UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO INTERDISCIPLINAR DE NOVAS TECNOLOGIAS NA EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA NA EDUCAÇÃO

Vanessa Faria de Souza

O QUANTO EU QUERO ESTE CERTIFICADO? CAÇADORES DE CERTIFICADOS NO LÚMINA

Porto Alegre
2022

Vanessa Faria de Souza

O QUANTO EU QUERO ESTE CERTIFICADO? CAÇADORES DE CERTIFICADOS NO LÚMINA

Tese apresentada ao Programa de Pós-Graduação em Informática na Educação, do Centro Interdisciplinar de Novas Tecnologias na Educação, da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do título de Doutor em Informática na Educação.

Orientadora: Profa. Dra. Gabriela Trindade Perry

Linha de pesquisa: Interfaces Digitais em Educação, Arte, Linguagem e Cognição

Porto Alegre
2022

CIP - Catalogação na Publicação

Souza, Vanessa Faria de
O QUANTO EU QUERO ESTE CERTIFICADO? CAÇADORES DE CERTIFICADOS NO LÚMINA / Vanessa Faria de Souza. -- 2022.
165 f.
Orientadora: Gabriela Trindade Perry.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Centro de Estudos Interdisciplinares em Novas Tecnologias na Educação, Programa de Pós-Graduação em Informática na Educação, Porto Alegre, BR-RS, 2022.

1. Massive Open Online Courses. 2. Mineração de Dados Educacionais . 3. Agrupamento. 4. Regressão Logística. I. Perry, Gabriela Trindade, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Vanessa Faria de Souza

O QUANTO EU QUERO ESTE CERTIFICADO? CAÇADORES DE CERTIFICADOS NO LÚMINA

Tese apresentada ao Programa de Pós-Graduação em Informática na Educação, do Centro Interdisciplinar de Novas Tecnologias na Educação, da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do título de Doutor em Informática na Educação.

Orientadora: Profa. Dra. Gabriela Trindade Perry

Aprovada em: Porto Alegre, 01 de novembro de 2022.

BANCA EXAMINADORA:

Profa. Dra. Gabriela Trindade Perry (orientadora)
UFRGS / Programa de Pós-graduação em Informática na Educação (PPGIE)

Prof. Dr. Sílvio César Cazella
UFRGS / Programa de Pós-Graduação em Informática na Educação (PPGIE)

Profa. Dra. Ana Luísa Petersen Cogo
UFRGS / Programa de Pós-Graduação em Enfermagem (PPGE)

Profa. Dra. Ilka Márcia Ribeiro de Souza Serra
UEMA / Programa Profissional de Pós-Graduação em Educação Inclusiva (PROFEI)

Ao meu esposo Leandro
companheiro de uma vida,
presente em todos os momentos.
Por acreditar mais em mim que eu mesma!

AGRADECIMENTOS

Agradeço à minha Orientadora Gabriela Trindade Perry, acima de tudo pela paciência e por muitas vezes ser mais psicóloga que orientadora (foi um tempo difícil para fazer doutorado) e por me ensinar muito mais do que a pesquisar e escrever o necessário para a Tese, pois desenvolver um pensamento crítico, bem como a busca contínua para ser uma pesquisadora cada vez melhor e com trabalhos de relevância, vai muito além da mera orientação.

Agradeço ao meu Pai João Batista de Souza e à minha Mãe Inez Faria de Souza, pelo apoio incondicional e de sempre, pelo incentivo em todos os momentos e por nunca me deixarem desistir.

Agradeço à minha irmã Mayara Faria de Souza por sempre ser um ouvido amigo e atencioso, mesmo quando eu passava horas falando sobre coisas que ela muitas vezes nunca tinha ouvido falar, e a maioria delas, ela nem queria!

Agradeço à Professora Liane Margarida Rockenbach Tarouco por ter sido uma excelente coordenadora, na maior parte do meu processo de doutoramento, e por estar sempre disposta a ajudar.

Agraço aos Colegas Tony Bignardi e Gilson Saturnino pelo companheirismo, trabalhos em conjunto e ótimos momentos de descontração.

À todos os meus professores no PPGIE, que com muita dedicação, não mediram esforços para continuar proporcionando um ensino de excelência, mesmo em meio a todas as dificuldades e barreiras impostas pela pandemia do novo Coronavírus.

Finamente, agradeço ao meu esposo Leandro Aparecido de Aguiar, por sempre ter algo bom a me dizer, por sempre fazer piadas e por ser um “chato”, além de ter muita maturidade emocional para ser um apoio e suporte nos momentos que mais precisei, meu muito obrigada!

A presente Tese foi desenvolvida com apoio do *Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)* e do *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)*.

"O educador se eterniza em cada ser que educa"

"A educação é um ato de amor, por isso, um ato de coragem"

Paulo Freire

RESUMO

Com o grande aumento do número de alunos inscritos em *Massive Open Online Course*s (MOOCs), e o crescimento na quantidade de plataformas de distribuição, a oferta deste tipo de curso está cada vez maior, pois representam uma possibilidade de disseminação de conhecimento especializado, de forma flexível e aberta. Todavia, com o aumento do uso de plataformas on-line de aprendizagem o estudo da desonestidade acadêmica se torna relevante, neste contexto, pois estes cursos podem ser mais facilmente burlados do que cursos presenciais. Por isso, o objetivo geral desta Tese é identificar parâmetros de configuração de MOOCs que desestimulem estudantes que têm comportamentos de “caçadores de certificados” a obter certificações, ao mesmo tempo em que não desestimulem estudantes engajados, na Plataforma de MOOCs da Universidade Federal do Rio Grande do Sul (UFRGS), o Lúmina. Para tanto, é preciso identificar e caracterizar o perfil dos “caçadores de certificados”, estudantes que buscam explorar características da plataforma e dos cursos, para obter um certificado sem se dedicar a sua aprendizagem. Nesta tese, levantou-se a hipótese de haver um “comportamento de caçador” (independente do aluno) e um perfil que pode ser chamado de “estudante-caçador” (um indivíduo que sempre exibe este comportamento). Para caracterizar o comportamento de caçador e dos estudantes-caçadores foi desenvolvido um processo metodológico iterativo com as seguintes etapas: Seleção e Processamento dos dados; Aplicação de Técnicas de Mineração de Dados Educacionais; e Considerações sobre o Processo. Como técnica de Mineração de Dados Educacionais para identificar este comportamento e estes estudantes, foram utilizados algoritmos de Aprendizagem de Máquina não supervisionados, mais especificamente algoritmos de agrupamento hierárquico. Em relação à identificação do "comportamento de caçador", o algoritmo de agrupamento não foi capaz de identificar características que permitam identificar os usuários, pois os grupos formados apresentam níveis parecidos na maioria das variáveis utilizadas, à exceção das variáveis "curso tem mais de 10 questões", que é um indicador de dificuldade do curso. Em relação à identificação de estudantes caçadores, entende-se que a obtenção de pelo menos 3 certificados em menos de 35 dias é um bom indicador para classificar um estudante como caçador de certificados. Em relação ao modelo que ajusta a presença de caçadores às configurações dos cursos, conclui-se que não há indícios suficientes para indicar que as restrições nas configurações sejam eficazes para inibir caçadores de certificados.

Palavras-Chave: *Massive Open Online Courses (MOOCS)*. Mineração de Dados Educacionais. Agrupamento. Regressão Logística.

ABSTRACT

With the large increase in the number of students enrolled in Massive Open Online Courses (MOOCs), and the growth in the number of distribution platforms, the offer of this type of course is increasing, as they represent a possibility of disseminating specialized knowledge, of flexible and open way. However, with the increased use of online learning platforms, the study of academic dishonesty becomes relevant in this context, as these courses can be more easily circumvented than face-to-face courses. Therefore, the general objective of this Thesis is to identify MOOCs configuration parameters that discourage students who have “certificate hunter” behaviors to obtain certifications, while not discouraging engaged students, in the MOOCs Platform of the Federal University of Rio de Janeiro. Grande do Sul (UFRGS), the Lúmina. To do so, it is necessary to identify and characterize the profile of “certificate hunters”, students who seek to explore features of the platform and courses, in order to obtain a certificate without dedicating themselves to learning. In this thesis, it was hypothesized that there is a “hunter behavior” (independent of the student) and a profile that can be called “student-hunter” (an individual who always exhibits this behavior). To characterize the behavior of hunters and student-hunters, an iterative methodological process was developed with the following steps: Selection and Processing of data; Application of Educational Data Mining Techniques; and Process Considerations. As an Educational Data Mining technique to identify this behavior and these students, unsupervised Machine Learning algorithms were used, more specifically hierarchical clustering algorithms. Regarding the identification of "hunter behavior", the grouping algorithm was not able to identify characteristics that allow identifying users, since the groups formed have similar levels in most of the variables used, with the exception of the variables "course has more than 10 questions", which is an indicator of course difficulty. Regarding the identification of student hunters, it is understood that obtaining at least 3 certificates in less than 35 days is a good indicator to classify a student as a certificate hunter. Regarding the model that adjusts the presence of hunters to the course settings, it is concluded that there is not enough evidence to indicate that the restrictions in the settings are effective in inhibiting certificate hunters.

Keywords: Massive Open Online Courses (MOOCS). Educational Data Mining. Clustering. Logistic Regression.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVEATURAS E SIGLAS

AA	Análise de Aprendizagem
AVA	Ambiente Virtual de Aprendizagem
AM	Aprendizagem de Máquina
AP	Aprendizagem Profunda
EaD	Educação à Distância
KNN	K-Nearest Neighbor
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MOOCS	*Massive Open Online Courses*
MVS	Máquinas de Vetores de Suporte
UFRGS	Universidade Federal do Rio Grande do Sul

SUMÁRIO

1 INTRODUÇÃO

Esta Tese tem como objetivo principal identificar como as configurações de MOOCs (*Massive Open Online Courses*) podem desestimular estudantes que burlam os ambientes computacionais de ensino para obter certificados, ao mesmo tempo em que não desestimulam estudantes engajados. O contexto é a plataforma Lúmina, da Universidade Federal do Rio Grande do Sul (UFRGS). Para tanto, é preciso caracterizar o perfil dos “caçadores de certificados”, estudantes que buscam explorar características da plataforma e dos cursos, para obter um certificado sem se dedicar à sua aprendizagem. A terminologia “caçadores de certificados” foi cunhada no decorrer do desenvolvimento desta Tese, não foi ainda relatada na literatura. Corresponde a um perfil de usuários cujas ações foram inicialmente percebidas pela equipe gestora da plataforma Lúmina, não havia no início das pesquisas uma nomenclatura para designar tal comportamento, porém com o desenvolvimento e aprofundamento dos estudos que levaram a esta Tese, além da percepção das particularidades deste perfil de aluno, esta terminologia foi definida para especificar um conjunto de ações inadequadas, que levam um aluno obter um certificado de forma rápida e sem estar comprometido com a aprendizagem dos conteúdos do MOOC em que se matriculou.

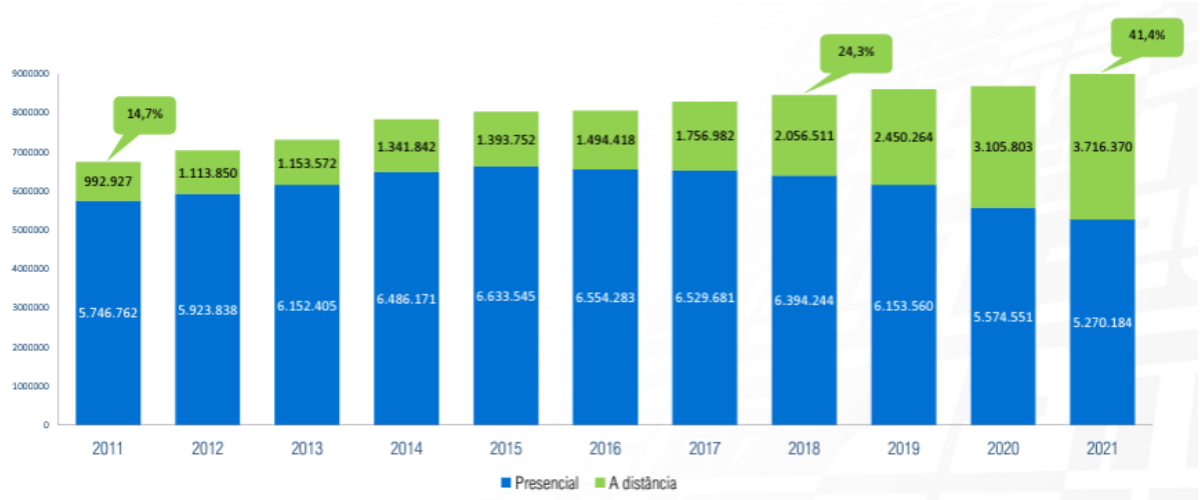
Dessa forma, essa Tese tem um contexto de desenvolvimento interdisciplinar, envolvendo o estudo dos comportamentos dos alunos durante a realização dos MOOCs, a partir dos dados originados de suas trajetórias na plataforma. Tais dados são analisados de forma quantitativa a partir de métodos computacionais de análise de dados, a fim de sistematizar um conjunto de configurações para os MOOCs, buscando desestimular práticas inadequadas, por meio da modulação do esforço necessário para conseguir a certificação.

Realizar um estudo sobre MOOCs em meio a grande notoriedade dada ao ensino remoto foi gratificante, pois percebeu-se as possibilidades que esses tipos de cursos oportunizam, ao incluir mais pessoas em grandes instituições de ensino, disponibilizar materiais de qualidade produzidos por professores especialistas em suas áreas, democratizando o ensino. Todavia, mesmo antes das mudanças emergenciais do ensino tradicional para o ensino remoto, em função da pandemia do vírus Sars-Covid 19, a chegada da “Indústria 4.0” trouxe grandes mudanças para a

sociedade e para o desenvolvimento científico, que afetaram fortemente a maneira como as pessoas aprendem e ensinam.

Neste sentido, a Educação a Distância (EaD), que tem crescido em vários países, no Brasil tem se estabelecido de forma abrangente. De acordo com Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) entre 2011 e 2021, o número de ingressantes em cursos superiores de graduação, na modalidade de educação a distância (EaD), aumentou 474%. No mesmo período, a quantidade de ingressantes em cursos presenciais diminuiu 23,4%. Se, em 2011, os ingressos por meio de EaD correspondiam a 18,4% do total, em 2021, esse percentual chegou a 62,8%. Os dados, que refletem a expansão do ensino a distância no Brasil, fazem parte dos resultados do Censo da Educação Superior 2021[1], divulgados pelo Inep e pelo Ministério da Educação (MEC), uma melhor visualização destes dados pode ser observada na Figura 1. Além disso, é interessante observar a evolução do número de cursos de graduação a distância no Brasil na década supracitada, esta informação pode ser visualizada na Figura 2.

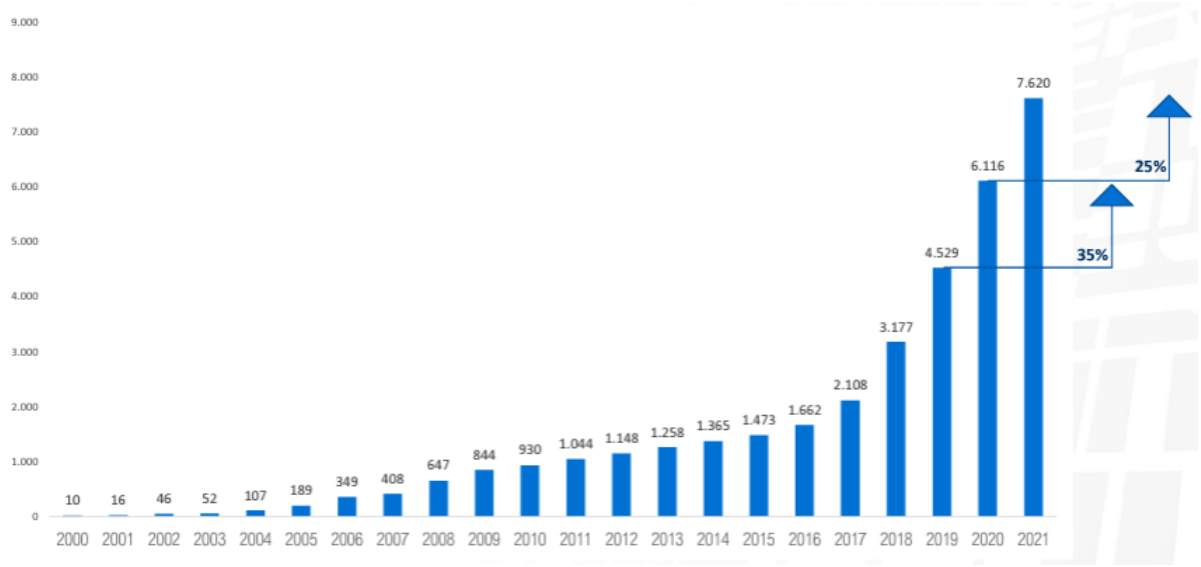
Figura 1 – Número de Matrículas em Cursos de Graduação, por modalidade de Ensino



Fonte: Censo da Educação Superior 2021 – Inep

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados/2021

Figura 2 – Evolução do número de cursos de graduação a distância



Fonte: Censo da Educação Superior 2021 – Inep

Com o aumento da demanda por cursos EaD, também se observa um número crescente de estudantes atraídos para os *Massive Open Online Course*s (MOOCs), com números expressivos de matrículas nos últimos anos (ROMERO; VENTURA, 2017). Há mais de 10 anos atrás, em 2012, a *edX*, uma startup sem fins lucrativos de *Harvard* e do *Massachusetts Institute of Technology*, teve 370 mil alunos em seus primeiros cursos oficiais; o *Coursera*, fundado em janeiro de 2011, alcançou 1,7 milhão de alunos registrados, em um ano, e está crescendo rapidamente (PAPPANO, 2012), registrando uma evolução anual em número de usuários maior que o Facebook; um curso de Inteligência Artificial, de *Stanford* oferecido em 2011, on-line e gratuitamente atraiu 160 mil estudantes (LI; ZHOU, 2018).

Pesquisadores como Romero e Ventura (2017); Greene, Oswald e Pomerantz (2015); Hew, Qiao e Tang (2018); Wang, Hu e Zhou (2018); e Xing *et al.* (2016) salientam que o modelo MOOC de ensino e aprendizagem corresponde a um exemplo bastante aprimorado de educação sustentável. Além disso, ajuda a estabelecer uma educação personalizada e flexível, na qual o aluno tem mais autonomia na gerência de seu tempo de estudo; bem como nos conteúdos e no ambiente que deseja estudar, fatores considerados como tendo grande potencial para promover também um desenvolvimento sustentável. O interesse nesse formato de educação resultou na declaração expressa no *The New York Times*[2] de que 2012 foi “O ano do MOOC”.

---

2 https://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html

Mesmo assim, até o momento, há questões sem resposta, no que diz respeito aos MOOCS. Um exemplo é a grande proporção de desistências, apontado como um importante fator limitante na consolidação desse tipo de curso. O forte contraste entre a quantidade de inscrições e de abandono aumenta o ceticismo sobre a sustentabilidade deste formato, e levanta questões sobre como aumentar a quantidade de concluintes, e de que forma aperfeiçoar a qualidade do engajamento dos alunos, componentes chave para consolidação deste modelo educacional (GREENE; OSWALD; POMERANTZ, 2015; WANG; HU; ZHOU, 2018; XING *et al.*, 2016).

Em relação à questão do abandono e a falta de engajamento, uma característica dos MOOC que pode auxiliar a busca por soluções é a grande quantidade de dados gerada pelas interações nas plataformas, o que abre novas possibilidades para estudar e compreender estas interações. Segundo Romero e Ventura (2017), os MOOCs oferecem um laboratório e uma janela única para compreender os processos de aprendizagem de uma população grande e diversificada de estudantes. A possibilidade de analisar volumes sem precedentes de dados sobre estudantes em MOOCs é de grande relevância e interesse para pesquisadores da ciência de dados e da educação. Tais plataformas armazenam, como salientado, um grande volume de dados, mas é inviável analisá-los manualmente. Verdadeiramente, um dos desafios que as instituições de ensino enfrentam na atualidade é o crescimento exponencial de dados e como transformá-los em novas ideias que podem beneficiar alunos, professores e gestores (BAKER, 2015).

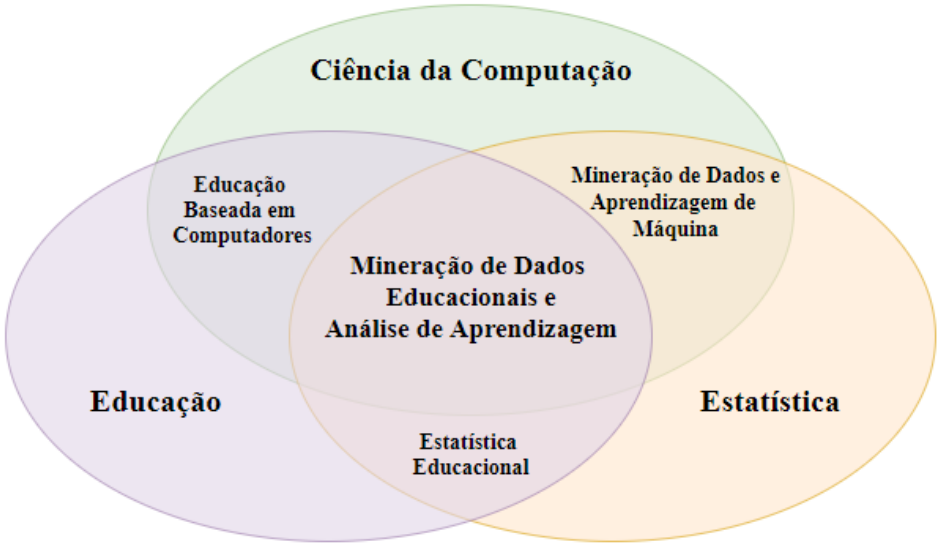
À vista disso, percebeu-se potencial para novos campos e estudos, e abordagens baseadas em métodos computacionais, que podem ser explorados em benefício da educação e aprimoramento dos processos de ensino e aprendizagem, como a Mineração de Dados Educacionais (MDE) e Análise de Aprendizagem (AA) ou *Learning Analytics* (LA). A Mineração de Dados Educacionais abrange o desenvolvimento e aplicação de métodos para explorar esses tipos únicos de dados provenientes de ambientes educacionais (BAKHSHINATEGH *et al.*, 2018), podendo ser definida como a aplicação de técnicas de mineração de dados para abordar questões relativas à educação (ROMERO; VENTURA, 2013). A Análise de Aprendizagem pode ser definida como a medição, coleta, análise e relatório de dados

sobre alunos e seus contextos, com o objetivo de entender e otimizar a aprendizagem e os ambientes em que ocorre (LANG *et al.*, 2017).

Tais abordagens compartilham um interesse em comum – métodos intensivos em dados para pesquisa educacional – e compartilham o objetivo de aprimorar a prática educacional (LIÑÁN; PÉREZ, 2015). Por um lado, AA está focada no desafio educacional e a MDE no desafio tecnológico. A AA analisa os dados para orientar a tomada de decisão e integração das dimensões técnica, social e pedagógica da aprendizagem, aplicando conhecimentos e modelos preditivos, enquanto a MDE geralmente procura novos padrões nos dados e desenvolve novos algoritmos e/ou modelos (LIÑÁN; PÉREZ, 2015).

Por fim, as diferenças entre elas são mais baseadas em foco, na investigação de forma geral, questões de pesquisa e eventual uso de modelos, do que sobre as técnicas utilizados (BAKER; INVENTADO, 2014). Independentemente das diferenças entre a AA e a MDE (que são sutis), as duas áreas têm sobreposição significativa, tanto nos objetivos dos pesquisadores, como nos métodos e técnicas utilizados na investigação (BAKER; INVENTADO, 2014). Elas são áreas interdisciplinares, incluindo, entre outros: recuperação de informações, sistemas de recomendação, análise de dados visuais, mineração de dados orientada por domínio, análise de redes sociais, psicopedagogia, psicologia cognitiva, psicometria (ROMERO; VENTURA, 2020). De fato, elas podem ser compreendidas como a combinação de três áreas principais (Figura 3): Ciência da Computação, Educação e Estatística. A interseção dessas três áreas também forma subáreas intimamente relacionadas – a Educação Baseada em Computadores; Mineração de Dados e Aprendizagem de Máquina, e Estatística Educacional (ROMERO; VENTURA, 2020).

Figura 3 – Principais áreas relacionadas a MDE e AA



Fonte: Adaptado de Romero e Ventura (2020)

Todavia, mesmo que a aplicação de processos de MDE e AA possa gerar muitas descobertas com relação aos processos de aprendizagem dos alunos, não são todos os aspectos que podem ser analisados somente com a coleta e análise dos dados das plataformas de ensino e aprendizagem. Nessa direção, Souza e Perry (2019) em sua revisão sistemática de literatura, sobre a identificação do comportamento de alunos em MOOCs, apontaram como um grande desafio no entendimento desses comportamentos; sobretudo no que diz respeito a persistência e engajamento nos MOOCs, a influência de fatores externos que não estão particularmente ligados ao curso, mas que interfere de forma determinante na decisão de um aluno em prosseguir na realização do curso, ou na desistência. Como citado por Souza e Perry (2019) fatores deste tipo não podem ser identificados por meio das interações com a plataforma e são bastante heterogêneos e difíceis de tratar, por esse motivo não é possível obter informações totalmente precisas com métodos baseados unicamente na mineração de dados, o que caracteriza uma limitação em estudos que incluem exclusivamente tais técnicas.

Diante do contexto apresentado, destaca-se que esta Tese utilizou como estratégia para a modelagem dos comportamentos dos alunos principalmente a MDE, com uma abordagem exploratória que empregou algoritmos de Aprendizagem de Máquina não supervisionados. Ademais, para identificar como as configurações dos MOOCs podem desestimular estudantes caçadores de certificados, foram utilizados também algoritmos de Aprendizagem de Máquina do tipo supervisionado. Além disso,

análises quantitativas foram aplicadas no decorrer de todo o desenvolvimento dos processos destacados. Destaca-se que a MDE foi empregada para analisar em nível micro as atividades e interações dos alunos com os recursos educacionais, que estão presentes nos MOOCs do Lúmina.

1.1 JUSTIFICATIVA

A desonestidade acadêmica – definida nesta Tese como qualquer tipo de ação fraudulenta em trabalhos acadêmicos – é um sério problema na educação. Com o aumento do uso de plataformas on-line de aprendizagem, o estudo da desonestidade acadêmica também aumentou, de forma que alguns pesquisadores fazem uma divisão entre os métodos de trapaça “tradicionais”, como aqueles que são usados em sala de aula, e métodos “contemporâneos”, que incorporaram a internet ou novos dispositivos eletrônicos (CORRIGAN-GIBBS *et al.*, 2015).

Se em ambientes presenciais é difícil detectar trapaças, na educação on-line este problema é ainda mais desafiador, pois não há (ou é muito difícil haver) confirmação de identificação sobre quem fez o exame, ou o que os estudantes fizeram durante o exame, ou se estão realmente fazendo as atividades, como: leituras e assistindo as vídeo aulas – problema que se acentua com a difusão do ensino remoto.

Considerando os MOOCs, a desonestidade acadêmica pode ser ainda mais difícil de detectar, pois as interações com professores e colegas são pouco frequentes. Exemplo disso é o relato de Webley (2012) que afirmou que “desde o início, os provedores de MOOCs têm lutado com o problema da trapaça”, e ressalta que instrutores manifestaram preocupações sobre várias formas de trapaça em seus MOOCs, como: plágio em redações, colaborações ilícitas em exames, postagens de soluções para perguntas de testes on-line, ou envio de respostas por e-mail aos colegas. Essa situação não tem uma solução simples e destaca-se que um dos principais obstáculos ao explorar essa questão em MOOCs é a falta geral de dados confiáveis, o que torna difícil encontrar uma prova definitiva da trapaça, pois os alunos relutam em admitir seu mau comportamento (WEBLEY, 2012).

Um dos principais métodos para trapacear em MOOCs foi inicialmente relatado por Northcutt, Ho e Chuang (2016), este método foi chamado de CAMEO (*Copying Answers using Multiple Existences On-line*), o mesmo consiste em procurar respostas usando múltiplas contas na plataforma de MOOCs. Depois destes autores,

outros como: Ruiperez-Valiente *et al.* (2016); Alexandron *et al.* (2017), Ruiperez-Valiente *et al.* (2017); e Bao, Chen e Hauff (2017), também se dedicaram a desenvolver abordagens que identificassem alunos que se utilizam do CAMEO para obter seus certificados de forma irregular, principalmente em plataformas como o *Coursera* e *edX*.

Contudo, as trapaças acadêmicas em sistemas educacionais informatizados vêm sendo relatadas há muito mais tempo, muito antes da aparição dos MOOCs. O primeiro registro encontrado foi mencionado por Tait, Hartley e Anderson (1973), dentro do contexto de instruções assistidas por computador administradas por teletipo. Na década de 1990 e 2000, foram novamente relatadas pelos autores Schofield (1995); Aleven e Koedinger (2000), Aleven e Koedinger, (2002). Em 2004, dois artigos usaram pela primeira vez a expressão “enganar o sistema” (*gaming the system*) para distinguir essas más atitudes em ambientes de aprendizagem baseados em computador: Baker *et al.* (2004) e Baker, Corbett e Koedinger (2004), e em seguida o interesse em pesquisas sobre esse tema aumentou, levando ao aumento da produção de artigos sobre comportamentos inadequados nesse contexto (BAKER, 2011).

Muitas pesquisas foram desenvolvidas com o objetivo de identificar trapaças em ambientes de aprendizado baseados em computador, como as investigações de Baker *et al.* (2004); (2010); (2008); Baker (2007); Bevan e Hood (2006); Cocea, Hershkovitz e Baker (2009), tais pesquisas tinham como foco um software específico, o Tutor Cognitivo. Embora o software de tutor cognitivo fosse usado para aumentar o envolvimento e esforço dos alunos em sala de aula, alguns alunos se utilizavam de estratégias orientadas para trapacear na utilização da ferramenta. Esse conjunto de estratégias que visavam “enganar o sistema”, foram descritas por Baker, Corbett e Koedinger (2004) como: uma tentativa de obter sucesso em um ambiente educacional explorando propriedades do sistema/software, em vez de tentar aprender o material e usar esse conhecimento para responder corretamente.

“Trapacear” ou “enganar” sistemas informatizados de ensino e aprendizagem, como percebido, não é um desafio novo, contudo, não foi encontrada solução para esse problema – que foi exacerbado no contexto de cursos à distância e dos MOOCs. Trapacear ou não, é uma característica que depende do perfil de cada aluno, de uma avaliação sobre o *“risco versus oportunidade”*, dos diferentes objetivos ao se envolver na realização de um curso e da plataforma utilizada pelos alunos, em especial sobre como suas particularidades permitem que esses comportamentos ocorram.

Nesse sentido, segundo relatos dos gestores da plataforma Lúmina, desde que os certificados passaram a ser oferecidos e um canal de atendimento aos usuários via e-mail foi estabelecido, percebeu-se que alguns estudantes exibiam um perfil de “caçadores de certificados”, seja para ampliar o currículo, seja para apresentar no seu curso de graduação a fim de obter horas complementares, ou para outros fins. Esta estratégia baseia-se na análise de custo/benefício por parte do estudante trapaceiro, uma vez que o risco de ser apontado pela plataforma como “trapaceiro” é nulo – pois não há como provar de forma definitiva este tipo de comportamento – deseja-se descobrir qual “custo” que alunos com este perfil estão dispostos a pagar para obter o certificado sem estudar.

Porém, é importante dosar este “custo” – que pode ser, por exemplo, reduzir a carga horária dos certificados, ou aumentar a quantidade ou dificuldade de exercícios avaliativos – para não afugentar os estudantes que querem estudar. Dessa forma, destaca-se que cursos que já sejam difíceis por natureza de conteúdo devem equilibrar as restrições de configurações, para não desmotivar alunos realmente interessados, e cursos que tenham conteúdos mais simples devem prezar por restringir certas configurações, para que alunos que tentem trapacear sejam inibidos.

Com base nessa perspectiva é que esta Tese foi desenvolvida com o intuito de identificar o perfil dos alunos “caçadores de certificados”, para poder entender qual é a melhor forma de desestimular este comportamento, sem prejudicar os demais alunos. Além disso, busca definir quais são as características dos MOOCs que mais impactam no comportamento dos alunos. A identificação dos alunos que tem atitudes indicativas de “caçadores de certificados” foi realizada a partir dos dados gerados pelos relatórios da plataforma, pois fornecem informações sobre todas as ações efetuadas pelos alunos, e a análise do comportamento desses estudantes foi desenvolvida com técnicas exploratórias de MDE. Para a definição das características dos MOOCs, técnicas supervisionadas de MDE foram empregadas.

Diante desses apontamentos, destaca-se a originalidade dessa proposta de tese, que visa, diferentemente das pesquisas encontradas até o momento, verificar se existe uma relação entre a persistência na burla e os parâmetros de configurações dos MOOCs da plataforma Lúmina, por meio da análise das características do padrão de comportamento de “caçadores de certificados”. Pois, ainda que os MOOCs tenham diversas características semelhantes, existem diferenças relevantes o suficiente para

que as soluções encontradas em um contexto não possam ser aplicadas *“ipsis literis”* em outros.

No caso do Lúmina, uma grande fração dos inscritos são também alunos da UFRGS, de forma que os certificados têm valor como horas complementares. Ademais, podem ser usados como títulos em concursos públicos ou como carga horária em cursos de capacitação. Outro fator que especifica as análises, é o formato dos cursos, que não é comum, pois eles não têm restrição de navegação nem período de realização definido. Outro aspecto obstante para soluções serem generalizadas neste contexto, é que o modelo de dados não é o mesmo de outras plataformas, de forma que as soluções de terceiros não são facilmente aplicáveis.

Quanto a isso, a partir da identificação dos principais atributos apresentados por alunos que não estão engajados, ou seja, a compreensão de como esses alunos se comportam quando realizam um curso na plataforma, serão sistematizadas propostas de configurações para os MOOCs do Lúmina, com potencial para provocar mudanças na quantidade de esforço necessário para obter o certificado. Isso pode estimular os estudantes não comprometidos com seu aprendizado a desistir, todavia essas alterações nos cursos não devem ser tão significativas a ponto de desmotivar os demais, equilibrando o esforço requerido para finalizar um MOOC e a recompensa ao final desse processo. Isso também evidencia a singularidade dessa proposta, pois esse limiar será desenvolvido especificamente para os MOOCs da plataforma Lúmina.

1.2 QUESTÕES DE PESQUISA

A partir das observações levantadas foram definidas as questões de pesquisa que orientaram a condução desse estudo:

Questão 1 – É possível identificar, dentre os estudantes que obtiveram certificados, quais têm o perfil de “caçador de certificados”?

Questão 2 – Quais as configurações de um MOOC desencorajam alunos que não estão comprometidos com a aprendizagem, sem desmotivar aqueles que queiram aprender com o curso?

1.3 OBJETIVOS

O objetivo geral desta Tese é identificar parâmetros de configuração de MOOCs que desestimulam estudantes que têm comportamentos de caçadores de certificados a obter certificações, ao mesmo tempo em que não desestimulam estudantes engajados, na Plataforma de MOOCs da Universidade Federal do Rio Grande do Sul (UFRGS), o Lúmina. Para tanto, é preciso identificar e caracterizar o perfil dos “caçadores de certificados”, estudantes que buscam explorar características da plataforma e dos cursos, para obter um certificado sem se dedicar a sua aprendizagem. Para alcançar o objetivo geral mencionado, os seguintes objetivos específicos são apontados:

1. Extrair conjuntos de dados relevantes presentes nos bancos de dados do Lúmina;
2. Transformar e processar dados presentes nos bancos de dados do Lúmina, para melhorar sua interpretabilidade;
3. Aplicar técnicas de MDE que apoiem na identificação de alunos caçadores de certificados;
4. Aplicar técnicas de MDE que apoiem na identificação de parâmetros de MOOCs que mais impactam nos comportamentos dos alunos;
5. Interpretar os resultados obtidos pelas técnicas de MDE utilizadas;
6. Entender os comportamentos dos alunos durante a realização dos MOOCs na plataforma Lúmina, caracterizando-os; e
7. Sistematizar um conjunto de configurações para os MOOCs do Lúmina, que desestimule comportamentos inadequados.

Atingir os objetivos descritos possibilitou, além de uma compreensão mais ampla do perfil comportamental dos alunos do Lúmina, um esclarecimento sobre quais configurações e níveis de dificuldade um MOOC deve possuir para desencorajar alunos que não estão engajados com sua aprendizagem, sem, no entanto, desmotivar aqueles que querem aprender com os conteúdos disponibilizados e receber uma certificação pelo empenho; oportunizando conhecer quais são os limites de complexidade ideais para um MOOC.

No que se refere às relações existentes entre as questões de pesquisa e os objetivos da Tese, a questão de pesquisa n° 1 – “É possível identificar, dentre os estudantes que obtiveram certificados, quais têm o perfil de “caçador de certificados?”

– está estreitamente relacionada ao objetivo principal da Tese (identificar parâmetros de configuração de MOOCs que desestimulam estudantes que tem comportamentos de caçadores de certificados a obter certificados), pois é necessário identificar esses perfis, para então poder entender quais configurações afetam mais a participação destes indivíduos. Além disso, essa questão está vinculada aos seguintes objetivos específicos: Extrair conjuntos de dados relevantes presentes nos bancos de dados do Lúmina; Transformar e processar dados presentes nos bancos de dados do Lúmina, para melhorar sua interpretabilidade; Aplicar técnicas de MDE que apoiem na identificação de alunos caçadores de certificados; Interpretar os resultados obtidos pelas técnicas de MDE utilizadas; e Entender os comportamentos dos alunos durante a realização dos MOOCs na plataforma Lúmina, caracterizando-os.

No que tange à questão de pesquisa n° 2 – “Quais as configurações de um MOOC desencorajam alunos que não estão comprometidos com a aprendizagem, sem desmotivar aqueles que queiram aprender com o curso?” – evidencia-se sua relação com objetivo geral da Tese, pois a identificação dos parâmetros de configuração dos MOOCs que desencorajam comportamentos indesejados, é um dos focos deste estudo. Além disso, essa questão também está vinculada aos seguintes objetivos específicos: Aplicar técnicas de MDE que apoiem na identificação de parâmetros de MOOCs que mais impactam nos comportamentos dos alunos; Interpretar os resultados obtidos pelas técnicas de MDE utilizadas; e Sistematizar um conjunto de configurações para: atividades, recursos e certificação dos MOOCs do Lúmina, que desestimule comportamentos inadequados na plataforma. Pois, além de caracterizar os comportamentos inadequados associados a plataforma, pretende-se dar alternativas para que este seja inibido.

Para finalizar o capítulo de Introdução, a seguir é apresentada a estrutura desta Tese.

1.4 ESTRUTURAÇÃO DA TESE

Para um melhor entendimento de como está estruturada esta Tese, nesse subcapítulo são feitas algumas considerações a respeito de como estão expostos os tópicos necessários ao seu desenvolvimento. Dessa forma, primeiramente é exposto o tema que irá permear o capítulo e depois uma breve contextualização do conteúdo abordado.

O capítulo 2 apresenta uma síntese sobre as principais definições de MOOC, salientado o crescimento de sua relevância e implicações para área da educação, bem como sistematiza as principais características da plataforma de MOOCs Lúmina, destacando os tipos de relatórios de dados disponíveis, a precisão e diversidade dos dados armazenados.

No capítulo 3 são apresentadas a definição e caracterização da MDE, ademais é descrito a evolução e consolidação dessa área de pesquisa, importante para o desenvolvimento desta Tese, cujos propósitos são a aplicação e amadurecimento de técnicas para a exploração de conjuntos de dados coletados em ambientes educacionais, visando o aperfeiçoamento dos processos ligados a educação. Ademais, são expostos alguns trabalhos que realizam a aplicação da MDE, por meio de duas das suas principais técnicas: Aprendizagem de Máquina e Aprendizagem Profunda.

Em consonância com o capítulo 3, no capítulo 4 é apresentado um mapeamento sistemático a respeito do tema de MDE em MOOCs. Tal mapeamento é significativo, pois, por meio dele foi propiciado identificar as áreas que já possuem considerável avanço dentro dessa temática e quais âmbitos precisam ainda de aprimoramento. Além do mais, por intermédio desse mapeamento, foi possível definir com mais propriedade a temática desta Tese, baseando-se nas investigações existentes e apresentando as devidas inovações indispensáveis, particularmente pela especificidade das bases de dados que são utilizadas e pela evolução tecnológica disponível.

O capítulo 5 aborda a temática principal desta Tese: “trapaças” em ambientes computacionais de ensino e aprendizagem. Neste capítulo são expostos trabalhos relacionados a este tema, iniciando com diversos estudos sobre o *“gaming the system”* (enganar o sistema); seguido dos estudos que tratam de trapaças em MOOCs, identificados no decorrer do mapeamento sistemático, em específico sobre o método CAMEO; por fim, é apresentado uma definição sobre os caçadores de certificados no Lúmina.

O capítulo 6 apresenta a metodologia para o desenvolvimento desta Tese, salientando os procedimentos necessários para sua construção e como a pesquisadora os realizou. Nesse sentido, primeiramente a pesquisa é caracterizada; na sequência é apresentada uma contextualização que permeia as hipóteses levantadas; é exposto o processo para identificação dos alunos caçadores de

certificados, que foi dividido em dois: análise do comportamento de caçador e análise dos estudantes-caçadores, além de apresentar a descrição das tabelas de dados utilizadas nestas análises; em seguida são destacados os procedimentos necessários à definição dos parâmetros de configuração dos MOOCs que impactam no comportamentos dos alunos; por fim, é detalhada a especificação das etapas de análises de dados.

No capítulo 7 são expostos os resultados desta Tese, primeiramente estatísticas descritivas sobre a amostra utilizada são apresentadas, posteriormente os resultados da aplicação das técnicas exploratórias de MDE são descritos, abordando o comportamento de caçador de certificado, em seguida expondo as análises que correspondem aos estudantes-caçadores. Por fim, são apresentados os resultados da identificação dos parâmetros de configuração de MOOCs que desestimulam estudantes que têm comportamentos de caçadores de certificados, este processo foi realizado, por meio da utilização de algoritmo de Aprendizagem de Máquina supervisionados.

O capítulo 8 refere-se as discussões acerca dos resultados alcançados, apresenta-se primeiro uma série de considerações sobre os processos desenvolvidos para identificar os caçadores de certificados, e sobre como os parâmetros de configuração de MOOCs influenciam nos comportamentos dos alunos.

Por fim, no capítulo 9 são apresentadas as conclusões desta Tese, que sistematizam as respostas para as duas questões de pesquisa levantadas neste capítulo, além de expor as principais limitações deste estudo e possibilidades de trabalhos futuros.

2 *MASSIVE OPEN ONLINE COURSES*

Este Capítulo aborda o contexto de desenvolvimento desta Tese, que são os MOOCs e a plataforma Lúmina. Nesse sentido, considerou-se importante destacar primeiramente a caracterização dos MOOCs de forma geral, como influenciaram na área educacional, temáticas de estudo desenvolvidas até o momento e estudos de revisão sobre este tipo de curso. Na sequência é descrita a plataforma Lúmina, em que são destacados os formatos dos MOOCs disponíveis, bem como suas características, além disso são apresentados os tipos de relatórios de dados que podem ser utilizados em processos de análise e mineração.

2.1 MOOCS – ORIGEM, CARACTERÍSTICAS E CONTRIBUIÇÕES

O termo “MOOC” foi cunhado por David Cormier em 2008 (CORMIER; SIEMENS, 2010) para descrever um curso on-line de doze semanas o “*Connectivism and Connected Knowledge”*, desenvolvido por George Siemens e Stephen Downes, oferecido pela Universidade de Manitoba, Canadá. Nesse curso houve dois tipos de oferta: presencial com 25 alunos e on-line 2.300 matriculas, um número surpreendente até então (HOLLANDS; TIRTHALI, 2014), sua forma digital foi disponibilizada diretamente pelo site da universidade[3], logo no início do movimento não haviam plataformas específicas para oferta de MOOCs.

Esse MOOC foi desenvolvido com base na filosofia pedagógica conectivista que tem como objetivo principal que as pessoas vivenciassem o que significa fazer parte de um sistema de aprendizagem social, em que a voz do professor não é o ponto central, mas, em vez disso, um nó em uma rede (HOLLANDS; TIRTHALI, 2014). De acordo com DOWNES (2019) o conectivismo pode ser caracterizado como uma teoria de rede de conhecimento e aprendizagem, com ênfase no uso de tecnologia digital para aprimorar e estender a interação on-line.

A partir de então, os MOOCs passaram a despertar grande interesse, particularmente pela sua ampla abrangência. Em 2011 o curso “Inteligência Artificial” oferecido por Sebastian Thrum da Universidade de Stanford atraiu mais de 160 mil participantes de 190 países, sendo que 23 mil deles concluíram (IQBAL *et al.*, 2014).

---

[3] http://ltc.umanitoba.ca/connectivism/

Entretanto, a estrutura e a filosofia pedagógica dos MOOCs oferecidos na Universidade de Stanford a partir de 2011 são muito diferentes dos MOOCs conectivistas, do início do movimento.

Considerando que os MOOCs conectivistas facilitam a aprendizagem por meio de interações dos participantes com uma rede de indivíduos e incentivam a criação, compartilhamento e construção sobre os artefatos uns dos outros (como vídeos e postagens), os MOOCs estilo Stanford foram projetados principalmente para fornecer educação em escala e envolvem uma transmissão direta de conhecimento, mais estruturada e sequencial (HOLLANDS; TIRTHALI, 2014). Nesse sentido, para diferenciar entre as duas abordagens, os termos “cMOOC” e “xMOOC” começaram a ser usados. cMOOC denota a essência no conectivismo, e xMOOC denota o exponencial, com foco nas matrículas massivas, ou extensão, por exemplo como o HarvardX sendo uma extensão da Universidade de Harvard, ou o MITx como uma extensão do Instituto de Tecnologia de Massachusetts (MIT) (HOLLANDS; TIRTHALI, 2014). Salienta-se que HarvardX e o MITx são plataformas MOOCs vinculadas as instituições de ensino, e a maioria dos cursos nelas disponibilizados tem como autores os professores de Harvard e do MIT.

Desde então, os MOOCs têm ganhado popularidade, e segundo King, Robinson e Vickers (2014) eles revolucionaram o setor educacional, oferecendo novas oportunidades na educação. Segundo Zhang (2016), promovem a acessibilidade global em larga escala com recursos on-line e abertos. Para Almatrafi, Johri e Rangwala (2018) essa nova forma de estudar é uma expansão do *e-learning* e da educação à distância, no sentido de que milhões de pessoas puderam ter acesso a esses cursos, mesmo sem vínculo com as instituições de ensino, o que no modelo *e-learning* tradicional não era possível.

Estima-se que os MOOCs tiveram mais de 58 milhões de usuários registrados após uma década de surgimento, e atualmente centenas de universidades oferecem milhares de cursos em plataformas diferentes, como: *Coursera*, *edX*, *Udacity* e *FutureLearn* – cujos cursos estão, em sua maioria, em inglês; *XeutangX*, em chinês; *MirindaX*, espanhol; *Lúmina* da Universidade Federal do Rio Grande do Sul, *Eskada* da Universidade Estadual do Maranhão (UEMA), *Moodle* do Instituto Federal do Rio Grande do Sul (IFRS) e o *PoCa* da Universidade Federal de São Carlos (UFSCAR), em português. Os MOOCs criaram uma grande expectativa sobre seu potencial de transformação do processo educacional, no sentido da democratização do ensino e

aprendizagem, por causa de suas características de serem (em sua maioria) gratuitos, abertos e permitirem aos alunos terem acesso a materiais e aulas de professores vinculados a instituições de ensino renomadas, bem como aprender em seu ritmo e a partir de qualquer lugar (ROMERO; VENTURA, 2017).

Embora uma das primeiras bandeiras dos MOOCs tenha sido a abrangência demográfica e social, oferecendo cursos gratuitamente e acesso a materiais e certificados on-line, isso mudou ao longo dos anos, pois algumas das principais plataformas, como *Coursera* e *edX*, gradualmente transformaram seus modelos de negócios, abandonando suas ofertas de certificados gratuitos e, em alguns cursos os usuários que não pagam podem não conseguir acessar o pacote completo de materiais do curso, incluindo a certificação (SHI *et al.*, 2018).

A popularização dos MOOCs criou uma nova área de pesquisa, delineando vários campos de estudo e oportunidades para instituições de ensino, pesquisadores e profissionais (DENG; BENCKENDORFF, 2017). Vários temas e questões surgiram e progressivamente foram reformulando e redefinindo os enfoques e interesses dos pesquisadores na área da educação. Por exemplo, houve um crescimento substancial na pesquisa referente aos problemas relacionados à desistência/conclusão de alunos (SUNAR *et al.*, 2017), à concepção pedagógica dos MOOCs (GARCÍA *et al.*, 2017), questões sobre a satisfação dos alunos com MOOCs em relação à aprendizagem (YOUSEF *et al.*, 2015) e sobre o envolvimento dos alunos com as plataformas (GUO; KIM; RUBIN, 2014). Além disso, existe um número considerável de estudos de revisão sobre MOOCs como os trabalhos apresentados por Fournier e Kop (2015); Duru, Dogan e Diri (2016); Davis *et al.* (2018) e Souza e Perry (2019), demonstrando o aumento gradual e o desenvolvimento de pesquisas neste sentido, oferecendo categorização e caracterização de vários temas emergentes, tópicos, tendências, interesses de pesquisa, métodos, questões, desafios, dentre outros.

Fournier e Kop (2015) por exemplo, realizaram uma revisão sistemática sobre as várias estruturas que visam orientar os esforços de pesquisa, desenvolvimento e avaliação em torno de MOOCs, revelando as seguintes áreas de interesse: 1) Análise de Aprendizagem (AA); 2) *Big Data* (BD); 3) Mineração De Dados Educacionais (MDE); 4) Questões de ética e privacidade em ambientes em rede; e 5) O uso de dados pessoais de aprendizado para alimentar o processo de pesquisa e desenvolvimento.

Na pesquisa de Duru, Dogan e Diri, (2016) foi realizada uma revisão de literatura com foco em pesquisas que realizaram análises de desempenho e aprendizado em MOOCs, enumerando descobertas sobre o uso da previsão de desempenho de alunos e AA em MOOCs. Segundo os autores, as áreas mais pesquisadas nesse segmento são: 1) Previsão de resultados acadêmicos; 2) Painéis de cursos e programas; 3) Avaliação Curricular; 4) Priorização de resultados de aprendizado; 5) Definição de políticas de curso e instrução; e 6) Definição de qualidade acadêmica.

O trabalho realizado por Davis *et al.* (2018) oferece uma síntese de descobertas anteriores no domínio das estratégias de aprendizado ativo, empiricamente avaliadas em ambientes digitais de aprendizagem. A principal preocupação do estudo foi avaliar esses achados visando a aprendizagem em MOOCs. Os autores consideraram pesquisas publicadas entre 2009 e 2017, que realizaram avaliações empíricas de estratégias de aprendizagem. Como resultados os autores identificaram três estratégias promissoras para alavancar efetivamente o aprendizado em MOOCs: 1) Aprendizado Cooperativo; 2) Simulações/Jogos; e 3) Multimídia Interativa.

Por fim, na revisão sistemática apresentada por Souza e Perry (2019), foram caracterizados os principais objetivos e desafios da identificação do comportamento de alunos em MOOCs, utilizando Aprendizagem de Máquina, uma das principais técnicas aplicadas na MDE. No que se refere aos principais objetivos de pesquisa nos artigos revisados pelas autoras, lista-se:

- As possibilidades da aplicação das técnicas de classificação e agrupamento, em que foram citados alguns sistemas, que empregam tais técnicas como: *DropoutSeer* – baseado em Aprendizado de Máquina, este software auxilia instrutores e pesquisadores a analisar a relação entre o desempenho do aluno e o abandono; e *Alaska* – software que estende o suporte de análise de aprendizagem da plataforma Khan Academy, para que as informações possam ser usadas para agrupar os alunos em classes.
- Otimização das metodologias utilizadas, como: aplicação de classificadores em cascata, realizada por meio de uma combinação de vários classificadores diferentes para prever o abandono; e a utilização da “Engenharia de Características”; um método aplicado para criar novos

atributos a partir dos dados coletados, o que leva a uma melhor previsão e desempenho.

Em relação aos desafios relatados nos artigos revisados por Souza e Perry (2019), os resultados da revisão da literatura apontaram para os seguintes itens:

- Incompatibilidade entre plataformas: diz respeito à estrutura dos MOOCs e às métricas de avaliação dos alunos, que são muito diversas e evoluem rapidamente; por este motivo, uma solução unificada, que poderia ser utilizada por vários provedores de MOOC, ainda não é viável.
- Complexidade da manipulação de dados*:* os dados disponíveis nos MOOCs precisam ser trabalhados antes de serem usados no treinamento de algoritmos e no processo de previsão, ou na geração de agrupamentos, isso demanda muito esforço, lidar com a grande diversidade de ações dos alunos que podem ser capturadas por plataformas on-line (por exemplo, postar uma pergunta em um chat e assistir a um vídeo) é uma tarefa complexa; e a variabilidade dos dados coletados dos MOOCs não a favorece.
- Problema de desequilíbrio de classe: como o número de evasões (maioria) é muito maior do que o número de alunos que concluem o curso (minoria), existe a necessidade de um modelo que mitigue o efeito de dados desequilibrados, para tornar os algoritmos menos tendenciosos.
- Influência de fatores externos: fatores externos têm forte influência na decisão de desistência, ou engajamento dos alunos e não podem ser identificados, por meio das interações com a plataforma, tais fatores são bastante heterogêneos e difíceis de tratar.
- Dificuldade em manipular os métodos por pessoal não treinado: os algoritmos de Aprendizado de Máquina são difíceis de entender por pessoas não treinadas, como administradores de plataforma, instrutores e professores. Este é um desafio porque para projetar uma rotina de previsão usando esses algoritmos é importante entender o problema, os usuários e o contexto; e esse conhecimento pode não estar disponível para os programadores, sendo necessária uma equipe multidisciplinar.

Com essa pesquisa as autoras puderam observar que os MOOCs têm potencial para se tornarem um modelo de aprendizagem sustentável, fornecendo aos

estudantes uma variedade de possibilidades educacionais e diferentes benefícios, contudo essa realidade ainda parece distante perante os desafios enfrentados por cursos ofertados nessa modalidade.

Conforme vem sendo destacado no decorrer desse estudo, os MOOCs ganharam notoriedade desde sua origem e se integraram ao processo de ensino de grandes instituições. No Brasil muitas dessas instituições, como a Universidade Estadual do Maranhão, o Instituto Federal do Rio Grande de Sul e a Universidade Federal de São Carlos possuem plataformas de oferta para esses tipos de cursos, que possibilitam uma ampliação dos conteúdos que podem ser ofertados pelos professores, beneficiando muitos alunos. Por consequência, a Secretaria de Educação a Distância da UFRGS também se mobilizou para viabilizar um ambiente virtual de aprendizagem para oferta de MOOCs, a Plataforma Lúmina[4], hoje responsável pela manutenção de diversos cursos com milhares de estudantes matriculados, que é o tema do Subcapítulo seguinte.

2.2 PLATAFORMA DE MOOCS – LÚMINA

A pluralidade de recursos da *Web* integrados em plataformas virtuais de aprendizagem oferece oportunidades para selecionar e adaptar informações, colaboração e recursos educacionais. A plataforma de aprendizado é uma maneira de estruturar as instruções, que promovem a organização do conteúdo, interação com os alunos e é usada pela maioria das universidades. As plataformas de distribuição de MOOCs são um ponto de acesso que também têm a função de gerenciamento, não sendo muito diferentes dos tradicionais Sistema de Gestão da Aprendizagem (*Learning Management System* – LMS), possuindo ferramentas para controlar a participação dos alunos e a distribuição do conteúdo do curso. A plataforma conecta professores com alunos e suporta todo o ciclo de vida dos MOOCs, proporciona as ferramentas para gerenciar e fornecer o conteúdo e hospeda todas as videoaulas, atividades e fóruns de discussão. Criar, registrar e produzir o material do curso são responsabilidades do instrutor.

Assim também é a plataforma Lúmina, que começou a oferta dos primeiros cursos em setembro de 2016, com a missão de democratizar o acesso à educação

[4] https://lumina.ufrgs.br/

pública, gratuita e de qualidade, disponibilizando sem nenhum custo aulas e conteúdos produzidos por professores da UFRGS, especialistas em suas áreas. O Lúmina foi uma iniciativa da Secretaria de Educação a Distância (SEAD) da UFRGS, e encontrou no NAPEAD – Produção Multimídia para Educação, um núcleo da SEAD – as condições técnicas e de recursos humanos para que a plataforma se efetivasse. No primeiro semestre de 2022, o Lúmina possuía 88 MOOCs ativos – sem contar os encerrados – distribuídos em 5 áreas.

Os cursos disponibilizados na plataforma Lúmina são desenvolvidos por professores, alunos ou servidores da UFRGS, e não possuem um desenho instrucional padronizado. O Lúmina é uma instalação do Moodle, com um tema personalizado. A configuração empregada nos MOOCs do Lúmina segue o seguinte padrão:

- O conteúdo é elaborado predominantemente no formato de vídeos, entretanto podem ser usados áudios, apresentações, textos, imagens e demais tipos de materiais que sejam possíveis de serem inseridos no Moodle;
- Todos os cursos têm um vídeo de apresentação, disponível mesmo sem o cadastro do usuário;
- Os MOOCs possuem módulos com informações sobre o curso e professores;
- Uma pesquisa sobre o perfil do aluno;
- As atividades avaliativas realizam-se no formato de fóruns ou questionários de múltipla escolha, dando acesso ao Certificado, que tem a quantidade de horas especificada pelo professor do curso e os tópicos abordados.

A imensa maioria dos MOOCs do Lúmina são auto formativos e não há interações com professores. Em vista disso, todas as ferramentas à disposição no Moodle que não requeiram a acompanhamento de um professor e/ou tutor, podem ser utilizadas. Exemplo disso, são os fóruns de discussão, dispositivo do Moodle que, em alguns MOOCs da plataforma, dispõem de propostas bem norteadas pelos professores/autores, ainda que esses não tenham interações com os alunos. Todavia, destaca-se que em alguns MOOCs há interações com os professores e/ou tutores dos MOOCs, o que aumenta a interatividade e participação.

De forma geral, os professores/autores dos MOOCs na plataforma têm grande autonomia pedagógica para o desenvolvimento de suas propostas. As restrições derivam prioritariamente do Ambiente Moodle, o que leva os proponentes dos cursos a um planejamento considerando o que é ou não possível fazer com as funções disponíveis. Todos os recursos dos MOOCs – textos, vídeos, áudios – são estruturados no Lúmina por um integrante da equipe do Napead, o que reduz a ocorrência de problemas de edição ou formatação. O formato dos MOOCs no Lúmina é não-linear, assim o estudante decide a ordem em que deseja acessar os materiais do curso.

O Lúmina, além do gerenciamento dos MOOCs, também armazena os dados dos alunos matriculados, esses referem-se ao perfil e as interações dentro dos cursos (por exemplo: notas e postagens em fóruns de discussão). Ao inscreverem-se na plataforma, os alunos concordam com um Termo de Uso, que traz uma sessão sobre a Política de Privacidade, que esclarece sobre a coleta de dados. Para visualizar esses dados a plataforma possibilita a geração dos seguintes relatórios – visíveis para os professores de cada curso e para os usuários com perfil de administrador:

- Conclusão do curso e Conclusão das atividades: informa nome e e-mail, traz todas as atividades, tanto acesso de recursos, como atividades avaliativas, seu status se foi concluída ou não e a data e horário de conclusão, assim como da geração do certificado;
- Logs e Logs ativos: compilação de todas as interações que o usuário teve com a plataforma, contendo: data, nome completo do usuário, usuário afetado, contexto do evento, componente, nome do evento, descrição, origem, endereço IP. Os Logs ativos, tem a mesma estrutura dos Logs e dizem respeito aos alunos que estão realizando o curso no momento do acesso;
- Atividade do curso: lista de todos os recursos de um MOOC, e o último acesso realizado a ele;
- Estatísticas Gerais: apresenta os números com relação a quantidade de alunos inscritos, por dia do mês, ou meses do ano;
- Participação do curso: contêm a quantidade de atividades realizadas pelos alunos no curso, o que proporciona uma visão geral de todos os inscritos e como está a participação de cada aluno; e

- Notas: agrupa as notas atribuídas a cada atividade avaliativa realizada pelos alunos.

Destaca-se também que, por meio da ferramenta *Configurable Reports* é possível gerar relatórios personalizados, usando *Structured Query Language* (SQL) diretamente na base de dados do Moodle.

Embora os dados disponibilizados pelos relatórios da plataforma possibilitem uma visão ampla do processo de aprendizagem dos estudantes, estes necessitam de aprimoramentos para que se possa ter uma compreensão mais precisa e detalhada do comportamento dos alunos no decorrer da realização de um MOOC, o que se reconhece como uma limitação para o desenvolvimento deste estudo.

Todavia, com a escolha dos métodos e técnicas de mineração adequados aos tipos de dados armazenados na plataforma Lúmina, é possível extrair informações relevantes sobre os comportamentos de aprendizagem dos alunos no decorrer da realização de um MOOC. Para isso, foi preciso realizar um estudo sobre as principais técnicas utilizadas em mineração de dados, e por se tratar de um estudo com dados provenientes de um contexto educacional, na literatura foi encontrado um grande corpo de estudos sobre a Mineração de Dados Educacionais. Uma abordagem associada as áreas de Educação, Estatística e Ciência da Computação, baseada em técnicas de mineração de dados utilizadas em diversos cenários, como: Aprendizagem de Máquina e Aprendizagem Profunda, mas dedicada a tratar desafios da área da Educação, levando em conta as particularidades dos dados extraídos desse contexto. Neste sentido, o capítulo seguinte apresenta uma caracterização da MDE, destacando a evolução desta área e estudos com aplicações em ambientes educacionais reais, como apresentado nesta Tese.

3 MINERAÇÃO DE DADOS EDUCACIONAIS

O volume de dados gerados pela navegação na internet aumentou expressivamente nos últimos anos, devido particularmente às redes sociais e à estratégia de buscadores de usar esses dados como forma de publicidade direcionada. Atentas a este potencial, diversas entidades têm buscado eficiência na captação, organização e armazenamento de bases de dados de grande volume. Entretanto, a chave para avançar é entender como converter estes dados em conhecimento sobre o usuário. À vista disso, a Mineração de Dados (*Data Mining*) está cada vez mais difundida como uma estratégia para extração de conhecimentos e informações a partir de dados, que tem potencial para apoiar o processo de tomada de decisão. Para Aggarwal (2015, p. 1) a Mineração de Dados (MD) é definida como:

> O estudo de coleta, limpeza, processamento, análise e obtenção de informações e ideias úteis de dados. Existe uma grande variação em termos de domínios problemáticos, aplicativos, formulações e representações de dados encontradas em aplicativos reais.

As técnicas de Mineração de Dados passaram a ser empregadas com sucesso também no contexto educacional, auxiliando em diversos cenários, sendo conhecida como Mineração de Dados Educacionais (MDE). A MDE foi a principal estratégia para investigação e análise do comportamento dos alunos do Lúmina, portanto é parte importante do desenvolvimento da metodologia desta pesquisa, por isso neste Capítulo são abordados seus principais aspectos, como: definição e objetivos, sua evolução como área de pesquisa e alguns estudos que apresentam aplicações desta abordagem a conjunto de dados reais.

3.1 DEFINIÇÃO E OBJETIVOS DA MDE

Nos últimos anos as práticas de ensino e aprendizagem têm se modificado em decorrência do avanço tecnológico, o que levou à maior instrumentação do setor educacional, em softwares voltados para o ensino, na administração digital dos registros acadêmicos pelos gestores das instituições e no uso da internet para a aprendizagem, em especial pela popularização do *e-learning*. Todos esses fatores impulsionaram um crescimento exponencial no volume de dados, e para analisá-los é

imprescindível contar com recursos computacionais, caso contrário a tarefa torna-se impraticável.

Dessa forma, as técnicas de mineração de dados estão ganhando cada vez mais importância no setor educacional, pois são uma forma de acompanhar, analisar e avaliar o processo de aprendizagem. Provavelmente as técnicas de mineração de dados podem fornecer aos formuladores de políticas educacionais modelos para apoiar seus objetivos de aprimorar a eficiência e a qualidade do ensino e da aprendizagem. Além disso, o uso de diferentes técnicas de mineração de dados pode ser visto como base para uma mudança sistêmica, capaz de impactar de maneira positiva nas soluções de problemas específicos das Instituições de Ensino; por exemplo, viabilizando soluções que envolvem a personalização dos ambientes educacionais ou fornecendo suporte para o processo de tomada de decisão no ambiente educacional.

Sendo assim, necessita-se de técnicas e ferramentas que auxiliem na tarefa de verificar, interpretar e relacionar esses dados, com o intuito de gerar conhecimento útil e relevante, o que, segundo De Los Reyes *et al.* (2019) já era um objetivo das técnicas de MD, empregadas para identificar padrões de comportamento e encontrar *insights* que provoquem melhorias em produtos e serviços.

O fluxo de trabalho de Mineração de Dados é iterativo e contém as seguintes fases: coleta de dados, processamento e análise. A etapa da coleta é específica da plataforma e geralmente fora do domínio do cientista de dados. O processamento refere-se às transformações que podem ser necessárias para adaptar o banco de dados aos algoritmos, tais como: renomear, selecionar ou filtrar colunas, juntar tabelas, converter tipos e criar variáveis (contagens, somas, médias). A análise é a etapa reflexiva do processo, quando o analista conclui sobre os dados ou busca outras formas de análise. Sendo um processo iterativo, pode ser que haja necessidade de voltar à etapa anterior para adequações.

A MDE, levando em consideração seus objetivos, técnicas e processo de funcionamento, forma uma importante metodologia para apoio a vários cenários educacionais. Em cursos baseados no *e-learning*, com muitos alunos, e que (na maioria) não têm acompanhamento de professores ou tutores – o que é o caso em MOOCs – tais técnicas tornam-se uma solução ainda mais interessante, pois talvez permitam acompanhar e compreender o processo de aprendizagem, bem como outros fatores que o influenciam. Como por exemplo, identificar que tipo de abordagem

instrucional pode propiciar mais ganhos ao aluno, analisando que atributos retratam melhor seu comprometimento com o curso. Ademais, estimula-se a oportunidade de averiguar se o aluno está compreendendo ou não os conteúdos, distinguir níveis de motivação, engajamento nas tarefas on-line, descoberta de fatores ou parâmetros comportamentais de finalização e êxito em um curso, reconhecer padrões de interação, encontrar técnicas ou métodos que colaborem para a continuidade dos estudantes nos cursos até a conclusão (PURSEL *et al.*, 2016), bem como detectar possíveis fraudes, ou trapaças no sistema de aprendizagem. Estes fatores podem ajudar a personalizar o ambiente e os métodos de ensino, para oferecer melhores condições de aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

Em suma, a MDE tem se desenvolvido e pode ser considerada como uma das formas mais promissoras para extração de informações de bases de dados educacionais e suas técnicas têm se tornado cada vez mais eficientes e eficazes, graças ao número crescente de dados disponíveis e dos avanços computacionais. Para entender melhor como se deu esse desenvolvimento e como a MDE tem sido empregada no decorrer do tempo, são descritos os aspectos de sua evolução no subcapítulo na sequência.

3.2 A EVOLUÇÃO DA MDE

A disponibilidade de grandes bases de dados educacionais, fomentada pelas modernas plataformas e mídias educacionais, combinadas com avanços na computação, formam a composição ideal para o surgimento da MDE. Embora existam relatos sobre publicações a respeito deste tema desde 1995 (ROMERO; VENTURA, 2007) o primeiro *workshop* foi realizado em 2005, em Pittsburgh, Pensilvânia, tendo sido seguido por várias oficinas e, em 2008, ocorreu a primeira Conferência Internacional sobre MDE realizada em Montreal, Quebec. As conferências anuais sobre MDE impulsionaram o surgimento do *Journal of Educational Data Mining*, que publicou sua primeira edição em 2009, e na sequência, o primeiro manual sobre MDE foi publicado em 2010 (ROMERO *et al.*, 2010). Posteriormente, em 2011, a Sociedade Internacional de Mineração de Dados Educacionais foi formada com o objetivo de promover pesquisa científica na área interdisciplinar da MDE, organizando as conferências e os periódicos. No campo das publicações, uma primeira revisão de literatura foi apresentada por Romero e Ventura (2007), seguido de um modelo teórico

proposto por Baker e Yacef (2009), e uma revisão bem mais abrangente sobre MDE foi desenvolvida por Romero e Ventura (2010). Na sequência, outras publicações iniciaram um amplo movimento de pesquisas nesse âmbito, e tiveram grande notoriedade.

Nesse sentido, no decorrer da pesquisa bibliográfica realizada durante a elaboração desta Tese, muitas publicações sobre MDE foram analisadas, e algumas delas eram revisões de literatura que reportavam como a MDE tinha sido aplicada em diversos contextos educacionais, seus objetivos, as técnicas mais utilizadas, verificação de resultados alcançados, validação dos benefícios proporcionados, identificação de avanços e desafios. Foram consideradas relevantes para esta pesquisa as revisões de Shahiri, Husain e Rashid (2015); Sukhija, Jindal e Aggarwal (2015); Aldowah, Al-Samarraie e Fauzy (2019); e Romero e Ventura (2020). Tais revisões foram fundamentais para um aperfeiçoamento da compreensão da evolução da MDE no decorrer de sua consolidação como área de pesquisa e são sintetizadas na sequência.

A revisão sobre MDE de Shahiri, Husain e Rashid (2015) forneceu uma visão geral das técnicas de mineração de dados que eram usadas para prever o desempenho dos alunos, em publicações datadas entre 2002 e 2015. O estudo também se concentrou em como os algoritmos de previsão poderiam ser usados para identificar os atributos mais importantes dentre a diversidade de dados dos alunos. Nessa revisão, os autores tinham duas questões de pesquisa: 1) Quais são os atributos mais importantes empregados na previsão do desempenho dos alunos? E 2) Quais as técnicas/algoritmos de previsão mais eficientes?

Quanto aos principais atributos, Shahiri, Husain e Rashid (2015) apontam que foram utilizadas com frequência a média cumulativa de notas e a avaliação interna, empregadas por 10 dos 30 artigos selecionados para a revisão. Os autores também chegaram à conclusão de que a Aprendizagem de Máquina era a técnica mais usada, citando as Redes Neurais como o algoritmo com maior precisão (98%) para previsão do desempenho dos alunos, seguida das Árvores de Decisão (91%), Máquinas de Vetores de Suporte e *K-Nearest Neighbors* (K-Vizinhos mais próximos – KNN) (83%) e Naïve Bayes (76%). No entanto, é importante ressaltar que os índices de eficácia são resultado da interação entre a complexidade da questão de pesquisa com a qualidade (e algumas vezes, a extensão) dos dados, não sendo uma avaliação a respeito dos métodos em si.

Na sequência foi analisada a revisão sistemática de Sukhija, Jindal e Aggarwal (2015) que descreveram a evolução da MDE, trazendo à tona os aspectos e resultados de vários estudos divididos em 3 gerações: 1ª geração de 2001 a 2005; 2ª geração de 2006 a 2010; e 3ª geração, de 2011 até 2015. No período de 2001 a 2005, os autores relatam que as pesquisas se basearam no uso da MDE como uma ferramenta para antecipar os padrões que ajudam na avaliação de cursos on-line. Os autores evidenciaram que as primeiras publicações relacionadas à MDE, produziram pesquisas com inclinação para o ambiente de aprendizagem baseado na *Web,* devido especialmente à grande disponibilidade de dados em cursos on-line. No final deste período, as pesquisas estavam com foco no uso de algoritmos evolutivos para mineração de dados da internet.

Referente ao período de 2006 a 2010, a MDE evoluiu e os estudos passaram a buscar a aplicação de algoritmos mais eficientes. Os bancos de dados usados se tornaram provenientes de sistemas de EaD, baseados na *Web* e vinculados à grandes instituições de ensino, bem como o tamanho desses bancos de dados aumentou. Além disso, houve uma inclinação dos pesquisadores para análises preditivas de dados, com relação a prever os problemas e identificar os alunos em potencial, com alta probabilidade de apresentar um desempenho acadêmico ruim, nesse sentido, sistemas de apoio à decisão para equilibrar a demanda e a oferta educacional também foram desenvolvidos. Durante esse período, a implementação da classificação baseada em Árvores de Decisão e Redes Neurais Artificiais se acentuou.

Finalmente, no que tange ao período de 2011 a 2015, a MDE evoluiu para incorporar técnicas melhores e mais eficientes, conseguindo integrar novas e mais eficientes regras de associações. Sukhija, Jindal e Aggarwal (2015) afirmam que os bancos de dados usados neste período ficaram consideravelmente maiores do que os anteriores e essa crescente no volume dos dados foi acompanhada do desenvolvimento de novas técnicas para MDE, como por exemplo, a retomada dos estudos com Aprendizagem Profunda *(Deep Learning)*.

Sukhija, Jindal e Aggarwal (2015) também apontaram cinco lacunas na área de MDE: 1) Indisponibilidade de conjuntos de dados consistentes que sejam grandes o suficiente para refletir o sistema educacional e seu funcionamento; 2) Necessidade de integração e versatilidade nos conjuntos de dados; 3) Grande parte das técnicas de mineração foram aplicadas isoladamente e poucos trabalhos foram realizados utilizando técnicas híbridas; 4) Havia falta de confiança das autoridades nos

resultados da MDE; e 5) Necessidade de comparar métodos. Pode-se dizer que embora as descobertas dos autores fossem fortemente fundamentadas, o cenário se modificou bastante deste então. No que se refere à primeira lacuna ressalta-se que com a evolução dos MOOCs, bases com milhões de dados estão disponíveis, como exemplo a pesquisa de Northcutt, Ho e Chuang (2016), que utilizou uma base de dados gerada a partir de uma plataforma MOOC com mais de 1,8 milhões de usuários, que produziram em média de 200 a 1500 interações com a plataforma, cada um, por curso realizado. No que tange à segunda limitação, salienta-se que em relação à integração das bases, ela se mantém, pois, não é possível integrar duas bases de forma simples, sem necessidade de um grande esforço de processamento. No que diz respeito à versatilidade dos dados – qualidade de não ser colinear – pode-se dizer que houve mudanças, pois, vários tipos diferentes de dados são usados nos modelos de MDE atuais.

A terceira lacuna, sobre uso de métodos híbridos, pode ser considerada a que menos coincide com a realidade dos experimentos realizados na área de MDE atualmente, pois muitos pesquisadores têm empregado técnicas de MDE em conjunto com outras ferramentas de pesquisa como Gallén e Caro (2017) que utilizaram algoritmos de agrupamento e um questionário respondido pelos alunos para analisar os motivos pelos quais uma pessoa se inscreve em um MOOC. Com relação à quarta limitação, supõe-se que com os grandes avanços tecnológicos disponíveis e a consolidação da Inteligência Artificial (IA), a aceitação da MDE como aporte para tomada de decisões no setor acadêmico tenha crescido. Enfim, quanto à quinta lacuna, é possível destacar que devido aos avanços tecnológicos muitos pesquisadores estejam se dedicado a comparar novas técnicas de MDE com outras mais consolidadas, para verificação da eficácia de modelos.

Na sequência, foi analisada a revisão de Aldowah, Al-Samarraie e Fauzy (2019) que teve como foco o tema MDE e AA para o século XXI no ensino superior. Os autores conduziram a revisão com o intuito de responder à duas questões: 1) Como usar MDE e AA para resolver desafios práticos na educação? E 2) Quais técnicas de mineração são mais adequadas para esses problemas? Para responder a essas perguntas, buscaram artigos publicados entre 2000 e 2017, disponibilizados nas bases: *Scopus*, *Web of Science, Google Scholar, ERIC, Science Direct, DBLP, ACM Digital Library, IEEEXplore e Springer*. Aldowah, Al-Samarraie e Fauzy (2019) afirmaram que as técnicas de MDE e AA podem ser agrupadas em 4 grandes

dimensões: Análise de Aprendizagem Suportada por Computador (*Computer-Supported Learning Analytics* – CSLA); Análise Preditiva Suportada por Computador (*Computer-Supported Predictive Analytics* – CSPA); Análise Comportamental Suportada por Computador (*Computer-Supported Behavioral Analytics* – CSBA); e Análise de Visualização Suportada por Computador (*Computer-Supported Visualization Analytics* – CSVA).

As pesquisas sobre CSLA (120 artigos) concentraram-se principalmente no uso de análise estatística de dados para executar tarefas analíticas sofisticadas, a fim de investigar os comportamentos de aprendizagem colaborativa e de busca de informações dos alunos no contexto de um curso. Os estudos sobre CSPA (253 artigos) focaram, em sua maioria, no uso de funções preditivas ou variáveis contínuas para sugerir maneiras eficazes de melhorar o aprendizado e o desempenho dos alunos, bem como avaliar a adequação do aprendizado. As publicações sobre a dimensão CSBA (80 artigos), em maior parte, criaram modelos de comportamento, ações e conhecimento. Os estudos sobre CSVA (38 artigos) concentraram-se em métodos para explorar visualmente os dados – usando gráficos interativos por exemplo – para destacar informações úteis e elaborar decisões precisas sobre as novas informações descobertas nos dados.

Para finalizar, destaca-se Romero e Ventura (2020), que efetuaram um estudo sobre a MDE e AA, atualizando revisões anteriores de sua autoria (ROMERO; VENTURA, 2007, 2010, 2013, 2017). A publicação de 2020 forneceu informações sobre o estado da arte, revisando as publicações da área no sentido de elucidar: os principais marcos; o ciclo de descoberta de conhecimento; os ambientes educacionais mais utilizados; as ferramentas específicas desenvolvidas; os conjuntos de dados disponíveis gratuitamente; os métodos e técnicas mais empregados; os principais objetivos; e por fim, as tendências futuras nessa área de pesquisa. Devido à grande amplitude da revisão, são abordados os itens considerados mais relevantes: as principais mudanças na MDE e na AA, desde o início de pesquisas na área pelos autores; e suas conclusões.

Romero e Ventura (2020) apresentaram apontamentos relacionados aos seguintes aspectos: 1) importância e evolução da MDE e AA; 2) avaliação de tendências de estudos anteriores; 3) principais desafios; e 4) tendências para novos estudos nessas áreas.

No que se refere à importância da MDE e a AA, os autores sugeriram que a área cresceu rapidamente nas últimas duas décadas, com duas conferências anuais, dois periódicos específicos e aumento do número de livros, artigos, pesquisas e resenhas relacionados. Os autores declararam, que em 2020, muitas pesquisas envolveram análise e mineração de dados. Isso indica que essas áreas se tornarão em breve maduras e amplamente utilizadas não apenas pelos pesquisadores, mas também por instrutores, administradores educacionais e empreendimentos relacionados, em todo o mundo.

Com relação à comparação entre as tendências levantadas em sua pesquisa anterior, os autores perceberam aumento no uso, disponibilidade e facilidade de acesso às ferramentas para MDE e AA, porém ressaltaram que ainda é necessário desenvolver ferramentas para uso geral. Isso pode acelerar a consolidação de uma “cultura baseada em dados”, que se utiliza de ferramentas de análise para tomar decisões e melhorar seus processos de ensino, aprendizagem e administrativos.

Segundo os autores, os principais desafios percebidos são: transferibilidade e generalização; eficácia e aplicabilidade; e interpretabilidade. Transferibilidade e generalização se referem à utilização de modelos comuns para vários contextos, o que ainda não ocorre na prática. Eficácia e aplicabilidade dizem respeito à realização de ações de intervenção a partir das análises de dados, dos quais constata-se que há uma grande diversidade de modelos desenvolvidos e relatados em publicações, contudo não se têm informações se foram ou são aplicados na prática das instituições de ensino, ou se são efetivos na resolução de problemas dessas instituições. Interpretabilidade concerne à capacidade dos usuários de compreender os modelos gerados para as análises de dados, o que quando não ocorre acarreta um mau aproveitamento de tais modelos.

Finalmente, os autores propuseram algumas ideias, que disseram ser “visionárias e pessoais” que, em suas opiniões, podem formar tendências e direcionamentos muito promissores para a áreas de MDE e AA, para a década de 2020: 1) levar em consideração todos os dados pessoais dos alunos durante toda a vida; 2) aplicação e integração da MDE e AA aos futuros ambientes educacionais tecnológicos; e 3) análise e Mineração de Dados coletados diretamente do cérebro dos alunos para uma melhor compreensão do aprendizado.

Em conclusão, com a análise e interpretação dessas publicações, que sistematizaram boa parte do estado da arte na área, em destaque dos últimos 20 anos,

foi possível compreender o início do processo de adoção de MDE, quais eram as principais técnicas, os dados utilizados para formação de bancos/bases e os resultados que foram alcançados, enfim, de entender o modo pelo qual MDE se estabeleceu como um campo de pesquisa. Além disso, constatou-se que sua evolução se deveu a dois fatores principais: 1) a adoção de grandes bases de dados na educação, impulsionado sobretudo pelo surgimento de cursos *e-learning,* como os do tipo MOOCs; e 2) o avanço das tecnologias computacionais, que são indispensáveis para aplicação e evolução das técnicas vinculadas a MD.

3.4 APLICAÇÕES DA MINERAÇÃO DE DADOS EDUCACIONAIS

Desde a popularização do *e-learning*, e do surgimento da análise automatizada de dados educacionais muitos esforços têm sido realizados para aprimorar a experiência da aprendizagem, por esse motivo a MDE ganhou notoriedade, pois um de seus interesses é explorar a maneira como as pessoas aprendem. Avanços nessa área permitiram coletar e analisar dados sobre os alunos e seus ambientes e explorar o comportamento das pessoas enquanto aprendem.

Nessa perspectiva, logo no início da disseminação e popularização da Inteligência Artificial e da Aprendizagem de Máquina (AM), Baker (2000) as visualizou como um conjunto promissor de mecanismos de software e tecnologias a serem empregadas na melhoria do processo educacional, antes mesmo da consolidação da MDE como um campo de pesquisa. Quando a Aprendizagem de Máquina é empregada como técnica de MDE, pode ajudar a entender situações educacionais e com isso dar apoio no processo de tomada de decisão. Nesse sentido, destacam-se as pesquisas de: Chui *et al.* (2020), Zhang e Wu (2019), Rodrigues *et al.* (2016) e Tan *et al.* (2018). Em que os dois primeiros estudos aplicam algoritmos supervisionados de AM e os demais aplicam algoritmos não supervisionados.

No estudo desenvolvido Chui *et al.* (2020) o objetivo foi desenvolver um modelo para prever o desempenho acadêmico de alunos da graduação, por meio da identificação de padrões de alunos com tendência a ter um baixo desempenho. A base de dados utilizada nessa pesquisa foi da *Open University Learning Analytics,* coletada entre 2013 e 2014*,* contendo dados de 7 cursos de graduação e dados de mais de 32 mil alunos sobre: atividades avaliativas, notas, perfil demográfico e uma apresentação pessoal desses estudantes. Durante a formatação da base foram definidas duas

categorias: Aprovação e Falha, o que configura um problema de classificação binária. Na sequência foi aplicado o algoritmo de Máquinas de Vetores de Suporte sobre a base de dados, tendo o modelo alcançado entre 92% e 94% de acurácia. Concluindo, os autores relataram que com previsões precisas de alunos que tendem a ter um desempenho ruim, ações podem ser projetadas para amenizar as desistências desses alunos e motivá-los a conclusão.

Zhang e Wu (2019), consideraram o contexto de MOOCs para o desenvolvimento de sua investigação, com objetivo de prever o desempenho (notas) nesses tipos de curso. Para isso, os autores utilizaram dados dos alunos que cursaram o MOOC de Programação em Linguagem C, usando os seguintes dados: informações básicas do perfil dos alunos, pontuação nas atividades avaliativas, número de questões solucionadas, pontuação final, postagens nos fóruns de discussão. Quanto à abordagem para a solução do problema, os autores dividiram os resultados dos alunos em classes de A a D (85 a 100 – A; de 70 a 85 – B; de 60 a 70 – C; e >60 – D), portanto tornando-se um problema de classificação. Após a formatação e classificação da base de dados, os autores realizaram a aplicação de 3 algoritmos para geração de modelos de previsão de notas: *ID3, C4.5* e *CART* – todos baseados em Árvore de Decisão. As precisões alcançadas pelos modelos sobre a base de dados de teste foram: IDE3 – 81%, C4.4 – 75%, CART – 76%. Zhang e Wu (2019) afirmaram que os modelos baseados em Árvores de Decisão são consideravelmente simples de serem implementados, e têm precisão relativamente satisfatória, por isso devem ser empregados para apoiar ações que induzam a permanência de alunos em MOOCs, ou em outros cenários educacionais.

Rodrigues *et al.* (2016) desenvolveram sua pesquisa também no contexto de MOOCs, com o objetivo de reconhecer perfis de engajamento. Para esse fim, os autores utilizaram uma base de dados formada por atributos de 5 mil alunos de um MOOC com o tema Nova Gramática da Língua Portuguesa da plataforma *Openredu*. Os dados utilizados pelos autores diziam respeito a 15 atributos que retratavam a frequência de diferentes categorias de postagens, assiduidade e notas. Na busca pelas categorias de engajamento, os autores utilizaram o agrupamento, por meio dos algoritmos *K-means e* Cluster Hierárquico, que em meio aos grupos gerados, possibilitou a identificação de três perfis de engajamento no MOOC investigado: Engajados – 16% do total de alunos; Esporádicos – 26% do total; e Desengajados – 58% do total.

- O perfil *Engajados* é descrito pelos autores como um grupo que possui uma ótima interação via fórum de discussão, mantém ritmo continuo na realização das atividades e pouca variabilidade entre as médias de notas, podendo ser considerado um grupo de alunos que realizam atividades do início ao fim do curso;
- O perfil *Esporádicos* foi detalhado como possuindo uma forte característica de mudanças, com picos repentinos durante o curso, uma das prováveis explicações é que alunos com esse perfil passam longos períodos sem entrar no ambiente e perdem alguns prazos de realização de atividades, embora possuam conhecimentos sobre o tema, que faz com que tenham notas elevadas quando as realizam; e
- O perfil *Desengajados* refere-se à alunos praticamente inativos na realização do MOOC, mais de 90% só realizaram duas atividades, as notas médias e a quantidade de interações nos fóruns são inferiores aos demais grupos. Os autores salientam que esses são indícios que este grupo tem características apenas de visualização de material didático e optam por participar apenas em alguns momentos, permanecendo envolvidos, entretanto não desejam ganhar um certificado, e por este motivo não realizam as atividades.

Rodrigues *et al.* (2016) destacaram em suas conclusões que a abordagem exploratória baseada em algoritmos de agrupamento pode ser usada para ajudar pesquisadores a identificar perfis comportamentais dos alunos em relação ao envolvimento em interações via fórum e durante atividades no decorrer de um MOOC.

Por fim, Tan *et al.* (2018) também aplicaram AM para o mapeamento de perfis de alunos, empregando o agrupamento com o algoritmo *K-means,* no reconhecimento de perfis de aprendizagem dos alunos no MOOC "*E-learning e Culturas Digitais"* que foi criado pela Universidade de Edimburgo. Os autores utilizaram na sua pesquisa três tipos de atributos de 87 alunos que se matricularam nesse MOOC: 1) Comportamento de aprendizagem on-line – número de postagens em fóruns, número de respostas, interações sociais, tempo de acesso a recursos; 2) Auto relatados – informações demográficas, proficiência em inglês, experiência de aprendizagem on-line, formação acadêmica e classificação das experiências de aprendizado com MOOCs; 3) Resultados de correções de questões dissertativas – o professor/tutor do MOOC realizou correções das questões dissertativas dos alunos, que foram incorporadas à

base de dados (as notas variaram entre 0 e 187). Os autores destacaram que a qualidade dos ensaios submetidos refletiu a compreensão dos alunos sobre o conteúdo do curso e habilidades de pensamento e, portanto, a avaliação desses ensaios foi utilizada para medir o desempenho de aprendizagem no MOOC.

Para a aplicação do algoritmo *K-means* Tan *et al*. (2018) formataram uma base de dados incluindo os atributos citados. A partir do processamento do algoritmo e análises sobre as informações geradas, os autores identificaram quatro perfis de aprendizagem entre os alunos que realizaram o MOOC: 1) *Estudantes competentes e ativos*; 2) *Estudantes competentes e inativos;* 3) *Estudantes incompetentes e inativos*; e 4) *Espectadores*. Os autores finalizaram o manuscrito afirmando que os resultados de seu estudo podem fornecer implicações úteis para projetos e implementações de MOOCs, principalmente no que tange a personalização.

A partir dessas pesquisas foi possível perceber como os conceitos apresentados podem ser desenvolvidos no contexto educacional. Não obstante, a AM seja uma das técnicas mais aplicadas a Mineração de Dados em diversos âmbitos, há também outras técnicas que podem ser utilizadas para exploração e análise de dados, como a Aprendizagem Profunda, que é uma abordagem da Aprendizagem de Máquina que trata exclusivamente de Redes Neurais Artificiais, tema tratado na sequência.

Desde 2006 a Aprendizagem Profunda (AP) tem atraído atenção e sido aplicada com sucesso em muitas áreas. No entanto, de acordo com Yang, Zhang e Su (2019) as aplicações da AP no contexto educacional são relativamente escassas, pelo menos até o momento, em comparação à AM. A sua utilização nesse contexto pode levar a um crescente corpo de pesquisa com foco na melhoria da modelagem do comportamento e desempenho dos alunos, ampliando os horizontes de estudos na MDE.

Dentro desse domínio, alguns estudos foram selecionados, por exemplo Lin *et al.*, (2019), que propuseram uma Rede Neural Convolucional treinada a partir de vídeo aulas de MOOCs, com o propósito de detectar e classificar de forma automática os conteúdos exibidos nesses vídeos, a fim de melhorar o desempenho da plataforma de aprendizado on-line. No que diz respeito à Rede Neural Convolucional, sua implementação foi executada com o *Framework Tensorflow* e foi treinada e testada em um conjunto de dados de 16 mil imagens e 72 horas de áudio, pertencentes a dois MOOCs diferentes.

Também destaca-se a investigação realizada por Guo *et al.*, (2019), na qual os autores implementaram uma Rede Neural Profunda Híbrida para a identificação de postagens “urgentes” que requerem atenção imediata de instrutores em fóruns de discussão em MOOCs. Quanto à estrutura, é uma Rede Neural Convolucional combinada com uma Rede Neural Recorrente, implementada para mineração de texto e funciona em 3 etapas: 1) Assimilar simultaneamente as informações semânticas e estruturais das sentenças de texto das postagens; 2) Utilizar as Redes Convolucionais em nível de caractere para capturar informações – isso foi necessário devido a muito ruído, como erros de ortografia e *emoticons* no texto das postagens; e 3) Associar as informações semânticas e estruturais com as informações de caracteres e assim chegar a representação final da frase. Guo *et al.*, (2019) chegaram a resultados que superam a precisão de soluções presentes no estado da arte em até 2,4%, e concluíram que sua pesquisa pode auxiliar professores e tutores a priorizar suas respostas e gerenciar melhor várias postagens, de modo que esses profissionais da educação possam responder às perguntas dos alunos em tempo hábil e ajudar a reduzir as taxas de evasão em MOOCs.

Wen *et al.* (2020) também utilizaram um modelo de AP para pesquisa no âmbito educacional, com o objetivo de identificar antecipadamente a desistência em MOOCs. Nesse sentido, depois que Wen *et al.* (2020) realizaram uma análise dos padrões de comportamento de aprendizagem dos alunos de um MOOC, relataram que esses estudantes geralmente exibem comportamentos de aprendizagem semelhantes em vários dias consecutivos (o *status* de aprendizagem de um aluno para o dia subsequente, provavelmente será semelhante ao do dia anterior). Embasados nessa premissa Wen *et al.* (2020) propuseram uma base de dados formada por atributos relacionados à correlação local de comportamentos de aprendizagem, sobre a qual aplicaram uma Rede Neural Convolucional, gerando um modelo para prever o abandono de alunos em MOOCs. Por fim, o modelo proposto obteve uma precisão que variou de 86% a 89%, e os autores destacaram que as principais contribuições da pesquisa foram: 1) Definição do conceito de *status de aprendizagem*, para encontrar a correlação local de comportamentos de aprendizagem; 2) Construção de uma base de dados formada a partir dos atributos da correlação local de comportamentos de aprendizagem; e 3) Implementação de um modelo construído a partir de uma Rede Neural Convolucional para previsão do abandono de alunos em MOOCs.

Finalizando, cita-se o estudo de Waheed *et al.*, (2020) que desenvolveram um modelo de Rede Neural Profunda com arquitetura Multilayer Perceptron (MLP), com o objetivo de prever o desempenho de alunos em MOOCs. Para isso os autores utilizaram relatórios de 7 MOOCs, com um total 32 mil alunos, e os atributos utilizados foram: perfil demográfico, fluxo de cliques e desempenho nas avaliações. O estudo foi conduzido com base na mineração de dois conjuntos de dados: 1) Notas das atividades avaliativas na plataforma e perfil demográfico; e 2) Atributos trimestrais do fluxo de cliques de cada aluno, o resultado foi um modelo para prever o risco de reprovação. Os autores relataram que em contraste com os métodos estatísticos, as Redes Neurais Profundas facilitam a generalização, o que possibilita inferir padrões escondidos nos dos dados, dando suporte a fazer suposições sobre eles. A precisão alcançada pelo modelo ficou entre 84% e 94% nos experimentos realizados, e Waheed *et al.*, (2020) concluíram que esses resultados demonstram a efetividade do modelo implementado para a previsão precoce do desempenho de alunos em MOOCs. Os autores ainda ressaltaram que estudos como esse, orientados a dados, são necessários para auxiliar instituições de ensino na formulação de uma estrutura de análise de aprendizagem, contribuindo para o processo de tomada de decisão.

Essas publicações propiciaram visualizar vários aspectos sobre estudos com análises de dados educacionais, o que despertou interesse em descobrir quais seriam as principais tendências temáticas de pesquisas com MDE aplicada no contexto dessa proposta de tese, os MOOCs. Para tanto, foi desenvolvido um mapeamento sistemático com o propósito de identificar essas tendências, e auxiliar no reconhecimento de tópicos de estudos, assim como encontrar quais as técnicas de MDE mais utilizadas e as principais oportunidades de pesquisa, como mostrado no capítulo 4.

4 MAPEAMENTO SISTEMÁTICO: TENDÊNCIAS DE PESQUISAS EM MINERAÇÃO DE DADOS EDUCACIONAIS NO CONTEXTO DOS MOOCS

Este Capítulo apresenta o desenvolvimento de um mapeamento sistemático de literatura realizado no primeiro trimestre de 2020, e publicado na Revista Brasileira de Informática na Educação com o título “Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático” (SOUZA; PERRY, 2020), que teve como principal objetivo a investigação na área a qual desenvolveu-se esta Tese. A partir desse mapeamento, foi possível identificar lacunas e oportunidades de pesquisa nos trabalhos publicados até a data de sua realização, e dessa forma, auxiliar a encontrar o melhor contexto para a implementação desta pesquisa.

4.1 EXECUÇÃO DO MAPEAMENTO

Para realização deste mapeamento, adotou-se um método usado em revisões sistemáticas de literatura, pois minimiza o enviesamento da escolha dos manuscritos, na medida em que é feita uma busca dos textos publicados sobre o tema em questão (DENYER; TRANFIELD, 2009). Ramos, Faria e Faria (2014, p. 21) salientam que de acordo com a página *Web* da associação *Campbell Collaboration*[5] “o intuito de uma revisão sistemática é compilar a melhor pesquisa à disposição a respeito de uma questão específica, o que é realizado, por meio da síntese dos resultados de diversos estudos”. Já um mapeamento, como o próprio nome indica, não busca avaliar como o conjunto da literatura responde determinadas questões de pesquisa, e sim fazer um apanhado geral sobre determinada área, apresentando um panorama que permita identificar oportunidades de pesquisa. Do mesmo modo que em uma revisão sistemática, em um mapeamento também é preciso utilizar procedimentos bem estabelecidos para recuperar, avaliar, qualificar e resumir os resultados de trabalhos relevantes na área em estudo.

Para a condução deste mapeamento utilizou-se como referência Ramos, Faria e Faria (2014) que propõem o seguinte protocolo: 1) Definir os objetivos; 2) Definir as *strings* de busca; 3) Definir bases de dados; 4) Definir critérios de inclusão;

[5] Organização sem fins lucrativos que promove decisões e políticas baseadas em evidências, por meio da produção de revisões sistemáticas e outros tipos de síntese de evidências.

5) Definir critérios exclusão; 6) Definir critérios de validade metodológica; 7) Tabular os dados; e 8) Realizar tratamento dos dados. Ressalta-se que Ramos, Faria e Faria (2014) apresentam este protocolo como método para conduzir revisões sistemáticas, mas considerou-se que ele poderia ser adaptado para um mapeamento. Na percepção de Ramos, Faria e Faria (2014) para realização de pesquisas desse tipo, é muito importante que sejam registradas todas as etapas da investigação, não só para que esta possa ser replicada por outro investigador, como também para se verificar que o processo em curso segue uma série de etapas previamente definidas e respeitadas.

O primeiro passo do processo é a definição dos objetivos do estudo, que neste mapeamento é responder à 4 questões de pesquisa: *QP1.* Quais as tendências temáticas e propósitos gerais da utilização de MDE em MOOCs? *QP2***.** Quais as técnicas e algoritmos mais utilizados? *QP3.* Quais as principais oportunidades de pesquisa? *QP4.* Quais os desafios mais relatados em pesquisas na área?

A *string* de busca utilizada foi: ((“educational data mining” OR “EDM”) AND (“massive open on-line course” OR “MOOC”)), e seu equivalente em português. As bases de dados foram: *IEEE Explore Digital Library; ACM Digital Library; ERIC; Science Direct; e Taylor e Francis,* por indexarem muitos trabalhos internacionais, que têm maior alcance. Além do que, os dois principais eventos e periódicos de MDE (*International Conference on Educational Data Mining* e *Journal of Educational Data Mining*) são indexados nessas bases*.*

Quanto aos critérios de inclusão e exclusão das pesquisas, definiu-se: 1) Inclusão – (a) em Inglês ou Português; (b) completos; (c) que tenham passado por um processo de revisão cega; (d) e publicados entre 2015 e 2020. 2) Exclusão – (a) duplicados; (b) revisões ou mapeamentos de literatura; (c) resumos ou artigos publicados em congressos nacionais; (d) publicados em revistas que não sejam “*peer review”*. Quanto a decisão de não incluir artigos apresentados em congressos nacionais ou regionais, entende-se que nesses fóruns as pesquisas tendem a estar em um estágio inicial, sem resultados completos. Além disso, exclui-se, por óbvio, artigos escritos em idiomas não compreendidos pela pesquisadora ou que não atendessem à temática “MDE em MOOCs”.

O sexto passo, relativo aos critérios de validade metodológica, deve assegurar a replicação do processo. Para este trabalho todos os artigos retornados da busca passaram por três níveis de triagem: 1) Os resumos de todos os artigos retornados

foram lidos, dos quais foram selecionados aqueles que atendem aos critérios de inclusão mencionados; 2) Todos os artigos que passaram pela primeira triagem tiveram seus resumos analisados, com a finalidade de gerar as categorias temáticas; 3) Por fim, com o intuito de verificar se as pesquisas selecionadas realmente satisfaziam aos critérios de inclusão, ou deveriam ser descartados por atender a algum critério de exclusão, a autora da Tese realizou a leitura dos artigos, enfocando a metodologia utilizada pelos autores, resultados alcançados e suas conclusões. A leitura da revisão bibliográfica e introdução foi feita de forma rápida, apenas se não fosse compreendido algo nos demais itens uma leitura mais apurada era conduzida. Nessa etapa também, foi efetuada a geração de categorias temáticas de pesquisa.

Sobre a tabulação e o tratamento dos dados, esses foram analisados em uma dimensão quantitativa (contagem de ocorrência para elaboração de estatísticas descritivas) e qualitativa (categorização). Todos os elementos essenciais dos artigos, incluídos nesse mapeamento, estão disponíveis no link que consta no rodapé[6], na sequência são apresentados os resultados do mapeamento.

4.2 RESULTADOS DO MAPEAMENTO

O mapeamento sistemático abrangeu o período de 2015 ao primeiro trimestre de 2020, sendo selecionados artigos completos que relataram o uso de MDE em MOOCs. A Tabela 1 apresenta o quantitativo de artigos retornados das bibliotecas digitais pesquisadas. No primeiro nível da triagem (resultado da *string*) foram selecionados 376 artigos, que em seguida foram avaliados quanto aos critérios de inclusão e exclusão, resultando em 158 manuscritos.

Tabela 1 – Totais de pesquisas encontradas e selecionadas, por base de dados

Bases	N° Artigos Retornados	N° Artigos Selecionados
IEEE Digital Library	104	67
ERIC	56	44
ACM Digital Library	96	28
Science@Direct	69	14
Taylor e Francis	27	5
SciElo	24	0
TOTAL	**376**	**158**

Fonte: Souza e Perry (2020)

6 https://docs.google.com/spreadsheets/d/1FPu4Cr39zWHU0JsmHapeAjXfFXOHDv0X/edit#gid=1746542574

A fonte de publicação com o maior quantitativo de manuscritos retornados e selecionados foi a *International Conference on Educational Data Mining,* com 40 artigos. Em seguida, as revistas *IEEE Access, Computers in Human Behavior e Computers & Education*, respectivamente com 5, 5 e 4 publicações. Os países de origem mais frequentes (considerando apenas primeiros autores) foram Estados Unidos (47) e China (40). Espanha, Suíça, Japão, Austrália, Índia, Equador, Brasil, Alemanha, Chile e Hungria tinham entre 7 e 2 artigos. As publicações sobre o tema tiveram seu ápice entre 2016 e 2018 – a série que vai de 2015 a 2020 indica 20, 38, 38, 40 e 19 publicações, respectivamente. Na sequência as questões levantadas – objetivo do mapeamento – são respondidas.

QP1 – Quais as tendências temáticas e propósitos gerais da utilização de MDE em MOOCs?

Foram identificadas 25 temáticas em meio aos artigos pesquisados, sendo que 4 delas (apresentadas no Quadro 1) respondem por 69% das pesquisas, sendo a Predição (de Desempenho, Abandono e Conclusão) o tema mais investigado. A síntese dos resultados do mapeamento pode ser vista no Quadro 1, que agrupa os artigos em função das tendências temáticas, listando: quantidade de artigos, oportunidades de pesquisa e desafios mais relatados. Este Quadro lista apenas as temáticas com pelo menos 6 artigos[7]. As temáticas mais presentes somaram 122 artigos, 77% do total.

[7] Corte arbitrário, definido pela pesquisadora.

Quadro 1 – Síntese dos Resultados Alcançados, agrupados pela temática.

Temática (Quantidade de Artigos)	Oportunidades de Pesquisa	Desafios
Predição: **de Desempenho (18)** **de Abandono (16)** **de Conclusão (6)**	1- Implementação de modelos de alunos que levem em consideração a leitura como um componente fundamental da aprendizagem. 2- Aplicação do método de empilhamento – abordagem para combinar várias técnicas de aprendizado de máquina em um modelo preditivo – é conhecido por ter um bom desempenho com dados desequilibrados.	1- Desequilíbrio de classe e grande variabilidade no formato de dados MOOCS. 2- Como a maioria dos alunos não conclui ou avança no curso, os algoritmos supervisionados aprendem de forma tendenciosa. 3- Demanda muito esforço na fase do processamento.
Análise de Comportamento (30)	1- Conclusão de curso, procrastinação, regularidade e interação. 2- Exploração de variáveis relacionadas à exibição de vídeos, autoteste e interação no fórum. 3- Análise de cliques em relação ao tempo na tarefa, para avaliar engajamento.	1- "*Cold start problem*" ou "arranque a frio" é um problema potencial em sistemas de modelagem automatizada de dados, diz respeito à questão de que o sistema não pode extrair inferências para usuários ou itens sobre os quais ainda não reuniu informações suficientes. 2- O Fluxo de Cliques pode ser uma medida de análise ruim, pois alunos podem somente clicar na tarefa, mas não estarem engajados em sua realização.
Mineração de Texto (28)	1- Análise da polaridade de palavras de um texto, com o objetivo de entender seu efeito sobre processos e estados de aprendizagem afetiva. 2- Integração de fontes de informação externa ao curso, isso permite uma análise mais detalhada do comportamento do aluno em atividades acadêmicas que devem ser cumpridas.	1- Na Mineração de texto, diferentes gramáticas e diferentes estilos linguísticos tornam a aplicação de algoritmos uma tarefa complexa, e os principais softwares de mineração de texto são proprietários e por isso têm pouca flexibilidade, e estão disponíveis na maioria das vezes para o inglês.
Sistemas de Recomendação (12)	1- Uso de algoritmos de agrupamento e redes neurais.	1- Desequilíbrio de classe e grande variabilidade no formato de dados dos MOOCs.
Identificação de Trapaças (6)	1- Identificação de Trapaças por meio de um algoritmo de detecção baseado na identificação de IPs iguais e no tempo decorrido das submissões de atividades.	1- Não se pode esperar que os estudantes sejam sinceros, sobre desonestidade acadêmica, por isso pesquisas com questionários não são muito confiáveis.
Análise de Redes Sociais (6)	1- Utilização da Análise de Rede Social para analisar fóruns de discussão em MOOCs, com o objetivo de prever desempenho, e traçar perfis de comportamento. 2- Avaliação da eficiência dos fóruns de discussão de MOOCs, comparando com uma rede de aprendizagem social ideal.	1- Muitos softwares de Análises de Redes Sociais são proprietários e pouco flexíveis

Fonte: Souza e Perry (2020)

Além destas 6 temáticas, também encontrou-se pesquisas sobre: Aprendizagem Auto Regulada (5), Análise de Sentimentos (4), Análise de Vídeos (4), Desenvolvimento de Plataformas MOOC (3), Avaliação por Pares (2), Gerenciamento de Personalização (2), Análise de Pré-Requisito (2), Sistemas de Gerenciamento de Dados (2), Análise de Trajetórias de Aprendizagem (2), Análise de Currículo (1), Análise de Design (1), *Digital Learning Ecosystem* (1), Análise do Engajamento (1), Geração de Usuários Artificiais (1), Análise de *Feedback* (1), *Gamificação* (1), Análise da Motivação (1), *Mind Wandering* (1), Análise de Dados de Recursos Educacionais (1).

No que diz respeito à categoria de Predição, quando o intuito é prever o Desempenho procura-se identificar com antecedência como será a performance do aluno no decorrer do curso, para poder intervir caso necessário e assim melhorar seu processo de aprendizagem. Por exemplo, Waheed *et al.* (2020) implementaram uma Rede Neural Artificial Profunda, a qual foi aplicada sobre um conjunto de dados extraídos do fluxo de cliques de uma plataforma MOOC, para prever o desempenho de estudantes e assim poder auxiliar aqueles que estivessem em risco. Artigos categorizados como Predição de Abandono, objetivam identificar alunos que pretendem desistir antes do encerramento do curso. Como no trabalho apresentado por Wen *et al.* (2020), em que os autores utilizaram uma Rede Neural Convolucional, que superou o desempenho de outros métodos mais tradicionais na previsão do abandono; nas conclusões, Wen *et al.* (2020) salientaram que em posse dessas previsões tutores, professores e gestores podem realizar ações que diminuam os índices de desistência em MOOCs. Sobre Predição de Conclusão, Pigeau, Aubert e Prié (2019) apresentaram um estudo de caso sobre um conjunto de dados fornecido pela plataforma francesa *OpenClassrooms*, que foi modelado de 8 formas diferentes, usando algoritmos de classificação e abordagens baseadas em sequência, como mineração de padrões de processo.

Análise de Comportamento é a segunda categoria mais numerosa em quantidade de publicações. O trabalho selecionado para apresentar como modelo deste tipo de pesquisa, é o de Lan *et al.* (2017) no qual foi proposto um modelo de aprendizado que relaciona o comportamento ao assistir vídeos e o envolvimento com o desempenho em atividades avaliativas. A maioria dos trabalhos com essa tendência tem foco no melhoramento da experiência educacional dos MOOCs.

A terceira categoria mais numerosa é a Mineração de Texto, que engloba Análise em Fóruns de Discussão, um tema muito recorrente. A Análise de Fóruns de Discussão possui vários propósitos, dentre os quais: detecção de erros dos alunos, relevância temática, engajamento, postagens que necessitam da atenção dos professores. Um exemplo é o trabalho desenvolvido por Guo *et al.* (2019), que apresentaram uma nova Rede Neural Híbrida para identificar postagens “urgentes” que requerem atenção imediata dos instrutores em fóruns de discussão. Geralmente, essas duas vertentes costumam possuir características interligadas, pois muitos estudos em fóruns são elaborados, por meio da mineração de texto. Além disso, foram encontrados artigos que tratam da mineração de texto em e-mails e redes sociais – como em Joksimović *et al.* (2015) – no qual realizaram análises de tópicos de discurso em redes sociais para descobrir o que alunos dos MOOCs postam, a respeito do curso.

O tópico Sistemas de Recomendação também é frequente e diz respeito à implementação de sistemas que fazem alguns tipos de sugestões para usuários de MOOCs, como por exemplo: recomendar contatos que possuam características semelhantes; recomendar conteúdos; e cursos dentro das plataformas. Em Labarthe *et al.* (2016) os autores tentaram aprimorar a experiência do MOOC com um sistema de recomendação que fornece a cada aluno uma lista individual de contatos com alto potencial de congruência.

No que se refere ao principal construto desta Tese – Identificação de Trapaças em MOOCs – 5 estudos, de um total de 6, se referem ao mesmo tipo de comportamento trapaceiro, o CAMEO (*Copying Answers using Multiple Existences Online*). Percebeu-se que há um grande esforço em estudar esse tipo específico de irregularidade, que vem acontecendo em muitos cursos MOOC e está preocupando Instituições renomadas que os ofertam, como o MIT, Harvard e Stanford. O CAMEO é uma estratégia que envolve um usuário que reúne soluções para perguntas de avaliação usando uma conta de “colheita – *harvester*” e envia respostas corretas usando uma conta “mestre – *master*” separada. Nesse sentido Ruipérez-Valiente *et al*. (2016) implementaram um método para identificação de trapaças deste tipo, com um algoritmo que utiliza MDE para rastrear os IP dos usuários e o tempo entre postagens de atividades.

Por fim, a Análise de Redes Sociais em MOOCs corresponde a uma linha de pesquisa que busca compreensão do papel social do MOOC como um espaço de

interação para aprendizagens dos alunos, como no estudo desenvolvido por Silva, Carvalho e Teixeira (2018). Neste os autores utilizam várias técnicas de MDE para verificar se as redes originadas em fóruns de discussão em MOOCs, podem ser consideradas redes sociais de aprendizagem realmente significativas.

QP2 – Quais as Técnicas e Algoritmos mais utilizados?

Diferentes dispositivos têm sido empregados em pesquisas sobre MDE, e eles baseiam-se na aplicação de rotinas, procedimentos e/ou algoritmos que podem detectar informações relevantes em meio a muitos dados, principalmente vinculados a técnica de Estatística Descritiva, Aprendizagem de Máquina e mais recentemente Aprendizagem Profunda. Entretanto, foi possível perceber que a Aprendizagem de Máquina é a técnica mais utilizada na MDE. Nas pesquisas sobre Análise de Comportamento e Sistemas de Recomendação, foram citados principalmente os seguintes algoritmos: K-Vizinhos mais Próximos (*K-Nearest Neighbor* – KNN), Agrupamentos, Máquinas de Vetores de Suporte (MVS), Naïve Bayes, Floresta Aleatória (*Randon Forest*), Árvores de Decisão*,* e Apriori, bem como modelos gerados a partir da Alocação Latente de Dirichlet e Regressão Logística, todos algoritmos de Aprendizagem de Máquina. Quando a temática era Predição (Desempenho, Abandono e Conclusão), Sistemas de Recomendação e Mineração de Textos, notou-se também o emprego da Aprendizagem Profunda, com alguns autores utilizando Redes Neurais Convolucionais e Redes Neurais Recorrentes. Constatou-se também uma predominância de algoritmos altamente especializados em pesquisas sobre Mineração de Texto.

Como exemplo das aplicações de algumas das técnicas destacadas, cita-se Balint (2016), que analisou o comportamento de resolução de problemas de física de alunos matriculados em um MOOC, por meio do algoritmo de agrupamento *K-means*. Além dele, Al-Shabandar *et al.* (2017) desenvolveram uma pesquisa com o objetivo de prever o desempenho em MOOCs, e utilizaram vários algoritmos de Aprendizagem de Máquina: Regressão Logística, Alocação Latente de Dirichlet, Naïve Bayes, MVS, Árvore de Decisão, Floresta Aleatória, Redes Neurais Simples e Mapa Auto-Organizado (SOM). Cita-se também Kashyap e Nayak (2018), que utilizaram os algoritmos Naïve Bayes, Floresta Aleatória, Árvores de Decisão e MVS, para prever o abandono em MOOCs. O Quadro 2 lista os 10 algoritmos mais empregados, considerando quantidade de temáticas abrangidas e sua funcionalidade.

Quadro 2 – Algoritmos de mineração de dados mais utilizados.

Algoritmo	No de Temáticas Abrangidas	Funcionalidade	Exemplo de Aplicação
Agrupamentos (*K-means* e Hierárquico)	**13**	Utilizam diferentes técnicas matemáticas, para colocar em um mesmo grupo instâncias que apresentem atributos semelhantes.	Identificação de perfis de aprendizagem em MOOCs - (RODRIGUES *et al.*, 2016).
Regressão Logística	**8**	Constrói, com base em um conjunto de análises, um modelo que possibilite a previsão de uma saída categórica para um conjunto de atributos. É um algoritmo usado para modelar a probabilidade de uma determinada classe ou evento existir.	Previsão de Desempenho em MOOCs - (AL-SHABANDAR *et al.*, 2017). Detecção de trapaças do tipo CAMEO em MOOCs - (RUIPEREZ-VALIENTE *et al.*, 2017).
Máquina de Vetores de Suporte (MVS)	**7**	Constrói um hiperplano ou conjunto de hiperplanos em um espaço de dimensão alta ou infinita, que pode ser usado para classificação, regressão ou outras tarefas, como detecção de outliers.	Previsão de abandono em MOOCs - (KASHYAP; NAYAK, 2018).
Árvore de Decisão	**5**	A árvore de decisão usa a representação em árvore para resolver um problema de classificação ou regressão, em que cada nó folha corresponde a um rótulo de classe e os atributos são representados nos nós internos.	Previsão desempenho em MOOCs - (XIAO; LIANG; MA, 2018).
KNN	**5**	É um algoritmo não paramétrico para classificação e regressão, que utiliza o cálculo da distância entre a entrada e os k exemplos de treinamento mais próximos, para atribuir uma classificação a uma instância.	Analisar as preferências dos alunos no ambiente instrucional dos MOOCs - (WANG *et al.*, 2018).
Redes Neurais	**5**	Realiza o aprendizado de máquina bem como o reconhecimento de padrões, atingindo uma solução generalizada para uma classe de problemas.	Previsão da conclusão em MOOCs - (PIGEAU; AUBERT; PRIÉ, 2019). Previsão de abandono em MOOCs - (WAHEED *et al.*, 2020).
Alocação Latente de Dirichlet	**5**	Descreve um conjunto de observações como uma mistura de categorias distintas. É mais comumente utilizado para descobrir um número de tópicos especificado pelo usuário.	Mineração de texto para entender melhor o discurso on-line no contexto dos MOOCs - (ATAPATTU; FALKNER; TARMAZDI, 2016).
Naïve Bayes	**4**	Utiliza dados históricos para prever a classificação de um novo dado.	Previsão de abandono em MOOCs - (COBOS; OLMOS, 2018).
Floresta Aleatória	**4**	Cria várias árvores de decisão, de maneira aleatória, formando como uma floresta, onde cada árvore será utilizada na escolha do resultado.	Identificação de trapaças do tipo CAMEO em MOOCs - (RUIPEREZ-VALIENTE *et al.*, 2017).

Fonte: Souza e Perry (2020)

Sobre as ferramentas mais utilizadas para a realização da MDE entre os estudos investigadas, foram encontrados trabalhos que utilizaram Python e R, como

Sunar *et al.*, (2020) e Cobos e Olmos (2018). Na pesquisa de Sunar *et al.* (2020) essas linguagens foram empregadas para tipificar os diferentes padrões de comportamento social dos participantes durante um MOOC, e testaram estatisticamente se existe uma correlação entre a conclusão do curso e os comportamentos modelados, com a finalidade de entender melhor o engajamento social dos alunos em uma plataforma MOOC e o impacto do engajamento na conclusão do curso. Cobos e Olmos (2018) implementaram um complexo modelo de predição chamado *EDX-MAS* que possui dois módulos: 1) Módulo de Importação – permite extrair, limpar, selecionar e pré-processar os dados do curso para detecção de abandono e selecionar atividades relevantes na coleta de dados, além disso, suporta também a criação de variáveis de entrada e gerenciamento de armazenamento de dados, codificado na linguagem Python; 2) Módulo de geração do modelo – este módulo suporta a geração de modelos preditivos do curso selecionados por dia ou por semana, codificado em R, e utiliza 10 algoritmos de Aprendizagem de Máquina – Regressão Logística, Floresta Aleatória, Reforço Estocástico de Gradiente, Naïve Bayes, Gradiente Extremo, Rede Neural Artificial, MVS, Modelo Linear Generalizado Bayesiano, KNN, Árvores de Decisão para Classificação e Regressão.

Ademais, duas ferramentas que já trazem os algoritmos implementados e são amplamente utilizadas em mineração de dados foram constatadas como as mais utilizadas para MDE, *Weka* e *RapidMiner*. O *Weka* é um software livre para MD, implementado na linguagem de programação Java, que se estabeleceu amplamente como a ferramenta de MD mais utilizada, por acadêmicos e docentes de diversas universidades, especialmente devido à grande facilidade na sua manipulação e aplicação, seu objetivo é agregar algoritmos provenientes de diferentes abordagens/paradigmas de Inteligência Artificial dedicados especialmente à AM. A pesquisa realizada por Brooks, Thompson e Teasley (2015) é um exemplo de sua aplicação. Nela os autores descreveram uma modelagem de perfis de alunos com base nos dados coletados dos ambientes de aprendizagem, onde implementaram modelos preditivos com Arvore de Decisão usando a ferramenta *Weka*, com a intenção de prever o desempenho. Em relação ao *RapidMiner*, é um sistema comercial também para análise de dados que utiliza algoritmos de AM, muito semelhante ao *Weka*. No trabalho desenvolvido por An, Krauss e Merceron, (2017) os autores o utilizaram na aplicação de algoritmos de Agrupamento com o intuito de analisar perfis comportamentais de alunos em MOOCs. Salienta-se que a maioria dos

estudos com objetivo analisar os comportamentos dos alunos, aplicam algoritmos de agrupamento para realizar essa tarefa.

QP3 – Quais as principais oportunidades de pesquisa identificadas?

Dentre as oportunidades identificadas no decorrer do mapeamento, avalia-se que algumas são mais proeminentes, pois ainda existem poucos trabalhos publicados e há muito espaço para evolução. Por exemplo, no âmbito da Análise de Redes Sociais, Brinton *et al.* (2018) usaram o conceito de Aprendizagem Social como plano de fundo para analisar as conexões realizadas nos MOOCs e assim avaliar a eficiência dos fóruns de discussão. Brinton *et al.* (2018) asseguram que a proliferação dos MOOCs apresentou uma infinidade de possibilidades para pesquisas em torno aprendizagem em redes sociais.

De mesmo modo, destaca-se a aplicação dos conceitos de Ecossistema de Aprendizagem Digital, que consiste em espécies, populações e comunidades interagindo entre si e com o ambiente. Segundo Morales *et al* (2019) esse modelo pode ajudar educadores e designers instrucionais a reunir informações baseadas em dados dos alunos, que permitam projetar estratégias e atividades inovadoras usando ferramentas diversas, que maximizam a aprendizagem em ambientes de ensino on-line.

Salienta-se também a Análise de *Mind Wandering* (MW), expressão que foi traduzida livremente como “devaneio”. Esta conceituação compreende a ideia de momentos em que a atenção e o conteúdo dos pensamentos se desviam do sentido original, ou da ação que está sendo realizada no instante em que esses pensamentos acontecem (HUTT *et al.*, 2017). Essa temática pode ser abordada de diferentes formas, como em Hutt *et al.* (2017), que investigaram o uso de rastreamento ocular no nível de usuário (alunos) para detectar automaticamente esse devaneio, enquanto assistiam a uma videoaula do MOOC. Os resultados apresentados pelos autores mostraram que essa detecção é exequível no contexto de assistir a uma videoaula, sendo possível alcançar precisão de 47%. Como pode ser observado, neste Mapeamento, ainda há poucas iniciativas como essa no contexto de MOOCs, dessa forma trabalhos nesse sentido podem ter muito a contribuir com a área.

Por último, evidencia-se a Identificação de Trapaças, uma tendência de pesquisa que ainda possui poucos trabalhas publicados sobre MOOCs, mas que detém muito interesse de grandes instituições de ensino, pelo fato de que não se deve

atribuir certificações a alunos que não possuem conhecimentos/habilidades para tal (RUIPEREZ-VALIENTE *et al.*, 2017a; RUIPÉREZ-VALIENTE *et al.*, 2017a). Neste sentido, autores como Alexandron *et al. (*2017), Ruiperez-Valiente *et al.* (2017) e Northcutt, Ho e Chuang (2016) empenharam esforços para implementar soluções computacionais que pudessem identificar automaticamente alunos matriculados em MOOCs de grandes instituições de ensino, que estivessem praticando irregularidades para conseguir certificados.

QP4 – Quais os desafios mais relatados nesses trabalhos em pesquisas na área?

Em referência aos principais desafios de pesquisas na área, pode-se apontar como um dos mais recorrentes o desequilíbrio de classe nas bases de dados analisadas, tendo sido citado em pelo menos 20 publicações. Uma das causas desse desbalanço é a baixa percentagem de atividades concluídas – desta forma, em abordagens supervisionadas os algoritmos acabam reconhecendo a solução de forma tendenciosa, pois existem mais instâncias da base de dados rotuladas como desistentes do que como concluintes. Em Xing *et al.* (2016) esse desafio é citado e eles implementaram uma solução para conseguir bons resultados, aplicando o método de empilhamento de algoritmos.

Outro desafio é a diversidade dos dados coletados em MOOCs, que é verificada em especial na predição de conclusão, mas aparece em várias publicações de outras categorias. É um problema que torna difícil a utilização de métodos estatísticos ou de agrupamento simples para criar um modelo preditivo. Korosi *et al.* (2018) apresentaram uma abordagem de MDE para analisar os dados de fluxo de cliques dos alunos para prever a conclusão em MOOCs, e mencionaram essa como a maior dificuldade encontrada para realização do processo, sendo necessária a aplicação de mais de 10 algoritmos para chegar na precisão desejada.

Finalmente, apresenta-se como um obstáculo para os pesquisadores a confiabilidade do fluxo de cliques dos alunos, pois muitos estudantes clicam em uma tarefa, mas não estão engajados em sua resolução. Deve-se, portanto, relacionar o clique do aluno com a quantidade de tempo que este permaneceu em cada tarefa, para validar seu engajamento, e desse modo conseguir informações mais confiáveis, o que requer uma etapa extra de processamento dos dados. He *et al.* (2018), por exemplo, desenvolveram uma investigação em dados de alunos de MOOCs para

extrair padrões de ritmos de aprendizagem e perceberam que ao utilizar apenas os dados de cliques dos alunos não estavam conseguindo os resultados esperados, com a confiabilidade desejada. Em suma, o número de cliques pode ser confiável, sob condição de estar associado a outros atributos coletados nos relatórios de dados dos alunos.

4.3 CONSIDERAÇÕES SOBRE O MAPEAMENTO

No início das pesquisas para o desenvolvimento dessa Tese, não havia a definição clara de seu objetivo, todavia já existia um interesse na investigação das trajetórias dos alunos, impressas nos relatórios de dados no decorrer da realização de um MOOC na plataforma Lúmina, por isso o mapeamento elaborado tinha como foco a Mineração de Dados Educacionais nesses tipos de cursos. A partir da sistematização dos resultados deste mapeamento foram identificadas quatro oportunidades de pesquisa, das quais a temática Identificação de Trapaças foi considerada promissora para o desenvolvimento desta Tese. Em primeiro Lugar, porque até momento foi pouco explorada; e em segundo lugar devido ao contexto de desenvolvimento desta pesquisa – os MOOCs do Lúmina – em que comportamentos inadequados dos alunos foram percebidos pelos gestores da plataforma. Portanto, pensando na aplicabilidade da solução que foi desenvolvida no decorrer deste estudo, a pesquisadora antepôs por esse tema, que pode ser abordado de várias formas, contudo optou-se pela utilização de técnicas exploratórias de MDE, mais especificamente algoritmos de agrupamento, para o reconhecimento desses comportamentos. Além de utilizar técnicas de aprendizagem de máquina supervisionada, para apoiar na identificação de características dos MOOCs que impactam no comportamento dos alunos.

Com relação aos estudos sobre Identificação de Trapaças em MOOCs foram encontrados seis artigos sobre essa temática, cinco tratam do mesmo tipo de comportamento trapaceiro, o CAMEO (*Copying Answers using Multiple Existence Online*), diferenciando-se apenas no que diz respeito a metodologia de identificação, são eles: Northcutt, Ho e Chuang, (2016); Ruiperez-Valiente *et al.* (2016); Ruiperez-Valiente *et al.* (2017a); Alexandron *et al*. (2017); Bao, Chen e Hauff (2017); e um desses estudos, desenvolvido por Ruipérez-Valiente *et al*. (2017b), aborda um método mais amplo para reconhecimento de comportamentos em MOOCs, mas também pode ser aplicado na detecção do CAMEO e outros tipos de condutas inadequadas.

A denominação CAMEO foi definida por Northcutt, Ho e Chuang, (2016) e se refere a um evento em que um usuário usa uma ou mais contas em uma plataforma MOOC, com o intuito de encontrar a resposta correta para uma pergunta, e então enviar esta resposta em sua conta oficial, pela qual o aluno pretende obter o certificado. Os autores que tratam desse tema se referem as contas utilizadas para encontrar a solução como contas de *"colheita"* e a conta principal como conta *"mestre",* encontrar a resposta na conta de colheita pode ser feito pedindo para ver a resposta correta depois de usar todas as tentativas – função "mostrar resposta" na plataforma *edX*, por exemplo – ou por pesquisa exaustiva, por exemplo adivinhação, para perguntas de múltipla escolha, até que a resposta correta seja encontrada. De acordo com Ruiperez-Valiente *et al.* (2016) alguns dos critérios para identificar CAMEO são:

1. A conta de colheita e a mestre pertencem ao mesmo grupo de IP.
2. A conta de colheita não tem motivação extrínseca – isso é operacionalizado porque ela não recebe um certificado, ganhar um certificado indica que o usuário busca uma recompensa pelo tempo investido nesta conta, reduzindo a probabilidade de que esta seja uma conta falsa.
3. As perguntas coletadas são realmente usadas na conta principal – a lógica subjacente a este critério é que se um aluno estabelece uma conta para colher as respostas, então a maioria das perguntas respondidas por esta conta será usada por uma conta mestre.
4. A conta mestre não é usada para realização de colheitas – o fundamento é semelhante ao do critério 2, na direção contrária.

Esse tipo de trapaça se justifica em plataformas como: *edX*, *MITx*, ou *Udacity* – vinculadas as Universidades de Harvard, MIT e a Stanford – sobretudo pela complexidade dos cursos MOOCs ofertados, em que muitos deles são utilizados até mesmo para convalidação de créditos nos cursos de graduação e pós-graduação ofertados presencialmente. Segundo o exemplo descrito por Webley (2012), alunos podem realizar MOOCs, por meio do *edX*, e receber certificados que são válidos como créditos na Pós-Graduação, pois diversas instituições de ensino concordaram em aceitar essas certificações como uma forma de desenvolvimento profissional, dessa forma alunos podem transformar cursos on-line gratuitos em créditos universitários tangíveis. Ruiperez-Valiente *et al.* (2016) ressaltam outro uso de certificados MOOC, como nas admissões acadêmicas, por exemplo a *Wharton Business School* anunciou que utilizaria seus MOOCs como uma ferramenta adicional para selecionar

candidatos. Isso pode explicar, pelo menos parcialmente, porque alguns alunos procuram maneiras mais fáceis e rápidas de obter certificados em MOOCs.

Esse tipo de comportamento tem causado preocupações em gestores destas plataformas, pois compromete a legitimidade das certificações emitidas, e sobretudo pela proporção de alunos que se utilizaram dessa estratégia, que de acordo com Northcutt, Ho e Chuang (2016) foi de 25% entre estudantes que completaram 20 cursos ou mais na plataforma *edX* no ano de 2016, por exemplo. Por isso, pesquisadores do contexto de MOOCs viram a necessidade de estudar mais a respeito de comportamentos indesejados nesse contexto, todavia até a época da realização do mapeamento sistemático descrito não havia muitos trabalhos sobre essa temática. De acordo com alguns autores (NORTHCUTT; HO; CHUANG, 2016; RUIPEREZ-VALIENTE *et al.*, 2016; WEBLEY, 2012) essa carência de estudos acontece, porque é difícil explorar a questão da trapaça em MOOCs, devido à falta geral de dados verdadeiros, já que os mantenedores de MOOCS relutam em confrontar os alunos – como é difícil encontrar uma prova definitiva da trapaça – e os alunos relutam em admitir seu mau comportamento.

Em contrapartida, a pequena quantidade de trabalhos sobre trapaças em MOOCs, há um amplo corpo de pesquisa sobre trapaças em Tutores Cognitivos, mais acessíveis de serem observados, pois são geralmente utilizados em salas de aula tradicionais, uma vertente denominada *"gaming the system"* (enganar o sistema). Para Baker (2011) depois do surgimento do construto de enganar o sistema, o interesse em pesquisas nesse sentido surgiu em uma escala maior, levando a dezenas, se não centenas, de artigos sobre comportamentos trapaceiros e construções estreitamente relacionadas.

Estes estudos tem grande influência no contexto de trapaças em sistemas computacionais de ensino e aprendizagem, autores como Alexandron *et al.* (2017) e Ruiperez-Valiente *et al*. (2016), apontaram em seus estudos uma relação entre o "*gaming the system"* e o CAMEO, salientando que este último pode ser considerado semelhante, ou um novo tipo de estratégia para enganar o sistema*,* caracterizado por alunos tentando obter sucesso em softwares educacionais, explorando propriedades do sistema, ao invés de aprender com o material – ambos em termos de motivação (melhorar as notas); e em termos do método (explorando recursos técnicos do sistema). Assim, a coleta de respostas CAMEO pode ser pensada como um novo caso

específico de enganar o sistema, pois suplanta os objetivos educacionais pretendidos e resultados de aprendizagem com o sistema.

Devido a relação entre as estratégias de enganar o sistema, o CAMEO e de forma geral a identificação de comportamentos inadequados em sistemas computacionais de ensino e aprendizagem, todos os estudos encontrados sobre o CAMEO no mapeamento sistemático, bem como alguns estudos relevantes sobre "*gaming the system*" são aprofundados no Capítulo seguinte, que trata dos trabalhos relacionados. Além disso, neste capítulo é caracterizado o contexto dos comportamentos indesejados no Lúmina.

5 ENGANAR O SISTEMA

Trapaças são ações destinadas a subverter as regras a fim de obter vantagens indevidas, e esse é um comportamento muito comum em ambientes escolares. Copiar dos colegas durante uma avaliação, por exemplo, é uma prática comum nos mais diversos níveis de ensino, mas nesse caso o aluno corre o risco de o professor perceber e punir os envolvidos. Todavia, em ambientes de ensino e aprendizagem on-line, esses comportamentos são mais difíceis de identificar, e se caracterizam como uma ameaça maior ainda, pois o professor não está visualizando a realização das atividades avaliativas.

Um dos problemas colaterais ocasionados pelas trapaças em ambientes informatizados de ensino e aprendizagem é que elas resultam na diminuição da confiabilidade da avaliação, reduzem a confiança de que um certificado é uma evidência válida de proficiência e, portanto, representam um desafio ao sistema de certificação. Portanto, é importante estudar como diminuir a prevalência de comportamentos inadequados e entender suas motivações.

A partir desses aspectos, considera-se importante sistematizar algumas pesquisas que tratam sobre comportamentos inadequados em ambientes virtuais de aprendizagem. Primeiramente abordando estudos sobre "*gaming the system*" em tutores cognitivos, precursores deste tipo de pesquisa em ambientes informatizados. Posteriormente são detalhados os estudos que têm como contexto os MOOCs, englobando o cenário desta Tese: caçadores de certificados no Lúmina.

5.1 ENGANANDO O SISTEMA EM TUTORES COGNITIVOS

Antes da popularização dos MOOC, nos formatos atuais, foram realizados muitos estudos com os chamados tutores inteligentes, ou tutores cognitivos, um tipo particular de sistema de tutoria que utiliza um modelo cognitivo para fornecer aos alunos feedback sobre a correção ou incorreção de suas respostas enquanto eles estão resolvendo atividades propostas. Os tutores cognitivos também têm a capacidade de fornecer dicas e instruções sensíveis ao contexto, para guiar os alunos em direção às próximas etapas das atividades propostas.

Segundo Baker, Corbett e Koedinger (2004) esse tipo de tutor era uma das abordagens mais bem-sucedidas e amplamente utilizadas para incorporar programas assistidos por computador e instrução em sala de aula. Em meados dos anos 2000 a utilização do tutor cognitivo combinado à instrução conceitual, ministrada por um professor, era uma tecnologia avançada, na qual cada aluno trabalhava individualmente com um sistema de tutoria e escolhia os exercícios e feedback com base em um modelo de execução de habilidades que ele próprio possuía (BAKER *et al.*, 2004). De acordo com Baker *et al.* (2004) em 2004 cerca 5% das escolas de ensino médio nos Estados Unidos usavam essa tecnologia. No entanto, embora o tutor cognitivo fosse usado para aumentar o envolvimento e esforço dos alunos em sala de aula, alguns desses alunos ao invés de aproveitar o software para melhoria da aprendizagem e desenvolvimento de habilidades, se utilizavam de estratégias orientadas para manipular a ferramenta.

As primeiras menções a esse tipo de comportamento em tutores cognitivos ocorreram entre 1990 e 2000, relatados pelos autores: Schofield (1995); Miller, Lehman e Koedinger (1999); Aleven e Koedinger (2000), Aleven e Koedinger (2002); e em 2004 dois artigos utilizaram pela primeira vez o termo *"gaming the system"* (enganando o sistema): Baker *et al.* (2004) e Baker, Corbett e Koedinger (2004). Essas publicações estabeleceram o construto de enganar o sistema, depois disso o interesse em pesquisas sobe este assunto aumentou (BAKER, 2011). Esse conjunto de estratégias foi definido por Baker, Corbett e Koedinger (2004, p. 532) como:

> Um comportamento destinado a obter respostas corretas e avançar dentro do currículo de tutoria, aproveitando as vantagens e regularidades nas respostas do software, sistematicamente a partir do uso indevido do feedback ou da ajuda do software, em vez de raciocinar ativamente sobre o material (BAKER; CORBETT; KOEDINGER, 2004, p. 532).

Um dos principais pesquisadores sobre este tema é Ryan Baker, que desenvolveu sua Tese (*"Designing Intelligent Tutors That Adapt to When Students Game the System"*) em 2005 com o objetivo de entender, detectar automaticamente e redesenhar um sistema de tutoria inteligente para se adaptar ao comportamento de enganar o sistema, termo definido e caracterizado por ele e seus orientadores em 2004 (BAKER *et al.*, 2004 e BAKER; CORBETT; KOEDINGER, 2004). Em sua tese Baker (2005) apresenta um conjunto de estudos voltados para a compreensão dos efeitos de enganar o sistema na aprendizagem, e no decorrer desses estudos

determina que este tipo de manipulação do sistema está associado ao baixo aprendizado.

Nos dados apresentados por Baker (2005) para mapear o perfil dos alunos com esse tipo de comportamento, ele mostra que esses estudantes têm um padrão consistente de afeto negativo em relação a muitos aspectos de sua experiência em sala de aula e nos estudos. No restante da tese ele destaca o desenvolvimento de um método, por meio de Aprendizagem de Máquina (AM), para detecção do comportamento de enganar o sistema, e por fim propõe a incorporação de um agente animado (“Scooter, o Tutor”) que tinha como intuito indicar ao aluno e ao professor se o aluno havia se comportado de forma inadequada. Scooter também tinha a função de oferecer exercícios complementares aos alunos, a fim de dar a eles uma segunda chance de aprender os conteúdos que haviam burlado. Baker (2005) evidenciou que Scooter foi capaz de reduzir a frequência desses comportamentos pela metade, e os exercícios suplementares estavam associados a um aprendizado substancialmente melhor.

Posteriormente, Baker continuou desenvolvendo inúmeras pesquisas relacionadas com a identificação do comportamento de enganar o sistema, como em Baker *et al.* (2008), neste estudo os autores propuseram o desenvolvimento de um algoritmo para identificação deste tipo de comportamento, aplicado ao tutor cognitivo. Os autores definiram neste estudo o comportamento de enganar o sistema, basicamente, da seguinte forma: “1) pedir ajuda rápida e repetidamente até que o tutor dê ao aluno a resposta correta; e 2) Inserir as respostas de forma rápida e sistemática” (BAKER *et al*., 2008, p. 289). Além disso, os autores destacaram haver dois tipos de comportamentos distintos entre os alunos que praticaram o *“game the system”*: *“gamed-hurt”* e *“gamed-not-hurt”*, denotando a existência do comportamento de enganar o sistema com efeitos prejudiciais e com efeitos não prejudiciais aos alunos, correlacionados as notas nos pré e pós testes realizados e nos ganhos de habilidades. Os alunos que enganaram o sistema, mas melhoraram suas habilidades, ou tiveram ganhos de aprendizagem eram caracterizados como *“gamed-not-hurt”*; enquanto àqueles que tiveram pontuações baixas tanto no pré-teste quanto no pós-teste foram rotulados como *“gamed-hurt”*, e estes últimos eram o foco do algoritmo desenvolvido.

Para isso foram utilizados 3 conjuntos de dados dos alunos: 1) observações quantitativas de campo, a fim de estimar a porcentagem de tempo que cada aluno enganou o sistema; 2) pré-testes e pós-testes para cada lição, para determinar o

quanto cada aluno aprendeu ao usar o tutor; 3) arquivos de registro detalhados das interações dos alunos com o tutor, os logs. O conjunto de dados de logs dos alunos era composto por 26 recursos que descreviam uma ação e consistiam basicamente em dados relacionados à: detalhes sobre a ação, avaliação de conhecimento, tempo e interações anteriores. Neste estudo os autores utilizaram um banco de dados com cerca de 129 mil ações.

Para o desenvolvimento do algoritmo os autores utilizaram o modelo de resposta latente (*latent response model)* como a base estatística, este modelo relaciona um conjunto de variáveis observáveis (manifestas), a um conjunto de variáveis latentes e têm a vantagem de integrar facilmente várias fontes de dados, em um único modelo. De forma geral, o algoritmo proposto para identificação de *"gamed-hurt"* foi construído baseado na identificação de 7 indícios, em que quatro são considerados *"gamed-hurt" e 3 "gamed-not-hurt".* Os indícios considerados *"gamed-hurt"* são: 1) cometer muitos erros em uma etapa específica do problema, entre os problemas; 2) acertar na primeira tentativa em alguns problemas e cometer muitos erros em outros problemas similares; 3) solicitar ajuda em várias etapas em sucessão curta, uma vez que o aluno atingiu uma alta probabilidade de saber pelo menos algumas etapas; 4) realizar ações muito rápidas, mas apenas se o aluno já cometeu um erro na etapa atual. Os indícios considerados *"gamed-not-hurt"* são: 1) cometer erros em etapas complexas, como durante a plotagem de pontos, uma atividade em que falhas são comuns, apesar dos alunos conhecerem bem a habilidade; 2) cometer erros extremamente rápidos e 3) pedidos de ajuda extremamente rápidos (ocorrendo em menos de um quinto de segundo), em muitos casos, as ações nessa velocidade consistem em ações rápidas idênticas, como cliques duplos acidentais na ajuda, ou apertar *enter* duas vezes.

Como resultados, Baker *et al.* (2008) destacaram que o algoritmo foi preciso nas suas classificações e a principal métrica utilizada pelos autores foi área sob a curva *Receiver Operating Characteristic* (ROC)[8], em que esta atingiu uma média de 86% nos dados de treinamento e 80% nos dados de teste. Isso significa que o modelo pode distinguir um aluno que tem o comportamento de enganar o sistema, de um aluno que não tem, 80% das vezes. Além disso, Baker *et al.* (2008) ressaltaram que

[8] A área sob a curva ROC mostra o quão bom o modelo criado pode distinguir entre duas categorias, essa métrica corresponde a área sob a curva ROC na teoria de detecção de sinal, ou também ao W, na estatística de Wilcoxon.

a utilização do algoritmo proposto pode auxiliar na condução de intervenções que melhoram a aprendizagem, focando naqueles alunos que tem comportamentos inadequados, bem como expande o conhecimento sobre a construção comportamental de “enganar o sistema”.

Em outro estudo desenvolvido em 2008, Baker e De Carvalho (2008) apresentaram um método mais rápido para classificação dos dados extraídos do tutor cognitivo, e também propuseram a aplicação de algoritmos de AM para geração de modelos para classificação de alunos com o comportamento de enganar o sistema. Visto que, este artigo foi publicado na primeira conferência sobre MDE, havia um interesse em propor um modelo de identificação reutilizável e que fosse simples de validar, por isso os autores utilizaram a ferramenta Weka para o seu desenvolvimento, um software bastante intuitivo e que foi massivamente empregado na área de MDE, além de aplicarem o modelo de resposta latente, que já havia sido utilizado no estudo anterior.

Neste artigo os autores apresentaram um método para rotular diretamente os arquivos de log dos alunos, extraídos dos tutores cognitivos: “*text replays*”. Essas repetições de texto representam um segmento do comportamento do aluno a partir dos arquivos de log em um formato textual. Uma sequência de ações de uma duração pré-selecionada (em termos de tempo) é mostrada em um formato textual que fornece informações sobre as ações e seu contexto. Desta forma, o especialista no contexto (professor, tutor, ou outro profissional que tenha domínio sobre a utilização de tutores cognitivos) observa a duração de cada ação, o contexto do problema, a entrada inserida, a habilidade relevante e como o sistema avaliou a ação (correta/incorreta, um pedido de ajuda, ou um erro). Então o especialista pode escolher uma, de um conjunto de categorias de comportamento (neste estudo, enganar ou não enganar o sistema), ou indicar que algo deu errado, tornando a ação especificada não categorizável. Este método foi proposto pela primeira vez em Baker, Corbett, e Wagner (2006, apud BAKER; DE CARVALHO, 2008), desde então foi amplamente empregado para classificação dos logs dos alunos em tutores cognitivos.

Os dados utilizados neste estudo são resultados das interações de 59 alunos durante um ano letivo completo, um total de 32 aulas com o Tutor Cognitivo de Álgebra, e estavam armazenados no repositório *DataShop*. Deste conjunto de dados 18 mil *text replays* foram categorizados seguindo o método descrito anteriormente. Para o desenvolvimento do modelo os autores utilizaram a ferramenta Weka com a

aplicação do algoritmo de Árvores de Decisão (J48), e o modelo de resposta latente. Os autores utilizaram novamente como principal métrica a área sob a curva ROC, em que o modelo de resposta latente obteve um valor de 96%, e o modelo baseado no algoritmo de Árvores de Decisão 69%. Todavia, a confiabilidade entre os modelos medido pelo coeficiente Kappa, foi bem maior para o modelo de Árvores de Decisão, com um valor de 4%, comparado com 0,4% do modelo de resposta latente.

Por fim, Baker e De Carvalho (2008) relatam que o método para categorização dos dados utilizado neste estudo agiliza o processo das pesquisas comportamentais em tutores cognitivos. Segundo eles, quando todos os fatores são levados em consideração, incluindo o tempo necessário para fazer rótulos individuais e a logística envolvida na realização de estudos nas escolas, o método de "*text replays"* acelera o processo de coleta de dados em cerca de 40 vezes em comparação com as observações quantitativas de campo, e pelo menos 6 vezes em comparação com reproduções de tela. Os autores ainda destacam que este método de categorização facilita a utilização de algoritmos de AM, igualmente os disponibilizados em softwares como o Weka. De acordo com eles, essa possibilidade deve tornar mais viável para pesquisadores e desenvolvedores sem experiência com AM desenvolver classificadores, que são úteis para análises de aprendizagem e para conduzir intervenções.

Outro aspecto investigado por Baker e seus colegas, além do comportamento de enganar o sistema foi o comportamento de ficar fora da tarefa (*off-task*), em tutores cognitivos. De acordo com Cocea, Hershkovitz e Baker (2009) o comportamento de "ficar fora da tarefa" acontece quando o aluno se envolve em ações que não compreendem o sistema, ou a tarefa de aprendizagem. Neste estudo, os autores investigam o quanto os comportamentos de enganar o sistema e ficar fora da tarefa podem levar os alunos a terem suas aprendizagens reduzidas. Para isso os autores consideraram duas variáveis para o comportamento de enganar o sistema: abuso no pedido de ajuda, e errar sistematicamente, em ambos os casos o tutor acaba por dar as respostas; e no que se refere ao comportamento de ficar fora da tarefa os autores denotaram que este pode ocorrer de diversas formas, e citam como exemplo: conversar com outros alunos sobre tópicos não relacionados, navegar na *web* e interromper outros alunos.

Para nortear o estudo, Cocea, Hershkovitz e Baker (2009) levaram em consideração duas hipóteses sobre esses comportamentos e como estes conduzem

a uma aprendizagem reduzida: 1) menos aprendizagem em etapas individuais, impacto prejudicial imediato, e 2) perda geral de aprendizagem devido a menos oportunidades de prática, impacto prejudicial agregado. Para testar as hipóteses os autores utilizaram os dados de 4 aulas do Tutor Cognitivo do Ensino Médio, referentes a disciplina de matemática. A metodologia utilizada pelos autores foi inspirada no método de decomposição de aprendizagem de Beck (2006, apud COCEA; HERSHKOVITZ; BAKER, 2009), onde a aprendizagem ao longo do tempo é avaliada em termos de eventos que ocorrem no processo de aprendizagem do aluno.

Nesse sentido, os autores conduziram análises em 4 etapas: 1) comportamento fora da tarefa e seu impacto na aprendizagem imediata; 2) comportamento de enganar o sistema e seu impacto na aprendizagem imediata; 3) comportamento fora da tarefa e seu impacto na aprendizagem agregada; e 4) comportamento de enganar o sistema e seu impacto na aprendizagem agregada. Para avaliar se o comportamento fora da tarefa, ou de enganar o sistema foi associado a uma aprendizagem imediata mais pobre, ou autores aplicaram um modelo de Regressão Logística, no qual o desempenho, em uma determinada habilidade, em um determinado momento, é previsto com base no número de passos nessa habilidade, em que aluno se envolveu nestes comportamentos anteriormente. Para verificar se o comportamento fora da tarefa, ou de enganar o sistema estava associado a um aprendizado agregado reduzido, os autores utilizaram duas técnicas: em primeiro lugar, a correlação entre o comportamento geral – fora da tarefa, ou enganar o sistema (medido como a porcentagem de etapas “fora da tarefa”, ou “enganar o sistema” em todas as etapas) – e o número total de etapas (calculado para todas as aulas e cada aula individualmente); e em segundo lugar, a regressão linear foi aplicada para estudar os fatores que contribuem para o desempenho no pós-teste, tendo o pré-teste e o número total de etapas como variáveis independentes.

Em suma, os autores chegaram aos seguintes resultados: enganar o sistema está associado tanto à aprendizagem mais pobre imediata (fortemente) quanto à aprendizagem mais pobre agregada (fracamente); e o comportamento fora da tarefa, por outro lado, está associado a um aprendizado mais pobre em um nível agregado (fortemente) somente. Para os autores o impacto de enganar o sistema, em um nível mais imediato, parece ser devido à falta de aprendizado na etapa em que o comportamento ocorreu; em outras palavras, ao burlar os conteúdos, uma oportunidade de aprender é desperdiçada. O impacto agregado aparente dos dois

comportamentos, consideravelmente mais forte no caso do comportamento fora da tarefa, é cumulativo, o pior desempenho parece ocorrer devido a menos oportunidades de aprendizagem. Por fim, Cocea, Hershkovitz e Baker (2009) salientam que este estudo pode contribuir para um projeto de melhores intervenções para lidar com essas formas potencialmente prejudiciais de interagir com ambientes de aprendizagem virtuais.

Além de modelo para identificação de comportamentos de enganar o sistema para o tutor cognitivo, Baker também propôs o desenvolvimento de um modelo para identificar este comportamento no SQL-Tutor (BAKER; MITROVIĆ; MATHEWS, 2010). O SQL-Tutor é um tutor baseado em restrições que auxilia estudantes de nível universitário a adquirir o conhecimento e as habilidades necessárias para criar consultas SQL; consiste em um sistema de tutoria inteligente e maduro, desenhado como um ambiente de prática, com o pré-requisito de que os alunos saibam os conceitos de SQL (BAKER; MITROVIĆ; MATHEWS, 2010). Nas justificativas para o desenvolvimento do estudo, os autores salientaram a dificuldade em se obter um modelo para identificação desses comportamentos que possa ser amplamente utilizado em diferentes tipos de ambientes de aprendizagem. Este problema pode ser definido como um problema geral em pesquisas com análises de dados, porque se o banco de dados não é o mesmo, não há como generalizar as soluções.

Baker, Mitrović e Mathews (2010) apontaram que no SQL-Tutor: solicitar a resposta e em seguida copiá-la e solicitar a resposta e em seguida desistir foram as categorias mais comuns de comportamentos de enganar o sistema, respectivamente responsáveis por 49,6% e 33,6% dos comportamentos desse tipo. Além desses, erros rápidos intencionais, adivinhação sistemática e outras estratégias foram significativamente mais raras, correspondendo à 8,6%, 5,7% e 2,4% respectivamente.

Para desenvolvimento do modelo proposto, os autores utilizaram os dados de logs de 61 alunos, e o conjunto total de dados consistiu em 4 mil tentativas de resolução de problemas, e as instâncias foram previamente rotuladas, pois o modelo proposto para detecção dos comportamentos de enganar o sistema foi baseado em algoritmos de AM supervisionados para classificação. A ferramenta utilizada para construção do modelo foi o software *RapidMiner*, com os algoritmos: Árvores de Decisão (J48), *Step Regression*, *Suporte Vector Machines*, *Naïve Bayes* e *Bagged Decision Stumps*; para avaliar os algoritmos os autores utilizaram o método de validação cruzada, empregando diversas métricas como: área sob a curva ROC,

precisão, *recall* e Kappa. O algoritmo de *Step Regression* gerou o modelo mais eficaz com uma precisão de 76%. Neste estudo os autores também tentaram desenvolver um modelo de classificação para diferenciar os comportamentos de enganar o sistema entre si, não chegando a bons resultados, como denotado pelos próprios autores.

Como salientado, muitos estudos foram desenvolvidos sobre o comportamento de enganar o sistema, e este foi analisado por diversos ângulos, considerando o aluno, ou o ambiente de aprendizagem como o foco dos comportamentos inadequados. Nesse sentido, Muldner *et al.* (2011) tinham como objetivo entender a motivação para essas atitudes identificando se este é um comportamento que o aluno adota porque o sistema favorece, ou se é um comportamento derivado de características do próprio estudante. De forma geral, o estudo incluiu 3 análises principais: 1) se as características do aluno, ou do sistema de tutoria preveem melhor os comportamentos de enganar o sistema; 2) o quanto e de que forma os alunos estão enganando o sistema; e 3) a utilidade dos recursos de ajuda de tutores inteligentes, incluindo o impacto do comportamento de enganar o sistema e vários tipos de ajuda na resolução de problemas e aprendizagem. Como o estudo envolve muitos aspectos e é bastante extenso, serão descritas em síntese as duas primeiras análises realizadas.

Para a realização da primeira análise os autores empregaram várias técnicas (Regressão Linear, Correlação, Histogramas de frequência e Redes Bayesianas) que chegaram às mesmas conclusões, por isso será apresentada apenas a que os autores desenvolveram um modelo baseado em Regressão Linear. Os dados utilizados pelos autores eram originados da interação de alunos com o Tutor Cognitivo Andes, durante 6 aulas da disciplina de física. No total foram utilizados 900 mil pares de interações aluno-tutor, uma quantidade demasiadamente grande de dados para serem rotulados manualmente. Devido a isso, os autores primeiro analisaram uma porção dos dados manualmente, para identificar os padrões indicativos do comportamento de enganar o sistema, em seguida, baseando-se nestes padrões, codificaram um algoritmo para realizar a categorização dos dados.

Tais padrões identificados pelos autores como enganar o sistema consistiam basicamente em: 1) ignorar a dica – o tutor apresenta a dica e o aluno pula a dica e pede outra rapidamente; 2) copiar uma dica – o tutor apresenta uma dica e o aluno gera rapidamente uma entrada de solução, sugerindo uma cópia superficial da dica; 3) adivinhação – o tutor sinaliza uma entrada incorreta e o aluno gera rapidamente

outra entrada incorreta; 4) falta de planejamento – o tutor sinaliza uma entrada correta e o aluno rapidamente pede uma dica, sugerindo confiança em dicas para o planejamento da solução.

Depois dos dados classificados, os autores empregaram a técnica de Regressão Linear, em que a porcentagem de comportamentos de enganar o sistema de um aluno em um problema foi considerada a variável dependente e as duas variáveis independentes foram: 1) aluno – o comportamento inadequado médio de um aluno em todos os problemas resolvidos por esse aluno; e (2) problema – o comportamento inadequado médio em um problema em todo o conjunto de alunos que resolveram esse problema. O modelo obtido foi significativo e com a realização dessa análise Muldner *et al.* (2011) chegaram ao seguinte resultado: se as variáveis independentes forem inseridas separadamente para analisar a variação explicada por cada uma, a variável estudante responde a aproximadamente 50% da variação, enquanto a variável problema responde a cerca de 19%, o que indica que as características do aluno são mais preditoras do comportamento de enganar o sistema, que o ambiente de aprendizagem (atividades propostas). Ainda no intuito de validar este modelo, os autores o aplicaram sobre um outro conjunto de dados originados no Tutor Cognitivo de Álgebra, esse conjunto de dados foi também utilizado por Baker e seus colegas em diferentes estudos, como em Baker e De Carvalho (2008). Os autores relataram que o modelo de produziu resultados semelhantes, em que a variável aluno respondeu por 30% da variação, enquanto a variável problema respondeu por 15%.

Os autores relataram que esta análise forneceu informações que a frequência do comportamento de enganar o sistema depende mais de “quem” e não “do que”, todavia, ela não forneceu uma visão sobre como este comportamento está ocorrendo nem suas consequências. Por isso, os autores indicaram a necessidade da realização da segunda análise, que visava identificar o quanto e de que forma, nos sistemas de tutoria, os alunos têm tais atitudes. De acordo com Muldner *et al.* (2011), a principal forma de comportamento de enganar o sistema foi realizada pelos alunos quando o tutor apresentava uma dica de alto nível (mais geral), em média 19% de todas as ações inadequadas identificadas corresponderam à má utilização dessas dicas. Depois dessa forma de enganar o sistema, as mais frequentes eram a falta de planejamento, em torno e 5%, e copiar as dicas era o padrão de comportamento mais raro.

As dicas correspondem a uma das principais funcionalidades dos tutores cognitivos. Esses sistemas começam dando dicas mais gerais, de alto nível, e à medida que os estudantes precisam de mais ajuda, vão dando dicas cada vez mais diretas, até que o tutor dá a solução. Muito do comportamento de enganar o sistema é baseado em pular as dicas mais gerais até forçar o tutor a dar as dicas diretas e, por fim a resposta. Um fator interessante identificado pelos autores, no que se refere as dicas de alto nível, foi que mesmo aqueles alunos com frequência baixa de comportamentos de enganar o sistema, possuíam um padrão de exploração inadequado dessas dicas, o que indica que quando a oportunidade aparece, mesmo quem não estaria sempre disposto à burla acaba aproveitando a oportunidade. Em suma Muldner *et al.* (2011) destacaram que este estudo trouxe como principais contribuições: 1) O comportamento do aluno é um preditor melhor de comportamentos inadequados do que os problemas propostos no sistema de tutoria, explicando 50% da variabilidade da variável dependente; e 2) Foi identificado diferenças individuais em termos de como os alunos se comportam e, por sua vez, se beneficiam (ou não) das características instrucionais do ambiente de aprendizagem.

Baker, como já salientado é um dos principais pesquisadores sobre o comportamento de enganar o sistema em tutores cognitivos. Dessa forma, também foi colaborador em diversos estudos, alguns dos mais recentes foram realizados com Paquette. Nos trabalhos feitos em colaboração entre estes pesquisadores, além dos modelos baseados em Aprendizagem de Máquina, para identificação de alunos que enganavam o sistema, os pesquisadores empregaram esforços na utilização da *"Knowledge Engineering"* (Engenharia do Conhecimento) para realização desta tarefa. A Engenharia do Conhecimento caracteriza-se quando um modelo é criado a partir do desenvolvimento de regras que capturam o conhecimento necessário para identificar uma característica, tais regras são criadas por um especialista no processo estudado, e as regras podem ser posteriormente generalizadas com a utilização de recursos computadorizados, como os algoritmos de AM.

Em um dos estudos desenvolvidos por estes pesquisadores, com enfoque na Engenharia do Conhecimento, Paquette, De Carvalho e Baker, (2014) buscaram compreender como os especialistas categorizam o descomprometimento do aluno na aprendizagem on-line, e a partir disso propor um modelo para identificação desse comportamento. De forma geral, os autores procuraram extrair conhecimento sobre como os especialistas codificam os alunos que enganam ou não o sistema no contexto

do Tutor Cognitivo de Álgebra. Os autores salientaram que modelos para identificação do comportamento de enganar o sistema podem ser elaborados tanto por meio da Engenharia do Conhecimento, como por meio algoritmos de AM, e ambos podem ser eficientes nesta tarefa.

Todavia, os autores destacaram que entender o contexto do comportamento de enganar o sistema vai além de gerar regras, como: “abuso de ajuda é um comportamento inadequado” e “adivinhação sistemática é comportamento inadequado”. Por exemplo, o abuso de ajuda foi modelado principalmente usando comportamentos que incluem copiar a resposta de uma dica, e solicitações de ajuda repetidas; porém, alunos com dificuldades muitas vezes olham para várias mensagens de ajuda e podem, às vezes, copiar as respostas das dicas como parte de seu processo de aprendizagem, e não necessariamente com intenções de burlar. Neste sentido Paquette, De Carvalho e Baker, (2014), salientaram que os especialistas contam com o contexto em que esses comportamentos ocorrem para determinar se o aluno está se comportamento de forma inadequada ou não, e estudar o contexto usado por estes especialistas para codificar tais comportamentos dos alunos, seria uma etapa potencialmente valiosa para obter uma compreensão mais profunda desse construto e modelá-lo melhor, seja usando AM, Engenharia do Conhecimento, ou ambas.

A fim de estudar como um especialista codifica os comportamentos de enganar o sistema, os autores utilizaram dados de 59 alunos que utilizaram o Tutor Cognitivo de Álgebra durante um ano letivo, como parte de seu currículo regular de matemática, em um total de 12 aulas. Destes dados foram selecionadas 10 mil ações dos alunos, em que cerca de 7% correspondiam ao comportamento de enganar o sistema. O especialista teve acesso a estes dados para que pudesse classificá-los, e para a obtenção de conhecimento sobre este processo os autores se utilizaram de uma abordagem denominada análise de tarefa cognitiva (COOKE 1994; CLARK *et al.* 2008; *apud* PAQUETTE; DE CARVALHO; BAKER, 2014) na qual o conhecimento foi obtido por meio de observações e entrevistas. Essas sessões foram gravadas e utilizadas para elaborar uma versão inicial de um modelo para identificação do comportamento de enganar o sistema, o modelo proposto rotulou as ações de cada aluno usando as regras (Quadro 1) definidas pelo especialista.

A versão inicial do modelo foi aplicada sobre um conjunto de dados de treinamento composto por 75% das ações classificadas com comportamento de

enganar o sistema e 75% das ações classificadas como não enganam o sistema (cerca de 500 e 7 mil ações respectivamente). Após ser treinado, o modelo foi avaliado sobre sua aplicação em um conjunto de dados de teste, composto pelos 25% restantes (cerca de 177 ações classificadas como enganam e 2 mil como não enganam o sistema). O modelo proposto obteve uma precisão de aproximadamente 64% nos dados de treinamento, e cerca de 52% no conjunto de dados de teste.

Embora o modelo seja importante, uns dos principais resultados deste estudo foi a identificação de um conjunto de componentes pelo especialista, que tipificam o comportamento de enganar o sistema. Estes itens foram identificados, principalmente, por meio da observação das pausas entre cada ação, desta forma o especialista pode construir uma interpretação provável do processo mental do aluno. Cada pausa antes ou depois de uma ação é uma oportunidade para o aluno pensar sobre o problema que está resolvendo naquele momento. O Quadro 3 fornece uma lista e a descrição dos componentes comportamentais dos alunos que foram identificados durante o processo de elicitação de conhecimento com o especialista.

Em outro estudo também publicado em 2014 (PAQUETTE *et al.*, 2014), os autores utilizaram o conjunto de componentes elaborados anteriormente pelo especialista (Quadro 3) com a utilização de Engenharia do Conhecimento, como base para desenvolver um modelo baseado em técnicas de Mineração de Dados Educacionais, neste estudo foram aplicados principalmente algoritmos de Aprendizagem de Máquina. Os autores contextualizaram seu problema de pesquisa salientando que à medida que a tecnologia educacional amadurece, as pesquisas debatem se os paradigmas de MDE, ou de Engenharia do Conhecimento são as melhores formas para modelar construções de aprendizagem complexas, nesse sentido Paquette *et al.* (2014) abordaram que um paradigma híbrido pode capturar pontos fortes de ambas as abordagens. Assim, propuseram a utilização das duas estratégias para criação de um modelo otimizado com o intuito de identificar o comportamento de enganar o sistema, em tutores cognitivos.

Quadro 3 – Lista das interpretações consideradas pelo especialista

Componente	Descrição
[não pensei antes do pedido de ajuda]	Pausa menor ou igual a 5 segundos antes de um pedido de ajuda
[pensamento antes do pedido de ajuda]	Pausa maior ou igual a 6 segundos antes de um pedido de ajuda
[ler mensagens de ajuda]	Pausa maior ou igual a 9 segundos por mensagem de ajuda após um pedido de ajuda
[digitalizando mensagens de ajuda]	Pausa entre 4 e 8 segundos por mensagem de ajuda após um pedido de ajuda
[procurando por uma dica final]	Pausa menor ou igual a 3 segundos por mensagem de ajuda após um pedido de ajuda
[pensamento antes da tentativa]	Pausa maior ou igual a 6 segundos antes da tentativa de passo
[planejado com antecedência]	A última ação foi uma tentativa de passo correto com uma pausa maior ou igual a 11 segundos
[acho]	Pausa menor ou igual a 5 segundos antes da tentativa de passo
[tentativa malsucedida, mas sincera]	Pausa maior ou igual a 6 segundos antes de um bug
[adivinhando com os valores do problema]	Pausa menor ou igual a 5 segundos antes de um bug
[ler mensagem de erro]	Pausa maior ou igual a 9 segundos após um bug
[não leu a mensagem de erro]	Pausa menor ou igual a 8 segundos após um bug
[pensou sobre o erro]	Pausa maior ou igual a 6 segundos após uma tentativa de etapa incorreta
[mesma resposta / diferente contexto]	A resposta foi a mesma da ação anterior, mas em um contexto diferente
[resposta semelhante]	A resposta foi semelhante à ação anterior (distância de Levenshtein de 1 ou 2)
[mudou o contexto antes da direita]	O contexto da ação atual não é o mesmo que o contexto da ação anterior (incorreta) (referida como "ponto fraco" em Baker, Mitrovic, e Mathews 2010)
[mesmo contexto]	O contexto da ação atual é o mesmo da ação anterior
[passo repetido]	A resposta e o contexto são os mesmos da ação anterior
[diferente. resposta E / OU diferente contexto]	A resposta ou o contexto não é o mesmo da ação anterior

Fonte: Paquette, De Carvalho e Baker, (2014)

Este estudo utilizou o mesmo conjunto de dados do anterior – composto por dados de 59 alunos, com 10 mil ações, em que cerca de 7% foram classificadas como comportamentos de enganar o sistema. Os 19 componentes identificados anteriormente, por meio da análise da tarefa cognitiva, foram utilizados como recursos para classificação das instâncias, dessa forma o número de vezes que cada regra apareceu foi calculado. Dessa forma, além da construção do modelo, os autores também estavam em busca de validar esses componentes, verificando se cada um se caracterizava como um padrão comportamental independente.

Para o desenvolvimento do modelo Paquette *et al.* (2014) utilizaram o software *RapidMiner*, empregando os algoritmos: Árvore de Decisão (J48), *JRip*, *Step Regression* e *Naïve Bayes*. Os autores realizaram três experimentos, em que utilizaram formas diferentes de classificações das instâncias, feitas a partir dos

componentes elicitados (Quadro 1), refinando os modelos nessas três etapas. Os modelos gerados foram uma combinação das estratégias de Engenharia do Conhecimento, que resultaram no desenvolvimento do Quadro 1, e da técnica de AM, por meio de seus algoritmos de classificação. No geral o melhor modelo foi originado pelo algoritmo *Naïve Bayes*, que alcançou uma precisão média nos três experimentos de 53% sobre as instâncias classificadas com o comportamento de enganar o sistema, e uma média de 85% na métrica referente a área sob a curva ROC.

Todavia, a principal contribuição deste estudo foi que com a implementação de um modelo de forma híbrida, com técnicas de MDE baseadas em AM e Engenharia do Conhecimento, houve uma melhoria na compreensão dos autores sobre os componentes, ações que tipificam o comportamento de enganar o sistema, no Tutor Cognitivo de Álgebra. Portanto, com a aplicação dos algoritmos e o desenvolvimento das experimentações, os autores chegaram à conclusão de que havia menos padrões comportamentais independentes entre os 19 componentes, isso indica que para modelar construções complexas, a combinação da Engenharia do Conhecimento e da MDE pode ser mais vantajosa, do que qualquer uma das estratégias isoladamente. Embora os autores relatem a identificação destes padrões mais gerais, não chegaram a listá-los, porém fazem isso em outros estudos, como o abordado na sequência.

Em uma continuação dos estudos de 2014, Paquette e Baker (2017) sumarizaram as regras apresentadas no Quadro 3 em 13 padrões de ações associados ao comportamento de enganar o sistema (Quadro 4), e compararam como esses padrões de ações são distribuídos entre diferentes populações de alunos usando o Tutor Cognitivo de Álgebra, e entre os alunos usando um dos seguintes ambientes de aprendizagem: Tutor Cognitivo de Álgebra, Tutor Cognitivo do Ensino Médio e o ASSISTments. Com o intuito de entender como (e se) os comportamentos de enganar o sistema se manifestam de formas diferentes nas populações de alunos, e em ambientes de aprendizagem distintos.

Na sumarização proposta (Quadro 4), os autores elencaram apenas as ações que estão estritamente associadas ao comportamento de enganar o sistema, e as formataram de modo generalista, excluindo as variáveis temporais. Os autores justificaram esta reformulação, porque segundo eles nenhum componente apresentado anteriormente era independente o suficiente para identificar o comportamento de enganar o sistema, mas que certas combinações desses componentes eram; então, por meio de entrevistas com especialistas, e dos

experimentos realizados em Paquette *et al.* (2014) identificaram 13 padrões essenciais dos 19 componentes anteriores, aos quais Paquette e Baker (2017) se referem como: características padrão do comportamento de enganar o sistema.

Quadro 4 – Descrição dos padrões do comportamento de enganar o sistema

N°	Padrão
1	O aluno insere uma resposta incorreta e, em seguida, insere rapidamente a mesma resposta incorreta em uma parte diferente do problema.
2	O aluno insere uma resposta incorreta, insere uma resposta semelhante e incorreta na mesma parte do problema, em seguida, insere outra resposta semelhante na mesma parte do problema.
3	O aluno insere uma resposta incorreta, seguida por uma resposta semelhante e incorreta e, finalmente, insere novamente a segunda resposta em uma parte diferente do problema.
4	O aluno insere rapidamente uma resposta incorreta, seguido por inserir rapidamente uma segunda resposta incorreta e, então, mais uma vez, inserir rapidamente uma resposta diferente.
5	O aluno insere uma resposta incorreta, seguida por uma resposta incorreta semelhante e, em seguida, insere rapidamente uma resposta diferente.
6	O aluno pede ajuda e rapidamente procura a resposta nas dicas fornecidas pelo ambiente de aprendizagem, insere uma resposta incorreta e depois insere uma resposta incorreta semelhante.
7	O aluno insere uma resposta incorreta, seguida pela mesma resposta incorreta em uma parte diferente do problema, em seguida o aluno tenta responder ou solicita ajuda para uma parte diferente do problema.
8	O aluno insere um erro conhecido (reconhecido pelo sistema como um "bug"), depois insere novamente a mesma resposta em uma parte diferente do problema, obtém a resposta certa e, em seguida, insere um novo bug para uma parte diferente do problema.
9	O aluno insere uma resposta incorreta, insere uma resposta incorreta semelhante e, em seguida, passa para uma parte diferente do problema e insere outra resposta incorreta.
10	O aluno insere uma resposta incorreta, passa para uma parte diferente do problema, mais uma vez insere uma resposta incorreta e, em seguida, insere uma resposta incorreta semelhante.
11	O aluno insere uma resposta incorreta, seguida por uma resposta incorreta semelhante, não pausa (espera um tempo) para pensar no erro antes de pedir ajuda e, finalmente, insere uma resposta incorreta semelhante.
12	O aluno pede ajuda, seguido por uma sequência de 3 respostas incorretas com pelo menos 2 das quais são semelhantes entre si.
13	O aluno insere uma sequência de 3 respostas incorretas, pelo menos 2 das quais são semelhantes entre si, e então pede ajuda rapidamente sem pensar nos erros.

Fonte: Paquette e Baker (2017)

Para comparar como os 13 padrões destacados (Quadro 4) podem se diferenciar em populações de alunos e em ambientes de aprendizagem, os autores aplicaram o modelo proposto em Paquette, De Carvalho e Baker, (2014) em 6 conjuntos de dados diferentes. Os três primeiros conjuntos de dados foram obtidos no repositório de dados *DataShop*, e foram todos coletados de alunos que utilizaram o Tutor Cognitivo de Álgebra, todavia cada conjunto foi coletado de uma escola e representa uma população diferente de alunos: estudantes rurais, suburbanos e urbanos. Os outros três conjuntos de dados foram coletados de diferentes ambientes de aprendizagem: Tutor Cognitivo de Álgebra, Tutor Cognitivo do Ensino Médio e ASSISTments. Para comparar as distribuições de porcentagens relativas do padrão de comportamento estudado, os autores empregaram análises estatísticas (*Kruskall-*

Wallis e *Mann-Whitney*), isso permitiu que eles investigassem quais padrões eram mais comuns em cada conjunto de dados.

Não são especificados todos os resultados encontrados pelos autores, mas em suma eles chegaram à conclusão de que havia diferenças significativas na natureza do comportamento de enganar o sistema demonstrado por diferentes populações de alunos e em diferentes ambientes de aprendizagem virtuais. No entanto, as diferenças entre as populações tendem a ser menos frequentes e mais fracas do que entre os ambientes. Além disso Paquette e Baker (2017) puderam verificar que para todos os conjuntos de dados, os padrões de comportamentos de enganar o sistema mais praticados pelos alunos foram o 2 e o 5 (Quadro 4), variando entre 20% e 36% das ações relacionadas a estes comportamentos. Por fim, os autores destacaram que o estudo realizado permitiu identificar que os comportamentos dos alunos, mais especificamente os comportamentos de enganar o sistema, podem variar em diferentes conjuntos de dados, dessa forma é importante considerar essa informação ao construir modelos para identificação de determinados comportamentos, seja para aqueles relacionados a manipulação do sistema ou para outras construções.

Em continuação aos três últimos estudos destacados, Paquette, Baker e Moskal (2018) propuseram uma aplicação que busca generalizar a detecção do comportamento de enganar o sistema, para tutores Cognitivos, como: Tutor Cognitivo de Álgebra, CTAT (*Cognitive Tutor Authoring Tools*) e ASSISTments. Os autores relataram que grande parte da dificuldade em se estudar dados educacionais é a ampla diferença nos dados armazenados em cada ambiente de aprendizagem, o que leva os pesquisadores a desenvolver modelos muito específicos para cada contexto estudado. Isso acaba impedindo a disseminação de modelos gerais o suficiente para funcionar em múltiplos contextos de tutoria por exemplo, exigindo que pesquisadores interessados em estudar intervenções pedagógicas relacionadas aos comportamentos dos alunos (como o de enganar o sistema) criem seu próprio modelo o que requer recursos consideráveis.

Nesse sentido, Paquette, Baker e Moskal (2018) apresentaram nesta pesquisa um modelo generalizável (para o âmbito de tutores cognitivos), para detectar alunos que enganam o sistema. Eles tinham como principal objetivo tornar essa ferramenta disponível a todos os interessados, reduzindo a quantidade de esforço necessária para conduzir pesquisas nesse contexto, e aumentar a adoção de modelos

computacionais para comportamentos inadequados por desenvolvedores de tutores inteligentes, seja para dirigir intervenções automáticas, ou para relatar tais atitudes aos professores, por meio de painéis.

O modelo proposto se baseia nos 13 padrões de ações relacionadas ao comportamento de enganar o sistema descritos em Paquette e Baker (2017), e é uma evolução do modelo inicial proposto em Paquette, De Carvalho e Baker, (2014). O modelo está contido em um arquivo *JavaScript* (*"gaming.js"*) que pode ser baixado por um *link* disponibilizado pelos autores. Uma vez baixado, o arquivo pode ser integrado a um tutor para detectar automaticamente comportamentos de enganar o sistema durante o tempo de execução, ou pode ser carregado em um repositório de dados como a *LearnSphere.* Para que seja possível utilizar este modelo em dados históricos de ambientes de aprendizagem, estes conjuntos de dados devem seguir o padrão de armazenamento dos tutores mencionados, não se aplicando, por exemplo aos dados de logs do Lúmina.

Uma vez que o modelo é incluído no tutor, ele gera automaticamente o diagnóstico do comportamento do aluno a cada cinco ações. Na aplicação em dados históricos, por exemplo do repositório *LearnSphere*, requer que o usuário crie um fluxo de trabalho, em que o usuário pode carregar qualquer arquivo de texto delimitado por tabulação, contendo os dados formatados de acordo com o padrão *DataShop*, bem como carregar o arquivo *"gaming.js"*, contendo a aplicação. Esses dois arquivos são então usados como entradas e a saída é um arquivo contendo todos os diagnósticos das ações. É importante notar que, embora seja possível aplicar o modelo a qualquer conjunto de dados armazenados no padrão *DataShop*, este foi validado apenas para os Tutores Cognitivos, Tutores CTAT e ASSISTments. Por fim, os autores pretendiam com a disponibilização desta ferramenta, contribuir para um aumento no uso desses modelos automatizados para identificação de comportamentos inadequados de alunos enquanto utilizam ambientes virtuais de aprendizagem, e apoiar no desenvolvimento de diferentes estratégias pedagógicas para amenizar os impactos causados na aprendizagem por este tipo de atitude.

Para finalizar as explanações de estudos sobre enganar os sistema, destaca-se uma pesquisa também realizada por Paquette e Baker (PAQUETTE; BAKER, 2019) em que os pesquisadores comparam a AM, a Engenharia do Conhecimento e a abordagem híbrida para modelagem desse tipo de comportamento. Neste artigo os autores compararam modelos implementados por eles e seus colegas em estudos

anteriores, que foram destacados anteriormente neste subcapítulo: Modelo baseado em AM proposto em Baker e De Carvalho (2008); Modelo baseado em Engenharia do Conhecimento proposto em Paquette, De Carvalho e Baker (2014); e Modelo Híbrido proposto em Paquette *et al.* (2014). Os autores compararam as abordagens em três dimensões principais: 1) precisão dos modelos nos dados originais, utilizados para implementá-los; 2) interpretabilidade do modelo; e 3) generalização do modelo para novos dados e contextos.

Quanto à precisão nos dados originais o modelo baseado na abordagem híbrida resultou no desempenho preditivo mais alto, com o valor médio da área sob a curva ROC acima de 80% e o modelo de Engenharia de Conhecimento obteve desempenho superior ao modelo de AM. No que se refere à interpretabilidade dos modelos, o modelo gerado a partir da Engenharia do Conhecimento é projetado para ser interpretável, visto que esta estratégia representa o resultado da elicitação explícita do conhecimento do especialista, tornando-o relativamente simples de interpretar; no entanto, não permite que um pesquisador estude o comportamento além do que o especialista foi capaz de elicitar. Quanto à generalização, o modelo baseado na abordagem de Engenharia do Conhecimento obteve o melhor desempenho quando aplicado a novos contextos, sendo um pouco melhor do que os modelos híbridos, apesar de ser inferior no contexto original; sugerindo que o desempenho do modelo de Engenharia de Conhecimento é mais estável.

Como conclusões, Paquette e Baker (2019) destacaram as vantagens e as desvantagens de cada uma das abordagens comparadas: Aprendizagem de Máquina – consistiu na menor quantidade de recursos e esforço necessários para desenvolver o modelo, no entanto originou um modelo mais difícil de ser interpretado e com baixa generalização para novos dados e contextos; Engenharia do Conhecimento – teve a maior robustez, apresentado um modelo bom ao ser generalizado para novos contextos e melhor interpretabilidade, entretanto essas vantagens tem como custo maiores esforços de desenvolvimento, exigindo um investimento de tempo considerável de um especialista na modelagem do comportamento do aluno e de um engenheiro de conhecimento experiente; e Abordagem Híbrida – teve um bom desempenho em todas as três dimensões estudadas: obteve alto desempenho no contexto original, é interpretável e se transferiu comparativamente bem entre os contextos, todavia a principal desvantagem desta abordagem são os maiores recursos necessários para desenvolver os modelos.

As pesquisas apresentadas sobre o comportamento de "*gaming the system*", embora não tratem do contexto específico pesquisado nesta Tese, são relevantes como precursoras para o desenvolvimento de estratégias para identificar comportamentos indesejados em ambientes de aprendizagem virtuais, como dos MOOCs. Um ponto negativo é que os modelos especificamente criados para os contextos dos Tutores Cognitivos não podem ser aplicados a outros ambientes de aprendizagem, pois estes têm recursos diferentes e por isso os dados armazenados não coincidem, impossibilitando o reaproveitamento. Além disso, devido a esta mesma diferença, nos recursos entre as plataformas, os comportamentos que são indicativos do comportamento de caçadores de certificados identificado no Lúmina, não são similares ao comportamento de enganar o sistema nos Tutores Cognitivos. Nesse sentido, foi elaborado um quadro comparativo (Quadro 5) especificando os comportamentos que são considerados como enganar o sistema nos tutores cognitivos, e se estes podem ser transpostos para o âmbito dos MOOCs do Lúmina.

Quadro 5 – Enganar o sistema nos tutores cognitivos e sua aplicação no Lúmina

Enganar o Sistema	Aplicação aos MOOCs do Lúmina
Dicas progressivamente mais objetivas (abuso de ajuda).	Não há possibilidade de análise, pois no sistema Moodle não existe este recurso.
Tempo muito rápido entre respostas.	Essa informação é possível de ser analisada utilizando o relatório de logs, observando o tempo entre as tentativas nos questionários.
Tempo fora da tarefa (off-task).	Não é possível, pois como os tutores são aplicados em sala de aula, esta é uma variável observável. No entanto, os MOOCs do Lúmina são on-line, os alunos podem estar fora da tarefa, mas estudando outros materiais do próprio curso, sem que seja possível esta diferenciação.
Cometer muitos erros em uma etapa específica da atividade, em meio a outras similares.	Parcialmente, pois tem-se a possibilidade de acessar a quantidade de tentativas e a quantidade de questionários, no relatório de logs. Uma variável que mede essa relação pode então ser criada, com a razão de quantidade de questionários pela quantidade de tentativas.
Acertar na primeira tentativa em alguns problemas e cometer muitos erros em outros problemas similares.	Não é possível saber o que são problemas similares nos MOOCs do Lúmina, a menos que isso seja dito pelo professor responsável pelo MOOC.
Comportamento *"gamed-not-hurt"* – enganar o sistema com consequências não prejudiciais ao aluno.	Há possibilidade de ser analisado, utilizando em especial o relatório de Notas. Todavia, há limitações, pois nos tutores há como analisar também as habilidades alcançadas e comparar as notas do pré e pós teste para saber se houve ganhos na aprendizagem; no Lúmina há apenas como analisar as notas nas atividades correntes do curso, se foram boas ou ruins, não há pré e pós teste.
Comportamento *"gamed-hurt"* – enganar o sistema com consequências prejudiciais ao aluno.	Há possibilidade de ser analisado, utilizando em especial o relatório de Notas, porém como detalhado no item anterior há limitações.

Fonte: Autora

Embora os modelos elaborados no contexto dos tutores não possam ser utilizados para os MOOCs, as estratégias utilizadas pelos autores para geração desses modelos, e classificação das bases de dados são válidas para qualquer contexto, podendo ser exploradas no âmbito do Lúmina.

5.2 ENGANANDO O SISTEMA EM MOOCS: CAMEO

Com a disseminação dos cursos on-line, sobretudo os MOOCs, houve uma evolução nas formas de enganar o sistema, uma migração dessas estratégias inadequadas para se adaptarem aos novos contextos de aprendizagem virtuais. Nesse sentido uma das principais estratégias identificadas na literatura foi o CAMEO (*Copying Answers using Multiple Existences Online*). Northcutt, Ho e Chuang (2016), os primeiros a caracterizar o CAMEO, notaram a velocidade com que muitos estudantes completavam um curso em plataformas como a *edX*, o que chamou atenção. Neste estudo os autores primeiramente explicaram como o CAMEO acontece: um usuário usa uma ou mais contas em uma plataforma MOOC, com o intuito de encontrar a resposta correta para uma pergunta, e então enviar esta resposta em sua conta oficial, pela qual o aluno pretende obter um certificado. Os autores se referem às contas utilizadas para encontrar a solução como contas de “colheita” e a conta principal como conta “mestre”.

Posteriormente, os autores propuseram um algoritmo para identificação dos usuários que praticam o CAMEO, baseado na distribuição das diferenças de tempo entre ações específicas do usuário entre pares. De forma geral o algoritmo considera 5 critérios, analisados nos relatórios de logs dos alunos, para identificação do CAMEO: 1) A conta de colheita deve escolher a resposta correta antes que a conta mestre envie essa resposta; 2) A conta de colheita deve fornecer respostas à conta mestre rapidamente (tempo inferior a 5 minutos); 3) A conta de colheita não deve ser certificada, enquanto que a conta mestre sim; 4) As contas de colheita e mestre devem compartilhar um endereço IP, ou ter compartilhado um em algum momento de seu histórico de cursos, o que aumenta a probabilidade de que o realmente as contas identificadas sejam de fato a mesma pessoa; 5) Deve haver poucas contas que compartilham ou compartilharam um endereço IP com uma conta de colheita e mestre, o que exclui cibercafés, redes de escolas e outros espaços comuns onde a coincidência casual de tempo de resposta e IP podem levar a uma detecção falsa.

Além de propor o algoritmo, os autores realizaram um experimento para validar seu funcionamento com aproximadamente 1,8 milhões de participantes de 115 MOOCs de duas universidades. Nos cursos investigados os autores estimaram que 1.237 certificados foram obtidos usando a estratégia CAMEO. Em alguns MOOCs os usuários CAMEO correspondiam a até 5% dos certificados obtidos. Os autores propuseram algumas medidas preventivas para a diminuição da incidência desse tipo de comportamento, como: restringir a opção “mostrar resposta”; e randomizar as questões avaliativas com soluções diferentes. Como este estudo foi pioneiro na questão da caracterização e detecção do CAMEO, justificam-se os esforços para o desenvolvimento de um algoritmo, por isso esse manuscrito foi bastante referenciado pelos seus sucessores.

O estudo apresentado por Ruiperez-Valiente *et al.* (2016) teve como objetivos primeiramente definir o fenômeno do CAMEO, e depois propor e validar um algoritmo para sua detecção em MOOCs. Para a validação os autores utilizaram um MOOC da área de física da plataforma *edX* com 13,5 mil participantes. O algoritmo desenvolvido utiliza como entrada os arquivos de Log dos usuários ordenados por data. Cada evento do relatório de Logs representa uma ação do aluno e contém o IP – similar ao do Lúmina. Uma submissão a uma pergunta cria um evento que contém, entre outras coisas, informações sobre a pergunta (o ID da pergunta), o nome do aluno, a resposta e se está correta ou não. O funcionamento do algoritmo segue as seguintes etapas: 1) processamento dos arquivos de Log do usuário; 2) construção de dois dicionários – um que mapeia cada aluno para todos os IPs usados, e um que mapeia um IP para todos os alunos que o usaram pelo menos uma vez durante o curso; 3) detecção de respostas bem sucedidas dos alunos, ou pedido para ver respostas dentre os grupos de IPs; 4) identificação de dez ou mais submissões bem sucedidas às questões com no máximo 20 segundos entre cada submissão consecutiva; 5) testa os critérios de definição para ser uma conta CAMEO; e 6) remoção das entradas que contenham itens que não cumpram estes critérios.

Como resultados, os autores demonstraram que dos ganhadores de certificados, cerca de 10% usaram CAMEO, também concluíram que os alunos que utilizam esse método tendem a ter alta taxa de sucesso e tempo de resposta rápido, em comparação com outros alunos. Descobriram ainda, que as características de uma pergunta como alto nível de complexidade, se correlacionam fortemente com a quantidade de trapaça. Como propostas para práticas de design instrucional que são

menos vulneráveis ao CAMEO, os autores propuseram atrasar o feedback e usar à randomização de problemas. No geral, este estudo traz grandes contribuições no que diz respeito à descrição e caracterização do CAMEO.

A pesquisa desenvolvida por Ruiperez-Valiente *et al.* (2017a) é uma continuação da descrita anteriormente, em que autores utilizaram a base de dados, já classificada na pesquisa de 2016 (com valores 0 e 1, para “não CAMEO” e para “CAMEO”, respectivamente) com o intuito de: 1) comparar a influência das características do aluno e dos problemas no CAMEO; e 2) construir um modelo com o algoritmo Floresta Aleatória que detecta as submissões com CAMEO, sem depender de IP. Os atributos utilizados pelos autores para compor a base de dados foram divididos em 3 partes:

1. A primeira relacionada aos *alunos,* que engloba: desempenho – porcentagem de problemas que foram apresentados corretamente na primeira tentativa do aluno; soma do tempo de vídeos assistidos; tentativas de responder as questões; tentativas com a resposta correta; tempo médio gasto para enviar uma resposta correta.
2. A segunda relacionada as *questões avaliativas,* formada por: atribuição de tipo – variável de fator que indica se o problema era um “Quizz”, “Dever de Casa” ou “Checkpoint”; tipo de resposta – variável de fator que define o tipo de resposta de cada problema, por exemplo, múltipla escolha, preencher o espaço em branco, ou fórmula; mostrar resposta – variável de fator que define a configuração do botão “mostrar resposta”; localização – do problema dentro do curso; random – variável binária indicando se o problema contém variáveis aleatórias ou não; tentativas máximas – especifica o máximo número de tentativas permitidas no problema.
3. A terceira se refere aos *recursos de envio*, que possui as variáveis: tempo até o prazo final; duração da tentativa; tentativas necessárias.

A base de dados possuía mais de 470 mil submissões corretas, das quais cerca de 27 mil (6,13%) foram rotuladas como CAMEO. O principal resultado desse estudo, segundo Ruiperez-Valiente *et al.* (2017a), foi um novo método para a detecção do CAMEO, com base em uma abordagem de Aprendizado de Máquina que não depende do uso de IPs. Ademais, os autores destacam que o modelo de classificação gerado a partir do algoritmo Floresta Aleatória alcançou resultados bastante positivos com níveis de sensibilidade e especificidade de 0,966 e 0,996, respectivamente. Esse

estudo realmente traz uma contribuição relevante na questão da metodologia para identificação do CAMEO, pois apresenta uma forma de detecção, por meio de Aprendizagem de Máquina, otimizando a solução proposta anteriormente (RUIPEREZ-VALIENTE *et al.*, 2016), pois a partir da criação do modelo este pode ser reaplicado nos Logs de qualquer aluno que esteja matriculado em um MOOC da *edX*. Assim, não é mais necessária a aplicação de um algoritmo específico, ademais, o algoritmo Floresta Aleatória utilizado, já está consolidado na literatura.

Dando continuidade às suas pesquisas sobre comportamentos, condutas inadequadas e trapaças em MOOCs, Ruipérez-Valiente *et al.* (2017b) apresentaram um estudo que busca entender de forma mais ampla os comportamentos de usuários de MOOCs. O estudo não iniciou com a intenção de distinguir um comportamento específico, mas visava fornecer uma abordagem geral para reconhecimento de diferentes tipos de associações entre os usuários que podem ser investigadas posteriormente. Para esse fim, os autores propuseram um algoritmo para detectar laços entre alunos com base em proximidade temporal de seus envios de tarefas, sua aplicação pode identificar laços genuínos entre dois estudantes, ou detectar irregularidades como o CAMEO.

Para testar a viabilidade do algoritmo, os autores utilizaram dados de dois MOOCs oferecidos pela Universidade de Edimburgo: Introdução à Filosofia e Teoria da Música, com cerca de 2,3 mil e 5 mil alunos matriculados, respectivamente; dos quais extraíram os seguintes atributos: nota final; lista de todas as submissões de atividades avaliativas; identificação se o aluno gerou certificado ou não; número total de tentativas nas atividades avaliativas; número total de dias que um aluno ficou ativo no curso; número total de vídeo aulas acessadas pelos alunos; número total de tópicos de discussão acessados; por fim, todos os acessos foram ordenados por data. O algoritmo proposto calcula a similaridade do aluno, por meio do cálculo da distância entre os tempos de envios das atividades, como por exemplo, é feito nos algoritmos KNN, ou *K-means*, só que nesse caso ao invés da distância entre dois pontos, comumente empregada nesses algoritmos, foram utilizadas a distância do desvio absoluto médio e o desvio médio quadrático.

Como resultados Ruipérez-Valiente *et al.* (2017b) ressaltaram que o algoritmo permitiu a identificação automática de alunos que tendem a estudar juntos e revelam diferentes tipos de colaborações – por exemplo, aqueles que verdadeiramente colaboram para aprender e obter um certificado e aqueles que talvez mostrem certos

comportamentos que poderiam ser caracterizados como desonestidade acadêmica, incluindo o CAMEO. Essas informações podem ser relevantes para alunos que procuram colaboradores em potencial, bem como professores, na forma de *insights* sobre vários comportamentos emergentes das interações dos alunos.

A pesquisa realizada por Alexandron *et al.* (2017) teve como objetivos determinar a prevalência de CAMEO, estudar as suas características e inferir a motivação para utilização dessa estratégia. Para determinar a prevalência do CAMEO os autores propuseram um algoritmo que tem como entrada os arquivos de Log dos alunos e é composto por duas etapas: a primeira etapa coleta eventos que aderem ao esquema geral do CAMEO, na qual uma conta obtém a solução para um problema e outra conta dentro do mesmo grupo de IPs envia uma resposta semelhante a essa pergunta logo depois; a segunda etapa concentra-se em filtrar os falsos positivos aplicando vários filtros. A fim de validar o algoritmo, os autores realizaram sua aplicação em dados de um MOOC da área de física da plataforma *edX*, obtendo um bom funcionamento. Quanto a estudar as características do CAMEO, os autores o descreveram e caracterizaram como ocorre, salientando suas particularidades e similaridades com práticas de desonestidade acadêmicas em outros ambientes.

Como resultados, Alexandron *et al.* (2017) apresentaram que 13% dos usuários investigados utilizaram o CAMEO, e como recomendações para sua diminuição os autores sugeriram: aumento no uso da randomização de questões quando possível; atrasar o feedback, especialmente em questões de alto risco; reconhecimento da importância deste problema e alocação de recursos (como tempo) para permitir que os instrutores projetem seus cursos de uma forma que sejam menos vulneráveis ao CAMEO; inclusão na plataforma de ferramentas para detectar CAMEO em tempo de execução; e desenvolvimento de métodos de identificação para garantir que os certificados indiquem conhecimentos e habilidades realmente desenvolvidas pelos alunos.

Para finalizar a descrição dos trabalhos relacionados, apresenta-se a pesquisa realizada por Bao, Chen e Hauff (2017) que replica o estudo realizado por Northcutt, Ho e Chuang (2016), aplicando o algoritmo desses autores a dez MOOCs da *edX*, criados pela *Delft University of Technology*. Os resultados expostos mostraram que em torno de 1,9% dos certificados foram provavelmente ganhos por meio da estratégia CAMEO, um número comparável à fração de trapaça observada no estudo de referência. Embora este estudo seja interessante, por comprovar as

alegações expostas na pesquisa de Northcutt, Ho e Chuang (2016), não traz nenhuma informação nova.

Tendo em vista os trabalhos relacionados, que dizem respeito ao CAMEO, uma das metodologias de identificação de trapaças mais relevante é a apresentada por Ruipérez-Valiente *et al.* (2017a), em que primeiramente os autores identificam os alunos que praticaram o CAMEO, depois classificaram uma base de dados e aplicaram um algoritmo de Aprendizagem de Máquina como técnica de MDE, no intuito de constatar esse tipo de trapaça. Desse estudo também é relevante a forma como a base de dados foi formatada, pois os autores fazem uma boa descrição sobre os atributos que foram utilizados para o processo de identificação do comportamento de trapaça. Outro estudo de relevância, em que destaca-se o método aplicado para identificar comportamentos em MOOCs e a formatação da base de dados, é Ruipérez-Valiente *et al.* (2017b), os autores utilizaram um algoritmo similar ao *K-means* (agrupamento) para analisar os dados dos alunos, e por meio de uma descrição detalhada presente no estudo, foi possível ter um bom entendimento de todos os dados que compunham suas bases de análise.

Nesta Tese, considera-se estes estudos como um ponto de partida, pois o objetivo desses autores era identificar um tipo específico de estratégia para trapacear, o CAMEO; entretanto, nesta pesquisa o objetivo é reconhecer um padrão de comportamento inadequado mais amplo, levando em consideração as características de acesso dos alunos, como será descrito em seguida.

5.3 ENGANANDO O SISTEMA EM MOOCS: CAÇADORES DE CERTIFICADOS

O tipo de comportamento indesejado identificado no Lúmina não é, necessariamente, similar ao CAMEO. O que os administradores da plataforma suspeitam[9] é que muitos alunos sejam “caçadores de certificados”, pois exibem um padrão: obtém certificados de cursos muito diferentes, em um intervalo curto e permanecem pouco tempo na plataforma. Analisando os logs, observa-se muitas tentativas realizadas nas atividades avaliativas e um intervalo pequeno na realização dos MOOCs, entre o primeiro acesso até a geração do certificado. Além disso, como

[9] Informações obtidas com a orientadora, que é uma das administradoras da plataforma, e tem contato direto com professores que criam materiais para a plataforma, e com os estagiários que dão suporte aos usuários e acompanham os dados do Lúmina.

a maioria dos cursos não têm atividades avaliativas exigentes ou numerosas, e os mesmos permitem até três oportunidades para resolução e envio, sendo que a maior nota é a considerada como nota final, as notas costumam ser muito altas, com médias próximas à 80%.

Embora os estudos sobre enganar o sistema tenham sido desenvolvidos para o contexto dos tutores cognitivos, as estratégias utilizadas pelos alunos do Lúmina, bem como o CAMEO, se enquadram na mesma definição, pois os alunos exploram propriedades do sistema, ao invés de aprender com o material, visando obter os certificados com menor esforço e o mais rápido possível. De forma geral, o comportamento de burlar o sistema identificado no Lúmina parece estar relacionado sobretudo às características do comportamento de acesso dos estudantes aos recursos, que não evidenciam um comprometimento com a aprendizagem quando realizam um MOOC, associada à exploração das funcionalidades disponíveis na plataforma, apenas para geração do certificado.

Devido às particularidades observadas nas condutas dos alunos com o comportamento descrito, presume-se que se os MOOCs do Lúmina forem configurados com atividades que requeiram mais esforço e sejam pré-requisitos para avançar no curso, os alunos não comprometidos tenderão a desistir. Ainda que as configurações do Lúmina estejam limitadas ao que o Moodle permite, é possível reduzir as oportunidades para enganar o sistema, tomando ações como por exemplo: reduzir a quantidade de tentativas de um questionário; configurar as questões para aparecerem de modo aleatório; incentivar os professores a elaborarem questionários com mais questões e com questões mais difíceis; dar o retorno sobre quais alternativas estão corretas apenas após o envio da última tentativa; calcular a nota usando a média das tentativas; e predefinir um tempo mínimo de permanência no curso para geração do certificado.

Outro fator que pode ser determinante para a desistência de alunos não comprometidos com a aprendizagem é a recompensa ao final do curso, a certificação. Acredita-se que MOOCs que apresentem cargas horárias reduzidas podem não ser tão atrativos, como outros com cargas horárias maiores, caso se aumente a quantidade de esforço exigido para concluir esse MOOC – tanto para quem tem intenção de burlar o sistema como para quem quer obter o certificado de forma honesta. Dessa forma, a relação esforço/recompensa parece ser um importante ponto de análise para o comportamento dos caçadores de certificados no Lúmina.

Considera-se como *"esforço"* o empenho que um estudante deve despender para realização de um MOOC na plataforma e tudo aquilo que ele deixa de fazer para estudar. A *"recompensa"* é a retribuição que se ganha ao terminar o curso, no caso do Lúmina a carga horária do certificado, a forma de validação dele (por exemplo, pela Pró-reitora de Extensão ou pela plataforma), o conhecimento adquirido e o reconhecimento social advindo desta especialização.

Cabe lembrar que a punição para a burla em plataformas de MOOCs é nula, de forma que mesmo alunos que a princípio não tentariam enganar o sistema podem se sentir tentados a reduzir a quantidade de esforço para obter as recompensas, dessa forma o melhor meio para inibir esses comportamentos é desmotivando o aluno a praticá-lo.

Em MOOCs ofertados por outras plataformas, fica um pouco mais evidente os fatores que podem motivar este tipo de comportamento, pois essas certificações representam, em muitos casos, créditos em cursos de graduação e Pós-Graduação de instituições de ensino. Embora os certificados ofertados pelo Lúmina não tenham validade direta como créditos em cursos de graduação ou Pós-Graduação ofertados pela UFRGS, estes podem ser utilizados como: comprovação de horas de atividades complementares em cursos de graduação (caso as coordenações dos cursos aceitem); comprovação para conhecimentos especificados nos currículos; contagem de horas de qualificação para professores do ensino Básico, o que colabora para progressão de carreira; em alguns processos seletivos como para oficiais temporários do Exército[10], certificados como os do Lúmina, são considerados na pontuação global dos inscritos; e contam pontos em concursos públicos. Nesse sentido, alguns desses fatores podem provavelmente ser indicativos das motivações de um indivíduo para se inscrever em um MOOC no Lúmina.

Todavia, de acordo com Muldner *et al.* (2011) um dos principais desafios consiste em compreender as causas desses comportamentos inadequados. Os autores relataram que os primeiros trabalhos realizados para entender por que os comportamentos de enganar o sistema ocorrem, focaram nas características dos alunos e como elas se correlacionam com esses comportamentos, levando em consideração os objetivos dos alunos, atitudes, afeto, dentre outros. Em um contraponto, o autor destaca que com a evolução dos estudos nesse âmbito, foi

[10] https://www.eb.mil.br/web/ingresso/militar-temporario

constatado que há uma associação entre os recursos das aulas, o design de interface do sistema e os comportamentos de burla, indicando que se o aluno está em um impasse, causado por um projeto de ambiente virtual de aprendizagem insatisfatório, e não por uma falta de conhecimento, ou motivação, ele irá usar de estratégias para contornar o impasse, não necessariamente de forma deliberada.

Por fim, Muldner *et al.* (2011) destacaram que diversas pesquisas que não envolveram tutores inteligentes mostraram que mesmo quando os alunos usam exatamente os mesmos materiais de instrução, eles variam acentuadamente na forma como escolhem processá-los (por exemplo, CHI *et al.* 1989; RENKL 1997; VANLEHN 1998, apud MULDNER *et al.*, 2011). Esses estudos sugerem que algum tipo de característica do aluno (por exemplo: conhecimento, motivação) pode ser mais importante do que os recursos instrucionais na determinação do comportamento de burla (MULDNER *et al.*, 2011), como os próprios autores comprovaram com o desenvolvimento de seu estudo.

Dessa forma, percebe-se o quão complexo é o processo de identificação desses comportamentos, que depende tanto das características dos alunos, como da plataforma utilizada, e no geral parece estar relacionado aos dois elementos. Levando em consideração esses dois aspectos de análise, neste estudo há um interesse em particular de verificar se existem configurações dos MOOCs que favorecem os comportamentos indesejados, e faz com que determinados cursos sejam mais procurados por alunos que tendem a se comportar desse modo; sem descartar uma investigação sobre as características dos alunos, relacionadas aos dados de suas trajetórias quando realizam um MOOC no Lúmina. Nesse sentido, foi desenvolvida uma metodologia para identificação dos caçadores de certificados, com foco nessas duas perspectivas, baseada em técnicas exploratórias de MDE. Esta metodologia é descrita no próximo capítulo.

6 METODOLOGIA

Nesta seção, estão descritos os procedimentos metodológicos que deram suporte à execução deste estudo. Inicia-se realizando uma caracterização da pesquisa. Posteriormente, é apresentada uma contextualização da Tese e das hipóteses investigadas, e como pretende-se responder cada questão de pesquisa proposta. Em seguida, são expostos, com detalhes, os procedimentos que foram utilizados para caracterizar o perfil de caçadores de certificados, e é a partir deste processo que é possível responder com maior assertividade a problemática desenvolvida nesta pesquisa. Neste mesmo subcapítulo, também é realizada uma descrição da amostra utilizada, em termos de quantidades de indivíduos, MOOCs analisados e tipos de variáveis que compuseram as tabelas de dados. Por fim, realiza-se uma explicação de como foram identificados os parâmetros dos MOOCs que mais impactam na participação de alunos caçadores.

6.1 CARACTERIZAÇÃO PESQUISA

Com base no objetivo desta Tese, pode-se definir este estudo como uma pesquisa exploratória, que tem como intuito proporcionar maior familiaridade com um problema, com vistas a torná-lo mais explícito ou a constituir hipóteses (GIL, 2002). Pode-se dizer que estas pesquisas têm como objetivo principal o aprimoramento de ideias ou a descoberta de intuições, neste caso, definidas pela equipe gestora do Lúmina. Seu planejamento é, portanto, bastante flexível, de modo que possibilite a consideração dos mais variados aspectos relativos ao fato estudado. De acordo com Gil (2002, p. 41) “na maioria dos casos, essas pesquisas envolvem: levantamento bibliográfico e análise de exemplos que estimulem a compreensão”.

Quanto aos procedimentos técnicos da pesquisa, este estudo se constitui como uma pesquisa *ex-post facto* (a partir do fato passado), já que foi realizada após a ocorrência de variações na variável dependente, no curso natural dos acontecimentos (GIL, 2002). O propósito básico desta pesquisa é o mesmo da pesquisa experimental: verificar a existência de relações entre variáveis. A diferença mais importante entre as duas modalidades é que, na pesquisa *ex-post facto,* o pesquisador não dispõe de controle sobre a variável independente, que constitui o

fator presumível do fenômeno, porque ele já ocorreu (GIL, 2002). O que o pesquisador procura fazer neste tipo de pesquisa é identificar situações que se desenvolveram naturalmente e trabalhar sobre elas como se estivessem submetidas a controles. Para Gil (2002), apesar das semelhanças com a pesquisa experimental, o delineamento *ex-post facto* não garante que as conclusões relativas a relações do tipo causa-efeito sejam totalmente seguras. O que geralmente se obtém nesta modalidade é a constatação da existência de relação entre variáveis. Por isso é que essa pesquisa muitas vezes é denominada correlacional. De acordo com Gil (2002), basicamente envolve as seguintes etapas:

a) Formulação do problema – exposto no capítulo 1;
b) Construção das hipóteses – exposto no subcapítulo 6.2;
c) Operacionalização das variáveis – exposta nos subcapítulos 6.3 e 6.4;
d) Localização dos grupos para investigação – exposta nos subcapítulos 6.3 e 6.4;
e) Coleta de dados – exposta no subcapítulo 6.3;
f) Análise e interpretação dos dados – exposta no capítulo 7; e
g) Apresentação das conclusões – exposta nos capítulos 8 e 9.

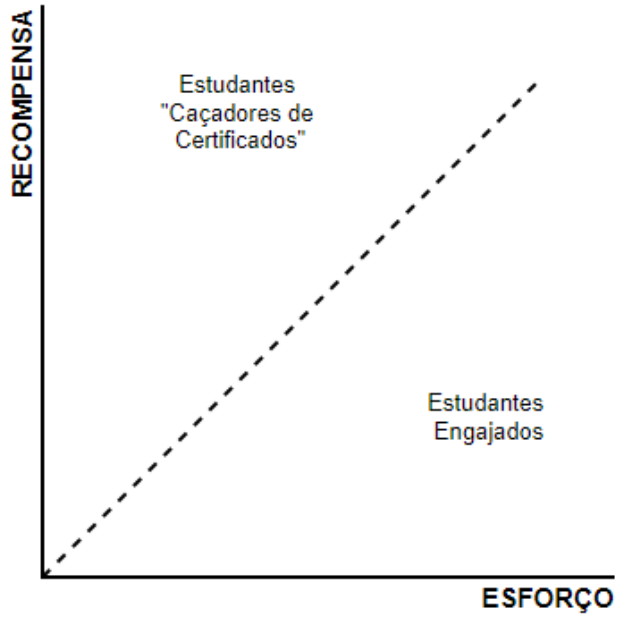
Quanto à abordagem para análise dos dados, utilizou-se principalmente a abordagem quantitativa. Uma vez que estudos com essa característica preveem a estruturação de dados em tabelas, as variáveis são quantitativas para fins de análises de correlação e/ou para testar hipóteses (GIL, 2002). Além da análise correlacional de variáveis, na dimensão quantitativa, foi empregado o processo de Mineração de Dados Educacionais, para caracterizar os perfis dos alunos caçadores de certificados e para identificar quais características dos MOOCs mais impactam na participação dos caçadores de certificados. Para isso, foram utilizadas técnicas exploratórias com algoritmos de Aprendizagem de Máquina não supervisionados e algoritmos de Aprendizagem de Máquina supervisionados, respectivamente.

6.2 CONTEXTUALIZAÇÃO DA PESQUISA

A partir da formulação do objetivo principal dessa pesquisa, considera-se que exista uma correlação entre a quantidade de esforço requerida para realização de um MOOC e a obtenção do certificado (a recompensa esperada pelo aluno). Por isso, espera-se que alunos não engajados evitem cursos que demandam altos níveis de

esforço para sua conclusão ou baixa compensação. Essa hipótese é representada na Figura 4, na qual observa-se a relação entre o esforço e a recompensa com a quantidade de alunos que têm comportamentos de caçadores de certificados, ou de engajamento. Supõe-se que alunos que não têm comprometimento com seus estudos tendem a buscar cursos que não exijam esforço, mas que apresentem uma recompensa satisfatória. Por outro lado, alunos engajados, mesmo que tenham que empenhar-se, tendem a permanecer no curso até o final, realizando todas as atividades. Por isso, na Figura 4, indica-se que quanto maior o esforço e menor a recompensa, a tendência é que permaneçam apenas os alunos engajados.

Figura 4 – Correlação entre esforço e recompensa.



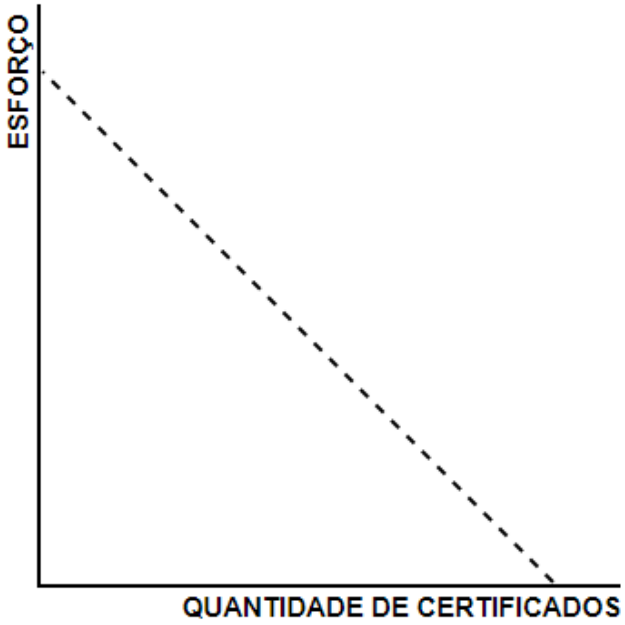
Fonte: Autora

Para tentar identificar esses caçadores de certificados, nesta Tese, foi elaborado um processo de Mineração de Dados Educacionais que pretende delinear como esses alunos se comportam, e identificar se esse comportamento tem relação com as configurações dos MOOCs analisados. A identificação dos alunos caçadores de certificado na plataforma Lúmina resultou de um processo, que se iniciou com a percepção desse comportamento pela equipe gestora da plataforma, indicando supostos alunos com uma postura inadequada. Depois disso, as trajetórias realizadas por esses estudantes quando realizam um MOOC, impressas em seus relatórios de navegação, foram observadas pela pesquisadora, que pode entender e indicar quais as principais características de alunos que se comportam desta forma. Por fim,

algumas técnicas computacionais para exploração de informações em dados foram aplicadas para aprimoramento do processo de identificação. Nesse sentido, ressalta-se que, mesmo com aplicação dessas técnicas, sobretudo por serem exploratórias, é importante o conhecimento que a pesquisadora desenvolveu a respeito da base de dados a ser analisada. Essa parte do estudo refere-se principalmente a como foi respondida a primeira questão de pesquisa.

A partir dos resultados das análises dos dados extraídos do Lúmina, e da caracterização dos caçadores de certificado, foi possível definir parâmetros de configurações para os MOOCs que reflitam um limiar, o mais próximo possível, do ideal entre a complexidade de conteúdo abordada no MOOC e as restrições de configuração, a fim de diminuir a incidência de alunos que se comportem dessa forma. Assim, espera-se que, com a aplicação de um conjunto de configurações que desestimulem os comportamentos indesejados no Lúmina, a quantidade de estudantes que evadem, sem obter certificados, aumente (pois não se espera que a quantidade de alunos persistentes aumente com a implementação destas restrições), como mostra a Figura 5. Nesse sentido, foi realizada uma sistematização de configurações que, presume-se, desencorajem alunos não comprometidos com a aprendizagem, sem desmotivar aqueles que queiram aprender com os conteúdos ofertados nos MOOCs. Essa parte do estudo está relacionada com a segunda questão de pesquisa, respondida com base nas análises dos dados, nas características encontradas em alunos caçadores de certificados e na aplicação de técnicas supervisionadas, bem como nos conhecimentos desenvolvidos sobre o que é possível ser configurado no Moodle (sistema que comporta a plataforma Lúmina) e sobre os níveis de dificuldades dos conteúdos dos MOOCs.

Figura 5 – Correlação entre esforço e quantidade de certificados



Fonte: Autora

6.3 IDENTIFICAÇÃO DOS CAÇADORES DE CERTIFICADOS

Salienta-se que um estudante pode se comportar como “caçador” em certas circunstâncias, ou seja, este rótulo não é usado para definir o comportamento daquela pessoa enquanto estudante em todas as situações. Sendo assim, a análise foi modelada de forma que seja incluída a investigação do comportamento de caçador (sem generalizar para cada estudante) e para estudantes caçadores (o que inclui todas as ações de uma pessoa específica). Isso se reflete na distinção entre “comportamento de caçador” e “estudante-caçador”.

O comportamento de caçador é uma caracterização associada a elementos de cada MOOC realizado por um aluno, com potencial de validar que existe um comportamento de caçador sem estar associado a um aluno em específico (independente do aluno), visto que um estudante pode apresentar comportamento de caçador quando realiza um determinado curso, mas em outro não. Em contrapartida, estudante-caçador está associado ao perfil dos alunos, referindo-se às suas ações na plataforma como um todo. Neste caso, o estudante se comporta seguindo o padrão de caçador de certificados em todos os MOOCs que realizar. Acredita-se que os indícios que podem identificar este comportamento e estes estudantes são: 1) Emissão de grande quantidade de certificados; 2) Pequeno intervalo entre obtenção

dos certificados; 3) Grande quantidade de cursos em que é inscrito, porém, sem realizar avaliações; e 4) Baixa visualização dos materiais instrucionais.

Com a análise da base de dados incluindo todos os alunos do Lúmina, foi possível observar que há estudantes com um número muito expressivo de certificados – os quais, quase certamente, são caçadores. Porém, não há como afirmar onde está o limite, pois para questões como “a obtenção de 10 certificados é o suficiente para qualificar um estudante como caçador?”, "a obtenção de 3 certificados em um dia qualifica o comportamento de caçador?” e “os caçadores exploram os cursos antes de decidir se vão tentar a certificação?” não há respostas definitivas. Neste ponto, é crucial compreender os limites dos relatórios de dados do Moodle, que não permitem que certas informações sejam extraídas. Na experiência alcançada com a manipulação destes relatórios, identificou-se que uma das maiores limitações é que o Moodle apenas registra o início da interação, como, por exemplo, o momento em que o estudante clica no link de um vídeo, não sendo possível saber quanto tempo do vídeo o estudante assistiu, nem quanto tempo uma página de texto foi a guia ativa (com foco) do navegador.

Sendo assim, é possível que um estudante faça um curso completo logando-se apenas uma vez, pois há cursos curtos na plataforma, o que indica que fazer um curso em dia não quer dizer, necessariamente, que o estudante é um caçador. Este estudante pode, por exemplo, usar várias guias abertas ao mesmo tempo, ou pode explorar os materiais instrucionais e depois decidir o que vai consumir. Estas informações não estão disponíveis, e considera-se que esta é uma limitação importante. Por esse motivo, acredita-se que utilizar um algoritmo não supervisionado pode mostrar como os dados podem evidenciar um padrão de navegação e melhor identificar os caçadores de certificados. Por isso, a técnica empregada para realização da MDE constitui-se por algoritmos de agrupamento, também conhecidos como algoritmos de clusterização.

Para o reconhecimento do padrão comportamental de caçadores de certificados, em meio aos estudantes usuários do Lúmina, foram utilizados os relatórios da plataforma, que contêm os seguintes dados: data de acesso a recursos (videoaulas ou materiais em texto), número de tentativas em questionários, realização de atividades avaliativas, notas, tipo de ação realizada e IP. Em linhas gerais, a representação destes perfis de alunos passou por ciclos iterativos, em que

otimizações no processo de identificação dos caçadores de certificados foram conduzidas com as seguintes etapas:

Seleção e Processamento dos dados: foram selecionados os relatórios e realizado o tratamento dessas variáveis com ações como, por exemplo: exclusão de linhas com valores ausentes, conversão de tipos, criação de novas colunas (variáveis), junção de tabelas.

Aplicação de Técnicas de Mineração de Dados Educacionais: foram utilizadas técnicas baseadas em algoritmos não supervisionados de Aprendizagem de Máquina. Os algoritmos selecionados foram os algoritmos de agrupamentos baseados nos princípios da clusterização hierárquica, também chamada de análise hierárquica de cluster. A clusterização hierárquica é um método de análise de cluster que procura construir uma hierarquia de clusters. As estratégias para o agrupamento hierárquico geralmente se enquadram em dois tipos: aglomerativos e divisivos. Os dois foram empregados nesta Tese. Na abordagem aglomerativa (de baixo para cima), cada observação começa em seu próprio cluster, e pares de clusters são mesclados à medida que se sobe na hierarquia. Na abordagem divisiva (de cima para baixo*),* todas as observações iniciam em um cluster e as divisões são executadas recursivamente à medida que se desce a hierarquia. Os algoritmos baseados nestas duas abordagens foram aplicados por meio da biblioteca cluster da linguagem de programação R. Para definir a quantidade de agrupamentos, cada um destes algoritmos foi executado cinco vezes, a fim de comparar o desempenho entre dois, três, quatro, cinco e seis grupos. Os critérios de escolha foram a quantidade de membros em cada grupo e a estabilização das medidas de Silhouette e da Soma Interna dos Quadrados (WSS, Within Sum of Squares). Silhouette é uma medida que varia entre -1 e 1, sendo que, quanto mais próxima de 1, melhor é o agrupamento, e quanto mais próxima de -1 pior. Esta medida representa as distâncias dentro dos grupos e entre os grupos. Por sua vez, WSS é a soma das distâncias entre cada ponto do grupo e a centróide do grupo, e quanto menor for esta medida, mais coeso é o grupo.

Considerações sobre o Processo: Eventuais modelos gerados podem não atender às expectativas e o processo executado pode requerer ajustes. Nesse caso, deve-se corrigir as falhas identificadas, melhorar a condução dos procedimentos e realizá-los novamente de forma otimizada.

6.3.1 Caracterização da Amostra

Para o desenvolvimento desta Tese, foram utilizados dados de todos os MOOCs da plataforma Lúmina (mesmo aqueles já encerrados), desde o seu lançamento, em setembro de 2016, até março de 2022. Foram, inicialmente, utilizados dados de cerca de 195 mil alunos, excluindo os perfis de administradores da plataforma e professores, matriculados em 103 cursos. Todavia, no decorrer do processamento destes dados, houve um refinamento, conforme os objetivos tornavam-se mais claros e o conhecimento sobre os dados aumentava. Dessa forma, a quantidade de dados foi sendo reduzida. Os dados extraídos do relatório geral do Lúmina foram processados e, a partir disso, foram criadas as seguintes tabelas de dados:

01. **Tabela comportamento nos cursos.** Permitirá a exploração do comportamento nos cursos (“comportamento de caçador”). Retrata um resumo das atividades dos alunos em cada curso, e em cada linha está representado um aluno, em um curso, com as seguintes variáveis: identificador único (chave externa[11]); média de atividades por dia; quantidade de dias ativo; quantidade de visualizações de perfil; obtenção ou não de certificado; quantidade de tarefas enviadas; quantidade de tentativas de questionário realizadas; tempo entre a primeira e a última tentativa em minutos; quantidade de postagens em fóruns; tempo entre a primeira e a última postagem em minutos; tempo total no curso; inscrição ativa ou inativa; e nome do curso (chave externa). Para categorizar uma inscrição como “inativa”, utilizou-se, como critério, um tempo de permanência no curso menor que 35 minutos e quantidade de tentativas de envio de questionário menor que duas. Considerando que vários estudantes se matricularam em mais de um curso, havia muitos nomes repetidos, e, por isso, a quantidade de linhas da tabela alcançou um número próximo a 335 mil.
02. **Tabela descrição do aluno.** Permitirá a exploração do comportamento na plataforma, de modo a determinar-se o perfil de um aluno (“estudante-caçador”). Retrata, em cada linha, o comportamento de um aluno mesmo

[11] Uma chave que permite a referência a registros oriundos de outras tabelas.

que ele tenha se inscrito em mais de um curso. Contém as seguintes variáveis: identificador único (chave externa); quantidade de certificados obtidos; tempo entre o primeiro e o último certificados (intervalo entre os certificados); quantidade de inscrições inativas; e quantidade de cursos em que está inscrito. Esta tabela é obtida após filtrar-se a quantidade de certificados para, pelo menos, um aluno, o que a caracteriza como possuindo um pouco menos de 93 mil linhas, quantidade de alunos únicos que obtiveram certificados no Lúmina.

03. **Tabela IPs dos alunos.** Sumariza a quantidade de IPs diferentes usados por cada aluno. Essa tabela foi unida às demais para realização de possíveis análises, e tem a mesma quantidade de linhas que número de alunos analisados. Contudo, ela não foi utilizada nas análises, pois verificou-se que muitos alunos têm mais de 10 IPs associados aos seus identificadores, o que denota o uso de VPNs ou atribuição dinâmica de IPs.
04. **Tabela descrição dos cursos.** Contém as somas das quantidades de cada tipo de atividade instalada no Moodle, que são: arquivos, fóruns, páginas, questionários, tarefas, *wikis*, *urls*, bases de dados, chats, diários, pastas, glossários, livros, escolha, lições, ferramentas externas, laboratórios de avaliação, elementos *bootstrap* e pacotes *scorm*. Além disso, foram agrupadas as quantidades de atividades não-avaliativas (pasta, páginas, urls, arquivos ou outras) e de atividades avaliativas (tarefa, questionário, laboratório de avaliação) em duas variáveis: uma com a soma das atividades e outra com a soma das avaliações. Ademais, foram inseridas informações sobre a área (conforme atribuída pela equipe gestora da plataforma) de cada MOOC, carga horária (conforme atribuída pelo professor do curso), se emite certificado e se possui mais de 10 questões. Os cursos também foram classificados quanto à sua configuração, que pode ser: *pouco restritiva (PR),* caracterizando-se por menos de 3 atividades avaliativas, nota final igual à nota mais alta, feedback sobre as alternativas de resposta mostrado após o envio da tentativa, 3 tentativas para realização dos questionários; *restritiva (R)*, constituída por mais de 3 atividades avaliativas, nota final igual à média das tentativas; feedback sobre as alternativas de respostas mostrado

apenas após o envio definitivo do questionário, 2 tentativas para realização dos questionários; *muito restritiva (MR)*, que caracteriza-se por configuração restritiva somada ao uso de métodos combinados de avaliação e/ou restrição de horas de permanência na plataforma. Essa tabela tem 103 linhas, sendo que cada uma representa um MOOC do Lúmina, e foi incorporada às demais para realização das análises.

Como o R não é uma linguagem que realiza processamento rápido e os processos de clusterização são computacionalmente caros, impôs-se a necessidade de reduzir a quantidade de linhas das tabelas mencionadas. Para tanto, além da seleção das colunas, visto que não é possível utilizar todas, decidiu-se por excluir os estudantes que não obtiveram certificados. Sendo assim, a quantidade foi reduzida de cerca de 335 mil para uma média de 165 mil linhas, o que corresponde à quantidade de certificados emitidos pela plataforma, pois, como já salientado, um aluno pode ter realizado mais de um curso, e obtido mais de um certificado. Por isso, a quantidade de alunos (únicos) que geraram um ou mais certificados na plataforma é de cerca 93 mil. Em seguida, para reduzir ainda mais a quantidade de linhas, a fim de verificar o desempenho de técnicas de exploração de mineração de dados, foi realizada uma partição das tabelas, sorteando aleatoriamente as linhas, sendo que, para identificação do comportamento de caçador, optou-se por utilizar 4% das linhas da tabela de dados e para a identificação dos estudantes-caçadores, decidiu-se utilizar 20% dos alunos.

6.4 IDENTIFICAÇÃO DOS PARÂMETROS DE CONFIGURAÇÃO DE MOOCS QUE IMPACTAM NA PARTICIPAÇÃO DE ALUNOS CAÇADORES

Para identificar se os fatores de configuração dos MOOCs têm influência no comportamento de caçador de certificados e estudante-caçador, foi selecionada a técnica de Regressão Logística. Após a identificação dos caçadores de certificados via exploração com algoritmos de agrupamento, foram geradas regras utilizadas para classificar todos os registros de alunos extraídos do Lúmina. Nesse sentido, foram formatadas duas tabelas: uma para regras geradas na exploração do comportamento de caçador, e outra resultando da aplicação das regras obtidas com a identificação dos estudantes-caçadores. Após, a técnica de Regressão Logística foi empregada para analisar quais variáveis pertencentes aos MOOCs mais impactam na

participação dos caçadores de certificados. Como essa técnica exige menos recursos computacionais, foi possível utilizar todos os registros que atendessem às restrições delimitadas neste estudo, em que foram empregados apenas os dados nos quais houve a geração de certificados.

Visto que a variável fim ou variável meta (variável dependente – é ou não caçador) é binária e categórica, a Regressão Logística torna-se uma técnica bastante indicada para analisar se o conjunto das variáveis preditoras é suficiente para prever a variável meta. Em comparação com as técnicas conhecidas de regressão, sobretudo a Regressão Linear, a Regressão Logística distingue-se pelo fato de a variável meta ser categórica. Enquanto método de predição para variáveis categóricas, a Regressão Logística é comparável às técnicas supervisionadas de Aprendizagem de Máquina (como: Árvores de Decisão, Redes Neurais, *Randon Forest*, *Suport Vector Machines*, dentre outras), podendo ser aplicada como uma técnica de MDE, ou, ainda, a análise discriminante preditiva em estatística exploratória. Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binominalmente distribuídas, sendo útil para modelar a probabilidade da ocorrência de um evento em função de outros fatores, como é o intuito desta Tese.

Um dos fatores que levaram à utilização da Regressão Logística, foi o fato de que esta técnica possibilita colocar as variáveis independentes em concorrência para escolha do modelo mais adaptado para um problema preditivo. Nesse sentido, para este estudo, foi utilizada a Regressão Logística *Stepwise*, que consiste em uma forma de aplicação da regressão usada nos estágios exploratórios da construção de modelos, para identificar um subconjunto útil de preditores. Em suma, é o processo de selecionar automaticamente um número reduzido de variáveis preditoras para construir o modelo de Regressão Logística de melhor desempenho. O processo adiciona sistematicamente a variável mais significativa ou remove a variável menos significativa durante cada etapa. Para tal, foi empregada a reamostragem com *Bootstrap*, também utilizada para garantir a consistência dos preditores. Neste processo, utiliza-se uma amostragem aleatória com substituição. Dessa forma, foi possível indicar com mais exatidão quais aspectos de um MOOC do Lúmina propiciam a ação de caçadores de certificados e quais inibem sua participação.

6.5 ESPECIFICAÇÃO DAS ETAPAS DE ANÁLISES DE DADOS

De forma geral, a metodologia utilizada nesta Tese, no que se refere à análise de dados e à aplicação das técnicas de MDE, foi realizada em três etapas pré-definidas:

1) Processamento dos dados: diz respeito ao processamento de todas as bases de dados do Lúmina em tabelas específicas (descritas na seção 6.3.1) e à análise por meio de estatísticas descritivas.

2) Identificação dos Caçadores de Certificados: refere-se ao desenvolvimento do processo de agrupamento (descrito na subcapítulo 6.3). A partir deste processo, algumas tabelas foram unidas e sintetizadas para que os algoritmos de agrupamento pudessem ser aplicados. Com isso, foi possível gerar regras para classificar todos os dados das tabelas, considerando dois aspectos: o comportamento de caçador e os estudantes-caçadores.

3) *Identificação dos Parâmetros de Configuração:* corresponde à aplicação do algoritmo de Regressão Logística (descrito no subcapítulo 6.4), a partir dos dados classificados, por meio das regras definidas na etapa 2. Com o desenvolvimento desta etapa, foi possível perceber quais fatores impactam na participação dos alunos com diferentes comportamentos e perfis nos MOOCs do Lúmina, também considerando aspectos relacionados ao comportamento de caçador e dos estudantes-caçadores.

Algumas dessas etapas possuem divisões em subetapas e tarefas/atividades vinculadas, as quais foram chamadas de passos. Além disso, todas as etapas geraram artefatos, utilizados nas etapas seguintes e/ou descritos como resultados desta Tese. Uma descrição detalhada dessas informações pode ser observada no Quadro 6.

Quadro 6 – Especificação das Etapas das Análises de Dados

Continua

Etapas/Sub Etapas		Passos	Artefatos Gerados
Etapa 1 – Processamento dos Dados		Passo 1 – Extração dos dados do Lúmina originados dos relatórios de Logs, Progresso e Notas	Tabelas de Dados (01, 02, 03 e 04)
		Passo 2 – Criação de variáveis: originadas de análises nos MOOCs e da combinação de variáveis já existentes	
		Passo 3 – Separação dos dados em tabelas: 01. Tabela comportamento nos cursos, 0.2 Tabela descrição do aluno, 0.3 Tabela IPs dos alunos, 0.4 Tabela descrição dos cursos	
		Passo 4 – Aplicação de métodos para geração de estatísticas descritivas	
Etapa 2 – Identificação dos Caçadores de Certificados	**2.1 Identificação do Comportamento de Caçador**	Passo 1 – União das Tabelas 01. Tabela comportamento nos cursos" e "04. Tabela descrição dos cursos, em uma tabela final	Regras para Classificar os dados
		Passo 2 – Seleção das variáveis mais significativas	
		Passo 3 – Seleção aleatória de 4% dos dados constantes na tabela	
		Passo 4 – Aplicação de dois algoritmos de agrupamento	
		Passo 5 – Testes com diferentes números de grupos para os dois algoritmos	
		Passo 6 – Aplicação de métricas para verificação do melhor número de grupos	
		Passo 7 – Escolha do melhor algoritmo de agrupamento e do melhor número de grupos	
		Passo 8 – Análise e descrição dos grupos gerados pelo algoritmo	
		Passo 9 – Definição de regras para classificação dos dados com base nos grupos gerados	
	2.2 Identificação dos Estudantes Caçadores	Passo 1 – Seleção das variáveis mais significativas da tabela 02. Tabela descrição do aluno	Regras para Classificar os dados
		Passo 2 – Seleção aleatória de 20% dos dados dos alunos constantes na tabela	
		Passo 3 – Aplicação de dois algoritmos de agrupamento	
		Passo 4 – Testes com diferentes números de grupos para os dois algoritmos	
		Passo 5 – Aplicação de métricas para verificação do melhor número de grupos	
		Passo 6 – Escolha do melhor algoritmo de agrupamento e do melhor número de grupos	
		Passo 7 – Análise e descrição dos grupos gerados pelo algoritmo	
		Passo 8 – Definição de regras para classificação dos dados com base nos grupos gerados	
Etapa 3 – Identificação dos Parâmetros de Configuração	**3.1 Regressão Logística na classificação do Comportamento de Caçador**	Passo 1– Junção das três tabelas principais deste estudo (01. Tabela comportamento nos cursos, 0.2 Tabela descrição do aluno e 0.4 Tabela descrição dos cursos)	1. Descrição dos Parâmetros que influenciam Caçadores 2. Tabela de Dados Categorizada
		Passo 2 – Classificação dos Dados com as regras definidas no processo de identificação do comportamento de caçador	
		Passo 3 – Aplicação de métodos associados a Regressão Logística, para a verificação de quais variáveis são mais significativas na predição do atributo meta	
		Passo 4 – Sintetização da tabela de dados com base nos resultados identificados	
		Passo 5 – Aplicação da Regressão Logística na Tabela de dados sintetizada	
		Passo 6 – Realização de Análise multivariada de Regressão Logística para identificar quais variáveis da tabela mais impactam nos comportamentos de caçadores	
		Passo 7 – Descrição destas variáveis e seu nível de correlação com os comportamentos de caçadores	

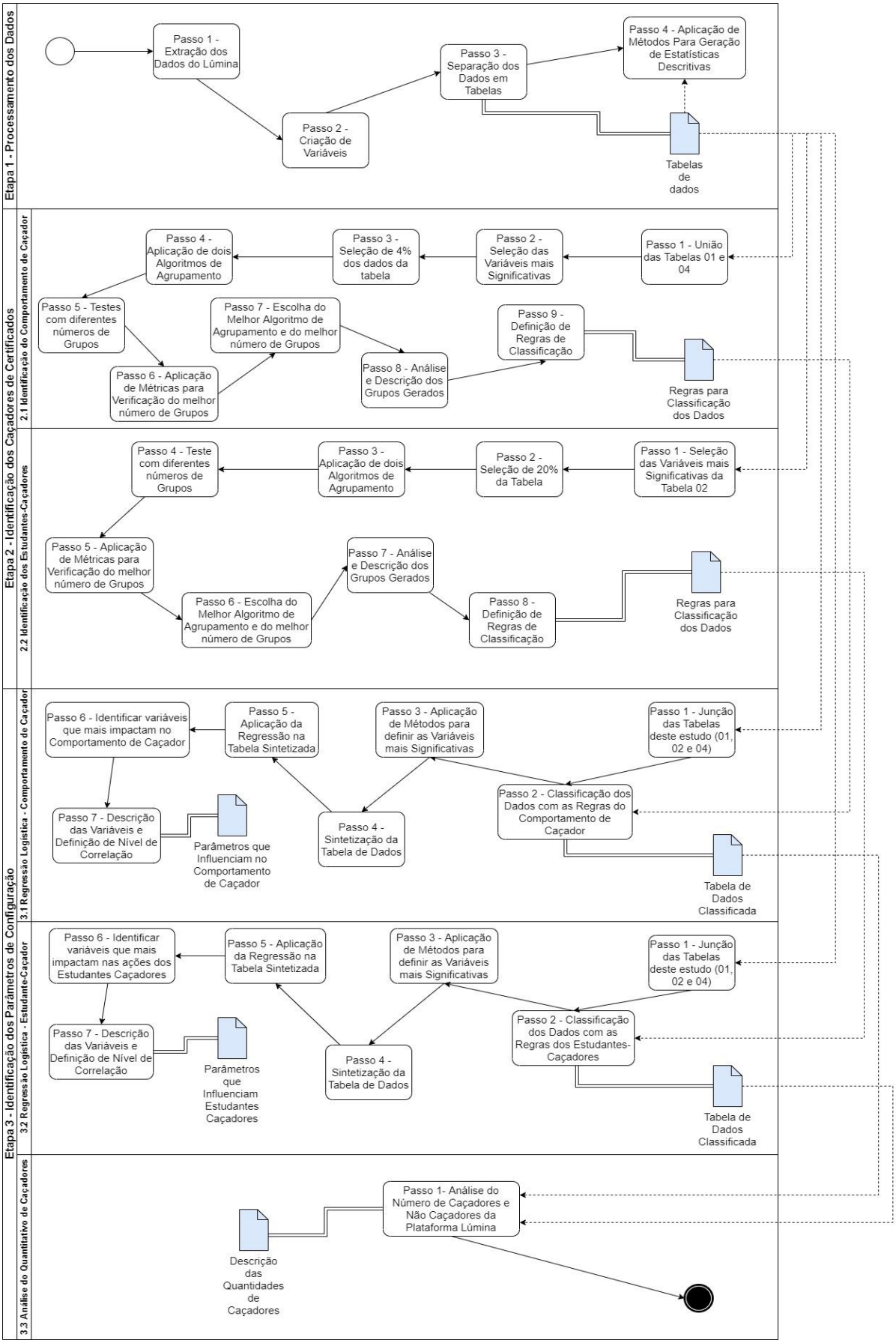
Conclusão

	3.2 Regressão Logística na classificação do Estudante-Caçador	Passo 1 – Junção das três tabelas principais deste estudo (01. Tabela comportamento nos cursos, 0.2 Tabela descrição do aluno e 0.4 Tabela descrição dos cursos)	1. Descrição dos Parâmetros que influenciam Caçadores 2. Tabela de Dados Categorizada
		Passo 2 – Classificação dos Dados com as regras definidas no processo de identificação dos estudantes-caçadores	
		Passo 3 – Aplicação de métodos associados a Regressão Logística, para a verificação de quais variáveis são mais significativas na predição do atributo meta	
		Passo 4 – Sintetização da tabela de dados com base nos resultados identificados	
		Passo 5 – Aplicação da Regressão Logística na Tabela de dados sintetizada	
		Passo 6 – Realização de Análise multivariada de Regressão Logística para identificar quais variáveis da tabela mais impactam nas ações dos estudantes-caçadores	
		Passo 7 – Descrição destas variáveis e seu nível de correlação com as ações dos estudantes-caçadores	
	3.3 Analise do quantitativo de caçadores	Passo 1 – Aplicação de métodos descritivos para análises dos quantitativos de caçadores e não caçadores vinculados a plataforma	Descrição da quantidade de Caçadores

Fonte: Autora

Para melhorar a visualização de como as etapas e os passos que as compõe estão interligados, bem como entender para quais fins foram aplicados os artefatos gerados no decorrer do processo, foi elaborado um Diagrama (Figura 6) que apresenta toda especificação da análise de dados implementada nesta Tese.

Figura 6 – Especificação das Etapas das Análises de Dados



Fonte: Autora

6.5 PROTEÇÃO DOS DADOS

As variáveis presentes nas tabelas que permitiam algum tipo de identificação foram anonimizadas, de forma randômica sendo substituídas por sequências alfa-numéricas, que impossibilitam distinguir os participantes. Destaca-se ainda, que o projeto desta Tese está cadastrado na plataforma Brasil com Certificado de Apresentação de Apreciação Ética número: 45733221.6.0000.5347; e foi aprovado segundo o Parecer Consubstanciado do Comitê de Ética em Pesquisa da UFRGS, de número: 4.709.452.

7 RESULTADOS

Neste capítulo, são descritos os resultados e as análises elaboradas a partir da aplicação dos algoritmos de agrupamento utilizados e de Regressão Logística. Como salientado, no que tange à aplicação dos algoritmos de agrupamento, tal processo foi conduzido, por meio de um processo iterativo, o qual possibilitou chegar à forma que estabeleceu as melhores respostas às perguntas levantadas nesta pesquisa. No que se refere à aplicação da regressão, foi utilizado o método de Regressão Lógística *Stepwise,* que possibilita identificar o melhor conjunto de variáveis preditoras para a variável meta. Primeiramente, considerou-se importante apresentar algumas estatísticas descritivas, para um melhor entendimento sobre a amostra estudada. Na sequência, são apresentadas as considerações sobre a caracterização do comportamento de caçador e, em seguida, são apontados os resultados sobre a identificação dos estudantes-caçadores. Por fim, os resultados da análise de regressão são expostos, primeiramente abordando uma base de dados classificada com as regras identificadas no modelo comportamento de caçador, e, posteriormente, empregando uma base de dados categorizada a partir das regras identificadas no modelo estudante-caçador. Dessa forma, torna-se possível perceber se as configurações dos MOOCs impactam sobre as ações dos caçadores de certificados.

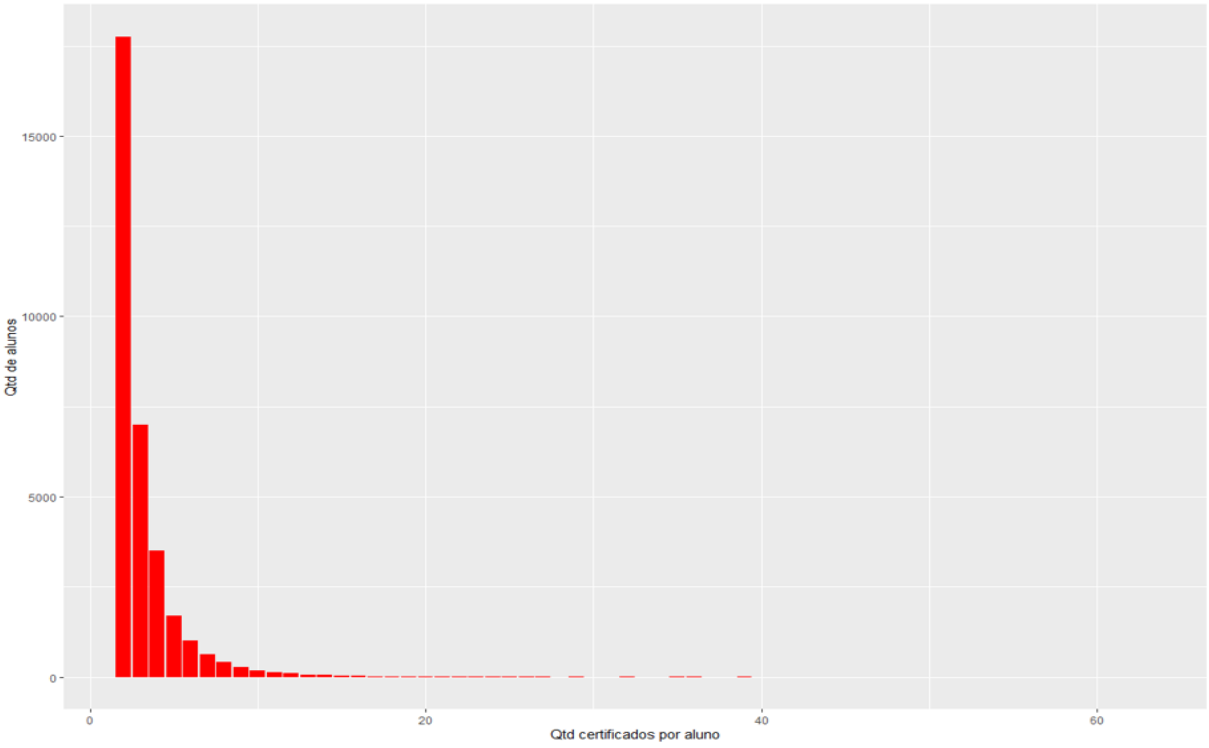
7.1 ESTATÍSTICAS DESCRITIVAS

Antes de apresentar o resultado das demais análises, julgou-se importante destacar algumas estatísticas descritivas sobre a amostra analisada, processada a partir do relatório geral do Lúmina com as informações de todos os alunos. Esta tarefa faz parte da etapa 1 das análises dos dados (Figura 6). Assim sendo, neste subcapítulo, são detalhadas as distribuições de certificados por aluno, os intervalos entre a primeira e a última certificação, a distribuição das inscrições nos MOOCs, as quantidades de atividades realizadas por aluno em cada MOOC, a quantidade de dias ativos por aluno e a quantidade de inscrições inativas por aluno. Além destes pontos, são apresentadas também algumas informações sobre os MOOCs que compuseram este estudo, especialmente sobre a área, a carga horária, a quantidade de atividades

e a quantidade de avaliações. O motivo da apresentação está na constatação de que, por mais sofisticadas que sejam as técnicas exploratórias, estas devem ser abastecidas com dados que representem o contexto da análise e permitam interpretação por parte dos especialistas.

A primeira análise realizada diz respeito às quantidades de certificados emitidos pelos alunos. Identificou-se, por esta análise, que cerca de 102,6 mil estudantes não possuem certificados, o que representa aproximadamente 52% da amostra. Enquanto isso, cerca de 77,7 mil possuem 1 ou 2, em torno de 12 mil alunos possuem de 3 a 5, cerca de 2,5 mil têm entre 6 e 10 certificados e 547 estudantes possuem mais de 10. Há, ainda, alguns casos extremos, como estudantes com 63 (1 caso), 52 (1 caso) e 42 (1 caso) certificados. Uma representação dessas informações é mostrada na Figura 7.

Figura 7 – Distribuição da quantidade de certificados por aluno

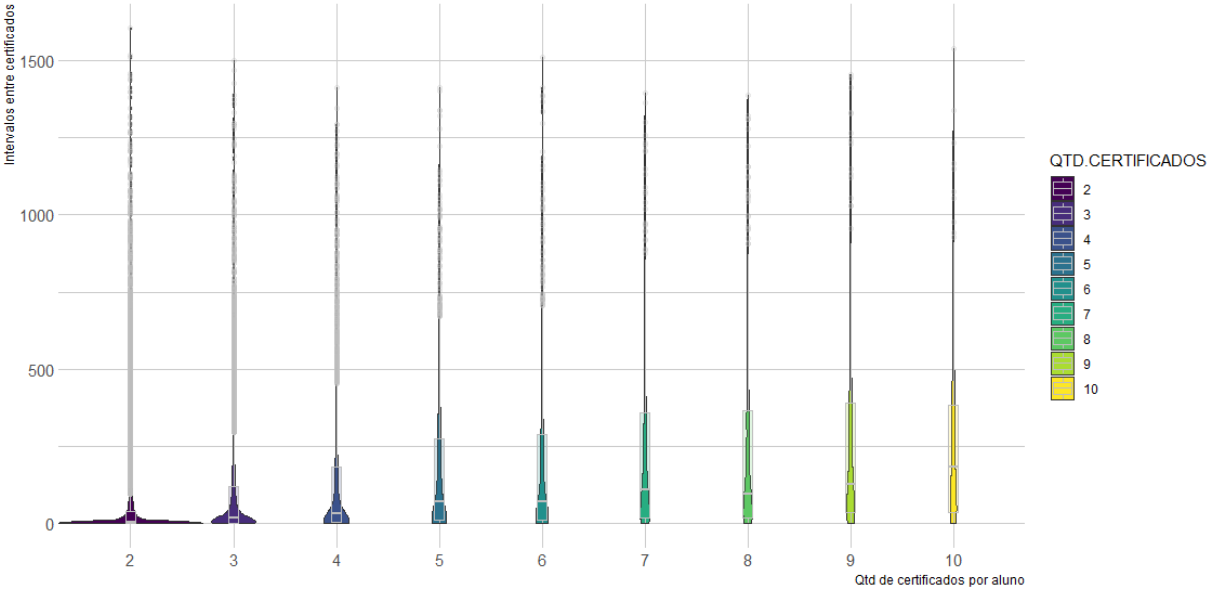


Fonte: Autora

Uma informação também relevante é o intervalo entre o primeiro e o último certificado. Essa informação foi extraída dentre os estudantes que obtiveram pelo menos 2 certificados. Em torno de 33 mil estudantes pertencentes à amostra atendem essa restrição. Dentre estes, cerca de 7 mil têm um intervalo igual a 0 (zero) dias, em torno de 4 mil, um intervalo de 1 ou 2 dias, 2,5 mil, um intervalo entre 3 e 5 dias e 2,6 mil, um intervalo entre 6 e 10 dias. Pouco mais de 17 mil estudantes têm intervalos

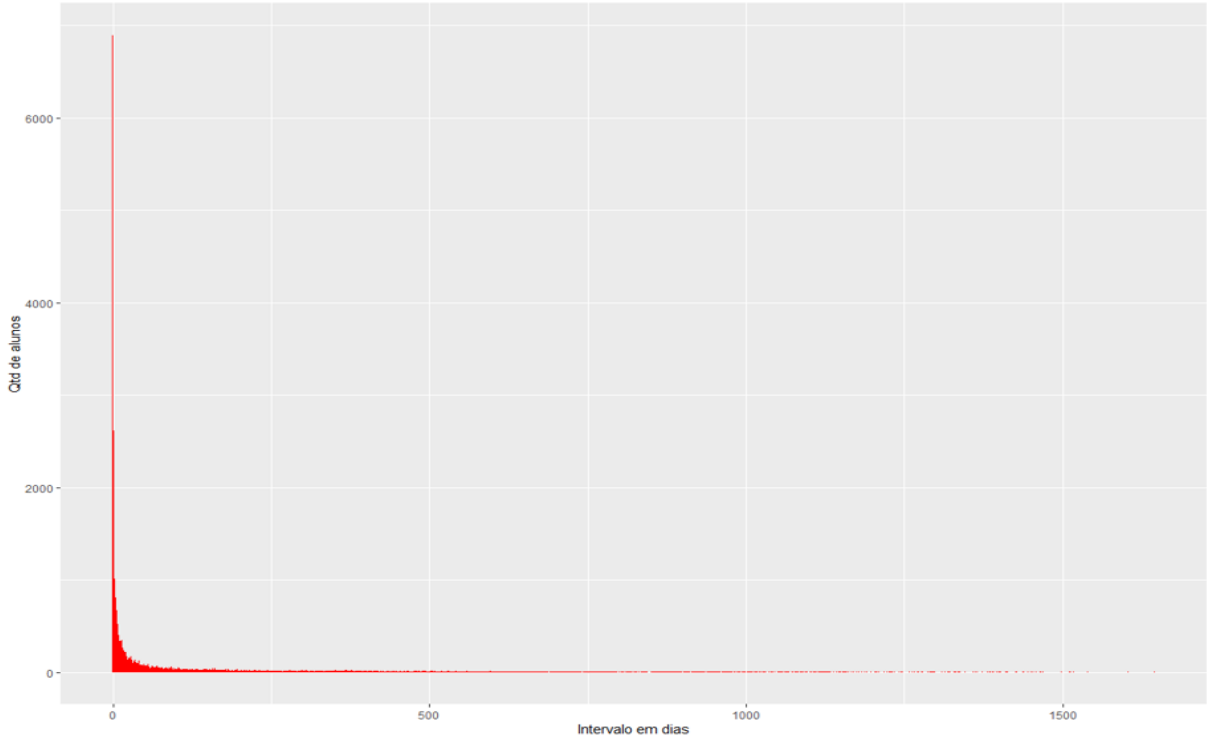
maiores do que 10 dias. Acredita-se que as informações sobres os intervalos sejam relevantes e, por isso, foi explorada a relação entre quantidade de certificados e intervalos, mostrada na Figura 8, que traz o gráfico de caixa sobreposto por um gráfico de violino, o qual permite visualizar a forma da distribuição. Além disso, na Figura 9, pode ser visualizada a distribuição destes intervalos em dias pela quantidade de alunos.

Figura 8 – Relação entre quantidade de certificados e intervalos de tempo



Fonte: Autora

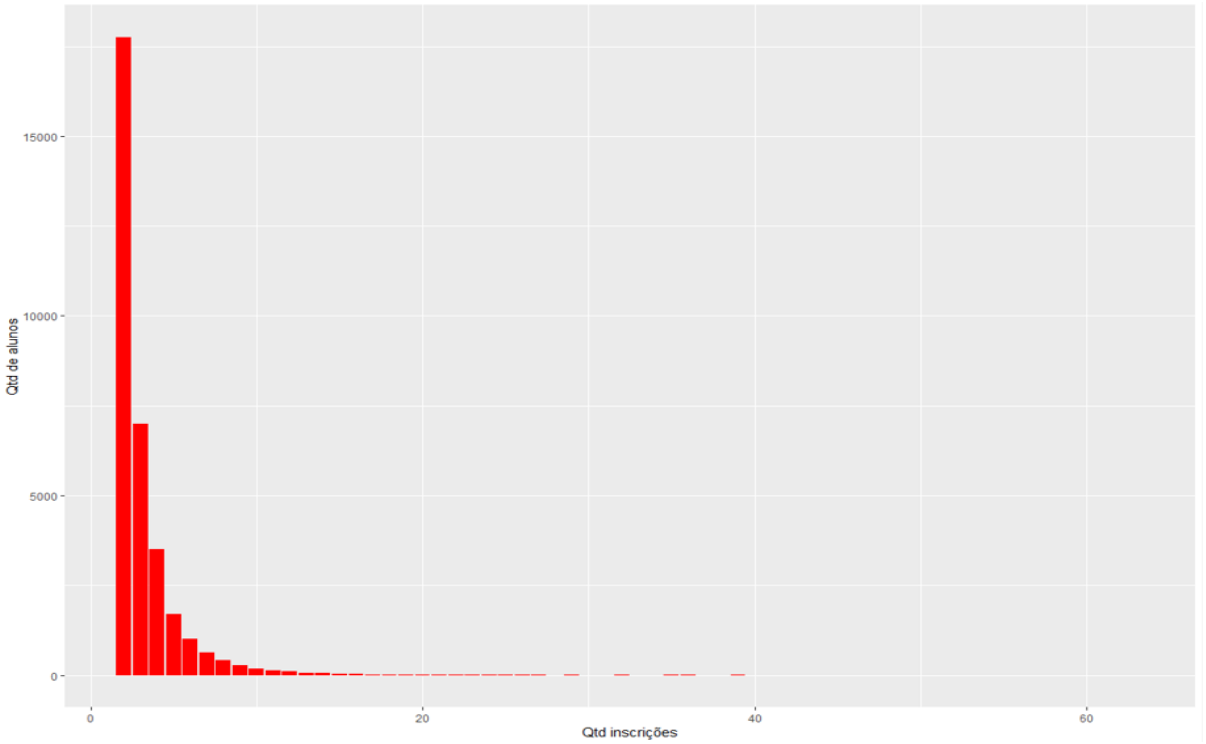
Figura 9 – Distribuição dos intervalos tempo entre o primeiro e o último certificado por aluno



Fonte: Autora

Outra informação importante é a distribuição das inscrições nos MOOCs, em que se levou em consideração todos os alunos que se matricularam em algum curso, independentemente de o terem concluído. A grande maioria dos alunos (cerca de 167 mil) se matriculou em 1 ou 2 MOOCs, em torno de 22 mil se matricularam entre 3 e 5, cerca de 5 mil alunos se matricularam entre 6 e 10 e em 10 MOOCs ou mais, uma média de 1,2 mil alunos. Há, ainda, alguns casos extremos, em que se observa alunos inscritos em 71 (1 caso) MOOCs, 70 (1 caso), 55 (1 caso) e 53 (1 caso). Na figura 10, pode ser observada a distribuição de inscrições pela quantidade de alunos. Destaca-se também que o MOOC com a maior quantidade de inscrições possuía em torno de 13,6 mil inscritos, o segundo, cerca de 5 mil inscritos e o terceiro, uma média de 2 mil inscrições. Dos 10 MOOCs com maior quantidade de inscritos, 4 são da área de Ciências da Saúde e Biológicas, 2 são da área de Tecnológicas, 2 são da área de Ciências Humanas e Sociais e 2 são da área de Ciências Exatas e da Terra.

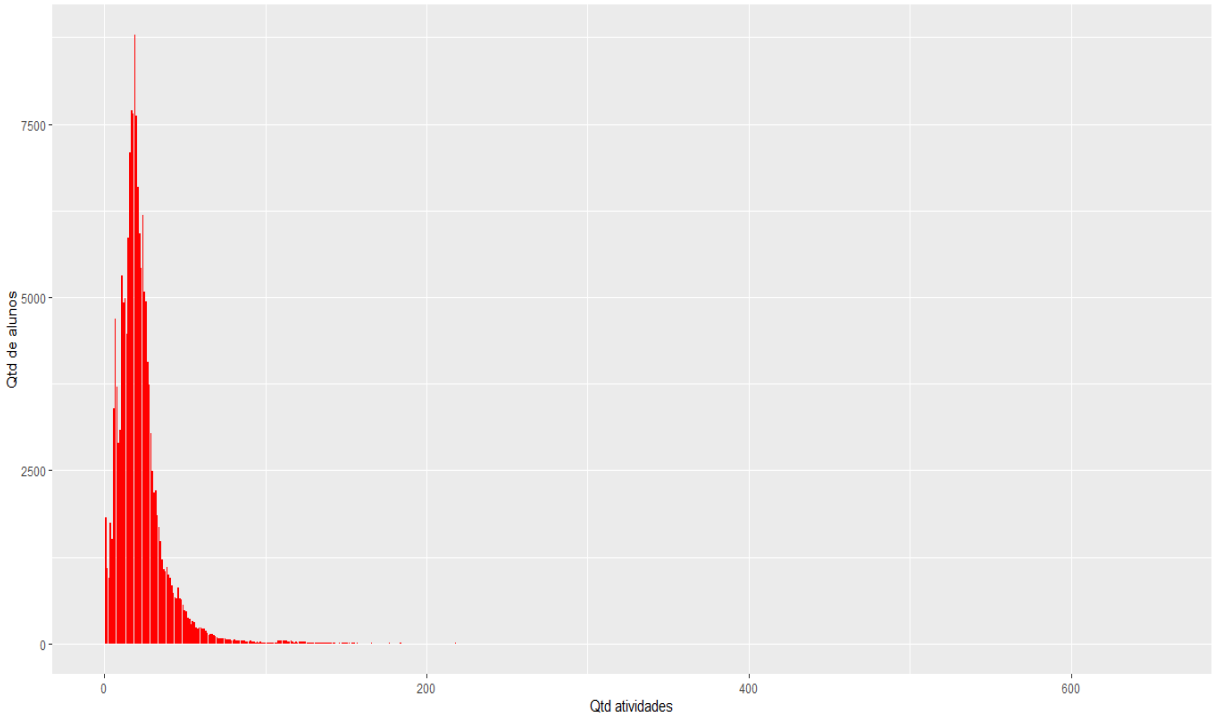
Figura 10 – Distribuição da quantidade inscrições por aluno



Fonte: Autora

Ademais, foram sistematizadas as quantidades de atividades realizadas pelos alunos no decorrer da realização dos MOOCs. Para isso, levou-se em conta apenas os registros de alunos que concluíram os cursos: aproximadamente 165 mil registros. Optou-se por realizar esta restrição, pois aqueles que não completaram geralmente apresentam a realização de poucas atividades, influenciando nos valores da distribuição. A média geral da quantidade de atividades realizadas por um aluno foi cerca de 22 atividades por MOOC. De forma mais detalhada, pode-se observar que cerca de 25 mil alunos realizaram entre 1 ou 10 atividades por MOOC, em torno de 64 mil alunos fizeram entre 11 e 20 atividades, cerca de 47 mil alunos efetuaram entre 21 e 30 atividades e aproximadamente 15 mil alunos fizeram entre 31 e 40 atividades por MOOC. Além destes, foram contabilizados 14 mil alunos que realizaram mais de 40 atividades por MOOC, e o valores máximos registrados foram 654 (1 caso), 608 (1 caso) e 418 (1 caso). Essa distribuição é destacada na Figura 11.

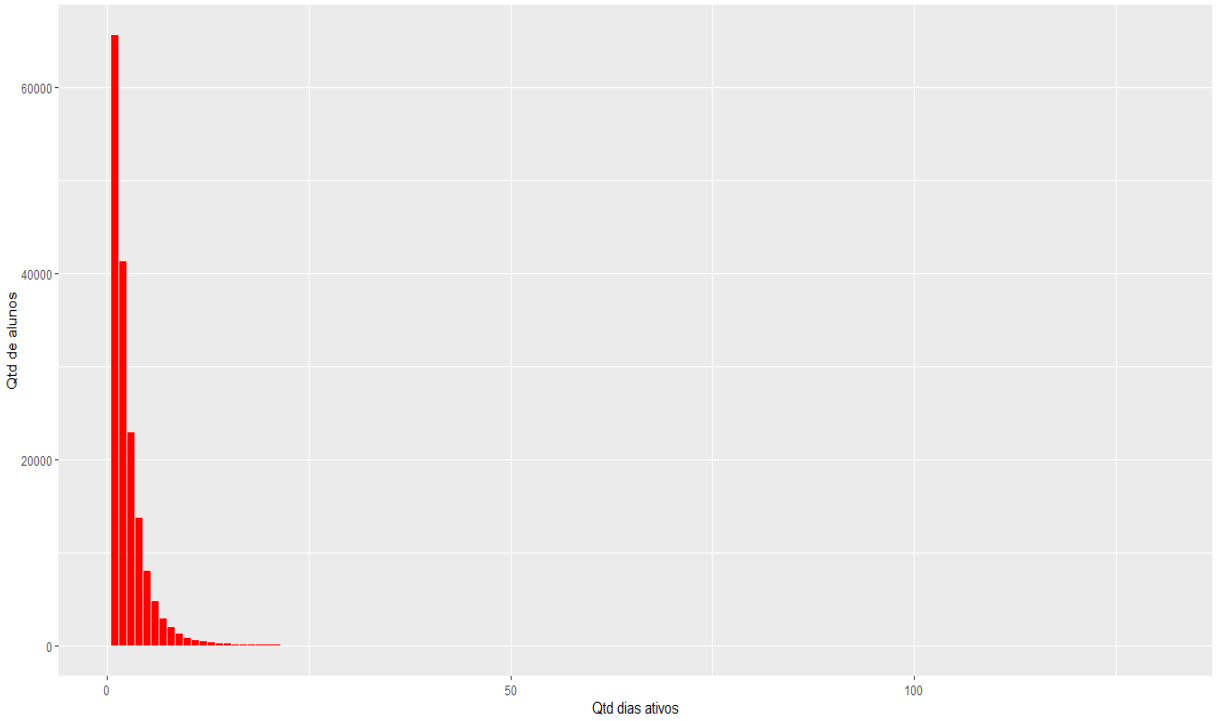
Figura 11 – Distribuição da quantidade atividades realizadas por aluno em cada MOOC



Fonte: Autora

Outra informação significativa é a quantidade de dias ativos dos alunos ao realizar um curso, isto é, os dias em que o aluno realmente se logou na plataforma, desde o início até a finalização do MOOC. A média geral da quantidade de dias ativos em que os alunos permanecem em um MOOC é de aproximadamente 2,6 dias. A maioria dos alunos (aproximadamente 107 mil) se concentra entre 1 e 2 dias ativos, em torno de 44 mil têm uma permanência entre 3 e 5 dias, cerca de 12 mil alunos realizaram seus MOOCs entre 6 e 10 dias e, em um período superior a 10 dias, contabilizou-se cerca de 2 mil alunos. Os maiores valores registrados foram: 130 dias (1 caso) e 71 dias (1 caso). A distribuição dos dias ativos, pela quantidade de alunos, pode ser visualizada na Figura 12.

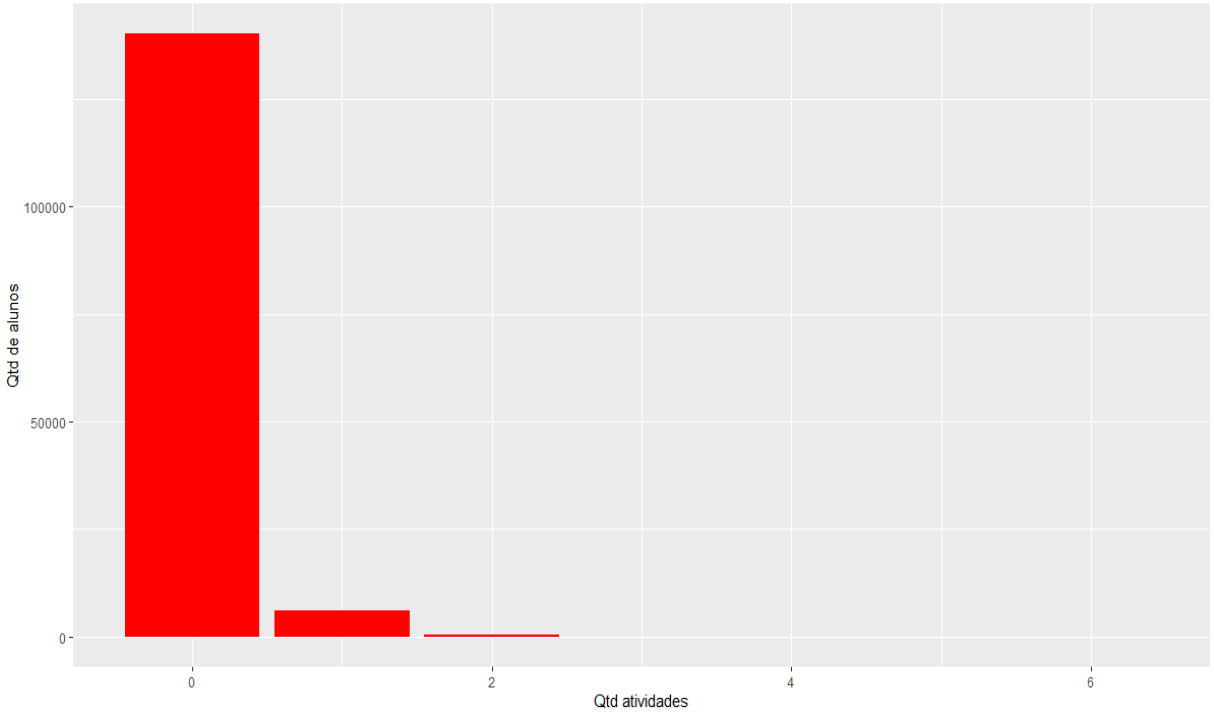
Figura 12 – Distribuição da quantidade de dias ativos por aluno em cada MOOC



Fonte: Autora

Por fim, uma última informação sobre os alunos, considerando todos os estudantes usuários que se matricularam em algum MOOC do Lúmina, é a quantidade de inscrições inativas. Sobre este dado, a maioria dos alunos (cerca de 189 mil, o que representa aproximadamente 96%) não tem nenhuma inscrição inativa, em torno de 6 mil têm uma1 inscrição inativa, e o restante tem de 2 a 6 inscrições inativas. Estes valores podem ser observados com mais detalhes na Figura 13.

Figura 13 – Distribuição da quantidade de inscrições inativas por aluno

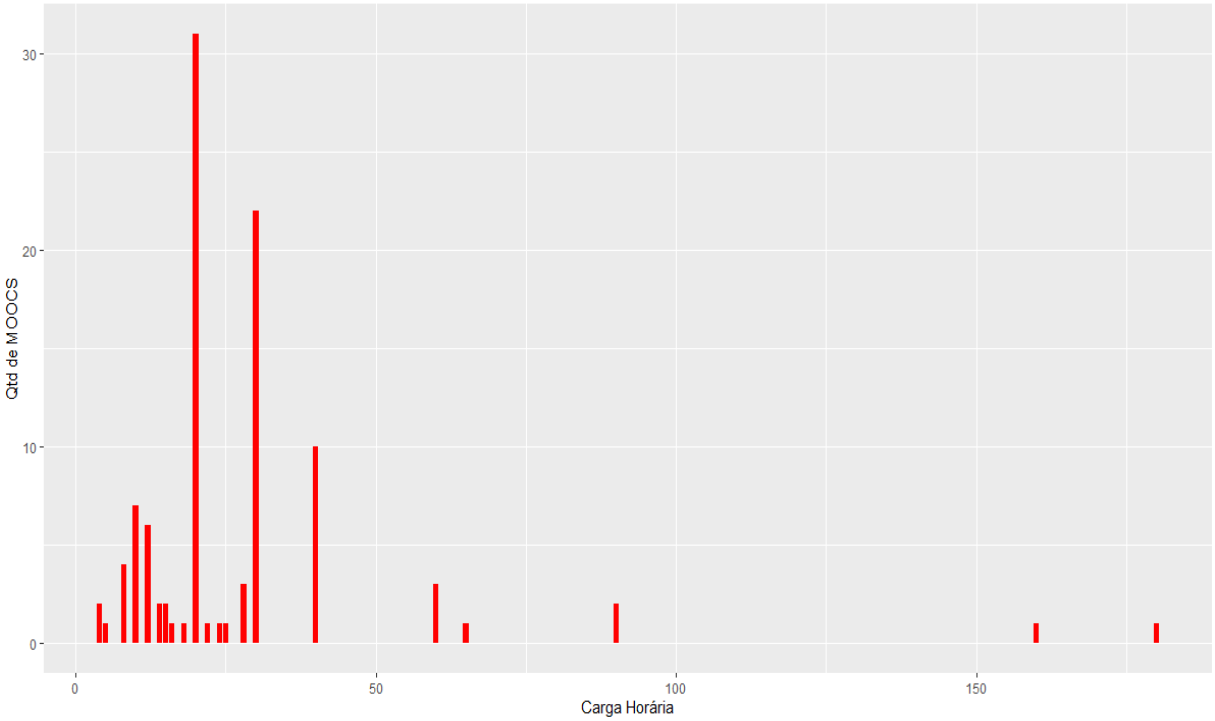


Fonte: Autora

Além das informações sobre os alunos, acredita-se ser importante apresentar algumas características dos MOOCs do Lúmina, considerando aqueles que efetivamente fizeram parte deste estudo. Os MOOCs analisados tem temáticas bem variadas e há uma boa representação de cada área da plataforma, sendo Ciências da Saúde e Biológicas (33) com a maior quantidade de MOOCs, seguida pela área de Tecnológicas (28). Na sequência, aparece Linguística, Letras e Artes (16), Ciências Humanas e Sociais (14) e por fim, Ciências Exatas e da Terra (12).

Uma informação bastante relevante refere-se às cargas horárias dos MOOCs, pois influenciam diretamente no tempo que um aluno deve dedicar para a conclusão do mesmo. Conforme foi observado, as cargas horárias dos MOOCs analisados vão de 4 a 180 horas. A maior parte deles se concentra em 20 e 30 horas (31 e 22 MOOCs respectivamente) contabilizando um pouco mais de 50% dos cursos. Há 2 MOOCs com carga horária de 4 horas, 24 MOOCs com cargas horárias que estão entre 5 e 18 e 18 MOOCs com carga horária acima de 30 horas. Dentre estes, destacam-se 2 com cargas horárias bem superiores à média: 160 e 180 horas. Uma melhor visualização desta distribuição é apresentada na Figura 14.

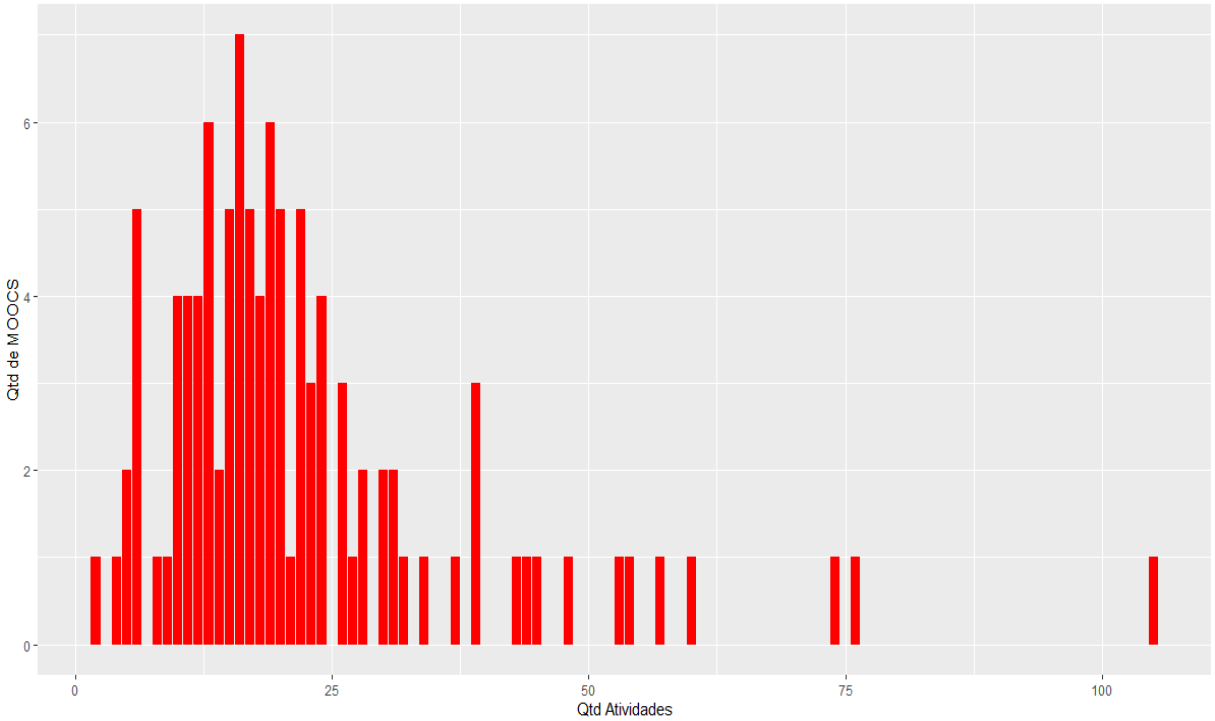
Figura 14 – Distribuição da quantidade de MOOCs por carga horária



Fonte: Autora

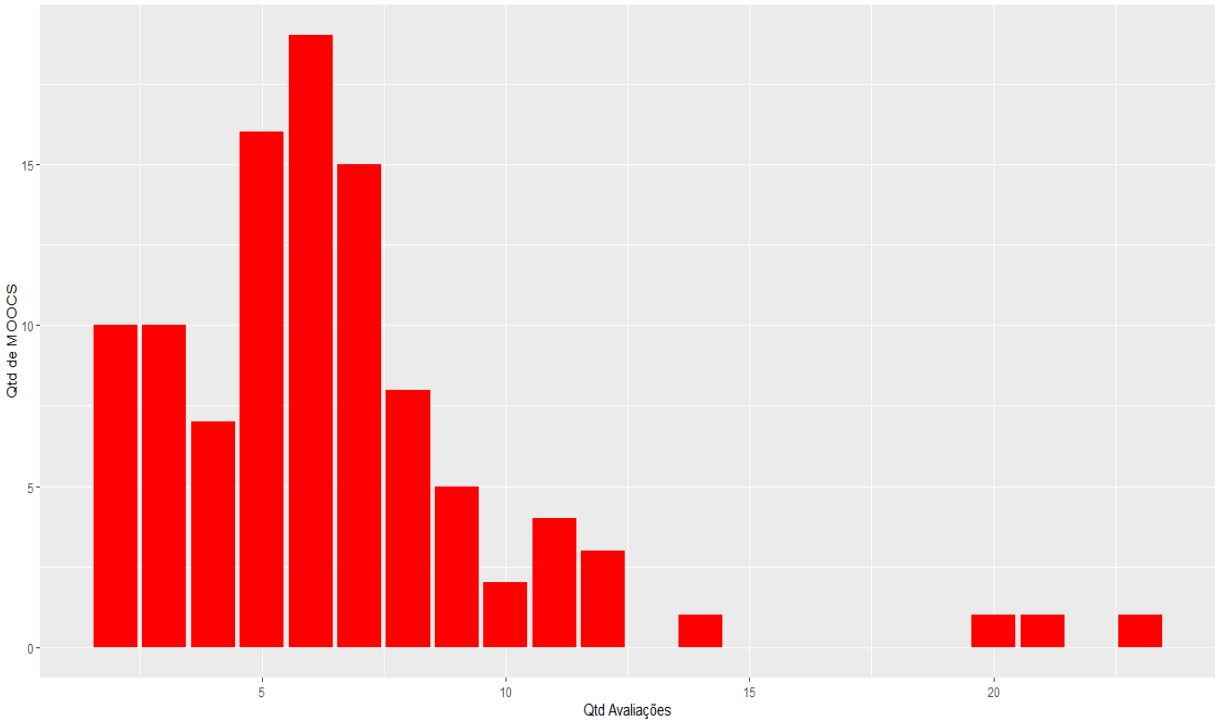
Com relação aos MOOCs, a quantidade de atividades — que aqui se refere a todos os tipos de recursos que podem haver no Moodle (descritos na tabela 04. Tabela descrição dos cursos) — também é uma característica importante para verificar o quanto um aluno está engajado na sua realização, bem como a quantidade de avaliações, que, nos MOOCs do Lúmina, são principalmente apresentadas no formato de questionários de múltipla escolha. As distribuições referentes à quantidade de atividades por MOOC e à quantidade de avaliações podem ser analisadas nas Figuras 15 e 16, respectivamente. A média de atividades por curso contabilizou-se como em torno de 22, e aproximadamente 50% dos MOOCs possuem entre 15 e 30 atividades. A média de avaliações por MOOC é cerca de 6, e a maioria dos cursos, 63%, possui entre 5 e 10 avaliações. Ainda sobre a quantidade de atividades/avaliações dos MOOCs, foi contabilizado que 74 MOOCs (72%) possuem mais de 10 questões.

Figura 15 – Distribuição dos MOOCs por quantidade de atividades



Fonte: Autora

Figura 16 – Distribuição dos MOOCs por quantidade de avaliações



Fonte: Autora

Após essa descrição sobre a amostra utilizada, são apresentados os resultados das demais análises de dados realizadas.

7.2 IDENTIFICAÇÃO DOS CAÇADORES DE CERTIFICADOS

A análise exploratória realizada para identificação dos caçadores de certificados corresponde à etapa 2 do processo de análise de dados desenvolvido nesta tese, para a qual foram aplicados algoritmos de cluster hierárquico de dois tipos: divisivo e aglomerativo. Como partiu-se do princípio de que pode haver um "comportamento de caçador" (independente do aluno) e um "estudante caçador" (que sempre exibe este comportamento), os resultados das análises são apresentados separadamente, nas próximas seções (estas correspondem as sub etapas 2.1 e 2.2 – Figura 6).

7.2.1 Identificação do Comportamento de Caçador

Por meio da fusão entre as tabelas "01. Tabela comportamento nos cursos" e "04. Tabela descrição dos cursos", foi criada uma tabela com quatro variáveis: se o curso tem mais de 10 questões, quantidade de dias ativo, persistência nas atividades e persistência nos questionários. As duas últimas colunas foram obtidas por meio do cálculo da razão entre a quantidade de atividades/questionários realizados e a soma de atividades/questionários do curso, para que se possa comparar os cursos. Embora tenha sido bastante demorado o processo de extração de todos os dados presentes nas tabelas (comportamentos nos cursos e descrição dos cursos), descobriu-se, por meio de testes, que não seria possível aplicar os algoritmos de cluster sobre todas essas variáveis, pois o tempo de processamento requerido e a capacidade computacional necessária estão além do que se tem à disposição. Por isso, tomou-se a decisão de utilizar apenas as variáveis que se julgou potencialmente mais relevantes para identificação do comportamento de caçadores.

Além da redução das variáveis, foi necessária uma redução da quantidade de linhas da tabela, visto que esta contém pouco mais de 165 mil linhas, uma para cada registro de aluno que gerou um certificado, ou seja, cada estudante pode estar representado por mais de uma linha. Como esta é uma quantidade muito grande para executar o algoritmo de agrupamento, devido às restrições já destacadas, foi utilizada uma partição de 4% destas linhas, chegando a, mais ou menos, 6.282 registros, escolhidos de forma aleatória.

Após a formatação desta tabela, foram feitos testes com diferentes números de agrupamentos (2, 3, 4, 5 e 6), com dois tipos de algoritmos de cluster hierárquico: divisivo e aglomerativo. Depois destes testes, decidiu-se que o algoritmo de cluster divisivo teve um melhor desempenho sobre a base de dados utilizada. Para cada teste, foram calculadas as métricas Silhouette e WSS. Como indicado na Tabela 2, as melhores soluções contêm 2 ou 3 grupos (embora as soluções com 4 e 5 grupos também tenham um valor alto) e, para a métrica WSS, entre 3 e 6 agrupamentos possuem valores próximos.

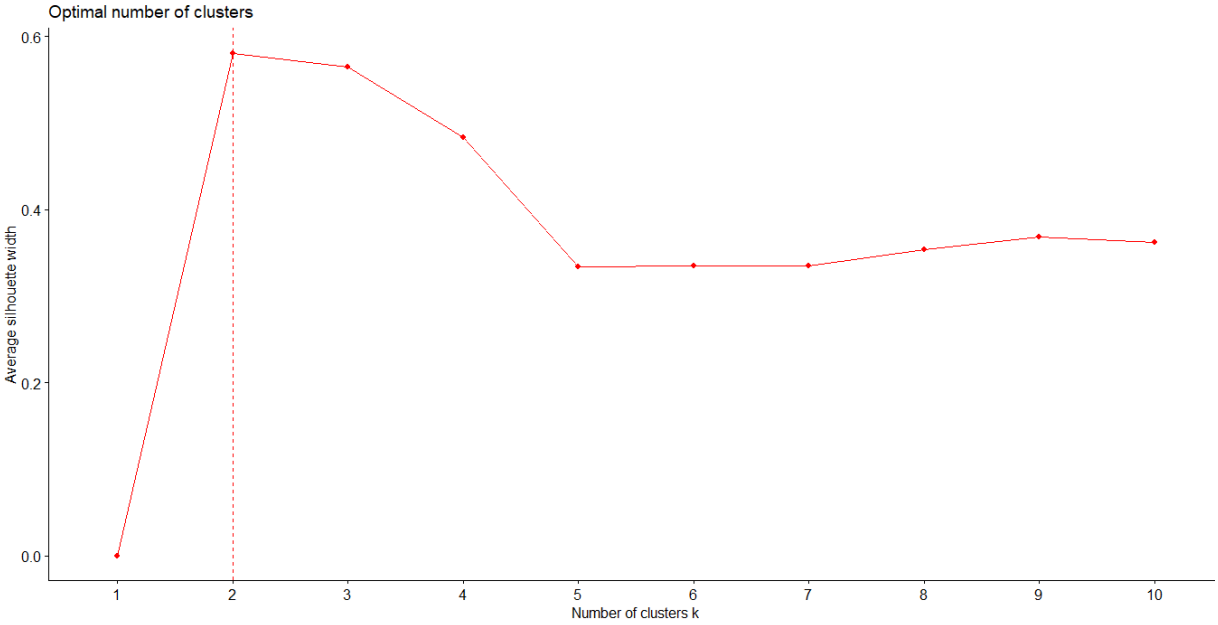
Tabela 2 – Valores das métricas Silhouette e WSS para cada teste

N° de Grupos	Cluster Divisivo	
	Silhouette	WSS
2	0,91	3,19
3	0,91	3,01
4	0,90	2,92
5	0,71	2,25
6	0,71	2,20

Fonte: Autora

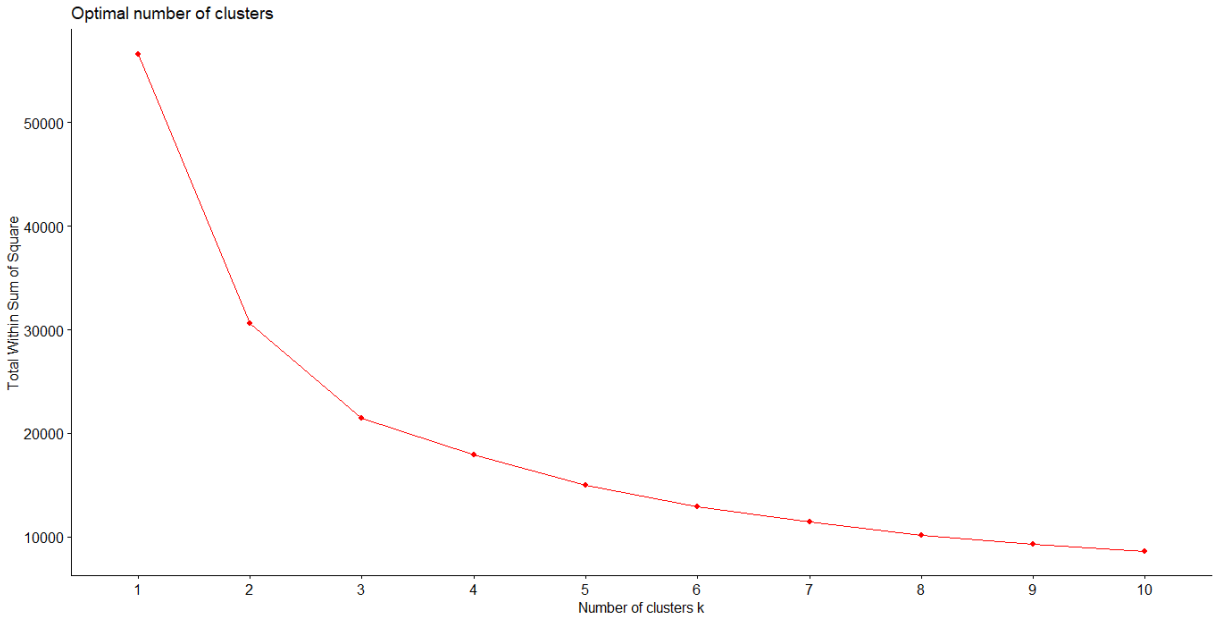
Uma outra forma para analisar a melhor quantidade de agrupamentos, visualmente, é utilizar o “*Elbow Method*” (método do cotovelo), que auxilia na definição da quantidade ideal de grupos, já que ele calcula a soma da variabilidade entre eles. Portanto, a a partir do ponto em que a linha tende a tornar-se plana, ou próximo disso, tem-se uma boa ideia da quantidade de grupos, pois, a partir dali a variabilidade e a distância são pequenas e muito provavelmente não valha a pena seguir adiante. Para gerar este gráfico (Figura 17), foi utilizada como parâmetro a tabela de dados formatada, descrita no início deste subcapítulo, e foi escolhida como métrica a Silhouette. O gráfico também foi gerado também com o auxílio da métrica WSS (Figura 18).

Figura 17 – Visualização do melhor número de agrupamentos – métrica Silhouette



Fonte: Autora

Figura 18 – Visualização do melhor número de agrupamentos – métrica WSS



Fonte: Autora

Como pode-se perceber pelas Figuras 16 e 17, os melhores números de agrupamentos seriam 2 ou 3. Entretanto, além destas métricas, um outro fator determinante é a quantidade de registros por grupo, já que não é possível dividir os grupos com valores muitos altos ou muito baixos de componentes. Nesse sentido, quando configurado para dois agrupamentos, o algoritmo de cluster utilizado fez a seguinte participação: um grupo com 62% dos registros da amostra e o outro com 38%. Todavia, quando configurado nos demais testes (entre 3 e 6 agrupamentos), não

foram obtidos grupos com quantitativos de registros que fizessem sentido. O algoritmo gerou grupos muito pequenos, não sendo representativos de um comportamento.

Por exemplo, quando configurado para particionar os dados em 3 grupos, o algoritmo divisivo chegou a 1 grupo com apenas 1 elemento; quando configurado para 4 grupos, ocorreu o mesmo: agora, 2 grupos ficaram com um número bem pequeno de componentes, variando de 1 a 2 itens apenas. Quando configurado para dividir a amostra em 5 ou 6 agrupamentos, 3 grupos ficaram com apenas 1 ou 2 registros. Visto que é necessário que se encontre, nesses grupos, um padrão de comportamento, julgou-se improvável que isso pudesse ocorrer com tão poucos registros pertencentes a eles. Isto fez com que fosse considerada, como o melhor número de agrupamentos, a quantia de 2 grupos, levando-se também em conta as métricas Silhouette e WSS, como destacado.

Com essa configuração, foram gerados 2 grupos: grupo 1, com 3.893 registros (linhas), e grupo 2, com 2.389 registros (linhas). Como já salientado, nesta formatação dos dados, cada linha não representa um único aluno, mas sim um comportamento de um aluno em MOOC. Para descrever de forma sintetizada algumas métricas importantes utilizadas para caracterizar os dois grupos particionados, foi elaborada uma tabela (Tabela 3) que sistematiza, por grupo, as seguintes informações: número máximo e mínimo da quantidade de dias ativos, do índice de persistência nas atividades e do índice de persistência nos questionários, além da média para cada uma dessas variáveis. Por último, foi incorporada a variável “mais de 10 questões”, que não é numérica e, por isso, não apresenta estas métricas.

Tabela 3 – Informações relevantes sobre os grupos formados

Grupos	Variável	Cluster Divisivo		
		Mínimo	Máximo	Média
1	Quantidade de dias ativos	1	47	2,7
	Persistência nas atividades	0	6	1,1
	Persistência nos questionários	1	35,5	1,6
	Mais que 10 questões		SIM	
2	Quantidade de dias ativos	1	20	2,3
	Persistência nas atividades	0	18	1,1
	Persistência nos questionários	0,2	6	1,5
	Mais que 10 questões		NÃO	

Fonte: Autora

O grupo 1 tem como principal característica o fato de que os cursos realizados possuem mais de 10 questões avaliativas. Cerca de 62% (2.416) dos registros apresentam uma quantidade de dias ativos como sendo um ou dois,

aproximadamente 23% (906) realizaram seus MOOCs em 3 ou 4 dias e em torno de 8% (333) fizeram os cursos em 5 ou 6 dias. Dessa forma, 93% deste grupo possui a quantidade de dias ativos variando entre 1 e 6, cerca de 5% está entre 7 e 12 dias e os demais estão acima de 12. Quanto à persistência nas atividades, a maior parte dos componentes (3.189) apresenta uma persistência correspondente ao intervalo de 1 a 2, (em torno de 81% dos registros) Em relação à persistência nos questionários, a maioria deste grupo (3.294) apresenta uma persistência em um intervalo de 1 a 2, (aproximadamente 84%). Salienta-se também uma boa concentração de alunos com persistência no questionário entre 2,1 e 2,9 cerca de 8% do grupo, e igual a 3 em torno de 5%.

O grupo 2 tem como principal característica o fato de que os cursos realizados não possuem mais de 10 questões avaliativas. Cerca de 70% (1.662) dos registros apresentam uma quantidade de dias ativos como equivalente a 1 ou 2, aproximadamente 20% (477) realizaram seus MOOCs em 3 ou 4 dias e em torno de 5% (147) fizeram os cursos em 5 ou 6 dias. Dessa forma, 95% deste grupo possui a quantidade de dias ativos variando entre 1 e 6, e cerca de 5% têm quantidade de dias ativos acima de 7. No que se refere à persistência nas atividades, a maior concentração de registros (2.016) apresenta uma persistência que variou de 0 a 1, o que corresponde a 84% do grupo. Em relação à persistência nos questionários, a maioria dos componentes (1.304) apresentou uma persistência que está entre 0,2 e 1,5, o que equivale a cerca de 54% do grupo.

O que chama a atenção nessas caracterizações é que os níveis das variáveis utilizadas para fazer o agrupamento foi muito semelhante, à exceção da quantidade que questões nos cursos, que separou perfeitamente os grupos, indicando que esta variável tem uma influência desproporcional em comparação à interação entre todas as demais. Este resultado indica que não é possível aceitar a hipótese de um “comportamento de caçador” com características do estudante (i.e. quantidade de dias ativos, persistência nas atividades e persistência nos questionários), e que este comportamento é determinado pela quantidade de questões do curso. Afastada essa hipótese, segue a análise do perfil de “estudante caçador”.

7.2.2 Identificação dos Estudantes Caçadores de Certificados

Para identificar os estudantes caçadores de certificados, foi utilizado o seguinte conjunto de dados: “02. Tabela descrição do aluno”. Para isso, foram excluídas as variáveis “chave de identificação” e “quantidade de inscrições inativas” (pouca variabilidade). Dessa forma, foram utilizadas as variáveis quantidade de certificados obtidos, intervalo entre os certificados e quantidade de inscrições. Como o objetivo é encontrar estudantes caçadores, a tabela foi filtrada para mostrar apenas os estudantes com, pelo menos, 2 certificados (pois um “caçador” deve ter, necessariamente, mais de um certificado), o que resultou em pouco mais de 33 mil estudantes. Como esta é uma quantidade muito grande para executar o algoritmo de agrupamento, foram utilizados 20% destas linhas, chegando em 6.609 estudantes, escolhidos de forma aleatória.

Para identificar os melhores agrupamentos, foram aplicados também dois tipos de algoritmos de cluster hierárquico, sendo eles o divisivo e o aglomerativo, e realizados testes com 2, 3, 4, 5 e 6 grupos. Após a realização destes testes, foi observado que o algoritmo divisivo teve um melhor desempenho que o aglomerativo. Dessa forma, este foi escolhido para realização do processo exploratório. Ademais, foram empregadas as métricas Silhouette e WSS (Tabela 4) como indicativas de quais seriam os melhores números de grupos para esta tabela de dados. No que diz respeito à Silhouette, os melhores números seriam 2 ou 3 grupos, e a métrica WSS aponta para 4, 5 ou 6, com valores próximos.

Tabela 4 – Valores das métricas Silhouette e WSS para cada teste

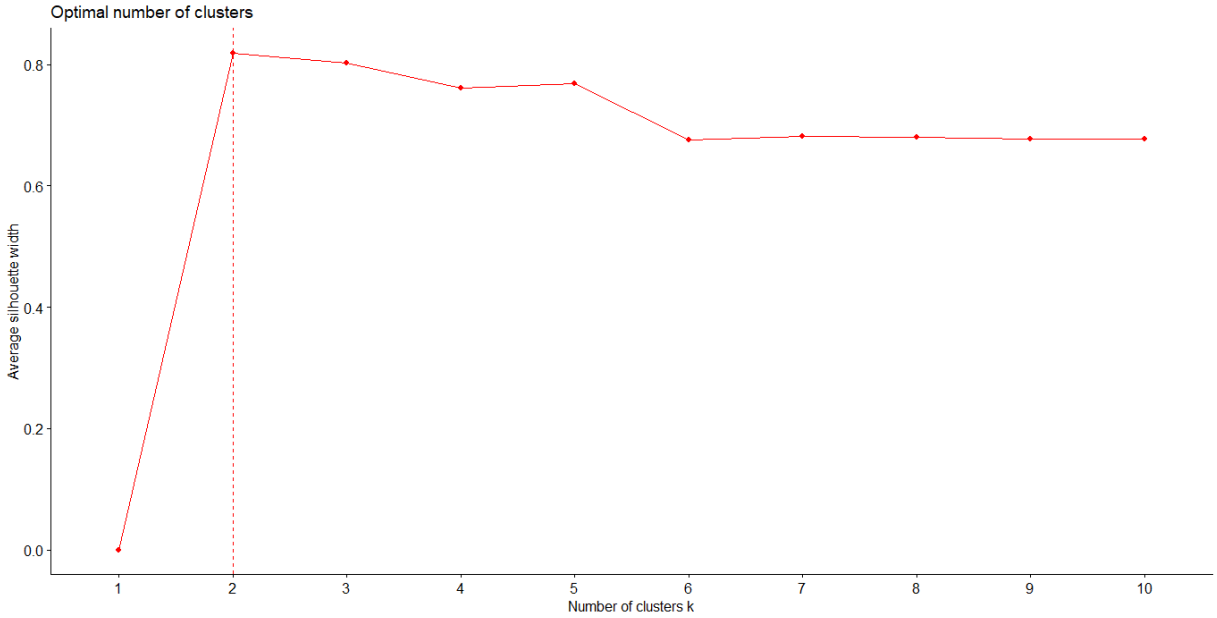
N° de Grupos	Cluster Divisivo	
	Silhouette	WSS
2	0,82	68465544,52
3	0,80	36525801,47
4	0,79	31192848,52
5	0,79	30565276,62
6	0,76	25173576,49

Fonte: Autora

Para deixar mais clara a percepção sobre essas métricas, foi gerada também uma visualização dos melhores valores de agrupamentos utilizando o *Elbow Method* (Figuras 19 e 20). Percebe-se que, quando utilizada a métrica Silhouette, a melhor quantidade de agrupamentos seria 2. Todavia, há uma diferença muito pequena para

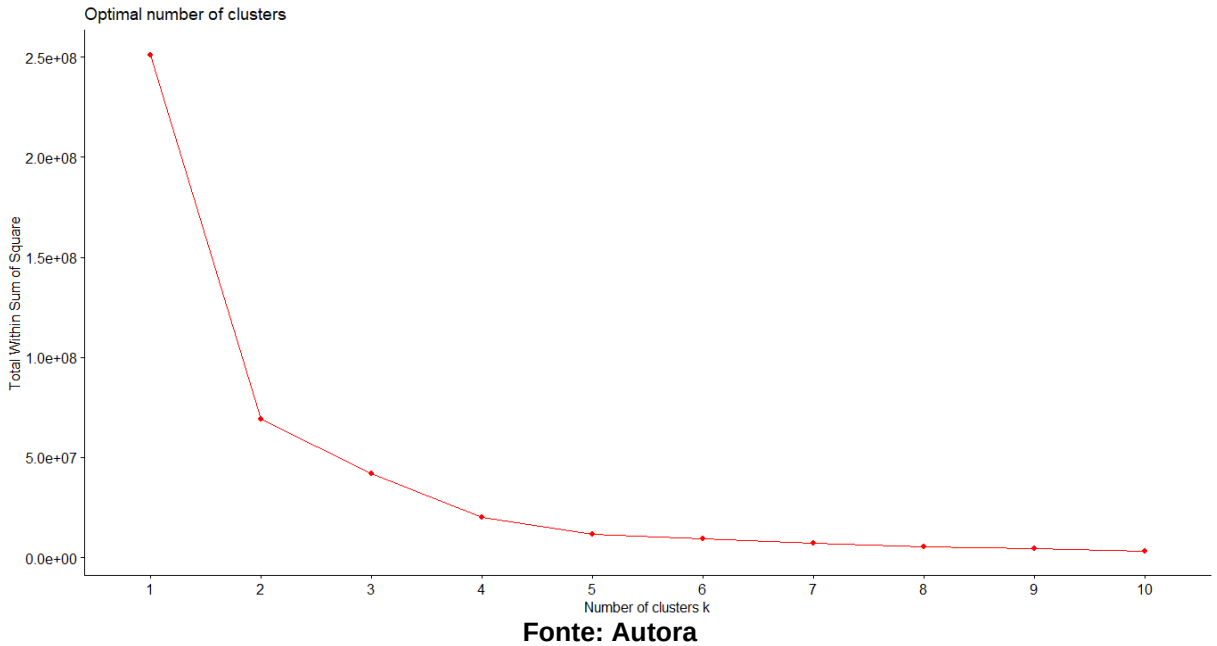
3 grupos, que não pode ser descartada como alternativa. Quando utilizada a métrica WSS, os melhores valores variam de 2 a 4.

Figura 19 – Visualização do melhor número de agrupamentos – métrica Silhouette



Fonte: Autora

Figura 20 – Visualização do melhor número de agrupamentos – métrica WSS



Fonte: Autora

A partir da observação das métricas como indicadores importantes para os melhores números de agrupamentos, partiu-se para realização de uma análise da quantidade de registros por grupo. Levando em consideração que, de acordo com as métricas, 2 ou 3 grupos seriam as melhores opções, focou-se nesses dois valores. Quando utilizado 2 grupos como parâmetro, o cluster divisivo realizou os

agrupamentos de forma que um grupo ficou com 5.739 alunos e o outro com 870. Quando utilizado como parâmetro 3 grupos, o cluster divisivo fez a seguinte partição: grupo 1 – 5.739 alunos, grupo 2 – 681 alunos e grupo 3 – 189 alunos. Devido a uma melhor distribuição entre os alunos no particionamento com 3 grupos, optou-se por este número de agrupamentos. Para uma melhor apresentação das características dos agrupamentos, foi elaborada uma tabela (Tabela 5) que sistematiza algumas informações relevantes, como número máximo e mínimo da quantidade de certificados emitidos, de cursos inscritos e do intervalo entre a geração do primeiro e do último certificados, para cada grupo, além da média de cada uma dessas variáveis.

Tabela 5 – Informações relevantes sobre os grupos formados

Grupos	Variável	Cluster Divisivo		
		Mínimo	Máximo	Média
1	Quantidade de certificados	2	36	3
	Intervalo entre os certificados	0	282	35,5
	Quantidade de cursos inscritos	2	36	3
2	Quantidade de certificados	2	42	4
	Intervalo entre os certificados	283	659	427
	Quantidade de cursos inscritos	2	42	4
3	Quantidade de certificados	2	26	4,8
	Intervalo entre os certificados	662	1540	891,6
	Quantidade de cursos inscritos	2	26	4,8

Fonte: Autora

O grupo 1 é formado pela maioria dos alunos, particionado com 5.739 alunos, o que representa uma parcela de aproximadamente 87% da amostra. Mesmo com quantitativos de certificados tendo números que variam de 2 a 36, a maioria dos alunos deste grupo tem entre 2 e 3 certificados (cerca de 77% do grupo). Mais de mil alunos possuem 4, 5 ou 6 certificados, correspondendo a um percentual de 17% do grupo, e uma boa parcela, cerca de 5% dos alunos, têm entre 7 e 15 certificados. Por fim, 1% obteve mais de 15 certificações. No que se refere ao número de MOOCs inscritos, para este grupo, há a particularidade de que todas as inscrições geraram certificados, ou seja, os alunos terminaram todos os cursos em que se matricularam. Apesar de haver uma similaridade com a quantidade de certificados emitidos, visto que foram utilizados apenas alunos que emitiram no mínimo 2 certificados como amostra, isso não é regra, pois pode haver alunos que possuem mais inscrições do que certificados. Dessa forma, neste grupo, também aproximadamente 94% dos alunos estão inscritos entre 2 e 5 MOOCs.

Por fim, quanto ao intervalo de tempo entre as certificações, a maioria dos alunos deste grupo apresenta intervalos entre 0 e 5 dias. Foi constatado que 1.401

alunos têm intervalo entre o primeiro e o último certificado de 0 (zero) dias, o que compreende cerca de 24% do grupo, e 1338 alunos possuem intervalos de tempo entre 1 e 5 dias, aproximadamente 23% dos alunos do grupo. Dessa forma, cerca de 47% dos alunos deste grupo possuem intervalos de tempo de 0 (zero) a 5 dias de diferença entre a emissão do primeiro e do último certificado, bem abaixo do apresentado nos demais agrupamentos. Evidencia-se, ainda, que cerca de 490 alunos têm intervalos entre 6 e 10 dias, 569 entre 11 e 20 e 519 entre 21 e 40 dias. Acima de 40 dias foi um período identificado para 1.422, com limite máximo de 282 dias. Para um período igual ou maior 100 dias, as taxas de alunos começam a ficar bastante reduzidas, representando menos de 10% do grupo.

O grupo 2 representa uma parcela de cerca de 10% dos registros (681 alunos). Neste grupo, a maior concentração de alunos possui entre 2 e 5 certificados (em torno de 78% da amostra) o que se refere a, mais especificamente, 533 alunos. Um número representativo de alunos, cerca de 6%, possuem 6 certificados e 5% possuem 7 certificados. Apenas 1% dos alunos obtiveram mais de 15 certificações. Como já descrito, devido à similaridade das variáveis, o quantitativo de cursos inscritos para cada aluno também se concentra entre 2 e 5, contabilizando 78% dos estudantes do grupo. Quanto ao intervalo do tempo entre a obtenção dos certificados, há uma distribuição bastante uniforme entre os intervalos identificados, sendo que a maior concentração de alunos está na casa dos 370 dias (11 alunos) seguida de 351 dias (9 alunos). Mas, no geral, há uma média de 2 a 4 alunos distribuídos entre os diversos períodos, sendo possível observar que os maiores quantitativos de alunos (70%) têm intervalos de tempo que vão de 300 a 550 dias, o que pode ser arredondado para, aproximadamente, um ano a um ano e meio de intervalo entre a primeira a última certificação.

O grupo 3 representa o menor grupo particionado, com 189 alunos. Mais de 90% deste grupo têm entre 2 e 8 certificados, correspondendo a 172 alunos. Um fator interessante é que cerca de 8% da amostra tem 8 certificados, enquanto nos limites superiores, acima de 8, foram identificados 17 casos. O mesmo ocorre para os números de cursos inscritos, em que 90% do grupo concentra de 2 a 8 inscrições. No que se refere ao intervalo de tempo entre as certificações, este é o grupo que possui os maiores intervalos identificados. A maioria possui mais de dois anos entre a primeira e a última emissão, e as distribuições dos intervalos são bastante homogêneas, sendo que os maiores valores incluem 3 alunos (intervalos de 705, 727

e 729 dias). O restante tem apenas um aluno por intervalo de tempo e assumem valores que vão de 662 a 1540 dias.

Com base nas análises dos agrupamentos, em primeiro lugar, foi possível perceber que, embora o número de certificados emitidos pelos alunos seja uma característica importante para identificação dos estudantes com perfil de caçadores de certificados, se este valor estiver desvinculado do intervalo de tempo entre essas obtenções, este atributo perde a relevância. Por exemplo, se um aluno tem 5 certificados, mas os obteve em um intervalo de 2 anos, este tem um comportamento que pode ser considerado habitual. Porém, se um aluno obtém 5 certificados em um intervalo de tempo de 0 (zero) dias, este apresenta um comportamento irregular. Em virtude disso, entende-se que o intervalo entre as certificações é uma variável que tem grande influência na separação dos grupos. A definição dos grupos a partir desta variável indica que um grupo de alunos obtém poucos certificados em curto intervalo (média de cerca de 3 certificados em cerca de 35,5 dias), enquanto os demais estudantes obtém suas certificações em intervalo longos cerca de 427 dias em média) ou muito longos (média de cerca de 891 dias).

Todos esses indícios pressupõem um comportamento característico de alunos que buscam obter uma certificação sem se comprometer de forma legítima com a realização do MOOC. Dessa forma, o grupo 1 tem como predominante o comportamento de estudante-caçador de certificados. Embora os limites superiores de intervalos de tempo identificados neste grupo, tenham valores de 282 e 338 dias, é uma parcela pequena deste grupo que se enquadra nessa situação (menos de 10% do grupo). No entanto, mesmo com baixas porcentagens, esse fator foi considerado para a classificação dos estudantes-caçadores.

Contudo, como já salientado na análise do comportamento de caçador, é importante frisar que embora de maneira mais proeminente tenha sido identificado o grupo 1 como tendo perfil de estudantes-caçadores, há muitos casos limítrofes que o algoritmo não é capaz de separar, e apenas com a realização de um estudo individual, caso a caso, torna-se possível a detecção. No entanto, a partir destas análises, houve a possibilidade de identificar-se quais são os principais atributos que caracterizam um estudante como caçador de certificados. Tais atributos serviram como base para geração de regras de classificação dos alunos como caçadores ou não caçadores de certificados. Essas definições foram utilizadas como apoio na determinação de quais

parâmetros de configurações dos MOOCs mais impactam nas ações dos caçadores. Esse processo é exposto no próximo subcapítulo.

7.3 PARÂMETROS DE CONFIGURAÇÃO DOS MOOCS E A INFLUÊNCIA NAS AÇÕES DE CAÇADORES DE CERTIFICADOS

A identificação dos Parâmetros de Configurações dos MOOCs do Lúmina que influenciam nas ações dos caçadores de certificados corresponde à etapa 3 das análises de dados desenvolvidas nesta Tese. Para sua implementação, escolheu-se utilizar uma técnica que pode ser conceituada como um algoritmo de Aprendizagem de Máquina do tipo supervisionado: a Regressão Logística, técnica muito empregada para os mais variados objetivos de MDE. Destaca-se que foi possível utilizar uma técnica supervisionada porque, a partir da aplicação dos algoritmos de agrupamento, foram geradas regras para classificar todos os registros das bases de dados do Lúmina.

Para processo de aplicação da Regressão Logística, foi realizada uma junção das três tabelas principais deste estudo (01. Tabela comportamento nos cursos, 0.2 Tabela descrição do aluno e 0.4 Tabela descrição dos cursos), obtendo-se uma tabela com 39 variáveis (colunas) e mais de 165 mil linhas. Cada registro corresponde a um certificado emitido, pois um aluno pode emitir mais de um certificado.

Após este processamento, os registros foram classificados, primeiramente utilizando-se as regras obtidas na identificação do comportamento de caçador e, em seguida, utilizando-se as regras definidas na identificação dos estudantes-caçadores. Assim, essas regras foram generalizadas para todos os alunos. Para isso, os dados foram filtrados da seguinte forma:

1. *Comportamento de caçador*: todos os registros com características semelhantes ao do grupo 2 foram categorizados com comportamento de caçador “TRUE”, e, semelhantes ao do grupo 1, com “FALSE”. Ressalta-se que não foi possível identificar um “comportamento de caçador”, apenas que se um curso tem mais de 10 questões, ele separa perfeitamente dois grupos;
2. *Estudantes-caçadores*: para esta classificação, apenas alunos com 2 ou mais certificados foram incluídos (como realizado no processo de agrupamento). Todos os alunos com características semelhantes às dos

> alunos que compunham o grupo 1 foram categorizados como caçadores “TRUE”, e os demais alunos, que se encaixam nos grupos 2 e 3 foram classificados como caçadores “FALSE”. No entanto, nesta última classificação, dos estudantes-caçadores, as partições feitas pelo algoritmo de agrupamento foram consideradas muito extensas, e muitos casos limítrofes foram encontrados. Por isso, resolveu-se estabelecer um corte menor do que o identificado no processo de clusterização, o que atenua a ocorrência de classificações indevidas. Dessa forma, foram filtrados para o grupo 1, de estudantes-caçadores, aqueles alunos que obtiveram seus certificados com intervalos de tempo de até dois meses.

A partir deste processo, foram criadas 2 tabelas (comportamento de caçador/estudantes-caçadores) categorizadas em alunos caçadores de certificados (TRUE) e alunos não caçadores de certificados (FALSE), tendo por base os cortes feitos pelos algoritmos de agrupamento e, também, as análises dos grupos identificados. Deve-se lembrar que, na tabela que corresponde ao comportamento de caçador, um tipo de comportamento nos MOOC foi categorizado, e, na tabela de estudantes-caçadores, o próprio aluno foi identificado como sendo ou não caçador.

Ainda, é importante frisar que somente registros que obtiveram os certificados foram considerados. Ademais, seguindo as mesmas regras delimitadas na identificação dos estudantes-caçadores, com os algoritmos de agrupamento, para a aplicação da Regressão Logística, foram utilizados apenas os registros de alunos que obtiveram 2 certificados ou mais. Os resultados da aplicação da regressão sobre as duas tabelas de dados são expostos na sequência (sub etapas 3.1 e 3.2 – Figura 6).

Posteriormente à apresentação dos resultados da análise de Regressão Logística, julgou-se importante apresentar algumas informações quantitativas a respeito dos números de caçadores de certificados distribuídos pelos MOOCs do Lúmina, com enfoque nos parâmetros de configuração, e a influência nas ações de caçadores de certificados. Nesse sentido, a última seção deste capítulo apresenta estas descrições (sub etapa 3.3 – Figura 6).

7.3.1 Regressão Logística na classificação do Comportamento de Caçador

Neste cenário, a tabela obtida possuía mais de 165 mil registros, e, por meio de análises e da aplicação da Regressão *Stepwise* combinada com a reamostragem

com *Bootstrap*, identificou-se que as variáveis mais significativas para este contexto eram *configuração* (definida na seção 6.3.1), em que a configuração poderia assumir os valores “muito restritiva”, “restritiva” ou “pouco restritiva”, e dificuldade pode assumir os níveis de 0 (zero) a 5, em que 0 (zero) indica um curso muito fácil e 5 indica um curso muito difícil. Desse modo, a Regressão Logística foi aplicada considerando esta variável e, também, a variável meta, que definia se o registro indicava um comportamento de caçador ou não.

A partir do modelo sintetizado com as 3 variáveis citadas (*é_caçador*, *categoria* e *dificuldade*), foi utilizada a razão de chances (odds ratio) derivada da análise multivariada de Regressão Logística para obter a probabilidade das estimativas. Explicando melhor, a razão de chances é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo; chance ou possibilidade é a probabilidade de ocorrência deste evento dividida pela probabilidade da não ocorrência do mesmo evento. Para exemplificar, uma razão de chances de 1 indica que a condição ou evento sob estudo é igualmente provável de ocorrer nos dois grupos. Uma razão de chances maior do que 1 indica que a condição ou evento tem maior probabilidade de ocorrer no primeiro grupo. Finalmente, uma razão de chances menor do que 1 indica que a probabilidade é menor no primeiro grupo do que no segundo. Dessa forma, foram obtidas as seguintes conclusões:

1. Quando o MOOC tem categoria pouco restritiva, as chances de um aluno ter um comportamento de caçador de certificados são 144% maiores que em cursos com categoria muito restritiva.
2. Quando o MOOC tem categoria restritiva, as chances de um aluno ter um comportamento de caçador de certificados são 4% menores do que em um curso de categoria muito restritiva.

De acordo com as razões de chances descritas, preliminarmente foi possível perceber que um curso pouco restritivo tem maior possibilidade de os alunos apresentarem comportamento de caçadores do que em MOOCs muito restritivos, o que corrobora a análise de agrupamentos, que indicou separação perfeita pela variável “mais de 10 questões”. Todavia, em contramão a esta constatação, de acordo com o modelo de regressão gerado, cursos restritivos tem menos possibilidades de haver alunos caçadores do que em cursos muito restritivos. Acredita-se que isso ocorreu devido às baixas possibilidades de alunos serem caçadores também em

MOOCs restritivos, pois os MOOCs restritivos geralmente também apresentam altos níveis de dificuldade.

O modelo gerado a partir do algoritmo de Regressão Logística apresentou uma precisão de classificação, ou acurácia (número de previsões corretas feitas como uma proporção de todas as previsões realizadas) de aproximadamente 58%, o que considera-se insuficiente para indicar que as restrições nas configurações sejam eficazes para inibir o comportamento de caçador de certificados, enquanto a sensibilidade foi de 81% e a especificidade 47%. Explicando melhor as métricas, as duas estão associadas a Área sob a Curva (AUC), uma métrica de desempenho para medir a capacidade de um classificador binário de discriminar entre classes positivas e negativas. A AUC pode ser dividida em sensibilidade e especificidade. Sensibilidade é a verdadeira taxa positiva, são as instâncias numéricas da classe positiva que realmente foram previstas como positivas. Especificidade é a verdadeira taxa negativa, ou seja, é o número de instâncias da classe negativa que foram realmente previstas como negativas. Sendo assim, destaca-se que o modelo obteve uma maior precisão ao classificar os comportamentos de caçadores de certificados (81%) do que os de não caçadores.

7.3.2 Regressão Logística na classificação do Estudante-Caçador

Para aplicar a regressão empregando as regras de estudantes-caçadores, a tabela foi filtrada para apresentar apenas os registros dos alunos que emitiram 2 ou mais certificados. À vista disso, a tabela obtida possuía um pouco mais de 105 mil registros (mais de 33 mil alunos). Depois de classificados, os registros foram unidos às demais tabelas citadas (01 e 04). Assim, foi obtida uma tabela que possui uma linha para cada certificado gerado por estes alunos, e, independentemente do comportamento em cada MOOC, cada aluno foi classificado conforme as regras obtidas no processo de agrupamento. Por meio de análises e da aplicação da Regressão *Stepwise* combinada com a reamostragem com *Bootstrap*, foi identificado, da mesma forma que na aplicação anterior, que as variáveis mais significativas para este contexto são: *configuração* e *dificuldade*. Então, a Regressão Logística foi aplicada considerando estas 2 variáveis e a variável meta, que definia se um aluno era ou não um caçador de certificados.

A partir do modelo sintetizado com as 3 variáveis citadas (*é_caçador*, *categoria* e *dificuldade*), foi também utilizada a razão de chances (*odds ratio*) para identificar quais os valores de variáveis mais influenciam nas ações dos estudantes caçadores de certificados. Com isso, foi possível encontrar as seguintes possibilidades:

1. Se MOOC apresenta categoria pouco restritiva, as chances de um aluno ser classificado como um estudante-caçador são 282% maiores do que em cursos com categoria muito restritiva.
2. Se MOOC apresenta categoria restritiva, as chances de um aluno ser classificado como um estudante-caçador são 261% maiores do que em um curso de categoria muito restritiva.

Preliminarmente, observou-se que, quando um MOOC possui uma menor restrição em suas configurações, as chances de haver alunos caçadores de certificados são aumentadas consideravelmente.

Uma outra conclusão interessante sobre a aplicação do processo de regressão é que, quanto menores os intervalos entre as emissões dos certificados pelos alunos, maior o impacto das restrições de configuração e do nível de dificuldade dos MOOCs. Como destacado no início deste subcapítulo, as regras de classificação dos caçadores de certificados foram obtidas com base nos cortes gerados pelos algoritmos de agrupamento entre os grupos, e para classificar os registros baseando-se nas regras dos estudantes-caçadores, leva-se em consideração, sobretudo, os intervalos entre os certificados. Nesse sentido, quando são considerados caçadores apenas os alunos com intervalos bem pequenos entre o primeiro e o ultimo certificado, o impacto das configurações e do nível de dificuldade aumentam.

Por exemplo: considerando como caçadores os estudantes com intervalos de tempo, entre a emissão do primeiro e do último certificado, de até 10 dias (zero a 10), foi identificado que, para MOOCs com categoria pouco restritiva, há 483% mais chances de um aluno ser classificado como um caçador de certificados do que em MOOCs muito restritivos, e 464% mais chances em MOOCs com categoria restritiva. Quando se determina como caçadores alunos que têm intervalos de tempo de até 15 dias (zero a 15), em MOOCs pouco restritivos, há 440% mais possibilidades de haver alunos caçadores do que em MOOCs muito restritivos, e 426% mais possibilidades em MOOCs restritivos, evidenciando uma diferença relevante da classificação original.

O modelo de regressão gerado apresentou uma precisão de classificação, de cerca de 55%, enquanto a sensibilidade foi de 52% e a especificidade 55%, o que também indica que as configurações não são suficientes para inibir a ação de caçadores.

8 DISCUSSÕES

Neste capítulo, algumas das descobertas elencadas no capítulo anterior, originadas das explorações feitas com os algoritmos de cluster e de Regressão Logística, são discutidas, com base, sobretudo, nos conhecimentos adquiridos no decorrer do processamento dos dados, construção das tabelas, aplicação dos algoritmos, associando-os a pesquisas correlatas ao tema que foram descritas no decorrer desta Tese.

8.2 DISCUSSÕES ACERCA DOS RESULTADOS

“Caçadores de certificados” é um termo que foi idealizado para indicar alunos que possuem as seguintes características: emissão de grande quantidade de certificados, pequeno intervalo de tempo entre obtenção dos certificados, inscrição em uma grande quantidade de MOOCs, porém, sem realizar avaliações e baixa visualização dos materiais instrucionais. Tais fatores indicam um baixo comprometimento com o curso, em uma tentativa de obter o certificado sem estudar todo o conteúdo disponibilizado. Após a descoberta de usuários que possuem este perfil, há interesse em que este tipo de aluno seja desestimulado de realizar MOOCs na plataforma, sobretudo porque o principal objetivo do Lúmina é democratizar o acesso à educação pública, gratuita e de qualidade, com foco na aprendizagem do aluno, e não no fornecimento de certificações. Tais certificados podem ter uma diminuição da confiabilidade de que sejam uma evidência válida de proficiência se comportamentos como estes não forem inibidos.

Há muitos aspectos dessa Tese que merecem destaque nessa discussão. Entretanto, alguns devem ser melhor detalhados, como: o intuito da seleção das variáveis que compõe as tabelas de dados utilizadas, a identificação dos caçadores de certificados realizada com duas perspectivas diferentes, imprecisões que decorrem de especificidades dos comportamentos dos alunos e da técnica de MDE utilizada, motivações para a escolha das técnicas aplicadas para identificação dos caçadores e na identificação dos parâmetros dos MOOCs, que influenciam os caçadores e, por fim, os resultados propriamente ditos e suas relações. Na sequência, esses tópicos são discutidos.

Por mais que muitos dados (variáveis) sobre os alunos do Lúmina pudessem ser sistematizadas, devido principalmente à limitação de recursos computacionais, optou-se por gerar tabelas de dados que tivessem informações que julgou-se relevantes para o contexto deste estudo, ou seja, possibilitar a identificação de alunos caçadores de certificados, e os parâmetros que os afetam, visto que, dependendo do objetivo, os dados sistematizados poderiam ser diferentes. Essa sistematização deu-se, em primeiro lugar, com base na experiência obtida pela observação da trajetória dos alunos, tanto pela pesquisadora, como pela orientadora desta Tese, e demais integrantes da equipe gestora do Lúmina. Sendo assim, aspectos como persistência, permanência, quantidade de certificados, intervalos de tempo entre certificados e tempo que os alunos permanecem logados na plataforma são indícios fortes para caracterizar os comportamentos indesejados observados na plataforma. Ademais, fatores como configuração e dificuldade de um MOOC impactam na participação dos alunos.

Em segundo lugar, muitos dos estudos analisados para composição desta Tese deram indícios de quais são os atributos mais relevantes de um aluno em um MOOC, especialmente se o objetivo for realizar a identificação de um comportamento em específico. Ruipérez-Valiente *et al*. (2017b) apresentam um estudo que busca entender de forma mais ampla os comportamentos de usuários de MOOCs. Para sua realização, os autores formataram uma base de dados com as seguintes variáveis: nota final, lista de todas as submissões de atividades avaliativas, identificação quanto à geração ou não de certificado pelo aluno, número total de tentativas nas atividades avaliativas, número total de dias em que um aluno ficou ativo no curso, número total de videoaulas acessadas pelos alunos e número total de tópicos de discussão acessados. Além deste estudo, em Ruiperez-Valiente *et al*. (2017ª), foi possível identificar uma descrição detalhada dos atributos que compuseram as bases de dados utilizadas pelos autores para identificar alunos que praticavam o CAMEO. Tais dados foram divididos em 3 partes: a primeira, relacionada aos alunos, a segunda, relacionada às questões avaliativas e a terceira, referente aos recursos de envio. Como se percebe, as tabelas de dados utilizadas nesta Tese apresentam influência de estudos como de Ruipérez-Valiente *et al*. (2017b) e Ruiperez-Valiente *et al*. (2017ª).

A decisão de analisar os caçadores de certificados sob duas óticas diferentes — uma que pressupõe um comportamento de caçador independentemente do aluno

e outra que indica um comportamento do estudante caçador de certificados, sempre exibindo este comportamento — originou-se da percepção de que o comportamento de enganar sistemas computacionais de aprendizagem depende especialmente de dois fatores: a plataforma e o aluno. Contudo, os resultados foram inconclusivos.

Baker *et al.* (2009) constatam que grande parte do comportamento de enganar o sistema, em tutores cognitivos (cerca de 56% da variância), está associado aos recursos dos sistemas de tutoria, como dicas e design. Todavia, alguns autores fora do contexto dos tutores cognitivos (por exemplo, Chi *et al.* 1989; Renkl 1997; Vanlehn 1998, apud Muldner *et al.*, 2011) alegam que, mesmo que os alunos utilizem materiais de instrução iguais, variam acentuadamente na forma como escolhem processá-los, indicando haver uma maior tendência de esses comportamentos dependerem do aluno. Da mesma forma, Muldner *et al.* (2011) destacam que o comportamento do aluno é um preditor melhor de comportamentos inadequados do que os problemas propostos no sistema de tutoria, explicando 50% da variabilidade da variável dependente. Os autores identificaram também diferenças individuais na forma como os alunos se comportam e, por sua vez, se beneficiam (ou não) das características instrucionais do ambiente de aprendizagem. Dessa forma, percebe-se a dualidade e a complexidade do processo de identificação desses comportamentos, surgindo, por isso, a necessidade de realizar análises sob duas perspectivas.

Além disso, outro fator influenciou na condução do processo da maneira que foi apresentada: a percepção de que um aluno pode se comportar como um “caçador” em determinadas circunstâncias. Contudo, isso não define seu perfil enquanto aluno, o que indica que muitos fatores podem influenciar na escolha em se dedicar realmente, ou, em determinado momento, não se dedicar. Esses fatores podem estar associados a alguma motivação externa, que não pode ser identificada, ou a algum fator interno da plataforma, que pode ser mais facilmente detectado com técnicas de MDE, por exemplo.

Quanto aos fatores externos, Souza e Perry (2019) alegam que estes não podem ser identificados por meio das interações com a plataforma e são bastante heterogêneos e difíceis de tratar. Em se tratando dos fatores internos do ambiente virtual de aprendizagem, de acordo com Muldner *et al.* (2011), há uma associação entre os recursos das aulas, o design de interface do sistema e os comportamentos de burla, indicando que, se o aluno está em um impasse causado por um projeto de ambiente virtual de aprendizagem insatisfatório, e não por uma falta de conhecimento

ou motivação, ele irá usar de estratégias para contornar o impasse, não necessariamente de forma deliberada. Tal circunstância reforça a ideia de que um aluno pode se comportar como caçador de certificado em um determinado MOOC, mas em outro não, apesar de haver o entendimento de que existem caçadores de certificados que agem dessa forma intencionalmente, sendo uma característica do perfil do aluno.

Assim, optou-se pelas duas investigações, que verificam um comportamento vinculado a determinadas características dos MOOCs e da plataforma com potencial para indicar que existe um comportamento de caçador de certificados no Lúmina. Além disso, torna-se possível uma verificação mais especifica sobre o perfil dos alunos do Lúmina (estudante-caçador), o que pode estabelecer quais alunos tem este tipo de comportamento de modo mais acentuado. Na primeira análise, as variáveis que compunham a tabela de dados tinham maior relação com os MOOCs, e na segunda, as variáveis se associavam mais diretamente com os alunos. Depois da realização destas análises, feitas por meio do processo de agrupamento, como destacado no capítulo anterior, foram geradas regras para classificar todos os alunos do Lúmina, e concluiu-se que, no caso do “comportamento do caçador”, o algoritmo de agrupamento não foi capaz de identificar características que permitam identificar os usuários, pois os grupos formados apresentam níveis parecidos na maioria das variáveis utilizadas, à exceção da variáveis "curso tem mais de 10 questões", que é um indicador de dificuldade do curso. Em relação à identificação de estudantes caçadores, entende-se que a obtenção de pelo menos 3 certificados em menos de 35 dias é um bom indicador para classificar um estudante como caçador de certificados. Em relação ao modelo que ajusta a presença de caçadores às configurações dos cursos, conclui-se que não há indícios suficientes para indicar que as restrições nas configurações sejam eficazes para inibir caçadores de certificados.

É importante destacar ainda que há muitos casos limítrofes, os quais as análises desenvolvidas neste estudo não são capazes de distinguir, como, por exemplo: como diferenciar um estudante muito aplicado de um caçador de certificados? Como diferenciar um estudante que obteve o certificado em um dia para usar em suas horas complementares, pois estava com pouco prazo, de um que fez o curso em pouco tempo de forma premeditada? Essa distinção é complexa. Contudo, Ruiperez-Valiente *et al*. (2016) apontam, acerca do CAMEO, que os alunos que utilizam esse método tendem a ter alta taxa de sucesso e tempo de resposta mais

rápido em comparação com outros alunos. Mesmo que os tipos de comportamentos inadequados não sejam iguais, quando indicam que o aluno tem o intuito de enganar o sistema, geralmente há maior tendência ao sucesso na obtenção do certificado, ou seja, este aluno mostra-se mais eficiente que os demais, o que foi considerado uma tendência para identificar os caçadores de certificados.

Contudo, também há alunos que apresentam este desempenho de forma real, ou que precisam obter um certificado rapidamente. Este limiar entre os comportamentos dos alunos pode ser analisado somente de modo individual ou de forma qualitativa, investigações que não estão no escopo desta Tese. Dessa forma, destaca-se que não há como ter certeza sobre o motivo da postura desenvolvida pelos alunos, e os resultados apresentados podem conter imprecisões. Isso acontece não somente por conta do próprio comportamento dos alunos, como destacado nas questões levantadas, mas, também, pela técnica exploratória utilizada, em que os grupos possuem limites muito próximos. Por isso, algumas regras obtidas no processo de agrupamento tiveram de ser revistas, a fim de melhor condizerem com os dados analisados, sobretudo para a análise de estudante-caçador.

Outro aspecto importante a ser discutido é a técnica utilizada neste estudo para a identificação dos caçadores. Muitos pesquisadores utilizam técnicas de MDE, baseadas em algoritmos de AM, para investigar comportamentos inadequados de alunos. Exemplos destes estudos podem ser vistos em Ruiperez-Valiente *et al.* (2017a), que empregaram o algoritmo *Randon Forest* para identificar alunos que praticavam o CAMEO e Paquette *et al.* (2014), que utilizaram os algoritmos Árvore de Decisão, *JRip*, *Step Regression* e *Naïve Bayes*, para caracterizar alunos que enganavam o sistemas em tutores cognitivos. Neste mesmo contexto, pode-se citar Baker e De Carvalho (2008), que utilizaram o algoritmo de Árvores de Decisão, e Baker, Mitrović e Mathews (2010), que aplicaram os algoritmos Árvores de Decisão, *Step Regression, Suporte Vector Machines, Naïve Bayes e Bagged Decision Stumps* para caracterizar alunos que enganam o sistema no SQL-Tutor.

O que essas pesquisas têm em comum é o fato de todos esses algoritmos serem supervisionados. Neste caso, uma base com dados já classificados é fornecida para os algoritmos de modo que eles sejam treinados. Assim, quando novos alunos utilizam a mesma plataforma, os modelos já treinados podem ser aplicados aos dados destes alunos, gerando, como resultado, uma nova classificação. Para que isso ocorresse, Ruiperez-Valiente *et al.* (2017a), em estudos anteriores (RUIPEREZ-

VALIENTE *et al.*, 2016), desenvolveram um algoritmo específico para categorizar alunos que praticavam o CAMEO. Nos diversos estudos citados, relativos ao engano do sistema em tutores, foi empregada a Engenharia do Conhecimento a fim de rotular os dados que compuseram a base de dados de treinamento. Tal processo foi desenvolvido por especialista em tutores cognitivos, que já estudava esse comportamento anteriormente.

Diferentemente deste estudo, os trabalhos citados já tinham um corpo de publicações e pesquisas, o que oportunizou a utilização de técnicas supervisionadas para identificação dos alunos com comportamentos inadequados. Estas pesquisas também possuíam uma certeza prévia de quais eram as características dos comportamentos a serem identificados, como o CAMEO, o que facilita a classificação de uma base de dados. Este estudo, por sua vez, além de não ter precursores, devido, sobretudo, às especificidades da plataforma objeto de estudo e a sua natureza exploratória, não estabelecia previamente, em totalidade, quais seriam os atributos que caracterizam um aluno como caçador de certificado. Como já exposto, há casos limítrofes difíceis de serem caracterizados, o que impossibilitou a utilização de algoritmos supervisionados inicialmente. Consequentemente, optou-se por aplicar técnicas exploratórias, como o agrupamento, que tem potencial para estabelecer quais seriam esses atributos.

Autores como Tan *et al.* (2018) e Rodrigues *et al.* (2016) utilizaram em suas pesquisas algoritmos de agrupamentos para identificar perfis comportamentais em MOOCs, e chegaram a resultados promissores sobre como os alunos se comportam quando realizam cursos deste tipo. Embora Tan *et al.* (2018) tivessem como foco a identificação de perfis de aprendizagem, e Rodrigues *et al.* (2016), o reconhecimento de perfis de engajamento, a estratégia pode ser estendida para qualquer tipo de comportamento, apenas alterando-se o foco dos dados analisados.

Além destes autores, no campo de estudos sobre comportamentos indesejados em MOOCs, os pesquisadores Ruipérez-Valiente *et al.* (2017b) propuseram um algoritmo semelhante ao *K-mens* (um popular algoritmo de agrupamento) com o objetivo de identificar os mais diversos tipos de comportamentos em MOOCs. Em particular, os autores tinham como intuito detectar contas de usuários que sempre enviam seus trabalhos muito próximos (em termos de tempo). Assim, ao observarem as atividades dos grupos identificados, buscavam entender melhor a intenção dos alunos quando trabalhavam em equipe, o que também permitiu

identificar alunos que estavam praticando o CAMEO. Portanto, utilizar algoritmos de agrupamento pareceu uma das melhores formas de caracterizar os caçadores de certificados.

Posteriormente à caracterização dos alunos em caçadores de certificado ou não, por meio dos algoritmos de agrupamento, pode-se generalizar as regras obtidas e classificar todos os alunos que emitiram certificados no Lúmina até março de 2022, o que possibilitou a aplicação de um algoritmo de Aprendizagem de Máquina supervisionado, o algoritmo de Regressão Logística, para identificar quais seriam os principais fatores relacionados aos MOOCs do Lúmina que influenciam na participação dos caçadores de certificados nestes cursos.

Além destas constatações acerca dos grupos categorizados no processo de agrupamento, após a classificação de todos os alunos do Lúmina e com a aplicação do algoritmo de Regressão Logística *Stepwise,* foi possível validar que fatores como configuração de um MOOC impactam na participação de alunos caçadores. Nesta aplicação, foram utilizados todos os registros de alunos que geraram certificados no Lúmina (de setembro de 2016 a março de 2022), e, após a identificação de um subconjunto útil de preditores, a tabela final manteve-se com duas variáveis preditoras (categoria e dificuldade) e a variável meta: se um registro apresentava ou não comportamento de caçador.

Com esta aplicação, foram obtidas as razões de chances, vinculadas ao modelo de regressão, em que foram constatadas que, em um MOOC pouco restritivo, há 144% mais chances de haver alunos caçadores do que em um MOOC muito restritivo. No entanto, com relação aos MOOCs restritivos, foi identificada uma informação divergente do esperado: em MOOCs com este tipo de configuração, há 4% menos chances de haver caçadores de certificados do que em MOOCs muito restritivos. Embora a porcentagem seja pequena, suscita alguns questionamentos sobre o motivo pelo qual isso ocorre. Acredita-se ser em função da pequena quantidade de MOOCs com este tipo de configuração, o que gera um baixo índice de comparação entre as categorias. Além disso, MOOCs restritivos apresentam altos níveis de dificuldade, fator que também impacta na participação dos caçadores. Contudo, como a capacidade de discernir falsos positivos e falsos negativos deste modelo foi muito baixa, conclui-se que o modelo não é adequado para indicar a capacidade das configurações inibirem o comportamento de caçador.

Com essa constatação, corroboram autores como Baker *et al.,* (2009) que afirmaram que o comportamento de enganar o sistema em tutores cognitivos está associado aos recursos dos sistemas de tutoria, como dicas e design. Além disso, estudos com CAMEO sugerem que alterações nas configurações dos MOOCs são eficazes para inibir este comportamento. Nesse sentido, Northcutt, Ho e Chuang (2016), Ruiperez-Valiente *et al.* (2016) e Alexandron *et al.* (2017) propõem, como principais recomendações para inibição do CAMEO, o aumento no uso da randomização de questões e o atraso no feedback das questões avaliativas. Com base nestas constatações, foram sistematizados alguns parâmetros que podem ser seguidos nas configurações dos MOOCs do Lúmina, que ajudariam a diminuir a incidência de alunos caçadores de certificados, sem dificultar a experiência dos demais alunos. Tais parâmetros são descritos no próximo subcapítulo.

Para encerrar, ainda é preciso realizar considerações sobre a segunda análise efetuada, que buscou identificar estudantes-caçadores de certificados, a qual pressupõe a existência de alunos com este perfil. Esta análise tinha como base uma tabela de dados mais resumida. As variáveis utilizadas para aplicação dos algoritmos de agrupamento foram: quantidade de certificados obtidos, intervalo entre os certificados e quantidade de inscrições, as quais estão associadas aos alunos. Cada linha da tabela representa um aluno do Lúmina. Nesta análise, identificou-se três tipos de perfis comportamentais para os alunos da plataforma: esporádicos – estes alunos realizam vários cursos na plataforma, mas em longos intervalos de tempo (2 a 4 anos); frequentes – engloba alunos que fazem de 2 a 4 MOOCs na plataforma por ano; e os estudantes-caçadores – dos quais grande parte obtém sua primeira e última certificação na plataforma com intervalos de tempo entre 0 (zero) e cinco dias. Por conta de a característica mais acentuada dos estudantes-caçadores ser, principalmente, o pequeno intervalo de tempo na obtenção dos certificados, considerando um corte mais compatível com os dados analisados, decidiu-se não seguir exatamente as partições feitas pelo algoritmo de agrupamento, o que amenizou a questão dos casos limítrofes para a classificação dos alunos caçadores, como destacado no subcapítulo 7.3.

Para melhor ilustrar como este comportamento é questionável, observe-se que há alunos que acumularam 10, 11, 12, 15, e, até mesmo, 23 certificados emitidos pelo Lúmina em intervalos de tempo de 0 (zero) dias, mais de 300 alunos têm entre 3 e 4 certificações neste mesmo intervalo de tempo, e também há registros de alunos

com 36 certificados em um intervalo de 5 dias, dentre outros muitos casos considerados inadequados. Por mais que uma pessoa precise de um certificado para compor seu banco de horas complementares em seu curso de graduação, ou de horas de formação para uma promoção, ou certificação para concursos públicos, este é um perfil comportamental bastante incomum.

Indícios de que estas duas características (muitos certificados em pouco tempo) são representativas de comportamentos de burla foram encontrados também na literatura sobre MOOCs, no que se refere à rapidez com que os alunos obtêm certificados. Northcutt, Ho e Chuang (2016), os primeiros a caracterizar o CAMEO, notaram uma velocidade fora do comum com que diversos estudantes completavam MOOCs na plataforma *edX*, e isso chamou a atenção. Assim, a partir dessa constatação, iniciaram suas pesquisas e descobriram como os alunos conseguiam certificações em tão pouco tempo: por meio do CAMEO. Quanto ao aspecto da obtenção de muitas certificações, os autores destacaram que, quanto mais certificados um aluno possuía, mais propenso a condutas inadequadas este aluno estava. Os autores relataram que, entre os estudantes que completaram 20 cursos ou mais na plataforma *edX*, 25% praticaram o CAMEO (NORTHCUTT; HO; CHUANG, 2016).

Da mesma forma que utilizado para o comportamento de caçador, também para a classificação de estudantes caçadores foi aplicado o algoritmo de Regressão Logística, depois da generalização das regras e da classificação de todos os alunos. Para esta aplicação, a tabela de dados também possuía 3 colunas, 1 variável preditora (categoria) e a variável meta: se o aluno é ou não um caçador de certificados. Este processo indicou que as configurações e a dificuldade dos MOOCs influenciam fortemente nas ações dos estudantes-caçadores, pois os MOOCs com configuração pouco restritiva têm 282% mais chances de ter caçadores de certificados do que MOOCs muitos restritivos, e MOOCs restritivos têm 261% mais chances em comparação aos muito restritivos.

Outra constatação interessante obtida com a aplicação da regressão é que, quanto menores os intervalos entre as emissões dos certificados, maior o impacto das configurações. Por exemplo: considerando como caçadores os estudantes com intervalos de tempo de até 10 dias (zero a 10), identificou-se que, para MOOCs com categoria pouco restritiva, há 483% mais chances de um aluno ser caçador de certificados do que em MOOCs muito restritivos, e 464% mais chances em MOOCs

com categoria restritiva do que em cursos muito restritivos. Contudo, devido ao baixo desempenho deste modelo em prever os perfis de estudantes com base nas configurações, conclui-se que ele também é insuficiente.

Independentemente da existência dos casos limítrofes, os curtos espaços de tempo e a quantidade elevada de certificados são um indicativo de práticas inadequadas, e não é intuito da plataforma Lúmina apenas uma distribuição de certificações, e sim a manutenção de um repositório com cursos de qualidade, disponibilizados para quem quer aprender com professores especialistas nas temáticas dos MOOCs e de forma acessível, visto que os cursos são gratuitos e necessitam de poucos recursos computacionais dos alunos. Por isso, tem-se um interesse em inibir esses comportamentos, os quais tendem a diminuir a confiabilidade da avaliação nestes tipos de cursos, pois podem abalar a imagem da plataforma, da instituição que a mantem e do formato de ensino.

9 CONCLUSÕES

A principal contribuição desta Tese foi identificar os perfis comportamentais predominantes entre os alunos do Lúmina e relacionar tais comportamentos aos parâmetros de configuração e características dos MOOCs. Dessa forma, com um maior conhecimento sobre os alunos proporcionado por este estudo, é possível por meio do desenvolvimento de adequações nos cursos da plataforma, promover maior engajamento e inibir comportamentos indesejados, não condizentes com os objetivos da plataforma.

Com o desenvolvimento de diversos procedimentos – de MDE e estatísticos, para preparação e análise dos dados – foi possível responder as duas questões de pesquisas levantadas nesta Tese:

Questão 1 – É possível identificar, dentre os estudantes que obtiveram certificados, quais têm o perfil de “caçador de certificados”?

De acordo com os resultados apresentados no Capítulo 7 e discutido no capítulo anterior, demonstrou-se não ser possível identificar estes tipos de alunos, tanto em um padrão comportamental associado aos MOOCs da plataforma, como relacionados aos perfis dos alunos do Lúmina.

Com a realização da segunda análise, identificou-se 3 perfis comportamentais dentre os alunos do Lúmina. O primeiro, é um perfil de alunos *esporádicos* que realizam seus MOOCs em longos períodos, revelando intervalos de aproximadamente 2 a 4 anos entre a obtenção do primeiro e do último certificado na plataforma. O segundo, caracteriza-se como um perfil de alunos *frequentes* na realização dos cursos, no qual a maioria dos estudantes realizam MOOCs no Lúmina em períodos semestrais ou trimestrais. Por fim, foi identificado um perfil que define os *estudantes-caçadores*, caracterizados principalmente pela obtenção de muitos certificados em intervalos de tempo bem pequenos. Boa parte dos alunos com este perfil tem intervalos de tempo entre o primeiro e o último certificado de 0 (zero) a 5 dias de diferença, um período considerado pequeno para um efetivo comprometimento com a realização de um único MOOC. Todavia, a maioria destes alunos possui entre 2 e 5 certificações, denotando um comportamento incompatível com os objetivos do Lúmina.

Destaca-se, também, que, como a quantidade de certificados gerada pelos estudantes-caçadores foi inferior aos registros identificados com comportamento de caçador, é possível dizer que, mesmo um aluno não tendo perfil de caçador de certificados, se este tiver oportunidade, pode se comportar desta forma. Sendo assim, é plausível dizer que há um comportamento de caçador vinculado aos MOOCs do Lúmina, e há também um perfil de estudante-caçador.

Questão 2 – Quais as configurações de um MOOC desencorajam alunos que não estão comprometidos com a aprendizagem, sem desmotivar aqueles que queiram aprender com o curso?

De acordo com as análises feitas, não foi possível estabelecer parâmetros que poderiam inibir os alunos caçadores de certificados a realizarem MOOCs no Lúmina e otimizariam a experiência dos alunos comprometidos com sua aprendizagem, tal como discutido no capítulo anterior.

9.1 LIMITAÇÕES DA PESQUISA

A contínua adaptação do Moodle às características de cursos no formato MOOC trouxe desafios inesperados, em especial com relação aos relatórios de dados gerados pela plataforma, subsídio para elaboração dessa pesquisa. Tais desafios relacionados ao uso desses relatórios dizem respeito principalmente à qualidade e à diversidade dos atributos mantidos. Com o formato nativo dos relatórios (pelo menos nas versões anteriores à 3.8), não é possível qualificar qual o comprometimento dos alunos com os recursos educacionais disponibilizados em cada curso, e isso limita a compreensão do comportamento desses estudantes enquanto realizam um MOOC na plataforma.

Por exemplo, não é possível visualizar quanto tempo os alunos permaneceram realizando uma atividade, somente a data e o horário em que clicaram no link com o nome da atividade. É possível fazer uma estimativa do tempo em cada atividade, considerando a diferença entre acessos em sequência, e contabilizando a quantidade de dias que os alunos se logaram (dias ativos). Contudo, esse método não permitiria diferenciar um estudante que usa várias abas abertas ao mesmo tempo de um que abriu e fechou a atividade apenas para marcá-la como “finalizada”. Da mesma forma, a interação com vídeos não é capturada, de forma que não se sabe, por

exemplo, quantos minutos foram visualizados, se o estudante pausou e retornou ou se aumentou a velocidade de reprodução.

Diante desses fatores, acredita-se que, da forma como os MOOCs estão organizados atualmente no Lúmina, somente as estimativas de tempo de realização das atividades podem ser indícios do comprometimento dos alunos, mas não uma confirmação definitiva para caracterização dos comportamentos destes estudantes. Outros indícios podem ser sistematizados a partir da contagem dos acessos aos recursos pelos alunos, o que complementa as informações obtidas. Todavia, algumas reformulações relacionadas ao método de captação das interações entre alunos e plataforma podem melhorar a compreensão do comportamento desses estudantes. Além disso, a inclusão de outros tipos de interações disponíveis na plataforma também poderia auxiliar no entendimento sobre o comprometimento dos estudantes usuários do Lúmina.

Outro limitador importante desse estudo corresponde, sobretudo, aos casos de comportamentos limítrofes entre os alunos, em virtude de não ser possível distinguir alunos caçadores de alunos muito aplicados, ou alunos que tenham prazos muito pequenos para entregas de certificações. Esse limite acentua-se pela técnica escolhida para identificação do comportamento estudado. Em razão de o contexto desta Tese ser, até o momento, inexplorado, a melhor alternativa para essa análise estava nas técnicas exploratórias de mineração de dados, como os algoritmos de agrupamento. Todavia, estes algoritmos não são capazes de diferenciar limites muito próximos, escolhendo tendências mais fortes para realizar os agrupamentos.

Uma forte limitação que diz respeito, sobretudo, à técnica de Regressão Logística está relacionada à distribuição de instâncias em cada uma das classes dos MOOCs analisados. Há classes que possuem poucos inscritos e concluintes, como nos cursos com maior dificuldade e maior restrição, e outras, como nos cursos com menor dificuldade e pouca restrição, em que há muitos inscritos e concluintes. Este fator pode limitar a generalização de regras de modelos gerados por diversos tipos de algoritmos, como o de Regressão Logística utilizado. De acordo com Periwal e Rana (2017), um dos grandes desafios ao se trabalhar com dados provenientes de MOOCs é a questão do desequilíbrio de classes. Além destes pesquisadores, Xing *et al.* (2016) também relatam este mesmo problema em seu estudo sobre desistência em MOOCs. De forma geral, em abordagens supervisionadas, os algoritmos acabam processando os dados de forma tendenciosa quando há um grande desequilíbrio, pois existem mais

dados categorizados em determinadas classes do que em outras (PERIWAL e RANA, 2017; XING *et al.*, 2016).

Em análises com conjuntos de dados que possuem natureza desbalanceada, como no caso deste estudo, os algoritmos demonstram certa dificuldade em diferenciar as classes. Na maior parte dos casos, há uma tendência em gerar modelos que favorecem as classes com maior número de instâncias, o que resulta em um reconhecimento menor para classes que possuem pequenos quantitativos. Este problema é recorrente na Mineração de Dados e ocorre principalmente porque os algoritmos atribuem, a erros diferentes, igual importância, considerando que as distribuições são relativamente equilibradas (HE e GARCIA, 2009). Dessa forma, neste estudo, este foi um impedimento para uma interpretação mais precisa sobre as características da amostra, especialmente sobre o impacto da configuração e do nível de dificuldade dos MOOCs. Todavia, foi possível obter-se um bom indicativo sobre como estes fatores influenciam no engajamento dos alunos.

É importante também destacar que o simples fato de se estudar dados educacionais é um fator limitante. Por exemplo, Paquette, Baker e Moskal (2018) relatam que grande parte da dificuldade em se estudar dados educacionais é a ampla diferença nos dados armazenados em cada ambiente de aprendizagem, o que leva os pesquisadores a desenvolver modelos muito específicos para cada contexto estudado. Isso acaba impedindo a disseminação de modelos gerais o suficiente para funcionarem em múltiplos contextos, exigindo que pesquisadores interessados em estudar intervenções pedagógicas relacionadas aos comportamentos dos alunos criem seu próprio modelo, o que requer recursos consideráveis. Consequentemente, para realização deste estudo, foi elaborado um processo para manipulação e análise de dados específico para o ambiente de ensino e aprendizagem utilizado o Lúmina, o que demandou esforços mais significativos, havendo também uma grande dificuldade em validá-lo em outros contextos, visto que isso não é possível, pois os modelos foram desenvolvidos especificamente para as bases de dados da plataforma analisada, o que impede a generalização para outras plataformas.

9.1 ESTUDOS FUTUROS

Como trabalho futuro, pretende-se investigar a relação esforço percebido *versus* valor do curso/certificado na realização de um MOOC o qual foi pouco

explorado nesta Tese, já que o foco esteve, de forma mais ampla, no esforço (configurações do Lúmina) em comparação à recompensa. Acredita-se que haja uma correlação entre a quantidade de esforço requerida para realização de um MOOC e a obtenção do certificado, ou do conhecimento proporcionado pelo curso (o valor do MOOC), modulada pelo engajamento do aluno. Além disso, a partir do estudo aqui apresentado, percebeu-se indicativos de que alunos não engajados tendem a evitar cursos que apresentem altos níveis de esforço para sua conclusão, ou seja, percebe-se uma relação entre esses dois aspectos

Sendo assim, para analisar a relação denotada, pretende-se utilizar como framework o modelo de Expectativas e Valores de Eccles *et al.* (1983). Este modelo especifica as ligações de desenvolvimento e causas entre fatores culturais, eventos históricos, expectativas e valores para os comportamentos de desempenho. O modelo tem elementos que frequentemente são objeto de pesquisas em Educação, tais como motivação, autoeficácia, autodeterminação, valor das experiências afetivas e adequação a papeis sociais. Contudo, identificou-se que o modelo apresenta um arranjo potencialmente esclarecedor para os comportamentos dos estudantes da plataforma Lúmina, e provavelmente para MOOCs em geral. Na opinião da pesquisadora, Eccles e seus colegas corretamente perceberam o relacionamento entre Expectativas de sucesso e Valor da tarefa influenciando o desempenho educacional, engajamento e escolhas.

REFERÊNCIAS BIBLIOGRÁFICAS

ABED. **CENSO EAD.BR - Relatório analítico da aprendizagem a distância no Brasil**. [*S. l.*], 2019. Disponível em: http://abed.org.br/arquivos/CENSO_DIGITAL_EAD_2018_PORTUGUES.pdf. Acesso em: 21 jan. 2020.

ABED. **CENSO EAD.BR - Relatório analítico da aprendizagem a distância no Brasil**. [*S. l.*], 2022. Disponível em: http://abed.org.br/arquivos/CENSO_EAD_2020_PORTUGUES.pdf. Acesso em: 10 jun. 2022.

AGGARWAL, Charu C. **Data Mining: The Textbook**. 1. ed. New York, USA: Springer, 2015. v. 1

AL-SHABANDAR, Raghad *et al.* Machine learning approaches to predict learning outcomes in Massive open online courses. *In*: , 2017, Alaska, USA. **International Joint Conference on Neural Networks (IJCNN)**. Alaska, USA: IEEE, 2017. p. 713–720.

ALDOWAH, Hanan; AL-SAMARRAIE, Hosam; FAUZY, Wan Mohamad. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. **Telematics and Informatics**, [*s. l.*], v. 37, p. 13–49, 2019. Disponível em: https://doi.org/10.1016/j.tele.2019.01.007.

ALEVEN, Vincent A.W.M.M.; KOEDINGER, Kenneth R. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. **Cognitive Science**, [*s. l.*], v. 26, n. 2, p. 147–179, 2002.

ALEVEN, Vincent; KOEDINGER, Kenneth R. Limitations of student control: Do students know when they need help?. *In*: , 2000, Berlin, Heidelberg. (Gauthier G., Frasson C., & VanLehn K., Org.)**5th International Conference on Intelligent Tutoring Systems (ITS 2000)**. Berlin, Heidelberg: Lecture Notes in Computer Science, Springer, 2000. p. 292–303. Disponível em: https://dl.acm.org/doi/abs/10.5555/648030.745996.

ALEXANDRON, Giora *et al.* Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC. **Computers and Education**, [*s. l.*], v. 108, p. 96–114, 2017.

ALMATRAFI, Omaima; JOHRI, Aditya; RANGWALA, Huzefa. Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums. **Computers and Education**, [*s. l.*], v. 118, p. 1–9, 2018.

AN, Truong Sinh; KRAUSS, Christopher; MERCERON, Agathe. Can typical behaviors identified in MOOCs be discovered in other courses?. *In*: , 2017, Hubei, China. **10th International Conference on Educational Data Mining (EDM 2017)**. Hubei, China: [*s. n.*], 2017. p. 220–225. Disponível em: http://educationaldatamining.org/EDM2017/proc_files/papers/paper_58.pdf.

ATAPATTU, Thushari; FALKNER, Katrina; TARMAZDI, Hamid. Topic-wise classification of MOOC discussions: A visual analytics approach. *In*: , 2016, North Carolina, USA. **Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)**. North Carolina, USA: [*s. n.*], 2016. p. 276–281.

BAKER, Ryan S.J.d. *et al.* Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. **International Journal of Human Computer Studies**, [*s. l.*], v. 68, n. 4, p. 223–241, 2010. Disponível em: http://dx.doi.org/10.1016/j.ijhcs.2009.12.003.

BAKER, Ryan S.J.D. **Big data and education**. New York, USA: A Massive Online Open Textbook (MOOT) - Teachers College, Columbia University, 2015. Disponível em: http://www.columbia.edu/~rsb2162/bigdataeducation.html.

BAKER, Ryan Shaun. **Designing Intelligent Tutors That Adapt to When Students Game the System**. 2005. 1–124 f. - Carnegie Mellon University, [*s. l.*], 2005. Disponível em: http://reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-05-104.pdf.

BAKER, Ryan S.J.D. *et al.* Developing a generalizable detector of when students game the system. **User Modeling and User-Adapted Interaction**, [*s. l.*], v. 18, n. 3, p. 287–314, 2008.

BAKER, Ryan S.J.D. *et al.* Educational software features that encourage and discourage "gaming the system". *In*: , 2009, Brighton, UK. **14th international conference on artificial intelligence in education**. Brighton, UK: [*s. n.*], 2009. p. 475–482.

BAKER, Ryan S. J. D. Gaming the System: A Retrospective Look. **Philippine Computing Journal**, [*s. l.*], v. 6, n. 2, p. 9–13, 2011. Disponível em: http://www.columbia.edu/~rsb2162/PSCS-gaming-v6.pdf.

BAKER, Ryan S. J. d. Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model. *In*: , 2007, Berlin, Heidelberg. **11th International Conference on User Modeling - Workshop on Data Mining for User Modeling**. Berlin, Heidelberg: [*s. n.*], 2007. p. 76–80. Disponível em: https://www.educationaldatamining.org/UM2007/Baker.pdf.

BAKER, Ryan Shaun *et al.* Off-task behavior in the cognitive tutor classroom: When students "game the system". *In*: , 2004, Vienna, Austria. **Conference on Human Factors in Computing Systems (CHI04)**. Vienna, Austria: [*s. n.*], 2004. p. 383–390. Disponível em: https://doi.org/10.1145/985692.985741.

BAKER, Ryan Shaun *et al.* Off-task behavior in the cognitive tutor classroom. *In*: , 2004, Vienna, Austria. **CHI 2004**. Vienna, Austria: [*s. n.*], 2004. p. 383–390.

BAKER, Michael J. The roles of models in Artificial Intelligence and Education research : a prospective view. **Journal of Artificial Intelligence and Education**, [*s. l.*], v. 11, p. 122–143, 2000. Disponível em: https://telearn.archives-ouvertes.fr/hal-00190395.

BAKER, Ryan Shaun; CORBETT, Albert T.; KOEDINGER, Kenneth R. Detecting Student Misuse of Intelligent Tutoring Systems. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, [*s. l.*], v. 3220, p. 531–540, 2004.

BAKER, Ryan S.J.D.; DE CARVALHO, Adriana M.J.A. Labeling student behavior faster and more precisely with text replays. *In*: , 2008, Montreal, Canada. **1st International Conference on Educational Data Mining (EDM 2008)**. Montreal, Canada: [*s. n.*], 2008. p. 38–47.

BAKER, Ryan Shaun; INVENTADO, Paul Salvador. Educational Data Mining and Learning Analytics. *In*: J.A. LARUSSON AND B. WHITE (EDS.) (org.). **Learning Analytics: From Research to Practice**. 1. ed. New York, USA: Springer, 2014. p. 1–195. *E-book*. Disponível em: https://link.springer.com/book/10.1007%2F978-1-4614-3305-7.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, [*s. l.*], v. 19, n. 02, p. 3–13, 2011.

BAKER, Ryan S.J.D.; MITROVIĆ, Antonija; MATHEWS, Moffat. Detecting gaming the system in constraint-based tutors. *In*: DE BRA, P.; KOBSA, A.; CHIN, D. (org.). **User Modeling, Adaptation, and Personalization. UMAP 2010. Lecture Notes in Computer Science**. Berlin, Heidelberg.: Springer, 2010. v. 6075, p. 267–278.

BAKER, Ryan S J D; YACEF, Kalina. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, [*s. l.*], v. 1, n. 1, p. 3–17, 2009. Disponível em: https://doi.org/10.5281/zenodo.3554657.

BAKHSHINATEGH, Behdad *et al.* Educational data mining applications and tasks: A survey of the last 10 years. **Education and Information Technologies**, [*s. l.*], v. 23, n. 1, p. 537–553, 2018.

BALINT, Trevor A. **An In-Depth Analysis of Problem-Solving Profiles of Students in Open Online Environments**. 199 p. Washington, USA: Tese (Doutorado). The Columbian College of Arts and Sciences - Universidade The George Washington, 2016. Disponível em: https://www.proquest.com/docview/1812057295.

BAO, Yingying; CHEN, Guanliang; HAUFF, Claudia. On the prevalence of multiple-account cheating in massive open online learning. *In*: , 2017, Wuhan, China. **10th International Conference on Educational Data Mining (EDM 2017)**. Wuhan, China: [*s. n.*], 2017. p. 262–265. Disponível em: http://educationaldatamining.org/EDM2017/proc_files/papers/paper_91.pdf.

BEVAN, Gwyn; HOOD, Christopher. What's measured is what matters: Targets and gaming in the English public health care system. **Public Administration**, [*s. l.*], v. 84, n. 3, p. 517–538, 2006.

BRINTON, Christopher G. *et al.* On the Efficiency of Online Social Learning Networks. **IEEE/ACM TRANSACTIONS ON NETWORKING**, [*s. l.*], v. 26, n. 5, p. 2076–2089, 2018. Disponível em: https://ieeexplore.ieee.org/document/8438545.

BROOKS, Christopher; THOMPSON, Craig; TEASLEY, Stephanie. A time series interaction analysis method for building predictive models of learners using log data. *In*: , 2015, New York, USA. **Fifth International Conference on Learning Analytics And Knowledge (LAK 2015)**. New York, USA: ACM, 2015. p. 126–135.

CHUI, Kwok Tai *et al.* Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. **Computers in Human Behavior**, [*s. l.*], v. 107, n. December 2017, p. 105584, 2020. Disponível em: https://doi.org/10.1016/j.chb.2018.06.032.

COBOS, Ruth; OLMOS, Lara. A Learning Analytics Tool for Predictive Modeling of Dropout and Certificate Acquisition on MOOCs for Professional Learning. *In*: , 2018, Bangkok, Thailand. **International Conference on Industrial Engineering and Engineering Management (IEEM 2018)**. Bangkok, Thailand: IEEE, 2018. p. 1533–1537.

COCEA, Mihaela; HERSHKOVITZ, Arnon; BAKER, Ryan S.J.D. The impact of off-task and gaming behaviors on learning: Immediate or aggregate. **Frontiers in Artificial Intelligence and Applications**, [*s. l.*], v. 200, n. 1, p. 507–514, 2009a.

COCEA, Mihaela; HERSHKOVITZ, Arnon; BAKER, Ryan S.J.D. The impact of off-task and gaming behaviors on learning: Immediate or aggregate. *In*: , 2009b. **Frontiers in Artificial Intelligence and Applications**. [*S. l.: s. n.*], 2009. p. 507–514.

CORMIER, Dave; SIEMENS, George. Through the Open Door: Open Courses as Research, Learning, and Engagement. **Educause Review**, [*s. l.*], v. 45, n. 4, p. 30–39, 2010. Disponível em: http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume45/ThroughtheOpenDoorOpenCoursesa/209320.

CORRIGAN-GIBBS, Henry *et al.* Measuring and maximizing the effectiveness of honor codes in online courses. **L@S 2015 - 2nd ACM Conference on Learning at Scale**, [*s. l.*], p. 223–228, 2015.

DAVIS, Dan *et al.* Activating learning at scale: A review of innovations in online learning strategies. **Computers and Education**, [*s. l.*], v. 125, n. April, p. 327–344, 2018.

DE LOS REYES, Daniel A. Guimarães *et al.* Predição de sucesso acadêmico de estudantes: uma análise sobre a demanda por uma abordagem baseada em transfer learning. **Revista Brasileira de Informática na Educação**, [*s. l.*], v. 27, n. 1, p. 1–25, 2019.

DENG, Ruiqi; BENCKENDORFF, Pierre. A Contemporary Review of Research Methods Adopted to Understand Students' and Instructors' Use of Massive Open Online Courses (MOOCs). **International Journal of Information and Education Technology**, [*s. l.*], v. 7, n. 8, p. 601–607, 2017.

DENYER, David; TRANFIELD, David. Producing a Systematic Review. *In*: BUCHANAN, David A.; BRYMAN, Alan (org.). **The SAGE Handbook of Organizational Research Methods**. 1. ed. Los Angeles, USA: Sage Publications Ltd, 2009. p. 671–689. Disponível em: https://www.cebma.org/wp-content/uploads/Denyer-Tranfield-Producing-a-Systematic-Review.pdf.

DOWNES, Stephen. Recent Work in Connectivism. **European Journal of Open, Distance and E-Learning**, [*s. l.*], v. 22, n. 2, p. 113–132, 2019.

DURU, Ismail; DOGAN, Gulustan; DIRI, Banu. An overview of studies about students' performance analysis and learning analytics in MOOCs. **Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016**, [*s. l.*], p. 1719–1723, 2016.

ECCLES, Jacqueline Parson *et al.* Expectancies, Values, and Adacemic Behaviors. *In*: SPENCE, Janet T. (org.). **Achievement and Achievement Motives: Psychological and Sociological Approaches**. 1. ed. San Francisco, USA: W. H. Freeman and Company, 1983. v. 97, p. 75–146.

FOURNIER, Hélène; KOP, Rita. MOOC learning experience design: Issues and challenges. **International Journal on E-Learning: Corporate, Government, Healthcare, and Higher Education**, [*s. l.*], v. 14, n. 3, p. 289–304, 2015.

GALLEN, Rosa Cabedo; CARO, Edmundo Tovar. An exploratory analysis of why a person enrolls in a Massive Open Online Course within MOOCKnowledge data collection. *In*: , 2017, Athens, Greece. **Global Engineering Education Conference, (EDUCON)**. Athens, Greece: IEEE, 2017. p. 1600–1605.

GARCÍA, Concepción Bueno *et al.* Designing and implementing a massive open online course: Lessons learnt. *In*: , 2017, Cádiz, Espanha. **5th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2017)**. Cádiz, Espanha: Association for Computing MachineryNew YorkNYUnited States, 2017. p. 1–6. Disponível em: https://dl.acm.org/doi/10.1145/3144826.3145431.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. 4. ed. São Paulo, Brasil: Atlas, 2002-. ISSN 85-224-3169-8.

GREENE, Jeffrey A.; OSWALD, Christopher A.; POMERANTZ, Jeffrey. Predictors of Retention and Achievement in a Massive Open Online Course. **American Educational Research Journal**, [*s. l.*], v. 52, n. 5, p. 925–955, 2015.

GUO, Shou Xi *et al.* Attention-Based Character-Word Hybrid Neural Networks With Semantic and Structural Information for Identifying of Urgent Posts in MOOC Discussion Forums. **IEEE Access**, [*s. l.*], v. 7, p. 120522–120532, 2019.

GUO, Philip J.; KIM, Juho; RUBIN, Rob. How video production affects student engagement: An empirical study of MOOC videos. *In*: , 2014, Atlanta, EUA. **1st ACM Conference on Learning at Scale (L@S 2014)**. Atlanta, EUA: ACM, 2014. p. 41–50.

HE, Jingjing *et al.* Analysis of MOOC Learning Rhythms. *In*: , 2018, Exeter, Reino Unido. **20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)**. Exeter, Reino Unido: IEEE, 2018. p. 1555–1562. Disponível em: https://ieeexplore.ieee.org/document/8622993.

HE, Haibo; GARCIA, Edwardo A. Learning from Imbalanced Data. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, [*s. l.*], v. 21, n. 9, p. 1263–1284, 2009.

HEW, Khe Foon; QIAO, Chen; TANG, Ying. Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. **International Review of Research in Open and Distance Learning**, [*s. l.*], v. 19, n. 3, p. 69–93, 2018.

HOLLANDS, Fiona M.; TIRTHALI, Devayani. MOOCs: Expectations and Reality Full Report. **Cost Studies of Education, Teachers College, Columbia University, NY.**, New York, USA, n. May, p. 211, 2014. Disponível em: https://files.eric.ed.gov/fulltext/ED547237.pdf.

HUTT, Stephen *et al.* Gaze-based detection of mind wandering during lecture viewing. *In*: , 2017, Hubei, China. **10th International Conference on Educational Data Mining (EDM 2017)**. Hubei, China: [*s. n.*], 2017. p. 226–231. Disponível em: https://files.eric.ed.gov/fulltext/ED596576.pdf.

IQBAL, Sajid *et al.* On the impact of MOOCs on engineering education. *In*: , 2014, Patiala, India. **International Conference on MOOCs, Innovation and Technology in Education (MITE 2014)**. Patiala, India: [*s. n.*], 2014. p. 101–104.

JOKSIMOVIĆ, Srećko *et al.* What do cMOOC participants talk about in social media?: a topic analysis of discourse in a cMOOC. *In*: , 2015, New York, USA. **Fifth International Conference on Learning Analytics And Knowledge (LAK 2015)**. New York, USA: ACM, 2015. p. 156–165.

KASHYAP, Avinash; NAYAK, Ashalatha. Different Machine Learning Models to Predict Dropouts in MOOCs. *In*: , 2018, Bangalore, India. **International Conference on Advances in Computing, Communications and Informatics (ICACCI 2018)**. Bangalore, India: IEEE, 2018. p. 80–85.

KING, Carolyn; ROBINSON, Andrew; VICKERS, James. Online education: Targeted MOOC captivates students. **Nature**, [*s. l.*], v. 505, n. 7481, p. 26, 2014.

KOROSI, Gabor *et al.* Clickstream-based outcome prediction in short video MOOCs. *In*: , 2018, Colmar, France. **International Conference on Computer, Information and Telecommunication Systems (CITS 2018)**. Colmar, France: IEEE, 2018. p. 1–5. Disponível em: https://ieeexplore.ieee.org/document/8440182.

LABARTHE, Hugues *et al.* Does a peer recommender foster students' engagement in MOOCs? *In*: , 2016, Raleigh, NC, USA. **9th International Conference on Educational Data Mining, (EDM 2016)**. Raleigh, NC, USA: [*s. n.*], 2016. p. 418–423. Disponível em: https://files.eric.ed.gov/fulltext/ED592665.pdf.

LAN, Andrew S. *et al.* Behavior-based latent variable model for learner engagement. *In*: , 2017, Hubei, China. **10th International Conference on Educational Data Mining (EDM 2017)**. Hubei, China: [*s. n.*], 2017. p. 64–71.

LANG, Charles *et al.* **Handbook of Learning Analytics**. 1. ed. [*S. l.*]: SOLAR - SOCIETY for LEARNING ANALYTICS RESEARCH, 2017. *E-book*. Disponível em: https://www.solaresearch.org/hla-17/.

LI, Chao; ZHOU, Hong. Enhancing the efficiency of massive online learning by integrating intelligent analysis into MOOCs with an Application to Education of

Sustainability. **Sustainability**, [*s. l.*], v. 10, n. 468, p. 1–16, 2018.

LIN, Jinjiao *et al.* Automatic Knowledge Discovery in Lecturing Videos via Deep Representation. **IEEE Access**, [*s. l.*], v. 7, p. 33957–33963, 2019.

LIÑÁN, Laura Calvet; PÉREZ, Juan Ángel Alejandro. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. **International Journal of Educational Technology in Higher Education**, [*s. l.*], v. 12, n. 3, p. 98–112, 2015. Disponível em: http://dx.doi.org/10.7238/rusc.v12i3.2515.

MILLER, CRAIG S.; LEHMAN, JILL FAIN; KOEDINGER, KENNETH R. Goals and learning in microworlds. **Cognitive Science**, [*s. l.*], v. 23, n. 3, p. 305–336, 1999.

MORALES, Miguel *et al.* Applying a Digital Learning Ecosystem to Increase the Effectiveness of a Massive Open Online Course. *In*: , 2019, Milwaukee, WI, USA. **IEEE Learning With MOOCS, (LWMOOCS 2019)**. Milwaukee, WI, USA: IEEE, 2019. p. 69–74. Disponível em: https://ieeexplore.ieee.org/document/8939636.

MULDNER, Kasia *et al.* An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. **User Modeling and User-Adapted Interaction**, [*s. l.*], v. 21, n. 1–2, p. 99–135, 2011.

NORTHCUTT, Curtis G.; HO, Andrew D.; CHUANG, Isaac L. Detecting and preventing "multiple-account" cheating in massive open online courses. **Computers and Education**, [*s. l.*], v. 100, p. 71–80, 2016. Disponível em: http://dx.doi.org/10.1016/j.compedu.2016.04.008.

PAPPANO, Laura. The Year of the MOOC. **The New York times**, New York, USA, 2 nov. 2012. p. 2012. Disponível em: https://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html.

PAQUETTE, Luc *et al.* Reengineering the Feature Distillation Process : A Case Study in the Detection of Gaming the System. *In*: , 2014. **7th International Conference on Educational Data Mining (EDM 2014)**. [*S. l.: s. n.*], 2014. p. 284–287.

PAQUETTE, Luc; BAKER, Ryan S. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. **Interactive Learning Environments**, [*s. l.*], v. 27, n. 5–6, p. 585–597, 2019. Disponível em: https://doi.org/10.1080/10494820.2019.1610450.

PAQUETTE, Luc; BAKER, Ryan S. Variations of gaming behaviors across populations of students and across learning environments. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, [*s. l.*], v. 10331 LNAI, p. 274–286, 2017.

PAQUETTE, Luc; BAKER, Ryan S.; MOSKAL, Michal. A system-general model for the detection of gaming the system behavior in CTAT and LearnSphere. *In*: PENSTEIN, Rosé et al. (org.). **Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science**. 1. ed. [*S. l.*]: Springer International Publishing, 2018. v. 10948 LNAI, p. 257–260. Disponível em: http://dx.doi.org/10.1007/978-3-319-93846-2_47.

PAQUETTE, L.; DE CARVALHO, A.; BAKER, R. S. Towards understanding expert coding of student disengagement in online learning. *In*: , 2014, Quebec, Canadá. **36th Annual Cognitive Science Conference**. Quebec, Canadá: Cognitive Science Society, 2014. p. 1126–1131.

PERIWAL, Nidhi; RANA, Keyur. An empirical comparison of models for dropout prophecy in MOOCs. *In*: , 2017, Greater Noida, India. **IEEE International Conference on Computing, Communication and Automation (ICCCA 2017)**. Greater Noida, India: [*s. n.*], 2017. p. 906–911.

PIGEAU, Antoine; AUBERT, Olivier; PRIÉ, Yannick. Success Prediction in MOOCs: A Case Study. *In*: , 2019, Montreal, Canada. **12th International Conference on Educational Data Mining (EDM 2019)**. Montreal, Canada: [*s. n.*], 2019. p. 390–395. Disponível em: https://files.eric.ed.gov/fulltext/ED599219.pdf.

PURSEL, B. K. *et al.* Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. **Journal of Computer Assisted Learning**, [*s. l.*], v. 32, n. 3, p. 202–217, 2016.

RAMOS, Altina; M. FARIA, Paulo; FARIA, Ádila. Revisão sistemática de literatura: contributo para a inovação na investigação em Ciências da Educação. **Revista Diálogo Educacional**, [*s. l.*], v. 14, n. 41, p. 17, 2014.

RODRIGUES, Rodrigo Lins *et al.* Discovery engagement patterns MOOCs through cluster analysis. **IEEE Latin America Transactions**, [*s. l.*], v. 14, n. 9, p. 4129–4135, 2016.

ROMERO, Cristóbal *et al.* **Handbook of Educational Data Mining**. 1. ed. Boca Raton, USA: CRC Press - Taylor & Francis, 2010. *E-book*. Disponível em: https://www.taylorfrancis.com/books/handbook-educational-data-mining-cristobal-romero-sebastian-ventura-mykola-pechenizkiy-ryan-baker/e/10.1201/b10274.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [*s. l.*], v. 3, n. 1, p. 12–27, 2013.

ROMERO, Cristbal; VENTURA, Sebastin. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews**, [*s. l.*], v. 40, n. 6, p. 601–618, 2010.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, [*s. l.*], v. 33, n. 1, p. 135–146, 2007.

ROMERO, Cristobal; VENTURA, Sebastian. Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [*s. l.*], v. 10, n. 3, p. 1–21, 2020.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data science in massive open online courses. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [*s. l.*], v. 7, n. 1, 2017.

RUIPEREZ-VALIENTE, Jose A. *et al.* Using Machine Learning to Detect “Multiple-

Account” Cheating and Analyze the Influence of Student and Problem Features. **IEEE Transactions on Learning Technologies**, [*s. l.*], v. 12, n. 1, p. 112–122, 2017.

RUIPEREZ-VALIENTE, Jose A. *et al.* Using multiple accounts for harvesting solutions in MOOCs. *In*: , 2016, Edinburgh, Scotland Uk. **3rd Conference on Learning at Scale (ACM - L@S 2016)**. Edinburgh, Scotland Uk: ACM, 2016. p. 63–70.

RUIPÉREZ-VALIENTE, José A. *et al.* A data-driven method for the detection of close submitters in online learning environments. *In*: , 2017, Perth, Austrália. **26th International World Wide Web Conference 2017 (WWW 2017)**. Perth, Austrália: [*s. n.*], 2017. p. 361–368.

SCHOFIELD, Janet Ward. **Computers and Classroom Culture**. 1. ed. Pittsburgh, USA.: Cambridge University Press, 1995. *E-book*. Disponível em: https://www.cambridge.org/core/books/computers-and-classroom-culture/16168247162BDF16A982D04D7FB84E7D.

SHAHIRI, Amirah Mohamed; HUSAIN, Wahidah; RASHID, Nur’Aini Abdul. A Review on Predicting Student’s Performance Using Data Mining Techniques. **Procedia Computer Science**, [*s. l.*], v. 72, p. 414–422, 2015. Disponível em: http://dx.doi.org/10.1016/j.procs.2015.12.157.

SHI, Yingnan *et al.* Knowledge pricing structures on MOOC platform - A use case analysis on edX. *In*: , 2018, Yokohama, Japão. **Twenty-Second Pacific Asia Conference on Information Systems (PACIS 2018)**. Yokohama, Japão: [*s. n.*], 2018. p. 1–13.

SILVA, Patricia Grasel Da; CARVALHO, Marie Jane Soares; TEIXEIRA, Adriano Canabarro. Study on MOOC for social learning. *In*: , 2018, São Paulo, Brasil. **3th Latin American Conference on Learning Technologies (LACLO 2018)**. São Paulo, Brasil: [*s. n.*], 2018. p. 444–449.

SOUZA, Vanessa Faria; PERRY, Gabriela. Identifying student behavior in MOOCs using Machine Learning. **International Journal of Innovation Education and Research**, [*s. l.*], v. 7, n. 3, p. 30–39, 2019. Disponível em: https://ijier.net/index.php/ijier/article/view/1318.

SOUZA, Vanessa Faria de; PERRY, Gabriela Trindade. Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático. **Revista Brasileira de Informática na Educação**, [*s. l.*], v. 28, n. 1, p. 491–508, 2020. Disponível em: https://www.br-ie.org/pub/index.php/rbie/article/view/v28p491/6730.

SUKHIJA, Karan; JINDAL, Manish; AGGARWAL, Naveen. The recent state of educational data mining: A survey and future visions. *In*: , 2015, Amritsar, India. **3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)**. Amritsar, India: IEEE, 2015. p. 354–359.

SUNAR, Ayse Saliha *et al.* How learners’ interactions sustain engagement: A MOOC case study. **IEEE Transactions on Learning Technologies**, [*s. l.*], v. 10, n. 4, p. 475–487, 2017.

SUNAR, Ayse Saliha *et al.* Modelling MOOC learners’ social behaviours. **Computers**

in Human Behavior, [*s. l.*], v. 107, p. 1–12, 2020. Disponível em: https://doi.org/10.1016/j.chb.2018.12.013.

TAIT, K.; HARTLEY, J. R.; ANDERSON, R. C. Feedback procedures in computer assisted arithmetic instruction. **British Journal of Educational Psychology**, [*s. l.*], v. 43, n. 2, p. 161–171, 1973. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8279.1973.tb00752.x.

TAN, Yueying *et al.* Learning Profiles, Behaviors and Outcomes: Investigating International Students' Learning Experience in an English MOOC. *In*: , 2018. **Proceedings - 2018 International Symposium on Educational Technology, ISET 2018**. [*S. l.*]: IEEE, 2018. p. 214–218.

WAHEED, Hajra *et al.* Predicting academic performance of students from VLE big data using deep learning models. **Computers in Human Behavior**, [*s. l.*], v. 104, p. 1–13, 2020.

WANG, Tai *et al.* Rating MOOCs: Implications from gamification. *In*: , 2018, Osaka, Japan. **Proceedings - 6th International Conference of Educational Innovation Through Technology (EITT 2017)**. Osaka, Japan: [*s. n.*], 2018. p. 142–147.

WANG, Ling; HU, Gongliang; ZHOU, Tiehua. Semantic analysis of learners' emotional tendencies on online MOOC education. **Sustainability**, [*s. l.*], v. 10, n. 6, p. 1–19, 2018. Disponível em: https://doi.org/10.3390/su10061921.

WEBLEY, Kayla. **Mooc brigade: Can online courses keep students from cheating?**. New York, USA, 2012. Disponível em: https://nation.time.com/2012/11/19/mooc-brigade-can-online-courses-keep-students-from-cheating/. Acesso em: 13 jul. 2020.

WEN, Yimin *et al.* Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. **Tsinghua Science and Technology**, [*s. l.*], v. 25, n. 3, p. 336–347, 2020.

XIAO, Bing; LIANG, Meiping; MA, Junliang. The application of cart algorithm in analyzing relationship of mooc learning behavior and grades. *In*: , 2018, Xi'an, China. **Proceedings - 2018 International Conference on Sensor Networks and Signal Processing (SNSP 2018)**. Xi'an, China: IEEE, 2018. p. 250–254.

XING, Wanli *et al.* Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. **Computers in Human Behavior**, [*s. l.*], v. 58, p. 119–129, 2016. Disponível em: http://dx.doi.org/10.1016/j.chb.2015.12.007.

YANG, Jian; ZHANG, Xiao Ling; SU, Peng. Deep-Learning-Based Agile Teaching Framework of Software Development Courses in Computer Science Education. **Procedia Computer Science**, [*s. l.*], v. 154, p. 137–145, 2018. Disponível em: https://doi.org/10.1016/j.procs.2019.06.021.

YOUSEF, Ahmed Mohamed Fahmy *et al.* The effect of peer aßeßment rubrics on learners' satisfaction and performance within a blended mooc environment. *In*: , 2015, Lisboa, Portugal. **7th International Conference on Computer Supported Education (CSEDU - 2015)**. Lisboa, Portugal: [*s. n.*], 2015. p. 148–159.

ZHANG, Jie. Can MOOCs be interesting to students? An experimental investigation from regulatory focus perspective. **Computers and Education**, [*s. l.*], v. 95, p. 340–351, 2016. Disponível em: http://dx.doi.org/10.1016/j.compedu.2016.02.003.

ZHANG, Yaling; WU, Bei. Research and application of grade prediction model based on decision tree algorithm. *In*: , 2019, Chengdu, China. **Turing Celebration Conference (ACM TURC 2019)**. Chengdu, China: ACM, 2019. p. 1–6. Disponível em: https://doi.org/10.1145/3321408.3322857.