UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA APLICADA

José Orlando da Silva Brandão

UMA ABORDAGEM COM *LEARNING ANALYTICS* E SÉRIES TEMPORAIS NA ANÁLISE DE DADOS EDUCACIONAIS

Dissertação de Mestrado

RECIFE
2018

José Orlando da Silva Brandão

UMA ABORDAGEM COM *LEARNING ANALYTICS* E SÉRIES TEMPORAIS NA ANÁLISE DE DADOS EDUCACIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Informática Aplicada da Universidade Federal Rural de Pernambuco como exigência parcial à obtenção do título de Mestre em Informática Aplicada.

Orientador: Prof. Dr. Adenilton José da Silva
Coorientadora: Profª. Drª. Roberta Macêdo Marques Gouveia

RECIFE
2018

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca Central, Recife-PE, Brasil

B817a Brandão, José Orlando da Silva.
Uma abordagem com learning analytics e séries temporais na análise de dados educacionais / José Orlando da Silva Brandão. – Recife, 2018.
97 f.: il.

Orientador(a): Adenilton José da Silva.
Coorientador(a): Roberta Macêdo Marques Gouveia.
Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2018.
Inclui referências e apêndice(s).

1. Procrastinação 2. Educação - Processamento de dados 3. Learning analytics 4. Análise de séries temporais I. Silva, Adenilton José da, orient. II. Gouveia, Roberta Macêdo Marques, coorient. III. Título

CDD 004

Dissertação de mestrado apresentada por **José Orlando da Silva Brandão** ao programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco, sob o título **Uma Abordagem com Learning Analytics e Séries Temporais na Análise de Dados Educacionais**, orientada pelo **Prof. Dr. Adenilton José da Silva**, coorientada pela **Profª. Drª Roberta Macêdo Marques Gouveia** e aprovada pela banca examinadora formada pelos professores:

_____

Prof. Dr. Adenilton José da Silva (DC/UFPE)
Presidente

_____

Prof. Dr. Rodrigo Gabriel Ferreira Soares (DEINFO/UFRPE)
Examinador Externo

_____

Prof. Dr. Rodrigo Lins Rodrigues (DEd/UFRPE)
Examinador Externo

DEDICATÓRIA

À minha querida esposa Luciana Brandão e a meus amados filhos Brenno Brandão e Anna Luiza Brandão, pelas demonstrações de amor, carinho e afeto durante essa minha longa caminhada no mestrado. Sem a compreensão, paciência e apoio de vocês eu não teria concluído este trabalho. Amo-os incondicionalmente.

Aos meus pais, Seu Paulo e Dona Gerodite; meus irmãos, Osmar, Josivaldo, Genivaldo e Paloma, pelos incentivos e palavras de apoio. Sei que mesmo distantes, estão sempre torcendo pelo meu sucesso, muito obrigado. Vocês são o meu porto seguro.

Aos meus sobrinhos e sobrinhas, Arthur, Caroline, Diego, Érika, Geovana e Gabriel, pela amizade, respeito e carinho que vocês têm como todos da família. Peço que continuem assim e perpetuem esse laço de amizade que une a família Brandão. Que esse trabalho, mesmo tardio, seja uma inspiração para todos vocês.

Ao meu sogro Jairo Mota Romeu e minha sogra Selma Trigueiro, minhas cunhadas Uemia Trigueiro, Elisângela Brandão, Jaciara Brandão e Ely Brandão, pelas palavras de incentivo, apoio e pela compreensão de algumas vezes não poder estar juntos de vocês, nesses últimos dois anos e meio, durante alguma comemoração da família. Muito obrigado por tudo.

AGRADECIMENTOS

Ao Meu Bom DEUS, fonte de inspiração em minha vida. Agradeço imensamente pela saúde, pela força, pela coragem e pela perseverança que me destes para conduzir este trabalho. Devo dizer-LHE que não foi fácil, mas o Senhor em SUA infinita bondade sempre me mostrou qual o melhor caminho para trilhar e concluir este mestrado. Agradeço-TE por tudo de bom que TENS feito na minha vida e na minha família.

Aos meus orientadores, Professores Dr. Adenilton José da Silva e Drª Roberta Macêdo Marques Gouveia, pela maestria com que conduziram essa orientação e pela paciência que tiveram com um mestrando que além de estudante é um trabalhador, e não podia dedicar-se exclusivamente à vida acadêmica. Meus sinceros agradecimentos e muito obrigado pela orientação.

Em especial, ao Professor Dr. Rodrigo Lins que, mesmo não sendo oficialmente meu orientador, apresentou orientações fundamentais para o sucesso deste trabalho. Meus sinceros agradecimentos e muito obrigado por tudo.

Aos meus Chefes e Companheiros de trabalho, pelo incentivo e pela força que sempre me deram durante esse período de vida acadêmica. Especialmente aos Coronéis Carneiro e Pessanha, Capitão Sílvio, Tenente Santanelli e Sargento Ernando. O apoio de vocês foi decisivo para o sucesso deste mestrado.

“Se você torturar os dados por tempo suficiente, eles confessarão.”

Ronald Harry Coase, economista britânico.

RESUMO

A modalidade de educação a distância (EaD), antes discriminada por estudantes dos mais variados seguimentos sociais, vem se firmando como uma excelente alternativa à educação tradicional. A sua evolução tem estreitos laços com os avanços em tecnologia da informação e comunicações. A expansão da banda larga para os lugares mais distantes do país, oferecendo acessos cada vez mais velozes à internet, favorece a disseminação de cursos de EaD oferecidos por instituições educacionais da iniciativa privada e pública. Essa mudança de paradigma na educação trouxe transformações no comportamento de gestores e professores, que cada vez mais fazem do uso da tecnologia para criação de conteúdos didáticos mais interessantes e interativos, bem como, no comportamento dos estudantes, que devem se adaptar à nova realidade, deixando de ser agentes passivos no processo educacional para se tornarem agentes ativos de sua própria aprendizagem, por meio de comportamentos de autorregulação. Para que ocorra a interação *on-line* entre estudantes e professor, é necessário a implantação de um ambiente virtual de aprendizagem (AVA), a exemplo do Moodle. Esse ambiente é fundamental para a comunicação síncrona e assíncrona entre os atores da EaD, armazenando em seu banco de dados todas as interações que estudantes, professores e tutores realizam durante as atividades *on-line*. Tais interações tornaram-se campo fértil para pesquisadores de mineração de dados educacionais e *learning analytics* estudarem o comportamento desses estudantes por meio de atributos derivados dessas interações. Neste contexto, esta pesquisa apresenta uma abordagem de aprendizado não supervisionado de máquina, com o algoritmo de agrupamentos *k-means,* para descobrir padrões de comportamentos de engajamento e procrastinação de estudantes de um curso de licenciatura a distância. As interações de estudantes e professores foram extraídas de arquivos de *logs* do Moodle, AVA utilizado pela Instituição de Ensino Superior que oferece o curso, e transformadas em atributos usados na criação das séries temporais que compõem o conjunto de dados de entrada do algoritmo de agrupamento. Encontrando como resultado grupos de estudantes com níveis baixo, intermediário e alto de engajamento, que apresentam relação entre o comportamento de procrastinação e o desempenho ao final da disciplina.

Palavras-chave: engajamento; procrastinação; mineração de dados educacionais; *learning analytics*; séries temporais

ABSTRACT

The mode of distance education, previously discriminated by students from a wide range of social segments, has been established as an excellent alternative to traditional education. Its evolution has close ties with advances in information technology and communications. The expansion of broadband to the most distant places in the country, offering fast access to the Internet, favors the dissemination of distance education courses offered by private and public companies. This paradigm shift in education has brought about transformations in the behavior of managers and teachers, who become increasingly dependent on the use of technology to create more interesting and interactive didactic content, as well as on student behavior, which should adapt to the newness, from being passive agents in the educational process to becoming active agents of their own learning through self-regulating behaviors. In order for online interaction between students and teachers to occur, it is necessary to implement a virtual learning environment (VLE), such as Moodle. This environment is fundamental for communication between the actors of the e-learning, storing in their database all the interactions that students, teachers and tutors perform during the activities online. Such interactions have become a fertile field for educational data mining researchers and learning analytics to study the behavior of these students through the attributes derived from these interactions. In this context, this research presents an approach of unsupervised learning of machines, through the algorithm of k-means clustering, to discover patterns of engagement behaviors and procrastination of students of a e-learning graduation course. Student and teacher interactions were extracted from Moodle log files, VLE used by the Institution of Higher Education that offers the course, being transformed into attributes used in the creation of the time series that compose the data set of input data of the clustering algorithm. Finding as results groups of students with low, intermediate and high levels of engagement that present a relationship between procrastination behavior and performance at the end of the course.

Keywords: engagement; procrastination; educational data mining; learning analytics; time series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

EaD	Educação a Distância
ABED	Associação Brasileira de Educação a Distância
TIC	Tecnologia da Informação e Comunicação
IES	Instituições de Ensino Superior
LDB	Lei de Diretrizes e Bases da Educação Nacional
AVA	Ambiente Virtual de Aprendizagem
MOODLE	Modular Object-Oriented Dynamic Learning Environment
LA	Learning Analytics
SOLAR	Society for Learning Analytics Research
PROINFO	Programa Nacional de Informática na Educação
UAB	Universidade Aberta do Brasil
e-TEC	Programa Escola Aberta do Brasil
IoT	Internet of Things
DW	Data Warehouse
OLAP	Online Analytical Processing
KNN	K-Nearest Neighbor
SOM	Self-OrganIzed Maps
MLP	MutiLayer Perceptron
JEDM	Journal of Educational Data Mining
EDM-TF	IEEE Task Force of Educational Data Mining
MOOC	Massive Open Online Courses
KDD	Knowledge Discovery in Databases
SSE	Sum of Squared Errors

Sumário

1 INTRODUÇÃO

A Educação a Distância (EaD) vem se afirmando como uma alternativa viável à educação na modalidade presencial em todo o mundo. No Brasil, essa é uma modalidade presente desde cursos profissionalizantes e técnicos até a educação de nível superior. O último Censo EaD de 2016 da Associação Brasileira de Educação a Distância (ABED)[1] informa que houve um crescimento de 65% de instituições privadas e 35% de instituições públicas (Censo EAD.Br, 2016). Segundo Baker *et al*. (2011), o Brasil é um dos países que possui os maiores crescimentos em número de cursos ofertados na modalidade de EaD.

A evolução na tecnologia de telecomunicações favorece a expansão da internet para os lugares mais distantes, alcançando uma quantidade maior de pessoas. Tal fato, associado às novas áreas de pesquisa de tecnologia da informação e comunicações (TIC) na educação (Iglesias-Pradas *et al.*, 2014; Rodrigues *et al.*, 2014), bem como, ao aumento da demanda de aprendizagem em todo planeta (Almeida *et al.*, 2013 *apud* Hanna, 2003) têm proporcionado um grande avanço na EaD, fazendo com que muitas Instituições de Ensino Superior (IES) ofereçam cursos de graduação nessa modalidade. Fenômeno observado pela ABED no seu relatório analítico de aprendizagem a distância no Brasil de 2016 (Censo EAD.Br, 2016).

Em 1996, com o lançamento da Lei de Diretrizes e Bases da Educação Nacional (LDB), o Governo Federal iniciou a regulamentação da educação a distância no Brasil. Essa mudança de paradigmas trouxe novos desafios para instituições, educadores e alunos, deixando-os “profundamente preocupados com o desenvolvimento de novos métodos e ferramentas que venham a melhorar o desempenho acadêmico dos alunos”, conforme Iglesias-Prada *et al.* (2014).

Segundo Almeida *et al.* (2013) “um dos principais fatores impeditivos ou complicadores do êxito de cursos a distância, seria a dificuldade apresentada pelos professores de manterem seus aprendizes atentos e engajados”, pois no ambiente virtual de aprendizagem (AVA) os estudantes possuem flexibilidade de horário e espaço para construção do seu próprio aprendizado, estando “circundado por um

1 http://www.abed.org.br

número maior de elementos distratores do que quando presente em uma sala de aula".

Outro comportamento bastante recorrente em alunos de cursos a distância é a procrastinação, caracterizada pela postergação na finalização de tarefas, exercícios e avaliações disponibilizados no AVA, sendo causada principalmente pela inabilidade do estudante em gerir o tempo dedicado ao seu estudo efetivo (TRUEMAN e HARTLEY, 1996). Para estudantes da EaD, tal capacidade pode ser considerada fator determinante para o sucesso ao final do curso, estando negativamente relacionada ao desempenho acadêmico (MICHINOV *et al.*, 2011 *apud* BADULF, 2009).

Dessa forma, há de se considerar o grande desafio de gestores, professores e tutores da EaD para mensurar os comportamentos de engajamento e procrastinação de estudantes a partir de um ambiente virtual de aprendizagem, que armazena em banco de dados todas as interações decorrentes das ações realizadas no AVA por aqueles atores do sistema. Tal volume de dados "fornece mais que informações gerenciais ou relatórios de desempenho" (Rodrigues *et al.*, 2014), sendo objeto de pesquisa para descoberta de conhecimentos úteis na área de Mineração de Dados Educacionais (MDE) e *Learning Analytics* (LA).

Diante desse contexto, surge o desafio de extrair atributos relevantes dos arquivos de *logs* gerados pelo AVA, bem como, identificar quais desses atributos possuem relação com os citados comportamentos. Assim, esta pesquisa apresenta um trabalho de mineração de dados educacionais de uma disciplina obrigatória do primeiro semestre do curso de licenciatura EaD, realizado por uma Instituição de Ensino Superior (IES) que utiliza o Moodle[2] (*Modular Object-Oriented Dynamic Learning Environment*) como plataforma virtual dos seus cursos de nível superior a distância.

Utilizou-se uma abordagem de agrupamento de séries temporais geradas por atributos extraídos das interações dos estudantes no AVA, com o objetivo de encontrar padrões dos comportamentos de engajamento do aluno com o curso e da

2 https://moodle.org

procrastinação acadêmica na finalização das atividades *on-line* obrigatórias, que influenciem o desempenho dos estudantes na disciplina.

1.1 Questões de pesquisa

Para desenvolver o presente trabalho, as seguintes questões de pesquisa foram elaboradas:

I) Como encontrar padrões comportamentais de engajamento e de procrastinação de estudantes a partir de um ambiente virtual de aprendizagem?

II) Quais atributos derivados de interações registradas em arquivos de *logs* desse ambiente virtual de aprendizagem melhor descrevem engajamento e procrastinação acadêmicos de estudantes?

1.2 Objetivos

1.2.1 Objetivo Geral

Apresentar uma abordagem de aprendizagem de máquinas para identificar grupos de estudantes com padrões de comportamento similares de engajamento e procrastinação que influenciem o desempenho acadêmico, a partir das interações registradas em arquivos de *logs* de um AVA, visando apoiar ações proativas de professores, tutores e gestores para o aperfeiçoamento contínuo de estratégias pedagógicas adequadas ao ambiente de educação a distância.

1.2.2 Objetivos Específicos

Tendo em vista a consecução do objetivo geral, são definidos os seguintes objetivos específicos:

- Realizar revisão da literatura na área de mineração de dados educacionais e *learning analytics* que contemplem a criação de atributos das interações de estudantes, professores e tutores derivados dos arquivos de *logs* do Moodle;
- Consolidar os arquivos de *logs* do Moodle com as planilhas de dados de histórico escolar provenientes do sistema de gestão acadêmica da Instituição de Ensino Superior pesquisada; e

- Extrair os atributos de interações dos arquivos de *logs* do Moodle considerados relevantes para a pesquisa, bem como, os atributos elencados pelos trabalhos relacionados, para a consolidação do conjunto de dados das séries temporais relativas às interações dos estudantes, professores e tutores.

1.3 Estrutura do Trabalho

Este trabalho de pesquisa acadêmica está estruturado em mais 5 capítulos:

No capítulo 2, Fundamentação Teórica, são apresentados os conceitos, definições e exemplos dos temas abordados na pesquisa, como educação a distância, os comportamentos de engajamento e procrastinação acadêmicos, mineração de dados educacionais, *learning analytics*, aprendizado de máquina e séries temporais.

No capítulo 3, Trabalhos Relacionados, encontram-se os trabalhos pesquisados e selecionados para essa dissertação, que possuem estreito relacionamento com os temas estudados. Também apresenta um mapeamento sistemático, cujo objetivo foi encontrar trabalhos de autores que utilizam conceitos de *leaning analytics* para criar atributos derivados das interações de estudantes no Moodle.

No capítulo 4, Metodologia e Experimentos, são apresentados os conceitos da metodologia de descoberta do conhecimento de bases de dados utilizada na pesquisa, bem como, a execução de suas etapas na realização dos experimentos. São relatados o pré-processamento dos dados, seleção de atributos, transformação dos dados educacionais, criação das séries temporais, mineração dos dados com o algoritmo de agrupamento *k-means* e validação do resultado.

No capítulo 5, Resultados e Discussões, são apresentados os resultados decorrentes da seção anterior e feitas as considerações quanto aos agrupamentos gerados pelo *k-means*, levando em consideração os comportamentos pesquisados: engajamento e procrastinação acadêmicos.

No capítulo 6, Conclusão, são apresentadas as conclusões a que chegaram esta pesquisa. E, as sugestões de Trabalhos Futuros, bem como as Limitações encontradas durante a realização da pesquisa estão no capítulo 7.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Educação a distância

A história da educação a distância no Brasil traça um paralelo com o avanço da tecnologia de comunicação de massa no nosso país. Kenski (2018) relata que ao contrário do que todos pensam, a primeira experiência de EaD no Brasil foi realizada pelas ondas do rádio e não por via impressa, como a utilizada pelo antigo Instituto Universal Brasileiro, que “oferecia ensino a distância em caráter supletivo, através de correspondências”. Seguindo o avanço tecnológico, na década de 70 a educação a distância passou a fazer uso do meio de comunicação televisivo, através dos Telecursos de 1º e 2º graus, que tinham um alcance nacional.

Com o grande avanço das redes de telecomunicações através do uso de fibras óticas, entregando serviços de comunicação de dados cada vez mais velozes e alcançando os lugares mais distantes do país, pode-se dizer que a educação a distância deixou de ser uma inovação ou uma tendência e passou a ser tratada como uma forma de ensino regular, tal qual a educação na modalidade presencial. Assim, a Lei de Diretrizes e Bases da Educação Nacional (LDB), em 1996, legisla pela primeira vez sobre a educação a distância no Brasil, conceituando-a como:

> […] a modalidade educacional na qual mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempo diversos.

Desde então o Governo Federal vem normatizando a EaD no país para instituições da iniciativa privada ou pública, estendendo inicialmente essa modalidade para o ensino fundamental, médio e profissionalizante. Atualmente, já existe legislação que regulamenta a educação a distância para os níveis de graduação e pós-graduação no nível especialização e, segundo Kenski (2018), “os programas de mestrado e doutorado na modalidade a distância, no Brasil, ainda são objeto de regulamentação específica”.

Foram várias as iniciativas do poder público para favorecer o acesso, principalmente aos professores da rede pública, a essa nova modalidade de ensino. Dentre elas, segundo Portal (2016), destacam-se o Programa Nacional de

Informática na Educação - PROINFO, a Universidade Aberta do Brasil – UAB e o Programa Escola Aberta do Brasil – e-TEC.

Segundo o site do Ministério de Educação, o PROINFO foi criado em 1997 com o objetivo de promover o uso da tecnologia como ferramenta de enriquecimento pedagógico no ensino público fundamental e médio. Trata-se do fornecimento de equipamentos de informática, acesso à internet e conteúdos educacionais para professores e alunos daquele nível de educação, onde o município ou o estado, em contrapartida, devem garantir a infraestrutura adequada dos laboratórios de informática e capacitar os professores para utilização da tecnologia fornecida.

Segundo Portal (2016), o programa da UAB representa uma das principais políticas públicas no campo da educação de ensino superior a distância no Brasil. A UAB é um sistema integrado por universidades públicas que oferece cursos de nível superior a distância para o público em geral. Porém, o site do Ministério da Educação enfatiza que a prioridade é para os professores da educação básica pública que ainda não possuem o nível superior, assim como, para dirigentes, gestores e outros profissionais daquele nível de educação. O programa procura diminuir as desigualdades na oferta de cursos do ensino superior, ampliando a interiorização e democratização do ensino superior público e gratuito.

Para estender a modalidade da EaD para a área do ensino técnico e profissionalizante, o Governo Federal criou em 2007, o Programa Escola Aberta do Brasil, cujos pressupostos, segundo Barbosa *et al.* (2012), foram “a democratização e o acesso ao ensino técnico dos jovens das periferias dos grandes centros urbanos e dos municípios do interior do país”. Pacheco e Maia (2010) destaca que é notória a implantação do e-TEC no contexto atual da educação brasileira, enfatizando que o programa representa uma mudança de paradigma na busca de “soluções para ofertar a sociedade brasileira uma importante e diversificada lista de cursos no universo da educação profissional”.

Na modalidade de educação a distância, alunos e professores não se encontram no mesmo ambiente físico, é necessário a criação de um sistema virtual que faça a intermediação entre esses agentes. A necessidade de expansão de programas de incentivo a EaD, bem como, o avanço tecnológico facilitou a

implementação desse ambiente virtual de aprendizagem. Atualmente existem várias plataformas que possibilitam um ambiente dinâmico e interativo, permitindo que professores possam criar conteúdos didáticos atrativos, interessantes e atuais para estudantes da EaD cada vez mais àvidos por novos conhecimentos e tecnologias.

Nos tópicos seguintes são apresentadas algumas ferramentas e temas que exploram essa importante área, as quais foram utilizadas nesta pesquisa.

2.2 Engajamento e procrastinação de estudantes

O engajamento de estudantes com o curso ao qual estão matriculados vem sendo pesquisado já há algum tempo. Inicialmente as pesquisas eram focadas em "estudantes do ensino fundamental e médio, onde a falta de engajamento normalmente se torna uma preocupação" (Willms *et al.*, 2009). Segundo Trowler (2010), no contexto da educação superior, a literatura passou a dar uma "considerável atenção ao tema em meados da década de 1990, apesar do início dos estudos uma década antes".

Desde então, surgiram várias definições para o termo engajamento de estudantes. Segundo Krause *et al.* (2008), o engajamento é a qualidade do esforço que os alunos dedicam a atividades educacionais, que contribuem diretamente para os resultados desejados. Chen *et al.* (2008) enfatiza que engajamento mostra o grau no qual os estudantes estão envolvidos com suas atividades educacionais, estando positivamente ligado a resultados, como notas altas, satisfação e perseverança. Adicionalmente, Beer *et al.* (2010) *apud* Stoval (2003) descreve que o engajamento pode ser definido pelo tempo que o aluno gasta ao executar uma tarefa e sua vontade de participar das atividades.

Parson e Taylor (2011) enfatizam que, historicamente, as estratégias para melhorar o engajamento foram implementadas primariamente em estudantes em riscos de reprovação ou evasão. Mas, atualmente o conceito está ligado à participação ativa dos estudantes na realização de tarefas e no contexto educacional que os envolvem, estando intimamente ligado ao sucesso acadêmico podendo,

inclusive, ser utilizado como um indicador da qualidade de ensino de uma instituição (BEER *et al.*, 2010 *apud* KUH, 2001).

Alguns pesquisadores descrevem o engajamento como um fenômeno multidimensional (Parson e Taylor, 2011; Rodrigues *et al.*, 2016), que deve ser estudado sobre os vários aspectos educacionais, comportamentais e psicológicos. Nesse sentido, Fredericks *et al*. (2004) classificou o engajamento em três categorias: comportamental, emocional e cognitivo.

Resumindo Fredericks *et al*. (2004), o engajamento comportamental é tido como a participação do estudante em atividades acadêmicas, sociais e extracurriculares. O engajamento emocional aparece quando o estudante tem atitudes e reações positivas em relação à universidade, professores, outros estudantes e com o aprendizado. O engajamento cognitivo é caracterizado pelo investimento pessoal do estudante em uma aprendizagem focada, estratégica e autorregulada.

Outro comportamento que pode levar a um deficiente gerenciamento do processo de aprendizagem é a procrastinação acadêmica. Tema bastante pesquisado no contexto de aprendizado autorregulado, como em (Rothblum *et al.*, 1986; Klassen *et al.*, 2008; Kim e Seo, 2015, Cerezo *et al.*, 2017), a procrastinação acadêmica é definida como uma “tendência para postergar uma atividade sob controle até o último minuto possível ou até mesmo para não realizá-la” (GAFNI e GERI, 2010).

A maioria das pesquisas que relacionam o tema procrastinação a desempenho acadêmico, utilizam uma abordagem subjetiva por meio da medição de uma escala de procrastinação predefinida por perguntas em questionários, que o estudante é solicitado a responder após o término do curso. As perguntas vão desde “Em geral, o quanto a procrastinação influencia negativamente suas funções acadêmicas?” até “Quais os tipos de tarefas nas quais você mais procrastina, normalmente?” (KLASSEN *et al.*, 2008).

Uma das escalas de medida de procrastinação mais utilizadas em pesquisas sobre o tema, segundo Kim e Seo (2015), é a *Procrastination Assessment Scale-Students* (*PASS*; Solomon e Rothblum, 1984), que “consiste de perguntas para os

estudantes relatarem a frequência com que eles procrastinam". Tuckman (1991) criou a *Tuckman Procrastination Scale* (*TPS*) que avalia a "procrastinação acadêmica como uma incapacidade do estudante de autorregular ou controlar os horários das tarefas". Essas escalas trazem uma abordagem de correlação negativa entre procrastinação e desempenho ou nota final do aluno.

Em sentido contrário às abordagens das escalas *PASS* e *TPS*, Choi e Moran (2009) apresentaram a *Active Procrastination Scale* (*APS*) que contrabalanceia as várias pesquisas neste contexto acadêmico e introduz o conceito de procrastinação ativa "sugerindo efeitos positivos da procrastinação" (GODA *et al.*, 2015). Eles identificaram quatro dimensões que caracterizam a procrastinação ativa: capacidade para alcançar resultados satisfatórios, preferência pela pressão do tempo, decisão intencional pela procrastinação e a capacidade em cumprir prazos com esse comportamento.

Para Rotenstein *et al.* (2009), essas abordagens de comportamento postergante dos estudantes através de medidas autorrelatadas de procrastinação "baseadas nas respostas de alunos, são frequentemente medidas fracas da procrastinação real". Eles utilizaram as variáveis data de início e data de finalização das tarefas no AVA para averiguar o relacionamento entre procrastinação e desempenho de estudantes nas tarefas *on-line* de aulas de contabilidade de um curso de graduação.

2.3 Mineração de dados educacionais e *Learning Analytics*

Han *et al.* (2012) enfatizam que a mineração de dados é a "evolução natural da tecnologia da informação", tal evolução surge da necessidade cada vez maior de armazenamento de grandes quantidades de dados gerados por sistemas, pessoas, equipamentos e, mais recentemente, com o advento da Internet das Coisas, até por produtos elétricos, como uma geladeira, por exemplo.

Ainda segundo Han *et al.* (2012), embora as ferramentas de *On-Line Analytical Processing* (OLAP) suportem análises multidimensionais e tomadas de decisão, para uma análise "mais profunda" dos dados, necessita-se de outras ferramentas de

mineração que sejam capazes de tarefas como predição, classificação ou agrupamento de dados, para a descoberta de algum padrão que se encontra escondido nos dados e as ferramentas usuais não conseguem encontrar. É neste ínterim que se inserem as técnicas de mineração de dados.

No contexto da educação, a mineração de dados educacionais surge como uma nova área de pesquisa, abrangendo o desenvolvimento, pesquisa e aplicação de métodos computacionais para detectar padrões em base de dados educacionais, com o objetivo de melhor entender como o aluno aprende, segundo Romero e Ventura (2013). E, através desse entendimento, melhorar o seu desempenho educacional previamente ao término do curso, aplicando outros métodos didáticos, personalizando o ambiente virtual para proporcionar melhores condições de aprendizagem.

De acordo com Baker e Yacef (2009), a MDE surgiu como uma área de pesquisa para analisar dados oriundos de ambientes educacionais para resolver problemas de pesquisa educacional. Romero e Ventura (2013) corroboram com essa afirmativa, acrescentando que tal análise é realizada com a aplicação de técnicas de mineração de dados tradicionais, como classificação, agrupamento, modelagem *bayesiana*, dentre outras. Ainda, segundo Baker e Yacef (2009), a MDE é uma combinação das áreas ciência da computação, educação e estatística, conforme mostrado na Figura 1.

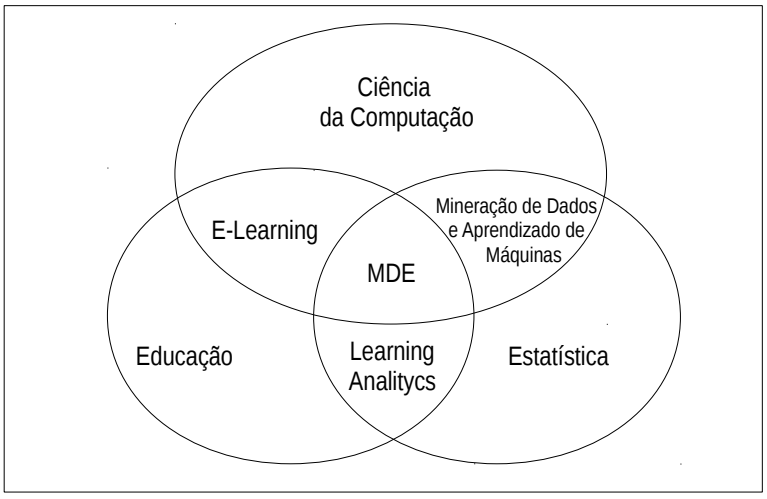

Figura 1 – Principais áreas relacionadas à mineração de dados educacionais
Adaptado de Romero e Ventura (2013)

O tema *Learning Analytics* surge neste contexto como uma área de estudo que, segundo Romero e Ventrura (2013), de todas as áreas surgidas da interseção da ciência da computação, educação e estatística, a área de *Learning Analytics* é a que está mais relacionada com a mineração de dados educacionais. Sendo definida pela *Society for Learning Analytics Research* (SoLAR)[3] como a “medição, coleta, análise e comunicação de dados sobre estudantes e seus contextos, para compreender e otimizar a aprendizagem e o ambiente em que ela ocorre” (SoLAR, 2011; Siemens e Baker, 2012; Khalil e Ebner, 2015; Ramos *et al.*, 2015).

2.4 Aprendizado de Máquina

Aprendizado de máquina é uma área da Inteligência Artificial (IA) que objetiva adquirir conhecimento automático através do aprendizado computacional. O conceito está relacionado com o aprendizado humano, onde programas são criados com o objetivo de aprender automaticamente determinados comportamentos a partir de exemplos ou observações (RUSSELL e NORVIG, 2013).

Segundo Mitchell (1997), o aprendizado de máquina é inerentemente um campo multidisciplinar, que se baseia em resultados das áreas de inteligência artificial, probabilidade e estatística, teoria da complexidade computacional, teoria da informação, filosofia, psicologia, neurobiologias, dentre outros. Han *et al.* (2012) enfatizam que essa é uma área de pesquisa onde os programas aprendem automaticamente a reconhecer padrões complexos e tomar decisões inteligentes baseadas nos dados.

Para Russel e Norvig (2013), o aprendizado de máquina é utilizado em diversas aplicações, destacando-se o reconhecimento de padrões de imagens médicas para auxiliar na identificação de determinadas doenças, reconhecimento de padrões de faces para identificação em meio à multidão ou em banco de dados de faces de pessoas procuradas na justiça, mineração de textos para identificação de padrões de assinaturas, classificação de e-mails indesejáveis ou de notícias falsas, identificação de fraudes financeiras, dentre muitas outras áreas.

3 https://solaresearch.org/

Conforme Rezende (2005), existe uma hierarquia do aprendizado, onde o aprendizado de máquina utiliza a indução, cuja inferência lógica permite a “obtenção de conclusões genéricas sobre um conjunto particular de exemplos”. Assim, na indução o conhecimento é adquirido efetuando-se a inferência sobre um determinado conjunto de dados, sendo que as hipóteses levantadas, podem ou não ser verdaeiras (REZENDE, 2005).

O aprendizado indutivo divide-se em aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi-supervisionado, segundo Han *et al.* (2012) e Rezende (2005). A Figura 2 exibe essa classificação.

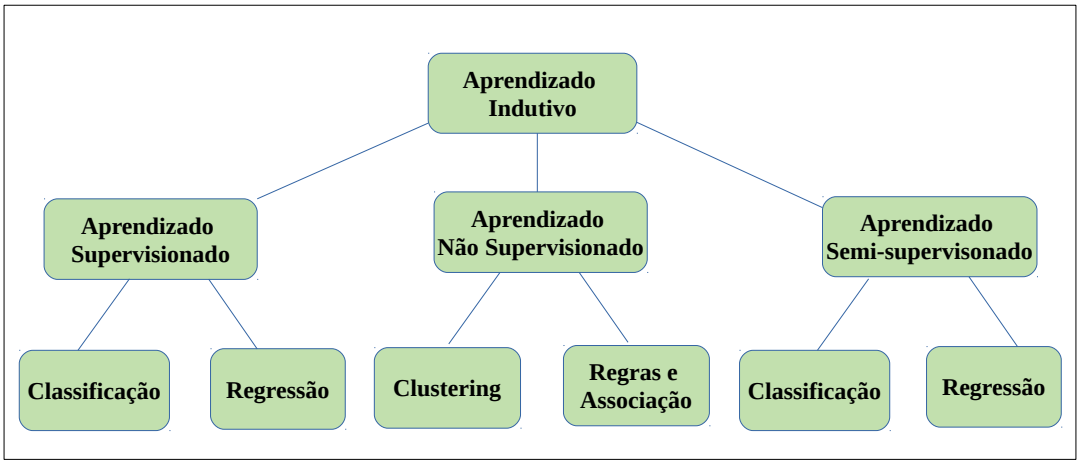

Figura 2- Hierarquia de aprendizado de máquina.
Adaptado de Rezende (2005) e Han *et al.* (2012)

No aprendizado supervisionado, as observações ou exemplos do conjunto de dados são acompanhadas por rótulos, que caracterizam as classes às quais os exemplares pertencem. Quando esses rótulos possuem um conjunto finito de valores (como bom, ótimo ou excelente), ou seja, apresentam valores discretos, o problema de aprendizagem será chamado de classificação, e será denominado de problema de classificação binária se houver apenas dois valores. Quando os rótulos são compostos por valores contínuos (como a temperatura do dia) o problema de aprendizagem supervisionada é chamado de regressão (RUSSELL e NORVIG, 2013).

No aprendizado não supervisionado, os atributos do conjunto de dados não possuem rótulos ou classes. O conjunto de observações a ser processado no aprendizado não supervisionado é utilizado com o propósito de se estabelecer a

existência de grupos de exemplares potencialmente úteis para o problema em questão. A tarefa mais comum para esse tipo de aprendizado é o agrupamento e a associação.

O aprendizado supervisionado objetiva construir um classificador (indutor) que possa determinar a classe de novos exemplos a partir das observações do conjunto de treinamento com classes rotuladas, enquanto no aprendizado não supervisionado, o indutor analisa os atributos do conjunto de dados sem rótulos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando os grupos ou *clusters* (REZENDE, 2005).

O aprendizado semi-supervisionado utiliza conceitos das outras duas técnicas de aprendizado de máquina. Segundo Russell e Norvig (2013), são fornecidos alguns poucos exemplares com os rótulos definidos e o indutor deve “fazer o que puder de uma grande coleção de exemplos não rotulados”. Ou seja, o aprendizado semi-supervisionado une as duas principais características dos outros dois aprendizados para propor um processo de classificação.

Existem várias técnicas usadas no aprendizado supervisionado de máquina, dentre eles destacam-se as árvores de decisão, regressão linear e logística, classificação bayesiana, redes neurais artificias e as máquinas de vetores de suporte, os quais são usados na predição de classes de atributos utilizando as técnicas de classificação ou regressão. Segundo Peña-Ayala (2014), os algoritmos mais utilizados no aprendizado não supervisionado são *k-means*, *fuzzy c-means*, *k-medoid*, *k-nearest neighbor* (*KNN*) e mapas auto-organizáveis.

2.5 Séries Temporais

Segundo Box e Jenkins (1994), uma série temporal pode ser entendida como um conjunto de observações sobre uma ou mais variáveis obtidas sequencialmente ao longo do tempo, podendo ser contínua, quando as observações são obtidas em qualquer intervalo de tempo, ou discretas, quando as observações forem obtidas em intervalos de tempos discretos ou equidistantes. Ehlers (2007) enfatiza que,

enquanto nos modelos de regressão a ordem é irrelevante para a análise, “em séries temporais a ordem dos dados é crucial”.

Uma série temporal pode ser representada por $Z_t = (Z_1 , Z_2 , Z_3 ,..., Z_n)$, onde Z corresponde à variável observável e $t = (1, 2, 3, .., m)$ equivale ao parâmetro de tempo, que determina o tamanho da série. Além disso, uma série temporal pode ser classificada como determinística, quando seus valores podem ser escritos por uma função matemática do tempo $y = f(tempo)$, ou estocástica, quando existe um componente aleatório (e) associado a essa função do tempo $y = f(tempo, e)$.

Muitas são as áreas onde os conceitos de séries temporais são mais comumente empregados, alguns exemplos são as previsões do tempo, de valores de ações de empresas nas bolsas de valores, de índices de correções diários ou mensais como o IPC-A (Índice Nacional de Preço ao Consumidor Amplo), de números mensais de novo casos de doenças ou de taxas mensais de desemprego num país. Conforme visto nos exemplos apresentados, o conceito de séries temporais está muito ligado à previsão, Box e Jenkins (1994) destaca que existe um sistema causal relacionado com o tempo, que exerceu ou exerce influência nos dados no passado ou presente e podem continuar a fazer o mesmo com os dados no futuro.

Conforme Ehlers (2007), ao se analisar uma ou mais séries temporais a representação gráfica dos dados sequencialmente ao longo do tempo é fundamental e pode revelar padrões comportamentais importantes. Essa abordagem descritiva de séries temporais pode ser feita através da decomposição clássica, método de análise de séries temporais cujos principais componentes são: (I) tendência, que são os movimentos subjacentes que caracterizam a evolução do nível médio da série; (II) a sazonalidade, que correspondem aos movimentos estritamente periódicos; (III) os ciclos, que caracterizam os movimentos oscilatórios do tipo recorrente e (III) o componente de irregularidade, que são os movimentos aleatórios de natureza imprevisível (MORETTIN e TOLOI, 2004).

3 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos que possuem relação com esta pesquisa e com os temas por ela abordados, como mineração de dados educacionais, *learning analytics*, séries temporais e comportamento de estudantes da EaD.

Em uma das pesquisas mais recentes sobre a relação entre engajamento e desempenho final de estudantes de cursos a distância, Li e Baker (2018) identificaram quatro subgrupos "(*disengagers*, *auditors*, *quiz-takers* e *all-rounders*)" com padrões de comportamentos distintos em três MOOCs (*Massive Open Online Courses*) da plataforma Coursera[4].

Li e Baker (2018) utilizaram as interações dos alunos com os vídeos disponibilizados nos cursos (pausas, avanços, recuos e buscas) como medidas para inferir o engajamento cognitivo. Quanto ao engajamento comportamental, utilizaram as interações dos alunos com a resolução dos questionários dos cursos. Para encontrar a relação entre engajamento e desempenho dos alunos, as pesquisadoras conduziram uma análise de regressão para predição das notas dos alunos em cada um dos quatro subgrupos. Os resultados da pesquisa indicam que a mesma medida de engajamento pode estar opostamente associada ao desempenho de diferentes subgrupos de estudantes e que algumas medidas predizem o desempenho para um subgrupo e para outros não.

Com o objetivo de analisar a efetividade do primeiro MOOC de língua portuguesa realizado no Brasil, Rodrigues *et al.* (2016a) utilizaram MDE para encontrar padrões de engajamento de alunos não só do Brasil, mas também da Colômbia, Portugal e Argentina. Foram definidas como medidas de engajamento comportamental quinze variáveis referentes à "frequência de diferentes categorias de postagens, assiduidade e notas", extraídas do ambiente virtual de aprendizagem.

Rodrigues *et al.* (2016a) utilizaram a técnica de análise de agrupamentos hierárquico e não-hierárquico para encontrar grupos de estudantes com padrões de engajamento semelhantes. Como resultado, identificaram três grupos distintos de estudantes que representam as categorias de engajamento enquadradas como engajados, aqueles que realizam atividades do início ao fim do curso e possuem

4 https://www.coursera.org

pouca variabilidade; esporádicos, os que passam longos períodos sem entrar no ambiente e perdem alguns prazos de realização de atividades; desengajados, formado por 90% dos alunos que realizam poucas atividades e possuem notas inferiores aos demais grupos.

Khalil *et al.* (2016) empregaram técnicas de *Learning Analytics* para retratar o engajamento de estudantes do MOOC Aspectos Sociais de Tecnologia da Informação dos cursos de Ciência da Computação e Engenharia da Computação. Eles utilizaram a técnica de agrupamentos com o algoritmo *k-means* para identificar grupos de estudantes com comportamentos comuns, usando variáveis extraídas de leitura/postagem do fórum de discussão, vídeos assistidos pelos estudantes e quantidade de tentativas de resolução dos questionários do curso.

Como resultado do emprego da técnica de agrupamento, Khalil *et al.* (2016) encontraram quatro grupos de estudantes com comportamentos de engajamento distintos. Os grupos 1 e 3 foram classificados como “evasivos” por serem muito inativos durante o curso e apresentarem baixa taxa de certificação no curso; o grupo 4 foi nomeado como “sociáveis” pois foi formado por estudantes com muito alta atividade nos fóruns; o grupo 2 foi chamado de “perfeitos” porque foi agrupado com estudantes com alto comprometimento nas aulas e altas taxas de certificação no curso.

Para encontrar padrões de engajamento de estudantes em MOOCs da Universidade de Stanford oferecidos pela plataforma Coursera, Anderson *et al*. (2016) propuseram um *framework* no ambiente on-line que atribui um cartão com tonalidades bronze, prata, ouro ou diamante para o estudante como incentivo para aumentar as atividades nos fóruns dos cursos. Para caracterização dos “estilos” de engajamento, os autores extraíram do AVA variáveis derivadas das interações dos estudantes com a visualização das aulas, resolução de exercícios, questionários e avaliações, bem como, de participação dos fóruns.

Usando técnica de agrupamento, Anderson *et al.* (2016) apresentaram como resultado cinco grupos com “estilos distintos de engajamento”, classificados como (I) *Viewers*, estudantes que assistem às aulas e fazem poucos exercícios; (II) *Solvers*, alunos que resolvem alguns exercícios, porém assistem poucas ou nenhuma aula;

(III) *All-rounders*, formado por alunos com características dos dois grupos anteriores; (IV) *Colletors*, estudantes que baixam as aulas e resolvem poucos exercícios; (V) *Bystanders*, grupo de estudantes que se matriculam nos cursos, mas realizam pouquíssimas atividades no AVA.

Com o aumento de MOOCs oferecendo os mais variados tipos de cursos, tornou-se comum o mesmo estudante está matriculado em diferentes MOOCs de uma mesma instituição. Segundo Kovanovic *et al.* (2016) esse é um “campo fértil” para o estudo do comportamento de estudantes da EaD. Os autores utilizaram os dados de 11 diferentes MOOCs da Universidade de Edinburgh ofertados pela plataforma Coursera.

Kovanovic *et al.* (2016) empregaram a técnica de análise de agrupamento com o *k-means*, utilizando um conjunto de dados composto por variáveis extraídas das interações dos estudantes com os módulos do AVA, como quantidade de exercícios submetidos, quantidade de mensagens lidas e postadas nos fóruns, quantidade de tentativas de resolução dos questionários, quantidade de vídeos assistidos, dentre outras.

Dessa forma, Kovanovic *et al.* (2016) encontraram cinco grupos de estudantes com comportamentos distintos, classificados como (I) Matriculados, composto por estudantes matriculados que não realizam qualquer atividade no ambiente virtual; (II) Engajamento Baixo, aqueles que realizam poucas atividades; (III) Vídeos, grupo com estudantes que primariamente assistem vídeos; (IV) Vídeos e Questionários, agrupam estudantes engajados em assistir vídeos e resolver exercícios; (V) Social, grupo caracterizado por estudantes que participam ativamente dos fóruns de discussão.

Outro tema que tem despertado o interesse dos pesquisadores da EaD é a procrastinação de estudantes durante o período de entrega dos exercícios, questionários e outras atividades propostas do ambiente virtual. A procrastinação é caracterizada pela tendência do estudante em postergar a entrega da tarefa até os momentos finais da data de vencimento da atividade, sendo que muitos estudantes de cursos a distância extrapolam essa data, fazendo a entrega das tarefas em dias posteriores, podendo até mesmo deixar de realizá-las (GAFNI e GERI, 2010).

Klassen *et al.* (2008) conduziram dois estudos para explorar a procrastinação de 456 estudantes de graduação. No primeiro estudo, os autores utilizaram algumas escalas (e.g. Tuckman, 1991; Pintrich *et al.*, 1993; Rosenberg, 1979 e Zimmerman, 1992) que mensuram as chamadas “auto” variáveis de estudantes da EaD, para encontrar a média, desvio-padrão e coeficiente alpha das variáveis de autorregulação, auto-eficácia, auto-estima, auto-eficácia da autorregulação, procrastinação, utilizadas no estudo. Na sequência, utilizaram regressão múltipla hierárquica para encontrar a correlação entre a procrastinação e as outras variáveis.

No segundo estudo da pesquisa, Klassen *et al.* (2008) utilizaram mais duas escalas (e.g. Ackerman e Gross, 2005; Solomon e Rothblum, 1984) para criar as variáveis de procrastinação diária e procrastinação na tarefa, para em conjunto com as “auto” variáveis citadas anteriormente predizer o impacto da procrastinação nos estudantes. Eles chegaram a conclusão que a procrastinação tem influência negativa em alguns estudantes e que a variável da auto-eficácia da regulação é o mais forte preditor da tendência à procrastinação acadêmica.

Segundo Kandemir (2014) estudantes que não mostram bons desempenhos no processo de aprendizagem falham por causa do comportamento procrastinante durante o curso e, entender as razões que levam a tal comportamento, pode ser a chave para o sucesso acadêmico. O autor utilizou várias escalas para criar as variáveis autorrelatadas que foram utilizadas na pesquisa como auto-eficácia, autorregulação, satisfação de vida, esperança, ausência do ambiente virtual, uso da internet, procrastinação e média de sucesso acadêmico.

Kandemir (2014) também usou regressão múltipla hierárquica para determinar o nível predição da procrastinação acadêmica dos estudantes em termos das auto-variáveis mencionadas. O autor chegou à conclusão que as variáveis média de sucesso acadêmico, autorregulação, auto-eficácia, esperança e satisfação de vida possuem correlação negativa com a procrastinação, enquanto as variáveis uso da internet e ausência do ambiente virtual possuem correlação positiva com a procrastinação.

Para Michinov *et al.* (2011) a procrastinação acadêmica tem forte influência no processo de aprendizado de estudantes de cursos a distância e que isso pode ser

explicado pelo nível de participação do estudante nos fóruns de discussão do curso. Os autores administraram no curso a escala de procrastinação de Tuckman (Tuckman, 1991) para criar as seguintes variáveis autorrelatadas: vontade de (re)começar a trabalhar remotamente, vontade de desistir do curso, motivação para trabalhar remotamente, procrastinação, participação nos fóruns e desempenho no curso.

Para análise da correlação das variáveis com a procrastinação, os autores usaram o coeficiente de correlação de Pearson, onde concluíram que a procrastinação está relacionada negativamente com a participação nos fóruns e, consequentemente, com o desempenho final dos estudantes no curso. Adicionalmente, Michinov *et al.* (2011) analisaram que existe correlação positiva entre participação nos fóruns de discussão e desempenho final, sugerindo que quanto mais os estudantes participam dos fóruns, melhores são seus desempenhos finais.

Em contrapartida aos trabalhos que utilizaram medidas indiretas para predição da procrastinação acadêmica, Rotenstein *et al.* (2009) usaram variáveis obtidas diretamente do AVA, para averiguar o relacionamento entre procrastinação e desempenho de estudantes nas tarefas *on-line* de aulas de contabilidade de um curso de graduação. Os autores enfatizaram que as medidas autorrelatadas de procrastinação “baseadas nas respostas de alunos, são frequentemente medidas fracas da procrastinação real”.

Rotenstein *et al.* (2009) utilizaram os coeficientes de correlação de Pearson, coeficiente de correlação de Spearman e regressão linear múltipla, utilizando como variáveis a nota das tarefas, a nota final do estudantes no curso, a data de início de realização das tarefas e a data de finalização das mesmas. O resultado da pesquisa alinhou-se com o de outros trabalhos, que sugerem que a procrastinação acadêmica possui correlação negativa com o desempenho de estudantes de cursos a distância.

A abordagem de séries temporais realizada por Santos *et al.* (2016) deu-se com a divisão do período semestral de oferta da disciplina escolhida em oito períodos de quinze dias, o qual contêm duas avaliações presenciais. Eles realizaram uma seleção de atributos através do “método de cápsula (*wrapper*)*”* para compor as

séries e foram aplicando os algoritmos de classificação em cada um dos “cortes temporais” de forma incremental, onde os dados de um período era incrementado pelos dados dos períodos antecedentes, objetivando avaliar o desempenho dos estudantes através das citadas etapas.

Mlynarska *et al.* (2017) utilizaram o conceito de agrupamento de séries temporais para determinar se os “agrupamentos produzidos com tal abordagem produzem grupos de estudantes que compartilham comportamentos e desempenhos semelhantes”. Utilizaram dados de interações de alunos de treze cursos presenciais de Ciência da Computação, que usavam o Moodle para submissão de tarefas, fórum de discussão e *download* de materiais didático.

As séries temporais criadas por Mlynarska *et al.* (2017) possuíam três semanas de *timeline,* compreendendo “duas semanas antes da data de submissão das tarefas e uma semana após esse *deadline”*, sendo divididas em períodos de 12 horas contendo todas as interações dos estudantes no Moodle. Eles usaram *k-means* na clusterização das séries, com a medida de distância *Dynamic Time Warping* (*DTW*) e valor de *K* igual a 4, encontrando 7 padrões de comportamento entre os estudantes.

Os trabalhos citados estudam o fenômeno do engajamento de estudantes da EaD de forma independente ou isolada, em cursos no estilo MOOCs, que possuem uma abrangência demográfica de alcance mundial. A proposta deste trabalho é fazer um estudo para classificar estudantes por níveis de engajamento (e.g. Li e Baker, 2018; Rodrigues *et al.*, 2016; Khalil *et al.,* 2016; Kovanovic *et al.,* 2016) e, posteriormente, analisar a relação do comportamento de procrastinação acadêmica (e.g. Rotenstein *et al.,* 2009) desses mesmos estudantes com o desempenho final na disciplina de um curso “fechado” do tipo semi-presencial, com momentos a distância e momentos presenciais, disponível para um universo mais restrito de pessoas, de um estado ou região do país.

3.1 Mapeamento Sistemático

Este mapeamento sistemático foi realizado com a finalidade de fazer um levantamento dos atributos criados por pesquisadores de MDE e LA a partir da fonte de dados dos arquivos de *logs* do Moodle, adquiridos pela interface gráfica ou diretamente do banco de dados desse AVA. O objetivo desta atividade foi consolidar em um único lugar a maior quantidade possível de atributos mencionados na literatura congênere, para servir de referência para esta pesquisa.

Definiu-se como espaço temporal dos trabalhos a serem revisados o período de 2013 a 2017, para busca de pesquisas mais recentes que tenham abordado os temas mineração de dados educacionais e *learning analytics*. Utilizou-se as seguintes *strings* de busca: *educational data mining and learning analytics and moodle*. Foram utilizadas as bases acadêmicas a partir da base de dados dos periódicos da Capes, que redirecionou várias pesquisas para bases como Springer, ACM Digital Library, Science Direct, IEEE e Google Scholar. A Tabela 1 exibe uma amostra de seis ações registradas no banco de dados do Moodle referentes às interações realizadas por alunos, professores ou tutores na disciplina de primeiro semestre do curso de licenciatura em estudo.

Tabela 1- Ações do AVA Moodle

AÇÃO	DESCRIÇÃO
assign submit	Aluno submeteu um trabalho no AVA
chat talk	Aluno, Professor ou Tutor postou uma mensagem no Chat
course view	Aluno, Professor ou Tutor fez um login no AVA
forum add	Professor criou um Fórum no AVA
forum add post	Aluno, Professor ou Tutor postou uma mensagem no Fórum
quiz attempt	Aluno fez uma tentativa de responder o questionário, exercício ou prova

A Tabela 2 exibe uma amostra dos atributos criados a partir das ações de *logs* do Moodle, mencionados pelos vários pesquisadores de MDE e LA, resultantes deste mapeamento sistemático. A planilha completa com todos os atributos pesquisados encontra-se no Apêndice A.

Tabela 2- Atributos criados por outros pesquisadores

ATRIBUTO	DESCRIÇÃO	UTILIZADOS POR
Logins de acesso	Quantidade de logins realizadas pelo aluno	Khalil e Ebener (2015); Iglesias-Prada et al. (2015); Joksimovic et al. (2015); Gasevic et al. (2015); Palmer (2013); Martin e Whitmer (2016); Gottardo et al. (2012); Santos et al. (2016); Rodrigues et al. (2017);
Postagens em Fóruns	Quantidade de postagens em fóruns	Greller e Drachsler (2012); Khalil e Ebener (2015); Iglesias-Prada et al. (2015); Agudo-Peregrina et al. (2014); Joksimovic et al. (2015); Gasevic et al. (2015); Zacharis (2015); Palmer (2013); Anderson et al.
Postagens em Chats	Quantidade de postagens em chats	Agudo-Peregrina et al. (2014); Gasevic et al. (2015); Palmer (2013); Gottardo et al. (2012); Santos et al. (2016);
Tarefas concluídas	Quantidade de tarefas submetidas	Khalil e Ebener (2015); Joksimovic et al (2015); Gasevic et al (2015); Zacharis (2015); Anderson et al. (2014); Santos et al. (2016); Rodrigues et al. (2017);
Frequência de logins	Frequência diária com que o aluno acessa o AVA	Joksimovic et al. (2015); Zacharis (2015); Gottardo et al. (2012);
Tempo total online	Tempo total gasto nas sessões *web online* no AVA durante o semestre	Joksimovic et al. (2015); Palmer (2013); Martin e Whitmer (2016); Gottardo et al. (2012); Rodrigues et al. (2017);

O próximo Capítulo trata da metodologia adotada na pesquisa, bem como dos experimentos realizados para alcançar os objetivos propostos.

4 METODOLOGIA E EXPERIMENTOS

4.1 Contexto

Esta pesquisa utiliza conceitos de aprendizado de máquina não-supervisionado, com uma abordagem na tarefa de agrupamento (*clustering*) de dados, objetivando encontrar grupos de estudantes com padrões comportamentais caracterizados pelo nível de engajamento durante o curso e pela procrastinação acadêmica na finalização das atividades realizadas no ambiente virtual de aprendizagem.

Os dados são oriundos do histórico escolar fornecidos pelo sistema de gestão acadêmica da IES e dos arquivos de *logs* do Moodle, plataforma utilizada pelo curso. São dados referentes a 420 estudantes de uma disciplina de 1º período de um curso de licenciatura na modalidade a distância, gerados em formato de planilha. Tendo em vista que a pesquisa utiliza o conjunto de dados levando em consideração as características temporais de *logs* do Moodle, utilizou-se apenas algumas variáveis do histórico escolar dos estudantes: nome, matrícula, desempenho e resultado final.

Dessa forma, após a revisão sistemática da literatura, que resultou na planilha de atributos constantes do Apêndice A, criou-se os novos atributos que foram selecionados e ajustados no formato de série temporal para serem utilizados pelo algoritmo de partição *k-means*. O algoritmo agrupou os dados em 7 *clusters* que foram então validados e, posteriormente, interpretados de acordo com os padrões comportamentais de engajamento.

A presente pesquisa utiliza *softwares* de uso livre, como o aplicativo de planilhas LibreOffice Calc, a linguagem de programação Python com o conjunto de bibliotecas do Pandas e a linguagem de programação R Statistic com os pacotes TSdist, k-means, factoextra, mlbench, caret, dentre outros, fazendo uma abordagem metodológica através do processo de descoberta de conhecimento em bases de dados, descrito nos próximos tópicos.

4.2 Processo de descoberta de conhecimentos em bases de dados

A metodologia utilizada nesta pesquisa segue o processo *Knowledge Discovery in Databases* (KDD) apresentado por Frawley *et al.* (1992), amplamente utilizado em pesquisas de mineração de dados, sendo definido como um “processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados, de forma a produzir conhecimentos” (QUEIROGA *et al.*, 2017 apud FAYYAD *et al.*, 1996). A Figura 3 abaixo mostra as etapas que compõem o processo.

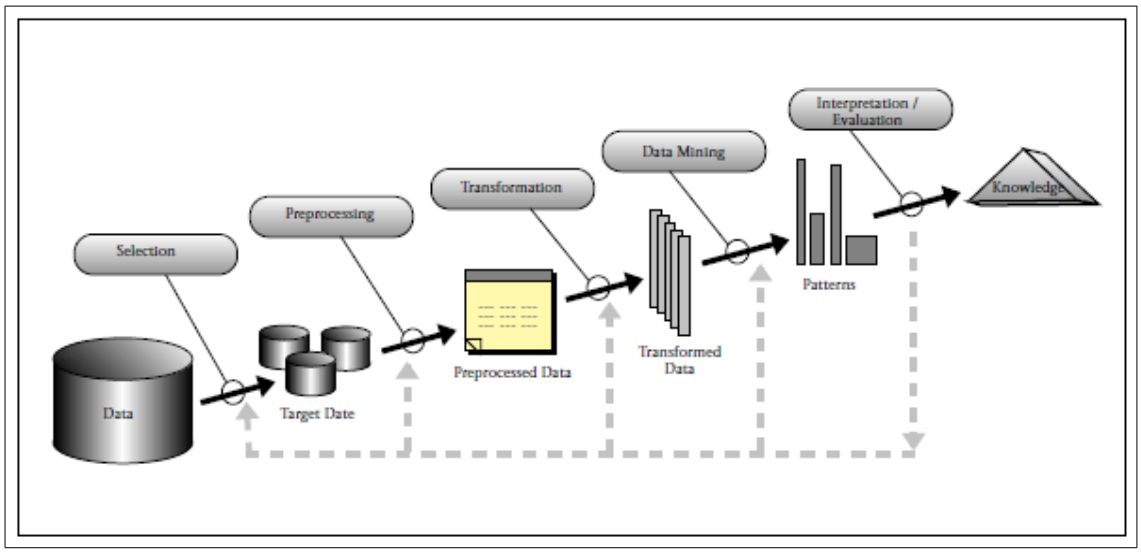

Figura 3 – Processo KDD (Autores: Fayyad *et al.*, 1996)

A fase de seleção de dados consiste da extração de dados de uma determinada base, que sejam relevantes para a pesquisa; o pré-processamento dá-se com a organização dos dados em uma base livre de inconsistências, como dados faltosos e atípicos; transformação é a fase de adequação dos dados para serem utilizados pelo algoritmo de mineração de dados escolhido na pesquisa; na etapa de mineração é realizada a identificação de padrões que se encontravam ocultos nos dados; por último, a interpretação/avaliação consiste da análise do pesquisador sobre os resultados, conduzindo-o à descoberta de conhecimentos sobre os dados em questão. Esta é uma definição resumida das etapas do processo de KDD, que serão mais detalhadas nos tópicos seguintes.

Seleção de dados educacionais

Esta fase do processo de KDD pode ser considerada complexa, dependendo do tipo de fonte disponibilizada para coleta dos dados utilizados na pesquisa, pois pode ser necessário desenvolver algum código para realizar a extração desses dados ou criar um *script* SQL, por exemplo.

Neste contexto, cabe ressaltar que os dados referentes aos históricos escolares dos estudantes, oriundos do sistema de gestão acadêmico da IES do período de julho de 2013 a dezembro de 2014,foram selecionados e formatados em planilhas. Conforme citado, devido a pesquisa fazer uma abordagem de *clustering* em séries temporais, apenas algumas variáveis desse conjunto de dados foram utilizadas, objetivando o conhecimento do domínio e interpretação dos resultados apresentados ao final do processo.

A Tabela 3 abaixo mostra as variáveis desse conjunto de dados que foram utilizadas na pesquisa, com seus tipos, descrições e categorias.

Tabela 3- Dados de histórico escolar usados na pesquisa

VARIÁVEL	TIPO	DESCRIÇÃO	CATEGORIAS
MATRÍCULA	Numérica	Número da matrícula do aluno	420 Matrículas
NOME	Categórica	Nome do aluno	420 Nomes de Alunos
DESEMPENHO	Categórica	Desempenho final do aluno	Insuficiente, Regular, Bom, Ótimo e Excelente
RESULTADO	Binária	Define a aprovação ou não do aluno	Aprovado e Reprovado

A classificação da variável DESEMPENHO foi feita a partir da média final dos estudantes na disciplina, da seguinte maneira: (I) Insuficiente – de 0 a 4,99; (II) Regular – de 5,00 a 6,49; (III) Bom – de 6,50 a 7,99; (IV) Ótimo – de 8,00 a 8,99; (V) Excelente – de 9,00 a 10,00.

Os dados referentes à segunda fonte de dados, *logs* do Moodle, foram extraídos através da sua interface gráfica, totalizando em 799 planilhas referentes aos estudantes de uma disciplina de curso EaD, sendo 263 planilhas referentes aos alunos da Turma 2013.2, 276 planilhas referentes à Turma 2014.1 e 260 planilhas da

Turma 2014.2. Resultando em 213.743 registros de interações realizadas pelos estudantes, professores e tutores no ambiente virtual de aprendizagem.

A Tabela 4 abaixo exibe uma parte de uma dessas planilhas geradas pelo Moodle com cinco ações realizadas no AVA pelo(a) Estudante 23 (os dados referentes ao nome do curso, nome do(a) estudante e URL da IES foram alterados para preservar a identidade dos estudantes).

Tabela 4- Exemplo de uma planilha de *log* do Moodle

CURSO	HORA	ENDEREÇO IP	NOME COMPLETO	AÇÃO	INFORMAÇÃO
Licenciatura	04/11/2014 19:15:25	X.X.X.X	Estudante 23	course view (http://xxx/view.php?id=1469)	Disciplina do 1º período de curso EaD
Licenciatura	03/11/2014 19:23:01	X.X.X.X	Estudante 23	forum view discussion (http://xxx/mod/forum/discuss.php?d=84498)	Deixe sua contribuição neste tópico!
Licenciatura	03/11/2014 19:22:04	X.X.X.X	Estudante 23	forum view forum (http://xxx/mod/forum/view.php?id=121268)	Desafio 4 (Fórum de Discussão)
Licenciatura	02/11/2014 15:03:35	X.X.X.X	Estudante 23	quiz view (http://xxx/mod/quiz/view.php?id=118599)	Desafio 2 (Questionário)
Licenciatura	02/11/2014 19:23:01	X.X.X.X	Estudante 23	assign view (http://xxx/mod/assign/view.php?id=116051)	Ver própria página de status de envio.

Pré-processamento de dados educacionais

A etapa de pré-processamento é fundamental para o processo de descoberta de conhecimento em mineração de dados, pois, a qualidade dos dados resultantes desta etapa vai determinar a eficiência do algoritmo de mineração escolhido para a pesquisa. Em geral, os dados chegam para o especialista com algum tipo de “ruído”

(dados incompletos, errados, ausentes, desatualizados ou inconsistentes) e precisam ser “limpos” e organizados.

Tendo em vista que os dados dos históricos escolares dos estudantes das três turmas encontravam-se consolidados em uma planilha, com pouca incidência de dados ausentes, incompletos ou inconsistentes, assim como, com inexistência de dados fora dos padrões, as atividades realizadas nesta fase concentraram-se na consolidação das 799 planilhas de *logs* do Moodle, que resultou em três planilhas referentes aos estudantes das três turmas.

A parte final da integração das informações das duas fontes de dados deu-se com a junção de cada uma das planilhas do Moodle referentes às três turmas, com a planilha consolidada dos dados de histórico escolar. Resultando em três planilhas, uma para cada turma, contendo apenas os alunos que constavam concomitantemente das duas fontes de dados. Cabe ressaltar, que com essa operação as interações de professores e tutores foram excluídas do conjunto de dados.

A realização desta atividade resultou na planilha da Turma 1 contendo 143 estudantes, da Turma 2 com 137 estudantes e da Turma 3 com 140 estudantes, que foram consolidas em uma única planilha com informações de 420 estudantes. É importante frisar que essa atividade deve ser feita com uma constante revisão dos resultados para evitar inconsistências no conjunto de dados final.

Em seguida, criou-se uma nova coluna com codificação referente a cada um dos estudantes objetivando preservar-lhes a identidade, sendo excluída a coluna com os nomes e matrículas.

Análise exploratória dos dados

Antes de partir para a próxima fase de transformação dos dados, convém fazer uma análise exploratória dos dados com a finalidade de obter-se um entendimento mais aprofundado de aspectos importantes do conjunto de dados explorado na pesquisa. Nesta primeira análise, foram considerados os dados referentes ao histórico escolar. O Gráfico 1 mostra os percentuais dos tipos de variáveis desse conjunto de dados.

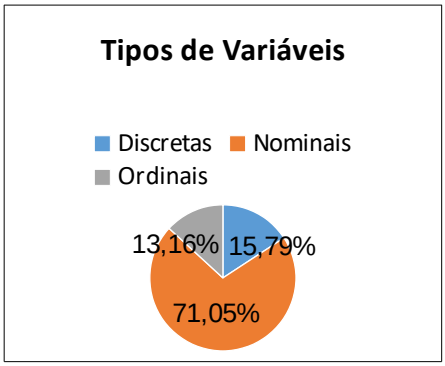

Gráfico 1– Tipos de variáveis do conjunto de dados histórico escolar

Como pode ser visto, a grande maioria das variáveis desse conjunto de dados é nominal, pouco mais de 71%. As variáveis discretas e ordinais possuem quase o mesmo percentual dentro do conjunto de dados, 15,79% e 13,16% respectivamente. O Gráfico 2 mostra a quantidade de alunos que ingressaram em cada turma, os percentuais de aprovados e reprovados ao final da disciplina.

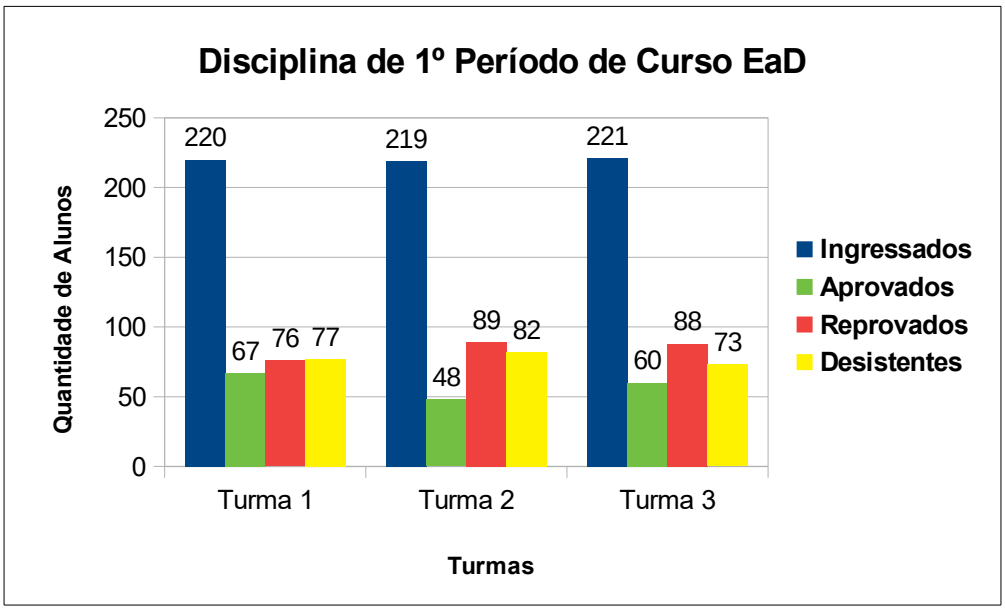

Gráfico 2- Disciplina de primeiro período de curso EaD

É possível perceber que, para as três turmas, existem muitos alunos reprovados, seja por nota ou por falta, bem como bastantes alunos desistentes.

Transformação de dados educacionais

Segundo Han *et al.*, 2012, nesta fase os dados devem ser transformados, formatados, derivados ou consolidados adequadamente para que os algoritmos possam ser aplicados com eficiência.

Nesta etapa são criados os atributos resultantes da revisão sistemática da literatura descritos no item 3.1. A Figura 4 apresenta uma parte do *script* Python utilizado para criar os doze primeiros atributos quantitativos consolidados na Tabela 6, utilizando as bibliotecas *pandas*, *datetime* e *dateutil*.

```
6  df = pd.read_csv("datasetfinal.csv", encoding='latin-1', low_memory=False)
7  df['DATA'] = pd.to_datetime(df['DATA'], format = '%d/%m/%Y')
8
9  acoes = pd.read_csv("acoes.csv", encoding='latin-1', low_memory=False)
10
11 for i in acoes.index:
12     atributo = acoes.loc[i:'ATRIBUTO']
13     list_acoes = acoes.loc[i:'ACAO']
14     l = list_acoes['ACAO'].values
15     l = [l.split(',') for l in l][0]
16     ds = df.loc[df['ACAO'].isin(l)]
17     ds1 = ds.groupby(['NOME'], as_index=False)['ACAO'].count()
18     ds1.to_csv(atributo['ATRIBUTO'].values[0]+ '.csv')
```

Figura 4 – Script de criação dos atributos derivados do Moodle

Inicialmente o *script* faz a leitura do arquivo *datasetfinal.csv* (linha 6) que contém todas as interações dos alunos no Moodle, em seguida converte a coluna *DATA* para o tipo *datetime.* O próximo arquivo lido *acoes.csv* (linha 9) contêm duas colunas, uma chamada *ATRIBUTO*, que contém os nomes dos atributos a serem criados, e a outra com nome *ACAO*, que contêm os nomes das ações do Moodle que devem ser extraídas para compor tal atributo. Por exemplo, para criação do atributo *QuestTentv*, que armazena a quantidade de tentativas feitas pelos estudantes para resolver os questionários, a coluna referente à *ACAO* possui os seguintes valores: *quiz attempt* e *quiz continue attempt*.

Por fim, o laço *for* cria um objeto do tipo Lista contendo os valores da coluna *ACAO* do arquivo *acoes.csv* (linha 14) comparando-os, através do método *isin* do *dataframe* (linha 16), com os valores da coluna *ACAO* do arquivo *datasetfinal.csv*. Depois, agrupa o *dataframe* por NOME e faz a contagem das referidas ações (linha 17), salvando em arquivo *csv* (linha 18) com o nome do atributo criado.

O arquivo *datasetfinal.csv* utilizado neste código é uma planilha do tipo *csv* equivalente à Tabela 4, que passou por um pré-processamento, perdendo a coluna INFORMAÇÃO, e mantendo a coluna ACAO apenas com o registro da ação realizada pelo estudante. A Tabela 5 abaixo mostra o arquivo *acoes.csv* usado nesta parte de criação dos atributos.

Tabela 5- Exemplo do arquivo acoes.csv

ATRIBUTO	ACAO
TarefEnviadas	assign submit
TarefAcessadas	assign view, assign view all, asign view grade form, assign view submit assignment form
RecurAcessados	book print, book print chapter, book view, book view all, book view chapter, resource view, resource view all
MsgEnvChat	chat talk
MsgLidasChat	chat view, chat view all, chat report
Logins	course view
ArqBaixados	folder view, folder view all, imscp view all
TopAddForum	forum add discussion
MsgEnvForum	forum add post, forum update post
MsgLidasForum	forum view discussion, forum view forum, forum view forums, forum user report
QuestFinalz	quiz close attempt
QuestTentv	quiz attempt, quiz continue attempt

Dessa forma, criou-se os doze primeiros atributos constantes da Tabela 6, sendo uma planilha para cada atributo contendo os nomes dos alunos e a quantidade de ações realizadas por eles para aquele atributo. Posteriormente, as doze planilhas foram consolidadas em apenas uma, a qual foi integrada com os sete atributos restantes, criados individualmente.

A Figura 5 abaixo apresenta outro *script* utilizado para criar o atributo *DiasAcesso* para cada aluno, o atributo faz a contagem da quantidade de dias únicos (distintos) que o aluno acessou o AVA.

```
22  df = pd.read_csv("datasetfinal.csv", encoding='latin-1', low_memory=False)
23  df['DATA'] = pd.to_datetime(df['DATA'], format = '%d/%m/%Y')
24
25  aggregations = dict()
26  aggregations['DATA'] = 'nunique'
27  resultado = df.groupby(['NOME'], as_index = False).agg(aggregations)
28  resultado.to_csv('AtrbDiasAcesso-T1.csv')
```

Figura 5- Script de criação do atributo DiasAcesso

Mais uma vez o arquivo *datasetfinal.csv* é lido pelo método *read_csv* do *pandas* (linha 22) e armazenado em um objeto do tipo *dataframe*, a coluna *DATA* é convertida para o tipo *datetime* (linha 23), depois cria-se um objeto do tipo Dicionário com a função *aggregations* com o objetivo de agregar o *dataframe* pela coluna *DATA* com valores únicos (distintos) (linhas 25 e 26). Depois o *dataframe* é agrupado por *NOME* e agregado com a quantidade de dias pelo método *agg* (linha 27), finalizando com a criação do arquivo *csv* referente ao atributo extraído.

Os atributos *InicioAcesso* e *PrimUltimLogin* também foram criados usando *scripts* em Python. Os atributos referentes a acessos por período do dia, *AcessoManha*, *AcessoTarde*, *AcessoNoite* e *AcessoMadrug* foram criados com funções do LibreOffice Calc. A Tabela 6 abaixo apresenta esses 19 atributos que foram criados a partir dos dados originais de *logs* do Moodle, gerados em forma de planilhas.

Tabela 6- Atributos criados nesta pesquisa

ATRIBUTO	OBJETO DO MOODLE	DESCRIÇÃO
Logins	Course	Quantidade de acessos realizadas no curso
ArqBaixados	Folder	Quantidade de arquivos baixados no curso
MsgEnvChat	Chat	Quantidade de mensagens enviadas em chats
MsgEnvForum	Forum	Quantidade de mensagens enviadas em fóruns
MsgLidasChat	Chat	Quantidade de mensagens lidas em chats
MsgLidasForum	Forum	Quantidade de mensagens lidas em fóruns
TopAddForum	Forum	Quantidade de tópicos adicionados em fóruns
QuestFinalz	Quiz	Quantidade de questionários/exercícios/provas finalizados
QuestTentv	Quiz	Quantidade de tentativas para resolver questionários/exercícios/ provas
TarefEnviadas	Assign	Quantidade de tarefas enviadas no AVA
TarefAcessadas	Assign	Quantidade de vezes que o aluno acessou as tarefas
RecursAcessados	Book, Resources, Imscp	Quantidade de recursos multimídias acessados no AVA
InicioAcesso	Course	Quantidade de dias para fazer o primeiro acesso
DiasAcesso	Course	Quantidade de dias distintos que o aluno acessou o AVA
AcessoManha	Course	Quantidade de vezes que o aluno acessou o AVA de manhã
AcessoTarde	Course	Quantidade de vezes que o aluno acessou o AVA à tarde
AcessoNoite	Course	Quantidade de vezes que o aluno acessou o AVA à noite
AcessoMadrug	Course	Quantidade de vezes que o aluno acessou o AVA de madrugada
PrimUltimLogin	Course	Quantidade de dias entre o primeiro e o último login no AVA

A maioria dos atributos constantes do Apêndice A, criados por outros pesquisadores de mineração de dados educacionais, foram extraídos através de acesso direto ao banco de dados do Moodle ou através de *surveys*, entrevistas ou questionários *on-line* realizados com os estudantes dos cursos em questão. Tais opções são determinantes para criação dos vários atributos encontrados na revisão sistemática da literatura deste trabalho, pois ampliam o escopo de fontes de dados disponíveis para os pesquisadores.

Séries Temporais

Durante a condução do mapeamento sistêmico da literatura observou-se que, apesar dos arquivos de *logs* do Moodle apresentarem características de temporalidade, com uma variável que registra a data, hora, minuto e segundo das interações realizadas no AVA, existiam poucos trabalhos na literatura que tratavam o

tema mineração de dados educacionais através do conceito de séries temporais, sendo essa uma das contribuições importantes desta pesquisa.

Tendo em vista que uma série temporal pode ser definida como uma sequência de realizações ou observações de uma variável ao longo do tempo e tendo sido criados dezenove atributos com essa característica, utilizou-se uma ferramenta para avaliar a importância de cada atributo criado com o desempenho e o resultado final dos alunos.

Para isso, utilizou-se a biblioteca R Caret, acrônimo para ***C**lassification **A**nd **RE**gression **T**raining* com o objetivo de validar a importância estatística dos atributos com relação ao desempenho final dos estudantes. Essa biblioteca é um *wrapper* para mais de duzentos algoritmos de aprendizado supervisionado de máquina, possuindo diversas funcionalidades, dentre elas, métodos para seleção de atributos. A seleção é feita através do ranqueamento da importância dos atributos em relação a um atributo-alvo.

A Figura 6 abaixo é o resultado do processamento da biblioteca, configurada com o classificador *K-Nearest Neighbors* (*KNN*), mostrando o *ranking* de importância dos atributos criados com relação ao resultado final (aprovado/reprovado) dos alunos, considerado como atributo classe.

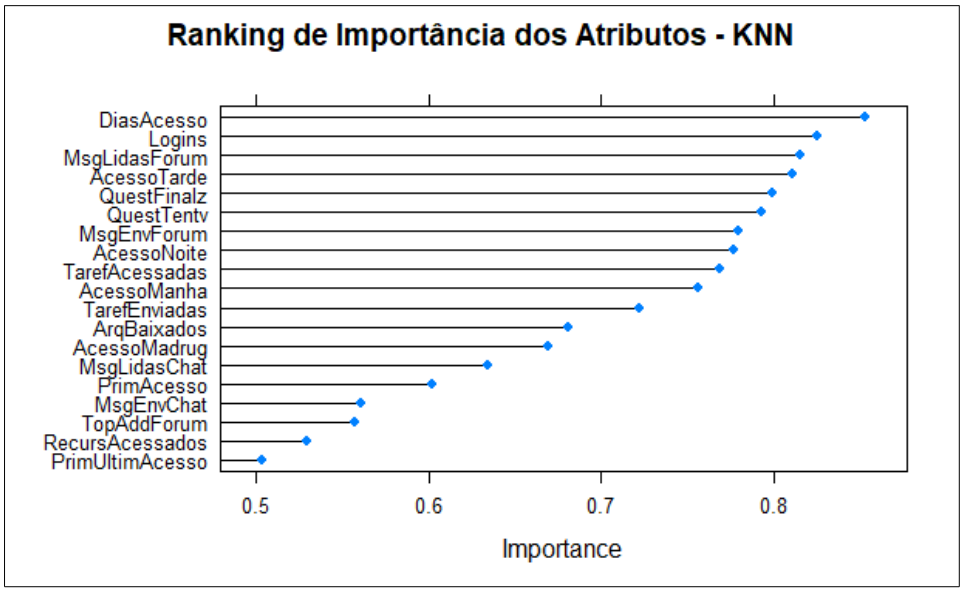

Figura 6– Importância dos atributos em relação ao resultado final

Neste primeiro teste com a configuração descrita acima, nota-se que o atributo *DiasAcesso*, que armazena a quantidade de dias distintos que o estudante acessou o Moodle, tem a importância mais alta, seguida pela quantidade de *Logins*. Verifica-se também que os atributos referentes aos objetos de aprendizagem questionários, tarefas e fórum possuem importância com valores acima de 70%, são eles, *TarefEnviadas*, *TarefAcessadas* e *MsgEnvForum*, sendo que os atributos *MsgLidasForum, QuestFinalz* e *QuestTentv* possuem valores acima de 80% de importância em relação ao resultado final.

No segundo teste, configurou-se a biblioteca com o classificador *Random Forest* (*RF*). A Figura 7 mostra o resultado da importância dos atributos em relação ao desempenho final dos estudantes (insuficiente, regular, bom, ótimo e excelente) como atributo classe.

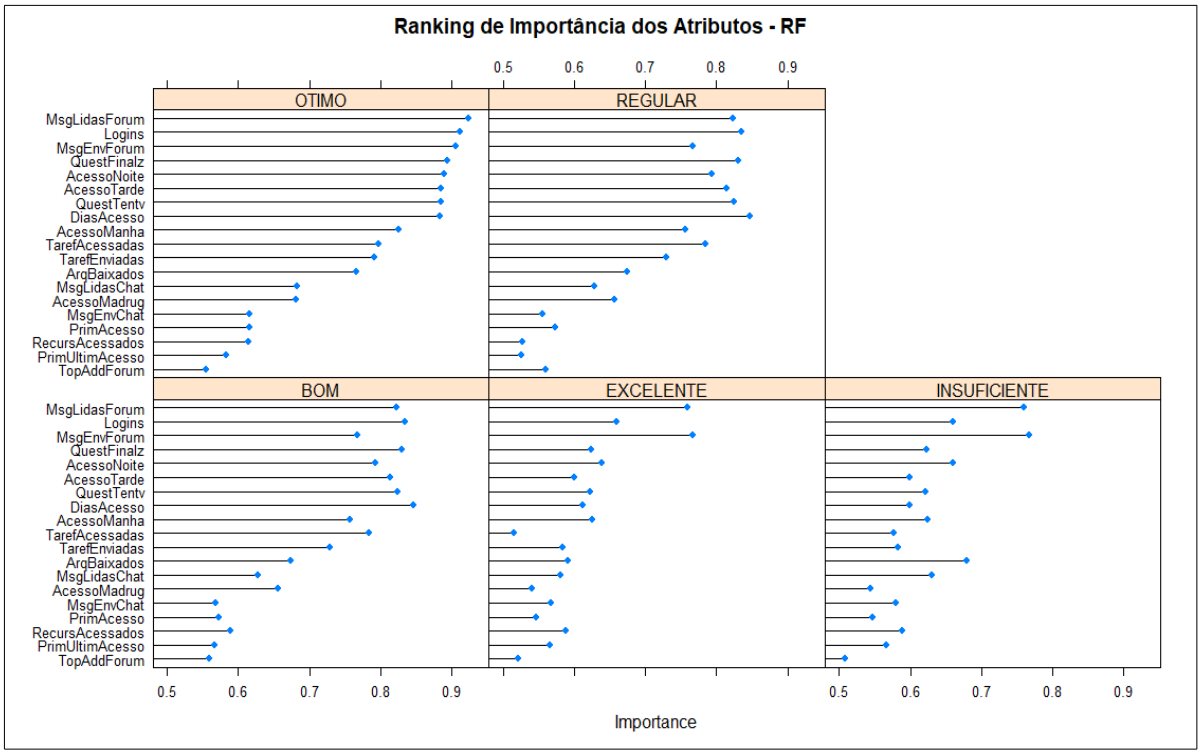

Figura 7- Importância dos atributos em relação ao desempenho final

Nesta abordagem, o programa mostra a importância dos atributos com cada uma das classes do atributo-alvo. Neste resultado, verifica-se também que quase todos os atributos relativos aos objetos de aprendizagem citados anteriormente

apresentam importância acima de 70%, exceto o atributo *TarefEnviadas* que, para as classes excelente e insuficiente, possui resultado em torno de 60%. Porém, todos os outros atributos também tiveram o desempenho reduzido em relação a essas duas classes.

Com esses resultados e tendo em vista que os objetos de aprendizagem do questionário, tarefas e fórum de discussão são utilizados como avaliação *on-line* da disciplina, as séries temporais foram criadas com as interações referentes aos atributos *QuestFinalz*, *QuestTentv*, *TarefFinalz*, *TarefAcessadas*, *MsgEnvForum* e *MsgLidasForum.* Tais atributos correspondem aos objetos *quiz (*cinco questionários), *assign (*uma produção textual por meio de envio de arquivo) e *forum* do AVA Moodle, (postagem em um fórum de discussão).

Criação das Séries Temporais

As séries temporais foram criadas seguindo o período de um ano e meio, referente a um semestre por turma, que compreende ao período de 19 de agosto de 2013 a 31 de dezembro de 2014. Abrangem todas as interações dos estudantes referentes aos atributos selecionados, sendo divididas em intervalos de tempo de 24 horas. A Figura 8 mostra uma parte do *script* de criação das referidas séries.

```
resultado = df.groupby(['NOME'], as_index = False).agg(aggregations)
resultado.to_csv('AtrbDiasAcesso-T1.csv')


df = pd.read_csv(DatasetFinal.csv, encoding='latin-1', low_memory=False)

df['DATA/HORA'] = pd.to_datetime(df['DATA/HORA'], format = '%Y-%m-%d %H:%M:%S')

df = df.sort_values(by=['DATA/HORA', 'ESTUDANTE' ], ascending=[True, True])

df = df.groupby(['DATA/HORA', 'ESTUDANTE'])['ACAO'].count()

df1 = pd.pivot_table(df, index=['DATA/HORA'], columns=['ESTUDANTE'], values='QTD')

df1.set_index(['DATA/HORA'], inplace=True)

df2 = df1.resample('D', label='right', closed='right').sum()

df2.fillna(0, inplace=True)

df2.to_csv('SeriesFinais-B24H_TQF.csv')

# Imputando zeros aos NaN Values
```

Figura 8- Criação das séries temporais

Inicialmente o arquivo d*atasetfinal.csv*, contendo apenas as interações dos atributos selecionados, é lido pelo método *read_csv* do *pandas*, a coluna *DATA/HORA* é convertida para o formato *datetime* (linha 33), depois é realizada uma ordenação crescente pelas colunas *DATA/HORA* e *ESTUDANTE* (linha 35), com o objetivo de ordenar as datas dentro do período considerado. Em seguida, é feito um *groupby* nestas duas colunas e realizada a contagem das interações constantes na variável *ACAO* (linha 37). O resultado do método *groupby* não é um *dataframe*, portanto, usa-se o método *pivot_table* para ajustar o conjunto de dados neste formato, atribuindo o índice à variável *DATA/HORA,* a coluna de dados à variável *ESTUDANTE* e os valores desses dados à *QTD* (linha 39).

Até o momento foram feitos apenas ajustes no *dataframe* para que as séries fossem de fato criadas. O que é feito com o método *resample* do *pandas* (linha 43), que cria as séries referentes a cada estudante em intervalos de tempo de 24 horas (parâmetro “D”, de Dia), somando os valores da variável *ESTUDANTE* dentro de cada um desses períodos. O método *fillna* usado na linha 45 faz o preenchimento das células *NaN* (*Not a Number*) que aparecem por ocasião da criação das séries, com valores zeros. A Tabela 7 exibe uma amostra das séries temporais criadas.

Tabela 7- Matriz das séries temporais de cada estudante

DATA/HORA	E1	E10	E100	E101	E102	E103	E104	E105	E106	E107
2014-03-18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2014-03-19	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
2014-03-20	0.0	0.0	0.0	2.0	0.0	4.0	0.0	0.0	0.0	3.0
2014-03-21	0.0	0.0	0.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0
2014-03-22	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2014-03-23	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
2014-03-24	0.0	0.0	4.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0
2014-03-25	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
2014-03-26	0.0	0.0	2.0	0.0	0.0	0.0	4.0	0.0	2.0	0.0
2014-03-27	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Conforme pode ser observado nesta pequena amostra das séries temporais de dez estudantes (E1 a E107) e dez dias, após a criação surgem muitas células com valores ausentes (que foram preenchidos com zeros), transformando o conjunto de dados das séries em uma matriz esparsa. Neste caso, quanto menor for a periodização dos intervalos das séries (6 e 12 horas) mais a matriz se torna esparsa, influenciando na escolha do período de 24 horas das séries estudadas e da medida de distância utilizada pelo algoritmo de agrupamento, descrita no próximo tópico. Cabe ressaltar, que nesta pesquisa foram realizados experimentos com as séries temporais divididas em intervalos de 6, 12 e 24 horas.

Segundo Tan (2006), uma matriz esparsa é um caso especial de matriz de dados “com atributos assimétricos, onde somente os valores não-zero são importantes”. Esse tipo de conjunto de dados é comumente encontrado em pesquisas de mineração de dados nas áreas de medicina, bioinformática e agrupamento de textos. A Figura 9 abaixo mostra a representação gráfica das séries temporais referentes a oito estudantes.

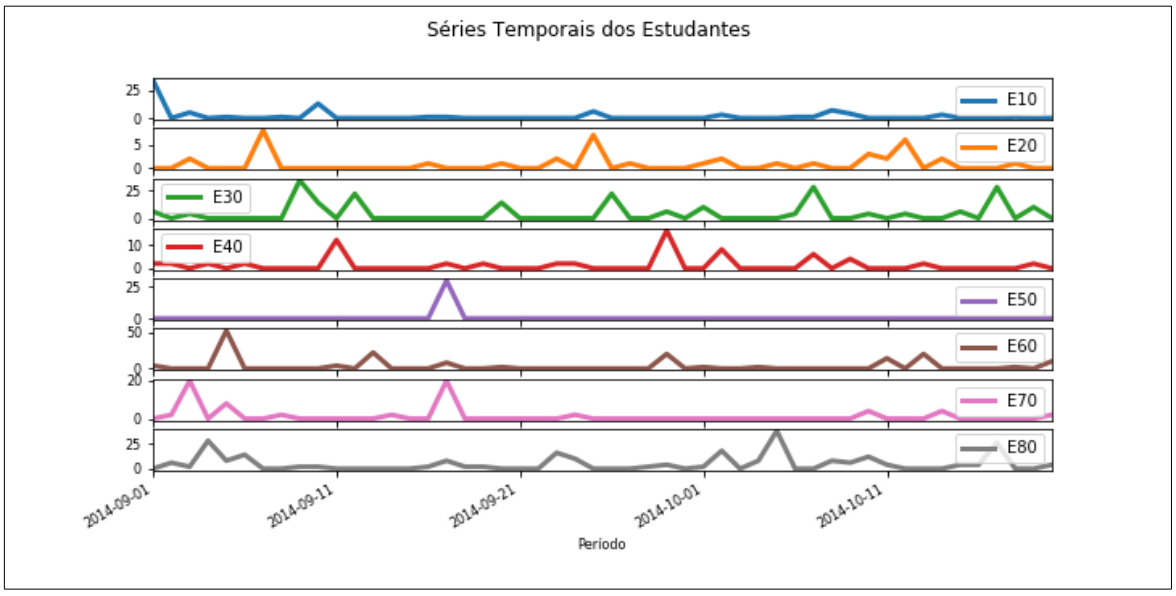

Figura 9- Séries temporais de oito estudantes

Neste período plotado, pode-se notar que as séries são compostas de “picos” de atividades e “vales” de pouca ou nenhuma atividade. É possível perceber ainda que alguns estudantes mantêm as atividades com mais frequência no AVA, como

E20, E30, E40 e E80, sendo que outros interagem com menos frequência, apresentando picos de atividades mais isolados na série, como E10 e E50.

4.3 Mineração de dados educacionais

Esta penúltima fase da descoberta do conhecimento consiste na principal etapa do processo, pois as atividades realizadas nas fases anteriores foram uma preparação do conjunto de dados para ser processado pelo algoritmo de mineração de dados nesta etapa da pesquisa.

O algoritmo escolhido para esta pesquisa foi o *k-means*, que foi apresentado em 1967 por MacQueen (1967), sendo bastante utilizado para tarefas de agrupamento em análise de perfis de usuário, análise de padrões de comportamento de multidões, reconhecimento de padrões em imagens médicas, descoberta de padrões de cliques em páginas de internet, dentre outros (SILVA *et al.*, 2016).

O *k-means* tem por objetivo encontrar agrupamentos no conjunto de dados de forma que *k* grupos distintos de exemplares sejam descobertos, sendo *k* definido previamente no algoritmo. A busca do conjunto de *k* grupos é iterativa e iniciada a partir da escolha aleatória de *k* vetores, também distintos, que representam os centros dos grupos, conhecidos como centróides ou protótipos.

A partir dessa escolha inicial dos centróides, o algoritmo segue verificando quais exemplares são mais similares a quais centróides, ajustando os centróides aos exemplares mais similares a eles, estabelecendo dessa forma os agrupamentos ou partições.

O pseudocódigo do Quadro 1 mostra o funcionamento do *k-means*. Conforme Solange *et al.* (2011), o algoritmo apresenta grande simplicidade computacional, característica que o tornou no algoritmo de partição mais utilizado entre os pesquisadores de mineração de dados educacionais que utilizam tarefas de agrupamento, segundo Peña-Ayala (2014).

Quadro1- Pseudocódigo do *k-means* (adaptado de Solange *et al.*, 2011)

Algoritmo K-Means
Entrada: *X* = {*x1*, *x2*, …, *xn*}: conjunto de dados *k*: número de grupos **Saída:** *P* = {*G1*, *G2*,…, *Gk*}: partição com os *k* grupos selecionar aleatoriamente *k* exemplares como centróides *C* iniciais; **repita** **para cada** exemplar *x* E *X* **faça** computar a similaridade de *x* para cada centróide *C*; atribuir *x* ao centróide mais próximo; **fim** recomputar o centróide de cada grupo; **até** *atingir um critério de parada*;

Essa simplicidade computacional, o amplo uso em pesquisas de MDE com abordagens similares a desse trabalho e, principalmente, a forma de dispersão dos exemplares no conjunto de dados desta pesquisa, foram os motivos que levaram a escolha do *k-means* como algoritmo utilizado nesta etapa do processo.

Agrupamentos e medidas de distância

O *k-means* é um algoritmo de agrupamento particional, que difere dos algoritmos de agrupamentos hierárquicos pela necessidade de se determinar *a priori* o valor de *k,* a quantidade de grupos do conjunto de dados a ser particionado. Essa característica é considerada uma desvantagem do *k-means,* juntamente com a configuração inicial dos centróides, que podem “comprometer o resultado final do algoritmo” (SILVA *et al.*, 2016).

O objetivo do algoritmo é encontrar agrupamentos onde os exemplares dentro dos grupos sejam o mais parecidos possíveis (similaridade) e, em comparação com elementos de outros grupos, sejam o mais distintos possíveis (dissimilaridade). Tal função é quantificada pela medida de distância escolhida para o algoritmo, que especifica o quão próximo um exemplar está ou não de outro, servindo de parâmetro para atribuir o exemplar a um determinado grupo. A medida de distância padrão do algoritmo *k-means* é a distância euclidiana.

Para condução dos experimentos desta pesquisa com o *k-means*, a medida de similaridade utilizada foi o coeficiente de *Jaccard* (Jaccard, 1908), escolhido devido às características finais do conjunto de dados que, após a transformação em séries temporais, tornou-se uma matriz esparsa. Segundo Suryakant e Tripti (2016), a medida *Jaccard*, bem como, os coeficientes de correlação de *Pearson* e *Cosseno* possuem "boa acurácia em problemas de ambiente esparso", sendo considerado um dos coeficientes de similaridade "mais estáveis" (YIN e YASUDA, 2005). A similaridade de *Jaccard* é definida pela Equação 1 abaixo.

$$sim(u, u')^{Jaccard} = \frac{|I_u| \cap |I_{u'}|}{|I_u| \cup |I_{u'}|} \quad [1]$$

O coeficiente de similaridade de *jaccard* entre dois conjuntos de dados é o resultado da razão entre o número de elementos que são comuns aos dois conjuntos (intersecção) pela quantidade de todos os elementos dos dois conjuntos (união). Segundo Tan (2006), Yin e Yasuda (2005) e Meyer (2002), essa é uma medida de similaridade amplamente utilizada na área de mineração de textos.

Processamento do algoritmo e criação dos *clusters*

Além da definição da medida de distância, existe a necessidade de se determinar previamente qual é o melhor valor de *k* para o particionamento dos dados. Em geral, essa decisão não é trivial, tendo em vista que o algoritmo de *clustering* usa uma abordagem não-supervisionada, onde não se conhece as classes dos exemplares.

Felizmente existem alguns métodos que atuam justamente nesta parte da clusterização, fazendo uma análise dos dados de entrada e sugerindo o valor mais adequado para *k*. Nem sempre os resultados são satisfatórios, cabendo ao especialista decidir qual ou quais valores de *k* utilizar. Neste trabalho, utilizou-se o método *elbow* e o método hierárquico com dendograma, com o objetivo de auxiliar na escolha do melhor valor de *k* para processar o *k-means*.

O método *elbow* tem por finalidade sugerir um valor para a quantidade de *clusters* antes da utilização do algoritmo de agrupamento. Ele executa o *k-means*

com o *dataset* de entrada para um intervalo de valores de *k* (normalmente de 1 a 10), calcula a soma do quadrado dos erros (do inglês, *Sum of Squared Errors – SSE*) para cada valor de *k* e analisa o resultado a cada incremento de *k*.

Dessa forma, quando a diferença entre esses resultados parar de ser relevante, apresentando valores praticamente iguais, o método entra num "modelo platô" no qual o valor da diferença das distâncias entre os exemplares torna-se quase insignificante. Exatamente neste valor de *k,* onde o resultado da *SSE* não é significativo, forma-se um cotovelo (do inglês, *elbow*), daí o nome do método, sugerindo que aquele é um valor de *k* otimizado para o conjunto de dados em questão.

A Figura 10 exibe à esquerda a plotagem do resultado da execução do método *elbow* para um conjunto de dados qualquer e à direita mostra o resultado do conjunto de dados utilizado nesta pesquisa, após a transformação em séries temporais.

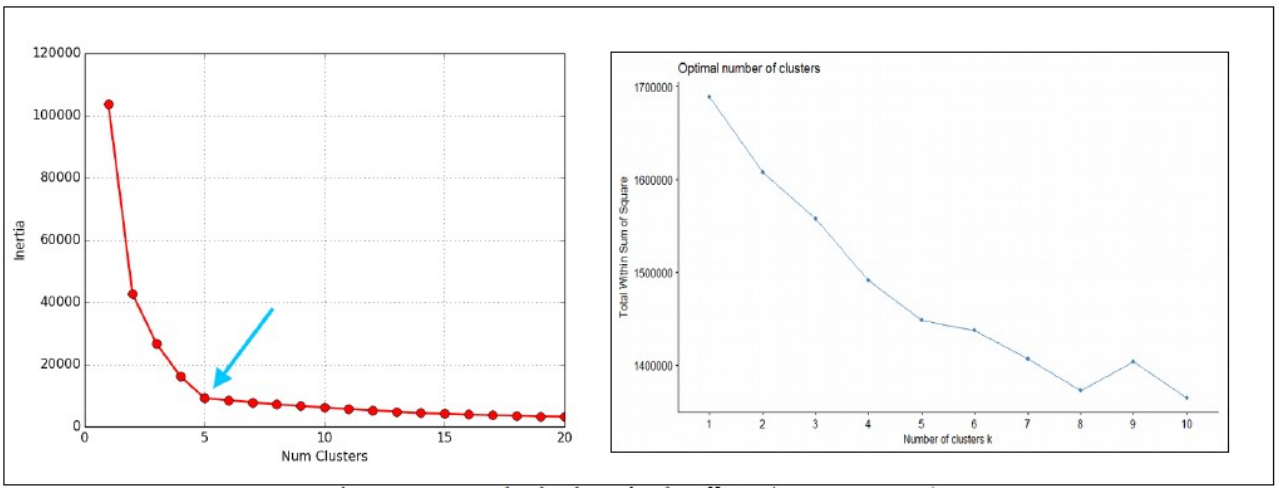

Figura 10- Plotagem do resultado do método *elbow*

Pode-se observar que para o conjunto de dados da esquerda o método apresenta resultado visualmente sugestivo. Neste caso, fica evidente que o valor de *k* sugerido pelo método é 5, pois a partir dele, a diferença da *SSE* para os próximos valores de *k* se torna praticamente contínua, mostrando o "platô" citado acima. Entretanto, o resultado para o conjunto de dados desta pesquisa não apresenta um valor de *k* tão evidente, mostrando que para o valor de *k* igual a 5, a *SSE* parece

estabilizar, porém volta a crescer para *k* igual a 8 em diante. Assim, houve a necessidade de utilizar outro método que complementasse o *elbow* para a escolha do melhor valor de *k*.

A estratégia de *clustering* hierárquico faz uma abordagem *top-down*, dividindo hierarquicamente o conjunto de dados inicial em grupos. Silva *et al*. (2016) aborda que o processo inicia “colocando todos os exemplares em um único grupo e, iterativamente, divide um grupo em grupos menores até que cada exemplar esteja isolado em um grupo”. Uma das representações gráficas da estratégia de partição hierárquica é o dendograma, que apresenta uma visualização bastante útil dos exemplares, organizando-os por similaridade.

Silva *et al.* (2016) aborda que a partir dessa organização, pode-se determinar a quantidade de grupos desejados (valor de *k*) para o agrupamento final. Assim, utilizou-se os métodos do pacote R *factoextra* atribuindo para *k* os valores 5, 6, 7 e 8, fez-se a plotagem dos dendogramas e análise dos resultados. Dessa forma, o valor de *k* igual a 7 apresentou desempenho mais significativo no agrupamento final das séries temporais, sendo esse o valor escolhido para a execução do *k-means*. O dendograma da Figura 11 mostra esse resultado.

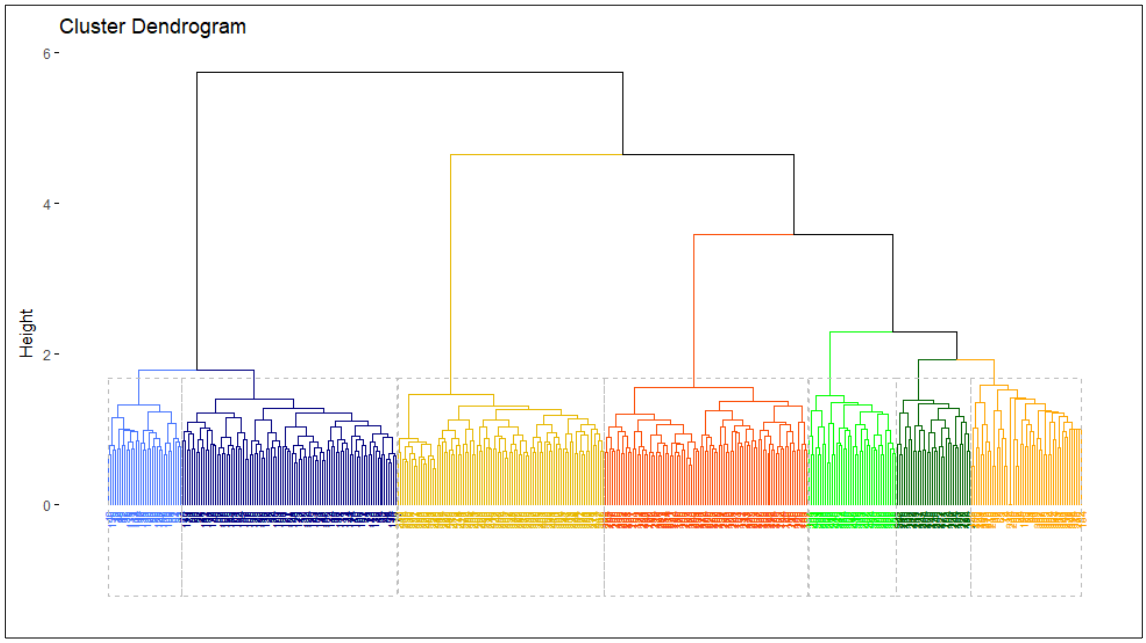

Figura 11- Dendograma com valor de *k* = 7

Neste gráfico pode-se notar que os exemplares foram agrupados em três *clusters* maiores representados pelas cores azul, amarelo e vermelho, e quatro *clusters* menores nas cores azul-claro, verde-claro, verde-escuro e laranja, contendo as séries temporais referentes a cada um dos estudantes.

Após a determinação da medida de distância ou de similaridade, e do valor mais adequado para a quantidade de grupos do conjunto de dados, utilizou-se o *k-means* do pacote R *Stats* conforme *script* mostrado na Figura 12 abaixo.

```
51  dfCurso <- read.csv("C:\\Mestrado_UFRPE\\SeriesFinais-B24H_TQF.csv", +
52      header=TRUE, sep=",", dec=".")
53  set.seed(1234)
54  library(TSdist)
55  dfCurso <- as.data.frame(dfCurso)
56  mdJacc <- dist(as.matrix(dfCurso[2:400]), method = "jaccard")
57  kmJacc7 <- kmeans(mdJacc, 7, iter.max = 500, nstart = 10)
58  resCursoJ7 <- data.frame(dfCurso, kmJacc7$cluster)
59  write.csv(resCursoJ7, "ResultadoCursoJ7.csv")
```

Figura 12- *Script* de execução do *k-means*

Inicialmente o arquivo das séries temporais é lido e armazenado na variável *dfCurso* (linha 51), em seguida usa-se a função *set.seed* (linha 53) para garantir a reprodutibilidade em novas execuções do *k-means*. Carrega-se a biblioteca *TSdist* que possui vários algoritmos de medidas de distância e de similaridade (linha 54), depois calcula-se a medida de similaridade do conjunto de dados de entrada usando o coeficiente de *jaccard* (linha 56).

Finalizando com a execução do *k-means* (linha 57), contendo os seguintes parâmetros: a matriz de resultado do índice de *jaccard* (*mdJacc*), o número de *clusters* (7), a quantidade máxima de iterações executadas pelo algoritmo para encontrar o tamanho otimizado dos *clusters* (*iter.max=500*) e o número de vezes que ele inicializa o processo com novos valores de centróides (*nstart=10*).

As linhas 58 e 59, da Figura 12, são funções utilizadas após a execução do *k-means*, objetivando criar uma nova coluna no *dataframe* das séries temporais, com o número do *cluster* atribuído a cada um dos exemplares (linha 58), assim como, para salvar esse resultado em um arquivo *csv* externo, analisado na próxima fase do

processo (linha 59). A Tabela 8 apresenta uma amostra desse resultado para dez estudantes (E1 a E10) e quatorze dias (D1 a D14).

Tabela 8- Atribuição dos *clusters* finais do *k-means*

ESTUDANTE	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	CLUSTER
E1	0	0	0	0	1	2	1	0	0	1	0	0	1	4	4
E2	0	0	0	0	0	0	0	0	0	0	0	1	2	0	4
E3	1	0	0	0	0	0	0	0	11	0	0	0	0	0	1
E4	10	0	13	5	1	5	0	2	1	2	1	1	0	1	3
E5	0	1	0	3	17	2	0	0	0	0	0	0	1	4	4
E6	0	0	0	0	0	1	0	1	0	1	9	2	0	1	3
E7	0	1	9	0	1	0	0	4	0	2	0	4	1	20	3
E8	0	4	0	1	0	0	0	0	0	0	0	0	0	0	1
E9	0	3	0	0	0	0	0	0	0	0	0	6	0	23	4
E10	1	5	10	11	2	5	0	6	3	0	0	4	0	39	3

A primeira coluna desta tabela identifica os dez primeiros estudantes, as colunas de D1 a D14 representam os dias de atividades no Moodle, com suas respectivas quantidades de interações. Essa substituição das datas pelos dias foi realizada com a finalidade de adequar o tamanho da tabela ao texto. A última coluna refere-se ao número do *cluster* atribuído a cada série temporal dos estudantes.

Logo após o processamento, conforme pode ser visto na Figura 13, o *k-means* mostra o tamanho de cada *cluster* (linha 61), os valores dos centróides finais para cada partição (linha 63), assim como o percentual de variância total do conjunto de dados após a clusterização, tal percentual é considerado um índice de validação interna das partições criadas (linha 64).

```
61  K-means clustering with 7 clusters of sizes 58, 80, 58, 69, 65, 48, 42
62  Within cluster sum of squares by cluster:
63  [1] 77.44166 78.64183 51.50341 79.64985 60.11349 55.64372 52.08916
64   (between_SS / total_SS =  63.9 %)
```

Figura 13- Tamanho dos *clusters* e percentual de WSS

Validação estatística da clusterização

Segundo Brock *et al.* (2008) *apud* Kassambara (2016) existem três categorias de validação estatística do resultado do algoritmo de partição: (I) validação interna, que

usa as informações internas dos *clusters* criados, para avaliar o nível de coesão da estrutura das partições, sem referências externas; (II) validação externa, que consiste na análise dos agrupamentos resultantes com alguma informação externa conhecida como, por exemplo, possíveis valores de atributos de classe; (III) validação relativa, que consiste na avaliação da estrutura dos *clusters* através da atribuição de diferentes parâmetros para o algoritmo.

A validação interna mede o nível de compactação ou coesão dos objetos dentro de uma mesma partição. Assim como, mensura o nível de separação dessas partições através da distância entre os centros dos *clusters* (centróides). O Quadro 2 abaixo apresenta o resultado da validação interna dos *clusters* criados pelo *k-means*, pelo índice de variância total após o agrupamento, índice de *Dunn* e índice de *Calinski Harabasz*.

Quadro 2- Índices de validação interna

Variância Total	**Índice de Dunn**	**Calinski Harabasz**
Between_SS/Total_SS	$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}.$	$CH(k) = \frac{B_c(k)}{(k-1)} / \frac{W_c(k)}{(n-1)}$
0,6390670	0,6666667	0,7068443

A validação relativa foi aplicada neste trabalho quando se utilizou, conforme apresentado no tópico anterior, os métodos *elbow* e hierárquico para se determinar o melhor valor de *k* para o algoritmo *k-means*. A validação externa, feita no próximo capítulo, apresenta a interpretação do resultado dos agrupamentos sob a perspectiva dos comportamentos de engajamento e procrastinação dos estudantes no curso em pauta.

5 RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados obtidos pelo algoritmo *k-means*, os agrupamentos obtidos, suas características, discussão sobre os resultados, tendo como foco os dois comportamentos analisados neste trabalho, engajamento e procrastinação acadêmicos.

A Tabela 9 abaixo apresenta os sete agrupamentos resultantes da aplicação do *k-means*, a quantidade de estudantes agrupadas em cada grupo, os percentuais de estudantes por classificação de desempenho e pelo resultado final.

Tabela 9- Características dos agrupamentos

Grupos	Quantidade de Estudantes	Desempenho					Resultado Final	
		Insuficiente	Regular	Bom	Ótimo	Excelente	Aprovado	Reprovado
Grupo 1	58	98,3%	-	1,7%	-	-	1,7%	98,3%
Grupo 2	80	28,8%	15,0%	30,0%	22,4%	3,8%	60,0%	40,0%
Grupo 3	58	36,2%	13,2%	24,1%	23,1%	3,4%	63,8%	36,2%
Grupo 4	69	56,5%	13,0%	20,4%	8,7%	1,4%	43,5%	56,5%
Grupo 5	65	16,9%	13,8%	36,9%	23,2%	9,2%	73,8%	26,2%
Grupo 6	48	77,1%	-	16,6%	4,2%	2,1%	22,9%	77,1%
Grupo 7	42	95,2%	4,8%	-	-	-	-	100,0%

Os Grupos 1 e 7 apresentam em suas quase totalidades estudantes REPROVADOS, 98,3% e 100%, respectivamente. Apenas um aluno do Grupo 1 teve desempenho BOM e foi aprovado. Os Grupos 2, 3 e 5 contém a maioria dos alunos APROVADOS, apresentando uma distribuição mais parecida entre todas as classificações de desempenho dos alunos, sendo que o Grupo 5 é o que apresenta maior percentual de alunos APROVADOS (73,8%) e o Grupo 2 o que tem o menor percentual (60,0%), dos três grupos. Os Grupos 4 e 6, apresentam percentuais de reprovação de 56,52% e 77,08%, respectivamente. Nesta primeira análise, pode-se notar que existem alguns grupos com características similares.

Nas duas próximas seções deste capítulo são apresentadas as análises dos agrupamentos resultantes quanto ao engajamento e procrastinação acadêmicos.

5.1 Análise dos agrupamentos quanto ao engajamento acadêmico

Tendo como referência vários trabalhos de pesquisa de EDM e LA (e.g. Li e Baker, 2018; Rodrigues *et al.*, 2016; Khalil *et al.,* 2016; Kovanovic *et al.,* 2016), este estudo busca encontrar grupos de estudantes que possuam características similares de engajamento comportamental em relação às interações dos objetos de aprendizagem fórum de discussão, questionários e tarefas realizadas no ambiente virtual de aprendizagem.

Cabe ressaltar que a mensuração do nível de engajamento comportamental abordado nesta pesquisa se limita a uma abordagem quantitativa das interações dos objetos de aprendizagem citados anteriormente. Dimensões emocionais, cognitivas, afetivas, psicológicas do engajamento, dentre outras, não são tratadas nesta pesquisa.

Seguindo a abordagem feita por Rodrigues *et al.* (2016), houve a necessidade de fazer um teste de comparação de médias grupo a grupo para verificar a "aceitação, ou não, de possíveis igualdades ente os grupos encontrados", tendo em vista que, realizando uma primeira análise sobre os grupos resultantes da clusterização, encontram-se algumas similaridades entre alguns grupos.

O teste de comparação de médias entre grupos é um teste de hipótese estatística, utilizado para tomada de decisões. Para desenvolver o teste de hipótese estatística, deve-se formular duas hipóteses: (I) Hipótese nula (H0), que diz que não há diferença significativa entre os parâmetros da amostra e da população, isto é, a afirmação é considerada verdadeira até que, com base em uma outra amostra, conclua-se que essa afirmação deve ser rejeitada; (II) Hipótese alternativa (H1), quando a hipótese nula é rejeitada, ou seja, a afirmação é considerada como sendo falsa, deve-se aceitar a hipótese alternativa como sendo verdadeira.

Para realizar o teste inferencial de comparação de médias dos grupos, utilizou-se o pacote R *dplyr* com o teste *t-student*, parametrizando o nível de significância do teste em 0,05, isto é, tem-se cerca de 5% de chance da hipótese nula ser rejeitada quando deveria ser aceita. A Tabela 10 abaixo exibe o resultado desse teste.

Tabela 10- Resultado do teste de comparação de médias

Grupos	1	2	3	4	5	6
1	-	-	-	-	-	-
2	0,000000	-	-	-	-	-
3	0,000000	**0,529900**	-	-	-	-
4	0,000000	0,000000	0,000013	-	-	-
5	0,000000	**0,788600**	**0,393200**	0,000005	-	-
6	0,000000	0,007827	0,008073	**0,095310**	0,013520	-
7	**0,792800**	0,000127	0,000001	0,004798	0,000000	0,000488

Como mostrado na tabela acima, os valores de *p-values* abaixo de 0,05 rejeitam a hipótese nula de igualdade entre os agrupamentos. Assim, pode-se afirmar que, estatisticamente, os grupos 1 e 7 são similares, os grupos 2, 3 e 5 também são semelhantes, bem como, os grupos 4 e 6.

Após a verificação das características comportamentais dos estudantes em relação aos atributos em análise, nos grupos que apresentam similaridades, constata-se que existem três níveis de engajamento acadêmico nos agrupamentos: Baixo (Grupos 1 e 7), Intermediário (Grupos 4 e 6) e Alto (Grupos 2, 3 e 5).

A medição foi realizada comparando todos os grupos com o grupo 5, considerado o mais engajado devido a sua maior quantidade de interações dos estudantes no fórum de discussão, questionários e tarefas. Dessa forma, analisou-se os estudantes que estavam acima ou abaixo da média de interações do grupo 5.

Cabe ressaltar que, os estudantes dos Grupos 2, 3 e 5 foram aqueles que mais postaram mensagens no fórum de discussão, caracterizando o comportamento de participação ativa nas tarefas do curso e comprovando o alto nível de engajamento acadêmico com que foram classificados.

Para todos os gráficos das seções seguintes, as barras na cor amarela são dos alunos que apresentam a contagem de “Mensagens Lidas/Enviadas” no fórum de discussão, bem como, de “Acessos/Submissões” dos questionários e tarefas abaixo da média; as barras na cor verde correspondem aos alunos que possuem a quantidade acima da média; a linha horizontal na cor vermelha representa a média do atributo no conjunto de dados e a linha horizontal na cor azul corresponde à

mediana. Para o fórum de discussão a média é 24 e mediana 15, e para questionários e tarefas a média é 19 e mediana 16.

(I) Estudantes com baixo engajamento

Fórum de discussão

O Gráfico 3 mostra a quantidade de mensagens lidas/enviadas no fórum de discussão pelos alunos classificados com baixo nível engajamento. Esse agrupamento é formado por 99% de alunos reprovados (99).

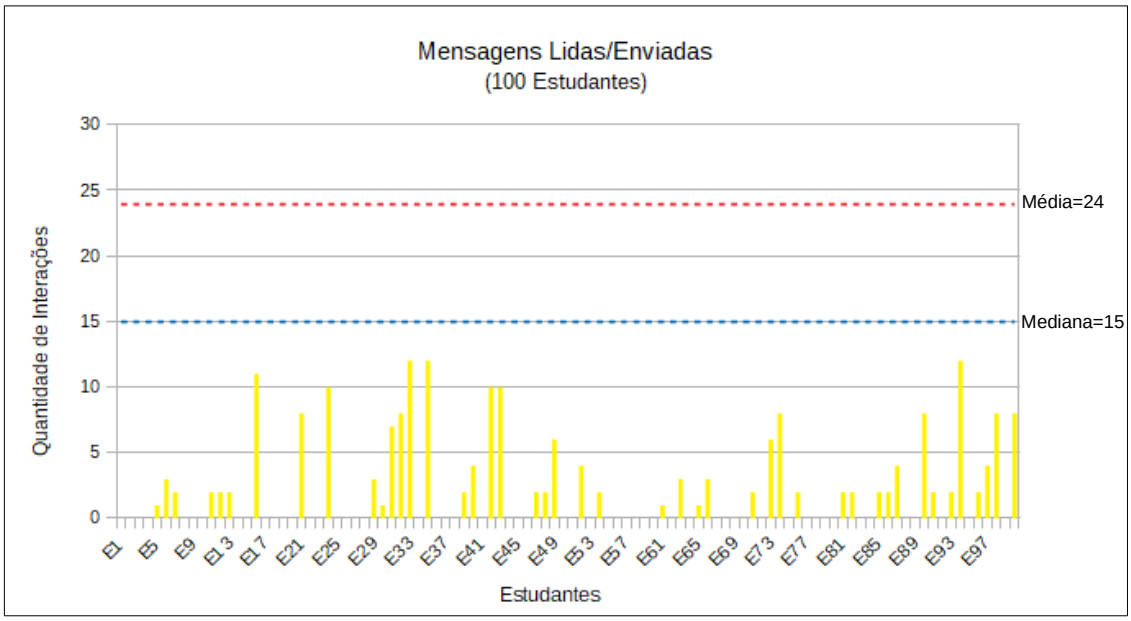

Gráfico 3- Mensagens lidas/enviadas no fórum de discussão (Baixo Engajamento)

Como mostram os gráficos acima, fica notório que os estudantes classificados com baixo engajamento apresentam valores abaixo da média ou mediana em relação às mensagens lidas/enviadas no fórum de discussão. Na realidade, esse nível de engajamento não possui estudantes acima da média ou mediana neste quesito, sendo que alguns alunos nem entraram no fórum, como por exemplo, os alunos E1, E9, E37, E57, dentre outros.

Questionários e Tarefas

O Gráfico 4 exibe a quantidade de acessos/submissões realizados pelos estudantes nos questionários e das tarefas *on-line*.

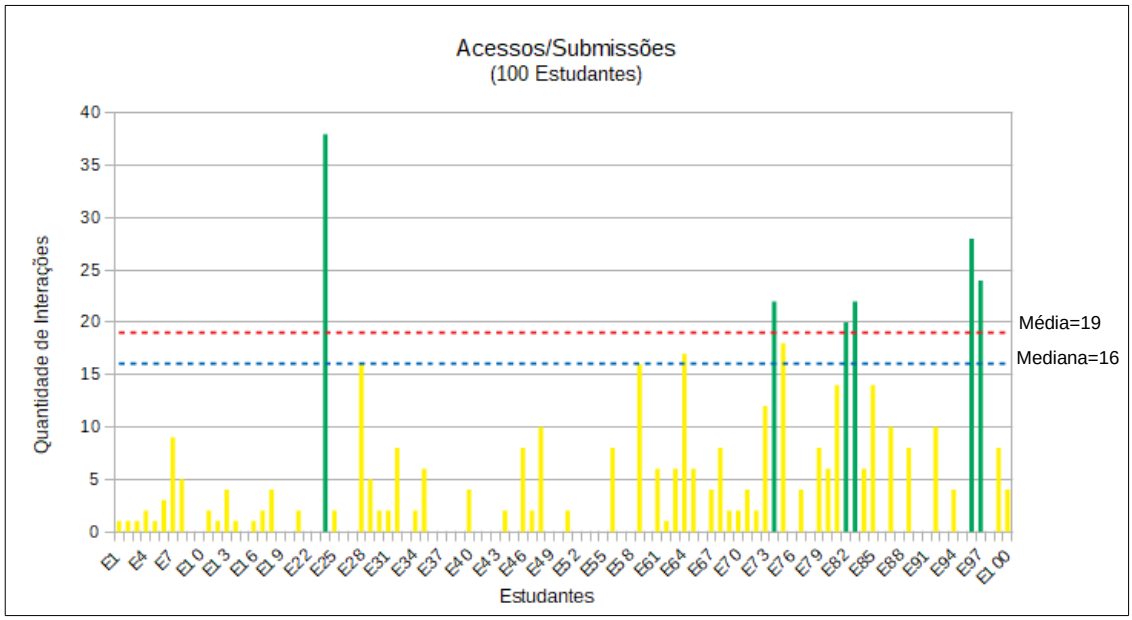

Gráfico 4- Acessos e submissões dos questionários/tarefas (Baixo Engajamento)

Para a leitura das instruções dos questionários e da tarefa a ser submetida no AVA, a média de acesso é 19 e a mediana 16. Neste caso, o grupo com baixo nível de engajamento possui 94% de estudantes (94) que acessaram o objeto questionário e tarefas abaixo da média e 6% (6) que acessaram acima da média, com alunos que não acessaram a tarefa nenhuma vez: E10, E22, E55, E91, dentre outros.

(II) Estudantes com alto engajamento

Fórum de Discussão

O Gráfico 5 mostra a quantidade de mensagens lidas/enviadas no fórum de discussão pelos alunos classificados com alto nível engajamento. Esse grupo é formado por 69,45% de estudantes aprovados (141).

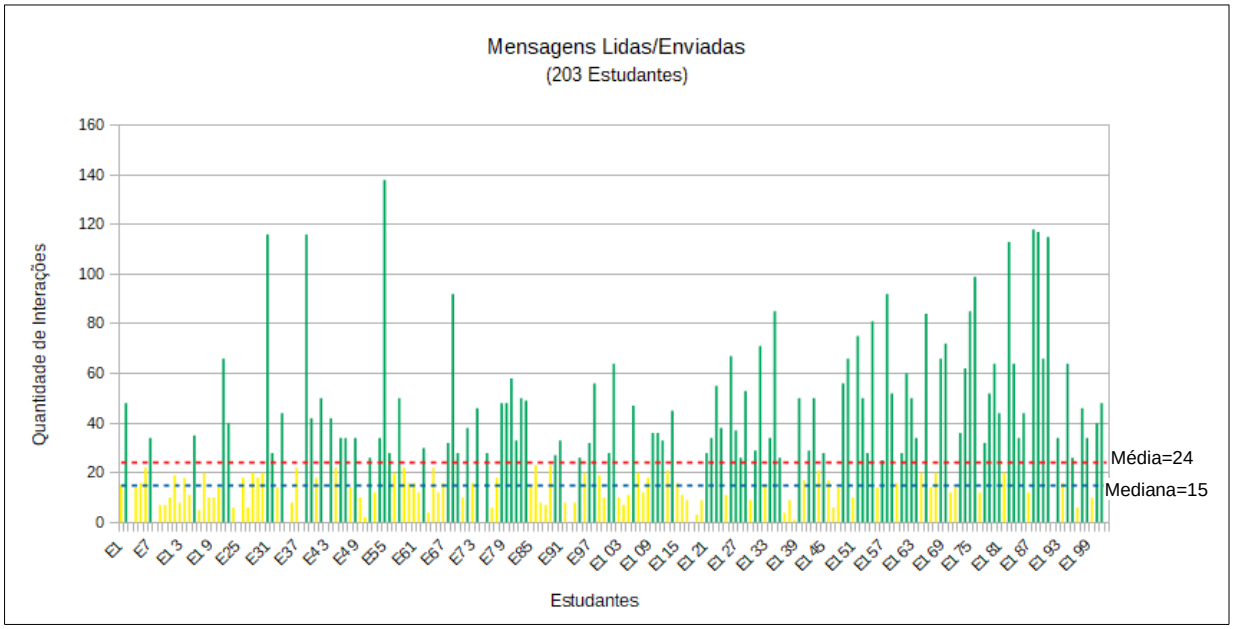

Gráfico 5- Mensagens lidas/enviadas no fórum de discussão (Alto Engajamento)

Diferentemente dos estudantes classificados com baixo nível de engajamento, percebe-se que este grupo possui vários estudantes, com quantidade de leitura e postagem de mensagens no fórum de discussão, bem acima da média e mediana de referência, sendo 65,17% de estudantes com esse perfil, que equivale a 135 estudantes. Fato que ratifica o alto nível de engajamento com que esses alunos foram classificados em relação ao objeto de aprendizagem fórum de discussão.

Questionários e Tarefas

O Gráfico 6 exibe a quantidade de vezes que os estudantes do acessaram ou submeteram os questionários/tarefa no AVA.

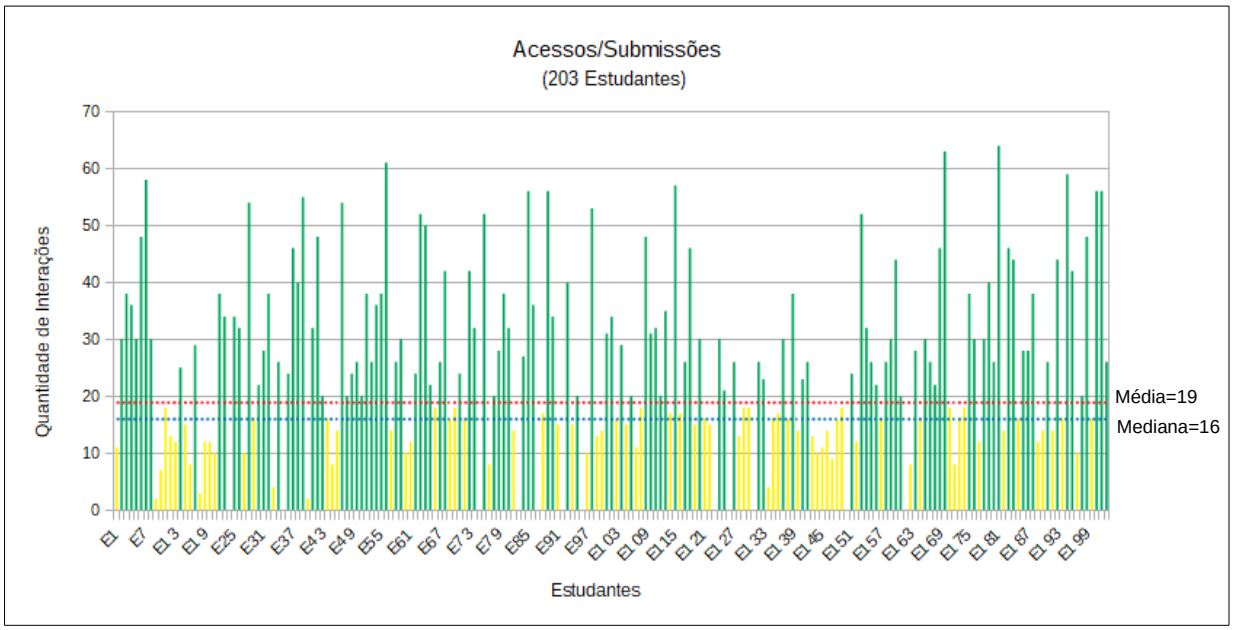

Gráfico 6- Acessos e submissões dos questionários/tarefas (Alto Engajamento)

Quanto a esse objeto de aprendizagem, percebe-se também que este nível de engajamento possui estudantes com quantidade de acessos/submissões de questionários e tarefas superiores ao grupo de estudantes com baixo engajamento, sendo 64,53% de estudantes acima da média e mediana de referência, equivalente a 131 estudantes.

(III) Estudantes com engajamento intermediário

Fórum de Discussão

Esse agrupamento foi classificado com estudantes que apresentam nível intermediário de engajamento, tendo mais alunos reprovados do que aprovados, porém em números menores do que o agrupamento de estudantes com baixo engajamento. O Gráfico 7 mostram a quantidade de mensagens lidas/enviadas no fórum de discussão pelos alunos desse nível de engajamento.

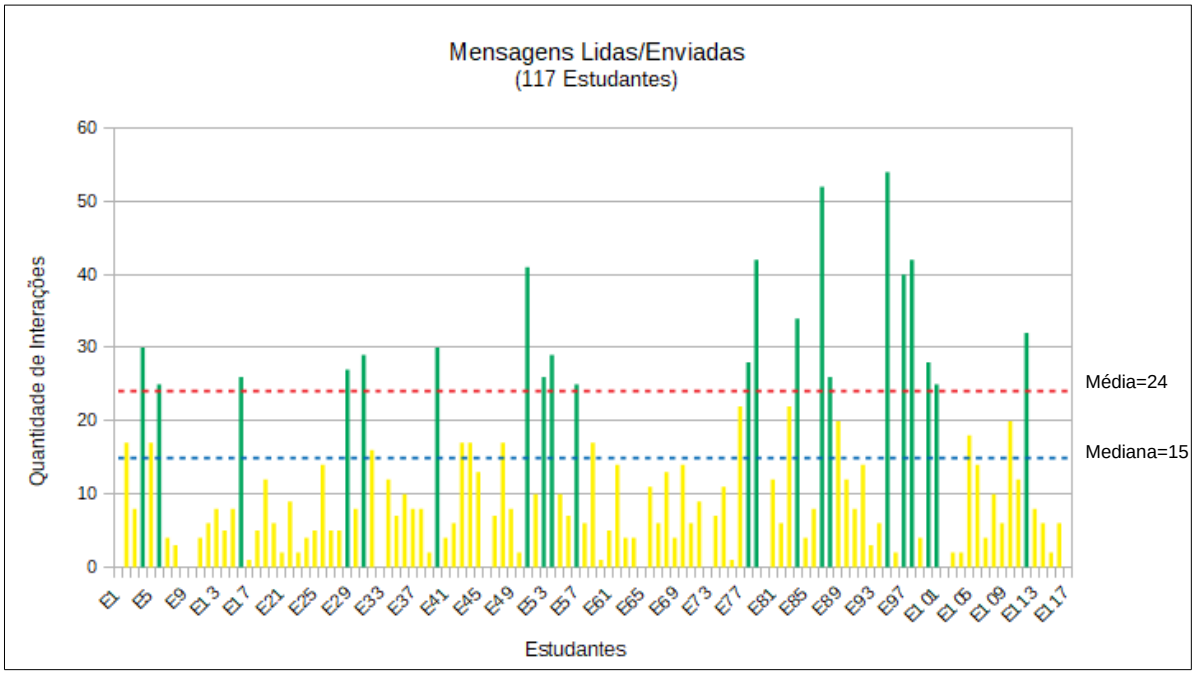

Gráfico 7- Mensagens lidas/enviadas no fórum de discussão (Engajamento Intermediário)

Aqui, percebe-se que o engajamento dos estudantes na visualização e envio das mensagens do objeto de aprendizagem em tela é superior ao dos alunos do agrupamento de nível baixo e inferior ao dos alunos do nível alto de engajamento, destacando o caráter intermediário desse agrupamento. Apresenta 18,80% de estudantes (22) que leram/enviaram mensagens no fórum de discussão acima da média e mediana, com 81,19% de estudantes (95) abaixo desses índices.

Questionários e Tarefas

O Gráfico 8 exibem a quantidade de vezes que os estudantes visualizaram ou submeteram os questionários e tarefas no ambiente virtual de aprendizagem.

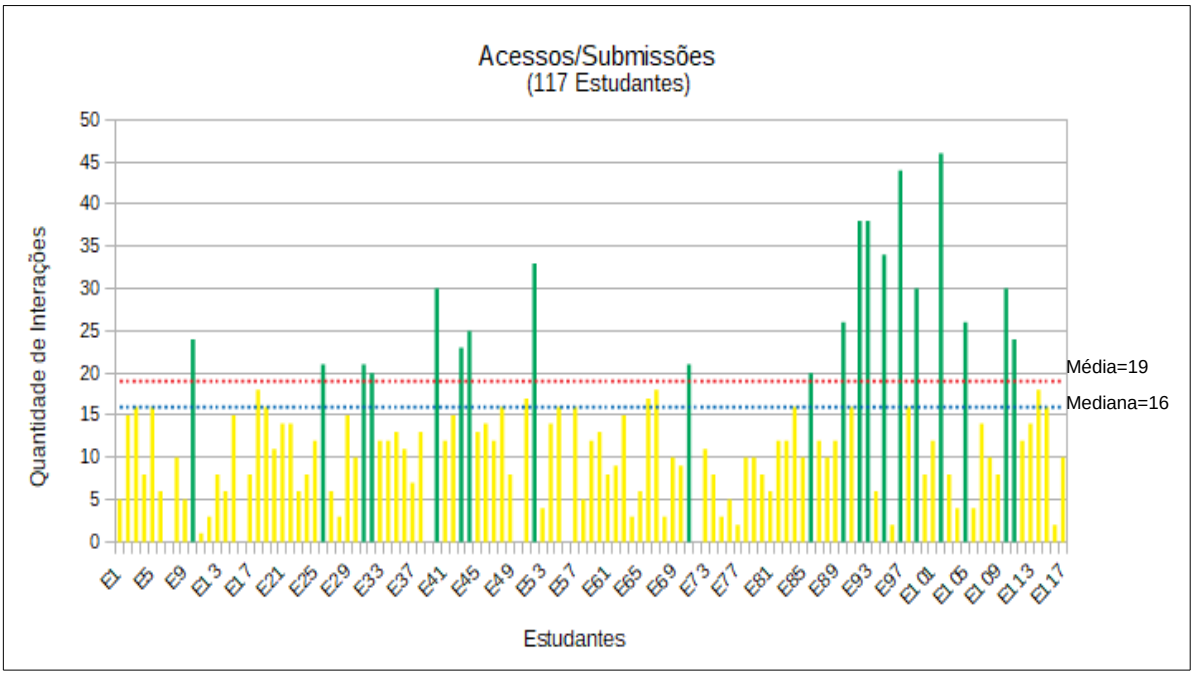

Gráfico 8- Acessos e submissões dos questionários/tarefas (Engajamento Intermediário)

O comportamento dos alunos desse agrupamento se repete neste atributo, mantendo um cenário de alunos acima ou abaixo da média bem parecido com a análise do fórum de discussão, confirmando o caráter intermediário do nível de engajamento dos estudantes desse agrupamento neste recurso de aprendizagem. Apresenta 17,09% de estudantes (20) que fizeram acessos/submissões de questionários e tarefas acima da média e mediana, com 82,90% de estudantes (97) abaixo desses índices.

O Gráfico 9 mostra algumas características de medidas estatísticas, como variabilidade, média e mediana, das interações dos estudantes no fórum de discussão em cada nível de engajamento.

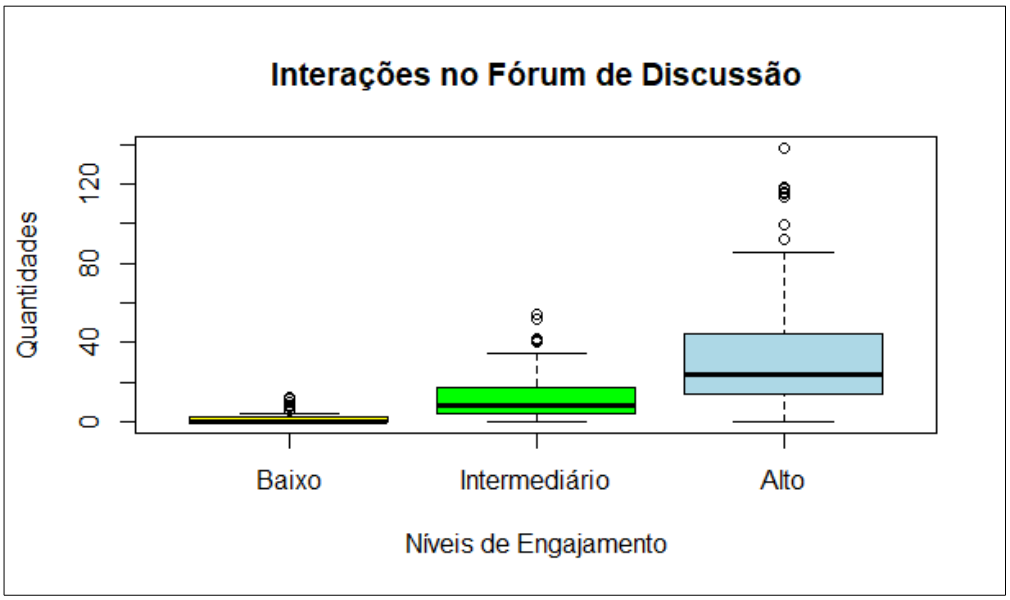

Gráfico 9- Interações no fórum de discussão por nível de engajamento

A quantidade mínima de interações dos estudantes dos três níveis de engajamento é zero, o valor máximo para o nível baixo é 12 acessos (DP 3,32), para o intermediário é 56 acessos (DP 11,70) e para o nível alto é 138 (DP 26,40) . Nota-se que os níveis intermediários e alto possuem alguns valores considerados fora do padrão (*outliers*). A média de acessos é 2,1 para os estudantes do nível baixo, 11,70 para o nível intermediário e 32,29 para o nível alto de engajamento.

O Gráfico 10 mostra algumas características de medidas estatísticas, como variabilidade, média e mediana, das interações dos estudantes nos questionários e tarefas em cada nível de engajamento.

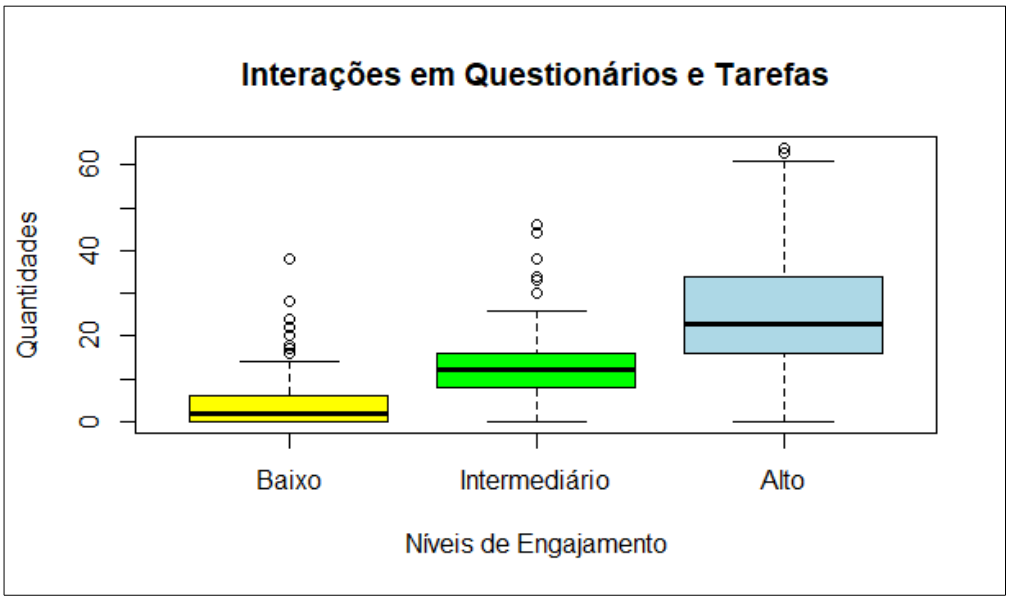

Gráfico 10- Interações em questionários/tarefas por nível de engajamento

A quantidade mínima de interações dos estudantes em questionários e tarefas dos três níveis de engajamento é zero, o valor máximo para o nível baixo é 38 (DP 7,04), para o intermediário é 46 (DP 9,06) e para o nível alto é 64 (DP 14,30). Percebe-se que os três níveis de engajamento apresentam alguns valores considerados fora do padrão (*outliers*). A média de interações é 4,71 para os estudantes do nível baixo, 13,05 para o nível intermediário e 25,69 para o nível alto de engajamento.

A próxima seção descreve a análise dos resultados com o foco no comportamento de procrastinação acadêmica dos estudantes durante o curso.

5.2 Análise dos agrupamentos quanto à procrastinação acadêmica

Seguindo várias pesquisas que estudaram a procrastinação acadêmica (Klassen *et al.*, 2008; Kandemir, 2014; Michinov *et al.*, 2001; Kim e Seo, 2015) e, em complemento à análise do engajamento comportamental dos estudantes, foi verificado se o comportamento procrastinante dos alunos tem relação com o resultado final dos mesmos, durante o período de realização dos questionários e tarefas, que devem ser submetidos pelo AVA.

Assim, fez-se a contagem de quantas vezes os estudantes finalizaram durante ou após o prazo proposto pelo professor da disciplina para cada uma dessas atividades, classificando-os de acordo com os níveis de procrastinação descritos na Tabela 11. Cabe ressaltar, que esta pesquisa utiliza uma abordagem similar a de Rotenstein *et al.* (2009), que não usa variáveis de procrastinação autorrelatadas pelos estudantes, mas sim, variáveis extraídas diretamente do ambiente virtual de aprendizagem.

Tabela 11– Níveis de procrastinação

Nível de Procrastinação	Descrição
Nunca	Estudantes que nunca procrastinaram em atividades do curso, incluindo aqueles reprovados por faltas (desistentes)
Baixo	Estudantes que procrastinaram em 1 ou 2 atividades
Médio	Estudantes que procrastinaram em 3 atividades
Alto	Estudantes que procrastinaram em 4 ou 5 atividades
Sempre	Estudantes que procrastinaram em todas as atividades do curso

Dessa forma, a hipótese que se formula é que não existe relação entre os níveis de procrastinação dos estudantes e seu respectivo resultado final (aprovado ou reprovado). Caso a hipótese nula esteja correta, não existe relação entre procrastinação e resultado final. Se encontrarmos associação, então existe relacionamento entre esses dois atributos, a hipótese alternativa é aceita e a hipótese nula é rejeitada.

A Tabela 12 mostra o resultado desta atividade, utilizando os testes de qui-quadrado de Pearson e Spearman com os estudantes agrupados pelos níveis de engajamento, resultantes da análise anterior.

Tabela 12– Resultados dos testes de qui-quadrado de Pearson e Spearman

Níveis de Engajamento	Grupos	Quantidade de Estudantes	Pearson *p-value* < 0,05	Spearrman *p-value* < 0,05
Engajamento Baixo	Grupos 1 e 7	100	0,1801	0,2131
Engajamento Intermediário	Grupos 4 e 6	117	0,0211	0,0428
Engajamento Alto	Grupos 2, 3 e 5	203	0,0145	0,0127

Para a categoria de estudantes classificados com baixo engajamento, o teste de Pearson resultou em *p-value* = 0,2576 e o teste de Spearman apresentou *p-value* = 0,2153, que representa um nível de significância maior do que 5%, sugerindo que não existe relação entre a procrastinação dos estudantes dessa categoria e os seus respectivos resultados finais na disciplina. Este resultado era esperado devido à grande quantidade de alunos reprovados que compõem a categoria de alunos com baixo engajamento, dos 100 alunos desse agrupamento, 99 apresentam a reprovação como resultado final.

Em contrapartida, os estudantes das categorias de engajamento intermediário e alto apresentam resultados estatisticamente significantes. Para a categoria intermediária, o teste de Pearson apresentou *p-value* = 0,0211 e o de Spearman resultou em *p-value* = 0,0428. Em relação aos estudantes da categoria de engajamento alto, o teste de Pearson apresentou *p-value* = 0,0145 e o teste de Spearman teve *p-value* = 0,0127. Ou seja, em ambas as categorias o nível de significância ficou menor do que 5%, sugerindo que há 95% de confiabilidade de que existe relação entre os níveis de procrastinação dos estudantes dessas duas categorias e os seus resultados finais (aprovado ou reprovado).

Para melhor ilustração e análise do comportamento procrastinante ou não procrastinante dos estudantes das turmas em estudo, necessitou-se fazer uma análise de tal comportamento ao longo das semanas de realização da disciplina. O Gráfico 11 abaixo apresenta uma amostra de onze estudantes que nunca procrastinaram no período do 1º e 2º questionários.

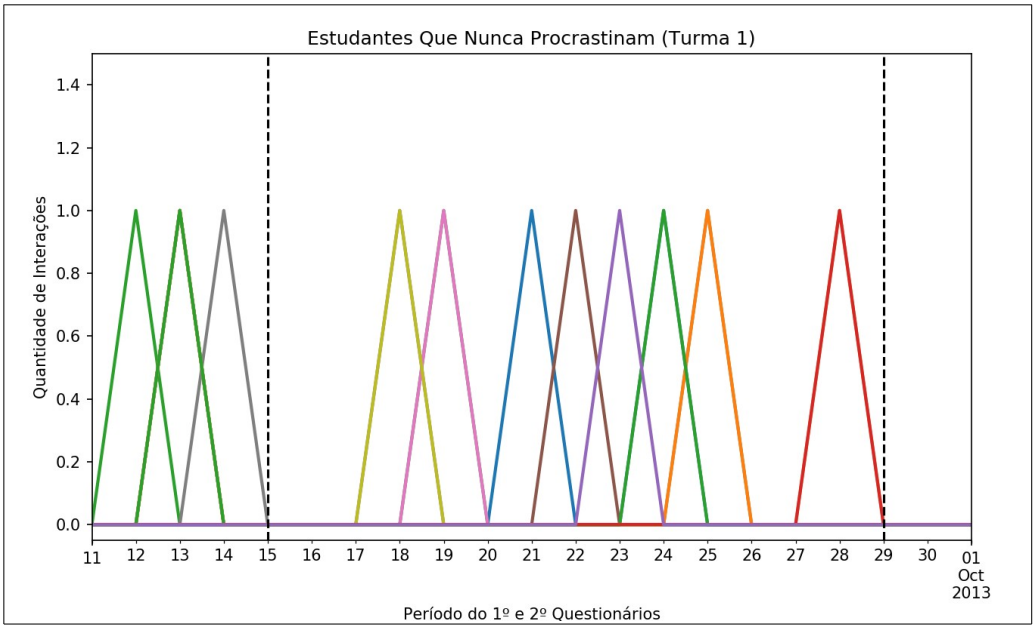

Gráfico 11- Alunos que nunca procrastinam – 1º e 2º Questionários

O gráfico da figura acima compreende o período de 11 de setembro a 1 de outubro de 2013, que corresponde às semanas de finalização do 1º e do 2º questionários do curso, cujas datas de vencimento são 15 Set e 29 Out, respectivamente, estando destacadas pelas linhas serrilhadas na vertical. Percebe-se claramente que esses estudantes não deixam para finalizar os questionários na data de vencimento ou após a mesma, chegando no máximo a finalizá-los no dia anterior ao vencimento (14 Set e 28 Out). Outro detalhe percebido, é que a maioria desses estudantes submeteram a finalização do 2º questionário bem antes da data de vencimento, entre 18 e 25 de setembro.

O Gráfico 12 mostra o mesmo período de atividades do 1º e 2º questionários, porém com uma amostra de sete estudantes possuem o comportamento procrastinante.

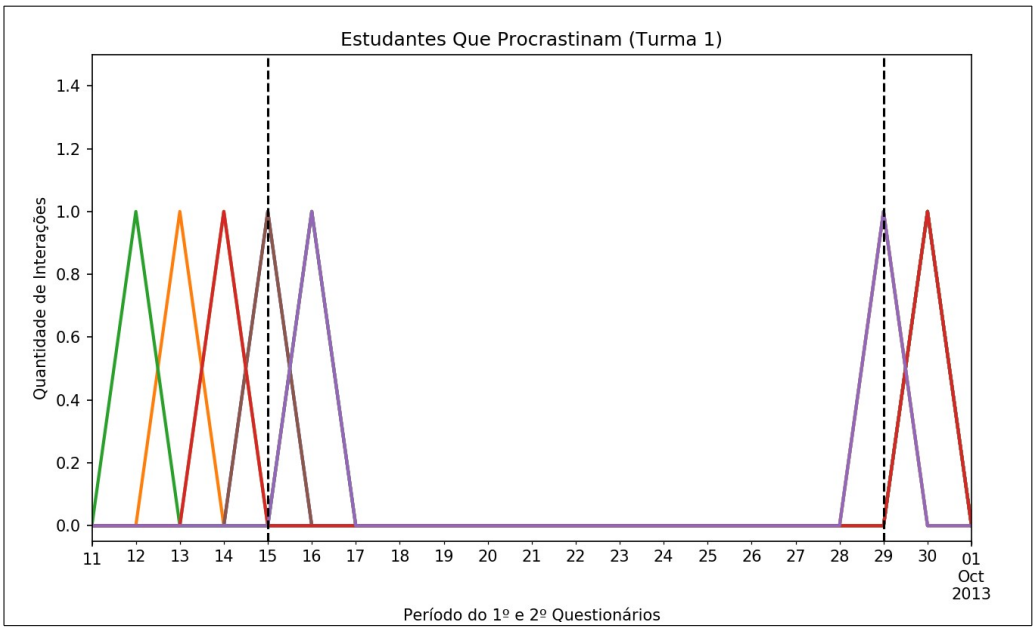

Gráfico 12- Alunos que procrastinam – 1º e 2º Questionários

Nota-se que o comportamento desses estudantes, quanto à finalização dos questionários em tela, é diferente do grupo de estudantes anterior. Esses alunos deixam para finalizar os questionários bem próximo aos dias de vencimento, nos próprios dias de vencimento (15 Set e 29 Out) e após esses dias (16 Set e 30 Out).

O Gráfico 13 apresenta uma amostra de outros quinze estudantes que nunca postergaram a finalização dos 3º e 4º questionários do curso, no período de 14 de novembro a 3 de dezembro de 2013. As duas linhas serrilhadas marcam a data de vencimento dos questionários no AVA, que correspondem a 17 de novembro e 01 de dezembro, respectivamente.

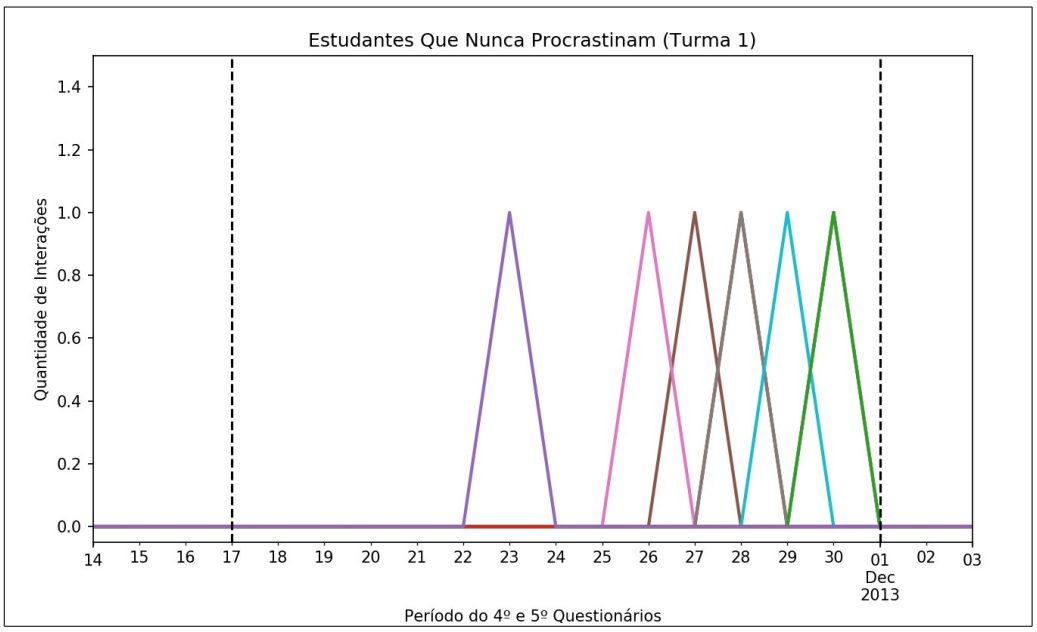

Gráfico 13- Alunos que nunca procrastinam – 4º e 5º Questionários

É possível perceber que os estudantes plotados neste gráfico apresentam comportamento semelhante aos do Gráfico 16, ilustrando o comportamento não-procrastinante destes alunos.

O Gráfico 14 abaixo, correspondente ao período de atividades dos 3º e 4º questionários, exibe uma amostra de quinze estudantes da mesma turma, classificados com níveis alto e sempre de procrastinação.

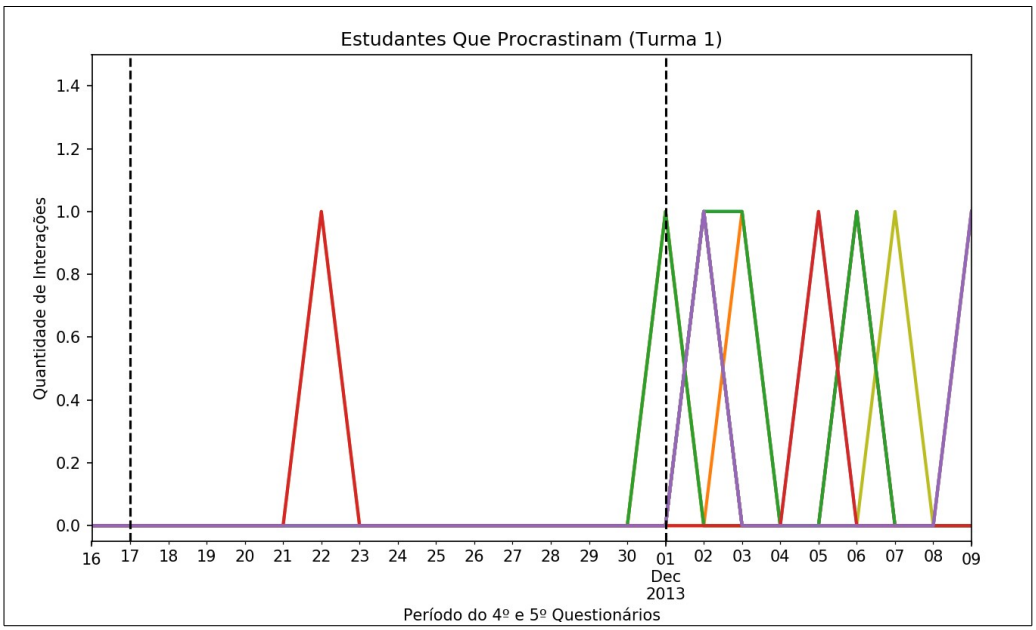

Gráfico 14- Alunos que procrastinam – 4º e 5º Questionários

Pode-se notar que o comportamento diferente entre os dois grupos de estudantes também se repete nas semanas do 3º e 4º questionários. Os estudantes que não procrastinam, conseguem finalizar os questionários antes da data de vencimento ou, no máximo, um dia antes desta data. Enquanto os estudantes que procrastinam, chegam a finalizar o questionário após a data de vencimento, como ilustrado nesta figura, onde a maioria deles finalizou o 5º questionário após a data de vencimento, que foi 1º de dezembro.

6 CONCLUSÃO

Esta pesquisa utiliza a abordagem de aprendizado de máquina não-supervisionado com o objetivo de para agrupar estudantes da disciplina de 1º período do curso de licenciatura a distância, levando em conta os comportamentos de engajamento e procrastinação acadêmicos de alunos.

Segundo a literatura, o engajamento possui relação direta e/ou positiva com o desempenho acadêmico dos estudantes. Já quanto ao comportamento procrastinante de estudantes, encontram-se na literatura trabalhos que concluem que esse comportamento tem relação negativa, positiva e nenhuma relação com o desempenho final dos alunos, conforme apresentado no referencial teórico desta pesquisa.

Inicialmente é realizada a revisão da literatura para fazer um levantamento dos atributos derivados de *logs* do Moodle citados em pesquisas de MDE, objetivando a criação da maior quantidade possível de tais atributos a partir do conjunto de dados disponível para a pesquisa, utilizando o conceito de *learning analytics*.

Depois os atributos foram dispostos em formato de séries temporais para utilização do método de clusterização das séries, com o objetivo de agrupar os estudantes em níveis similares de engajamento comportamental. Para, em seguida, analisar a correlação entre o comportamento de procrastinação do estudante na disciplina.

Quanto ao agrupamento de estudantes classificados com baixo nível de engajamento, formado pelos Grupos 1 e 7, os quais contêm a grande maioria dos estudantes reprovados do conjunto de dados, 98,3% para o Grupo 1 e 100% para o Grupo 7, nota-se que o algoritmo agrupou os estudantes com quantidade muito baixa de interações no fórum de discussão da disciplina e nos questionários e tarefas do ambiente virtual, podendo haver associação com o mal desempenho apresentado por esses alunos na disciplina.

Em relação aos Grupos 2, 3 e 5, classificados com estudantes com alto nível de engajamento, verificou-se que os três grupos contêm a maioria de estudantes aprovados do conjunto de dados, Grupo 2 com 60%, Grupo 3 com 68,3% e Grupo 5 com 73,8%, os quais apresentam os maiores quantitativos de interações no fórum

de discussão da disciplina e nos questionários e tarefas do AVA, sendo o Grupo 5 classificado como o grupo mais engajado, de onde foram tiradas as médias de interações utilizadas para classificação dos níveis de engajamento.

Ainda, nesta análise foi possível perceber que o Grupo 2 possui 71,2% de alunos com desempenho acima de BOM, Grupo 3 tem 63,8% de alunos com esse mesmo desempenho, bem como, o Grupo 5 possui 83,1%. Dessa forma, pode existir associação entre o alto desempenho apresentado pelos estudantes desses grupos e as interações dos objetos de aprendizagem considerados.

Quanto ao agrupamento de estudantes classificados com nível intermediário de engajamento comportamental, composto pelos Grupos 4 e 6, nota-se que ambos possuem tanto estudantes aprovados quanto reprovados, sendo que em sua maioria são estudantes reprovados. O Grupo 4 possui 56,5% e o Grupo 4 tem 77,1% de estudantes reprovados, valores equivalentes aos alunos com desempenho INSUFICIENTE. Em relação ao desempenho final dos alunos, verifica-se que o Grupo 4 apresenta 30,43% de alunos com desempenho acima de BOM e o Grupo 6 tem 22,92%, evidenciando que esse agrupamento possui índices, tanto de aprovação quanto de desempenho, posicionados entre os agrupamentos com baixo e alto nível de engajamento.

Quanto à análise da relação entre o comportamento procrastinante dos estudantes e o resultado apresentado ao final da disciplina (aprovado ou reprovado), verifica-se que não há relação entre o agrupamento de alunos com baixo engajamento, muito provavelmente devido à amostra conter 99% de estudantes reprovados. Desses, grande parte reprovados por falta (desistentes), isto é, sem nenhuma interação no AVA.

Em contrapartida, para os agrupamentos de estudantes com níveis de engajamento intermediário e alto o resultado apresentou relação negativa, podendo-se inferir que os estudantes desses dois agrupamentos que possuem comportamento com níveis baixos de procrastinação apresentam melhores desempenhos e, consequentemente, melhores resultados (aprovação) ao final da disciplina.

Dessa forma, o método proposto nesta pesquisa apresenta as seguintes contribuições: (I) levantamento para concentrar em um único trabalho a maior quantidade possível de atributos derivados das interações de professores, tutores e alunos, registradas em *logs* do Moodle e citados em outros trabalhos de pesquisadores de MDE e LA; (II) abordagem do tema mineração de dados educacionais com aprendizado de máquina não-supervisionado, analisando numa mesma pesquisa o engajamento comportamental e a procrastinação acadêmica de estudantes da EaD, estendendo o estudo para uma área demográfica da região de um país, mais especificamente, um estado da região nordeste do Brasil; (II) utilização da técnica de aprendizado de máquina não-supervisionado dispondo em formato de séries temporais os atributos derivados do AVA, haja vista o caráter temporal dos dados armazenados nos *logs* do Moodle, abordagem ainda muito pouco explorada na área de MDE e LA.

7 TRABALHOS FUTUROS E LIMITAÇÕES

Como trabalhos futuros sugere-se aplicar a mesma abordagem de aprendizado de máquinas não supervisionado desta pesquisa, levando em consideração as notas parciais dos estudantes na realização das tarefas e dos questionários quinzenais, para analisar se os comportamentos de engajamento e procrastinação acadêmicos se repetem ou não ao longo do período da disciplina. De tal forma, que proporcione um melhor acompanhamento por parte dos professores, bem como a aplicação de metodologias e/ou estratégias didático-pedagógicas nas salas de aula virtuais, com vistas a motivar o estudante durante o seu percurso na disciplina, para melhorar o seu desempenho ao final do semestre.

Outra sugestão apresentada é fazer uma abordagem supervisionada usando previsão de séries temporais, utilizando as notas parciais dos objetos de aprendizagem em pauta, como rótulos parciais das séries para criação de um modelo de predição que possa prever tanto o desempenho parcial do aluno na próxima quinzena quanto o desempenho final na disciplina, fazendo uma comparação dos resultados de cada quinzena para averiguar qual apresentou a melhor acurácia na tarefa de predição.

Como fatores limitantes desta pesquisa, pode-se destacar a falta de acesso direto ao banco de dados do AVA Moodle; não disponibilização das notas dos estudantes das avaliações *on-line* referentes às tarefas, questionários e fóruns de discussão; e falta de contato com os estudantes para entrevistas ou *surveys*. Esses fatores foram desafiadores em relação à criação de atributos necessários para a presente pesquisa.

REFERÊNCIAS

ACKERMAN, D. S., & GROOS, B. L. **My instructor made me do it: Task characteristics of procrastination**. Journal of Marketing Education, v. 27, pp. 5–13, 2005.

ÁGUDO-PEREGRINA, A. F.; IGLESIAS-PRADA, S.; CONDE-GONZÁLEZ, M. A.; HERNÁNDEZ-GARCÍA, A. **Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-Supported F2F and online learning**. Computers in Human Behavioral, v. 31, pp. 542-550. https://doi.org/10.1016/j.chb.2013.05.031, 2014.

ALMEIDA, R. O. M. **Três ensaios com aplicações de redes neurais em séries financeiras**. Tese de Doutorado defendida na Faculdade de Economia, Administração e Contabilidade -FEA, 2004.

ALMEIDA, O. C. S.; ABBAD, G.; MENESES, P. P. M.; ZERBINI, T. **Evasão em Cursos a Distância: Fatores Influenciadores**. Revista Brasileira de Orientação Profissional, Vol. 14, Nº 1, pág. 19-33, 2013.

ANDERSON, A.; HUTTENLOCHER, D.; KLEINBERG, J.; LESKOVEC, J. **Engaging with Massive Online Courses**. https://arxiv.org/abs/1403.3100, 2014.

BAKER, R. S. J. D.; YACEF, K. **The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining**, Article 1, v.1, n.1, 2009.

BAKER, R.; ISOTANI, S.; CARVALHO, A. **Mineração de Dados Educacionais: Oportunidades para o Brasil**, 2011.

BARBOSA, M. A.; FERREIRA, A.; PACHECO, M. M. **Programa e-TEC Brasil: a experiência do Instituto Federal do Paraná/EAD**. http://www.sinect.com.br/anais2012/html/artigos/educacao%20tec/1.pdf. Acessado em 16 de junho de 2018, 2012.

BEER, C.; CLARK K.; JONES, D. **Indicators of engagement**. ASCILITE 2010 - The Australasian Society for Computers in Learning in Tertiary Education. pp. 75-86, 2010.

BOX, G. E. P.; JENKINS, G. M. **Times series analysis: forecasting and control**. 3 ed. New Jersey: Prentice Hall, 1994.

CENSO EAD.BR 2016. **Relatório analítico da aprendizagem a distância no Brasil**. Associação Brasileira de Educação a Distância (ABED), 2016. Acessado em 23/01/2018.

CEREZO, R.; ESTEBAN, M.; SÁNCHEZ-SANTILLÁN, M.; NUÑES, J. C. **Procrastinating behavior in computer-based learning environments to predict performance: a case study in moodle**. Frontiers in Psychology 8:1403. doi: 10.3389/fpsyg.2017.01403, 2017.

CHOI, J. N.; MORAN, S.V. **Why not procrastinate? Development and validation of a new active procrastination scale**. The Journal of Social Psychology, v. 149(2), pp. 195-211, 2009.

EHLERS, R. S. **Análise de séries temporais**. 4 ed. Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná, 2007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. AI magazine, [S.l.], v.17, n.3, p.37, 1996.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. **Knowledge discovery in databases: An overview**. AI magazine, v. 13, n. 3, p. 57, 1992.

GAFNI, R.; GERI, N. **Time management: procrastination tendency in individual and collaborative tasks**. Interdisciplinary Journal of Information, Knowledge, and Management, v. 5, pp. 115-125, 2010.

GASEVIC, D.; DAWSON, S.; ROGERS, T.; GASEVIC, D. **Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success**. Internet and Higher Education, v. 28, pp. 68-84. https://doi.org/10.1016/j.iheduc.2015.10.002, 2015.

GODA, Y.; YAMADA, M.; KATO, H.; MATSUDA, T.; SAITO, Y.; MIYAGAWA, H. **Procrastination and other learning behavioral types in e-learning and their relationship with learning outcomes**. Learning and Individual Differences, v. 37, pp. 72–80. http://dx.doi.org/10.1016/j.lindif.2014.11.001, 2015.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. **Predição de desempenho de estudantes em cursos EaD utilizando mineração de dados: uma estratégia baseada em séries temporais**. In Anais do 23º Simpósio Brasileiro de Informática na Educação (SBIE 2012), 2012.

GRELLER, W.; DRACHSLER, H. **Translating learning into numbers: a generic framework for learning analytics**. Educational Technology & Society, 2012, v.15(3), pp.42-57, 2012.

IGLESIAS-PRADA, S.; RUIZ-DE-AZCÁRATE, C.; AGUDO-PEREGRINA, A. F. **Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies**. Computers in Human Behavioral, v. 47, pp. 81-89. https://doi.org/10.1016/j.chb.2014.09.065, 2015.

JACCARD, P. **Novelles recgerches sur la distribution florale**. Bulletin de la Societé Vaudoise des Sciences Naturelles, 44, 223–270, 1908.

JOKSIMOVIC, S.; GASEVIC, D.; LOUGHIN, T. M.; KOVANOVIC, V.; HATALA, M. **Learning at distance: efects of interaction traces on academic achievement**. Computers and Education, v. 87, pp. 204-217. https://doi.org/10.1016/j.compedu.2015.07.002, 2015.

KANDEMIR, M. **Reasons of academic procrastination: self- regulation, academic self-efficacy, life satisfaction and demographics variables**. Procedia – Social andd Behavioral Sciences, v. 152, pp. 188-193, 2014.

KASSAMBARA, A. **Pratical Guide to Cluster Analysis in R**. Publicado em STHDA (http:\\www.sthda.com), 2017.

KENSKI, V. **O desafio da educação a distância no Brasil**. https://www.researchgate.net/publication/267697506. Acessado em 16 de junho de 2018, 2018.

KHALIL, M.; EBNER, M. **Learning analytics: Principles and Constraints**. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2015. pp1326-1336. Chesapeake, VA, AACE, 2015.

KHALIL, M.; KASTL, C.; EBNER, M. **Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics**. Proceedings of the European Stakeholder Summit on experiences and best practices in and around MOOCs (EMOOCS 2016) , Graz, Áustria, pp.265-278, 2016.

KIM, K. R.; SEO, E. H. **The relationship between procrastination and academic performance: a meta-analysis**. Personality and Individual Differences 82 (2015), pp. 26-33. http://dx.doi.org/10.1016/j.paid.2015.02.038, 2015.

KIRCHGASSNER, G.; WOLTERS, J. **Introduction to Modern Time Series Analysis**. Springer, 2007.

KLASSEN, R. M.; KRAWCHUK, L. L.; RAJANI, S. **Academic procrastination of undergraduates: Low self-efficacy to self regulate predicts higher levels of procrastination**. Contemporary Educational of Psychology, 33 (2008), pp. 915-931. http://www.sciencedirect.com. Acessado em janeiro de 2018, 2008.

KOVANOVIC, V.; JOKSIMOVIC, S.; GASEVIC, D.; OWERS, J.; SCOTT, A. M.; WOODGATE, A. **Profiling MOOC Course Returners: How Does Student Behavior Change Between Two Course Enrollments?**. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale, pp. 269-272, 2016.

LI, Q. & BAKER R. **The different relationships between engagement and outcomes across participants subgroups in Massive Open Online Courses**. Computers & Education (2018), doi:10.106/j.compedu.2018.08.005, 2018.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bsmsp/1200512992. Acessado em fevereiro de 2018.

MARTIN, F.; WHITMER, J. C. **Tech Know Learn 21: 59**. https://doi-org.ez19.periodicos.capes.gov.br/10.1007/s10758-015-9261-9, 2016.

MEYER, A. S. **Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes**. Dissertação de Mestrado da Escola Superior de Agricultura “Luiz Queiroz”, Universidade de São Paulo, 2002.

MICHINOV, N.; BRUNOT, S.; LE BOHEC, O.; JUHEL, J.; DELAVAL, M. **Procrastination, participation and performance in online learning environments**. Computer and Education 56 (2011), pp. 243-252. https://doi:10.1016/j.compedu.2010.07.025, 2011.

MITCHELL, T. M. **Machine Learning**. 1 ed. Redmond,WA. Editora McGraw-Hill Science, 1997.

MLYNARSKA, E.; GREENE, D.; CUNNINGHAM, P. **Time Series Clustering of Moodle Activity Data**, 2016.

MORETTIN, P. A., TOLOI, C. M. C. **Análise de Série Temporais**. 1 ed. São Paulo, SP. Editora Edgard Blucher, 2004.

PACHECO, J. C.; MAIA, S. F. **Escola Técnica Aberta do Brasil: Desenvolvimento Local?**. http://leg.ufpi.br/subsiteFiles/ppged/arquivos/files/VI.encontro.2010/GT.17/GT_17_07_2010.pdf. Acessado em 17 de junho de 2018, 2010.

PALMER, S. **Modelling Engineering Student Academic Performance Using Academic Analytics**. International Journal of Engineering Education. v. 29, pp. 132-138, 2013.

PARSONS, J., & TAYLOR, L. **Student Engagement: What do we know and what should we do?**. University of Alberta, 2011.

PEÑA-AYALA, A. **Educational data mining: A survey and data mining-based analysis of recent works**. Expert Systems with Applications, v. 41, I. 4, Part 1, 2014, pp. 1432-1462, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2013.08.042, 2014.

PINTRICH, P. R.; SMITH, D. A. F.; GARCIA, T.; MCKEACHIE, W. J. **Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ)**. Educational and Psychological Measurement, v. 53, pp. 801–813, 1993.

QUEIROGA, E. M., CECHINEL, C., ARAÚJO, R. M. **Predição de estudantes com risco de evasão em cursos técnicos à distância**. Anais do XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017) do VI Congresso Brasileiro de Informática na Educação (CBIE 2017), 2017.

PORTAL, C. **Estratégia para minimizar a evasão e potencializar a permanência em EaD a partir de sistema que utiliza mineração de dados educacionais e *learning analytics***. Dissertação de Mestrado do Programa de Pós-graduação da Universidade do Vale do Rio dos Sinos - UNISINOS, 2016.

RAGA Jr. R. C.; RAGA, J. D. **Monitoring class activity and predicting student performance using moodle action log data**. In International Journal of Computing Sciences Research, v. 1, n. 3, pp. 1-16, 2017. https://stepacademic.net. Acessado em janeiro de 2018.

RAMOS, J. L. C.; SILVA, R. F. P.; SILVA, J. C. S.; GOMES, A. S. **Estudo comparativo entre ambientes virtuais para uso em *Blended Learning* na Universidade Federal do Vale do São Francisco**. In Anais do XXI Workshop de Informática na Escola (WIE 2015), 2015.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. 1 ed. Barueri, SP. Editora Manole Ltda, 2005.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. **O uso da mineração de texto para extração e organização não supervisionada de conhecimento**. Revista de Sistemas de Informação da FSMA, n. 7 (2011), pp. 7-21, 2011.

RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S.; GOMES, A. S. **A Literatura brasileira sobre mineração de dados educacionais**. In Anais do III Congresso Brasileiro de Informática na Educação (CBIE, 2014), 2014.

RODRIGUES, R. L. **Uma abordagem de Mineração de Dados Educacionais para previsão de desempenho a partir de padrões comportamentais de Autorregulação da Aprendizagem**. Tese de Doutorado, 2016.

RODRIGUES, R. L.; RAMOS, J. C. L.; SILVA, J. C. S.; GOMES, A. S. **Discovery Engagement Patterns MOOCs Through Cluster Analysis**. IEEE Latin America Transactions, Vol. 14, N. 9, September 2016, 2016(a)

ROMERO, C.; VENTURA, S. **Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 2013.

ROSENBERG, M. J. **Conceiving The Self**. New York: Basic Books, 1979.

ROTENSTEIN, A.; DAVIS, H. Z.; TATUM, L. **Early Birds versus Just-in-Timers: The effect of procrastination on academic performance of accounting students**. Journal of Accounting Education, v. 27, i. 4, pp. 223-232. https://doi.org/10.1016/j.jaccedu.2010.08.001, 2009.

ROTHBLUM, E. D.; SOLOMON, L. J.; MURAKAMI, J. **Affective, cognitive, and behavioral differences between high and low procrastinators**. Journal of Counseling Psychology, 33, pp. 387-394, 1986.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3 ed. Rio de Janeiro, RJ. Editora Elsevier Campus, 2013.

SAEL N.; MARZAK A.; BEHJA, H. **Multilevel clustering and association rule mining for learners profiles analysis**. In IJCSI International Journal of Computer Science Issues, v. 10, i. 3, n. 1, 2013.

SANTOS, R. M. M.; PITANGUI, C. G.; ANDRADE, A. V.; ASSIS, L. P. **Uso de séries temporais e seleção de atributos em mineração de dados educacionais para previsão de desempenho acadêmico**. Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016) do V Congresso Brasileiro de Informática na Educação (CBIE 2016), 2016.

SIEMENS, G.; BAKER, R. S. J. D. **Learning analytics and educational data mining: towards communication and collaboration**, 2012.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados com aplicações em R**. 1ª edição. Rio de Janeiro. Elsevier Editora Ltda., 2016.

SOLAR 2011. **Open Learning Analytics: an integrated & modularized platform**. Society for Learning Analytics Research. https://solaresearch.org/wp-content/uploads/2011/12/OpenLearningAnalytics.pdf, 2011. Acessado em 23 de janeiro de 2018.

SOLOMON, L. J., & ROTHBLUM, E. D. **Academic procrastination: Frequency and cognitive-behavioral correlates**. Journal of Counseling Psychology, 31, pp. 503–509. http://dx.doi.org/10.1037//0022-0167.31.4.503, 1984.

SURYAKANT M.; TRIPTI M. **A new similarity measure based on mean measure of divergence for collaborative filtering in sparse environment**. Procedia Computer Science, 89 (2016), pp. 450-456. https://doi: 10.1016/j.procs.2016.06.099, 2016.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Editora Pearson Addison-Wesley. 1 ed. Boston-MA, 2006.

TUCKMAN, B. W. **The development and concurrent validity of the procrastination scale**. Educational and Psychological Measurement, v. 51, pp. 473-480. http://dx.doi.org/10.1177/0013164491512022, 1991.

TROWLER, V. **Student Engagement Literature Review**. 2010.

TRUEMAN, M. & HARTLEY, J. **A comparison between the time management skills and academic performance of mature and traditional-entry university students**. Higher Education, v. 32, pp. 199–215, 1996.

WITTEN, I. H.; FRANK, E. **Data Mining: Pratical Machine Learning Tools and Techniques**. 2ª Edição, San Francisco - CA, 2005.

YIN, Y.; YASUDA, K. **Similarity coefficient method applied to the cell information problem: a comparative investigation**. Computers and Industrial Engineering 48 (2005), pp. 471-489. https://doi:10.1016/j.cie.2003.01.001, 2005.

ZACHARIS, N.; **A multivariate approach to predicting student outcomes in web-enabled blended learning courses**. Internet and Higher Education, v. 27, pp. 44-53, https://doi.org/10.1016/j.iheduc.2015.05.002, 2015.

ZIMMERMAN, B. J.; BANDURA, A.; MARTINEZ-PONS, M. **Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal-setting**. American Educational Research Journal, v. 29, pp. 663–676, 1992.

APÊNDICE A

ATRIBUTO	DESCRIÇÃO	AUTORES QUE CITARAM
Logins de acesso	Quantidade de logins realizadas pelo aluno no AVA	Khalil e Ebener (2015); Iglesias-Prada *et al.* (2015); Joksimovic *et al.* (2015); Gasevic *et al.* (2015); Palmer (2013); Martin e Whitmer (2016); Gottardo *et al.* (2012); Santos *et al.* (2016); Rodrigues *et al.* (2016); Raga Jr. e Raga (2017)
Clicks do mouse	Quantidade de clicks do mouse do aluno quando acessou o AVA	Khalil e Ebener (2015); Zacharis (2015)
Recursos acessados	Quantidade dos módulos diferentes do AVA acessados pelo aluno.	Khalil e Ebener (2015); Agudo-Peregrina *et al.* (2014); Gasevic *et al.* (2015); Palmer (2013); Martin e Whitmer (2016); Sael *et al.* (2013)
Tarefas concluídas	Quantidade de tarefas concluídas e entregues no AVA	Khalil e Ebener (2015); Joksimovic *et al.* (2015); Gasevic *et al.* (2015); Zacharis (2015); Anderson *et al.* (2014); Santos *et al.* (2016); Rodrigues *et al.* (2016)
Visualização de tarefas	Quantidade de vezes que o aluno apenas acessou/visualizou a tarefa	Anderson *et al.* (2014); Santos *et al.* (2016); Raga Jr e Raga (2017)
Visualização de vídeos	Quantidade de vídeos vistos pelo aluno	Khalil e Ebener (2015); Agudo-Peregrina *et al.* (2014); Santos *et al.* (2016)
Documentos acessados	Quantidade de documentos vistos pelo aluno	Khalil e Ebener (2015); Agudo-Peregrina *et al.* (2014); Gasevic *et al.* (2015); Zacharis (2015)
Arquivos baixados	Quantidade de arquivos baixados do AVA pelo aluno	Khalil e Ebener (2015)
Material didático visualizado	Quantidade de livros-textos acessados	Agudo-Peregrina *et al.* (2014); Gasevic *et al.* (2015); Santos *et al.* (2016)

Recursos multimídias visualizados	Quantidade de arquivos de áudios acessados	Agudo-Peregrina *et al.* (2014); Gasevic *et al.* (2015); Santos *et al.* (2016)
Mensagens enviadas em fóruns	Contagem das mensagens enviadas em fórum de discussão	Joksimovic *et al.* (2015); Gasevic *et al.* (2015); Gottardo *et al.* (2012); Santos *et al.* (2016); Rodrigues *et al.* (2016); Greller e Drachsler (2012); Khalil e Ebener (2015); Iglesias-Prada *et al.* (2015); Agudo-Peregrina *et al.* (2014); Zacharis (2015); Palmer (2013); Anderson *et al.* (2014); Raga Jr. e Raga (2017)
Mensagens lidas nos fóruns	Contagem de posts visualizados/lidos nos fóruns	Joksimovic *et al.* (2015); Palmer (2013); Anderson *et al.* (2014); Gottardo *et al.* (2012); Santos *et al.* (2016); Rodrigues *et al.* (2016); Raga Jr e Raga (2017); Nick Zacharis (2015)
Mensagens enviadas em chats	Contagem das mensagens enviadas em chats	Joksimovic *et al.* (2015); Gasevic *et al.* (2015); Gottardo *et al.* (2012); Santos *et al.* (2016); Rodrigues *et al.* (2017); Agudo-Peregrina *et al.* (2014); Palmer (2013)
Mensagens lidas em chats	Quantidade de mensagens lidas Chats	Joksimovic *et al.* (2015); Nick Zacharis (2015); Gottardo *et al.* (2012); Santos *et al.* (2016)
Mensagens enviadas por e-mail	Contagem das mensagens enviadas por e-mail	Joksimovic *et al.* (2015); Gasevic *et al.* (2015)
Mensagens lidas no e-mail	Quantidade de mensagens lidas em e-mail	Joksimovic *et al.* (2015); Nick Zacharis (2015)
Tamanho das mensagens do e-mail	Tamanho das mensagens enviadas por e-mail	Joksimovic *et al.* (2015)
Sessões on-line	Quantidade total de sessões web iniciadas no AVA.	Joksimovic *et al.* (2015); Palmer (2013); Gottardo *et al.* (2012); Santos *et al.* (2016); Rodrigues *et al.* (2016); Sael *et al.* (2013)
Início de acesso ao AVA	Refere-se a quantos dias, após iniciado o curso, o aluno fez o seu primeiro acesso ao AVA.	Joksimovic *et al.* (2015); Palmer (2013); Rodrigues *et al.* (2016); Sael *et al.* (2013); Gottadro *et al.* (2012)

Criação de conteúdo em wiki	Quantidade de criações de conteúdo em wiki	Zacharis (2015); Santos *et al.* (2016)
Criação de conteúdo em blog	Quantidade de criações de conteúdo em blog	Zacharis (2015); Santos *et al.* (2016)
Início de acesso a um recurso disponibilizado pelo professor	Refere-se a quantos dias, após o material ser disponibilizado pelo professor, o aluno o acessa pela primeira vez.	Martin e Whitmer (2016); Rodrigues *et al.* (2016)
Interações com outros alunos nos fóruns	Registra se o aluno interagiu com o professor ou outros alunos no fórum, bem como se outros alunos, ao menos, visualizaram a mensagem no fórum, contabilizando a quantidade de alunos diferentes que interagiram com ele.	Rodrigues *et al.* (2016); Gottardo *et al.* (2012)
Interações com Professores e Tutores nos fóruns	Quantidade de interações feitas pelo aluno com professores e tutores nos fóruns	Rodrigues *et al.* (2016); Gottardo *et al.* (2012)
Interações com outros alunos em chats	Quantidade de interações com outros alunos em chats	Gottardo *et al.* (2012)
Interações com professores e tutores em chats	Quantidade de interações feitas pelo aluno com professores e tutores nos chats	Gottardo *et al.* (2012)
Tópicos criados nos fóruns	Registra a quantidade de tópicos criados pelo aluno nos fóruns	Rodrigues *et al.* (2016)
Questionários finalizados	Quantidade de questionários que o aluno finalizou e enviou no AVA	Raga Jr. e Raga (2017)
Tentativas para resolver os questionários	Quantidade de vezes que o aluno tentou resolver o questionário antes de enviar	Raga Jr. e Raga (2017)
Dias entre o	Quantidade de dias entre o primeiro e o último acesso no AVA	Sael *et al.* (2013)

primeiro e o último login		
Endereços IP de onde o aluno acessou o Moodle	Quantidade de endereços IP diferentes de onde o aluno fez login no AVA	Rodrigues *et al.* (2016)
Mensagens enviadas ao professor ou tutor	Quantidade de mensagens enviadas ao professor ou tutor usando o AVA	Rodrigues *et al.* (2016)
Visualização da seção de conteúdos	Quantidade de vezes que o aluno visualizou a seção de conteúdos no AVA	Rodrigues *et al.* (2016)
Tarefas feitas pelo aluno após a data de vencimento	Quantidade de tarefas feitas pelo aluno após a data de vencimento das mesmas.	Rodrigues *et al.* (2016)
Visualizações de notas das atividades	Quantidade de vezes que o aluno visualizou o relatório de suas notas das atividades feitas no AVA	Rodrigues *et al.* (2016)
Timeouts de sessões	Quantidade de timeouts de sessões no AVA	Rodrigues *et al.* (2016)
Dias diferentes de acesso ao AVA	Quantidade de dias diferentes que o aluno acessou o AVA	Rodrigues *et al.* (2016)
Acesso ao AVA por período do dia	Quantidade de vezes que o aluno acessou o AVA, por período do dia (manhã, tarde, noite ou madrugada)	Rodrigues *et al.* (2016)
Respostas do professor a perguntas do estudante no fórum	Quantidade de respostas do professor a perguntas do estudante no fórum	Rodrigues *et al.* (2016); Gottardo *et al.* (2012)
Questões respondidas	Quantidade de questões respondidas corretamente nos questionários	Gottardo *et al.* (2012)

corretamente		
Frequência de logins	Frequência diária com que o aluno acessa o AVA	Joksimovic *et al.* (2015); Zacharis (2015); Gottardo *et al.* (2012)
Tempo médio por login	Tempo médio gasto pelo aluno por acesso durante o semestre	Martin e Whitmer (2016); Gottardo *et al.* (2012); Rodrigues *et al*. (2016); Sael *et al.* (2013)
Tempo entre duas mensagens	Quantidade de tempo entre o envio de duas mensagens	Joksimovic *et al.* (2015)
Tempo total on-line	Quantidade total de tempo gasta em todas as sessões	Joksimovic *et al.* (2015); Palmer (2013); Martin e Whitmer (2016); Gottardo *et al.* (2012); Rodrigues *et al.* (2016)
Tempo gasto na leitura de mensagens postadas no fórum	Tempo gasto durante a leitura das mensagens postadas no fórum.	Joksimovic *et al.* (2015); Rodrigues *et al.* (2016)