Análise Exploratória de Dados Educacionais para o Letramento Estatístico: Um Estudo de Caso

Elvis Medeiros de Melo[1], Isabel Dillmann Nunes[1], Luiz Affonso Henderson Guedes de Oliveira[2]

[1]Instituto Metrópole Digital – Universidade Federal do Rio Grande do Norte (UFRN)
Av. Sen. Salgado Filho, 3000 – Lagoa Nova, CEP: 59.078-970 – Natal – RN – Brasil

[2]Departamento de Engenharia da Computação e Automação – UFRN – Natal – RN – Brasil

elvismedeiros.mm@gmail.com, bel@imd.ufrn.br, affonso@dca.ufrn.br

Abstract. *Statistical literacy as well as reading and interpretation skills (HLIT) are essential for 21st Century citizens. Therefore, this paper aims to perform an exploratory analysis on the data of a diagnostic instrument applied to students of a public school in Natal/RN in a census way, configuring itself in a case study. With the data, we conducted tests with Python 3.0 libraries, performing data based HLIT interpretation, and manipulating the data in Orange tools. We observed a promising environment for research and prospecting activities aimed at groups of students and school years.*

Resumo. *O letramento estatístico, assim como habilidades sobre leitura e interpretação de tabelas (HLIT) são essenciais para os cidadãos do Século XXI. Por isso, este trabalho tem o objetivo de realizar uma análise exploratória sobre os dados de um instrumento diagnóstico aplicado a estudantes de uma escola pública de Natal/RN de maneira censitária, se configurando em um estudo de caso. Com os dados em mãos, realizamos testes com bibliotecas do Python 3.0, realizando a interpretação das HLIT com base nos dados, além de manipulação dos dados em ferramentas do Orange. Observamos um ambiente promissor de pesquisa e prospecção de atividades direcionadas a determinados grupos de estudantes e anos escolares.*

1. Introdução

Quando nos deparamos com o ensino da Matemática, um dos principais problemas encontrados é a significância de seus conteúdos [Unesco 2016]. Entre os seus conteúdos, encontra-se a Estatística, com seu ensino atrelado ao bloco de conteúdo do Tratamento da Informação [Brasil 1998], mas a ênfase destas aulas ainda tem estado nas tabelas, gráficos e alguns cálculos com medidas de posição destoantes da realidade do aluno [Lopes e Carvalho 2009]. Apesar desse cenário, o estudo contextualizado e a proficiência em Estatística é importante para o cidadão do Século XXI na sociedade da informação [Unesco 2015].

A Estatística está presente no nosso cotidiano, e para exercício da plena cidadania, se faz necessário conhecê-la e gozar das habilidades atreladas aos seus conteúdos, entre elas destacamos a habilidade de leitura interpretativa de uma tabela (HLIT). No âmbito do projeto *Desenvolvimento Profissional de Professores que Ensinam Matemática* (D-ESTAT), a pesquisa acontece no contexto da Universidade-Escola, com fundamento principal sendo a colaboração por meio de interações questionadoras sobre as práticas

DOI: 10.5753/cbie.wcbie.2019.129

educativas sobre Estatística que os docentes da escola desenvolvem juntamente com seus alunos.

O projeto D-ESTAT trabalha com a formação de professores sobre conceitos estatísticos relacionados ao cotidiano. Em seu instrumento diagnóstico, aplicado com estudantes de uma escola pública do município de Natal/RN, foram abordados conteúdos estatísticos como medidas de tendência central, leitura, interpretação e construção de gráficos e tabelas de distribuição de frequência relativa, absoluta, entre outras habilidades fundamentais para o cidadão do Século XXI. Com o intuito de analisar os resultados dos estudantes, mapeando as HLIT que permeiam as questões, o objetivo deste trabalho é realizar uma Análise Exploratória de Dados Educacionais (AEDE) dos com ênfase nas HLIT, importantes habilidades relacionadas ao letramento estatístico, que ajudará na tomada de decisão sobre quais abordagens os professores deverão tomar sobre esses resultados.

O artigo está divido em 5 sessões, das quais se faz presente esta introdução, a fundamentação teórica, na qual discorremos brevemente sobre o letramento estatístico e o método da AEDE; uma sessão na qual discorremos sobre os procedimentos metodológicos adotados para a AEDE; seguida dos resultados encontrados, considerações e referências consultadas.

2. Letramento Estatístico e Análise Exploratória de Dados Educacionais

Com vista a formação estatística na sociedade da informação, Lopes (2002) define o letramento estatístico como a capacidade de reconhecer e classificar dados como quantitativos ou qualitativos, discretos ou contínuos, saber como o tipo de dado conduz a um tipo específico de tabela, gráfico, ou medida estatística. Além disso, saber ler e interpretar tabelas e gráficos, entender as medidas de posição e dispersão, usar as ideias de aleatoriedade, chance e probabilidade para fazer julgamentos sobre eventos incertos e relacionar a amostra com a população. Para a autora, é muito mais do que possuir competências de cálculo, é preciso adquirir habilidades para compreender a leitura e a interpretação numérica necessária para o exercício pleno da cidadania com responsabilidade social na tomada de decisões [Lopes 2002]. Essas habilidades se fazem necessárias e estão contempladas no instrumento diagnóstico do D-ESTAT.

Gal (2002) aponta que conhecimentos estatísticos devem ser intrínsecos a uma educação para a cidadania, e entre eles a competência de leitura e interpretação de tabelas faz parte de um letramento estatístico. Para que seja realizada uma AEDE, é necessário possibilitar a geração de situações de aprendizagem contextualizadas em temas que sejam de interesse, lançar mão de um forte apoio às representações gráficas (que facilitam a percepção da variabilidade no conjunto de dados observados), empregar as estatísticas de ordem (que aportam maior facilidade na atribuição de significado pelo aluno da Escola Básica) e utilizar diferentes escalas de categorização das variáveis para o estudo dos dados observados [Gal 2002].

A análise exploratória tem o objetivo de entender quais são os dados, quais as tendências a partir de todas as perspectivas e com todas as ferramentas possíveis, incluindo as já existentes [Cox 2017]. O propósito é extrair toda a informação possível, gerar novas hipóteses no sentido de construir conjecturas sobre as observações que dispomos [Batanero, Estepa e Godino 1991]. Com o intuito de identificar quais variáveis estão relacionadas às HLIT no questionário diagnóstico do D-ESTAT, além do processo delimitado para a AEDE, descrevemos os procedimentos metodológicos na próxima seção.

3. Procedimentos Metodológicos

Devido à natureza dos dados e do estudo neste artigo, alinhamos a pesquisa como um Estudo de Caso, pois tem o objetivo de investigar um fenômeno contemporâneo sob o qual o pesquisador não tenha controle, comprovando ou contrastando as relações evidenciadas no caso e compreendendo o evento em estudo, desenvolvendo inferências a respeito do fenômeno observado [Yazan 2016]. Para sua análise, Yin (2005) recomenda que a partir dos dados provenientes sejam criados e analisados dados estatísticos. No caso, faremos inferências estatísticas sobre os instrumentos diagnósticos dos estudantes de uma escola pública natalense que está inserida no projeto D-ESTAT.

Para análise do *dataset* com as respostas de todos os estudantes da escola, utilizamos bibliotecas da linguagem de programação *Python 3.0*, denominadas *Pandas* e *Numpy* para análises exploratórias iniciais, além do *seabourn*, *matplot* para plotagem de gráficos e *missingno* para análise da que permitem obter uma visão geral rápida da integridade do conjunto de dados. A princípio, o *dataset* possui os dados de todos os estudantes, aos quais responderam o teste diagnóstico em seu conjunto contendo 4 perguntas contextualizadas, configurando-se em 27 itens dissertativos, atribuindo-lhes valores de 1 ou 0 para acertos e erros, respectivamente.

Dentro do escopo deste instrumento, analisamos as questões que trabalhavam diretamente os conceitos que o D-ESTAT propõe. Para tanto, levantamos as habilidades referentes ao letramento estatístico presente nos itens das questões e fizemos uma tabulação de quais questões trabalhavam determinadas competências. Dentre as 27 questões dos instrumentos, 6 questões tratam sobre a HLIT. Calculamos a média geral do aluno no diagnóstico, levando em consideração que cada questão tem peso igual, e a média sobre as questões referentes a HLIT (no caso, são as questões *q11*, *q22*, *q27*, *q28*, *q32* e *q35* do diagnóstico). Neste estudo, focaremos a análise nesta competência. Ela também está presente em matrizes de referências como a Avaliação Nacional da Educação Básica (Aneb): “Resolver problema envolvendo informações apresentadas em tabelas e/ou gráficos” (D36) e “Associar informações apresentadas em listas e/ou tabelas simples aos gráficos que as representam e vice-versa” (D37) [Brasil 2011]; na Base Nacional Comum Curricular (BNCC): “Interpretar e resolver situações que envolvam dados de pesquisas sobre contextos ambientais, sustentabilidade, trânsito, consumo responsável, entre outros, apresentadas pela mídia em tabelas e em diferentes tipos de gráficos e redigir textos escritos com o objetivo de sintetizar conclusões” (EF06MA32) [Brasil 2017]; e na matriz do PISA: “Interpretar dados e evidências cientificamente” [OECD 2015]. Vale salientar que tratamos dados de 307 estudantes dos anos finais do Ensino Fundamental e que o instrumento diagnóstico foi aplicado no dia 25 de setembro de 2018 a todos os anos de escolaridade e turmas da escola.

Para a categorização dos estudantes, possuímos dados de turma, ano de escolaridade e o seu número identificador. Além disso, para cada aluno, as respostas de cada pergunta encontram-se tabuladas com zeros, uns e espaços vazios, no qual zero significa que o aluno errou a questão, 1 que acertou e um espaço vazio significa que o aluno deixou a questão em branco. Ao utilizar a ferramenta *missingno*, observamos que existiam muitos dados faltando. Como opção metodológica, optamos por substituir os espaços em branco por zeros, considerando que o aluno que não respondeu recebeu a nota zero (ver Figura 1a e Figura 1b).

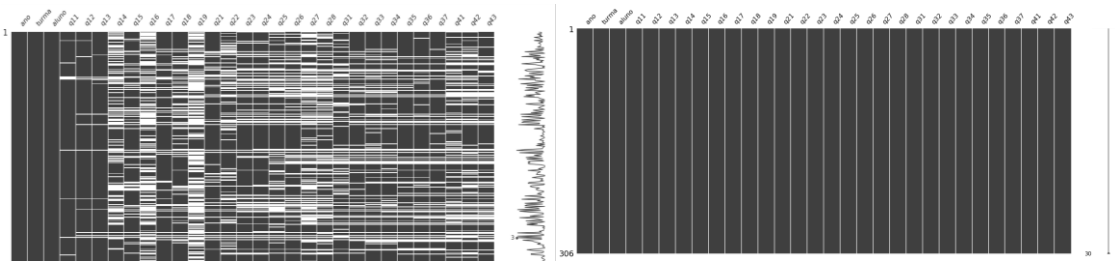

Figura 1. Integridade do *dataset* (a) antes - esquerda e (b) direita - abaixo

Como análise inicial, além da correção da integridade do *dataset*, traçamos correlações entre as variáveis utilizando *Pandas* e *Numpy*. Para tal, a correlação de *Spearman* se faz mais adequada a natureza dos dados, com o intuito de medir a intensidade da relação entre duas variáveis categóricas. Plotamos gráficos de correlação com ajuda do *seabourn* e do *matplot*. Vale salientar que essas técnicas foram escolhidas de acordo com o que propõe Barros, Silva e Guedes (2018) em Revisão Sistemática de Literatura sobre como proceder em uma AEDE.

Para testagem de modelos na AEDE dos dados do diagnóstico, utilizamos *Orange*. Trata-se de uma ferramenta *open source* baseado em programação visual que permite carregar bases de dados, fazer transformações e pré-processamento, visualizar os dados de forma interativa e executar algoritmos de *Machine Learning*. Com os dados, optamos pelo método de *k-Means* para identificação de 5 principais grupos (*clusters*) de estudantes (por *default*), segundo suas respostas no teste diagnóstico por sugestão da própria ferramenta. Essa técnica, neste escopo, consiste em gerar de agrupamentos de alunos com desempenhos semelhantes.

Para visualização dos grupos em relação as HLIT no diagnóstico e quais atributos tinham maior influência sobre, utilizamos o *FreeViz*[1], que é um método de otimização para encontrar uma melhor projeção linear dos dados em espaços para uma representação bidimensional, de forma a aproximar ao máximo os dados de mesma classe, e afastar ao máximo os dados de classes distintas. Para plotagem no *FreeViz*, cruzamos as questões relacionadas as HLIT e os *clusters* com o intuito de qualificar os agrupamentos e identificar características comuns aos participantes de cada agrupamento.

Com o intuito de verificar qual variável mais influencia nos resultados dos *clusters* para as HLIT, utilizamos o algoritmo *Tree* implementado no software *Orange*, usando *Machine Learning* para uma árvore de decisão simples com base nos dados e nos *clusters* delineados pelo *k-Means*. Para a caracterização dos *clusters*, utilizamos o *silhouette plot* e *scatter plot* para análise visual dos agrupamentos e sua completude, identificando-os dentro das turmas e anos de escolaridade. Com isso, prospectamos atividades específicas que os professores realizaram no âmbito do D-ESTAT em formação continuada. Os resultados da AEDE estão relatados a seguir.

4. Resultados e Discussões

Ao traçar o gráfico de correlação de *Spearman* entre todos os atributos do *dataset*, observamos que há pouca uma fraca correlação entre as respostas dos estudantes e os itens do diagnóstico (ver Figura 2), o que nos mostra que os itens do diagnóstico possuem determinados graus independência entre si.

[1] Documentação disponível em: <https://bit.ly/2mgoU3x>. Acesso em: 30 set. 2019

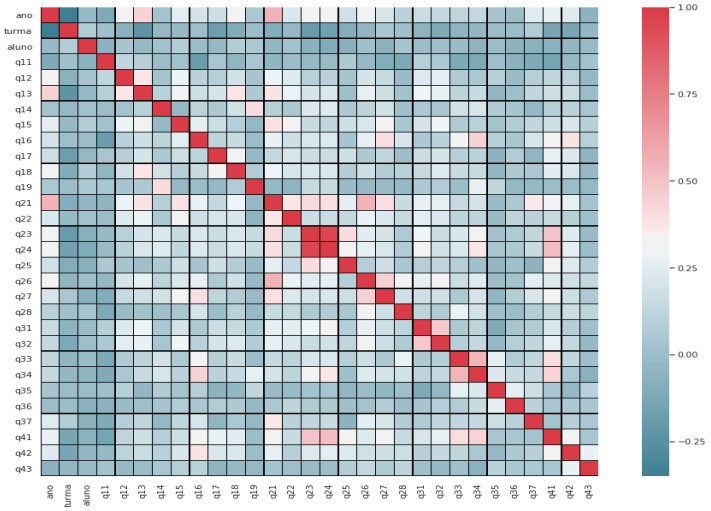

Figura 2. Correlação entre todos os itens do *dataset*

Apesar disso, ao analisar mais detalhadamente, notamos que existem questões pontuais que têm uma forte correlação no diagnóstico, como por exemplo, a questão *q23* e *q24*, na qual ambas solicitam o cálculo de uma média. Pode-se inferir que os estudantes que acertaram a *q23* também acertaram a *q24*. Decidimos cruzar apenas os dados específicos relacionados diretamente com as HLIT para se ter uma visão mais específica do objeto de estudo.

Filtrando os dados para as perguntas que tratavam da HLIT, observamos na figura 4 que é fraco o grau de correlação entre as variáveis, atentando para os resultados da questão *q27* e da *q21*, que possuem correlação positiva de 0,23 e da *q27* com a *q28*, com 0,18. Isso mostra que essas variáveis se movem juntas, porém o grau de intensidade é fraco.

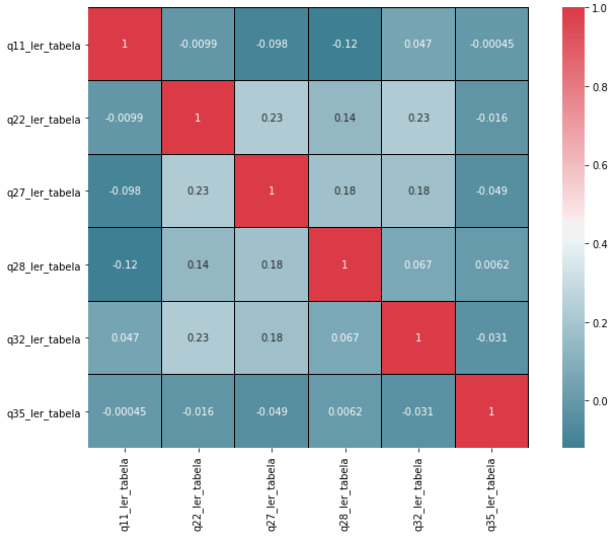

Figura 3. Correlações das HLIT no *dataset*

Quando inserimos as variáveis de média geral e da média das questões referentes ao HLIT, observamos (ver Figura 4) que o cenário muda. Há muita relação entre todas as variáveis da HLIT tanto com a média geral tanto com a média da HLIT, o que sugere uma exploração nesse caminho.

media_geral

1	+0.765	media_geral	media_ler_tabela
4	+0.552	media_geral	q22
5	+0.548	media_geral	q32
6	+0.412	media_geral	q27
10	+0.263	media_geral	q28
16	+0.142	media_geral	q11
18	+0.126	media_geral	q35

media_ler_tabela

1	+0.765	media_geral	media_ler_tabela
2	+0.650	media_ler_tabela	q22
3	+0.607	media_ler_tabela	q32
4	+0.398	media_ler_tabela	q27
5	+0.389	media_ler_tabela	q28
6	+0.337	media_ler_tabela	q11
7	+0.257	media_ler_tabela	q35

Figura 4. Correlação quando insere média geral e média da HLIT

Para analisar quais outros fatores influenciam os alunos sobre o desempenho em HLIT, carregamos no *Orange* o *dataset* modelado, seguindo o *workflow* abaixo (ver Figura 5).

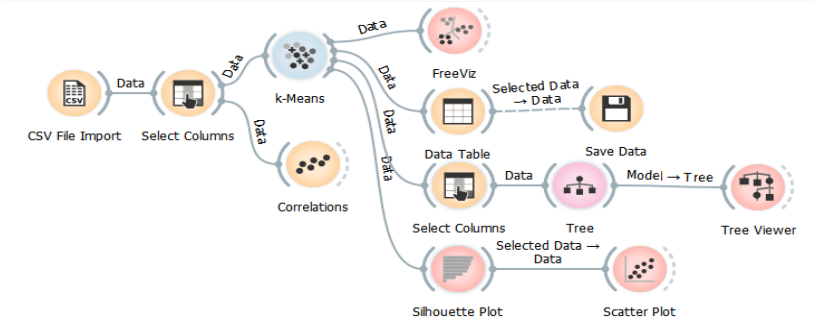

Figura 5. Workflow das análises no *Orange*

Propomos agrupar os estudantes com o algoritmo de clusterização *k-Means*, criando um agrupamento de alunos por características similares e analisar quais as variáveis que interferem em determinados grupos de estudantes (ver Figura 6). Para tanto, para fins de otimização desse trabalho e das análises, foram definidos 5 *clusters* na medida que o *silhouette plot* ficasse o mais próximo possível de 1. Com a aplicação do algoritmo *k-Means*, foram formados os *clusters* C1 (45 alunos), C2 (56 alunos), C3 (46 alunos), C4 (111 alunos) e C5 (48 alunos). Utilizamos a distância euclidiana para o *silhouette plot*, assinalada por *default* pela ferramenta.

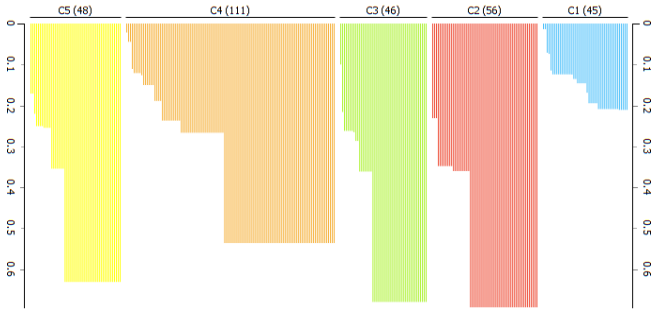

Figura 6. *Silhouette plot* dos *Clusters* delineados através do k-Means

Ao selecionar os atributos que interessam para a análise, selecionamos as questões *q11, q22, q27, q28, q32* e *q35* que tratam sobre as HLIT, os *clusters* delineados como variável *target* e como meta atributos as variáveis com as informações do *ano*, *media_geral*, *media_ler_tabela* e a *turma*.

Ao usar o *FreeViz* para o *plot* dos *clusters* no espaço, ao aumentar o raio de significância dos dados, observamos que os resultados das questões *q11* e *q22* são os que mais influenciam na distribuição dos alunos nos *clusters*, respectivamente. Além disso, observamos uma baixa relevância das demais variáveis na configuração do *FreeViz*.

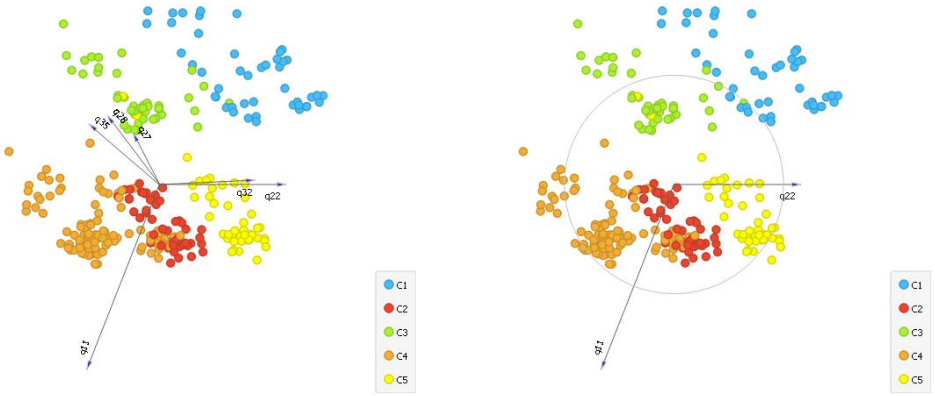

Figura 7. Visualização das HLIT no *FreeViz* por *clusters*

De acordo com a figura 7, podemos inferir que os alunos do C4 foram mais influenciados pelas respostas *q11* que os alunos do C1, por exemplo, que ficaram opostos na distribuição espacial do *FreeViz*. Além disso, os alunos do C3 podem ter sido mais influenciados pelos demais atributos das HLIT, visto que estão tendenciosos a três variáveis que não estão em evidência. Vale salientar que a ferramenta faz uma otimização da melhor visualização dos dados por *clusters* em um plano 2D.

Com o intuito de caracterizar os *clusters* e identificar quais as necessidades formativas relacionadas às HLIT, levando em consideração os acertos e erros da *q22* e *q11*, conforme apontou o *FreeViz* como variáveis de maior importância na tendência dos dados, observamos que os indivíduos de C2 e C5 tiveram um bom aproveitamento (ver Figura 8).

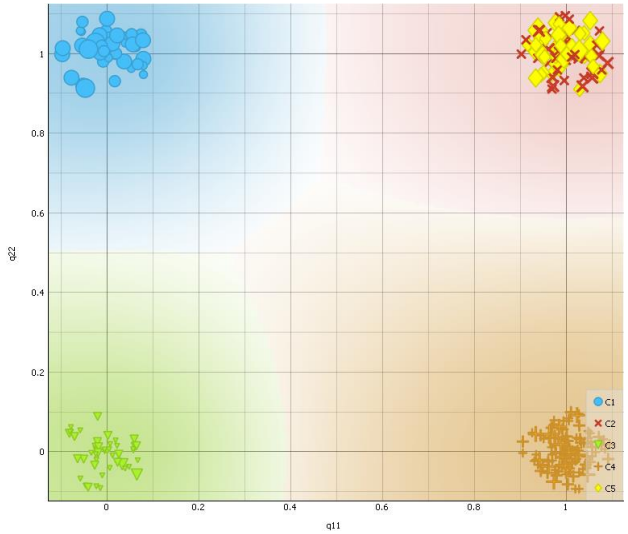

Figura 8. Aproveitamento em *q22* e q11 por *cluster*

Observamos um cenário contrário preocupante para os indivíduos de C3. Ao cruzar os dados da média geral com a média de aproveitamento nas HLIT (ver Figura 9), percebemos um bom agrupamento dos estudantes do C5 em relação as suas médias geral e das HLIT, os alunos de C1, C2 e C4 possuem comportamentos que precisam ser

investigados. Aqui, queremos destacar que os estudantes de C3 tiveram problemas com a média de HLIT.

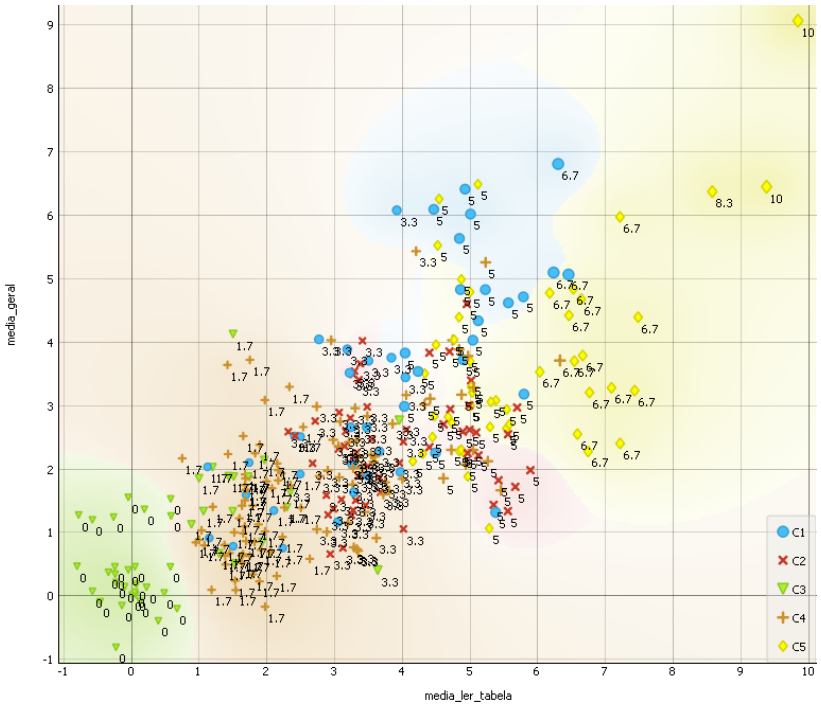

Figura 9. Aproveitamento em média geral e média em HLIT, por *cluster*, sendo destacada a média em HLIT

Ao analisar o comportamento dos *clusters*, relacionando q11 e q22 com a média de HLIT (ver Figura 10), observamos que os estudantes de C3 erraram as duas questões que mais influenciam no agrupamento dos estudantes pelo *k-Means*.

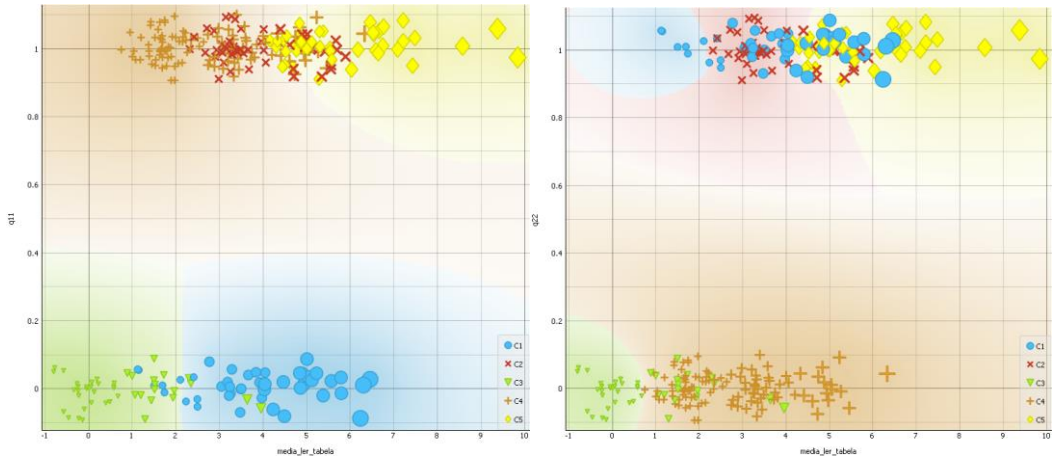

Figura 10. Média geral dos *clusters* relacionado a (a) q11 a esquerda e (b) q22 a direita

Ao verificar quais demais fatores podem ter influenciado esses estudantes, plotamos o *scatter plot* dos estudantes do C3 em relação às demais perguntas sobre as HLIT (ver Figura 11) em relação ao seu ano de escolaridade e turma, sugerindo ações específicas para cada professor. Ao analisar, observamos que as questões *q32* e *q35* possuem maior influência na nota dos alunos do 6° e do 9° ano. Isso mostra que os professores desses anos de escolaridade podem trabalhar questões para as HLIT com o contexto similar ao da segunda questão do diagnóstico, obtendo apenas 2 acertos em seus itens.

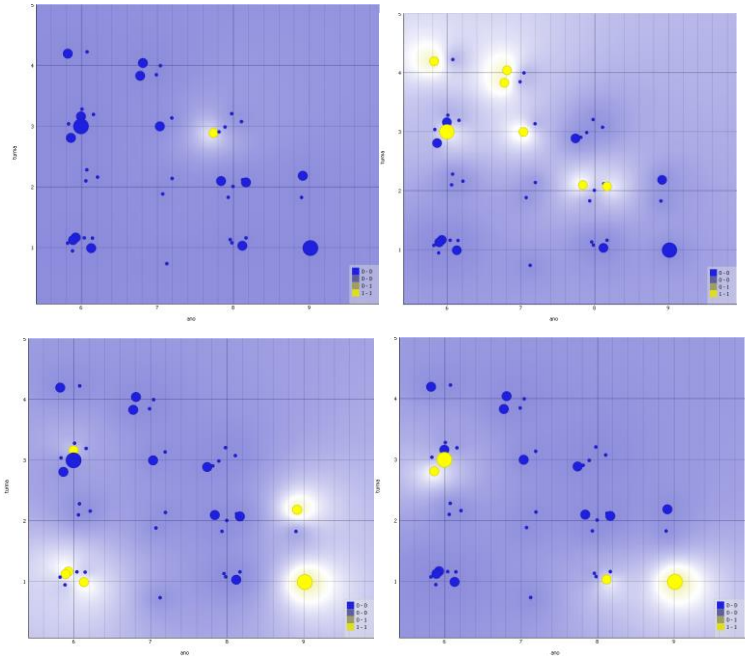

Figura 11. Desempenho de C3 em *q27*, *q28*, *q32* e *q35*, respectivamente

Com o intuito de facilitar a identificação dos estudantes nos *clusters*, utilizamos o algoritmo *Tree* implementado no *Orange* para traçar uma árvore de decisão, colocando como *target* os *clusters* e meta atributos as HLIT. Assim, identificamos que a principal característica (a que está no nó da árvore de decisão) para saber se um aluno é de um agrupamento é o acerto ou erro de *q22*. Ao descer os níveis, observamos que o segundo critério é *q11*. Para o algoritmo ter certeza, ele pergunta a resposta em *q32*. Esses resultados confirmam as predições do *FreeViz* pelo algoritmo *k-Means*.

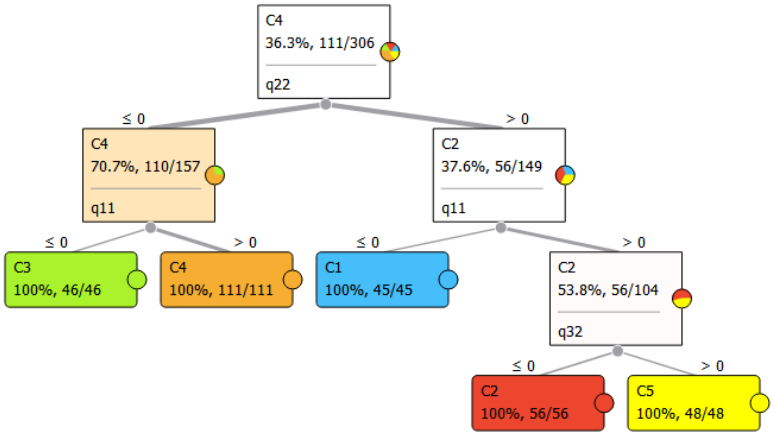

Figura 12. Algoritmo de *Machine learning Tree* para identificação dos alunos nos *clusters*

Isso nos mostra que as HLIT espalhadas em determinadas questões em contextos diferentes interferem no agrupamento e na forma de diagnóstico do D-ESTAT. Além disso, o modelo dá a possibilidade de encontrar o aluno do *cluster* com apenas três verificações, no máximo, e trabalhar determinados problemas para um grupo de alunos específico.

5. Considerações

A utilização dos métodos delineados, principalmente se forem tomados de forma automática, pode ajudar gestores e professores a pensarem em estratégias mais

localizadas para a formação de professores ou aulas para um grupo específico, levando em consideração o seus níveis de letramento estatístico, que é proposto em estratégias nas matrizes de referência da Aneb, nas habilidades da BNCC e matriz do PISA. Observamos também uma grande aplicabilidade na tomada de decisões sobre o que ensinar e ao tipo de questão que deve ser trabalhada para cada *cluster* delineado.

Como propostas para trabalhos futuros, pretendemos expandir as análises para as demais habilidades mapeadas no instrumento diagnóstico do D-ESTAT, validar os modelos a partir de dados provenientes do instrumento diagnóstico aplicado em 2019 e pensar em atividades que trabalhem as HLIT nas turmas e agrupamentos diagnosticados nas formações do D-ESTAT.

Referências

Barros, T. M.; Silva, I.; Guedes, L.A. (2018) "Uso da Técnica de Análise de Correspondência para Análise Exploratória de Dados no Contexto Educacional". In: VII Congresso Brasileiro de Informática na Educação (CBIE 2018). Anais dos Workshops do VII Congresso Brasileiro de Informática na Educação (WCBIE 2018).

Base Nacional Comum Curricular (BNCC, 2017). Disponível em: <http://basenacionalcomum.mec.gov.br>. Acesso em: 23 set. 2019.

Brasil (1998). Ministério da Educação (MEC). Conselho Nacional de Educação (CNE). Diretrizes Curriculares Nacionais para o Ensino Fundamental. Resolução n. 2, de 7 abril de 1998. Institui as Diretrizes Curriculares Nacionais para o Ensino Fundamental. Diário Oficial da União, Brasília/DF, 1998.

Batanero, C.; Estepa, A.; Godino, J. D. (1991). "Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria". Suma, v. 9, p. 25-31.

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

Gal, I. (2002) "Adult statistical literacy: meaning, components, responsibilities". International Statistical Review, v. 1, n. 70, p. 1-25, 2002.

Lopes, C. A. E. (2002). "Literacia estatística e o INAF 2002". In: FONSECA, M. C. F. R. Letramento no Brasil – habilidades matemáticas: reflexões a partir do INAF". São Paulo: Global: Ação educativa Assessoria, Pesquisa e Informação: Instituto Paulo Montenegro, p. 187-197.

Lopes, C. E.; Carvalho, C. (2009). "Escritas e Leitura na Educação Matemática". Belo Horizonte: Autêntica.

OECD (2015). "PISA 2015 - Programa Internacional de Avaliação de Estudantes". In: Matriz de Avaliação de Ciências: Science Framework 2013.

Unesco (2015). "Educação para a cidadania global: preparando alunos para os desafios do século XXI". Brasília.

Unesco (2016). "Os desafios do ensino de matemática na educação básica". São Carlos: EdUFSCar, 2016, 114 p.

Yazan, B. (2016). Três abordagens do método de estudo de caso em educação: Yin, Merriam e Stake. Revista Meta: Avaliação, , v. 8, n. 22, p. 149-182, may 2016.

Yin. R. K. (2005). Estudo de caso: planejamento e métodos. 3 ed., Porto Alegre: Bookman.