UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RAFAEL CASTRO DE SOUZA

APLICAÇÃO DE LEARNING ANALYTICS PARA AVALIAÇÃO DO DESEMPENHO DE TUTORES A DISTÂNCIA

MOSSORÓ – RN

2016

RAFAEL CASTRO DE SOUZA

APLICAÇÃO DE LEARNING ANALYTICS PARA AVALIAÇÃO DO DESEMPENHO DE TUTORES A DISTÂNCIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação – associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Francisco Milton Mendes Neto – UFERSA.

Coorientador: Prof. Dr. Araken de Medeiros Santos – UFERSA.

MOSSORÓ – RN

2016

719a Souza, Rafael Castro de.
APLICAÇÃO DE LEARNING ANALYTICS PARA AVALIAÇÃO DO DESEMPENHO DE TUTORES A DISTÂNCIA / Rafael Castro de Souza. - 2016.
108 f. : il.

Orientador: Francisco Milton Mendes Neto.
Coorientador: Araken de Medeiros Santos.
Dissertação (Mestrado) - Universidade Federal Rural do Semi-árido, Programa de Pós-graduação em Ciência da Computação, 2016.

1. Learning Analytics. 2. Avaliação de Tutores. 3. Indicadores de desempenho comportamental na educação. I. Neto, Francisco Milton Mendes, orient. II. Santos, Araken de Medeiros, co-orient. III. Título.

RAFAEL DE CASTRO SOUZA

APLICAÇÃO DE LEARNING ANALYTICS PARA AVALIAÇÃO DO DESEMPENHO DE TUTORES A DISTÂNCIA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 27 / 09 / 2016

Prof. Dr. Francisco Milton Mendes Neto
(Orientador - UFERSA)

Prof. Dr. Araken de Medeiros Santos
(Coorientador - UFERSA)

Prof. Dr. Rogério Patrício Chagas do Nascimento
(Examinador Externo - UFS)

Prof. Dr. Patrício de Alencar Silva
(Examinador Interno - UFERSA)

DEDICATÓRIA

Dedico este trabalho a Jesus Cristo. O Senhor da minha vida.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, o Autor da vida, pela sua companhia, apoio, e por ter me dado a sua benção para que mais uma etapa da minha vida pudesse ser concluída.

Aos meus pais Marcos Batista de Souza e Otaciana Maria pelo carinho, apoio e incentivo. Saibam que todos vocês me ensinaram ao longo da vida que o trabalho e dedicação são as chaves para a vitória.

Aos meus avós João Batista de Souza *in memorian*, Terezinha Luzia, João de Deus e Maria Salete que sempre me apoiaram, estiveram na torcida por mim, e ensinaram não só a mim ao longo da vida, como também aos meus pais.

Agradeço a compreensão de toda a minha família por entenderem que em alguns momentos não pude estar nas reuniões da família devido estar aplicado aos trabalhos, mas saibam que a torcida e compreensão de vocês formaram as bases de sustentação para que eu pudesse finalizar o quanto antes minhas tarefas de pesquisador.

A minha namorada Sâmua Lene, pela paciência comigo, pelo entendimento de que nem toda hora eu poderia estar presente. Obrigado meu amor, você é uma linda na minha vida!

Ao meu professor orientador Francisco Milton Mendes Neto, meu grande mentor no mundo acadêmico que desde a graduação vem sempre me apoiando, me dando dicas, e mostrando a toda hora o caminho das pedras para que eu pudesse atingir meus objetivos com mais facilidade.

Gostaria de também agradecer ao meu professor coorientador Araken de Medeiros Santos, por ter aceitado ser meu coorientador neste projeto, que mesmo morando em Angicos, à 106 km de Mossoró, se dispôs a deixar de lado o conforto de sua residência e toda semana viajava para Mossoró, a fim de ser o professor da disciplina de Aprendizado de Máquina na qual ampliou e muito meus conhecimentos sobre a área da Inteligência Artificial.

A Laysa Mabel e Efraim Rodrigues pela ajuda e auxílio que me foi dado, pelos questionamentos levantados, podem ter a certeza que a concretização deste projeto se deve também ao esforço e ajuda que vocês me deram. Muito obrigado!

Ao prof. Ricardo Valentim e a SEDIS por terem confiado em mim, e permitido que eu pudesse trabalhar sobre a base de dados educacionais. Gostaria também de fazer menção aos analistas do NEAD da UFERSA, Ulisses Guimarães e Adriana Mara, pela atenção que me foi dada e por serem prestativos em todas as vezes que fui conversar sobre a organização da base de dados do Moodle. Saibam que vocês têm a minha gratidão!

A todos os professores e os técnicos administrativos que compõe programa de Mestrado de associação ampla da UERN e UFERSA, pois esse projeto maravilhoso vem abrindo portas nas vidas de várias pessoas, que Deus abençoe todos vocês.

Gostaria de agradecer em especial ao prof. Milton, prof. Araken, prof. Rommel, prof. Lima Jr., prof. Heitor e profa Cicília, pelos ensinamentos que me foram dados nas disciplinas do mestrado.

Não poderia deixar de mencionar também os técnicos administrativos Rosita e Carlos Magno, agradeço a vocês por estarem sempre de prontidão para tirar minhas dúvidas e me ajudarem nas documentações, muito obrigado.

A todos os professores que aceitaram fazer parte tanto da banca de qualificação, como também da banca de dissertação do mestrado, que foi o prof. Milton Mendes e prof. Araken como meu orientador e coorientador respectivamente, o prof. Rommel, o prof. Danniel Lopes, o prof. Patrício Alencar, e o prof. Rogério Patrício como membros internos e/ou externos.

Aos amigos que fiz e também aqueles que eu já conhecia e que me acompanharam durante essa jornada, que foi o Alisson Maurício, Bruno Elvis, Mariza Sousa, Alex Trindade, Rennê Santos, Wilamis Kleiton, entre tantos outros. Obrigado pela amizade, pelas risadas, por estarem comigo nos momentos bons e ruins, os momentos ruins, diga-se de passagem, foram os dias que precederam a prova de PAA.

Em especial gostaria de destacar meus dois amigos, Rennê Santos (Rennêzilas) e Wilamis Kleiton (Kleitonzilas), os grandes barões dos engenhos de cana-de-açúcar e capitãos-mor da casa da moeda do Delta do Parnaíba, e que são os maiores publicadores de artigos do Piauí.

Por fim, gostaria de agradecer a todas as pessoas que me ajudaram de forma direta ou indireta, e dizer que agradeço sempre a Deus por ter colocado pessoas maravilhosas a minha volta, e finalizar dizendo que:

- "*... até aqui nos ajudou o Senhor." 1 Samuel 7:12.*

EPÍGRAFE

“Os cientistas não dependem das ideias de nenhum homem em particular, mas da sabedoria acumulada de milhares de homens”

Ernest Rutherford

RESUMO

A crescente evolução da tecnologia, em conjunto com seus recursos computacionais, tem propiciado novas expectativas em várias áreas de pesquisa, tais como na indústria, saúde e educação. A inserção da tecnologia nesses ambientes, combinada com sua larga utilização, tem gerado um aumento no volume dos dados armazenados. Diante disso, pesquisadores perceberam a possibilidade de analisar esse grande volume de informações a fim de extrair conhecimento, de modo que essas informações possam, por exemplo, auxiliar no suporte à tomada de decisão. Quando aplicadas no âmbito educacional, a coleta, a medição e a análise de dados educacionais, a fim de identificar fatores que possam impactar positivamente ou negativamente o processo de ensino, são denominadas *Learning Analytics*. Diante dessa perspectiva, o presente trabalho apresenta uma ferramenta de avaliação das ações comportamentais dos tutores de disciplinas da modalidade de ensino a distância, de modo que, por meio desta, seja possível avaliar os comportamentos de tutores e turmas, bem como identificar quais os comportamentos do tutor que podem ou não estarem relacionados com os comportamentos da turma. Com as informações resultantes deste trabalho, pode-se compreender melhor o impacto dos comportamentos dos tutores nas turmas da modalidade de ensino a distância, além de possibilitar intervenções pedagógicas pautadas em informações objetivas, a fim de atenuar os problemas enfrentados por essa modalidade de ensino.

Palavras-Chave: *Learning Analytics*, Avaliação de Tutores, Indicadores de desempenho comportamental na educação.

ABSTRACT

The growing evolution of technology in conjunction with its computing resources has provided new expectations in several areas of research such as industry, health and education. Such integration of technology in these environments combined with its widespread use has led to an increase in the volume of stored data. Therefore, researchers then realized the possibility of analyzing this large volume information in order to extract knowledge, so that this information can, for example, help to support on making decision. When applied in the education area, the collection, measurement and analysis of educational data to identify factors that may impact positively or negatively the teaching process are called Learning Analytics. Given this perspective, this work presents an assessment tool of behavioral actions of tutors teaching mode disciplines distance, so that through this be possible to evaluate the behavior of tutors and classes, as well as identify tutor behaviors that may or may not be related to the behavior of the class. With the information resulting from this work, one can better understand the impact of the behavior of tutors in the teaching mode groups at a distance, and enables pedagogical interventions guided in concise information in order to alleviate the problems faced by this type of education.

Keywords: Learning Analytics, Tutors Rating, Behavioral performance indicators in education.

LISTA DE TABELAS

Tabela 1 - Evolução da quantidade de trabalhos submetidos e aceitos da Conferência Internacional de LA.....29

Tabela 2 - Resultado da correlação do atributo Número de Tópicos Criados pelo Tutor com os atributos da turma.....86

Tabela 3 - Resultado da correlação do atributo Número de Tópicos Criados pelo Tutor com os atributos da turma.....87

Tabela 4 - Resultado da correlação do atributo Média de Postagens em Tópicos dos Fóruns do tutor com os atributos da turma.....87

Tabela 5 - Resultado da correlação do atributo Taxa de Visualizações em Fóruns do tutor com os atributos da turma.....88

Tabela 6 - Resultado da correlação do atributo Taxa de Visualizações em Tópicos dos Fóruns do tutor com os atributos da turma.....88

Tabela 7 - Resultado da correlação do atributo Número de Atividades Criadas pelo tutor com os atributos da turma.....88

Tabela 8 - Resultado da correlação do atributo Número de Atividades Avaliadas pelo tutor com os atributos da turma.....88

Tabela 9 - Resultado da correlação do atributo Média de Postagens em Chats pelo tutor com os atributos da turma.....89

Tabela 10 - Resultado da correlação do atributo Número de Cliques do Tutor com os atributos da turma.....89

Tabela 11 - Resultado da correlação do atributo Número de URLs Criadas pelo tutor com os atributos da turma.....89

Tabela 12 - Resultado da correlação do atributo Número de Arquivos Criados pelo tutor com os atributos da turma.....90

Tabela 13 - Quadro comparativo entre os trabalhos relacionados de acordo com os critérios de inovação, técnica de análise empregada e alvo da análise.....98

LISTA DE FIGURAS

Figura 1 - Quadro comparativo entre a LA e a AA 27

Figura 2 - Processo Básico da LA.....28

Figura 3 - Modelo de LA proposto por Chatti. *et. al.* (2012).....33

Figura 4 - Dimensões Críticas da LA.....37

Figura 5 - Camada dos níveis de interessados da LA.....38

Figura 6 - Diagrama em grafos de uma rede social.....41

Figura 7 - Exemplo de um modelo de Aprendizado Supervisionado.....46

Figura 8 - Hierarquia das técnicas de aprendizado de máquina.....47

Figura 9 - Exemplo de cluster do *K-Means*.....48

Figura 10 - Exemplo da organização dos clusters.....49

Figura 11 - Arquitetura da ferramenta de análise.....60

Figura 12 - Relação entre as tabelas que contêm informações da organização dos cursos.....62

Figura 13 - Arquitetura resumida das relações entre as tabelas que armazenam informações referentes ao relacionamento dos usuários e as turmas criadas.....63

Figura 14 - Relacionamento das tabelas que formam o módulo de *quizzes* do Moodle.....66

Figura 15 - Relacionamento das tabelas que formam o módulo fórum do Moodle.....68

Figura 16 - Relacionamento das tabelas que compõem o módulo *chat*.....70

Figura 17 - Cabeçalho exemplo de um *dataset* no formato de organização do WEKA.....75

Figura 18 - Organização dos valores dos atributos no arquivo ARFF.....75

Figura 19 - Arquitetura detalhada do sistema de avaliação de tutores e turmas.....76

Figura 20 - Listagem dos cursos a serem avaliados.....77

Figura 21 - Tela inicial da categoria de avaliação do tutor.....77

Figura 22 - Tela inicial da categoria de avaliação da turma.....78

Figura 23 - Tela de avaliação do atributo Taxa de Visualização em Tópicos do tutor.....79

Figura 24 - Opção para escolha de visualização das informações do atributo Taxa de Visualização de Tópicos.....80

Figura 25 - Gráfico comparativo em barras do atributo Taxa de Visualização em Tópicos do tutor usando o critério de valores relativos.....80

Figura 26 - Gráfico comparativo em barras do atributo Taxa de Visualização em Tópicos do tutor usando o critério de valores absolutos.....81

Figura 27 - Resultado da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *K-Means*.....82

Figura 28 - Detalhes da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *K-Means*.....83

Figura 29 - Resultado da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *Farthest First*.....84

Figura 30 - Detalhes da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *Farthest First*.....84

Figura 31 - Apresentação dos resultados da correlação do atributo Taxa de Visualização em Fóruns com os atributos da turma.....86

LISTA DE SIGLAS

AA - *Academics Analytics*

ABED - Associação Brasileira de Educação a Distância

AL - Alunos

AM - Aprendizado de Máquina

AMNS - Aprendizado de Máquina Não Supervisionado

AMS - Aprendizado de Máquina Supervisionado

API - *Application Programming Interface*

APT - *Applied Predictive Technologies*

CCV - Coeficientes de Correlação entre Variáveis

CNF - Criação de Nova Ferramenta

CPF - Cadastro de Pessoa Física

CS - *Course Signals*

DA - Dados Abertos

EaD - Ensino a Distância

JSF - *JavaServer Faces*

GNU - *General Public License*

IDE - *Integrated Development Environment*

IES - Instituição de Ensino Superior

LA - *Learning Analytics*

LAK - *Learning Analytics & Knowledge*

LMS - *Learning Management System*

MOODLE - *Modular Object-Oriented Dynamic Learning*

NEA - *New Enterprise Associates*

NEAD - Núcleo de Educação a Distância

NNM – Normalização Min-Max

PC - *Personal Computer*

RG - Registro Geral

RN – Rio Grande do Norte

SEDIS - Secretaria de Educação a Distância

SOLAR - *Society and Learning Analytics Research*

SQL - *Structured Query Language*

TD - Tutores a Distância

TI - Tecnologia da Informação

UFE - Uso de Ferramenta Existente

UFRN - Universidade Federal do Rio Grande do Norte

SUMÁRIO

1 INTRODUÇÃO 19
1.1 CONTEXTUALIZAÇÃO 19
1.2 PROBLEMÁTICA 22
1.3 OBJETIVOS 23
1.4 ORGANIZAÇÃO DA DISSERTAÇÃO 23

2 REFERENCIAL TEÓRICO 25
2.1 LEARNING ANALYTICS 25
2.1.1 Metodologias de Learning Analytics 32
2.1.1.1 Modelo "What? Who? Why? How? 33
2.1.1.1.1 *What?* (O quê?) 34
2.1.1.1.2 *Who?* (Quem?) 34
2.1.1.1.3 *Why?* (Por quê?) 34
2.1.1.1.4 *How?* (Como?) 36
2.1.1.2 Modelo de 6 Dimensões 37
2.1.1.2.1 *Stakeholders* (Interessados) 37
2.1.1.2.2 *Objective* (Objetivos) 39
2.1.1.2.3 *Data* (Dados) 39
2.1.1.2.4 *Instruments* (Instrumentos) 40
2.1.1.2.5 *External Constraints* (Restrições Externas) 40
2.1.1.2.6 *Internal Limitations* (Limitações Internas) 41
2.1.2.3 Estudo Comparativo 41
2.2 LEARNING MANAGEMENTS SYSTEM 42
2.2.1 MOODLE 44
2.3 APRENDIZADO DE MÁQUINA 45
2.3.1 K-MEANS 48
2.3.2 FARTHEST FIRST 49
2.4 COEFICIENTE DE CORRELAÇÃO 50
2.4.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON 50

3 METODOLOGIA 53

4 FERRAMENTA DE EXTRAÇÃO E ANÁLISE DOS DADOS 60
4.1 EXTRAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS 61
4.1.1 Extração dos Dados 61
4.1.2 Pré-processamento dos Dados 71
4.1.2.1 Eliminação Manual de Atributos 72
4.1.2.2 Integração de Dados 72
4.1.2.3 Amostragem de Dados 73
4.1.2.4 Balanceamento de Dados 73
4.1.2.5 Limpeza de Dados 73
4.1.2.6 Transformação de Dados 74
4.1.3 Construção do *Dataset* 74
4.2 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS 76
4.2.1 Análise e Resultados da Normalização Min-Max 78
4.2.2 Análise e Resultados da Clusterização 81
4.2.3 Análise e Resultados da Correlação entre Variáveis 85
4.2.3.1 Número de Questionários Criados 90
4.2.3.2 Número de Tópicos Criados 91
4.2.3.3 Média de Postagens em Tópicos dos Fóruns 91
4.2.3.4 Taxa de Visualização em Fóruns 92
4.2.3.5 Taxa de Visualização em Tópicos dos Fóruns 93
4.2.3.6 Número de Atividades Criadas 93
4.2.3.7 Número de Atividades Avaliadas 93
4.2.3.8 Média de Postagens em *Chats* 94
4.2.3.9 Número de Cliques do Tutor 94
4.2.3.10 Número de URL Criadas 94
4.2.3.11 Número de Arquivos Criados 95

5 TRABALHOS RELACIONADOS 96

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 100

7 REFERÊNCIAS BIBLIOGRÁFICAS 103

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Ao longo da história é possível percebemos que por muito tempo a informação produzida foi apenas armazenada em lugares físicos, comumente denominados bibliotecas, que tinham por objetivo armazenar, catalogar e dispor os registros para uma determinada comunidade de usuários. No entanto, essas bibliotecas tinham limites físicos no que tange a esse armazenamento, devido ao limitado número de estantes, instalações físicas, etc. e também na forma de acesso, pois o acesso ao conhecimento só poderia ser feito de modo presencial (Sayão e Marcondes, 2008).

Com o surgimento dos sistemas computacionais e posteriormente da Internet, tornou-se possível um armazenamento e compartilhamento em larga escala desse conhecimento, que dessa vez eram armazenados em mídias digitais, tais como: áudios, vídeos, imagens e textos; que podem ser acessadas remotamente utilizando o computador, como ferramenta mediadora.

Uma vez que esses sistemas computacionais provêem uma alta capacidade no processamento de dados e armazenamento seguro, confiável, barato e rápido da informação, isso resultou na inserção da tecnologia nos mais diferentes tipos de ambientes, tais como empresarial, educacional, industrial, médico, etc (Souza e Mendes Neto, 2014).

O barateamento desses recursos, combinado com sua capacidade de alto processamento, grande volume de armazenamento e confiabilidade, acarretaram em sua popularização, que consequentemente gerou um grande aumento na utilização dos serviços oferecidos pelos sistemas computacionais. Sendo assim, a prática do armazenamento e gerenciamento de informações por meio dos sistemas computacionais se tornou uma conduta regular em instituições e empresas que têm buscado novas tecnologias, de modo que valores em seus negócios sejam agregados.

Segundo McAFee e Brynjolfsson (2012), estima-se que 2,5 *exabytes* de dados são criados por dia, e que esse número tende a convergir para o dobro a cada 40 meses. Todo esse grande volume de dados gerado recebe a nomenclatura de *Big Data*, que é um termo empregado para referir-se a grandes volumes de dados e que vem aumentando cada dia mais.

Diante da intensificação do volume de dados gerados, pesquisadores começaram então a estudar sobre formas de como manusear esses dados, de modo a obter conhecimento a partir de análises sobre estes, de forma que esse conhecimento possa: auxiliar na tomada de

decisões, realizar predições futuras, etc (Mayer-Schönberger e Cukier, 2013; Chen, Chiang e Storey, 2012).

Assim sendo, surgiram então novas áreas na ciência da computação direcionadas ao estudo de meios para obtenção desse conhecimento a partir desse grande volume de dados, tal fato se deve à inviabilidade da análise manual desse grande volume de dados, enquanto que essa análise pode ser realizada, em alguns casos, em poucos segundos por um sistema computacional.

A utilização dessas técnicas tem se tornado frequente por instituições públicas e privadas, pois estas podem ser utilizadas nas mais diversas áreas, dentre as quais podemos citar: saúde (Shah e Lipscombe, 2015), educação (Campagni *et. al.*, 2015), análise de imagens (Thepade e Kalbhor, 2015), detecções de fraudes (Dilla e Raschke, 2015), análise de dados meteorológicos (Pessoa *et. al.*, 2012), etc.

Quando aplicada no âmbito educacional, a análise sobre dados educacionais, objetivando a extração de conhecimento de modo que se possa proporcionar às partes interessadas (alunos, educadores, administradores e financiadores) uma melhor informação e um profundo conhecimento sobre os fatores dentro do processo de aprendizagem que contribuem para o sucesso do aprendiz, recebe a nomenclatura de *Learning Analytics* (LA) (Siemens *et. al.*, 2011).

A utilização da LA no meio educacional para obtenção do conhecimento pode gerar diversas oportunidades, dentre as quais podemos citar:

I. Auxiliar na tomada de decisão dos tutores no que se refere ao modelo de avaliações sobre uma determinada turma;
II. Auxiliar na inovação ou até mesmo transformar o sistema de uma universidade, bem como seus modelos acadêmicos e abordagens pedagógicas (Siemens e Long, 2011);
III. Permitir a identificação de alunos com baixo rendimento acadêmico, de forma a recomendar intervenções sobre esses, de forma que possam alcançar a aprovação na disciplina. Essa análise pode ser feita através de verificações das mensagens postadas em fóruns de discussão, número de exercícios concluídos, notas de provas, entre outras, permitindo assim com que os educadores possam identificar os alunos que estão em risco de reprovação (Macfadyen e Dawson, 2010);
IV. Apoiar o aumento da produtividade e eficácia organizacional fornecendo atualizações de informações e permitindo uma resposta mais rápida aos desafios e mudanças (Siemens e Long, 2011).

Sendo assim, pode-se considerar que a matéria-prima para a aplicação da LA é o grande volume de dados educacionais, visto que o conhecimento só pode ser extraído uma vez que esses dados estejam disponíveis para análise.

Uma das principais formas de se obter esses dados é por meio dos *Learning Management System* (LMS). Os LMS são sistemas de *software* concebidos para apoiar o ensino por meio da administração de um ou mais cursos com um ou mais alunos. Por meio deles, torna-se possível a realização de uma gama de atividades que antes, na maior parte das vezes, somente era possível de serem realizadas de forma presencial, tais como: aplicação de atividades, divulgação de notas, entrega de trabalhos, contato com professores, entre outras (Berking e Gallagher, 2011).

Devido sua praticidade, os LMS vêm sendo utilizados cada vez mais por instituições de ensino. E esta utilização torna-se ainda mais acentuada nas modalidades do Ensino a Distância (EaD). Visto que o gerenciamento de cursos, matrículas, lançamento de notas, e a própria comunicação entre professores e alunos pode ser feita em qualquer lugar, desde que se tenha um aparato computacional conectado a Internet, como, por exemplo, um computador, *tablet*, *smartphone*, ou qualquer outro dispositivo congênere.

Desta forma, uma vez que os LMS tendem a registrar notas de alunos, horários de acessos, visualizações e participações em fóruns de discussão da turma, quantidade de exercícios resolvidos e demais ações tanto de alunos quanto de professores, estes contêm amplas informações desse processo de ensino, tornando seus bancos de dados ideais para a aplicação das técnicas da LA. Isto possibilita a descoberta de informações do processo de ensino e de aprendizagem que podem estar associadas, mas não limitadas, à motivação dos estudos, evasão de alunos, êxito em determinadas metodologias pedagógicas e aprovação ou reprovação dos estudantes.

Sendo assim, devido essas informações do processo de ensino nesses ambientes estarem armazenadas nessas bases de dados, torna-se possível a extração e análise computacional dessas informações a fim de identificar possíveis padrões do impacto de comportamentos dos mentores dos alunos nestas turmas.

1.2 PROBLEMÁTICA

Embora a tecnologia tenha permitido maiores possibilidades e também um maior leque de recursos a serem utilizados no suporte do EaD, este continua sendo bem diferente do modelo tradicional de ensino em salas de aula, pois, em um LMS, o número de iterações do estudante com seus tutores e colegas são bastante reduzidas se comparadas com o ensino tradicional. Vale ainda ressaltar que a conversa entre o estudante e seu mentor nem sempre se dá de forma instantânea.

Segundo o último censo realizado pela Associação Brasileira de Educação a Distância (ABED), o número de matrículas em 2014 (desde o ensino fundamental até a pós-graduação: *stricto sensu* - doutorado) somaram 3.868.706 matrículas, com 519.839 (13%) nos cursos regulamentados totalmente a distância, 476.484 (12%) nos cursos regulamentados semipresenciais ou disciplinas de cursos presenciais e 2.872.383 (75%) nos cursos livres (Associação Brasileira de Educação a Distância, 2014).

Por outro lado, o maior obstáculo enfrentado em 2014 foi a evasão dos estudantes, seguido pela resistência dos educadores à modalidade do EaD, combinada aos desafios organizacionais de uma instituição presencial que passa a oferecer essa modalidade de ensino. De modo que, em todos os tipos de cursos analisados, a evasão dos alunos se concentra na faixa de até 25%, sendo que uma das suas causas é a falta de participação nos cursos (Associação Brasileira de Educação a Distância, 2014).

Outro problema deste ambiente de ensino é que, embora exista a possibilidade de contato entre alunos de uma mesma turma, devido estes geralmente não se conhecerem pessoalmente, iterações entre os mesmos não acontecem frequentemente e, como conseqüência disso, há uma redução na aprendizagem informal no que se refere à troca de conhecimento entre colegas. Consequentemente estes fatores tendem a gerar falta de motivação nos estudantes, que muitas vezes resultam no abandono dos cursos a distância.

Sob a perspectiva do tutor a distância, nem sempre estes são capazes de perceber que essa modalidade de ensino inspira: i) mais cuidados na forma de ensino do que o modelo tradicional; ii) métodos de ensino precisamente selecionados e adaptados para as circunstâncias dos alunos; e iii) material de estudo diferenciado do modelo tradicional de ensino. Além de que nem sempre estes estão dispostos a aprender e/ou usar novas ferramentas e funcionalidade dos LMS.

Diante do exposto, há uma dificuldade de inferir se o tutor está desempenhando bem ou não o seu trabalho, e se a turma está motivada ou não a partir de seus comportamentos no LMS devido à complexidade de se analisar todas essas informações que comumente se encontram dispersas pela base de dados do sistema.

Sob essa perspectiva exposta, o presente trabalho apresenta uma ferramenta de análise de comportamentos de tutores e turmas objetivando a coleta, medição, monitoramento e análise sobre esses dados, a fim de que, por meio da extração do conhecimento obtido sobre o processo de ensino nesses ambientes, resulte em uma melhor compreensão dos fatores que estejam associados de forma positiva ou negativa sobre o ensino na modalidade a distância.

Além disso, a referida ferramenta é capaz de avaliar tutores e turmas de acordo com seus respectivos comportamentos, a fim de que, por meio dessas informações, seja possível traçar novos modelos pedagógicos de ensino, facilitar a tomada de decisões, e permitir o monitoramento e a avaliação do desempenho da turma e do tutor desta em tempo real.

1.3 OBJETIVOS

O objetivo deste trabalho consiste na construção de um sistema que seja capaz de selecionar e analisar os dados dos tutores e turmas da modalidade do ensino a distância que estejam dispersos no banco de dados de um LMS, de modo que, por meio dessa análise, se possa obter informações sobre quais comportamentos estão mais propensos a resultar em uma maior participação ou não dos estudantes nestas turmas, além de possibilitar a classificação de tutores e turmas a partir de seus comportamentos no LMS.

1.4 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada da seguinte forma: o Capítulo 2 descreve a fundamentação teórica dos temas que compõe este trabalho; no Capítulo 3 são descritas as hipóteses de pesquisa, bem como a metolodogia de *Learning Analytics* que foi empregada neste trabalho; o Capítulo 4 apresenta a ferramenta construída; o Capítulo 5 apresenta os

trabalhos relacionados com o tema desta dissertação; e, por fim, o Capítulo 6 expõe as considerações finais e as perspectivas de trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os temas relevantes a esta dissertação. Na subseção 2.1 são descritos conceitos no campo da *Learning Analytics*. A subseção 2.2 apresenta conceitos relacionados aos *Learning Management System*. A subseção 2.3 expõe a fundamentação teórica da área de Aprendizado de Máquina. E, por fim, a subseção 2.4 apresenta o referencial teórico dos coefientes de correlação entre variáveis.

.

2.1 LEARNING ANALYTICS

O uso das tecnologias como forma de solução no que tange a comunicação entre pessoas tem sido cada vez mais frequente. Hoje em dia, instituições bancárias, educacionais, militares, e tantas outras fazem uso de todo esse aparato computacional. Tal uso pode ser justificado por uma gama de fatores, dentre os quais podemos citar a facilidade de armazenamento, transmissão, processamento e busca de informações em um curto intervalo de tempo, que, além proporcionar uma maior velocidade na condução dos processos operacionais dessas instituições, ainda reduz os custos destes.

Do ponto de vista educacional, o barateamento da tecnologia, associada à grande difusão da Internet, promoveu uma facilidade na disseminação de conteúdos educacionais que antes estavam comumente restritos aos livros. Essa inserção tecnológica no processo de ensino promoveu o surgimento de pelo menos três novos conceitos no se refere ao aprendizado utilizando aparatos tecnológicos como ferramenta mediadora entre aprendiz e conteúdo educacional.

Esses três conceitos denominam-se: i) *Eletronic Learning* (*E-Learning*), que consiste basicamente em um modelo de aprendizagem suportado pelo computador como ferramenta mediadora (Rosenberg, 2001; Downes, 2005; Garrison, 2011); ii) *Mobile Learning*, quando as ferramentas mediadoras do conhecimento são os dispositivos móveis (Motiwalla, 2007; Jacob e Isaac, 2014); e iii) *T-Learning*, que se apóia na utilização da TV analógica ou digital como o meio utilizado para difusão em massa do conhecimento (Mendes Neto, 2013; Mendes Neto *et al*., 2015).

A ampla utilização dessas tecnologias no ensino provocou um aumento na demanda em termos de recursos computacionais, pois, com o barateamento desses recursos, as quantidades de usuários cresciam cada vez mais e o volume de dados relacionados ao uso de sistemas acadêmicos acompanhava esse crescimento.

Devido ao grande crescimento das bases de dados por conta do armazenamento de informações referentes a registros de acessos (*logs*), conversas entre alunos e tutores, registros de notas, atividades, questionários, criação de fóruns, grupos etc, pesquisadores perceberam então a oportunidade de extrair informações que pudessem estar associadas ao processo de ensino, com a finalidade de compreender melhor os fatores que podem estar associados ao sucesso ou fracasso de cursos, turmas, alunos e práticas pedagógicas.

Diante disso, surgiu então o campo de pesquisa denominado *Learning Analytics* (LA), que pode ser definido como sendo a medição, coleta, análise e comunicação de dados sobre os alunos e os seus contextos, para fins de compreensão e otimização da aprendizagem nos ambientes em que esse processo ocorre (Siemens *et. al.*, 2011).

Já Lias e Elias (2011) definem a LA de uma forma mais sucinta, como sendo um campo emergente em que ferramentas de análise são utilizadas a fim de melhorar a aprendizagem e a educação.

Outros autores, como Johnson *et. al.* (2011), por exemplo, definem a LA da seguinte forma:

> "*A Learning Analytics refere-se à interpretação de uma ampla gama de dados produzidos e coletados por interesse dos estudantes, a fim de avaliar o progresso acadêmico, prever o desempenho futuro e identificar possíveis problemas*".

Sendo assim, em concordância com as definições dos demais pesquisadores, podemos definir a *Learning Analytics* como sendo a coleta, análise e compreensão das informações relacionadas ao processo de ensino, seja presencial, semipresencial ou a distância, e ao ambiente em que este processo ocorre, a fim de proporcionar para as partes envolvidas, seja de forma direta (estudantes, professores, etc) ou indiretas (instituições de ensino, comunidade científica, etc), a percepção de fatores que possam influenciar positivamente ou negativamente o ensino, independentemente de práticas pedagógicas que possam estar sendo utilizadas.

Conforme já citado, a área de pesquisa da LA é ainda um campo investigativo emergente, tendo ocorrido sua primeira conferência internacional no ano de 2011, na cidade de Banff, Alberta, Canadá (Siemens, 2011), o termo *Learning Analytics*, pertencente ao idioma inglês, que traduzido para a língua portuguesa, corresponde ao termo de "Análise da Aprendizagem".

Ainda no mesmo ano da primeira conferência de LA foi criada a SOLAR (*Society and Learning Analytics Research*), que é uma rede interdisciplinar de pesquisadores internacionais que vem pesquisando sobre o papel e o impacto que essas análises podem gerar sobre o processo de ensino e aprendizagem (Siemens *et. al.*, 2011).

A SOLAR faz parte da organização da Conferência Internacional sobre *Learning Analytics & Knowledge* (LAK), atuando também no suporte ao lançamento de várias iniciativas no apoio à pesquisa colaborativa sobre a área (Siemens *et. al.*, 2011).

Outra área de pesquisa que também está relacionada com a LAK é o campo da *Academics Analytics*. A *Academics Analytics* (AA) reflete na análise de dados a nível institucional de forma que as informações resultantes possam apoiar o processo de tomada de decisões estratégicas, de modo que universidades possam identificar quais seus pontos fracos e fortes, e áreas onde podem ser aplicadas melhorias (Academics Analytics, 2015).

Portanto, uma diferença importante entre a LA e a AA é que, enquanto a LA foca em aperfeiçoar os processos de ensino e de aprendizagem, a AA se concentra na descoberta de informações no campo estratégico dos setores administrativos, dentre os quais podemos citar: número de cursos oferecidos pela instituição, número de estudantes matriculados por curso, número de laboratórios disponíveis, níveis dos produtores acadêmicos por curso, etc.

Tipo de Análise	Nível ou Objeto da Análise	Beneficiados
Learning Analytics	Nível de Curso: Análise de conversas em *chat*	Alunos e Professores
	Nível Departamental: padrões de sucesso ou falha	Alunos e Professores
Academics Analytics	Nível Institucional: Fluxos de Conhecimento	Administradores, Financiadores e Profissionais de Marketing.
	Nível Regional: Comparação entre sistemas	Administradores e Financiadores
	Nível Nacional e Internacional	Governos Nacionais e Autoridades na Educação

Figura 1. Quadro comparativo entre a LA e a AA (Siemens e Long, 2011). Figura adaptada pelo autor.

Além da AA, existem ainda outras áreas que também podem ser consideradas correlatas com a LA, são elas: *Action Analytics*, *Web Analytics*, *Bussiness Analytics*, *Predictive Analytics* e outras.

Segundo Chatti *et. al.* (2012), o processo da LA é basicamente dividido em três etapas distintas. A Figura 2 ilustra as etapas desse processo, que são:

I. Coleta de dados e pré-processamento: Essa etapa é responsável pela coleta dos dados que serão analisados. Este é um passo crítico, pois a escolha dos dados que servirão para a análise pode implicar diretamente nos resultados apresentados pelo algoritmo que irá extrair informações a partir desse grande volume de dados. Além do processo de coleta, essa etapa conta ainda com um pré-processamento dos dados, que consiste em uma conversão dos dados para um formato que possa ser utilizado pelas técnicas da LA;

II. Análises e Ações: Com base nos dados coletados e pré-processados, o objetivo desse estágio é explorar os dados de modo que se possa extrair informações a partir deles. Este estágio não se resume apenas nas análises, mas também em ações que podem ser tomadas a partir das informações obtidas. Dentre essas ações, podemos citar como exemplos: o monitoramento, a intervenção, a predição, entre outras;

III. Pós-processamento: Essa etapa foca na melhoria contínua do processo da LA. Essa melhoria pode ser feita por meio da compilação de novos dados através da adição de novos atributos, de modo que novos indicadores e análises possam ser obtidos, ou então melhorar os já existentes.

Figura 2. Processo Básico da LA (Chatti *et. al.*, 2012). Figura Adaptada pelo autor.

Vale ressaltar que cada etapa do ciclo de vida gera uma saída para a etapa subseqüente. A etapa I define quais dados serão analisados e os coloca em um formato para que análises possam ser feitas. A etapa II, de posse da saída da etapa anterior, analisa os dados objetivando o descobrimento de padrões e informações relacionadas ao processo de ensino, que, consequentemente, servirá no auxílio da tomada de decisões.

E, por fim, com o *feedback* das ações tomadas na etapa antecedente, a etapa III focará em melhorias que poderão ser planejadas de modo a contemplar um maior número de variáveis que podem estar relacionadas ao ensino. Sendo assim, esse *feedback* servirá como entrada para a seleção dos dados que serão novamente compilados na etapa I, para que assim se possa alcançar novas oportunidades nas novas análises que ocorrerão. Portanto, pode-se considerar que o processo global da LA é freqüentemente um ciclo de vida iterativo (Chatti *et. al.*, 2012).

Embora a LA seja considerada uma área recente de pesquisa, é possível percebermos que seus objetivos de pesquisa e resultados trazem benefícios com finalidades aplicadas para educação, e que, de fato, ainda existem muitas subáreas desse campo a serem exploradas. Tal fato pode ser observado a partir de uma análise sobre as tendências de trabalhos voltados para a área da LA evidenciadas nas produções científicas da Conferência Internacional de *Learning Analytics*, em todas as edições do evento até o presente momento, realizadas de 2011 a 2015.

A Tabela a seguir apresenta a quantidade de trabalhos submetidos e aceitos da Conferência Internacional de *Learning Analytics*.

	Submetidos	Aceitos	Taxa de aceitação (%)
LAK'11	38	17	45%
LAK'12	36	14	39%
LAK'13	58	16	28%
LAK'14	44	13	30%
LAK'15	74	20	27%

Tabela 1. Evolução da quantidade de trabalhos submetidos e aceitos da Conferência Internacional de LA (Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, LAK 2013: Third International Conference on Learning Analytics and Knowledge, LAK 2014: Fourth International Conference on Learning Analytics And Knowledge, LAK 2015: Fifth International Conference on Learning Analytics And Knowledge).

Vale ressaltar que dentre os trabalhos aceitos não foram considerados os *papers* de autores convidados a escrever para o evento, *short papers*, *workshops*, tutoriais e *posters*.

Na LAK'11, os artigos aceitos ficaram basicamente divididos em: i) artigos conceituais, cujos conteúdos estavam mais focados nos aspectos teóricos da LA; ii) artigos com estudos de caso, que apresentavam um estudo de caso sobre análises que foram realizadas no âmbito da LA; e iii) artigo sobre ferramentas de análise de dados.

Já na LAK'12, pode-se perceber uma evolução nos trabalhos publicados em termos análises do campo investigativo da LA. Os artigos publicados tinham como foco: i) *Social Learning Analytics*; ii) LA nas perspectivas institucionais; iii) Análises para a aprendizagem reflexiva; iv) Intervenções educacionais; v) além de outras áreas correlatas com a LA, como, por exemplo, a *Visual Analytics*.

Na LAK'13, o campo investigativo dos temas dos artigos publicados se dissolveu ainda mais, mostrando que existem inúmeros subcampos neste campo de pesquisa a serem estudados. Os artigos aceitos nessa conferência se enquadravam nas trilhas de: i) Reflexões da LA; ii) Visualização de dados para apoiar a conscientização e reflexão; iii) Comunicação e colaboração; iv) Análise de dados sociais; v) Análise de diálogos; vi) Análise de Comportamentos; vii) Análise de emoções; viii) Análise preditiva; ix) Análise seqüenciais de dados; x) MOOCs; xi)Avaliações; xii) Suporte a professores; xiii) Desafios da LA; xiv) Arquiteturas de Análises; e xv) *Design briefings*.

As conferências da LAK'11, LAK'12 e LAK'13 apresentaram uma significativa mudança dos temas abordados dos artigos. A LAK'11 teve trabalhos mais conceituais da área e alguns trabalhos aplicados no que se refere à análise de dados. Porém, na LAK'12 evidenciou-se uma evolução nos temas abordados pelos trabalhos, além de que foi possível perceber alguns sub-campos investigativos de pesquisa da LA. E essa evolução se acentuou mais ainda na LAK'13, que apresentou vários trabalhos enquadrados em 14 sub-campos de pesquisa sobre o tema.

Já na LAK'15 pode-se perceber de forma mais evidenciada a evolução e consolidação de suas subáreas investigativas de pesquisa, sendo que foi dada a continuidade na adição de mais algumas trilhas ao evento, a fim de facilitar o enquadramento da temática proposta pelos seus sub-campos de pesquisa. Nesta edição, os artigos completos aceitos foram divididos em 22 temas específicos da área. Além disso, podemos também mencionar, principalmente na LAK'15, o crescente número de trabalhos que propunham ferramentas de análise de dados educacionais na identificação de fatores comportamentais e do desempenho de estudantes.

De fato, é notável que os resultados obtidos por meio da *Learning Analytics* venham a contribuir para o surgimento de novas oportunidades para o ensino, e também para o planejamento estratégico de metodologias pedagógicas. Pois com essas informações pode-se

aplicar ações que estarão fundamentadas em dados de experiências anteriores da aprendizagem, portanto, além de poder auxiliar na compreensão do ensino, a LA também ajudar a melhorar esse processo (Ferguson e Shum, 2012).

Em 2012, a Desire2Learn Incorporated, que é uma empresa bastante conceituada no provimento de soluções *eLearning*, recebeu 80 milhões de dólares de financiamento da New Enterprise Associates (NEA) e OMERS Ventures (The Globle and Mail, 2015; GIGAOM, 2015).

No ano seguinte, a APT (*Applied Predictive Technologies*), que é uma empresa no campo da análise preditiva baseada em nuvem, recebeu um investimento financeiro de 100 milhões de dólares do grupo Goldman Sachs. Dentre as empresas que fazem parte de sua lista de clientes, pode-se citar: Walmart, Hilton, a Anheuser-Busch InBev, McDonald, e outras, que tem adquirido cada vez mais os produtos da APT para tomar decisões críticas em todas as suas principais áreas estratégicas (ATP, 2015).

No ano de 2014, a Google Capital investiu 40 milhões de dólares na empresa Renaissance Learning, que é uma empresa conhecida por trabalhar na avaliação de dados educacionais (Education Week, 2015; Re/code, 2015). A tendência é que esses investimentos se tornem cada vez mais comum, pois eles trazem um alto retorno sobre investimento (do inglês, *Return on Investment* ou ROI) a médio e longo prazo para empresas e instituições.

Vale ressaltar também que com o alto investimento na área da LA, novos projetos serão concebidos, e, como conseqüência disso, novas metodologias e abordagens na construção de sistemas para essa área, considerada até então recente, irão surgir. Estima-se também que não somente os LMS, mas também os demais aplicativos que apoiam o processo de ensino, irão dispor das técnicas da LA de forma integrada como sendo parte de seu próprio sistema. A LA ainda contém muitas questões que estão em aberto, tais como:

- Lidar com o aumento no volume de dados;
- O problema da heterogeneidade entre dados e a interoperabilidade entre sistemas;
- Processamento em tempo hábil dos dados;
- Definição das diretrizes das políticas de ética e de privacidade dos dados;
- Métricas quantitativas e qualitativas sobre os resultados das análises e predições.

Diante do exposto, podemos concluir que a LA tem um grande potencial para ajudar a entender quais fatores podem estar mais associados do que outros no sucesso do ensino, assim como recomendar intervenções pedagógicas sobre alunos que se encontram em potencial risco de evasão, de modo que se possa melhorar o processo de ensino e de aprendizagem tanto da modalidade do ensino presencial, como da modalidade a distância.

2.1.1 METODOLOGIAS DE LEARNING ANALYTICS

É notável que os sistemas de *software* estão cada vez mais incluídos dia-a-dia de empresas e pessoas. Em muitos casos, a utilização desses sistemas se torna imprescindível para a realização de determinada ação e, para que esses sistemas possam apresentar desempenho satisfatório aos seus utilizadores, atributos, tais como confiabilidade, disponibilidade, eficiência e usabilidade, são características essenciais para o sucesso de seu uso.

Entretanto, a construção desses sistemas pode se tornar uma tarefa onerosa e bastante complexa, visto que o *software* é abstrato e intangível (Sommerville, 2007). Segundo Sommerville (2007), no ano de 1968 foi realizada uma conferência para debater a chamada "crise de *software*". Esta crise foi gerada devido à nova introdução dos circuitos integrados no *hardware* que ampliou as possibilidades do processamento de aplicações pelos computadores, sendo assim, tarefas até então consideradas não realizáveis tornaram-se possíveis. No entanto, apenas a experiência informal no desenvolvimento desses sistemas não foi capaz de acompanhar o rápido avanço do *hardware*. Pois atrasos em projetos, aumento nos custos, baixa confiabilidade e outros problemas estavam se tornando cada vez mais comuns na construção desses sistemas.

Diante disso, pesquisadores perceberam que novas técnicas e metodologias no desenvolvimento desses sistemas eram essenciais para o controle dos mesmos. Sendo assim, ficou que definido que as metodologias, os processos e os documentos de um *software* são tão importantes quanto o código-fonte da aplicação em si. As metodologias são de grade importância para que se possa atingir o sucesso na construção de um sistema (Sommerville, 2007).

Assim como na engenharia de *software* existem diversas metodologias no desenvolvimento de *software*, alguns estudiosos criaram também algumas metodologias para a área de *Learning Analytics*, porém essas metodologias não estão relacionadas com as

técnicas de programação, mas sim com o auxílio na compreensão do problema que se deseja resolver.

As subseções a seguir apresentam os dois principais modelos existentes até o momento desta pesquisa, bem como um estudo de comparativo entre estes.

2.1.1.1 MODELO “WHAT? WHO? WHY? HOW?”

Esse modelo foi proposto por Chatti *et. al*. (2012), os autores apresentam um modelo de LA que visa identificar oportunidades e possibilidades em cada dimensão desse modelo. A Figura 3 ilustra essas quatro dimensões, que são:

“O quê?”, do inglês *What?* - Essa dimensão é responsável pela definição de quais dados serão utilizados para análise;

“Quem?”, do inglês *Who?* - Responsável pela identificação dos alvos ou os interessados da análise;

“Por quê?”, do inglês *Why?* - Avalia os objetivos dos tipos de cada análise;

“Como?”, do inglês *How?* - Responsável pela identificação e definição das técnicas que farão parte da análise dos dados.

Figura 3. Modelo de LA proposto por Chatti. *et. al.* (2012). Figura adaptada pelo autor.

As subseções a seguir apresentam os conceitos de cada uma das dimensões desse modelo.

2.1.1.1.1 *What*? (O quê?)

Esta é a dimensão responsável pela abordagem sobre os dados que servirão para a análise. Esses dados podem ser oriundos de dois tipos de categorias de sistema, que são: i) Os sistemas educacionais centralizados, que são representados pelos *Learning Management System* (LMS), ou em português "sistemas de gestão da aprendizagem", como o Moodle, por exemplo; ii) Ambientes de aprendizagem distribuídos, que são sistemas interativos e facilitados pela tecnologia ubíqua no apoio da aprendizagem de modo informal do estudante.

Como na maioria das vezes esses dados ficam espalhados em diferentes sistemas, na maioria das vezes esses estão em diferentes formatos, distribuídos ao longo do espaço, tempo e meios de comunicação, o grande desafio é integrar esses dados que são obtidos de fontes heterogêneas e transformá-los em tempo hábil para formatos que sejam aceitáveis pelas técnicas de análise de dados que serão utilizadas (Chatti *et. al*., 2012).

2.1.1.1.2 *Who*? (Quem?)

O objetivo dessa dimensão é identificar todas as partes interessadas que podem ser beneficiadas com o uso da LA. Dentre as quais podemos mencionar: alunos, professores, pesquisadores, tutores e outros. Para cada uma dessas partes o sistema pode dar uma contribuição específica, por exemplo, no módulo aluno o sistema pode recomendar aos alunos determinadas ações que podem possivelmente implicar em um aumento de sua nota, enquanto que, no módulo professor, o sistema pode dar orientações pedagógicas de modo a aumentar a eficácia de suas metodologias de ensino (Chatti *et. al*., 2012).

2.1.1.1.3 *Why*? (Por quê?)

Nessa dimensão definem-se os possíveis objetivos da análise sobre os dados. Dentre entre esses objetivos, podemos citar (Chatti *et. al*., 2012):

- Previsão e Intervenção: O objetivo da previsão é tentar prever o desempenho do estudante com base em suas ações. A partir dos dados de saída da previsão, pode-se então recomendar a intervenção de forma pró-ativa sobre esses alunos que provavelmente estarão precisando de uma assistência adicional;
- Monitoramento e Análise: O objetivo principal é de realizar o monitoramento das atividades dos alunos e gerar relatórios sobre estes, a fim de auxiliar na tomada de decisão dos professores;
- Tutoria: Este item está relacionado em ajudar os alunos com suas atribuições de aprendizagem sob um determinado domínio e limitado ao contexto de um curso. Um professor, por exemplo, apóia os alunos em sua orientação e introdução em novos módulos de aprendizagem, bem como instruções de áreas específicas dentro de um curso. Em processos de tutoria, o controle é com o tutor e o foco está sob determinados itens de determinado domínio do curso;
- Avaliação e *Feedback*: O objetivo desse item é de auxiliar na auto-avaliação da eficiência e eficácia do processo de ensino, de forma que o sistema possa retornar *feedbacks* inteligentes tanto para professores quanto para alunos;
- Adaptação: Este item está associado a instrução adaptativa de recursos de aprendizagem e atividades de acordo com as necessidades do aluno individual;
- Personalização e recomendação: Na personalização, a LA é altamente centrada no aluno, focando em como ajudar os alunos a decidir sobre a sua própria aprendizagem e continuamente moldar suas preferências pessoais de aprendizagem para alcançar seus objetivos de aprendizagem. Já os sistemas de recomendação podem desempenhar um papel crucial para promover a aprendizagem auto-dirigida. Neste caso, o objetivo é ajudar os alunos, seja por meio da recomendação de conhecimento explícito (recursos de aprendizagem) ou através de nós de conhecimento tácito (pessoas), com base em suas preferências ou atividades de outros alunos com preferências semelhantes.

2.1.1.1.4 *How*? (Como?)

Esta fase é responsável pela aplicação das técnicas objetivando o descobrimento de padrões existentes nos dados que serão analisados. Dentre elas, podemos mencionar (Chatti *et. al*., 2012):

- Estatísticas: As técnicas estatísticas podem ser utilizadas para fornecer relatórios com estatísticas básicas sobre iterações do aluno com o sistema ou, em casos mais complexos, podem ser utilizadas para indicarem correlações entre comportamentos de alunos, tutores, professores ou elementos pedagógicos de ensino, que possam influenciar, ou não, o processo de aprendizagem e o sucesso dos alunos no curso, por exemplo;
- Visualização de Informação: Devido às informações estatísticas em forma de relatórios ou tabelas de dados nem sempre serem fáceis de interpretar. Este item tem por objetivo a representação das informações obtidas pelos métodos de análise de dados de uma forma visual com fácil interpretação, a fim de poder facilitar o entendimento e a análise dos dados educacionais. Para isso, diferentes técnicas podem ser utilizadas como, por exemplo: i) gráficos de dispersão; ii) representações 3D; iii) mapas conceituais; iv) entre outros. E que podem ser usadas para representar as informações de forma clara e concisa;
- Mineração de Dados: Este item consiste na obtenção do conhecimento a partir das técnicas de mineração de dados e de aprendizado de máquinas. Em termos gerais, as técnicas mais conhecidas são: Regras de Associação, Árvores de Decisão, Máquinas de Vetor de Suporte, Redes Neurais etc.;
- Análise de Redes Sociais: A análise de redes sociais consiste no estudo quantitativo das relações entre indivíduos ou organizações. Para isso, uma rede social é modelada por um grafo G = (V, E), onde V é o conjunto de nós (também conhecido como vértices), que representa os atores (pessoas ou organizações, por exemplo), e E é um conjunto de arestas (também conhecidas como arcos, links, ou laços), que representam um certo tipo de ligação entre atores.

2.1.1.2 Modelo de 6 Dimensões

Greller e Drachsler (2012) propuseram um modelo de seis dimensões no domínio e aplicação da LA. Este modelo considera seis dimensões críticas, conforme ilustrado na Figura 4, que são:

1. Interessados (do inglês *Stakeholders*);
2. Objetivos (do inglês *Objective*);
3. Dados (do inglês *Data*);
4. Instrumentos (do inglês *Instruments*);
5. Restrições Externas (do inglês *External Constraints*);
6. Limitações Internas (do inglês *Internal Limitations*).

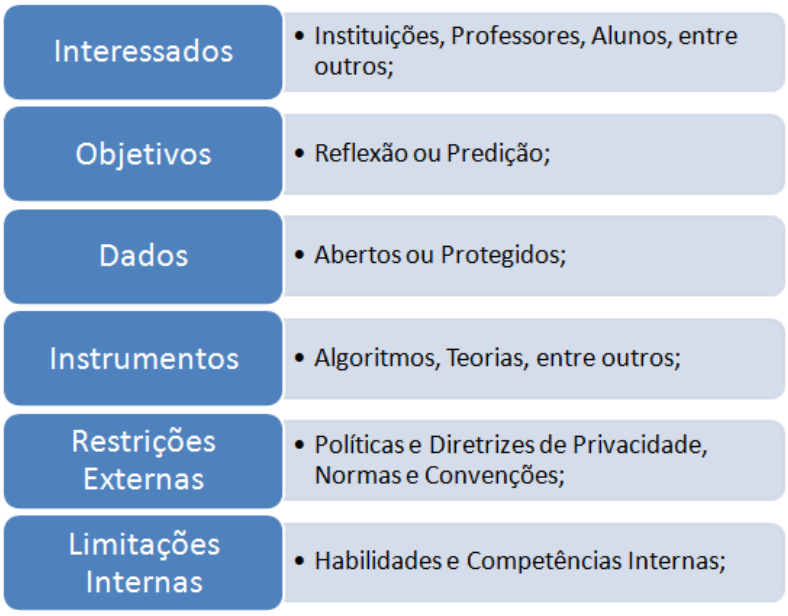

Figura 4. Dimensões Críticas da LA. Figura adaptada pelo autor.

Cada uma dessas dimensões é descrita a seguir.

2.1.1.2.1 *Stakeholders* (Interessados)

A dimensão dos interessados inclui todos os envolvidos de forma direta ou indireta com o uso da LA. Greller e Drachsler (2012) citam como sendo os principais grupos de interessados da LA os alunos, professores e instituições educacionais. Estes podem ser expandidos ou

substituídos por outros grupos de interesse, tais como pesquisadores ou ainda agências governamentais.

A Figura 5 apresenta um resumo da camada de interessados em sua forma mais direta, partindo do nível estudante até o nível das instituições governamentais. Vale ressaltar que cada um desses grupos pode ter diferentes necessidades no que se refere à informação extraída, conforme os exemplos apresentados a seguir:

I. Nível Professor - Os professores podem usar as informações para planejarem intervenções pedagógicas específicas em tempo real ou ainda ajustar as suas estratégias de ensino antes do início do curso;
II. Nível Instituição de Ensino – As instituições de Ensino podem, por exemplo, utilizar informações a partir de dados dos alunos e professores, a fim de promover novas oportunidades no planejamento de políticas e processos organizacionais objetivando melhorar a qualidade no ensino;
III. Nível Instituições Governamentais – As Instituições Governamentais podem coletar dados dessas instituições de ensino a fim de avaliar as necessidades de cada Instituição de Ensino Superior (IES);
IV. Todos os níveis – Todos os interessados, tais como: estudantes, professores e instituições de ensino podem se beneficiar por meio de uma auto-reflexão por meio de uma avaliação sobre seu desempenho individual.

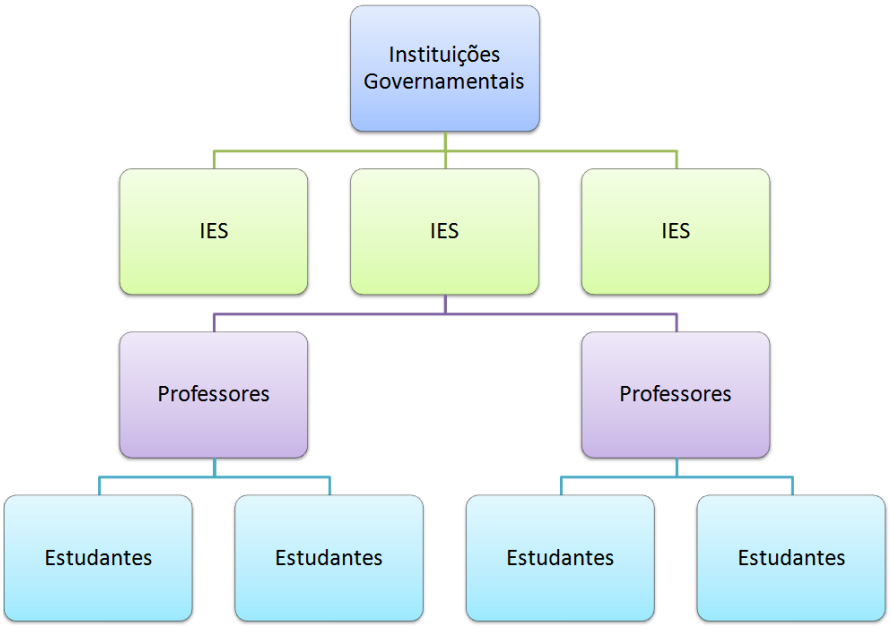

Figura 5. Camada dos níveis de interessados da LA (Greller e Draschler, 2012). Figura adaptada pelo autor.

2.1.1.2.2 *Objectives* (Objetivos)

Esta dimensão divide os objetivos em duas grandes áreas, que é a reflexão e a predição. A Reflexão é tida normalmente como sendo a auto-avaliação crítica de um usuário sobre seus próprios conjuntos de dados, a fim de obter um autoconhecimento. No entanto, esta área não está somente limitada como sendo auto-avaliação do usuário sobre seus próprios dados, pois esta avaliação pode também ser realizada com base em conjuntos de dados de outras partes interessadas. Por exemplo, um professor pode ser levado a refletir sobre o seu estilo de ensino, sendo indicado pelos conjuntos de dados dos seus alunos. No entanto, este também pode refletir sobre estilos de ensino com base nas informações obtidas de dados de outros professores.

Já a predição pode ser utilizada para prever e modelar as atividades do aluno, de modo a recomendar intervenções sobre estes com a finalidade de atenuar possíveis futuros problemas que possam acontecer com este durante o curso de uma disciplina.

2.1.1.2.3 *Data* (Dados)

Essa dimensão é responsável pela origem dos dados que serão analisados. Comumente, os dados educacionais utilizados para objeto de estudo são extraídos dos LMS. Um dos grandes desafios dessa dimensão está relacionado com a disponibilidade dos conjuntos de dados disponíveis publicamente para avaliar os métodos da LA, pois a maioria dos dados estão protegidos pelas instituições de ensino, e essa proteção é uma tarefa de alta prioridade para os departamentos de Tecnologia da Informação (TI).

No entanto, atualmente já se discute sobre a possibilidade do acesso aberto para esses conjuntos de dados educacionais. Uma forma de se fazer isso é por meio do anonimato, que é um meio de criar o acesso aos chamados Dados Abertos (DA).

O anonimato se dá através da exclusão dos identificadores da pessoa que geraram estas informações, ou que possa gerar um constrangimento a esta, tais como: número de matricula, nome completo, nome de pai e mãe, endereços residenciais, Cadastro de Pessoa

Física (CPF), Registro Geral (RG), conversas, e qualquer outra informação que a possa identificar ou causar-lhe constrangimento.

2.1.1.2.4 *Instruments* (Intrumentos)

Este dimensão está relacionada com as diferentes tecnologias que podem ser aplicadas no desenvolvimento de aplicações de cunho educacional a fim de suportar os objetivos de entidades ligadas à educação.

Neste caso, com o auxílio das tecnologias, a LA pode contribuir com sistemas de apoio à decisão, sendo este adaptado de acordo com os objetivos das partes interessadas, por exemplo. Ou ainda, contribuir para a diminuição das taxas de abandonos de aluno por meio do desenvolvimento de um sistema que possa notificar em tempo real ao professor de um curso que determinados estudantes estão em risco de abandonar o curso e/ou a disciplina.

Isso poderia ser feito através da utilização de conjuntos de dados de LMS e treinar uma determinada tecnologia de análise de informação (usando um classificador, por exemplo) nos conjuntos de dados para aprender padrões de comportamento de estudantes que abandonaram o curso anteriormente. O uso da LA com o auxílio das tecnologias pode levar a diferentes perspectivas no que se refere à tomada de decisão, predição de resultados e estratégias pedagógicas de ensino na educação.

2.1.1.2.5 *External Constraints* (Restrições Externas)

Esta dimensão está relacionada com os diferentes tipos de restrições que podem limitar a aplicação benéfica dos processos de LA, que são as restrições éticas, legais e sociais que podem estar envolvidas com a origem dos dados a serem analisados.

Essas restrições compõem um tema bastante sensível no que se refere à privacidade das informações. Por exemplo, um usuário poderia interpor processos judiciais caso suas informações, tais como: notas, conversas, etc venham ao conhecimento público e isso lhe cause danos morais. Portanto, uma solução para este caso seria adicionar uma subetapa na fase de pré-processamento dos dados, com o objetivo de remover identificadores pessoais dos usuários que irão compor o *dataset* a ser analisado.

2.1.1.2.6 *Internal Limitations* (Limitações Internas)

Este item está relacionado com as habilidades e competências internas para que as informações sejam corretamente compreendidas de modo que as mudanças venham estar fundamentadas sobre o conhecimento extraído.

Por exemplo, a Figura 6 apresenta um diagrama em grafos de rede uma social onde tal informação, embora tenha um visual atraente, apresenta risco na interpretação de seu significado, pois nem sempre os usuários têm as habilidades necessárias para interpretação dos resultados das técnicas de análise de dados.

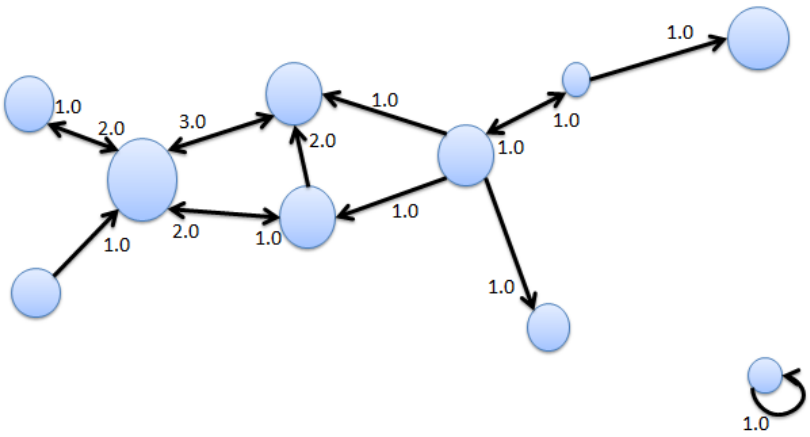

Figura 6. Diagrama em grafos de uma rede social (Greller e Draschler, 2012). Figura adaptada pelo autor.

É por este motivo que o usuário não pode somente se basear pela leitura dos *logs* resultantes da análise dos dados. Uma forma de solucionar este problema é por meio da tradução, para linguagem natural do usuário, dos resultados obtidos pelas técnicas de análise de dados, de modo que este possa compreender claramente as informações resultantes da análise, eliminando, assim, interpretações errôneas.

A fim de tornar a LA eficaz para a prática educacional, é importante considerar a apresentação e a interpretação dos resultados obtidos, pois a avaliação interpretativa das informações resultantes pode exigir, algumas vezes, competências de alto nível.

2.1.1.3 Estudo Comparativo

A fim de identificar qual das metodologias supracitadas mais se adéqua para a delimitação e solução do problema em questão, ambas as metodologias foram estudadas e, após realizar um estudo comparativo, pôde-se perceber que o modelo de seis dimensões é mais adequado pelos seguintes motivos que se seguem:

I. Primeiramente, o modelo de seis dimensões abrange, de forma geral, os itens do modelo de quatro dimensões;
II. O modelo de seis dimensões define melhor os objetivos de cada fase;
III. E, além dos itens anteriores, o modelo de seis dimensões tem um tópico exclusivo para discutir as questões éticas, morais e legais sobre os dados que serão analisados, de forma que as informações contidas sejam protegidas por políticas e diretrizes de privacidade. Esse fator foi um ponto crítico para este trabalho, pois a liberação da base de dados, utilizada nesta pesquisa, passou por análise de coordenadorias da instituição de ensino que cedeu essas informações, onde foi tratada toda a questão ética, moral e de privacidade, principalmente pelo fato da base de dados armazenar dados pessoais de todos os usuários do sistema.

Diante disso, a metodologia escolhida para a delimitação do objeto de estudo e proposta de solução foi a metodologia proposta por Greller e Drachsler (2012), isto é, o modelo de seis dimensões.

Todas as informações relacionadas à aplicação desta metodologia para este trabalho são apresentadas no Capítulo 3.

2.2 LEARNING MANAGEMENT SYSTEM

De acordo com Litto e Formiga (2012), o EaD surgiu no ano de 1728, quando o novo método de ensino de Caleb Philips, através de aulas por correspondência, foi anunciado. Ao longo do tempo, o avanço da tecnologia permitiu que outros meios fossem também utilizados no EaD, como o rádio e até mesmo a TV Educacional, que teve seu início nas décadas de 60 e 70 no Brasil (Litto e Formiga, 2012). No entanto, estes meios de comunicação não permitem muita interação entre os alunos e os professores, o que pode gerar uma desmotivação por parte do aluno.

O EaD, na década de 90, foi marcado pela integração do computador e das estações de multimídias no processo de ensino e de aprendizagem, o que acabou gerando novas

expectativas, pois possibilitou o acesso a informações sistematizadas e uma interação entre os alunos e professores (Faria e Salvadori, 2010). Para Litto e Fomiga (2012), houve uma maximização nas vantagens do EaD, pois a utilização dos novos meios de comunicação, técnicas e metodologias de ensino que obedecem padrões específicos geram uma maior qualidade.

Logo, a introdução de novos sistemas de comunicação, mediados pelo computador, possibilitou a multiplicação de tecnologias com a finalidade de apoiar o ensino a distância. A evolução da tecnologia, presente nos sistemas computacionais, promoveu a criação de *softwares* de cunho educacional que pudessem oferecer para seus usuários uma gama de recursos e funcionalidades.

Atualmente, esses *softwares* já permitem a criação de comunidades virtuais, dispõem de ferramentas de busca de conteúdos digitais, permitem acessos aos seus sistemas em qualquer local com conexão à Internet e, além disso, possibilitam a criação de videoconferência, envio de e-mail, gerenciamento de cursos e turmas, dentre outras facilidades. Esses ambientes virtuais de ensino são denominados de *Learning Management System* (LMS).

Esses sistemas têm sido bastante utilizados como objeto de estudo no que engloba seus aspectos conceituais ou teóricos, como também seus aspectos práticos, como, por exemplo, influência na aprendizagem, utilização pela comunidade acadêmica, e outros fatores, como pode ser visto em (Hunt, Davies e Pittard, 2007).

Segundo Kats (2013), os LMS são sistemas de *software* concebidos para apoiar o processo de ensino e ajudar os alunos a aprenderem por meio do gerenciamento da aprendizagem. Isso pode ser feito por meio da complementação de recursos que podem ser utilizados no ensino, tais como treinamentos, simulações, aulas ao vivo, entre outros.

Já de acordo com Berking e Gallagher (2011), os LMS são sistemas de *software*, baseados em servidor, usados para gerenciar e entregar (através de um navegador web) a aprendizagem de muitos tipos, particularmente de forma assíncrona. Geralmente, esses sistemas possuem a capacidade de rastrear e gerenciar vários tipos de dados dos alunos, especialmente os dados referentes ao seu desempenho.

Os LMS podem ser considerados também como uma tecnologia essencial ao acesso à aprendizagem de conteúdos e administração dos cursos em qualquer hora e qualquer lugar. Dentre as características em comum aos LMS, destacam-se:

- Permitem a administração de cursos de uma instituição;
- Permitem o gerenciamento de matrículas, notas e tarefas de alunos;

- Facilitam a criação, gerenciamento e publicação dos calendários acadêmicos dos cursos;
- Possibilitam a comunicação entre alunos, professores e tutores por meio de fóruns, e-mails, bate-papos etc;
- Proveem acesso ao sistema independe da hora e local;
- Contêm métodos de avaliação e testes, como, por exemplo, os *quizzes*;
- Propiciam integração e criação de repositórios de conteúdos educacionais;
- Armazenam registros da utilização do ambiente, que podem servir para fins de segurança ou até mesmo para a análise de comportamento dos usuários, a fim de permitir uma melhor compreensão de como se dá o processo de ensino e aprendizagem.

Como já mencionado, os LMS permitem armazenar, em seus bancos de dados, informações sobre o processo de ensino, como, por exemplo, informações sobre conversas, registros de acessos e visualizações de conteúdos das disciplinas, resoluções de questionários, envio de atividades, horário em que os usuários acessam o sistema, informações sobre metodologias de ensino, entre outros. Por isso, esses bancos de informações se tornam ideais para uma análise sobre fatores que podem influenciar, tanto positivamente quanto negativamente, o processo de ensino na modalidade a distância.

Os dados utilizados nesta pesquisa foram extraídos dos cursos oferecidos pela Secretaria de Educação a Distância (SEDIS) pertencente à Universidade Federal do Rio Grande do Norte (UFRN). Esta instituição utiliza o LMS MOODLE para a oferta de seus cursos.

A subseção a seguir apresenta os aspectos teóricos, bem como as aplicações práticas facilitadas por este LMS.

2.2.1 MOODLE

A palavra MOODLE é um acrônimo para *Modular Object-Oriented Dynamic Learning Environment*, que se refere a um sistema de gestão de aprendizagem gratuito, que permite que educadores criem seus próprios sites privados, preenchidos com cursos dinâmicos, a fim de estender a aprendizagem (Moodle, 2016).

O projeto MOODLE é liderado e coordenado pelo MOODLE HQ, que é uma empresa australiana de 30 desenvolvedores que é financeiramente apoiada por uma rede de mais de 60 empresas parceiras. Ao todo, estima-se que existe mais de 79 milhões de usuários em todo o mundo que usa este LMS para fins acadêmicos e empresariais, tornando esta a plataforma de aprendizagem mais utilizada do mundo (Moodle, 2016).

O MOODLE é fornecido gratuitamente como software *Open Source*, sob a licença GNU (*General Public License*), permitindo que possa modificar o MOODLE para projetos comerciais e não comerciais, sem quaisquer taxas de licenciamento (Moodle, 2016).

Este foi concebido segundo uma concepção guiada por uma pedagogia sócio-construtivista, dispondo de um leque de opções para seus usuários, como, por exemplo, ferramentas e atividades colaborativas, calendários, envio de notificações, verificação de progressos, capacidade multilíngue, alta interoperabilidade, além de outras características (Moodle, 2016).

2.4 APRENDIZADO DE MÁQUINA

A área de Aprendizado de Máquina (AM) é uma área da Ciência da Computação, comumente considerada uma subárea da Inteligência Artificial, cujo objetivo de pesquisa é o desenvolvimento de métodos e técnicas a fim de encontrar padrões, regularidades ou conceitos em conjuntos de dados (Goldschmidt e Bezerra, 2015).

Já Faceli *et. al.* (2011) cita esta como sendo a programação de computadores para aprender a partir de experiências passadas por meio do princípio de inferência, denominado também de indução, que obtém respostas genéricas resultantes da análise de um conjunto particular de exemplos. De forma sucinta, podemos dizer que o Aprendizado de Máquina é a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência (Mitchell, 1997).

Conforme Faceli *et. al.* (2011), existem vários problemas reais cujas técnicas de AM podem ser aplicadas, dentre os quais podem ser citados:

I. Predicação da taxa de cura de pacientes;
II. Reconhecimento de palavras;
III. Identificação de fraudes em cartões de crédito;
IV. Diagnóstico de câncer ou outros tipos de doenças; etc.

Ainda conforme Faceli *et. al.* (2011), as tarefas de aprendizado de máquina se dividem em duas categorias, que são a predição e a descrição. A predição consiste em encontrar uma fórmula ou gerar um modelo que seja capaz de prever ou rotular um dado a partir de um conjunto de dados para treinamento. Os algoritmos que pertencem a essa categoria são denominados de algoritmos de aprendizado supervisionado (preditivo).

O termo supervisionado advém da presença de um rótulo ou valor de saída existente nos dados que foram utilizados para o treinamento do conjunto. Dessa forma, as técnicas de aprendizado supervisionado usam o valor do rótulo de saída do conjunto de dados de treinamento para avaliar a capacidade da hipótese induzida de predizer os valores de saída para novos exemplos.

A Figura 7 apresenta um exemplo da estrutura de um algoritmo supervisionado.

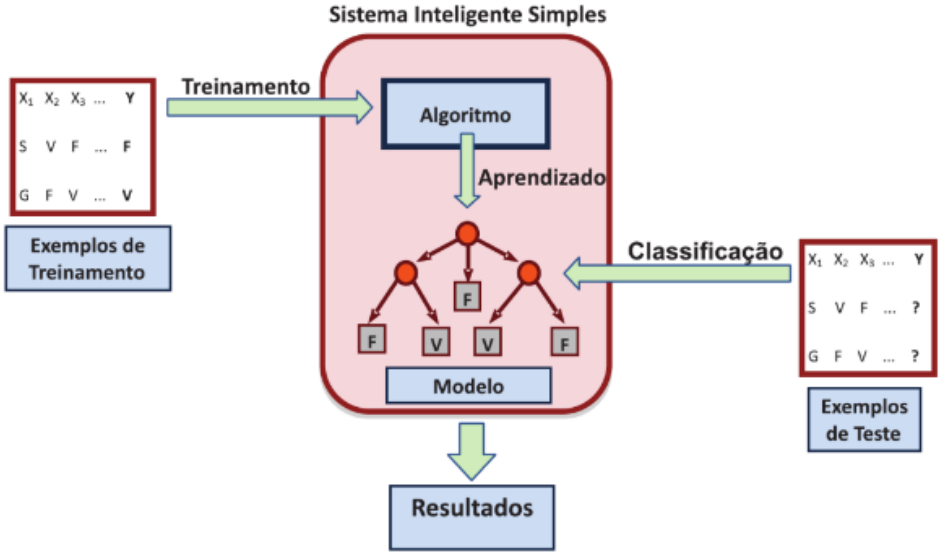

Figura 7. Exemplo de um modelo de Aprendizado Supervisionado (Quilici-Gonzalez e Zampirolli, 2014).

Pode-se observar na Figura 7 que os exemplos de treinamento são as informações de entrada para o algoritmo computacional de aprendizado de máquina e que, a partir destes exemplos, o algoritmo irá criar um modelo de classificação. Uma vez gerado o modelo de classificação, é possível utilizá-lo para classificar novas informações.

Já nas tarefas de descrição, o objetivo é explorar o conjunto de dados. Os algoritmos pertencentes a essa categoria são denominados de algoritmos de aprendizado não supervisionado (descritivo). O termo não supervisionado vem da inexistência de um rótulo de saída nos dados do conjunto de treinamento. Dessa forma, uma vez que não há rótulo de saída, esses algoritmos buscam explorar os dados a partir de sua regularidade.

De acordo com Faceli *et. al.* (2011), os algoritmos de aprendizado não supervisionado podem ser genericamente divididos em três tipos:

I. Agrupamento, em que os dados são agrupados de acordo com sua similaridade;
II. Sumarização, cujo objetivo é encontrar uma descrição simples e compacta para um conjunto de dados; e
III. Associação, que consiste em encontrar padrões frequentes de associações entre os atributos de um conjunto de dados.

A Figura 8 apresenta a hierarquia das técnicas de aprendizado de máquina de acordo com os tipos de seus modelos.

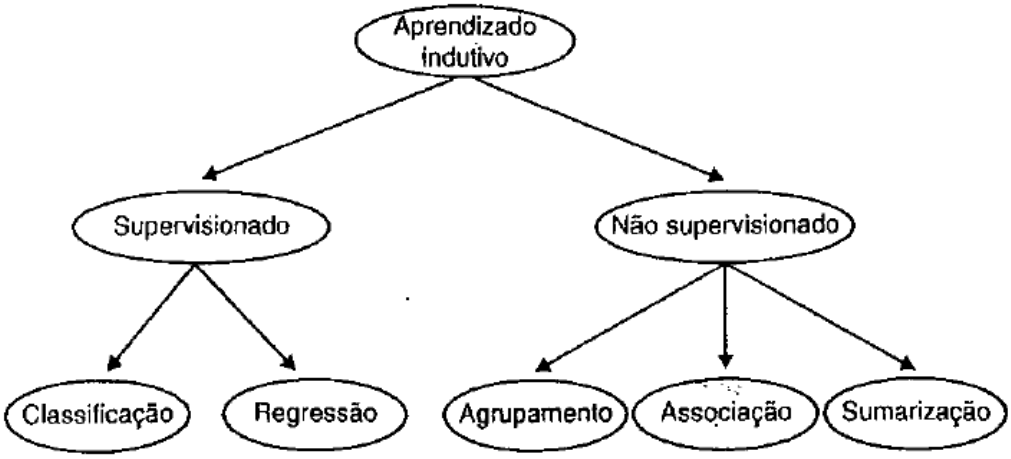

Figura 8. Hierarquia das técnicas de aprendizado de máquina (Faceli *et. al.*, 2011).

Tendo em vista que não há rótulo de saída nos dados comportamentais dos alunos e tutores, as técnicas que se adéquam ao enfoque deste trabalho, que é a classificação de comportamentos e ações das turmas e dos tutores do ensino a distância, são as técnicas descritivas (aprendizado não supervisionado).

Dessa forma, para a escolha da técnica que seria empregada na ferramenta proposta, foram levados em consideração três critérios: (i) desempenho (verificação da taxa de erro); (ii) tempo de execução; e (iii) simplicidade para interpretação dos resultados.

Após um estudo comparativo, as técnicas que apresentaram os melhores resultados, segundo os critérios supracitados, foram as técnicas *K-Means* e *Farthest First*, pertencente à categoria de agrupamento do aprendizado de máquina não supervisionado.

As subseções a seguir apresentam a fundamentação teórica dessas duas técnicas.

2.4.1 K-MEANS

O algoritmo *K-Means*, proposto por J. MacQueen (MacQueen, 1967), foi concebido para particionar uma população *n*-dimensional em *k* conjuntos, com base em uma amostra. Primeiro, o algoritmo inicia com a escolha de *k* elementos que formarão os *clusters* iniciais. O número *k* refere-se ao número de *clusters* definidos. Por exemplo, para *k* = 2, o algoritmo agrupará os elementos em dois *clusters*. A escolha inicial da posição dos *clusters* pode obedecer aos seguintes critérios: (i) por meio da seleção das *k* primeiras instâncias; (ii) selecionando *k* instâncias de forma aleatória; e (iii) por meio da seleção de *k* instâncias que possuam alto grau de dissimilaridade.

Após terem sido escolhidos os *clusters* iniciais, são calculadas as distâncias dos demais elementos em relação aos *clusters*, de modo que os elementos que possuírem a menor distância serão agrupados ao *cluster*. Em seguida, é recalculado o centroide deste *cluster*, levando em conta os novos elementos que foram agrupados. Esse processo é repetido até que todas as instâncias pertençam a um *cluster* tenham sido analisadas (Wu, 2012).

O resultado final desse processo é o agrupamento dos dados em *clusters*. A Figura 9 apresenta um exemplo desse resultado de forma gráfica.

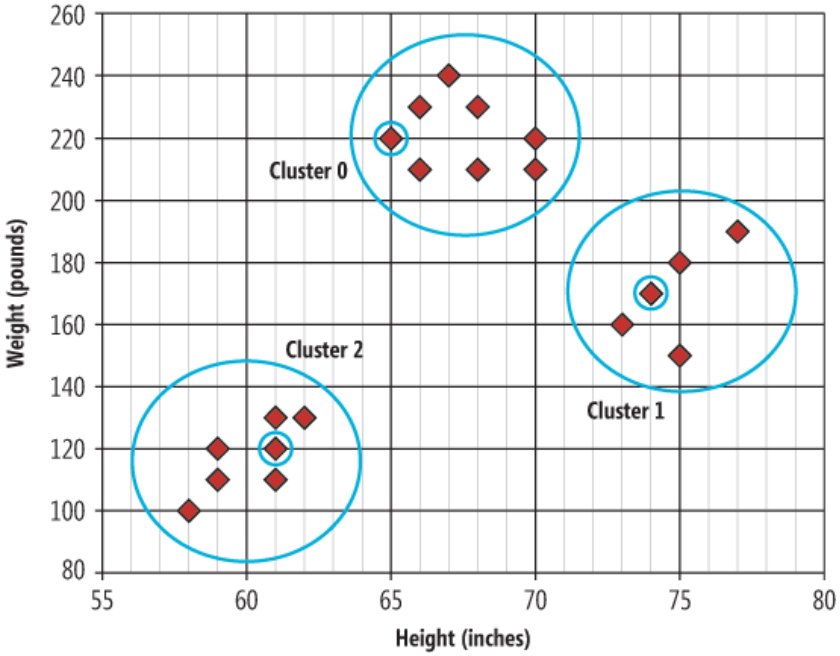

Figura 9. Exemplo de cluster do *K-Means* (McCaffrey, 2013).

Vale salientar que o item de dado circulado dentro de cada cluster representa o centroide do *cluster*.

2.4.2 FARTHEST FIRHST

Outra abordagem para seleção de centros de um *cluster* é escolher pontos de dados de distância máxima um do outro, maximizando assim raio do *cluster* (Dasgupta, 2002). Esta abordagem foi adotada pela primeira vez por Hochbaum e Shmoys (1985), e o nome do algoritmo foi denominado *Farthest First*, que em português equivale a "o primeiro mais distante".

Assim como o *K-Means*, o *Farthest First* opera em duas fases, que é a seleção do centroide e a atribuição de *cluster*.

A seleção de centroide começa por selecionar um ponto de dados aleatório como o centro do conjunto original e, em seguida, é escolhido o próximo centro como sendo o ponto mais distante (de acordo com a distância métrica) a partir do primeiro centro.

Vale salientar que os centros subsequentes são escolhidos de forma semelhante, como sendo os mais distantes do conjunto de centros previamente escolhidos.

A Figura 10 exemplifica a formação de clusters após a execução do algoritmo *Farthest First*.

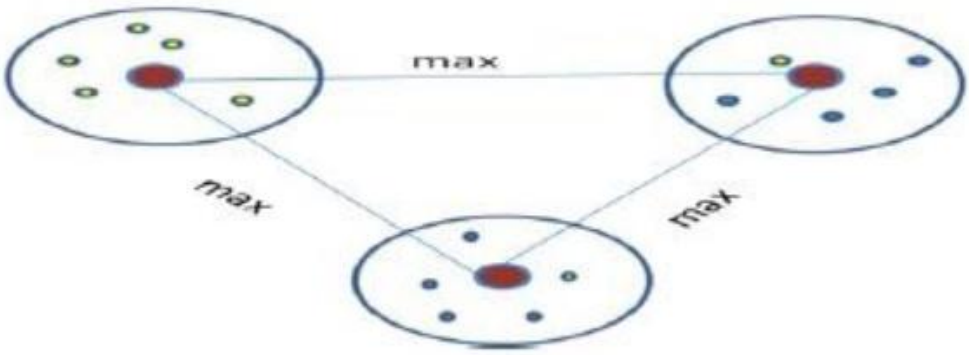

Figura 10. Exemplo da organização dos *clusters* (Vadeyar e Yogish, 2014).

Vale frisar que o *Farthest First* difere do *K-Means*, pois todos os seus centroides são pontos de dados reais e não os centros geométricos de clusters (Vadeyar e Yogish, 2014).

2.5 COEFICIENTES DE CORRELAÇÃO

Dado X e Y como sendo duas amostras, como podemos estabelecer uma relação entre X e Y? Essa é uma pergunta que está presente no cotidiano das pessoas. Muitas vezes nos perguntamos: Se a taxa de juros aumentar, será que as vendas irão cair? Ou então, se o preço do dólar subir, será que as importações vão cair?

O campo investigativo que objetiva responder essas questões são os métodos estatísticos de análise de correlação e a análise de regressão entre variáveis. A análise de correlação, também conhecida como coeficiente de correlação, indica o grau de variação conjunta entre duas variáveis. Esse grau representa a intensidade e a direção da relação linear ou não-linear entre estas. Esse método atende à necessidade de se estabelecer a existência ou não de uma relação entre variáveis, sem a necessidade de aplicar uma função matemática, pois não existe a distinção entre a variável explicativa e a variável resposta. Em outras palavras, o grau de variação conjunta entre X e Y é igual ao grau de variação entre Y e X (O'Rourke, Hatcher e Stepanski, 2005; Sharma, 2012; Schumaker, 2014).

Já a análise de regressão, além de medir a associação entre a variável explicativa e a variável resposta, também estima os parâmetros do comportamento sistemático entre estas (Sharma, 2005). No entanto, quando se deseja quantificar somente a força da relação entre as variáveis, nem sempre é necessário um detalhamento como o da análise de regressão, mas apenas determinar o grau de relacionamento entre as variáveis analisadas.

Conforme Meissner (2013), os coeficientes de correlação se dividem em três tipos: (i) o coeficiente de correlação de Pearson; (ii) o coeficiente de correlação de Spearman; e (iii) o coeficiente de correlação de Kendall.

Tendo em vista que a natureza dos dados analisados neste trabalho pertence à categoria das variáveis quantitativas, o coeficiente de correlação que mais se adéqua para essa análise é o coeficiente de correlação de Pearson (LeBlank, 2004; Sharma, 2005; Rubin, 2012).

2.5.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON

O coeficiente de correlação de Pearson tem como origem o trabalho em conjunto de Karl Pearson e seu professor Francis Galton. De forma sucinta, o coeficiente de correlação de

Pearson (r) é uma medida de associação linear entre variáveis (LeBlank, 2004; Sharma, 2012), e é calculado da seguinte forma:

$$r = \frac{1}{n-1} \times \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \times \left(\frac{y_i - \bar{y}}{s_y} \right) \quad (1)$$

Onde,

I. n é o número das amostras;

II. x_i e y_i são os elementos de cada amostra;

III. $\bar{x}$ e $\bar{y}$ representam a média dos elementos das amostras x e y; e

IV. S_x e S_y indicam o valor do desvio padrão das amostras x e y, respectivamente.

O resultado retornado deverá estar entre +1 e -1. O sinal indica a direção, ou seja, se a correlação é positiva ou negativa, e o valor numérico indica a força da correlação.

Segundo Weinberg e Abramowitz (2002) *apud* Cohen[1] (1988), quando os objetos de estudo são fatores comportamentais, a interpretação para a correlação de Pearson deve ser feita da seguinte forma:

I. Se $1.0 \leq r \leq 0.5$, sendo positivo ou negativo, indica uma forte correlação;

II. Se $0.3 \leq r < 0.5$, sendo positivo ou negativo, indica correlação moderada;

III. Se $0.1 \leq r < 0.3$, sendo positivo ou negativo, indica fraca correlação;

IV. Em último caso, para $0 \leq r < 0.1$, sendo positivo ou negativo, pode-se considerar nula.

Todavia, o uso da correlação de Pearson, por si só, não é suficiente para validar determinado resultado. Por exemplo, digamos que a correlação entre as amostras X e Y obteve um grau de r = 0,955. Como podemos afirmar que essa correlação não se deu por coincidência? Em outras palavras, a pergunta que se faz é: o quão significante é o valor de r?

A fim de solucionar essa questão, na estatística, utiliza-se o conceito de nível de significância de um resultado, fazendo-se uso do conceito de hipótese nula. A hipótese nula (H_0) simplesmente assume que um dado resultado estatístico foi obtido apenas por coincidência, devido a flutuações probabilísticas dos eventos medidos. Por outro lado, caso a hipótese nula seja rejeitada, o resultado não ocorreu por mera coincidência, e, portanto, deve-se aceitar a hipótese concorrente, que é chamada de hipótese alternativa (H_1).

[1] Jacob Cohen foi o primeiro pesquisador a introduzir a categorização do grau de relação de magnitudes no âmbito da análise comportamental, sendo hoje amplamente utilizada nas pesquisas das ciências comportamentais (Weinberg e Abramowitz , 2002).

O nível de significância é denotado por alfa (α) e indica a probabilidade de se cometer um erro do tipo I. O erro do tipo I consiste na possibilidade de se rejeitar a hipótese nula, quando esta é verdadeira. Logo, se α = 0,05, então a chance de se cometer um erro do tipo I é de 5%. Diante disso, o nível de confiança, que indica a probabilidade de decisão correta, baseada na hipótese nula, é de 95%, pois este é calculado como sendo 1 - α (Schlotzhauer, 2007; Rubin, 2012).

Sendo assim, quando se deseja utilizar o coeficiente de correlação de Pearson, em conformidade com o nível de significância, os seguintes passos devem ser realizados (LeBlanc, 2004):

I. No primeiro passo, deve-se definir as hipóteses:
   - H_0: $r = 0$;
   - H_1: $r \neq 0$.

II. No segundo passo, deve-se escolher um valor α para a significância. Por exemplo, α = 0,05;

III. No terceiro passo, como as amostras contêm *n* pares de dados, deve-se consultar na tabela de distribuição *t* de *Student* o valor de *t*(*gl*) para o valor de α escolhido, onde $gl = n - 2$. A tabela de distribuição *t* de *Student* fornece os valores críticos do intervalo de confiança a partir da probabilidade unicaudal ou bicaudal e do número de graus de liberdade (Keller, 2011);

IV. No quarto passo, deve-se calcular t_0:

$$t_0 = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \tag{2}$$

Caso $t_0 > t(gl)$ ou $t_0 < -t(gl)$, então H_0 deve ser rejeitada e H_1 deve ser aceita. Senão, H_0 deve ser aceita e H_1 rejeitada.

Se H_0 for rejeitada, deve-se concluir que o valor de *r*, obtido para a amostra é significante e que existe correlação *r* entre as variáveis X e Y, com nível de significância igual a α e nível de confiança de 1 – α.

Se H_0 for aceita, deve-se concluir que o valor obtido de *r* não é significante, ou seja, tanto pode haver correlação *r*, como não haver correlação.

3 METODOLOGIA

A partir de uma perspectiva de *Design Science* (Hevner *et. al*, 2004; Wieringa, 2014), a questão do conhecimento central abordada por esta pesquisa é:

- *A participação do tutor nos LMS pode afetar o desempenho em termos de participação efetiva dos alunos matriculados em cursos à distância?*

Esta pergunta central de pesquisa, em caso de resposta afirmativa, nos leva a seguinte questão:

- *Se a participação do tutor nos LMS afeta o desempenho em termos de participação efetiva dos alunos nos cursos à distância, como então podemos avaliar o compormento do tutor e dos alunos nestes ambientes?*

Estas questões de pesquisa, portanto, tem nos levado à configuração de um projeto de pesquisa baseada principalmente em dois métodos de pesquisa: revisão da literatura e análise de dados (do mundo real).

Para orientar a revisão da literatura, consideramos duas metodologias específicas de investigação no domínio da *Learning Analytics* para definir melhor o nosso problema de pesquisa. Por isso, consideramos o modelo de referência para delimitação do problema propostas por Chatti *et. al.* (2012), e por Greller e Drachsler (2012).

Nós escolhemos a última abordagem, devido à sua integralidade (se comparado com o anterior), e por exigir aspectos éticos, morais e legais sobre a análise de dados, com a recomendação de que a divulgação de dados educacionais serão regidos pelas políticas de privacidade.

Além disso, estes aspectos foram especialmente relevantes para esta pesquisa, uma vez que se baseiam em acordos de não divulgação de dados educacionais privados liquidados pelos nossos parceiros académicos.

A elaboração sobre as dimensões propostas pela metodologia escolhida para delimitação do escopo do problema é apresentada nas subseções a seguir.

3.1 INTERESSADOS

Os beneficiados com este trabalho se dividem em duas categorias, que são os beneficiados de forma direta e indireta. Os beneficiados de forma direta são:

i. Os alunos: Com a possibilidade de análise da participação efetiva da turma, torna-se possível propor alternativas no que se refere ao modelo de ensino, de modo que a participação destes possa melhorar, além da identificação de problemas na participação da turma de forma antecipada;
ii. Os tutores a distância: A análise dos dados e avaliação dos tutores resultará com que estes possam fazer uma auto avaliação sobre seus desempenhos, além da possibilidade de acompanhar o desempenho da turma em tempo real;
iii. Os coordenadores de tutores e de cursos de EaD: As facilidades promovidas pela ferramenta permitem com que os coordenadores possam avaliar o desempenho dos tutores e da turma de uma forma mais rápida e confiável, a partir do uso de técnicas consolidadas e amplamente indicadas e utilizadas pela comunidade científica.

Já os beneficiados de forma indireta são:

i. As instituições de ensino: Estas podem fazer uso da ferramenta para avaliar o desempenho de modelos pedagógicos, por exemplo, aplica-se um modelo pedagógico A em uma turma de Inteligência Artificial e, no outro semestre, aplica-se um modelo B sobre esta mesma turma, onde, ao final, faz-se uso da ferramenta para avaliar o desempenho no que se refere a participação efetiva da turma e dos tutores para os modelos pedagógicos A e B. Dessa forma, a instituição pode, por exemplo, catalogar informações sobre o desempenho de turmas a partir do uso de vários modelos de ensino e, dessa forma, melhorar os modelos de ensino e as práticas organizacionais, objetivando sempre a melhoria da aprendizagem.
ii. Pesquisadores: Estes podem se beneficiar de diversas maneiras, uma delas seria por meio da divulgação dos resultados obtidos com este trabalho, ou ainda, por exemplo, seria aplicar uma variação deste trabalho para alguma outra área, como na análise de dados médicos, análise da bolsa de valores, entre outras, ou seja, aplicar as técnicas e metodologias, que foram empregadas neste trabalho, como tentativa de solucionar algum outro problema.

3.2 OBJETIVO

O objetivo deste trabalho é a construção de uma ferramenta que seja capaz de extrair informações de alunos e tutores, e:

i. Seja capaz de analisar quais comportamentos dos tutores a distância tendem a refletir positivamente ou negativamente na participação efetiva dos alunos, de modo que seja possível traçar novos modelos, ou planos pedagógicos, pautadas em informações concisas, objetivando a atenuação das deficiências relacionadas à falta de participação dos alunos em suas turmas;
ii. Seja capaz de analisar os dados dos tutores e da turma, e permitir a inferência de desempenho dos comportamentos analisados em questão.

A ferramenta irá extrair e analisar as seguintes informações sobre os comportamentos dos tutores:

i. Número de questionários (*quizzes*) criados;
ii. Número de tópicos criados nos fóruns;
iii. Média de postagens em tópicos dos fóruns;
iv. Taxa de visualizações em fóruns;
v. Taxa de visualizações em tópicos dos fóruns;
vi. Número de atividades criadas;
vii. Número de atividades avaliadas;
viii. Média de postagens em *chats*;
ix. Número de cliques do tutor;
x. Número de URL criadas;
xi. Número de páginas criadas;
xii. Número de arquivos criados.

Já em relação às turmas, a ferramenta irá verificar os seguintes atributos das turmas:

i. Taxa de participação da turma nos questionários;
ii. Tempo médio para finalização dos questionários;
iii. Média de tópicos criados;
iv. Média de postagens em tópicos por alunos;
v. Média de visualização em fóruns por alunos da turma;

vi. Média de visualização em tópicos dos fóruns por alunos da turma;
vii. Taxa de submissão em tarefas da turma;
viii. Média de postagens em *chats* por aluno da turma;
ix. Média de cliques dos alunos da turma;
x. Média de páginas visualizadas pelos alunos da turma;
xi. Média de arquivos visualizados pelos alunos da turma.

3.3 DADOS

Neste trabalho, os dados foram cedidos pela Secretaria de Educação a Distância (SEDIS) pertencente à Universidade Federal do Rio Grande do Norte (UFRN). Essa instituição usa o Moodle como LMS. Portanto, todas as informações que compõem o *dataset* foram extraídas desse LMS.

Ao todo, foram extraídos e analisados informações de 62 turmas do EaD, pertencentes a 10 cursos de graduação, que são: (i) Bacharelado em Administração Pública; (ii) Licenciatura em Ciências Biológicas; (iii) Licenciatura em Educação Física; (iv) Licenciatura em Física; (v) Licenciatura em Geografia; (vi) Licenciatura em Letras; (vii) Licenciatura em Matemática; (viii) Licenciatura em Pedagogia; (ix) Licenciatura em Química; e (x) Licenciatura em História.

A base de dados histórica cedida corresponde aos anos de 2012 a 2013. No total, o *dataset* foi composto por informações de 2.227 alunos e 38 tutores a distância.

3.4 INSTRUMENTOS

Essa dimensão está associada às tecnologias e técnicas que foram empregadas para a análise dos dados. As subseções a seguir apresentam as informações de tudo o que foi utilizado para a construção da ferramenta.

3.4.1 Linguagens de Programação

A ferramenta foi construída utilizando como base a linguagem de programação JAVA (Oracle , 2016). Além da linguagem JAVA, para consulta das informações na base de dados, foi utilizada a linguagem de consulta estruturada SQL (do inglês *Structured Query Language*) (W3Schools, 2016), e, para a construção da interface para o usuário via servidor Web, foi utilizada a linguagem *JavaServer Faces* (JSF) (Oracle b, 2016), que utiliza o conceito de componentes para a criação de interfaces de sistemas Web. As *taglibs* oferecidas pelo JSF encapsulam o uso de *tags* HTML, permitindo ao desenvolvedor trabalhar em um nível mais alto de abstração.

3.4.2 Ferramentas e APIs Utilizadas

Como IDE (*Integrated Development Environment*), que em português significa ambiente de desenvolvimento integrado, que é uma aplicação de *software* que fornece um conjunto de instalações para que programadores possam desenvolver programas, foi utilizado o IDE NetBeans (NetBeans, 2016).

Já para o armazenamento da base de dados do Moodle que foi cedida pela SEDIS, foi utilizado o PostgreSQL (PostgreSQL, 2016). Em relação às APIs (*Application Programming Interface*), foram utilizadas: (i) a API WEKA, que fornece um conjunto de algoritmos de Aprendizado de Máquina para tarefas de mineração de dados, onde os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados a partir de um código Java (Weka, 2016); e (ii) a API Prime Faces, que é uma coleção de componentes baseados na API padrão do JSF de código aberto, com mais de 100 componentes, permitindo criar interfaces ricas para aplicações web de forma simplificada e eficiente (PrimeFaces, 2016).

3.4.3 Técnicas de Análise de Dados

A análise sobre as informações obtidas consistem em dois tipos, o primeiro tipo é a análise via correlação de variáveis que estabelece o quanto um dado comportamento X pode

influenciar positivamente ou negativamente em um comportamento Y. Para esta análise foi utilizado o coeficiente de correlação de Pearson e, para o nível de significância desta análise, foi utilizado o valor para $\alpha = 0{,}05$. Portanto, o nível de confiança dos resultados obtidos pela correlação de variáveis é de 95%.

Devido à correlação de Pearson ser sensível a *outliers*, para a remoção de *outliers* foi utilizada a técnica de Z-Score, tendo a margem aceitável de Z definida para o intervalo de $-2.1 < Z < 2.1$, sendo o valor Z calculado por meio da Equação 3 (Warner, 2012):

$$Z = \frac{(x_i - \bar{x})}{\sigma} \tag{3}$$

Onde:

I. x_i representa um valor da amostra;
II. $\bar{x}$ representa a média da amostra;
III. σ representa o desvio padrão da amostra;
IV. Z representa o quanto o valor da amostra se afasta da média amostral em termos de desvio padrão.

Sendo que pela tabela da distribuição normal, o intervalo de $-2.1 < Z < 2.1$ abrange 96.42% da área sob a curva da distribuição normal de valores. Dessa forma, se um dado tem 96.42% de chances de pertencer a margem de $-2.1 < Z < 2.1$, e ainda assim ficou de fora dessa margem, ao que tudo indica, esse dado é um *outlier* (Warner, 2012).

Já a segunda técnica de análise consiste no uso das técnicas de Aprendizado de Máquina, mais especificamente as técnicas de agrupamento (*clustering*) do Aprendizado de Máquina Não Supervisionado, onde foram utilizadas as técnicas *K-Means* (MacQueen, 1967) e *FarthestFirst* (Hochbaum e Shmoys, 1985). Tal uso se deu pelos seguintes motivos: i) bom desempenho no que se refere a tempo de processamento; ii) facilidade de interpretação dos resultados; e iii) facilidade na integração do algoritmo para a plataforma Web.

3.5 LIMITAÇÕES EXTERNAS

Esta perspectiva está relacionada com as questões éticas, morais e legais sobre os dados analisados, de forma que as informações sejam protegidas por políticas e diretrizes de privacidade.

A obtenção da cópia da base de dados do LMS utilizado no gerenciamento dos cursos a distância da SEDIS, que é um órgão pertencente à UFRN, se deu através de uma reunião, na qual ficou acordada a possibilidade de extrair somente os dados que estivessem relacionados com a utilização do sistema, por parte de seus usuários, e com restrição na recuperação de informações de registros pessoais, tais como CPF, RG, nome de pai e mãe, endereço residencial, etc.

Além disso, ficou decidido que os resultados da análise da ferramenta proposta neste trabalho poderiam ser publicados e que o *dataset* ficará totalmente à disposição da SEDIS para futuras análises.

3.6 LIMITAÇÕES INTERNAS

Este aspecto compreende as habilidades e competências internas para que as informações sejam corretamente compreendidas, de modo que as mudanças venham estar fundamentadas com base no conhecimento extraído.

A compreensão correta das informações analisadas é de responsabilidade da equipe de especialistas que construiu a ferramenta. E, após a compreensão das informações resultantes desse processamento de dados, as informações são compartilhadas com toda a equipe pedagógica, que ficará a cargo de uso da ferramenta de análise.

4 FERRAMENTA DE EXTRAÇÃO E ANÁLISE DOS DADOS

Neste capítulo será descrita a arquitetura da ferramenta de análise. Esta foi construída sob uma arquitetura de três núcleos (*cores*) de processamento de dados. Essa arquitetura é ilustrada na Figura 11.

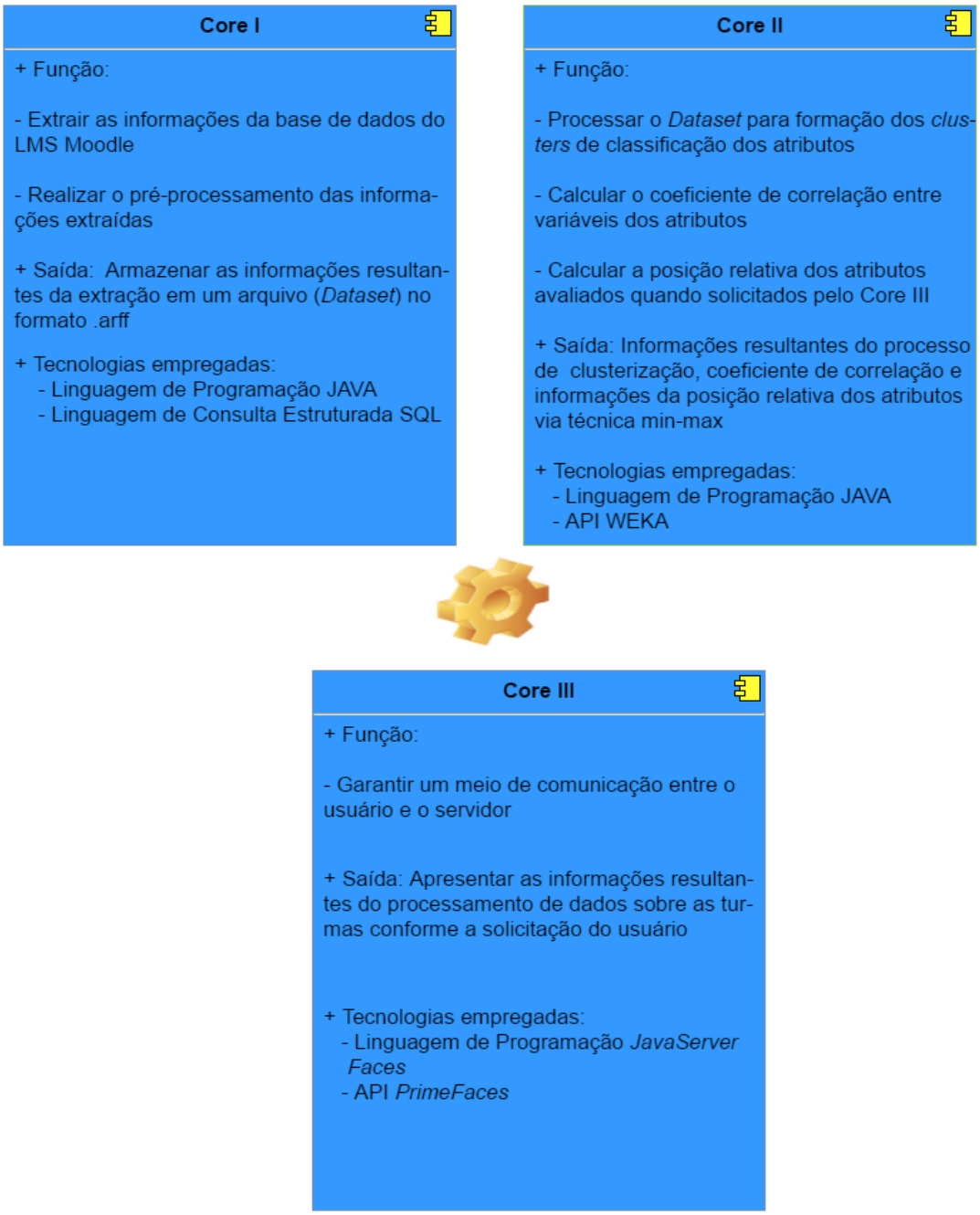

Figura 11. Arquitetura da ferramenta de análise. Fonte: Autoria Própria.

As subseções a seguir apresentam o detalhamento do processamento para cada núcleo de processamento de dados.

4.1 EXTRAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

Esta subseção apresenta todos os procedimentos realizados no núcleo de extração e pré-processamentos dos dados (*core* I). A fim facilitar o entendimento, optamos por dividir as informações referentes a este núcleo em três subseções que se seguem.

4.1.1 Extração dos Dados

Inicialmente, para que as informações dos tutores e alunos pudessem ser extraídas do LMS, foi necessário realizar um estudo sobre como se dá o armazenamento destes dados na base de dados.

Como referência, foi utilizada a documentação oficial do Moodle, além do apoio de profissionais do NEAD (Núcleo de Educação a Distância) da UFERSA e da SEDIS da UFRN, pois ambas instituições utilizam este LMS, e também por meio do estudo exploratório no que se refere ao armazenamento dos dados na base, pois as próprias tabelas da base de dados do Moodle contêm descrições sobre elas.

Vale salientar que, na instalação padrão do Moodle, este cria todas as suas tabelas com as iniciais de "*mdl_"*, porém isto é passível de mudança, e a sua mudança não reflete em nenhuma alteração estrutural da organização das tabelas.

Na base de dados que nos foi cedida para análise, todas as tabelas continham as iniciais de "*mdlacademico_*", dessa forma, se nos referirmos a tabela "*mdlacademico_course*", esta corresponde a tabela "*mdl_course*" por ser a nomenclatura inicial padrão do Moodle, da mesma forma para as demais tabelas.

O primeiro passo para a extração dos dados é identificar quais são os alvos da análise, neste caso, os alvos da análise são informações de alunos e tutores. Dessa forma, deve-se, portanto, identificar qual o ID referente a estes usuários.

Essa informação pode ser obtida na tabela "*mdlacademico_role*". Uma simples consulta como "select * *from mdlacademico_role*" retornará uma lista dos IDs de todos os usuários do LMS, tais como: alunos, professores, tutores, moderadores, visitantes, entre outros tipos de usuários.

Após a execução da SQL supracitada, foi identificado que, na base de dados que foi cedida para análise, os alunos são representados pelo ID de número 5, e os tutores a distância pelo ID de número 10.

Uma importante nota a ser frisada é que o Moodle aloca de forma padrão identificadores para os tipos dos usuários, porém estes identificadores são passíveis de mudanças. Neste caso, utilizar a documentação oficial sem averiguar esta tabela pode resultar em erros, pois nem sempre os IDs dos usuários serão os mesmos informados pela documentação oficial, dado que, conforme já mencionado, isto é passível de mudança.

De posse das informações dos IDs dos alunos e tutores a distância, a primeira extração a ser realizada é dos cursos existentes na base de dados. Para facilitar a identificação das consultas de extração de dados, todas as consultas que se seguem serão numeradas. A consulta 1 "*select a.id, a.fullname , a.format , b.id as idcat , b.name as cat from mdlacademico_course a join mdlacademico_course_categories b on a.category = b.id order by a.id;*" retorna as informações de ID de uma turma, nome da turma, formato da turma, ID da categoria do turma, e o nome da categoria da turma. Essas informações são armazenadas para que, em seguida, se possam extrair informações de cada turma existente na base de dados.

A Figura 12 apresenta a relação das tabelas que contêm informações da organização das turmas. Vale salientar que, para a extração de informações, não necessariamente é preciso utilizar todas as tabelas de um módulo.

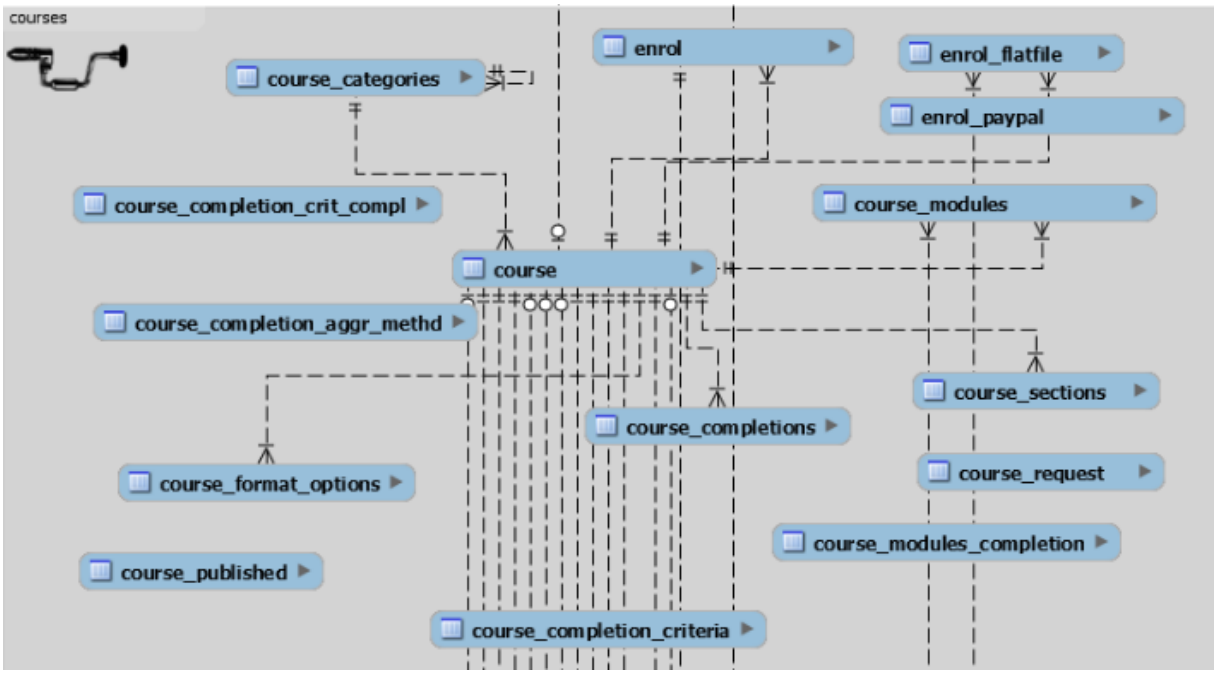

Figura 12. Relação entre as tabelas que contêm informações da organização dos cursos (Examulator, 2016).

De posse da lista das turmas armazenadas na base de dados, o próximo passo é a identificação dos tutores e dos alunos de cada turma. Primeiro foram extraídas as informações dos tutores e, em seguida, foram extraídas as informações dos alunos.

A Figura 13 esboça, de forma resumida, as relações entre as tabelas que armazenam informações referentes ao relacionamento dos usuários e as turmas criadas.

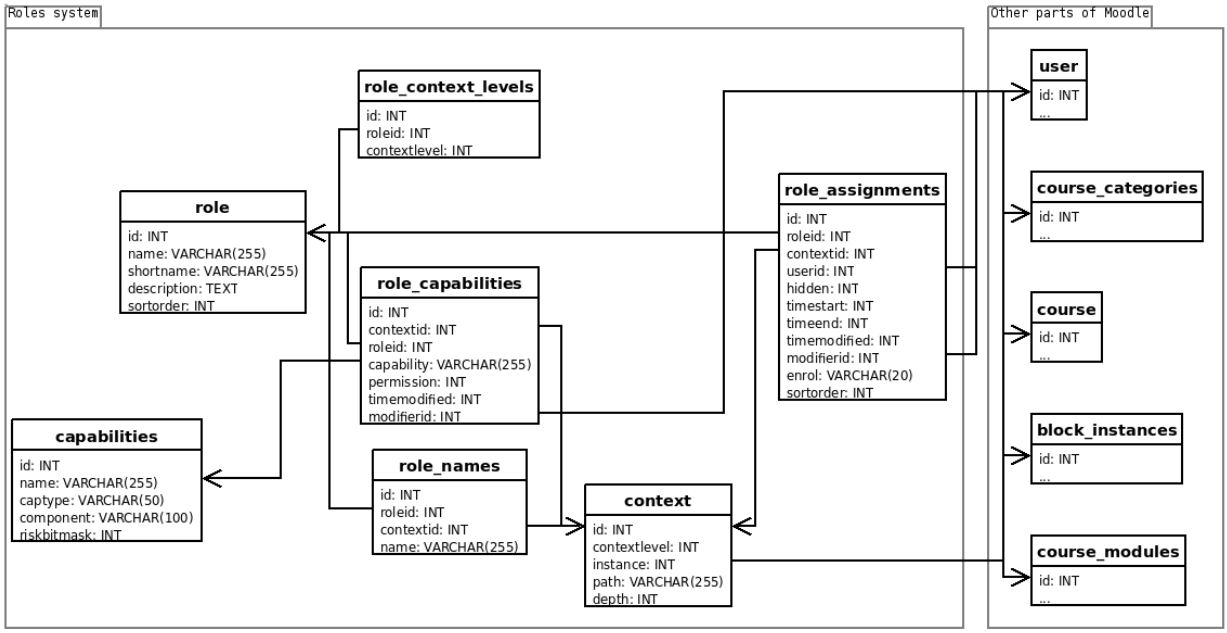

Figura 13. Arquitetura resumida das relações entre as tabelas que armazenam informações referentes ao relacionamento dos usuários e as turmas criadas (Moodle b, 2016).

A consulta 2 "*select c.id, c.fullname, u.id as t_id, u.firstname as t_name from mdlacademico_role_assignments rs inner join mdlacademico_context e on rs.contextid = e.id inner join mdlacademico_course c on c.id = e.instanceid inner join mdlacademico_user u ON u.id=rs.userid where e.contextlevel = 50 and rs.roleid = 10 order by c.id , u.id;*" obtém a informação do ID da turma, nome da turma, ID do tutor e nome do tutor.

De posse das informações dos tutores de cada turma, as consultas 3, 4, 5 e 6 obtêm as informações de:

i. Número de cliques: "*select count (ml.id) as total_de_cliques from mdlacademico_log ml inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+curso.get(i).listaTutor.get(0).id+"' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_cliques desc*";

ii. Número de páginas criadas: "*select count (ml.id) as total_de_paginas_criadas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+curso.get(i).listaTutor.get(0).id+"' and ml.module = 'page' and ml.action = 'add' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_paginas_criadas desc*";

iii. Número de URLs criadas: "*select count (ml.id) as total_de_urls_criadas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+curso.get(i).listaTutor.get(0).id+"' and ml.module = 'url' and ml.action = 'add' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_urls_criadas desc*";

iv. Número de arquivos criados: "*select count (ml.id) as total_de_arquivos_criados from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+curso.get(i).listaTutor.get(0).id+"' and ml.module = 'resource' and ml.action = 'add' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_arquivos_criados desc*".

Uma nota importante sobre as consultas 3, 4, 5 e 6, é que estas fazem uso da tabela *mdlacademico_log*, pois é nesta tabela que a maioria das ações dos usuários ficam armazenadas.

Depois de obtidas as informações sobre qual tutor estão responsáveis por uma determinada turma, a consulta 7 "*select c.id, c.fullname, u.id as u_id, u.firstname as u_name from mdlacademico_role_assignments rs inner join mdlacademico_context e on rs.contextid = e.id inner join mdlacademico_course c on c.id = e.instanceid inner join mdlacademico_user u ON u.id=rs.userid where e.contextlevel = 50 and rs.roleid = 5 and c.id = '"+curso.get(i).id+"' order by u.id;*" irá retornar a lista dos alunos que estão matriculados na turma.

E para cada aluno matriculado, as consultas 8, 9, 10 e 11 irão retornar as seguintes informações dos alunos:

i. Número de cliques: "*select count (ml.id) as total_de_cliques from mdlacademico_log ml inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 5 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+rs.getInt("u_id")+"' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_cliques desc*";

ii. Porcentagem de páginas visualizadas: "*select ml.userid as id_do_aluno, ml.course as id_do_curso, coalesce((select count (distinct mm2.cmid)::float from mdlacademico_log mm2 inner join mdlacademico_course_modules cmm on cmm.id = mm2.cmid where mm2.userid = ml.userid and mm2.course = ml.course and mm2.action = 'view' and mm2.module = 'page'),0) / coalesce(nullif((select count (mcd.id)::float from mdlacademico_course_modules mcd where mcd.course = ml.course and mcd.module = 15 and mcd.visible = 1), 0)) * 100 as porcentagem_de_paginas_visualizadas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 5 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+rs.getInt("u_id")+"' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid, ml.course order by porcentagem_de_paginas_visualizadas desc*";

iii. Porcentagem de URLs visualizadas: "*select ml.userid as id_do_aluno, ml.course as id_do_curso, coalesce((select count (distinct mm2.cmid)::float from mdlacademico_log mm2 inner join mdlacademico_course_modules cmm on cmm.id = mm2.cmid where mm2.userid = ml.userid and mm2.course = ml.course and mm2.action = 'view' and mm2.module = 'url'),0) / coalesce(nullif((select count (mcd.id)::float from mdlacademico_course_modules mcd where mcd.course = ml.course and mcd.module = 20 and mcd.visible = 1), 0)) * 100 as porcentagem_de_urls_visualizadas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 5 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+rs.getInt("u_id")+"' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid, ml.course order by porcentagem_de_urls_visualizadas desc*";

iv. Porcentagem de arquivos visualizados: "*select ml.userid as id_do_aluno, ml.course as id_do_curso, coalesce((select count (distinct mm2.cmid)::float from mdlacademico_log mm2 inner join mdlacademico_course_modules cmm on cmm.id = mm2.cmid where mm2.userid = ml.userid and mm2.course = ml.course and mm2.action = 'view' and mm2.module = 'resource'),0) / coalesce(nullif((select count (mcd.id)::float from mdlacademico_course_modules mcd where mcd.course = ml.course and mcd.module = 17 and mcd.visible = 1), 0)) * 100 as porcentagem_de_arquivos_visualizados from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 5 and ml.course = '"+curso.get(i).id+"' and ml.userid = '"+rs.getInt("u_id")+"' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid, ml.course order by porcentagem_de_arquivos_visualizados desc*".

As consultas 12, 13, 14 e 15 estão relacionadas aos *quizzes* (questionários) da turma. A Figura 14 apresenta o relacionamento das tabelas que formam o módulo de *quizzes* do Moodle.

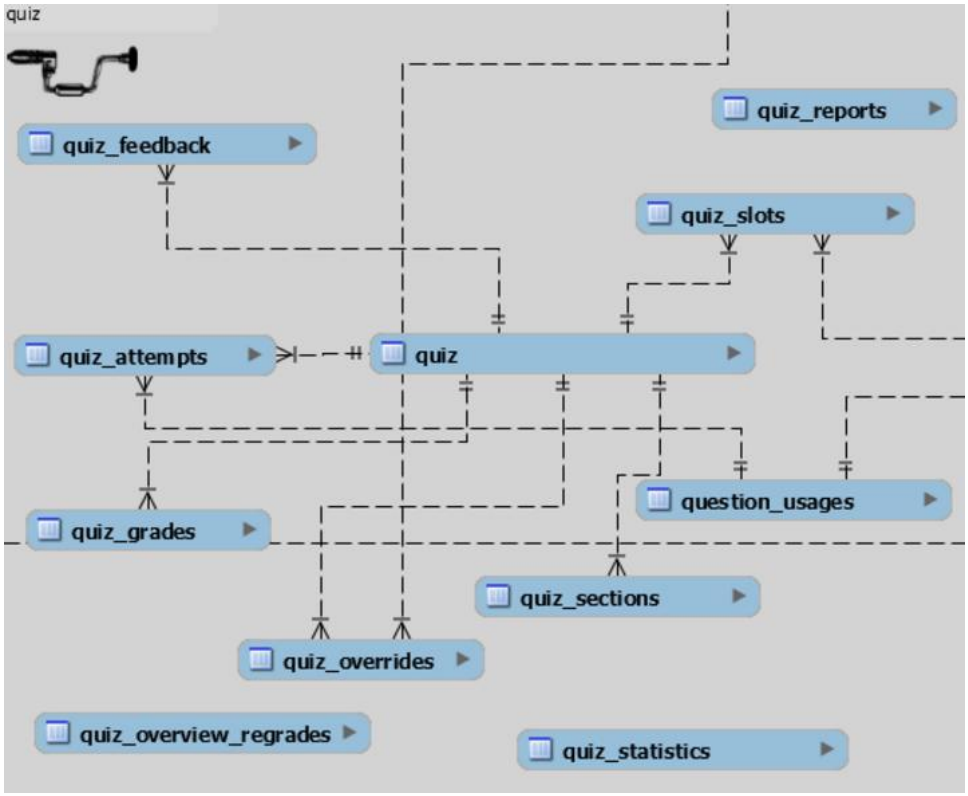

Figura 14. Relacionamento das tabelas que formam o módulo de *quizzes* do Moodle (examulator, 2016).

A consulta 12 “*select * from mdlacademico_quiz where course = '"+curso.get(i).id+"'*” retorna informações gerais sobre os quizzes que foram passados em uma turma. A partir desta consulta, é obtido o ID dos quizzes que foram passados e, em seguida, para cada quiz passado, são extraídas as informações sobre a quantidade de submissões dos alunos nos quizzes, conforme é apresentado na consulta 13 “*select distinct userid from mdlacademico_quiz_attempts where quiz = '"+rs.getInt("id")+"' and state = 'finished' and userid in (select userid from mdlacademico_role_assignments where roleid = 5) order by userid;*"

Já a consulta 14 irá retornar a quantidade de quizzes que foram criados pelo tutor da turma “*select ml.userid as id_do_tutor, count (ml.id) as total_de_questionarios from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.module = 'quiz' and ml.action = 'add' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_questionarios desc*”.

Por fim, a consulta 15 extrai as informações do tempo de finalização dos quizzes pelos alunos: “*SELECT distinct userid, quiz, state , TO_CHAR('epoch'::timestamp + (timestart ||' seconds')::interval, 'YYYY-MM-DD-HH24-MI-SS') as timestart, TO_CHAR('epoch'::timestamp + (timefinish ||' seconds')::interval, 'YYYY-MM-DD-HH24-MI-SS') as timefinish FROM mdlacademico_quiz_attempts where quiz = '"+curso.get(i).listaQuiz.get(j).id +"' and state = 'finished' and userid in (select userid from mdlacademico_role_assignments where roleid = 5) order by userid;*”.

Basicamente, esta consulta retorna o ID do aluno, o ID do *quizz*, o estado do *quizz* (finalizado), a data de criação do *quizz* e a data final da submissão do *quizz*, ambas no formato de ano, mês, dia, hora, minuto e segundo.

Dessa forma, a partir da data de criação e a data final da submissão, as informações extraídas são ainda processadas para que o tempo de finalização do *quizz* pelo aluno esteja em horas.

As consultas 16, 17, 18, 19 e 20 extraem informações sobre os fóruns, tópicos e postagens da turma. A Figura 15 apresenta o relacionamento das tabelas que formam o módulo fórum do Moodle.

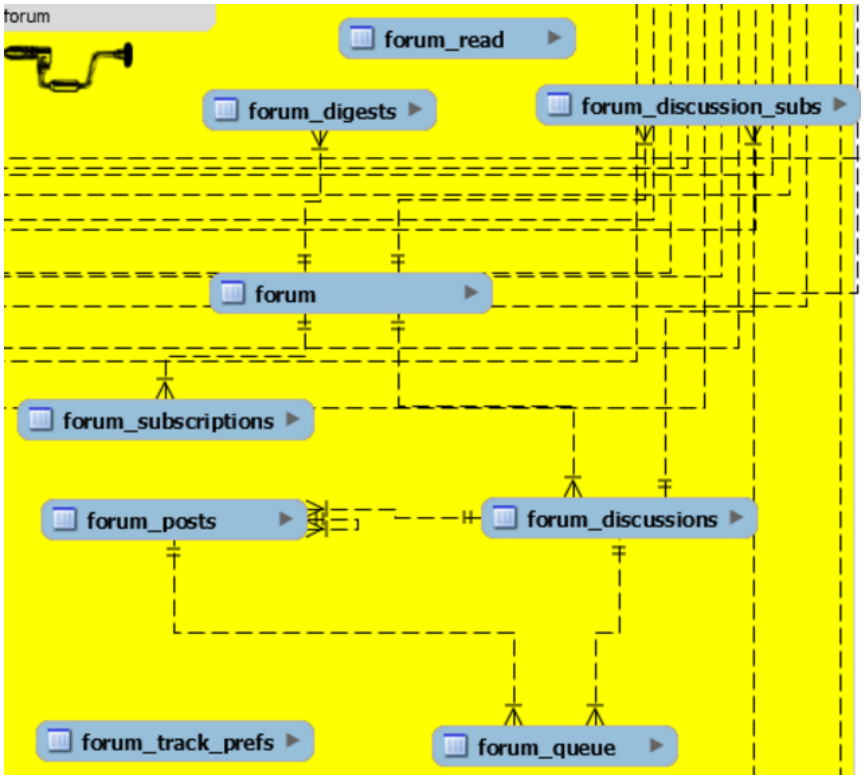

Figura 15. Relacionamento das tabelas que formam o módulo fórum do Moodle (examulator, 2016).

No Moodle, uma turma pode ter *n* fóruns e, em cada fórum, pode-se ter *n* tópicos, e para cada tópico pode-se ter *n* postagens.

Primeiramente são extraídas as informações dos fóruns de uma turma, conforme é apresentado na consulta 16 "*select * from mdlacademico_forum where course = '"+curso.get(i).id+"' order by id*".

Para cada fórum da turma, são extraídos os tópicos que foram criados no fórum, conforme é exposto na consulta 17 "*select * from mdlacademico_forum_discussions where course = '"+curso.get(i).id+"' and forum = '"+curso.get(i).listaForum.get(j).id+"' order by id;*" .

E, para cada tópico, são extraídas a quantidade total de postagens no tópico, a quantidade total de postagens do tutor no tópico, e a quantidade total de postagens dos alunos no tópico, conforme é exposto nas consultas 18, 19 e 20:

i. Consulta 18: "*select count(discussion) from mdlacademico_forum_posts where discussion = '"+rs.getInt("id")+"';*";
ii. Consulta 19: "*select count(discussion) from mdlacademico_forum_posts where discussion = '"+rs.getInt("id")+"' and userid in (select userid from mdlacademico_role_assignments where roleid=10);*";

iii. Consulta 20: “*select count(discussion) from mdlacademico_forum_posts where discussion = '"+rs.getInt("id")+"' and userid in (select userid from mdlacademico_role_assignments where roleid=5);*”.

As consultas 21, 22, 23 e 24 extraem informações relacionadas às visualizações de fóruns e tópicos pelos alunos e tutores. Estas são apresentadas a seguir:

i. Consulta 21 retorna *logs* da visualização de fóruns pelo tutor: “*select distinct l.info as info, l.module, l.course, l.action from mdlacademico_log l inner join mdlacademico_forum f on l.userid = '"+curso.get(i).listaTutor.get(0).id+"' and l.module = 'forum' and l.course = '"+curso.get(i).id+"' and l.action ='view forum' and NULLIF(l.info, '')::int = f.id order by l.info;*”;

ii. Consulta 22 retorna a quantidade de visualização nos fóruns da turma pelos alunos: “*select count(l.action) from mdlacademico_log l inner join mdlacademico_forum f on l.module = 'forum' and l.course = '"+curso.get(i).id+"' and l.action ='view forum' and NULLIF(l.info, '')::int = f.id and l.userid in (select userid from mdlacademico_role_assignments where roleid=5) ;*”

iii. Consulta 23 retorna os *logs* da visualização de tópicos tutor: “*select distinct l.info, l.module, l.course, l.action from mdlacademico_log l inner join mdlacademico_forum_discussions f on l.userid = '"+curso.get(i).listaTutor.get(0).id+"' and l.module = 'forum' and l.course = '"+curso.get(i).id+"' and l.action ='view discussion' and NULLIF(l.info, '')::int = f.id;*”;

iv. Consulta 24 retorna a quantidade de visualização nos tópicos da turma pelos alunos: “*select count(l.action) from mdlacademico_log l inner join mdlacademico_forum f on l.module = 'forum' and l.course = '"+curso.get(i).id+"' and l.action ='view discussion' and NULLIF(l.info, '')::int = f.id and l.userid in (select userid from mdlacademico_role_assignments where roleid=5) ;*”.

As consultas 26 e 27 retornam as quantidades de tópicos criados pelos tutores e alunos respectivamente. Estas são apresentadas a seguir:

i. Consulta 26: “*select count(forum) from mdlacademico_forum_discussions where course = '"+curso.get(i).id+"' and forum = '"+curso.get(i).listaForum.get(j).id+"' and userid in (select userid from mdlacademico_role_assignments where roleid=10);*”;

ii. Consulta 27: “*select count(forum) from mdlacademico_forum_discussions where course = '"+curso.get(i).id+"' and forum = '"+curso.get(i).listaForum.get(j).id+"' and userid in (select userid from mdlacademico_role_assignments where roleid=5);*”.

As consultas 28, 29, 30 e 31 são relacionadas aos *chats* da turma. A Figura 16 apresenta o relacionamento das tabelas que formam o módulo *chat*.

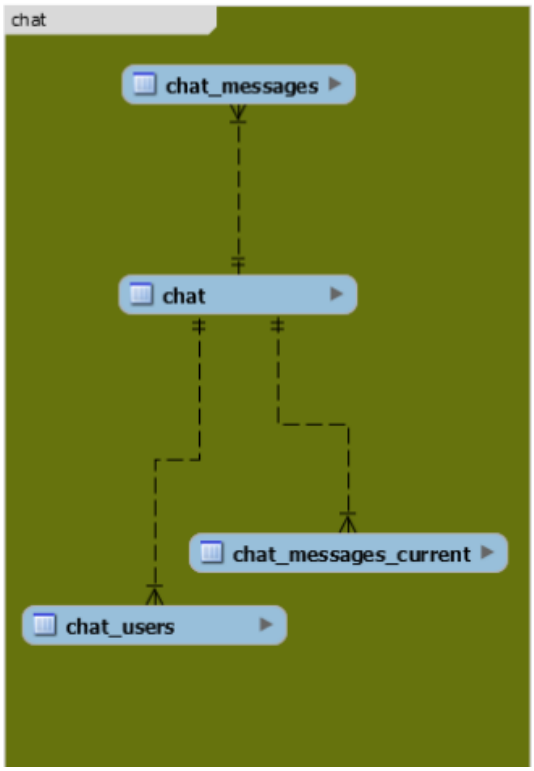

Figura 16. Relacionamento das tabelas que compõem o módulo *chat* (examulator, 2016).

As consultas 28, 29, 30 e 31 extraem informações relacionadas aos *chats* da turma. Estas são apresentadas a seguir:

i. Consulta 28: seleciona todos os *chats* da turma “*select * from mdlacademico_chat where course = '"+curso.get(i).id+"'*”;

ii. Consulta 29: extrai o total de mensagens postadas no *chat* “*select count(chatid) from mdlacademico_chat_messages where chatid = '"+rs.getInt("id")+"' and message !='exit' and message !='enter' and message not like'beep %'*”;

iii. Consulta 30: extrai o total de mensagens postadas pelos alunos no *chat* “*select count(chatid) from mdlacademico_chat_messages where chatid = '"+rs.getInt("id")+"' and message !='exit' and message !='enter' and message not like'beep %' and userid in (select userid from mdlacademico_role_assignments where roleid = 5)*”;

iv. Consulta 31: extrai o total de mensagens postadas no *chat* pelo tutores "*select count(chatid) from mdlacademico_chat_messages where chatid = '"+rs.getInt("id")+"' and message !='exit' and message !='enter' and message not like'beep %' and userid='"+curso.get(i).listaTutor.get(0).id+"'*".

As consultas 32, 33 e 34 extraem informações relacionadas às atividades da turma. Estas são apresentadas a seguir:

i. Consulta 32: extrai a quantidade de atividades criadas pelo tutor "*select ml.userid as id_do_tutor, count (distinct ml.cmid) as total_de_tarefas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.module = 'assign' and ml.action = 'add' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_tarefas desc*";
ii. Consulta 33: extrai a quantidade de atividades avaliadas pelo tutor "*select ml.userid as id_do_tutor, count (ml.id) as total_de_tarefas_avaliadas from mdlacademico_log ml inner join mdlacademico_course_modules cm on cm.id = ml.cmid inner join mdlacademico_role_assignments ra on ra.userid = ml.userid where ra.roleid = 10 and ml.course = '"+curso.get(i).id+"' and ml.module = 'assign' and ml.action = 'grade submission' and ra.contextid in (select id from mdlacademico_context c where c.instanceid = ml.course) group by ml.userid order by total_de_tarefas_avaliadas desc*";
iii. Consulta 34: extrai a quantidade de submissão dos alunos para a atividade "*select count(itemid) as n from mdlacademico_grade_grades where itemid = '"+rs.getInt("id")+"'*".

Após a extração das informações, dá-se início a etapa de pré-processamento das informações, que é exposta na subseção a seguir.

4.1.2 Pré-processamento dos Dados

Uma vez que todas as informações são extraídas, a próxima etapa é realizar o pré-processamento dos dados.

Conforme cita Faceli *et. al.* (2011), a etapa de pré-processamento dos dados consiste na realização das seguintes atividades:

i. Eliminação manual de atributos;
ii. Integração de Dados;
iii. Amostragem de Dados;
iv. Balanceamento de Dados;
v. Limpeza de Dados; e
vi. Transformação de Dados.

As informações referentes a cada uma dessas etapas, bem como os procedimentos adotados, são descritos a seguir.

4.1.2.1 Eliminação Manual de Atributos

A eliminação manual de atributos corresponde aos atributos que não contribuem para a solução do problema em questão. Em relação a esta etapa, vários atributos foram eliminados. Isto ocorreu porque várias consultas SQL contêm o símbolo * (*all*), isto se dá porque algumas informações das tabelas são necessárias para novas consultas, como, por exemplo, os IDs das atividades, dos questionários, dos *chats*, entre outras informações, porém estas informações não têm relação direta com a análise comportamental de alunos e tutores.

Dessa forma, somente foram armazenadas as informações correspondentes aos comportamentos dos tutores e alunos, enquanto que as demais informações, que são necessárias para novas consultas, porém não têm valor comportamental dos alunos e tutores foram excluídas.

4.1.2.2 Integração de Dados

Uma vez que os dados podem vir de diferentes fontes, esta etapa está relacionada com a integração de diferentes conjuntos de dados. Neste caso, tendo em vista que as informações

são extraídas de uma única base de dados, não houve nenhum processamento específico para esta etapa.

4.1.2.3 Amostragem de Dados

Esta etapa trata a questão do grande número de objetos no *dataset* que pode resultar na saturação de memória e no aumento do tempo computacional para ajustar os parâmetros do modelo computacional da técnica de análise de dados.

Para este trabalho, não houve a necessidade de aplicar processamento para esta etapa, pois o *dataset* resultante não é tão extenso ao ponto de saturar a memória de um PC (*Personal Computer*) ou ainda em resultar em uma alta carga de processamento computacional para a análise das informações.

4.1.2.4 Balanceamento de Dados

Esta etapa está relacionada com a variação da quantidade de objetos para as diferentes classes. Uma vez que exista um grande desbalanceamento do número de instâncias para as classes, isto pode implicar diretamente na acurácia preditiva de classificador, de modo que favoreça a classificação na classe majoritária.

No entanto, o modelo de classificação utilizado por este trabalho não é o modelo preditivo, e sim o modelo descritivo, no qual não há rótulos de saída para as informações comportamentais que serão classificadas por meio de agrupamentos. Desta forma, este modelo se torna bem menos sensível ao desbalanceamento do que os modelos preditivos de classificação.

4.1.2.5 Limpeza de Dados

Esta etapa está relacionada com a qualidade dos dados, nos quais podem conter alguns problemas, tais como:

i. Ruídos: erros ou valores diferentes do esperado;
ii. Inconsistências: não combinam/contradizem valores de outros atributos no mesmo objeto;
iii. Redundâncias: objetos/atributos com mesmos valores.

Para esta etapa, foi realizada uma minuciosa análise de todas as informações das turmas extraídas, a fim de verificar a existência de inconsistência de informação. No entanto, na análise realizada não foram encontradas inconsistência nos dados.

4.1.2.6 Transformação de Dados

Esta etapa corresponde à dificuldade que vários algoritmos de AM têm em usar os dados em seu formato original, por exemplo, algumas técnicas são limitadas à manipulação de valores de determinado tipo, tais como valores apenas numéricos ou simbólicos.

Como todas as informações analisadas são valores numéricos, pois correspondem a informações comportamentais, inexistiram problemas para as técnicas que foram empregadas neste trabalho.

4.1.2 Construção do *Dataset*

Finalizada as etapas de extração e pré-processamento, os dados são armazenados em um arquivo no formato ARFF, que é utilizado na organização dos dados que é aceita pelo *software* WEKA. O arquivo ARFF no formato de organização do WEKA tem duas seções distintas.

A primeira seção se refere ao cabeçalho do *dataset*, enquanto que na segunda seção são inseridas as informações propriamente ditas. O cabeçalho do arquivo ARFF contém o nome da relação, uma lista dos atributos (as colunas nos dados) e seus tipos. A Figura 17 apresenta um cabeçalho exemplo sobre o conjunto de dados IRIS.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Figura 17. Cabeçalho exemplo de um *dataset* no formato de organização do WEKA (Weka b, 2016).

As linhas que começam com '%' são comentários. O "@RELATION" e "@Attribute" correspondem ao nome da base (*dataset*) e ao nome dos atributos, respectivamente. Vale salientar que o último atributo da Figura 17 tem por nome *class* e representa o rótulo de saída dos atributos, ou seja, as informações dos atributos correspondem às classes de saída Iris-setosa, Iris-versicolor, e Iris-virginica.

Devido às informações comportamentais dos tutores e alunos não terem um rótulo de saída, logo, na formação do *dataset* deste trabalho, não há um atributo que corresponda à classe de saída.

Já a segunda seção do arquivo contém os valores dos atributos que foram definidos na primeira seção. A Figura 18 apresenta o exemplo da organização dos dados referente ao que foi definido na Figura 17.

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Figura 18. Organização dos valores dos atributos no arquivo ARFF (Weka b, 2016).

Dessa forma, tendo sido feita a extração, pré-processamento e construção do *dataset* em um formato adequado para a análise, pode-se dar início então a etapa de análise e processamento dos dados, que é realizada pelo núcleo de processamento dos dados (*core* II) e é descrita a seguir.

4.2 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

A análise e interpretação dos resultados se dão através da atuação em conjunto do núcleo de processamento dos dados (*core* II) e núcleo de apresentação de dados (*core* III), que é responsável pela exibição das informações processadas no *core* II.

O *core* II tem o objetivo de processar as informações que foram extraídas pelo *core* I. Ao todo são realizados três tipos de processamento no *core* II, que são:

i. Cálculo da posição relativa, via técnica de normalização Min-Max, das informações dos cursos em comparação com as outras turmas;
ii. Criação do modelo de agrupamento por meio do uso das técnicas *K-Means* e *Farthest First*;
iii. Cálculos dos coeficientes de correlação entre os atributos do tutor com os atributos da turma.

A Figura 19 apresenta, de forma detalhada, a arquitetura do sistema de avaliação de tutores e turmas.

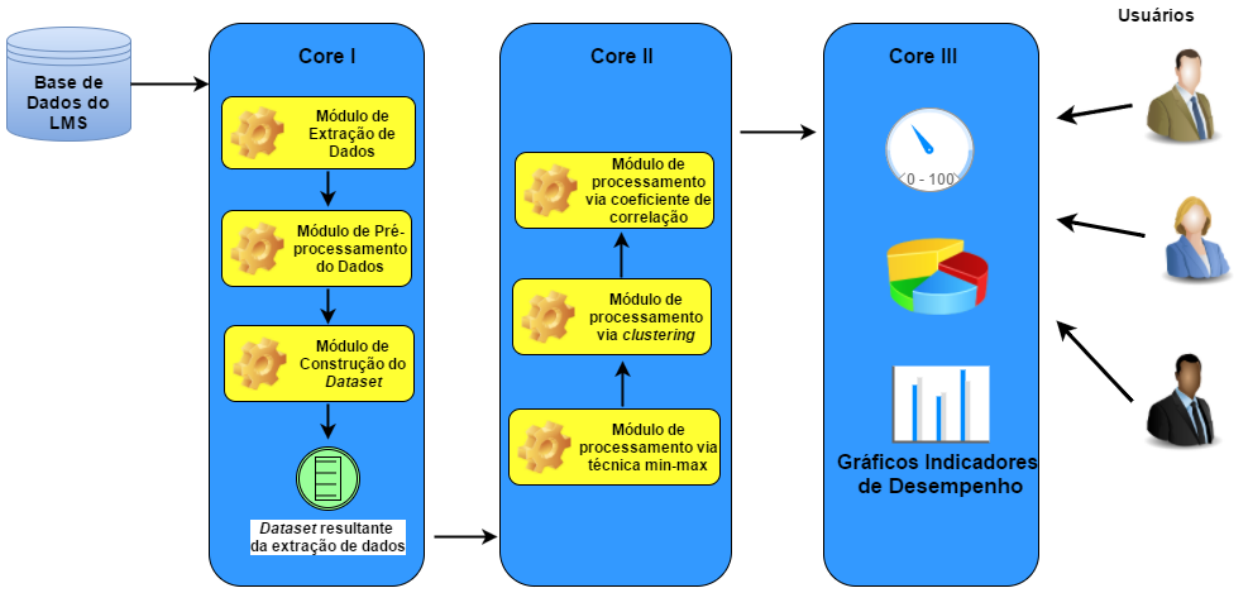

Figura 19. Arquitetura detalhada do sistema de avaliação de tutores e turmas. Fonte: Autoria Própria.

A Figura 20 apresenta a listagem dos cursos a serem avaliados.

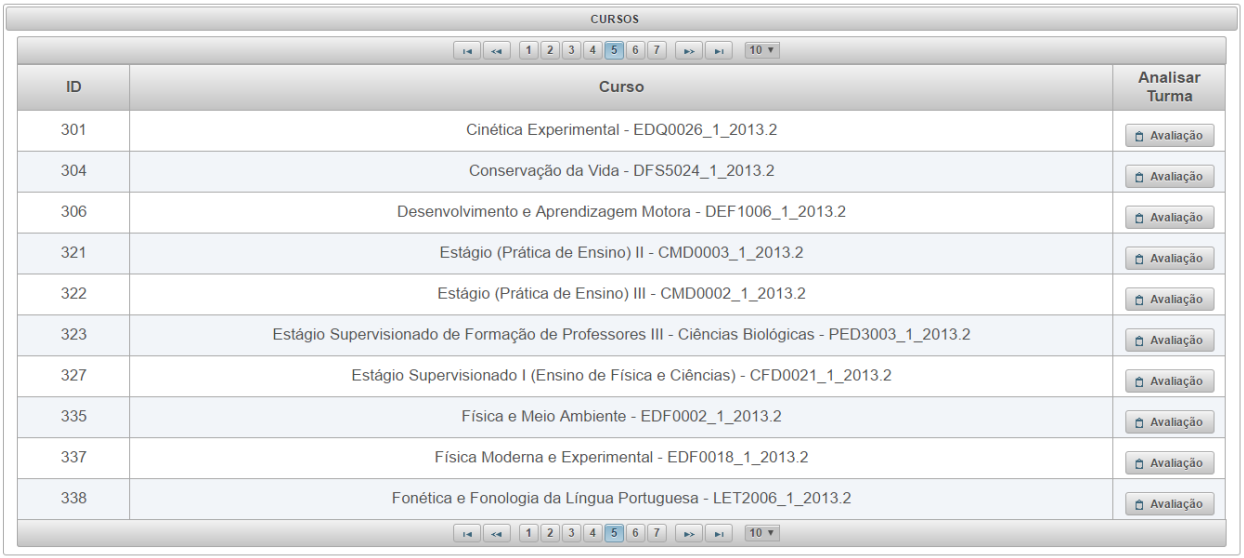

Figura 20. Listagem dos cursos a serem avaliados. Fonte: Autoria Própria.

Após a escolha dos cursos a serem avaliados, os dados são extraídos em tempo real (*core* I), e, em seguida, são processados no *core* II, e, por fim, as respectivas informações de classificação e desempenho relativo são apresentadas (*core* III).

As informações extraídas são divididas em duas categorias de apresentação. A primeira se refere às informações dos tutores, enquanto que a segunda se refere às informações da turma. As Figuras 21 e 22 apresentam as telas iniciais da categoria de avaliação do tutor e da turma, respectivamente.

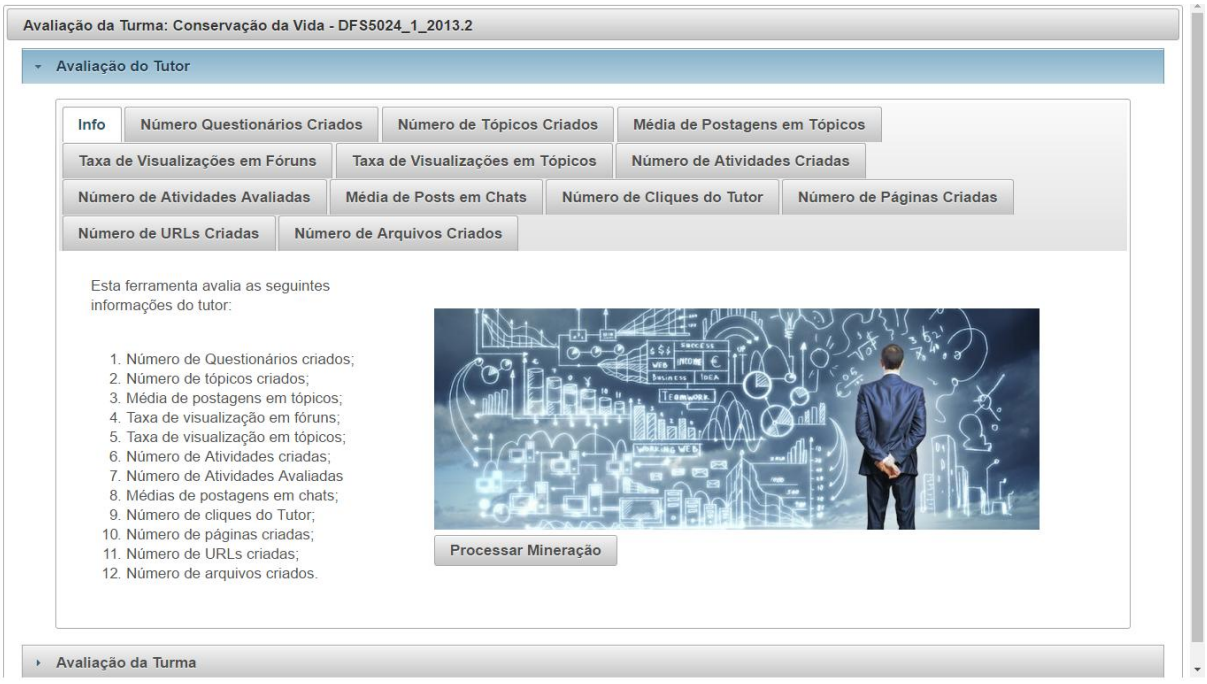

Figura 21. Tela inicial da categoria de avaliação do tutor. Fonte: Autoria Própria.

Figura 22. Tela inicial da categoria de avaliação da turma. Fonte: Autoria Própria.

A fim de facilitar o entendimento, optamos por dividir as informações referentes aos três módulos de processamento do *core* II, bem como os resultados obtidos, por cada módulo de processamento.

4.2.1 Análise e Resultados da Normalização Min-Max

O objetivo do cálculo da posição relativa de cada atributo é permitir a comparação com os dados das demais turmas. Para isso, é utilizada a técnica de normalização Min-Max, que é definida pela Equação 4 (Faceli *et. al.*, 2011).

$$v' = \frac{(v - min)}{(\max - min)} \times (nmax - nmin\ ) + nmin \tag{4}$$

Onde:

i. *v* é o valor a ser normalizado;
ii. *min* é o valor mínimo encontrado no *dataset*;
iii. *max* é o valor máximo encontrado no *dataset*;
iv. *nmax* é o valor máximo do novo intervalo;
v. *nmin* é o valor mínimo do novo intervalo.

Desse modo, isto irá permitir uma avaliação de cada atributo do tutor e da turma em comparação direta com o desempenho obtido por outras turmas. Para o calculo da posição relativa, os valores foram convertidos para uma escala entre 0 e 100, fazendo-se do uso da Equação 4.

A Figura 23 apresenta o gráfico do cálculo do valor relativo do atributo taxa de visualizações em tópicos do tutor.Vale frisar que as informações dos valores relativos foram divididas em três tipos, que podem ser identificados pelas cores vermelha, amarela e verde. Se o ponteiro estiver apontado para a cor vermelha, então significa que o valor relativo k está na faixa de $0 \leq k < 33$ do intervalo de 0 a 100. Caso o ponteiro esteja apontando para a cor amarela, então o valor relativo k está na faixa de $33 \leq k < 66$ do intervalo de 0 a 100, e se o ponteiro estiver apontando para a cor verde, então significa que o valor relativo k está na faixa de $66 \leq k \leq 100$ do intervalo de 0 a 100.

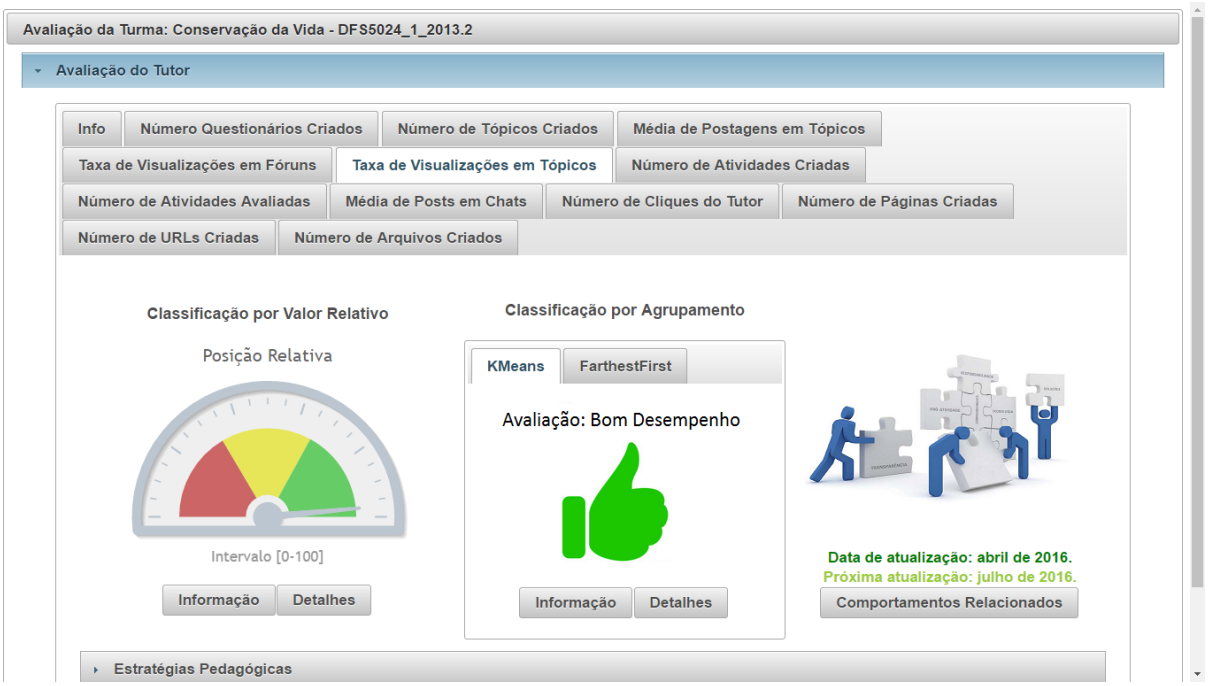

Figura 23. Tela de avaliação do atributo Taxa de Visualização em Tópicos do tutor. Fonte: Autoria Própria.

Além disso, foi disponibilizada também a opção de verificar a posição da turma avaliada em comparação com as demais turmas tanto em valores relativos como também em valores absolutos.

A Figura 24 apresenta a opção supracitada após o usuário clicar no botão ‘Detalhes’ referente ao conteúdo de Posição Relativa.

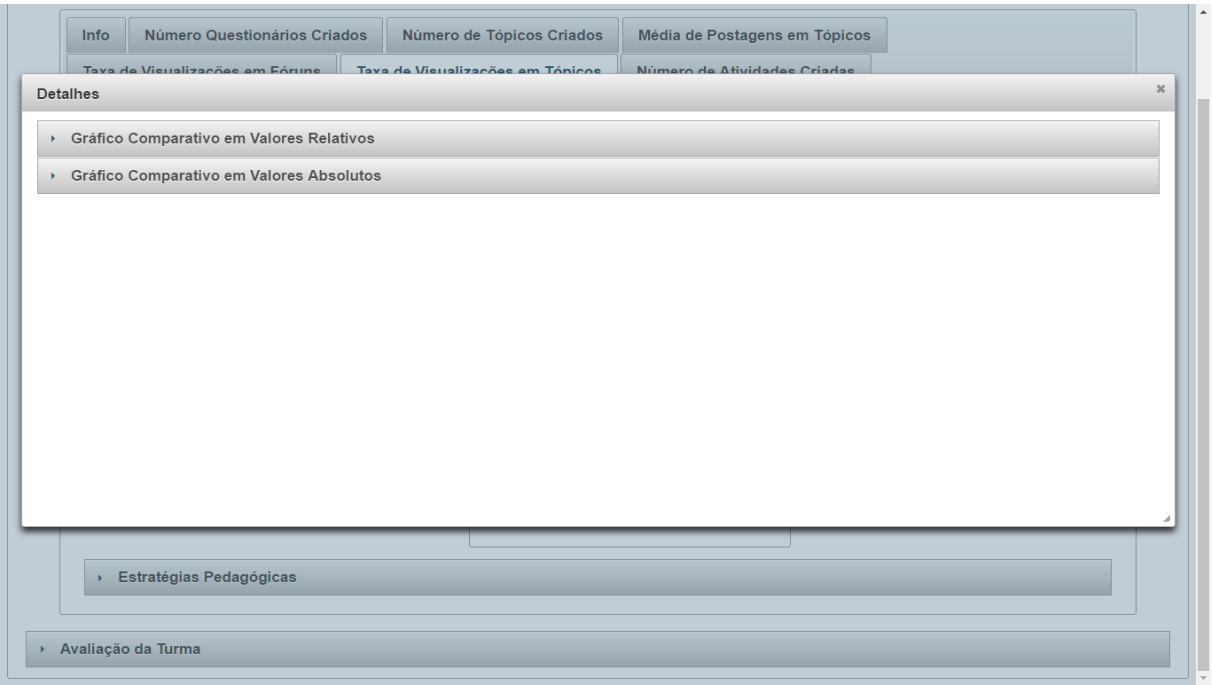

Figura 24. Opção para escolha de visualização das informações do atributo Taxa de Visualização de Tópicos. Fonte: Autoria Própria.

Caso o usuário clique na aba "Gráfico Comparativo em Valores Relativos", será mostrado o gráfico em barras do comportamento do tutor em comparação com os tutores das outras turmas, conforme ilustra a Figura 25.

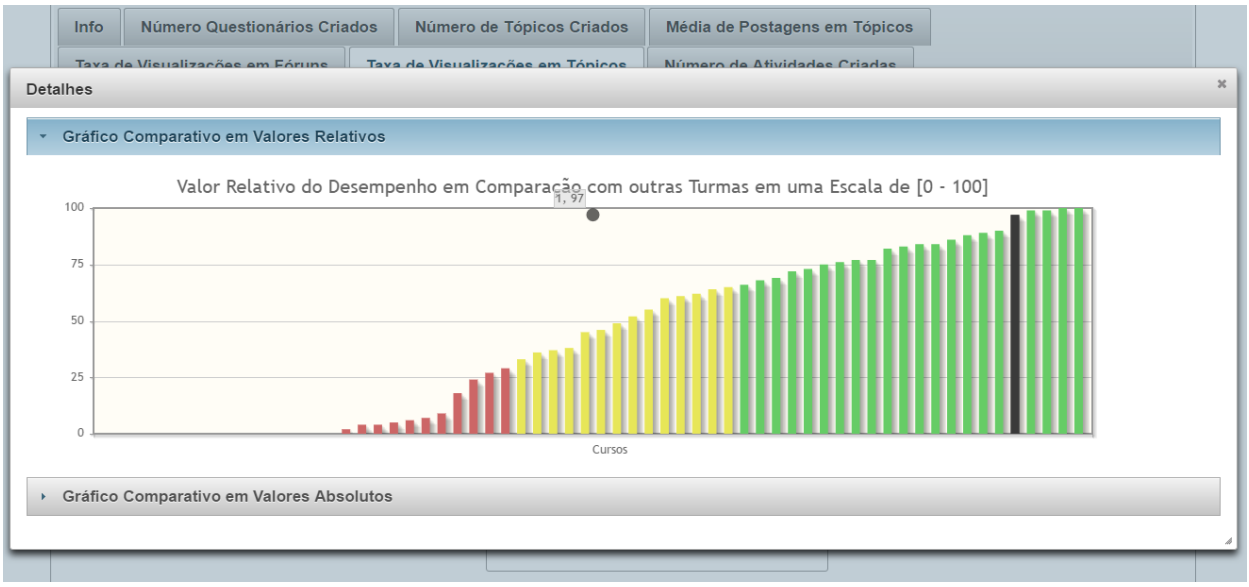

Figura 25. Gráfico comparativo em barras do atributo Taxa de Visualização em Tópicos do tutor usando o critério de valores relativos. Fonte: Autoria Própria.

É importante frisar que a turma em questão está destacada pela cor preta, e que ao passar o *mouse* sobre barra (cor preta) que representa a turma avaliada, é informado seu valor relativo de forma numérica, conforme pode ser visto na Figura 25.

Já se a escolha for a aba “Gráfico Comparativo em Valores Absolutos”, os valores serão mostrado sem a normalização via técnica Min-Max, como pode ser visto na Figura 26.

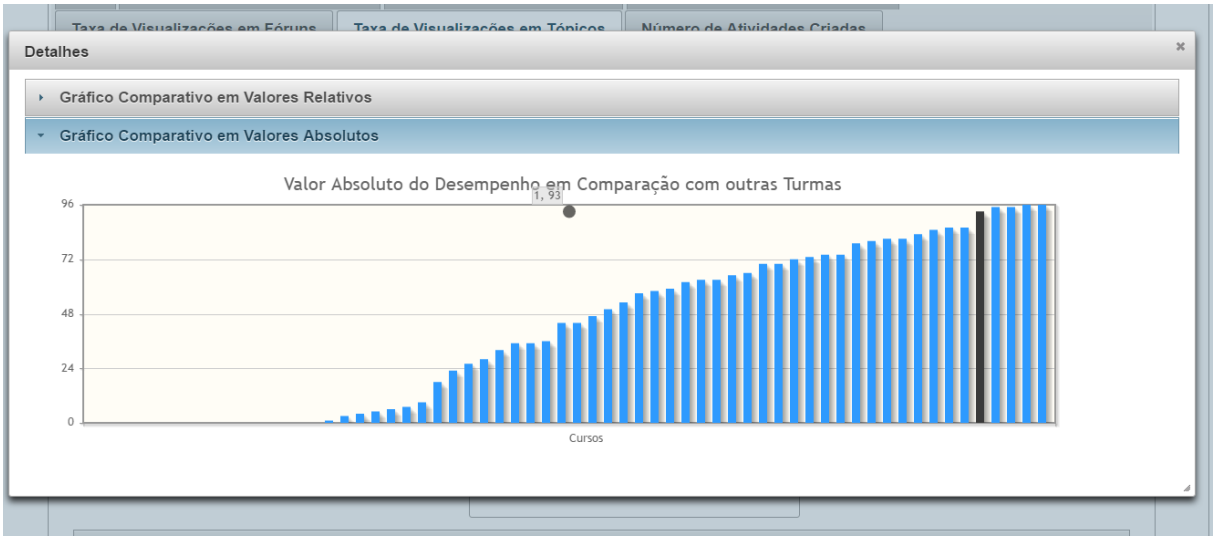

Figura 26. Gráfico comparativo em barras do atributo Taxa de Visualização em Tópicos do tutor usando o critério de valores absolutos. Fonte: Autoria Própria.

Devido as informações estarem em valores absolutos, não foi realizada a categorização for faixas de valores. Perceba que na Figura 25 as informações estão em uma escala de 0 a 100, porém na Figura 26 os valores estão em uma faixa de 0 a 96. Dessa forma, significa que de todas as turmas analisadas o pior comportamento obtido foi dos tutores que tiveram 0% de visualização de tópicos, enquanto que o melhor comportamento obtido foi dos tutores que visualizaram até 96% dos tópicos.

Na turma avaliada na figura, o tutor obteve um percentual de 93% de visualização dos tópicos, conforme é exposto na Figura 26, que se comparado com o melhor comportamento de todas as turmas, que foi de 96% e o pior que foi 0%, em valores relativos numa escala de 0 a 100, seu valor seria de 96,875 ou $\cong 97$. É importante mencionar que os valores fracionários no gráfico em barras não puderam ser apresentados devido a limitações da API Prime Faces. Dessa forma, optou-se pelo arredondamento das casas decimais para este gráfico, conforme pode ser visto na Figura 25.

4.2.2 Análise e Resultados da Clusterização

A análise de dados por agrupamento tem por objetivo criar para cada atributo do tutor e da turma a sua própria rede de classificação de acordo com os agrupamentos dos comportamentos das outras turmas.

Desse modo, além da classificação por posição relativa, o sistema dispõe de outra forma de avaliação a partir de duas técnicas de aprendizado não supervisionado, que, embora pertençam à mesma categoria de agrupamento, estas diferem entre si.

A grande vantagem desta forma de avaliação em comparação com a técnica de normalização Min-Max é que, além de permitir outra forma de avaliação sobre os resultados, estes modelos são bem menos sensíveis a *outliers* se comparados com a técnica de normalização Min-Max.

Vale salientar que a aplicação da técnica de remoção de *outliers* Z-Score só pode ser aplicada no módulo de análise dos coeficientes de correlação entre variáveis, pois, se essa técnica fosse aplicada no módulo de clusterização e de cálculo da posição relativa, as turmas cujos dados fossem considerados *outliers* pela técnica Z-Score ficariam sem a classificação dos seus atributos, devido suas informações terem sido removidas pela técnica Z-Score.

Sendo assim, a avaliação dos atributos do tutor e da turma foi realizada por meio das técnicas *K-Means* e *Farthest First*, utilizando-se do critério da distância euclidiana para o *K-Means* entre os centroides, tendo sido definido o número *k* de clusters iguais a três para as duas técnicas, ou seja, foram criados três agrupamentos que se referem à categoria dos que tiveram baixo desempenho, desempenho moderado ou um bom desempenho.

A Figura 27 apresenta o resultado da classificação do atributo ‘Taxa de Visualizações em Fóruns’ por meio da técnica *K-Means*.

Figura 27. Resultado da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *K-Means*. Fonte: Autoria Própria.

Caso se deseje uma avaliação mais detalhada do resultado, a opção ‘Detalhes’ referente à classificação do *K-Means* apresenta a posição da turma avaliada em relação aos centroides resultantes do agrupamento, conforme é exposto na Figura 28.

Pode-se observar na Figura 28 que existem três *clusters*, sendo o *cluster* de cor vermelha referente ao baixo desempenho, o *cluster* de cor amarela referente ao desempenho moderado e o *cluster* de cor verde referente ao bom desempenho. O ponto preto se refere ao valor comportamental da turma avaliada, e os pontos azuis se referem aos valores comportamentais das outras turmas.

Note que o ponto preto, que se refere ao valor comportamental da turma em questão, está mais próximo do centroide do *cluster* de bom desempenho, se comparado com os outros *clusters*. Desse modo, a indicação é de bom desempenho para este atributo segundo a análise por meio da técnica *K-Means*.

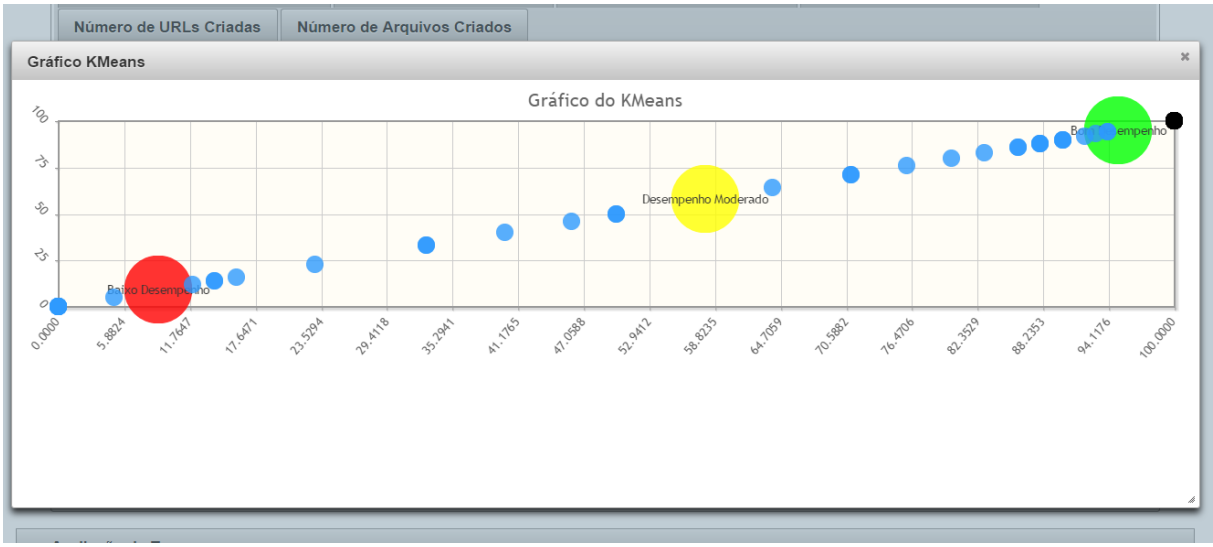

Figura 28. Detalhes da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *K-Means*. Fonte: Autoria Própria.

O mesmo conceito é utilizado para a avaliação por meio do algoritmo *Farthest First*. As Figuras 29 e 30 apresentam o resultado da classificação do atributo ‘Taxa de Visualizações em Fóruns’ por meio da técnica *Farthest First*, bem como o detalhamento do resultado para esta técnica.

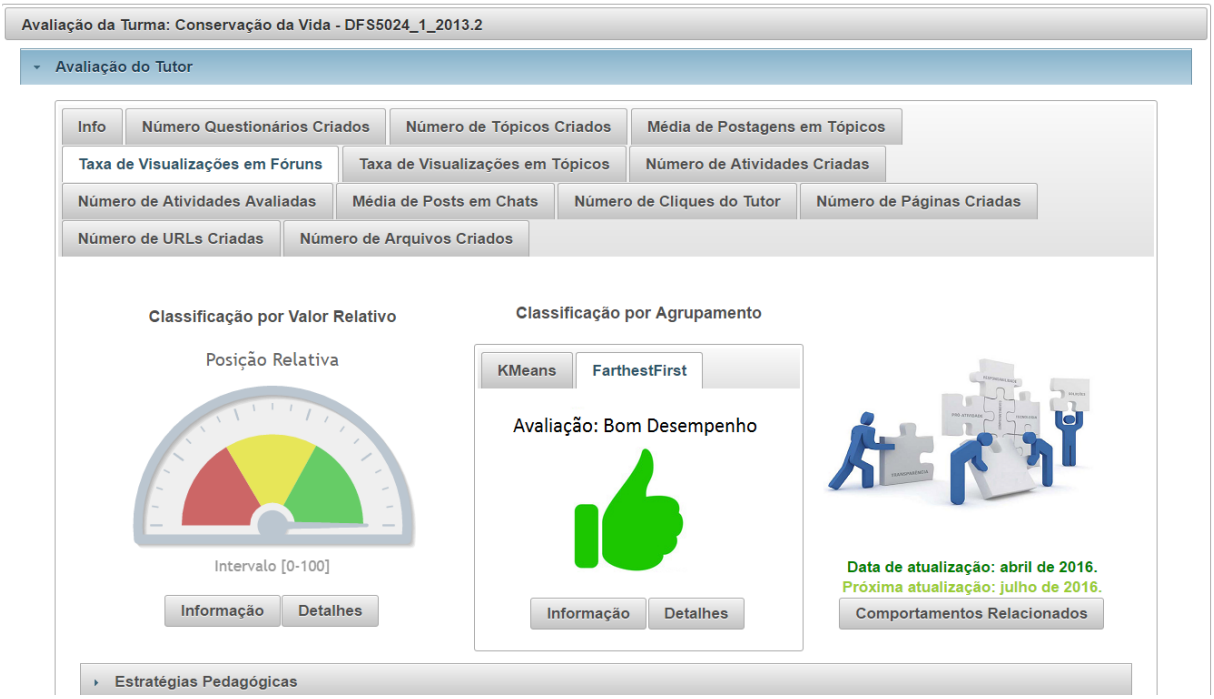

Figura 29. Resultado da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *Farthest First*. Fonte: Autoria Própria.

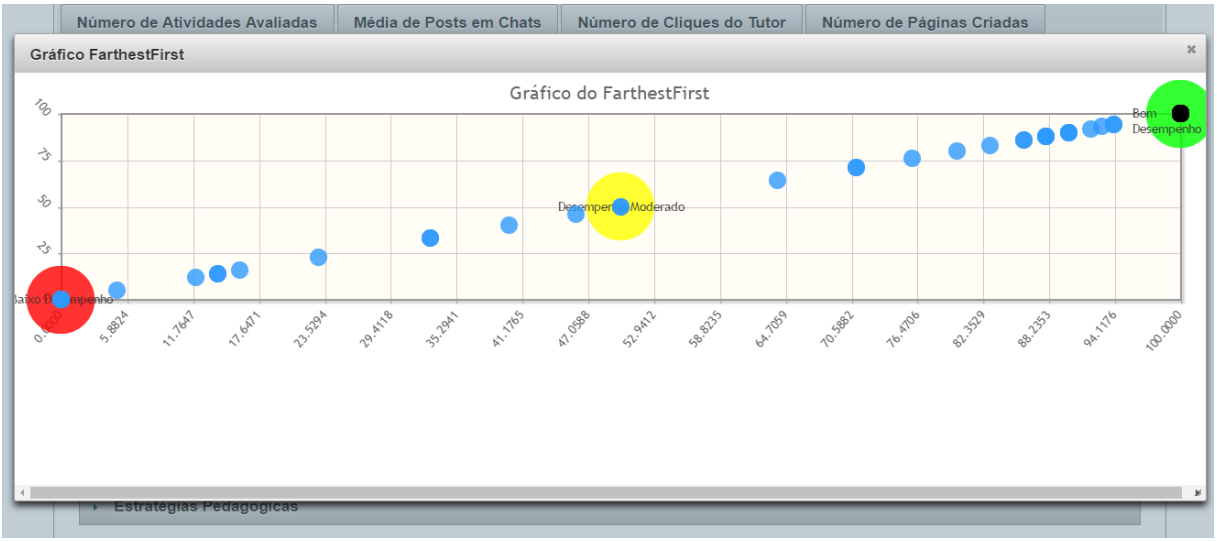

Figura 30. Detalhes da classificação do atributo Taxa de Visualizações em Fóruns por meio da técnica *Farthest First*. Fonte: Autoria Própria.

Pode-se notar que muito embora as técnicas *K-Means* e *Farthest First* tenham indicado a mesma avaliação, que foi bom desempenho, a posição dos seus clusters não foi igual, conforme pode ser visto por meio da comparação da Figura 28 com a Figura 30.

Além disso, vale ressaltar que para a avaliação do atributo 'Taxa de visualização de Tópicos' do tutor, o indicador foi de bom desempenho segundo a técnica de normalização Min-Max, e agrupamento *K-Means* e *Farthest First*. Porém, não necessariamente os

resultados irão sempre coincidir, pois a coincidência dos indicadores de desempenho não depende por si só das técnicas envolvidas na avaliação, mas também do conjunto de informações que serão analisadas.

4.2.3 Análise e Resultados da Correlação entre Variáveis

O módulo de coeficiente e correlação de variáveis tem por objetivo calcular os resultados das correlações entre os atributos comportamentais do tutor e da turma, de modo que seja possível identificar quais comportamentos podem ser associados tanto de forma positiva ou negativa com os comportamentos da turma.

Para a execução do coeficiente de correlação entre as variáveis, é disposta uma opção na aba ‘info’ do módulo tutor, conforme pode ser visto na Figura 21. Caso o usuário clique nesta opção, o módulo irá requisitar do *core* I a extração, pré-processamento e construção do *dataset*, que consequentemente resultará na formação do *dataset* com informações das turmas atualizadas.

Em seguida, este módulo irá calcular o coeficiente de correlação de todos os atributos do tutor em relação com os atributos da turma, e, para todos os resultados cujas informações sejam de recomendação da hipótese alternativa (H_1), o módulo criará textos de forma automática com as informações resultantes da correlação entre variáveis. Sendo os resultados exibidos na opção ‘Comportamentos Relacionados’, que está dentro da aba de cada um dos atributos analisados do tutor.

Por exemplo, na Figura 29, que apresenta os resultados da análise no que se refere ao atributo taxa de visualizações em fóruns do tutor, caso o usuário clique na opção ‘Comportamentos Relacionados’, serão expostos todos os comportamentos da turma que são influenciados pelo comportamento taxa de visualizações em fóruns do tutor.

Vale salientar que muito embora a extração dos dados e o processamento da correlação possam ser feitos a qualquer hora, o ideal é que o uso dessa opção seja utilizado somente no término do semestre. Pois usar dados de turmas já finalizadas em conjunto com dados de turmas ainda em andamento no cálculo da correlação entre variáveis resultará em resultados não confiáveis da correlação, devido à utilização de informações que não estão em um mesmo contexto, ou seja, as turmas já finalizadas e as turmas ainda em andamento, que

estão sujeitas a alterações de valores no que se refere aos comportamentos dos alunos e tutores.

A Figura 31 apresenta a caixa de diálogo que apresenta o resultado da correlação entre o atributo taxa de visualizações em fóruns do tutor com os atributos da turma que resultaram na aceitação da hipótese alternativa.

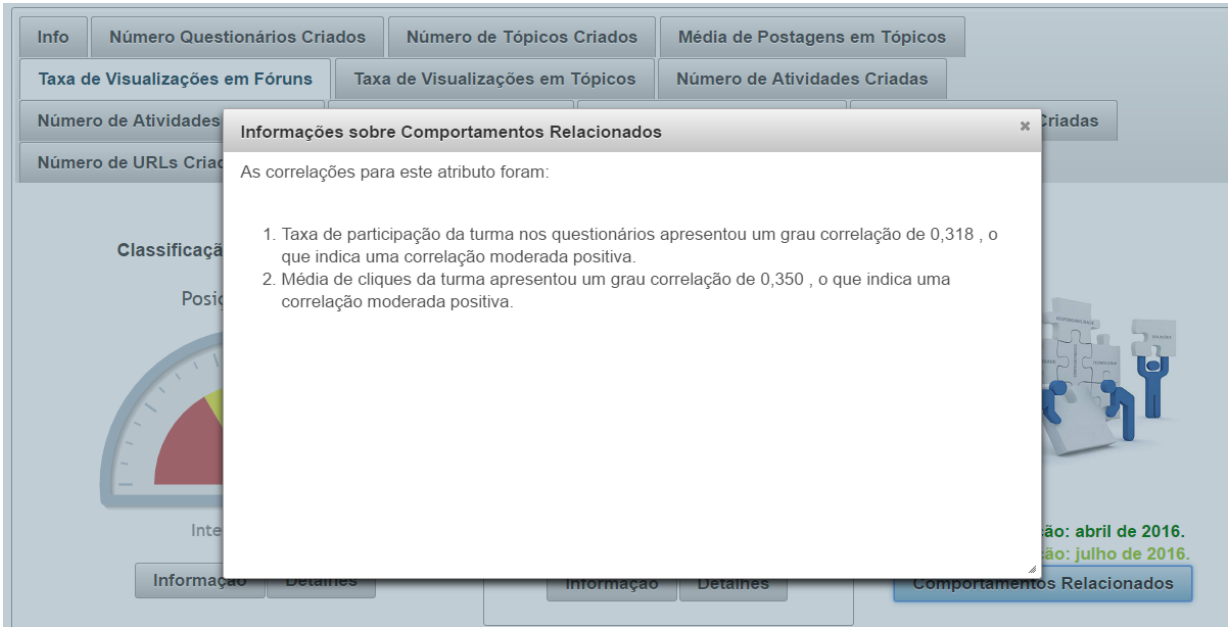

Figura 31. Apresentação dos resultados da correlação do atributo Taxa de Visualização em Fóruns com os atributos da turma. Fonte: Autoria Própria.

A análise realizada abrangeu doze comportamentos dos tutores e onze comportamentos das turmas, cujos atributos foram descritos no Capítulo 3, gerando, portanto, cento e trinta e um (12 x 11) resultados de correlações.

Porém, a fim de simplificar a exposição dos resultados, as tabelas a seguir irão apresentar as informações das correlações cujas hipóteses alternativas (H_1) foram aceitas. Logo, as demais possibilidades que não constam nas tabelas a seguir indicam que a recomendação foi de aceitar a hipótese nula (H_0).

As Tabelas 2, 3, 4, 5, 6, 7, 8, 9,10, 11 e 12 apresentam os resultados da correlação para cada comportamento do tutor em relação com o comportamento da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação da turma nos questionários	62	58	0.468	Correlação Moderada Positiva

Taxa de submissão de atividades	62	60	0.333	Correlação Moderada Positiva
Média de páginas visualizadas	62	57	0.543	Correlação Forte Positiva

Tabela 2. Resultado da correlação do atributo Número de Tópicos Criados pelo Tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Média de tópicos criados por estudante	62	57	0.371 (Aceita H1)	Correlação Moderada Positiva
Média de visualização em tópicos	62	57	0.276 (Aceita H1)	Correlação Fraca Positiva
Média de postagens em *chats*	62	57	0.252 (Aceita H1)	Correlação Fraca Positiva
Média de cliques dos estudantes	62	57	0.240 (Aceita H1)	Correlação Fraca Positiva

Tabela 3. Resultado da correlação do atributo Número de Tópicos Criados pelo Tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	57	0.400 (Aceita H1)	Correlação Moderada Positiva
Média de tópicos criados por estudante	62	55	0.260 (Aceita H1)	Correlação Fraca Positiva
Média de postagens em *chats*	62	55	0.334 (Aceita H1)	Correlação Moderada Positiva

Tabela 4. Resultado da correlação do atributo Média de Postagens em Tópicos dos Fóruns do tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	60	0.318 (Aceita H1)	Correlação Moderada Positiva
Média de cliques dos estudantes	62	58	0.350 (Aceita H1)	Correlação Moderada Positiva

Tabela 5. Resultado da correlação do atributo Taxa de Visualizações em Fóruns do tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	60	0.412 (Aceita H1)	Correlação Moderada Positiva
Taxa de submissão em atividades	62	62	0.262 (Aceita H1)	Correlação Fraca Positiva
Média de cliques dos estudantes	62	58	0.270 (Aceita H1)	Correlação Fraca Positiva

Tabela 6. Resultado da correlação do atributo Taxa de Visualizações em Tópicos dos Fóruns do tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Média de postagens em *chats*	62	56	0.223	Correlação Fraca Positiva

Tabela 7. Resultado da correlação do atributo Número de Atividades Criadas pelo tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação da turma nos questionários	62	60	0.344	Correlação Moderada Positiva

Média de páginas visualizadas	62	57	0.319	Correlação Moderada Positiva

Tabela 8. Resultado da correlação do atributo Número de Atividades Avaliadas pelo tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Média de postagens em *chats*	62	55	0.590 (Aceita H1)	Correlação Forte Positiva

Tabela 9. Resultado da correlação do atributo Média de Postagens em Chats pelo tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	57	0.463 (Aceita H1)	Correlação Moderada Positiva
Taxa de submissão em atividades	62	58	0.293 (Aceita H1)	Correlação Fraca Positiva
Média de páginas visualizadas pelos estudantes	62	55	0.345 (Aceita H1)	Correlação Moderada Positiva

Tabela 10. Resultado da correlação do atributo Número de Cliques do Tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	59	0.357 (Aceita H1)	Correlação Moderada Positiva
Média de visualização em tópicos	62	56	0.269 (Aceita H1)	Correlação Fraca Positiva
Taxa de submissão em atividades	62	59	0.311 (Aceita H1)	Correlação Moderada Positiva

Média de páginas visualizadas pelos estudantes	62	55	0.511 (Aceita H1)	Correlação Forte Positiva

Tabela 11. Resultado da correlação do atributo Número de URLs Criadas pelo tutor com os atributos da turma.

Atributos da turma	Número de instâncias	Número de instâncias sem *outliers*	Resultado da correlação	Interpretação para a correlação
Taxa de participação em questionários	62	57	0.493 (Aceita H1)	Correlação Moderada Positiva

Tabela 12. Resultado da correlação do atributo Número de Arquivos Criados pelo tutor com os atributos da turma.

Note que inexiste a tabela relacionada ao atributo do tutor número de páginas criadas. Isto se deu devido às recomendações de aceitação da hipótese nula (H_0) para todos os resultados de correlação deste atributo.

As subseções a seguir apresentam uma discussão sobre os resultados obtidos.

4.2.3.1 Número de Questionários Criados

As correlações resultantes cuja indicação foi de aceitação da hipótese alternativa entre o número de questionários criados e os comportamentos da turma foram:

i. Taxa de participação da turma nos questionários apresentou um grau correlação de 0.468, o que indica uma correlação moderada positiva, onde, à medida que o tutor propõe a resolução de mais questionários, a participação da turma nestes tende a melhorar;
ii. Taxa de submissão de atividades apresentou um grau correlação de 0.333, o que indica uma correlação moderada positiva, desse modo, a adição de questionários tende a melhor de forma moderada os índices de submissão de atividades da turma;
iii. Média de páginas visualizadas apresentou um grau correlação de 0.543, o que indica uma correlação forte positiva, dessa forma, percebe-se que a criação de questionários melhora fortemente a visualização das páginas do curso por parte dos alunos.

4.2.3.2 Número de Tópicos Criados

O grau de correlação entre o número de tópicos criados pelo tutor gerou quatro correlações positivas, que foram:

i. Em relação à média de tópicos criados pela turma, o grau foi de $r = 0.371$, o que indica uma correlação moderada positiva, onde a medida que mais tópicos são criados pelos tutores, mais tópicos também são criados pelos alunos, de modo que se pode perceber que a criação de tópicos pelos tutores estimulam também aos alunos criarem seus tópicos de discussão;
ii. Em relação à média de visualização de tópicos pela turma, o grau foi de $r = 0.276$, indicando uma correlação fraca positiva, onde o aumento da criação dos tópicos por parte dos tutores influi para uma melhora na taxa de visualização dos tópicos nos fóruns pelos alunos;
iii. Em relação à média de postagens em *chats*, o grau foi de $r = 0.252$, indicando uma correlação fraca positiva, dessa forma, pode-se perceber que a criação de tópicos em fóruns pelos tutores melhora, ainda que fracamente, os índices da média de postagens em *chats* pelos alunos;
iv. Em relação à média de cliques da turma, o grau foi de $r = 0.240$, indicando uma correlação fraca positiva, desse modo, pode-se inferir que, à medida que tópicos são criados pelos tutores, de fato, há uma maior participação no LMS.

Uma vez que o número de tópicos criados pelos tutores reflete no aumento da criação de tópicos pela turma, aumento na taxa de visualização de tópicos e aumento na média de postagens em *chats*, desse modo se explica o aumento leve no número de cliques pelos estudantes.

4.2.3.3 Média de Postagens em Tópicos dos Fóruns

O grau de correlação entre o número de postagens em tópicos dos fóruns gerou três correlações positivas, que foram:

i. Em relação à taxa de participação de questionários pela turma, o grau foi de $r = 0.400$, o que indica uma correlação moderada positiva. Desse modo, com este resultado pode-se inferir que a criação de tópicos de discussão por parte dos tutores reflete em uma melhora significativa na participação dos estudantes nos questionários passados;

ii. Em relação à média de criação de tópicos por parte da turma, o grau foi de $r = 0.260$, o que indica uma correlação fraca positiva. A interpretação para este resultado é que, à medida que os tutores fazem mais postagens em tópicos de fóruns, a média de criação de tópicos por parte da turma tende levemente a aumentar. Isto pode ser refletido devido à atuação dos tutores nos tópicos, no qual os alunos percebem a participação dos tutores nos tópicos dos fóruns, e criam mais tópicos de discussão, expondo suas dúvidas, com a convicção que seu tópico criado não será ignorado pelo tutor;

iii. Em relação à média de postagens em *chats* pela turma, o grau foi de $r = 0.334$, o que indica uma correlação moderada positiva. A partir deste resultado pode-se inferir que a criação de tópicos de discussão nos fóruns por parte dos tutores tendem a refletir positivamente no aumento moderado de postagens em *chats* pelos alunos.

4.2.3.4 Taxa de Visualização em Fóruns

O grau de correlação entre a taxa de visualização de fóruns e os demais atributos da turma apresentou duas correlações positivas, que foram:

i. Taxa de participação em questionários, com $r = 0.318$, que indica correlação moderada positiva;

ii. Média de cliques dos estudantes, com $r = 0.350$, que indica correlação moderada positiva.

A taxa de visualização de fóruns consiste em identificar o acompanhamento dos tutores nos fóruns de um curso. Por exemplo, se um curso tem 20 fóruns e o tutor visualizou apenas 12 fóruns, isso significa que a taxa de visualização em fóruns desse tutor foi de 60%, pois ele acessou apenas 12 dos 20 fóruns possíveis. Sendo assim, esse atributo reflete o acompanhamento dos tutores nos fóruns que são criados na turma.

O resultado da correlação indica que o acompanhamento dos tutores nos fóruns do curso implica em aumento moderado na taxa de submissão dos questionários, e também aumenta moderadamente a média de cliques dos estudantes da turma.

Vale salientar que a média dos cliques dos estudantes de uma turma em um LMS está diretamente relacionada ao grau de utilização do LMS. Em outras palavras, se a média de cliques de uma turma está bastante alta, significa que os alunos estão acessando bastante o ambiente, enquanto que se a média de cliques estiver muito baixa, significa que o acesso ao LMS acontece de forma casual.

4.2.3.5 Taxa de Visualização em Tópicos dos Fóruns

O grau de correlação entre a taxa de visualização de fóruns e os demais atributos da turma apresentou três correlações positivas, que foram:

i. Taxa de participação em questionários, com r = 0.412, que indica correlação moderada positiva;
ii. Taxa de submissão de tarefas, com r = 0.262, que indica correlação fraca positiva;
iii. Média de cliques dos estudantes, com r = 0.270, que indica correlação fraca positiva.

A taxa de visualização de tópicos consiste em identificar o acompanhamento dos tutores nos tópicos que foram criados dentro dos fóruns de um curso. Por exemplo, se um curso tem 10 fóruns e cada fórum tem 10 tópicos, no total, esse curso possui 100 tópicos. Se tutor visualizou apenas 50 tópicos, isso significa que a taxa de visualização de tópicos desse tutor é de 50%, pois ele acessou apenas 50 dos 100 tópicos possíveis. Sendo assim, esse atributo reflete o acompanhamento dos tutores nos tópicos dos fóruns da turma.

Os resultados da correlação indicam que o acompanhamento dos tutores nos fóruns do curso implica em um moderado aumento na taxa de participação em questionários. Além disso, ainda indica um aumento leve na taxa de submissão de atividade e média de cliques dos alunos da turma.

4.2.3.6 Número de Atividades Criadas

O grau de correlação entre o número de atividades criadas gerou apenas uma correlação positiva, que foi em relação à média de postagens em *chats* pela turma, no qual apresentou um grau de correlação de r = 0.223, o que indica uma correlação fraca positiva.

4.2.3.7 Número de Atividades Avaliadas

O grau de correlação entre o número de atividades avaliadas geraram duas correlações positivas, que foram:

i. Taxa de participação da turma nos questionários apresentou um grau correlação de 0.344, o que indica uma correlação moderada positiva;

ii. Média de páginas visualizadas apresentou um grau correlação de 0.319, o que indica uma correlação moderada positiva.

4.2.3.8 Média de Postagens em *Chats*

O referido atributo apresentou um grau de correlação forte positiva com o atributo média de postagens em *chats* da turma, com $r = 0.506$. Dessa forma, o resultado da correlação indica que a quantidade de postagens dos tutores nos *chats* está fortemente ligada com a quantidade de postagens dos alunos nos *chats*.

4.2.3.9 Número de Cliques do Tutor

O número de cliques do tutor em um LMS representa, de forma básica, a efetividade de sua participação no sistema.

A análise desse atributo com os atributos da turma apresentaram três correlações positivas, que foram:

i. Taxa de participação em questionários, com $r = 0.463$, que indica uma correlação moderada positiva;
ii. Taxa de submissão em tarefas, com $r = 0.293$, que indica uma correlação fraca positiva;
iii. Média de páginas visualizadas pelos estudantes, com $r = 0.345$, que indica uma correlação moderada positiva.

Dessa forma, a partir desses resultados, sugere-se que o frequente uso do sistema contribui para uma melhora efetiva na participação dos alunos, no que se refere aos questionários, as tarefas, bem como em mais acessos nas páginas criadas pelos tutores.

4.2.3.10 Número de URL Criadas

O grau de correlação entre o número de URL criadas e os atributos analisados da turma gerou quatro correlações positivas, que foram:

i. Em relação à taxa de participação em questionários, com $r = 0.357$, o que indica uma correlação moderada positiva;
ii. Em relação à média de visualização de tópicos pela turma, com $r = 0.269$, o que indica uma correlação fraca positiva;

iii. Em relação à taxa de submissão de atividades, com r = 0.311, o que indica uma correlação moderada positiva;
iv. Em relação à média de páginas visualizadas pelos estudantes, com r = 0.511, o que indica uma correlação forte positiva.

Os resultados obtidos para este atributo mostram que a disponibilização de URL de acesso para os alunos tendem a refletir positivamente em uma maior participação efetiva dos alunos no que se refere à submissão de questionários e atividades, e também visualização de páginas e tópicos por parte dos alunos.

4.2.3.11 Número de Arquivos Criados

O número de arquivos criados apresentou correlação positiva com o atributo taxa de participação em questionários, com r = 0.493, que indica uma correlação moderada positiva. Este resultado indica que o aumento no número de arquivos criados pelos tutores melhora a taxa de participação nos questionários.

5 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados alguns trabalhos relacionados a esta proposta.

Arnold e Pistilli (2012) propuseram um sistema, denominado *Course Signals* (CS), que fornece um *feedback* significativo para o aluno, com base em modelos preditivos. A ideia desse sistema é utilizar a riqueza de dados encontrada em uma instituição de ensino, e, por meio de análises sobre os dados extraídos, determinar, em tempo real, quais os momentos em que esses alunos podem estar em risco de reprovação.

O algoritmo de predição utilizado realiza uma avaliação dos dados com base em quatro categorias: (i) desempenho, que é medido por um percentual sobre as notas obtidas; (ii) esforço, que é definido pela interação do estudante com o LMS, em comparação aos outros estudantes; (iii) histórico acadêmico; e (iv) características dos alunos, que são definidas pela idade, quantidade de créditos matriculados, dentre outras.

Após processar o perfil de cada estudante, o algoritmo exibe um, dentre três possíveis sinais, na página do perfil do aluno, que são: (i) luz vermelha, que indica uma alta probabilidade de insucesso no curso; (ii) luz amarela, que indica que o aluno pode ter problemas para obter sucesso no curso; e (iii) luz verde, que indica uma elevada probabilidade do aluno obter sucesso no curso.

Com base nisso, os tutores, que também têm acesso a essas informações, podem programar uma agenda de intervenções para os alunos que se encontram em níveis mais críticos, como, por exemplo, marcar reuniões presenciais ou encaminhar novos materiais de estudo.

Dimopoulos *et. al.* (2013) desenvolveram um *plugin* para o Moodle (versão 2.2+), denominado LAe-R tool, que é capaz de extrair uma série de competências dos alunos, como, por exemplo, participação em fóruns ou *chats*, interação com outros usuários, além de outros critérios, que podem ser definidos pelo professor, no intuito de ajudá-lo na identificação de elementos que possam estar associados à aprendizagem dos alunos.

Uma vez definidos esses critérios, o *plugin* retorna os dados dos alunos correspondentes às características definidas pelo professor e, em seguida, forma uma tabela de apresentação para este, onde as linhas correspondem aos critérios e as colunas correspondem aos respectivos níveis de desempenho dos alunos.

Charleer, Klerkx e Duval (2014) projetaram uma ferramenta que explora técnicas de visualização de informações educacionais, a fim de ajudar professores e estudantes a atingirem seus objetivos. Essa ferramenta projeta, para os seus usuários, os níveis de suas atividades em relação aos fatores que podem ajudá-los a atingirem seus objetivos. Por exemplo, caso um aluno esteja realizando poucos acessos ao LMS, a ferramenta permite que esse aluno possa verificar seus níveis de acessos ao sistema e perceber que, em comparação aos outros alunos, ele está bem aquém do esperado.

Hernández-García *et. al.* (2015) propuseram a utilização da LA para verificação de fatores de interações sociais, tais como uso de fóruns, respostas a mensagens, etc.; que podem influenciar o processo de aprendizado dos alunos. As interações sociais estudadas foram aquelas que aconteceram dentro do ambiente de gerenciamento dos cursos a distância.

O estudo foi realizado com dados reais, advindos de 10 turmas de Introdução à Informação Financeira, onde cada turma tinha mais de 60 alunos e com diferentes professores. Para a análise dos dados, foi utilizada a ferramenta Gephi 0.8.2., que é uma ferramenta projetada para análise de dados, cujo objetivo é ajudar o pesquisador no levantamento de hipóteses, na descoberta de padrões e na descoberta de singularidades entre dados (Gephi, 2015).

Já Fidalgo-Blanco *et. al.* (2015) propuseram a utilização da LA para analisar o progresso individual de um aluno, dentro do contexto do trabalho em equipe. Para que isso pudesse ser feito, foram coletados dados referentes à troca de mensagens entre estudantes no LMS, número de visualizações de mensagens, dentre outros. Além disso, o sistema identifica os estudantes autores de tópicos e o número de mensagens enviadas por estes, a fim de identificar quais seriam os estudantes ativos e passivos no grupo. O LMS utilizado foi o Moodle e, ao todo, foram analisados dados de 110 estudantes da Universidade Técnica de Madri. Para a análise dos dados, foi utilizado o coeficiente de correlação de Pearson.

Similarmente ao trabalho Dimopoulos *et. al.* (2013), Zielinski e Schmitt (2015) apresentaram um *plugin* para o Moodle, porém, neste caso, o *plugin* permite o acesso a visualizações do conteúdo acessado, das submissões e da participação da turma por meio de gráficos.

Vários são os fatores que diferem este trabalho dos demais. A Tabela 13 apresenta um comparativo entre os trabalhos relacionados e o trabalho proposto nesta dissertação, de acordo com os seguintes critérios:

I. Inovação: Este critério se refere ao grau inovador do trabalho, onde é verificado se o trabalho em questão realizou a criação de uma nova ferramenta de análise ou se fez uso de alguma ferramenta já existente, sendo:
   a. CNF a sigla correspondente a Criação de Nova Ferramenta;
   b. UFE a sigla correspondente a Uso de Ferramenta Existente.

II. Técnica de Análise Empregada: Este critério se refere às técnicas de análise de dados que foram empregadas no trabalho, sendo:
   a. AMS como sendo a sigla no que se refere às técnicas de Aprendizado de Máquina Supervisionado;
   b. AMNS como sendo a sigla no que se refere às técnicas de Aprendizado de Máquina Não Supervisionado;
   c. CCV como sendo a sigla no que se refere aos Coeficientes de Correlação entre Variáveis.
   d. NMM como sendo a sigla referente à técnica de normalização Min-Max;

III. Alvo da Análise: Este critério se refere ao objeto de estudo em questão, sendo:
   a. AL como sendo a sigla para Alunos;
   b. TD como sendo a sigla para os Tutores a Distância.

Trabalhos	Inovação		Técnica de Análise Empregada				Alvo da Análise	
	CNF	UFE	AMS	AMNS	CCV	NMM	AL	TD
Arnold e Pistilli (2012)	X		X				X	
Dimopoulos *et. al.* (2013)	X						X	
Charleer, Klerkx e Duval (2014)	X						X	
Hernández-García *et. al.* (2015)		X			X		X	
Fidalgo-Blanco *et. al.* (2015)	X				X		X	
Zielinski e Schmitt (2015)	X						X	
Ferramenta de Avaliação de Turmas e Tutores a Distância	X			X	X	X	X	X

Tabela 13. Quadro comparativo entre os trabalhos relacionados de acordo com os critérios de inovação, técnica de análise empregada e alvo da análise.

Neste cenário, muito embora a ferramenta apresentada possua algumas características existentes dos trabalhos encontrados na literatura, como a análise de dados de alunos e a aplicação dos coeficientes de correlação entre variáveis, esta se diferencia das ferramentas supracitadas por possuir as características de: i) uso de técnicas de aprendizado de máquina não supervisionado; ii) uso da técnica de normalização Min-Max; iii) a análise de dados dos tutores a distância.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste trabalho, procurou-se abordar a questão de como a participação do tutor pode afetar o desempenho dos alunos em cursos à distância. Ao adotar uma perspectiva de *Design Science*, temos tratado esta questão propondo uma ferramenta que permite investigar fatores comportamentais dos tutores que podem, ou não, estar associada com a participação efetiva dos alunos em turmas da modalidade de ensino a distância, além de permitir inferências de desempenho sobre ações comportamentais de tutores e alunos das turmas do ensino a distância.

A arquitetura do sistema tem componentes para extração, pré-processamento, processamento, análise e apresentação de dados. A contribuição principal desta arquitetura é uma combinação das técnicas:

i. Análise de correlação de variáveis, cujo objetivo é descobrir quais comportamentos dos tutores podem ou não impactar em determinados comportamentos da turma;
ii. Análise por meio da normalização de dados, cujo objetivo é permitir a comparação do desempenho entre turmas; e
iii. Análise por meio de técnicas de agrupamento, cujo objetivo é permitir inferências do desempenho comportamental dos tutores e das turmas.

Para a realização da análise, foram obtidos dados de 62 turmas do ensino a distancia, onde o *dataset* reuniu informações comportamentais de 2.227 alunos e 38 tutores a distância. Os dados analisados neste trabalho foram cedidos pela Secretaria de Educação a Distância (SEDIS), que é um órgão vinculado a Universidade Federal do Rio Grande do Norte (UFRN).

A técnica de análise de correlação de variáveis revelou correlações relevantes entre os tutores e os alunos. Dessa forma, estes resultados respondem a primeira pergunta central de pesquisa deste trabalho que é apresentada no Capítulo 3, cuja conclusão é:

- *Sim, os tutores a distância influenciam na participação efetiva dos alunos das turmas do ensino a distância*;

A lógica por trás da análise foi identificar quais indicadores do tutor correlacionam-se positivamente ou negativamente com as medições dos estudantes. Nosso objetivo imediato da pesquisa é fornecer uma visão sobre a forma como esta informação correlational vai permitir que os atores locais de educação (bem como estudiosos globais interessados) possam tomar

medidas no sentido de melhorar o desempenho dos alunos matriculados em cursos de ensino à distância.

Já a análise via *Clustering* permite uma inferência de desempenho comportamental das ações dos tutores e das turmas com base no histórico de outras turmas. O objetivo dessa análise é vialibilizar a atribuição de desempenho comportamental das ações dos tutores e da participação efetiva das turmas, de modo que: i) Seja possível avaliar em termos quantitativos a atuação dos tutores nos LMS; e ii) Seja possível avaliar em termos quantitativos a taxa de participação efetiva das turmas nesses ambientes de ensino.

Com os resultados obtidos por esta análise, respondemos então a segunda pergunta central desta pesquisa, cuja conclusão é:

- *Sim, é possível realizar inferências de desempenho comportamental dos tutores no que se refere a suas ações no LMS, bem como realizar inferências de desempenho comportamental das turmas do ensino a distância no que se refere à participação efetiva*;

É interessante notar que o estudo foi limitado à realidade educacional da região Nordeste do Brasil, mais especificamente para o estado do Rio Grande do Norte, com base em dados produzidos durante o período entre 2012 e 2013. A análise poderia, portanto, produzir resultados completamente diferentes para outros estados da mesma região do país, ou ainda para outros estados de outras regiões.

A partir deste ponto, esta pesquisa tem pelo menos duas direções distintas. Em primeiro lugar, pretendemos aumentar o nosso escopo de dados. Rio Grande do Norte é um estado relativamente pequeno em comparação com as dimensões do Brasil. Daí, uma extensão interessante de nossa pesquisa será integrar mais bases de dados de outros estados, por uma questão de compreender quais aspectos influenciam a aprendizagem dessa modalidade de ensino nessas regiões.

Em segundo lugar, visamos enriquecer a arquitetura do sistema com tecnologias de Web Semântica (por exemplo, ontologias) para ajudar a encontrar relações significativas entre os indicadores aparentemente disjuntos para aprender desempenho, além da inclusão de técnicas de análises de sentimentos e de interações sociais, a fim de extrair mais informações que posam impactar ou não o processo de ensino e aprendizagem que ocorre nessa modalidade de ensino.

Por fim, este trabalho também apresenta uma contribuição secundária, que é a exposição de conceitos da área de *Learning Analytics* no idioma em português, no qual, até o

momento da escrita deste trabalho, poucos são os artigos e trabalhos disponíveis que abordam esta área de estudo neste idioma, devido esta ainda ser um campo recente de investigação.

6.1 PRODUÇÕES CIENTÍFICAS

Os resultados obtidos ao longo deste trabalho foram submetidos e/ou aprovados nas seguintes conferências e revistas:

- **Título:** Um Ambiente Inteligente de Avaliação de Comportamentos de Tutores e Turmas no Ambiente Virtual de Aprendizagem Moodle;
  Autores: Rafael Castro de Souza, Francisco Milton Mendes Neto, Araken de Medeiros Santos, Laysa Mabel de Oliveira Fontes, Efraim N. H. D. Rodrigues e Ricardo Alexsandro de Medeiros Valentim;
  Veículo: Anais dos Workshops do CBIE 2016 - Workshop sobre Avaliação e Acompanhamento da Aprendizagem em Ambientes Virtuais (WAvalia 2016);
  Resultado: Aceito.
- **Título:** Investigação acerca do Impacto dos Comportamentos dos Tutores em Turmas de Ensino a Distância;
  Autores: Rafael Castro de Souza, Francisco Milton Mendes Neto, Araken de Medeiros Santos, Laysa Mabel de Oliveira Fontes e Ricardo Alexsandro de Medeiros Valentim;
  Veículo: Anais dos Workshops do CBIE 2016 - Workshop sobre Mineração de Dados Educacionais 2016 (WMDE 2016);
  Resultado: Aceito.
- **Título:** Evaluating Online Tutoring Behavior with Learning Analytics;
  Autores: Rafael Castro de Souza, Francisco Milton Mendes Neto, Araken de Medeiros Santos, Laysa Mabel de Oliveira Fontes, Ricardo Alexsandro de Medeiros Valentim e Patrício de Alencar Silva;
  Veículo: Computers in Human Behavior;
  Resultado: Submetido.

REFERÊNCIAS BIBLIOGRÁFICAS

Arnold, K.E., Pistilli, M.D. Course signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics & Knowledge, 2012. ACM Retrieved Oct 15, 2015, from:< https://www.itap.purdue.edu/learning/docs/research/Arnold_PistilliPurdue_University_Course_Signals-2012.pdf>.

Associação Brasileira de Educação a Distância (ABED). Disponível em < http://www.abed.org.br/censoead2014/CensoEAD2014_portugues.pdf> Acesso em: 15 de março de 2016.

Academics Analytics. Academics Analytics Benchmarking for academic excellence**.** Disponível em <http://www.academicanalytics.com/Public/About> Acesso em: 10 de junho de 2015.

ATP. *Applied Predictive Technologies Receives $100 Million Investment from Goldman Sachs*. Disponível em <http://www.predictivetechnologies.com/newsroom/press-releases/2013/applied-predictive-technologies-receives-$100-million-investment-from-goldman-sachs.aspx> Acesso em: 11 de abril de 2016.

Berking, P., Gallagher, S. Choosing a learning management system. *Advanced Distributed Learning (ADL) Co-Laboratories,(2.4)*, 2011.

Campagni R., Merlini D., Srupnoli R., Verri M. C. Data mining models for student careers. Expert Systems with Applications, v. 42, n. 13, p. 5508-5521, 2015.

Charleer, S., Klerkx, J., Duval, E. Learning Dashboards. Journal of Learning Analytics, 1(3), 199-202, 2014.

Chatti M. A., Dyckhoff A. L., Schroeder U., Thus H. A Reference Model for Learning Analytics. International Journal of Technology Enhanced Learning, v. 4, n. 5, p. 318-331, 2012.

Chen, H., Chiang, R. H., Storey, V. C.. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, vol. *36*, n. 4, p. 1165-1188, 2012.

Dasgupta, S. Performance guarantees for hierarchical clustering. In Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT ’02), Sydney, Australia, July 8-10, pp. 351-363, 2002.

Dilla W. N., Raschke R. L. Data visualization for fraud detection: Practice implications and a call for future research. International Journal of Accounting Information Systems, v. 22, p. 1-22, 2015.

Dimopoulos, I., Petropoulou, O., Boloudakis, M., Retalis, S. Using Learning Analytics in Moodle for assessing students’ performance. In 2 nd Moodle Research Conference, 2013.

Downes, S. Feature: E-learning 2.0. *Elearn magazine*, vol. *1*, 2005.

Education Week. Billion-Dollar Deal Heats Up Ed-Tech Market. Disponível em<http://www.edweek.org/ew/articles/2014/03/26/26acquisition.h33.html> Acesso em: 11 de abril de 2016.

Examulator. Disponível em < http://www.examulator.com/er/> Acesso em: 5 de maio de 2016.

Faceli K., Lorena A. C., Gama J., de Carvalho A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 2011.

Faria A. A., Salvadori A. Educação a Distância e seu movimento histórico no Brasil. Revistas das Faculdades Santa Cruz, Curitiba, v. 8, n. 1, 2010.

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., Conde, M. Á. Using Learning Analytics to improve teamwork assessment.Computers in Human Behavior, 47, 149-156, 2015.

Ferguson, R., Shum, S. B. Social Learning Analytics: Five Approaches. ACM. 2012.

Garrison, D. R. E-learning in the 21st century: A framework for research and practice. Taylor & Francis, 2011.

Greller, W., Drachsler. H. Translating Learning into Numbers: a generic framework for learning analytics. Education Technology & Society, 2012.

GIGAOM. *As ed tech heats up, Desire2Learn raises $80M in its first VC round* . Disponível em<https://gigaom.com/2012/09/04/as-ed-tech-heats-up-desire2learn-raises-80m-in-its-first-vc-round/> Acesso em: 11 de abril de 2016.

Goldschmidt R., Bezerra E. Data mining : conceitos, técnicas, algoritmos, orientações e aplicações. Elsevier Brasil, 2015.

Hernández-García, Á., González-González, I., Jiménez-Zarco, A. I., Chaparro-Peláez, J. Applying social learning analytics to message boards in online distance learning: A case study. Computers in Human Behavior, 47, 68-80, 2015.

Hevner, A., March, S., Park, J., Ram, S. Design Science in Information Systems Research. *MIS Quarterly* 28 (1), 75-105, 2004.

Hochbaum, D. S., and Shmoys, D.B. A best possible heuristic for the k-center problem. Mathematics of Operations Research, vol. 10, no. 2, pp. 180-184, 1985.

Hunt, M., Davies, S., e Pittard, V. Harnessing technology review 2007. Progress and impact of technology in education. Full Report, 2007.

Jacob, S. M., Issac, B. The mobile devices and its mobile learning usage analysis. *arXiv preprint arXiv:1410.4375*, 2014).

Johnson, L., Smith, R., Willis, H., Levine, A., Haywood, K. The 2011 Horizon Report. The New Media Consortium, 2011.

Kats, Y. Learning Management Systems and Instructional Design: Best Practices in Online Education. Idea Group Inc (IGI), 2013.

Keller, G. Statistics for Management and Economics. Cengage Learning, 2011.

LAK 2013: Third International Conference on Learning Analytics and Knowledge 8 – 12 April 2013, Leuven, Belgium. ACM ISBN: 978-1-4503-1785-6.

LAK 2014: Fourth International Conference on Learning Analytics And Knowledge 24—28 March 2014, Indianapolis, IN, USA ACM ISBN: 1-59593-036-1.

LAK 2015: Fifth International Conference on Learning Analytics And Knowledge 21—20 March 2015, Poughkeepsie, NY, USA ACM ISBN: 1-59593-036-1.

LeBlanc, D. C. Statistics: Concepts and Applications for Science, Volume 2. Editora: Jones & Bartlett Learning, 2004.

Lias, T. E., Elias, T. Learning Analytics: The Definitions, the Processes, and the Potential, (2011).

Litto F. M., Formiga M. Educação a Distância: o estado da arte. São Paulo: Person Education do Brasil. 2 ed. 2012.

Macfadyen L. P., Dawson S. Mining LMS Data to Develop an 'Early Warning System for Educators: A Proof of Concept'. Computers & Education, v. 54, n. 2, p. 588-599, 2010.

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, 1967.

Mayer-Schönberger, V., Cukier, K. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

Meissner, G. Correlation Risk Modeling and Management: An Applied Guide including the Basel III Correlation Framework - With Interactive Models in Excel / VBA. John Wiley & Sons, 2013.

McAFee A., Brynjolfsson E. Big Data: The Management Revolution. Harvard Business Review, 2012.

McCaffrey J. Data Clustering - Detecting Abnormal Data Using k-Means Clustering. Microsoft Magazine. Disponível em <https://msdn.microsoft.com/en-us/magazine/jj891054.aspx> Acesso em: 28 de maio de 2016.

Mitchell, T. M. Machine Learning. McGraw-Hill, Inc., 1 ed.,1997.

Moodle. Disponível em<https://docs.moodle.org/> Acesso em: 11 de abril de 2016.

Moodle b. Disponível em <https://docs.moodle.org/dev/Database_schema_introduction#Question_types> Acesso em 5 de maio de 2016.

Motiwalla, L. F. Mobile learning: A framework and evaluation.*Computers & Education*, vol. *49*, n. 3, p. 581-596, 2007.

NetBeans. An Introduction to NetBeans. Disponível em <https://netbeans.org/about/> Acesso em 5 de maio de 2016.

Neto, F. M. M. (Ed.). *Technology Platform Innovations and Forthcoming Trends in Ubiquitous Learning*. IGI Global, 2013.

Neto, F. M. M., de Carvalho Muniz, R., Burlamaqui, A. M. F., de Souza, R. C. An Agent-Based Approach for Delivering Educational Contents Through Interactive Digital TV in the Context of T-Learning. *International Journal of Distance Education Technologies (IJDET)*, vol. *13*, n. 2, 73-92, 2015.

Oracle. Learn about Java Technologies. Disponível em <https://java.com/en/about/> Acesso em 5 de maio de 2016.

Oracle b. JavaServer Faces Technologies. Disponível em< http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html> Acesso em 5 de maio de 2016.

O'Rourke, N., Hatcher L. , Stepanski. E. J. A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics SAS Institute, 2005.

Pessoa A. S. A., De Lima G. R. T., Da Silva J. D. S., Stephany S., Strauss C., Caetano M., Ferreira N. S. Mineração de dados meteorológicos previsão de eventos severos. Revista Brasileira de Meteorologia, v. 27, n. 1, p. 61-74, 2012.

PostgreSQL. Disponível em <http://www.postgresql.org/about/> Acesso em 5 de maio de 2016.

PrimeFaces. Disponível em <http://www.primefaces.org/whyprimefaces> Acesso em 5 de maio de 2016.

Proceedings of the 1st International Conference on Learning Analytics and Knowledge. February 27–March 1, 2011, Banff, Alberta, Canada ACM ISBN: 978-1-4503-1057-4/11/02.

Proceedings of the 2nd International Conference on Learning Analytics and Knowledge April 29–May 2, 2012, Vancouver, British Columbia, Canada ACM ISBN: 978-1-4503-1111-3/12/04.

Quilici-Gonzalez J. A., Zampirolli F. A. Sistemas Inteligentes e Mineração de Dados. Triunfal Gráfica e Editora, 2014.

Re/code. *Google Capital Invests $40M in Renaissance Learning, Valued at $1B* Disponível em <http://recode.net/2014/02/19/google-capital-invests-40m-in-renaissance-learning-valued-at-1b/> Acesso em: 11 de abril de 2016.

Rosenberg, M. J. E-learning: Strategies for delivering knowledge in the digital age (Vol. 3). New York: McGraw-Hill, 2001.

Rubin, A. Statistics for Evidence-Based Practice and Evaluation. Cengage Learning, 2012.

Sayão L. F., Marcondes C. H. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais. TransInformação, Campinas, 2008.

Schlotzhauer, S. D. Elementary Statistics Using JMP. SAS Institute, 2007.

Schumacker, R. E. Learning Statistics Using R. SAGE Publications, 2014.

Shah B. R., Lipscombe L. L. Clinical Diabetes Research Using Data Mining: A Canadian Perspective. Canadian Journal of Diabetes, v. 39, n. 3, p. 235-238, 2015.

Sharma, A.K. Text Book Of Correlations And Regression. Discovery Publishing House, 2005.

Sharma, J.K. Business Statistics. Pearson Education India, 2012.

Sommerville, I. Engenharia de Software. 8ª Edição. São Paulo: Pearson Addison-Wesley. 2007.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., Baker, R. S. J. D.. Open Learning Analytics: an integrated & modularized platform. *Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques*, 2011.

Siemens, G., Long, P.. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review*, vol. *46*, n. 5, p. 31-40, 2011.

Souza R. C., Mendes Neto F. M. Construção de um Repositório de Recursos Educacionais Abertos Baseado em Serviços Web para Apoiar Ambientes Virtuais de Aprendizagem. Revista Novas Tecnologias na Educação, v. 12, n. 2, 2014.

The Globe and Mail. *Desire2Learn finally accepts venture investment.* Disponível em<http://www.theglobeandmail.com/globe-investor/desire2learn-finally-accepts-venture-investment/article4516924/> Acesso em: 11 de abril de 2016.

Thepade S. D., Kalbhor M. M. Novel Data Mining based Image Classification with Bayes, Tree, Rule, Lazy and Function Classifiers using Fractional Row Mean of Cosine, Sine and Walsh Column Transformed Images. In: International Conference on Communication, Information & Computing Technology, 2015.

Vadeyar, D. A., Yogish, H. K. Farthest First Clustering in Links Reorganization. Int. J. Web Semantic Technol, 5(3), 2014.

W3Schools. SQL Introduction. Disponível em <http://www.w3schools.com/sql/sql_intro.asp> Acesso em 5 de meio de 2016.

Warner, R. M. Applied Statistics: From Bivariate Through Multivariate Techniques, SAGE, 2012.

Weinberg, S. L., Abramowitz, S. K. Data Analysis for the Behavioral Sciences Using SPSS. Cambridge University Press, 2002.

Weka. Disponível <http://www.cs.waikato.ac.nz/ml/weka/> Acesso 5 de maio de 2016.

Weka b. Disponível em <http://weka.wikispaces.com/ARFF+%28book+version%29> Acesso em 11 de maio de 2016.

Wieringa, R. J. Design Science Methodology for Information Systems and Software Engineering. Springer, 2014.

Wu J. Advances in K-means Clustering: A Data Mining Thinking. Springer Science & Business Media, 2012.

Zielinski, F. D. C., Schmitt, M. A. R. Uma ferramenta gráfica para suporte à atividade docente no Moodle. RENOTE, 13(1), 2015.