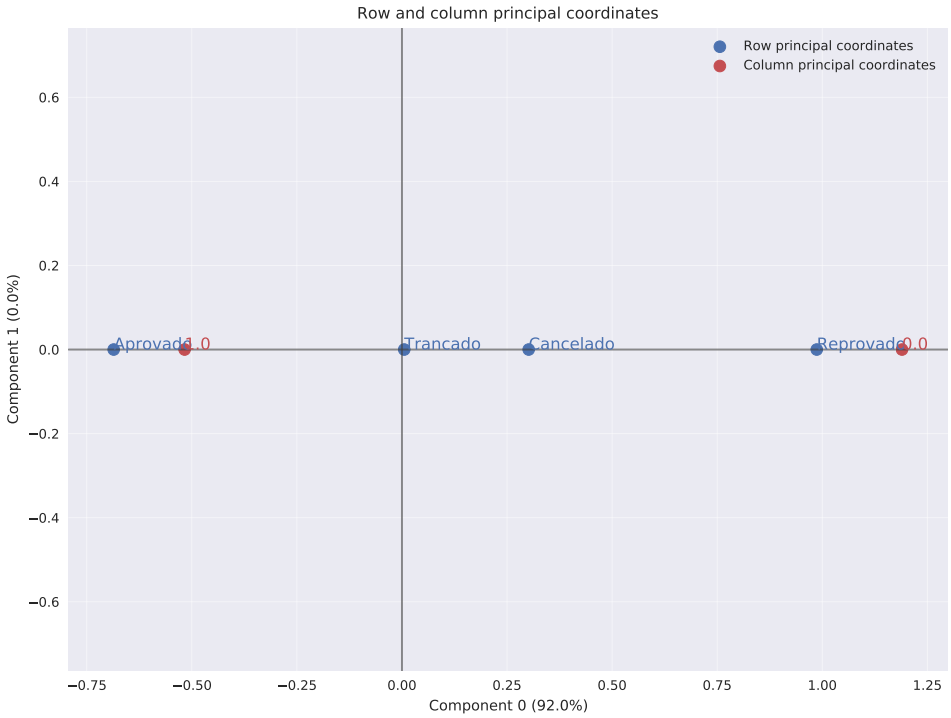

Figura 6. Mapa perceptual gerado a partir da análise de correspondência

O primeiro grupo (provavelmente dos alunos aprovados) se ajusta bem a reta, logo, a distribuição pode ser bem modelada pela distribuição Normal. Já os dados dos alunos possivelmente reprovados, não se ajusta tão bem, padrão encontrado em várias disciplinas analisadas. É possível interpretar que entre os alunos reprovados há mais de um grupo.

Acreditasse que esses dois grupos possam ser representados pelos alunos que tentaram até o final da disciplina, mas não conseguiram êxito, e os alunos que ao longo da disciplina evadiram.

Após o entendimento que há, pelo menos, dois grupos bem distintos que compõe a distribuição das notas, é utilizado a técnica de análise de correspondência, a fim de certificar a informação que os alunos que possuem o label "Aprovado"na base de dados original são os alunos classificado pelo *k-means* como classe 1, e os alunos com label "Reprovado"são os classificados pela classe 0. A figura 6 mostra o mapa perceptual gerado a partir da análise de correspondência, a qual apresenta de forma visual indícios de similaridade entre os valores de duas variáveis com escalas nominais. Neste mapa perceptual mostra que a classe 0 está mais próxima ao valor "reprovado", em contraposição da classe 1 que está mais próximo ao valor "aprovado", corroborando o resultado do *k-means*.

5. Conclusão

A partir dos gráficos apresentados foi evidenciado o potencial do uso de técnicas de visualização de dados para detectar padrões nos dados que não seria possível analisando apenas as tabelas de dados. O uso do gráfico violino permitiu a detecção de padrão bimodal corroborados pelo gráfico Q-Q, o que não seria possível se fosse utilizado apenas o *boxplot*. Uma vez que foi detectado o aparecimento de mais de um grupo nos dados, a técnica de *k-means* mostrou eficaz em separar esses grupos, o qual foi corroborado pela técnica de análise de correspondência.

Como trabalho futuro será feita uma melhor investigação relacionada ao grupo de alunos reprovados, a partir da hipótese que há dois grupos: os alunos evadidos e os alunos que de fato cursaram toda a disciplina, mas não obtiveram êxito.

Referências

Barros, T. M. (2017). Modelo ifrn. `https://github.com/tmedeirosb/modelo_ifrn`.

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 –.

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

Hintze, J. L. and Nelson, R. D. (1998). Violin plots: A box plot-density trace synergis. *The American Statistician*, 52(2):181–184.

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.

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.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python*. O'REILLY, 1th edition.

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.

WILK, M. B. and GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1.

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.