



Pós-Graduação em Ciência da Computação

Rodrigo Lins Rodrigues

**Uma abordagem de Mineração de Dados Educacionais para
previsão de desempenho a partir de padrões comportamentais
de Autorregulação da Aprendizagem**

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2016

Rodrigo Lins Rodrigues

**Uma abordagem de Mineração de Dados Educacionais para
Previsão de Desempenho a partir de Padrões Comportamentais
de Autorregulação da Aprendizagem**

Tese de Doutorado apresentada ao Programa
de Pós-Graduação em Ciência da Computação
da Universidade Federal de Pernambuco, como
requisito parcial para a obtenção do título de
Doutor em Ciência da Computação.

Orientador: Alex Sandro Gomes

Coorientador: Paulo Jorge Leitão Adeodato

RECIFE
2016

Catalogação na fonte

Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

R696a Rodrigues, Rodrigo Lins

Uma abordagem de mineração de dados educacionais para previsão de desempenho a partir de padrões comportamentais de autorregulação da aprendizagem / Rodrigo Lins Rodrigues. – 2016.

230 f.: il., fig., tab.

Orientador: Alex Sandro Gomes.

Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.

Inclui referências, apêndices e anexos.

1. Mineração de dados. 2. Educação à distância. I. Gomes, Alex Sandro (orientador). II. Título.

006.312

CDD (23. ed.)

UFPE- MEI 2017-30

Rodrigo Lins Rodrigues

Uma abordagem de Mineração de Dados Educacionais para Previsão de Desempenho a partir de Padrões Comportamentais de Autorregulação da Aprendizagem

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovada em 21/12/2016

Orientador: Prof. Dr. Alex Sandro Gomes

BANCA EXAMINADORA

Prof. Fernando da Fonseca de Souza
Centro de Informática / UFPE

Prof. Dr. Giordano Ribeiro Eulalio Cabral
Centro de Informática / UFPE

Prof. Dr. Marcelo Brito Carneiro Leão
Departamento de Química / UFRPE

Prof. Dr. Francisco Milton Mendes Neto
Departamento de Ciências Exatas e Naturais / UFERSA

Profa. Dra. Filomena Maria Gonçalves da Silva Cordeiro Moita
Departamento de Educação / UEPB

*Dedico este trabalho aos meus pais:
Jairo Rodrigues da Silva e Socorro Lins Rodrigues
A minha irmã Ravenna Lins Rodrigues
E a minha esposa Simone Aparecida da Silva Lins.*

Agradecimentos

Este trabalho teve a participação e colaboração de várias pessoas e entidades. Todas elas contribuíram de alguma forma para a construção desta Tese. Sinto a necessidade de agradecer às pessoas que direta ou indiretamente contribuíram para sua realização:

Agradeço a banca avaliadora desta tese: Dr. Fernando da Fonseca de Souza, Dra. Filomena Maria G. da S. C. Moita, Dr. Francisco Milton Mendes Neto, Dr. Giordano Ribeiro Eulalio Cabral e o Dr. Marcelo Brito Carneiro Leão.

Aos professores da UFPE, em particular aos do Centro de Informática, pelos conhecimentos compartilhados. Em especial, ao orientador e amigo Alex Sandro Gomes que com muita maestria soube guiar o desenvolvimento deste trabalho, ajudando e apoiando nos momentos necessários.

Com carinho, a amiga e orientadora de iniciação científica, desde tempos de graduação, Filomena Maria G. da S. C. Moita que me apresentou os caminhos da pesquisa científica.

A minha esposa Simone Aparecida da Silva Lins que foi uma verdadeira companheira, e além de compreender todo o esforço, apoiou e torceu pelo sucesso do trabalho.

Aos meus familiares, em especial aos meus pais Jairo Rodrigues da Silva e Socorro Lins Rodrigues e minha irmã Ravenna Lins Rodrigues, que mesmo distantes fisicamente foram presentes durante todo o tempo demonstrando o carinho, o amor e a atenção de sempre. Agradeço muito ao esforço dos meus pais por tornarem o sonho do doutorado uma realização possível, incentivando e contribuindo durante toda minha formação acadêmica.

Aos meus amigos de trabalho do Departamento de Educação da Universidade Federal Rural de Pernambuco, em especial a Diretora Monica Folena e a área II de Métodos e Técnicas de Ensino.

Aos amigos parceiros de pesquisa da Universidade de Pernambuco: Alexandre Fonseca (NEAD) e Alexandre Magno Maciel (POLI).

Agradeço, em particular, aos meus amigos de grupo de pesquisa CCTE, Luma Seixas, Ivanildo Melo, Hugo Vieira, Josy Brito, Ana Luiza Rolim, Rosângela Carvalho, Kelsen Oliveira, Vânia e em especial aos amigos Jorge Cavalcanti

Ramos e João Sedraz pelos inúmeros dias e madrugadas de trabalho em uma parceria frutífera de pesquisa.

Ao Núcleo de Educação a Distância da UPE pelo fornecimento da base de dados histórica dos cursos de educação à distância.

Ao órgão de fomento FACEPE que por meio do projeto de financiamento PRONEM/FACEPE/CNPq 08/2014, forneceu recursos necessários para o desenvolvimento da pesquisa.

E a todos aqueles que, direto ou indiretamente, colaboraram para que este trabalho chegasse a atingir os objetivos propostos.

*“Paciência e perseverança tem o efeito mágico
de fazer as dificuldades desaparecerem
e os obstáculos sumirem.”*
(John Quincy Adams)

Resumo

O uso crescente dos sistemas de LMS, do inglês *Learning Management System*, tem gerado um crescente volume de dados provenientes de interações entre professores e alunos. Esse volume de dados, se devidamente explorado, pode fornecer o entendimento de como os alunos autorregulam-se no desenvolver da sua aprendizagem e as relações destas habilidades com o desempenho acadêmico. Diante deste contexto, delimitamos o objetivo de pesquisa desta tese como sendo a modelagem preditiva do desempenho de estudantes por meio de indicadores comportamentais de autorregulação da aprendizagem em ambientes de LMS. O método utilizado nesta pesquisa foi baseado na metodologia CRISP-DM, percorrendo pelas fases de Entendimento do domínio em educação, Entendimento dos dados educacionais, Preparação dos dados, Modelagem, Avaliação do modelo e Implementação da solução em formato de software de visualização de dados. O processo de modelagem foi realizado por meio de dados históricos de uma instituição de ensino superior, armazenados durante um período de sete anos. Como resultado da fase de modelagem, foi realizado o desenvolvimento, avaliação e comparativo entre quatro modelos de previsão, por meio dos algoritmos de Árvore de Decisão (CART), Regressão Logística, SVM e Naive Bayes. O modelo de regressão logística apresentou-se como sendo o mais satisfatório, sendo capaz de predizer o desempenho acadêmico dos estudantes, de acordo com variáveis de autorregulação da aprendizagem, com uma taxa de acurácia de 0,893 e área sobre a curva ROC de 0,9574. Por fim, foi realizada a concepção, implementação e avaliação de uma solução de software, para visualização de dados, capaz de fornecer indícios de previsibilidade do desempenho do aluno a partir do seu registro comportamental armazenado na plataforma de LMS.

Palavras-chave: Mineração de Dados Educacionais. Autorregulação da Aprendizagem. Regressão Logística. Educação a Distância.

Abstract

The growing use of Learning Management System has generated an increasing volume of data from interactions between teachers and students. This volume of data, if properly explored, can provide an understanding of how students self-regulate themselves in the development of their learning and the relationships of these skills to academic performance. Given this context, we delimit the research objective of this thesis as the predictive modeling of student performance through behavioral indicators of self-regulation of learning in LMS environments. The method used in this research was based on the CRISP-DM methodology, through the Understanding phases of the domain in education, Understanding of educational data, Data preparation, Modeling, Evaluation of the model and Implementation of the solution in data visualization software format. The modeling process was performed through historical data from a higher education institution, stored over a period of seven years. As a result of the modeling phase, the development, evaluation, and comparison between four prediction models were performed through the Decision Tree (CART), Logistic Regression, SVM and Naive Bayes algorithms. The logistic regression model was the most satisfactory, being able to predict students' academic performance, according to self-regulated learning variables, with an accuracy rate of 0.893 and area under the ROC curve of 0.9574. Finally, the design, implementation, and evaluation of a software solution for data visualization were carried out, able to provide evidence of predictability of student performance from the behavioral register stored in the LMS platform.

Keywords: Educational Data Mining. Self-Regulated Learning. Learning Management System. Logistic Regression. Distance Education.

Lista de Figuras

Figura 1: Ciclo de autorregulação da aprendizagem	30
Figura 2: Principais áreas relacionadas a EDM.....	41
Figura 3: Modelo de Base de Classificação	45
Figura 4: Exemplo de Fluxograma Árvore de Decisão	46
Figura 5: Exemplo de representação gráfica da curva logística	49
Figura 6: Exemplo de um classificador SVM no hiperplano	50
Figura 7: Área abaixo das curvas ROC A e B	53
Figura 8: Exemplo de Dendrograma	56
Figura 9: Quantidade de publicações na área nos últimos dez anos	62
Figura 10: Principais periódicos com publicações na área nos últimos dez anos	63
Figura 11: Quantidade de publicações na área nos últimos seis anos.....	64
Figura 12: Principais periódicos com publicações na área nos últimos seis anos....	64
Figura 13: Tipo de pesquisa em SRL	66
Figura 14: Forma de coleta de dados.....	67
Figura 15: Técnica utilizada para mensuração de SRL.....	68
Figura 16: Área de pesquisa declarada na análise dos dados	70
Figura 17: Fases do modelo de referência CRISP-EDM	72
Figura 18: Fluxo do entendimento do domínio	74
Figura 19: Fluxo do entendimento dos dados	75
Figura 20: Fluxo para a construção da tabela de análise	76
Figura 21: Fluxo do processo de modelagem	78
Figura 22: Fluxo da arquitetura	84
Figura 23: Fluxo do processo de avaliação do software	85
Figura 24: Resultado da aplicação do questionário OSLQ.....	93
Figura 25: Bases de dados do curso de Administração	97
Figura 26: Bases de dados do curso de Pedagogia	97
Figura 27: Bases de dados do curso de Biologia	98
Figura 28: Bases de dados do curso de Letras	98
Figura 29: Gráfico de comparação da categorização da variável alvo	102
Figura 30: Gráfico comparativo de algoritmos de classificação.....	104
Figura 31: Curva ROC para comparativo entre classificadores.....	105

Figura 32: Curva ROC do algoritmo de Regressão Logística	111
Figura 33: Módulo Análise de Desempenho dos dados da aba Geral	112
Figura 34: Módulo Análise de Desempenho dos dados da aba Indicadores.....	113
Figura 35: Módulo Análise de Desempenho dos dados da aba Alunos	114
Figura 36: Arquitetura da solução de software desenvolvida	115
Figura 37: Interface do módulo de Visão geral dos dados	116
Figura 38: Módulo de Avaliação de Desempenho Gráfico Geral.....	117
Figura 39: Fluxo da geração dos percentuais de classificação	117
Figura 40: Módulo de Avaliação de Desempenho gráfico Indicadores.....	118
Figura 41: Módulo de Avaliação de Desempenho gráfico Alunos	118
Figura 42: Gravação da interação dos usuários na utilização da aplicação	121
Figura 43: Avaliação com professores da UPE	121
Figura 44: Avaliação com professores da UFPE e UNIVASF	122
Figura 45: Avaliação com tutores da UFRPE	122
Figura 46: Valores médios das dimensões e retângulos de confiança.....	125
Figura 47: Valores médios das dimensões	126
Figura 48: Comparativo de médias entre Expectativa e Experiência	127

Lista de Quadros

Quadro 1: Fases e áreas do modelo de SRL	30
Quadro 2: Abordagens em EDM	42
Quadro 3: Mensuração de tarefas de DM utilizadas em EDM	43
Quadro 4: Matriz de confusão para um problema binário	50
Quadro 5: Primeira String de Busca	62
Quadro 6: Segunda String de Busca	63
Quadro 7: Pacotes utilizados no processo de modelagem	79
Quadro 8: Questionário sobre User Experience	86
Quadro 9: Descrição das tarefas para o teste de usabilidade	88
Quadro 10: Variáveis do questionário de autorregulação da aprendizagem	91
Quadro 11: Variáveis de autorregulação da aprendizagem extraídas do Moodle ...	95
Quadro 12: Variáveis de desempenho acadêmico	100
Quadro 13: Resultados de significância estatística por variável	107
Quadro 14: Resultados de significância estatística por variável após o stepwise...	108
Quadro 15: Matriz de confusão para o algoritmo de Regressão Logística	110
Quadro 16: Identificação de problemas de usabilidade	123
Quadro 17: Médias e valores do testes de Wilcoxon	129
Quadro 18: Relatos dos professores sobre a utilização da ferramenta	138

Lista de Tabelas

Tabela 1: Variáveis de interação coletadas	93
Tabela 2: Quantitativo geral de alunos por curso	98
Tabela 3: Comparação entre algoritmos de classificação	103
Tabela 4: Matriz de confusão com percentuais	136

Lista de Abreviaturas e Siglas

AVA:	Ambiente Virtual De Aprendizagem
BD:	Banco De Dados
CSV:	Comma-Separated Values
CRISP-DM:	Cross Industry Standard Process for Data Mining
DM:	Data Mining
EDM:	Educational Data Mining
EAD:	Educação à Distância
ETL:	Extract Transform Load
IA:	Inteligência Artificial
LA:	Learning Analytics
LMS:	Learning Management Systems
LASSI:	Learning and Study Strategies Inventory
MD:	Mineração de Dados
MOODLE:	Modular Object-Oriented Dynamic Learning Environment
MSLQ:	Motivated Strategies for Learning Questionnaire
NEAD:	Núcleo De Educação a Distância
OLAP:	Online Analytical Process
ROC:	Receiver operating characteristic
SNA	Social Network Analysis
SRLIS:	Self-Regulated Learning Interview Scale
SQL:	Structured Query Language
SRL:	Self-Regulated Learning
SVM:	Máquina de Vetor de Suporte
TI:	Tecnologia da Informação
TIC:	Tecnologias da Informação e da Comunicação
UPE:	Universidade de Pernambuco
UX:	User Experience

Sumário

1. INTRODUÇÃO	18
1.1 Justificativa	18
1.2 Contexto e Motivação	20
1.3 Questões de pesquisa	21
1.4 Objetivos	21
1.4.1 Geral	22
1.4.2 Específicos.....	23
1.5 Método de Pesquisa	23
1.6 Organização e estrutura da tese.....	24
2. AUTORREGULAÇÃO DA APRENDIZAGEM.....	26
2.1 Conceito de autorregulação da aprendizagem	26
2.2 Modelos teóricos de autorregulação da aprendizagem	28
2.2.1 O modelo de Zimmerman.....	29
2.2.2 O modelo de Pintrich.....	30
2.3 Características da autorregulação em ambientes de EAD	32
2.4 Relação entre autorregulação e desempenho acadêmico.....	33
2.5 Abordagens de mensuração da autorregulação	34
2.5.1 Instrumentos de coleta	35
2.5.2 Abordagens de mensurações	36
2.6 Considerações sobre o capítulo	38
3. MINERAÇÃO DE DADOS EDUCACIONAIS (EDM)	40
3.1 Contextualização sobre Mineração de Dados Educacionais	40
3.2 Áreas relacionadas a EDM	41
3.4 Tarefas de Mineração de Dados em contextos Educacionais	43
3.4.1 Classificação	44
3.4.2 Agrupamento.....	54
3.4.3 Regressão Linear.....	57
3.4.4 Regras de associação.....	58
3.5 Considerações sobre o capítulo	58
4. MAPEAMENTO SISTEMÁTICO DA LITERATURA	60
4.1 Processo do Mapeamento Sistemático.....	60

4.1.1 Objetivo do mapeamento	60
4.1.2 Questões de pesquisa.....	61
4.1.3 Critérios de Inclusão e Exclusão	61
4.1.4 Estratégia de busca	62
4.1.5 Análise e Interpretação dos Resultados.....	65
4.2 Considerações sobre o capítulo	70
5. MÉTODO DA PESQUISA	72
5.1 Fases do Processo	72
5.2 Entendimento do Domínio	73
5.3 Entendimento dos dados	74
5.4 Preparação dos dados.....	76
5.4.1 Seleção dos dados.....	76
5.4.2 Categorização de Variáveis	77
5.5 Modelagem	78
5.5.1 Algoritmos de Classificação	79
5.6 Avaliação dos Modelos	80
5.6.1 Avaliação da Precisão de Classificação.....	81
5.7 Implementação da solução	82
5.7.1 Prototipagem.....	82
5.7.2 Arquitetura e Implementação	83
5.7.3 Avaliação da solução de software desenvolvida	84
5.8 Considerações sobre o capítulo	89
6. RESULTADOS	91
6.1 Resultados sobre o entendimento do domínio.....	91
6.1.1 Público alvo e Amostra.....	92
6.2 Resultados sobre o entendimento dos dados	94
6.2.1 Identificação das variáveis significativas	94
6.2.2 Extração das bases de dados históricas da plataforma Moodle	96
6.3 Resultados sobre a preparação dos dados	99
6.3.1 Transformação e limpeza das variáveis de autorregulação	99
6.3.2 Extração das variáveis relacionadas ao desempenho dos estudantes	99
6.4 Resultados da Modelagem e Validação.....	101
6.4.1 Construção dos modelos de classificação	101
6.4.2 Validação dos modelos de classificação	103

6.4.3 Modelo de Regressão Logística	106
6.5 Resultados do desenvolvimento e implementação	112
6.5.1 Protótipos de baixa fidelidade	112
6.5.2 Arquitetura e Implementação	114
6.5.3 Experimento de validação com professores.....	120
6.7 Considerações sobre o capítulo	130
7. DISCUSSÕES.....	132
7.1 Lacunas de pesquisa identificadas	132
7.2 Capacidade preditiva de desempenho acadêmico	133
7.2.1 Comparativo entre modelos de classificação	134
7.2.2 Significância e poder de explicação das variáveis	134
7.2.3 Previsão do desempenho dos alunos	136
7.3 Solução de software desenvolvida.....	137
7.3.1 Percepção dos professores.....	138
8. CONSIDERAÇÕES FINAIS	140
8.1 Conclusão.....	140
8.2 Dificuldades e Limitações	140
8.3 Contribuições	141
8.4 Trabalhos Futuros.....	142
8.5 Produção Científica Desenvolvida	142
Referências	147
Apêndices.....	165
Anexos	218

1. INTRODUÇÃO

Este capítulo apresenta uma visão geral desta tese. São explicitados os argumentos que justificam o desenvolvimento do trabalho, o contexto e motivação, os objetivos, o método de pesquisa e a organização e estruturação da tese.

1.1 Justificativa

Com a evolução no desenvolvimento de artefatos tecnológicos para a educação, os ambientes computacionais de ensino e os conteúdos digitais passaram a ser vistos como tecnologias essenciais na condução dos processos educacionais. Esse aumento no desenvolvimento e utilização de plataformas de aprendizagem online tem impactado diretamente o aumento dos cursos na modalidade a distância (WANG, DOLL, *et al.*, 2013). Este crescimento tem possibilitado a geração de alta carga de dados, provindas de interações dos usuários, permitindo a geração de grandes bases de dados com informação sobre o comportamento dos estudantes. Estes dados, se devidamente explorados, tem potencial para fornecer importantes indícios de como os alunos interagem e aprendem. (KOEDINGER, BAKER, *et al.*, 2010).

Acompanhar essas interações, nas diferentes ferramentas de comunicação e plataformas de aprendizagem, é um dos desafios mais intensos e exaustivos que os professores da modalidade de Educação a Distância enfrentam atualmente. Turmas cada vez maiores e heterogêneas e o tempo limitado imposto aos professores, induzem a adotar uma postura passiva, apenas respondendo mensagens e corrigindo atividades. Perceber as individualidades de cada aluno, ou até mesmo intervir no seu processo de aprendizagem, sem contatos face-a-face, torna-se, portanto, uma tarefa desafiadora.

Por outro lado, os estudantes nesta modalidade se deparam com a flexibilidade de tempo, local e recursos dos ambientes virtuais de aprendizagem, onde é exigido um maior nível de responsabilidade e autonomia. O aluno passa a ter grande controle sobre o seu próprio processo de aprendizagem (PINTRICH, 1999). Este controle faz parte de habilidades socioemocionais.

Conforme o relatório publicado pela UNESCO (2015), um dos aspectos que devem ser estimulados no aluno do Século XXI é o desenvolvimento de competências socioemocionais.

Dentro deste contexto, estudos vêm sendo desenvolvidos e autores como Chen e Paul (2003); Kizilcec, Pérez e Maldonado (2016) afirmam que essas habilidades socioemocionais interferem na capacidade de aprender. Esses estudos buscam identificar habilidades individuais que afetam os resultados da aprendizagem de acordo com o perfil comportamental de cada estudante nas plataformas.

Problemas de pesquisa como esses são abordados por uma área de pesquisa que busca entender o fenômeno da autorregulação da aprendizagem. Constitui uma área de estudo dentro da psicologia cognitiva, da qual os aspectos estudados são, as estratégias de gestão de recursos, gestão do tempo, do esforço, do ambiente, da interação e da busca por ajuda (CHEN, 2002; PINTRICH, 1999; ZIMMERMAN e MARTINEZ-PONS, 1988).

É necessário considerar que existem diferenças significativas nas habilidades de cada aluno em estabelecer estratégias de autorregulação da aprendizagem. Estudantes que detém essas habilidades conseguem controlar o ritmo, o direcionamento do processo de aprendizagem e gerenciar um conjunto de características do ambiente de aprendizagem, realizando escolhas que podem auxiliar o desenvolvimento de sua estrutura cognitiva (BROADBENT e POON, 2015) .

Por outro lado, muitos estudantes têm dificuldades com as habilidades de autorregulação e o automonitoramento diante de tarefas rotineiras de aprendizagem. Tais alunos, na maioria dos casos, foram formados por meio de uma dependência regulatória por parte de professores e gestores que o modelo tradicional normalmente provê. Muitos desses alunos, familiarizados com o ensino presencial, dependem de um acompanhamento personalizado para gerenciar atividades em tempos flexíveis, com a supervisão direta dos professores.

Diante dessa realidade, um dos problemas ainda em aberto, que justifica o tema desta tese, é o desafio de identificar a relação de influência dos fatores de

autorregulação da aprendizagem no desempenho de alunos pertencentes à modalidade de educação a distância. Como desdobramento deste problema de pesquisa temos a seguinte indagação: *As diferenças individuais de padrões comportamentais de autorregulação da aprendizagem influenciam o desempenho de alunos na modalidade de EAD?*

Perguntas como esta podem ser respondidas por meio de uma área de pesquisa intitulada como Mineração de Dados Educacionais (MDE, do inglês, *Educational Data Mining*, ou EDM). É uma área recente, ainda emergente e pouco explorada, mas com bastante potencial para aplicações educacionais, possibilitando a descoberta de informações que auxiliem na melhoria, tanto do ensino quanto da aprendizagem, entre professores e alunos nas diversas modalidades.

Por meio do uso de EDM é possível apontar os diferentes fatores que influenciam o comportamento do aluno e identificar aspectos sutis, muitas vezes imperceptíveis por professores e gestores de cursos em EAD (BAKER, CARVALHO e ISOTANI, 2011).

A EDM torna possível a descoberta de padrões comportamentais de alunos e suas respectivas habilidades de aprendizagem. Esta descoberta de conhecimento pode subsidiar o desenvolvimento de *mecanismos capazes de auxiliar na tomada de decisão de professores para identificar comportamentos dos alunos*, visando traçar caminhos de aprendizado, que façam sentido para o aluno durante o decorrer de sua formação e os ajudem na tomada de decisões sobre o caminhar da sua própria aprendizagem.

1.2 Contexto e Motivação

São perceptíveis as mudanças que vêm ocorrendo no papel e nas competências dos professores ao longo dos anos. Essas mudanças são ainda mais acentuadas quando se trata de professores da modalidade de educação a distância.

Em cursos promovidos por meio dessa modalidade, o acompanhamento do aprendizado do aluno é dificultado pela falta de contato face-a-face entre professores e alunos e pela falta de percepção da análise de aspectos comportamentais da interação dos usuários (WANG, DOLL, et al., 2013). Por outro lado as inúmeras possibilidades de interação, bem como o registro das mesmas, oferecem um leque valioso para o enriquecimento do processo de aprendizagem e consequentemente da prática de avaliações formativas (LONN e TEASLEY, 2009).

Analisar e monitorar esse volume de interações entre professor, alunos e conteúdo é uma importante e complexa tarefa do professor no processo de aprendizagem *online*. A estrutura atual dos ambientes virtuais de aprendizagem não possibilita ao professor, de forma amigável, o acompanhamento, a avaliação e a percepção do comportamento dos alunos na realização das atividades propostas durante a execução de um curso (MEDEIROS, GOMES, *et al.*, 2013).

Além disso, a análise e visualização de dados comportamentais, fornecidas pelas atuais plataformas de educação a distância, conhecidas como LMS (Sistema de Gestão da Aprendizagem), geralmente limitam-se à geração de tabelas de *logs* (registro automático de eventos) ou apenas quantificações sobre o acesso dos participantes, limitando-se a informações retrospectivas. Uma das motivações desta tese é a utilização dessas informações para desenvolver instrumentos de previsões futuras, com a capacidade de auxiliar na compreensão implícita de informações sobre tendências futuras da utilização de recursos ou comportamentos dos estudantes.

1.3 Questões de pesquisa

Neste sentido, emerge a questão central de pesquisa desta tese, que busca responder: *É possível desenvolver um modelo que permita realizar a previsão do desempenho acadêmico de estudantes, a partir do seu comportamento autorregulatório?* As questões abaixo são norteadoras da questão central desta tese.

- Quais as principais variáveis, registradas em ambientes de LMS, que melhor descrevem as estratégias comportamentais de autorregulação da aprendizagem?
- Quais as relações existentes entre fatores de autorregulação e desempenho acadêmico de estudantes?

1.4 Objetivos

Nesta seção são apresentados os objetivos desta tese, ou seja, as metas que o trabalho procurou alcançar.

1.4.1 Geral

O objetivo geral desta tese é a modelagem preditiva do desempenho de estudantes por meio de indicadores comportamentais de autorregulação da aprendizagem em ambientes de LMS.

1.4.2 Específicos

- Mapear fatores da literatura de autorregulação da aprendizagem com variáveis comportamentais presentes em sistemas de LMS;
- Identificar as relações existentes entre variáveis comportamentais de autorregulação e desempenho acadêmico;
- Construir uma modelagem capaz de predizer o desempenho acadêmico de acordo com variáveis de autorregulação em plataformas de LMS;
- Desenvolver uma solução de software que promova a identificação de padrões comportamentais de autorregulação como subsídios para a prática docente de professores; e
- Avaliar a eficácia da interface para condução da ação docente;

1.5 Método de Pesquisa

O método utilizado nesta tese foi fundamentado no processo CRISP-DM (*Cross Industry Standard Process For Data Mining*). O processo CRISP-DM é descrito de forma hierárquica, composto por seis fases: entendimento do negócio/domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação (CRISP-DM, 1999).

Na fase de **entendimento do domínio** foi realizada a revisão da literatura e entrevistas de campo com especialistas no domínio com o objetivo de aprofundamento no problema de pesquisa.

Na fase do **entendimento dos dados** foi realizado um acordo de cooperação científica com o NEAD/UPE para a disponibilização da base de dados histórica do sistema de LMS da instituição. Foram obtidos dados referentes aos últimos sete anos de cursos referente aos registros comportamentais dos alunos nas plataformas de LMS. Após o entendimento dos dados foi iniciada a fase de escolha das variáveis que representassem aspectos comportamentais de autorregulação dos alunos no seu processo de aprendizagem.

Na fase de **preparação dos dados** foi realizada a extração das variáveis, construção da base de dados para análise e a junção das mesmas. Para a realização da extração das variáveis, foi necessário um estudo da organização do banco de dados. O estudo envolveu inicialmente a modelagem das entidades no banco de

dados, no qual foi necessário entender o contexto em que a plataforma estava sendo utilizada e suas especificidades.

Na fase de **modelagem**, responsável por atingir o objetivo geral desta tese, foram definidas as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de classificação, bem como seus ajustes de parâmetros. Esta fase teve como objetivo modelar os padrões comportamentais de autorregulação por meio de técnicas de mineração de dados educacionais em variáveis comportamentais.

A quinta fase foi responsável por a **validação** da modelagem. Nesta fase foi verificado como os modelos desenvolvidos com as técnicas de mineração são executados nos dados reais. Nesta tese a avaliação foi realizada por meio de técnicas que buscam identificar o poder de previsão dos modelos por meio de métricas de precisão de classificadores.

Por fim, a sexta e última fase do método foi a fase de **implementação**, a qual buscou materializar os resultados encontrados na fase de modelagem. Esta fase de implementação foi direcionada à construção de uma interface de software *dashboard* que proporcionasse ao professor/tutor uma visualização dos indicativos comportamentais de alunos que se encontram em estado crítico de autorregulação da aprendizagem e consequentemente em risco de mau desempenho acadêmico.

Todas as fases e instrumentos utilizados em cada fase descrita acima foram detalhados no Capítulo 5 desta tese.

1.6 Organização e estrutura da tese

Esta tese está dividida em oito capítulos. O Capítulo 2 descreve o referencial teórico que dá suporte a este trabalho, descreveu os principais modelos conceituais sobre autorregulação da aprendizagem existentes na literatura. O Capítulo 3 aborda a revisão bibliográfica relacionada à área de pesquisa intitulada Mineração de Dados Educacionais, mostrando as principais abordagens e técnicas utilizadas no âmbito da educação.

O Capítulo 4 descreve um mapeamento sistemático da literatura sobre as principais abordagens quantitativas de mensuração do fenômeno de autorregulação da aprendizagem, buscando identificar lacunas e oportunidades que justifiquem o desenvolvimento desta tese.

O Capítulo 5 descreve a proposta de método utilizada, levando em consideração a caracterização do problema, entendimento sobre os dados, os participantes envolvidos, o processo de mineração de dados e por fim a concepção de uma solução de software. No Capítulo 6 são descritos os resultados alcançados em todas as fases do método. No Capítulo 7 é feita uma discussão sobre os resultados, alinhando os fundamentos teóricos que embasaram esta tese com os resultados encontrados e, por fim, o Capítulo 8 discorre sobre as considerações finais, limitações encontradas e trabalhos futuros.

2. AUTORREGULAÇÃO DA APRENDIZAGEM

Neste capítulo iniciamos a seção referente à fundamentação, na qual apresentamos os conceitos fundamentais. Este capítulo apresenta algumas teorias existentes sobre autorregulação da aprendizagem e trabalhos relacionados, que fornecem a base teórica para posicionar este trabalho no campo da pesquisa. Entre os conceitos a serem explorados estão: o significado e conceito de autorregulação da aprendizagem, teorias e modelos de autorregulação da aprendizagem e por fim as abordagens de mensuração existentes.

2.1 Conceito de autorregulação da aprendizagem

A regulação é o termo utilizado em referência aos processos específicos que visam criar, implementar e ajustar estratégias de ensino à aprendizagem dos sujeitos. A regulação tem como objetivo contribuir diretamente para a progressão de aprendizagens, o que significa acompanhar o aprendiz em seu desenvolvimento (BRAGAGNOLOG e LOURDES , 2006).

Nesse sentido, a ação de regular pode ser definida como regularizar, estabelecer ordem ou ajustar (FERREIRA, 2012). A regulação da aprendizagem pode ser conduzida pelo professor ou pelo próprio estudante, sendo, nesse caso, denominada de Autorregulação.

A autorregulação da aprendizagem (*do inglês Self-Regulated Learning ou SRL*) apresenta-se como um processo de construção de capacidades para desenvolver e direcionar estratégias diante de tarefas e obstáculos inerentes a elas. Nesta concepção, a participação autônoma, ativa e planejada do indivíduo nos processos de ensino e de aprendizagem são componentes indispensáveis.

Apesar de existir um vasto consenso entre os investigadores relativamente à importância do papel desempenhado pelas estratégias de aprendizagem na construção de aprendizagens significativas, esses tardam em convergir na sua definição. Tal como foi referido por Weinstein e colaboradores (1987), parece não existir uma definição consensual de estratégias de aprendizagem, sendo, por vezes,

utilizadas designações diferenciadas para variáveis similares, ou inversamente, aspectos e dimensões distintas do construto são designados pelo mesmo termo.

A autorregulação da aprendizagem é definida por teóricos como o processo em que os sujeitos, após estabelecerem metas que interagem com suas expectativas, desenvolvem estratégias para alcançá-las, criando condições para que a aprendizagem efetive-se (ZIMMERMAN, 1998; ZEIDNER, BOEKERTS e PINTRCH, 2000; SILVA, 2004; VEIGA SIMÃO, 2006).

Gonçalves (2010) defende que os alunos devem ser explicitamente estimulados a deixar a atitude tradicional de espectadores da apresentação de conhecimentos e se tornarem personagens principais na construção das suas próprias competências.

Em síntese, podemos considerar que a aprendizagem autorregulada é um processo de aquisição de conhecimento: ativo, construtivo e orientado para objetivos sob a tutela da interação dos recursos cognitivos, motivacionais e emocionais do aluno. Significa também a competência dos aprendizes para desenvolverem o conhecimento, as competências estratégicas e as atitudes necessárias para incrementar e facilitar as futuras aprendizagens não só no contexto escolar, mas também em outros contextos.

Zimmerman (1989) afirma que para um aluno ser considerado autorregulado este deve ser participante ativo metacognitivo, motivacional e comportamental em sua própria aprendizagem. O autor afirma que dependendo do desenvolvimento sócio cognitivo do indivíduo vão existir alunos inexperientes e que apresentam uma baixa capacidade de autorregular seu aprendizado, enquanto outros alunos vão ser mais experientes e possuem capacidades mais desenvolvidas para regular o seu processo de aprendizado. Zimmerman (1998) destacou as principais diferenças entre os alunos autorregulados e os não regulados.

De acordo com a taxonomia de McKeachie et al. (apud FILCHER e MILLER, 2000) pode-se dividir as estratégias de autorregulação da aprendizagem em três grupos: as estratégias cognitivas, as metacognitivas e as de gestão dos recursos de aprendizagem. Segundo Filcher e Miller (2000) as estratégias cognitivas são importantes para a compreensão de como a informação é processada e codificada em um ambiente de aprendizagem. As estratégias metacognitivas são relacionadas

ao planejamento e monitoramento de seu desempenho. Por fim as estratégias de gestão de recursos de aprendizagem consistem em estratégias utilizadas para gerir e controlar o ambiente físico e social e os recursos disponíveis.

Dentre as estratégias de gestão de recursos as principais são a gestão do tempo e a gestão do ambiente de estudo (PITRICH, 1999; FILCHER E MILLER, 2000). A gestão do tempo envolve o processo de estabelecimento de objetivos bem definidos e de programação do curso para obtermos os melhores resultados. Programação neste sentido significa o processo no qual os estudantes definem um tempo específico e organizam o dia ou a semana para realizarem as atividades de aprendizagem necessárias para um bom desempenho no curso (FICHER e MILLER, 2000).

Autores como Eilam e Aharon (2003) consideram o tempo um aspecto crucial do planejamento e da regulação. A limitação de tempo adiciona uma dimensão extra que os estudantes devem considerar ao tomar decisões e escolher entre alternativas de ações. De acordo com Testa (2005), relacionado com o fator tempo está a taxa de progresso do estudante em direção aos objetivos. Uma taxa de progresso diferente das expectativas interfere sobre o tipo de regulação adotada, assim como as emoções individuais em relação ao ritmo percebido.

Diante dos diversos construtos e definições sobre a conceituação da autorregulação da aprendizagem foram desenvolvidos modelos teóricos que apresentam informações específicas sobre o processo de autorregulação. A seção seguinte apresenta os principais modelos teóricos encontrados na literatura.

2.2 Modelos teóricos de autorregulação da aprendizagem

Os modelos de autorregulação da aprendizagem são fundamentados em preceitos, como o papel ativo do aluno, a reflexão sobre as ações realizadas e sobre seu comportamento, e uma posterior mudança de atitudes buscando melhorar seu desempenho (SOUZA e GOMES, 2012). Embora existam diversos outros modelos teóricos sobre autorregulação da aprendizagem (BUTLER e WINNE, 1995; BORKOWSKI, 1996; WINNE e HADWIN, 1998; PINTRICH, 2000; ZIMMERMAN, 2000), neste trabalho iremos apresentar os modelos de Zimmerman e Pintrich, tendo em vista que estes são os principais modelos encontrados na literatura.

2.2.1 O modelo de Zimmerman

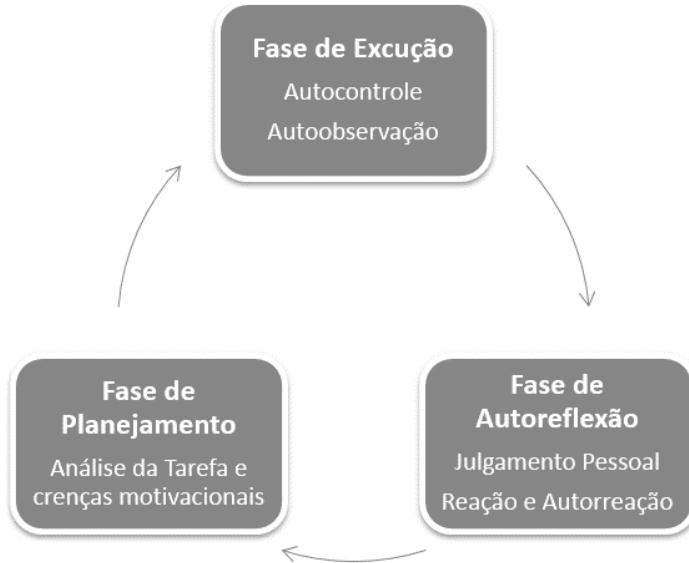
Este modelo é um dos mais abrangentes e apresenta informações específicas sobre os processos de trabalho do indivíduo por meio de um modelo cíclico (ZIMMERMAN, 2009). O modelo leva em consideração a interação entre fatores pessoais, comportamentais e do ambiente. De acordo com Zimmerman (2009) a autorregulação é um processo cíclico que envolve as habilidades cognitivas, metacognitivas e as habilidades do indivíduo em auto gerenciar as demandas do ambiente. De acordo com Kizilcec, Pérez-Sanagustín e Maldonado (2016) esta característica de agência pessoal é a capacidade de o indivíduo exercer um papel ativo em sua própria história.

Zimmerman e Martinez-Pons, 1986 e Zimmerman E Bandura, 1994 ao definirem seu modelo de autorregulação identificaram os principais fatores envolvidos no processo a partir da realização de vários estudos. O modelo proposto é constituído de três fases: fase de planejamento, fase de execução da atividade e fase de autorreflexão. As três fases formam um processo cíclico e acompanham toda a realização da atividade (Figura 1).

A primeira é a fase de planejamento. Nesta fase o indivíduo define as metas que pretende alcançar e simultaneamente analisa sua própria capacidade de atingir esses resultados.

A fase de execução envolve os processos que irão ocorrer durante a atividade de aprendizagem. Nesta fase o indivíduo faz uso de estratégias adequadas para o contexto no qual está inserido.

Figura 1: Ciclo de autorregulação da aprendizagem



Fonte: Zimmerman (1986)

Por fim, a fase de autorreflexão, que inicia após a conclusão da atividade de aprendizagem. Envolve a avaliação do processo de aprendizagem que o indivíduo vivenciou. Esta é uma fase de autorreflexão, na qual o sujeito irá avaliar a eficiência das estratégias de aprendizagem utilizadas no alcance de suas metas e na obtenção dos resultados desejados, sendo possível fazer reflexões e ajustes, e retornar ao ciclo.

2.2.2 O modelo de Pintrich

O modelo proposto por Pintrich, Smith, *et al.* (1993) é composto por quatro etapas, o planejamento e ativação, o monitoramento, o controle e, por fim, reação e reflexão. Estas etapas podem acontecer de modo simultâneo e dinâmico durante a realização das tarefas.

Em cada uma das fases, devem ocorrer atividades de regulação em quatro áreas: cognição, motivação e afeto, comportamento e contexto. O Quadro 1 apresenta as etapas do modelo e as áreas de regulação.

Quadro 1: Fases e áreas do modelo de autorregulação de aprendizagem.

	Áreas para regulação			
	Cognição	Motivação	Comportamento	Contexto
Planejamento e Ativação	Estabelecimento de metas. Ativação do conhecimento prévio. Ativação do conhecimento metacognitivo.	Adoção de metas. Julgamento da autoeficácia. Ativação de crenças sobre valor da tarefa. Percepção das dificuldades da tarefa. Ativação do interesse pessoal.	Planejamento do tempo e esforço. Planejamento de como será monitoramento do comportamento.	Percepção da tarefa. Percepção do contexto.
Monitoramento	Percepção metacognitiva. Monitoramento da cognição.	Percepção e monitoramento da motivação e dos aspectos afetivos.	Consciência e auto-observação do esforço, do tempo utilizado e da necessidade de ajuda.	Monitoramento das condições da tarefa e do contexto.
Controle	Seleção e adaptação das estratégias cognitivas para o aprendizado.	Seleção e adaptação de estratégias para gerenciamento, motivação e aspectos afetivos.	Aumento e redução de esforço. Comportamento de persistência e busca de ajuda.	Alterar ou renegociar a tarefa. Alterar ou desistir do contexto.
Reação e Reflexão	Julgamento cognitivo.	Reações afetivas.	Escoa do comportamento.	Avaliação da tarefa. Avaliação do contexto.

Fonte: Pintrich (2005)

A primeira fase se refere ao estabelecimento de metas. Nesta fase os alunos ativam seus conhecimentos prévios, examinam suas expectativas de sucesso e analisam as expectativas e normas sociais em relação à execução da tarefa.

Na segunda fase são realizados julgamentos acerca da própria compreensão da tarefa pelo aluno, do autoexame dos afetos e por uma contínua avaliação da tarefa em relação às demandas externas e recompensas.

A terceira fase se caracteriza pela aplicação de estratégias para solucionar a tarefa. Esta é uma das fases mais importantes para esse trabalho, é nesse momento que os estudantes empregam estratégias para o controle de seus afetos e motivações e o esforço a ser dedicado para execução da tarefa. Nesta fase são tomadas decisões sobre prosseguir ou desistir.

A última fase do modelo caracteriza-se pela autoavaliação do desempenho do estudante na tarefa, pela reflexão acerca de suas reações emocionais e pela análise da adequação do investimento de seu esforço, as quais acabam sendo muito úteis para o engajamento em tarefas futuras.

Os modelos de autorregulação apresentados nas seções 2.2.1 e 2.2.2 apresentam maneiras de buscar desenvolver no aluno estratégias que apoiam o direcionamento de seu processo de aprendizagem. A próxima seção menciona os aspectos específicos da autorregulação da aprendizagem em ambientes característicos de Educação a Distância.

2.3 Características da autorregulação em ambientes de EAD

Estudos vêm sendo desenvolvidos com objetivo de analisar a importância de habilidades da autorregulação da aprendizagem em ambientes de EAD. Diversos pesquisadores acreditam que os ambientes de aprendizagem online exigem do aluno uma maior responsabilidade para o processo de aprendizagem (SCHUNK e ZIMMERMAN, 1998; DABBAGH e KITSANTAS, 2005).

Estes autores acreditam que as habilidades de autorregulação são essenciais para o sucesso em situações de aprendizagem autônomas e que o desenvolvimento dessas habilidades pode ser suportado por ferramentas pedagógicas baseadas na Web (DABBAGH e KITSANTAS, 2004; AZEVEDO, 2005).

Estas habilidades podem ser importantes para alunos que participam de cursos nesta modalidade (SCHUNK e ZIMMERMAN, 1998; HARTLEY e BENDIXEN, 2001;

GARRISON, 2003; DABBAGH e KITSANTAS, 2004). Por exemplo, Dabbagh e Kitsantas (2004) argumentaram que, "em um ambiente de aprendizagem baseado na Web, os alunos devem exercer um alto grau de competência de autorregulação para realizar seus objetivos de aprendizagem, enquanto que em ambientes tradicionais face-a-face em sala de aula, o instrutor exerce um controle significativo sobre o processo de aprendizagem e é capaz de monitorar a atenção do aluno e do seu progresso". Da mesma forma, autores como Kinzie (1990) afirmam que a habilidade de autorregulação é um dos três requisitos críticos para o sucesso do aluno. A utilização eficaz das estratégias de SRL é essencial em situações de aprendizagem flexíveis, devido ao alto grau de autonomia do aluno, resultante da ausência física do professor.

2.4 Relação entre autorregulação e desempenho acadêmico

Dado o rápido crescimento da modalidade de aprendizagem online, na última década, há uma necessidade de compreender como os alunos podem utilizar melhor as estratégias de SRL para alcançar melhores níveis de desempenho acadêmico.

O desempenho acadêmico pode ser definido como o ato de alcançar um resultado esperado em uma atribuição, atividade, prova ou grau de entendimento sobre um determinado assunto, no qual é normalmente expresso em termos de pontuação numérica (RICHARDSON, ABRAHAM e BOND, 2012). Diversas pesquisas vem revelando a relação positiva entre o uso de estratégias SRL e os resultados acadêmicos em contextos de aprendizagem presencial (PINTRICH, 2004; DIGNATH e BUTTNER, 2008; ZIMMERMAN, 2008; BEISHUIZEN e STEFFENS, 2011; RICHARDSON, ABRAHAM e BOND, 2012; WANG, DOLL, *et al.*, 2013; AGUSTIANI, CAHYAD e MUSA, 2016). Estas pesquisas dão ênfase em ações de metacognição, gestão do tempo, e regulação de esforço (RICHARDSON, ABRAHAM e BOND, 2012).

No entanto, ainda são poucas as pesquisas conduzidas sobre o uso de SRL na aprendizagem online e suas relações com desempenho acadêmico (BROADBENT e POON, 2015). A maioria dos trabalhos são estudos prospectivos (HODGES e JOHNSON, GUEUTAL e FALBE, 2009; KIM, 2010; CARSON, 2011; MICHINOV, BRUNOT, *et al.*, 2011; CHANLIN, 2012; CHO e SHEN, 2013), seguidos por estudos

experimentais (MERRIËNBOER, 2007; CHANG, 2010) e transversais (KLINGSIECK, FRIES, et al., 2012; WANG, SHANNON e ROSS, 2013).

Em relação aos estudos prospectivos, o trabalho de Carson (2011) analisou o grau de sucesso de estudantes aplicando a escala de aprendizagem online LASSI (LLO). O autor fornece evidências de previsibilidade de desempenho, baseado nos itens da escala LASSI com uma acurácia global de 67%.

No trabalho de Cho e Shen (2013) os autores discutem a importância da orientação para as metas individuais dos estudantes e a relação com seu desempenho acadêmico. Os resultados mostraram uma relação positiva entre a quantidade de tempo que o aluno passa na plataforma, gestão do tempo, e seu desempenho acadêmico.

Em relação aos estudos transversais, um dos trabalhos que merece destaque é o trabalho de Klingsieck, Fries, et al (2012). O estudo dá ênfase ao problema de procrastinação dos estudantes e suas implicações no desempenho de atividades acadêmicas. Embora existam pesquisas sobre a procrastinação de estudantes, é pouco comum estudos destinados a entender a procrastinação em ambientes de ensino a distância. Dentre as pesquisas de autorregulação da aprendizagem, a procrastinação é ainda pouco explorada, sendo uma característica ainda mais necessária de estudos em ambientes de educação a distância.

Dentre os estudos mencionados anteriormente que buscam a relação entre estratégias de autorregulação da aprendizagem e desempenho acadêmico, os instrumentos de medida utilizados foram: o questionário (MSLQ), o questionário LASSI e a escala de procrastinação Tuckman.

A maioria destes instrumentos de coleta foram desenvolvidos para ambientes presenciais de ensino. Na seção seguinte explicitamos as diferentes abordagens de mensuração que vêm sendo desenvolvidas ao longo dos anos, bem como os diversos instrumentos de coleta.

2.5 Abordagens de mensuração da autorregulação

Diversas abordagens de mensuração da autorregulação vêm sendo desenvolvidas ao longo dos anos. Quantificar medidas de aptidão de habilidades de

autorregulação do aluno, durante as tarefas específicas de aprendizagem, é um desafio. Estas medidas de aptidões são normalmente desenvolvidas por meio de escalas de medidas que os alunos respondem por meio de questionários.

2.5.1 Instrumentos de coleta

Uma das primeiras medidas desenvolvidas foi a *LASSI - Learning and Study Strategies Inventory* desenvolvida por Weinstein, Schulte E Palmer (1987). Trata-se de um questionário de autorrelato, com 77 variáveis formuladas para avaliar estratégias de aprendizagem de estudantes do ensino superior, composto por dez subescalas: **atitude, motivação, organização do tempo, ansiedade, concentração, processamento da informação, seleção de ideias principais, uso de técnicas e materiais de apoio, autovalorização e estratégias de avaliação.**

Posteriormente Pintrich, Smith, et al. (1993) desenvolveram a escala de medida MSLQ - *Motivated Strategies for Learning Questionnaire*. Este instrumento de autorrelato, com 81 itens, visa a **avaliação da orientação motivacional e uso de estratégias de aprendizagem de estudantes em um curso ou matéria específica**. A motivação analisa componentes de expectativas (crenças de controle e autoeficácia), de valor (metas intrínsecas, extrínsecas e valor da tarefa) e afetivos (ansiedade nos exames). Quanto ao uso das estratégias de aprendizagem, avaliam-se estratégias cognitivas e metacognitivas (repetição, elaboração, organização, pensamento crítico e metacognição) e estratégias de gerência de recursos (tempo e lugar de estudo, regulação do esforço, modelação e busca de ajuda).

Diferentemente dos instrumentos anteriores Zimmerman E Pons (1986) desenvolveram a medida de mensuração intitulada SRLIS - Self-Regulated Learning Interview Scale. Neste caso o procedimento de coleta sobre a autorregulação ocorre por meio de entrevista estruturada. **São avaliadas as 14 estratégias: autoavaliação, organização e transformação da informação, estabelecimento de objetivos e planejamento, busca de informação, anotação e monitoramento, estruturação do ambiente, auto consequência, memorização, busca de ajuda dos pares, busca de ajuda dos professores, busca de ajuda de adultos, revisão de notas, revisão de testes e revisão de livros.**

Embora alguns instrumentos de medição são utilizados igualmente no ensino presencial e no ensino a distância, se formos considerar as diferenças entre os dois ambientes de aprendizagem, um instrumento que funciona bem na sala de aula presencial pode não ser válido em situações de aprendizagem online (TALLENT-RUNNELS, THOMAS, et al., 2006). É importante, então, que estudos que pretendam mensurar SRL em ambiente online utilizem instrumentos de pesquisa adequados para tal contexto.

Um instrumento que vêm sendo utilizado na literatura, com foco na aprendizagem online, é resultado do trabalho de Barnard, Lan, et al. (2009), intitulado OSLQ. *Este instrumento* foi desenvolvido e projetado para a avaliação de habilidades de autorregulação que são importantes na aprendizagem online. Suas sub-escalas incluem estruturação do ambiente, estabelecimento de metas, gestão do tempo, a procura de atendimento e estratégias de tarefas de auto-avaliação.

É interessante observarmos que, embora o fenômeno de autorregulação da aprendizagem venha sendo estudado há décadas, existe na literatura apenas um único instrumento de coleta com foco na mensuração em ambientes de EAD. Este instrumento foi validado em contexto brasileiro (RODRIGUES, RAMOS, et al., 2016). Na seção seguinte são discutidas as formas de mensuração que vem sendo utilizadas para a coleta de dados de autorregulação.

2.5.2 Abordagens de mensurações

As abordagens mais utilizadas para mensuração da autorregulação, de acordo com a literatura, são: a observação direta, o método do pensamento em voz alta, a microanálise, agenda estruturada e caminhos da aprendizagem (WINNE e PERRY, 2000).

No método de observação direta (PERRY, VANDEKAMP, et al., 2002) o pesquisador observa os alunos em um ambiente de aprendizagem, tais como uma sala de aula e registra ocorrências de características da autorregulação. A força do método é que não há nenhum viés de resposta a questionários ou entrevistas e os dados são da situação atual, em vez de retrospectiva.

Os protocolos do Pensamento em voz alta, definidos por Azevedo, Cromley, et al., (2005) podem ser usados para adquirir uma visão mais detalhada dos processos

autorregulatórios do estudante durante a aprendizagem. Neste método, o pesquisador fica ao lado de um aluno e pede para explicar continuamente os pensamentos e ações. As transcrições são codificadas, a fim de categorizar as estratégias utilizadas.

Outra abordagem é a Microanálise utilizada por Cleary (2011). É uma forma de mensuração que tem como alvo os comportamentos ou processos como eles ocorrem em tempo real em contextos autênticos (CLEARY, 2011). Eles diferem da abordagem do pensamento em voz alta, pois a microanálise envolve questões altamente estruturadas, enquanto o pensamento em pensar em voz alta tem forma mais livre e sem estruturação pré-definida. Na microanálise o pesquisador se senta ao lado dos alunos e apresenta perguntas específicas sobre processos de autorregulação (por exemplo, "O que você precisa fazer para cumprir estes objetivos?"). As transcrições são codificadas e registadas com rótulos para categorizar as estratégias. Nos métodos de observações diretas, tanto o método de pensamento em voz alta quanto no método microanálise, os alunos são observados durante as atividades de aprendizagem, permitindo a condução de estudos com um grande número de estudantes.

O método de agenda estruturada (SCHMITZ e WIESE, 2006) é um método utilizado para coletar dados de autorregulação sem exigir que o pesquisador esteja presente durante a tarefa de aprendizagem dos alunos. Eles são comparáveis às escalas de autorrelato, em que ambos se baseiam em autorrelatos retrospectivos, no entanto, os diários estruturados permitem que informações sejam coletadas de tarefas específicas de aprendizagem em vez de relatos de comportamentos gerais.

Por fim, o método de caminhos de aprendizagem é a técnica de coleta de dados dos alunos por meio dos traços que os alunos naturalmente deixam no ambiente de aprendizagem ou nos materiais didáticos. Estes traços ou caminhos, poderiam ser, por exemplo sublinhados no livros, esquemas e fluxogramas. No que se refere à aprendizagem online, este método pode ser bastante proveitoso pois as plataformas de aprendizagem online possibilitam o registro dos **traços comportamentais dos alunos** diariamente, tais como **quantidades de cliques, tempo de permanência online nas plataformas, trocas de mensagens, acesso a materiais didáticos, discussões fóruns, entre outros.**

Dentre todas as abordagens mencionadas anteriormente, alguns autores questionam a validade da abordagem de autorrelato. Esta abordagem caracteriza-se

pelo autorrelato retrospectivo do aluno sobre suas ações de aprendizagem. Autores como Winne e Perry (2000), Winne e Jamieson-Noel (2002), têm duvidado da eficácia desta abordagem, pois afirmam que em alguns casos os alunos podem não ser capazes de lembrar com precisão suas atividades no passado, em outros casos podem subestimar seu próprio comportamento dando respostas tendenciosas e em casos mais extremos podem não necessariamente responder tais perguntas com honestidade.

Dentro desta perspectiva, Winne e Jamieson-Noel (2002) estudaram a precisão da abordagem auto relato comparada com a abordagem de caminhos de aprendizagem por meio do rastreamento de *logs* do usuário. Neste estudo os alunos foram submetidos a estudar um determinado conteúdo por meio de um software educacional e em seguida executar uma autoavaliação de desempenho respondendo um questionário sobre o uso de estratégias de autorregulação. Estes foram comparados a um pós-teste e foram coletados os logs do software.

Os autores descobriram existirem diferenças significativas entre as afirmações de autorregulação dos alunos e seus traços coletados por meio dos logs do software educacional. O resultado mostrou que os alunos podem ter uma visão muito distorcida do seu próprio comportamento. No entanto, Winne e Perry (2000) afirmam ser mais seguro mensurar comportamentos de autorregulação por meio de uma triangulação, utilizando várias abordagens diferentes.

2.6 Considerações sobre o capítulo

Nesta seção explicitamos os principais instrumentos de coleta e formas de mensuração utilizados na literatura. É importante observarmos que grande parte dos trabalhos que envolvem a mensuração dos construtos de autorregulação da aprendizagem priorizam instrumentos de coleta pontuais, por meio de métodos de observação ou aplicação de questionários.

Estas abordagens não permitem uma mensuração em tempo real sobre dados de interação do estado autorregulatório dos alunos.

Em ambientes de EAD, especificamente em plataformas virtuais de aprendizagem, essa necessidade de indícios em tempo real é acentuada e

abordagens como mineração de dados educacionais podem contribuir de forma substancial.

O próximo capítulo aborda a descrição sobre as principais técnicas que podem contribuir para mensurar comportamentos em plataformas de EAD por meio de dados de interações dos usuários.

3. MINERAÇÃO DE DADOS EDUCACIONAIS (EDM)

Neste capítulo explicitamos as principais técnicas de Mineração de Dados Educacionais, nos quais são apresentados os conceitos necessários para a fundamentação desta tese. Entre os conceitos a serem explorados estão: contextualização sobre EDM, áreas relacionadas e abordagens, as principais tarefas, intercalando com os principais trabalhos desenvolvidos na literatura.

3.1 Contextualização sobre Mineração de Dados Educacionais

Com a ampla difusão do uso de sistemas informatizados nas escolas e universidades, cresce a cada dia o volume de dados gerados e armazenados em bases de dados (RIGO, CAMBRUZZI, *et al.*, 2014). Alguns exemplos destes fatores de crescimento são as Tecnologias de Informação e Comunicação, o suporte computacional para Ambientes de Aprendizagem, a consolidação da modalidade de Educação a Distância e do *Blended Learning*, a incorporação de Sistemas Integrados de Gestão em instituições educacionais, as comunidades virtuais e suas ferramentas de relacionamento, compartilhamento e publicação online (CAMBRUZZI, W. L, DE MORAES, R, *et al.*, 2012).

Este aumento no volume de dados, tem fomentado o interesse na sua utilização pela comunidade científica. Muitos pesquisadores da área de Informática na Educação, em particular, **Inteligência Artificial Aplicada à Educação**, têm mostrado interesse em utilizar **técnicas de mineração de dados** para investigar perguntas científicas na área educacional (BAKER, CARVALHO e ISOTANI, 2011). Dentro deste contexto, diversas pesquisas vêm sendo desenvolvidas na área de “Mineração de Dados Educacionais” (do inglês, “*Educational Data Mining*”, ou EDM). A EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais.

Atualmente, existem várias definições para o termo Mineração de Dados Educacionais ou EDM. O Termo é definido pela **sociedade internacional de mineração**

de dados educacionais¹ como “É uma área emergente, preocupada com o desenvolvimento de métodos para explorar tipos de dados provindos de contextos educacionais, e usando métodos para entender os alunos e suas configurações de aprendizagem” (MCGAW, PETERSON e BAKER, 2010).

Romero e Ventura (2010) definem a EDM com sendo " a aplicação de técnicas de mineração de dados (DM) para tipos específicos de conjuntos de dados provenientes de ambientes educacionais para abordar questões educacionais importantes ".

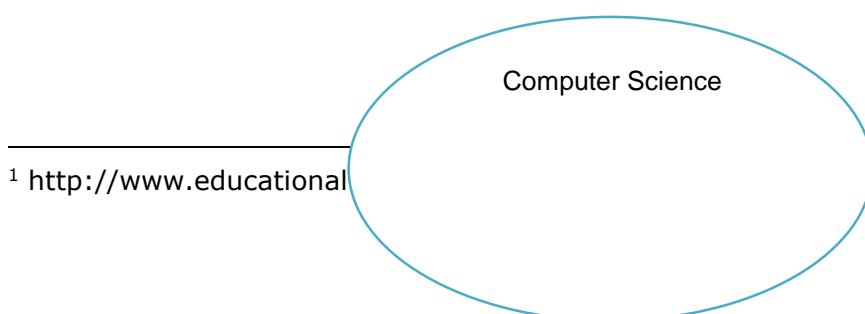
A EDM demanda de adaptações das técnicas já existentes e o desenvolvimento de novas técnicas. A natureza dos dados analisados em EDM é mais diversa do que a observada nos dados tradicionalmente utilizados em abordagens clássicas de mineração de dados. Ao mesmo tempo, esta diversidade nos dados representa um potencial de implementação de recursos fundamentais para auxílio na melhoria da Educação (ROMERO, VENTURA, et al., 2010).

Na seção seguinte explicitamos as diversas áreas e suas intersecções que compõem a área de *Educational Data Mining*.

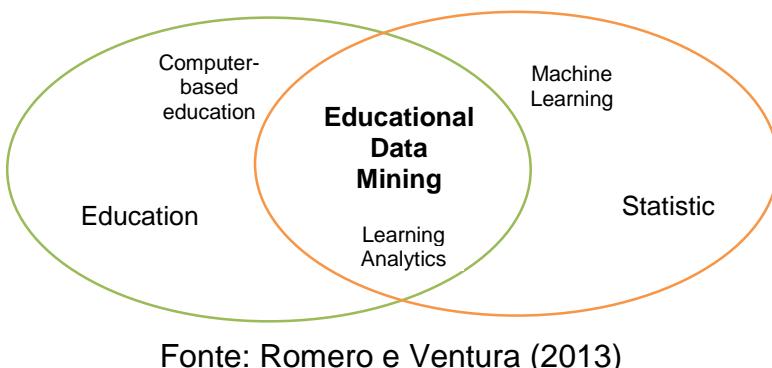
3.2 Áreas relacionadas a EDM

Mineração de dados educacionais é uma área interdisciplinar, incluindo áreas como a recuperação de informação, sistemas de recomendação, visualização de dados, análise de redes sociais (SNA), psicopedagogia, psicologia cognitiva, psicometria, dentre outras. Em uma visão mais abrangente, definida por Romero e Ventura (2013), os autores afirmam que EDM pode ser visualizada como a combinação de três principais áreas (Figura 2): Ciência da Computação, Educação e Estatística. Na intersecção dessas três áreas surge a relação com algumas subáreas estreitamente relacionadas com EDM, como a área de educação baseada em computador (*E-learning*), Mineração de Dados e Aprendizagem de Máquina e Aprendizagem Analítica (LA).

Figura 2: Principais áreas relacionadas a EDM



¹ <http://www.educational>



Fonte: Romero e Ventura (2013)

De todas as áreas mencionadas, a mais relacionada com EDM é LA, em *Learning Analytics* (LA), as técnicas mais utilizadas são a aplicação de inferência estatística, visualização de dados, análise de redes sociais (SNA), análise de sentimento, análise de influência e análise de discurso.

Na área de EDM, são utilizadas as principais tarefas tradicionais de mineração de dados, tais como: classificação, regressão, agrupamento e análise de associação. Estas tarefas têm sido aplicadas com sucesso no domínio da educação (PEÑA-AYALA, 2014).

Embora boa parte dos trabalhos em EDM utilizem destas tarefas, os sistemas educacionais têm características especiais que exigem o desenvolvimento de novas abordagens de mineração de dados (ROMERO, VENTURA, et al., 2010). Pesquisadores envolvidos com EDM utilizam não apenas tarefas de mineração de dados tradicionais, mas também propõem desenvolver e aplicar métodos e técnicas extraídas de uma variedade de áreas relacionadas com a EDM (estatísticas, aprendizado de máquina, mineração de texto, análise de logs web, psicometria).

No trabalho de Rodrigues, Ramos et al.(2014) foi realizada uma análise das principais pesquisas, que vêm sendo desenvolvidas na área de Mineração de Dados Educacionais, publicados no Brasil desde o ano de 2006. Onde obtiveram como resultado, que a tarefa mais utilizada em trabalhos de EDM é a classificação, seguida por trabalhos que utilizam regras de associação.

Diante deste contexto, diversas abordagens, com a utilização destas tarefas, vêm sendo desenvolvidas no âmbito educacional. O Quadro 2 mostra as abordagens listadas por Romero e Ventura (2013).

Quadro 2: Abordagens em EDM.

Aplicação	Descrição
Predição do desempenho do estudante	Estimação de desempenho de alunos baseado em variáveis comportamentais
Investigação científica	Testes e comprovações de teorias de aprendizagem, com o objetivo de formular novas hipóteses científicas, e assim por diante.
Fornecendo feedback para apoio a professores	Fornecimento de feedback para apoiar educadores no processo de decisão sobre a melhor forma de aprendizagem dos alunos e permitir-lhes tomar / ou ação apropriada proativa e de reparação.
Aprendizagem personalizada/adaptativa para alunos	Para adaptar-se automaticamente a diversos perfis de aprendizagem, formas de navegação e conteúdo.
Recomendação para estudantes	Para fazer recomendações aos estudantes com relação a suas atividades ou tarefas, links para visitas, atividades a serem feitas e cursos a serem realizados.
Criação de alertas para estudantes	Para monitorar o progresso da aprendizagem dos alunos, para a detecção em tempo real do comportamento indesejável dos alunos, tais como baixa motivação, uso indevido, abandono, e assim por diante.
Modelagem do usuário/estudante	Para o desenvolvimento de modelos cognitivos dos estudantes, representando suas competências e seus conhecimentos.
Modelagem do domínio	Para descrever o domínio de instrução em termos de conceitos, habilidades, itens e aprender e suas inter-relações.
Agrupamento de perfis de alunos	Para criar grupos de estudantes de acordo com suas características pessoais, dados de aprendizagem.
Construindo material didático	Para ajudar a instrutores e desenvolvedores para realizar o processo de construção / desenvolvimento de material didático e aprendizagem de conteúdos automaticamente.
Estimação de parâmetros	Para inferir parâmetros de modelos probabilísticos a partir de determinados dados para prever a probabilidade de eventos de interesse

Fonte: Adaptado de Romero e Ventura (2013).

É importante mencionar que um trabalho na área de EDM nem sempre se classifica em apenas uma das abordagens mencionadas no Quadro 2. Geralmente os trabalhos perpassam por mais de uma abordagem educacional.

Além disto, cada abordagem educacional é desenvolvida por meio de uma ou mais tarefas de Mineração de Dados. A seção a seguir descreve as principais tarefas e seu percentual de utilização em trabalhos encontrados na literatura.

3.4 Tarefas de Mineração de Dados em contextos Educacionais

A implementação de um modelo de EDM é feito por meio de tarefas. Algumas destas tarefas são reconhecidas nas áreas clássicas da mineração de dados.

Normalmente são divididas em **modelos exploratórios**, tais como **clustering** (BERKHIN, 2006), **as regras de associação** (HONG, LIN e WANG, 2003), **a análise de correlação** (HARDOON, SHAWE-TAYLOR, & SZEDMAK, 2004) e **modelos preditivos**, tais como a **classificação** (CHAU, CHENG, *et al.*, 2006), e a **regressão** (WU e LI, 2007). Nos estudos de Peña-Ayala (2014), o autor elaborou um quadro (Quadro 3) com a frequência das tarefas mais comuns em EDM.

Quadro 3: Mensuração de tarefas de DM utilizadas em EDM

Tarefas	Quantidade	Percentual (%)
Classificação	102	42,15
Clustering/Agrupamento	65	26,86
Régressão	37	15,29
Regras de associação	16	6,61
Outros	22	9,09
Total	242	100

Fonte: Peña-Ayala (2014)

Dentre as tarefas listadas, a **classificação e agrupamento** foram as mais citadas em trabalhos da área. Juntas, ambas atingiram um total de 69% das tarefas de mineração de dados utilizados por abordagens EDM. Nas seções seguintes são mostradas a descrição sobre cada uma das tarefas mencionadas no Quadro 3.

3.4.1 Classificação

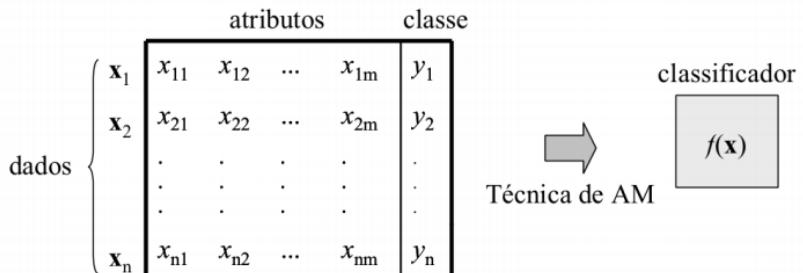
Uma das tarefas mais populares em mineração de dados é a tarefa de classificação. Em EDM a tarefa de classificação é geralmente utilizada para **classificar perfis de alunos, classificação de estilos de aprendizagem e previsão de desempenho** (PAI, LYU e WANG, 2010; PARDOS, WANG e TRIVEDI, 2012).

Essa tarefa pode ser compreendida como a busca por uma **função** que permita associar corretamente cada registro X_i de uma base de dados, a um **único rótulo categórico** Y_j , denominado **classe**. Uma vez identificada, essa função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram (TAN, STEINBACH e KUMAR, 2009).

A construção de um classificador, a partir do **aprendizado supervisionado**, pode ser representado, em sua forma simplificada pela Figura 3. Esta figura mostra um conjunto com n dados. Cada observação x_i possui m atributos. As variáveis y_i representam as classes ou rótulos. A partir dos dados de exemplos e as classes, o

algoritmo de classificação constrói uma função de classificação intitulada de classificador.

Figura 3: Modelo de Base de Classificação



Fonte: Lorena e Carvalho (2007)

A tarefa de classificação pode ser dividida em duas categorias, a classificação binária e a classificação multiclasse. Na primeira categoria a classe alvo possui apenas dois valores, sendo portanto intitulada como uma variável binária. Na segunda, a classe alvo é uma variável categórica com mais de duas classes.

Para a construção de um classificador, diversos algoritmos vêm sendo desenvolvidos ao longo de pesquisas, tais como classificadores por árvores de decisão, classificadores baseados em regras, redes neurais, máquinas de vetor de suporte, classificadores de Bayes, dentre outros. Cada técnica emprega um algoritmo de aprendizagem para identificar um modelo que seja mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo da classe dos dados de entrada.

O modelo gerado pelo algoritmo de classificação deve se adaptar bem aos dados de entrada e ter a capacidade de prever corretamente os rótulos de classes de registros desconhecidos. Portanto, o objetivo chave do algoritmo é construir modelos com boa capacidade de generalização, modelos que prevejam com precisão os rótulos de classes de registros não conhecidos previamente. A seguir descrevemos as características dos principais algoritmos de classificação utilizados nesta tese.

✓ Árvore de Decisão

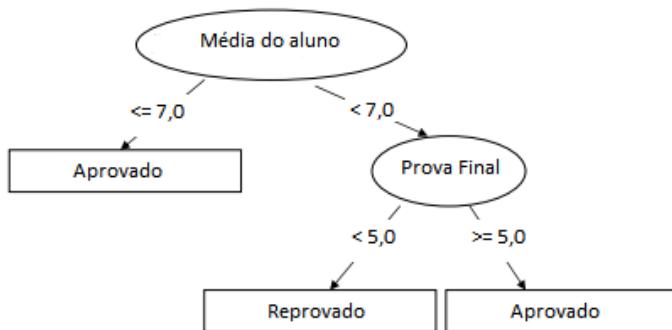
Os algoritmos de árvore de decisão são os mais utilizados em tarefas de classificação de estudantes. No trabalho de Narlı, Özgen E Alkan (2011) os autores

buscaram identificar a relação entre as várias áreas de **inteligência dos estudantes** e seus **estilos de aprendizagem** utilizando o conceito de conjuntos irregulares.

Uma das razões para que os algoritmos de árvore de decisão sejam bastante utilizados é o fato do **conhecimento adquirido poder ser representado visualmente por um conjunto de regras** (TAN, STEINBACH e KUMAR, 2009). Essas regras podem ser expressas em linguagem natural, facilitando o entendimento por parte das pessoas.

Existem diversas abordagens para a construção de classificadores baseados em árvores de decisão. Estas abordagens tem sido desenvolvidas, por meio da construção de algoritmos, para produzir árvores de decisão razoavelmente precisas, embora não perfeitas devido aos tamanhos de espaços de buscas e custo computacional deste tipo de classificador (TAN, STEINBACH e KUMAR, 2009). Estes algoritmos geralmente empregam uma estratégia que cresce uma árvore de decisão por meio de uma série de estratégias locais, sobre qual atributo usar para particionar os dados. A Figura 4 mostra um exemplo de uma árvore de decisão para a classificação de um aluno em situação de aprovado ou reprovado, de acordo com as notas de avaliações (provas).

Figura 4: Exemplo de Fluxograma Árvore de Decisão



Fonte: Elaborada pelo Autor (2016)

As folhas indicam uma classe e os nós de decisão, os quais definem algum teste condicional sobre o valor de um atributo específico (ou de um subconjunto de atributos), com um ramo e sub-árvore para cada um dos valores possíveis do teste. Na construção da árvore de decisão, programam associar **cada nó de decisão ao atributo "mais informativo"** entre aqueles que ainda não foram utilizados no caminho desde a raiz da árvore.

Existem vários algoritmos para a implementação de árvores de decisão. Um deles é o algoritmo de Hunt, que é a base de muitos algoritmos de indução de árvores de decisão existentes na literatura, incluindo o ID3, C4.5 e CART. Nesta tese foi utilizado o algoritmo CART (*Classification and Regression Trees*), foi proposto por Breiman, Friedman, et al.,(1984) e consiste de uma técnica não-paramétrica que induz tanto árvores de classificação quanto árvores de regressão, dependendo se o atributo é nominal (classificação) ou contínuo (regressão).

Dentre as principais virtudes do CART está a grande capacidade de pesquisa de relações entre os dados, mesmo quando elas não são evidentes, bem como a produção de resultados sob a forma de árvores de decisão de grande simplicidade e legibilidade (FONSECA, 1994).

✓ Naive Bayes

Outro algoritmo de classificação muito utilizado é o Naive Bayes. No trabalho de Anaya e Boticario (2011), os autores utilizaram o algoritmo de classificação Naive Bayes para modelar o comportamento colaborativo dos estudantes. O Nayve Bayes é um classificador probabilístico e tem suas raízes no teorema de Bayes, com princípios estatísticos para combinar conhecimento prévio das classes com novas evidências colhidas dos dados (TAN, STEINBACH e KUMAR, 2009).

Um classificador Bayesiano avalia a probabilidade condicional de classe, supondo que os atributos sejam condicionalmente independentes, dado o rótulo de classe y . A suposição de independência condicional pode ser declarada formalmente da seguinte maneira:

$$P(X|Y = y) = \prod_{i=1}^d P(X_i|Y = y) \quad (3.1)$$

onde cada conjunto de atributos $X = \{X_1, X_2, \dots, X_d\}$ consiste de d atributos. Dada a suposição de independência condicional, é estimada a probabilidade condicional de cada X_i , dado Y . Uma das vantagens é que esse classificador não requer um conjunto de treinamento grande para obter estimativas boas de probabilidade.

Desta forma, para classificar um registro de teste, o classificador de Bayes calcula a probabilidade posterior para cada classe Y :

$$P(X|Y = y) = \frac{P(Y) \prod_{i=1}^d P(X_i|Y)}{P(X)} \quad (3.2)$$

Já que $P(x)$ é fixo para cada Y , escolhe-se a classe que maximiza o numerador, $P(Y) \prod_{i=1}^d P(X_i|Y)$.

O desempenho do classificador de Naive Bayes é satisfatório em situações nas quais as variáveis de entrada são de natureza categóricas. As situações nas quais as variáveis são numéricas exigem a suposição de normalidade dos dados.

✓ Regressão Logística

Em aplicações de mineração de dados educacionais os modelos de regressão logística têm atingido bons resultados. Nos trabalhos de Sabourin, Mott E Lester (2012); Sabourin, Shores, et al., (2012) e Marbouti, Diefes-Dux E Madhavan (2016), os autores demonstram resultados promissores quanto ao uso do classificador de regressão logística para previsão de desempenho de estudantes, baseado em suas variáveis comportamentais.

A técnica de regressão logística foi especificamente elaborada para prever a probabilidade de um evento binário ocorrer. Esta técnica destina-se a investigar o efeito das variáveis pelas quais os indivíduos, objetos ou sujeitos, estão expostos sobre a probabilidade de ocorrência de determinado evento de interesse (HAIR, BLACK, et al., 2009). Daí a popularidade do uso desta técnica, pois há uma infinidade de eventos de interesse que poderiam ser modelados pela regressão logística.

Uma das grandes vantagens da regressão logística, diante de outras técnicas de classificação probabilísticas, é a flexibilidade de seus pressupostos, o que amplia sua aplicabilidade (FAVEIRO, BELFIORE, et al., 2009). A regressão logística deriva seu nome da transformação *logit* usada como variável dependente. Um modelo é definido como logístico se a função segue a seguinte equação:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (3.3)$$

Em que p_i indica a probabilidade de ocorrência de determinado evento de interesse, x_1, \dots, x_n representa o vetor de variáveis explicativas (ou independentes) e β_0 e β_x indicam os coeficientes do modelo encontrados por meio do método de máxima verossimilhança.

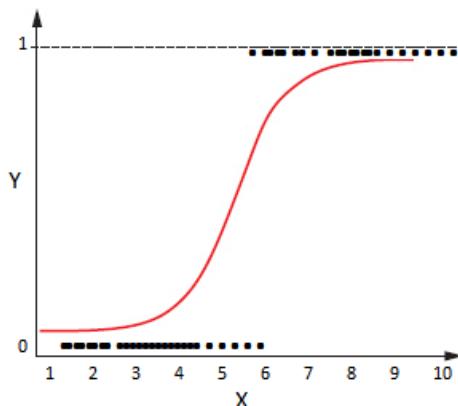
Os coeficientes logísticos são difíceis de interpretar em sua forma original (3.3), pois são expressos em termos de logaritmos quando usamos a função *logit* (HAIR,

BLACK, et al., 2009). Desta forma é possível aplicar a transformação de anti-logaritmo por meio da exponenciação dos coeficientes originais, gerando a razão de desigualdades, onde o modelo logístico pode ser expresso pela seguinte equação:

$$\text{Razão de Desigualdades}_i (\text{odds}) = \left(\frac{p_i}{1-p_i} \right) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n} \quad (3.4)$$

Para cada observação, é previsto um valor de probabilidade entre 0 e 1, onde os valores previstos para todos os valores da variável independente geram a curva logística mostrada na Figura 5.

Figura 5: Exemplo de representação gráfica da curva logística



Fonte: Elaborada pelo Autor (2016)

Se a probabilidade prevista é maior do que 0,50, então a previsão é de que o resultado seja 1 (evento ocorreu); caso contrário o resultado é previsto como sendo 0 (evento não ocorreu).

✓ Máquina de Vetor de Suporte (SVM)

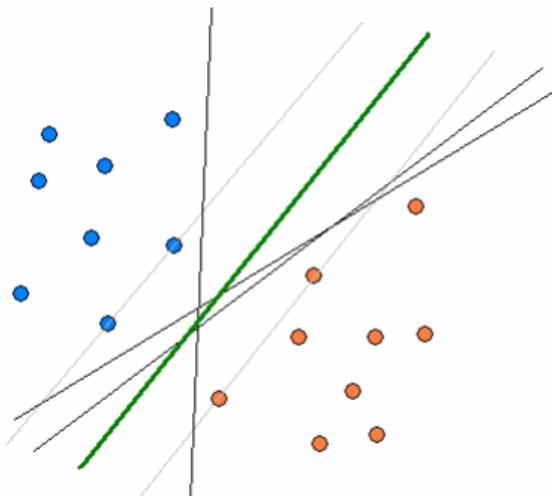
A técnica de Máquina de Vetor de Suporte (SVM, do Inglês *Support Vector Machine*) é uma técnica de aprendizado que vem recebendo crescente atenção da comunidade de Aprendizagem de Máquina e tem mostrado resultados promissores em muitas aplicações práticas (TAN, STEINBACH e KUMAR, 2009).

Em trabalhos de Mineração de Dados Educacionais, a técnica de SVM foi aplicada no artigo de Cetintas, Si, et al., (2014). Os autores buscaram estimar o nível de dificuldade de problemas matemáticos implementados em um software educacional por meio de um classificador SVM.

É uma técnica que funciona bem com **dados de alta dimensionalidade** e apresenta limite de decisão usando um **subconjunto** de exemplos de treinamento, conhecido como os vetores de suporte.

Foi desenvolvida inicialmente para a classificação de alvo binário, busca construir um hiperplano como superfície de decisão, maximizando a separação entre as observações. Considerando o exemplo da Figura 6, vemos que existem várias possibilidades de classificações lineares que separam as observações das classes azul e laranja, mas somente um hiperplano maximiza a separação. Este classificador linear é chamado “**hiperplano de separação ótimo**”.

Figura 6: Exemplo de um classificador SVM no hiperplano



Fonte: Elaborada pelo Autor (2016)

Este exemplo é restrito aos casos nos quais os **dados de treinamento são separáveis linearmente**. Para casos nos quais os dados **não são linearmente separáveis**, existe uma solução, que consiste em mapear **os dados para um espaço de dimensão maior (espaço de características)**. Utilizando uma função **Kernel** (CRISTIANINI e SHAWE-TAYLOR., 1999), é possível fazer o mapeamento no espaço de características, onde as SVM constroem um hiperplano de separação ótimo neste espaço.

3.4.1.1 Métricas de Avaliação de classificadores

Para entender os erros gerados por um classificador é possível visualizar por meio da construção de uma matriz de erros denominada **matriz de confusão**. A partir desta matriz é possível obter métricas de qualidade para a avaliação do desempenho de um sistema de aprendizagem. A seguir são apresentadas as principais métricas de avaliação de classificadores.

- **Matriz de Confusão**

A avaliação de um modelo de classificação é baseada na análise da matriz de confusão. Esta matriz (Quadro 4) resume o número de instâncias previstas corretas ou incorretas por um modelo de classificação. As contagens podem ser expressas em termos de porcentagens.

Quadro 4: Matriz de confusão para um problema binário

Matriz de Confusão		Classe Atual	
		Negativa (-)	Positiva (+)
Classe Prevista	Negativa (-)	$f_{--} (TN)$	$f_{+-} (FN)$
	Positiva (+)	$f_{-+} (FP)$	$f_{++} (TP)$

Fonte: Faveiro, Belfiore, et al., (2009)

Onde as seguinte terminologias são usadas para o entendimento da matriz de confusão:

- ✓ **Positivo verdadeiro (TP):** é relacionado ao número de instâncias positivas previstas corretamente pelo classificador.
- ✓ **Negativo falso (FN):** é o número de instâncias previstas erroneamente como negativos pelo classificador.
- ✓ **Positivo falso (FP):** é o número de exemplos negativos previstos erroneamente como positivos pelo classificador.
- ✓ **Falso verdadeiro (TN):** é o número de exemplos negativos previstos corretamente pelo classificador.

Uma das maneiras mais comuns de avaliar modelos é por meio da derivação de medidas que, de alguma maneira, tentam medir a “qualidade” do modelo. Essas medidas geralmente podem ser obtidas a partir da matriz de confusão, reduzindo suas quatro células principais a um único índice numérico de qualidade. A seguir descrevemos algumas das principais medidas derivadas da matriz de confusão.

A taxa de *positivos verdadeiros* (TPR) ou **sensibilidade** é definida como a fração das instâncias positivas previstas corretamente pelo classificador.

$$TPR = \frac{TP}{(TP + FN)} \quad (3.5)$$

A taxa de negativos verdadeiros (TNR) ou **especificidade** é definida como a fração de instâncias negativas previstas corretamente pelo classificador.

$$TRN = \frac{TN}{(TN - FP)} \quad (3.6)$$

A taxa de **positivos falsos** (FPR) é a fração de instâncias negativas previstas como uma classe positiva.

$$FPR = \frac{FP}{(TN - FP)} \quad (3.7)$$

A taxa de **negativos falsos** (FNR) é a fração de instâncias positivas previstas como uma classe negativa.

$$FNR = \frac{FN}{(TP - FN)} \quad (3.8)$$

A **precisão** e a **lembraça** são duas outras métricas muito utilizadas em aplicações nos quais existe o interesse em uma classe específica. A precisão determina o percentual de registros que são positivos no grupo que o classificador previu como classe positiva.

$$Precisão = \frac{TP}{TP + FP} \quad (3.9)$$

Quanto maior o percentual de precisão, menor será o número de erros positivos falsos pelo classificador.

A lembrança (*Recall*) mede o percentual de instâncias positivas previstas corretamente pelo classificador.

$$Recall = \frac{TP}{TP + FN} \quad (3.10)$$

O valor do *Recall* é equivalente a taxa de positivos verdadeiros. O grande desafio dos algoritmos de classificação é obter métricas com o melhor valor, tanto da precisão quanto da lembrança (*Recall*) (TAN, STEINBACH e KUMAR, 2009).

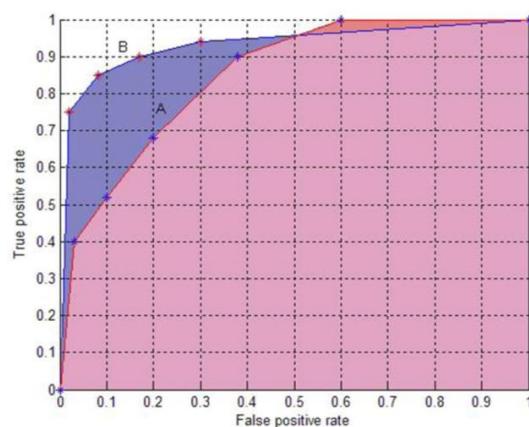
- Área da Curva ROC

Os gráficos ROC foram originalmente utilizados em detecção de sinais, para se avaliar a qualidade de transmissão de um sinal em um canal com ruído (EGAN, 1975). É uma representação gráfica que vem sendo bastante utilizada para descrever o desempenho de um sistema classificador binário e, além disso, permite estudar a variação da sensibilidade e especificidade para diferentes valores de corte (FAWCETT, 2006).

Uma das primeiras aplicações dos gráficos de curva ROC em aprendizado de máquina foi apresentado em Spackman (1989), que demonstrou a utilidade das curvas ROC na avaliação e comparação entre algoritmos. Estes gráficos são baseados na taxa de verdadeiros positivos $TPR = (P/X)$, e na taxa de falsos positivos $FPR = (P/\bar{X})$, onde a construção do gráfico é feita com a plotagem de FPR no eixo das ordenadas (eixo y) e TPR no eixo das abscissas (eixo x).

Esta plotagem forma uma demonstração bidimensional da performance de um classificador (SCHIAVONI, 2015). Como a área da curva ROC é uma porção da área do quadrado unitário (espaço ROC) seus valores variam entre zero e um. Estes valores compararam as performances de classificadores, quanto maior a área AUC melhor a performance global do classificador. A Figura 7 mostra a área abaixo de duas curvas ROC, A e B.

Figura 7: Área abaixo das curvas ROC A e B



Fonte: Schiavoni (2015)

Em algumas situações é possível que um classificador mostre-se melhor que outro devido às regiões do espaço ROC. Na Figura 7, o classificador B demonstra ser melhor que o classificador A, exceto para fatores de falso positivos acima de 0.6, onde

o classificador *A* tem uma pequena vantagem. De acordo com Faveiro, Belfiore, *et al.* (2009), existem quatro faixas de valores de referência para classificar a área da curva ROC. O classificador é tido como excelente se o valor da área é maior do que 0,8. Tem discriminação aceitável se estiver entre 0,7 e 0,8 e não tem poder de discriminação se estiver abaixo de 0,5.

- **Kolmogorov-Smirnov**

Outra medida de avaliação da qualidade de ajuste de modelo bastante utilizada é o *K-S* (*Kolmogorov-Smirnov*), que mede o grau de segregação de dois grupos, cujo valor pode ser interpretado conforme Faveiro, Belfiore, *et al.* (2009), onde os autores afirmam que valores de *K-S* acima de 0,5 indicam uma discriminação excelente. Valores entre 0,3 e 0,5 indicam uma boa discriminação e valores abaixo de 0,3 indicam uma baixa discriminação.

- **Utilização de métricas conjuntamente**

É comum encontrar situações nas quais um modelo aparenta ser melhor que outro para algumas das métricas, mas pior com relação a outras. Nesses casos, utilizar uma única medida pode dar a falsa impressão de que o desempenho pode ser avaliado utilizando-se apenas essa medida (PRATI, BATISTA e MONARD, 2008).

Desta forma, métricas que tenham como objetivo reduzir a avaliação de um modelo de classificação a um único valor tendem, em maior ou menor grau, a uma perda de informação e a avaliação de um modelo utilizando uma única medida pode levar a conclusões errôneas. Para uma avaliação mais precisa é ideal que seja utilizado um conjunto de métricas levando em consideração o objetivo da pesquisa.

3.4.2 Agrupamento

Em EDM técnicas de mineração de dados, tais como o agrupamento são frequentemente usadas para criação de perfis dos alunos com base nas suas atividades em ambientes de aprendizagem online (BIENKOWSKI, FENG e MEANS, 2012).

O agrupamento é normalmente um processo não supervisionado para agrupar elementos semelhantes em clusters. Esses grupos identificados podem ser usados

para classificar novas instâncias e permitir adaptações em tempo real (HARLEY, TREVORS e AZEVEDO, 2013).

Vários trabalhos sobre o uso da técnica de agrupamento na análise de dados educacionais foram publicados recentemente, dentre os quais destacamos: (NUGENT, DEAN e AYERS, 2010), (ANTONENKO, TOY e NIEDERHAUSER, 2012), (LOPEZ, LUNA, *et al.*, 2012), (RAD, NADERI e SOLTANI, 2011), (YANTO, VITASARI, *et al.*, 2012), (TRIVEDI, PARDOS, *et al.*, 2012) e (AKCAYNAR, ALTUN e COSGUN, 2014).

O trabalho de Akcapynar, Altun e Cosgun (2014) identificou grupos de alunos que interagem com um ambiente de aprendizagem online de maneira semelhante. Os resultados da análise revelaram a existência de três grupos distintos de alunos, marcados pela interação (não-ativo, ativo, muito ativo) e pelo seu sucesso no curso (baixa aprendizagem, média aprendizagem e alta aprendizagem).

López et al. (2012) conceberam uma abordagem de classificação via tarefa de agrupamento para prever as notas finais de um curso com base em dados do fórum de discussão. Os resultados do trabalho apontaram que o algoritmo de agrupamento expectation–maximization (EM) produziu resultados semelhantes aos algoritmos de classificação.

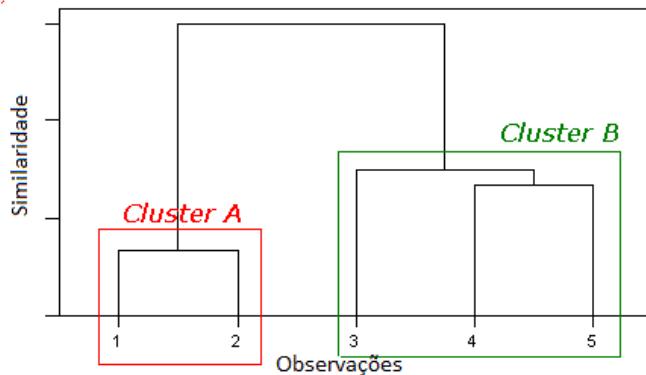
Antonenko (2012) mineraram logs de clicks que refletem o uso do estudante de ambientes de aprendizagem online. Eles fizeram uso de um método de agrupamento hierárquico (Método de Ward) e um método de agrupamento não-hierárquico (*k*-Means) para analisar características do comportamento de aprendizagem, enquanto os alunos se envolveram em uma atividade de resolução de problemas no ambiente de aprendizagem online. Basicamente as técnicas de agrupamento se dividem em dois tipos: as técnicas de agrupamento hierárquico, mais utilizadas em data mining e as técnicas não-hierárquicas mais utilizadas em abordagens estatísticas.

➤ Agrupamento Hierárquico

Os métodos hierárquicos, como o próprio nome diz, envolvem a construção de uma hierarquia aglomerativa ou divisiva, nos quais as observações vão sendo combinadas passo a passo, e que não há um número predefinido de grupos que serão formados (WARD, 1963). Os resultados finais desses agrupamentos podem ser apresentados por uma árvore de classificação chamada de dendrograma. Existem

duas versões para a realização de agrupamento hierárquico. A aglomerativa, que opera criando conjuntos a partir de parcelas isoladas e a divisiva, que começa com um grande conjunto e vai quebrando-o em partes até chegar a parcelas isoladas. A Figura 8 mostra um exemplo de um dendrograma.

Figura 8: Exemplo de Dendrograma



Fonte: Ramos, Silva, et al., (2016)

Para Mingoti (2005), as técnicas hierárquicas aglomerativas partem do princípio de que, no início do processo de agrupamento, tens n grupos, ou seja, cada parcela do conjunto de dados observado é considerado como sendo um agrupamento isolado. O propósito é encontrar o número k que esteja associado à "partição natural" das parcelas que estão sendo comparadas e agrupadas.

Muitas vezes, os algoritmos hierárquicos são usados de forma exploratória, e a solução resultante é submetida a um método não hierárquico para refinar ainda mais a solução.

➤ Agrupamento Não-Hierárquico

As técnicas não-hierárquicas têm alcançado crescente aceitabilidade e são cada vez mais aplicadas em diversas áreas. Os métodos não-hierárquicos diferem dos hierárquicos em vários aspectos. Primeiramente, requerem que o usuário especifique, a priori, o número de agrupamentos k desejado, onde o pesquisador pode optar por meio de algum conhecimento, pela conveniência, por simplicidade ou pelo método hierárquico.

Os métodos não hierárquicos dividem o conjunto de objetos de uma base de dados em um número pré-definido de subconjuntos disjuntos e procuram iterativamente o melhor particionamento até atingir uma condição de parada (RAMOS, SILVA, et al., 2016). Esse método exige a prefixação de critérios que produzam

medidas relativas à qualidade da partição produzida e não é utilizado, a construção de dendrogramas.

Quando comparado com o método hierárquico, o não hierárquico é mais rápido, porque nele não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade (FERREIRA, 2008). Existem diversos algoritmos de agrupamento não-hierárquico, um dos mais utilizados é o algoritmo *K-means* (MACQUEEN, 1967).

3.4.3 Regressão Linear

A regressão linear é uma tentativa de modelar uma equação matemática linear que descreva o relacionamento entre duas variáveis (CURRAL, 1994). Em EDM a técnica de **Regressão Linear é frequentemente usada para a construção de modelos preditivos relacionados ao desempenho de alunos.**

Por exemplo, um pesquisador em EDM pode tentar explicar as variações do desempenho de alunos em função do **aumento da carga horária diária de estudos.** Deve-se notar, entretanto, que a lógica de uma relação causal deve advir de fenômenos externos ao âmbito da estatística. A análise estatística de regressão apenas modela qual relacionamento matemático pode existir, se existir algum (HOFFMANN, 2015).

No trabalho de González-Brenes e Mostow (2010), os autores buscaram desenvolver um modelo de regressão linear para estimar o **tempo de conclusão de uma determinada atividade escolar, bem como estimar o tempo que um aluno irá desistir da realização da mesma.** Os autores buscaram detectar os **fatores que influenciam a conclusão da tarefa**, proporcionando assim uma orientação útil para o envolvimento do professor na intervenção com os alunos.

Embora as relações de regressão possam assumir uma grande diversidade de formas, demonstramos neste trabalho apenas as equações lineares. Estas equações são importantes porque servem para aproximar muitas relações da vida real, e são relativamente fáceis de lidar e de interpretar.

Duas importantes características da equação de regressão linear são: **o coeficiente angular e o coeficiente linear da reta em determinado ponto.** Uma equação linear tem a forma:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \mu \quad (3.11)$$

Onde Y é a variável que representa o fenômeno em estudo (variável dependente), α representa o intercepto (constante), β_k ($k = 1, 2, \dots, n$) são os coeficientes de cada variável (coeficientes angulares), X_k são as variáveis explicativas e μ é o termo do erro ou resíduos, ou seja, é a diferença entre o valor real de Y e o valor previsto de Y .

É importante mencionar que nem todas as situações são bem aproximadas por uma equação linear (HOFFMANN, 2015). Por isso, em geral é necessário desenvolver um trabalho preliminar para determinar se um modelo linear é adequado. O processo mais simples consiste em fazer o diagrama de dispersão dos dados, isto é, verificar se os valores de X e de Y , mostram uma tendência visualmente linear.

3.4.4 Regras de associação

As regras de associação, também denominada de regras associativas, foram introduzidas por Agrawal, Imielinski e Swami (1993). Intuitivamente essa tarefa consiste em encontrar conjuntos de itens que ocorram simultaneamente e de forma frequente em um banco de dados. A tarefa de Descoberta por Associação pode ser definida formalmente como a busca por regras de associação frequentes e válidas em um banco de dados, a partir da especificação dos parâmetros de suporte e confiança mínimos (GOLDSCHMIDT e PASSOS, 2005).

Existem diferentes técnicas que são classificadas como técnicas de regras de associação, tais como a mineração de regras de associação (qualquer relação entre as variáveis), mineração de padrões sequenciais (associações temporais entre variáveis), mineração de correlações (correlação linear positiva ou negativa entre as variáveis), e mineração de causas (relação de causalidade entre as variáveis). Em EDM, a mineração de associação tem sido usada para identificar relacionamentos em padrões de comportamento de alunos, com o objetivo de identificar perfis de alunos com dificuldades de aprendizagem (MERCERON e YACEF, 2005).

3.5 Considerações sobre o capítulo

Esta capítulo discutiu as diferentes tarefas que vêm sendo utilizadas em pesquisas na área de Mineração de Dados Educacionais. Percebemos, por meio deste capítulo, que o maior número de pesquisas em EDM utilizam tarefas de classificação, seguidos por trabalhos que utilizam tarefas de agrupamento.

Diversas outras tarefas vêm sendo utilizadas em pesquisas em EDM, tais como técnicas de SNA, técnicas provindas da estatística, sistemas de recomendação, dentre outras. Esta seção focou nas principais técnicas provindas da área de Mineração de Dados, devido à extensa quantidade de técnicas que vem sendo empregadas em pesquisas com dados educacionais.

No próximo capítulo explicitaremos os principais trabalhos e técnicas, identificados por meio de um mapeamento sistemático da literatura, o que buscou identificar um conjunto de trabalhos que utilizaram abordagens quantitativas para a mensuração de comportamentos de autorregulação da aprendizagem.

4. MAPEAMENTO SISTEMÁTICO DA LITERATURA

Neste capítulo, explicitamos as etapas e resultados do mapeamento sistemático realizado nesta tese, com o objetivo de identificar trabalhos relacionados e lacunas de pesquisa que tenham relação com a aplicação de métodos para a mensuração de características da autorregulação da aprendizagem em ambientes de LMS.

4.1 Processo do Mapeamento Sistemático

O estudo de Mapeamento Sistemático (MS) é um método projetado para fornecer uma visão ampla de uma determinada área de pesquisa, permitindo identificar, quantificar e analisar os resultados, estabelecendo evidências da mesma e lacunas de pesquisa sobre um determinado tema (KITCHENHAM, 2007; ARKSEY e O'MALLEY, 2005).

O MS é considerado como um estudo secundário, visto que é uma revisão mais ampla dos estudos primários, tornando-o dependente da realização desses estudos para revelar as evidências da pesquisa (KITCHENHAM, 2007). A metodologia utilizada para fazer um estudo de MS é apoiada na mesma visão da metodologia de uma Revisão Sistemática, pois é conduzida de maneira rigorosa e transparente (ARKSEY e O'MALLEY, 2005).

As próximas seções mostram o detalhamento da metodologia utilizada na construção do mapeamento sistemático desta tese, bem como os resultados encontrados e conclusões.

4.1.1 Objetivo do mapeamento

O objetivo deste mapeamento sistemático foi identificar, catalogar, e classificar os métodos utilizados para mensurar características de autorregulação da aprendizagem em ambientes de LMS, com o intuito de identificar lacunas de pesquisa que subsídinem e reforcem a questão de pesquisa desta tese.

4.1.2 Questões de pesquisa

Uma das etapas essenciais do Mapeamento Sistemático é a definição das questões de pesquisa que conduzem a busca dos documentos relevantes, permitindo posteriormente a triagem de documentos, *keywording* de *abstracts* e extração dos dados (PETERSEN, 2007). Partindo do objetivo, buscamos responder a questão central do estudo, descrita como: “*Quais os principais métodos utilizados na literatura para a mensuração das características da autorregulação da aprendizagem em ambientes de LMS ?*”.

Foram elaboradas quatro questões secundárias que tiveram como intuito auxiliar nas seguintes indagações:

- **Q1:** Qual a natureza das principais pesquisas que vêm sendo desenvolvidas com o tema de Autorregulação da Aprendizagem?
- **Q2:** Quais as principais formas de coleta, no que diz respeito aos construtos de SRL, em ambientes de LMS?
- **Q3:** Quais as principais técnicas que estão sendo utilizadas para analisar e mensurar as características de SRL?
- **QS4:** Quais as principais abordagens de pesquisa que dedicam esforços para o desenvolvimento e aplicação de técnicas de mensuração de SRL?

Estas questões serviram de direcionamento para a elaboração de uma *string* de busca que será descrita na Subseção 4.1.4. A subseção seguinte descreve os critérios de inclusão e exclusão que foram utilizados para a seleção dos trabalhos.

4.1.3 Critérios de Inclusão e Exclusão

Para a inclusão de um trabalho na pesquisa foi determinada sua relevância em relação às questões de investigação por meio da análise do título, palavras-chave e resumo. Dois critérios de inclusão foram levados em consideração nessa análise: (1) o primeiro é que o trabalho deveria ter explicitamente o método/técnica que foi utilizado para mensurar características de autorregulação da aprendizagem, e o segundo (2) que o trabalho apresentasse uma abordagem de mensuração com foco em ambientes de LMS. Trabalhos que não apresentaram tais características foram excluídos no processo de seleção. Na busca retornada nesta segunda *string* foram excluídos apenas dois trabalhos.

4.1.4 Estratégia de busca

A fonte de pesquisa utilizada para a busca dos estudos primários foi a base *Elsevier Scopus* (<http://www.scopus.com>), uma vez que este repositório também inclui as bases da IEEE Digital Library, ACM Digital Library e ScienceDirect. Foram realizadas duas interações por meio de uma primeira e uma segunda *string* de busca. O Quadro 5 mostra a primeira *string* de busca:

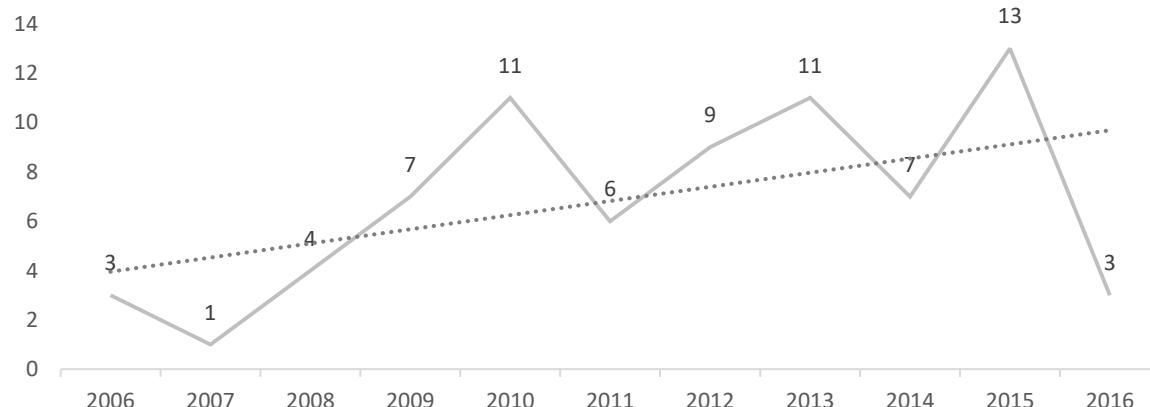
Quadro 5: Primeira String de Busca

1ª String de Busca
TITLE-ABS-KEY(("Self-Regulated Learning" OR "SRL") AND ("learning management system" OR "online learning environment" OR "computer supported collaborative learning" OR "LMS" OR "eLearning") AND (EXCLUDE (LANGUAGE , "Chinese") OR EXCLUDE (LANGUAGE, "Croatian")) AND (EXCLUDE (DOCTYPE, "cr") OR EXCLUDE (DOCTYPE,"ip") OR EXCLUDE (DOCTYPE, "re")))

Fonte: Elaborada pelo Autor (2016)

Esta busca teve como resultado 75 trabalhos primários. A Figura 9 mostra o gráfico com os trabalhos distribuídos por ordem cronológica de publicações nos últimos dez anos.

Figura 9: Quantidade de publicações na área nos últimos dez anos

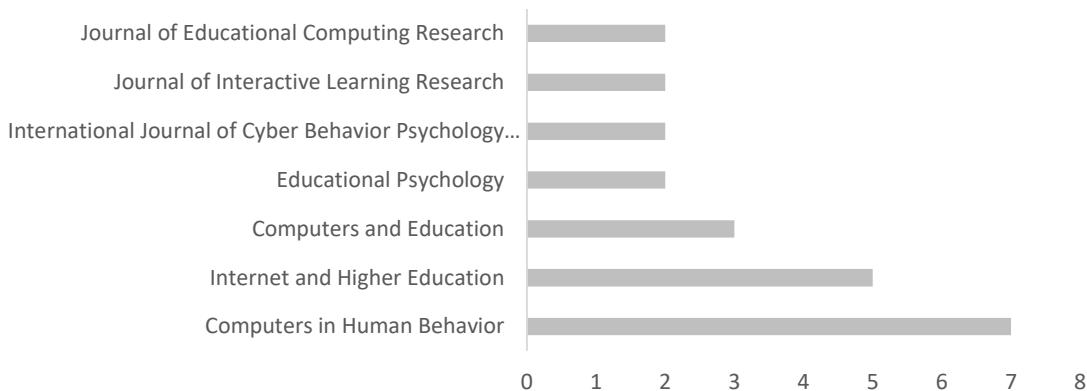


Fonte: Elaborada pelo Autor (2016)

Dentre os trabalhos retornados pelo engenho de busca Scopus, 49 (65,3%) foram artigos em periódicos, 23 (30,7%) foram artigos publicados em conferências e 3 (4%) capítulos de livros.

O principal periódico foi o *Journal Computers in Human Behavior*, o qual teve sete publicações na área de SRL aplicados em ambientes de LMS. A Figura 10 mostra a frequência de cada um dos periódicos.

Figura 10: Principais periódicos com publicações na área nos últimos dez anos



Fonte: Elaborada pelo Autor (2016)

O segundo periódico que teve mais publicações no tema foi o jornal *Internet and Higher Education*, obtendo um total de cinco publicações. Os demais periódicos tiveram menos de duas publicações nos últimos dez anos.

Segunda String de busca

Para especificar ainda mais a busca, limitando apenas a trabalhos que utilizaram abordagens quantitativas para mensuração de SRL, foi necessário desenvolver uma segunda *string* de busca adicionando os seguintes termos: *statistic*, *data mining*, *learning analytics* e *Machine Learning*.

Para a construção dos termos chaves da pesquisa foram utilizadas as palavras-chave da questão central. O Quadro 6 mostra a segunda *string* de busca.

Quadro 6: Segunda String de Busca

2ª String de Busca
<pre>TITLE-ABS-KEY(("Self-Regulated Learning" OR "SRL") AND ("learning management system" OR "online learning environment" OR "computer supported collaborative learning" OR "LMS" OR "eLearning")) AND ("questionnaire" OR "log" OR "data base") AND("statistic" OR "data mining" OR "learning analytics" OR "Machine Learning") AND(EXCLUDE(LANGUAGE,"Chinese") OR EXCLUDE(LANGUAGE,"Croatian")) AND(EXCLUDE(DOCTYPE,"cr") OR EXCLUDE(DOCTYPE,"ip") OR EXCLUDE(DOCTYPE,"re")))</pre>

Fonte: Elaborada pelo Autor (2016)

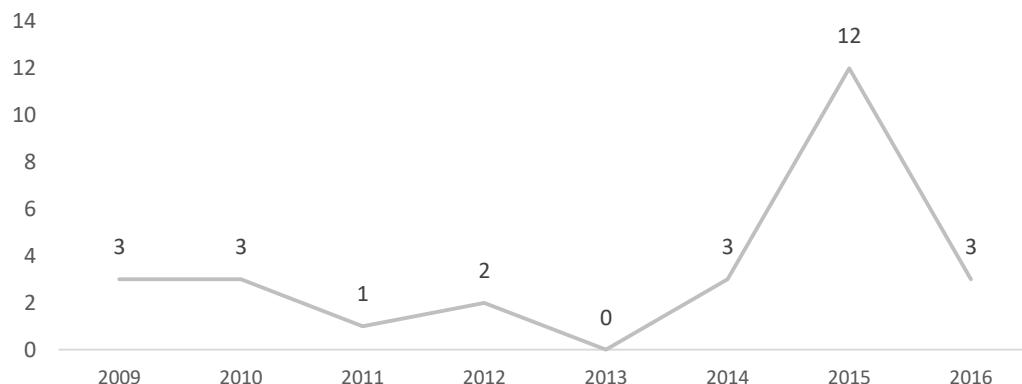
Nesta segunda *string* de busca tivemos como resultado 18 trabalhos primários. Dos 75 trabalhos encontrados na primeira *string* de busca, apenas 18 trabalhos utilizaram algum tipo de abordagem quantitativa para mensurar construtos de autorregulação da aprendizagem.

É importante mencionar que nesta segunda String, a pesquisa foi complementada com sete artigos de uma série especial (*Self-regulated learning and*

learning analytics) do *Journal of Learning Analytics* e mais dois trabalhos, um do periódico *Lecture Notes in Computer Science* e outro trabalho da *International conference on educational data mining*.

Na Figura 11 é possível visualizar os trabalhos distribuídos por ordem cronológica de publicações nos últimos oito anos. É interessante observarmos que o ano de 2015, foi o ano que obteve um maior número de trabalhos na área.

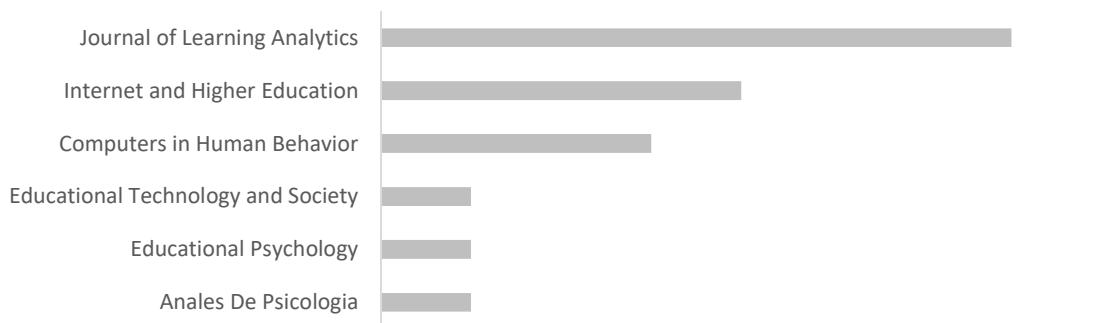
Figura 11: Quantidade de publicações na área nos últimos seis anos



Fonte: Elaborada pelo Autor (2016)

Dentre todos os trabalhos, 22 (91,66%) foram artigos em periódicos e apenas 2 (8,3%) artigos publicados em conferências.

Figura 12: Principais periódicos com publicações na área nos últimos seis anos



Fonte: Elaborada pelo Autor (2016)

O periódico que teve maior quantidade de publicações na área foi o *Journal of Learning Analytics*, com sete publicações, seguido pelos periódicos *Internet and Higher Education* e o *Computers in Human Behavior* com um total de sete publicações.

4.1.5 Análise e Interpretação dos Resultados

O processo foi iniciado com a leitura dos vinte e quatro artigos selecionados na segunda interação. Inicialmente foi feita a identificação dos objetivos e das metodologias de cada um dos artigos, selecionando informações relevantes para responder cada uma das questões de pesquisa.

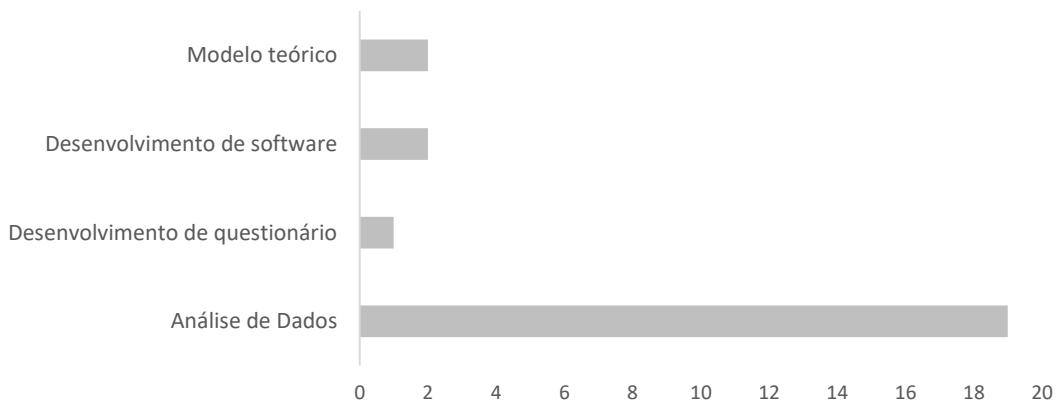
Primeira questão de pesquisa

A primeira questão de pesquisa focou na identificação da natureza das pesquisas que vem sendo desenvolvidas na área. A grande maioria das pesquisas que foram desenvolvidas foca exclusivamente em análise de dados de autorregulação da aprendizagem. Estas pesquisas tiveram como temas principais a aprendizagem colaborativa autorregulada, como pode ser visto nos trabalhos de Gašević, Adesope, *et al.*, (2015); Malmberg, Järvelä, *et al.*, (2015); Panadero, Kirschner, *et al.*, (2015) e no estudo das relações entre autorregulação da aprendizagem e desempenho acadêmico, onde os principais trabalhos foram Bondareva, Conati, *et al.* (2013); Ali, Hatala, *et al.* (2014) e Park e Lim (2015).

Uma característica interessante das pesquisas que trabalharam com análise de dados, especificamente com foco na previsibilidade do desempenho acadêmico de alunos, é que as pesquisas foram realizadas por meio de abordagens de coleta com questionários, nos quais foram utilizados questionários já validados na literatura, tais como o *MSLQ* e o *OSLQ* para a realização de inferências sobre o desempenho de estudantes.

Na Figura 13 é apresentado os principais tipos de pesquisas em autorregulação da aprendizagem.

Figura 13: Tipo de pesquisa em SRL



Fonte: Elaborada pelo Autor (2016)

Os três outros tipos de pesquisas encontradas na literatura foram, a validação de modelos teóricos de SRL, o desenvolvimento de software para suporte a SRL e o desenvolvimento de questionários.

Os artigos sobre validação de modelos teóricos de SRL de ROLL e WINNE (2015) e SANCHEZ-SANTILLAN, PAULE-RUIZ, *et al.* (2016) buscaram entender o estado atual da investigação de SRL, buscando identificar temas e desafios de apoio a SRL em ambientes de LMS. Apesar destes artigos não aplicarem nenhuma técnica de mensuração, foram considerados nesta pesquisa pois abordam diversos artigos que utilizam.

Dentre os artigos que trabalharam com desenvolvimento de software, um que merece destaque é o artigo desenvolvido por Nussbaumer, Hillemann, *et al.* (2015). Neste artigo, os autores apresentam uma abordagem conceitual e um serviço baseado na Web para apoio à aprendizagem autorregulada em ambientes virtuais. O trabalho fornece uma descrição da implementação técnica do modelo conceitual e uma validação de usabilidade com alunos de graduação.

Em relação a trabalhos que se propuseram a desenvolver questionário específico para mensurar SRL, tivemos o trabalho de Barnard, Lan, *et al.* (2009). Essa pesquisa é justificada pela ausência de um instrumento de coleta que aborde os construtos de SRL especificamente para aprendizagem online, na qual é realizado a construção de um instrumento com vinte e quatro questões que abordam seis construtos, tais como: estruturação do ambiente, busca por ajuda, estratégias para realização da tarefa, auto avaliação, gerenciamento do tempo e definição de metas.

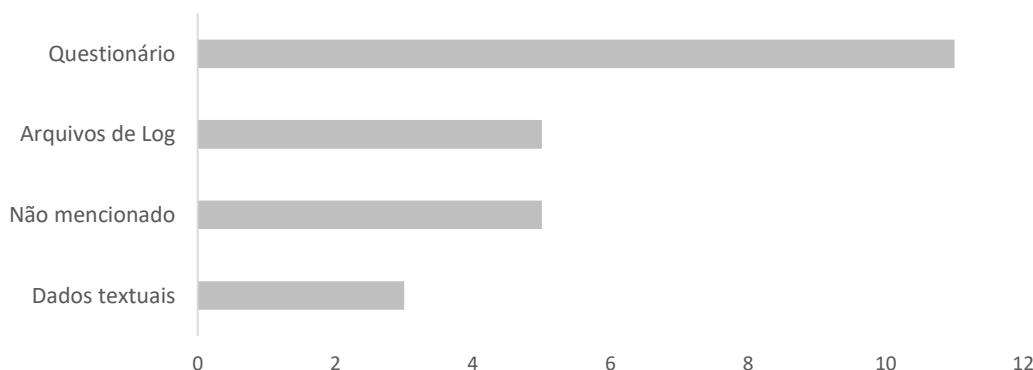
Estes construtos foram validados por meio de análise fatorial confirmatória que

buscou avaliar as propriedades psicométricas, onde resultados indicam que o instrumento OSLQ é uma medida aceitável de autorregulação nos ambientes de aprendizagem online.

Segunda questão de pesquisa

A segunda questão de pesquisa, desenvolvida neste trabalho, foi em relação a *quais as principais formas de coleta, no que diz respeito aos construtos de SRL, em ambientes de LMS?* Os resultados encontrados nesta segunda questão de pesquisa mostraram que, a grande maioria dos trabalhos que se propõem a mensurar SRL, utiliza questionários. Dos onze trabalhos que utilizaram esta abordagem, seis deles utilizaram o questionário MSLQ (YUKSELTURK e BULUT, 2009; ALI, HATALA, et al., 2014; ADESOPE, ZHOU e NESBIT, 2015; PANADERO, KIRSCHNER, et al., 2015; SANCHEZ-SANTILLAN, PAULE-RUIZ, et al., 2016). Este é o questionário mais utilizado na literatura, embora tenha sido construído para ser utilizado em ensino presencial. Outro instrumento que foi utilizado nas pesquisas com questionário foi o OSLQ, um questionário desenvolvido por Barnard, Lan, et al. (2009) com foco em características de SRL em ambientes de LMS. A Figura 14 apresenta as principais formas de coleta.

Figura 14: Forma de coleta de dados



Fonte: Elaborada pelo Autor (2016)

Uma segunda forma de coleta utilizada nos trabalhos foi por meio de arquivos de log. Os principais trabalhos que utilizaram essa abordagem para coleta de dados foram Schoor e Bannert (2012); Sabourin, Mott e Lester (2012); Bondareva, Conati, et al. (2013); Nussbaumer, Hillemann, et al. (2015); SEGEDY, Kinnebrew e Biswas (2015); Sonnenberg e Bannert (2015); You (2016). A maioria destes trabalhos utilizaram dados de softwares educacionais, de conteúdo específico para e-Learning.

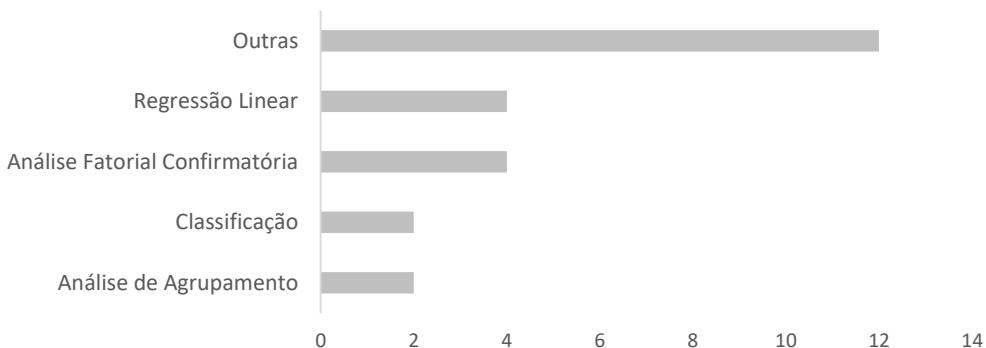
Um dos trabalhos que merece destaque é o trabalho do You (2016). Este foi o único dos trabalhos que utilizou arquivos de log provindos de ambientes de LMS. O estudo procurou identificar indicadores comportamentais para a aprendizagem autorregulada em sistemas de LMS como forma de previsão de sucesso dos alunos no término de um curso. Para tal, utilizou a regressão linear para construção de modelos e verificação de variáveis significativas. O trabalho possui algumas limitações com relação à escolha das variáveis explicativas. Foram utilizadas apenas variáveis relacionadas a quantidade de acesso, tempos de visualizações e atraso na entrega de atividades.

Terceira questão de pesquisa

A terceira questão de pesquisa foi elaborada com o objetivo de identificar *quais as principais técnicas que estão sendo utilizadas para analisar e mensurar as características de SRL?* As técnicas de análise mais utilizadas foram análise de regressão e análise fatorial, sendo quatro trabalhos que utilizaram regressão linear (YUKSELTURK e BULUT, 2009), (PARK e LIM, 2015), (YOU, 2016) e quatro trabalhos que utilizaram análise fatorial confirmatória (BARNARD, LAN, *et al.*, 2009), (CHO e JONASSEN, 2009), (ADESOPE, ZHOU e NESBIT, 2015), (ALI, HATALA, *et al.*, 2014).

A maioria dos trabalhos que utilizaram regressão linear como técnica principal, tiveram como foco o desenvolvimento de modelos de previsão, especificamente desempenho de alunos. Todos os trabalhos que utilizaram análise fatorial como técnica principal, tiveram como objetivo a validação de questionários ou a identificação de pesos ou cargas fatoriais dos construtos de SRL. A Figura 15 apresenta as principais técnicas.

Figura 15: Técnica utilizada para mensuração de SRL



Fonte: Elaborada pelo Autor (2016)

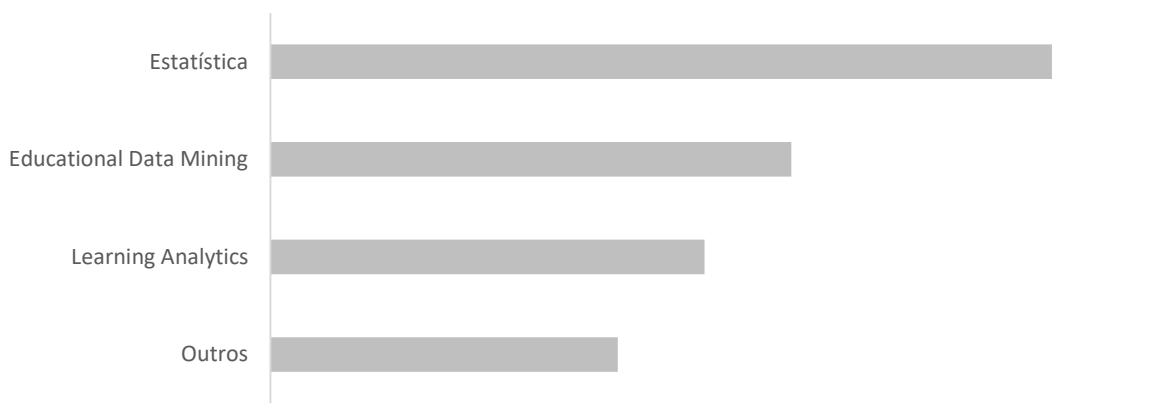
Outras duas técnicas que foram utilizadas foram a classificação e a análise de agrupamento, ambas com dois trabalhos encontrados na literatura. Os trabalhos de classificação (BONDAREVA, CONATI, *et al.*, 2013; SABOURIN, MOTT e LESTER, 2012) tiveram como foco o comparativo de múltiplos algoritmos de classificação de dados provindos de sistemas de tutores inteligentes e os trabalhos de análise de agrupamento (LAWANTO, SANTOSO, *et al.*, 2014; COLTHORPE, ZIMBARDI, *et al.*, 2015) tiveram o objetivo de identificar padrões ou perfis comportamentais de autorregulação.

Quarta questão de pesquisa

A quarta questão de pesquisa foi a identificação de *quais as principais abordagens de pesquisa que dedicam esforços para o desenvolvimento e aplicação de técnicas de mensuração de SRL?* A abordagem baseada em técnicas provindas da estatística foi a que teve o maior número de trabalhos, os quais utilizaram técnicas como Testes de comparação de médias (GAŠEVIĆ, ADESOPE, *et al.*, 2015; SEGEDY, KINNEBREW e BISWAS, 2015), Análise fatorial (BARNARD, LAN, *et al.*, 2009; CHO e JONASSEN, 2009; ALI, HATALA, *et al.*, 2014 e ADESOPE, ZHOU e NESBIT, 2015;) e Análise de regressão linear (YUKSELTURK e BULUT, 2009; PARK e LIM, 2015; YOU, 2016).

Em seguida tiveram trabalhos que utilizaram as abordagens baseada em Mineração de Dados Educacionais, os quais utilizaram técnicas como Mineração de Processos (SCHOOR e BANNERT, 2012; SONNENBERG e BANNERT, 2015), Análise de agrupamento (LAWANTO, SANTOSO, *et al.*, 2014; COLTHORPE, ZIMBARDI, *et al.*, 2015) e Algoritmos de Classificação (SABOURIN, MOTT e LESTER, 2012; BONDAREVA, CONATI, *et al.*, 2013).

Figura 16: Área de pesquisa declarada na análise dos dados



Fonte: Elaborada pelo Autor (2016)

Outros quatro trabalhos utilizaram técnicas declaradas como sendo provindas da área de *Learning Analytics*. É importante mencionarmos que uma mesma técnica pode ser considerada pertencer a mais de uma abordagem de pesquisa, como por exemplo, a técnica de Análise de agrupamento tem suas origens na Estatística, no entanto a área de mineração de dados e *Learning Analytics* têm evoluído e desenvolvido características e algoritmos diferenciados para sua aplicação.

A seção seguinte descreve as principais considerações obtidas com este mapeamento sistemático, bem como lacunas de pesquisa que serão abordados neste trabalho de tese.

4.2 Considerações sobre o capítulo

A análise dos trabalhos pesquisados e selecionados neste mapeamento sistemático, sobre os métodos utilizados para mensurar características de autorregulação da aprendizagem em ambientes de LMS, teve como objetivo identificar lacunas de pesquisa que subsidiem e reforcem a questão de pesquisa desta tese.

Na primeira interação observou-se que a área de pesquisa que busca entender o fenômeno de autorregulação da aprendizagem em ambientes de LMS tem aumentado ao longo dos anos, embora a grande maioria dos trabalhos encontrados na primeira interação tinha como abordagens de mensuração métodos qualitativos de análise. A segunda interação focou a busca por artigos que utilizassem apenas abordagens quantitativas de mensuração, dentro das áreas de estatística, mineração de dados, *learning analytics* e aprendizagem de máquina.

Um ponto que merece destaque nos resultados encontrados é a predominância de dois tipos de coleta de dados: as coletas baseadas em questionários e as coletas baseadas em arquivos de log. Os trabalhos que utilizaram questionário como instrumento principal de coleta, limitaram-se a utilizar adaptações de questionários desenvolvidos especificamente para o ensino presencial, com exceção dos trabalhos que utilizaram o único questionário desenvolvido para a aprendizagem em ambientes online.

Outra limitação, dos trabalhos encontrados, refere-se à transposição dos modelos teóricos para as pesquisas com análise de dados. A literatura de SRL, no decorrer dos anos, vem desenvolvendo modelos teóricos e geração de construtos de autorregulação da aprendizagem. No entanto, observamos que as pesquisas que buscam confirmar estas características, por meio de análises quantitativas, não utilizam os construtos teóricos baseados nos modelos já desenvolvidos na literatura.

A grande maioria dos artigos que trabalharam com mensuração de desempenho utilizam questionários como o instrumento principal de coleta dos construtos de SRL. Apenas um dos trabalhos utilizou uma abordagem de coleta por meio de arquivos de *log* dos ambientes de LMS especificamente. No entanto, esse trabalho apresentou limitações no que diz respeito à seleção das variáveis que explicam os construtos de SRL. Foram utilizadas apenas variáveis relacionadas a quantidade de acesso, tempos de visualizações e atraso na entrega de atividades. Estas variáveis não compõem nenhum dos modelos teóricos desenvolvidos na literatura para os construtos de SRL.

Não foi encontrado nenhum artigo com foco no desenvolvimento de soluções de software que deem suporte para professores ou tutores, no que diz respeito à identificação de alunos com dificuldade de autorregulação. Identificar estados autorregulatórios de aprendizagem e fornecer subsídios para professores e tutores a partir da análise de dados em tempo real parece ser uma lacuna de pesquisa para este trabalho. É um tema desafiador especialmente considerando o número cada vez maior de cursos a distância, número de alunos e volume de dados gerados.

No próximo capítulo explicitaremos o método deste trabalho, bem como a tarefa a ser utilizada para a análise dos dados.

5. MÉTODO DA PESQUISA

Neste capítulo, explicitamos as etapas do método da pesquisa, relatando a caracterização, o entendimento do domínio, o entendimento dos dados, o processo de preparação dos dados, a fase de modelagem e aplicação de técnicas de EDM para descoberta de padrões comportamentais de autorregulação da aprendizagem, a avaliação dos modelos desenvolvidos e, por fim, a concepção, implementação e avaliação da solução proposta.

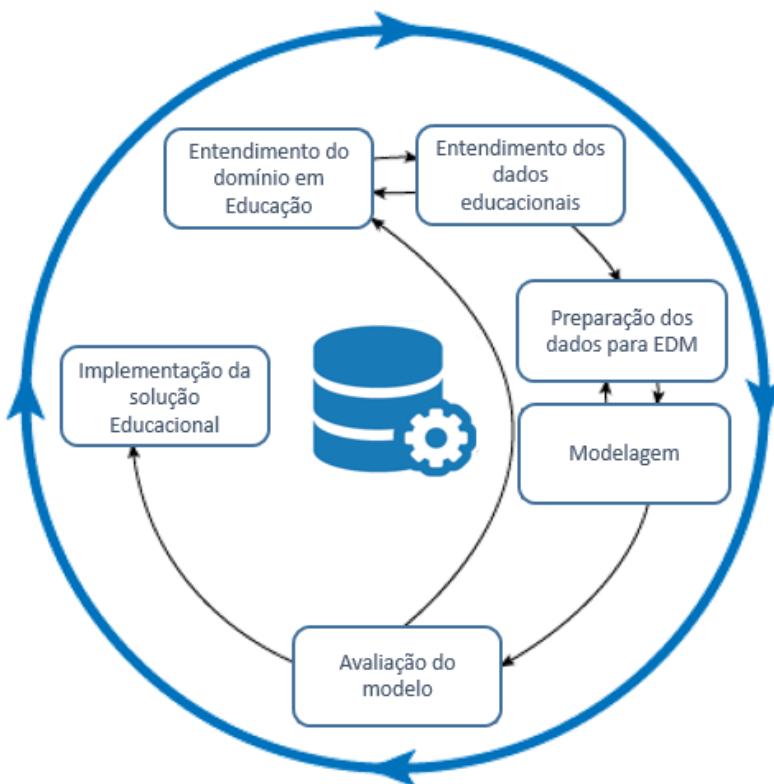
5.1 Fases do Processo

O método utilizado nesta tese foi fundamentado no processo CRISP-DM (*Cross Industry Standard Process For Data Mining*). Este processo foi desenvolvido por um consórcio formado por NCR Systems Engineering Copenhagen, Daimler Chrysler AG, SPSS Inc. e OHRA Verzekeringen en Bank Groep B.V em 1996 (CHAPMAN, CLINTON, *et al.*, 2000).

O processo CRISP-DM é descrito de forma hierárquica e composto por seis fases: entendimento do negócio/domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação (CRISP-DM, 1999). No artigo de Sheth e Patel (2010), os autores apresentam o CRISP-DM como um processo adequado para problemas de mineração de dados educacionais.

Para a melhor adequação ao contexto da mineração de dados educacionais, o processo CRISP-DM foi adaptado para o contexto de dados educacionais e intitulado de CRISP-EDM. A Figura 17 é mostra o ciclo de vida de um projeto de mineração de dados educacionais, a qual consiste de 6 (seis) fases de acordo com o processo CRISP-EDM. A sequência de fases não é obrigatória, podendo ocorrer a transição para diferentes fases, dependendo do resultado de cada fase. As setas indicam as mais importantes e mais frequentes dependências entre as fases.

Figura 17: Fases do modelo de referência CRISP-EDM



Fonte: Adaptador de CHAPMAN, CLINTON, *et al.*, (2000)

O ciclo externo na figura simboliza o ciclo natural da mineração de dados. Um processo de mineração de dados continua após a solução ter sido desenvolvida, como por exemplo com a validação e atualização dos modelos desenvolvidos. As próximas seções detalham cada uma das fases, as quais estão estruturadas em várias tarefas gerais e específicas, onde são definidas as ações que são desenvolvidas para situações específicas.

5.2 Entendimento do Domínio

Esta é uma das fases mais importantes do processo. É nesta fase que é feito o levantamento das problemáticas educacionais da instituição. Para esta tese, o entendimento do domínio foi realizado a partir de três etapas. O fluxo da Figura 18 demonstra cada uma das três etapas, divididas em: Revisão da literatura, Entrevista com especialistas e Aplicação de questionário de autorregulação da aprendizagem (OSLQ).

Figura 18: Fluxo do entendimento do domínio



Fonte: Elaborada pelo Autor (2016)

Na primeira etapa foi realizado o entendimento da literatura sobre as problemáticas da EAD, no que se refere ao fenômeno de autorregulação da aprendizagem e às diversas abordagens de mensuração (Capítulos 2 e 4).

A segunda etapa foi a realização de entrevistas com especialistas no domínio, por meio de encontros periódicos com os envolvidos do Núcleo de Educação a Distância da Universidade de Pernambuco (NEAD/UPE)², afim de entender o funcionamento e a utilização da plataforma de ensino Moodle por todos os envolvidos.

A terceira etapa foi a aplicação do questionário de autorregulação da aprendizagem OSQL (BARNARD, LAN, et al., 2009) (*Apêndice B*) e (*Anexo A*). Este questionário vem sendo aplicado em diversas pesquisas na comunidade científica e busca realizar a captura dos aspectos de autorregulação da aprendizagem em contextos de aprendizagem online.

Nesta pesquisa aplicamos o instrumento para um quantitativo de 466 alunos da instituição. De posse desses dados, foi realizado uma validação do instrumento por meio da técnica estatística de análise fatorial confirmatória (AFC) (*Apêndice E*) e posteriormente análises descritivas dos dados fornecendo um entendimento exploratório sobre o perfil comportamental autorregulatório dos alunos na instituição.

5.3 Entendimento dos dados

Inicialmente foi realizado um acordo de cooperação científica com o NEAD/UPE para a disponibilização das bases de dados históricas do sistema de LMS Moodle utilizado na instituição nos últimos sete anos (de 2009.2 à 2016.1), levando em consideração os aspectos de anonimato e segurança das informações.

De posse das bases de dados, foi realizado o entendimento sobre a arquitetura do sistema de banco de dados do LMS Moodle. Em seguida foram identificadas e

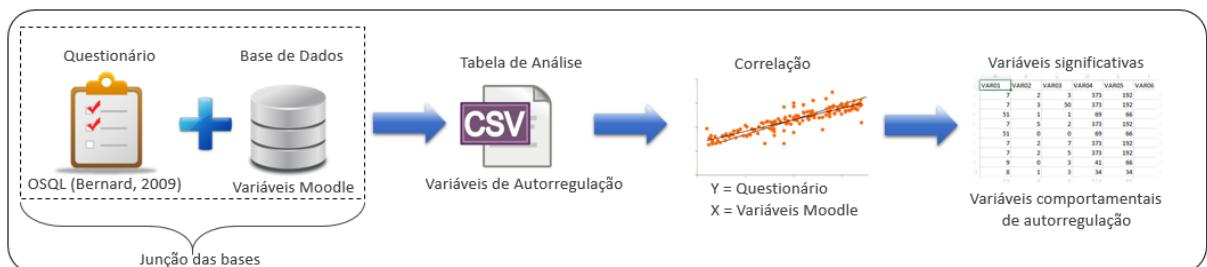
² www.nead.upc.br

extraídas, por meio dos scripts desenvolvidos no *Apêndice D*, as variáveis comportamentais armazenadas no banco de dados da plataforma Moodle, que representaram os comportamentos dos alunos de acordo com os construtos definidos pela teoria de autorregulação da aprendizagem (ZEIDNER, BOEKERTS e PINTRCH, 2000).

Esta identificação das variáveis levou em consideração os seis construtos propostos pelo modelo teórico: (1) Estruturação do ambiente; (2) Busca por Ajuda; (3) Estratégias de realização da tarefa; (4) Gerenciamento do tempo; (5) Estabelecimento de metas; e (6) Autoavaliação. O objetivo desta fase foi identificar os aspectos comportamentais de autorregulação que são registrados na base de dados da plataforma de LMS Moodle.

Em seguida foi realizada uma junção dos dados coletados por meio do questionário OSLQ (Seção 5.3) e os dados comportamentais dos alunos extraídos do banco de dados da plataforma Moodle do semestre 2014.1. A Figura 19 mostra o fluxo do processo da identificação das variáveis comportamentais significativas.

Figura 19: Fluxo do entendimento dos dados



Fonte: Elaborada pelo Autor (2016)

Após a junção das bases foi gerada uma tabela de análise contendo as variáveis de ambas as bases de dados. Em seguida, foi aplicada a técnica de correlação, com o objetivo de identificar as variáveis comportamentais que mais se correlacionavam com os construtos de autorregulação da aprendizagem presentes no questionário.

A função desta fase foi analisar se a afirmação dos alunos, respondida por meio do questionário, teriam relação com o seu comportamento dentro da plataforma de LMS, onde posteriormente foi escolhido um conjunto de variáveis comportamentais, na base de dados Moodle, que fossem suficientes para representar o fenômeno sem a necessidade da utilização de questionários.

Como produto final desta fase identificamos um conjunto de trinta e três variáveis comportamentais, armazenadas na plataforma, que representassem o fenômeno de autorregulação da aprendizagem, possibilitando a eliminação do questionário, utilizando apenas o conjunto de variáveis comportamentais armazenadas em banco de dados.

Para o prosseguimento da pesquisa, todas as análises seguintes foram realizadas apenas com a utilização das variáveis registradas na base de dados do sistema de LMS Moodle.

5.4 Preparação dos dados

Nesta fase foi realizada a junção das bases de dados, extração, transformação e limpeza das variáveis para a construção da tabela de análise. Para a realização da extração das variáveis, foi necessário um estudo da estrutura arquitetural do banco de dados. O estudo envolveu inicialmente a modelagem das entidades no banco de dados, no qual foi necessário entender o contexto em que a plataforma estava sendo utilizada e suas especificidades. Após a fase de entendimento da estrutura dos bancos de dados, foi realizada a construção da tabela de análise. A Figura 20 mostra o fluxo de integração das bases de dados e a construção da tabela de análise.

Figura 20: Fluxo para a construção da tabela de análise



Fonte: Elaborada pelo Autor (2016)

Esta tabela foi construída a partir da junção das bases de dados referente a um período de sete anos de cursos de EAD na instituição.

5.4.1 Seleção dos dados

Essa etapa foi relacionada a quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas para o estudo. Em geral dados provindos de sistemas computacionais encontram-se organizados em bases de dados transacionais. Assim foi necessário fazer uma junção para compor uma só base apropriada para a fase de mineração.

Desta forma todos os registros da base de dados foram incluídos em uma nova tabela. Para tal, foi realizado a construção de um conjunto de scripts SQL (*Apêndice D*) com o objetivo de extrair as variáveis comportamentais mapeadas na fase de entendimento dos dados. Foi realizada a extração de trinta e três variáveis agrupadas por construto de autorregulação da aprendizagem, adicionadas às variáveis relacionadas ao desempenho dos alunos.

Algumas dificuldades foram encontradas durante a fase de seleção dos dados, tais como: mudança da estrutura relacional nas diversas versões do Moodle, sistemas de SGBD diversos, divergência na identificação única de alunos, alto volume de dados e divergência de sintax nos *scripts* de extração dos dados, devido às diversas versões dos sistemas de SGBD.

Após a extração das variáveis foram aplicados processos de redução de variáveis, categorização e limpeza dos dados. Inicialmente foram mapeadas e coletadas 40 variáveis comportamentais de SRL, dos quais 7 variáveis foram retiradas por motivos de autocorrelação e alta presença de frequência zero.

Posteriormente foi feito uma análise das disciplinas que tinham comportamento diferenciado das demais e verificamos que as disciplinas de Estágio Supervisionado deveriam ser retiradas da base de análise. Estas disciplinas tinham comportamento diferenciado e não seguiam o padrão de avaliação estabelecido para as demais disciplinas.

5.4.2 Categorização de Variáveis

Diversos algoritmos de mineração de dados, especialmente os algoritmos de classificação, requerem que os dados estejam na forma de atributos categorizados, especificamente o atributo alvo (variável de interesse) (TAN, STEINBACH e KUMAR, 2009). Assim, muitas vezes é necessário transformar um atributo contínuo em um atributo categorizado ou binário.

Para a realização desta fase foi feita a categorização da variável *Desempenho* e validada de três formas. A primeira categorização foi baseada na construção de três classes, dos quais a primeira classe foi categorizada no intervalo de nota de 0,0 a 4,0; a segunda classe acima de 4,0 e abaixo de 7,0; e a terceira classe com notas maior ou igual a 7,0.

A segunda categorização teve como meta transformar a variável contínua em uma variável binária. A primeira classe recebeu valores de notas abaixo de 4,0 e a

segunda classe notas iguais ou maiores que 4,0. Nesta binarização o objetivo foi dividir os grupos de alunos em situação de risco de reprovação direta e alunos com chance de aprovação.

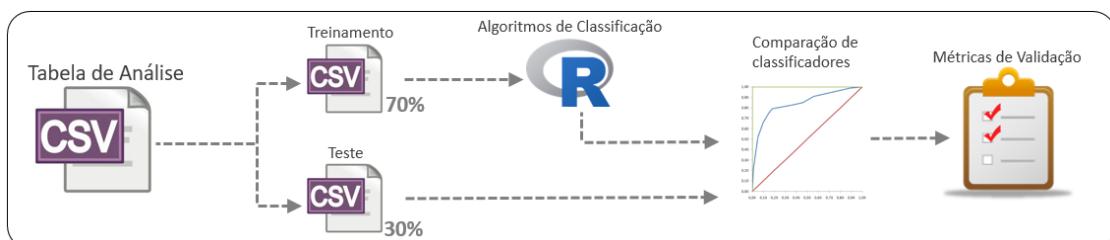
A terceira categorização foi também do tipo transformação binária, com valores de notas abaixo de 7,0 para a primeira classe e notas acima de 7,0 para a segunda classe. Nesta binarização o objetivo foi dividir os grupos em alunos aprovados direto e alunos reprovados com chance de recuperação.

Estas três formas de categorização foram testadas e validadas na fase de modelagem e por fim escolhemos a que demonstrou melhores resultados com os algoritmos de classificação.

5.5 Modelagem

Nesta fase definimos as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de classificação, bem como seus ajustes de parâmetros. Esta fase teve como objetivo modelar os padrões comportamentais de autorregulação, por meio de técnicas de mineração de dados educacionais, utilizando variáveis comportamentais como preditoras do desempenho acadêmico dos alunos. A Figura 21 mostra o fluxo do processo de modelagem.

Figura 21: Fluxo do processo de modelagem



Fonte: Elaborada pelo Autor (2016)

Para esta fase utilizamos uma base de dados com um total de 30.217 alunos, no qual foi realizada a separação de 21.154 (70%) observações (alunos) para treinamento dos modelos de classificação e 9.063 (30%) de observações (alunos) para a fase de teste e validação, de acordo com Hair, Black, et al. (2009). Em seguida a base de treinamento foi utilizada para realizar a modelagem e construção dos classificadores.

Esta construção dos classificadores foi realizada por meio da técnica de *Cross-Validation* ou **validação cruzada**. Essa técnica realiza o particionamento do conjunto

de dados em subconjuntos mutualmente exclusivos, onde posteriormente, utiliza-se alguns dos subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento) e o restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo. Para esta tese, aplicamos *10-fold cross-validation* para cada combinação de parâmetros obtida a partir de um conjunto de parâmetros candidatos. Em seguida selecionamos a combinação ótima de parâmetros observando o melhor desempenho médio estimado a partir dos subconjuntos (folds) de validação. (CHAPMAN, CLINTON, *et al.*, 2000).

Posteriormente realizamos um comparativo entre os diversos modelos de classificação por meio da base de teste. Por fim, escolhemos o melhor modelo por meio dos resultados das métricas de validação.

Essa fase ocorreu de forma cíclica com a fase anterior (5.5). Em alguns momentos foi necessário retornar à fase de preparação dos dados para fazer tratamentos e manipulações nas variáveis originais, afim de obter os melhores resultados em termos de modelagem.

5.5.1 Algoritmos de Classificação

Foram utilizados quatro algoritmos de classificação: Regressão Logística, Árvore de Decisão, SVM e Naive Bayes (descritos na Seção 3.4.1). Toda a fase de modelagem foi realizada por meio da construção de algoritmos utilizando a linguagem de programação R³ e diversos pacotes para a implementação dos algoritmos de classificação. O Quadro 7 mostra as referências dos pacotes que foram utilizados:

Quadro 7: Pacotes utilizados no processo de modelagem.

Nome do Pacote	Descrição	Documentação
kernlab	Pacote utilizado para a implementação do algoritmo de classificação SVM.	https://cran.r-project.org/web/packages/kernlab/kernlab.pdf
party	Pacote utilizado para a implementação do algoritmo de árvore de decisão.	https://cran.r-project.org/web/packages/party/party.pdf
e1071	Pacote utilizado para a implementação do algoritmo de classificação Naive Bayes.	https://cran.r-project.org/web/packages/e1071/e1071.pdf

³ <https://cran.r-project.org>

dplyr	Pacote utilizado para manipulação e limpeza da base de dados.	https://cran.r-project.org/web/packages/dplyr/dplyr.pdf
caret	Pacote utilizado para a geração da matriz de confusão e métricas de validação.	https://cran.r-project.org/web/packages/caret/caret.pdf
ROCR	Pacote utilizado para a geração do gráfico de curva ROC e valor de área.	https://cran.r-project.org/web/packages/ROCR/ROCR.pdf
ggplot2	Pacote utilizado para a geração de gráficos.	https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
DMwR	Pacote utilizado para balanceamento de bases de treinamento.	https://cran.r-project.org/web/packages/DMwR/DMwR.pdf

Fonte: Elaborado pelo Autor (2016)

A aplicação dos algoritmos de classificação tiveram como meta produzir, a partir de um conjunto de observações, um modelo que permitisse a predição de valores tomados por uma variável categórica (desempenho), a partir de uma série de variáveis explicativas numéricas e/ou categóricas (variáveis de autorregulação da aprendizagem).

Dentre alguns critérios de escolha, sobre a utilização dos algoritmos de classificação, foi necessário selecionar os algoritmos que desse suporte a variáveis numéricas e categóricas como preditoras, onde fosse possível modelar a probabilidade de um evento ocorrer em função de variáveis preditoras, e descriminar o poder de influência de cada variável no modelo. O Apêndice G contém todos os scripts desenvolvidos para a construção dos modelos de classificação desta tese.

5.6 Avaliação dos Modelos

A avaliação é o processo de verificar como os modelos desenvolvidos com as técnicas de mineração são executados nos dados reais. É importante validar os modelos de mineração entendendo suas qualidades e características antes de implantá-los em um ambiente de produção (CHAPMAN, CLINTON, et al., 2000). Nesta tese a avaliação foi realizada por meio de técnicas que buscaram identificar o poder de previsão dos modelos por meio de métricas de precisão preditiva de classificadores.

5.6.1 Avaliação da Precisão de Classificação

Foi realizado a avaliação dos modelos desenvolvidos em uma base de teste de 9.063 alunos pertencentes a quatro cursos distintos.

Para entender os erros gerados pelo classificador foi feita a construção da matriz de erros denominada matriz de confusão. A partir desta matriz foi possível obter métricas de qualidade para a avaliação do desempenho dos algoritmos de classificação.

As seguinte terminologias foram usadas para a validação do poder de precisão dos classificadores a partir da matriz de confusão:

- **Positivo verdadeiro (TP):** é relacionado ao número de instâncias positivas previstas corretamente pelo classificador;
- **Negativo falso (FN):** é o número de instâncias previstas erroneamente como negativos pelo classificador;
- **Positivo falso (FP):** é o número de exemplos negativos previstos erroneamente como positivos pelo classificador;
- **Falso verdadeiro (TN):** é o número de exemplos negativos previstos corretamente pelo classificador.

A partir destas terminologias foram utilizadas as seguintes métricas para validação: Acurácia, Sensibilidade, Especificidade, Precisão e Curva ROC.

A **acurácia** é definida como sendo o número de instâncias corretas divididas pelo número total de instâncias;

A **sensibilidade** buscou verificar número de instâncias positivas (alunos em situação de risco) previstos corretamente pelo classificador.

A **especificidade** é definida como a fração de instâncias negativas corretamente (alunos que não se encontram em situação de risco) previstas pelo classificador.

A **precisão** determinou o percentual de registros que são positivos no grupo que o classificador previu como classe positiva, ou seja, de todos os alunos que o modelo classificou como sendo alunos de situação de risco, aqueles que realmente são.

Por fim a última validação foi por meio da construção do gráfico que esboça a área sob a curva ROC. Esta foi uma das principais métricas utilizadas nesta fase de avaliação. Esta métrica é uma representação gráfica que ilustra a performance de um

sistema classificador binário e como o seu limiar de discriminação é variado, permitindo estudar a variação da sensibilidade e especificidade, para diferentes valores de corte (FAWCETT, 2006).

Para a plotagem do gráfico da curva ROC e o cálculo de área, foram utilizados os pacotes *ROCR*⁴, *plotROC*⁵, *ggplot2*⁶da plataforma R.

As métricas analisadas serviram para realizar a escolha do modelo de classificador que melhor se ajustou aos dados sobre o fenômeno de autorregulação da aprendizagem. A partir da escolha do modelo com os melhores resultados, iniciamos a fase de concepção de uma interface de software dashboard para professores/tutores de Educação a Distância que utilizassem do modelo desenvolvido para predizer o desempenho de alunos. A próxima seção detalha o processo de concepção e implementação da solução de software.

5.7 Implementação da solução

Como forma de materializar os resultados encontrados na fase de modelagem, esta fase foi direcionada à construção de uma interface de software *dashboard* que proporcionasse ao professor/tutor uma visualização dos indicativos comportamentais de alunos que se encontram em estado crítico de autorregulação da aprendizagem e consequentemente em risco de mau desempenho acadêmico.

O software desenvolvido buscou aplicar o modelo de classificação em um dashboard, possibilitando que professores e tutores possam utilizar de modelos de mineração de dados por meio de visualizações e sem exigir conhecimentos técnicos. Esta fase de concepção da solução foi composta de três etapas, inicialmente foram feitos protótipos de baixa fidelidade, seguido pelo desenvolvimento da arquitetura e protótipos de alta fidelidade e por fim foi realizado uma validação da solução com professores e tutores.

5.7.1 Prototipagem

O processo de prototipagem fez parte do ciclo de desenvolvimento da solução de software desta tese, sendo utilizado para antecipar ao usuário final características da interface que puderam ser testadas, validadas e modificadas.

⁴ <https://cran.r-project.org/web/packages/ROCR/ROCR.pdf>

⁵ <https://cran.r-project.org/web/packages/plotROC/plotROC.pdf>

⁶ <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>

Segundo Preece, Rogers e Sharp (2005) os protótipos de baixa fidelidade são úteis porque tendem a ser simples, baratos e de rápida produção. Possuem características importantes, os quais podem ser rapidamente modificadas, oferecendo, portanto, suporte à exploração de alternativas de *design*.

Inicialmente foram desenvolvidos protótipos de baixa fidelidade como forma de validação com especialistas (professores e tutores). Os protótipos iniciais tiveram como objetivo modelar uma interface de *dashboard* que possibilitasse ao professor visualizar indícios de previsibilidade para tomada de decisão em termos didáticos e metodológicos.

Foram desenvolvidas um total de oito telas de baixa fidelidade e seus respectivos casos de uso (*Apêndice L*), os quais passaram por um processo de refinamento até chegar a uma versão final e possível de ser implementada.

5.7.2 Arquitetura e Implementação

De posse dos protótipos de baixa fidelidade, foi realizado o desenvolvimento da arquitetura, implementação do protótipo de alta fidelidade e sua integração com a plataforma de LMS.

Tendo em vista que a fase de mineração de dados foi realizada por meio da plataforma e linguagem de programação R, foi necessário buscar uma solução que pudesse atender o desenvolvimento da aplicação preservando o poder analítico da plataforma R e a implementação dos modelos de previsão desenvolvidos.

Desta forma, a aplicação foi implementada utilizando o framework Shiny⁷ e duas bibliotecas gráficas do R: rcharts⁸ e plotly⁹. O framework Shiny é um pacote do software estatístico R, criado para facilitar o desenvolvimento de aplicações web com HTML, CSS e JavaScript. Por meio deste framework foi possível construir um *dashboard* interativo que pudesse consumir o poder analítico do R de forma integrada. Isso pôde ser alcançado por meio de componentes que facilitam a implementação da interface com redução de esforços e facilidade de integração com modelos estatísticos e de aprendizagem de máquina (BAVARESCO e ROSA, 2015). A Figura 22 apresenta um fluxo simples da arquitetura que será melhor detalhada na Seção 6.6.

⁷ <http://shiny.rstudio.com/>

⁸ <http://ramnathv.github.io/rCharts/>

⁹ <https://plot.ly/r/>

Figura 22: Fluxo da arquitetura



Fonte: Elaborada pelo Autor (2016)

O primeiro módulo da aplicação é responsável por realizar a coleta dos dados em instâncias de sistemas Moodle. Esta coleta é realizada baseada nas variáveis comportamentais de autorregulação da aprendizagem.

O segundo módulo é responsável pela inteligência da aplicação. Ele recebe as variáveis comportamentais e realiza o processo de treinamento do modelo de classificação, fazendo a estimativa dos parâmetros do modelo de decisão por meio do servidor Shiny, que executa os diversos pacotes de estatística e aprendizagem de máquina.

O terceiro módulo é responsável por apresentar ao usuário as informações sobre o estado autorregulatório da aprendizagem da turma, bem como de cada aluno individualmente. Este módulo apresenta para o professor/tutor informações de previsibilidade do desempenho dos alunos baseado nas suas variáveis comportamentais.

Este módulo cliente é composto pelas seguintes informações: Indicativo de proporção de alunos com baixo desempenho, lista de alunos em estado crítico de autorregulação, descrição gráfica de cada uma das variáveis de autorregulação da aprendizagem (construtos), visualização de cada uma das variáveis por aluno e percentual preditivo de alunos em situação de baixo desempenho.

5.7.3 Avaliação da solução de software desenvolvida

Como forma de avaliar a solução com professores e tutores, foi selecionada uma amostra de 20 professores e tutores das instituições de ensino UFRPE, UPE, UNIVASF e UFPE. Os critérios de inclusão dos professores na amostra foram: experiência de no mínimo um ano com EAD e experiência com a plataforma de LMS Moodle.

A fase de avaliação da solução foi realizada seguindo o fluxo apresentado na Figura 23. Inicialmente foram apresentados aos professores/tutores os protótipos de

baixa fidelidade. Na segunda etapa foi aplicado um questionário para a mensuração da expectativa do usuário em relação ao protótipo apresentado.

Figura 23: Fluxo do processo de avaliação do software



Fonte: Elaborada pelo Autor (2016)

Em seguida, na terceira etapa, foi apresentada a versão implementada (protótipo de alta fidelidade), os usuários tiveram dois minutos para a familiarização com o software. A quarta etapa buscou realizar o teste de usabilidade, no qual os professores/tutores foram submetidos a realizar um total de seis tarefas (Quadro 9) na ferramenta.

Por fim, a quinta etapa, buscou avaliar a experiência do usuário (*User eXperience-UX*). Aplicamos um questionário de experiência com o objetivo de avaliar a diferença significativa entre as expectativas e experiências de uso dos professores/tutores em relação à solução proposta.

Estas etapas, conjuntamente, objetivaram mensurar tanto a compreensão de como o usuário se sente e sua satisfação em relação à aplicação (UX), quanto ao sucesso e tempo necessário para a realização de determinadas atividades na aplicação (usabilidade). A seguir descrevemos com maior riqueza de detalhes o processo e os instrumentos utilizados.

Avaliação da Experiência do Usuário

Diversas abordagens vêm sendo desenvolvidas para avaliar a experiência do usuário durante o processo de desenvolvimento de software (VERMEEREN, LAW, et al., 2010). A abordagem que adotamos nesta tese foi baseada no instrumento de coleta *AttrakDiff*¹⁰. Este instrumento é composto de 28 pares de adjetivos opostos (diferencial semântico), onde cada par de palavras representa um item no questionário e é baseado em uma escala de sete níveis variando de -3 até 3.

¹⁰ <http://www.attrakdiff.de/>

Os itens do questionário são subdivididos em categorias ou dimensões, onde os sete primeiros itens avaliam a dimensão Qualidade Pragmática (PQ), catorze itens avaliam a Qualidade Hedônica (HQ) destes, sete avaliam o estímulo (HQ-S) e sete avaliam a Identidade (HQ-I), e sete itens avaliam a Atratividade (ATT) (FIEBIG, HALBRÜGGE e KRAUS, 2016). O questionário é apresentado no Quadro 8.

Quadro 8: Questionário sobre User Experience

		Qualidade Pragmática (PQ)							
		-3	-2	-1	0	1	2	3	
Qualidade Pragmática (PQ)	Técnico								Humanizado
	Complicado								Simples
	Impraticável								Prático
	Rebuscado								Direto
	Imprevisível								Previsível
	Confuso								Claramente estruturado
	Desorganizado								Gerenciável
		Qualidade hedônica (HQ)							
		-3	-2	-1	0	1	2	3	
HQ-I. (Identidade)	Isolador								Conectivo
	Amador								Profissional
	Deselegante								Elegante
	Inferior								Alto Nível
	Segregador								Integrador
	Afasta-me das pessoas								Aproxima-me das pessoas
	Não apresentável								Apresentável
HQ-S. (Estímulo)	Convencional								Inventivo
	Sem imaginação								Criativo
	Cauteloso								Arrojado
	Conservador								Inovador
	Entediante								Cativante
	Pouco exigente								Desafiador
	Ordinário								Original
		Atratividade (ATT)							
		-3	-2	-1	0	1	2	3	
ATT. (Atratividade)	Desagradável								Agradável
	Feio								Atraente
	Enfadonho								Simpático
	Rejeitável								Convidativo
	Mau								Bom
	Repulsivo								Atrativo
	Desmotivador								Motivador

Fonte: <http://www.attrakdiff.de/>

A PQ refere-se à usabilidade da interface, ou seja, avalia se os usuários conseguem realizar tarefas e atingir os objetivos utilizando o sistema; a HQ-S avalia o quanto o produto possui características inovadoras, e apresenta interações estimulantes; a HQ-I avalia o quanto a interface permite que o usuário se identifique. Por fim a ATT descreve o valor global do produto baseado na percepção do usuário em relação à qualidade do produto (HASSENZAHL, 2004).

A aplicação do instrumento (questionário) foi dividida em dois momentos. No primeiro momento mostramos os protótipos de baixa fidelidade e aplicamos o primeiro questionário com o objetivo de capturar a expectativa do usuário quanto a solução de baixa fidelidade (protótipo). Posteriormente, aplicamos seis tarefas que foram realizadas na solução implementada e aplicamos novamente o questionário.

Por fim, realizamos uma análise gráfica comparativa que buscou identificar, para os 28 itens dos questionários, as diferenças entre as expectativas dos professores e suas experiências durante a realização das tarefas. O teste de usabilidade, aplicado entre as duas coletas, é descrito com mais detalhes a seguir.

Avaliação de Usabilidade

Durante a realização das tarefas coletamos informações que puderam ser analisadas sobre o ponto de vista da usabilidade da ferramenta. A fase de usabilidade é tida como a participação direta dos usuários, os quais executam tarefas pré-definidas e comentam sobre a interface da aplicação, enquanto são registradas as dificuldades encontradas (PRATES e BARBOSA, 2003).

Para a realização do teste de usabilidade estabelecemos seis tarefas e coletamos as seguintes variáveis: contagem de erros na realização das tarefas, tempo de execução, quantidade de pedidos de ajuda, quantidade de cliques e questões abertas onde os usuários puderam relatar os problemas de usabilidade, sugestões e críticas. No Apêndice M mostramos o protocolo de aplicação. O Quadro 9 mostra a descrição de cada uma das tarefas.

Quadro 9: Descrição das tarefas para o teste de usabilidade

Nº	Descrição
1	[MÓDULO VISÃO GERAL DOS DADOS – ABA GERAL] a) No curso de Pedagogia, 3º período, visualizar o gráfico dos dados da disciplina Metodologia Científica. b) Descobrir no gráfico qual a média dessa turma em relação ao indicador "Quantidade de mensagens enviadas por aluno aos professores".
2	[MÓDULO VISÃO GERAL DOS DADOS - ABA INDICADORES] a) No curso de Biologia, 5º período, na disciplina Bioquímica II, visualizar o gráfico do indicador " Média semanal de acessos do aluno ao ambiente". b) Identificar qual (ou quais) aluno(s) de maior valor nesse indicador no gráfico e dar um zoom na região do gráfico que mostra esse(s) aluno(s).
3	[MÓDULO ANÁLISE DE DESEMPENHO - ABA GERAL] a) No curso de Biologia, 4º período, disciplina Genética Molecular, no construto ESTABELECIMENTO DE METAS, identifique o indicador “Quantidade geral de postagens de um(a) aluno(a) em fóruns” e verifique a média para o grupo insatisfatório; b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.
4	[Módulo Análise de Desempenho - aba Geral] a) Para o curso de Pedagogia, 4º Período, disciplina Planejamento Educacional, verifique o indicador “Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente” e informe a quantidade de mensagens enviadas do grupo Satisfatório e do grupo de Não Satisfatório; b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.
5	[MÓDULO ANÁLISE DE DESEMPENHO - ABA INDICADORES] a) Para o curso de Administração, 8º Período, disciplina Licitações, Contratos e Convênios, identifique o aluno que está em pior situação de desempenho em relação ao indicador “Quantidade de leituras feitas ao fórum (pageviews) ”. b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.
6	[MÓDULO ANÁLISE DE DESEMPENHO - ABA ALUNOS] a) Selecione o curso de Letras, 6º Período, disciplina Literatura Brasileira III, no construto GESTÃO DO TEMPO. Em seguida, procure a aluna Eveline <sobrenome omitido> e identifique os dois indicadores que a aluna se encontra em pior situação entre a frequência da aluna e a média da turma; b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.

Fonte: Elaborada pelo Autor (2016)

Todas as interações dos professores/tutores com a aplicação, por meio das tarefas realizadas, foram gravadas com o software Camtasia¹¹. Dentre estas interações, foram registradas a contagem de erros na realização das tarefas, tempo de execução, quantidade de pedidos de ajuda e sugestões, críticas e observações dos professores durante e ao final do teste.

¹¹ <https://www.techsmith.com/camtasia.html>

Aplicação de testes estatísticos

Após a aplicação dos dois questionários (*expectativa e experiência*) e do teste de usabilidade, realizamos a aplicação de testes estatísticos que buscaram analisar a diferença significativa entre *Expectativas* e *Experiência do Usuário*, por meio dos dados coletados com o instrumento AttrakDiff. Os testes foram realizados em busca de comprovar a existência, ou não, de diferenças significativas entre as respostas nos dois momentos.

Nesta análise optamos por utilizar o Teste de *Wilcoxon*, com correção de continuidade, a um nível de significância de 5%. Este teste substituiu o teste *t de Student* para amostras emparelhadas, quando os dados são medidos em escala ordinal baseando-se nos postos (*ranks*) das diferenças entre as médias (CAMPOS, 1987).

Este teste de comparação de médias foi realizado baseado na hipótese nula H_0 de não ter ocorrido diferença significativa entre a expectativa e a experiência do usuário, contra a hipótese alternativa H_1 de ter ocorrido diferença significativa. A formação da hipótese é mostrada abaixo:

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

O teste de hipótese foi aplicado para 28 variáveis: 7 relacionadas à Qualidade Pragmática (*PQ*), 7 à Qualidade Hendônica de Estímulo (*HQ-S*), 7 relacionadas à Qualidade Hendônica de Identidade (*HQ-I*) e 7 relacionadas à Atratividade (*ATT*).

5.8 Considerações sobre o capítulo

Nesse capítulo foram apresentadas as fases e etapas, com suas respectivas metas, técnicas e ferramentas que compõem o método de pesquisa desta tese. Na fase de entendimento do domínio foram realizadas a revisão da literatura e as entrevistas de campo com especialistas no domínio. Na fase do entendimento dos dados foi realizada a escolha das variáveis que representaram os aspectos comportamentais de autorregulação dos alunos.

Na fase de preparação dos dados foram realizadas a extração das variáveis e a construção da base de dados para análise. Em seguida foi realizada a modelagem,

onde foi aplicado um conjunto de quatro algoritmos de classificação. Estes algoritmos passaram por uma fase validação da modelagem, onde foi verificado um conjunto de métricas de validação.

Por fim foram realizadas a concepção, implementação e validação da solução de software desenvolvida. Esta fase buscou materializar os resultados encontrados na fase de modelagem por meio do desenvolvimento de uma ferramenta que possa ser usada por professores e tutores em cursos de EAD. Em seguida, no Capítulo 6, são apresentados os resultados alcançados no desenvolvimento deste método.

6. RESULTADOS

Neste capítulo, explicitamos os resultados do método apresentado no capítulo anterior, bem como as especificações que foram necessárias para atingir os resultados e suas respectivas discussões.

6.1 Resultados sobre o entendimento do domínio

A primeira fase da realização desta tese, de acordo com a metodologia CRISP-EDM (Seção 5.3), foi o entendimento do domínio. Esta fase foi realizada com o objetivo de compreender os problemas em relação à capacidade de autorregulação da aprendizagem dos alunos na modalidade de educação a distância. Nesta fase foi realizado o entendimento sobre o funcionamento do Núcleo em Educação a Distância NEAD/UPE. Este entendimento foi realizado por meio de entrevistas com Gestores e administradores da plataforma de EAD utilizada no NEAD e a aplicação de um questionário de autorregulação da aprendizagem aos estudantes da instituição.

As entrevistas foram necessárias para o entendimento dos diferentes perfis que compõem a EAD da UPE, bem como o entendimento da estrutura pedagógica e gestão de recursos. Em seguida, para ter uma visão exploratória sobre o comportamento de autorregulação dos alunos, foi aplicado o instrumento de coleta *Self-Regulated Learning Questionnaire (OSQL)* desenvolvido por Barnard, Lan, et al. 2008 e Barnard, Lan, et al. 2009. Este instrumento é uma escala de 24 itens com um formato de resposta do tipo escala de *Likert* de 5 pontos com valores variando de ‘concordo totalmente’ (5) até ‘discordo totalmente’ (1). O Quadro 10 apresenta todos os itens do questionário.

Quadro 10: Variáveis do questionário de autorregulação da aprendizagem.

N	Item do questionário	Construtos
1	Eu defino metas para a realização das minhas tarefas em cursos online.	Estabelecimento de metas (EM)
2	Eu defino metas de curto prazo (diário ou semanal), bem como metas de longo prazo.	
3	Eu mantendo um alto padrão de aprendizagem no meu curso.	
4	Eu defino metas que me ajudam com o tempo de estudo dedicado para os meus cursos à distância.	
5	Eu não comprometo a qualidade do meu trabalho porque é em EAD.	
6	Eu escolho o local onde eu estudo para evitar distrações.	
7	Eu procuro um lugar confortável para estudar.	
8	Eu sei onde posso estudar de forma mais eficiente quando me dedico a EAD.	Estruturação do Ambiente (EA)

9	Eu escolho um horário do dia para estudar que tenha poucas distrações.	
10	Eu tento fazer esquemas e anotações relacionadas com os conteúdos.	Estratégia para realização de tarefas (ET)
11	Eu costumo ler os materiais em voz alta para não sofrer distrações.	
12	Eu preparam minhas perguntas antes de entrar nos chats e fóruns de discussões.	
13	Costumo utilizar material extra do que foi disponibilizado na plataforma.	
14	Eu reservo tempo extra para estudar para minhas disciplinas da EAD.	Gerenciamento do Tempo (GT)
15	Agendo dias específicos da semana para estudar para meu curso	
16	Tento distribuir meu tempo estudando uniformemente todos os dias.	Busca por ajuda (BA)
17	Procuro amigos para tirar dúvidas sobre os conteúdos quando preciso.	
18	Compartilho meus problemas com os meus colegas descobrir como resolver os nossos problemas.	
19	Quando necessário, eu tento encontrar meus colegas de curso presencialmente.	
20	Eu costumo pedir ajuda ao tutor ou professor através de e-mail ou mensagem via plataforma.	Auto avaliação (AA)
21	Eu costumo fazer um resumo do meu aprendizado para refletir sobre o que aprendi na disciplina.	
22	Faço reflexão e questionamentos sobre o material do curso disponibilizado.	
23	Me comunico com meus colegas para refletir como está meu andamento no curso.	
24	Costumo fazer comparativos entre o que estou aprendendo e o que meus colegas estão aprendendo.	

Fonte: Adaptado de Barnard, Lan, *et al.* 2009

Como as competências de autorregulação da aprendizagem parecem ser “altamente dependentes de contexto” (SCHUNK, 2005), o desenvolvimento de um instrumento adaptado para os ambientes de aprendizagem online é uma etapa necessária em pesquisas que avaliam o fenômeno na modalidade de ensino a distância.

Inicialmente realizamos a validação deste instrumento em contexto da Educação a Distância no Brasil e publicado em Rodrigues, Ramos, *et al.*, 2016. Este trabalho analisou a adequação do instrumento para capturar as características e habilidades de autorregulação dos alunos na instituição a qual esta pesquisa foi desenvolvida. A próxima seção descreve as características da amostra coletada e os respectivos resultados.

6.1.1 PÚBLICO ALVO E AMOSTRA

Para a aplicação do instrumento de coleta OSLQ, utilizamos uma amostra coletada da Universidade de Pernambuco (UPE) que promove, a mais de dez anos, cursos de graduação na modalidade de EAD por meio do programa Universidade Aberta do Brasil (UAB)¹².

Os dados foram coletados por meio de um processo de amostragem aleatória simples de acordo com Bolfarine e Bussab, 2005, totalizando uma amostra de 466

¹² <http://uab.capes.gov.br/>

alunos, com idade média de 30 ($SD=18,23$) anos (*Apêndice B*). Os alunos respondentes foram estudantes dos cursos de Bacharelado em Administração Pública e das Licenciaturas em Biologia, Letras e Pedagogia. A Tabela 1 mostra a frequência do quantitativo de alunos por curso.

Tabela 1: Variáveis de interação coletadas.

Cursos	Frequência (fi)	Percentual (%)
Administração Pública	41	8,80 %
Licenciatura em Biologia	121	25,97 %
Licenciatura em Letras	107	22,96 %
Licenciatura em Pedagogia	197	42,27 %
Total	466	100

Fonte: Elaborada pelo Autor (2016)

Dos alunos que fizeram parte da amostra 57,7% possuíam apenas ensino médio, 24,22% e já tinham realizado algum curso superior, seja completo ou parcialmente e 7,35% possuíam pós-graduação, nos níveis de especialização ou mestrado. A Figura 24 mostra o gráfico com os resultados da predominância de cada um dos construtos para a amostra selecionada.

Figura 24: Resultado da aplicação do questionário OSLQ



Fonte: Elaborada pelo Autor (2016)

De acordo com o gráfico, podemos observar que o construto que obteve maiores valores médios foi o construto de *Estruturação do Ambiente*, do qual a variável V7 (Eu procuro um lugar confortável para estudar) obteve o maior resultado médio. Isto representa uma predominância na habilidade dos alunos na organização do seu ambiente de estudo.

Os construtos que apresentaram menores médias, em relação à afirmação dos alunos, foi o construto relacionado à *Gestão do Tempo* e o construto de *Auto Avaliação*. Isto nos mostra que, dentre todas as habilidades de autorregulação da

aprendizagem, a habilidade do aluno em organizar e gerir seu tempo de estudo é uma das mais deficientes, seguida pela baixa habilidade de auto avaliar seu processo de aprendizagem.

Os resultados desta fase serviram para ter uma visão inicial e exploratória sobre o comportamento autorregulatório dos alunos da instituição, na qual a pesquisa foi realizada. A próxima seção descreve os resultados da identificação das variáveis comportamentais armazenadas em plataformas de LMS; o entendimento das bases de dados nas quais são armazenadas e suas respectivas extrações.

6.2 Resultados sobre o entendimento dos dados

Esta fase buscou identificar as principais variáveis comportamentais, armazenadas na plataforma Moodle que representam o fenômeno de autorregulação da aprendizagem, sem a necessidade da utilização de questionário.

6.2.1 Identificação das variáveis significativas

Para a identificação destas variáveis foi realizada uma junção dos dados coletados por meio do questionário OSLQ (Seção 5.3) e os dados comportamentais dos alunos extraídos do banco de dados da plataforma Moodle do semestre 2014.1. Ou seja, para todos os alunos que responderam o questionário, buscou-se os seus respectivos comportamentos armazenados no banco de dados do sistema de LMS Moodle.

Realizamos o mapeamento de um conjunto, inicialmente de 42 variáveis que representavam as ações dos alunos relacionadas a comportamentos relacionados aos construtos do fenômeno de autorregulação da aprendizagem. Em seguida desenvolveu um conjunto de scripts SQL (*Apêndice B*) com o objetivo de extrair as variáveis selecionadas na base de dados 2014.1.

Posteriormente criamos uma tabela de análise contendo os dados do questionário e os dados comportamentais para os 466 alunos que responderam ao questionário.

Em seguida aplicamos a técnica de análise de correlação com o objetivo de identificar as variáveis comportamentais que apresentavam os maiores valores de correlação com os construtos de autorregulação. A função desta fase foi analisar se as afirmações dos alunos, respondidas por meio do questionário, teriam relação com

o seu comportamento dentro da plataforma de LMS e a validação da representatividade das variáveis para o fenômeno em estudo.

Após a aplicação da análise de correlação, a quantidade de variáveis com valores de correlação acima de 0,4 foi reduzida de 42 para 33 variáveis. As variáveis comportamentais foram classificadas nas seguintes seis categorias, de acordo com os construtos estabelecidos pelo modelo teórico proposto por Zeidner, Boekaerts e Pintrch (2000): (0) Dados Gerais; (1) Estruturação do ambiente; (2) Busca por Ajuda; (3) Estratégias de realização da tarefa; (4) Gerenciamento do tempo; (5) Estabelecimento de metas; (6) auto avaliação.

O Quadro 11 apresenta o conjunto de variáveis comportamentais que apresentaram correlação significativa.

Quadro 11: Variáveis de autorregulação da aprendizagem extraídas do Moodle

Variável	Descrição sobre as variáveis	Construto	Sig.
VAR01	Quantidade de diferentes locais (IP's) a partir dos quais a(o) aluna(o) acessou o ambiente.	Estrut. do ambiente	0,43
VAR02	Quantidade de mensagens enviadas por aluna(o) às(os) Professoras(es) pelo ambiente.		
VAR03	Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente.		
VAR04	Quantidade geral de mensagens enviadas pela(o) aluna(o) dentro do ambiente.		
VAR05	Quantidade geral de mensagens recebidas pela(o) aluna(o) dentro do ambiente.		
VAR06	Quantidade de tópicos criados pelo(a) aluno(a) em fórum do tipo "tira-dúvidas".		
VAR07	Quantidade de postagens no "Fórum tira dúvidas";		
VAR08	Quant. de postagens de um(a) aluno(a) em fóruns que foram respondidas por outros(as) alunos(as).		
VAR09	Quantidade de postagens de um(a) aluno(a) em fóruns que foram respondidas pelo(a) professor(a) ou tutor(a).		
VAR10	Quantidade de colegas diferentes para quem o(a) aluno(a) enviou mensagens dentro do ambiente.		
VAR12	Quantidade de visualizações da aba "Conteúdo" do curso, onde constam os arquivos com o conteúdo programático do curso		
VAR13	Horário que mais realizou atividades;		
VAR14	Turno do dia em que realizou mais atividades.		
VAR16	Quantidade de atividades entregues por um(a) aluno(a) fora do prazo, por disciplina;		
VAR17	Tempo médio entre a abertura da atividade e sua submissão;		
VAR18	Quantidade de leituras feitas ao fórum (pageviews);		
VAR20	Quantidade de respostas ao tópico principal (refazer opinião em fórum);		
VAR21	Quantidade de pageviews ao quadro de notas;		
		Autoavaliação	0,43

VAR22	Quantidade de vezes que o aluno visualiza o (Checklist Atividades)		
VAR23	Quantidade de visualizações de notas por atividade;		
VAR24	Média semanal da quantidade de acessos de um(a) aluno(a) ao ambiente.		
VAR25	Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno;		
VAR28	Quantidades de Time Out;		
VAR31	Quantidade de acessos do(a) aluno(a) ao ambiente.		
VAR31b	Quantidade de dias distintos que o aluno entrou na disciplina		
VAR31c	Quantidade de dias distintos que o aluno entrou na plataforma		
VAR32a	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Manhã).		
VAR32b	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Tarde).		
VAR32c	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Noite).		
VAR32d	Quantidade de acessos do(a) aluno(a) ao ambiente por turno (Madrugada).		
VAR33	Quantidade de atividades entregues por um(a) aluno(a) no prazo, por disciplina;		
VAR34	Quantidade geral de postagens de um(a) aluno(a) em fóruns.		
VAR35	Quantidade de respostas de um(a) professor(a) para as dúvidas de alunos(as) em fóruns.		

Fonte: Elaborada pelo Autor (2016)

Como produto final desta fase, identificamos um conjunto de trinta e três variáveis comportamentais significativas, armazenadas na plataforma, os quais representaram o fenômeno de autorregulação da aprendizagem, possibilitando a eliminação do questionário, sendo possível utilizarmos apenas o conjunto de variáveis comportamentais registrados na plataforma para o prosseguimento das análises.

A próxima seção aborda o processo de extração de todas as bases de dados referente a um período de sete anos de curso.

6.2.2 Extração das bases de dados históricas da plataforma Moodle

As bases coletadas referem-se a quatro cursos superiores, dos quais foram obtidos dados de sessenta e seis semestres e duzentas e duas disciplinas distintas, totalizando 18,9 Gigabytes de dados provindos de doze plataformas diferentes com sistemas de SGBD variando entre Postgres¹³ e Mysql¹⁴.

¹³ <https://www.postgresql.org/>

¹⁴ <https://www.mysql.com/>

A Figura 25 mostra os dados coletados referente ao curso de Bacharelado em Administração Pública. Ao todo foram disponibilizados dados de quatro instâncias de plataformas LMS distintas. Foram coletados seis semestres de uma turma referente a uma primeira oferta que iniciou em 2010.1

Figura 25: Bases de dados do curso de Administração

Administração	2010.1	2010.2	2011.1	2011.2	2013.2	2014.2	2015.1	2015.2	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	8º Per.				
2ª Oferta					3º Per.	5º Per.	6º Per.	7º Per.	8º Per.
3ª Oferta						1º Per.	2º Per.	3º Per.	4º Per.

Posteriormente foram coletados quatro períodos referentes a uma segunda oferta e em seguida quatro períodos referentes a uma terceira oferta.

A Figura 26 mostra a coleta de dados referente ao curso de Licenciatura em Pedagogia. Ao todo foram coletados dezesseis semestres distribuídos ao longo de cinco anos divididos em quatro ofertas. A primeira oferta teve duração de sete períodos que ocorreram em duas instâncias de plataformas distintas.

Figura 26: Bases de dados do curso de Pedagogia

Pedagogia	2010.2	2011.1	2011.2	2012.1	2013.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	6º Per.	7º Per.	8º Per.			
2ª Oferta					2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
3ª Oferta								1º Per.	2º Per.	4º Per.

Referente a segunda oferta foram disponibilizadas quatro instâncias distintas de LMS e um total de seis semestres. A terceira instância teve três semestres divididos em três instâncias distintas.

A Figura 27 mostra a disponibilização das bases de dados referentes ao curso de Licenciatura em Biologia. Para este curso foram coletados um total de dezoito semestres distribuídos ao longo de cinco anos. Ao longo destes anos foram utilizadas seis instâncias de LMS distintas. Para a terceira oferta foram coletados oito semestres.

Figura 27: Bases de dados do curso de Biologia

Biologia	2011.1	2011.2	2012.1	2012.2	2013.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta										
2ª Oferta										
3ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	7º Per.	8º Per.		
4ª Oferta				1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
5ª Oferta								1º Per.	2º Per.	4º Per.

Em seguida coletamos sete semestres para a quarta oferta e três semestres para a quinta oferta de cursos. Por fim o último curso disponibilizado foi o curso de Licenciatura em Letras. Para este curso coletamos um total de vinte semestres durante um período de cinco anos e meio divididos em cinco instâncias de LMS distintas.

Os dados referentes à primeira oferta e os dados referentes a reoferta foram armazenados em uma única instância de LMS. Para a segunda oferta tiveram três semestres letivos em duas instâncias distintas.

Figura 28: Bases de dados do curso de Letras

Letras	2009.2	2010.1	2010.2	2011.1	2011.2	2012.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.					
Reoferta (1ª)				1º Per.	2º Per.	3º Per.					
2ª Oferta				1º Per.	2º Per.	3º Per.	6º Per.	7º Per.	8º Per.		
3ª Oferta							3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
3ª Oferta									1º Per.	2º Per.	4º Per.

A terceira oferta foi a que teve o maior número distinto de instâncias da plataforma: foram quatro instâncias distintas e um total de sete semestres.

Após a coleta de toda as bases de dados realizamos uma verificação para identificar o número de alunos referentes a cada um dos quatro cursos. A Tabela 2 sumariza o quantitativo de alunos por curso e suas respectivas porcentagens. Após a integração de todas a bases de dados obtivemos um total de 30.217 alunos.

Tabela 2: Quantitativo geral de alunos por curso.

Cursos	Frequência (fi)	Percentual (%)
Administração Pública	2892	9,57%
Licenciatura em Biologia	6526	21,60%
Licenciatura em Letras	6297	20,84%
Licenciatura em Pedagogia	14502	47,99%
Total	30217	100,00%

Fonte: Elaborada pelo Autor (2016)

O curso disponibilizado com o maior percentual de alunos foi a Licenciatura em Pedagogia, seguido pelo curso de Licenciatura em Biologia. A subseção seguinte descreve o processo de seleção de variáveis bem como as fases de preparação dos dados para a modelagem.

6.3 Resultados sobre a preparação dos dados

De posse da identificação do conjunto de variáveis e as bases de dados coletadas, esta fase buscou aplicar técnicas de pré-processamento para a transformação das variáveis para a fase de criação dos modelos preditivos (modelagem).

6.3.1 Transformação e limpeza das variáveis de autorregulação

Após a extração do conjunto de variáveis de autorregulação foi necessário realizar dois tipos de pré-processamentos nas variáveis originais. Estas variáveis passaram por um processo de identificação e remoção de *outliers* e preenchimento de dados faltantes, especificamente frequências sem ocorrências que foram preenchidas com valores mínimos.

Optamos por não aplicar nenhuma transformação do tipo discretização nas variáveis de autorregulação (explicativas), pois tivemos como meta desenvolver todo o processo de modelagem com as variáveis originais, diminuindo assim a carga de preprocessamento e facilitando a comunicação da aplicação proposta (software dashboard) diretamente com os dados originais provindos das plataformas de LMS.

A variável responsável por compor o desempenho dos alunos (variável alvo) passou por um processo de discretização com o objetivo de transformação de contínua para binária. A próxima subseção descreve o processo de composição da variável *desempenho*, bem como suas formas de discretização e binarização.

6.3.2 Extração das variáveis relacionadas ao desempenho dos estudantes

Tendo posse da base de dados com as variáveis comportamentais, iniciamos a coleta das variáveis relacionadas às atividades avaliativas dos alunos, que fizeram parte da composição da variável alvo (*desempenho*). Esta variável foi composta por provas, reposições de provas, discussões avaliativas via fórum e atividades em formato de webquest.

O desempenho de cada um dos alunos foi contabilizado com a realização de duas provas presenciais, quatro interações avaliativas por meio do fórum de discussão e duas atividades no formato de webquest.

O Quadro 12 mostra a descrição de cada um dos instrumentos de avaliação, bem como sua descrição e suas reposições.

Quadro 12: Variáveis de desempenho acadêmico

Variável	Descrição sobre as variáveis	Categoria
PROVA01	Nota da primeira prova presencial	
PROVA01_2CHAMADA	Segunda chamada da primeira prova presencial	
PRIMEIRA_PROVA	Nota da primeira prova ou da segunda chamada	
PROVA02	Nota da segunda prova presencial	
PROVA02_2CHAMADA	Segunda chamada da segunda prova presencial	
SEGUNDA_PROVA	Nota da segunda prova ou da segunda chamada	
MEDIA_PROVAS	Média geral das duas provas presencial	
FORUM01	Nota do primeiro fórum	
FORUM02	Nota do segundo fórum	
FORUM03	Nota do terceiro fórum	
FORUM04	Nota do quarto fórum	
MEDIA_FORUM	Média geral dos quatro fóruns	
WEBQUEST01	Nota da primeira atividade (webquest)	
WEBQUEST02	Nota da segunda atividade (webquest)	
MEDIA_WEBQUEST	Média geral das duas atividades	
DESEMPENH0	Desempenho final	Média final

Fonte: Elaborada pelo Autor (2016)

A variável “DESEMPENH0”, utilizada como sendo a variável alvo, foi composta pela ponderação das notas referentes aos seguintes pesos: prova = 5,5, fóruns = 2,0 e webquests (atividades) = 2,5. Levamos em consideração provas de reposição, caso o aluno tenha perdido uma das avaliações.

A fórmula seguinte sumariza a equação utilizada para calcular o desempenho de cada aluno:

$$\text{Desempenho} = (\text{MEDIA}_{\text{PROVAS}} * 5,5) + (\text{MEDIA}_{\text{FORUM}} * 2,0) + (\text{MEDIA}_{\text{WEBQUEST}} * 2,5)$$

Após a construção da variável *desempenho*, de natureza contínua, a mesma passou por um processo de transformação e categorização. O processo de categorização aconteceu de três formas distintas.

A primeira categorização foi baseada na construção de três classes, onde a primeira classe foi categorizada no intervalo de nota de 0,0 a 4,0, a segunda classe acima de 4,0 e abaixo 7,0 e a terceira classe com notas maior ou igual a 7,0.

A segunda categorização teve como meta transformar a variável contínua em uma variável binária. A primeira classe recebeu valores de notas abaixo de 4,0 e a segunda classe notas iguais ou maiores que 4,0. Nesta binarização o objetivo foi dividir os grupos em alunos em situação de risco de reprovação direta e alunos com chance de aprovação.

A terceira categorização foi também do tipo transformação binária com valores de notas abaixo de 7,0 para a primeira classe e notas acima de 7,0 para a segunda classe. Nesta binarização o objetivo foi dividir os grupos em alunos aprovados direto e alunos reprovados com chance de recuperação.

6.4 Resultados da Modelagem e Validação

Nesta fase são abordados os resultados referentes à construção dos modelos preditivos e suas respectivas validações. Utilizamos a tarefa de classificação, que segundo Silva, Peres e Boscarioli (2016) é o processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence qualquer exemplar de um domínio sob análise, com base em um conjunto de dados já classificados (treinamento).

A criação de um modelo de classificação depende da qualidade dos dados e dos algoritmos utilizados para aprendizagem (ou treinamento). O modelo criado precisa ser validado, verificando se atende a condições impostas para solucionar o problema.

O propósito desta fase foi obter modelos de classificação de desempenho do estudante, com base nas variáveis de autorregulação da aprendizagem.

6.4.1 Construção dos modelos de classificação

Construímos quatro scripts (*Apêndice G*) de algoritmos de classificação. O primeiro *script* foi referente ao modelo de Regressão Logística, o segundo foi o algoritmo de Árvore de Decisão CART, o terceiro foi o algoritmo de Máquina de Vetor de Suporte (SVM) e o quarto foi o algoritmo de Naive Bayes.

Para cada um dos quatro algoritmos de classificação foi feito um experimento com a variável alvo categorizada de três formas de categorização distintas: (A) Binaizada com a primeira classe sendo valores de notas menores do que 4,0 e igual ou acima de 4,0, (B) binaizada com a primeira classe sendo valores de notas menores

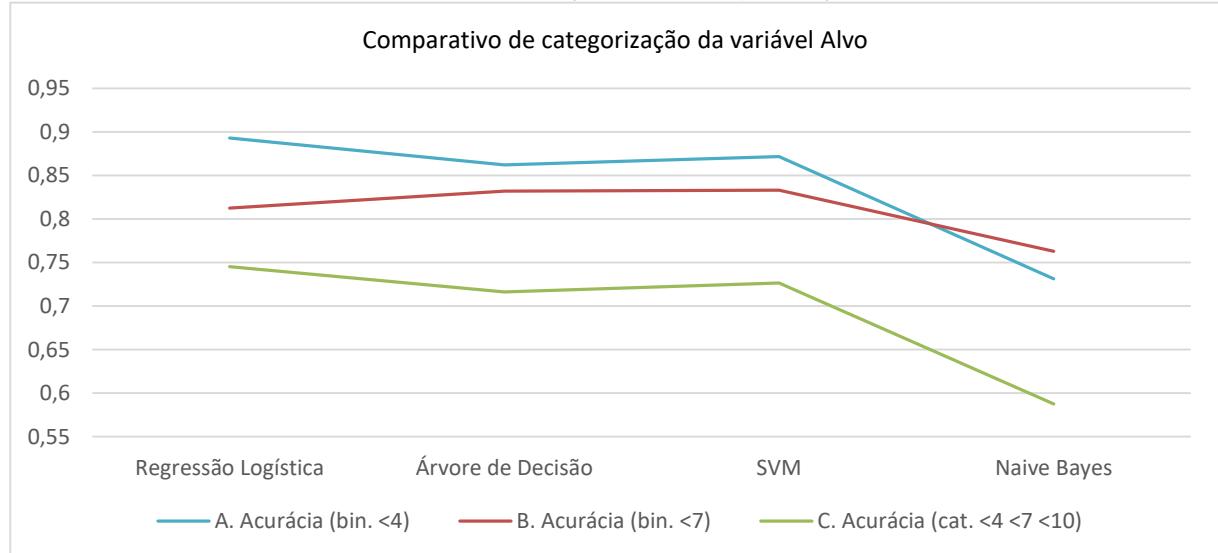
do que 7,0 e igual ou acima de 7,0 e (C) categorizada em três classes, abaixo de 4,0, entre 4,0 e 7,0 e acima de 7,0. O objetivo desta fase foi entender o melhor limiar de categorização da variável alvo de acordo com o resultado da acurácia para cada um dos classificadores.

Para o treinamento e testes dos modelos, os casos de análise foram divididos em duas bases de dados, a primeira com 70% dos casos reservados para o fase de treinamento do classificador e 30% reservados para a fase de teste.

A base de treinamento foi utilizada para construir três modelos para cada um dos quatro algoritmos de classificação. Cada um dos modelos foi construído levando em consideração os três tipos de categorização.

A Figura 29 mostra o gráfico referente à acurácia de cada um dos classificadores por categorização.

Figura 29: Gráfico de comparação da categorização da variável alvo



Fonte: Elaborada pelo Autor (2016)

A Categorização A (binalização < 4) foi a que obteve melhores resultados de acurácia, exceto para o algoritmo Naive Bayes. Esta categorização obteve resultados de acurácia considerados excelentes (próximo dos 89,6%) para o algoritmo de regressão logística.

A Categorização B (binalização < 7) obteve bom resultado apenas para o algoritmo Naive Bayes que obteve um valor de acurácia de 76%. Esta categorização se mostrou inferior à categorização B se levar em consideração os demais algoritmos.

A categorização que obteve piores resultados foi a C, com três classes. O melhor resultado para esta categorização foi com o algoritmo de regressão logística que obteve valor próximo a 75% de acurácia.

Como forma de prosseguirmos com as análises, estabelecemos que as próximas fases do processo de modelagem teriam como padrão a categorização A, ou seja, que a variável alvo (Desempenho) seria categorizada de forma binária seguindo os limiares da categorização A. A próxima seção descreve a validação dos quatro modelos de classificação.

6.4.2 Validação dos modelos de classificação

A validação buscou verificar como os modelos são executados nos dados reais. É importante validar os modelos de mineração entendendo suas qualidades e características antes de implantá-los em um ambiente de produção (CHAPMAN, CLINTON, *et al.*, 2000). Para a validação utilizamos a base de treinamento que foi composta por 11.824 observações pertencentes à classe 0 (alunos pertencentes a classe de desempenho insatisfatório) e 9.330 observações pertencentes à classe 1 (alunos em situação de desempenho insatisfatório).

A validação foi realizada a partir dos valores de Acurácia, Sensibilidade, Especificidade, Precisão e curva ROC. A Tabela 3 apresenta os resultados das métricas de avaliação para cada um dos classificadores (Regressão Logística, Árvore de Decisão, SVM e Naive Bayes).

Tabela 3: Comparação entre algoritmos de classificação

Métricas de Avaliação	Algoritmos de Classificação			
	Regressão Logística	Árvore de Decisão	SVM	Naive Bayes
Accuracy	0,893	0,8622	0,8717	0,7312
No Information Rate	0,5561	0,5572	0,5634	0,5596
P-value [Acc > NIR]	< 2,2e-16	< 2,2e-16	<2e-16	< 2,2e-16
Kappa	0,7825	0,7197	0,7388	0,4755
Mcnemar's Test P-Value	3,44E-02	0,000119	0,197	< 2,2e-16
Sensitivity (Recall)	0,8633	0,8273	0,8474	0,8815
Specificity	0,9167	0,8899	0,8905	0,613
Pos Pred Value (Precision)	0,8921	0,8566	0,8571	0,6419
Neg Pred value	0,8936	0,8664	0,8827	0,868
Prevalence	0,4439	0,4428	0,4366	0,4404
Detection Rate	0,3832	0,3663	0,37	0,3882
Detection Prevalence	0,4295	0,4277	0,4316	0,6048
Balanced Accuracy	0,8900	0,8586	0,8689	0,7472

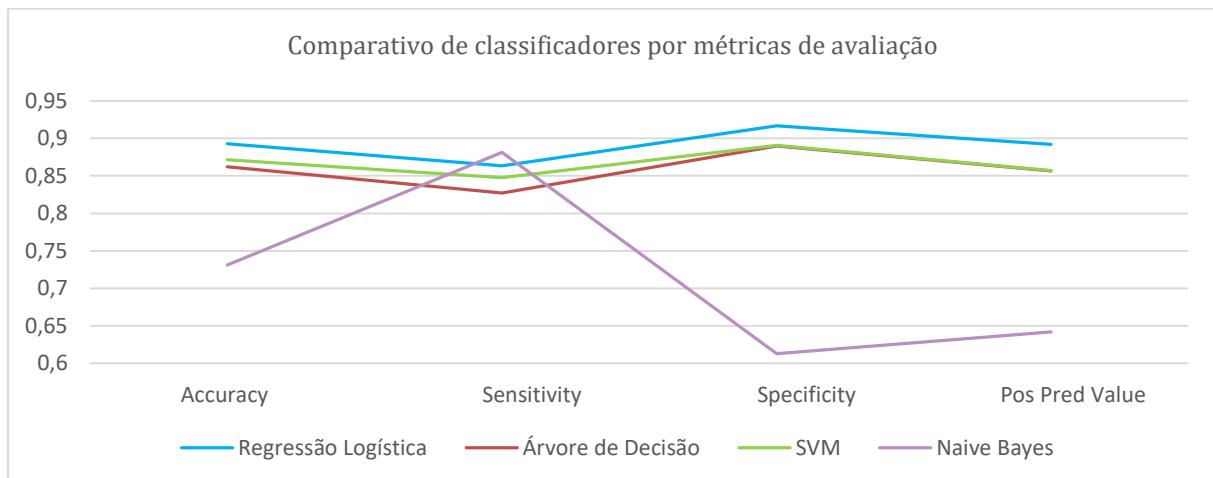
Fonte: Elaborada pelo Autor (2016)

A Figura 30 exibe o gráfico que mostra os resultados das quatro principais métricas de avaliação de modelos utilizadas nesta tese. Em relação à métrica de Acurácia, a Regressão Logística teve o melhor resultado (0,893) seguida pelo classificador SVM (0,8717).

Em relação a métrica sensibilidade (*recall*) o classificador que obteve melhor desempenho foi Naive Bayes (0,8815), seguido pelo classificador e Regressão Logística (0,8633). Esta foi a única métrica na qual o classificador Naive Bayes teve resultados satisfatórios em relação aos outros classificadores.

As duas últimas métricas utilizadas foram a especificidade e precisão (*Pos pred Vale*). O modelo de Regressão Logística obteve resultados superiores em ambas as métricas. Para a especificidade obteve (0,9167) e em relação à precisão, o classificador obteve (0,8921).

Figura 30: Gráfico comparativo de algoritmos de classificação

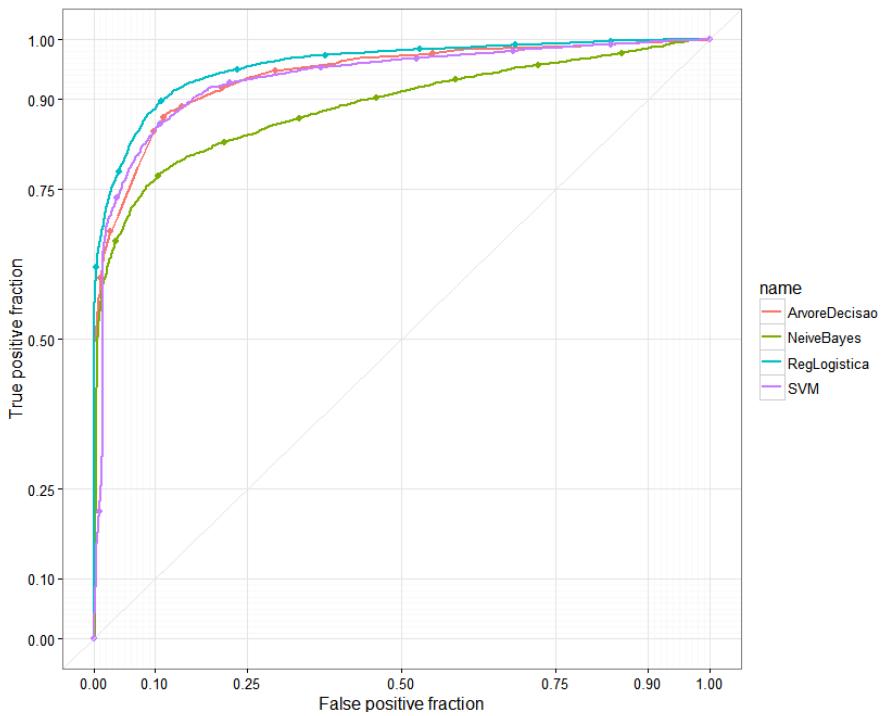


Fonte: Elaborada pelo Autor (2016)

Outra medida que foi utilizada nesta tese para a escolha do melhor modelo de classificação foi o cálculo da curva ROC. Esta medida é visualizada por meio de uma representação gráfica que esclarece o desempenho de um sistema classificador binário e, além disso, permite estudar a variação da sensibilidade e especificidade para diferentes valores de corte (FAWCETT, 2006).

A Figura 31 apresenta o gráfico da curva ROC, no qual é possível visualizar o comparativo entre os quatro modelos de classificação desenvolvidos nesta tese. Todo o código referente à geração do comparativo entre os classificadores econtra-se no *Apêndice I*.

Figura 31: Curva ROC para comparativo entre classificadores



Fonte: Elaborada pelo Autor (2016)

Em relação à curva ROC para os quatro classificadores, obtivemos os seguintes resultados: Árvore de Decisão (0,9157), Naive Bayes (0,8850), Regressão Logística (0,9574), SVM (0,9291). Sendo possível verificar, além das métricas mostradas anteriormente, a métrica de área sob a curva ROC mostra que o classificador logístico foi o que teve o melhor resultado em termos de poder de classificação.

Estes resultados corroboram com os estudos de Sabourin, Mott e Lester (2012) e Sabourin, Shores, *et al.*, (2012), nos quais os autores obtiveram resultados promissores quanto ao uso de classificador de regressão logística, para previsão de desempenho de alunos do ensino fundamental, baseados em comportamentos autorregulatórios, provindos de dados da plataforma Crystal Island, um ambiente de aprendizagem baseado em jogos.

Tendo em vista o bom desempenho do modelo logístico, na próxima seção será realizada uma descrição mais detalhada do processo de modelagem deste classificador, bem como uma análise detalhada do processo de validação e ajuste de parâmetros.

6.4.3 Modelo de Regressão Logística

Por meio do modelo de regressão logística foi possível entendermos o comportamento da variável dependente binária (Desempenho) com base em um conjunto de variáveis independentes numéricas ou categóricas (variáveis de autorregulação da aprendizagem):

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \quad (6.1)$$

Os modelos de Regressão Logística são formulados por (6.1), onde $\beta_0, \beta_1, \dots, \beta_n$, são os coeficientes das variáveis que explicam a ocorrência de um determinado evento.

A grande vantagem da regressão logística, para esta tese, em relação a outros algoritmos de classificação, está no fato da mesma permitir não somente a previsão de ocorrência de um determinado evento de interesse, mas promovem a capacidade de apresentar a probabilidade de sua ocorrência, bem como as razões de chance (*Odd Ratios*) para cada uma das variáveis de autorregulação. Isto permite explicar o poder de influência de cada uma das variáveis em relação ao desempenho do aluno.

Para a construção do modelo de classificação logística foi utilizada uma base de treinamento de 21.154 observações, o que correspondeu a 70% do total de dados da base. Esta base de treinamento foi composta por 11.824 observações pertencentes a classe 0 (alunos pertencentes a classe de desempenho insatisfatório) e 9.330 observações pertencentes à classe 1 (alunos em situação de desempenho insatisfatório). Foi utilizada a técnica de Validação Cruzada (*Cross-Validation*) com 10 *folds*. Essa técnica possibilitou aumentar a capacidade de generalização do modelo treinado, a partir do conjunto de dados. Esta técnica é amplamente empregada em problemas nos quais o objetivo da modelagem é a predição (STONE, 1974)

Avaliação do ajuste de estimação do modelo

No modelo de regressão logística uma das validações que observamos inicialmente é o teste de hipóteses sobre os coeficientes individualmente. O valor do teste da estatística *Wald* é utilizado para avaliar a significância de cada coeficiente estimando. Se o coeficiente logístico é estatisticamente significativo ($< 0,05$), podemos interpretá-lo em termos de como o mesmo impacta na previsão de pertinência ao grupo.

Na saída do Quadro 13 é possível observarmos que as variáveis *VAR06*, *VAR07*, *VAR08*, *VAR09*, *VAR13*, *VAR14*, *VAR17* e *VAR23* não apresentaram significância estatística e por sua vez tiveram que ser retiradas da construção do modelo final.

Quadro 13: Resultados de significância estatística por variável

	Coefficients:	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.5961457	0.0665924	54.002	< 2e-16	***	
VAR01	0.0053145	0.0012980	4.094	4.23e-05	***	
VAR02	0.0971106	0.0272446	3.564	0.000365	***	
VAR03	-0.0140579	0.0045620	-3.082	0.002060	**	
VAR04	0.0025958	0.0013660	1.900	0.057381	.	
VAR05	0.0024620	0.0006952	3.541	0.000398	***	
VAR06	-0.2819494	0.2156452	-1.307	0.191054	.	
VAR07	0.0582210	0.1272019	0.458	0.647164	.	
VAR08	0.0173138	0.0212034	0.817	0.414182	.	
VAR09	0.0280463	0.0183993	1.524	0.127430	.	
VAR10	-0.0669658	0.0139889	-4.787	1.69e-06	***	
VAR12	-0.0689171	0.0289311	-2.382	0.017214	*	
VAR13	0.0008412	0.0055663	0.151	0.879882	.	
VAR14	-0.0393113	0.0371802	-1.057	0.290366	.	
VAR16	-1.0816371	0.0592768	-18.247	< 2e-16	***	
VAR17	0.0013450	0.0017969	0.749	0.454144	.	
VAR18	0.0110846	0.0018030	6.148	7.86e-10	***	
VAR20	0.3991621	0.0787712	5.067	4.03e-07	***	
VAR21	0.1089689	0.0137360	7.933	2.14e-15	***	
VAR22	-0.0190252	0.0096252	-1.977	0.048086	*	
VAR23	-0.0165666	0.0211643	-0.783	0.433767	.	
VAR24	-0.1338515	0.0721790	-1.854	0.063677	.	
VAR25	-0.0060239	0.0004758	-12.661	< 2e-16	***	
VAR28	0.0013602	0.0007590	1.792	0.073119	.	
VAR31	0.0175758	0.0061181	2.873	0.004069	**	
VAR31b	-0.0156180	0.0027218	-5.738	9.58e-09	***	
VAR31c	-0.0106327	0.0021396	-4.969	6.72e-07	***	
VAR32a	-0.0224103	0.0056532	-3.964	7.36e-05	***	
VAR32b	-0.0146981	0.0055969	-2.626	0.008636	**	
VAR32c	-0.0151420	0.0057026	-2.655	0.007924	**	
VAR33	-1.6689937	0.0412373	-40.473	< 2e-16	***	
VAR34	-0.3363416	0.0189809	-17.720	< 2e-16	***	
VAR35	0.2908787	0.1120174	2.597	0.009412	**	
<hr/>						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Fonte: Elaborada pelo Autor (2016)

Existem algumas abordagens para auxiliar na seleção de variáveis mais significativas para a construção do modelo. Neste trabalho, utilizamos a abordagem *stepwise*. O método *stepwise* é o processo de estimativa de modelos estatísticos em que as variáveis independentes são adicionadas ou retiradas do modelo de acordo com o poder de discriminação que agregam ao grupo de variáveis explicativas (HAIR, BLACK, et al., 2009). O Quadro 14 mostra as variáveis significativas após o *stepwise*.

Quadro 14: Resultados de significância estatística por variável após o stepwise

Coefficients:		Estimate	Std. Error	z	value	Pr(> z)	
(Intercept)	3.6022762	0.0659374	54.632	< 2e-16	***		
VAR01	0.0052704	0.0012947	4.071	4.69e-05	***		
VAR02	0.0984237	0.0271719	3.622	0.000292	***		
VAR03	-0.0137607	0.0045174	-3.046	0.002318	**		
VAR04	0.0025634	0.0013643	1.879	0.060256	.		
VAR05	0.0024759	0.0006843	3.618	0.000296	***		
VAR10	-0.0661682	0.0139603	-4.740	2.14e-06	***		
VAR12	-0.0686481	0.0288113	-2.383	0.017187	*		
VAR16	-1.0815550	0.0590862	-18.305	< 2e-16	***		
VAR18	0.0109902	0.0017967	6.117	9.55e-10	***		
VAR20	0.4143110	0.0764235	5.421	5.92e-08	***		
VAR21	0.1057263	0.0133887	7.897	2.86e-15	***		
VAR22	-0.0189974	0.0095872	-1.982	0.047530	*		
VAR24	-0.1308503	0.0719442	-1.819	0.068946	.		
VAR25	-0.0060779	0.0004704	-12.921	< 2e-16	***		
VAR28	0.0013142	0.0007570	1.736	0.082548	.		
VAR31	0.0177954	0.0060928	2.921	0.003492	**		
VAR31b	-0.0156337	0.0026765	-5.841	5.18e-09	***		
VAR31c	-0.0108233	0.0021315	-5.078	3.82e-07	***		
VAR32a	-0.0226223	0.0056171	-4.027	5.64e-05	***		
VAR32b	-0.0149143	0.0055568	-2.684	0.007276	**		
VAR32c	-0.0154217	0.0056644	-2.723	0.006477	**		
VAR33	-1.6658986	0.0384280	-43.351	< 2e-16	***		
VAR34	-0.3320120	0.0178559	-18.594	< 2e-16	***		
<hr/>							
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1							

Fonte: Elaborada pelo Autor (2016)

Após a aplicação da redução de variáveis, por meio do método *stepwise*, o modelo final ficou com 23 variáveis significativas pertencentes ao modelo final. Dentre estas variáveis, uma variável faz parte do construto **Estruturação do ambiente** (VAR01), cinco variáveis fizeram parte do construto **Busca por Ajuda** (VAR02, VAR03, VAR04, VAR05, VAR10), três variáveis relacionadas ao construto **Estratégia de Realização de Tarefas** (VAR12, VAR16, VAR18), três variáveis fizeram parte do construto **Autoavaliação** (VAR20, VAR21, VAR22), nove variáveis fizeram parte do construto **Gestão do Tempo** (VAR24, VAR25, VAR28, VAR31, VAR31b, VAR31c, VAR32a, VAR32b, VAR32c) e duas variáveis fizeram parte do construto **Estabelecimento de Metas** (VAR33, VAR34).

É possível verificarmos que os construtos que obtiveram a maior quantidade de variáveis significativas foram os construtos Busca por Ajuda e Gestão do Tempo.

Em relação ao construto Busca por Ajuda, as variáveis que tiveram o maior poder de influência, em relação ao desempenho do aluno, foram as variáveis VAR02 e VAR10, onde a primeira representou a Quantidade de mensagens enviadas por aluna(o) às(os) Professoras(es) pelo ambiente, seguida pela segunda variável mais

significativa que representou a Quantidade de colegas diferentes para quem o(a) aluno(a) enviou mensagens dentro do ambiente.

Em relação ao construto Gestão do Tempo, as variáveis que apresentaram o maior poder de influência em relação ao desempenho foram as varáveis VAR25 e VAR32b. A variável VAR25 representou o Tempo médio entre a criação de um tópico no fórum temático e a primeira postagem do aluno, ou seja, essa variável buscou registrar o tempo que o aluno levou para responder a atividade, após a mesma ter sido disponibilizada no ambiente. É importante observarmos que essa variável teve um poder de influência negativo, ou seja, existe uma relação entre o aumento do tempo para a realização da atividade e o baixo desempenho do aluno.

Por fim, foi possível construir o modelo de regressão logística, representando uma função *logit* que englobou todas as variáveis significativas apresentadas no Quadro 13. A função representativa do modelo de classificação logístico é dada por:

$$\text{logit}(p_{des.}) = 3,602 + var_1 * 0,005 + \dots + var_{34} * -0,332 \quad (6.2)$$

Após o teste de significância das variáveis e construção do modelo foram obtidas as medidas de ajuste geral do modelo, por meio do *pseudo R²* e a medida de *R² de Cox e Snell*. Estas medidas de ajuste foram necessárias para verificar o quanto o modelo desenvolvido se ajusta aos dados da pesquisa.

Estas medidas são interpretadas como refletindo a quantia de variação explicada pelo modelo logístico, com valor 1 indicando ajuste perfeito. Para o modelo desenvolvido nesta tese obtivemos valores de *R²* e *R² de Cox e Snell* de 0,785 e 0,776 respectivamente, nos indicando que o modelo tem bom ajuste aos dados e que podemos utilizar o modelo de regressão logística para a construção do classificador.

A seguir realizamos outra avaliação, por meio da precisão preditiva, que buscou verificar o quanto o modelo consegue classificar os estudantes (em desempenho satisfatório ou não) de forma correta.

Avaliação da precisão preditiva

Nesta fase avaliamos a precisão preditiva do modelo de regressão logística. Foi gerada a matriz de confusão e os resultados das métricas de avaliação da precisão preditiva. Para esta fase de avaliação da precisão foi utilizada uma base de teste com um total de 9.063 observações, que representou 30% do total de observações da base

total. A matriz de confusão gerada (Quadro 15) representou os níveis de precisão preditiva atingidos pelo modelo logístico.

Quadro 15: Matriz de confusão para o algoritmo de Regressão Logística

		Reference
Prediction	0	1
0	4620	550
1	420	3473
 Accuracy : 0.893 95% CI : (0.8864, 0.8993) No Information Rate : 0.5561 P-Value [Acc > NIR] : < 2.2e-16		
 Kappa : 0.7825 McNemar's Test P-Value : 3.444e-05		
 Sensitivity : 0.8633 Specificity : 0.9167 Pos Pred Value : 0.8921 Neg Pred Value : 0.8936 Prevalence : 0.4439 Detection Rate : 0.3832 Detection Prevalence : 0.4295 Balanced Accuracy : 0.8900		
 'Positive' Class : 1		

Fonte: Elaborada pelo Autor (2016)

A matriz de confusão gerada mostra que o classificador conseguiu um valor de **acurácia** de 89,3% de previsão correta, o que indica uma ótima taxa de previsão, de acordo com os limiares estabelecidos por Faveiro, Belfiore, *et al.* (2009). Além da métrica de acurácia, outras métricas foram utilizadas para verificar a precisão de classificação do modelo, tais como: *Sensibilidade*, *Especificidade*, *Precisão* e a área sobre a curva ROC.

Para a métrica de **sensibilidade** (*Sensitivity*), o modelo atingiu um valor de 0,8633 (86,33%). Essa métrica é definida como a fração das instâncias positivas previstas corretamente pelo classificador, ou seja, é a quantidade de alunos em situação de risco que foram previstos corretamente. É possível afirmar que dos 4.023 alunos em situação de risco, que passaram pelo classificador, o modelo conseguiu acertar 3.474 alunos corretamente.

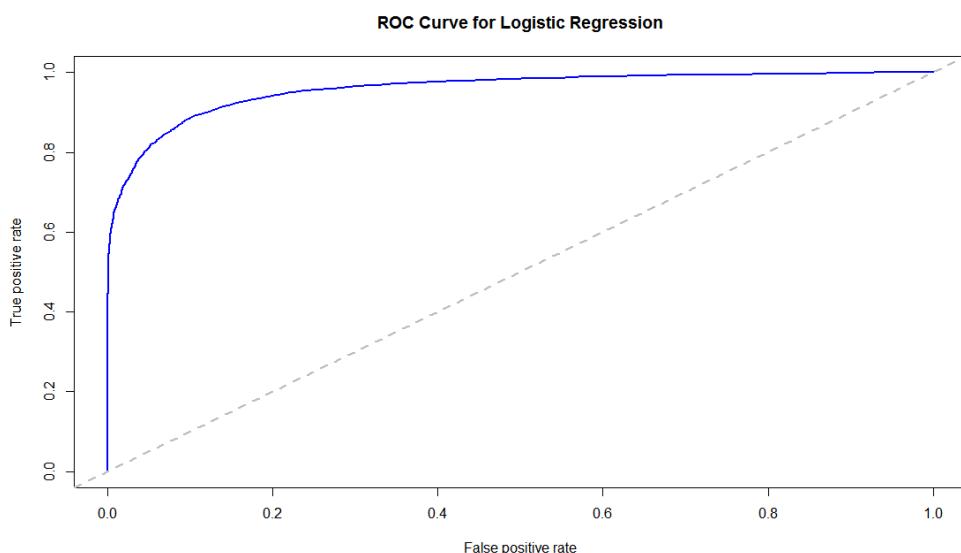
Outra métrica avaliada foi a **especificidade** (*Specificity*). Esta métrica é definida como a fração de instâncias negativas corretamente previstas pelo classificador, ou seja, é a quantidade de alunos que não se encontram em situação de risco de desempenho que foram previstos corretamente. Para esta métrica

obtivemos um valor de 0,9167 (91,67%) onde podemos afirmar que dos 5.040 alunos que se encontravam em situação satisfatória de desempenho, 4.620 deles foram classificados corretamente pelo modelo logístico.

Para a métrica **precisão** (*Pos Pred Value*) obtivemos um valor de 0,8921 (89,21%). Esta métrica determina o percentual de registros que são positivos no grupo que o classificador previu como classe positiva. Quanto maior o percentual de precisão, menor será o número de erros falsos positivos pelo classificador, ou seja, quanto maior a precisão menor será a quantidade de alunos que estão em situação de risco, mas que o classificador errou e os classificou como alunos em uma situação de desempenho satisfatório.

A última métrica avaliada foi a **área sobre curva ROC**. Para esta métrica obtivemos um valor de 0,9574. A construção do gráfico de curva ROC é baseada na taxa de verdadeiros positivos e na taxa de falsos positivos, onde a construção do gráfico é feita por meio da plotagem dos falsos positivos no eixo das ordenadas (eixo x) e verdadeiros positivos no eixo das abscissas (eixo y). A Figura 37 mostra a imagem do gráfico plotado por meio do cálculo da área ROC.

Figura 32: Curva ROC do algoritmo de Regressão Logística



Fonte: Elaborada pelo Autor (2016)

A área da curva é uma porção da área do quadrado unitário (espaço ROC) seus valores variam entre 0 e 1. Quanto maior a área, melhor a performance global do classificador.

De acordo com a construção do modelo e a avaliação das métricas de avaliação da precisão positiva, podemos concluir que a modelagem realizada por meio da construção do modelo de regressão logística apresentou-se como sendo o melhor dos quatro classificadores utilizados e avaliados nesta tese.

Como forma de materializar o conhecimento adquirido nesta fase de mineração dos dados, especificamente na modelagem de um classificador de desempenho. Projetamos uma interface de software *dashboard* que buscou tornar transparente para o professor a utilização do modelo de classificação, desenvolvido por meio da equação (6.2). A próxima seção aborda os resultados do processo de desenvolvimento e avaliação da solução desenvolvida.

6.5 Resultados do desenvolvimento e implementação

Para a geração das alternativas de soluções de software, concebidas a partir da descoberta de conhecimento nas fases anteriores, utilizamos a técnica de prototipagem de baixa fidelidade e prototipagem de alta fidelidade com envolvimento dos usuários na fase de avaliação.

O objetivo desta fase foi desenvolver uma interface de *dashboard*, no qual o professor/tutor possa utilizar dos benefícios da mineração de dados, especificamente modelagem preditiva, sem a necessidade de conhecimentos técnicos na área. As próximas subseções mostram o processo de concepção e validação da solução.

6.5.1 Protótipos de baixa fidelidade

Nesta subseção são demonstrados os resultados do desenvolvimento dos protótipos de baixa fidelidade. Foram desenhados casos de uso e protótipos de baixa fidelidade que puderam ser refinados ao longo do desenvolvimento da solução (Apêndice L).

A Figura 33 mostra a primeira tela do protótipo desenvolvido. Nesta tela o professor tem a possibilidade de selecionar o módulo que deseja utilizar. Tendo as opções entre: Visão geral dos dados, Análise de Desempenho e Análise de Evasão. Para esta tese foram prototipados os módulos de Visão geral dos dados e Análise de Desempenho (previsão de desempenho), como é mostrado na Figura 33.

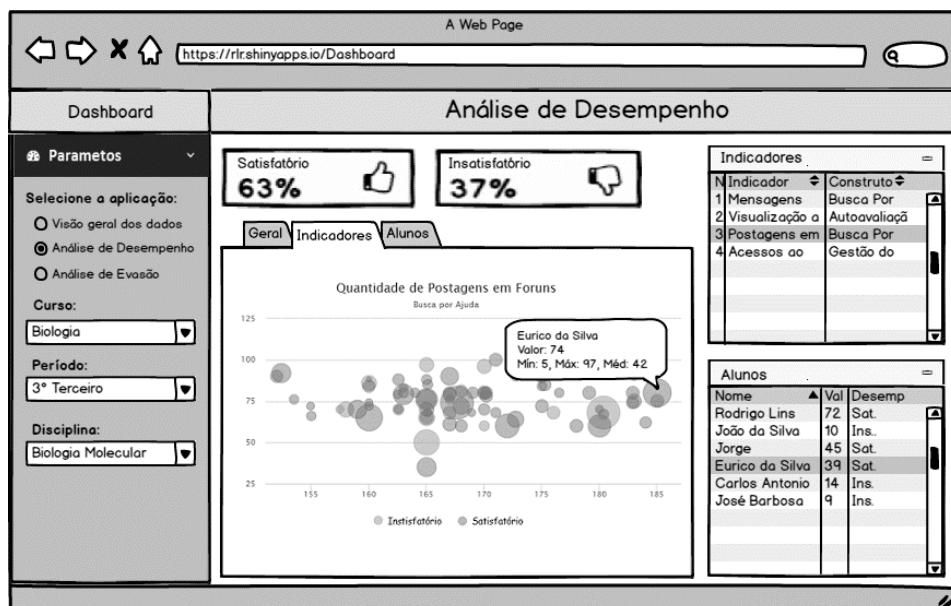
Figura 33: Módulo Análise de Desempenho dos dados da aba Geral



Fonte: Elaborada pelo Autor (2016)

A Figura 34 mostra o protótipo relacionado a aba Indicadores. Esta aba apresenta o gráfico do tipo bolhas. Nesta aba o usuário visualiza apenas um único indicador por vez e tem a possibilidade de visualizar a frequência do comportamento de cada um dos alunos, onde cada aluno é representado por uma bolha.

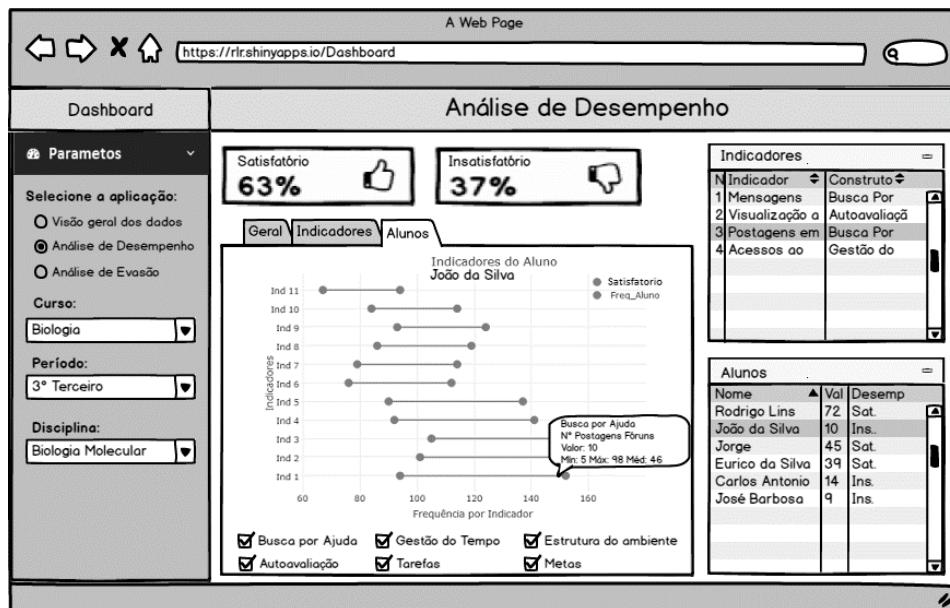
Figura 34: Módulo Análise de Desempenho dos dados da aba Indicadores



Fonte: Elaborada pelo Autor (2016)

A Figura 35 apresenta o protótipo do gráfico da aba Alunos. Este gráfico apresenta o distanciamento da frequência do aluno em relação à média do grupo satisfatório, ou seja, mostra o quanto a frequência do aluno, em cada um dos indicadores de autorregulação da aprendizagem se distancia do ideal.

Figura 35: Módulo Análise de Desempenho dos dados da aba Alunos



Fonte: Elaborada pelo Autor (2016)

Com a técnica de prototipagem de baixa fidelidade foi possível materializar a concepção das funcionalidades iniciais. Os protótipos de baixa fidelidade foram de extrema importância, pois permitiram antecipar ao usuário final características da interface que podem ser testadas, validadas e modificadas pelos usuários. Neste processo, criamos uma interface semelhante à interface final, partindo de requisitos iniciais e, por meio de um processo iterativo, essas interfaces foram sendo refinadas (*Apêndice L*).

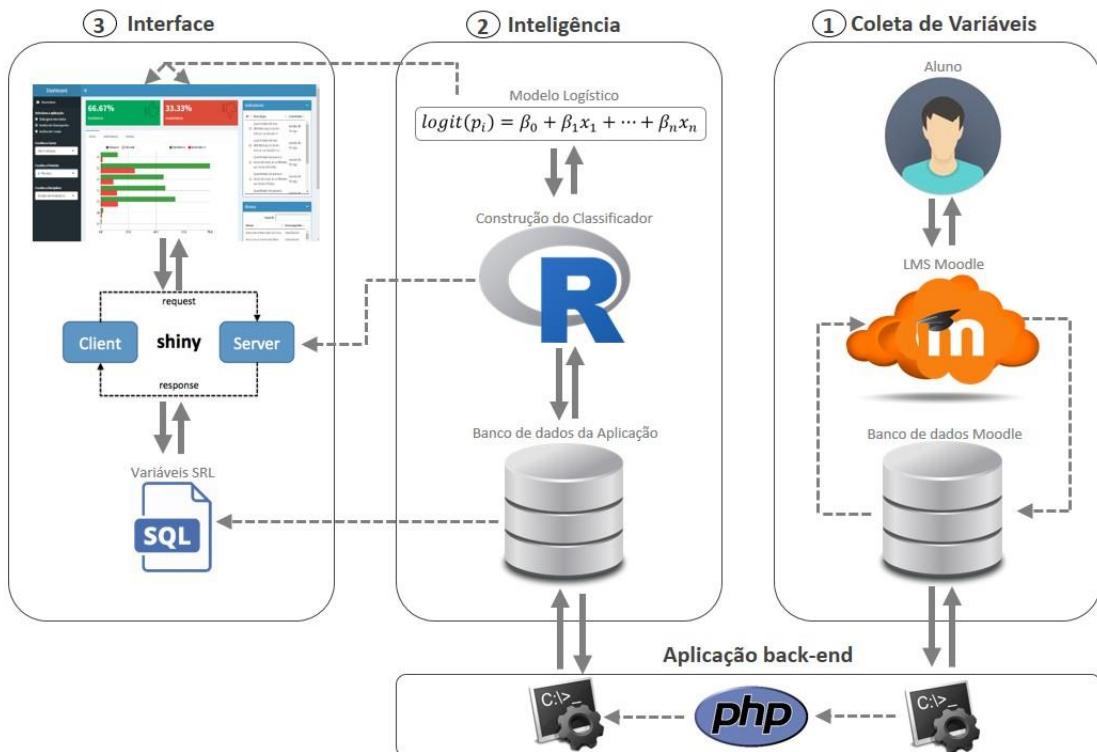
6.5.2 Arquitetura e Implementação

A partir da especificação feita e construção de protótipos de baixa fidelidade, que demonstraram as alternativas de *design*, realizamos o desenvolvimento e implementação do protótipo em uma versão interativa, versão esta que pôde ser testada pelos usuários finais.

- **Arquitetura**

Inicialmente desenvolvemos uma arquitetura de software que especifica a conexão entre os três módulos do sistema. A Figura 36 apresenta a arquitetura dividida em três módulos: (1) Modulo de coleta de variáveis; (2) Modulo de Inteligência; e (3) o Modulo de Interface.

Figura 36: Arquitetura da solução de software desenvolvida



Fonte: Elaborada pelo Autor (2016)

O **módulo de coleta de variáveis** é responsável por coletar as variáveis de autorregulação da aprendizagem em uma instância do Moodle, hospedada em um servidor externo. O processo de coleta é realizado em *batch* e executado uma vez por dia coletando os registros comportamentais dos alunos e enviando para o módulo de inteligência.

O **módulo de inteligência** tem a função de criar e executar o modelo preditivo. Este modelo foi desenvolvido levando em consideração as variáveis de autorregulação e o desempenho dos alunos em disciplinas passadas. Na Seção 6.4.3 foi ilustrado o processo de construção do modelo de classificação que pode ser incorporado à arquitetura do módulo de inteligência. Esse modelo é sempre atualizado à medida que uma disciplina finaliza e é possível ter acesso aos registros comportamentais dos alunos e seus respectivos desempenhos.

Por fim o **módulo de interface** é responsável por consumir os resultados do módulo de inteligência e apresentar para o usuário a interface em formato de *dashboard*. A seguir apresentamos os resultados de interface relacionados a este módulo, onde será descrito cada um dos componentes e suas respectivas funcionalidades.

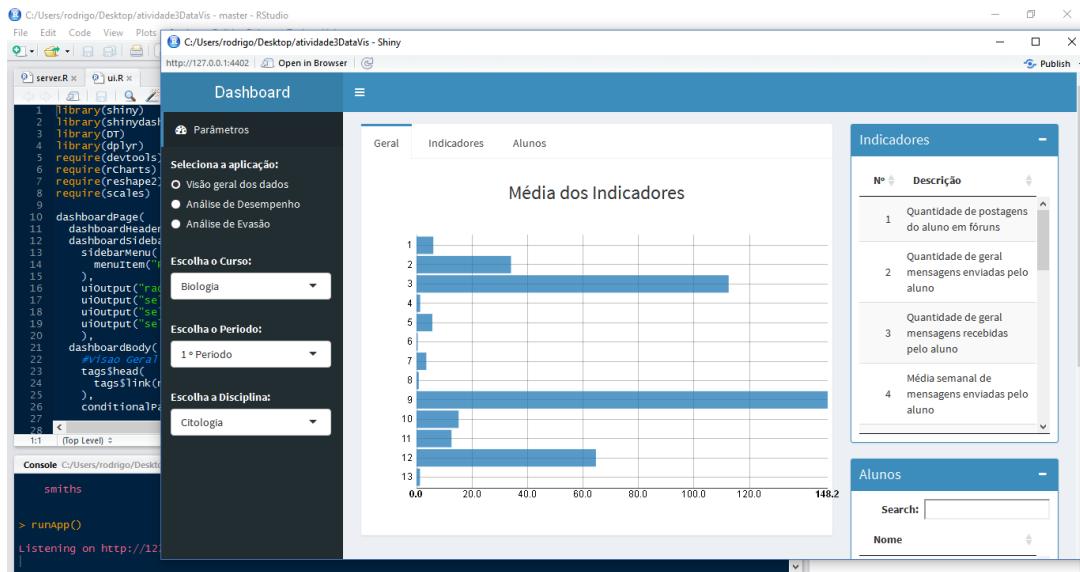
- **Implementação**

Nesta etapa foi possível implementar as funcionalidades propostas na fase de prototipagem, bem como a arquitetura descrita na fase anterior. A aplicação foi desenvolvida utilizando o framework Shiny e teve seu desenvolvimento distribuído e com código aberto disponibilizado em (<https://github.com/grupoccte>).

O motivo da escolha do framework Shiny se deu pela possibilidade de modelagem preditiva e implementação utilizando uma mesma tecnologia, a plataforma e linguagem R. Desta forma foi possível desenvolver todo o processo de descoberta do conhecimento, por meio da mineração, e desenvolvimento do *dashboard* utilizando um mesmo arcabouço tecnológico.

Na Figura 37 é apresentada a tela do primeiro módulo da aplicação. Neste módulo é possível visualizar apenas gráficos descritivos, sem a inserção dos modelos preditivos.

Figura 37: Interface do módulo de Visão geral dos dados

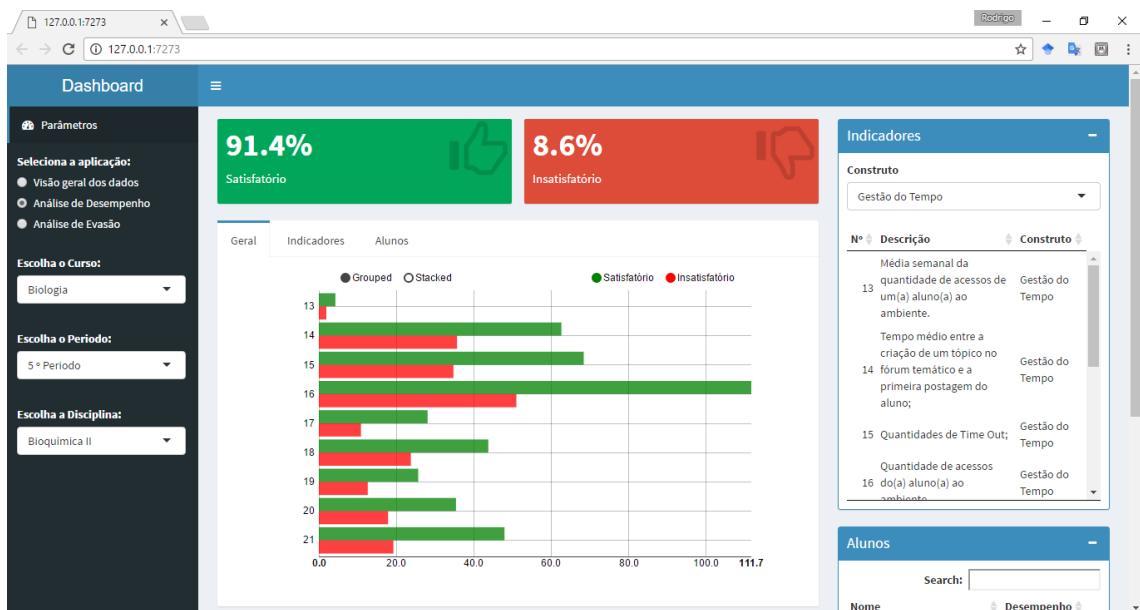


Fonte: Elaborada pelo Autor (2016)

Do lado esquerdo é a área de configuração de parâmetros. É possível selecionar o curso, período e a disciplina da qual desejamos informações.

O segundo módulo é referente a Análise de Desempenho (predição). Neste módulo é possível o usuário visualizar a previsão do percentual de alunos em situação de desempenho satisfatório e o percentual de alunos em situação insatisfatória, para cada turma e período selecionado.

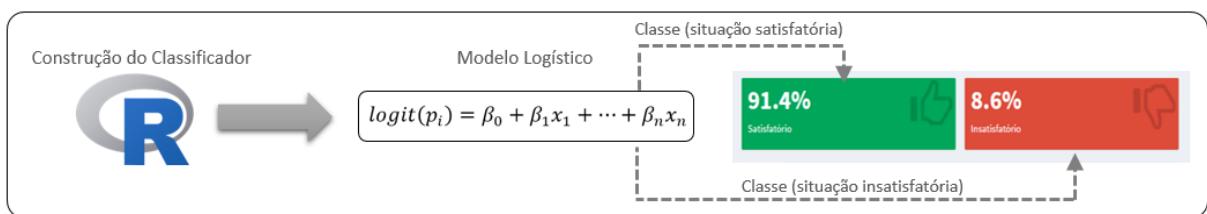
Figura 38: Módulo de Avaliação de Desempenho Gráfico Geral



Fonte: Elaborada pelo Autor (2016)

Este percentual preditivo é gerado a partir do modelo desenvolvido na fase de modelagem desta tese (Seção 6.4). A Figura 39 mostra o fluxo da visualização dos percentuais que são gerados por meio do arquivo *modelo.rda* construído no processo de treinamento do modelo de regressão logística.

Figura 39: Fluxo da geração dos percentuais de classificação



Fonte: Elaborada pelo Autor (2016)

Estes percentuais são gerados a partir da implementação do algoritmo de classificação que recebe as variáveis comportamentais de autorregulação da aprendizagem e faz uma previsão do aluno em um das duas categorias relacionadas ao desempenho.

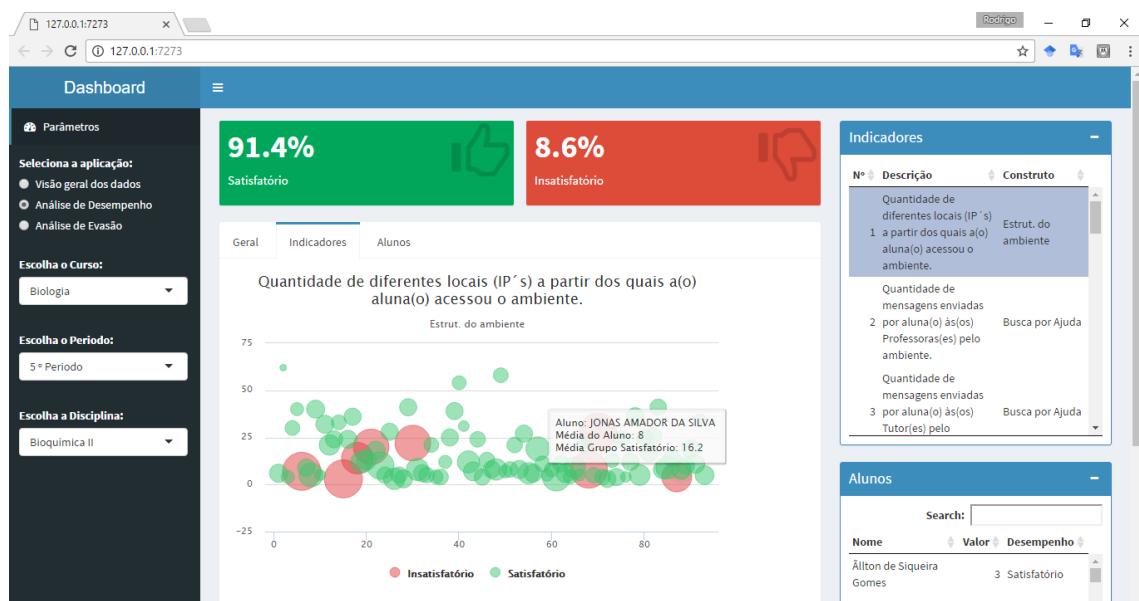
Em seguida, na região central da interface (Figura 38), é exibido o gráfico com as frequências de todas as variáveis. Este gráfico mostra um comparativo entre a frequência de ambos os grupos (satisfatório e insatisfatório). O usuário poderá selecionar na tabela de indicadores qual indicador (variável comportamental) deseja visualizar no gráfico.

Podem ser selecionados mais de um indicador. É possível o usuário selecionar os indicadores por grupos, cada grupo possui um conjunto de variáveis de autorregulação. Na medida em que o usuário for selecionando os indicadores, sua barra é mostrada no gráfico e ao passar o mouse sobre a barra são mostradas informações de nome do indicador, construto ao qual pertence, valor mínimo, valor máximo e média.

A Figura 40 apresenta a interface relacionada à aba de Indicadores. Esta aba é responsável por apresentar a frequência de todos os alunos da turma, selecionada de acordo com informações sobre um indicador específico.

A informação é mostrada por meio do gráfico de bolhas. As bolhas são representadas em duas cores, a cor vermelha, representa os alunos que estão em situação Insatisfatória, e as bolhas verdes, representam os alunos em situação Satisfatória.

Figura 40: Módulo de Avaliação de Desempenho gráfico Indicadores

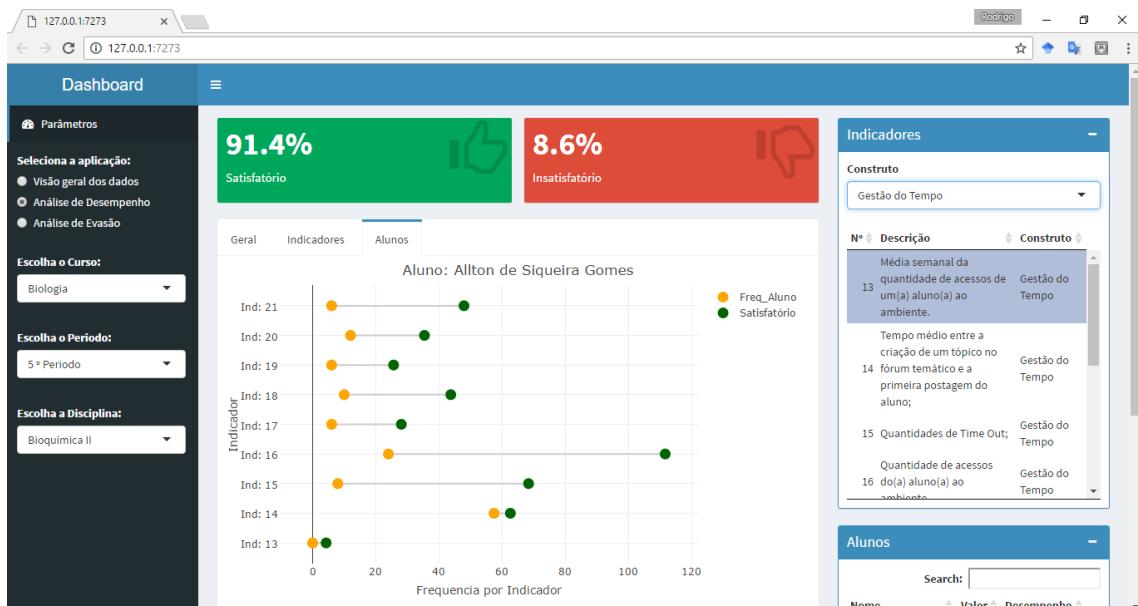


Fonte: Elaborada pelo Autor (2016)

Ao passar o mouse sobre a circunferência é mostrado um hit com informações sobre a descrição da variável (indicador), valor do aluno, valor mínimo, valor máximo, e média da variável na disciplina selecionada.

A Figura 41 apresenta a interface para a aba Alunos. Nesta aba o professor poderá selecionar, na tabela Alunos, qual aluno deseja visualizar no gráfico. Para isto, a tabela troca informações com o gráfico.

Figura 41: Módulo de Avaliação de Desempenho gráfico Alunos



Fonte: Elaborada pelo Autor (2016)

Ao selecionar um aluno na tabela, é plotado o gráfico, com o título referente ao nome do aluno. Neste gráfico é possível ter uma visão do distanciamento do aluno selecionado em relação à média do grupo satisfatório.

Com a finalidade de avaliar a expectativa e a experiência do usuário na utilização desta solução desenvolvida, e identificar problemas de usabilidade que implicam na sua aceitação por parte dos professores/tutores. Na próxima subseção apresentamos os resultados de dois tipos de avaliações realizadas: a avaliação da experiência do usuário e a avaliação de usabilidade.

6.5.3 Experimento de avaliação com professores

Para a realização deste experimento selecionamos um grupo de 20 professores/tutores, cujo critério de escolha foi o conhecimento da plataforma de LMS Moodle e experiência, de mais de um ano, como tutor ou professor em cursos de educação a distância.

Inicialmente apresentamos o protótipo de baixa fidelidade e suas possíveis funcionalidades. Posteriormente, os usuários responderam a uma escala, relacionada com suas expectativas em relação a apresentação e visualização dos protótipos. Em seguida aplicamos seis tarefas que os usuários realizaram no protótipo funcional. Por fim aplicamos um segundo questionário que buscou mensurar características da experiência do usuário na realização das tarefas na ferramenta.

A seguir apresentamos os resultados das avaliações de usabilidade e a avaliação comparativa entre a expectativa e a experiência do usuário.

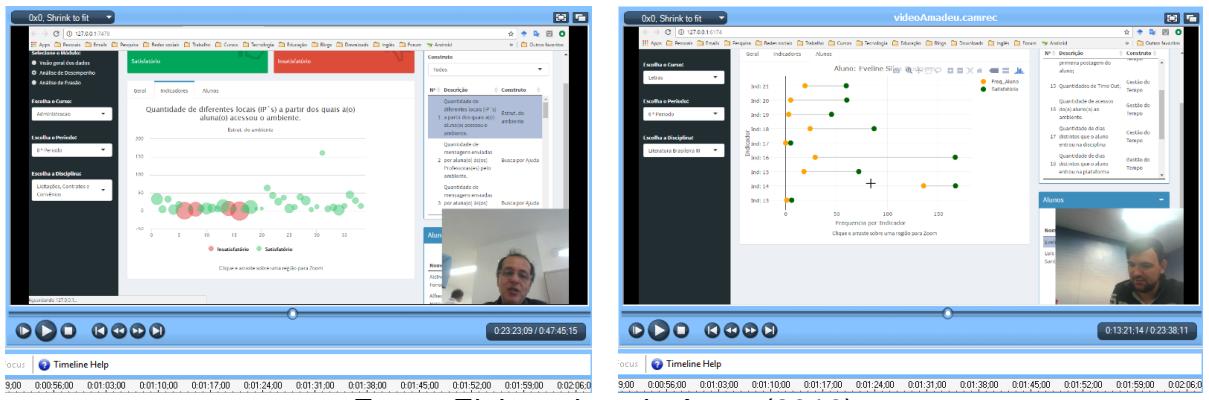
Resultados da Avaliação de Usabilidade

A maioria das dificuldades em projetar interfaces com usabilidade, relaciona-se à interpretação dos usuários frente às informações apresentadas para a realização da tarefa em um sistema (ROGERS, SHARP e PREECE, 2013).

A usabilidade define-se como capacidade e facilidade de uso de um produto, portanto, corresponde a uma característica inerente ao *design* (ROYO, 2008). Entenda-se que não se pode planejar o processo de *design* sem considerar profundamente todas as questões relacionadas, direta ou indiretamente, à usabilidade.

Para a realização do teste de usabilidade foram definidas seis tarefas realizadas pelos usuários. As tarefas foram gravadas e analisadas. As seguintes variáveis foram coletadas: quantidade de erros, quantidade de pedidos de ajuda, tempo de execução das tarefas e quantidade de cliques. Além dessas variáveis, foram coletadas opiniões dos usuários em relação às funcionalidades da ferramenta e possíveis sugestões de melhoria. Na Figura 42 são apresentados dois casos de testes gravados durante a interação dos usuários.

Figura 42: Gravação da interação dos usuários na utilização da aplicação



Fonte: Elaborada pelo Autor (2016)

Ao final da realização das tarefas aplicamos o segundo questionário, o qual teve como objetivo avaliar a experiência do usuário após o uso da aplicação.

As figuras 43, 44 e 45 mostram a realização do teste com professores e tutores. Ao total aplicamos o teste com 20 professores/tutores, com uma média de 41,88 ($SD=8,84$) anos e uma média de 3,97 ($SD=2,62$) anos de experiência com EAD. A Figura 43 mostra o teste realizado com os professores da Universidade de Pernambuco, UPE.

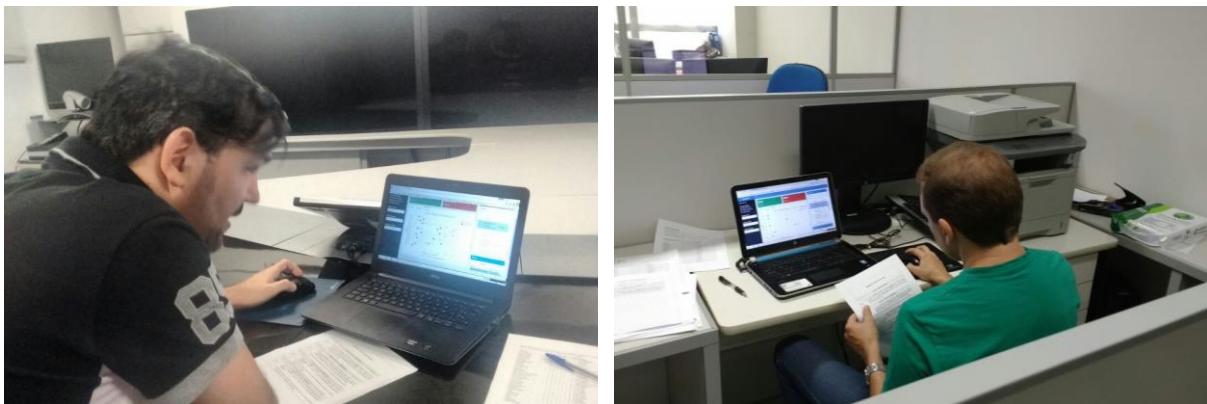
Figura 43: Avaliação com professores da UPE



Fonte: Elaborada pelo Autor (2016)

Os participantes do teste foram professores e tutores das diversas instituições públicas de Pernambuco. Ao todo participaram professores da UPE, UFPE, UNIVASF e UFRPE. A Figura 44 apresenta o teste realizado com professores da UFPE e UNIVASF.

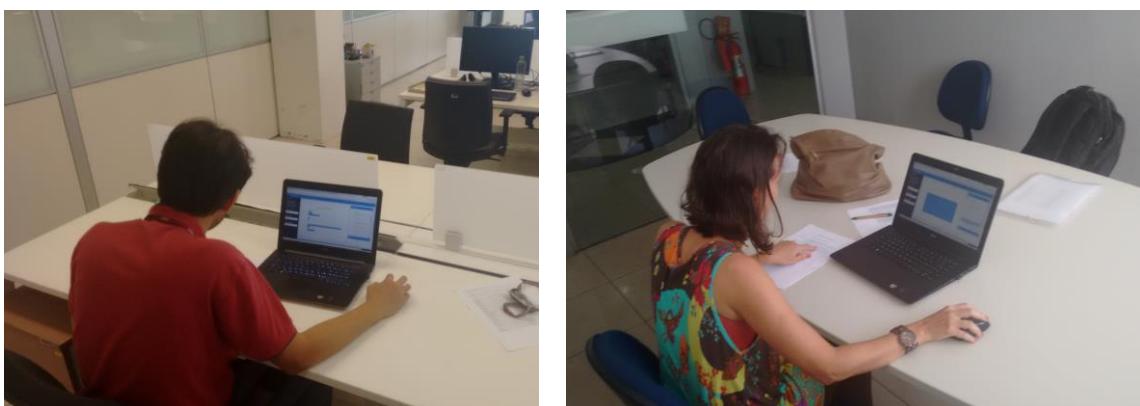
Figura 44: Avaliação com professores da UFPE e UNIVASF



Fonte: Elaborada pelo Autor (2016)

Durante a realização do teste de usabilidade, realizamos gravações de todas as interações dos usuários com a utilização da ferramenta. Cada teste teve uma duração média de 40 minutos. A Figura 45 mostra a avaliação realizada com tutores da UFRPE.

Figura 45: Avaliação com tutores da UFRPE



Fonte: Elaborada pelo Autor (2016)

Durante a realização dos testes identificamos uma média de 0,89 ($SD=0,79$) erros por atividades, com um tempo médio de 87,57 ($SD=44,18$) segundos para a realização de cada tarefa; uma média de 0,84 ($SD=0,71$) pedidos de ajuda e uma média de 10,63 ($SD=0,58$) cliques.

Ao final de cada uma das seis atividades foi possível identificar, de acordo com a realização das tarefas, problemas de usabilidade na interface. O Quadro 16 mostra a identificação dos problemas de usabilidade de acordo com o teste realizado com os professores e tutores.

Quadro 16: Identificação de problemas de usabilidade.

Atividade Realizada	Descrição do problema
Atividade 1.1: Descobrir no gráfico qual a média da turma em relação ao indicador "Quantidade de mensagens enviadas por aluno aos professores".	CO01: Durante a atividade o usuário confundiu a tabela de indicadores" com a "aba indicadores. CO02: O usuário confundiu o eixo Y com o eixo X. Inicialmente ficou confuso qual informação estava mostrando o eixo x.
Atividade 1.2: Identificar e falar o nome do aluno de maior valor nesse indicador no gráfico	CO03: Em (aba > Indicadores) o título do gráfico é média semanal, mas a informação no hit é "valor do aluno" o usuário se confundiu na divergência das informações. O usuário acha desnecessário a informação no eixo x.
Atividade 2.1: Identificar o indicador "Quantidade geral de postagens de um(a) aluno(a) em fóruns" e verifique a média para o grupo insatisfatório.	CO04: O usuário confundiu a seleção de construtos com a coluna que mostra o construto por variáveis. CO05: Quando o usuário seleciona um construto, mas já tem selecionado uma variável antes, a função de selecionar o construto não tem efeito. CO06: No infobox (que mostra o percentual da turma) deveria ter alguma informação deixando claro que esses percentuais são referente a turma, pois confunde dando a entender que esse percentual é do aluno.
Atividade 2.2: Verificar o indicador "Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente" e informe a quantidade de mensagens enviadas do grupo Satisfatório e do grupo Não Satisfatório.	CO07: O usuário apresentou dúvidas com relação a informação mostrada no hit. Ficou na dúvida se a informação apresentada era mesmo relacionada com a média. CO08: O usuário teve problemas para selecionar a disciplina, pois não observou que precisaria utilizar a barra de rolagem da página, utilizou apenas a barra de rolagem do selectbox. CO09: A aplicação precisa resetar a seleção de indicadores sempre que o usuário selecionar um novo construto
Atividade 2.3: identifique os alunos que estão em pior situação de desempenho em relação ao indicador "Quantidade de leituras feitas ao fórum (pageviews).	CO10: Está tendo uma inconsistência entre a seleção do construto e a seleção do indicador. Se o usuário selecionar primeiro um indicador e depois selecionar o construto o gráfico desaparece. CO11: Os nomes das disciplinas, no box de escolha, está muito grande dificultando a visualização. CO12: Na tabela de indicadores deveria ter um separador entre cada indicador.
Atividade 2.4: Em seguida, procure a aluna Eveline <sobrenome omitido> e identifique os dois indicadores que a aluna se encontra em pior situação entre a frequência da aluna e a média da turma.	CO13: No gráfico da aba Aluno o usuário ficou confuso pois não aparecem todos os itens, ficando interpolados alguns itens. Se o usuário selecionar "todos" construtos, então deverá mostrar todos os indicadores. CO14: É necessário fazer o tratamento do clique. Está dando falha quando o usuário dá um duplo clique. CO15: A barra de ferramenta que está aparecendo para o gráfico da aba Alunos está confundindo o usuário, deveria ser retirada.

Fonte: Elaborada pelo Autor (2016)

Ao todo foram coletados quinze problemas de usabilidade da interface. Embora tenhamos identificado pequenos problemas de usabilidade, podemos afirmar que a

solução desenvolvida apresentou bons resultados. As dificuldades encontradas foram semelhantes entre os usuários, incidindo invariavelmente nos mesmos problemas de interface.

Podemos considerar que a proposta de interface desenvolvida apresentou resultados promissores, considerando a reação positiva dos usuários, sobretudo a sua satisfação em relação à facilidade de utilização da ferramenta desenvolvida.

Naturalmente, também foram detectadas algumas dificuldades relativamente simples de corrigir, tais como: mudança de título de componentes de interface, legenda de gráficos, posição dos títulos e tratamento de cliques. A seguir são apresentados os resultados da aplicação dos questionários de expectativa e experiência do usuário.

Resultados da Avaliação da Experiência do Usuário

Considerando o usuário como figura central no processo de *design* de interface, os aspectos formais, funcionais e estéticos devem favorecer à experiência do usuário.

O dever do *design* da interface é tornar essa experiência bem-sucedida para um público alvo. De acordo com Royo (2008), experiência do usuário corresponde ao “conjunto de sensações, valores e conclusões obtidas a partir da utilização de um produto ou equipamento”, e que resulta dos objetivos do próprio usuário, do seu repertório cultural e da maneira como elabora seu modelo mental da tarefa.

Para a avaliação da experiência do usuário foi realizada a metodologia proposta por Hassenzahl (2004) por meio do instrumento *AttrakDiff*¹⁵. Este instrumento permite avaliar a atratividade de uma aplicação por meio de diferentes aspectos.

O instrumento é dividido em três dimensões: (1) Qualidade Pragmática (PQ); (2) Qualidade Hedônica (Estímulo (HQ-S) e Identidade (HQ-I)); e (3) Atratividade (ATT). A PQ refere-se à usabilidade da interface, ou seja, avalia se os usuários conseguiram atingir os objetivos das tarefas durante a utilização do produto; a HQ-S avalia o quanto o produto possui características inovadoras, e se apresenta interações estimulantes; a HQ-I avalia o quanto a interface permite que o usuário se identifique; e, Por fim, a ATT descreve o valor global do produto baseado na qualidade de percepção do usuário (HASSENZAHL, 2004).

¹⁵ <http://www.attrakdiff.de/>

A Figura 46 apresenta os resultados do comparativo entre a expectativa e a experiência. Podemos observar que com relação à expectativa dos usuários (retângulo laranja), o protótipo apresentado foi considerado como desejado. E com relação à experiência dos usuários (retângulo azul), relacionado ao protótipo funcional, também foi considerado como desejado.

Figura 46: Valores médios das dimensões e retângulos de confiança



Fonte: Elaborada pelo Autor (2016)

É importante observar que os valores de confiança em relação à expectativa e à experiência do usuário teve maiores valores de confiança para a Experiência no que diz respeito a Qualidade Pragmática ($PQ = 0,41$). Para a Qualidade Hedônica a Expectativa teve valor de confiança superior ($HQ = 0,26$).

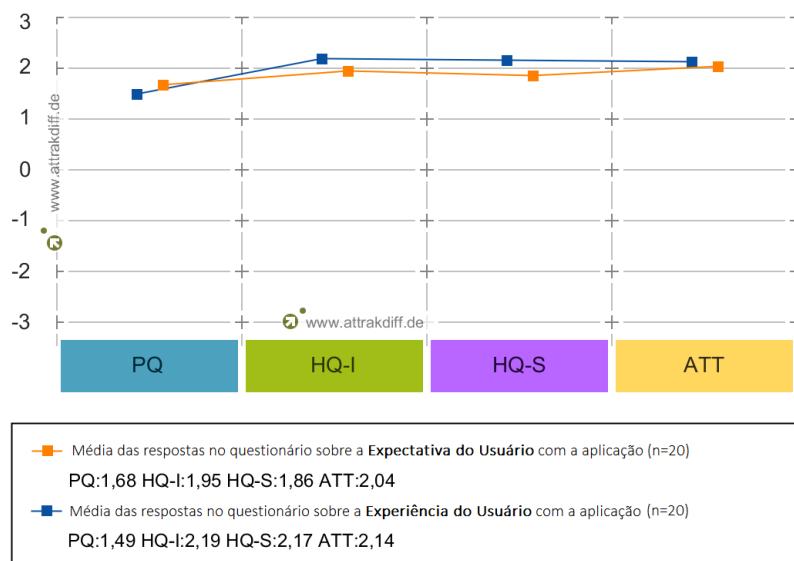
Os resultados exibidos nesse gráfico apontam que o produto foi classificado como "Desejado" nos dois momentos. Suas qualidades pragmáticas e hedônicas são claramente destacadas nesse quadrante, indicando que o produto auxilia, desperta o interesse e estimula os usuários.

O gráfico da Figura 47 apresenta os valores médios por dimensões. Apenas para a dimensão Qualidade Pragmática (PQ) obtevou-se valor de expectativa maior

do que experiência do usuário. As demais dimensões Qualidade Hedônica de Estímulo (HQ-S), Identidade (HQ-I) e Atratividade (ATT) apresentaram médias superiores para experiência do usuário.

Um ponto a destacarmos nesta análise é que todos os pontos (médias) situaram-se na região acima da média (acima de zero). Isto significa que tanto em relação à expectativa, quanto em relação à experiência do usuário, a interface proposta apresentaram bons resultados no que diz respeito as quatro categorias de análise, de acordo com o metodologia proposta por Hassenzahl (2004).

Figura 47: Valores médios das dimensões

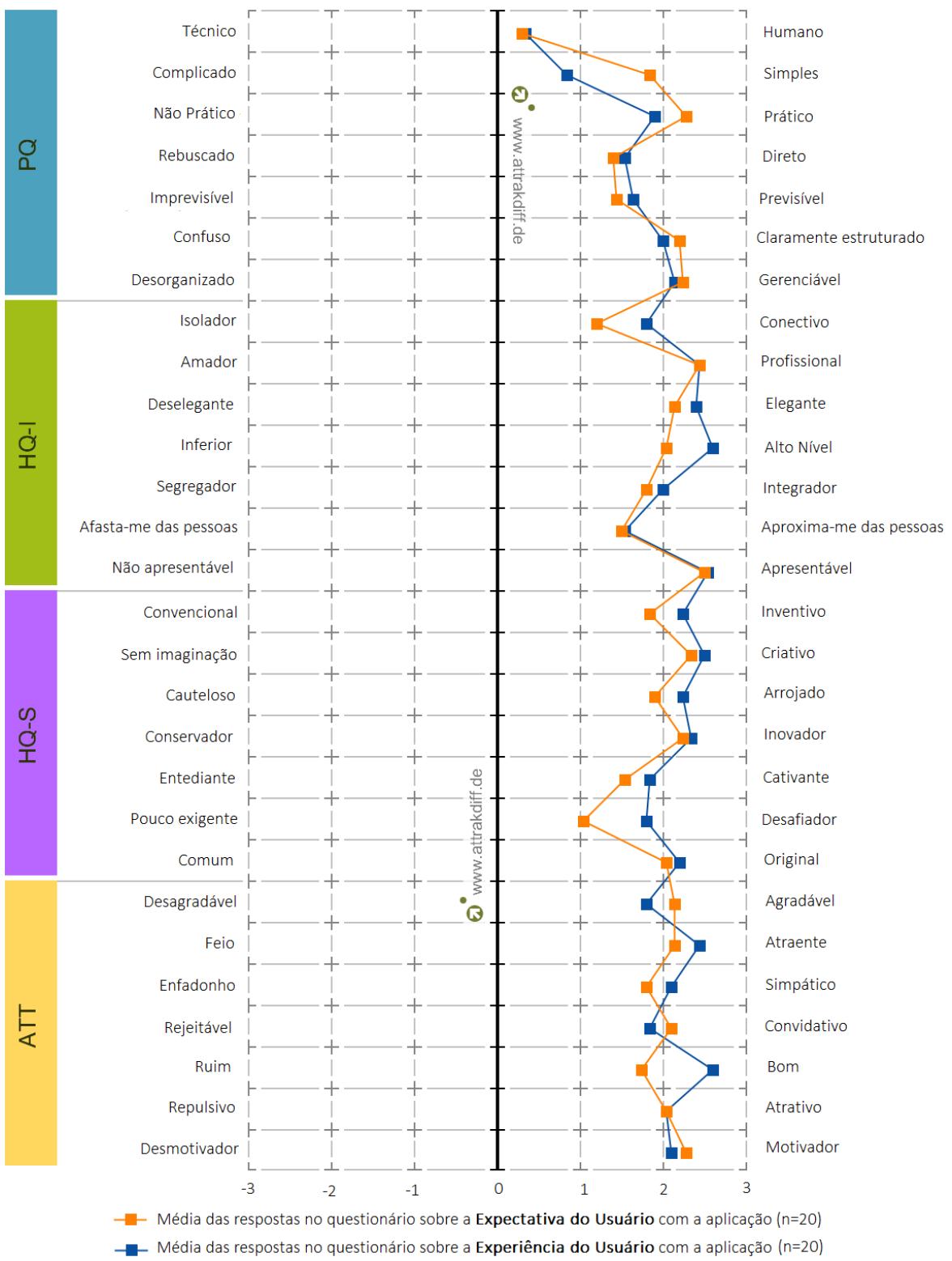


Fonte: Elaborada pelo Autor (2016)

Para uma visualização mais detalhada sobre cada um dos itens das quatro categorias de análise, calculamos os valores médios dos pares de palavras. A Figura 48 apresenta o gráfico comparativo das médias dos pares de palavras disponibilizados no questionário para os usuários.

Este gráfico mostra cada um dos 28 itens do questionário. Em relação à interpretação gráfica, podemos afirmar que para a categoria Qualidade Pragmática (PQ), dos sete itens do questionário, quatro deles tiveram médias maiores em relação à experiência e três tiveram médias maiores em relação à expectativa.

Figura 48: Comparativo de médias entre Expectativa e Experiência



Fonte: Elaborada pelo Autor (2016)

Em relação à Qualidade Hendônica de Identidade (HQ-I), seis itens tiveram média maior em relação à expectativa. A Qualidade Hendônica de Estímulo (HQ-S) teve todos os itens com média maior para o questionário de expectativa do usuário. A

atratividade teve três itens com média maior para experiência, três itens com média maior para expectativa e um item com valor de média igual para ambos. É importante destacarmos que todos os pontos (médias), dos pares de palavras, situaram-se na região acima da média (acima de zero).

Embora graficamente possamos perceber as diferenças entre expectativas e experiência do usuários, apenas resultados gráficos não são suficientes para concluirmos sobre a existência de diferença significativa entre expectativa e experiência do usuário. Para tal foi necessário aplicar testes estatísticos com o objetivo de verificar diferenças significativas do ponto de vista estatístico.

Testes de comparação de médias entre *Expectativa* e *Experiência*

Nesta etapa analisamos a diferença significativa entre *Expectativa* e *Experiência do Usuário*, por meio dos dados coletados com o instrumento AttrakDiff. Aplicamos testes estatísticos para avaliar a existência, ou não, de diferenças significativas entre as respostas nos dois momentos.

Para esta análise utilizamos o teste de comparação de médias para dados pareados, com a aplicação do questionário no primeiro e no segundo momento, onde cada observação da primeira amostra (Questionário I) foi pareada com cada observação da segunda amostra (Questionário II). Tal delineamento ocorreu, através de medidas feitas antes (expectativa) e depois do indivíduo ter utilizado a interface (experiência). Este teste de comparação de médias foi realizado baseado na hipótese nula H_0 de não ter ocorrido diferença significativa entre a expectativa e a experiência do usuário, contra a hipótese alternativa H_1 de ter ocorrido diferença significativa. A formação da hipótese é mostrada abaixo:

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

O teste de hipótese foi aplicado para 28 variáveis, 7 relacionadas à Qualidade Pragmática (*PQ*), 7 à Qualidade Hendônica de Estímulo (*HQ-S*), 7 relacionadas a Qualidade Hendônica de Identidade (*HQ-I*) e 7 relacionadas à Atratividade (*ATT*).

Para a realização das análises optamos por utilizar o Teste de *Wilcoxon*, com correção de continuidade, a um nível de significância de 5%. Este teste substitui o teste *t* de *Student* para amostras emparelhadas, quando os dados são medidos em

escala ordinal baseando-se nos postos (*ranks*) das diferenças entre pares (CAMPOS, 1987).

Foi necessário fazer uma transformação na escala das variáveis. Foi realizado uma categorização de 0 à 7, permitindo assim a aplicação do teste de *Wilcoxon*. No Quadro 17 temos os resultados dos testes de *Wilcoxon* (*p*-valor) para cada uma das amostras pareadas.

Quadro 17: Médias e valores dos testes de *Wilcoxon*.

Médias dos itens e P-valor					
	Indicador	Média Questionário I	P-valor	Média Questionário II	Indicador
Qualidade Pragmática (PQ)	Técnico	4,35	0,7188	4,30	Humanizado
	Complicado	5,80	0,3239	5,20	Simples
	Impraticável	5,80	0,6243	6,15	Prático
	Rebuscado	6,25	0,1623	5,85	Direto
	Imprevisível	4,85	0,0567	5,85	Previsível
	Confuso	6,45	0,7959	6,45	Claramente estruturado
	Desorganizado	6,45	0,3132	6,15	Gerenciável
HQ-I. (Identidade)	Isolador	5,90	0,2429	6,30	Conectivo
	Amador	6,10	0,4512	5,80	Profissional
	Deselegante	5,55	0,7555	5,40	Elegante
	Inferior	6,40	0,3425	6,15	Alto Nível
	Segregador	5,65	0,8990	5,45	Integrador
	Afasta-me das pessoas	6,60	0,0221*	6,05	Aproxima-me das pessoas
	Não apresentável	6,00	0,7309	5,80	Apresentável
HQ-S. (Estímulo)	Convencional	5,5	0,9662	5,50	Inventivo
	Sem imaginação	6,55	0,6618	6,50	Criativo
	Cauteloso	5,85	0,4083	6,10	Arrojado
	Conservador	6,50	0,5150	6,35	Inovador
	Entediante	6,60	0,4083	5,75	Cativante
	Pouco exigente	6,00	0,5150	6,20	Desafiador
	Ordinário	6,05	0,1853	6,05	Original
ATT. (Atratividade)	Desagradável	6,25	0,1568	6,90	Agradável
	Feio	6,35	0.3789	6,25	Atraente
	Enfadonho	5,85	0,4033	5,55	Simpático
	Rejeitável	5,80	0.1397	5,05	Convidativo
	Mau	6,10	0,8030	6,30	Bom
	Repulsivo	6,20	0,1535	6,05	Atrativo
	Desmotivador	6,15	0,6725	6,25	Motivador

Fonte: Elaborada pelo Autor (2016)

De acordo com Campos (1987), testes com *p*-valor acima de 0,05 significam que a hipótese H_0 tende a ser aceita, ou seja, pode-se afirmar que não existiu diferença significativa entre a expectativa e a experiência dos usuários.

No Quadro 16 é possível observarmos que apenas um dos indicadores teve diferença significativa entre expectativas e experiência (Afasta-me das pessoas vs Aproxima-me das pessoas), ou seja, embora visualmente, por meio de análise gráfica, tenhamos encontrado diferenças entre expectativas e habilidades, por meio da análise inferencial foi possível afirmar que não existiu diferença significativa entre o que os usuários esperavam da ferramenta, de acordo com a apresentação inicial dos protótipos e a real utilização da mesma por meio das tarefas realizadas.

Embora não tenhamos encontrado diferença significativa entre expectativas e habilidades, podemos afirmar que, em ambos os casos, todos os itens foram avaliados pelos professores/tutores dentro da faixa de valores positivos (de 0 a 3). Isto nos indica que o potencial da ferramenta, do ponto de vista dos usuários, foi avaliado com bons resultados para ambos os casos, e que a expectativa que o professor/tutor teve inicialmente, se manteve após a utilização e realização das tarefas de usabilidade.

6.7 Considerações sobre o capítulo

Este capítulo apresentou os resultados alcançados nesta tese, de acordo com os objetivos propostos no Capítulo 1, onde o objetivo geral foi a construção de uma modelagem preditiva do desempenho de estudantes, por meio de indicadores comportamentais de autorregulação da aprendizagem em ambientes de LMS. Estes resultados permitirão que professores e tutores possam identificar o comportamento autorregulatório dos alunos em plataformas de LMS como indicadores de desempenho.

Para atingir estes resultados, utilizamos a metodologia CRISP-EDM. O processo foi composto por seis fases: Entendimento do domínio em educação, Entendimento dos dados educacionais, Preparação dos dados para EDM, Modelagem, Avaliação do modelo, Implementação e avaliação da solução educacional. Como resultados destas fases, podemos destacar:

- Validação de um instrumento de coleta (*questionário OSLQ*) em contexto da EAD no Brasil;
- Mapeamento e avaliação de um conjunto de variáveis comportamentais, por meio da técnica de análise fatorial confirmatória AFC, representativas do fenômeno de autorregulação da aprendizagem armazenadas em sistemas de LMS;
- Construção de um conjunto de scripts para extração de variáveis de autorregulação em sistemas de banco de dados de plataformas de LMS;
- Construção de um conjunto de scripts para a construção de modelos de classificação em bases de dados educacionais;
- Modelagem de um classificador com boas taxas de desempenho para a previsão do desempenho de alunos de EAD; e
- Desenvolvimento e avaliação de uma interface para que professores e tutores, sem experiência com mineração de dados, possam utilizar dos benefícios dos modelos de classificação para previsibilidade do desempenho de alunos e visualização dos seus respectivos comportamentos autorregulatórios;

No próximo capítulo serão discutidas as implicações dos resultados alcançados neste capítulo. Estes resultados serão discutidos levando em consideração suas relações com trabalhos encontrados na literatura, bem como os impactos que estes resultados desempenharão no contexto da Educação a Distância e as respectivas contribuições para a área de pesquisa.

7. DISCUSSÕES

Este capítulo apresenta as discussões e as implicações sobre os resultados alcançados nesta tese. Além disso, abordamos as descobertas que enriqueceram os resultados, como também suas implicações no âmbito da educação a distância.

7.1 Lacunas de pesquisa identificadas

Um dos primeiros passos para a construção do tema de pesquisa desta tese foi a identificação das lacunas de pesquisa com foco no fenômeno de autorregulação da aprendizagem. Para esta identificação, inicialmente desenvolvemos um mapeamento sistemático da literatura. Neste mapeamento buscamos identificar, catalogar, e classificar um conjunto de artigos (*Apêndice A*) e seus respectivos métodos para mensurar características de autorregulação em ambientes de LMS.

Observamos que a área de pesquisa relacionada ao fenômeno de autorregulação da aprendizagem tem aumentado ao longo dos anos, embora a grande maioria dos trabalhos são realizados baseados na utilização de questionários como instrumentos principal de coleta de dados. Estes trabalhos limitaram-se a utilizar adaptações de questionários desenvolvidos especificamente para o ensino presencial.

Não encontramos nenhum artigo com foco no desenvolvimento de soluções de software que deem suporte para professores ou tutores no que diz respeito à identificação de alunos com dificuldade de autorregulação e consequentemente previsão de desempenho acadêmico. O fornecimento de subsídios preditivos para professores e tutores a partir da análise de dados em tempo real foi uma das principais lacunas de pesquisa abordadas nesta tese. O tema se mostrou desafiador especialmente considerando o número cada vez maior de cursos a distância, número de alunos e volume de dados gerados.

Diante destes achados e lacunas de pesquisa identificadas na revisão da literatura e no mapeamento sistemático, esta tese buscou desenvolver um processo no qual fosse possível mensurar o fenômeno de autorregulação da aprendizagem sem a utilização de questionários e baseado na extração de variáveis comportamentais.

7.2 Capacidade preditiva de desempenho acadêmico

Atualmente, diversas pesquisas vêm sendo desenvolvidas com o objetivo de elevar o sucesso acadêmico dos estudantes do ensino superior, especificamente em relação ao desempenho acadêmico e a retenção dos alunos (MARABOUTI, DIEFES-DUX e MADHAVAN, 2016).

De acordo com Rates (2013) quase um terço dos estudantes abandonam a universidade após o primeiro ano de graduação. O desempenho acadêmico destes estudantes é visto como um dos fatores mais influentes deste abandono (ESSA e AYAD, 2012). Assim, uma forma de diminuir esta retenção é identificar os problemas de desempenho com antecedência, permitindo que professores possam usar uma variedade de estratégias para se comunicar com os estudantes em situações de risco e fornecer-lhes caminhos para melhorar seu desempenho.

O primeiro passo para aumentar o sucesso acadêmico é a identificação de alunos em situações de risco de desempenho. Com a utilização de métodos de modelagem preditiva é possível a identificação estes alunos e seus respectivos comportamentos. A análise preditiva permite que professores, tutores e gestores de EAD possam tomar medidas preventivas, à medida que conseguem identificar as ações dos possíveis alunos em situações críticas.

Para a construção dos modelos preditivos desta tese coletadamos variáveis comportamentais dos alunos armazenadas na plataforma de LMS Moodle. A escolha destas variáveis foi fundamentada no fenômeno de autorregulação da aprendizagem, definido por teóricos como sendo o processo no qual os sujeitos desenvolvem capacidades de autogestão das suas ações comportais para a realização de tarefas (VEIGA SIMÃO, 2006); (ZIMMERMAN, 1989); (SILVA, 2004); (ZEIDNER, BOEKERTS e PINTRCH, 2000).

Estas variáveis foram utilizadas como indicadores comportamentais capazes de predizer o desempenho acadêmico de alunos de EAD. Para a construção dos modelos preditivos utilizamos uma base de treinamento de 21.154 observações, o que correspondeu a 70% do total de dados da base. Esta base de treinamento foi composta por 11.824 alunos com desempenho insatisfatório e 9.330 alunos em situação de desempenho satisfatório. A seguir é discutido um comparativo entre os modelos de classificação.

7.2.1 Comparativo entre modelos de classificação

Na construção dos modelos preditivos utilizamos quatro dos seis algoritmos mais empregados em mineração de dados educacionais para previsão de desempenho, de acordo com Romero & Ventura (2010). Utilizamos os algoritmos de Árvore de Decisão (CART), Regressão Logística, Maquina de Vetor de Suporte (SVM) e Naive Bayes. Buscamos realizar um comparativo entre os modelos de classificação que melhor se ajustassem as varáveis em estudo. Este comparativo foi realizado em função das principais métricas utilizadas na literatura, tais como valores de Acurácia, Sensibilidade, Especificidade, Precisão e área sobre a curva ROC.

Em três das quatro métricas utilizadas, o algoritmo que apresentou melhor resultado, dentre os quatro, foi o algoritmo de Regressão Logística: acurácia (0,893); especificidade (0,9167); precisão (0,8921) e área sob a curva ROC (0,9574). Estes resultados vêm a contribuir com os trabalhos desenvolvidos por Hendel (2007), Braunstein, Lesser e Pescatrice (2008), Eckles e Stradley (2012) e Marbouti, Diefes-Dux e Madhavan (2016) que apresentam a modelagem por meio de regressão logística como sendo um algoritmo de classificação adequado para mineração de dados educacionais.

Os resultados alcançados na fase de modelagem mostraram-se satisfatórios, de acordo com os limiares propostos por Hair, Black, *et al.* (2009). Estes valores de índices de avaliação nos indicam que o modelo de regressão logística desenvolvido nesta tese mostrou-se como sendo o mais adequado dentre os quatro modelos avaliados.

7.2.2 Significância e poder de explicação das variáveis

A escolha de variáveis é o processo de seleção de um subconjunto de dados (isto é, variáveis) a serem usados no processo de treinamento de um modelo de classificação. De acordo com James, Witten, *et al.* (2013) a escolha das variáveis é um dos fatores mais importantes para o sucesso na construção de um classificador.

Os métodos de seleção tentam selecionar as variáveis que têm o maior poder preditivo e estão mais relacionadas com a variável predita, excluindo as que não contribuem para o poder preditivo do classificador.

Inicialmente o modelo de regressão logística desenvolvido nesta tese foi composto por 32 variáveis explicativas (Quadro 12-Seção 6.4.3). Após a aplicação do método de redução de variáveis (*Stepwise*) foi possível manter o valor de acurácia com a redução para 23 variáveis explicativas (Quadro 13 - Seção 6.4.3). Estas 23 variáveis fizeram parte dos seis construtos de autorregulação da aprendizagem (Estruturação do ambiente, Busca por Ajuda, Estratégia de Realização de Tarefa, Auto avaliação, Gestão do Tempo e Estabelecimento de metas).

Das 23 variáveis que foram selecionadas para compor a construção do modelo, algumas tiveram um maior poder de influência na predição do desempenho dos alunos. Dentre elas, a variável 33, que fez parte do construto *Estabelecimento de Metas* foi a que teve maior valor de significância estatística no modelo. Esta variável extraiu a quantidade de atividades entregues no prazo por aluno. Este comportamento se mostrou o mais significativo de todos, ou seja, existe uma relação entre alunos que costumam enviar atividades acadêmicas dentro do prazo estabelecido pelo professor e o aumento do seu desempenho.

Outra variável que mostrou o segundo maior valor de significância estatística foi a variável 34, também pertencente ao construto *Estabelecimento de metas*. Esta variável buscou extrair a quantidade geral de postagens dos alunos nos fóruns. A interação dos alunos via fórum de discussão foi outra variável que teve uma forte influência no poder de discriminação do classificador.

Por fim, outra variável que merece destaque, quanto à significância estatística é a variável 25. Esta variável fez parte do construto *Gestão do tempo*. Esta variável buscou extrair o tempo médio entre a criação de um tópico, no fórum, e a primeira postagem do aluno, ou seja, o tempo que o aluno levou para começar a interagir no fórum da disciplina.

Além destas três variáveis discutidas, mais vinte variáveis fizeram parte do modelo de previsão desenvolvido. Embora tenhamos discutido as três que tiveram o maior valor de significância estatística, não é interessante olharmos para cada uma destas variáveis isoladamente, tendo em vista que o poder de previsão de um algoritmo de classificação depende da composição conjunta de todas as variáveis incluídas no modelo. A seguir discutimos a capacidade de previsão do modelo desenvolvido.

7.2.3 Previsão do desempenho dos alunos

A tabela de matriz de confusão é uma das referências mais utilizadas para identificar o poder preditivo de um classificador. A Tabela 4 apresenta o quantitativo de alunos previstos corretamente e incorretamente por meio dos resultados obtidos para 9063 alunos submetidos ao modelo desenvolvido com o algoritmo de regressão logística.

Tabela 4: Matriz de confusão com percentuais.

Matriz de Confusão		Classe Atual		Total
		Desempenho Satisfatório	Desempenho Insatisfatório	
Classe Prevista	Desempenho Satisfatório	4620 (50,98 %)	550 (6,07 %)	5170 (57,04 %)
	Desempenho Insatisfatório	420 (4,63 %)	3473 (38,32 %)	3893 (42,96 %)
Total		5040 (55,61 %)	4023 (44,39)	9063 (100 %)

Fonte: Elaborada pelo Autor (2016)

Algumas discussões sobre esta tabela que merecem destaque são:

- De todos os alunos presentes na base de dados de teste, 9.063 alunos, o modelo conseguiu classificar 8.093 alunos corretamente em ambas as classes (desempenho satisfatório e insatisfatório). Com um percentual de erro de apenas 10,7%;
- Dos 5.040 alunos que pertenciam à classe de desempenho satisfatório, o modelo conseguiu classificar 4.640 (91,66%) corretamente;
- Dos 4.023 alunos que pertenciam a classe de desempenho insatisfatório, o modelo conseguiu classificar 3.473 (86,32%) corretamente; e
- Para esta tese, um dos casos mais críticos de erro de classificação é o erro do tipo (falso positivo), ou seja, prever um aluno na classe de alunos com desempenho satisfatório, mas na verdade esse aluno pertenceria a classe de alunos com desempenho insatisfatório. Para esta situação tivemos um total de 550 alunos, representando apenas 6,07% dos dados de teste. Este percentual de erro é considerado baixo e consequentemente bastante aceitável.

Estes resultados se mostram bastante aceitáveis, tendo em vista que um dos critérios que adotamos nesta tese foi a busca por um aumento de generalização do

modelo de previsão desenvolvido. Este aumento de generalização teve como meta desenvolver um classificador que pudesse ser utilizado por diversos cursos de EAD e não apenas para um curso específico.

A busca por um modelo com alto poder de generalização, foi justificada pelo objetivo específico desta tese, o qual propôs o desenvolvimento e implementação de um sistema de previsão, como apoio ao professor para tomada de decisão, em relação à previsibilidade do desempenho do aluno, de acordo com seu comportamento auto regulatório. A próxima seção mostra uma discussão sobre os resultados do desenvolvimento desta solução.

7.3 Solução de software desenvolvida

Como forma de materializar os resultados da modelagem descritos na fase anterior, um dos objetivos específicos desta tese foi o desenvolvimento de uma solução de visualização de dados que pudesse consumir o modelo preditivo de desempenho para a geração de visualizações de dados.

Esta solução se mostrou inovadora, pois foi possível integrar componentes de visualização de dados com modelos preditivos, possibilitando ao professor visualizar indicativos de previsão de desempenho acadêmico, por meio do monitoramento de indicadores de autorregulação da aprendizagem.

O motivo da escolha do framework Shiny para o desenvolvimento da solução, foi a possibilidade de modelagem preditiva e implementação utilizando uma mesma tecnologia, a plataforma e linguagem R. Desta forma foi possível realizar a mineração e desenvolvimento da solução de visualização utilizando um mesmo arcabouço tecnológico. Isto contribuiu com a curva de aprendizagem da linguagem.

Um dos requisitos principais, no desenvolvimento da ferramenta, foi a facilidade de uso por parte dos professores e tutores. Buscou-se, após a fase de desenvolvimento, realizar um processo de avaliação da expectativa e experiência do usuário com avaliações de usabilidade para identificar inconsistências na interface.

Essa avaliação identificou um conjunto de quinze problemas de usabilidade que foram catalogados e apresentados como requisitos de implementações para trabalhos futuros desta tese.

Em relação à avaliação da expectativa e experiência do usuário, os resultados mostraram valores excelentes em ambas as avaliações. O produto foi classificado como "Desejado" nos dois momentos pelos usuários. Suas qualidades foram destacadas pelos usuários, indicando que o produto auxilia, desperta o interesse e estimula os usuários na sua utilização.

7.3.1 Percepção dos professores

Os discursos dos professores, expressos durante a fase de avaliação de usabilidade, indicaram que a solução de software desenvolvida nesta tese se mostrou com um bom nível de aceitação e potencial para utilização. O Quadro 18 mostra três relatos de professores durante o teste de usabilidade.

Quadro 18: Relatos dos professores sobre a utilização da ferramenta.

<p><i>"A ferramenta vai permitir individualizar o problema de cada aluno. Isso possibilita um foco maior na situação de dificuldade de cada aluno. Quando for entrar em contato com o aluno, o tutor já vai com o diagnóstico do mesmo, permitindo um atendimento específico. Isso é bem diferente das ferramentas atuais que não permitem esse tipo de diagnóstico."</i></p>
<p><i>"Essa ferramenta vai facilitar a visualização dos dados do Moodle que o professor não consegue entender."</i></p>
<p><i>"Nós não temos essa cultura de tomar ações antes que as coisas aconteçam. Essa ferramenta nos dá essa informação, o professor precisa criar a cultura de, com essa informação, tomar uma ação. Essa ferramenta deve estar também com as pessoas que se preocupem com a gestão do curso."</i></p>

Fonte: Elaborada pelo Autor (2016)

É importante destacar a contribuição deste trabalho no sentido de que ele poderá ser evoluído por outros pesquisadores. Isso se deve ao fato de que os artefatos de software serão registrados e distribuídos sob licença de software livre.

Uma vez que este trabalho desenvolveu e avaliou uma solução de software com propósitos de utilização no contexto da educação a distância, sua evolução e validação em contextos reais torna-se uma necessidade. Além disso, com base neste trabalho, outros pesquisadores poderão se concentrar na aplicabilidade da solução para turmas reais e verificar sua robustez e por consequência a geração de novos

requisitos, por meio de experimentos de maior duração e com outras instituições de ensino.

8. CONSIDERAÇÕES FINAIS

O último capítulo desta tese apresenta as considerações finais, dificuldade e limitações, contribuições, trabalhos futuros e por fim apresenta as principais produções científicas desenvolvidas durante o período de doutorado.

8.1 Conclusão

Esta tese buscou identificar relação de influência dos fatores de autorregulação da aprendizagem no desempenho de alunos pertencentes à modalidade de educação a distância. Tendo como questão de pesquisa a seguinte indidagação “*As diferenças individuais de padrões comportamentais de autorregulação da aprendizagem influenciam o desempenho de alunos na modalidade de EAD?*”

Para tal realizamos uma busca exaustiva de comportamentos de alunos em contextos de cursos a distância que pudessem responder a esta questão de pesquisa, considerando que existem diferenças significativas nas habilidades de cada aluno em estabelecer estratégias de autorregulação da aprendizagem.

O uso de técnicas de mineração de dados educacionais mostrou-se como uma abordagem satisfatória para o entendimento e mensuração do femômeno em estudo. Durante as fases e procedimentos desta tese buscamos desenvolver um modelo que permitisse realizar a previsão do desempenho acadêmico de estudantes, a partir do seu comportamento autorregulatório.

Os achados desta pesquisa contribuiram de forma satisfatória, tanto do ponto de vista da modelagem, quanto da implementação da solução de software desenvolvida, contribuindo para facilitar o acompanhamento e a tomada de decisão por parte de professores na modalidade de ensino a distância.

8.2 Dificuldades e Limitações

Uma das principais dificuldades, durante o desenvolvimento desta tese, foi a extração e pré-processamento das bases de dados utilizadas para na fase de modelagem. A complexidade da arquitetura do banco de dados Moodle, com um conjunto de aproximadamente 400 tabelas, tornou a extração das bases de logs comportamentais uma das fases com maior dedicação de tempo e esforço durante o método CRISP-DM. Ao longo dos sete anos de cursos de EAD, a instituição que

forneceu as bases de dados utilizou diversas versões do sistema Moodle, estas versões tiveram que ser trabalhadas separadamente, e posteriormente foi feita a sua integração e extração dos dados relevantes, tornando esse processo exaustivo e custoso.

Após a fase de extração das bases de dados, outra etapa que demandou bastante esforço foi a identificação de variáveis representativas do fenômeno em estudo desta tese. Neste sentido, o processo empregado nesta tese mostra-se como inovador pois permitiu mapear um conjunto satisfatório de variáveis comportamentais de autorregulação, armazenadas em plataformas de LMS.

Embora a solução de software desenvolvida tenha sido testada com professores e tutores, os quais possibilitaram o levantamento de pontos de melhoria, uma das limitações desta tese foi a inexistência de um experimento temporal com alunos durante um semestre. Desta forma, temos consciência da necessidade, em trabalhos futuros, da condução de um experimento controlado, com um curso superior na modalidade de EAD, em um contexto real de ensino, para observar possíveis problemas de aceitação por parte dos alunos e a robustez da ferramenta durante a sua utilização. Podemos considerar esta como sendo a principal limitação desta tese.

8.3 Contribuições

Durante o desenvolvimento desta tese, diversas contribuições emergiram de acordo com a execução do método proposto. Algumas contribuições que merecem destaque são: (1) Geração de um conjunto de scripts para extração de variáveis em plataformas Moodle; (2) Geração de um conjunto de scripts em R para análise de dados e construção de classificadores; e (3) Construção de uma arquitetura de software de visualização de dados em Shiny.

O entendimento das arquiteturas dos bancos de dados históricos e as respectivas extrações das variáveis da plataforma Moodle puderam ser considerados como uma das fases mais exaustivas do método de desenvolvimento desta tese. A construção e disponibilização dos scripts SQL (*Apêndice D*) pode ser considerado como uma das contribuições para a comunidade acadêmica de mineração de dados educacionais, podendo ser aplicados em outras pesquisas e adaptados para outros tipos de variáveis.

Tendo em vista o crescimento de utilização da linguagem de análise de dados R, outra contribuição é a disponibilização de todos os scripts de análise de dados construídos durante a fase de modelagem desta tese (*Apêndices E, F, G, I e J*). Ao todo foram desenvolvidos um script para análise factorial de dados, um script para análise descritiva, quatro scripts para construção de modelos de classificação e dois scripts para avaliação e comparação de classificadores.

Por fim, uma das contribuições com grande potencial para replicação em outras pesquisas foi o desenvolvimento da arquitetura de software e ferramenta de visualização de dados, desenvolvida e avaliada nesta tese. Esta aplicação foi desenvolvida por meio de um processo distribuído e está disponível no Github (<https://github.com/grupoccte>), podendo ser adaptada e evoluída para atender diversas instituições, bem como ser integrada em diversos sistemas de LMS.

8.4 Trabalhos Futuros

Este trabalho possibilitou o desenvolvimento de um arcabouço tecnológico para previsão de desempenho de alunos, baseado em características de autorregulação da aprendizagem. Como possibilidades de trabalhos futuros, que podem ser explorados a partir desse estudo, temos:

- Evolução da interface desenvolvida de acordo com as possibilidades de melhorias identificadas na fase de avaliação de usabilidade;
- Integração com outras plataformas de LMS;
- Validação do modelo de classificação em outras instituições de ensino superior e em cursos diversificados;
- Realização de testes da efetividade da interface em um ambiente real de Educação a Distância; e
- Construção de novos modelos preditivos, relacionados a outros fenômenos de aprendizagem, a serem acoplados à interface.

8.5 Produção Científica Desenvolvida

Durante o desenvolvimento do doutorado, foram desenvolvidos artigos científicos, publicados em conferências e periódicos, com temáticas afins a esta proposta, no sentido de promover um aperfeiçoamento do proponente enquanto

pesquisador, além de proporcionar uma apropriação de conteúdos e ferramentas necessárias ao bom andamento desta tese.

8.4.1 Artigos publicados em Anais de Conferências Científicas

N	Artigo
1	RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; GOMES, A. S.; SEDRAZ SILVA, JOÃO CARLOS; MACIEL, A. M. A. Discovering level of participation in MOOCs through clusters analysis. In: IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, Texas, USA. IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016.
2	SEDRAZ SILVA, JOÃO CARLOS ; RAMOS, J. L. C. ; RODRIGUES, RODRIGO LINS ; SOUZA, F. F. ; GOMES, A. S. ; MACIEL, A. M. A. An EDM Approach to the Analysis of Students' Engagement in Online Courses from Constructs of the Transactional Distance. In: IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, Texas, USA. IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016.
3	RICHARD MACHADO, C. J.; LIMA, B. R. B. ; MACIEL, A. M. A. ; RODRIGUES, RODRIGO LINS. An investigation of students behavior in discussion forums using Educational Data Mining. In: The 28th International Conference on Software Engineering and Knowledge Engineering, 2016, California, USA. The 28th International Conference on Software Engineering and Knowledge Engineering, 2016.
4	RODRIGUES, RODRIGO LINS; SILVA, J. C. S.; RAMOS, J. L. C.; SOUZA, F. F.; GOMES, A.S. Uma Abordagem de Regressão Múltipla para Validação de Variáveis de Autorregulação da Aprendizagem em Ambientes de LMS. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXVII Simpósio Brasileiro de Informática na Educação, 2016.
5	RAMOS, J. L. C.; SILVA, J. C. S.; RODRIGUES, RODRIGO LINS; GOMES, A. S.; SOUZA, F.F. Mapeamento de dados de um LMS para medida de construtos da distância transacional. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXVII Simpósio Brasileiro de Informática na Educação, 2016.
6	ISEPPON, D.; GOMES, A. S.; RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; SEDRAZ SILVA, JOÃO CARLOS. Impact of teaching action on student interaction in virtual learning environments: canonical correlation analysis. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXVII Simpósio Brasileiro de Informática na Educação, 2016.
7	SOARES, F.; RICHARD MACHADO, C. J.; DINIZ, D.; MACIEL, A. M. A.; RODRIGUES, RODRIGO LINS. Educational Data Mining to support Distance Learning students with difficulties in the Portuguese Grammar. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXVII Simpósio Brasileiro de Informática na Educação, 2016.
8	SILVA, J. C. S.; GOMES, A. S.; RAMOS, J. L. C.; RODRIGUES, RODRIGO LINS; SOUZA, F. F. Análise da efetividade de componente curricular para motivar colaboradores em projetos de Software Público. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXII Workshop de Informática na Escola, 2016.
9	SEDRAZ SILVA, JOÃO CARLOS; GOMES, A. S.; RAMOS, J. L. C.; SOUZA, F. F.; RODRIGUES, RODRIGO LINS. Dificuldades e sugestões para a adoção de Software Público em atividades acadêmicas. In: V Congresso Brasileiro de Informática na Educação, 2016, Uberlândia. XXVII Simpósio Brasileiro de Informática na Educação, 2016.

10	CAVALCANTI, A. G. G.; GOMES, A. S.; RODRIGUES, RODRIGO LINS; RAMOS, J. L. C. Apresentação e avaliação de ferramenta de visualização gráfica das interações dos estudantes no ambiente Moodle. In: Workshop de Trabalhos de Iniciação Científica e Graduação - WTICGBASE, 2016, Maceió. ERBASE - Escola Regional de Computação Bahia-Alagoas/Sergipe, 2016.
11	SOUZA, V. L.; MELO FILHO, I. J.; GOMES, A. S.; RODRIGUES, RODRIGO LINS. O Processo de Autoavaliação em Ambientes Colaborativos e sua Relação com a Mineração de Dados Educacionais. In: XXXVI Congresso da Sociedade Brasileira de Computação, 2016, Porto Alegre/RS. V - Workshop de Desafios da Computação Aplicada à Educação, 2016.
12	SILVA, R.; RAMOS, J. L. C.; RODRIGUES, RODRIGO LINS; GOMES, A. S.; FONSECA, A. Mineração de dados educacionais na análise das interações dos alunos em um Ambiente Virtual de Aprendizagem. In: IV Congresso Brasileiro de Informática na Educação, 2015, Maceió. Simpósio Brasileiro de Informática na Educação, 2015.
13	AMARAL, Y.; RODRIGUES, RODRIGO LINS; MACIEL, A. M. A.; SOARES, F.; GONCALVES, A. F. D. Analysis of the linearity relation between the tutors actions and the learners interactions in Distance Education courses. In: IV Congresso Brasileiro de Informática na Educação, 2015, Maceió. Simpósio Brasileiro de Informática na Educação, 2015.
14	AMARAL, Y.; MACIEL, A. M. A.; RODRIGUES, RODRIGO LINS. Development of a Virtual Assistant for Alerts and Notifications in a Learning Environment. In: IV Congresso Brasileiro de Informática na Educação, 2015, Maceió. Simpósio Brasileiro de Informática na Educação, 2015.
15	RAMOS, JORGE LUIS CAVALCANTI; SILVA, JOAO CARLOS SEDRAZ; RODRIGUES, RODRIGO LINS; GOMES, ALEX SANDRO. Analysis of students' expectations and skills in e-learning. In: 2014 9th Iberian Conference on Information Systems and Technologies (CISTI), 2014, Barcelona. 2014 9th Iberian Conference on Information Systems and Technologies (CISTI). p. 1.
16	CAVALCANTI, L.; MACIEL, A. M. A.; RODRIGUES, R. L. Avaliação do Perfil de Uso no Ambiente Moodle Utilizando Técnicas de Mineração de Dados. In: Simpósio Brasileiro de Informática na Educação, 2014, Dourados. Congresso Brasileiro de Informática na Educação, 2014.
17	MACIEL, A. M. A.; RODRIGUES, R. L.; CARVALHO FILHO, E. C. B. Desenvolvimento de um Assistente Virtual Integrado ao Moodle para Suporte a Aprendizagem Online. In: Simpósio Brasileiro de Informática na Educação, 2014, Dourados. Congresso Brasileiro de Informática na Educação.
18	RAMOS, J. L. C.; RODRIGUES, R. L.; GOMES, A. S.; SILVA, J. C. S. Analizando Fatores que Afetam o Desempenho de Estudantes Iniciantes em um Curso a Distância. In: Simpósio Brasileiro de Informática na Educação, 2014, Dourados. Congresso Brasileiro de Informática na Educação, 2014.
19	ADEODATO, P. J. L.; SANTOS FILHO, M. M.; RODRIGUES, R. L. Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: Congresso Brasileiro de Informática na Educação, 2014, Dourados. Simpósio Brasileiro de Informática na Educação, 2014.
20	RAMOS, J. L. C.; GOMES, A. S.; RODRIGUES, R. L.; SILVA, J. C. S. A literatura brasileira sobre mineração de dados educacionais. In: Congresso Brasileiro de Informática na Educação, 2014, Dourados. Workshop de Mineração de Dados em Ambientes Virtuais do Ensino/Aprendizagem, 2014.
21	SEIXAS, L. R.; GOMES, A. S.; MELO FILHO, I. J.; RODRIGUES, R. L. Gamificação como Estratégia no Engajamento de Estudantes do Ensino Fundamental. In: Simpósio Brasileiro de Informática na Educação, 2014, Dourados. Congresso Brasileiro de Informática na Educação, 2014.

22	PERRIS, PAULO A. R.; RODRIGUES, RODRIGO LINS; GOMES, ALEX SANDRO. Concepção de ferramentas de percepção, representando dados do fluxo de atividades discentes em ambiente colaborativo de aprendizagem. In: XXIV Simpósio Brasileiro de Informática na Educação, 2013.
23	RODRIGUES, RODRIGO LINS; PERRIS, PAULO A. R.; GOMES, ALEX SANDRO. Percepção do uso de jogos integrados ao LMS Amadeus. In: XXIV Simpósio Brasileiro de Informática na Educação, 2013.
24	RODRIGUES, RODRIGO LINS; DE MEDEIROS, FRANCISCO P. A.; GOMES, ALEX SANDRO. Modelo de Regressão Linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In: XXIV Simpósio Brasileiro de Informática na Educação, 2013.

7.2.2 Artigos publicados em Periódicos Científicos

N	Artigo
1	RAMOS, J. L. C.; SILVA, R.; RODRIGUES, RODRIGO LINS; SEDRAZ SILVA, JOÃO CARLOS; GOMES, A. S. A Comparative Study between Clustering Methods in Educational Data Mining. Revista IEEE América Latina, v. 4, p. 34, 2016.
2	RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; SEDRAZ SILVA, JOÃO CARLOS; GOMES, A. S. Discovery engagement patterns MOOCs through cluster analysis. Revista IEEE América Latina, v. 4, p. 34, 2016.
3	SEDRAZ SILVA, JOÃO CARLOS; RAMOS, J. L. C. ; RODRIGUES, RODRIGO LINS ; SOUZA, F. F. ; GOMES, A. S. Analysis of engagement of students based on Transactional Distance from the Educational Data Mining. RENOTE. Revista Novas Tecnologias na Educação, v. 1, p. 1, 2016.
4	SEDRAZ SILVA, JOÃO CARLOS; RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; SOUZA, F. F.; GOMES, A. S. Educational Data Mining in Oriented Learning Activities. RENOTE. Revista Novas Tecnologias na Educação, v. 1, p. 1, 2016
5	RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; SEDRAZ SILVA, JOÃO CARLOS; SOUZA, V. L.; GOMES, A. S.; SOUZA, F. F. Systematic mapping about approaches measurement of Self-Regulated Learning. RENOTE. Revista Novas Tecnologias na Educação, v. 1, p. 1, 2016.
6	RODRIGUES, RODRIGO LINS; RAMOS, J. L. C.; SEDRAZ SILVA, JOÃO CARLOS; GOMES, A. S.; TENORIO, A. C.; SOUZA, F. F. Validation of a self-regulated learning measurement instrument in the Brazilian context using confirmatory factor analysis. RENOTE. Revista Novas Tecnologias na Educação, v. 1, p. 1, 2016.
7	MACIEL, A. M. A.; RODRIGUES, RODRIGO LINS; CARVALHO FILHO, E. C. B. Desenvolvimento de uma Ferramenta para a Construção e Integração de Personagens Virtuais Animados com Voz Sintética a Materiais Didáticos para EAD. Revista Brasileira de Informática na Educação, v. 23, p. 161, 2015.
8	MELO FILHO, IVANILDO JOSÉ; CARVALHO, ROSÂNGELA SARAIVA; GOMES, ALEX SANDRO; PERRIS, PAULO ANDRÉ DA ROCHA; RODRIGUES, RODRIGO LINS; FELICIANO, FÁBIO DENILSON DE OLIVEIRA. Análise comparativa da Usabilidade dos Ambientes de Gestão da Aprendizagem Amadeus e Moodle. Revista Brasileira de Informática na Educação, v. 22, p. 107-120, 2014.

Cumprindo o planejamento desta tese, esperamos a submissão de mais dois artigos científicos (Qualis A1 ou A2) para periódicos reconhecidos internacionalmente, assim como também mais dois artigos para conferências internacionais nas áreas de *Educational Data Mining* (EDM) e *Technology Enhanced Learning* (TEL).

Referências

- ADESOPE, O. O.; ZHOU, M.; NESBIT, J. C. Achievement Goal Orientations and Self-Reported Study Strategies as Predictors of Online Studying Activities. **Journal of Educational Computing Research**, p. 436-458, 2015.
- AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. **ACM SIGMOD Conference Management of Data**, Washington, 1993.
- AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. **ACM SIGMOD Conference Management of Data**, Washington, 1993.
- AGUSTIANI, H.; CAHYAD, S.; MUSA, M. Self-efficacy and Self-Regulated Learning as Predictors of Students Academic Performance. **The Open Psychology Journal**, 2016.
- AKCAYNAR, G.; ALTUN, A.; COSGUN, E. Investigating Students' Interaction Profile in an Online Learning Environment with Clustering. In **Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on IEEE**, p. 109-111, 2014.
- ALI, L. et al. Leveraging MSLQ Data for Predicting Students Achievement Goal Orientations. **Journal of Learning Analytics**, p. 157-160, 2014.
- ALLY, M. The Theory and Practice of Online Learning. Edmonton. **CA: Athabasca University Press**, 2004.
- ANAYA, A. R.; BOTICARIO, J. G. Content-free collaborative learning modeling using data mining. **User Modeling and User-Adapted Interaction**, p. 181-216, 2011.
- ANTONENKO, P. D.; TOY, S.; NIEDERHAUSER, D. S. Using cluster analysis for data mining in educational technology research. **Educational Technology Research and Development**, p. 383-398, 2012.

- ARKSEY, H.; O'MALLEY, L. Scoping Studies: Towards A Methodological Framework, 2005.
- AZEVEDO, R. Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. **Educational Psychologist**, p. 199-209, 2005.
- AZEVEDO, R. et al. Adaptive human scaffolding facilitates adolescents' self-regulated learning. **Instructional Science**, p. 381–412, 2005.
- BAKER, R. S. J. Data mining for education. In: McGaw, B., Peterson, P., Baker, E. (eds.) International Encyclopedia of Education. **Elsevier, Amsterdam**, v. 7, n. 3, p. 112–118, 2010.
- BAKER, S. J. D.; CARVALHO, M. J. B. D.; ISOTANI,. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, Agosto 2011.
- BARNARD, L. et al. **The relationship of epistemological beliefs with self-regulatory skills in the online course environment**. [S.I.]: Journal of Online and Learning Teaching, 2008.
- BARNARD, L. et al. Measuring self-regulation in online and blended learning environments. **The Internet and Higher Education**, p. 1-6, 2009.
- BARNARD, L. et al. **Measuring self-regulation in online and blended learning environments**. [S.I.]: The Internet and Higher Education, 2009.
- BARNARD, L. et al. Measuring self-regulation in online and blended learning environments. **The Internet and Higher Education**, p. 1-6, 2009.
- BAVARESCO, J.; ROSA, R. Redes Neurais com Neuralnet e Shiny. In: **II Simpósio de Informática IFSUL Passo Fundo**, Passo Fundo, p. 55-64, 2015.
- BEISHUIZEN, J.; STEFFENS, K. A conceptual framework for research on self-regulated learning, 2011.
- BERKHIN, P. Survey of clustering data mining techniques. In J. Kogan, C. Nicholas, &

- M. Teboulle (Eds.), Grouping multidimensional data. **Heidelberg: Springer.**, p. 25–71, 2006.
- BIENKOWSKI, M.; FENG, M.; MEANS, B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. **US Department of Education, Office of Educational Technology**, p. 1-57, 2012.
- BOLFARINE, H.; BUSSAB, O. H. **Elementos de amostragem**. São Paulo: Edgard Blücher, 2005.
- BONDAREVA, D. et al. Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In **Artificial Intelligence in Education**, Springer Berlin Heidelberg, p. 229-238, 2013.
- BORKOWSKI, J. G. Metacognition: Theory or chapter heading? *Learning and Individual Differences*, p. 391–402, 1996.
- BRAGAGNOLOG, F.; LOURDES , M. Auto-regulação da Aprendizagem: Atuação do pedagogo em espaços não-escolares. **Tese de Doutorado**, 2006.
- BRAUNSTEIN, A. W.; LESSER, M. H.; PESCATRICE, D. R. The impact of a program for the disadvantaged on student retention. **College Student Journal**, 2008.
- BREIMAN, L. et al. Classification and Regression Trees. **Wadsworth**, 1984.
- BROADBENT, J.; POON, W. L. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. **The Internet and Higher Education**, n. 27, p. 1-13, 2015.
- BROADBENT, J.; POON, W. L. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. **The Internet and Higher Education**, 2015. 1-13.
- BUTLER, D. L.; WINNE, P. H. Feedback and Self-Regulated Learning: A Theoretical Synthesis. **Review of Educational Research**, p. 245–281, 1995.
- CAMBRUZZI, W. L et al. **Um Modelo para Gerenciamento de Multiplas Trilhas Aplicado a Sistemas de Apoio a Educação**. Simpósio Brasileiro de Informática

- na Educação. [S.I.]: [s.n.]. 2012.
- CAMERON, A. C.; TRIVEDI, P. K. **Regression analysis of count data**. [S.I.]: Cambridge University Press, 1998.
- CAMPOS, H. Estatística experimental não-paramétrica. **Esalq. Piracicaba-SP**, p. 230p, 1987.
- CARSON, A. D. Predicting student success from the LASSI for learning online (LLO). **Journal of Educational Computing Research**, 2011. 399–414.
- CETINTAS, et al. A joint probabilistic classification model of relevant and irrelevant sentences in mathematical word problems. **arXiv preprint arXiv**, 2014.
- CHANG, M. M. Effects of self-monitoring on web-based language learner's performance and motivation. **CALICO Journal**, 2010. 298–310.
- CHANLIN, L. J. Learning strategies in web-supported collaborative project. **Innovations in Education and Teaching International**, 2012. 319–331.
- CHAPMAN, P. et al. CRISP-DM 1.0 Step-by-step data mining guide, 2000.
- CHAU, M. et al. Uncertain data mining: an example in clustering location data. In W. K. Ng, M. Kitsuregawa, J. Li, & K. Chang (Eds.). **Heidelberg: Springer**, v. 3918, 2006.
- CHO, M. H.; JONASSEN, D. Development of the human interaction dimension of the Self-Regulated Learning Questionnaire in asynchronous online learning environments. **Educational Psychology**, p. 117-138, 2009.
- CHO, M. H.; SHEN, D. Self-regulation in online learning. **Distance Education**, 2013. 290–301.
- CLEARY, T. J. Emergence of self-regulated learning microanalysis: Historical overview, essential features, and implications for research and practice. **Zimmerman, B. J. and Schunk, D. H., editors, Handbook of Self-Regulation of Learning and Performance**, v. chapter 21, p. 329–345, 2011.
- COLTHORPE, K. et al. Know thy student! Combining learning analytics and critical

- reflections to develop a targeted intervention for promoting self-regulated learning. **Journal of Learning Analytics**, p. 134-15, 2015.
- CORMACK, R. M. A review of classification. **Journal of the Royal Statistical Society. Series A (General)**, p. 321-367, 1971.
- CRISTIANINI, N.; SHawe-Taylor., J. An introduction to support Vector Machines: and other kernel-based learning methods. **New York: Cambridge University**, 1999.
- CURRAL, J. **Statistics Packages: A General Overview**. [S.I.]: Universidade de Glasgow, 1994.
- DA SILVA, L. A.; PERES, S. M.; BOSCAROLI, C. **Introdução à Mineração de Dados com Aplicações em R**. 1. ed. Rio de Janeiro: ELSEVIER, 2016.
- DABBAGH, N.; KITSANTAS, A. Supporting self-regulation in student-centered Web-based learning environments. **International Journal on E-Learning**, p. 40-47, 2004.
- DABBAGH, N.; KITSANTAS, A. Supporting self-regulation in student-centered web-based learning environments. **International Journal on E-learning**, p. 40-47, 2004.
- DABBAGH, N.; KITSANTAS, A. Using web-based pedagogical tools as scaffolds for self-regulated learning. **Instructional Science**, p. 513-540, 2005.
- DIAS, M. M. et al. Aplicação de Técnicas de Mineração de dados no Processo de Aprendizagem na Educação a Distância. **Anais do Simpósio Brasileiro de Informática na Educação**, 2008. 105-114.
- DIGNATH, C.; BUTTNER, G. Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. **Metacognition Learning**, 2008. 231–264.
- ECKLES, J. E.; STRADLEY, E. G. A social network analysis of student retention using archival data. **Social Psychology of Education: An International Journal**, p. 165-180, 2012.

- EGAN, J. P. Signal detection theory and ROC analysis. **Academic Press**, New York, 1975.
- EILAM, B.; AHARON, I. Students' planning in the process of self-regulated learning. **Contemporary Educational Psychology**, v. 28, p. 304-334, 2003.
- ESSA, A.; AYAD, H. Student success system: Risk analytics and data visualization using ensembles of predictive models. In **The 2nd International conference on learning analytics and knowledge**, Vancouver, BC, Canada, 2012.
- FAVEIRO, L. et al. Análise de dados: modelagem multivariada para tomada de decisão. **São Paulo: Campus.**, 2009.
- FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, p. 861-874, 2006.
- FERREIRA, A. B. H. Dicionário da Língua Portuguesa. **Nova Fronteira**, 2012.
- FERREIRA, D. F. Estatística multivariada. **Editora UFLA**, 2008.
- FICHER, C.; MILLER, G. Learning strategies for distance education students. **Journal of Agricultural Education**, v. 41, p. 60-68, 2000.
- FIEBIG, A.; HALBRÜGGE, M.; KRAUS, L. User Experience Measurement of a Static Website Compared to a Responsive Website Using AttrakDiff Mini. **Interaction**, 2016.
- FONSECA, J. Indução de árvores de decisão. **Tese de Mestrado**, Lisboa, 1994.
- FREITAS, A. L. P.; RODRIGUES, S. G. **A avaliação da confiabilidade de questionário: uma análise utilizando o coeficiente alfa de Cronbach.** [S.I.]: XII SIMPEP, 2005.
- GARRISON, D. R. Self-directed learning and distance education. **Handbook of distance education**, p. 161-168, 2003.
- GARTON, ; HAYTHORNTHWAITE, C.; WELLMAN, B. Studying Online Social Networks. **JCMC**, 3(1), p. <http://jcmc.indiana.edu/vol3/issue1/garton.html>, 1997.

GAŠEVIĆ, D. et al. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. **The Internet and Higher Education**, v. 24, p. 53-65, 2015.

GOLDSCHMIDT, R.; PASSOS, E. Data Mining: um guia prático. **Campus**, 2005.

GONÇALVES, I. C. Contributos dos Modelos da Auto-Regulação da Aprendizagem para a formação de Alunos e Professores no Ensino Superior, 2010.

GONZÁLEZ-BRENES, J. P.; MOSTOW, J. Predicting task completion from rich but scarce data, 2010.

HAIR, J. F. . B. W. C. . B. B. J. . A. R. E. . & T. R. L. Multivariate data analysis (Vol. 6). **Upper Saddle River, NJ: Pearson Prentice Hall.**, 2006.

HAIR, J. F. et al. **Análise multivariada de dados**. [S.I.]: Bookman Editora, 2009.

HAIR, J. F. et al. **Análise multivariada de dados**. [S.I.]: Bookman Editora, 2009.

HARLEY, J. M.; TREVORS, G. J.; AZEVEDO, R. Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. **JEDM-Journal of Educational Data Mining**, p. 104-146, 2013.

HARTLEY, K.; BENDIXEN, L. D. Educational research in the Internet age: Examining the role of individual characteristics. **Educational Researcher**, p. 22-26, 2001.

HASSENZAHL, M. The interplay of beauty, goodness, and usability in interactive products. **Human-computer interaction**, p. 319-349, 2004.

HENDEL, D. D. Efficacy of participating in a first-year seminar on student satisfaction and retention. **Journal of College Student Retention: Research, Theory & Practice**, p. 413-423, 2007.

HIARAGI,. Mineração de Dados em Base de Germoplasma. **Dissertação (Mestrado em Informática) - Universidade de Brasília - UnB, Brasília**, 2008.

HODGES, C. B.; KIM, C. Email, self-regulation, self-efficacy, and achievement in a college online mathematics course. **Journal of Educational Computing Research**, 2010. 207–223.

- HOFFMANN, R. **Análise de regressão:** uma introdução à econometria. [S.I.]: [s.n.], 2015.
- HONG, T. P.; LIN, K. Y.; WANG, S. L. Fuzzy data mining for interesting generalized association rules. **Journal of Fuzzy Sets and systems**, v. 2, n. 138, p. 255–269, 2003.
- JAMES, G. et al. Linear model selection and regularization an introduction to statistical learning. **Springer**, 2013.
- JOHNSON, R. D.; GUEUTAL, H.; FALBE, C. M. Technology, trainees, metacognitive activity and e-learning effectiveness. **Journal of Managerial Psychology**, 2009. 545–566.
- KINZIE , M. B. Requirements and benefits of effective interactive instruction: Learning control, self-regulation, and continuing motivation. **Educational Technology Research and Development**, p. 5-21, 1990.
- KITCHENHAM, B. Guidelines for performing Systematic Literature Reviews in , 2007.
- KIZILCEC, R. F.; PÉREZ-SANAGUSTÍN, M.; MALDONADO, J. J. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. **Computers & Education**, 2016.
- KLEINBAUM, D. G.; KLEIN, M. Analysis of Matched Data Using Logistic Regression. **Springer New York**, p. 389-428, 2010.
- KLINGSIECK, K. B. et al. Procrastination in a distance university setting. **Distance Education**, 2012. 295–310.
- KOEDINGER, K. R. et al. A data repository for the EDM community: The PSLC DataShop. **Handbook of educational data mining**, p. 43, 2010.
- LARANJEIRO, J. B. Contributos para a Análise e Caracterização de. **Tese de Mestrado, Faculdade de Ciências da Universidade do Porto**, Porto, 2008.
- LAWANTO, O. et al. Self-Regulated Learning Skills and Online Activities between Higher and Lower Performers on a Web-Intensive Undergraduate Engineering

- Course. **Journal of Educators Online**, 2014.
- LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, p. 18-36, 2009.
- LONN, S.; TEASLEY, S. D. Saving Time or innovative practice: Investigating perceptions and uses of Learning Management Systems. **Computers and Education**, p. 686-694, 2009.
- LOPEZ, M. I. et al. Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums. **International Educational Data Mining Society**, 2012.
- LORENA, A. C.; CARVALHO, A. C. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, p. 43-67, 2007.
- MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, p. 281-297, 1967.
- MALMBERG, J. et al. Promoting socially shared regulation of learning in CSCL: Progress of socially shared regulation among high-and low-performing groups. **Computers in Human Behavior**, 2015.
- MARBOUTI, F.; DIEFES-DUX, H. A.; MADHAVAN, K. Models for early prediction of at-risk students in a course using standards-based grading. **Computers & Education**, p. 1-15, 2016.
- MARINHO, et al. Um Framework para Mineração de Dados Educacionais Baseado em Serviços Semânticos. **Anais do Simpósio Brasileiro de Informática na Educação**, 2009.
- MATO, C. A.; VEIGA, R. T. **Avaliação da qualidade percebida de serviços**: um estudo em uma organização não-governamental. São Paulo: Caderno de Pesquisa em Administração, 2000.
- MCGAW, B.; PETERSON, P.; BAKER, E. **International Encyclopedia of Education**. Amsterdam: Elsevier, 2010.

- MEDEIROS, F. et al. Architecture for Social Interactions Monitoring in Collaborative Learning Environments as a Support for the Teacher's Awareness. In **2013 IEEE 13th International Conference on Advanced Learning Technologies**, p. 123-127, 2013.
- MERCERON, A.; YACEF, K. Educational Data Mining: a Case Study. **AIED**, p. 467-474, 2005.
- MICHINOV, N. et al. Procrastination, participation, and performance in online learning environments. **Computers & Education**, 2011. 243–252.
- MIGUEL, P. A. C.; SALOMI, G. E. **Uma revisão dos modelos de medição de qualidade em serviços**. Rio de Janeiro: Revista Produção, 2004.
- MIKUT, R.; REISCHL, M. Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. 431-443, 2011.
- MILAN, G. S.; TREZ, G. **Pesquisa de satisfação**: um modelo para planos de saúde. [S.I.]: RAE Eletrônica, Revista de Administração de Empresas, 2005.
- MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. **Editora UFMG**, 2005.
- MOHAMAD, S. K.; TASIR, Z. Educational Data Mining: A Review. **Procedia-Social and Behavioral Sciences**, p. 220-224, 2013.
- NARLI, S.; ÖZGEN, K.; ALKAN, H. In the context of multiple intelligences theory, intelligent data analysis of learning styles was based on rough set theory. **Learning and Individual Differences**, p. 613-618, 2011.
- NUGENT, R.; DEAN, N.; AYERS, E. Skill set profile clustering: the empty K-means algorithm with automatic specification of starting cluster centers. In **Educational Data Mining**, 2010.
- NUSSBAUMER, A. et al. A competence-based service for supporting self-regulated learning in virtual environments. **Journal of Learning Analytics**, p. 101-133, 2015.
- OVIEDO, H. C.; CAMPO-ARIAS, A. **Aproximación al uso del coeficiente alfa de**

- Cronbach. [S.I.]: Revista Colombiana de Psiquiatría, 2005.
- PAI, P. F.; LYU, Y. J.; WANG, Y. M. Analyzing academic achievement of junior high school students by an improved rough set model. **Computers & Education**, p. 889-900, 2010.
- PANADERO, E. et al. How Individual Self-Regulation Affects Group Regulation and Performance A Shared Regulation Intervention. **Small Group Research**, 2015.
- PARDOS, Z. A.; WANG, Q. Y.; TRIVEDI, S. The Real World Significance of Performance Prediction. **International Educational Data Mining Society**, 2012.
- PARK, Y.; LIM, K. Effects of Environmental and Human Constructs on E-Learning Effectiveness in Online University Settings. **Indian Journal of Science and Technology**, p. 103-109, 2015.
- PASSOS, E.; GOLDSCHMIDT, R. Data Mining: um guia prático. **Editora Campus**, Rio de Janeiro, 2005.
- PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert systems with applications**, p. 1432-1462, 2014.
- PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert systems with applications**, v. 4, n. 41, p. 1432-1462, 2014.
- PERES, S. M. et al. Tutorial sobre Fuzzy-c-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamento e Classificação. **Revista de Informática Teórica e Aplicada**, 2012. 120-163.
- PERRY, N. E. et al. Investigating teacher-student interactions that foster self-regulated learning. **Educational Psychologist**, v. 37(1), p. 5–15, 2002.
- PETERSEN, K. E. A. Systematic Mapping Studies in Software Engineering, 2007.
- PICCOLI, G.; AHMAD, R.; IVES, B. Web-based virtual learning environments: a research framework and a preliminary assessment of effectiveness in basic IT skills training. **MIS Quarterly**, p. 401-426, 2001.

- PILATI, R.; LAROS, J. A. **Modelos de equações estruturais em psicologia: conceitos e aplicações.** [S.I.]: Psicologia: teoria e pesquisa, 2007.
- PIMENTEL, P. E.; OMAR, N. **Descobrindo Conhecimentos em Dados de Avaliação da Aprendizagem com Técnicas de Mineração de Dados.** [S.I.]: [s.n.]. 2006.
- PINTRICH, P. R. The role of motivation in promoting and sustaining self-regulated learning. **International Journal of Educational Research**, v. 31, p. 459-470, 1999.
- PINTRICH, P. R. The Role of Goal Orientation in Self-Regulated Learning. **Handbook of Self-Regulation**, San Diego: Academic, p. 451–502, 2000.
- PINTRICH, P. R. A conceptual framework for assessing motivation and self-regulated learning in college students. **Educational Psychology Review**, 2004. 385–407.
- PINTRICH, P. R. The Role of Goal Orientation in Self-regulated Learning. **Monique Boekaerts, Paul R. Pintrich and Moshe Zeidner (Ed.) Handbook of Selfregulation**, San Diego, p. 452-502, 2005.
- PINTRICH, P. R. et al. Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). **Educational and psychological measurement**, n. 53(3), p. 801-813, 1993.
- PRATES, R. O.; BARBOSA, S. Avaliação de Interfaces de Usuário - Conceitos e Métodos. **Jornada de Atualização em Informática**, Campinas, 2003.
- PRATI, R. C.; BATISTA, E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, p. 215-222, 2008.
- PREECE, J.; ROGERS, Y.; SHARP, H. Design de interação. [S.I.]: Bookman, 2005.
- RAD, A.; NADERI, B.; SOLTANI, M. Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. **Expert Systems with Applications**, p. 755-763, 2011.
- RAMOS, J. L. C. et al. A Comparative Study between Clustering Methods in Educational Data Mining. **IEEE Latin America Transactions**, 2016.
- RATES, F. T. S. Y. R. National Collegiate Retention and Persistence to Degree Rates,

2013.

RICHARDSON, M.; ABRAHAM, C.; BOND, R. Psychological correlates of university students' academic performance. A systematic review and meta-analysis. **Psychological Bulletin**, 2012. 353–387.

RIGO, S. J. et al. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, 8 março 2014.

RODRIGO, R. et al. Mapeamento Sistemático sobre Abordagens de Mensuração de Autorregulação da Aprendizagem. **RENOTE - Revista Novas Tecnologias na Educação**, 2016.

RODRIGUES, et al. A literatura brasileira sobre mineração de dados educacionais. **Workshop de Mineração de Dados Educaionais**, Dourados, Novembro 2014.

RODRIGUES, R. L. et al. Validação de um instrumento de mensuração de autorregulação da aprendizagem em contexto brasileiro usando análise fatorial confirmatória. **RENOTE**, n. 14(1), 2016.

ROGERS, Y.; SHARP, H.; PREECE, J. **Design de Interação**. [S.I.]: Bookman Editora, 2013.

ROLL, I.; WINNE, P. H. Understanding, evaluating, and supporting self-regulated learning using learning analytics. **Journal of Learning Analytics**, p. 7-12, 2015.

ROMERO, ; VENTURA,. Data mining in education. **WIREs Data Mining Knowl Discov**, February 2013. 12-27.

ROMERO, C. et al. **Handbook of Educational Data Mining**. [S.I.]: [s.n.], 2010.

ROMERO, C. et al. Handbook of educational data mining. CRC Press., 2010.

ROMERO, C.; VENTURA, C. Educational data mining: a review of the state of the art. **IEEE Trans. Syst. Man Cybern. Part C Appl. Rev**, 2010. 601-618.

ROYO, J. Design digital. Rosari, São Paulo, 2008.

- SABOURIN, J. et al. Predicting student self-regulation strategies in game-based learning environments. **International Conference on Intelligent Tutoring Systems**, Springer Berlin Heidelberg, p. 141-150, 2012.
- SABOURIN, J. L.; MOTT, B. W.; LESTER, J. C. Early Prediction of Student Self-Regulation Strategies by Combining Multiple Models. **International Educational Data Mining Society**, 2012.
- SALOMI, G. E.; MIGUEL, P. A. C.; ABACKERLI, A. J. **SERVQUAL x SERVPERF: comparação entre instrumentos para avaliação da qualidade de serviços internos.** [S.I.]: Gestão & Produção, 2005.
- SANCHEZ-SANTILLAN, M. et al. MeL: Modelo de adaptación dinámica del proceso de aprendizaje en eLearning. **anales de psicología**, p. 106-114, 2016.
- SCHIAVONI, A. S. Um estudo comparativo de métodos para balanceamento do conjunto de treinamento em aprendizado de redes neurais artificiais. **Dissertação de Mestrado - UFLA**, 2015.
- SCHMITZ, B.; WIESE, B. S. New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. **Contemporary Educational Psychology**, n. 31(1), p. 64–96, 2006.
- SCHOOR, C.; BANNERT, M. Exploring regulatory processes during a computer-supported collaborative learning task using process mining. **Computers in Human Behavior**, p. 1321-1331, 2012.
- SCHUNK, D. H. **Self-regulated learning The educational legacy of Paul R. Pintrich.** [S.I.]: Educational Psychologist, 2005.
- SCHUNK, D. H.; ZIMMERMAN, B. J. Self-regulated learning: From teaching to self-reflective practice. **New York: The Guilford Press**, 1998.
- SCHUNK, D. H.; ZIMMERMAN, B. J. Self-regulated learning: From teaching to self-reflective practice. **Guilford Press**, 1998.
- SEGEDY, J. R.; KINNEBREW, J. S.; BISWAS, G. Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning

- environments, p. 13-48, 2015.
- SERDYUKOV, P.; HILL, R. Flying with clipped wings: are students independent in online college classes? **Journal of Research in Innovative Teaching**, 2013. 54–67.
- SHETH, J.; PATEL, B. Best practices for adaptation of Data mining techniques in Education Sector. **National Journal of System and Information Technology**, p. 186, 2010.
- SILVA, A. L. A auto-regulação na aprendizagem: a demarcação de um campo de estudos e de intervenção. **Aprendizagem auto-regulada pelo estudante: perspectivas psicológicas e educacionais**, p. 17-39, 2004.
- SONNENBERG, C.; BANNERT, M. Discovering the Effects of Metacognitive Prompts on the Sequential Structure of SRL-Processes Using Process Mining Techniques. **Journal of Learning Analytics**, p. 72-100, 2015.
- SOUZA, F. V. C.; GOMES, A. S. ESTRATÉGIAS DE AUTORREGULAÇÃO DE APRENDIZAGEM. **Tese de Doutorado**, Recife, p. 1-268, Agosto 2012.
- STONE, M. Cross-validatory choise and assessment of statistical predictions. **Journal of the Royal Statistical Society**, London, p. 111-148, 1974.
- TALLENT-RUNNELS, M. K. et al. Teaching courses online: A review of the research. **Review of educational research**, p. 93-135, 2006.
- TAN, P. N.; STEINBACH, M.; KUMAR, V. Introdução ao datamining: mineração de dados. **Ciencia Moderna**, 2009.
- TAN, P. N.; STEINBACH, M.; KUMAR, V. Introdução ao datamining: mineração de dados. **Ciencia Moderna**, 2009.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining: Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.
- TESTA, M. G. . & F. H. uto-Regulação da Aprendizagem: analisando o perfil do estudante de Administração. **ENCONTRO NACIONAL DA ASSOCIAÇÃO**

- NACIONAL DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO, p. 29-45, 2005.
- TRIVEDI, S. et al. Co-Clustering by Bipartite Spectral Graph Partitioning for Out-of-Tutor Prediction. **International Educational Data Mining Society**, 2012.
- UNESCO. **Relatório de Monitoramento Global de EPT**. UNESCO. [S.I.]. 2015.
- URDAN, A. T. **Qualidade de Serviços médicos na perspectiva do cliente**. [S.I.]: Revista de Administração de Empresas, 2001.
- VAN DEN BOOM, G.; PAAS, F.; VAN MERRIËNBOER, J. J. Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. **Learning and Instruction**, 2007. 532–548.
- VEIGA SIMÃO, A. M. Auto-regulação da aprendizagem: um desafio para a formação de professores. **Formação de professores de línguas estrangeiras: reflexões, estudos e experiências**, p. 192-206, 2006.
- VERMEEREN, A. P. et al. User experience evaluation methods: current state and development needs. **Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries**, p. 521-530, 2010.
- WANG, C. H.; SHANNON, D.; ROSS, M. Students' characteristics, self-regulated learning, technology self-efficacy, and course outcomes in online learning. **Distance Education**, 2013. 302–323.
- WANG, C. H.; SHANNON, D.; ROSS, M. Students' characteristics, self-regulated learning, technology self-efficacy, and course outcomes in online learning. **Distance Education**, 2013. 302–323.
- WANG, J. et al. The impact of faculty perceived reconfigurability of learning management systems on effective teaching practices. **Computers & Education**, v. 61, p. 146-157, 2013.
- WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of American Statistical**, 1963.
- WAZLAWICK, S. Metodologia de Pesquisa para Ciência da Computação. **Elsevier**,

- Rio de Janeiro, 2008.
- WINNE, P. H.; HADWIN, A. F. Studying as self-regulated learning. **Metacognition in Educational Theory and Practice**, p. 27–30, 1998.
- WINNE, P. H.; JAMIESON-NOEL, D. Exploring students' calibration of self reports about study tactics and achievement. **Contemporary Educational Psychology**, p. 551–572, 2002.
- WINNE, P. H.; PERRY, N. E. **Measuring self-regulated learning**. [S.l.]: [s.n.], 2000. 531–566 p.
- WU, Z.; LI, C. H. L0-constrained regression for data mining. In W. K. Ng, M. Kitsuregawa, J. Li, & K. Chang (Eds.), **Advances in knowledge discovery and data**. Heidelberg: Springer, v. 4426, 2007.
- YANTO, I. T. R. et al. Applying variable precision rough set model for clustering student suffering study's anxiety. **Expert Systems with Applications**, p. 452-459, 2012.
- YOU, J. W. Identifying significant indicators using LMS data to predict course achievement in online learning. **The Internet and Higher Education**, p. 23-30, 2016.
- YUKSELTURK, E.; BULUT, S. Gender differences in self-regulated online learning environment. **Journal of Educational Technology & Society**, p. 12-22, 2009.
- ZEIDNER, M.; BOEKAERTS, P.; PINTRCH, P. Handbook of self-Regulation. **New York: Academic Press**, p. 13-39, 2000.
- ZIMMERMAN, B. J. . & M. A. R. Self-regulation: Where metacognition and motivation intersect. **Handbook of metacognition in education**, p. 299-315, 2009.
- ZIMMERMAN, B. J. A social cognitiveview of self-regulated learning. **J. Educ. Psychol**, p. 329-339, 1989.
- ZIMMERMAN, B. J. Developing sel-fulfilling cycles of academic regulation: an analysis of exemplary instructional models. **New York, The Guilford Press**, p. 1-19, 1998.
- ZIMMERMAN, B. J. Attaining Self-Regulation: a social cognitive perspective.

Handbook of SelfRegulation, p. 13–39, 2000.

ZIMMERMAN, B. J. Investigation self-regulation and motivation: historical background, methodological developments, and future prospects. **American Educational Research Journal**, 2008. 166–183.

ZIMMERMAN, B. J.; BANDURA, A. Impact of self-regulatory influences on writing course attainment. **American Educational Research Journal**, p. 845-862, 1994.

ZIMMERMAN, B. J.; MARTINEZ-PONS, M. Development of a structured interview for assessing student use of self-regulated learning strategies. **American Educational Research Journal**, p. 614-628, 1986.

ZIMMERMAN, B. J.; PONS, M. M. Development of a structured interview for assessing student use of self-regulated learning strategies. **American educational research journal**, v. 23(4), p. 614-628, 1986.

Apêndices



Lista de artigos do mapeamento sistemático

Trabalho	Periódico/Conferência	Autor	Ano
Measuring self-regulation in online and blended learning environments	The Internet and Higher Education	L Barnard, WY Lan, YM To, VO Paton	2009
Gender Differences in Self-Regulated Online Learning Environment	Educational Technology & Society	E Yukselturk, S Bulut	2009
Profiles in Self-Regulated Learning in the Online Learning Environment	InternationalReview of Research in Open andDistanceLearning	L Barnard-Brak, VO Paton, WY Lan	2010
Exploring regulatory processes during a computer-supported collaborative learning task using process mining	Computers in Human Behavior	C Schoor, M Bannert	2012
Open group-sourcing: a social learning model based on peer production and feedback	Eighth International Conference on Complex, Intelligent and Software Intensive Systems	L Moccozet, C Tardy	2014
Self-Regulated Learning Skills and Online Activities Between Higher and Lower Performers on a Web-Intensive Undergraduate Engineering Course	Journal of Educators Online	O Lawanto, HB Santoso, KN Lawanto, W Goodridge	2014
Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions	Internet and Higher Education	D Gašević, O Adesope, S Joksimović, V Kovanović	2015
Investigation of the Relationship between Learning Process and Learning Outcomes in E-Learning Environments	Eurasian Journal of Educational Research	H YURDUGÜL, NM ÇETİN	2015
Promoting socially shared regulation of learning in CSCL: Progress of socially shared regulation among high- and low-performing groups	Computers in Human Behavior	J Malmberg, S Järvelä, H Järvenoja, E Panadero	2015

Development of the human interaction dimension of the Self-Regulated Learning Questionnaire in asynchronous online learning environments	Educational Psychology	MH Cho, D Jonassen	2009
Achievement Goal Orientations and Self-Reported Study Strategies as Predictors of Online Studying Activities	Journal of Educational Computing Research	OO Adesope, M Zhou, JC Nesbit	2015
How Individual Self-Regulation Affects Group Regulation and Performance: A Shared Regulation Intervention	Small Group Research	E Panadero, PA Kirschner, S Järvelä, J Malmberg	2015
Identifying significant indicators using LMS data to predict course achievement in online learning	Internet and Higher Education	JW You	2016
MeL: Modelo de adaptación dinámica del proceso de aprendizaje en eLearning	Anales de Psicología	M Sanchez-Santillan, MP Paule-Ruiz, R Cerezo	2016
Effects of environmental and human constructs on e-learning effectiveness in online university settings	Indian Journal of Science and Technology	Y Park, K Lim	2015
Leveraging MSLQ Data for Predicting Students	Journal of Learning Analytics	L Ali, M Hatala, D Gašević, PH Winne	2014
Know Thy Student! Combining Learning Analytics and Critical	Journal of Learning Analytics	K Colthorpe, K Zimbardi, L Ainscough, S Anderson	2015
A Competence-based Service for Supporting Self-Regulated Learning	Journal of Learning Analytics	A Nussbaumer, EC Hillemann, C Gütl, D Albert	2015
Understanding, evaluating, and supporting self-regulated learning	Journal of Learning Analytics	I Roll, PH Winne	2015
Using Coherence Analysis to Characterize Self-Regulated Learning	Journal of Learning Analytics	JR Segedy, JS Kinnebrew, G Biswas	2015
Discovering the Effects of Metacognitive Prompts on the Sequential Structure of SRL	Journal of Learning Analytics	C Sonnenberg, M Bannert	2015
Identifying significant indicators using LMS data to predict course achievement	Journal of Learning Analytics	JW You	2015
Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning	Lecture Notes in Computer Science	D Bondareva, C Conati, R Feyzi-Behnagh, JM Harley	2011
Early Prediction of Student Self-Regulation Strategies by Combining Multiple Models	International conference on educational data mining	JL Sabourin, BW Mott, JC Lester	2012

Resultados do Questionário OSLQ

A figura abaixo mostra a coleta dos dados do questionário. Foram coletados dados de 466 respondentes de alunos do NEAD/UPE durante um período de dois meses.

Respostas submetidas: 466

Questões: 24

(1) Eu defino estratégias para a realização das minhas tarefas em cursos online.

- 1 Discordo totalmente:  15 (3,22 %)
- 2:  13 (2,79 %)
- 3:  92 (19,74 %)
- 4:  100 (21,46 %)
- 5 Concordo totalmente:  246 (52,79 %)

(2) Eu defino metas de curto prazo (diário ou semanal), bem como metas de longo prazo (mensal ou para o semestre).

- 1 Discordo totalmente:  29 (6,22 %)
- 2:  19 (4,08 %)
- 3:  104 (22,32 %)
- 4:  125 (26,82 %)
- 5 Concordo totalmente:  189 (40,56 %)

(3) Eu mantendo um alto padrão de estudo e aprendizagem no meu curso.

- 1 Discordo totalmente:  14 (3,00 %)
- 2:  30 (6,44 %)
- 3:  115 (24,68 %)
- 4:  148 (31,76 %)

- 5 Concordo totalmente:  159 (34,12 %)

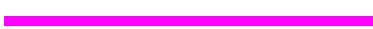
(4) Eu defino metas que me ajudam com o tempo de estudo dedicado para os meus cursos à distância.

- 1 Discordo totalmente:  6 (1,29 %)

- 2:  14 (3,00 %)

- 3:  75 (16,09 %)

- 4:  116 (24,89 %)

- 5 Concordo totalmente:  255 (54,72 %)

(5) Eu não comprometo a qualidade do meu trabalho porque é em EAD.

- 1 Discordo totalmente:  112 (24,03 %)

- 2:  24 (5,15 %)

- 3:  55 (11,80 %)

- 4:  62 (13,30 %)

- 5 Concordo totalmente:  213 (45,71 %)

(6) Eu escolho o local onde realizo meus estudos para evitar distrações.

- 1 Discordo totalmente:  2 (0,43 %)

- 2:  7 (1,50 %)

- 3:  34 (7,30 %)

- 4:  74 (15,88 %)

- 5 Concordo totalmente:  349 (74,89 %)

(7) Eu procuro um lugar confortável para estudar.

- 1 Discordo totalmente:  2 (0,43 %)

- 2:  7 (1,50 %)

- 3:  24 (5,15 %)

- 4:  54 (11,59 %)

- 5

Concordo totalmente:  379 (81,33 %)

(8) Eu sei onde posso estudar de forma mais eficiente quando me dedico a EAD.

- 1 Discordo totalmente:  4 (0,86 %)

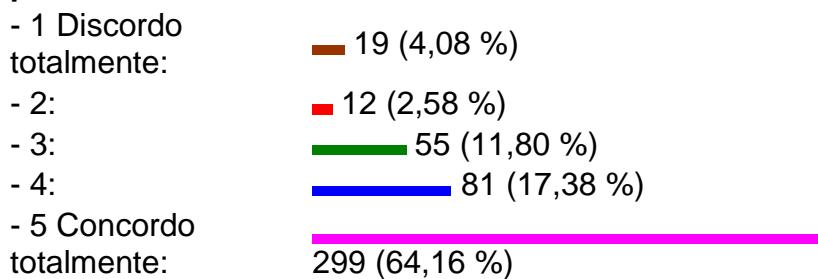
- 2:  10 (2,15 %)

- 3:  36 (7,73 %)

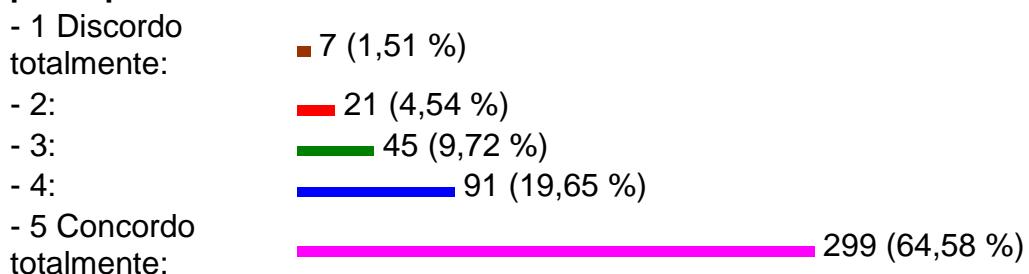
- 4:  91 (19,53 %)

- 5 Concordo totalmente:  325 (69,74 %)

(9) Eu escolho um horário do dia que tenha poucas distrações para realizar meus estudos da EAD.



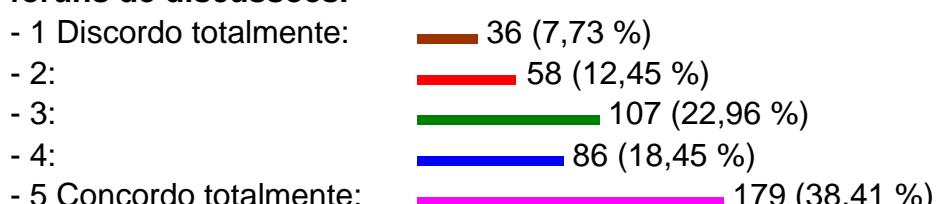
(10) Eu tento fazer esquemas e anotações relacionadas com os conteúdos das disciplinas, porque as anotações são importantes para aprender em EAD



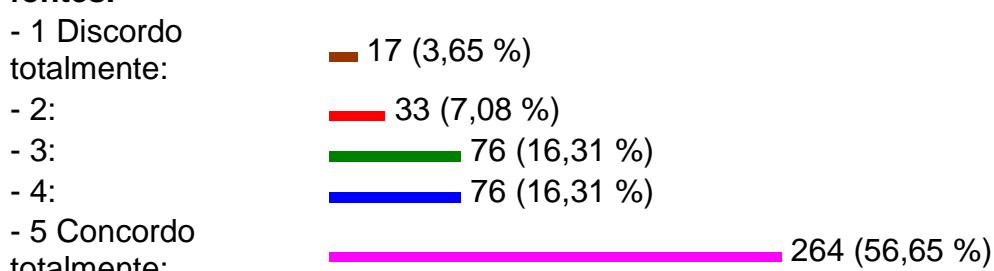
(11) Eu costumo ler os materiais em voz alta para não sofrer distrações



(12) Eu preparam minhas perguntas antes de entrar nos chats e fóruns de discussões.

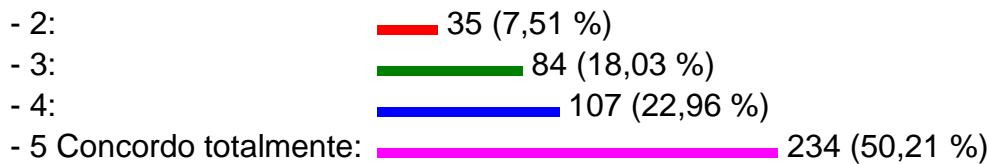


(13) Eu realizo questões extra no meu cursos online, além dos materiais disponibilizados na plataforma eu pesquiso em outras fontes.

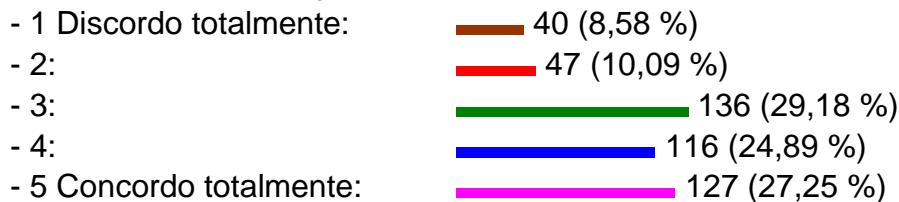


(14) Eu reservo tempo extra para estudar para minhas disciplinas da EAD.

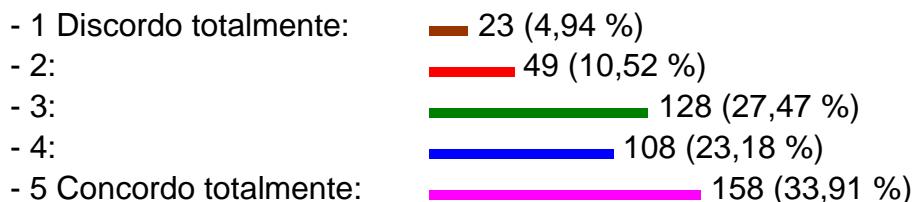




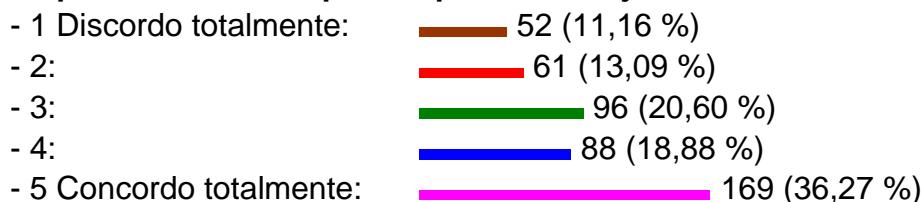
(15) Eu tento agendar o mesmo tempo todos os dias ou todas as semanas para estudar para meus cursos da EAD, e eu cumpro com essa programação.



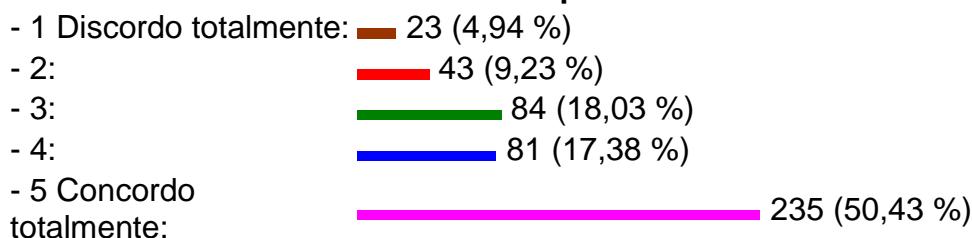
(16) Apesar na modalidade em EAD eu não ter aulas diárias, eu tentar distribuir meu tempo estudando uniformemente todos os dias.



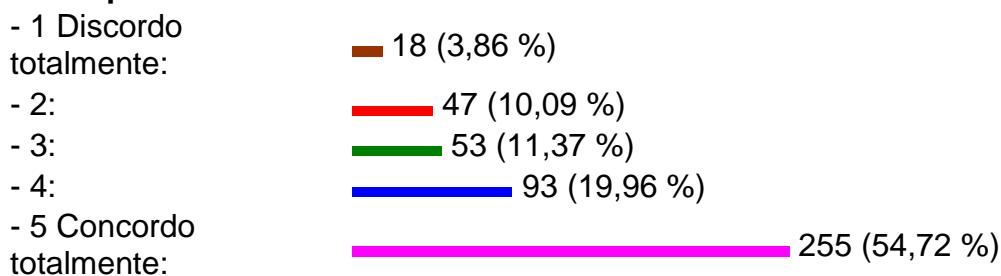
(17) Eu procuro alguém que sabe do conteúdo do curso para que eu possa consultar quando preciso de ajuda.



(18) Eu compartilho meus problemas com os meus colegas para descobrir como resolver os nossos problemas.



(19) Quando necessário, eu tento encontrar meus colegas de curso presencialmente.



(20) Eu costumo pedir ajuda ao tutor ou professor através de e-mail ou mensagem via plataforma

- 1 Discordo totalmente:  16 (3,43 %)
- 2:  45 (9,66 %)
- 3:  83 (17,81 %)
- 4:  100 (21,46 %)
- 5 Concordo totalmente:  222 (47,64 %)

(21) Eu costumo fazer um resumo do meu aprendizado para refletir a minha compreensão do que eu aprendi na disciplina.

- 1 Discordo totalmente:  28 (6,01 %)
- 2:  46 (9,87 %)
- 3:  119 (25,54 %)
- 4:  107 (22,96 %)
- 5 Concordo totalmente:  166 (35,62 %)

(22) Faço questionamentos sobre o material do curso online.

- 1 Discordo totalmente:  39 (8,37 %)
- 2:  73 (15,67 %)
- 3:  132 (28,33 %)
- 4:  96 (20,60 %)
- 5 Concordo totalmente:  126 (27,04 %)

(23) Me comunico com meus colegas para refletir como está meu andamento no curso.

- 1 Discordo totalmente:  29 (6,22 %)
- 2:  47 (10,09 %)
- 3:  93 (19,96 %)
- 4:  86 (18,45 %)
- 5 Concordo totalmente:  211 (45,28 %)

(24) Me comunico com meus colegas para fazer um comparativo entre o que estou aprendendo e o que eles estão aprendendo.

- 1 Discordo totalmente:  40 (8,58 %)
- 2:  64 (13,73 %)
- 3:  104 (22,32 %)
- 4:  95 (20,39 %)
- 5 Concordo totalmente:  163 (34,98 %)

Apêndice

C

Panorama de Bases de Dados Coletadas

Bases de dados coletadas de 2010 até 2016.

Administração	2010.1	2010.2	2011.1	2011.2	2013.2	2014.2	2015.1	2015.2	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	8º Per.				
2ª Oferta					3º Per.	5º Per.	6º Per.	7º Per.	8º Per.
3ª Oferta						1º Per.	2º Per.	3º Per.	4º Per.

Pedagogia	2010.2	2011.1	2011.2	2012.1	2013.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	6º Per.	7º Per.	8º Per.			
2ª Oferta					2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
3ª Oferta								1º Per.	2º Per.	4º Per.

Biologia	2011.1	2011.2	2012.1	2012.2	2013.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta										
2ª Oferta										
3ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	7º Per.	8º Per.		
4ª Oferta				1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
5ª Oferta								1º Per.	2º Per.	4º Per.

Letras	2009.2	2010.1	2010.2	2011.1	2011.2	2012.1	2013.2	2014.1	2014.2	2015.1	2016.1
1ª Oferta	1º Per.	2º Per.	3º Per.	4º Per.	5º Per.	6º Per.					
Reoferta (1ª)				1º Per.	2º Per.	3º Per.					
2ª Oferta				1º Per.	2º Per.	3º Per.	6º Per.	7º Per.	8º Per.		
3ª Oferta							3º Per.	4º Per.	5º Per.	6º Per.	8º Per.
3ª Oferta									1º Per.	2º Per.	4º Per.



Documentação de Scripts SQL

Scripts desenvolvidos para a extração das variáveis de autorregulação da aprendizagem referente a base de dados do MOODLE. A extração total foi dividida em quatro scripts:

1. **Script Base:** Lista de todos os alunos da base de dados
2. **Views:** Cria uma lista de views de perfis e disciplinas
3. **Select SRL:** Seleciona as variáveis de SRL
4. **Select Notas:** Seleciona as notas dos alunos referente ao fórum, atividades e provas.

1. Documentação de script a tabela base (lista de disciplinas e alunos)

```
/*
Neste script é feita a extração de categorias de cursos, disciplinas e
lista de alunos */

SELECT u.id, u.username, u.firstname, u.lastname, u.email, c.id AS
courseid, c.fullname
FROM mdl_role_assignments rs
INNER JOIN mdl_context e ON rs.contextid=e.id
INNER JOIN mdl_course c ON c.id = e.instanceid
INNER JOIN mdl_user u ON u.id=rs.userid
WHERE e.contextlevel=50 AND rs.roleid=5 AND c.id=5
ORDER BY u.firstname
```

2. Documentação de script de views

```
/*
Neste script são realizadas as seguintes tarefas:
- Criar a view 'alunos'
- Criar view 'disciplinas'
- criar a view 'professores'
- Criar a view 'tutores'
- Criar view 'posts'
- Criar view 'topicos'
- Criar view 'atividades nota'
```

```

    - Criar view 'num_foruns_nota'
    - Criar view 'provas'
    - Criar view 'id_alunos'
    - Criar view 'id_disciplinas'
    - Criar view 'log_reduzido'
 */

-- Criar a view 'alunos'
CREATE OR REPLACE VIEW alunos AS
SELECT c.id AS disciplina_id, u.id AS aluno_id
FROM mdl_role_assignments rs
INNER JOIN mdl_context e ON rs.contextid=e.id
INNER JOIN mdl_course c ON c.id = e.instanceid
INNER JOIN mdl_user u ON u.id=rs.userid
WHERE e.contextlevel=50 AND rs.roleid=5
ORDER BY c.id, u.id;

-- Criar view 'disciplinas'
CREATE OR REPLACE VIEW disciplinas AS
SELECT distinct(disciplina_id), data_inicio, data_final FROM base
ORDER BY disciplina_id;

-- Criar a view 'professores'
CREATE OR REPLACE VIEW professores AS
SELECT c.id AS disciplina_id, u.id AS professor_id
FROM mdl_role_assignments rs
INNER JOIN mdl_context e ON rs.contextid=e.id
INNER JOIN mdl_course c ON c.id = e.instanceid
INNER JOIN mdl_user u ON u.id=rs.userid
WHERE e.contextlevel=50 AND rs.roleid IN (3,19)
ORDER BY c.id, u.id;

-- Criar a view 'tutores'
CREATE OR REPLACE VIEW tutores AS
SELECT c.id AS disciplina_id, u.id AS tutor_id
FROM mdl_role_assignments rs
INNER JOIN mdl_context e ON rs.contextid=e.id
INNER JOIN mdl_course c ON c.id = e.instanceid
INNER JOIN mdl_user u ON u.id=rs.userid
WHERE e.contextlevel=50 AND rs.roleid IN (11,12)
ORDER BY c.id, u.id;

-- Criar a view 'posts'
CREATE OR REPLACE VIEW posts AS
SELECT fd.course AS disciplina_id, p2.created AS data, f.name AS nome_forum, p2.id AS post, p2.parent, p2.userid AS emissor, pl.userid AS receptor
FROM mdl_forum_posts p1
INNER JOIN mdl_forum_posts p2 ON p1.id=p2.parent
INNER JOIN mdl_forum_discussions fd ON p1.discussion=fd.id
INNER JOIN mdl_forum f ON fd.forum=f.id
ORDER BY fd.course, p2.userid;

-- Criar view 'topicos'
CREATE OR REPLACE VIEW topicos_temp AS
SELECT f.course AS disciplina_id, f.id AS id_forum, f.name AS nome_forum, d.id AS id_topico, d.firstpost, pl.created AS inicio_topico
FROM mdl_forum f
INNER JOIN mdl_forum_discussions d ON f.id=d.forum
INNER JOIN mdl_forum_posts pl ON d.id=pl.discussion

```

```

WHERE pl.parent=0 AND (f.name ILIKE '%temático%' OR f.name ILIKE '%Temático%' OR f.name ILIKE '%Tematico%' OR f.name ILIKE '%tematico%')
ORDER BY disciplina_id,id_forum,id_topico;

CREATE OR REPLACE VIEW topicos AS
SELECT a.disciplina_id,p2.userid AS
id_usuario,a.id_forum,a.nome_forum,a.id_topico,a.inicio_topico,
p2.created AS data_post,(p2.created - a.inicio_topico) AS dif
FROM topicos_temp a
INNER JOIN mdl_forum_posts p2 ON p2.discussion =a.id_topico WHERE
p2.parent>0
ORDER BY disciplina_id,id_usuario,id_forum,id_topico,data_post;

-- Criar view 'id_alunos'
CREATE OR REPLACE VIEW id_alunos AS
SELECT distinct(aluno_id) FROM base
ORDER BY aluno_id;

-- Criar view 'id_disciplinas'
CREATE OR REPLACE VIEW id_disciplinas AS
SELECT distinct(disciplina_id) FROM base
ORDER BY disciplina_id;

-- Criar view 'log_reduzido'
CREATE OR REPLACE VIEW log_reduzido AS
SELECT ( row_number() OVER(ORDER BY userid,timecreated)) AS id,
timecreated,userid,courseid,component,action
FROM
(SELECT * FROM mdl_logstore_standard_log WHERE "courseid" IN (SELECT *
FROM id_disciplinas) AND "userid" IN (SELECT * FROM id_alunos)
UNION
SELECT * FROM mdl_logstore_standard_log WHERE "action"='loggedout' AND
"userid" IN (SELECT * FROM id_alunos)
UNION
SELECT * FROM mdl_logstore_standard_log WHERE "action"='loggedin' AND
"userid" IN (SELECT * FROM id_alunos)) log;

```

3. Documentação de script que seleciona variáveis de SRL

```

/*
Neste script são realizadas as seguintes tarefas:
- Criar o cabeçalho para o arquivo com o resultado da consulta
- Definir as variáveis que serão utilizadas na análise
- Junções que determinam os valores das variáveis para cada aluno
- Ordenar o resultado por curso, semestre, período, nome da
disciplina e nome do estudante
- Salvar o resultado no arquivo c:/resultado_SRL.csv
*/

-- Criar o cabeçalho para o arquivo com o resultado da consulta
COPY (
SELECT
base.curso AS "Curso",
base.semestre AS "Semestre",
base.periodo AS "Período",
base.disciplina_nome AS "Nome da Disciplina",
base.disciplina_id AS "ID da Disciplina",
base.data_inicio AS "Data de Início",
base.data_final AS "Data de Final",

```

```

base.aluno_nome AS "Nome do Aluno",
base.aluno_id AS "ID do Aluno",
COALESCE(VAR01.VAR01,0) AS "VAR01",
COALESCE(VAR02.VAR02,0) AS "VAR02",
COALESCE(VAR03.VAR03,0) AS "VAR03",
COALESCE(VAR04.VAR04,0) AS "VAR04",
COALESCE(VAR05.VAR05,0) AS "VAR05",
COALESCE(VAR06.VAR06,0) AS "VAR06",
COALESCE(VAR07.VAR07,0) AS "VAR07",
COALESCE(VAR08.VAR08,0) AS "VAR08",
COALESCE(VAR09.VAR09,0) AS "VAR09",
COALESCE(VAR10.VAR10,0) AS "VAR10",
-- COALESCE(VAR11.VAR11,0) AS "VAR11",
-- COALESCE(VAR12.VAR12,0) AS "VAR12",
VAR13.VAR13 AS "VAR13",
VAR14.VAR14 AS "VAR14",
COALESCE(VAR15.VAR15,0) AS "VAR15",
COALESCE(VAR16.VAR16,0) AS "VAR16",
COALESCE(VAR17.VAR17,0) AS "VAR17",
COALESCE(VAR18.VAR18,0) AS "VAR18",
COALESCE(VAR19.VAR19,0) AS "VAR19",
COALESCE(VAR20.VAR20,0) AS "VAR20",
COALESCE(VAR21.VAR21,0) AS "VAR21",
COALESCE(VAR22.VAR22,0) AS "VAR22",
COALESCE(VAR23.VAR23,0) AS "VAR23",
COALESCE(VAR24.VAR24,0) AS "VAR24",
VAR25.VAR25 AS "VAR25",
-- COALESCE(VAR26.VAR26,0) AS "VAR26",
-- COALESCE(VAR27.VAR27,0) AS "VAR27",
COALESCE(VAR28.VAR28,0) AS "VAR28",
COALESCE(VAR17.VAR17,0) AS "VAR29",
COALESCE(VAR17.VAR17,0) AS "VAR30",
COALESCE(VAR31.VAR31,0) AS "VAR31",
-- COALESCE(VAR31a.VAR31a,0) AS "VAR31a",
COALESCE(VAR31b.VAR31b,0) AS "VAR31b",
COALESCE(VAR31c.VAR31c,0) AS "VAR31c",
COALESCE(VAR32a.VAR32a,0) AS "VAR32a",
COALESCE(VAR32b.VAR32b,0) AS "VAR32b",
COALESCE(VAR32c.VAR32c,0) AS "VAR32c",
COALESCE(VAR32d.VAR32d,0) AS "VAR32d",
COALESCE(VAR15.VAR15,0) AS "VAR33",
COALESCE(VAR19.VAR19,0) AS "VAR34",
COALESCE(VAR35.VAR35,0) AS "VAR35"

-- Junções que determinam os valores das variáveis para cada aluno
FROM

(SELECT * FROM base) AS base

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR01 <<<<<<< OK */

    LEFT OUTER JOIN
    (SELECT temp.disciplina_id,temp.aluno_id, count(*) AS "var01"
     FROM (SELECT b.disciplina_id,b.aluno_id, split_part(ip,'.',1)||'.'||split_part(ip,'.',2)||'.'||split_part(ip,'.',3) AS ip_reduzido, count(*) AS Num_Acesso_IP
           FROM (SELECT * FROM mdl_logstore_standard_log WHERE
"action"='loggedin') l
           INNER JOIN base b
           ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio AND
b.data_final

```

```

        GROUP BY b.disciplina_id,b.aluno_id,ip_reduzido) AS temp
        GROUP BY temp.disciplina_id, temp.aluno_id) AS VAR01
        ON VAR01.disciplina_id = base.disciplina_id AND VAR01.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR02 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var02"
    FROM mdl_message_read r
    INNER JOIN base b ON b.aluno_id=r.useridfrom AND r.timecreated
BETWEEN b.data_inicio AND b.data_final
    INNER JOIN professores p ON p.professor_id=r.useridto AND
p.disciplina_id=b.disciplina_id
        WHERE r.subject ILIKE '%Nova mensagem de%'
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR02
        ON VAR02.disciplina_id = base.disciplina_id AND VAR02.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR03 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var03"
    FROM mdl_message_read r
    INNER JOIN base b ON b.aluno_id=r.useridfrom AND r.timecreated
BETWEEN b.data_inicio AND b.data_final
    INNER JOIN tutores p ON p.tutor_id=r.useridto AND
p.disciplina_id=b.disciplina_id
        WHERE r.subject ILIKE '%Nova mensagem de%'
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR03
        ON VAR03.disciplina_id = base.disciplina_id AND VAR03.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR04 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var04"
    FROM mdl_message_read r
    INNER JOIN base b ON b.aluno_id=r.useridfrom AND r.timecreated
BETWEEN b.data_inicio AND b.data_final
        WHERE r.subject ILIKE '%Nova mensagem de%'
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR04
        ON VAR04.disciplina_id = base.disciplina_id AND VAR04.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR05 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var05"
    FROM mdl_message_read r
    INNER JOIN base b ON b.aluno_id=r.useridto AND r.timecreated
BETWEEN b.data_inicio AND b.data_final
        WHERE r.subject ILIKE '%Nova mensagem de%'
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR05
        ON VAR05.disciplina_id = base.disciplina_id AND VAR05.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR06 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var06"
    FROM mdl_forum_posts p
    INNER JOIN mdl_forum_discussions d ON (d.id = p.discussion)
    INNER JOIN mdl_forum f ON d.forum=f.id
    INNER JOIN base b ON b.disciplina_id=d.course AND
b.aluno_id=p.userid AND p.created BETWEEN b.data_inicio AND b.data_final

```

```

    WHERE p.parent=0 AND (f.name ILIKE '%duvida%' OR f.name ILIKE
    '%dúvida%' OR f.name ILIKE '%Dúvida%' OR f.name ILIKE '%Dúvida%')
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR06
        ON VAR06.disciplina_id = base.disciplina_id AND VAR06.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR07 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var07"
    FROM mdl_forum_posts p
    INNER JOIN mdl_forum_discussions d ON (d.id = p.discussion)
    INNER JOIN mdl_forum f ON d.forum=f.id
    INNER JOIN base b ON b.disciplina_id=d.course AND
b.aluno_id=p.userid AND p.created BETWEEN b.data_inicio AND b.data_final
    WHERE f.name ILIKE '%duvida%' OR f.name ILIKE '%dúvida%' OR f.name
ILIKE '%Dúvida%' OR f.name ILIKE '%Dúvida%'
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR07
        ON VAR07.disciplina_id = base.disciplina_id AND VAR07.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR08 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT p1.disciplina_id,emissor, count(*) AS "var08"
    FROM posts p1
    INNER JOIN alunos p2 ON p2.disciplina_id=p1.disciplina_id AND
p2.aluno_id=p1.receptor
    INNER JOIN base b ON b.disciplina_id=p1.disciplina_id AND
b.aluno_id=p1.emissor AND p1.data BETWEEN b.data_inicio AND b.data_final
        GROUP BY p1.disciplina_id, emissor) AS VAR08
        ON VAR08.disciplina_id = base.disciplina_id AND VAR08.emissor =
base.aluno_id

/*>>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR09 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT p1.disciplina_id,receptor, count(*) AS VAR09
    FROM posts p1
    INNER JOIN ((SELECT * FROM professores) UNION (SELECT
disciplina_id,tutor_id AS professor_id FROM tutores)) p2
        ON p2.disciplina_id=p1.disciplina_id AND
p2.professor_id=p1.emissor
    INNER JOIN alunos p3 ON p3.disciplina_id=p1.disciplina_id AND
p3.aluno_id=p1.receptor
    INNER JOIN base b ON b.disciplina_id=p1.disciplina_id AND
b.aluno_id=p1.receptor AND p1.data BETWEEN b.data_inicio AND b.data_final
        GROUP BY p1.disciplina_id, receptor) AS VAR09
        ON VAR09.disciplina_id = base.disciplina_id AND VAR09.receptor =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR10 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT temp.disciplina_id, temp.aluno_id, count(*) AS "var10"
    FROM (SELECT b.disciplina_id, b.aluno_id, r.userid AS
"VAR10_temp"
    FROM mdl_message_read r
    INNER JOIN base b ON b.aluno_id=r.userid AND r.timecreated
BETWEEN b.data_inicio AND b.data_final
    INNER JOIN alunos a ON a.aluno_id=r.userid AND
a.disciplina_id=b.disciplina_id
    WHERE r.subject ILIKE '%Nova mensagem de%' AND
r.userid=b.userid) AS temp
    GROUP BY b.disciplina_id, b.aluno_id,r.userid) AS temp
    
```

```

        GROUP BY temp.disciplina_id, temp.aluno_id) AS VAR10
        ON VAR10.disciplina_id = base.disciplina_id AND VAR10.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR12 <<<<<<<
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var12"
    FROM base b
    INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE "action"
= 'view' AND info = '2') l
        ON b.aluno_id=l.userid AND b.disciplina_id=l.course AND l.time
BETWEEN b.data_inicio AND b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR12
    ON VAR12.aluno_id = base.aluno_id AND VAR12.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR13 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT DISTINCT ON (b.course,a.userid)
b.course,a.userid,COUNT(extract(hour from to_timestamp(a.timecreated))) AS count,
            extract(hour from to_timestamp(a.timecreated)) AS "var13" FROM
mdl_assign_submission a
        INNER JOIN mdl_assign b ON a.assignment=b.id
        INNER JOIN base c ON c.disciplina_id=b.course AND
c.aluno_id=a.userid WHERE a.timecreated BETWEEN c.data_inicio and
c.data_final
        GROUP BY b.course,a.userid, "var13"
        ORDER BY b.course,a.userid ASC, count DESC, "var13" ASC) AS VAR13
    ON VAR13.course = base.disciplina_id AND VAR13.userid =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR14 <<<<<<< OK */
LEFT OUTER JOIN
    (SELECT DISTINCT ON (b.course,a.userid) b.course, a.userid,
COUNT(
            CASE WHEN extract(hour from to_timestamp(a.timecreated)) >=
6 AND extract(hour from to_timestamp(a.timecreated)) < 12 THEN '1-manha'
                WHEN extract(hour from
to_timestamp(a.timecreated)) >= 12 AND extract(hour from
to_timestamp(a.timecreated)) < 18 THEN '2-tarde'
                    WHEN extract(hour from
to_timestamp(a.timecreated)) >= 18 AND extract(hour from
to_timestamp(a.timecreated)) < 24 THEN '3-noite'
                        ELSE '4-madrugada' END) as count,
            CASE WHEN extract(hour from to_timestamp(a.timecreated)) >=
6 AND extract(hour from to_timestamp(a.timecreated)) < 12 THEN '1-manha'
                WHEN extract(hour from
to_timestamp(a.timecreated)) >= 12 AND extract(hour from
to_timestamp(a.timecreated)) < 18 THEN '2-tarde'
                    WHEN extract(hour from
to_timestamp(a.timecreated)) >= 18 AND extract(hour from
to_timestamp(a.timecreated)) < 24 THEN '3-noite'
                        ELSE '4-madrugada' END AS "var14"
            FROM mdl_assign_submission a
        INNER JOIN mdl_assign b ON a.assignment=b.id
        INNER JOIN base c ON c.disciplina_id=b.course AND
c.aluno_id=a.userid WHERE a.timecreated BETWEEN c.data_inicio and
c.data_final
        GROUP BY b.course,a.userid, "var14"
)

```

```

        ORDER BY b.course,a.userid ASC, count DESC,"var14" ASC) AS VAR14
        ON VAR14.course = base.disciplina_id AND VAR14.userid =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR15 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var15"
    FROM mdl_assign a
    INNER JOIN mdl_assign_submission s
    ON a.id=s.assignment AND a.duedate >= s.timecreated
    INNER JOIN base b
    ON b.disciplina_id= a.course AND b.aluno_id=s.userid AND
s.timecreated BETWEEN b.data_inicio AND b.data_final
    GROUP BY b.disciplina_id,b.aluno_id) AS VAR15
    ON VAR15.aluno_id = base.aluno_id AND VAR15.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR16 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var16"
    FROM mdl_assign a
    INNER JOIN mdl_assign_submission s
    ON a.id=s.assignment AND a.duedate <= s.timecreated
    INNER JOIN base b
    ON b.disciplina_id= a.course AND b.aluno_id=s.userid AND
s.timecreated BETWEEN b.data_inicio AND b.data_final
    GROUP BY b.disciplina_id,b.aluno_id) AS VAR16
    ON VAR16.aluno_id = base.aluno_id AND VAR16.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR17 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT temp.course, temp.userid,ROUND(AVG(temp.dif),2) AS "var17"
FROM
    (SELECT b.course,a.userid,a.assignment,(b.duedate -
a.timecreated)/(3600*24) AS dif
    FROM mdl_assign_submission a
    INNER JOIN mdl_assign b ON a.assignment=b.id
    INNER JOIN base c ON c.disciplina_id=b.course AND
c.aluno_id=a.userid
    WHERE a.timecreated < b.duedate AND a.timecreated BETWEEN
c.data_inicio AND c.data_final
    GROUP BY b.course,a.userid, a.assignment,b.duedate,a.timecreated)
AS temp
    GROUP BY temp.course,temp.userid) AS VAR17
    ON VAR17.course = base.disciplina_id AND VAR17.userid =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR18 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var18"
    FROM base b
    INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
("component"='mod_forum' AND "action"='viewed' AND
"target"='course_module')) l
    ON b.disciplina_id=l.courseid AND b.aluno_id=l.userid AND
l.timecreated BETWEEN b.data_inicio AND b.data_final
    GROUP BY b.disciplina_id,b.aluno_id) AS VAR18
    ON VAR18.disciplina_id = base.disciplina_id AND VAR18.aluno_id =
base.aluno_id

```

```

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR19 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id, b.aluno_id, count(*) AS "var19"
    FROM mdl_forum_posts p
    INNER JOIN mdl_forum_discussions d ON d.id = p.discussion
    INNER JOIN base b ON d.course=b.disciplina_id AND
p.userid=b.aluno_id AND p.created BETWEEN b.data_inicio AND b.data_final
        GROUP BY b.disciplina_id, b.aluno_id) AS VAR19
    ON VAR19.disciplina_id = base.disciplina_id AND VAR19.aluno_id =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR20 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT p1.disciplina_id,emissor,count(*) AS "var20"
    FROM posts p1
    INNER JOIN base b ON b.disciplina_id=p1.disciplina_id AND
b.aluno_id=p1.emissor AND p1.data BETWEEN b.data_inicio AND b.data_final
        INNER JOIN alunos p2 ON p2.disciplina_id=p1.disciplina_id AND
p2.aluno_id=p1.receptor AND emissor=receptor
        GROUP BY p1.disciplina_id, emissor) AS VAR20
    ON VAR20.disciplina_id = base.disciplina_id AND VAR20.emissor =
base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR21 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var21"
    FROM base b
    INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
component='mod_assign' AND action='viewed' AND target='grading_table') l
        ON b.aluno_id=l.userid AND b.disciplina_id=l.courseid AND
l.timecreated BETWEEN b.data_inicio AND b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR21
    ON VAR21.aluno_id = base.aluno_id AND VAR21.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR22 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var22"
    FROM base b
    INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
component='mod_checklist' AND action='viewed') l
        ON b.aluno_id=l.userid AND b.disciplina_id=l.courseid AND
l.timecreated BETWEEN b.data_inicio AND b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR22
    ON VAR22.aluno_id = base.aluno_id AND VAR22.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR23 <<<<<< OK */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var23"
    FROM base b
    INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
component='mod_assign' AND action='viewed' AND
target='submission_confirmation_form') l
        ON b.aluno_id=l.userid AND b.disciplina_id=l.courseid AND
l.timecreated BETWEEN b.data_inicio AND b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR23
    ON VAR23.aluno_id = base.aluno_id AND VAR23.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR24 <<<<<< OK */

```

```

LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*)/26 AS "var24"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
action='loggedin') l
        ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio and
b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR24
        ON VAR24.aluno_id = base.aluno_id AND VAR24.disciplina_id =
base.disciplina_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR25 <<<<<< OK*/
LEFT OUTER JOIN
    (SELECT temp.disciplina_id,
temp.id_usuario,ROUND(AVG(temp.tempo_primeiro_post)/(3600*24),2) AS
"var25"      FROM
        (SELECT a.disciplina_id, a.id_usuario, a.id_topico, MIN(a.dif) AS
tempo_primeiro_post
         FROM topicos a
         INNER JOIN base b ON b.disciplina_id=a.disciplina_id AND
b.aluno_id=a.id_usuario AND a.data_post BETWEEN b.data_inicio and
b.data_final
         GROUP BY a.disciplina_id, a.id_usuario, a.id_topico) AS temp
        GROUP BY temp.disciplina_id, temp.id_usuario) AS VAR25
        ON VAR25.disciplina_id = base.disciplina_id AND VAR25.id_usuario =
base.aluno_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR28 <<<<<< */
LEFT OUTER JOIN
    (SELECT login.disciplina_id, login.aluno_id, (num_login -
num_logout) AS "var28" FROM
        (SELECT b1.disciplina_id, b1.aluno_id, count(*) AS num_logout
         FROM base b1
         INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
action='loggedout') l1
            ON b1.aluno_id=l1.userid AND l1.timecreated BETWEEN b1.data_inicio
and b1.data_final
            GROUP BY b1.disciplina_id,b1.aluno_id) AS logout
        INNER JOIN
        (SELECT b2.disciplina_id, b2.aluno_id, count(*) AS "num_login"
         FROM base b2
         INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
action='loggedin') l2
            ON b2.aluno_id=l2.userid AND l2.timecreated BETWEEN b2.data_inicio
and b2.data_final
            GROUP BY b2.disciplina_id,b2.aluno_id) AS login
        ON logout.aluno_id=login.aluno_id AND
logout.disciplina_id=login.disciplina_id) AS VAR28
        ON VAR28.aluno_id = base.aluno_id AND VAR28.disciplina_id =
base.disciplina_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR31 <<<<<< OK*/
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var31"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
action='loggedin') l
        ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio and
b.data_final
        GROUP BY b.disciplina_id,b.aluno_id) AS VAR31

```

```

    ON VAR31.aluno_id = base.aluno_id AND VAR31.disciplina_id =
base.disciplina_id

/*
>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR31b <<<<<< OK */
-- Quantidade de dias distintos que o aluno entrou na disciplina
LEFT OUTER JOIN
    (SELECT disciplina_id, aluno_id, count(*) AS "var31b" FROM
     (SELECT distinct ON(date_trunc('day',
to_timestamp(timecreated)),b.disciplina_id, b.aluno_id) b.disciplina_id,
      b.aluno_id, date_trunc('day', to_timestamp(timecreated)) AS "data"
FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
component='core' AND action='viewed' AND contextlevel>49 AND
target='course') AS l
        ON b.aluno_id=l.userid AND b.disciplina_id=l.courseid AND
l.timecreated BETWEEN b.data_inicio AND b.data_final) AS temp
      GROUP BY disciplina_id, aluno_id) AS VAR31b
    ON VAR31b.aluno_id = base.aluno_id AND VAR31b.disciplina_id =
base.disciplina_id

/*
>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR31c <<<<<< OK */
-- Quantidade de dias distintos que o aluno entrou na plataforma
LEFT OUTER JOIN
    (SELECT disciplina_id, aluno_id, count(*) AS "var31c" FROM
     (Select distinct ON(date_trunc('day', to_timestamp(timecreated)),
b.aluno_id,b.disciplina_id) b.disciplina_id,
      b.aluno_id,date_trunc('day', to_timestamp(timecreated)) AS "data"
FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
action='loggedin') AS l
        ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio AND
b.data_final) AS temp
      GROUP BY disciplina_id, aluno_id) AS VAR31c
    ON VAR31c.aluno_id = base.aluno_id AND VAR31c.disciplina_id =
base.disciplina_id

/*
>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR32 <<<<<< OK TODAS*/
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var32a"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
"action"='loggedin' AND extract(hour from to_timestamp(timecreated)) >= 6
AND extract(hour from to_timestamp(timecreated)) < 12) l
        ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio AND
b.data_final
      GROUP BY b.disciplina_id,b.aluno_id) AS VAR32a
    ON VAR32a.aluno_id = base.aluno_id AND VAR32a.disciplina_id =
base.disciplina_id

LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var32b"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
"action"='loggedin' AND extract(hour from to_timestamp(timecreated)) >=
12 AND extract(hour from to_timestamp(timecreated)) < 18) l
        ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio AND
b.data_final
      GROUP BY b.disciplina_id,b.aluno_id) AS VAR32b
    ON VAR32b.aluno_id = base.aluno_id AND VAR32b.disciplina_id =
base.disciplina_id

```

```

LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var32c"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
"action"='loggedin' AND extract(hour from to_timestamp(timecreated)) >=
18 AND extract(hour from to_timestamp(timecreated)) < 24) l
      ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio and
b.data_final
      GROUP BY b.disciplina_id,b.aluno_id) AS VAR32c
      ON VAR32c.aluno_id = base.aluno_id AND VAR32c.disciplina_id =
base.disciplina_id

LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id, count(*) AS "var32d"
     FROM base b
     INNER JOIN (SELECT * FROM mdl_logstore_standard_log WHERE
"action"='loggedin' AND extract(hour from to_timestamp(timecreated)) >= 0
AND extract(hour from to_timestamp(timecreated)) < 6) l
      ON b.aluno_id=l.userid AND l.timecreated BETWEEN b.data_inicio and
b.data_final
      GROUP BY b.disciplina_id,b.aluno_id) AS VAR32d
      ON VAR32d.aluno_id = base.aluno_id AND VAR32d.disciplina_id =
base.disciplina_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL VAR35 <<<<<<< */
LEFT OUTER JOIN
    (SELECT p1.disciplina_id,receptor, count(*) AS "var35"
     FROM posts p1
     INNER JOIN professores p2 ON p2.disciplina_id=p1.disciplina_id AND
p2.professor_id=p1.emissor
     INNER JOIN alunos p3 ON p3.disciplina_id=p1.disciplina_id AND
p3.aluno_id=p1.receptor
     INNER JOIN base b ON b.disciplina_id=p1.disciplina_id AND
b.aluno_id=p1.receptor AND p1.data BETWEEN b.data_inicio and b.data_final
     WHERE p1.nome_forum ILIKE '%duvida%' OR p1.nome_forum ILIKE
'%dúvida%' OR p1.nome_forum ILIKE '%Duvida%'
     OR p1.nome_forum ILIKE '%Dúvida%' OR p1.nome_forum ILIKE '%DÚVIDA%'
OR p1.nome_forum ILIKE '%DUVIDA%'
     GROUP BY p1.disciplina_id, receptor) AS VAR35
     ON VAR35.disciplina_id = base.disciplina_id AND VAR35.receptor =
base.aluno_id

-- Ordenar o resultado por curso, semestre, período, nome da disciplina e
nome do estudante
ORDER BY
base.curso, base.semestre, base.periodo, base.disciplina_nome,
base.aluno_nome

-- Salvar o resultado no arquivo c:/resultado_SRL.csv
) TO 'c:/variaveis postgres_NEAD2016_SRL.csv' DELIMITER ';' CSV HEADER;

```

4. Documentação de script de coleta de notas

```

/*
Neste script são realizadas as seguintes tarefas:
- Cria views auxiliares
- Criar o cabeçalho para o arquivo com o resultado da consulta
- Definir as variáveis que serão utilizadas na análise

```

```

    - Junções que determinam os valores das variáveis para cada aluno
    - Ordenar o resultado por curso, semestre, período, nome da
disciplina e nome do estudante
    - Salvar o resultado no arquivo c:/notas_SRL.csv
*/

-- Cria views auxiliares
/* >>>>> View prova <<<<<< */
-- DROP VIEW prova01;
CREATE OR REPLACE VIEW prova01 AS
SELECT disciplina_id,id_prova01 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, p.id)) AS id, b.disciplina_id, p.id AS id_prova01
FROM disciplinas b
INNER JOIN mdl_quiz p ON b.disciplina_id=p.course AND p.timeclose
BETWEEN b.data_inicio AND b.data_final
WHERE name not like '%chamada%' AND name not like '%Chamada%' AND
name not like '%Final%' AND name not like '%final%') AS temp
WHERE id=1;

-- DROP VIEW prova02;
CREATE OR REPLACE VIEW prova02 AS
SELECT disciplina_id,id_prova02 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, p.id)) AS id, b.disciplina_id, p.id AS id_prova02
FROM disciplinas b
INNER JOIN mdl_quiz p ON b.disciplina_id=p.course AND p.timeclose
BETWEEN b.data_inicio AND b.data_final
WHERE name not like '%chamada%' AND name not like '%Chamada%' AND
name not like '%Final%' AND name not like '%final%') AS temp
WHERE id=2;

/* >>>>> View segunda_chamada <<<<<< */
-- DROP VIEW segunda_chamada01;
CREATE OR REPLACE VIEW segunda_chamada01 AS
SELECT disciplina_id,id_segunda_chamada01 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, p.id)) AS id, b.disciplina_id, p.id AS
id_segunda_chamada01
FROM disciplinas b
INNER JOIN mdl_quiz p ON b.disciplina_id=p.course AND p.timeclose
BETWEEN b.data_inicio AND b.data_final
WHERE name like '%chamada%' OR name like '%Chamada%') AS temp
WHERE id=1;

-- DROP VIEW segunda_chamada02;
CREATE OR REPLACE VIEW segunda_chamada02 AS
SELECT disciplina_id,id_segunda_chamada02 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, p.id)) AS id, b.disciplina_id, p.id AS
id_segunda_chamada02
FROM disciplinas b
INNER JOIN mdl_quiz p ON b.disciplina_id=p.course AND p.timeclose
BETWEEN b.data_inicio AND b.data_final
WHERE name like '%chamada%' OR name like '%Chamada%') AS temp
WHERE id=2;

/* >>>>> View forums_disciplina <<<<<< */
-- DROP VIEW forum01;
CREATE OR REPLACE VIEW forum01 AS
SELECT disciplina_id,id_forum01 FROM

```

```

(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_forum01
FROM disciplinas b
INNER JOIN mdl_forum f ON b.disciplina_id=f.course AND f.assess timefinish
BETWEEN b.data_inicio AND b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='forum') AS temp
WHERE id=1;

-- DROP VIEW forum02;
CREATE OR REPLACE VIEW forum02 AS
SELECT disciplina_id,id_forum02 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_forum02
FROM disciplinas b
INNER JOIN mdl_forum f ON b.disciplina_id=f.course AND f.assess timefinish
BETWEEN b.data_inicio AND b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='forum') AS temp
WHERE id=2;

-- DROP VIEW forum03;
CREATE OR REPLACE VIEW forum03 AS
SELECT disciplina_id,id_forum03 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_forum03
FROM disciplinas b
INNER JOIN mdl_forum f ON b.disciplina_id=f.course AND f.assess timefinish
BETWEEN b.data_inicio AND b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='forum') AS temp
WHERE id=3;

-- DROP VIEW forum04;
CREATE OR REPLACE VIEW forum04 AS
SELECT disciplina_id,id_forum04 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_forum04
FROM disciplinas b
INNER JOIN mdl_forum f ON b.disciplina_id=f.course AND f.assess timefinish
BETWEEN b.data_inicio AND b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='forum') AS temp
WHERE id=4;

/* >>>>> View webquest <<<<<< */
-- DROP VIEW webquest01;
CREATE OR REPLACE VIEW webquest01 AS
SELECT disciplina_id,id_webquest01 FROM
(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_webquest01
FROM disciplinas b
INNER JOIN mdl_assign f ON b.disciplina_id=f.course AND f.duedate BETWEEN
b.data_inicio AND b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='assign') AS temp
WHERE id=1;

-- DROP VIEW webquest02;
CREATE OR REPLACE VIEW webquest02 AS
SELECT disciplina_id,id_webquest02 FROM

```

```

(SELECT (row_number() OVER(PARTITION BY b.disciplina_id ORDER BY
b.disciplina_id, f.id)) AS id, b.disciplina_id, f.id AS id_webquest02
FROM disciplinas b
INNER JOIN mdl_assign f ON b.disciplina_id=f.course AND f.duedate BETWEEN
b.data_inicio and b.data_final
INNER JOIN mdl_grade_items i ON i.iteminstance=f.id WHERE
i.itemmodule='assign') AS temp
WHERE id=2;

-- Criar o cabeçalho para o arquivo com o resultado da consulta
COPY (
SELECT
base.curso AS "Curso",
base.semestre AS "Semestre",
base.periodo AS "Período",
base.disciplina_nome AS "Nome da Disciplina",
base.disciplina_id AS "ID da Disciplina",
base.data_inicio AS "Data de Início",
base.data_final AS "Data de Final",
base.aluno_nome AS "Nome do Aluno",
base.aluno_id AS "ID do Aluno",
COALESCE(PROVA01.PROVA01,0) AS "PROVA01",
COALESCE(PROVA01_2CHAMADA.PROVA01_2CHAMADA,0) AS "PROVA01_2CHAMADA",
COALESCE(PROVA02.PROVA02,0) AS "PROVA02",
COALESCE(PROVA02_2CHAMADA.PROVA02_2CHAMADA,0) AS "PROVA02_2CHAMADA",
COALESCE(FORUM01.FORUM01,0) AS "FORUM01",
COALESCE(FORUM02.FORUM02,0) AS "FORUM02",
COALESCE(FORUM03.FORUM03,0) AS "FORUM03",
COALESCE(FORUM04.FORUM04,0) AS "FORUM04",
COALESCE(WEBQUEST01.WEBQUEST01,0) AS "WEBQUEST01",
COALESCE(WEBQUEST02.WEBQUEST02,0) AS "WEBQUEST02"

-- Junções que determinam os valores das variáveis para cada aluno
FROM

(SELECT * FROM base) AS base

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL PROVA01 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"prova01"
    FROM base b
    INNER JOIN prova01 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE
"itemmodule"='quiz') i ON i.iteminstance=f.id_prova01
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
    ORDER BY b.disciplina_id,b.aluno_id) AS PROVA01
    ON PROVA01.disciplina_id = base.disciplina_id AND PROVA01.aluno_id
= base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL PROVA01_2CHAMADA <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"prova01_2chamada"
    FROM base b
    INNER JOIN segunda_chamada01 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='quiz')
i ON i.iteminstance=f.id_segunda_chamada01
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id

```

```

        ORDER BY b.disciplina_id,b.aluno_id) AS PROVA01_2CHAMADA
        ON PROVA01_2CHAMADA.disciplina_id = base.disciplina_id AND
PROVA01_2CHAMADA.aluno_id = base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL PROVA02 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"prova02"
    FROM base b
    INNER JOIN prova02 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='quiz')
i ON i.iteminstance=f.id_prova02
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS PROVA02
        ON PROVA02.disciplina_id = base.disciplina_id AND PROVA02.aluno_id
= base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL PROVA02_2CHAMADA <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"prova02_2chamada"
    FROM base b
    INNER JOIN segunda_chamada02 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='quiz')
i ON i.iteminstance=f.id_segunda_chamada02
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS PROVA02_2CHAMADA
        ON PROVA02_2CHAMADA.disciplina_id = base.disciplina_id AND
PROVA02_2CHAMADA.aluno_id = base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL FORUM01 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"forum01"
    FROM base b
    INNER JOIN forum01 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='forum')
i ON i.iteminstance=f.id_forum01
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS FORUM01
        ON FORUM01.disciplina_id = base.disciplina_id AND FORUM01.aluno_id
= base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL FORUM02 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"forum02"
    FROM base b
    INNER JOIN forum02 f ON b.disciplina_id=f.disciplina_id
    INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='forum')
i ON i.iteminstance=f.id_forum02
    INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS FORUM02
        ON FORUM02.disciplina_id = base.disciplina_id AND FORUM02.aluno_id
= base.aluno_id

/* >>>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL FORUM03 <<<<<< */

```

```

LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"forum03"
        FROM base b
        INNER JOIN forum03 f ON b.disciplina_id=f.disciplina_id
        INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='forum')
i ON i.iteminstance=f.id_forum03
        INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS FORUM03
        ON FORUM03.disciplina_id = base.disciplina_id AND FORUM03.aluno_id
= base.aluno_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL FORUM04 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"forum04"
        FROM base b
        INNER JOIN forum04 f ON b.disciplina_id=f.disciplina_id
        INNER JOIN (Select * FROM mdl_grade_items WHERE itemmodule='forum')
i ON i.iteminstance=f.id_forum04
        INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS FORUM04
        ON FORUM04.disciplina_id = base.disciplina_id AND FORUM04.aluno_id
= base.aluno_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL WEBQUEST01 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"webquest01"
        FROM base b
        INNER JOIN webquest01 f ON b.disciplina_id=f.disciplina_id
        INNER JOIN (Select * FROM mdl_grade_items WHERE
itemmodule='assign') i ON i.iteminstance=f.id_webquest01
        INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS WEBQUEST01
        ON WEBQUEST01.disciplina_id = base.disciplina_id AND
WEBQUEST01.aluno_id = base.aluno_id

/* >>>>> SCRIPT DE EXTRAÇÃO DA VARIÁVEL WEBQUEST02 <<<<<< */
LEFT OUTER JOIN
    (SELECT b.disciplina_id,b.aluno_id,round(g.finalgrade,2) AS
"webquest02"
        FROM base b
        INNER JOIN webquest02 f ON b.disciplina_id=f.disciplina_id
        INNER JOIN (Select * FROM mdl_grade_items WHERE
itemmodule='assign') i ON i.iteminstance=f.id_webquest02
        INNER JOIN mdl_grade_grades g ON g.itemid=i.id AND
g.userid=b.aluno_id
        ORDER BY b.disciplina_id,b.aluno_id) AS WEBQUEST02
        ON WEBQUEST02.disciplina_id = base.disciplina_id AND
WEBQUEST02.aluno_id = base.aluno_id

-- Ordenar o resultado por curso, semestre, período, nome da disciplina e
nome do estudante
ORDER BY
base.curso, base.semestre, base.periodo, base.disciplina_nome,
base.aluno_nome

```

```
-- Salvar o resultado no arquivo c:/notas_SRL.csv
)
TO 'c:/notas_postgres_SRL.csv' DELIMITER ';' CSV HEADER;
```

E

Script de Análise Fatorial Confirmatória

Scripts desenvolvidos para a análise fatorial confirmatória referente a aplicação do questionário. Este script valida os itens do questionário OSLQ. Os resultados desta análise foram publicados em:

- Rodrigues, R. L., Ramos, J. L. C., Silva, J. C. S., Gomes, A. S., Fonseca, J. A. V., & de Souza, F. D. F. (2016) Validação de um instrumento de mensuração de autorregulação da aprendizagem em contexto brasileiro usando análise fatorial confirmatória. *RENOTE*, 14(1).

Script de Análise Fatorial Confirmatória em R

```
# carregando os pacotes que necessários para a análise
library(lavaan)
library(semPlot)

# importando o banco de dados
dados <- read.table("dados.txt", head=T)

# especificando o modelo (variáveis latentes e variáveis observadas)
model <- '
fator1 =~ v1 + v2 + v3 + v4
fator2 =~ v5 + v6 + v7 + v8 + v9
fator3 =~ v10 + v11 + v12
fator4 =~ v13 + v14 + v15 + v16
fator5 =~ v17 + v18 + v19 + v20
fator6 =~ v21 + v22 + v23 + v24
'
# construindo o modelo de análise fatorial
fit <- cfa(model, data=dados)

# sumário do cálculo de todos os parâmetros de ajuste
summary(fit, fit.measures=TRUE)
parameterEstimates(fit)

# plotando o diagrama através do pacote semPlot
```

```
semPaths(fit,"std", rotation = 2, layout = "tree3", nCharNodes = 0,
sizeLat = 12, sizeLat2 = 7,
curvePivot = TRUE, edge.label.cex=1.2, residuals = F)

# salvando a imagem do diagrama de equação estrutural (diagrama SEM)
dev.print(png, "fig_1_measurement.png", width=6, height=4, res=300,
units="in")
```



Script Estatísticas Descritivas

Este documento refere-se ao script utilizado para gerar a sumarização nos dados utilizados neste tese. Mostra o script de import da base de dados, bem como o script de sumarização e a tabela geral sumarizada.

Script de Estatísticas Descritivas

```
#Carregando dados
dados <- data.frame(read.csv2("baseGeral.csv", header = TRUE, sep = ";",
dec = ","))
dados <- na.exclude(dados[,c(10:46,62)])
nomevariaveis <- colnames(dados)
```

Script tabela sumário

O código mostra o script construído para gerar a tabela de sumarização das variáveis:

```
# Gerando Tabela Sumário
linhas <- length(dados)
matrizsumario <- matrix(NA, nrow=linhas, ncol=7)
dimnames(matrizsumario) = (list(nomevariaveis, c("Mín.", "1º Qt.", "Media
na", "Média", "Variância","3º Qt.","Máx.")))
for (i in 1:linhas) {matrizsumario[i,] <- round(c(min(dados[,i]), quantil
e(dados[, i], 0.25), median(dados[, i]), mean(dados[, i]), var(dados[, i]
),quantile(dados[, i], 0.75), max(dados[, i])),2)}
```

Tabela sumário

A tabela abaixo mostra os resultados da sumarização de todas as variáveis:

	Mín.	1º Qt.	Mediana	Média	Variância	3º Qt.	Máx.
## VAR01	0	4.00	8.00	15.77	479.57	21.00	413.00
## VAR02	0	0.00	0.00	0.20	1.03	0.00	51.00
## VAR03	0	0.00	0.00	2.30	60.48	1.00	210.00
## VAR04	0	0.00	3.00	14.62	2338.49	15.00	2139.00
## VAR05	0	4.00	22.00	39.02	3711.30	52.00	2011.00
## VAR06	0	0.00	0.00	0.03	0.06	0.00	14.00

## VAR07	0	0.00	0.00	0.06	0.31	0.00	60.00
## VAR08	0	0.00	0.00	0.59	3.94	0.00	229.00
## VAR09	0	0.00	0.00	0.98	2.80	1.00	28.00
## VAR10	0	0.00	0.00	1.14	8.91	1.00	96.00
## VAR12	0	0.00	0.00	0.15	1.67	0.00	85.00
## VAR13	0	0.00	12.00	10.34	63.20	17.00	23.00
## VAR14	0	0.00	2.00	1.51	1.40	3.00	4.00
## VAR15	0	0.00	1.00	1.16	0.83	2.00	4.00
## VAR16	0	0.00	0.00	0.10	0.15	0.00	5.00
## VAR17	0	0.00	1.16	8.59	177.90	12.00	81.37
## VAR18	0	2.00	13.00	18.32	542.35	24.00	680.00
## VAR19	0	1.00	3.00	3.10	10.04	4.00	247.00
## VAR20	0	0.00	0.00	0.05	0.13	0.00	14.00
## VAR21	0	0.00	0.00	1.00	3.45	2.00	63.00
## VAR22	0	0.00	0.00	0.61	10.13	0.00	105.00
## VAR23	0	0.00	0.00	0.62	1.69	1.00	81.00
## VAR24	0	0.00	2.00	2.94	10.21	4.00	26.00
## VAR25	0	0.00	68.76	69.16	4229.69	82.96	288.98
## VAR28	7	5.00	23.00	41.91	3208.67	56.00	561.00
## VAR29	0	0.00	1.16	8.59	177.90	12.00	81.37
## VAR30	0	0.00	1.16	8.59	177.90	12.00	81.37
## VAR31	0	24.00	68.00	84.89	7098.66	120.00	692.00
## VAR31b	0	0.00	7.00	11.77	213.22	19.00	125.00
## VAR31c	0	16.00	40.00	43.50	1185.85	64.00	178.00
## VAR32a	0	5.00	14.00	22.48	783.23	30.00	299.00
## VAR32b	0	6.00	21.00	29.22	1098.76	41.00	378.00
## VAR32c	0	6.00	23.00	31.36	1090.49	46.00	350.00
## VAR32d	0	0.00	0.00	1.83	20.61	2.00	77.00
## VAR33	0	0.00	1.00	1.16	0.83	2.00	4.00
## VAR34	0	1.00	3.00	3.10	10.04	4.00	247.00
## VAR35	0	0.00	0.00	0.04	0.09	0.00	14.00
## DESEMPENHO	0	1.25	4.50	4.29			



Scripts R para algoritmos de classificação

Apendice contendo os quatro scripts de classificação em R referente aos seguintes algoritmos: Regressão Logística, Árvore de Decisão, Naive Bayes e SVM.

Script de Regressão Logística em R

```

library(caret) # Para a Matriz de Confusão, Acurácia, Sensibilidade e Especificidade
library(ROCR) # Para a Curva ROC

# Importando e tratando a base de dados

dados <- na.exclude(read.csv2("baseGeral.csv", header = T))

# Separação dos dados de treinamento e de teste com o pacote caret
split <- createDataPartition(y = dados$VAR01, p = 0.7, list = FALSE)
treinamento <- dados[split,]
teste <- dados[-split,]

# Construção do modelo de Regressão Logística
#"cv" = cross-validation, 10-fold
tc <- trainControl("cv", 10, savePredictions=T)
modelo <-
train(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+
      VAR10+VAR12+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
      VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
      VAR32d+VAR33+VAR34+VAR35,data=treinamento, method = "rf",
      family=binomial, trControl = tc)

# Sumário do modelo inicial com todas as variáveis
summary(modeloInicial)

# Escolhendo as variáveis mais significativas através do método Stepwise
modelo = step(modeloInicial)

# Sumário do modelo
summary(modelo)

# Aplicação dos dados de teste para avaliação do modelo

```

```

classificacaoProb <- predict(modelo, newdata = teste)
classificacaoBinaria <- ifelse(classificacaoProb > 0.5, 1, 0)

# Geração da Matriz de confusão e métricas para a análise do modelo
MatrizDeConfusao<-
confusionMatrix(data=classificacaoBinaria, reference=teste$DESEMPENHOBINARIO, positive = "1")
print(MatrizDeConfusao)

# Curva ROC
FG <- prediction(classificacaoProb, teste$DESEMPENHOBINARIO)
pFG <- performance(FG, measure = "tpr", x.measure = "fpr")
plot(pFG,col="blue",lwd=2,main="ROC Curve for Logistic")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
aucFG <- performance(FG, measure = "auc")
aucFG <- aucFG@y.values[[1]]
aucFG

```

Script do algoritmo de Árvore de Decisão

```

library(rpart) # Pacote para o algoritmo árvore de decisão
library(caret) # Pacote para construção da matriz de confusão
library(ROCR) # Para a Curva ROC

# carregando a base de dados
dados <- na.exclude(read.csv2("baseGeral.csv", header=T))
# Separação dos dados de treinamento e de teste com o pacote caret
split <- createDataPartition(y = dados$VAR01, p = 0.7, list = FALSE)
treinamento <- dados[split,]
teste <- dados[-split,]

# contruindo o modelo
modeloArvoreDecisao<- rpart
(DESEMPENHOBINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35, data=treinamento)

# sumarizando o modelo
summary(modeloArvoreDecisao)
# construindo o vetor de predições
classificacaoProb <- predict(modeloArvoreDecisao,teste)
classificacaoBinaria <- ifelse(classificacaoProb > 0.5, 1, 0)

# Gerando a matriz de confusão
MatrizDeConfusao<-
confusionMatrix(data=classificacaoBinaria, reference=teste$DESEMPENHOBINARIO, positive = "1")
print(MatrizDeConfusao)

# Curva ROC
pred <- prediction(classificacaoProb, teste$DESEMPENHOBINARIO)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf,col="blue",lwd=2,main="ROC Curve for Decision Tree")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
auc

```

Script do algoritmo Naive Bayes

```

library(caret)           # Pacote para gerar matriz de confusão
library(ROCR)            # Para a Curva ROC
library(dplyr)           # Pacote para manipulação de dados
library(e1071)           # Pacote com o algoritmo Naive Bayes
# carregando a base de dados
dados <- na.exclude(read.csv2("baseGeral.csv", header=T))

# Separação dos dados de treinamento e de teste com o pacote caret
split <- createDataPartition(y = dados$VAR01, p = 0.7, list = FALSE)
treinamento <- dados[split,]
teste <- dados[-split,]

# construção do modelo
modeloNaiveBayes<-
naiveBayes(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+V
AR08+VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35, data=treinamento)

# sumarizando o modelo
summary(modeloNaiveBayes)
# construindo o vetor de predições e adicionando a base de teste
classificacaoProb <- predict(modeloNaiveBayes, teste, type = "raw")
classificacaoProb <- classificacaoProb[,2]
classificacaoBinaria <- ifelse(classificacaoProb > 0.5, 1, 0)

# Geração da Matriz de confusão
MatrizDeConfusao<-
confusionMatrix(data=classificacaoBinaria, reference=teste$DESEMPENHO_BINARIO, positive = "1")
print(MatrizDeConfusao)

# Curva ROC
pred <- prediction(classificacaoProb, teste$DESEMPENHO_BINARIO)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf, col="blue", lwd=2, main="ROC Curve for Naive Bayes")
abline(a=0, b=1, lwd=2, lty=2, col="gray")
auc <- performance(pred, measure ="auc")
auc <- auc@y.values[[1]]
auc

```

Script do algoritmo SVM

```

library(kernlab)    # Pacote para o algoritmo SVM
library(caret)      # Pacote para a construção da matriz de confusão
library(ROCR)       # Para a Curva ROC

# carregando a base de dados
dados <- na.exclude(read.csv2("baseGeral.csv", header=T))

# Separação dos dados de treinamento e de teste com o pacote caret
split <- createDataPartition(y = dados$VAR01, p = 0.7, list = FALSE)
treinamento <- dados[split,]
teste <- dados[-split,]

# Construindo o modelo SVM
modeloSVM<-
ksvm(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35, data=treinamento)

```

```
# summarizando o modelo
summary(modeloSVM)

# construindo o vetor de predições e adicionando a base de teste
classificacaoProb <- predict(modeloSVM,newdata=teste,type="response")
classificacaoBinaria <- ifelse(classificacaoProb > 0.5,1,0)

# Geração da Matriz de confusão
MatrizDeConfusao<-
confusionMatrix(data=classificacaoBinaria,reference=teste$DESEMPENHO_BINARIO,positive = "1")
print(MatrizDeConfusao)

# Curva ROC
pred <- prediction(classificacaoProb, teste$DESEMPENHO_BINARIO)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf,col="blue",lwd=2,main="ROC Curve for SVM")
abline(a=0,b=1,lwd=2,lty=2,col="gray")
auc <- performance(pred, measure ="auc")
auc <- auc@y.values[[1]]
auc
```



Tabelas comparativas da Regressão Logística

Quadros de comparativo entre os resultados do algoritmo de regressão logística por curso e semestre.

Biologia	Acurácia	Curva ROC
1º Semestre	0.9159	0.9720033
2º Semestre	0.9362	0.9464286
3º Semestre	0.8457	0.8973039
4º Semestre	0.8676	0.9476584
5º Semestre	0.9184	0.9442446
6º Semestre	0.9773	0.9827586
7º Semestre	0.7872	0.7104072
8º Semestre	0.9298	0.9626249
Todos	0.8747	0.9408973

Pedagogia	Acurácia	Curva ROC
1º Semestre	0.8847	0.9294992
2º Semestre	0.9517	0.9844569
3º Semestre	0.9175	0.9693735
4º Semestre	0.8485	0.9249766
5º Semestre	0.9548	0.9785928
6º Semestre	-	-
7º Semestre	0.9327	0.9460966
8º Semestre	0.8981	0.9592837
Todos	0.8763	0.9396812

Letras	Acurácia	Curva ROC
1º Semestre	0.913	0.9393434
2º Semestre	0.7867	0.9097331
3º Semestre	0.8881	0.9404314
4º Semestre	0.9066	0.9394536
5º Semestre	0.829	0.9136971
6º Semestre	0.8721	0.9217949
7º Semestre	0.8148	0.65
8º Semestre	0.8900	0.9265734
Todos	0.8506	0.9132089

Administração	Acurácia	Curva ROC
1º Semestre	0.8986	0.9672446
2º Semestre	0.886	0.9458937
3º Semestre	0.8913	0.9594542
4º Semestre	0.8986	0.8964286
5º Semestre	0.913	0.9222222
6º Semestre	-	-
7º Semestre	-	-
8º Semestre	0.8261	0.9213675
Todos	0.8717	0.9345426

Todos	Acurácia	Curva ROC
1º Semestre	0.8772	0.9447889
2º Semestre	0.9429	0.9677876

Cursos	Acurácia	Curva ROC
Biologia	0.8676	0.9302967
Pedagogia	0.8735	0.9414979

3º Semestre	0.8744	0.9250426
4º Semestre	0.8439	0.9266285
5º Semestre	0.8573	0.9182223
6º Semestre	0.9237	0.9607981
7º Semestre	0.9311	0.968108
8º Semestre	0.893	0.9465806
Todos	0.8642	0.932377

Letras	0.841	0.8985874
Administração	0.8936	0.9567689
Todos	0.8642	0.932377

Comparativo entre classificadores

M1: Comparação entre classificadores binários, onde a binalização foi feita 1 (alunos com notas abaixo de 4,0) e 0 (alunos com notas acima de 4,0)

Métricas de Avaliação (M1)	Regressão Logística	Árvore de Decisão	SVM	Naive Bayes
Accuracy	0,893	0,8622	0,8717	0,7312
No Information Rate	0,5561	0,5572	0,5634	0,5596
P-Value [Acc > NIR]	< 2,2e-16	< 2,2e-16	<2e-16	< 2,2e-16
Kappa	0,7825	0,7197	0,7388	0,4755
McNemar's Test P-value	3,44E-02	0,000119	0,197	< 2,2e-16
Sensitivity	0,8633	0,8273	0,8474	0,8815
Specificity	0,9167	0,8899	0,8905	0,613
Pos Pred Value	0,8921	0,8566	0,8571	0,6419
Neg Pred Value	0,8936	0,8664	0,8827	0,868
Prevalence	0,4439	0,4428	0,4366	0,4404
Detection Rate	0,3832	0,3663	0,37	0,3882
Detection Prevalence	0,4295	0,4277	0,4316	0,6048
Balanced Accuracy	0,8900	0,8586	0,8689	0,7472
ROC	0,9574502	0,915771	0,9291459	0,8850649

M2: Comparação entre classificadores binários, onde a binalização foi feita 1 (alunos com notas abaixo de 7,0) e 0 (alunos com notas acima de 7,0)

Métricas de Avaliação	Regressão Logística	Árvore de Decisão	SVM	Naive Bayes
Accuracy	0,8124	0,8321	0,8332	0,763
No Information Rate	0,7674	0,7595	0,7605	0,7634
P-Value [ACC > NIR]	< 2,2e-16	< 2e-16	< 2,2e-16	0,545132
Kappa	0,4303	0,5335	0,5174	0,3272
McNemar's Test P-value	< 2,2e-16	0,01713	< 2,2e-16	0,000963
Sensitivity	0,9116	0,8963	0,9138	0,8559

Specificity	0,4853	0,6294	0,5772	0,4632
Pos Pred Value	0,8539	0,8842	0,8728	0,8373
Neg Pred Value	0,6245	0,6577	0,6784	0,499
Prevalence	0,7674	0,7595	0,7605	0,7634
Detection Rate	0,6995	0,6807	0,6949	0,6534
Detection Prevalence	0,8193	0,7698	0,7962	0,7804
Balanced Accuracy	0,6984	0,7628	0,7455	0,6595
ROC	0,8822181	0,8965511	0,8780679	0,826732

M3: Comparação entre classificadores multinominais, onde a categorização foi feita 1 (alunos com notas abaixo de 4,0) e 2 (alunos com notas acima de 4,0 e abaixo de 7,0) e 3 (alunos com notas acima de 7,0)

Métricas de Avaliação	Regressão Logística			Árvore de Decisão		
Accuracy	0,7452			0,7162		
95% CI	(0,7315, 0,7533)			(0,7068, 0,7255)		
No Information Rate	0,4566			0,4365		
P-Value [Acc > NIR]	< 2,2e-16			< 2,2e-16		
Kappa	0,6312			0,5635		
McNemar's Test P-Value	2,07E-03			1,565E-08		
Classes	1,Class[0<4]	2,Class[4<7]	3,Class[7<=10]	1,Class[0<4]	2,Class[4<7]	3,Class[7<=10]
Sensitivity	0,862	0,6324	0,72	0,8531	0,5342	0,7119
Specificity	0,9115	0,8268	0,9	0,89	0,82	0,8685
Pos Pred Value	0,883	0,6384	0,69	0,8573	0,5857	0,632
Neg Pred Value	0,895	0,823	0,91	0,8867	0,7871	0,9048
Prevalence	0,4366	0,3259	0,24	0,4365	0,3226	0,2409
Detection Rate	0,3764	0,2061	0,17	0,3724	0,1723	0,1715
Detection Prevalence	0,4262	0,3229	0,25	0,4344	0,2943	0,2713
Balanced Accuracy	0,8867	0,7296	0,81	0,8715	0,6771	0,7902

Métricas de Avaliação	SVM			Naive Bayes		
Accuracy	0,7267			0,5877		
95% CI	(0,7174, 0,7358)			(0,5774, 0,5978)		
No Information Rate	0,4351			0,4372		
P-Value [Acc > NIR]	< 2,2e-16			< 2,2e-16		
Kappa	0,5802			0,3395		
McNemar's Test P-Value	< 2,2e-16			< 2,2e-16		
Classes	1,Class[0<4]	2,Class[4<7]	3,Class[7<=10]	Class1:[0<4]	Class2: [4<7]	Class2: [7<=10]
Sensitivity	0,8565	0,5454	0,73	0,8728	0,3732	0,35727

Specificity	0,893	0,8329	0,87	0,6414	0,793	0,90852
Pos Pred value	0,8604	0,6077	0,64	0,6541	0,4538	0,56211
Neg Pred value	0,8898	0,7943	0,91	0,8665	0,733	0,81134
Prevalence	0,4351	0,3219	0,24	0,4372	0,3155	0,24738
Detection Rate	0,3726	0,1755	0,18	0,3816	0,1177	0,08838
Detection Prevalence	0,4331	0,2889	0,28	0,5834	0,2594	0,15723
Balanced Accuracy	0,8747	0,6892	0,8	0,7571	0,5831	0,63289



Comparativo Curva ROC entre classificadores

Script Markdown para o comparativo entre classificadores por curva ROC

Pacotes utilizados

```
library(kernlab)          # pacote svm
library(party)            # Pacote para Árvore de decisão
library(caret)             # pacote para criar a matriz de confusão
library(plotROC)           # Para a Curva ROC
library(ggplot2)            # Pacote para visualização gráfica
library(e1071)              # Pacote com o algoritmo Naive Bayes
```

1. Importando e tratando a base de dados

```
# carregando a base de dados
dados <- na.exclude(read.csv2("baseGeral.csv", header=T))
```

2. Construindo base de treinamento e teste

```
# Separação dos dados de treinamento e de teste com o pacote caret
split <- createDataPartition(y = dados$VAR01, p = 0.7, list = FALSE)
treinamento <- dados[split,]
teste <- dados[-split,]
```

3. Construindo os modelos de classificação

```
# Criação do Modelo para o algoritmo SVM
modeloSVM<-ksvm(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VA
R07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35, data=treinamento)
```

Criação do Modelo para a Regressão Logística

```
modeloLogistico <- glm(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+V
AR06+VAR07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35, family=binomial(link="logit"), data=treinamento)
```

```
#modeloLogistico <- step(modeloLogisticoInicial)
```

```
# Criação do Modelo ArvoreDecisao
modeloArvoreDecisao<-ctree(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35,data=treinamento)
```

Criação do Modelo Naive Bayes

```
modeloNaiveBayes<-naiveBayes(DESEMPENHO_BINARIO~VAR01+VAR02+VAR03+VAR04+VAR05+VAR06+VAR07+VAR08+
VAR09+VAR10+VAR12+VAR13+VAR14+VAR16+VAR17+VAR18+VAR20+VAR21+VAR22+
VAR23+VAR24+VAR25+VAR28+VAR31+VAR31b+VAR31c+VAR32a+VAR32b+VAR32c+
VAR32d+VAR33+VAR34+VAR35,data=treinamento)
```

4. Avaliando o classificador

```
# Gerando o vetor de probabilidades para o algoritmo SVM
classificacaoProbSVM <- predict(modeloSVM,newdata=teste,type="response")
```

Gerando o vetor de probabilidades para o algoritmo Arvore de decisão

```
classificacaoProbLogistico <- predict(modeloLogistico,newdata=teste,type="response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

Gerando o vetor de probabilidades para o algoritmo Arvore de Decisão

```
classificacaoProbArvore <- predict(modeloArvoreDecisao,teste)
colnames(classificacaoProbArvore)<- "ArvoreDecisao"
```

Gerando o vetor de probabilidades para o algoritmo Naive Bayes

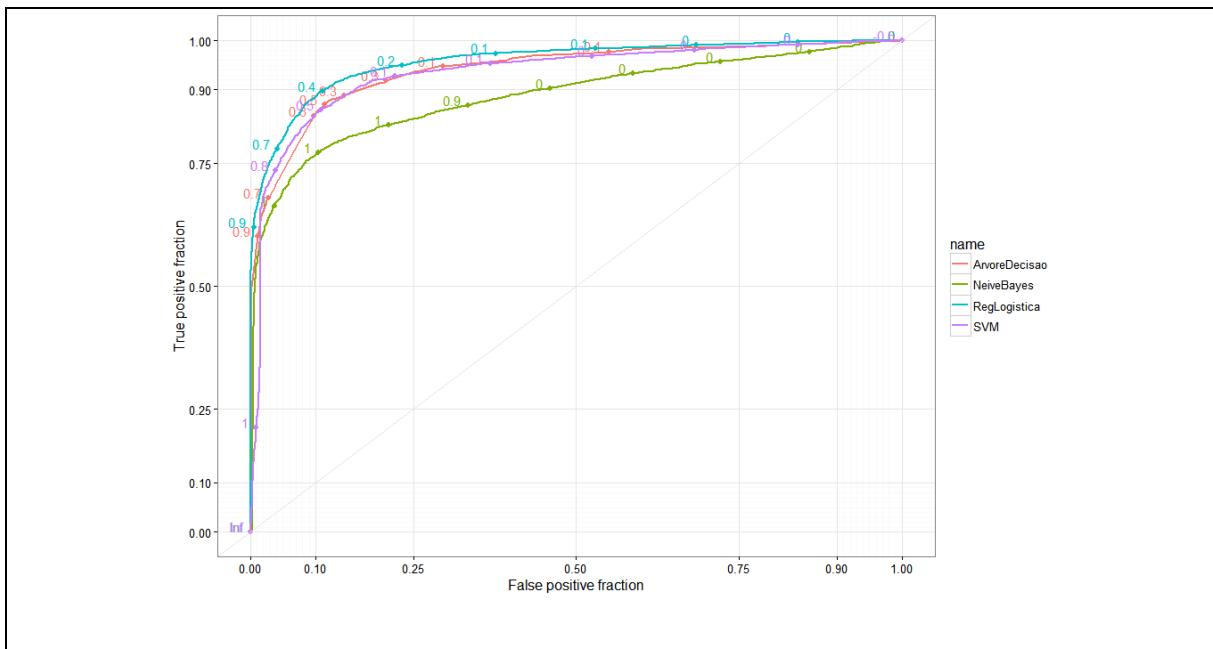
```
classificacaoProbNaiveBayes <- predict(modeloNaiveBayes,teste, type = "raw")
classificacaoProbNaiveBayes <- classificacaoProbNaiveBayes[,2]
```

5. Gerando a curva ROC

Curva ROC

```
test <- data.frame(D = teste$DESEMPENHO_BINARIO, SVM = classificacaoProbSVM,
Logistico = classificacaoProbLogistico, ArvoreDecisao = classificacaoProbArvore,
NaiveBayes = classificacaoProbNaiveBayes, stringsAsFactors = FALSE)
```

```
longtest <- melt_roc(test, "D", c("SVM","Logistico", "ArvoreDecisao","NaiveBayes"))
ggplot(longtest, aes(d = D, m = M, color = name)) + geom_roc() + style_roc()
```



Apêndice



Casos de Uso e Protótipos

Casos de Uso 1: Módulo Visão geral dos dados

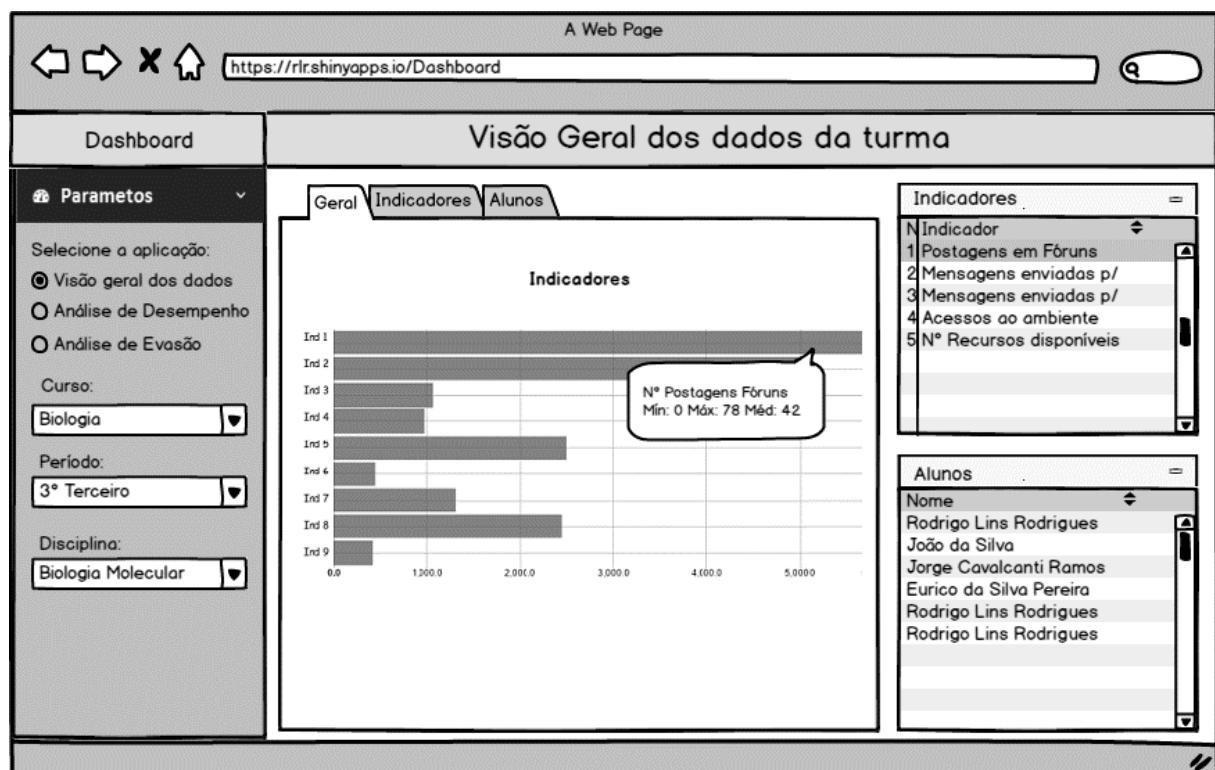
Fluxo Principal de Eventos

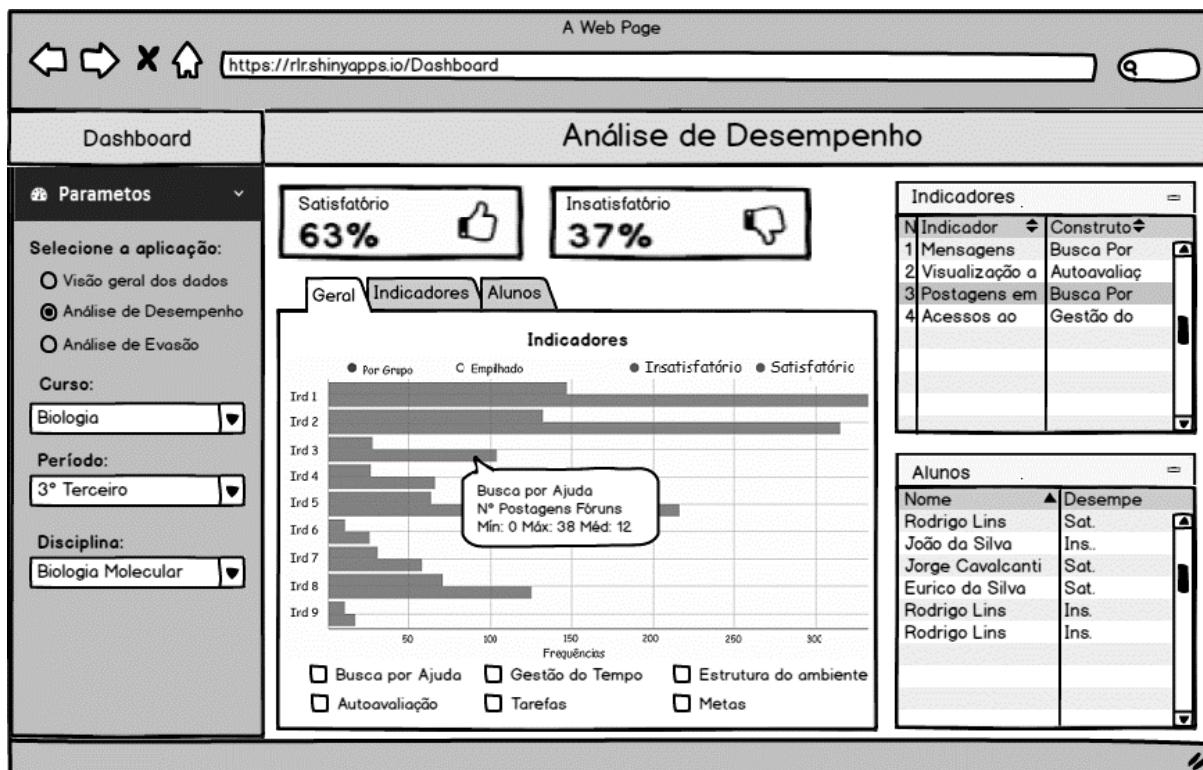
Passos	Ações
1	O professor seleciona o módulo ao qual deseja fazer utilização. Tendo as opções entre: Visão geral dos dados e Análise de Desempenho
2	O Professor seleciona o período ao qual deseja informações
3	O professor Seleciona a Disciplina a qual deseja informações
4	O professor visualizará o gráfico da figura 1 por default a primeira aba (Geral)
5	Ao passar o mouse em cima da barra aparecerá um hit com informações sobre Descrição da variável (indicador), valor mínimo [min(x)], valor máximo [max(x)], e média [mean(x)].
6	O usuário poderá selecionar na tabela de indicadores (Figura 1 - direita) qual indicador deseja mostrar no gráfico, para isto, a tabela deverá ser um objeto reativo trocando informações com o gráfico. Podem ser selecionadas mais de um indicador. Na medida em que o usuário for selecionando uma variável, sua barra é mostrada no gráfico.
7	Ao selecionar a variável na tabela reativa, deverá ser plotado o gráfico, com o título referente a descrição da variável.
8	O professor seleciona o módulo Análise de Desempenho .
9	Ao selecionar as informações aparecerá uma tela [figura 2] contendo dois objetos visuais com os percentuais de alunos na classe 0 (Satisfatório) e da classe 1 (Insatisfatório), de acordo com a variável (DESEMPENHO_BINARIO) da base.
10	Aparecerá o componente tabBox que servirá de container para exibir o gráfico. O tabBox deverá ter três abas (figura 2)
11	O professor poderá selecionar (Por grupo ou Empilhado).
12	O professor poderá selecionar Insatisfatório ou Satisfatório,

13	O professor poderá selecionar os grupos de variáveis que será mostrado (busca por ajuda, gestão do tempo, estrutura do ambiente, autoavaliação, tarefas, metas) de acordo com o dicionário.
----	---

Fluxo Alternativo

Passos	Ações
1.1	Se o professor selecionar qualquer um dos três módulos o sistema deverá setar (Curso, Período e Disciplina) como o primeiro curso da base, o primeiro período da base e a primeira disciplina.
13.1	Caso o professor selecione Insatisfatório mostrará apenas as barras referente ao grupo insatisfatório, caso o usuário selecione satisfatório será mostrado apenas as barras referente ao grupo satisfatório.





Casos de Uso 2: Módulo Indicadores

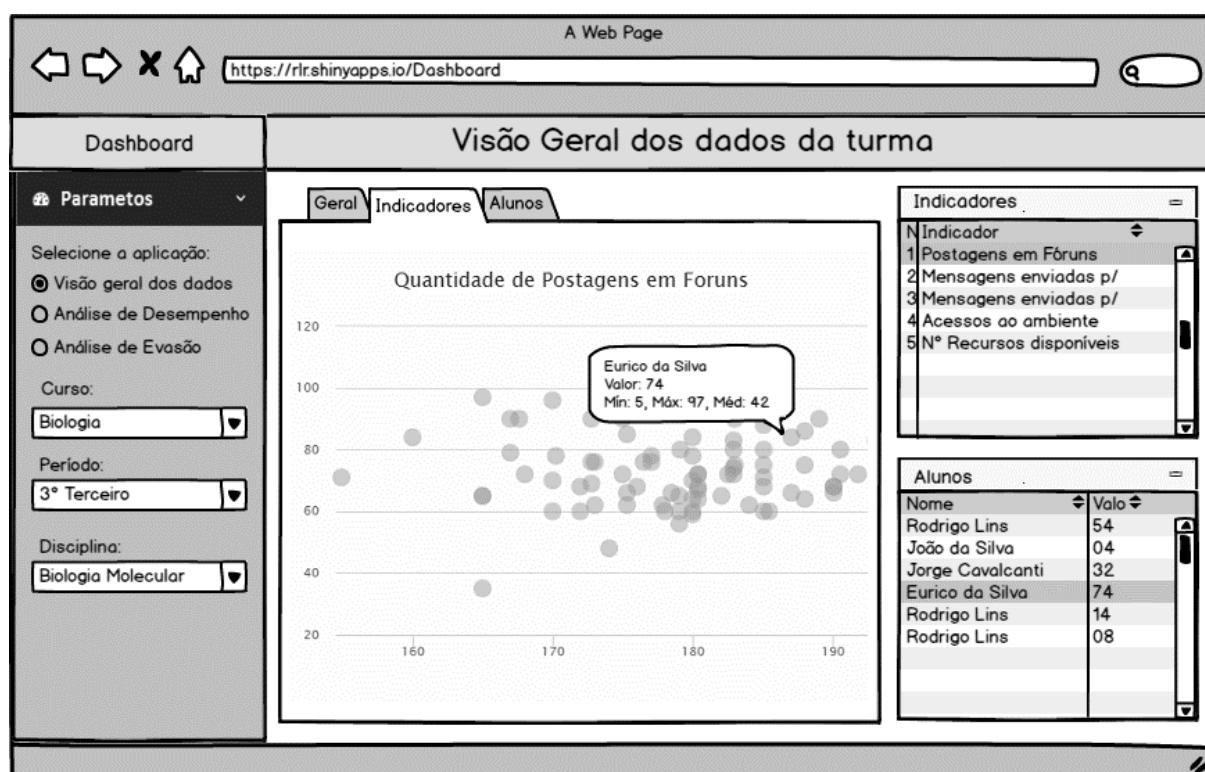
Fluxo Principal de Eventos

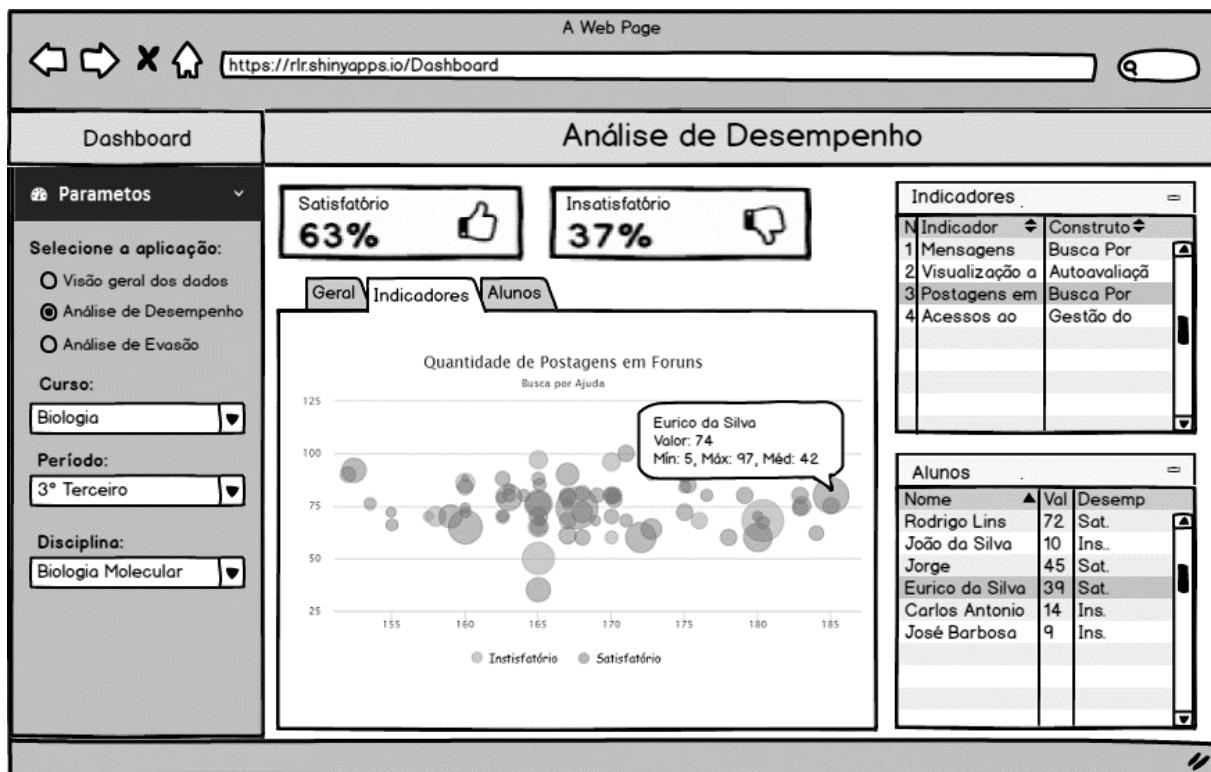
Passos	Ações
1	O Usuário seleciona o módulo ao qual deseja fazer utilização. Tendo as opções entre: Visão geral dos dados e Análise de Desempenho .
2	O Usuário seleciona o período ao qual deseja informações
3	O Usuário Seleciona a Disciplina a qual deseja informações
4	O Usuário visualizará o gráfico da figura 1 por default a primeira aba (Geral)
5	O Usuário Clica na aba 2 (Indicadores)
6	O Usuário visualizará um gráfico com a primeira variável da tabela Indicadores . No rodapé de todos os gráficos deve ter uma observação Clique e arraste sobre uma região para Zoom .
7	Ao passar o mouse em cima de alguma bolha aparecerá um hit com o nome do aluno, valor desse indicador para o aluno, valor mínimo [min(x)], valor máximo [max(x)], e média [mean(x)] da variável na disciplina selecionada.
8	O usuário poderá selecionar na tabela de indicadores (Figura 1 - direita) qual indicador deseja mostrar no gráfico, para isto, a tabela deverá ser um objeto reativo trocando informações com o gráfico.
9	Ao selecionar a variável na tabela de indicadores, deverá ser plotado o gráfico, com o título referente a descrição da variável.
10	Ao selecionar um aluno, a bolha que o representa também será destacada no gráfico, com seu respectivo hit.
11	O Usuário seleciona o módulo Análise de Desempenho .

12	Ao selecionar as informações aparecerá uma tela [figura 2] contendo dois objetos visuais (valueBox), que serão carregados progressivamente, com os percentuais de alunos na classe 0 (Satisfatório) e da classe 1 (Insatisfatório), de acordo com a variável (DESEMPENHOBINARIO) da base.
11	Aparecerá o componente tabBox que servirá de container para exibir o gráfico. O tabBox deverá ter três abas (figura 2)
12	O Usuário poderá selecionar Insatisfatório , satisfatório ou ambos .

Fluxo Alternativo

Passos	Ações
1.1	Se o Usuário selecionar qualquer um dos três módulos o sistema deverá setar (Curso, Período e Disciplina) como o primeiro curso da base, o primeiro período da base e a primeira disciplina.
12.1	Caso o Usuário selecione Insatisfatório mostrará apenas as bolhas referente ao grupo insatisfatório, caso o usuário selecione satisfatório será mostrado apenas as barras referente ao grupo satisfatório. Ambas as opções também podem ser selecionadas. Idem para a ação 17 (Evasão).





Casos de Uso 3: Módulo por Alunos

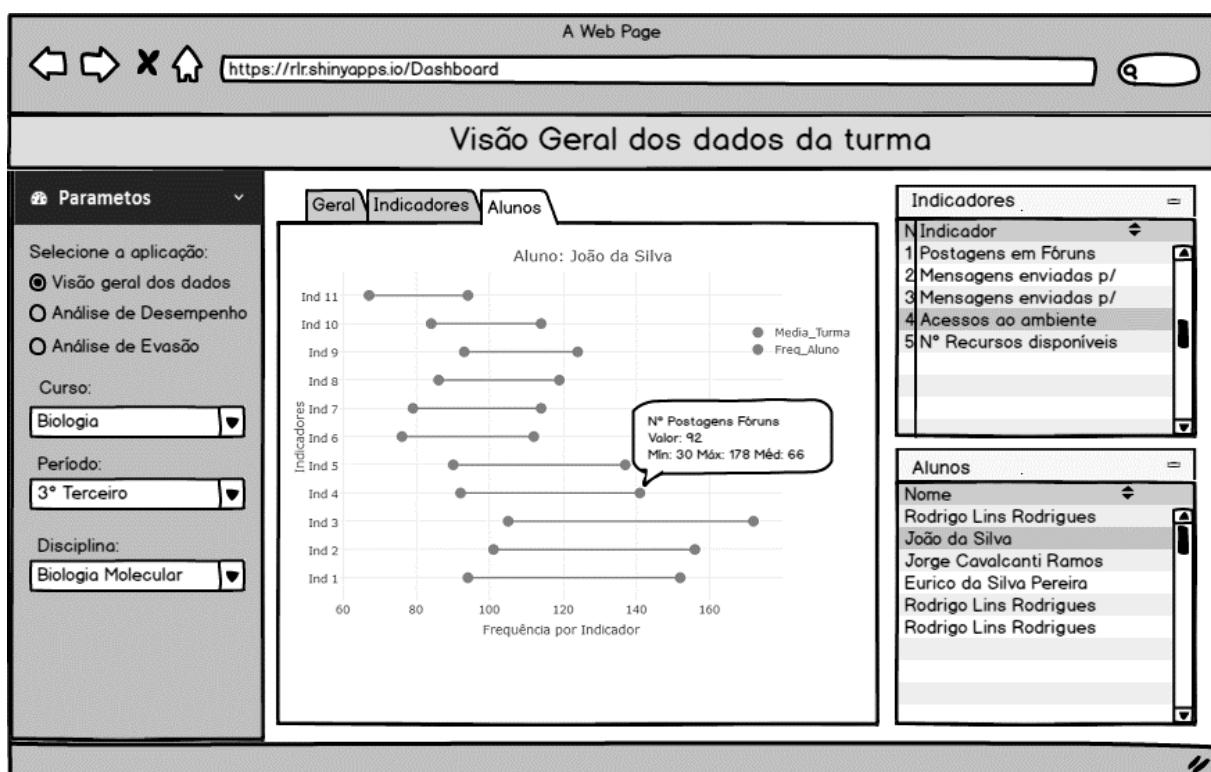
Fluxo Principal de Eventos

Passos	Ações
1	O professor seleciona o módulo ao qual deseja fazer utilização. Tendo as opções entre: Visão geral dos dados e Análise de Desempenho
2	O Professor seleciona o período ao qual deseja informações
3	O professor Seleciona a Disciplina a qual deseja informações
4	O professor visualizará o gráfico da figura 1 por default a primeira aba (Geral)
5	Ao passar o mouse em cima de qualquer bolinha no gráfico aparecerá um hit com informações sobre Descrição da variável (indicador), valor do aluno, valor mínimo [min(x)], valor máximo [max(x)], e média [mean(x)].
6	O usuário poderá selecionar na tabela Alunos qual aluno deseja mostrar no gráfico, para isto, a tabela deverá ser um objeto reativo trocando informações com o gráfico.
7	Ao selecionar um aluno na tabela reativa, deverá ser plotado o gráfico, com o título referente ao nome do aluno.
8	O professor seleciona o módulo Análise de Desempenho .
9	Ao selecionar as informações aparecerá uma tela [figura 2] contendo dois objetos visuais com os percentuais de alunos na classe 0 (Satisfatório) em verde e da classe 1 (Insatisfatório) em vermelho , de acordo com a variável (DESEMPENHO_BINARIO) da base.
10	Aparecerá o componente tabBox que servirá de container para exibir o gráfico. O tabBox deverá ter três abas (figura 2)
11	O professor poderá selecionar clicando na legenda somente satisfatório, somente a frequência do aluno, ambos ou nenhum.

12	O professor poderá selecionar os grupos de variáveis que serão mostrados (busca por ajuda, gestão do tempo, estrutura do ambiente, autoavaliação, tarefas, metas) de acordo com o dicionário.
----	---

Fluxo Alternativo

Passos	Ações
1.1	Se o professor selecionar qualquer um dos três módulos o sistema deverá setar (Curso, Período e Disciplina) como o primeiro curso da base, o primeiro período da base e a primeira disciplina.
11.1	Caso o professor selecione Insatisfatório mostrará apenas as barras referente ao grupo insatisfatório, caso o usuário selecione satisfatório será mostrado apenas as bolinhas referente à média do grupo satisfatório.



A Web Page
<https://rlr.shinyapps.io/Dashboard>

Dashboard

Análise de Desempenho

Parametros

Selecionar a aplicação:

- Visão geral dos dados
- Análise de Desempenho
- Análise de Evasão

Curso:

Biologia

Período:

3º Terceiro

Disciplina:

Biologia Molecular

Satisfatório

63%

Insatisfatório

37%

Geral Indicadores Alunos

Indicadores do Aluno João da Silva

Ind 11 Ind 10 Ind 9 Ind 8 Ind 7 Ind 6 Ind 5 Ind 4 Ind 3 Ind 2 Ind 1

60 80 100 120 140 160

Frequência por Indicador

● Satisfatório ● Freq_Aluno

Busca por Ajuda N° Postagens Fóruns Valor: 10 Min: 5 Máx: 98 Méd: 46

Busca por Ajuda Gestão do Tempo Estrutura do ambiente
 Autoavaliação Tarefas Metas

Indicadores

Indicador	Construto
1 Mensagens	Busca Por
2 Visualização a	Autoavaliaçā
3 Postagens em	Busca Por
4 Acessos ao	Gestão do

Alunos

Nome	Val	Desemp
Rodrigo Lins	72	Sat.
João da Silva	10	Ins.
Jorge	45	Sat.
Eurico da Silva	39	Sat.
Carlos Antonio	14	Ins.
José Barbosa	9	Ins.



Roteiro da Avaliação de Usabilidade



Universidade Federal de Pernambuco
CIN - CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

INSTRUÇÕES PARA PARTICIPAÇÃO NO EXPERIMENTO DE AVALIAÇÃO DE PROTÓTIPOS E VALIDAÇÃO DE APLICAÇÃO

Caro(a) colega professor(a) ou tutor(a):

Convidamos você para contribuir com duas pesquisas de doutorado em ciência da computação pelo Centro de Informática/UFPE. As pesquisas tratam, em linhas gerais, da definição de modelos de previsão de **desempenho** de alunos na Educação a Distância. Para tanto, os estudos foram baseados em duas importantes teorias: A *Autorregulação da Aprendizagem* (Zimmerman, 1990).

Sua contribuição se dará a partir da participação em uma sessão experimental com execução de tarefas e respostas a questionários. Estimamos em cerca de 20 minutos a realização da sessão e seus resultados serão usados única e exclusivamente no desenvolvimento das análises e definições das pesquisas. Será resguardada a identificação de cada participante, assim como seus resultados individuais.

Esta sessão será composta das seguintes etapas:

- 1 - *Explicações gerais sobre o experimento e assinatura do Termo de Consentimento Livre e Esclarecido.*
- 2 - *Apresentação das linhas gerais das pesquisas, com uma breve introdução sobre as duas teorias predominantes nos projetos.*
- 3 - *Apresentação sobre os objetivos da previsão de desempenho.*
- 4 - *Apresentação dos protótipos da aplicação: Visão Geral dos Dados e Análise de Desempenho.*

5 - Resposta a um breve questionário sobre as expectativas dos usuários quanto à aplicação, conforme protótipos apresentados.

6 - Apresentação da aplicação: Visão Geral dos Dados e Análise de Desempenho.

7 - Execução de algumas tarefas utilizando a aplicação.

8 - Resposta a um breve questionário sobre as experiências dos usuários quanto à aplicação.

Durante todo o experimento, um aplicador poderá esclarecer dúvidas gerais. Por questões da análise posterior da execução das tarefas, a ajuda se limitará a esclarecer pontos relacionados com a interpretação da tarefa e não da sua execução.

Antes de iniciarmos o experimento, será necessário ler e assinar o **Termo de Consentimento Livre e Esclarecido** de participação, onde estão expressas as diretrizes e recomendações acerca dos procedimentos a serem utilizados antes, durante e após o experimento.

A execução das tarefas a seguir será toda gravada para análises posteriores, por isso, o voluntário deve, sempre que possível, falar sobre o procedimento adotado, dúvidas que surgirem, resultados obtidos etc.

Antes de começar o teste você terá **2 minutos** para explorar à vontade a aplicação. Após esse tempo, você receberá as tarefas a serem realizadas na aplicação.

EXECUÇÃO DE TAREFAS NA APLICAÇÃO	
Nº	Descrição
1	<p>[MÓDULO VISÃO GERAL DOS DADOS – ABA GERAL]</p> <p>a) No curso de Pedagogia, 3º período, visualizar o gráfico dos dados da disciplina Metodologia Científica.</p> <p>b) Descobrir no gráfico qual a média dessa turma em relação ao indicador "Quantidade de mensagens enviadas por aluno aos professores".</p>
2	<p>[MÓDULO VISÃO GERAL DOS DADOS - ABA INDICADORES]</p> <p>a) No curso de Biologia, 5º período, na disciplina Bioquímica II, visualizar o gráfico do indicador "Média semanal de acessos do aluno ao ambiente".</p> <p>b) Identificar qual (ou quais) aluno(s) de maior valor nesse indicador no gráfico e dar um zoom na região do gráfico que mostra esse(s) aluno(s).</p>
3	<p>[MÓDULO ANÁLISE DE DESEMPENHO - ABA GERAL]</p> <p>a) No curso de Biologia, 4º período, disciplina Genética Geral, no construto ESTABELECIMENTO DE METAS, identifique o indicador "Quantidade geral de postagens de um(a) aluno(a) em fóruns" e verifique a média para o grupo insatisfatório;</p> <p>b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.</p>
4	<p>[MÓDULO ANÁLISE DE DESEMPENHO - ABA GERAL]</p> <p>a) Para o curso de Pedagogia, 4º Período, disciplina Planejamento Educacional, verifique o indicador "Quantidade de mensagens enviadas por aluna(o) às(os) Tutor(es) pelo ambiente" e informe a quantidade de mensagens enviadas do grupo Satisfatório e do grupo de Não Satisfatório;</p> <p>b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.</p>
5	<p>[MÓDULO ANÁLISE DE DESEMPENHO - ABA INDICADORES]</p> <p>a) Para o curso de Administração, 8º Período, disciplina Licitações, Contratos e Convênios, identifique o aluno que está em pior situação de desempenho em relação ao indicador "Quantidade de leituras feitas ao fórum (pageviews) ".</p> <p>b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.</p>

[MÓDULO ANÁLISE DE DESEMPENHO - ABA ALUNOS]	
6	<p>a) Selecione o curso de Letras, 6º Período, disciplina Literatura Brasileira III, no construto GESTÃO DO TEMPO. Em seguida, procure a aluna Eveline <sobrenome omitido> e identifique os dois indicadores que a aluna se encontra em pior situação entre a frequência da aluna e a média da turma;</p> <p>b) Fale quais ações pedagógicas devem ser realizadas em razão deste resultado.</p>

Anexos

Anexo

Questionário OSLQ

Questionário OSLQ desenvolvido por (BARNARD, LAN, et al., 2009).

Item	Subscale
1. I set standards for my assignments in online courses.	Goal Setting
2. I set short-term (daily or weekly) goals as well as long-term goals (monthly or for the semester).	
3. I keep a high standard for my learning in my online courses.	
4. I set goals to help me manage studying time for my online courses.	
5. I don't compromise the quality of my work because it is online.	
6. I choose the location where I study to avoid too much distraction.	Environment Structuring
7. I find a comfortable place to study.	
8. I know where I can study most efficiently for online courses.	
9. I choose a time with few distractions for studying for my online courses.	
10. I try to take more thorough notes for my online courses because notes are even more important for learning online than in a regular classroom.	Task Strategies
11. I read aloud instructional materials posted online to fight against distractions.	
12. I prepare my questions before joining in the chat room and discussion.	
13. I work extra problems in my online courses in addition to the assigned ones to master the course content.	
14. I allocate extra studying time for my online courses because I know it is time-demanding.	Time Management
15. I try to schedule the same time every day or every week to study for my online courses, and I observe the schedule.	
16. Although we don't have to attend daily classes, I still try to distribute my studying time evenly across days.	
17. I find someone who is knowledgeable in course content so that I can consult with him or her when I need help.	Help Seeking
18. I share my problems with my classmates online so we know what we are struggling with and how to solve our problems.	
19. If needed, I try to meet my classmates face-to-face.	
20. I am persistent in getting help from the instructor through e-mail.	
21. I summarize my learning in online courses to examine my understanding of what I have learned.	Self Evaluation
22. I ask myself a lot of questions about the course material when studying for an online course.	
23. I communicate with my classmates to find out how I am doing in my online classes.	
24. I communicate with my classmates to find out what I am learning that is different from what they are learning.	

Questionário da Experiência do usuário (UX)

Fonte: <http://www.attrakdiff.de/>

Desinteressante						Convidativo
Mau						Bom
Repulsivo						Atrativo
Desmotivador						Motivador