Políticas para Adoção de *Learning Analytics*: Uma Proposta Baseada nas Opiniões dos Estudantes

Thiago Kelvin[1], **Flávio Leandro**[1], **Roberta Fagundes**[2], **Elyda Freitas**[1]

[1]Instituto de Informática – Universidade de Pernambuco (UPE)
- 55.002-971 - Caruaru – PE – Brasil

[2]Departamento de Engenharia da Computação - Universidade de Pernambuco
- 50.720-001 - Recife - PE.

{thiago.kelvin, flavio.leandromorais, roberta.fagundes, elyda.freitas}@upe.br

Abstract. *Learning Analytics (LA) aims to analyze educational data to support teachers and students in the teaching and learning process. Aspiring LA effective adoption, it is essential to consider the opinion of the stakeholders. For that so, this paper aims to have knowledge about the expectation of students from a Brazilian public Higher Education Institution (HEI) concerning the use of their data. The ultimate goal is to propose guidelines to define policies that support the adoption of LA and attend to the expectation of these students. To achieve this goal, a case study was conducted using the SHEILA project questionnaire to collect the data. Data were analyzed by statistical techniques and Educational Data Mining.*

Resumo. *Learning Analytics (LA) visa a análise de dados educacionais para melhorar o processo de ensino e aprendizagem. Para sua efetiva adoção, é essencial considerar a opinião dos stakeholders. Assim, este artigo tem por objetivo conhecer as expectativas dos estudantes de uma Instituição de Ensino Superior (IES) pública brasileira sobre o uso de seus dados, com o objetivo final de propor diretrizes para a definição de políticas que apoiem a adoção de LA e atendam às expectativas desses estudantes. Para isso, conduziu-se um estudo de caso com a utilização do questionário do projeto SHEILA para coleta de dados; a análise de dados foi realizada por meio de técnicas estatísticas e Mineração de Dados Educacionais (MDE).*

1. Introdução

O crescimento do ensino *online* trouxe consigo a necessidade de apoiar o desenvolvimento dos estudantes. Assim, *Learning Analytics* (LA) emerge como uma disciplina focada em coletar e analisar dados educacionais sobre as ações dos estudantes e seus contextos, procurando entender e otimizar o aprendizado e o ambiente em que ocorre [1]. Além disso, serve como solução para abordar problemas sobre a conservação, desenvolvimento e aperfeiçoamento do sucesso do estudante [Ferguson 2012].

[Avella et al. 2016] explicam que LA permite objetivos como: identificar estudantes em risco de desistência ou baixo desempenho; auxiliar na motivação do estudante, por meio de suporte na definição de seus próprios objetivos, além de organizar o tempo e os

[1]https://www.solaresearch.org/about/what-is-learning-analytics/

DOI: 10.5753/sbie.2021.218599

recursos utilizados no aprendizado. Apesar das possibilidades, [Tsai and Gasevic 2017] afirmam que LA ainda é uma disciplina imatura e há pouca evidência empírica de sucesso e vários desafios em sua adoção. [Falcao et al. 2019] também explicam que LA não é uma disciplina bem conhecida entre estudantes brasileiros e regulamentos no uso de dados educacionais não são tão transparentes e restritos quanto os da Europa.

Para ajudar a sobrepor essas dificuldades, pode ser utilizado o *framework* SHEILA[2] (*Supporting Higher Education to Integrate Learning Analytics*, em português, Apoiando Educação Superior para a Integração de *Learning Analytics*), que visa auxiliar na implementação e pesquisa de LA [Tsai et al. 2018b]. Para o SHEILA, três fatores devem ser considerados para a adoção correta de LA: (i) demanda de recursos necessários, como recursos humanos, financeiros ou infraestrutura tecnológica disponível; (ii) problemas de ética e privacidade de dados; e (iii) engajamento dos *stakeholders*. O SHEILA utiliza entrevistas, grupos focais, questionários e pesquisas com estudantes, professores e gestores para apoiar as instituições na adoção de LA.

Desse modo, [Tsai et al. 2018b] explicam que é essencial ouvir os *stakeholders*, sendo eles a principal fonte de dados para metodologias de LA. É preciso entender se estes estão de acordo sobre o modo como seus dados são usados, quais dados deveriam ou não ser coletados e sobre a necessidade de consentimento nesses assuntos. São informações essenciais para a correta e ética implementação de LA [Tsai and Gasevic 2017], [Hoel et al. 2015]. [Lim and Tinio 2018] explicam ainda que o contexto histórico, cultural, político, social e econômico influencia nas metodologias que podem ser aplicadas numa instituição. Assim, os instrumentos do SHEILA auxiliam na coleta de informações e adaptação de práticas pedagógicas ao contexto dos *stakeholders*.

Portanto, este artigo visa entender, por meio de um estudo de caso com estudantes dos cursos de computação de uma Instituição de Ensino Superior (IES) pública brasileira, suas opiniões quanto ao uso de seus dados em projetos de LA. O propósito final é a indicação de diretrizes que apoiem a definição de políticas para a adoção de LA, tendo em vista suas necessidades. Utilizando-se do questionário do SHEILA, os estudantes foram indagados quanto às suas expectativas (o que gostariam que acontecesse); e sua percepção da realidade (o que acreditam que ocorrerá de fato em cada situação apresentada).

Os dados obtidos foram analisados por meio de técnicas estatísticas e de Mineração de Dados Educacionais (MDE), de modo que o presente trabalho busca responder à seguinte questão de pesquisa principal (PP): Quais diretrizes podem ajudar na definição de políticas para adoção de LA nos cursos de computação da universidade estudada, tendo em vista as opiniões de seus estudantes? Como também, três perguntas de pesquisa secundárias (PS), PS1: Quais as opiniões dos estudantes sobre o uso de seus dados pela universidade em projetos de LA? PS2: Qual a diferença entre as expectativas e a realidade apresentada nas opiniões dos estudantes? PS3: Como pode-se definir os perfis dos estudantes de acordo com as respostas?

1.1. Trabalhos Relacionados

O estudo das expectativas dos estudantes no contexto de LA é uma preocupação recente. [Falcao et al. 2019] afirmam que apesar de pesquisas dessa natureza terem sido realizadas, elas ainda são limitadas, especialmente em países em desenvolvimento. Um dos

[2]https://sheilaproject.eu/sheila-framework/

trabalhos que busca entender as opiniões de professores e estudantes sobre o tema é o de [Tsai et al. 2018a]. Para isso, os autores coletaram dados em 4 países, por meio de grupos focais e do questionário SHEILA, e utilizaram técnicas estatísticas para análise dos dados quantitativos. O estudo identificou desafios para a implementação de LA, bem como mostrou que os *stakeholders* têm interesse no uso de LA para sobrepor os desafios da educação. Em [Hilliger et al. 2020], os autores estudaram, utilizaram os mesmos instrumentos de coleta, a opinião dos professores e estudantes de universidades latino-americanas. Eles identificaram que há oportunidade para implementar os serviços de LA, visto que os *stakeholders* compreendem sua importância no desenvolvimento estudantil.

Por fim, o estudo de [Falcao et al. 2019] apresenta o resultado da condução de entrevistas e grupos focais com estudantes de uma IES pública brasileira. Quanto ao uso de LA, estes opinaram que seus dados deveriam ser utilizados para desenvolver perfis de seu aprendizado, identificar vulnerabilidades e sugerir meios de melhorar. Além disso, eles confiam na universidade no que se refere ao uso e à gestão de seus dados.

O presente estudo se diferencia dos demais pelo seu método, um estudo de caso que se utiliza do questionário SHEILA; e objetivo, a proposição de diretrizes para definição de políticas, com base na opinião de estudantes brasileiros, aplicando MDE para análise dos dados. Especificamente quanto ao trabalho de [Falcao et al. 2019], este é diferente pois utiliza-se de dados quantitativos para atingir seus objetivos. Além disso, enquanto o trabalho de [Falcao et al. 2019] foca na compreensão do contexto dos estudantes, o presente trabalho avalia as questões de ética e privacidade sobre o uso de dados.

2. Estudo de Caso

Para a condução do estudo de caso, utilizou-se a metodologia proposta em [Runeson and Höst 2009]. As etapas estão descritas na Figura 1 e detalhadas a seguir.

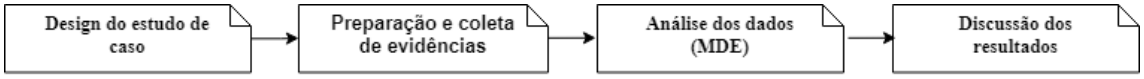

Figura 1. Etapas do estudo de caso, inspirado em [Runeson and Höst 2009].

2.1. *Design* do Estudo de Caso

Nesta etapa, os autores dedicaram-se a entender a importância de envolver os estudantes na adoção de LA, compreendendo-se como fundamental considerar suas opiniões quando da definição de políticas, a fim de assegurar seus interesses e sua aceitação no emprego dessa nova tecnologia. Também observou-se como essas políticas foram estabelecidas em outras IES. Posteriormente, planejou-se o estudo de caso, verificando-se quais instrumentos poderiam apoiar a coleta e compreensão das opiniões dos estudantes. Definiu-se, então, pela utilização do questionário SHEILA, dada a importância e abrangência desse projeto e o desenho bem fundamentado do referido instrumento [Tsai et al. 2018a]. As questões de pesquisa, que orientam o estudo de caso, estão dispostas na introdução deste trabalho.

2.2. Preparação e Coleta de Evidências

Na etapa de **preparação**, o questionário foi traduzido e adaptado. Houve a exclusão de uma de suas sentenças, a qual descreve uma conduta considerada inesperada e improvável

para uma universidade pública no Brasil (”A Universidade solicitará meu consentimento antes de terceirizar meus dados educacionais para análise por empresas terceirizadas”). As 11 (onze) questões restantes foram mantidas. Em seguida, dois testes piloto do questionário adaptado (com dois e três participantes) foram realizados. Os testes apresentaram a necessidade de adaptações semânticas e de usabilidade, as quais foram efetuadas.

Cada questão é dividida na opinião do estudante quanto ao que desejaria que ocorresse (identificada como x.1); e na opinião do estudante quanto ao que acredita que será a decisão da universidade caso a situação questionada ocorra (x.2). As questões foram organizadas como no exemplo abaixo, em um Formulário Google[3], com 7 opções de respostas, organizadas na escala Likert [Likert 1932] entre ”discordo totalmente”e ”concordo totalmente”:

1.1) A universidade solicitará meu consentimento antes de usar qualquer dado identificável sobre mim mesmo (por exemplo, etnia, idade, sexo).

Lembre-se: isso é o que, idealmente, você gostaria que acontecesse.

1.2) A universidade solicitará meu consentimento antes de usar qualquer dado identificável sobre mim mesmo (por exemplo, etnia, idade, sexo).

Lembre-se: isso é o que você acha que sua universidade fará na realidade.

Por fim, na fase de **coleta de evidências**, o questionário validado foi distribuído entre os estudantes dos cursos de computação (Sistemas de Informação, Licenciatura em Computação, Engenharia de *Software* e Engenharia da Computação) de uma universidade pública localizada no estado de Pernambuco. O período de aplicação está compreendido entre os meses de fevereiro e julho de 2021, alcançando um quantitativo de 132 respostas.

2.3. Análise dos Dados

Nesta seção, serão apresentadas as etapas da análise dos dados, realizada através da utilização da metodologia de MDE e baseada no *Cross-Industry Standard for Data Mining* (CRISP-DM) [Chapman et al. 2000]. As etapas do estudo estão descritas a seguir, renomeadas de acordo com o contexto deste trabalho.

Compreensão do Contexto e dos Dados

A **compreensão do contexto** se refere à compreensão dos objetivos da análise de dados, a fim de apoiar a escolha das técnicas adequadas. Nesse caso, objetiva-se interpretar qual a opinião dos estudantes sobre o uso de seus dados em projetos de LA, o que servirá de base para a proposição de políticas para a adoção de LA na instituição estudada.

No que se refere à **compreensão dos dados**, tem-se uma base de dados composta de 28 atributos, dos quais 5 (carimbo de data e hora, confirmação de leitura do texto introdutório e consentimento de participação na pesquisa, faixa etária e gênero) foram considerados irrelevantes para os objetivos do estudo. Os demais atributos são variáveis numéricas associadas às respostas dos estudantes e a variável curso, que foi utilizada para identificar a região do estudantes. Assim, para avaliar se a localização geográfica afeta a percepção dos estudantes sobre LA, os dados foram divididos quanto à região de Pernambuco onde os cursos se encontram: Capital (Engenharia da Computação) e Interior

[3]encurtador.com.br/cmFIU

(Sistemas de Informação, Licenciatura de Computação e Engenharia de *Software*). Os dados de cada região foram divididos ainda em dois grupos: o primeiro é composto das respostas às questões x.1, ou seja, as expectativas dos estudantes, nomeado como "Expectativa"; e o segundo composto das respostas às questões x.2, referentes às opiniões realistas do estudantes, nomeado como "Realidade".

Modelagem

Nesta etapa, são selecionados e aplicados os métodos e algoritmos para análise dos dados. Para isso, buscou-se entender as opiniões dos estudantes e como estas opiniões podem ser agrupadas em cada região, bem como suas diferenças. Foram utilizados métodos de análise descritiva que, segundo [Silvestre 2007], é composta de processos e técnicas que procuram identificar nos dados os atributos relevantes para organizar, analisar e interpretar uma síntese sobre a população investigada.

Também foi utilizada clusterização, que [Han et al. 2011] definem como agrupamentos de objetos em classes (do inglês *cluster*), no qual cada *cluster* é composto por diversos objetos que possuem alta similaridade. Portanto, o presente estudo utilizou o algoritmo de clusterização *k-means*, conhecido como uma técnica de aprendizagem simples, escolhido para trazer maior confiabilidade e precisão para o agrupamento, sem a interferência de especialistas da área. Segundo [Memarsadeghi and O'Leary 2003], o *k-means* consiste em fixar um centróide k, de maneira aleatória para cada grupo de *cluster*, associando cada indivíduo (x_i) ao seu centróide mais próximo e recalculando os centróides baseado nos indivíduos classificados (x_i). Em síntese, há a escolha dos centros iniciais $[c1, c2, \cdots, ck]$, depois repete-se até que os centros parem de mudar, como mostra a equação: $[i = 1, 2, \cdots, n]$, em seguida atribui-se cada ponto de dados x_i ao *cluster* C_j, cujo centro c_j está mais próximo do seu *cluster*, onde $[i = 1, 2, \cdots, k]$. Por fim, calcule novamente o centro c_j, para ser a média (centróide) dos pontos no *cluster*, de acordo com a Equação 1:

$$c_j = \frac{1}{n_j} \sum_{i:x_i \epsilon C_j} x_i \tag{1}$$

onde n_j é o número de pontos de dados em k. O *k-means* utiliza a média no cálculo das distâncias e centróides, o que o torna sensível a *outliers*. Por isso, foi utilizado também o algoritmo *k-medoid*, definido por [Sheng and Liu 2006] como uma técnica de agrupamento para aprendizagem não supervisionada similar em alguns aspectos ao *k-means*. Porém, nos cenários estudados, o *k-means* obteve melhor rendimento, por isso, serão apresentados apenas os resultados desse algoritmo. Por fim, para a escolha da quantidade de grupos foi utilizado o método Elbow [Bholowalia and Kumar 2014].

Avaliação dos Agrupamentos

Uma das funções utilizadas na avaliação do algoritmo foi o Coeficiente de Silhueta $s(i)$, que é usado para determinar a qualidade da alocação dos objetos (indivíduos) nos grupos por um método de agrupamento, onde $i = 1, \cdots, n$. No qual, cada objeto do *cluster* é representado por i. E para cada objeto i o valor $s(i)$ é calculado de acordo com a Equação 2:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2}$$

onde $a(i)$ mede a distância média desse ponto com todos os outros pontos nos mesmos *clusters* e $b(i)$ é a distância média dele com os pontos dos *clusters* mais próximos ao seu *cluster*. Entretanto, $s(i)$ calcula o coeficiente pela média das distâncias e o define numa pontuação de -1 a +1: Quanto maior a pontuação (próximo à +1), mais denso e correto é o *cluster*; pontuação 0 indica que os *clusters* podem estar sobrepostos; e pontuação negativa indica que o *cluster* ou objetos podem estar erroneamente agrupados.

3. Discussão dos Resultados

Esta seção apresenta a análise descritiva, o resultado da aplicação do algoritmo *k-means* e sua respectiva avaliação, bem como as diretrizes para definição de políticas para adoção de LA, com base no resultado das análises. A Tabela 1 expõe os resultados da média e desvio padrão na análise das opiniões dos estudantes, utilizadas na análise descritiva.

3.1. Análise Descritiva

Nas questões relacionadas às expectativas, verificou-se que a maioria das respostas foram próximas do valor máximo da escala *Likert* (”concordo totalmente”), havendo um alinhamento entre estudantes da capital e interior. Considerando a média e desvio padrão, observa-se que os estudantes são otimistas e têm uma expectativa confiante de que a IES solicite consentimento no uso de seus dados pessoais (questão 1.1) e na coleta e análise dos mesmos (4.1). Além disso, esperam que a IES deverá solicitar o consentimento novamente, caso os dados sejam usados para propósitos diferentes do que os anteriormente acordados (5.1).

Existe uma expectativa otimista com relação à manutenção da segurança de seus dados pela IES (2.1). Similarmente, os estudantes esperam que os professores sejam capazes de incorporar as análises no *feedback* em seu benefício (9.1) e que eles sejam obrigados a agir com base no resultado dessas análises (10.1). Além disso, a universidade os atualizará sobre seu progresso com regularidade (3.1), relacionando tal progresso aos objetivos de aprendizagem e da disciplina (7.1). Quanto aos objetivos, os estudantes esperam que LA os apoie na tomada de decisões (6.1), na promoção das habilidades acadêmicas e profissionais (11.1) e na definição do seu perfil de aprendizado (8.1).

Tabela 1. Média e desvio padrão (SD) da opinião dos estudantes da capital e do interior em termos de expectativa e realidade.

Questão	Expectativa (x.1)				Realidade (x.2)			
	Capital		Interior		Capital		Interior	
	Média	SD	Média	SD	Média	SD	Média	SD
1	6,45	1,23	6,57	1,04	5,25	1,67	4,95	1,81
2	6,94	0,42	6,89	0,45	5,57	1,50	5,37	1,50
3	6,65	0,98	6,69	0,77	5,06	1,84	4,81	1,84
4	6,27	1,31	6,42	1,17	5,06	1,91	4,94	1,74
5	6,84	0,42	6,65	1,02	5,04	1.98	4,77	1,88
6	6,59	0,80	6,42	1,05	5,53	1,51	5,10	1,66
7	6,53	0,95	6,54	0,84	5,45	1,50	5,45	1,52
8	6,51	0,81	6,47	1,27	5,65	1,44	5,27	1,73
9	6,47	0,95	6,45	1,04	4,92	1,90	4,71	1,62
10	6,18	1,28	6,39	1,08	4,49	1,91	4,74	1,70
11	6,65	0,66	6,60	0,94	5,25	1,58	5,06	1,68

Na análise estatística, constatou-se uma diferença entre a expectativa e realidade. Foram identificadas incertezas nas opiniões dos estudantes sobre a como a IES irá agir

na realidade quanto à solicitação de consentimento sobre o uso de seus dados pessoais (1.2), tendo em vista o desvio padrão, que mostra dispersão nas respostas. Os estudantes têm preocupação moderada quanto à proteção de dados (2.2) e em relação à solicitação de consentimento da IES para coleta e uso de dados educacionais (4.2). Há uma incerteza por parte dos estudantes se haverá uma nova solicitação de consentimento caso os dados sejam utilizados para um propósito diferente do acordado (5.2). Adicionalmente, identificou-se grande heterogeneidade nas respostas dos estudantes, porém ainda mantendo a opinião de que as seguintes ações podem não ocorrer na prática: uso das análises pelo corpo docente e incorporação das mesmas no *feedback* (9.2); e a obrigação dos docentes de agir (10.2).

Quanto à aplicação de LA, os estudantes apresentam dúvidas se a mesma será utilizada, na realidade, com os propósitos de: atualização regular do progresso, com uma média moderada e com alto desvio padrão (3.2); promoção da tomada de decisão dos estudantes (6.2) e para relacionar o progresso acadêmico do estudante a seus objetivos e da disciplina (7.2); criação e apresentação de um perfil de aprendizado do estudante (8.2); desenvolvimento das habilidades acadêmicas e profissionais (11.2).

3.2. Aplicação do Algoritmo *k-means*

Tendo por base as médias de cada *cluster* em relação às respostas, este estudo classifica cada agrupamento em: otimista (quando as médias são em sua maioria maiores que 5); moderado (médias entre 3,5 e 5); e pessimistas (médias menores que 3,5).

Clusterização - Expectativa e Realidade

A Figura 2(a) apresenta os *clusters* resultantes da análise das expectativas dos 51 estudantes da capital. Nela, é possível identificar que o *cluster* 1 obteve o menor agrupamento, com 7 estudantes, e as médias das suas respostas foram superiores a 6,07, indicando que o grupo é composto de estudantes com expectativas otimistas. O *cluster* 2, composto de 30 estudantes, mostrou expectativas otimistas em todos os itens, com a menor média das repostas igual a 6,70. De modo semelhante, o *cluster* 3, com 14 estudantes, também apresentou expectativas otimistas com relação a algumas questões e moderada com outras, com a menor média (4,64) e a mais alta (6,78). Considerando que o maior dos três *clusters* apresenta médias altas e os outros dois, médias moderadas, a capital demonstra uma concordância alta entre os estudantes bem como uma expectativa alta sobre a coleta, uso, análise e segurança de seus dados.

Já a Figura 2(b) representa as expectativas dos 80 estudantes do interior, agrupados pelo algoritmo *k-means*. O *cluster* 1 (com 15 alunos) apresenta uma percepção entre moderada e otimista (médias variando entre 4,86 e 6,80). O *cluster* 2 é o menor (com 6 estudantes), com as médias mais baixas (variando entre 3,66 e 6,66), o que ainda representa uma percepção moderada. E o *cluster* 3 (com 59 estudantes) apresenta uma percepção otimista (médias entre 6,54 e 6,98). Assim, o interior demonstra a mesma expectativa alta e concordância entre os estudantes, como ocorreu na capital, com o maior *cluster* sendo otimista e os outros dois moderados. Assim, pode-se concluir que tanto os estudantes da capital quanto do interior têm expectativas otimistas sobre o uso de seus dados em projetos de LA.

A Figura 3(a), sobre a realidade da capital, mostra o *cluster* 1, composto de 21 estudantes, que possui 6,04 como a média mais baixa das respostas. Portanto, identifica-se uma visão da realidade otimista com relação à coleta, uso e segurança de seus dados. Já

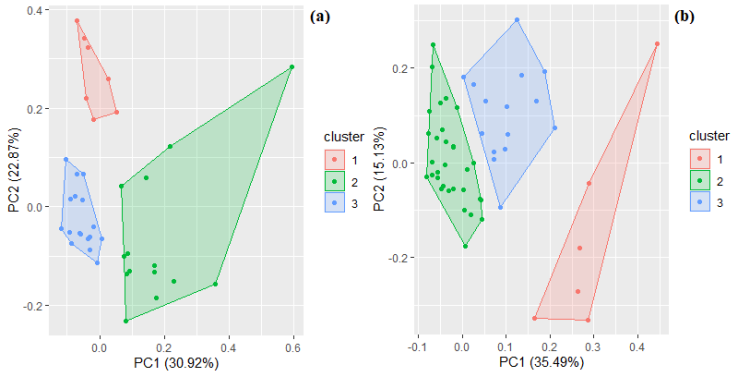

Figura 2. *K-means* aplicado às expectativas da capital (a) e interior (b)

no *cluster* 2, que agrupa 25 estudantes (com médias variando entre 3,8 e 5,3) foi possível identificar uma percepção realista moderada com relação à escala utilizada. As respostas dos 5 estudantes do *cluster* 3, porém, possuem médias baixas na maioria das questões (entre 1,60 e 3,60), apresentando assim um possível sentimento pessimista quanto à maioria dos assuntos tratados no questionário, visto que a média da escala utilizada é 3,5. Entretanto, a questão 2.2 (média 5,00), a questão 5.2 (média 4,60) e a questão 11.2 (média 4,00) podem ser consideradas exceções, por se aproximarem à classificação moderada deste estudo.

A Figura 3(b) representa os *clusters* criados pelo algoritmo *k-means*, agrupados pela percepção da realidade de 80 estudantes do interior. O *cluster* 1 (com 43 estudantes), pode ser definido como moderado, por apresentar médias variando entre 4,11 e 4,97. O *cluster* 2 (com 30 estudantes), apresenta respostas mais otimistas (médias variando entre 5,86 e 6,56). O *cluster* 3 (com 7 estudantes), apresenta uma percepção mais pessimista (médias variando entre 1,71 e 2,42, exceto na questão 2.2, com média 4,14).

Percebe-se que há um alinhamento entre os estudantes da capital e interior, ou seja, nos dois grupos há uma maioria composta por estudantes otimistas ou moderados. Porém, há uma mudança de cenário em comparação aos resultados das expectativas, devido à existência de estudantes que se deslocaram para grupos moderados ou pessimistas. Esses resultados apontam para uma incerteza desses estudantes sobre as decisões que a IES pode tomar no que se refere ao uso de seus dados.

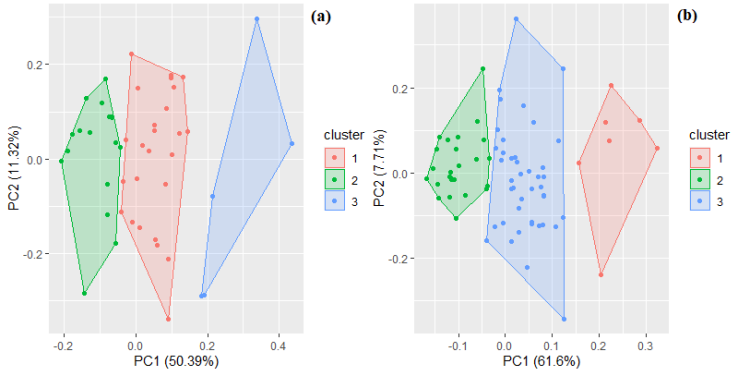

Figura 3. *K-means* aplicado à realidade da capital (a) e interior (b)

Avaliação dos *Clusters*

O Coeficiente de Silhueta foi utilizado para avaliar a qualidade da clusterização realizada pelo algoritmo *k-means*. Para isso, foram utilizados critérios definidos por [Kaufman and Rousseeuw 2009], que considera a similaridade entre os objetos dos *clusters*. A Tabela 2 apresenta os resultados da análise avaliativa da individualidade do algoritmo *k-means* sobre as expectativas e realidade dos estudantes da capital e interior. Os *clusters* com estrutura de agrupamento mais robusta, com coeficiente de silhueta moderado (0,58 e 0,57, respectivamente), são o *cluster* 3 (capital) e o *cluster* 2 (interior). Isso significa que a relação entre os indivíduos de cada *cluster* é razoável. Além disso, observa-se que todos os agrupamentos realizados pelo *k-means* foram corretos, com exceção dos *clusters* 1 e 3 (interior) e o *cluster* 2 (capital), que apresentaram coeficiente de silhueta negativos, indicando a alocação incorreta de alguns indivíduos. Os resultados da avaliação demonstram que, apesar do tamanho da amostra, o algoritmo utilizado se comportou de maneira aceitável, produzindo resultados de acordo com a literatura.

Tabela 2. Coeficiente de silhueta, aplicado ao algoritmo *k-means*.

Cluster	Expectativa (x.1)				Realidade (x.2)			
	Capital		Interior		Capital		Interior	
	Qª	Média	Qª	Média	Qª	Média	Qª	Média
1	7	0,15	6	-0,06	25	0,11	7	0,31
2	14	-0.05	59	0,57	21	041	30	0,43
3	30	0,58	15	-0,03	5	0,07	43	0,22

Diretrizes para o Estabelecimento de Políticas para LA

Após a análise dos dados, é possível compreender as opiniões e necessidades dos estudantes analisados, tornando viável indicar à IES em questão diretrizes que apoiam o desenvolvimento de políticas. O presente artigo realizou esta orientação, com base em [Tsai and Gasevic 2017]. Os resultados foram organizados em 4 categorias:

Estratégia: O propósito para o uso de LA deve ser bem estabelecido, de modo a ficar claro no que LA deveria ou não ser empregada, bem como suas limitações. A instituição deve ser transparente quanto aos tipos de dados que serão coletados e analisados, garantindo sua qualidade, e quanto às intervenções que podem ser realizadas. A IES deve promover intervenções personalizadas para cada aluno, viabilizando a tomada de decisão do estudante e atualizando-o com regularidade sobre seu progresso acadêmico, que deve estar relacionado com os seus objetivos de aprendizagem e das disciplinas. Os usuários de LA precisam ser informados sobre o seu perfil de aprendizado e sobre como os seus resultados podem afetá-los e quais são seus direitos e obrigações. Deverá ser fornecido treinamento para toda a instituição. Os professores deverão realizar intervenções sempre que identificarem estudantes com dificuldades ou em risco de reprovação.

Obrigações Legais e Organizacionais: A IES brasileira deverá seguir a legislação para proteção dos usuários. No Brasil, isto se refere à Lei Geral de Proteção de Dados Pessoais (LGPD - Lei Nº 13.853, de 8 de Julho de 2019).

Proteção da privacidade: Os dados devem ser mantidos seguros pela universidade. Os alunos devem ser notificados sobre como seus dados serão divulgados e a IES adotará medidas para proteger a segurança e integridade. O consentimento deve ser obtido antes que os dados sejam coletados e solicitado novamente quando estes são usados para

propósitos diferentes que os anteriormente acordados. Os estudantes devem ter a opção de não participar dos projetos de LA ou se retirarem, se assim desejarem.

Gestão e Governança dos Dados: Os usuários devem ter o direito a acessar seus próprios dados e serem capazes de gerenciá-los e atualizá-los. Além disso, eles devem ter a opção de cancelar o processo de coleta ou uso, ou retirar dados já coletados. Os alunos devem ter acesso aos dados sobre seu aprendizado de forma que isso seja benéfico para seu desenvolvimento profissional e de aprendizagem.

4. Conclusões e Ameaças à Validade

Este artigo apresentou um estudo de caso a respeito das opiniões dos estudantes de uma IES brasileira sobre o uso de seus dados em projetos de LA. Os dados foram coletados utilizando-se o questionário SHEILA e analisados por meio de técnicas estatísticas e de Mineração de Dados Educacionais. Desse modo, apresentam-se, de forma resumida, os resultados que respondem às perguntas de pesquisa secundárias e, em seguida, à principal.

Em resposta à questão de pesquisa PS1 (Quais as opiniões dos estudantes sobre o uso de seus dados pela universidade em projetos de LA?) tem-se, em suma, que os estudantes, tanto do interior quanto da capital, têm expectativas altas de que LA seja utilizada em seu benefício e de que a IES agirá de modo a garantir a ética e privacidade no uso de seus dados. Quanto à pergunta de pesquisa PS2 (Qual a diferença entre as expectativas e a realidade apresentada nas opiniões dos estudantes?), os estudantes demonstram incerteza quanto ao que eles acreditam que a instituição realizará na prática. Apesar de terem expectativas altas de que ocorra, os estudantes demonstraram incertezas especialmente no que se refere à incorporação de LA na prática pedagógica dos professores, na obrigação dos docentes de agirem caso o estudante tenha dificuldades e quanto à solicitação de um novo consentimento caso os dados venham a ser utilizados para um propósito diferente do inicialmente acordado. Porém, apesar dessas incertezas, os estudantes demonstram entender que LA pode trazer benefícios para seu aprendizado. Respondendo à pergunta de pesquisa PS3 (Como pode-se definir os perfis dos estudantes de acordo com as respostas?), as respostas dos estudantes foram agrupadas utilizando-se o algoritmo *k-means*. Desse modo, foi possível dividir os perfis dos estudantes em otimistas, pessimistas e moderados, sendo os otimistas e moderados maioria neste estudo.

No que se refere à questão de pesquisa principal, PP (Quais diretrizes podem ajudar na definição de políticas para adoção de LA nos cursos de computação da universidade estudada, tendo em vista as opiniões de seus estudantes?), foi possível indicar diretrizes que permitam à IES brasileira estabelecer políticas para a adoção de LA, as quais atendam às expectativas e necessidades dos estudantes. Foram abordadas questões relacionadas à estratégia, legislação, privacidade e gestão e governança dos dados.

Por fim, cabe ressaltar as ameaças à validade identificadas na condução deste estudo, quais sejam: (a) o tamanho da amostra, que se limitou a 132 indivíduos, e que também inviabiliza a realização de análises entre os diferentes cursos de computação considerados; (b) O desbalanceamento dos dados, com uma diferença de 30 respostas entre o interior e a capital, o que pode beneficiar os resultados da região com maior número de dados; (c) a ausência das opiniões de outros profissionais da educação, em especial os professores, visto que estes fazem parte dos *stakeholders* de LA, e suas opiniões também devem ser consideradas na definição de políticas.

Referências

Avella, J. T., Kebritchi, M., Nunn, S. G., and Kanai, T. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning*, 20(2):13–29.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9):17–24.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9:13.

Falcao, T. P., Ferreira, R., Rodrigues, R. L., Diniz, J., and Gasevic, D. (2019). Students' perceptions about learning analytics in a brazilian higher education institution. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 204–206. IEEE.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hilliger, I., Ortiz, M., Pesántez-Cabrera, P., Scheihing, E., Tsai, Y.-S., Merino, P., Broos, T., Whitelock-Wainwright, A., Gasevic, D., and Pérez-Sanagustín, M. (2020). Towards learning analytics adoption: A mixed methods study of data-related practices and policies in latin american universities: Data practices and policies in latin america. *British Journal of Educational Technology*, 51.

Hoel, T., Mason, J., and Chen, W. (2015). Data sharing for learning analytics–questioning the risks and benefits. In *Proceedings of the 23rd International Conference on Computers in Education*. China: Asia-Pacific Society for Computers in Education.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lim, C. and Tinio, V. (2018). Learning analytics for the global south. *Quezon City, Philippines: Foundation for Information Technology Education and Development*.

Memarsadeghi, N. and O'Leary, D. P. (2003). Classified information: the data clustering problem. *Computing in Science & Engineering*, 5(5):54–60.

Runeson, P. and Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164.

Sheng, W. and Liu, X. (2006). A genetic k-medoids clustering algorithm. *Journal of Heuristics*, 12(6):447–466.

Silvestre, A. L. (2007). *Análise de dados e estatística descritiva*. Escolar editora.

Tsai, Y.-S. and Gasevic, D. (2017). Learning analytics in higher education—challenges and policies: a review of eight learning analytics policies. In *Proceedings of the seventh International Learning Analytics & Knowledge Conference*, pages 233–242.

Tsai, Y.-S., Gasevic, D., Whitelock-Wainwright, A., Moreno-Marcos, P. M., Fernandez, A. R., Muñoz-Merino, P. J., Kloos, C. D., Tammets, K., Kollom, K., Scheffel, M.,

and Drachsler, H. (2018a). Teacher and student perspectives on learning analytics – executive summary.

Tsai, Y.-S., Moreno-Marcos, P. M., Tammets, K., Kollom, K., and Gašević, D. (2018b). Sheila policy framework: informing institutional strategies and policy processes of learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 320–329.