UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO
EM MODELAGEM COMPUTACIONAL DO CONHECIMENTO

Cheops Araujo Malta

# Um modelo computacional para classificação da motivação de estudantes em educação on-line

Maceió - AL
2016

CHEOPS ARAUJO MALTA

# Um modelo computacional para classificação da motivação de estudantes em educação on-line

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Modelagem Computacional do Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Alan Pedro da Silva
Coorientador: Prof. Dr. Ig Ibert Bittencourt Santana Pinto

Maceió - AL
2016

**Catalogação na fonte**
**Universidade Federal de Alagoas**
**Biblioteca Central**
**Divisão de Tratamento Técnico**
Bibliotecária responsável: Helena Cristina Pimentel do Vale

M261c Malta, Cheops Araujo.
Um modelo computacional para classificação da motivação de estudantes em educação on-line / Cheops Araujo. 2017.
123 f. : il.

Orientador: Alan Pedro da Silva.
Coorientador: Ig Ibert Bittencourt Santana Pinto.

Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação, Maceió, 2017.

Bibliografia: f. 97-98.
Apêndices: 99-117.
Anexos: f. 118-123.

1. Programação de computadores – Ensino. 2. Ferramenta de auxílio a aprendizagem. 3. Motivação na educação. 4. Mineração de dados Educacionais I. Título.

CDU: 004.42:378

**UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL**
**Programa de Pós-Graduação em Modelagem Computacional de Conhecimento**
Avenida Lourival Melo Mota, Km 14, Bloco 09, Cidade Universitária
CEP 57.072-900 – Maceió – AL – Brasil
Telefone: (082) 3214-1364

Membros da Comissão Julgadora da Dissertação de Mestrado de Cheops Araújo Malta, intitulada: "Um modelo computacional para classificação da motivação de estudantes em educação on-line", apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas, em 27 de junho de 2016, às 9h00min, na sala 30 do Instituto de Computação da Ufal.

**COMISSÃO JULGADORA**

**Prof. Dr. Alan Pedro da Silva**
Ufal – Instituto de Computação
Orientador

**Prof. Dr. Ig Ibert Bittencourt Santana Pinto**
Ufal – Instituto de Computação
Coorientador

**Prof. Dr. Jorge Artur Peçanha de Miranda Coelho**
Ufal – Faculdade de Medicina
Examinador

**Profa. Dra. Patrícia Augustin Jaques Maillard**
Unisinos – Centro de Ciências Exatas e Tecnológicas
Examinadora

Maceió, junho de 2016.

*Dedico essa dissertação aos meus pais por sempre me mostrarem que a educação é a chave para a realização das transformações, sejam elas pessoais ou sociais. À minha namorada Mariana pela compreensão em minhas ausências e pelos ombros e abraços afetuosos nos momentos mais difíceis. A todos aqueles que, de modo velado ou explícito, torceram e vibraram pelo meu sucesso ao longo dessa jornada... Sem vocês, eu não teria conseguido.*

# AGRADECIMENTOS

A **Deus**, Senhor de todas as coisas, Criador e Inspirador de todos os momentos, pelas bênçãos e graças derramadas em abundância em minha vida e, em especial, pelo fato de ter me permitido encontrá-lo verdadeiramente quando eu mais precisava.

À Santa **Tereza D'Ávila**, espanhola, Doutora da Igreja, que me acalmou em muitas aflições por meio de sua oração: "Nada te perturbe, nada te assuste, tudo passa. Deus não muda. A paciência tudo alcança. Quem a Deus tem nada lhe falta. Só Deus basta!"

Aos meus pais, **Malta** e **Vera**, meus maiores incentivadores, pessoas de valor em seu sentido mais profundo. Vocês me mostraram pelos seus erros e acertos que a vida é feita para ser vivida, e que, só se vale a pena vivê-la se for para fazer o melhor para si e para os que estão ao nosso redor. Tudo que fiz e que faço é para que vocês se orgulhem de mim. Obrigado por plantarem, adubarem e regarem essa semente que hoje está crescendo. Tenho certeza de que vocês ainda hão de ver bons frutos.

À minha namorada **Mariana** pelo incentivo, pela torcida e pelo amor que tem me dado ao longo dos últimos anos. Sei que muitas vezes é difícil entender uma vida tão corrida, quilômetros de distância, breves contatos, mas, sempre que nos encontramos, tudo isso se dissipa e as coisas se resolvem. Acredite, tudo isso um dia fará sentido, todas as nossas lutas de hoje serão o insumo do nosso sucesso de amanhã.

À tia **Lúcia** e à prima **Taciana**, família que eu escolhi e que me escolheu. Obrigado por dividirem comigo o seu espaço, obrigado por me acolherem em Maceió sempre que eu precisei. Vou lembrar-me disso para sempre!

Aos meus orientadores, **Alan Pedro** que sempre me tranquilizou nos momentos de maior apreensão e **Ig Bittencourt** por ter me ensinado por meio de palavras e atos o verdadeiro sentido da palavra "resiliência". Obrigado por todos os momentos, pela amizade e confiança. E, sobretudo, por terem acreditado em mim, por terem me mostrado que eu era capaz sempre que achei que não era. Orientar não é só guiar por caminhos acadêmicos; é mostrar rumos para uma vida. Vocês fizeram isso muito bem. Por isso e por muitas outras coisas serei eternamente grato.

À professora **Patrícia Ospina**, "anjo" que me apareceu no momento preciso. Sua presença em minha vida me fez acreditar ainda mais na força do Criador e ter ainda mais certeza de que ele não abandona os seus filhos. Obrigado pela sua calma, afeto, companheirismo e presença constante. Você foi, é, e será minha eterna Orientadora. Seu amor de mãe para com seus alunos é algo ímpar. Continue sempre assim!

Aos meus amigos de perto ou de longe, do meio acadêmico ou não, que nunca hesitaram em me ajudar e comemorar os meus sucessos. Com menção especial a **Mateus**

**Teles** e **Natália Franco**.

Aos membros do **NEES** e da **Universidade São Francisco**, amigos e grandes companheiros nesse processo. Com menção especial a **Ranilson Paiva**, **Ananias Queiroga**, **Diego Dermeval**, os quais perturbei em excesso, mas que prontamente me ajudaram em todas as "pelejas".

À nossa querida Universidade Federal de Alagoas – **UFAL**, casa do saber, que me acolheu e me acolhe, a ela e a todo o seu corpo de técnicos, professores e alunos o meu agradecimento.

A todos aqueles que, de alguma forma, contribuíram para a conclusão desse trabalho, o meu mais sincero obrigado!

*"Um pouco de ciência nos afasta de Deus. Muito, nos aproxima."*

***Louis Pasteur (1822-1895)***

# RESUMO

Em 2014, o Brasil contou com uma oferta de 25.166 cursos por meio da educação a distância. Este e outros dados mostram que a EAD não é um modismo, mas sim parte de um amplo e contínuo processo de mudança, bem como atestam a fase de consolidação desta modalidade de ensino, principalmente no ensino superior, onde seu crescimento tem sido expressivo e sustentado. No entanto, ainda são muitos os desafios sejam eles pedagógicos e/ou tecnológicos. Um destes desafios está relacionado com a motivação para aprender. Estudos evidenciam que conhecer os motivos e as metas que levam os alunos a envolver-se ou não com a aprendizagem é importante, tanto do ponto de vista motivacional, quanto ao fato de ser uma questão-chave para ajudar a compreender os processos de aprendizagem e as variáveis que os determinam. Neste sentido, propomos um modelo para a classificação automática da motivação dos estudantes da educação on-line, gerado com o auxílio de Instrumentos Psicométricos e Mineração dos Dados Educacionais. O modelo é desenvolvido em três etapas que ocorrem de modo sequencial, iniciando com a "construção da base de dados", onde são coletados os dados dos alunos por meio de instrumentos psicométricos (questionários) e logs (registros de interação no ambiente virtual de aprendizagem). Na etapa seguinte, é realizado um experimento para "seleção do algoritmo para classificação" a ser utilizado na construção dos modelos. Finalmente, na etapa de "construção do modelo", é construído e validado o modelo para classificação da motivação dos estudantes.

**Palavras-chaves**: Educação a Distância; Motivação para Aprender; Metas de Realização; Classificação da Motivação; Mineração de dados Educacionais.

# ABSTRACT

In 2014, Brazil offered 25,166 courses through distance education. Data show that the distance education (Educação a Distância - EAD, in portguese) is not a fad, but rather part of a broad and continuous process of change, and attest the consolidation phase of this type of education, especially in higher education, showing significant and sustainable growth. However, there are still many pedagogical and technological challenges. One of these challenges is related to the motivation to learn. Studies show that the reasons and the goals that lead students whether to engage learning, or not, is important from a motivational point of view, as well as a key issue to help educators understanding the processes of learning and their most important constructs. In this sense, we propose a model to automatically classify the motivation to learn from students of online/distance education. The model is generated with the aid of Psychometric instruments and Educational Data Mining. It is developed in three stages that occur sequentially, starting from the "construction of the database," where students' data are collected through psychometric instruments (questionnaires) and logs (recorded actions and interactions in the virtual learning environment). In the next step, we carried out an experiment to "algorithm selection for classification" to be used in the construction of models. Finally, at the stage of "model building" is constructed and validated the model for student motivation classification.

**Key-words**: Distance Education; Motivation to Learn; Achievement Goals; Classification of Motivation; Educational Data Mining.

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AVA | Ambiente Virtual de Aprendizagem; |
| CEP | Comitê de Ética em Pesquisa; |
| CIED | Coordenadoria Institucional de Educação a Distância; |
| DAM | Desvio Absoluto Médio; |
| EAD | Educação a Distância; |
| EDM | *Educational Data Mining*; |
| EMAPRE | Escala de Motivação para a Aprendizagem; |
| EQM | Erro Quadrático Médio; |
| LDB | Lei de Diretrizes e Bases do Ensino Nacional; |
| MEC | Ministério da Educação; |
| NTI | Núcleo de Tecnologia da Informação; |
| OBL | *Online and Blended Learning*; |
| RNA | Rede Neural Artificial; |
| SVM | *Support Vector Machine*; |
| SEED | Secretaria de Educação a Distância; |
| TICs | Tecnologias da Informação e da Comunicação; |
| TCLE | Termo de Consentimento Livre e Esclarecido; |
| UAB | Universidade Aberta do Brasil; |
| UFAL | Universidade Federal de Alagoas; |
| WEKA | *Waikato Environment for Knowledge Analysis*; |

# *SUMÁRIO*

# 1 INTRODUÇÃO

O presente trabalho trata sobre a proposição, criação e avaliação de um modelo para classificação da motivação de estudantes de educação on-line, estando situada na linha de pesquisa de Modelagem Computacional em Educação do Mestrado em Modelagem Computacional de Conhecimento, do Instituto de Computação da Universidade Federal de Alagoas. Trata-se de um tema inerentemente interdisciplinar, envolvendo áreas do conhecimento como Informática na Educação, Estatística, Inteligência Artificial e Avaliação Psicológica.

O resultado desta dissertação visa auxiliar psicólogos, professores e tutores de cursos ofertados por meio da educação on-line a identificar o perfil motivacional dos seus alunos de forma automática. O modelo que permitirá tal identificação será construído a partir de um conjunto de dados que contém informações obtidas através da aplicação de questionários e informações colhidas nos registros de interação dos alunos dentro do ambiente virtual de aprendizagem (AVA).

## 1.1 MOTIVAÇÃO E CONTEXTUALIZAÇÃO DO TRABALHO

O mundo vem passando por grandes mudanças em todas as esferas, sejam elas políticas, econômicas, sociais, culturais e tecnológicas. Encontramo-nos em meio à sociedade da informação e do conhecimento, uma sociedade pós-industrial que não se preocupa em simplesmente produzir “mais” e de “maneira eficiente”, mas sim em produzir diferenciais pelo tratamento destes novos e volumosos insumos (informação e conhecimento).

Neste cenário extremamente mutável, as incertezas são grandes e estas incertezas, por sua vez, trazem inseguranças que se transformam em barreiras para profissionais de todas as áreas, inclusive educadores acostumados com modelos tradicionais de disseminação de conhecimento.

Assim, emergem as Tecnologias da Informação e Comunicação (TICs), trazendo consigo novas oportunidades e desafios para este cenário onde o conhecimento é fluido, flexível e em constante expansão. Dentre estas tecnologias temos a educação a distância apoiada por computador.

Quando falamos em educação a distância apoiada por computador, podemos pensar que a educação a distância (EAD) seja algo relativamente novo, entretanto, relatos mostram seu início já no século XIX.

A primeira instituição a utilizar a educação a distância foi o Instituto Líber Hermondes, fundado em 1829, na Suécia. Este instituto possibilitou que mais de 150.000 pessoas

realizassem cursos através da EAD (ALVES, 2011).

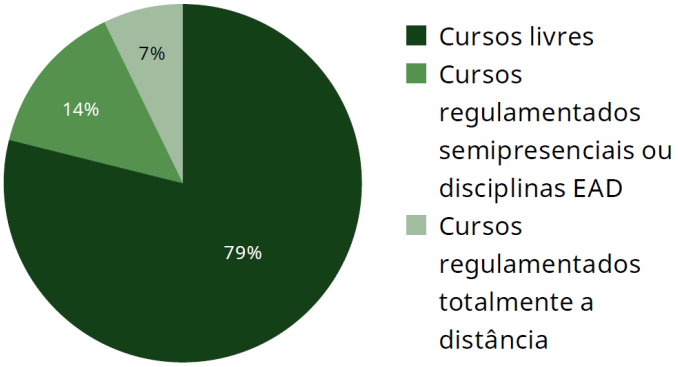
No Brasil, assim como no resto do mundo, não se sabe exatamente quando a EAD teve início. Romão (2008) afirma que ela começou no início do século XX (1922-1945) com a criação da Rádio Sociedade do Rio de Janeiro. Já Alves (2011), apresenta registros que indicam que ela começou quando o Jornal do Brasil registra, em 1904, na primeira edição da seção de classificados, um anúncio que oferece profissionalização por correspondência para datilógrafo.

Entre as décadas de 1970 e 1980, fundações privadas e organizações não governamentais iniciaram a oferta de cursos supletivos a distância, utilizando a teleducação, com aulas via satélite complementadas com materiais impressos. As gerações da EAD por meio do rádio, correspondência e televisão foram se passando, mas somente na década de 1990 é que a maior parte das Instituições de Ensino Superior brasileiras mobilizou-se para a adoção da EAD com o uso de novas tecnologias de informação e comunicação (NTIC's) (ALVES, 2011).

Em 2005, o governo federal regulamenta a EAD pelo Decreto n° 5.622 de 20 de dezembro (BRASIL, 2005). Após este ano, a evolução da oferta e das matrículas de cursos ofertados por meio da EAD vem crescendo ano após ano, segundo dados do CensoEaD.BR 2014/2015, publicação de periocidade anual que busca investigar o setor da educação a distância no Brasil.

Em 2014, o Brasil contou com uma oferta de 25.166 cursos por meio da educação a distância. Dos cursos ofertados neste ano, os cursos livres foram os mais comuns - 19.873 cursos; em seguida, estão os cursos regulamentados semipresenciais ou disciplinas EAD - 3.453 cursos, e os cursos regulamentados totalmente oferecidos por meio da EAD - 1.840 cursos (ABED, 2014). Os percentuais de cada categoria podem ser vistos na Figura 1.

Figura 1 – Proporção de cursos oferecidos em 2014 por tipo.




Fonte: (ABED, 2014)

Neves (2006) afirma que a EAD não é um modismo, mas sim parte de um amplo e contínuo processo de mudança que inclui não só a democratização do acesso a níveis

crescentes de escolaridade e atualização permanente, mas também a adoção de novos paradigmas educacionais. No entanto, ainda são muitos os desafios sejam eles pedagógicos e/ou tecnológicos.

Um destes desafios está relacionado ao problema da evasão. A desistência dos alunos em cursos ofertados pela EAD tem sido evidenciada em diversos estudos e muitos fatores endógenos (relativos ao aluno) e exógenos (relativos ao curso) foram levantados. Autores como Pinto (2010) destacam que a evasão é um fenômeno causado primariamente pela combinação de características dos alunos e suas circunstâncias de vida. Uma destas características é a motivação para aprender.

Neste sentido, diversos estudos têm apontado a importância da motivação para uma aprendizagem efetiva, dada a sua forte relação com resultados de aprendizagem (BOICHé; STEPHAN, 2013). Um aluno desmotivado pode apresentar dificuldades para resolver problemas ou tomar decisões acertadas (CUNHA; BORUCHOVITCH, 2012), e, em casos mais graves, pode até desistir do curso.

Apesar das diferenças existentes entre os vários enfoques teóricos para motivação (Teoria da Autodeterminação, Teoria da Atribuição de Causalidade, Teoria de Metas de Realização, entre outros), dentro dos ambientes escolares Cabanach et al. (1996) comentam que a maioria considera a motivação como o conjunto de processos que implicam na ativação, direção e persistência da conduta.

Dentre os fatores que dirigem essa conduta, encontra-se a percepção que o sujeito tem de si mesmo e das tarefas que irá realizar, as atitudes, os interesses, as expectativas e as diferentes representações mentais que geram os tipos de metas.

Lüftenegger et al. (2012) afirmam que os estudantes que estão motivados a aprender são mais propensos a persistir com a sua educação. Nesta mesma linha, Steinmayr e Spinath (2009) evidenciam que conhecer os motivos e as metas que levam os alunos a envolver-se ou não com a aprendizagem é importante, tanto do ponto de vista motivacional, quanto ao fato de ser uma questão-chave para ajudar a compreender os processos de aprendizagem e as variáveis que os determinam.

Entretanto, entender os fatores motivacionais dos alunos da EAD, especialmente no que se refere à motivação para aprender, pode se tornar uma tarefa bastante complexa. Isso se dá, não somente pelas características dos alunos e suas circunstâncias de vida, mas também pelo processo complexo de interação entre estes fatores (pessoais) e o ambiente educacional (on-line).

## 1.2 PROBLEMÁTICA

A educação a distância apoiada por computador oferece diversas ferramentas para que tanto alunos quanto professores possam desenvolver suas atividades. Por meio da EAD, é possível construir conhecimento de modo não presencial, ou seja, fora de uma sala de aula tradicional, bem como desenvolver habilidades nos alunos que lhe permitam atuar na sociedade do conhecimento.

Nessa modalidade de ensino, o aluno deixa de ser um sujeito passivo, aquele que simplesmente escuta e aplica o que o professor apresenta, e passa a ser um sujeito ativo dentro do processo de aprendizagem, tornando-se o centro do processo e o principal responsável pela busca do conhecimento definindo seu ritmo de estudos e etc. Para Jacobsen et al. (2011), os alunos devem aceitar este fato e assumir maior responsabilidade na condução de seu próprio aprendizado, ou seja, devem ter autonomia dentro do processo.

Nos ambientes de educação a distância, especialmente os que possuem grande número de alunos, os professores nem sempre podem interagir de maneira direta com os alunos, fazendo surgir a presença de um novo sujeito, o professor-tutor. O tutor é o responsável por intermediar toda a comunicação entre os professores e os alunos e vice-versa.

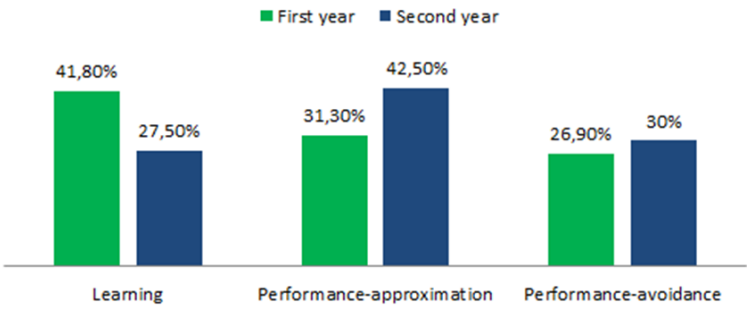
Mesmo com a presença da tutoria e com a conscientização da maior responsabilidade do aluno na busca do conhecimento, Vieira (2015), REIS (2009), Pacheco (2007), Neves (2006) e Biazus (2004) relatam que os alunos, muitas vezes, por não terem a presença física no mesmo ambiente dos colegas e professores, sentem-se solitários, desmotivados e algumas vezes chegando a desistir do curso.

O trabalho realizado por Malta et al. (2015) teve como um dos objetivos classificar a motivação dos alunos em relação à aprendizagem de acordo com a Teoria de Metas de Realização e avaliar as mudanças na classificação da motivação conforme o curso avança. O estudo foi conduzido com alunos de um curso de Sistemas de Informação a distância e os dados foram coletados usando um questionário impresso.

Os resultados mostraram que os alunos do primeiro ano do curso tendem a ter *scores*[1] mais elevados para a meta Aprender que está relacionada a uma postura direcionada para a busca do crescimento intelectual, valorização do esforço pessoal, enfrentamento dos desafios, persistência em relação às atividades acadêmicas e tendência ao uso de estratégias de aprendizagem mais efetivas (AMES, 1992; CLAYTON; BLUMBERG; AULD, 2010; MIDDLETON; MIDGLEY, 1997; ZENORINI; SANTOS, 2010). Por outro lado, os alunos do segundo ano tendem a ter *scores* mais elavados para a meta Performance-Aproximação que está relacionada a uma postura de preocupação em demonstrar a sua capacidade para os demais, menor engajamento e evitação de desafios (ARCHER, 1994; BZUNECK, 2004; ZENORINI; SANTOS, 2010). Um gráfico que apresenta estes resultados pode ser visto na Figura 2.

[1] Em português: pontuações

Figura 2 – Classificação de orientação para a meta dos alunos por ano de curso.




Fonte: (MALTA et al., 2015)

Estes resultados indicam uma mudança no perfil motivacional dos alunos conforme o curso avança e que estas mudanças podem impactar diretamente a sua aprendizagem. Isso atesta ainda mais a importância da compreensão dos perfis motivacionais dos alunos e, neste caso em especial, dos alunos da EAD.

Quando falamos em EAD, não podemos deixar de pensar nos AVAs, ou seja, no espaço onde os alunos desenvolvem suas atividades on-line. Atualmente existe uma grande variedade destes ambientes (podemos citar plataformas como o Moodle[2], Teleduc[3], e-Proinfo[4]).

Esses ambientes, por sua vez, fornecem várias ferramentas que, se bem usadas, podem potencializar os processos de interação, colaboração e cooperação entre professores, alunos e tutores seja de maneira síncrona (atividades em que os participantes estão conectados no ambiente simultaneamente, e.g., chats, videoconferências) ou assíncrona (atividades que não precisam ocorrer em dia e horário determinados, e.g., fóruns, blogs, wikis, glossários e etc).

Levando-se em consideração que os AVAs são para os alunos da educação a distância como a sala de aula convencional dos alunos da educação presencial, temos que as interações dos alunos na plataforma representam as interações que os mesmos teriam com seus professores e colegas de classe, o que se constitui um pilar fundamental para o processo de ensino-aprendizagem. Corroborando com o citado, Moore e Kearsley (2007) afirmam que fatores determinantes do sucesso da EAD são a quantidade e a qualidade dos diálogos entre os professores e os estudantes.

Ainda tomando como referência o processo de classificação dos perfis motivacionais realizado por Malta et al. (2015), mesmo se tratando de um estudo realizado com uma turma da EAD, o método utilizado para a identificação foi manual, o que o torna caro, lento e dependente do usuário.

Desta forma, levando-se em consideração as várias possibilidades dos AVAs e os problemas relacionados às aplicações recorrentes de métodos manuais para classificação dos

[2] Moodle - Modular Object Oriented Dynamic Learning Environment, disponível em: https://moodle.org/
[3] Teleduc, disponível em: http://www.teleduc.org.br/
[4] E-Proifo, disponível em: http://e-proinfo.mec.gov.br

perfis motivacionais dos alunos da EAD, surge a nossa primeira questão: **Como classificar a motivação dos alunos da educação a distância por meio da análise de interações em ambientes virtuais de aprendizagem?**

Os AVAs atuais, em sua maioria, possuem sistemas de armazenamento automático dos registros (*logs*[5]) das atividades realizadas pelos estudantes no que concerne às entradas no sistema, materiais consultados, postagens realizadas em fóruns, participação em *chats*, realização de atividades, *hyperlinks* visitados e etc. A existência desses registros pode ser de grande importância para se monitorar o progresso dos estudantes bem como para detectar, por parte do professor/tutor, casos de possível desmotivação dos alunos e potencial abandono.

Entretanto, dado o número crescente dos alunos em cursos on-line, pode-se tornar humanamente impossível para os professores e tutores uma análise dos registros de cada aluno em separado e em intervalos curtos de tempo.

Com isso, temos a nossa segunda questão: **É possível classificar a motivação dos alunos da educação a distância de modo automático usando seus registros de interação com o sistema?**

Para Pintrich e Schunk (2001) a motivação deve ser encarada como um processo e não como um produto, uma vez que não é possível observá-la diretamente, mas somente inferi-la diante dos comportamentos e dos efeitos que ela produz. Neste sentido, a psicologia tem desenvolvido, nas últimas décadas, instrumentos de avaliação específicos sobre a motivação baseados em modelos teóricos adotados por vários pesquisadores na área, tais como a Teoria da Autodeterminação e a Teoria de Metas de Realização.

A Teoria de Metas de Realização tem trazido grandes contribuições para o entendimento dos fatores motivacionais que influenciam o comportamento do aluno, pois busca explicar a motivação para aprender focalizando o aspecto qualitativo do envolvimento do aluno em seu processo de aprendizagem. As metas referem-se a um conjunto de padrões, pensamentos, propósitos, crenças, percepções, atribuições e conceitos que a pessoa tem de sua capacidade, fatores estes que acabam conduzindo a diferentes consequências cognitivas, afetivas e comportamentais. (AMES, 1992).

Assim sendo, chegamos a nossa terceira questão: **É possível classificar a motivação dos alunos da educação a distância de modo automático usando como arcabouço instrumentos desenvolvidos pela psicologia, baseados na teoria de metas de realização?**

Esses problemas e questões nos guiaram na busca de uma forma de auxiliar, especialmente os professores/tutores da educação on-line, no processo de ensino dentro do contexto da informática na educação.

Apesar de existirem trabalhos relacionados que, por um lado, aplicaram técnicas

[5] Registros de atividades gerados por programas de computador

de mineração de dados para a detecção de padrões educacionais e motivacionais (COCEA; WEIBELZAHL, 2007; ZHANG; CHENG, 2003) e, por outro lado, classificaram a motivação de estudantes por meio de testes baseados em teorias da psicologia (VANSLAMBROUCK et al., 2015; BELUCE; OLIVEIRA, 2015), a literatura da área ainda carece de trabalhos que abordem, no contexto da educação on-line, a classificação da motivação utilizando a mineração de dados educacionais tendo como arcabouço as teorias e os testes desenvolvidos pela psicologia.

Neste sentido, temos como problema de negócio a necessidade de classificar a motivação dos estudantes da educação a distância. Para resolver este problema, e partindo das limitações encontradas nos trabalhos relacionados, desejamos, numa perspectiva técnica, gerar um modelo que dispobilize uma solução computável baseada em análises dos ambientes virtuais de aprendizagem e teorias da psicologia.

## 1.3 OBJETIVOS

O objetivo geral desta dissertação é a criação de um modelo para **classificação da motivação dos alunos de cursos mediados por computador, na perspectiva da teoria de metas de realização**. Essa classificação será obtida a partir da aplicação de testes da psicologia e posterior mineração de dados educacionais oriundos das interações dos alunos no ambiente virtual de aprendizagem.

Nossos objetivos específicos são:

1. Fazer um levantamento bibliográfico sobre motivação para aprender, teoria de metas de realização e mineração de dados educacionais;
2. Levantar e analisar dados para classificação da motivação dos estudantes;
3. Realizar um experimento para detectar qual técnica/algoritmo é mais eficiente para construção do modelo de classificação;
4. Propor um modelo de classificação que possibilite que professores e tutores possam classificar a motivação de seus estudantes, em um dado momento, sem a necessidade de aplicação de instrumentos de avaliação psicológica tradicionais (questionários).
5. Validar o modelo proposto.

## 1.4 CONTRIBUIÇÕES DO TRABALHO

O presente trabalho contribui com a comunidade de Informática na Educação e Psicologia Educacional apresentando um modelo para a classificação da motivação de estudantes da educação on-line baseado em testes da psicologia e em mineração de dados

educacionais. Além disso, outra contribuição significativa do trabalho é a apresentação de um estudo empírico que avaliou as técnicas/algoritmos de classificação que melhor se adequam à construção deste modelo.

## 1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação contém 5 capítulos e encontra-se organizada da seguinte maneira:

- Capítulo 1: Neste capítulo são apresentados a motivação, contextualização, problemática, objetivos, a proposta e as contribuições do trabalho aqui proposto.
- Capítulo 2: É retratada neste capítulo a fundamentação teórica que aborda os principais conhecimentos utilizados nesta dissertação, enfatizando os temas principais desta pesquisa.
- Capítulo 3: São apresentados trabalhos relacionados à classificação da motivação de estudantes em ambientes on-line.
- Capítulo 4: Neste capítulo, apresenta-se de forma detalhada, a proposta da dissertação, abordando todas as etapas do processo de contrução do modelo, desde a obtenção dos dados, avaliação/seleção das técnicas/algoritmo, escolha dos modelos até a etapa final de validação..
- Capítulo 5: Por fim, este capítulo expõe as conclusões acerca do trabalho apresentado, bem como algumas sugestões de trabalhos futuros.
- No final, são evidenciadas as referências utilizadas na elaboração da pesquisa, os apêndices e anexos.

# 2 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo, apresentaremos a fundamentação teórica que serviu de base para esta pesquisa e abordaremos os conceitos necessários para o entendimento desta dissertação. Nesse sentido, distribuímos as seções da forma como segue: na seção 2.1, será apresentado um breve panorama da educação a distância no Brasil e em Alagoas; na sequência, abordaremos a motivação para aprender na seção 2.2. Na seção 2.3, faremos uma introdução à teoria de metas de realização. Na seção, 2.4 será apresentada uma introdução à mineração de dados e, logo após, nas seções 2.5 e 2.6, respectivamente, abordaremos o processo de mineração de dados e as técnicas de mineração de dados. Por fim, na seção 2.7, será apresentada a mineração de dados no contexto da educação, conhecida como mineração de dados educacionais.

## 2.1 EDUCAÇÃO A DISTÂNCIA

Historicamente, a educação a distância é caracterizada por um modelo de ensino-aprendizagem em que não há o contato face a face entre professor e aluno. Nesse sentido, há registros do início da EAD no mundo a partir do século XVIII, com o anúncio de aulas por correspondências (VASCONCELOS; GOUVêA, 2010, 2006 apud ALVES, 2011). No entanto, no que diz respeito à legislação brasileira, as bases legais para a modalidade da EAD ocorreram a partir de 1996, com a legalização da oferta de cursos a distância por meio da Lei de Diretrizes e Bases do Ensino Nacional (LDB). A partir disso, houve um aumento na procura por esta modalidade de ensino. Esse crescimento reflete a busca por condições mais flexíveis de acesso à educação, necessidade de um grande contingente da população brasileira.

A EAD ofertada pela Internet prevê a ausência de contato físico frequente entre professor e aluno. O contato acontece via ferramentas de comunicação on-line, síncronas ou assíncronas. A forma síncrona permite a comunicação entre as pessoas em tempo real, onde o emissor envia uma mensagem para o receptor e este a recebe quase que instantaneamente (e.g. chat e videoconferência). Na forma assíncrona é dispensada a participação simultânea das pessoas, e o emissor pode enviar uma mensagem ao receptor, o qual poderá ler e responder esta mensagem em outro momento (e.g. e-mail, fórum e lista de discussão).

São diversos os benefícios associados à EAD, tais como: flexibilidade de tempo, economia no deslocamento até o local de estudos, vários meios de aprendizagem, interação com pessoas de diferentes culturas e experiências profissionais, além da oportunidade de estudar a partir de novas metodologias e tecnologias (ALVES, 2011; ABED, 2014; SANTOS; WECHSLER, 2009).

Os avanços tecnológicos tornaram mais visíveis as possibilidades de desenvolvimento da EAD, favorecendo, ainda no final do século XIV e no início do século XX, a multiplicação

de iniciativas em muitos países da Europa, África e América. Países como Suécia, Inglaterra, França, Canadá e EUA e, mais recentemente, o Brasil, são considerados grandes propulsores da metodologia da educação a distância.

Devido ao crescimento da EAD tanto no sistema formal quanto não formal de ensino, o Ministério da Educação (MEC) criou, em 1996, a Secretaria de Educação a Distância (SEED), com o objetivo de inovar a área tecnológica nos processos de ensino e aprendizagem, promovendo a pesquisa e o desenvolvimento voltados para a introdução de novos conceitos e práticas em escolas públicas brasileiras. A SEED foi criada com a missão de "atuar como agente de inovação dos processos de ensino-aprendizagem, fomentando a incorporação das Tecnologias de Informação e Comunicação (TICs) e da Educação a Distância aos métodos didático-pedagógicos das escolas públicas" (LEITE; DIAS, 2010).

Antes de sua extinção, em janeiro de 2011, a SEED passou a desenvolver numerosos projetos de EAD voltados para os diferentes níveis de ensino. Dentre esses projetos, está a Universidade Aberta do Brasil (UAB), criada para ofertar cursos e programas de educação continuada de nível superior, na modalidade a distância, pelas universidades públicas brasileiras. A UAB não é uma nova instituição de ensino; ela se articula com os governos estaduais, municipais e instituições públicas de Ensino Superior, com a ação prioritária na formação inicial e continuada de professores para a educação básica. Cabe aos estados e municípios a responsabilidade pela implementação e sustentação de seus polos, onde se desenvolvem as atividades presenciais (LEITE; DIAS, 2010).

As buscas por melhores condições de ensino em relação à EAD são um grande avanço para a educação no Brasil. Essa consolidação no meio educacional é importante para alcançar patamares de qualidade, com criatividade e inovação dos cursos a distância, garantindo que os alunos que buscam sua formação por meio desta modalidade possam ter efetividade na sua aprendizagem e formação adequada.

No estado de Alagoas, a Universidade Federal de Alagoas (UFAL) foi pioneira em investir nesta modalidade de ensino, iniciando sua jornada em 1998, no Centro de Educação, através das ações do Programa de Assessoria Técnica aos Municípios Alagoanos (PROMUAL) junto aos municípios alagoanos. Ainda neste mesmo ano, a universidade deu os primeiros passos para a oferta do curso de licenciatura em Pedagogia, sendo este curso o primeiro a ser reconhecido pelo MEC em EAD no Estado.

O ano de 2006 é considerado como um divisor na história da EAD na UFAL, pois as ações deixaram de ser quase que exclusivas do NEAD/Cedu e entraram na ordem do dia de várias unidades acadêmicas e outras áreas, o que contribuiu para aprovação de projetos de polos de apoio presencial e cursos de bacharelado que passaram a funcionar em 2007.

Ainda em 2006, a UFAL em parceria com a UAB e empresas estatais como o Banco do Brasil, em um projeto piloto, ofertou 500 vagas no estado de Alagoas para o curso de

administração a distancia através da Faculdade de Economia, Administração e Contabilidade (FEAC).

Hoje a UFAL conta com 10 polos de educação a distância nas cidades de Maceió, Arapiraca, Maragogi, Olho dÁgua das Flores, Santana do Ipanema, Palmeira dos Índios, São José da Lage, Penedo, Delmiro Gouveia e Matriz do Camaragibe, oferecendo 9 cursos de graduação, 6 cursos de especialização e 2 cursos de aperfeiçoamento (CIED, 2015).

## 2.2 MOTIVAÇÃO PARA APRENDER

A preocupação com a motivação no ambiente educacional tem sido evidenciada cada vez mais, considerando o crescente número de pesquisas sobre o tema nas últimas décadas. Boruchovitch e Bzuneck (2010) afirmam que as pesquisas sobre a motivação escolar têm aumentado mundialmente, sendo que o construto vem sendo estudado sob diferentes abordagens teóricas, mostrando a complexidade do assunto.

Cabanach et al. (1996) comentam que apesar das diferenças existentes entre os vários enfoques teóricos, a maioria considera a motivação como o conjunto de processos que implicam na ativação, direção e persistência da conduta.

Maehr e Meyer (1997) defendem que estudantes motivados são aprendizes permanentes, que, pela vida toda, continuarão a investir na construção de novos conhecimentos. Autores, como Clayton, Blumberg e Auld (2010), afirmam que a motivação tem sido constantemente associada a uma aprendizagem bem sucedida. Ainda segundo os mesmos autores, a compreensão da motivação dos estudantes é o caminho para a promoção da aprendizagem efetiva.

É inegável que os problemas motivacionais podem interferir na aprendizagem dos estudantes. Muitos estudos têm demonstrado a relação entre o sucesso acadêmico e a motivação (BZUNECK, 2004; BZUNECK, 2005). Neste contexto, diversos teóricos apresentam muitas variáveis que podem interferir na motivação do estudante. Entre elas, destacam-se o ambiente da sala de aula, as ações do professor, os aspectos emocionais, as questões relacionadas à falta de envolvimento do aluno com situações de aprendizagem, o uso inadequado de estratégias de aprendizagem, entre outras (ZENORINI; SANTOS; MONTEIRO, 2011).

Nos ambientes educacionais, a motivação aparece como um elemento impulsionador, que ajuda o indivíduo a alcançar um determinado objetivo. Um aluno desmotivado pode apresentar dificuldades para resolver problemas ou tomar decisões acertadas (CUNHA; BORUCHOVITCH, 2012).

Investigações de Accorsi, Bzuneck e Guimarães (2007), Boruchovitch e Bzuneck (2001) e Goya, Bzuneck e Guimarães (2008) apontam que, no caso do aluno, a falta de motivação

para aprender pode se reverter em um baixo desempenho escolar, tendo em vista o pouco investimento no próprio aprendizado. Desta forma, os problemas motivacionais podem ocasionar dificuldades de aprendizagem até em alunos tidos como muito inteligentes, já que a baixa motivação pode gerar descrença quanto a sua própria capacidade de realizar tarefas escolares com sucesso.

Para BRENELLI et al. (2001), a motivação é uma variável-chave para a aprendizagem. Para eles, a motivação para aprendizagem é o ponto de partida e a manutenção de comportamento com o objetivo de se atingir uma determinada meta. Tanto ela como os fatores associados a um bom desempenho têm estado na pauta de educadores e psicólogos.

Vale-se ressaltar que a motivação para aprender vem sendo entendida pelos teóricos contemporâneos como um constructo multidimensional caracterizado por diversas teorias. Dentre estas teorias, encontram-se a Teoria da Autodeterminação, Teoria da Atribuição de Causalidade e Teoria de Metas de Realização. A primeira tem como foco a motivação intrínseca e extrínseca; a segunda, as crenças individuais que influenciam a motivação do individuo para aprender e, a última, as metas ou objetivos que os alunos buscam enquanto aprendem.

Não é nosso objetivo nesse trabalho detalhar essas teorias, entretanto, como a Teoria de Metas de Realização faz parte do enfoque teórico deste trabalho, iremos apresentá-la de modo resumido na seção que se segue.

## 2.3 TEORIA DE METAS DE REALIZAÇÃO

As metas de realização referem-se a proposições ou razões de indivíduos que possuem uma tarefa a realizar. Esse construto parece ser melhor operacionalizado no que diz respeito atividades de aprendizagem acadêmica, embora possa ser aplicado a outros contextos de realização.

Ames (1992) conceituou as “metas de realização” como um conjunto de pensamentos, crenças, propósitos e emoções que traduzem as expectativas dos alunos em relação a determinadas tarefas que deverão executar, ou seja, as metas são representadas por modos diferentes de enfrentar as tarefas acadêmicas. As primeiras pesquisas desenvolvidas nesta perspectiva descrevem apenas dois tipos de metas: Aprender e Performance.

Quando orientado para a meta Aprender, o aluno busca o crescimento intelectual, valoriza o esforço pessoal, enfrenta os desafios, é persistente em relação às atividades acadêmicas e tende a utilizar estratégias de aprendizagem mais efetivas (AMES, 1992; CLAYTON; BLUMBERG; AULD, 2010; MIDDLETON; MIDGLEY, 1997; ZENORINI; SANTOS, 2010). Os alunos com esse tipo de meta compreendem que o sucesso nas realizações acadêmicas consiste em aprimorar os conhecimentos e habilidades, progredir e dominar com criatividade os conteúdos. O esforço empenhado nas atividades promove orgulho e realização, enquanto que as situações de fracassos e erros são estímulos para a busca de novas

estratégias para atingir os objetivos (BZUNECK, 2004).

Por outro lado, os alunos orientados para a meta Performance estão preocupados em demonstrar a sua capacidade para os demais. Estes alunos são menos engajados e evitam desafios. Quando em situação de fracasso, atribuem esse resultado à falta de capacidade e apresentam emoções negativas, tais como vergonha e raiva (ARCHER, 1994; BZUNECK, 2004; ZENORINI; SANTOS, 2010).

Bzuneck (1999) diz que, embora as metas Aprender e Performance tenham características contrastantes, o aluno pode, de forma simultânea e em diferentes graus, apresentar uma orientação para as metas Aprender e Performance. A possibilidade de aspectos positivos da meta Performance foi demonstrada em estudos de Elliot e Harackiewicz (1996), Elliot e Church (1997), Elliot, McGregor e Gable (1999), entre outros, principalmente quando acompanhada da meta Aprender. Estes estudos mostraram em seus resultados dois componentes independentes na meta Performance: Aproximação e Evitação.

Na meta Performance-Aproximação, o aluno busca parecer inteligente e quer estar entre os melhores da classe, enquanto que, na meta Performance-Evitação, o aluno evita qualquer situação que possa mostrar a sua incapacidade. Dessa forma, as pesquisas mais recentes incluem na meta Performance os componentes Aproximação e Evitação.

## 2.4 INTRODUÇÃO À MINERAÇÃO DE DADOS

Desde o surgimento dos computadores, um dos grandes objetivos tem sido o armazenamento de dados. Especialmente, nos últimos anos, com a queda dos custos das tecnologias para transmissão e armazenamento, muitos processos têm sido informatizados, deixando de lado o papel, transformando os materiais anteriormente físicos em digitais.

Atualmente, os dados que, no passado, poderiam ser considerados desnecessários e eliminados pelos custos de armazenamento são armazenados, mesmo que não necessários, no momento, para usos futuros, e, para tal, foram desenvolvidas novas e mais complexas estruturas de armazenamento, tais como bancos de dados, *Data Warehouses*, entre outras.

Um exemplo dessa grande quantidade de dados armazenados são os satélites de observação da NASA que geram cerca de um terabyte de dados por dia, ou os dados do projeto Genoma que são constituídos de milhares de bytes para cada uma das bilhões de bases genéticas (BRAMER, 2007) .

Estima-se que são gerados 2,5 quintilhões de bytes diariamente. Somente no Facebook, são feitas pelo menos 10 milhões de uploads de fotos, no Twitter são gerados em torno de 12 terabytes de informações diariamente e o Google processa, por sua vez, 24 pentabytes de dados todos os dias e esse número não para de crescer (CHEDE, 2013).

Com esse crescente volume de dados armazenados, extrair informações por meio

de técnicas tradicionais pode não ser o meio mais adequado. Neste caso, faz-se necessário encontrar formas de se analisar, classificar, sumarizar, descobrir e caracterizar tendências nesses dados de uma forma automática e relativamente precisa.

Com a finalidade de atacar essa problemática, surgiu, por volta de 1990, a Mineração de Dados, do inglês *Data Mining*. Vários teóricos definiram a Mineração de Dados e, por ser considerada uma área multidisciplinar, as definições variam de acordo com o campo de atuação destes autores. De modo geral, a Mineração de Dados é o processo de exploração de grandes quantidades de dados com o objetivo de encontrar anomalias, padrões e correlações para suportar a tomada de decisões. Essa descoberta é feita por meio do uso de algoritmos para essa finalidade.

Para cada tipo de padrão que se deseja encontrar em um conjunto de dados, podem ser realizadas tarefas de mineração de dados. Estas tarefas podem ser classificadas em descritivas (aprendizado não-supervisionado) e preditivas (aprendizado supervisionado). Segundo Han, Kamber e Pei (2011), as tarefas descritivas caracterizam as propriedades gerais dos dados no repositório; já as preditivas realizam inferências sobre esses dados, com o objetivo de fazer previsões.

O conjunto de funcionalidades de mineração de dados, listadas a seguir, é o resultado da junção do conteúdo sobre o tema, da forma como é abordado em Witten, Frank e Hall (2011), Wu et al. (2007) e Han, Kamber e Pei (2011):

- **Caracterização e Discriminação:** sumariza conceitos e classes de forma concisa, mas precisa.
- **Mineração de Padrões Frequentes, Associações e Correlações:** detecta itens que ocorram com frequência em um conjunto de dados, bem como verifica se existem associações e correlações entre itens.
- **Classificação e Previsão:** gera modelos capazes de distinguir classes e conceitos.
- **Análise de Agrupamentos:** agrupa itens semelhantes cuja classe não é conhecida.
- **Análise de Outliers:** analisa os dados considerados "anormais"e detecta padrões por trás do surgimento desses dados.
- **Análise de Evolução:** analisa a forma como os dados evoluem em busca de padrões relevantes nesse processo.

## 2.5 O PROCESSO DE MINERAÇÃO DOS DADOS

O processo de mineração de dados pode ser dividido em três partes: pré-processamento dos dados, mineração dos dados e pós-processamento dos dados. Esse processo é, muitas

vezes, referenciado como Descoberta de Conhecimento a partir de Dados[1] (HAN, 2011)[2].

Por ser um processo com considerável complexidade, atualmente diversos processos definem e padronizam as fases e atividades da mineração de dados, com estratégias orientadas à solução de problemas. Um destes processos é o CRISP-DM - (*Cross-Industry Standard Process of Data Mining*), uma solução não proprietária e livremente disponível para organização do ciclo de vida de aplicação de mineração de dados em um projeto.

O processo CRISP-DM é composto por seis fases organizadas de maneira cíclica. Além disto, apesar de ser composto por fases, o fluxo não é unidirecional, podendo ir e voltar entre as fases. Na Figura 3 demonstramos uma imagem desse ciclo e a relação entre as suas fases.

1. **Fase de entendimento do negócio:** nessa etapa, o foco é entender qual é o objetivo que se deseja atingir com a mineração de dados, busca-se entender as demandas do projeto do ponto de vista do ambiente no qual ele está inserido. O entendimento do negócio e a definição de um problema de mineração de dados irão ajudar nas próximas etapas do processo.

2. **Fase de entendimento dos dados:** as fontes fornecedoras dos dados podem vir de diversos locais e podem ter diversos formatos. Após definir os objetivos, é necessário conhecer os dados com a finalidade de se descrever, de forma clara, o problema, identificar os dados relevantes para o problema em questão e certificar-se de que as variáveis relevantes para o projeto não são interdependentes (OLSON; DELEN, 2008). Também é nesta fase onde se busca realizar uma primeira análise dos dados, familiarizando-se com algumas de suas qualidades e, ao mesmo tempo, formulando possíveis hipóteses para extrair deles informações relevantes.

3. **Fase de preparação dos dados:** devido às diversas origens possíveis dos dados, nem sempre esses dados estão preparados para que os métodos de mineração sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser necessárias. Essa fase tem por objetivo tratar os dados, através de tarefas de pré-processamento, de forma que o resultado seja o conjunto final de dados, o qual alimentará a ferramenta de modelagem.

4. **Fase de modelagem:** nessa fase, selecionam-se as técnicas (algotimos) de mineração mais apropriadas para o problema e para os dados analisados nas fases anteriores.

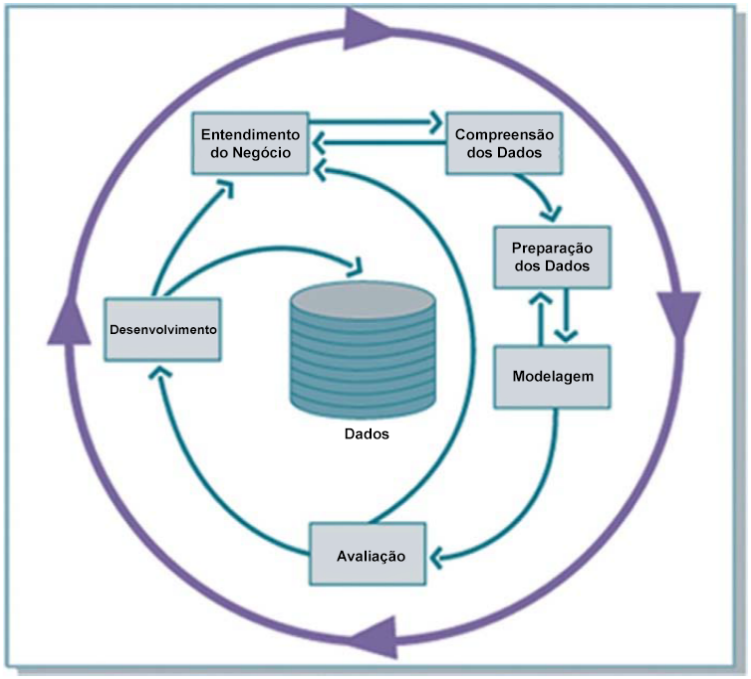
5. **Fase de avaliação:** considerada uma fase crítica do processo de mineração, a meta dessa fase é analisar se os resultados são capazes de atingir os objetivos pretendidos. Busca-se determinar se algum ponto importante para o problema de mineração não foi considerado de forma apropriada.

[1] Em Inglês: *Knowledge Discovery from Data* - KDD

[2] Nessa dissertação utilizaremos o termo mineração de dados nesse sentido amplo, ou seja, como sinônimo do Processo de Mineração de Dados.

6. **Fase de distribuição:** nessa fase, após executado o modelo com os dados reais e completos, organiza-se o conhecimento obtido para que o mesmo seja disponibilizado de uma forma que os envolvidos conheçam os resultados.

Figura 3 – Representação do processo CRISP-DM




Fonte: (NISBET; ELDER; MINER, 2009)

## 2.6 TÉCNICAS DE MINERAÇÃO DE DADOS

Nessa seção, abordaremos as técnicas que utilizaremos para o processo de mineração dos dados. Apresentaremos, de forma geral, as técnicas relacionadas aos processos de regressão, foco deste estudo.

Os processos de mineração de dados possuem duas grandes atividades: a predição e a descrição. No contexto da predição ou mineração de dados preditiva, é utilizada a inferência indutiva para examinar exemplos que possuam algum rótulo e, a partir desses, obter uma generalização que permita a previsão dos rótulos de novos exemplos. Essa atividade de mineração apresenta dois tipos de problemas, de acordo com os valores que esses rótulos assumem: se os rótulos assumem valores nominais (discretos ou categóricos), o problema é denominado classificação, já se os dados assumem valores contínuos, o problema é chamado de regressão (WEISS; INDURKHYA, 1998).

A tarefa de regressão, de um modo geral, consiste na obtenção de um modelo baseado em um conjunto de exemplos que descrevem uma função não conhecida. Este modelo, por sua vez, é utilizado para predizer o valor de um atributo desejado em novos exemplos.

O objetivo da regressão, portanto, é encontrar uma relação entre um conjunto de dados de entrada (conjunto de variáveis de entrada ou variáveis preditoras) e um atributo-meta contínuo (variável de saída ou variável resposta).

Os modelos gerados pelos métodos de regressão possuem diferentes formatos de representação, dado que esses modelos expressam o conhecimento obtido durante o processo de mineração e cada método pode expressar o conhecimento de uma forma diferenciada. Nas subseções seguintes são apresentados, de forma breve, os modelos de regressão e alguns dos principais métodos de regressão utilizados nesse trabalho.

### 2.6.1 Modelos de regressão

Os valores observados de variáveis envolvidas em problemas reais são, de fato, resultados de um experimento que pode ser descrito através de um modelo matemático, através do qual estas variáveis estejam relacionadas.

Este modelo matemático objetiva principalmente reproduzir o verdadeiro processo gerador dos dados. No entanto, em todo processo gerador de dados existem fatores que não podem ser controlados, ou são desconhecidos, os quais podem ser representados pelos erros aleatórios.

Este é o contexto dos modelos de regressão.

Considerando o seguinte modelo linear:

$$y = X\beta + \epsilon =$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}}_{\begin{pmatrix} x_1 & x_2 & \dots & x_k \end{pmatrix}} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Onde, $y$ é um vetor de $n$ observações da variável aleatória dependente (ou de interesse), $X$ é uma matriz $n \times k$ formada pelas covariadas, em que cada coluna é um conjunto de $n$ observações da covariada $x_t$, $t = 1, \dots, k$, totalizando $k$ covariadas. As covariadas não são variáveis aleatórias. Ainda temos $\beta$ que é um vetor de $k$ parâmetros também fixos e desconhecidos (não são variáveis aleatórias) e $\epsilon$ um vetor de $n$ erros aleatórios.

O objetivo de um modelo de regressão é explicar o máximo possível o caráter aleatório da variável resposta. O que não é possível ser explicado deve estar contido no erro aleatório

$\epsilon$. De fato, espera-se que ele seja zero, o que conduz a uma das principais suposições de modelos de regressão:

$$\mathrm{E}(\epsilon) = \mu_\epsilon = 0.$$

Consequentemente,

$$\mathrm{E}(y) = \mathrm{E}(X\beta) + \mathrm{E}(\epsilon) \Leftrightarrow \mathrm{E}(y) = X\beta \Leftrightarrow \mu = X\beta.$$

Logo, o modelo final é

$$\mu = X\beta.$$

Tem-se que $\mathrm{E}(X\beta) = X\beta$ por que nem $X$ nem $\beta$ são variáveis aleatórias (valor esperado de uma constante é uma constante).

Com relação à variância da resposta, segue que

$$\mathrm{var}(y) = \underbrace{\mathrm{var}(X\beta)}_{0} + \underbrace{\mathrm{var}(\epsilon)}_{\sigma^2} \Leftrightarrow \mathrm{var}(y) = \sigma^2.$$

Pois, $\mathrm{var}(\epsilon) = \sigma^2$ e $\mathrm{var}(X\beta) = 0$, variância de uma constante é zero.

A representação do modelo considerando a $i$-ésima observação é dada por

$$\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Para que o modelo acima seja conhecido, é preciso estimar $\beta_1, \beta_2, \ldots, \beta_k$. Tipicamente isto é realizado usando o método de máxima verossimilhança. Assim, $\widehat{m}u_i$ é obtido quando são obtidos: $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$, tal que

$$\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \ldots + \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n.$$

Para usar o método de máxima verossimilhança, é preciso definir o tipo de distribuição de probabilidades que a variável aleatória $y$ (nossa resposta) segue. Apesar da distribuição normal ser a mais conhecida, na prática, essa distribuição não é adequada para diversos tipos de variáveis aleatórias.

- Se $\mu$ (a média da variável resposta) pode assumir tanto valores positivos quanto valores negativos e a curva de densidade de $y$ é próxima da forma de sino, então se justifica pensar na distribuição normal.
- Se $\mu$ só pode assumir positivos e a curva de densidade de $y$ é simétrica positiva, devemos pensar na distribuição gama.
- Se $y \in (0,1)$, podemos pensar na distribuição beta ou na distribuição simplex.

Neste ponto, é necessário generalizar o modelo linear com o objetivo de permitir o uso de outras distribuições além da normal. Isso é feito considerando a expressão:

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}.$$

$\triangleright$ Aqui $g(\mu_i)$ é uma função de ligação, que conecta a média da variável resposta e o modelo envolvendo as covariadas e os $\beta$'s.

Quando $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, temos que $g(\mu_i) = \mu_i$, ou seja, $g$ é o que chamamos de função identidade. Isto acontece no modelo normal porque assim como a resposta que pertence a todos os reais, $y \in (-\infty, +\infty)$, o mesmo ocorre com sua média $\mu \in (-\infty, +\infty) = IR$. O mesmo deve ocorrer com $\widehat{\mu_i}$, $\widehat{\mu_i} \in (-\infty, +\infty) = IR$. Neste caso, da distribuição normal, $\widehat{\mu_i}$ pode assumir qualquer valor real e consequentemente, $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$ estão livres para também assumir qualquer valor.

Isto não acontece, por exemplo, se a variável resposta segue uma distribuição gama, já que $y \in (0, +\infty)$, $\mu \in (0, +\infty) = IR^+$ e $\widehat{\mu} \in (0, +\infty) = IR^+$. Com esta restrição os $\widehat{\beta}$'s não estão livres, pois deve ser garantido que $\widehat{\beta}X$ só assuma valores reais positivos. Como $X$ é fixa, então o processo de estimação do $\beta$'s deve considerar tal restrição, para garantir que $\widehat{\beta}X \in (0, +\infty) = IR^+$, processo que pode ser bastante complicado.

A alternativa é aplicar uma função $g$ em $\mu_i$ de forma que $g(\mu_i) \in (-\infty, +\infty) = IR$. Então, os $\widehat{\beta}$*'s* estão liberados.

### 2.6.2 Regressão Linear dos Mínimos Quadrados

O modelo de regressão linear é uma ferramenta muito interessante e poderosa para análise de dados. Hair et al. (2009) afirmam que "a análise de regressão múltipla é uma técnica

estatística que pode ser usada para analisar a relação entre uma única variável dependente e múltiplas variáveis independentes."

Por meio da regressão linear, é possível estimar o grau de associação entre uma variável dependente e um conjunto de variáveis independentes. Ou seja, nesses modelos, o objetivo é resumir a correlação entre $X_i$ (Variáveis preditoras) e $Y$ (variável dependente), em termos de direção e magnitude.

Um exemplo clássico e bastante utilizado desse tipo de abordagem é o modelo paramétrico global utilizando o critério de erros dos mínimos quadrados. Tecnicamente, dizer que um modelo é ajustado utilizando este critério significa que uma reta que minimiza a soma dos erros quadrados será utilizada para resumir a relação linear entre $Y$ e $X_i$

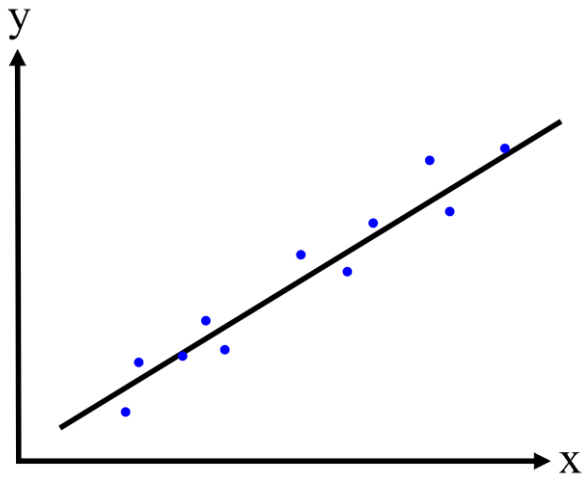
De modo didático, a notação do modelo de regressão linear pode ser dada por:

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

Onde $Y$ representa a variável dependente, aquilo que desejamos explicar. $X_1$ representa a variável independente, aquilo que se acredita que pode ajudar a explicar $Y$. O intercepto ($\alpha$), ou constante, representa o valor de $Y$ quando $X_1$ assume o valor zero. O coeficiente ($\beta$) representa a mudança em $Y$ associada ao incremento em $X_1$ e o termo estocástico ($\varepsilon$) representa o erro ao tentar explicar $Y$ a partir de $X_1$.

A Figura 4 mostra um exemplo de um modelo de regressão linear com dados bem adaptados.

Figura 4 – Exemplo de Regressão Linear




### 2.6.3 Regressão Beta

Os modelos de regressão beta são utilizados normalmente para modelar variáveis aleatórias que assumem valores em um intervalo (0, 1). Eles tomam como base o pressuposto

de que a variável dependente é modelada por uma distribuição beta e que a sua média está relacionada com um conjunto de regressores através de um preditor linear com coeficientes desconhecidos e uma função de ligação.

O modelo de regressão beta foi proposto por Ferrari e Cribari-Neto (2004) e modela a variação contínua que os valores assumem em um intervalo de unidade padrão, por exemplo, taxas, proporções ou índices de concentração. A principal motivação do modelo de regressão beta reside na flexibilidade do fato de se assumir a lei da distribuição beta, cuja densidade pode assumir diferentes formas dependendo da combinação dos valores dos parâmetros. Neste modelo, os parâmetros de regressão são interpretáveis em termos de média $y$ (a variável de interesse) e o modelo é naturalmente heterocedástico e facilmente acomoda assimetrias.

### 2.6.4 Aprendizado Baseado em Exemplos

O aprendizado baseado em exemplos, ou memória, (IBL – *Instance Based-Learning*) consiste na classificação de um exemplar baseado em outro similar cuja classe é conhecida assumindo que o novo exemplo terá a mesma classe (MONARD; BARANAUSKAS, 2003).

Esse tipo de método de aprendizado é denominado *Lazy*[3] e possui três características/componentes muito importantes que devem ser consideradas:

- Conjunto de treinamento armazenado na memória;
- Medida de similaridade que consiste nas métricas para realizar comparações;
- Determinação do número $k$ de exemplos mais próximos que serão utilizados para predição ou cardinalidade do relacionamento entre os exemplos.

Neste tipo de método, todos os exemplos de treinamento são armazenados em memória e não existe uma fase inicial de treinamento, por isso a grande importância do tamanho do conjunto de treinamento, bem como a qualidade dos dados no que se refere a ruídos.

O algoritmo de classificação baseado no vizinho mais próximo (*Nearest Neighbor* – NN) é uma das técnicas mais simples e amplamente empregada no aprendizado baseado em exemplos. O centro de seu funcionamento está em descobrir o vizinho mais próximo de uma dada instância calculando a média dos valores dos atributos-meta dos exemplos do conjunto de treinamento mais similares, entretanto, o algoritmo do vizinho mais próximo simples classifica um caso baseado em um único exemplo similar, o que pode acarretar em erros prematuros.

Já o algoritmo (*k-Nearest Neighbor* – $k$-NN) é uma versão aperfeiçoada do NN. Nele são encontrados os $k$ vizinhos mais próximos do padrão de consulta, ao invés de apenas o

[3] Preguiçoso

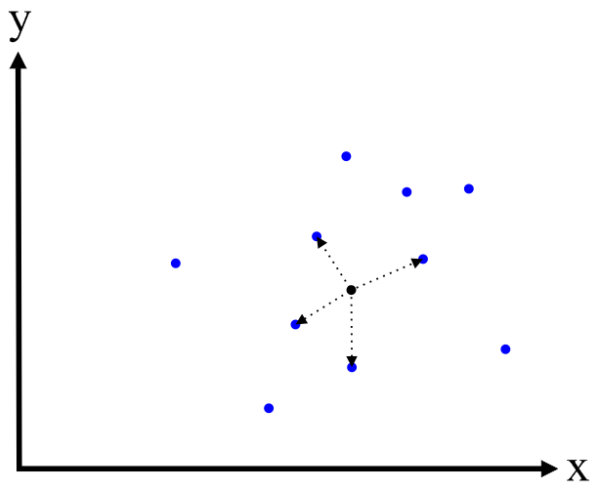
vizinho mais próximo. O $k$-NN classifica um dado elemento de acordo com as respectivas classes dos $k$ ($k \geq 1$) vizinhos mais próximos – pertencentes a uma base de treinamentos dada. O algoritmo calcula a distância do elemento dado para cada elemento da base de treinamento e então ordena os elementos da base de treinamento do mais próximo ao de maior distância. Dos elementos ordenados, selecionam-se apenas os $k$ primeiros, que servem de parâmetro para a regra de classificação.

Neste caso, 1-NN é um $k$-NN onde $k = 1$, ou seja, seleciona-se apenas o elemento do treinamento mais próximo da instância que se pretende classificar. 5-NN vai usar os cinco elementos mais próximos da instância e, baseado nas classes dos cinco elementos, infere-se a classe do atributo-meta.

Uma grande desvantagem do aprendizado baseado em exemplos repousa no fato de que ele não produz modelos que permitam a interpretação do conhecimento obtido a partir de sua execução.

A Figura 5 apresenta a seleção de exemplos mais similares em um conjunto de treinamento.

Figura 5 – Exemplo de vizinhos mais próximos




### 2.6.5 Indução de Regras de Regressão

Um fácil entendimento dos modelos gerados é algo considerado muito importante quando se realiza uma tarefa de regressão. Buscando fornecer soluções mais facilmente interpretáveis, a comunidade de aprendizado de máquina passou a desenvolver métodos de aprendizado simbólico que variam de acordo com a linguagem escolhida para representar as hipóteses (DOSUALDO; REZENDE, 2003).

Uma das linguagens bastante utilizada é a lógica proposicional. A lógica proposicional permite determinar a validade de proposições compostas por meio de conectivos a partir da validade de fatos e da interpretação destes conectivos.

Vários podem ser os conectivos utilizados para compor as proposições. Se o conectivo utilizado é o operador OR, a notação proposicional é denominada Forma Normal Disjuntiva (FND); e, se o conectivo é o AND, então, a notação é chamada de Forma Normal Conjuntiva (FNC) (DOSUALDO; REZENDE, 2003).

As regras de regressão na Forma Normal Conjuntiva são compostas por duas partes: A primeira consiste na parte condicional das regras e a segunda, na parte conclusiva que contém a função para explicar o atributo-meta.

Assim sendo, uma regra de regressão na FNC possui a seguinte forma:

$$\textbf{if} < condição > \textbf{then} < y = f(x_i) >$$

Onde $f(x_i)$ é uma função de regressão como as já apresentadas nesse trabalho, com suas variáveis preditoras e $< condição >$ são as condições da regra que estão relacionadas a um atributo preditor ($x_i$), um operador ($op \in \{=, \neq, <, \leq, >, \geq\}$ e um valor constante válido para o atributo em questão.

A principal diferença entre uma regra de decisão e uma regra de regressão está no fato de que, em sua parte conclusiva, as regras de decisão tratam de atributos-meta discretos, enquanto que, nas regras de regressão, os atributos-meta são contínuos.

### 2.6.6 Indução Top-Down de Árvores de Regressão

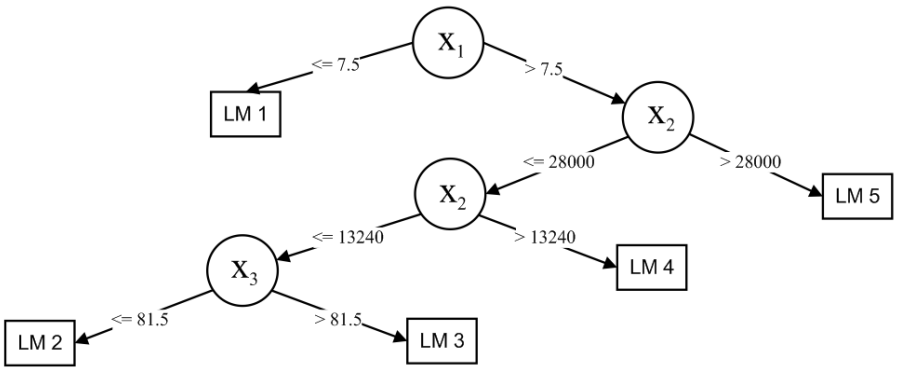
Uma Árvore de Regressão é uma estrutura hierárquica de nós e arestas utilizada para solução de problemas onde o atributo-meta é contínuo e apresenta em seus nós folhas funções matemáticas, o que já caracteriza a sua diferença para as árvores de decisão onde os atributos-meta são discretos.

Estas árvores são construídas baseadas em uma estratégia gulosa, *top-down* e com particionamento recursivo. Deste modo, partindo do nó raiz, são realizados testes lógicos sob determinados atributos particionando a árvore em ramos com novos nós. O processo é então repetido até que seja alcançado um último nó (nó folha - variável dependente ou função de regressão linear) onde o valor da predição é alcançado.

Há dois tipos de árvores que predizem atributos-meta contínuos, as Árvores de Regressão (*Regression Tree*) que guardam em seus nós folhas as médias dos valores presentes nestes nós, e as Árvores Modelo (*Model Tree*) que guardam planos de regressão linear em seus nós folhas.

A Figura 6 apresenta um exemplo de Árvore Modelo (*Model Tree*). Esta árvore foi construída usando três atributos de entrada ($x_1$, $x_2$ e $x_3$) de um determinado conjunto de dados. Os nós folhas, LM 1, LM 2, LM 3, LM 4 e LM 5, representam os modelos lineares gerados em cada um destes nós.

Figura 6 – Exemplo de Árvore Modelo (*Model Tree*)

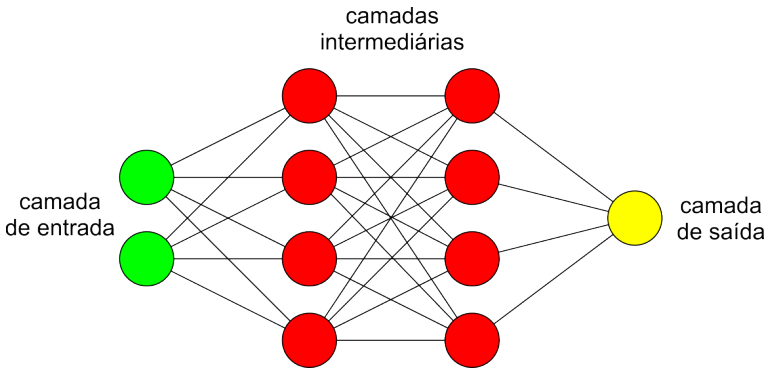



### 2.6.7 Redes Neurais Artificiais

As Redes Neurais Artificias (RNAs) são modelos computacionais inspirados no funcionamento do cérebro humano. Elas são compostas por várias unidades de processamento (neurônios), interligados por um grande número de conexões (sinapses).

Graficamente o modelo é composto de nós (que atuam como entradas, saídas ou processadores intermediários) conectados ao próximo conjunto de nós por uma série de arcos (caminhos ponderados), semelhantes aos pesos em um modelo de regressão (HAYKIN, 2001). Uma representação genérica de uma RNA pode ser vista na Figura 7.

Figura 7 – Exemplo de uma Rede Neural Artificial




Similar às redes biológicas, uma RNA pode ser organizada de vários modos diferentes (topologias), isto é, os neurônios podem ser conectados de vários modos. Portanto, as RNAs aparecem representadas por diversas possíveis configurações. No processamento da informação, muitos dos elementos da rede executam seus cálculos ao mesmo tempo. Este processamento paralelo assemelha-se ao modo como o cérebro trabalha e difere do processamento serial dos cálculos tradicionais.

Apesar de as RNAs terem como vantagens a possibilidade de não estarem restritas a um único atributo de saída, possuírem alta precisão na predição dos valores e robustez diante de dados com ruídos, elas também possuem desvantagens, como por exemplo, dependendo

do modelo de rede e do algoritmo usado, elas podem apresentar lenta convergência, além do fato de que as soluções fornecidas não são de fácil interpretação pelos usuários, visto que o conhecimento está embutido nos pesos e conexões da rede.

### 2.6.8 Support Vector Machines

SVM (*Support Vector Machine*) é um método de aprendizado de máquina usado para tarefas de classificação e regressão de dados lineares e não lineares. O princípio das SVMs consiste em encontrar um hiperplano ótimo que separe membros e não membros de uma classe em um espaço abstrato, denominado *feature space*. Nesse espaço, as classes presentes no conjunto de treinamento se tornam linearmente separáveis e o hiperplano ótimo é definido como aquele para o qual a margem de separação entre as mesmas é maximizada (DOSUALDO; REZENDE, 2003).

No método de SVM para regressão, a ideia básica é encontrar uma função que aproxima bem os pontos de treinamento por meio da minimização dos erros de predição.

As máquinas de vetores-suporte normalmente têm a habilidade de resolver problemas de classificação de padrões de modo próximo ao ótimo, entretanto, assim como as Redes Neurais, geram modelos que não explicam de forma explícita e clara o processo pelo qual sua saída é obtida o que o torna sua compreensão dificultada.

## 2.7 MINERAÇÃO DE DADOS EDUCACIONAIS

A mineração de dados tem se tornado uma importante ferramenta para diversas áreas do conhecimento e uma delas é a educação. Desde meados de 2005, tem-se notado o aumento do interesse e da aplicação da mineração de dados na educação (ROMERO; VENTURA, 2007). Abaixo, na tabela 1, listamos 10 áreas onde técnicas de mineração de dados são aplicadas. Essas áreas correspondem às mais votadas na enquete: Onde você aplicou Técnicas Analíticas/Mineração de Dados em 2012? [4] [5]

Notamos o destaque no uso da mineração de dados na educação, que é o contexto de interesse nessa dissertação. Notamos que seu percentual de votos está próximo ao de áreas onde a aplicação de técnicas de mineração de dados já é tradicional, como saúde, varejo e operações bancárias.

Isto se deve, em parte, pelo recente aumento da oferta de cursos a distância mediados por ambientes com suporte computacional, onde pesquisadores vêm adaptando métodos provenientes da mineração de dados (BAKER, 2010) e aplicando-os à grande quantidade de

[4] Em Inglês: Where did you apply Analytics/Data Mining in 2012?

[5] Realizada pelo site KDNuggets entre 26 de Novembro e 11 de Dezembro de 2012 (NUGGETS, 2012)

Tabela 1 – Resultado da enquete Onde você aplicou técnicas analíticas/mineração de dados em 2012.

| Posição | Área de Aplicação | Percentual de votos |
|---|---|---|
| 1º | Relacionamento com Clientes | 28.6% |
| 2º | Saúde | 16.3% |
| 3º | Varejo | 14.8% |
| 4º | Operações Bancárias | 14.3% |
| **5º** | **Educação** | **14.3%** |
| 6º | Propaganda | 13.3% |
| 7º | Detecção de Fraudes | 12.8% |
| 8º | Mídias Sociais / Redes Sociais | 12.2% |
| 9º | Ciências | 11.7% |
| 10º | Finanças | 10.2% |

Fonte: http://www.kdnuggets.com/polls/2012/where-applied-analytics-data-mining.html.

dados oriundos desses ambientes, dando origem a uma nova área de pesquisa, a Mineração de Dados Educacionais - EDM[6].

A EDM utiliza conceitos e técnicas da mineração de dados tradicional com o objetivo de aprimorar os processos de ensino e da aprendizagem, bem como a gestão desses processos, com base em dados de meios educacionais. Dados desse contexto podem ser provenientes de sistemas tutores inteligentes, sistemas educacionais clássicos baseados em computação, dados administrativos da própria escola, testes padronizados, dentre outros (ROMERO et al., 2010).

De acordo com Romero et al. (2010), a EDM vem sendo utilizada para a obtenção de diversos objetivos educacionais. Essas tarefas podem ser agrupadas nas seguintes categorias:

- **Comunicação com *stakeholders*:**[7] visa prover auxílio a administradores de cursos e professores para avaliar as atividades realizadas pelos alunos, bem como a participação dos mesmos no curso. Mineração de processos, geração de relatórios, visualização de dados e a análise estatística de dados, são as técnicas mais utilizadas para esse grupo de aplicações.

- **Realizar melhorias e manutenções em cursos:** tem como objetivo ajudar gestores de curso e educadores sobre quais estratégias utilizar para obter melhorias. Associação, agrupamento e classificação. São as técnicas mais usuais nesse grupo.

- **Gerar recomendações:** objetiva recomendar conteúdo apropriado para o momento educacional vivenciado pelo estudante. Segundo Brito et al. (2012), as recomendações devem atender às necessidades dos alunos e levar em consideração seu nível de

[6] Em Inglês: Educational Data Mining - EDM

[7] Stakeholders, nesse contexto, se refere a pessoas diretamente interessadas no ambiente educacional e no desempenho dos alunos.

conhecimento. Associação, sequenciação, agrupamento e classificação, são as técnicas mais utilizadas nesse grupo.

- **Prever resultados de atividades/provas ou de avaliações de aprendizado:** busca antever o resultado de testes e de outras avaliações educacionais, com base na análise das atividades realizadas pelos estudantes. Mais uma vez, as técnicas de associação, agrupamento e classificação, são as mais utilizadas.

- **Criar modelos de alunos:** o objetivo é estudar determinadas características dos alunos. Para esse tipo de aplicação, há uma demanda maior e as técnicas mais utilizadas são: Análises estatísticas, redes Bayesianas, modelos psicométricos e aprendizado por reforço.

- **Análise da estrutura do domínio:** busca avaliar a estrutura do domínio (por exemplo: do ambiente de aprendizado), analisando seu desempenho, ou seja, como o domínio realiza uma determinada tarefa (por exemplo: quão eficiente é a detecção de desistentes do ambiente avaliado?). As técnicas mais utilizadas incluem regras de associação, métodos de agrupamento e algoritmos de busca.

# 3 TRABALHOS RELACIONADOS

Neste capítulo são abordados os trabalhos que consideramos relacionados a esta pesquisa. Foram pesquisados trabalhos que abordaram a classificação da motivação de estudantes no contexto da educação on-line.

Inicialmente são apresentados os trabalhos, com uma breve descrição sobre eles, relatando-se no que a proposta desse trabalho difere de cada um deles. Ao final do capítulo, é apresentada por meio da Tabela 2 uma sumarização das comparações.

## 3.1 MOTIVATIONAL PROFILES OF ADULT LEARNERS IN ONLINE AND BLENDED LEARNING

O objetivo do estudo realizado por pesquisadores da *Vrije Universiteit Brussel*[1] e da *Universiteit Gent*[2] foi analisar a existência de perfis motivacionais entre alunos no contexto da *online and blended learning*[3] (OBL) na educação de adultos (VANSLAMBROUCK et al., 2015).

Para atingir este objetivo, inicialmente, foi aplicado um questionário, tendo como alvo os alunos inscritos em um programa de OBL na educação de adultos e, posteriormente, foi realizada uma análise de *cluster*[4] das pontuações dos participantes retiradas do questionário.

Os participantes foram 180 alunos de cursos on-line ou mistos. 65% dos participantes eram do sexo feminino. 62,8% dos participantes estavam inscritos em uma formação de professores, 28,9% eram do ensino secundário adulto e apenas 8,3% estavam inscritos no ensino superior profissional de adultos.

Inicialmente, foi desenvolvido um questionário a fim de recolher informações dos alunos sobre sua formação, suas situações sociodemográfica e socioeconômica e as suas características psicológicas, tais como a sua motivação para aprender.

Para coleta dos dados relativos à motivação para aprender, foi utilizada a *Academic Motivation Scale* (AMS) de Vallerand et al. (1992). O instrumento foi traduzido para o holandês e os dados foram submetidos a uma análise fatorial confirmatória para testar a validade da escala.

Após a validação e aplicação do instrumento, foi conduzida a análise dos dados. Inicialmente, foi realizada a análise de normalidade para cada uma das variáveis, bem como a verificação da ocorrência de valores atípicos para evitar distorções na formação

[1] Disponível em: http://www.vub.ac.be
[2] Em português: Universidade de Gante, disponível em: http://www.ugent.be
[3] Em português: Aprendizagem on-line e mista
[4] Em português: Grupo

dos *clusters*. Em seguida, foi realizada a análise do agrupamento em função das características motivacionais dos alunos utilizando um procedimento em duas etapas: a primeira etapa consistiu na realização do agrupamento hierárquico para explorar o número de grupos que surgiram naturalmente usando o método de Ward e distância euclidiana quadrada; o segundo passo utilizou-se o procedimento *K-means* para efetivamente formar os agrupamentos.

Os resultados da análise descritiva detectaram 8 valores extremos que foram excluídos da amostra. Todos os participantes tiveram uma amotivação muito baixa (média = 1,57, DP = 0,68) e pontuaram alto em motivação intrínseca (média = 3,95, DP = 0,83) e em regulação identificada (média = 3,98; DP = 0,67).

Para a formação dos *clusters*, tomando como base os coeficientes de aglomeração encontrados anteriormente, foram testadas soluções utilizando 5-, 4- e 3-*clusters*, sendo que a melhor solução encontrada foi a de 3-*clusters*. Para rotular os grupos, foram utilizadas características de cada sub-escala, sendo estes batizados de “extrínsecos”, “autônomos” e “motivados”.

Também foram realizados testes para verificar a relação entre os *clusters* e as variáveis base. As variáveis idade, estado civil, emprego e experiência de OBL não influenciaram a associação dos *clusters* dos alunos. Um teste qui-quadrado de associação revelou resultados significativos para as variáveis "gênero", “mais alto grau atingido”, "nível educacional"e "escolaridade".

Por fim, uma análise de regressão logística multivariada foi realizada para identificar a influência das variáveis gênero, “mais alto grau atingido” e nível de escolaridade sobre a associação do *cluster*. O modelo mostrou um coeficiente de determinação (R2) de 0,23 (Nagelkerke), o que significa que 23% dos membros dos *clusters* podem ser explicados pelas três variáveis independentes.

Este trabalho apresenta a classificação do perfil motivacional dos alunos da educação on-line ou mista tomando como base um questionário e a definição de *clusters* em função das respostas deste questionário. Está fundamentado nos estudos da psicologia para a classificação da motivação e apresenta um processo de validação em sua etapa de tradução do instrumento original desenvolvido por Vallerand et al. (1992), entretanto, não é possível realizar esta classificação futuramente de modo automático, ou seja, sem a aplicação dos questionários.

## 3.2 STUDENTS' MOTIVATION FOR LEARNING IN VIRTUAL LEARNING ENVIRONMENTS

O estudo desenvolvido por pesquisadores da Universidade Estadual de Londrina[5] buscou identificar a motivação dos alunos para a aprendizagem em ambientes virtuais de aprendizagem. Para tanto, utilizou-se a Escala de Estratégia de Ensino, de Aprendizagem e Motivação para Aprender em Ambientes Virtuais de Aprendizagem - EEAM-AVA (BELUCE; OLIVEIRA, 2015).

Para a realização da pesquisa, foram selecionados 572 alunos matriculados em cursos de graduação e pós-graduação mediados através de ambientes virtuais de aprendizagem. As mulheres representavam 95,8% (n = 548) e os homens de 4,2% (n = 24).

A média de idade dos estudantes foi de 40 anos e oito meses (DP = 7,96), com a idade mínima de 23 anos de idade e a máxima de 67 anos. Os alunos eram do último ano do curso de graduação em Pedagogia - Grupo 1 (n = 544; 95,1%), de um curso de extensão universitária em História - Grupo 2 (n = 7; 1,2%) e de formação contínua para professores a partir de uma rede municipal de ensino - Grupo 3 (n = 21; 3,7%). As amostras foram selecionadas por conveniência.

A Escala de Estratégia de Ensino, de Aprendizagem e Motivação para Aprender em Ambientes Virtuais de Aprendizagem - EEAM-AVA é composta por 32 itens com uma estrutura de seis dimensões, a saber: estratégias de ensino (9 itens), motivação autônoma (5 itens), motivação controlada (6 itens), desmotivação (4 itens), estratégias cognitivas e metacognitivas de aprendizagem (6 itens) e monitoramento da aprendizagem (2 itens). As alternativas usam uma escala de Likert de três pontos estabelecidos como "sempre", "às vezes" e "nunca". O valor 2 foi atribuído à opção "sempre", o valor de 1 para a opção "às vezes" e o valor 0 para a opção "nunca".

Para o estudo, apenas os itens da escala correspondentes à motivação para a aprendizagem foram analisados. Os dados foram coletados por meio da aplicação do instrumento disponibilizado na web em 2013. Após a análise, foram obtidos resultados relativos às dimensões motivacionais pesquisadas.

Para a dimensão da motivação autônoma, criada com 5 itens da escala e com um total de pontos que poderia variar entre 0 e 15, as taxas indicaram uma pontuação máxima de 10 (n = 346, 60,5%), uma pontuação mínima de 2 pontos (n = 1, 0,2%) e uma pontuação média de 9,22 (DP = 1,24). Os resultados também revelaram que 75,9% (n = 434) dos estudantes selecionaram a opção "sempre" para as perguntas que tratavam da motivação intrínseca dos estudantes para participar de cursos on-line / disciplinas. Constatou-se também que 2,9% (n = 17) dos estudantes escolheram a opção "nunca" referente a essas mesmas perguntas. Também foi dada ênfase às taxas relativas à motivação extrínseca através da regulação integrada, que

[5] Disponível em: http://www.uel.br

obtiveram o reconhecimento de 91,8% (n = 525) dos participantes.

A dimensão da motivação controlada, constituída por seis itens, obteve uma pontuação que variou de 0 a 18 pontos e uma média de 5,0 (DP = 2,62). A análise estatística evidenciou, para esta dimensão, a pontuação máxima de 12 (n = 4; 0,7%) e a pontuação mínima de 0 (N = 21; 3,7%).

Os resultados também indicaram que 35,8% (n = 204) dos alunos selecionaram a opção "nunca" para as questões que apresentaram exemplos de comportamentos regulados por motivação controlada. Em contraste com isso, 20,3% (n = 116) dos estudantes selecionaram a opção "sempre" para as perguntas que caracterizaram comportamentos regulados pela motivação extrínseca do tipo externo ou introjetado.

Números significativos também foram encontrados nos resultados decorrentes da análise dos dados referentes à dimensão da desmotivação. Esta dimensão foi composta por 4 itens, com escores variando entre 0 e 12, e os resultados alcançados indicaram uma pontuação máxima de 8 (n = 2, 0,3%), uma pontuação mínima de 0 (n = 449, 78,5%) e uma média significativa de 0,40 (DP = 0,95).

Os resultados também indicaram que 1,6% (n = 9) dos estudantes selecionaram a opção "sempre" para a dimensão da desmotivação, enquanto 91,5% (n = 532) selecionaram a opção "nunca" para declarações que descrevem comportamentos de desmotivação para aprender em situações de ensino mediadas por AVAs.

Este trabalho identifica a motivação para aprender de alunos em ambientes virtuais de aprendizagem por meio da aplicação de um questionário on-line. Apesar de o estudo apresentar um instrumento para a classificação da motivação dos estudantes com uma forte fundamentação em teorias da psicologia, em especial a teoria da autodeterminação (RYAN; DECI, 2000), o mesmo não apresenta um modelo que permita que essa classificação seja realizada de modo automático. Além disso, o trabalho apresenta indícios da validação do instrumento utilizado, entretanto não apresenta este processo de modo claro.

## 3.3 DEVELOPING A LOG-BASED MOTIVATION MEASURING TOOL

O estudo realizado no *Knowledge Technology Lab*[6] da Escola de Educação da Universidade de Tel Aviv, em Israel, teve como objetivo a construção de um framework conceitual e uma ferramenta para medir a motivação dos alunos (HERSHKOVITZ; NACHMIAS, 2008).

O framework sugerido na primeira fase foi construído a partir do conhecimento levantado em estudos prévios sobre reconhecimento de motivação baseado na interação aluno-computador e considera três dimensões:

(a) Engajamento - relaciona-se com a intensidade motivação;

[6] Laboratório de Tecnologia do Conhecimento, disponível em: http://muse.tau.ac.il/

(b) Energização - que se refere ao modo como a motivação é preservada e dirigida;

(c) Fonte de motivação (interna ou externa).

Na segunda fase, a fim de escolher e definir as variáveis relacionadas à motivação, foram utilizados como ferramenta principal de investigação *Learnograms* - representações visuais de variáveis de aprendizagem ao longo do tempo. Ao final desta fase, sete variáveis foram identificadas.

Já na terceira e última fase foi realizada a classificação das variáveis em função das dimensões de motivação propostas no framework. Para isso foi realizado um estudo empírico onde foram coletados logs de uma grande população (N = 2162), um filtro foi aplicado para manter os alunos com pelo menos 3 seções ativas (n = 1444); o conjunto de dados então foi pré-processado e foi definido o conjunto final de casos a serem analisados (N = 674). Finalmente, foi aplicado um *clustering*[7] hierárquico das variáveis usando SPSS com a distância de correlação de Pearson como a medida e relação entre-grupos como o método de agrupamento. Como resultado, as 7 variáveis foram classificadas em uma das três dimensões do framework proposto.

Este trabalho apresenta uma ferramenta que permite medir a motivação dos alunos usando apenas informações armazenadas em arquivos de log, tomando como base um framework definido pelos próprios autores. Entretanto, esse framework, por sua vez, não foi validado; não apresenta escalas claramente definidas e não possui ligações diretas com teorias psicológicas atuais que abordam o constructo da motivação.

## 3.4 ELICITING MOTIVATION KNOWLEDGE FROM LOG FILES TOWARDS MOTIVATION DIAGNOSIS FOR ADAPTIVE SYSTEMS

No trabalho realizado por pesquisadores da *National College of Ireland*[8], os autores estavam interessados no diagnóstico da motivação e na construção de um modelo de usuário de motivação dos alunos, com foco na elicitação do conhecimento da motivação a partir de arquivos de log (COCEA; WEIBELZAHL, 2007).

Com esta finalidade, foi proposta uma abordagem em duas etapas para o diagnóstico da motivação: na primeira etapa, o sistema monitora os alunos e detecta os que não estão engajados a partir dos arquivos de log; na segunda etapa, os alunos considerados desengajados serão envolvidos em um diálogo a fim de se detectar a sua autoeficácia, autorregulação (conceitos da Teoria Social Cognitiva) e outros conceitos relacionados à motivação. Entretanto, no trabalho, somente são apresentados resultados da primeira etapa.

[7] Técnica de mineração de dados para fazer agrupamentos automáticos de dados segundo seu grau de semelhança.

[8] Em Português: Faculdade Nacional da Irlanda, disponível em: https://www.ncirl.ie/

Para execução da primeira etapa, foram utilizadas ações e *timestamps*[9] registrados nos arquivos de log para classificar o nível de envolvimento do usuário. Foram criados vários subconjuntos de logs e estes, por sua vez, foram analisados por especialistas.

Para cada sequência de 10 minutos de interação, um valor foi atribuído pelos avaliadores mediante critérios pré-definidos: engajado, neutro ou desengajado. Os valores foram atribuídos em duas etapas, sendo a primeira uma avaliação informal e a segunda conduzida pelos especialistas. Os testes estatísticos de concordância entre os avaliadores mostraram bons resultados de média de concordância (92%), bem como de concordância Kappa (0,826 (p <0,01)) e alfa de Krippendorff (0,8449) o que indicou uma alta confiabilidade entre avaliadores.

Para as análises, foi utilizada a ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) e foram criados, a partir de 943 entradas obtidas, três conjuntos de dados diferentes: 1) todos os 30 atributos exceto ID de utilizador, chamado DS-30; 2) 10 atributos relacionados aos seguintes eventos: lendo páginas, testes, hyperlinks e glossário, chamado DS-10 e 3) seis atributos relacionados apenas para ler páginas e testes.

Foram utilizados 8 métodos diferentes e todos mostraram bons resultados de predição variando aproximadamente entre 84% e 88%, sendo que os melhores resultados foram obtidos via classificação por regressão em todos os *datasets*[10].

Apesar do modelo apresentado ter como foco a classificação da motivação, acreditamos ser um modelo limitado, pois foca somente em um aspecto motivacional: o nível de envovlvimento do usuário, bem como depende diretamente de uma avaliação – subjetiva – por parte de um especialista externo; além disso, apesar de mencionar a Teoria Social Cognitiva, o trabalho não chega a fazer uso dela.

## 3.5 A WWW-BASED LEARNER'S LEARNING MOTIVATION DETECTING SYSTEM

A intenção dos autores, pesquisadores da *University of Aizu*[11] foi criar um método para compreender os estados psicológicos dos alunos em um sistema de aprendizagem baseada na web usando a técnica de análise de fatores. Para tal, foram capturados estados psicológicos dos alunos, mais precisamente a motivação, descrita pelo modelo ARCS, sendo esta motivação analisada usando o histórico de aprendizado dos alunos (ZHANG; CHENG, 2003).

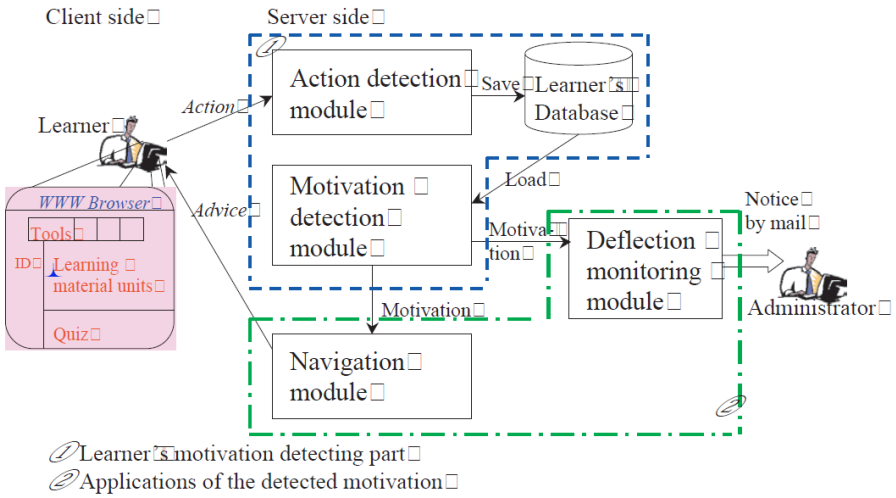
Em meio a teorias e modelos de motivação apresentados por diversos teóricos, os autores decidiram utilizar o modelo ARCS, pelo fato de, segundo eles, o mesmo está

[9] Em Português: Marca Temporal
[10] Em Português: Conjunto de dados
[11] Em português: Universidade de Aizu, disponível em: http://www.u-aizu.ac.jp/e-index.html

intimamente relacionado às ações de aprendizagem e pelo fato de mais facilmente detectado e usado. O modelo ARCS foi proposto por John M. Keller com base na teoria da Expectativa de Valor e utiliza quatro fatores para descrever a motivação, a saber: Atenção, Relevância, Confiança e Satisfação (KELLER, 1987).

O sistema de detecção de motivação proposto é constituído por 5 módulos, conforme exibido na Figura 8.

Figura 8 – Arquitetura do Sistema de Detecção de Motivação proposto por Zhang




Fonte: (ZHANG; CHENG, 2003)

1. Módulo de coleta de ação: Neste módulo, as ações de aprendizagem do aluno e situações de progresso durante a leitura de materiais de aprendizagem ou realização de exercícios são detectados. Em seguida, os dados detectados são armazenados no banco de dados do aluno. Estes dados serão utilizados para a análise da atenção e confiança do aluno.

2. Módulo de detecção de motivação: De acordo com as ações recolhidas do aluno, sua atenção e confiança são analisados neste módulo.

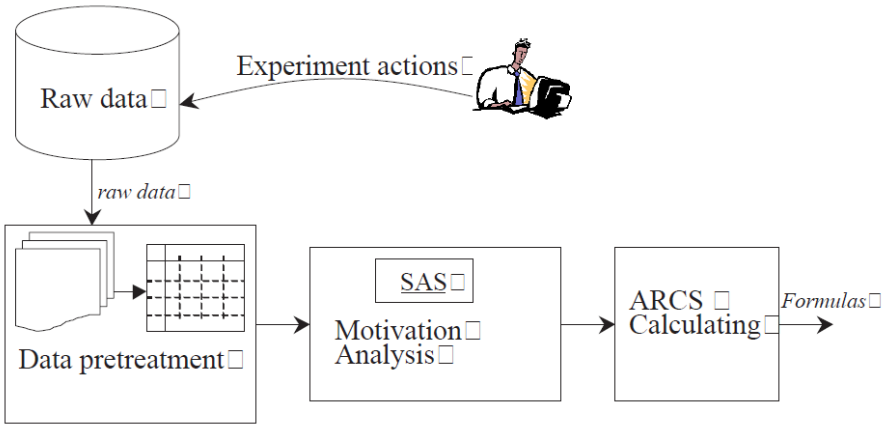
3. Base de dados do aprendiz: Este é um conjunto de dados detectados do aluno. É a base da análise de motivação.

4. Módulo de navegação: De acordo com os resultados analisados no módulo de detecção de motivação, neste módulo conselhos são dados para navegação do aluno a fim de promover a motivação para aprender.

5. Módulo de monitoramento de desvio: Este é um monitor da função de detecção de motivação. Para diferentes grupos de alunos, critérios de detecção de motivação devem ser diferentes quando o desvio dos valores detectados dos fatores de motivação são maiores do que o valor projetado; um e-mail de alerta é enviado para o administrador/professor deste módulo.

A partir desta arquitetura, foi proposto o método para detecção de motivação que consta de três passos descritos a seguir e cujo fluxo da análise de motivação pode ser visto na Figura 9.

Figura 9 – Fluxo de análise de motivação proposto por Zhang




Fonte: (ZHANG; CHENG, 2003)

1. Realização de experimento para coletar as ações do aluno;
2. Análise dos dados do experimento para se obter os fatores de motivação usando o emprego de análise fatorial e, em seguida, encontrar uma fórmula para calcular a motivação do aluno a partir de ações;
3. Execução das fórmulas obtidas na etapa anterior pelo sistema rodando em tempo real, para detectar a motivação do aluno em tempo hábil.

Os dados obtidos para o experimento foram coletados a partir do estudo de programação da linguagem basic realizado por 20 alunos do ensino médio. Os alunos foram solicitados a aprender livremente o conteúdo de um texto que continha explicações, exemplos e tarefas. A partir de então, os alunos testam os exemplos utilizando o ambiente VLB1 e, em seguida, tentam realizar as tarefas. No processo de aprendizagem, as ações de aprendizagem e um *timestamp* são salvos como dados brutos. Foram usados os dados brutos de 10 alunos para a análise fatorial.

Após uma etapa de pré-processamento dos dados, foram obtidos 6 itens que serviram de entrada para a análise fatorial. A partir da análise, foram selecionados seis itens cuja carga fatorial de cada fator foi maior do que 0,40. O primeiro fator pode ser descrito como atenção (Itens $I1$ a $I4$) e o segundo como confiança (Itens $I5$ e $I6$).

De um modo geral, o trabalho apresenta um método para compreender a motivação do aluno, bem como mostra um sistema para aplicar este método. Entretanto, apesar de utilizar como base o modelo ARCS, somente dois fatores do modelo foram utilizados

para detecção da motivação dos alunos (Atenção e Confiança). Além disso, o trabalho não apresenta um processo de validação do método de detecção da motivação.

## 3.6 TABELA COMPARATIVA

Nesta subseção, exibimos a tabela criada com o propósito de sumarizar as características dos modelos apresentados nos trabalhos relacionados, bem como compará-los com o modelo proposto nesta dissertação (tabela 2). Nesta sumarização, consideramos as seguintes características: o trabalho apresenta um modelo que permita a classificação automática da motivação dos alunos?; o trabalho usa como arcabouço teórico alguma das teorias da psicologia para motivação?; e, por fim, o trabalho apresenta algum processo de validação?

Tabela 2 – Tabela comparativa dos trabalhos relacionados

| | Apresenta um modelo para classificação automática? | Usa como arcabouço teorias da psicologia? | Apresenta processo de validação? |
|---|---|---|---|
| Modelo Proposto | SIM | SIM | SIM |
| Vanslambrouck et al., 2015 | NÃO | SIM | SIM |
| Beluce & Oliveira, 2015 | NÃO | SIM | NÃO |
| Hershkovitz & Nachmias, 2008 | NÃO | NÃO | NÃO |
| Cocea & Weibelzahl, 2007 | SIM | NÃO | SIM |
| Zhang & Cheng, 2003 | SIM | SIM, PARCIALMENTE | NÃO |

# 4 PROPOSTA

A proposta apresentada nesta dissertação é a criação de um modelo para **classificação da motivação de estudantes**, gerado com o auxílio de **Instrumentos Psicométricos** (i.e.: questionários) e **Mineração dos Dados** oriundos das interações dos alunos (i.e.: dados educacionais) com o ambiente de aprendizagem. Este modelo permite que os psicólogos, professores e tutores de cursos ofertados, por meio da educação on-line, utilizando o Moodle, possam realizar o processo de classificação do perfil motivacional de seus alunos de modo mais rápido e menos custoso.

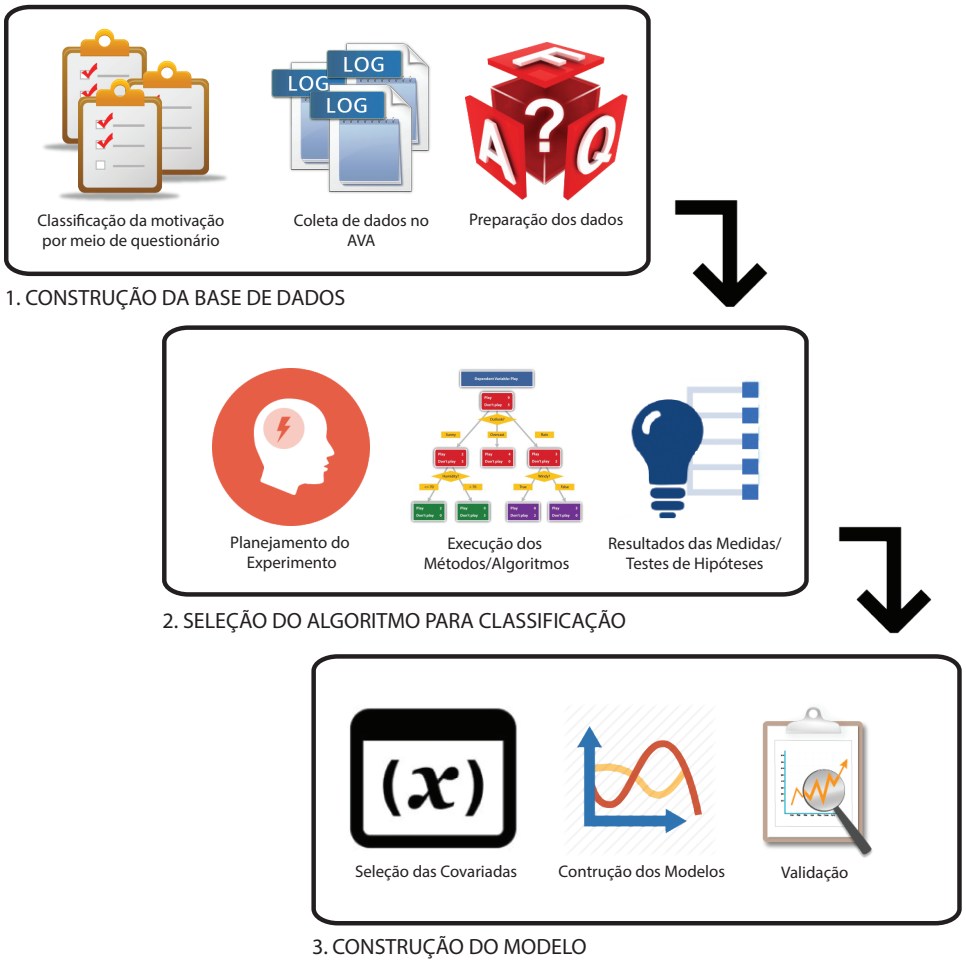
A proposta visa atender às questões de pesquisa (Seção 1.2). A aplicação de instrumentos de avaliação psicológica tem sido evidenciada por muitos autores para a identificação do perfil motivacional de alunos, entretanto as aplicações desses instrumentos podem ser bastante custosas e possíveis reavaliações podem demandar um tempo significativo para que seja reduzido o viés nos dados coletados. Neste cenário, a mineração de dados educacionais pode ser uma ferramenta para uma busca nos conjuntos de dados dos ambientes educacionais permitindo uma compreensão mais adequada dos alunos. Por exemplo, é possível verificar se o aluno está desmotivado ou confuso (BAKER; ISOTANI; CARVALHO, 2011).

Desta forma, justifica-se o uso da mineração de dados educacionais, o que atende à nossa primeira questão de pesquisa. Esta mineração, por consequência, irá produzir um ou mais modelos computacionais que permitam uma classificação automática da motivação dos estudantes, atendendo, também, a segunda questão de pesquisa. Por fim, a mineração de dados pode ser mais eficaz para o objetivo proposto, se tiver como base as teorias e os instrumentos desenvolvidos pela psicologia. O uso desse arcabouço atende a nossa terceira questão de pesquisa.

Nas seções 4.1, 4.2, 4.3 é desenvolvida esta proposta, partindo do processo de construção da base de dados (esta etapa está de acordo com as fases 2 e 3 do processo CRISP-DM[1]), seguido do processo de seleção da técnica/algoritmo de mineração e, por fim, a construção e validação do modelo de classificação da motivação dos estudantes (estas duas últimas etapas estão alinhadas com as fases 4 e 5 do processo CRISP-DM). Uma visão geral pode ser vista na Figura 10.

[1] mais detalhes sobre o processo CRISP-DM encontram-se na Seção 2.5

Figura 10 – Proposta do processo de classificação da motivação




## 4.1 CONSTRUÇÃO DA BASE DE DADOS

Esta seção apresenta as etapas para construção da base de dados. A subseção 4.1.1 apresenta a seleção da amostra, a subseção 4.1.2 apresenta o processo de seleção e aplicação do instrumento tradicional de classificação da motivação. O processo de coleta de logs é apresentado na seção 4.1.3 e, por fim, a seção 4.1.4 apresenta como foi construído o *dataset* final.

### 4.1.1 Etapas Iniciais e Seleção da Amostra

Antes do início do processo de coleta e construção da base de dados, foi estabelecida uma parceria com a Coordenadoria Institucional de Educação a Distância (CIED). Este órgão tem como missão coordenar os planos e ações de EAD no âmbito da UFAL. Após estabelecido o vínculo para a realização da pesquisa (Anexo A), este trabalho foi submetido e posteriormente aprovado pelo Comitê de Ética em Pesquisa (CEP) da UFAL (Anexos B e C).

Após a aprovação da pesquisa pelo CEP, foram selecionados por conveniência 179

alunos da EAD da UFAL nos polos de Arapiraca, Maceió, Maragogi, Olho d´Água das Flores e Santana do Ipanema, sendo estes 88 homens e 91 mulheres dos cursos de Ciências Sociais, Física, Geografia, Letras Espanhol, Português, Matemática, Pedagogia e Sistemas de Informação.

Os dados desses alunos foram obtidos por meio da aplicação de um instrumento de avaliação psicológica para classificação de motivação e pela coleta dos registros de interação (*logs*) no ambiente virtual de aprendizagem. Estas duas etapas, bem como a construção do *dataset* final, serão descritas com mais detalhes nas subseções a seguir.

### 4.1.2 Seleção e Aplicação do Instrumento de Avaliação Psicológica

Nesta etapa, inicialmente foi escolhido um instrumento de avaliação psicológica para classificação do perfil motivacional dos estudantes segundo a Teoria de Metas de Realização. Neste sentido, foi utilizada a Escala de Motivação para Aprendizagem – EMAPRE. A escala foi escolhida por apresentar precisão e fortes evidências de validade, testadas em um contexto brasileiro (SANTOS; ALCARÁ; ZENORINI, 2013).

A primeira versão da EMAPRE era composta por 67 itens (50 itens desenvolvidos pelos autores e 17 itens contidos na primeira versão da "Escala de Sensibilidade às Diferentes Metas de Realização"(MIDGLEY et al., 1998), sendo 20 referentes à meta Aprender, 22 à Performance-Aproximação e 25 à Performance-Evitação, agrupados em uma escala *Likert* com três opções de resposta – concordo (3 pontos), não sei (2 pontos) e discordo (1 ponto). Possui, em sua versão final, 28 itens apontados por uma análise fatorial exploratória e bons índices de fidedignidade avaliados pelo alfa de *Cronbach* ($\alpha$), tendo a meta Aprender ficado com 12 itens e $\alpha$ de 0,80; a meta Performance-Aproximação, com nove itens e $\alpha$ de 0,76; e a meta Performance-Evitação, sete itens e $\alpha$ de 0,73. As relações entre os fatores indicaram correlações consideradas fracas, entre a meta Performance-Aproximação e Evitação (r = 0,137), bem como entre a meta Aprender e meta Performance-Aproximação (r = 0,133) e negativa entre meta Aprender e meta Performance-Evitação (r = - 0,231).

Após a escolha do instrumento, foi desenvolvido um questionário (Apêndice A) dividido em duas partes, a primeira contém questões relativas ao perfil sóciodemográfico dos alunos e a segunda contém as questões da EMAPRE.

Posteriormente, o questionário foi aplicado presencialmente junto à amostra previamente selecionada - subseção 4.1.1. Antes, porém, os estudantes foram informados sobre os objetivos do estudo e convidados a participar voluntariamente da pesquisa. Também foi solicitado aos estudantes que aceitaram partipar do estudo que assinassem o Termo de Consentimento Livre e Esclarecido - TCLE (Apêndice B).

Para a classificação do perfil motivacional, a partir das respostas dos questionários foram atribuídos pontuações para cada questão e para cada fator. Sendo $X$ a nota para cada

questão e $Sb(Apr)$ a pontuação bruta para Aprender, $Sb(PerApr)$ a pontuação bruta para Performance-Aproximação e $Sb(PerEvit)$ a pontuação bruta para Performance-Evitação, as pontuações brutas para Aprender, Performance-Aproximação e Performance-Evitação são calculadas pelas Equações 4.1, 4.2 e 4.3 respectivamente.

$$Sb(Apr) = X_1 + X_2 + X_5 + X_7 + X_{10} + X_{12} + X_{14} + X_{19} + X_{21} + X_{23} + X_{25} + X_{28} \tag{4.1}$$

$$Sb(PerAprx) = X_3 + X_4 + X_8 + X_{11} + X_{13} + X_{15} + X_{17} + X_{20} + X_{24} \tag{4.2}$$

$$Sb(PerEvit) = X_6 + X_9 + X_{16} + X_{18} + X_{22} + X_{26} + X_{27} \tag{4.3}$$

Na sequência, foi calculada a média de cada sujeito em cada fator. Sendo $\overline{X}(Apr)$ a média para Aprender, $\overline{X}(PerAprx)$ a média para Performance-Aproximação e $\overline{X}(PerEvit)$ a média para Performance-Evitação. As médias para cada fator são calculadas conforme equação 4.4.

$$\overline{X}(Apr) = \frac{Sb(Apr)}{12} \qquad \overline{X}(PerAprx) = \frac{Sb(PerAprx)}{9} \qquad \overline{X}(PerEvit) = \frac{Sb(PerEvit)}{7} \tag{4.4}$$

Por fim, foi realizado o cálculo do percentual de cada fator em cada sujeito. Sendo $p(Apr)$ o percentual do fator Aprender, $p(PerAprx)$ o percentual do fator Performance-Aproximação e $p(PerEvit)$ o percentual do fator Performance-Evitação. O percentual para Aprender, Performance-Aproximação e Performance-Evitação de cada sujeito é calculado pelas Equações 4.5, 4.6 e 4.7, respectivamente.

$$p(Apr) = \frac{\overline{X}(Apr)}{\overline{X}(Apr) + \overline{X}(PerAprx) + \overline{X}(PerEvit)} \tag{4.5}$$

$$p(PerAprx) = \frac{\overline{X}(PerAprx)}{\overline{X}(Apr) + \overline{X}(PerAprx) + \overline{X}(PerEvit)} \tag{4.6}$$

$$p(PerEvit) = \frac{\overline{X}(PerEvit)}{\overline{X}(Apr) + \overline{X}(PerAprx) + \overline{X}(PerEvit)} \tag{4.7}$$

### 4.1.3 Coleta de Logs no Ambiente Virtual de Aprendizagem

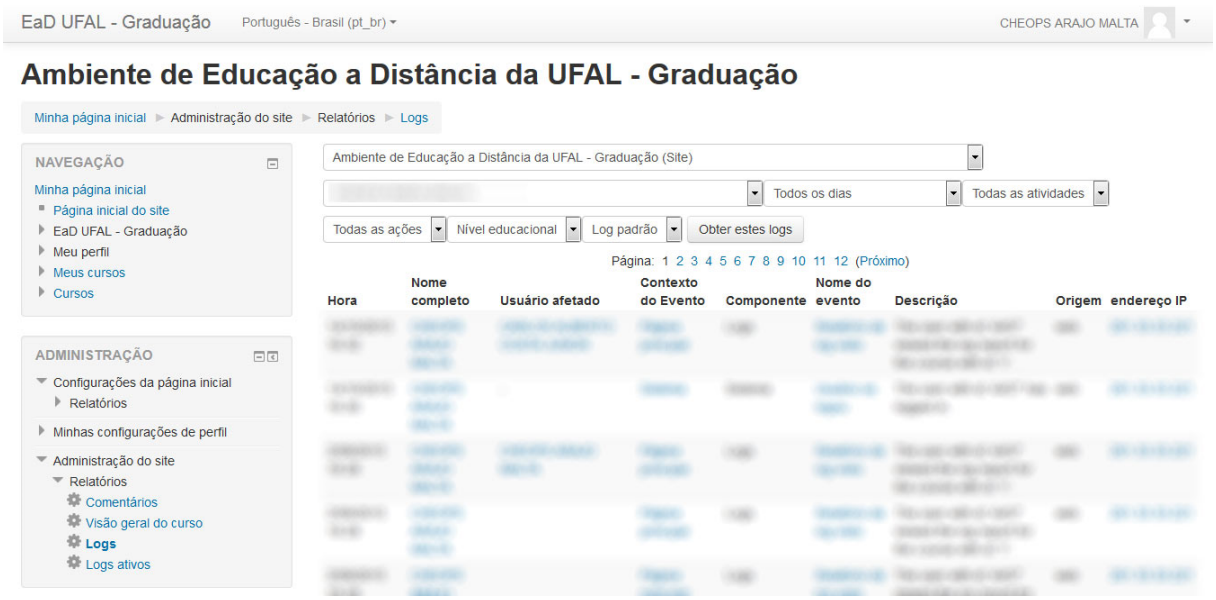
Neste etapa, como todos os alunos da amostra pertencem à UFAL, foi solicitado ao Núcleo de Tecnologia da Informação (NTI), responsável pelas soluções de tecnologia da

informação no âmbito da universidade, os logs dos alunos que haviam participado da etapa anterior.

A partir dessa solicitação, o núcleo analisou o projeto de pesquisa previamente aprovado pelo CEP e, após constatada a conformidade de toda a documentação, liberou uma conta para acesso aos dados na plataforma Moodle, conforme pode ser visto na Figura 11.

A partir do acesso ao Moodle, foram exportados os logs de cada um dos alunos para um arquivo CSV[2].

Figura 11 – Interface para acesso aos logs no Moode UFAL.




Fonte: NTI - UFAL

### 4.1.4 Conjunto de Dados (Dataset)

Após a aplicação dos questionários e da coleta dos logs, foi realizada a construção do *Dataset*. A primeira versão continha 154 variáveis para cada aluno, sendo estas 50 obtidas a partir do questionário e 89 a partir das ações registradas nos logs.

Num segundo momento, foram excluídas as variáveis com valores zerados e foi entrevistado um grupo de alunos para detectar quais as ações "eles acreditavam ser mais importantes". Após esse processo, foram selecionadas as 12 variáveis mais elencadas (user_login, course_view, resource_view, wiki_view, glossary_view, forum_view_discussion, forum_view_forum, forum_add_post, forum_update_post, wiki_comments, message_write, glossary_add_entry).

[2] *Comma Separated Values* – Valores separados por vírgula

Também foram agrupadas as ações coletadas nos logs em 13 variáveis (assign, blog, chat, choice, course, data, forum, glossary, message, quiz, scorm, user, wiki). Cada uma destas variáveis representa a quantidade de ações do usuário em cada um dos módulos do moodle.

Para as variáveis scorm, quiz e blog foram criadas variáveis auxiliares que receberam os nomes de IDscorm, IDquiz e IDblog. Estas variáveis identificam se o aluno realizou ou não atividades nos respectivos módulos. Esta decisão foi tomada após a análise inicial dos dados brutos, onde se percebeu que muitos alunos chegavam a não realizar nenhuma ação nestes módulos, da mesma forma que outros os utilizavam com frequência.

A versão final do *dataset* contém 158 observações com 32 variáveis, sendo 26 obtidas a partir dos logs e 6 a partir dos resultados sumarizados dos questionários. A Tabela 3 apresenta todas as variáveis do *dataset* final, bem como a descrição do que cada uma representa.

Tabela 3 – Variáveis e descrições

| Variável | Descrição |
|---|---|
| periodo | Período em que o aluno se encontra no curso |
| diasinteracao | Quantidade de dias em que interagiu no AVA |
| idade | Idade do aluno |
| genero | Gênero (0 - Masculino, 1 - Feminino) |
| user_login | Quantidade de vezes que fez login no AVA |
| course_view | Quantidade de vezes que visualizou uma disciplina |
| resource_view | Quantidade de vezes que visualizou um recurso |
| glossary_view | Quantidade de vezes que visualizou um glossário |
| forum_view_discussion | Quantidade de vezes que visualizou uma discussão em um fórum |
| forum_view_forum | Quantidade de vezes que visualizou um fórum |
| forum_add_post | Quantidade de vezes que adicionou uma postagem em um fórum |
| forum_update_post | Quantidade de vezes que atualizou uma postagem sua em um fórum |
| wiki_view | Quantidade de vezes que visualizou uma wiki |
| wiki_comments | Quantidade de vezes que adicionou um comentário em uma wiki |
| message_write | Quantidade de vezes que enviou uma mensagem para outra pessoa |
| glossary_add_entry | Quantidade de vezes que adicionou um registro ao glossário |
| assign | Quantidade de ações relacionadas ao módulo de tarefas |
| blog | Quantidade de ações relacionadas ao módulo de blogs |
| chat | Quantidade de ações relacionadas ao módulo de chats |
| choice | Quantidade de ações relacionadas ao módulo de questões de múltipla escolha |
| course | Quantidade de ações relacionadas ao módulo de cursos |
| data | Quantidade de ações relacionadas ao módulo de banco de informações |
| forum | Quantidade de ações relacionadas ao módulo de fóruns |
| glossary | Quantidade de ações relacionadas ao módulo de glossários |
| message | Quantidade de ações relacionadas ao módulo de envio de mensagens |
| quiz | Quantidade de ações relacionadas ao módulo de quizzes |
| scorm | Quantidade de ações relacionadas ao módulo de scorm |
| user | Quantidade de ações relacionadas as entrada e saída do usuário no AVA |
| wiki | Quantidade de ações relacionadas ao módulo de wiki |
| IDscorm | Identificador se realizou ou não atividades no módulo scorm (0 - não, 1 - sim) |
| IDquiz | Identificador se realizou ou não atividades no módulo quizz (0 - não, 1 - sim) |
| IDblog | Identificador se realizou ou não atividades no módulo blog (0 - não, 1 - sim) |
| per_aprender | Percentual da meta Aprender coletado a partir do questionário |
| per_aprox | Percentual da meta Performance-Aproximação coletado a partir do questionário |
| per_evit | Percentual da meta Performance-Evitação coletado a partir do questionário |

## 4.2 SELEÇÃO DA TÉCNICA/ALGORITMO PARA CLASSIFICAÇÃO DA MOTIVAÇÃO

Conforme mencionado na seção 1.3, o objetivo principal deste trabalho é a criação de um modelo para classificação da motivação dos alunos de cursos mediados por computador. Para este fim, propõe-se a aplicação de testes psicométricos (questionários) e posterior mineração nos *logs* dos alunos no AVA. Ou seja, a partir dos dados de interação, espera-se construir um modelo que permita a predição dos perfis motivacionais de novos alunos sem a necessidade da aplicação dos questionários.

Desta forma, podemos notar que existe uma necessidade de se avaliar quais técnicas/algoritmos descritos na literatura são mais eficientes para tratar desse problema partindo da perspectiva proposta neste estudo. No contexto da predição, ou mineração de dados preditiva, podemos encontrar dois tipos de problemas, de acordo com os valores que os dados que se deseja predizer assumem: se os dados assumem valores nominais (discretos ou categóricos), o problema é denominado classificação, já se os dados assumem valores contínuos, o problema é chamado de regressão.

Dado que os perfis motivacionais que serão preditos são representados pelos *scores* (valores contínuos) dos alunos em cada uma das três metas, chegamos às seguintes perguntas de investigação:

- Qual o melhor algoritmo/técnica para regressão no contexto da categorização da motivação de estudantes com base na interação entre logs e questionários?
- Qual destes algoritmos apresenta melhor acurácia?

A fim de obter estas respostas, foi realizado um experimento cujos detalhes são apresentado nas subseções a seguir.

### 4.2.1 Objetivos da investigação

A investigação a ser realizada é do tipo experimental e tem como objetivo comparar o comportamento dos algoritmos/técnicas de regressão quando utilizamos dados relativos à motivação de estudantes. Os dados foram coletados e organizados conforme descrito na seção 4.1. Na sequência estes dados foram utilizados como entrada para os algoritmos.

Formalmente, o objetivo deste experimento pode ser definido no formato GQM como **analisar** os algoritmos de regressão **com a intenção de** compará-los **a respeito** de sua acurácia **do ponto de vista** da classificação da motivação dos estudantes da educação on-line **no contexto** dos dados obtidos a partir de questionários e logs de alunos de uma única instituição de ensino superior.

### 4.2.2 Planejamento do Experimento

#### *4.2.2.1 Questão de investigação e hipóteses*

A eficácia de um sistema de classificação está relacionada com a capacidade de o mesmo predizer o valor de uma variável resposta em função dos valores das variáveis de entrada. Sendo assim, a principal questão de pesquisa deste experimento objetiva determinar e comparar a precisão dos algoritmos quando alimentados com dados de questionários e logs do moodle para a classificação da motivação de alunos da educação on-line.

**P1** - Os algoritmos de regressão apresentam diferenças na métrica de eficácia (precisão) quando temos como entrada dados de questionários e logs do moodle?

O que nos leva às seguintes hipóteses:

**H1-0:** O desvio médio absoluto para os algoritmos de classificação é igual.

**H1-1:** O desvio médio absoluto para os algoritmos de classificação é diferente.

**H2-0:** O erro médio quadrático para os algoritmos de classificação é igual.

**H2-1:** O erro médio quadrático para os algoritmos de classificação é diferente.

#### *4.2.2.2 Fatores e Variáveis Respostas*

As variáveis independentes (e também fatores) utilizados no experimento são:

- **Algoritmo:** Algoritmo de regressão que tem como entrada dados dos questionários respondidos pelos alunos e logs do moodle dos respectivos alunos;
- **Classe de motivação:** Classe final que se deseja obter a partir da execução dos algoritmos;

As variáveis de resposta são:

- **Valor do desvio absoluto médio - DAM**
- **Valor do erro quadrático médio - EQM**

À exceção do Algoritmo, todas as variáveis da pesquisa são quantitativas.

#### *4.2.2.3 Níveis dos fatores*

Os níveis dos fatores estão definidos de acordo com a Tabela 4

Tabela 4 – Experimento - Definição dos níveis dos fatores.

| Fator | Níveis |
|---|---|
| Técnica/Algoritmo | Regressão Linear/Linear Regression - seção 2.6.2<br>Regressão Beta/Betareg - seção 2.6.3<br>Aprendizado baseado em exemplos/KNN - seção 2.6.4<br>Indução de regras de regressão/M5Rules - seção 2.6.5<br>Indução Top-Down de árvores de regressão/M5P - seção 2.6.6<br>Redes Neurais Artificiais/MultilayerPerceptron - seção 2.6.7<br>Suport Vector Machines/SMOreg - seção 2.6.8 |
| Classe de motivação | Aprender<br>Performance-Aproximação<br>Performance-Evitação |

#### *4.2.2.4 Definição formal das hipóteses*

Formalmente, as hipóteses descritas anteriormente podem ser definidas conforme a Tabela 5.

Tabela 5 – Experimento - Definição formal das hipóteses.

| Hipótese | Hipótese Nula | Hipótese Alternativa |
|---|---|---|
| $H1$ | $H1_0 : D(A1) = D(A2)$ | $H1_1 : D(A1) \neq D(A2)$ |
| $H2$ | $H2_0 : E(A1) = E(A2)$ | $H2_1 : E(A1) \neq E(A2)$ |

Onde $D$ e $E$ são funções que retornam, respectivamente, os valores das métrica de DAM e EQM dos algoritmos $A1$ e $A2$. Estes algoritmos - e suas entradas - estão definidos na tabela 4.

#### *4.2.2.5 Unidades Experimentais*

Em nosso cenário de investigação, as unidades experimentais são os conjuntos de treinamento extraídos dos dados originais disponíveis. Cada conjunto de treinamento é uma amostra aleatória sob a qual o tratamento do experimento (conjunto de fatores) é aplicado para se obter a variável resposta citada anteriormente, portanto, a partir destas unidades, é que será possível obter a variação estatística na análise dos resultados da investigação.

#### *4.2.2.6 Design de experimento*

A fim de verificar o comportamento dos algoritmos sobre cada combinação possível dos dados de entrada e descobrir interações entre os fatores, o design utilizado será o Design Fatorial Completo com 10 replicações. Ou seja, no total 21 ensaios serão realizados, cada um 10 vezes, totalizando 210 execuções. Como cada execução tem custo mínimo – trata-se somente de execuções de algoritmos por máquina, o experimento não envolve pessoas – não há problemas em usar este tipo de design. A Tabela 6 descreve cada um dos tratamentos.

Tabela 6 – Experimento - Definição dos ensaios

| Tratamento | Algoritmo/Técnica | Classe de Motivação |
|---|---|---|
| 1 | Regressão Linear | Aprender |
| 2 | Regressão Linear | Performance-Aproximação |
| 3 | Regressão Linear | Performance-Evitação |
| 4 | Regressão Beta | Aprender |
| 5 | Regressão Beta | Performance-Aproximação |
| 6 | Regressão Beta | Performance-Evitação |
| 7 | Aprendizado baseado em exemplos | Aprender |
| 8 | Aprendizado baseado em exemplos | Performance-Aproximação |
| 9 | Aprendizado baseado em exemplos | Performance-Evitação |
| 10 | Indução de Regras de Regressão | Aprender |
| 11 | Indução de Regras de Regressão | Performance-Aproximação |
| 12 | Indução de Regras de Regressão | Performance-Evitação |
| 13 | Indução Top-Down de Árvores de Regressão | Aprender |
| 14 | Indução Top-Down de Árvores de Regressão | Performance-Aproximação |
| 15 | Indução Top-Down de Árvores de Regressão | Performance-Evitação |
| 16 | Redes Neurais Artificiais | Aprender |
| 17 | Redes Neurais Artificiais | Performance-Aproximação |
| 18 | Redes Neurais Artificiais | Performance-Evitação |
| 19 | Suport Vector Machines | Aprender |
| 20 | Suport Vector Machines | Performance-Aproximação |
| 21 | Suport Vector Machines | Performance-Evitação |

Em cada replicação do experimento, os dados de entrada serão representados por dois arquivos (um de treino e um de teste) selecionados aleatoriamente por meio do método *k-fold cross-validation*. Neste método, o conjunto de dados é aleatoriamente dividido em $k$ partições mutuamente exclusivas (*folds*), de tamanho aproximadamente igual a $\frac{n}{k}$. As ($k$-1) *folds* são usadas para treinamento e o *fold* restante para o teste. Este processo é repetido $k$ vezes, cada vez considerando um *fold* diferente para o teste (MATOS et al., 2009). Esta técnica foi utilizada para prover randomização e permitir que testes estatísticos sejam realizados.

#### *4.2.2.7 Medidas de precisão*

A precisão tem como objetivo avaliar o desempenho do classificador na predição do valor do atributo-meta de novos exemplos. Porém, quando o problema é de regressão, não é fácil medir o desempenho de predição do modelo, pois como o atributo-meta a ser predito assume valores numéricos, não se pode afirmar se o valor predito está correto ou não. Por isso, a maioria das medidas de precisão utilizadas em problemas de regressão são baseadas na diferença entre o valor predito pelo algoritmo e o valor real do atributo-meta.

Para tornar possível a comparação da precisão entre os diversos algoritmos utilizados neste experimento, serão calculadas algumas medidas sobre os resultados das execuções utilizando os conjuntos de testes de cada ensaio. Para tanto, tomaremos como base as medidas

utilizadas no relatório técnico realizado por Dosualdo e Rezende (2003).

A medida DAM (Desvio Absoluto Médio) é obtida por meio da média da diferença (em módulo) entre os valores reais do conjunto de dados e os valores preditos por um determinado atributo. Seja $h_i$ a hipótese construída pelo algoritmo na $i$-ésima partição. O valor da DAM calculada em cada uma das $i$ partições é obtido por meio da Equação 4.8, em que $n_{teste}$ corresponde ao número de exemplos do arquivo de teste; $y'_j$ corresponde ao valor predito pelo algoritmo no $j$-ésimo exemplo de teste; e $y_j$ é o valor real do atributo-meta desse mesmo exemplo.

$$DAM(h_i) = \frac{1}{n_{teste}} \sum_{j=1}^{n_{teste}} |y'_j - y_j| \tag{4.8}$$

A medida EQM (Erro Quadrático Médio) consiste na média da diferença ao quadrado entre os valores reais e os valores preditos para um atributo. Dado $h_i$ como sendo a hipótese gerada na $i$-ésima partição, o valor da EQM calculada para cada partição é obtido por meio da Equação 4.9

$$EQM(h_i) = \frac{1}{n_{teste}} \sum_{j=1}^{n_{teste}} (y'_j - y_j)^2 \tag{4.9}$$

Considerando A um algoritmo, $EQM(h_i)$ a medida EQM calculada sobre a $i$-ésima partição do conjunto de dados e $DAM(h_i)$ a medida DAM calculada sobre a $i$-ésima partição do conjunto de dados. As médias das medidas DAM e EQM calculadas sobre o algoritmo A são dadas pela Equações 4.10 e 4.11, em que $k$ corresponde ao número de partições do conjunto de dados.

$$mediaDAM(A) = \frac{1}{k} \sum_{i=1}^{k} DAM(h_i) \tag{4.10}$$

$$mediaEQM(A) = \frac{1}{k} \sum_{i=1}^{k} EQM(h_i) \tag{4.11}$$

Uma outra medida utilizada é a variância das medidas DAM e EQM de cada algoritmo. Para calculá-la, baseado nos valores das médias, utiliza-se as Equações 4.12 e 4.13, para DAM e EQM respectivamente.

$$varDAM(A) = \frac{1}{k-1} \sum_{i=1}^{k} (DAM(h_i) - media(A))^2 \tag{4.12}$$

$$varEQM(A) = \frac{1}{k-1} \sum_{i=1}^{k} (EQM(h_i) - media(A))^2 \tag{4.13}$$

Por fim, é calculado o desvio padrão das medidas de DAM e EQM obtidas sobre cada algoritmo tomando como base a variância. Este desvio é dado pela Equações 4.14 e 4.15.

$$dpDAM(A) = \sqrt{varDAM(a)} \tag{4.14}$$

$$dpEQM(A) = \sqrt{varEQM(a)} \tag{4.15}$$

### 4.2.3 Execução dos Métodos

Nesta subseção, é apresentado o processo de execução dos métodos de regressão. Como já foi mencionado anteriormente, o conjunto de dados foi particionado em 10 subconjuntos de treino e teste selecionados aleatoriamente por meio do método *k-fold cross-validation*. Desta forma, para cada técnica foram realizadas 10 execuções/ensaios para cada tratamento, sendo uma para cada subconjunto de dados.

A seguir é descrito como foi executado cada um dos métodos de regressão. A fim de exemplificar os modelos gerados por cada algoritmo, serão apresentados os modelos para obtenção do percentual de Aprender, gerados a partir do quinto subconjunto de dados.

#### *4.2.3.1 Regressão Linear*

Para a regressão linear, foi utilizado o classificador **LinearRegression** do WEKA. Como método de seleção de atributos, foi utilizado o método M5 e utilizada a eliminação de atributos colineares. No Código 4.1, é apresentada a função obtida.

Código 4.1 – Função de regressão linear gerada pelo WEKA

```
1  per_aprender =
2        0.0001 * diasinteracao +
3        0.0011 * idade +
4        0.0185 * genero +
5       -0.0002 * user_login +
6       -0.0001 * course_view +
7        0.0015 * glossary_view +
8        0.0021 * forum_update_post +
9        0.0099 * wiki_comments +
10      -0.0101 * glossary_add_entry +
11      -0.0002 * blog +
12       0.0001 * user +
13      -0.0033 * wiki +
14       0.4281
```

#### *4.2.3.2 Regressão Beta*

Para a regressão beta, foi utilizado o pacote **betareg** do R. Como o pacote não possui um método automático para a seleção de variáveis, foi conduzido um processo de seleção manual utilizando critérios equivalentes ao do método stepAIC[3]. No Código 4.2, é apresentada a função de regressão obtida por meio da regressão beta.

Código 4.2 – Função de regressão beta gerada pelo R

```
1 per_aprender =
2   exp(-0.1605+0.0032*idade+0.0440*log(forum_add_post+1.0)*genero-0.0325*log(scorm +
        1.0)*genero+0.0033*log(forum_update_post + 1.0)*genero)/
3   1+exp(-0.1605+0.0032*idade+0.0440*log(forum_add_post+1.0)*genero-0.0325*log(scorm +
        1.0)*genero+0.0033*log(forum_update_post + 1.0)*genero)
```

#### *4.2.3.3 Aprendizado Baseado em Exemplos*

Para o método de aprendizado baseado em exemplos, foi utilizado o classificador **IBk** do WEKA. Este classificador implementa o algoritmo *K-nearest neighbours*. O valor adotado para K foi 10, o que quer dizer que, para cada exemplo do conjunto de teste fornecido, o valor do atributo-meta foi calculado baseado nos 10 exemplos mais similares do conjunto de treinamento. Este algoritmo não fornece um modelo de saída que explique os padrões encontrados.

#### *4.2.3.4 Indução de Regras de Regressão*

Para a indução de regras de regressão, foi utilizado o classificador **M5Rules** do WEKA. O algoritmo foi executado com sua configuração padrão e o número de regras geradas em cada execução variou entre 1 e 3 regras.

O Código 4.3 apresenta um modelo baseado em regras gerados pelo M5Rules. Neste exemplo, foram geradas 3 regras. Na primeira regra, se a idade do aluno for menor ou igual a 23.5 (linhas 2 a 4), então é executado o primeiro modelo (linhas 6 a 22). Na segunda regra (linhas 25 a 28), se a quantidade de atividades relacionadas ao componente mensagens for menor ou igual a 56 e a quantidade de entradas do aluno no sistema for maior que 276.5, então é executado o segundo modelo (linhas 30 a 45). Já o terceiro modelo (linhas 49 a 52) é executado caso nenhuma das condições anteriores seja encontrada.

Código 4.3 – Regras de regressão geradas pelo algoritmo M5Rules

```
1 Rule: 1
2 IF
3         idade <= 23.5
4 THEN
```

[3] Mais detalhes sobre o stepAIC podem ser vistos na seção 4.3.2.1

```
5
6  per_aprender =
7          0 * diasinteracao
8          + 0.0002 * idade
9          + 0.0056 * genero
10         - 0.0001 * user_login
11         - 0 * course_view
12         + 0.0002 * glossary_view
13         + 0.0001 * forum_view_discussion
14         + 0 * forum_view_forum
15         - 0.0018 * forum_add_post
16         + 0.0042 * forum_update_post
17         + 0.0001 * message_write
18         - 0 * assign
19         + 0.0001 * data
20         - 0.0001 * message
21         + 0 * user
22         + 0.4794 [48/69.77%]
23
24 Rule: 2
25 IF
26         message <= 56
27         user_login > 276.5
28 THEN
29
30 per_aprender =
31         0 * diasinteracao
32         + 0.023 * genero
33         - 0 * user_login
34         - 0 * resource_view
35         + 0.0005 * glossary_view
36         - 0.0001 * forum_view_discussion
37         - 0.0001 * forum_view_forum
38         - 0.0002 * forum_add_post
39         + 0.0002 * message_write
40         + 0 * assign
41         + 0.0001 * forum
42         - 0.0002 * message
43         + 0.0001 * quiz
44         + 0 * user
45         + 0.4962 [43/55.308%]
46
47 Rule: 3
48
49 per_aprender =
50         -0.0001 * course_view
51         + 0.0002 * forum_view_forum
```

```
52          + 0.4877 [51/94.827%]
```

### *4.2.3.5 Indução Top-Down de Árvores de Regressão*

Para a obtenção das árvores de regressão, foi utilizado o classificador **M5P** do WEKA. Uma das árvores de regressão geradas de modo textual e visual pelo algoritmo podem ser vistas no Código 4.4 e na Figura 12 respectivamente.

No Código 4.4, é apresentada inicialmente a árvore de regressão de modo textual (linhas 1 a 6). Os nós folhas desta árvore representam modelos de regressão. Por exemplo, caso a idade do aluno seja menor ou igual a 23.5, temos como nó folha o modelo LM1 (linhas 9 a 25). Já se a idade do aluno for maior que 23.5 e, se a quantidade de atividades relacionadas ao componente mensagens for maior que 56, o nó folha e, consequentemente, o modelo a ser executado será o LM4 (linhas 70 a 84).

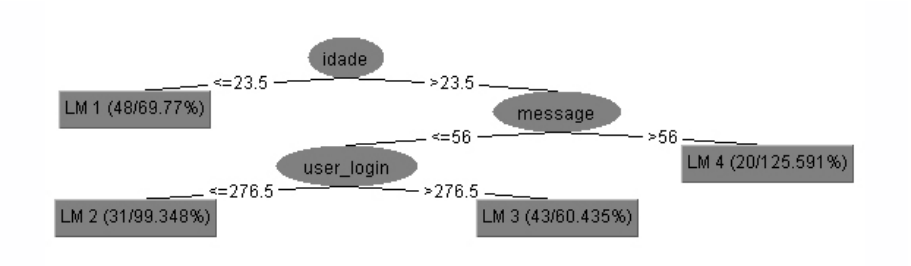
Código 4.4 – Árvore de regressão textual obtida pelo algoritmo M5P

```
1  idade <= 23.5 : LM1 (48/69.77%)
2  idade >  23.5 :
3  |   message <= 56 :
4  |   |   user_login <= 276.5 : LM2 (31/99.348%)
5  |   |   user_login >  276.5 : LM3 (43/60.435%)
6  |   message >  56 : LM4 (20/125.591%)
7
8  LM num: 1
9  per_aprender =
10         0 * diasinteracao
11         + 0.0002 * idade
12         + 0.0056 * genero
13         - 0.0001 * user_login
14         - 0 * course_view
15         + 0.0002 * glossary_view
16         + 0.0001 * forum_view_discussion
17         + 0 * forum_view_forum
18         - 0.0018 * forum_add_post
19         + 0.0042 * forum_update_post
20         + 0.0001 * message_write
21         - 0 * assign
22         + 0.0001 * data
23         - 0.0001 * message
24         + 0 * user
25         + 0.4794
26
27 LM num: 2
28 per_aprender =
29         0.0001 * diasinteracao
```

```
30          + 0.0001 * idade
31          + 0.0084 * genero
32          - 0.0001 * user_login
33          - 0 * course_view
34          - 0.0001 * resource_view
35          + 0.0005 * glossary_view
36          - 0.0001 * forum_view_discussion
37          - 0.0001 * forum_view_forum
38          - 0.0003 * forum_add_post
39          + 0.0002 * message_write
40          + 0 * assign
41          + 0.0001 * data
42          + 0.0001 * forum
43          - 0.0002 * message
44          + 0.0001 * quiz
45          + 0 * user
46          + 0.457
47
48  LM num: 3
49   per_aprender =
50          0 * diasinteracao
51          + 0.0001 * idade
52          + 0.0231 * genero
53          - 0.0001 * user_login
54          - 0 * course_view
55          - 0 * resource_view
56          + 0.0006 * glossary_view
57          - 0.0001 * forum_view_discussion
58          - 0.0001 * forum_view_forum
59          - 0.0002 * forum_add_post
60          + 0.0002 * message_write
61          + 0 * assign
62          + 0.0001 * data
63          + 0.0001 * forum
64          - 0.0002 * message
65          + 0.0001 * quiz
66          + 0 * user
67          + 0.4885
68
69  LM num: 4
70   per_aprender =
71          0 * diasinteracao
72          + 0.0001 * idade
73          + 0.0162 * genero
74          - 0 * user_login
75          - 0 * course_view
76          + 0.0004 * glossary_view
```

```
77          - 0.0001 * forum_view_discussion
78          + 0 * forum_view_forum
79          + 0.0007 * message_write
80          - 0 * assign
81          + 0.0001 * data
82          - 0.0005 * message
83          + 0 * user
84          + 0.4549
```

Figura 12 – Árvore de regressão visual obtida pelo algoritmo M5P




### *4.2.3.6 Redes Neurais Artificiais*

Para treinar a rede neural artificial, foi utilizado o algoritmo **MultilayerPerceptron** do WEKA, que utiliza a função de *backpropagation* para classificar as instâncias.

O algoritmo foi configurado com uma taxa de aprendizado de 0.4. Esta taxa geralmente varia entre 0.1 e 1.0. Uma taxa de aprendizado muito baixa torna o aprendizado da rede muito lento, da mesma forma que uma taxa de aprendizado muito alta provoca oscilações no treinamento e impede a convergência do processo de aprendizado.

Para o *momentum* que tem por objetivo o aumento da velocidade de treinamento da rede neural e a redução do perigo de instabilidade, foi utilizado o valor 0.2, em uma escala que varia entre 0.0 e 1.0.

Por fim, foram utilizados 2000 ciclos de treinamento para cada uma das redes. Parte da saída gerada pelo algoritmo pode ser vista no Código 4.5

Código 4.5 – Trecho da saída do algoritmo MultilayerPerceptron

```
1  Linear Node 0
2      Inputs      Weights
3      Threshold      0.2870854108192187
4      Node 1      0.8940257629754775
5      Node 2      -0.7911690883866411
6      Node 3      0.9951097627613372
7      Node 4      1.0043893884263853
8      Node 5      -0.8444110386030911
9      Node 6      0.9957375399778305
10     Node 7      1.1379424087339716
```

```
11      Node 8      0.6032131442930883
12      Node 9      -0.6308809561519966
13      Node 10      0.7367656591848779
14      Node 11      0.8987936927086803
15      Node 12      1.0346343591686658
16      Node 13      -1.067159497356275
17      Node 14      -0.8631899822768175
18      Node 15      -0.01071024723652884
19      Node 16      -0.07521960870688846
20  Sigmoid Node 1
21      Inputs      Weights
22      Threshold      -1.7093546296351578
23      Attrib periodo      0.27958953143424325
24      Attrib diasinteracao      1.8230322438808584
25      Attrib idade      1.5033135920675018
26
27  ....
```

### *4.2.3.7 Support Vector Machines*

Para obter o modelo usando a técnica de *Support Vector Machines*, foi utilizado o algoritmo **SMOReg** do WEKA que implementa a técnica de SVM para regressão. O algoritmo foi executado com sua configuração padrão. Um dos modelos gerados pode ser visto no Código 4.6

Código 4.6 – Modelo gerado pelo algoritmo SMOReg

```
1  weights (not support vectors):
2   +       0.1135 * (normalized) periodo
3   +       0.0729 * (normalized) diasinteracao
4   +       0.1711 * (normalized) idade
5   +       0.0538 * (normalized) genero
6   +       0.0377 * (normalized) user_login
7   -       0.2004 * (normalized) course_view
8   -       0.0708 * (normalized) resource_view
9   -       0.1914 * (normalized) wiki_view
10  +       0.2948 * (normalized) glossary_view
11  -       0.1344 * (normalized) forum_view_discussion
12  +       0.2009 * (normalized) forum_view_forum
13  -       0.0666 * (normalized) forum_add_post
14  +       0.1828 * (normalized) forum_update_post
15  +       0.2923 * (normalized) wiki_comments
16  -       0.153  * (normalized) message_write
17  -       0.2486 * (normalized) glossary_add_entry
18  -       0.0554 * (normalized) assign
19  -       0.3224 * (normalized) blog
20  -       0.041  * (normalized) chat
```

```
21   +       0.1141 * (normalized) choice
22   -       0.1882 * (normalized) course
23   +       0.025  * (normalized) data
24   +       0.0628 * (normalized) forum
25   +       0.2911 * (normalized) glossary
26   +       0.015  * (normalized) message
27   -       0.0792 * (normalized) quiz
28   -       0.0473 * (normalized) IDscorm
29   +       0.1065 * (normalized) scorm
30   +       0.2342 * (normalized) user
31   -       0.2328 * (normalized) wiki
32   +       0.0086 * (normalized) IDquiz
33   +       0.0163 * (normalized) IDblog
34   +       0.6195
```

### 4.2.4 Resultados das medidas de precisão

Para calcular a precisão dos métodos, foram utilizados os 10 conjuntos de treinamento e teste, sendo que em todos os casos o conjunto de teste foi fornecido após a construção dos modelos. Os resultados foram calculados em função da excução do método de regressão e do atributo-meta conforme descrito na subseção 4.2.2.3. As Tabelas 7, 8 e 9 apresentam as médias das medidas de DAM e EQM juntamente com seus respectivos desvios padrões calculados sobre cada método de regressão executado para os atributos-meta Aprender, Performance-Aproximação e Performance-Evitação respectivamente.

A partir desta subseção, passaremos a nos referir somente ao algoritmo executado. O método de regressão relacionado a cada algoritmo pode ser visto na tabela 4 (Níveis dos fatores).

Tabela 7 – Experimento - Média das medidas de DAM e EQM calculadas sobre cada método e atributo-meta Aprender

| **Algoritmo** | **DAM ± Desvio Padrão** | **EQM ± Desvio Padrão** |
|---|---|---|
| LinearRegression | 0.0467940 ± 0.00826030 | 0.00383993 ± 0.00177896 |
| Betareg | 0.0435646 ± 0.00937516 | 0.00302177 ± 0.00166926 |
| IBk | 0.0472346 ± 0.00980436 | 0.00359217 ± 0.00170456 |
| M5R | 0.0502242 ± 0.00881608 | 0.00419190 ± 0.00170293 |
| M5P | 0.0488087 ± 0.00742795 | 0.00397806 ± 0.00151880 |
| MultilayerPerceptron | 0.0926872 ± 0.01959824 | 0.01721225 ± 0.00934165 |
| SMOReg | 0.0454096 ± 0.00794146 | 0.00362249 ± 0.00156249 |

Tabela 8 – Experimento - Média das medidas de DAM e EQM calculadas sobre cada método e atributo-meta Performance-Aproximação

| **Algoritmo** | **DAM ± Desvio Padrão** | **EQM ± Desvio Padrão** |
|---|---|---|
| LinearRegression | 0.05062412 ± 0.01180310 | 0.00392216 ± 0.00168522 |
| Betareg | 0.04681171 ± 0.00852663 | 0.00319504 ± 0.00109278 |
| IBk | 0.04942871 ± 0.00966864 | 0.00377299 ± 0.00143985 |
| M5R | 0.05281724 ± 0.00987228 | 0.00422648 ± 0.00139473 |
| M5P | 0.05223644 ± 0.01001345 | 0.00424965 ± 0.00156310 |
| MultilayerPerceptron | 0.08800333 ± 0.01727902 | 0.01309186 ± 0.00670987 |
| SMOReg | 0.04798732 ± 0.00950707 | 0.00357321 ± 0.00135689 |

Tabela 9 – Experimento - Média das medidas de DAM e EQM calculadas sobre cada método e atributo-meta Performance-Evitação

| **Algoritmo** | **DAM ± Desvio Padrão** | **EQM ± Desvio Padrão** |
|---|---|---|
| LinearRegression | 0.05079510 ± 0.00950519 | 0.00431947 ± 0.002095642 |
| Betareg | 0.04017339 ± 0.00605248 | 0.00293337 ± 0.001063802 |
| IBk | 0.04629433 ± 0.00793948 | 0.00347658 ± 0.001263278 |
| M5R | 0.04682319 ± 0.01010104 | 0.00360443 ± 0.001488452 |
| M5P | 0.04631571 ± 0.00905021 | 0.00356966 ± 0.001420271 |
| MultilayerPerceptron | 0.09521446 ± 0.01564815 | 0.01488574 ± 0.005154536 |
| SMOReg | 0.04825964 ± 0.00690813 | 0.00417069 ± 0.001364221 |

### 4.2.5 Testes de Hipóteses

Mesmo sendo calculadas as medidas de DAM e EQM, em alguns casos não fica fácil perceber se um método executado é de fato melhor que outro.

Em geral, a comparação é feita tendo $A_P$ como algoritmo proposto e $A_S$ algoritmo padrão. Então, a média e o desvio padrão combinados são calculados de acordo com as Equações 4.16 e 4.17, respectivamente:

$$media(A_S - A_P) = media(A_S) - media(A_P) \tag{4.16}$$

$$dp(A_S - A_P) = \sqrt{\frac{dp(A_S)^2 + dp(A_P)^2}{2}} \tag{4.17}$$

Por fim, a diferença absoluta, em desvios padrões, é calculada por meio da Equação 4.18.

$$ad(A_S - A_P) = \frac{media(A_S - A_P)}{dp(A_S - A_P} \tag{4.18}$$

Desta forma, se $ad(A_S - A_P) > 0$ então $A_P$ supera $A_S$. Se $ad(A_S - A_P) \leq 0$, então $A_S$ supera $A_P$.

Para decidir se um método é melhor que um outro com significância estatística, inicialmente foi realizado um Teste de Anderson-Darling (ANDERSON; DARLING, 1952) para verificar se os dados atendiam à suposição de normalidade. Constatado que os dados não eram normais, foi utilizado o teste Wilcoxon Pareado (WILCOXON, 1945) para verificar se a diferença é significativa ou não.

As tabelas 10, 11 e 12 apresentam os resultados dos testes de hipóteses dos resultados das execuções dos algoritmos para os atributos-meta Aprender, Performance-Aproximação e Performance-Evitação respectivamente.

Tabela 10 – Experimento - Testes de Hipóteses dos resultados dos algoritmos para o atributo-meta Aprender

| $A_S$ | $A_P$ | $ad(A_S - A_P)$ | $Melhor$ | $p-value(DAM)$ | $p-value(EQM)$ |
|---|---|---|---|---|---|
| LinearRegression | Betareg | 0.3655087 | Betareg | 0.01953 * | 0.003906 ** |
| LinearRegression | IBk | -0.04859978 | LinearRegression | 1 | 0.7695 |
| LinearRegression | M5R | -0.4015365 | LinearRegression | 0.04883 * | 0.375 |
| LinearRegression | M5P | -0.2564811 | LinearRegression | 0.1602 | 0.4922 |
| LinearRegression | MultilayerPerceptron | -3.051675 | LinearRegression | 0.001953 ** | 0.001953 ** |
| LinearRegression | SMOReg | 0.1708632 | SMOReg | 0.6953 | 0.8457 |
| Betareg | IBk | -0.3825999 | Betareg | 0.02734 * | 0.003906 ** |
| Betareg | M5R | -0.7318314 | Betareg | 0.003906 ** | 0.003906 ** |
| Betareg | M5P | -0.6200316 | Betareg | 0.009766 ** | 0.005859 ** |
| Betareg | MultilayerPerceptron | -3.197657 | Betareg | 0.001953 ** | 0.001953 ** |
| Betareg | SMOReg | -0.2123613 | Betareg | 0.4316 | 0.1055 |
| IBk | M5R | -0.3206625 | IBk | 0.08398 · | 0.04883 * |
| IBk | M5P | -0.1809823 | IBk | 0.2754 | 0.1602 |
| IBk | MultilayerPerceptron | -2.93329 | IBk | 0.001953 ** | 0.001953 ** |
| IBk | SMOReg | 0.204556 | SMOReg | 0.4316 | 0.9219 |
| M5R | M5P | 0.1736478 | M5P | 0.1508 | 0.2049 |
| M5R | MultilayerPerceptron | -2.794418 | M5R | 0.001953 ** | 0.001953 ** |
| M5R | SMOReg | 0.5738406 | SMOReg | 0.03711 * | 0.1602 |
| M5P | Multilayer | -2.960757 | M5P | 0.001953 ** | 0.001953 ** |
| M5P | SMOReg | 0.4420748 | SMOReg | 0.08398 · | 0.1602 |
| MultilayerPerceptron | SMOReg | -2.960757 | SMOReg | 0.001953 ** | 0.001953 ** |

· siginificante a 10%. * siginificante a 5%. ** siginificante a 1%.

Tabela 11 – Experimento - Testes de Hipóteses dos resultados dos algoritmos para o atributo-meta Performance-Aproximação

| $A_S$ | $A_P$ | $ad(A_S - A_P)$ | $Melhor$ | $p-value(DAM)$ | $p-value(EQM)$ |
|---|---|---|---|---|---|
| LinearRegression | Betareg | 0.3702795 | Betareg | 0.1934 | 0.625 |
| LinearRegression | IBk | 0.1108012 | IBk | 0.8457 | 0.9219 |
| LinearRegression | M5R | -0.2015623 | LinearRegression | 0.5566 | 0.8457 |
| LinearRegression | M5P | -0.1473122 | LinearRegression | 0.4922 | 0.375 |
| LinearRegression | MultilayerPerceptron | -2.526206 | LinearRegression | 0.001953 ** | 0.001953 ** |
| LinearRegression | SMOReg | 0.2460447 | SMOReg | 0.1055 | 0.1309 |
| Betareg | IBk | -0.2870923 | Betareg | 0.08398 · | 0.1309 |
| Betareg | M5R | -0.6510746 | Betareg | 0.009766 ** | 0.08398 · |
| Betareg | M5P | -0.5833168 | Betareg | 0.01953 * | 0.03711 * |
| Betareg | MultilayerPerceptron | -3.023292 | Betareg | 0.001953 ** | 0.001953 ** |
| Betareg | SMOReg | -0.130187 | Betareg | 0.4922 | 1 |
| IBk | M5R | -0.3467949 | IBk | 0.1309 | 0.08398 · |
| IBk | M5P | -0.2852643 | IBk | 0.2324 | 0.06445 · |
| IBk | MultilayerPerceptron | -2.755164 | IBk | 0.001953 ** | 0.001953 ** |
| IBk | SMOReg | 0.1503297 | SMOReg | 0.4316 | 0.5566 |
| M5R | M5P | 0.05841226 | M5P | 0.4469 | 0.7998 |
| M5R | MultilayerPerceptron | -2.500483 | M5R | 0.001953 ** | 0.001953 ** |
| M5R | SMOReg | 0.498372 | SMOReg | 0.009766 ** | 0.01953 * |
| M5P | MultilayerPerceptron | -2.532796 | M5P | 0.001953 ** | 0.001953 ** |
| M5P | SMOReg | 0.4352027 | SMO | 0.01953 * | 0.01367 * |
| MultilayerPerceptron | SMOReg | 2.869476 | SMOReg | 0.001953 ** | 0.001953** |

· siginificante a 10%. * siginificante a 5%. ** siginificante a 1%.

Tabela 12 – Experimento - Testes de Hipóteses dos resultados dos algoritmos para o atributo-meta Performance-Evitação

| $A_S$ | $A_P$ | $ad(A_S - A_P)$ | $Melhor$ | $p-value(DAM)$ | $p-value(EQM)$ |
|---|---|---|---|---|---|
| LinearRegression | Betareg | 1.33303 | Betareg | 0.001953 ** | 0.003906 ** |
| LinearRegression | IBk | 0.5139394 | IBk | 0.03711 * | 0.1055 |
| LinearRegression | M5R | 0.4049812 | M5R | 0.1235 | 0.1235 |
| LinearRegression | M5P | 0.4826676 | M5P | 0.0972 · | 0.1235 |
| LinearRegression | MultilayerPerceptron | -3.431047 | LinearRegression | 0.001953 ** | 0.001953 ** |
| LinearRegression | SMOReg | 0.3051552 | SMOReg | 0.4316 | 0.4316 |
| Betareg | IBk | -0.8670733 | Betareg | 0.01953 * | 0.02734 * |
| Betareg | M5R | -0.7986243 | Betareg | 0.01367 * | 0.009766 ** |
| Betareg | M5P | -0.7978429 | Betareg | 0.01367 * | 0.009766 ** |
| Betareg | MultilayerPerceptron | -4.639433 | Betareg | 0.001953 ** | 0.001953 ** |
| Betareg | SMOReg | -1.24511 | Betareg | 0.001953 ** | 0.001953 ** |
| IBk | M5R | -0.05821388 | IBk | 0.7695 | 0.9219 |
| IBk | M5P | -0.002511461 | IBk | 0.7695 | 0.9219 |
| IBk | MultilayerPerceptron | -3.942737 | IBk | 0.001953 ** | 0.001953 ** |
| IBk | SMOReg | -0.2640946 | IBk | 0.1934 | 0.009766 ** |
| M5R | M5P | 0.05291748 | M5P | 1 | 1 |
| M5R | MultilayerPerceptron | -3.674368 | M5R | 0.001953 ** | 0.001953 ** |
| M5R | SMOReg | -0.1660036 | SMOReg | 0.3223 | 0.02734 * |
| M5P | MultilayerPerceptron | -3.825526 | M5P | 0.001953 ** | 0.001953 ** |
| M5P | SMOReg | -0.2414602 | M5P | 0.2324 | 0.009766 ** |
| MultilayerPerceptron | SMOReg | 3.88211 | SMOReg | 0.001953 ** | 0.001953 ** |

· siginificante a 10%. * siginificante a 5%. ** siginificante a 1%.

### 4.2.6 Análise dos Resultados

Por meio das Tabelas 10, 11 e 12, verifica-se que o algoritmo BetaReg apresentou os melhores resultados para Aprender, Performance-Aproximação e Performance-Evitação.

Entretanto, em alguns casos não houve significância estatística. Por exemplo, na comparação entre o algoritmo BetaReg e o algoritmo SMOReg para os atributos-meta

Aprender e Performance-Aproximação. Também não houve significância estatística na comparação entre o algoritmo BetaReg e o algoritmo LinearRegression para o atributo-meta Performance-Aproximação e entre o algoritmo BetaReg e o algoritmo IBk para a comparação da medida EQM para o atributo-meta Performance-Aproximação.

Para o atributo-meta Performance-Evitação, o algoritmo BetaReg foi superior com significância estatística em todas as situações, inclusive sendo superior os algoritmos LinearRegression e SMOReg com significância de 1%.

Por fim, o algoritmo MultilayerPerceptron apresentou os piores resultados para predição de todos os atributos-meta.

Desta forma, dado que o algoritmo BetaReg apresentou melhores resultados em todos os testes (superior com significância estatística em 83,33% dos casos) e que os modelos testados correspondem a modelos candidatos, podendo ainda ser melhorados, utilizaremos para a construção dos modelos finais a técnica de regressão beta.

## 4.3 CONSTRUÇÃO DO MODELO PARA CLASSIFICAÇÃO DA MOTIVAÇÃO

Nesta seção, é apresentado o processo de construção dos modelos para classificação dos perfis de motivação dos estudantes. Antes, porém, faz-se necessária uma explicação um pouco mais detalhada acerca do modelo de regressão beta e uma breve introdução ao processo de seleção de covariadas.

### 4.3.1 Modelo de Regressão Beta

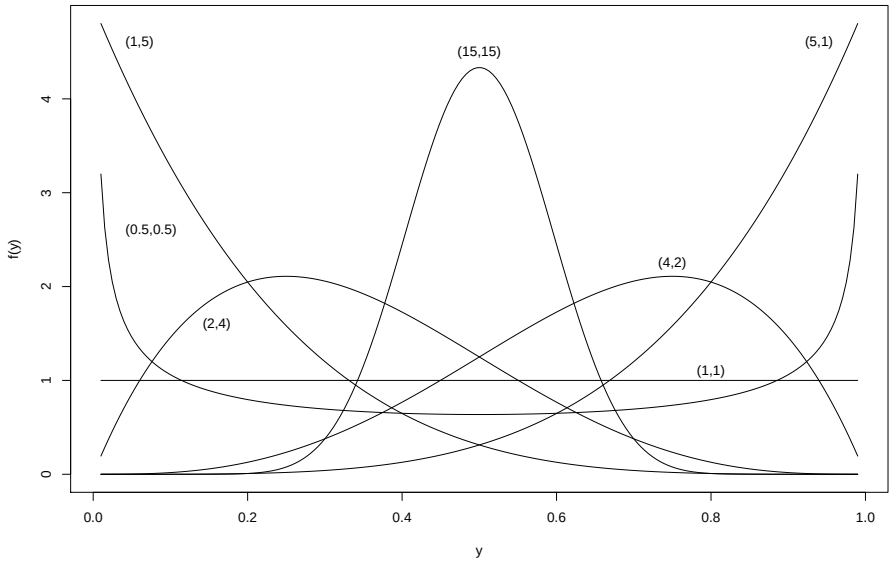
A distribuição beta é tipicamente usada para modelar variáveis aleatórias que se distribuem de forma contínua no $(0,1)$, tais como taxas, porcentagens e escores. A densidade beta é bastante flexível podendo assumir diversas formas, dependendo da combinação de valores de seus parâmetros. Inclusive, a distribuição beta pode assumir a forma de sino no $(0,1)$, própria da distribuição normal nos reais. Ver Equação 4.19 e a Figura 13.

Seja $y_1,\ldots,y_n$ uma amostra de variáveis independentes tal que cada $y_i$, $i = 1,\ldots,n$, segue a distribuição beta com densidade:

$$f(y;\mu_i,\phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i-1}(1-y)^{(1-\mu_i)\phi_i-1}, \quad 0<y<1, \tag{4.19}$$

onde $0<\mu_i<1$ e $\phi_i>0$. Aqui, $\mathrm{E}(y_i)=\mu_i$ e $\mathrm{var}(y_i)=(\mu_i(1-\mu_i))/(1+\phi_i)$, tem-se que $\phi$ representa um parâmetro de precisão, já que quanto maior $\phi$ menor a variância de $y_i$, consequentemente, $\phi^{-1}$ é um parâmetro de dispersão.

Figura 13 – Diferentes formas da densidade beta.




Ferrari e Cribari-Neto (2004) propõem que a média da variável resposta $y_i$, i.e., $\mu_i$ possa ser escrita como

$$g(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n.$$

Como o objetivo da função de ligação $g$ é permitir que os $\widehat{\beta}_i$'s possam assumir qualquer valor real e dado que $\mu_i \in (0,1)$, uma função de ligação que conduz essa média a todos os reais é a função de ligação logito, dada por

$$\mu_i \in (0,1) \leftrightarrow \log\left\{\frac{\mu_i}{(1-\mu_i)}\right\} \in (-\infty, +\infty) = IR.$$

Assim, como utilizamos um modelo matemático para tentar explicar a média da variável resposta $\mu_i$, que é um parâmetro desconhecido, também é possível e, às vezes, necessário fazer o mesmo com a variância da resposta que, neste caso, implica em modelar o parâmetro $\phi$. De fato, é sugerido um modelo para $\phi$ quando se suspeita que a dispersão não é constante para os dados, ou que é possível haver grupos com dispersões diferentes. Assim, Smithson e Verkuilen (2006) propõem um modelo de regressão beta em que

$$h(\phi_i) = \gamma_1 + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \ldots + \gamma_q z_{iq}, \quad i = 1, \ldots, n.$$

Neste caso, como $\phi > 0$ uma função de ligação adequada é $\log(\phi) \in IR$. Note que é necessário estimar os $\beta_t$'s ($t = 1, \ldots, k$) e os $\gamma_j$'s ($j = 1, \ldots, q$) para que $\phi_i$ e $\mu_i$ sejam estimados.

Isto é feito utilizando o método de máxima verossimilhança[4].

Pesquisadores podem usar o pacote `betareg`, que está disponível no software estatístico `R`.

Note que $\beta = (\beta_1, \ldots, \beta_k)^\top$ e $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ são vetores e $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)^\top$ e $\widehat{\gamma} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_q)^\top$ são os seus respectivos estimadores de máxima verossimilhança.

### 4.3.2 Seleção de covariadas

Uma das etapas mais importantes na construção de um modelo de regressão é a seleção de covariadas, ou seja, a escolha das variáveis que serão eleitas para tentar explicar a média da variável resposta. Existem vários critérios de seleção de covariadas na literatura estatística. No entanto, todos são muito suscetíveis a problemas que podem estar presentes na amostra estudada.

Assim, antes de utilizar qualquer algoritimo automático para escolhas de covariadas, é interessante realizar uma análise descritiva dos dados de forma, por exemplo, a avaliar o grau de correlação de cada covariada candidata com a variável resposta e também no sentido de verificar o problema de multicolinearidade. Esse problema ocorre quando em um modelo de regressão existem duas ou mais covariadas com um alto grau de correlação entre si. Do ponto de vista matemático, esse problema pode invalidar as conclusões estatísticas. Do ponto de vista intuitivo, se existem, por exemplo, duas covariadas consideravelmente correlacionadas, ou seja, covariadas que contêm informações semelhantes sobre a resposta, não faz sentido considerar as duas. Nesse caso, é muito mais eficaz escolher aquela que apresenta maior correlação com a resposta.

Outra questão importante são os gráficos de dispersão de cada covariada com a resposta. Esses gráficos podem indicar que tipo de função pode ser aplicada à covariada para que sua importância para a explicação da resposta seja mais evidente.

Funções aplicadas às covariadas também são úteis quando as escalas entre as mesmas são muito distintas, ou ainda quando uma covariada é muito importante em uma parte da amostra, apresentando valores extremos para este subconjunto de dados e valores menos expressivos para o restante da amostra. Nesses casos, uma transformação logarítmica é bastante indicada.

Adicionalmente, é interessante avaliar a possibilidade de interações entre as variáveis candidatas. Essas iterações, que matematicamente são expressas no modelo como o produto de covariadas, podem ser extremamente importantes para melhorar a qualidade do modelo.

A questão de multicolinearidade e a análise prévia dos dados são temas comuns na estatística e na mineração de dados. Um problema recorrente entre os algoritmos automáticos

[4] https://www.ime.usp.br/ giapaula/cursospos.htm

de seleção de covariadas, estatísticos ou não, é que os mesmos não conseguem identificar todos os aspectos acima citados. O modelo resultante da seleção automática pode ser um ponto de partida, para que então o pesquisador possa adaptar esse modelo com o objetivo de obter um modelo mais adequado para descrever a relação entre a resposta e as demais covariadas.

Ferramentas como o WEKA normalmente utilizam critérios para seleção automática de modelos e tipicamente as questões acima citadas não são consideradas em sua totalidade. Um desses critérios é o critério AIC, "Akaike information criterion"(AKAIKE, 1973; AKAIKE, 1974) presente no algoritmo LinearRegression para a escolha do melhor modelo.

Na subseção seguinte, explicaremos, de modo resumido, o Critério de Informação Akaike.

#### *4.3.2.1 Critério de Informação Akaike - stepAIC*

A ideia principal desse critério é encontrar o modelo mais parcimonioso, ou seja, o modelo que melhor se ajusta aos dados, com um menor número de parâmetros. Esse critério é usado sem envolver testes estatísticos.

Essa função foi proposta por Akaike (1974), onde ele se baseia na função de log-verossimilhança acrescida de uma penalidade pelo número de parâmetros do modelo. Entre vários modelos candidatos, deve-se escolher aquele que apresentar o menor valor da função AIC.

$$\text{AIC} = -2l(\widehat{\beta},\widehat{\gamma}) + 2p$$

onde $p = k + q$ é o número de parâmetros e $l(\widehat{\beta},\widehat{\gamma})$ é o logaritmo da função de verossimilhança avaliado nos estimadores de máxima verossimilhança de $\beta$ e $\gamma$.

No caso do modelo normal linear temos que

$$\text{AIC} = -2l(\widehat{\beta},\widehat{\sigma}^2) + 2p$$

e $p = k + 1$.

Um algoritmo de seleção de modelos que se baseia na medida AIC é o **stepAIC**. O algoritmo inicia com um modelo com uma única covariada (a mais correlacionada com a resposta), em seguida testa a entrada de outra covariada com base no valor do AIC. Caso o AIC do modelo adicional seja menor que o AIC do modelo inicial, o algoritmo mantém a covariada adicionada, em seguida ele testa a saída da covariada inicial e escolhe o próximo modelo com base no menor AIC. Ou seja, passo a passo o algoritmo testa a entrada de uma covariada e a saída de uma ou mais covariadas. No final, o algoritmo fornece um modelo, aquele que

apresentou o menor AIC entre todas as combinações possíveis da variável resposta com as variáveis candidatas.

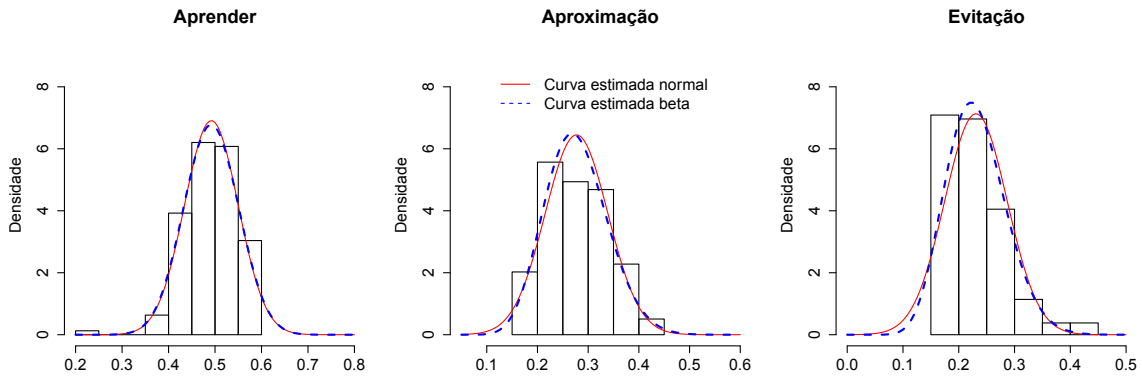
### 4.3.3 Seleção de covariadas para o modelo de classificação da motivação dos estudantes

Nesta subseção, apresentaremos o processo de seleção de covariadas (ou variáveis explicativas) que irão compor os modelos de regressão para classificação dos estudantes. Inicialmente, é apresentada a estatística descritiva das variáveis candidatas para a construção do modelo e, posteriormente, as seleções iniciais de covariadas para os modelos de Aprender, Performance-Aproximação e Performance-Evitação respectivamente.

#### *4.3.3.1 Estatística descritiva das variáveis candidatas para o modelo de classificação da motivação dos estudantes*

Inicialmente, foi realizada uma análise descritiva das três variáveis respostas, a saber: percentual de Aprender, percentual de Performance-Aproximação e percentual de Performance-Evitação. A Figura 14 apresenta um histograma das três variáveis, que foi gerado por meio de um Script do R que pode ser visto no Apêndice G.[5]

Figura 14 – Histograma das variáveis aleatórias




Nesta figura, percebe-se que a beta entendeu os dados como simétricos e, como a beta é bastante flexível, ela ajusta uma curva simétrica muito parecida como a normal. Este gráfico é muito importante pelas seguintes razões: primeiro que a proximidade da beta com a normal permite que seja utilizado com maior tranquilidade o procedimento stepAIC baseado no modelo normal transformado; segundo, apesar de estar no $(0,1)$, estes dados talvez fossem melhor ajustados por uma competidora direta da beta, a distribuição simplex.

[5] Esta etapa foi realizada após a seleção de covariadas utilizando o stepAIC, que será apresentada nas seções posteriores. Entretanto, é colocada aqui para motivar que, em trabalhos futuros, seja realizada inicialmente.

É importante notar que Aprender e Evitação apresentam valores atípicos, extremos. No caso de Aprender, existe um único percentual igual a aproximadamente 0.2, que torna os dados assimétricos à esquerda. Tanto a beta quanto a normal ignoram esse ponto e supõem simetria. No caso de Performance-Evitação, cinco valores extremos, entre 158 observações, atípicos à direita, pois estão entre 0.37 a 0.44 e causam uma certa assimetria (ignorada pelas duas distribuições).

O desafio é ajustar o modelo beta de forma a incluir esses pontos. Para isso, é preciso identificá-los e perceber quais características destes casos se destacam. Para observar melhor a distribuição dos três percentuais, foram construídos seu *boxplots* que são apresentados na Figura 15. Outro gráfico importante para entendimento dos valores extremos é o gráfico de dispersão das variáveis Aprender e Evitação que pode ser visto na Figura 16.

Figura 15 – Boxplots das variáveis respostas

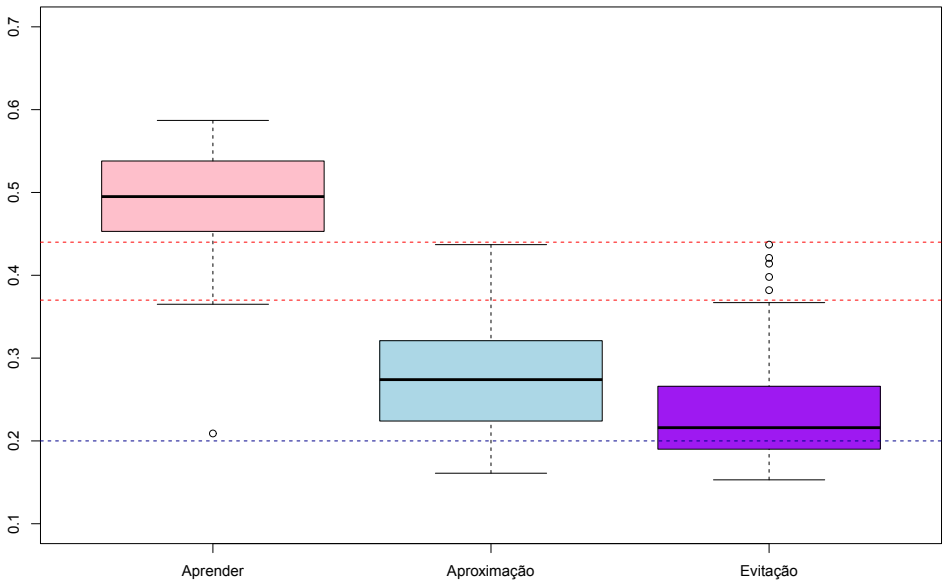
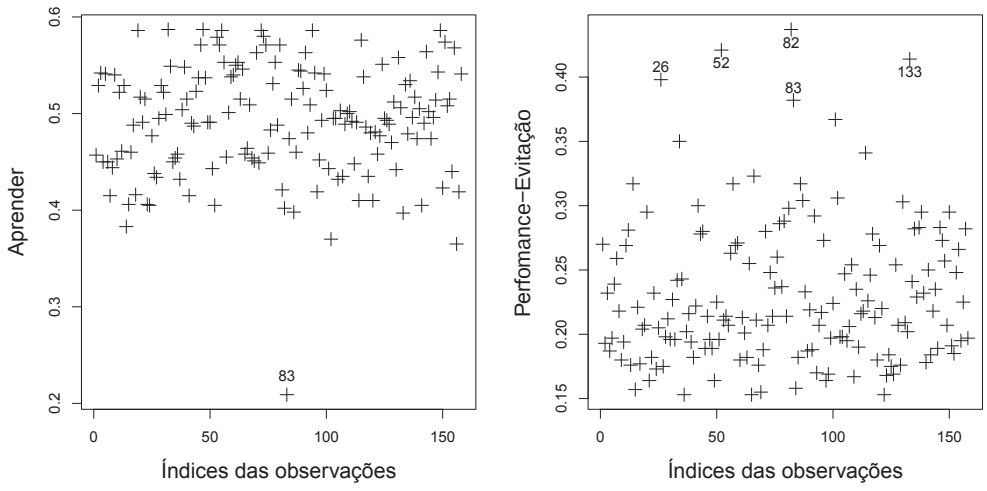


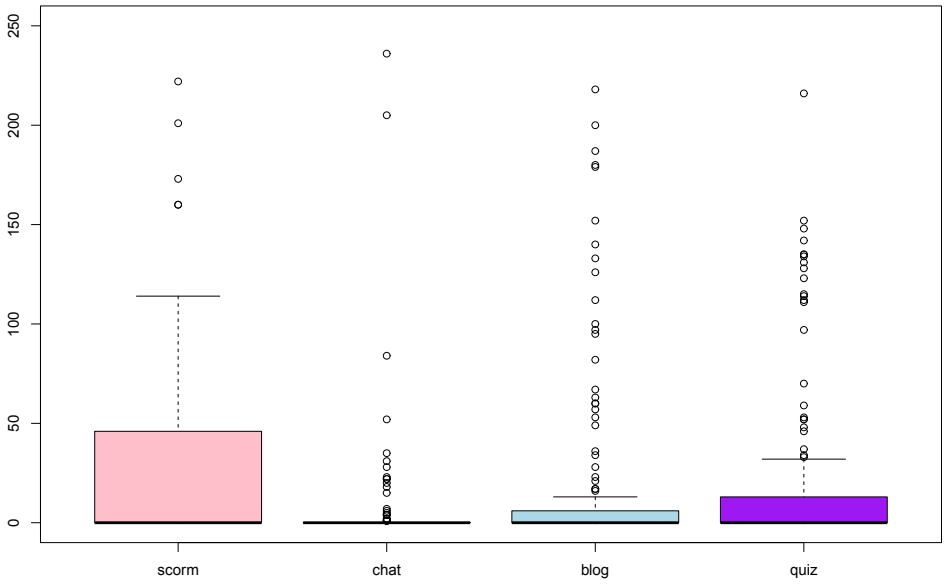

Figura 16 – Diagramas de dispersão das variáveis Aprender e Performance-Evitação



No caso de Aprender, a observação atípica é a de id=214 (Índice 83). Ao avaliar os valores das covariadas para esse caso, percebe-se que é um aluno que se destaca apenas em uma característica; é um dos maiores utilizadores do módulo de quiz, ou seja, um aluno com um percentual muito baixo de Aprender 0.209, mas com uma quantidade grande de realizações de quiz, 152. Isto é um sinal que quiz deve ser considerada no modelo de Aprender, de alguma forma, provavelmente, no modelo de dispersão.

Esse mesmo aluno tem um percentual de Evitação igual 0.382, um dos pontos atípicos no gráfico boxplot de Evitação. Outro ponto apresenta perfil parecido, ainda mais acentuado no caso de Evitação, percentual de Evitação igual 0.437, extremamente alto e com número de quiz realizado, igual a 142. Na Figura 17, é possível perceber como esses valores para quiz são extremamente altos, dado que a maioria dos alunos nem realiza quiz. Outro fator importante desses dois casos é que são alunos que também usam blog, mas não tanto quanto quiz. Assim, para o modelo do percentual de Evitação, é interessante considerar blog, quiz, IDquiz (se fez quiz) e IDblog (se fez blog). Notem que são cinco outliers para Performance-Evitação.

Figura 17 – Boxplots das variáveis candidatas a compor o modelo




O próximo caso apresenta um percentual Evitação de 0.414 e se destaca por ser um aluno do terceiro período, com baixo número de dias de interação, abaixo do primeiro quartil, e com 51 anos. Esse caso é interessante pois apesar de ser extremo, ele faz sentido. Se o indivíduo está no terceiro período e interagiu pouco é porque ele tem perfil de Evitação.

Os dois casos restantes também são de alunos mais velhos, no terceiro período. Esses três alunos não apresentam um percentual baixo de Aprender, apresentam percentual baixo de Performance-Aproximação.

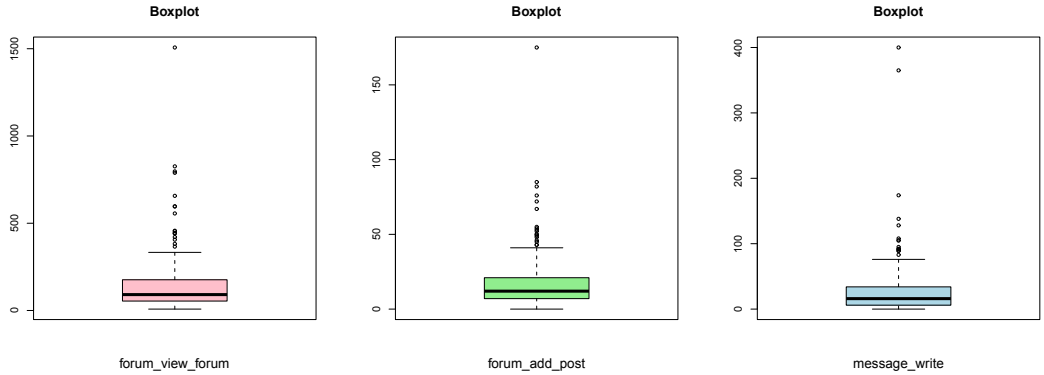
Na Tabela 13 são apresentadas as principais medidas estatísticas de algumas das variáveis candidatas utilizando o comando "summary"do R.

Tabela 13 – Principais medidas estatísticas das variáveis candidatas.

| Variável | Min. | 1º Qu. | Mediana | Média | 3º Qu. | Max. |
|---|---|---|---|---|---|---|
| diasinteracao | 95.0 | 281.0 | 299.5 | 369.1 | 462.0 | 587.0 |
| idade | 17.0 | 22.0 | 27.0 | 28.9 | 33.0 | 57.0 |
| genero | 0.0000 | 0.0000 | 0.0000 | 0.4937 | 1.0000 | 1.0000 |
| user_login | 10.0 | 209.2 | 358.5 | 424.9 | 538.8 | 1631.0 |
| course_view | 24.0 | 114.0 | 252.0 | 417.9 | 534.8 | 3265.0 |
| resource_view | 1.00 | 36.25 | 56.00 | 78.04 | 101.75 | 470.00 |
| wiki_view | 0.000 | 0.000 | 0.000 | 1.886 | 0.000 | 33.000 |
| forum_view_discussion | 0.00 | 11.00 | 64.00 | 96.53 | 119.00 | 657.00 |
| forum_view_forum | 8.00 | 54.25 | 91.00 | 157.37 | 175.75 | 1507.00 |
| forum_add_post | 0.00 | 7.00 | 12.00 | 18.54 | 21.00 | 175.00 |
| forum_update_post | 0.000 | 0.000 | 1.000 | 2.209 | 2.000 | 39.000 |
| wiki_comments | 0.0000 | 0.0000 | 0.0000 | 0.9494 | 0.0000 | 23.0000 |
| message_write | 0.00 | 6.00 | 16.00 | 31.01 | 34.00 | 400.00 |
| assign | 0.00 | 48.75 | 171.00 | 221.47 | 324.00 | 1134.00 |
| blog | 0.00 | 0.00 | 0.00 | 17.59 | 6.00 | 218.00 |
| chat | 0.000 | 0.000 | 0.000 | 5.247 | 0.000 | 236.000 |
| message | 0.00 | 9.00 | 22.00 | 37.65 | 46.00 | 412.00 |
| quiz | 0.00 | 0.00 | 0.00 | 18.19 | 12.50 | 216.00 |
| scorm | 0.00 | 0.00 | 0.00 | 23.49 | 45.75 | 222.00 |
| wiki | 0.000 | 0.000 | 0.000 | 3.481 | 0.000 | 70.000 |

Note que algumas variáveis apresentam distribuições com ocorrência de valores extremos, ocorrendo ao mesmo tempo valores muito pequenos e valores muito grandes. Os boxplots apresentados nas Figuras 17 e 18 evidenciam melhor as distribuições de valores extremos dessas variáveis.

Figura 18 – Boxplots das variáveis candidatas a compor o modelo - 1



Nestes casos, a transformação logarítmica das variáveis é bastante útil no sentido de amenizar essa ocorrência de valores extremos, o que poderia mascarar a importância ou não de uma variável para a explicação da resposta. Quando existem ocorrências de zero é necessário utilizar a transformação $\log(x_{tk} + 1.0)$.

#### *4.3.3.2 Seleção inicial de covariadas para o modelo de Aprender*

Para chegarmos às covariadas para o modelo Aprender, inicialmente carregamos todo o *dataset* e, em seguida, aplicamos a função logito à variável per_aprender de modo a transformar os dados desta variável aleatória do invervalo dos unitários para os reais. Esse processo é indicado para que possamos utilizar o stepAIC da distribuição normal. Na sequência, foi aplicada a regressão linear normal considerando todas as variáveis do *dataset* e, posteriormente, foi aplicado o método stepAIC de modo a selecionar as covariadas. Por fim, foi ajustado o modelo beta utilizando as covariadas selecinadas pelo stepAIC.

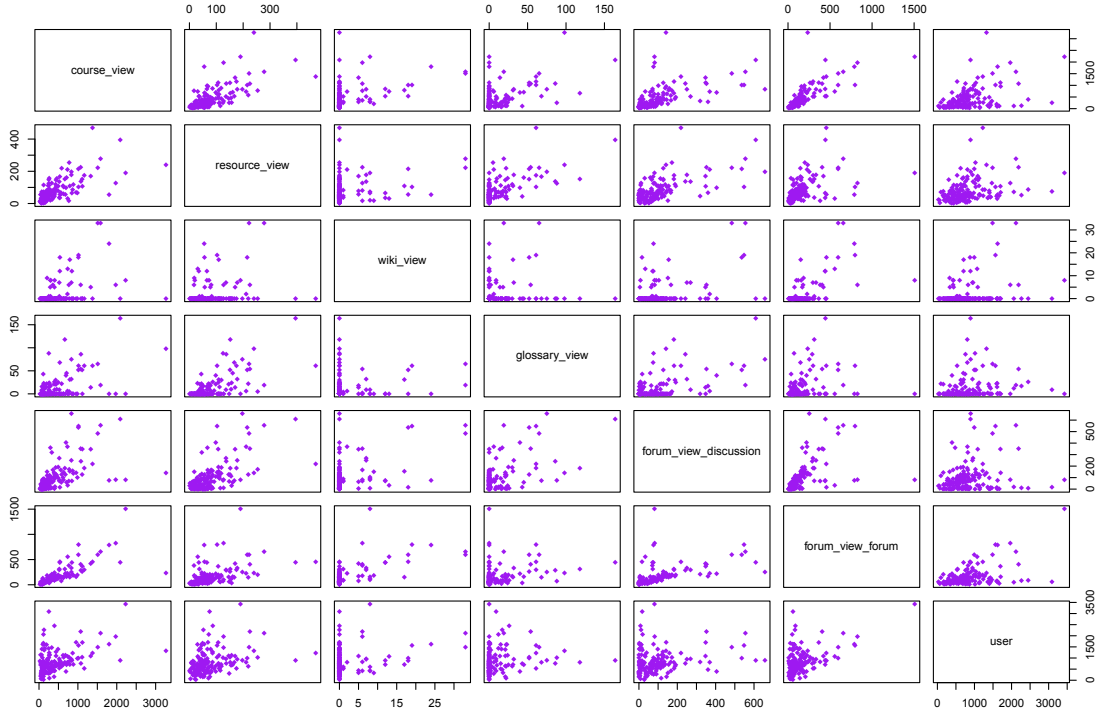
As covariadas selecionadas pelo stepAIC normal e os resultados de p-valor para estas variáveis via stepAIC Normal e Modelo de Regressão Beta podem ser vistos na Tabela 14. Com base nesta tabela, podem ser observadas várias questões estatísticas importantes que a seleção automática não consegue levar em consideração.

Tabela 14 – Variáveis selecionadas pelo stepAIC Normal e Modelo de regressão beta baseado nestas variáveis para o modelo Aprender.

| Covariadas | p-valor | |
|---|---|---|
| | stepAIC Normal | Modelo de regressão beta |
| (Intercept) | 0.002020 ** | 0.000795 |
| diasinteracao | 0.091055 · | 0.066064 · |
| idade | 0.065305 · | 0.049332 * |
| genero | 0.054148 · | 0.041689 * |
| user_login | 0.123618 | 0.096182 · |
| forum_view_discussion | 0.000180 *** | 0.000254 *** |
| forum_view_forum | 0.000139 *** | 0.000187 *** |
| forum_add_post | 0.092599 · | 0.100530 |
| forum_update_post | 0.020575 * | 0.016013 * |
| wiki_comments | 0.088028 · | 0.067782 · |
| glossary_add_entry | 0.002952 ** | 0.001254 ** |
| course | 0.018570 * | 0.012414 * |
| forum | 0.000125 *** | 0.000185 *** |
| glossary | 0.000913 *** | 0.000320 *** |
| message | 0.157027 | 0.134331 |
| user | 0.083384 · | 0.063076 · |
| wiki | 0.022250 * | 0.013985 * |
| · siginificante a 10%. * siginificante a 5%. ** siginificante a 1%. | | |
| *** siginificante a mais 1% | | |

Na Figura 19, estão apresentados os gráficos de dispersão entre algumas variáveis candidatas a compor o modelo. Com base nesta figura, é possível perceber como as variáveis originais course_view, resource_view, wiki_view, glossary_view, forum_view_discussion, forum_view_forum e user estão consideravelmente correlacionadas entre si, dado que nos diagramas de dispersão é possível identificar uma proximidade de uma tendência linear, ou seja uma reta do tipo $y = ax + b$. Esta reta aproximada que aparece nos diagramas de dispersão indica que uma variável pode explicar a outra. Isto é, as variáveis contêm o mesmo tipo de informação. Este é um sinal de grau de multicolinearidade o qual afeta as estimativas dos coeficientes, o sinal esperado para essas estimativas e a significância das covariadas (p-valor).

Figura 19 – Diagramas de dispersão entre variáveis candidatas a compor o modelo Aprender




Por exemplo, o sinal negativo de forum no modelo beta não faz muito sentido (vide Apêndice D, pois, desta forma, quanto maior a interação do aluno com o módulo fórum menor seria o seu percentual de Aprender. De fato, as variáveis: course_view, resource_view, wiki_view, glossary_view, forum_view_discussion, forum_view_forum não devem entrar nem conjuntamente, nem duas a duas em qualquer modelo. Deve-se escolher, entre estas todas, aquela que seja mais importante para a explicação da resposta de interesse e que dê mais sentido ao modelo. Por exemplo, escolher user, que está relacionado à quantidade de *logins* e *logouts* no sistema, faz menos sentido que escolher forum_view_discussion (quando um usuário visualiza uma discussão em um fórum), a qual diz respeito a uma ação mais efetiva no sentido de formar um perfil de Aprender.

Um modelo beta inicial para percentual de Aprender poderia considerar: idade, diasinteracao, genero, forum_update_post, glossary, wiki e as interacões entre essas covariadas, ou seja, o produto destas covariadas.

#### *4.3.3.3 Seleção inicial de covariadas para o modelo de Performance-Aproximação*

O processo para obtenção das covariadas seguiu os mesmos passos apresentados na Subseção 4.3.3.2 sendo, neste caso, transformada por meio da função logito a variável per_aprox.

As covariadas selecionadas pelo stepAIC normal e os resultados de p-valor para estas variáveis via stepAIC Normal e Modelo de Regressão Beta podem ser vistos na Tabela 15.

Tabela 15 – Variáveis selecionadas pelo stepAIC Normal e Modelo de regressão beta baseado nestas variáveis para o modelo Performance-Aproximação.

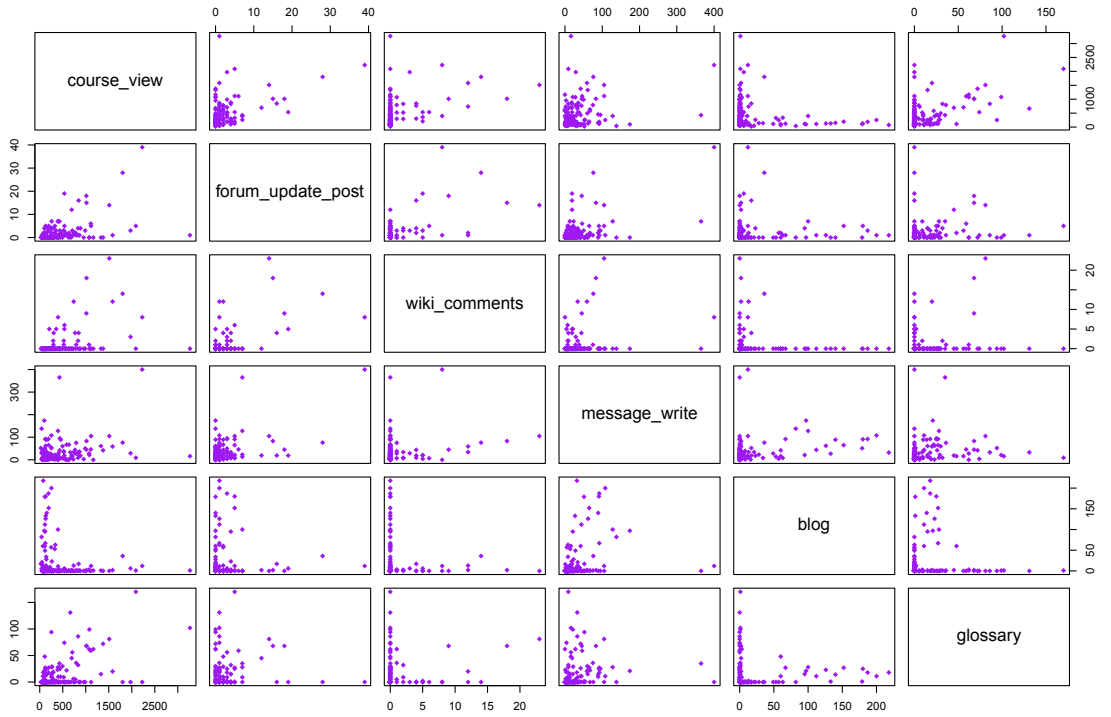
| Covariadas | p-valor | |
|---|---|---|
| | stepAIC Normal | Modelo de regressão beta |
| (Intercept) | 1.02e-13 *** | < 2e-16 *** |
| idade | 0.12188 | 0.143777 |
| genero | 0.00722 ** | 0.006117 ** |
| course_view | 0.02506 * | 0.015174 * |
| glossary_view | 0.01330 * | 0.010144 * |
| forum_add_post | 0.16063 | 0.117400 |
| forum_update_post | 0.05203 · | 0.043262 * |
| wiki_comments | 0.04574 * | 0.030292 * |
| glossary_add_entry | 0.00169 ** | 0.000853 *** |
| assign | 0.11940 | 0.085847 · |
| blog | 0.03701 * | 0.015488 * |
| glossary | 0.00900 ** | 0.006585 ** |
| message | 0.03434 * | 0.015709 * |
| IDscorm | 0.01064 * | 0.005022 ** |
| user | 0.09038 . | 0.044497 * |
| wiki | 0.02712 * | 0.014974 * |
| IDblog | 0.16096 | 0.143145 |
| · siginificante a 10%. * siginificante a 5%. ** siginificante a 1%. | | |
| *** siginificante a mais 1% | | |

Com base nos resultados do Modelo de Regressão Beta apresentado nesta tabela, as variáveis consideradas importantes para a explicação do percentual de perfil de Performance-Aproximação seriam: genero, course_view, glossary_view, forum_update_post, wiki_comments, glossary_add_entry, blog, glossary, message, IDscorm, user, wiki. No entanto, considerando o problema de multicolineariedade foi testado quais destas eram de fato importantes para o modelo.

Foi utilizado o critério equivalente ao AIC para escolher o melhor modelo entre modelos construídos com apenas uma das variáveis candidatas e como resposta o percentual da Performance-Aproximação. O modelo final selecionado incluiria as variáveis blog, IDblog, course_view e genero. É possível tentar incluir forum_update_post aplicando alguma transformação ou considerando alguma interação com outra covariável, de forma a diminuir sua alta correlação com course_view.

Na Figura 20, estão apresentados os gráficos de dispersão entre algumas das variáveis candidatas a compor este modelo.

Figura 20 – Diagramas de dispersão entre variáveis candidatas a compor o modelo Performance-Aproximação




### *4.3.3.4 Seleção inicial de covariadas para o modelo Performance-Evitação*

O processo para obtenção das covariadas para o modelo de Performance-Aproximação seguiu os mesmos passos apresentados nos modelos anteriores.

As covariadas selecionadas pelo stepAIC normal e os resultados de p-valor para estas variáveis via stepAIC Normal e Modelo de Regressão Beta podem ser vistos na Tabela 16.

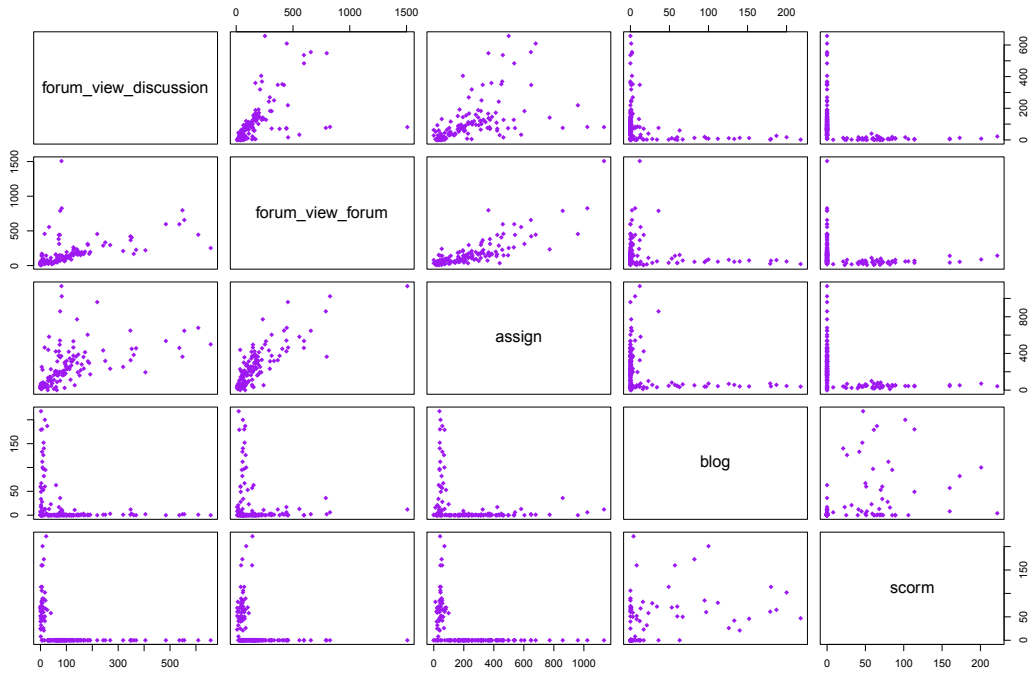
Tabela 16 – Variáveis selecionadas pelo stepAIC Normal e Modelo de regressão beta baseado nestas variáveis para o modelo de Performance-Evitação.

| Covariadas | p-valor | |
|---|---|---|
| | stepAIC Normal | Modelo de regressão beta |
| (Intercept) | < 2e-16 *** | < 2e-16 *** |
| genero | 0.0874 · | 0.0812 · |
| resource_view | 0.1309 | 0.1086 |
| forum_view_discussion | 3.22e-05 *** | 3.82e-06 *** |
| forum_view_forum | 7.41e-05 *** | 9.19e-06 *** |
| assign | 0.0190 * | 0.0141 * |
| blog | 0.0496 * | 0.0568 · |
| forum | 2.66e-05 *** | 2.51e-06 *** |
| message | 0.1124 | 0.1029 |
| scorm | 0.0466 * | 0.0474 * |
| · siginificante a 10%. * siginificante a 5%. ** significante a 1%. | | |
| *** siginificante a mais 1% | | |

Seguindo o princípio de exclusão de variáveis correlacionadas e escolha das variáveis mais importantes para a explicação do percentual de Performance-Evitação, as variáveis que ficam neste modelo inicial são forum_view_forum, blog e scorm.

Na Figura 21, estão apresentados os gráficos de dispersão entre algumas variáveis candidatas a compor este modelo.

Figura 21 – Diagramas de dispersão entre variáveis candidatas a compor o modelo Performance-Evitação



### 4.3.4 Modelos Finais

Após a execução das etapas anteriores, foram construídos os modelos beta finais para Aprender, Performance-Aproximação e Performance-Evitação.

Nesta etapa, foram utilizados todos os insumos das etapas anteriores, tais como análises de correlação, dados obtidos através do stepAIC para a normal e posteriores modelos de regressão beta de partida. A partir disso, foram realizados testes de ajustes utilizando funções como a transformação logarítmica de variáveis para amenizar valores extremos e interação entre variáveis de modo a obter modelos mais precisos.

Desta forma, os modelos finais foram obtidos por meio da execução dos Códigos 4.7, 4.8 e 4.9 no R. O código fonte completo, bem como os resultados da execução podem ser vistos no Apêndice H.

Código 4.7 – Modelo final para Aprender

```
1 > fitAprender<-betareg(per_aprender~idade:genero + log(forum_update_post + 1.0):genero
      + log(course_view):idade | exp(per_evit):IDquiz+log(per_evit) + forum_update_post
    :genero + log(forum_update_post+1.0):scorm)
```

Código 4.8 – Modelo final para Performance-Aproximação

```
1 > fitAproximacao<-betareg(per_aprox~idade:genero+log(scorm + 1.0):IDscorm + log(
    course_view):genero + blog:IDblog | scorm:idade + log(course_view):genero + scorm:
    IDscorm + genero:idade)
```

Código 4.9 – Modelo final para Performance-Evitação

```
1 > IDblogIDquiz<-IDblog+IDquiz
2 > fitEvitacao<-betareg(per_evit~quiz:IDquiz + log(message_write +1.0) + blog:IDblog +
    log(forum_add_post+1.0):scorm + log(forum_view_forum):scorm | forum_add_post:scorm
     + quiz:IDblogIDquiz + message + log(per_aprox) + forum_view_forum:genero)
```

Com isto, temos as expressões matemáticas para predizer os percentuais de Aprender, Performance-Evitação e Performance-Aproximação de cada aluno. Denotando p_Aprender, como percentual estimado de Aprender (Equação 4.20), p_Aprox, como percentual estimado de Performance-Aproximação (Equação 4.21) e p_Evit, como percentual estimado de Performance-Evitação (Equação 4.22), segue que:

$$p_{\text{Aprender}} = \frac{\exp(-0.144 + 0.003\text{idade} * \text{genero} + 0.06\log(\text{forum-update-post} + 1) * \text{genero} + 0.0005\log(\text{course-view}) * \text{idade})}{1 + \exp(-0.144 + 0.003\text{idade} * \text{genero} + 0.06\log(\text{forum-update-post} + 1) * \text{genero} + 0.0005\log(\text{course-view}) * \text{idade})} \tag{4.20}$$

$$p_{\text{Aprox}} = \frac{\exp(-0.96 - 0.01\text{idade} * \text{genero} + 0.02\log(\text{scorm} + 1) * \text{IDscorm} + 0.03\log(\text{course-view}) * \text{genero} + 0.001\text{blog} * \text{IDblog})}{1 + \exp(-0.96 - 0.01\text{idade} * \text{genero} + 0.02\log(\text{scorm} + 1) * \text{IDscorm} + 0.03\log(\text{course-view}) * \text{genero} + 0.001\text{blog} * \text{IDblog})} \tag{4.21}$$

$$p_{\text{Evit}} = \frac{\begin{array}{l}\exp(-1.2 + 0.002\text{quiz} * \text{IDquiz} - 0.02\log(\text{message-write} + 1) - 0.001\text{blog} * \text{IDblog} \\ \quad - 0.001\log(\text{forum-add-post}+1.0) * \text{scorm} - 0.001\log(\text{forum-view-forum}) * \text{scorm})\end{array}}{\begin{array}{l}1 + \exp(-1.2 + 0.002\text{quiz} * \text{IDquiz} - 0.02\log(\text{message-write} + 1) - 0.001\text{blog} * \text{IDblog} \\ \quad - 0.001\log(\text{forum-add-post}+1.0) * \text{scorm} - 0.001\log(\text{forum-view-forum}) * \text{scorm})\end{array}} \tag{4.22}$$

## 4.4 VALIDAÇÃO - ANÁLISE DE DIAGNÓSTICO

Verificar se um determinado modelo é uma representação adequada dos dados é um passo importante da análise estatística. A construção de um modelo de regressão envolve a definição da distribuição da variável de resposta, a escolha da função de ligação, a escolha das covariadas. Vários fatores podem levar um modelo ajustado pobre, por exemplo, a função de ligação inadequada, omissão de covariadas importantes, a escolha errada da distribuição da variável resposta, pontos influentes, especificação incorreta da variância entre outros fatores.

Ou seja, modelos estatísticos estão baseadas em certas suposições. A fim de ter confiança na análise, devemos verificar se os pressupostos associados são válidos. Isso pode ser alcançado por meio de análise de diagnóstico. Normalmente, esses diagnósticos são construídos em torno de resíduos e critérios de seleção como o $R^2$.

A maior parte dos resíduos é baseada nas diferenças entre as respostas observadas ($y$) e a média estimada ($\widehat{\mu}$). Por exemplo $r_i = y_i - \widehat{\mu}_i$, ou seja, o resíduo é uma medida de discrepância entre os dados reais e o modelo ajustado. Aqui vamos utilizar o resíduo proposto por Espinheira, Silva e Silva (2015), $r_{p,i}^{\beta\gamma}$ denominado resíduo combinado e baseado na diferença

$$(y_i^* - \widehat{\mu}_i^*), \quad \text{em que} \quad y_i^* = \log\left\{\frac{y_i}{(1-y_i)}\right\} \quad \text{e} \quad \mu_i^* = \text{E}(y_i^*).$$

Os gráficos de resíduos versus índices das observações ou versus valores preditos ($\widehat{\mu_i}$) são os mais básicos. Se um modelo está especificado corretamente, então estes gráficos não devem apresentar nenhuma tendência; os resíduos devem estar aleatoriamente distribuídos em torno do zero. A presença de quaisquer características sistemáticas tipicamente implica um falha de um ou mais pressupostos do modelo. Outro gráfico de resíduos importante é o gráfico de probabilidade normal com envelope simulado, que pode ser usado mesmo quando as distribuições empíricas dos resíduos não são normais. Se o modelo está adequado aos dados, esperamos que a maioria dos resíduos esteja aleatoriamente distribuída dentro das bandas do envelope.

As Figuras 22, 23 e 24 apresentam ambas as medidas: O resíduo combinado e baseado na diferença e o gráfico de probabilidade normal com envelope simulado respectivamente.

Figura 22 – Gráfico de Resíduos para o modelo de Aprender

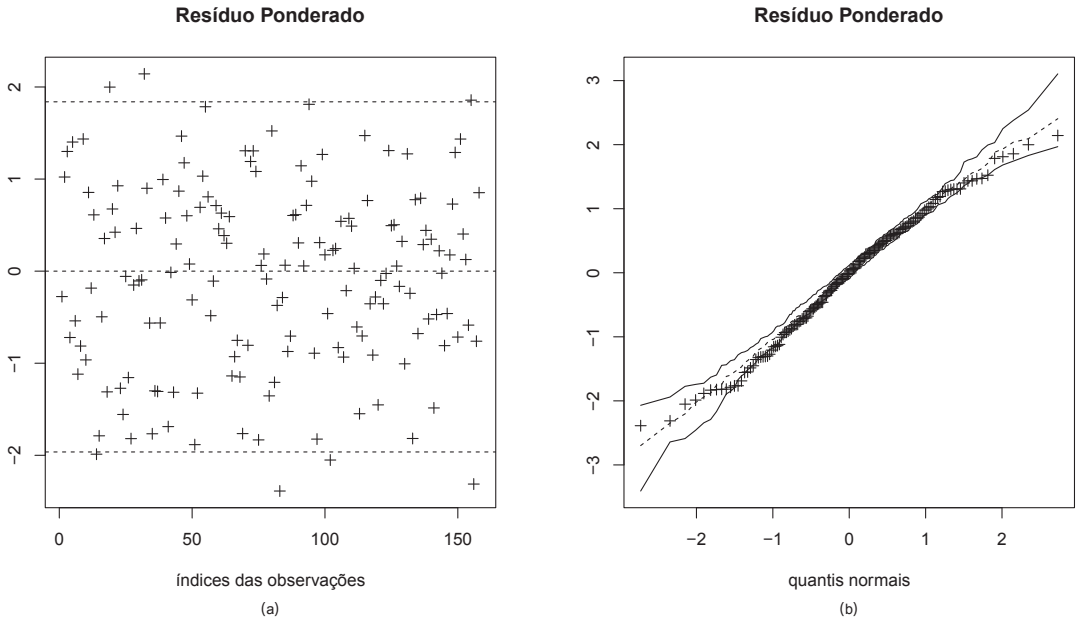



Figura 23 – Gráfico de Resíduos para o modelo de Performance-Aproximação

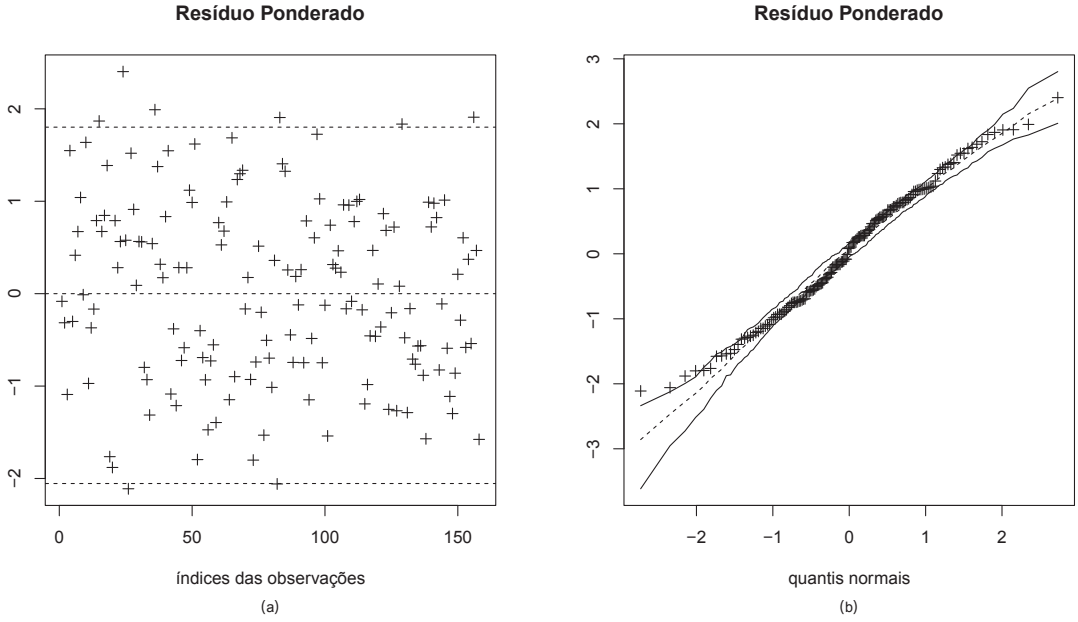
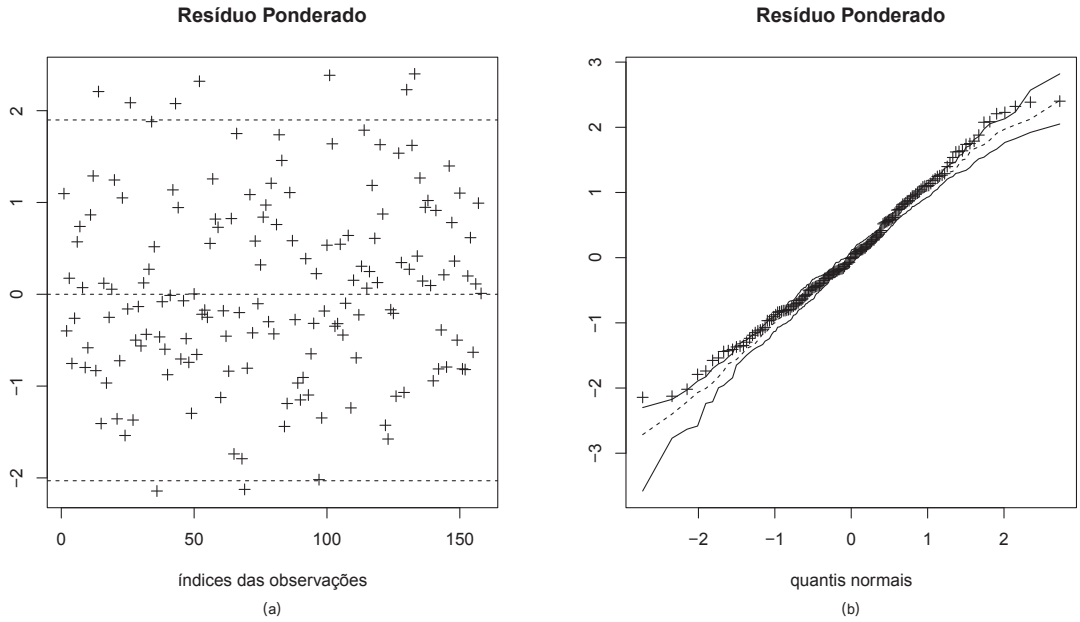


Figura 24 – Gráfico de Resíduos para o modelo de Performance-Evitação




Com base nestes gráficos de resíduos, percebe-se que os três modelos ajustados podem ser considerados adequados para descrever Aprender, Performance-Aproximação e Performance-Evitação com base nas variáveis extraídas dos logs do moodle e dos *scores* obtidos por meio dos questionários. Isto se dá pelo fato de que, nos seis gráficos apresentados, os resíduos encontram-se aleatoriamente distribuídos em torno do zero e também pela maioria dos resíduos estar aleatoriamente envolvido pelos envelopes.

## 4.5 DISCUSSÃO DOS RESULTADOS

De um modo geral, os resultados obtidos pelos modelos de regressão expressam o que é proposto na teoria de metas de realização.

Com base no modelo gerado para Aprender, percebe-se que forum_update_post é muito importante. De fato, o coeficiente positivo e alto quando comparado aos outros, indica que quanto mais atualizações um aluno faz em um post, o aumento, em seu percentual de Aprender, cresce, consideravelmente, muito mais do que se a ação for a de course_view. Isso faz sentido dado que, no contexto da EAD da UFAL, onde os dados foram coletados, a ferramenta Fórum é um dos principais objetos de aprendizagem. Por meio dele, alunos, tutores e professores discutem os conteúdos postados na plataforma, e, se um aluno está interessado em corrigir algo que fez errado, é porque entende que aquilo precisa ser melhorado e consequentemente está direcionado à meta Aprender.

Quanto à Performance-Aproximação, é notável a força das variáveis scorm e course_view; quanto maior scorm e course_view, maior é o percentual de Performance-Aproximação do

aluno. A presença de scorm neste modelo é bem interessante. SCORM ou *Sharable Content Object Reference Model* é um conjunto de padrões e especificações para e-learning de modo a prover interoperabilidade, acessibilidade e reutilização de conteúdo. O conceito é bastante amplo, entretanto, ao analisar os logs em seu formato RAW[6], percebemos que os objetos de aprendizagem marcados como scorm pelo sistema estão relacionados a formas lúdicas de se aprender, tais como: palavras cruzadas, associação de termos, textos embaralhados e etc. E estas atividades não são alvo principal do processo de aprendizagem no contexto da UFAL, contudo, os alunos da amostra tendem a executá-las com intensidade supostamente para "parecer inteligentes".

Finalmente, quanto à Performance-Evitação é notável como realizar quiz aumenta o percentual de Performance-Evitação do aluno. Esse percentual fica bastante explícito quando analisado o sujeito de id=214 (Índice 83). Ele possui um percentual de Aprender muito baixo e de Performance-Evitação muito alto. Este ponto bastante *outlier* para Aprender e também *outlier* para Performance-Evitação é um caso que descreve bem um aluno com Performance-Evitação. Ele gasta bastante tempo "brincando"no quiz e evita realizar as suas atividades acadêmicas, tais como, visualizar cursos, postar nos fóruns e etc. Isto pode ser um forte sinal que o uso do quiz deve ser reformulado ou que este módulo deve ter seu conteúdo melhorado, pois o aluno que realiza muito quiz está afetando negativamente o seu percentual de Aprender.

[6] Em Português: Cru

# 5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, foi apresentado um novo processo para a classificação da motivação de estudantes da educação on-line, mais especificamente, um processo que utiliza instrumentos de avaliação de motivação oriundos da psicologia e mineração de dados educacionais.

Foram listados trabalhos relacionados a esta proposta, trabalhos estes onde, na sua maioria, os pontos fracos das propostas convergiam para a falta de um método automático para classificação da motivação, uso claro de um arcabouço das teorias da psicologia e em alguns casos a ausência de um processo de validação.

Neste contexto, foi proposto uma nova abordagem para a classificação da motivação nos ambientes de aprendizagem on-line, baseado no uso de um instrumento de classificação de motivação, desenvolvido e validado por psicólogos, que tem como arcabouço a Teoria de Metas de Realização e na mineração de dados educacionais.

O desenvolvimento da proposta foi composto por três etapas. Na primeira etapa, foi construída a base de dados por meio da aplicação dos questionários e da coleta de logs. Na segunda etapa, foi realizado um experimento para seleção da técnica/algoritmo mais adequado para a construção do modelo. Por fim, na terceira etapa, foi realizado o processo de construção do modelo de classificação da motivação, bem como a sua validação.

Considerando a validação apresentada, a pesquisa atingiu o objetivo proposto neste trabalho, possibilitando a classificação da motivação dos alunos de cursos mediados por computador, na perspectiva da teoria de metas de realização. Entretanto, esta proposta possui algumas limitações. Uma delas repousa no fato de que a amostra coletada, embora tenha um número expressivo de observações, foi retirada de uma única instituição de ensino superior. Outra limitação é que, apesar de os modelos terem sido gerados e validados estatisticamente, não foi possível a execução de estudos de caso em ambientes reais.

Em suma, com este trabalho, pretendeu-se criar um novo método para a classificação da motivação de estudantes em ambientes on-line. Porém, para que ele de fato possa ser empregado em ambientes reais são necessários mais esforços no sentido de superar as limitações nele colocadas.

Desta forma, como trabalhos futuros, esta pesquisa propõe:

1. Ampliar o conjunto de dados por meio de parcerias com outras Universidades ou Instituições de Ensino Superior;
2. Verificar se os modelos se ajustam à generalização dos dados;
3. Desenvolver um plugin para o moodle que implemente os modelos propostos;

4. Validar o modelo por meio de estudos de caso reais;

5. Publicar artigos em periódicos relevantes visando disseminar o conhecimento produzido para a comunidade científica.

# REFERÊNCIAS

ABED. **Censo EAD.BR: Relatório Analítico da Aprendizagem a Distância no Brasil 2014**. Ibpex, 2014. 157 p. ISBN 9788541700566. Disponível em: <http://www.abed.org.br/censoead2014/CensoEAD2014_portugues.pdf>.

ACCORSI, D. M. P.; BZUNECK, J. A.; GUIMARãES, S. d. R. Envolvimento cognitivo de universitários em relação à motivação contextualizada. **Psico-USF (Impresso)**, v. 12, n. 2, p. 291–300, 2007. ISSN 1413-8271.

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: PETROV, B. N.; CSAKI, F. (Ed.). **Second International Symposium on Information Theory**. Budapest: Akadémiai Kiado, 1973. p. 267–281.

AKAIKE, H. A new look at the statistical model identification. **Automatic Control, IEEE Transactions on**, v. 19, n. 6, p. 716–723, Dec 1974. ISSN 0018-9286.

ALVES, L. Educaçâo a distância: conceitos e história no Brasil e no mundo. **Revista Brasileira de Aprendizagem e a Distância**, v. 10, n. 21, p. 83–92, 2011. ISSN 1086 - 1362.

AMES, C. **Classrooms: Goals, structures, and student motivation.** 1992. 261–271 p.

ANDERSON, T. W.; DARLING, D. A. Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 23, n. 2, p. 193–212, 1952. ISSN 00034851. Disponível em: <http://www.jstor.org/stable/2236446>.

ARCHER, J. Achievement goals as a measure of motivation in university students. **Contemporary Educational Psychology**, v. 19, n. 4, p. 430 – 446, 1994. ISSN 0361-476X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0361476X84710319>.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011. ISSN 1414-5685. Disponível em: <http://br-ie.org/pub/index.php/rbie/article/view/1301$\delimiter"026E3B2$nhttp://www.br-ie.org/pub/index.php/rbie/article/view/1301>.

BAKER, R. S. Data mining for education. In: ____. **International Encyclopedia of Education**. [S.l.]: Elsevier Science Limited, 2010. ISBN 9780080448930.

BELUCE, A. C.; OLIVEIRA, K. L. d. Students' Motivation for Learning in Virtual Learning Environments. **Paidéia (Ribeirão Preto)**, scielo, v. 25, p. 105 – 113, 04 2015. ISSN 0103-863X. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-863X2015000100105&nrm=iso>.

BIAZUS, C. A. **Sistema de fatores que influenciam o aluno a evadir-se dos cursos de graduação na UFSM e na UFSC: um estudo no cursos de Ciências Contábeis**. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2004.

BOICHé, J.; STEPHAN, Y. Motivational profiles and achievement: A prospective study testing potential mediators. **Motivation and Emotion**, v. 38, n. 1, p. 79–92, 2013. Disponível em: <http://dx.doi.org/10.1007/s11031-013-9361-6>.

BORUCHOVITCH, E.; BZUNECK, J. **A MOTIVAÇÃO DO ALUNO - CONTRIBUIÇÕES DA PSICOLOGIA CONTEMPORÂNEA**. [S.l.]: Vozes, 2001. ISBN 8532625436.

BORUCHOVITCH, E.; BZUNECK, J. A. Motivação para aprender no brasil: estado da arte e caminhos futuros. In: ____. **Motivação para Aprender - Aplicações no Contexto Educativo**. [S.l.]: Vozes, 2010. p. 231–250.

BRAMER, M. **Principles of Data Mining (Undergraduate Topics in Computer Science)**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007. ISBN 1846287650.

BRASIL. Decreto 5.622, de 19 de dezembro de 2005. Regulamenta o artigo 80 da Lei nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional. Brasília, DF, 2005. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5622.htm>.

BRENELLI, R. et al. **Dificuldades de aprendizagem no contexto psicopedagógico**. [S.l.]: Vozes, 2001. ISBN 9788532626202.

BRITO, P. H. S. et al. A systematic approach for designing educational recommender systems. educational recommender systems and technologies. In: ____. **Software Design and Development: Concepts, Methodologies, Tools, and Applications**. [S.l.]: Information Resources Management Association (USA), 2012.

BZUNECK, J. A. Uma abordagem sócio-cognitivista à motivação do aluno: a teoria de metas de realização. **Psico-USF**, v. 4, n. 2, p. 51–66, 1999.

BZUNECK, J. A. A motivação do aluno orientado a metas de realização. In: ____. **A MOTIVAÇÃO DO ALUNO - CONTRIBUIÇÕES DA PSICOLOGIA CONTEMPORÂNEA**. [S.l.]: Vozes, 2004. p. 58–77.

BZUNECK, J. A. A motivação dos alunos em cursos superiores. In: ____. **Questões do cotidiano universitário**. [S.l.]: Casa do Psicólogo, 2005. p. 217–237.

CABANACH, R. G. et al. Una aproximación teórica al concepto de metas académicas y su relación con la motivación escolar. **Psicothema**, v. 8, n. 1, p. 45–61, 1996. ISSN 02149915.

CHEDE, C. **Big Data: expectativas, benefícios e barreiras** . 2013. Disponível em: <https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/big_data_expectativas_beneficios_e_barreiras>.

CIED. **Cursos online disponibilizados**. 2015. Disponível em: <http://www.ufal.edu.br/cied/cursos>.

CLAYTON, K.; BLUMBERG, F.; AULD, D. P. The relationship between motivation, learning strategies and choice of environment whether traditional or including an online component. **British Journal of Educational Technology**, Blackwell Publishing Ltd, v. 41, n. 3, p. 349–364, 2010. ISSN 1467-8535. Disponível em: <http://dx.doi.org/10.1111/j.1467-8535.2009.00993.x>.

COCEA, M.; WEIBELZAHL, S. Eliciting motivation knowledge from log files towards motivation diagnosis for adaptive systems. In: ____. **User Modeling 2007**. [S.l.]: Springer Berlin Heidelberg, 2007. p. 197–206. ISBN 978-3-540-73078-1.

CUNHA, N. d. B.; BORUCHOVITCH, E. Estratégias de Aprendizagem e Motivação para Aprender na Formação de Professores. **Revista Interamericana de Psicología/Interamerican Journal of Psychology**, v. 46, n. 2, p. 247–257, 2012.

DOSUALDO, D. G.; REZENDE, S. O. **Análise da Precisão de Métodos de Regressão**. [S.l.], 2003. Disponível em: <http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_197.pdf>.

ELLIOT, A. J.; CHURCH, M. A. A hierarchical model of approach and avoidance achievement motivation. **Journal of Personality and Social Psychology**, v. 72, n. 1, p. 218–232, 1997. ISSN 0022-3514.

ELLIOT, A. J.; HARACKIEWICZ, J. M. Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. **Journal of Personality and Social Psychology**, v. 70, n. 3, p. 461–475, 1996.

ELLIOT, A. J.; MCGREGOR, H. a.; GABLE, S. Achievement goals, study strategies, and exam performance: A mediational analysis. v. 91, n. 3, p. 549–563, 1999. ISSN 0022-0663.

ESPINHEIRA, P. L.; SILVA, L. C. M. da; SILVA, A. d. O. Prediction Measures in Beta Regression Models. n. January 2015, 2015. Disponível em: <http://arxiv.org/abs/1501.04830>.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

GOUVêA, G. **Educação a distância na formação de professores: viabilidades, potencialidades e limites**. [S.l.]: Vieira & Lent, 2006. ISBN 9788588782341.

GOYA, A.; BZUNECK, J. A.; GUIMARãES, S. d. R. Crenças de eficácia de professores e motivação de adolescentes para aprender física. **Psicologia Escolar e Educacional (Impresso)**, v. 12, n. 1, 2008. ISSN 1413-8557.

HAIR, J. F. et al. **Multivariate Data Analysis**. [S.l.]: Prentice Hall, 2009. 785 p. ISBN 978-0138132637.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790, 9780123814791.

HAYKIN, S. S. **Redes neurais**. [S.l.]: Bookman, 2001.

HERSHKOVITZ, A.; NACHMIAS, R. Developing a log-based motivation measuring tool. In: **Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings**. Knowledge Technology Lab, School of Education, Tel Aviv University, Israel: [s.n.], 2008. (1st International Conference on Educational Data Mining, EDM 2008), p. 226–233. ISBN 9780615306292 (ISBN). Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-77957280754{&}partnerID=40{&}md5=8a9af51707a2e1804fc4a5cef5>.

JACOBSEN, A. de L. et al. Autonomia do aluno na educação a distância: o caso do curso de administração a distância da ufsc. **Revista Gestão Universitária na América Latina-GUAL**, v. 4, n. 2, p. 53–73, 2011.

KELLER, J. M. Development and use of the arcs model of instructional design. **Journal of instructional development**, Springer, v. 10, n. 3, p. 2–10, 1987.

LEITE, L. S.; DIAS, R. A. **Educação a Distância: Da Legislação ao Pedagógico**. [S.l.]: Vozes, 2010. 70 p. ISBN 9788532627292.

LüFTENEGGER, M. et al. Lifelong learning as a goal – do autonomy and self-regulation in school result in well prepared pupils? **Learning and Instruction**, v. 22, n. 1, p. 27 – 36, 2012. ISSN 0959-4752. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0959475211000430>.

MAEHR, M. L.; MEYER, H. a. Reflections on the Field Understanding Motivation and Schooling : Where We ' ve Been , Where We Are , and Where We Need to Go. **Educational Psychology Review**, v. 9, n. 4, p. 371–409, 1997. ISSN 1573-336X.

MALTA, C. A. et al. A survey analysis on goal orientation changes in an information systems distance course: A brazilian case study. In: **Proceedings of the 30th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2015. (SAC '15), p. 227–232. ISBN 978-1-4503-3196-8. Disponível em: <http://doi.acm.org/10.1145/2695664.2695775>.

MATOS, P. F. et al. **Conceitos sobre Aprendizado de Máquina**. [S.l.], 2009. Disponível em: <http://www.icmc.usp.br/~taspardo/techreportufscar2009b-matosetal.pdf>.

MIDDLETON, M. J.; MIDGLEY, C. Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. **Journal of Educational Psychology**, American Psychological Association, v. 89, n. 4, p. 710–718, 1997. Disponível em: <http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1997-43826-012>.

MIDGLEY, C. et al. The development and validation of scales assessing students' achievement goal orientations. **Contemporary Educational Psychology**, v. 23, n. 2, p. 113–131, 1998. ISSN 0361-476X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0361476X98909651>.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de maquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, p. 89–114, 2003.

MOORE, M.; KEARSLEY, G. **Educação a distância: uma visão integrada**. [S.l.]: Cengage Learning, 2007. ISBN 9788522105762.

NEVES, Y. P. d. C. e. S. **Evasão nos cursos a distância: curso de extensão TV na Escola e os desafios de hoje**. Dissertação (Mestrado) — Universidade Federal de Alagoas, 2006. Disponível em: <http://www.repositorio.ufal.br/bitstream/riufal/315/1/YaraPereiradaCostaeSilvaNeves.pdf>.

NISBET, R.; ELDER, J.; MINER, G. **Handbook of Statistical Analysis and Data Mining Applications**. [S.l.]: Academic Press, 2009. ISBN 0123747651, 9780123747655.

NUGGETS, K. **Where did you apply Analytics/Data Mining in 2012?** 2012. Disponível em: <http://www.kdnuggets.com/polls/2012/where-applied-analytics-data-mining.html>.

OLSON, D. L.; DELEN, D. **Advanced Data Mining Techniques**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISBN 3540769161, 9783540769163.

PACHECO, A. S. V. **Evasão: análise da realidade do curso de graduação em administração a distância da Universidade Federal de Santa Catarina**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, 2007.

PINTO, I. M. B. S. **EVASÃO NOS CURSOS NA MODALIDADE DE EDUCAÇÃO A DISTÂNCIA: estudo de caso do Curso Piloto de Administração da UFAL/UAB**. Dissertação (Mestrado) — Universidade Federal de Alagoas - UFAL, 2010.

PINTRICH, P. R.; SCHUNK, D. H. **Motivation in Education: Theory, Research and Application**. [S.l.]: Prentice-Hall, 2001. 460 p. ISBN 0130160091.

REIS, F. L. d. Do ensino presencial ao ensino a distância no contexto universitário. **Revista Científica da FAI**, v. 9, p. 81–94, 2009. Disponível em: <http://www.fai-mg.br/portal/download/revista_cientifica_2009/pub_dw_artigo_ensino.pdf>.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, v. 33, n. 1, p. 135–146, 2007. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417406001266>.

ROMERO, C. et al. **Handbook of Educational Data Mining**. [S.l.]: CRC Press, 2010. (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series). ISBN 9781439804582.

ROMãO, E. **A Relação Educativa por Meio de Falas, Fios e Cartas**. [S.l.]: EDUFAL, 2008. 249 p. ISBN 9788571774483.

RYAN, R. M.; DECI, E. L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. **American psychologist**, American Psychological Association, v. 55, n. 1, p. 68, 2000.

SANTOS, A. A. A. dos; ALCARÁ, A. R.; ZENORINI, R. d. P. C. Estudos psicométricos da escala de motivação para a aprendizagem de universitários. **Fractal : Revista de Psicologia**, v. 25, n. 3, p. 531–546, 2013. ISSN 1984-0292. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1984-02922013000300008{&}lng=en{&}nrm>.

SANTOS, E.; WECHSLER, S. M. Ensino à Distância : Uma Década das Publicações Científicas Brasileiras. **Revista Interamericcana de Psicologia**, v. 43, n. 3, p. 558–565, 2009.

SMITHSON, M.; VERKUILEN, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. **Psychological Methods**, v. 11, n. 1, p. 54–71, mar. 2006. ISSN 1939-1463. Disponível em: <http://dx.doi.org/10.1037/1082-989x.11.1.54>.

STEINMAYR, R.; SPINATH, B. The importance of motivation as a predictor of school achievement. **Learning and Individual Differences**, v. 19, n. 1, p. 80 – 90, 2009. ISSN 1041-6080. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1041608008000502>.

VALLERAND, R. J. et al. The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. **Educational and psychological measurement**, Sage Publications, v. 52, n. 4, p. 1003–1017, 1992.

VANSLAMBROUCK, S. et al. Motivational profiles of adult learners in online and blended learning. In: ____. **Proceedings of the 14th European Conference on e-Learning University of Hertfordshire Hatfield, UK**. [S.l.]: University of Hertfordshire, 2015. p. 754–761. ISBN 978-1-910810-70-5.

VASCONCELOS, S. P. G. d. **Educação à Distância: Histórico e Perspectivas**. 2010. Disponível em: <http://www.filologia.org.br/viiifelin/19.htm>.

VIEIRA, M. de F. Desafios na gestão de ead no contexto dos polos de apoio presencial da universidade aberta do brasil. **Discutindo a visibilidade da EaD Pública no Brasil**, p. 74, 2015.

WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining: A Practical Guide**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. ISBN 1-55860-403-0.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, [International Biometric Society, Wiley], v. 1, n. 6, p. 80–83, 1945. ISSN 00994987. Disponível em: <http://www.jstor.org/stable/3001968>.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123748569, 9780123748560.

WU, X. et al. Top 10 algorithms in data mining. **Knowl. Inf. Syst.**, Springer-Verlag New York, Inc., New York, NY, USA, v. 14, n. 1, p. 1–37, dez. 2007. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/s10115-007-0114-2>.

ZENORINI, R. d. P. C.; SANTOS, A. A. A. dos. Teoria de metas de realização: fundamentos e avaliação. In: ____. **Motivação para Aprender - Aplicações no Contexto Educativo**. [S.l.]: Vozes, 2010. p. 99–125.

ZENORINI, R. d. P. C.; SANTOS, A. A. A. dos; MONTEIRO, R. d. M. Motivação para aprender: relação com o desempenho de estudantes. **Paidéia (Ribeirão Preto)**, v. 21, n. 49, p. 157–164, 2011. ISSN 0103-863X. Disponível em: <http://www.scielo.br/pdf/paideia/v21n49/03.pdf>.

ZHANG, G.; CHENG, Z. A WWW-based learner's learning motivation detecting system. **... Information Systems ...**, 2003. Disponível em: <http://www.aise.ics.saitama-u.ac.jp/KEST/KEST03W/papers/36-zhang.pdf>.

# Apêndices

# APÊNDICE A – QUESTIONÁRIO

Prezado Sr. (a),

Sou mestrando do Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da UFAL/IC, da linha de pesquisa de Modelagem Computacional em Educação.
Estou realizando uma pesquisa para minha dissertação, a qual possui como objetivo analisar os fatores que motivam e/ou desmotivam alunos no contexto da educação a distância.
Todas as informações aqui apresentadas serão sigilosas, não havendo identificação do respondente em momento algum da pesquisa.
Agradeço desde já a atenção dispensada e enfatizo a importância de sua resposta para o sucesso desta pesquisa.

**NOME: _____**
**UNIVERSIDADE: _____**
**CURSO: _____ PERÍODO: _____**
**CIDADE: _____ ESTADO: _____**
**JÁ POSSUI CURSO SUPERIOR: ( ) Sim ( )Não Se sim, qual: _____**
**IDADE: _____**

***Perfil sócio-econômico***

**Gênero**
( ) Masculino ( ) Feminino

**Nível de renda familiar mensal**
( ) 01 a 03 salários mínimos ( ) 03 a 05 salários mínimos
( ) 05 a 08 salários mínimos ( ) 08 a 10 salários mínimos
( ) mais de 10 salários mínimos

**Estado civil**
( ) Solteiro ( ) Casado ou vivendo com companheiro(a)
( ) Separado ou divorciado ( ) Viúvo

***Formação acadêmica***

**Em sua formação acadêmica você já cursou outro curso na modalidade à distância?**
( ) Sim ( ) Não

**Como você avalia os materiais ofertados no seu curso (livros-textos, video aulas, etc.)**
( ) Excelente ( ) Muito bom
( ) Bom ( ) Razoável
( ) Ruim ( ) Péssimo

**Além do material do curso você teve acesso ou utilizou outras fontes para melhorar e aprofundar seus conhecimentos?**
( ) Sim ( ) Não
Se afirmativo, informe quais.
_____

**Quanto tempo você disponibiliza para seus estudos?**
( ) Até 5h semanais ( ) Entre 6h e 10h semanais
( ) Entre 11h e 15h semanais ( ) Entre 16h e 20h semanais
( ) Mais de 20h semanais

**Você trabalha? Se sim qual a carga horária que você dedica ao seu trabalho?**
( ) Não trabalho ( ) Dedico até 20h semanais ao trabalho
( ) Dedico entre 21h e 30h semanais ( ) Dedico mais de 30h semanais

***Uso das TIC***

**Em qual (is) desse (s) local (is) você tem acesso a computador com internet? (Assinale 1 ou mais alternativas)**
( ) No trabalho ( ) Em casa
( ) Na lan house ( ) Na casa de amigos/parentes
( ) Não tenho acesso a computador com internet

**Qual o tipo de acesso a internet você tem?**
( ) Telefone/Discada ( ) Via cabo
( ) Via TV a cabo ( ) Via modem de operadora de telefonia
( ) ADSL

---

**INSTRUÇÕES**. As questões a seguir referem-se à sua ***motivação*** e às suas ***atitudes*** em relação à aprendizagem. **Não há respostas certas ou erradas, o importante é que você seja sincero!** Marque com um X a opção que mais se ajusta a você: Marque **(X) 1** se você concorda com a afirmação, **(X) 2** se você não tem opinião a respeito e **(X) 3** se você discorda da afirmação.

| | ***1= Concordo 2= Não sei 3= Discordo*** | ***1*** | ***2*** | ***3*** |
|---|---|---|---|---|
| 1 | Quando vou mal numa prova, estudo mais para a próxima. | | | |
| 2 | Eu não desisto facilmente diante de uma tarefa difícil. | | | |
| 3 | Para mim, é importante fazer as coisas melhor que os demais. | | | |
| 4 | É importante para mim, fazer as tarefas melhor que os meus colegas. | | | |
| 5 | Faço minhas atividades acadêmicas porque estou interessado nelas. | | | |
| 6 | Não respondo aos questionamentos feitos pelo professor ou tutor, por medo de falar alguma "besteira". | | | |
| 7 | Gosto de trabalhos acadêmicos com os quais aprendo algo, mesmo que cometa uma porção de erros. | | | |
| 8 | Na minha turma, eu quero me sair melhor que os demais. | | | |

| | | | | |
|---|---|---|---|---|
| 9 | Não participo dos debates nos fóruns e em sala de aula, porque não quero que os colegas riam de mim. | | | |
| 10 | Uma razão pela qual eu faço minhas tarefas acadêmicas é que eu gosto delas. | | | |
| 11 | Sinto-me bem sucedido na aula quando sei que o meu trabalho foi melhor que dos meus colegas. | | | |
| 12 | Uma razão importante pela qual faço as tarefas acadêmicas é porque eu gosto de aprender coisas novas. | | | |
| 13 | Gosto de mostrar aos meus colegas que sei as respostas. | | | |
| 14 | Quanto mais difícil a matéria, mais eu gosto de tentar compreender. | | | |
| 15 | Para mim, é importante, conseguir concluir tarefas que meus colegas não conseguem. | | | |
| 16 | Não me posiciono nas discussões nos fóruns e em sala de aula, pois não quero que os professores achem que sei menos que os meus colegas. | | | |
| 17 | Sucesso na universidade é fazer as coisas melhor que os outros. | | | |
| 18 | Não participo das aulas e fóruns de discussão quando tenho dúvidas no conteúdo que está sendo trabalhado. | | | |
| 19 | Eu gosto mais das tarefas quando elas me fazem pensar. | | | |
| | ***1= Concordo 2= Não sei 3= Discordo*** | ***1*** | ***2*** | ***3*** |
| 20 | Gosto de participar de trabalhos em grupo sempre que eu possa ser o líder. | | | |
| 21 | Gosto quando uma matéria me faz sentir vontade de aprender mais. | | | |
| 22 | Uma razão pela qual eu não participo das aulas e fóruns de discussão é para evitar parecer ignorante. | | | |
| 23 | Uma importante razão pela qual eu estudo pra valer é porque eu quero aumentar meus conhecimentos. | | | |
| 24 | Ser o primeiro da classe é o que me leva a estudar | | | |
| 25 | Gosto de tarefas difíceis e desafiadoras | | | |
| 26 | Não questiono o professor ou tutor quando tenho dúvidas na matéria, para não dar a impressão de que sou menos inteligente que os meus colegas. | | | |
| 27 | Não participo das aulas e fóruns de discussão para evitar que meus colegas e professores me achem pouco inteligente. | | | |
| 28 | Sou perseverante, mesmo quando uma tarefa me frustra. | | | |

# APÊNDICE B – TCLE

Eu, _____, tendo sido convidad(o,a) a participar como voluntário(o,a) da pesquisa "Um modelo computacional para classificação da motivação de estudantes em educação online", recebi do Sr. Cheops Araújo Malta, do Prof. Dr. Alan Pedro da Silva e do Prof. Dr. Ig Ibert Bittencourt Santana Pinto, responsáveis por sua execução, as seguintes informações que me fizeram entender sem dificuldades e sem dúvidas os seguintes aspectos:

§ Que o estudo se destina a identificar fatores que motivam os alunos da educação a online.

§ Que a importância deste estudo está na construção de um modelo computacional que permita a classificação da motivação dos alunos da educação online na perspectiva da Teoria das Metas de Realização.

§ Que eu participei das seguintes etapas: coleta de dados com aplicação de questionários e análise de registros de interação no ambiente virtual de aprendizagem.

§ Que os incômodos que poderei sentir com a minha participação são os seguintes: Sentir-me inibido(a) em função da presença dos pesquisadores durante a aplicação dos questionários e pelo fato de saber que as minhas respostas estarão sendo avaliadas.

§ Que os possíveis riscos à minha saúde física e mental são: o desconforto emocional e psicológico ao responder o questionário e tomar consciência de que sua atuação está em processo de análise.

§ Que os benefícios que deverei esperar com a minha participação, mesmo que não diretamente são: contribuir para a construção de um modelo computacional que permita a classificação da motivação dos alunos da educação online.

§ Que a minha participação não terá qualquer impacto na minha avaliação nas disciplinas do curso.

§ Que, sempre que desejar, serão fornecidos esclarecimentos sobre cada uma das etapas do estudo.

§ Que, a qualquer momento, eu poderei recusar a continuar participando do estudo, e também, eu poderei retirar esse meu consentimento, sem que isso me traga qualquer penalidade ou prejuízo.

§ Que as informações conseguidas através da minha participação não permitirão a identificação da minha pessoa, exceto aos responsáveis pelo estudo, e que a divulgação das mencionadas informações só será feita entre os profissionais estudiosos do assunto.

§ Que atuarei como voluntário e que também não terei nenhuma despesa em virtude da pesquisa.

§ Que eu serei indenizado por qualquer dano que venha a sofrer com a participação na pesquisa.

§ Que eu receberei uma via do Termo de Consentimento Livre e Esclarecido.

Finalmente, tendo eu compreendido perfeitamente tudo o que me foi informado sobre a minha participação no mencionado estudo e estando consciente dos meus direitos, das minhas responsabilidades, dos riscos e dos benefícios que a minha participação implicam, concordo em dele participar e para isso DOU O MEU CONSENTIMENTO SEM QUE PARA ISSO EU TENHA SIDO FORÇADO OU OBRIGADO.

**Endereço d(o,a) participante-voluntári(o,a)**

Domícilio: (rua, praça, conjunto):

Bloco: /Nº: /Complemento:

Bairro: /CEP/Cidade: /Telefone:

Ponto de referência:

**Contato de urgência:**

Sr. Cheops Araújo Malta
Rua Delmiro Gouveia, 532, Camoxinga, Santana do Ipanema - AL
Email: cheopsmalta@gmail.com
Telefone: 82 99450595

**ATENÇÃO: Para informar ocorrências irregulares ou danosas durante a participação no estudo, dirija-se ao:**

**Comitê de Ética em Pesquisa da Universidade Federal de Alagoas**

**Prédio da Reitoria, 1º Andar , Campus A. C. Simões, Cidade Universitária**

**Telefone: 3214-1041**

| | |
|---|---|
| Assinatura ou impressão datiloscópica d(o,a) voluntári(o,a) ou responsável legal e rubricar as demais folhas | Nome e Assinatura do(s) responsável(eis) pelo estudo (Rubricar as demais páginas) |

# APÊNDICE C – FERRAMENTAS DE MINERAÇÃO DE DADOS

Trataremos aqui sobre as ferramentas escolhidas para realizar as tarefas de mineração de dados.

## C.1 R PROJECT

O R[1] é uma linguagem e um ambiente de desenvolvimento integrado para computação estatística e gráficos. É um projeto GNU que foi desenvolvido no Bell Laboratories (antiga AT & T, agora Lucent Technologies) por John Chambers e seus colegas.

A ferramenta fornece uma ampla variedade de técnicas estatísticas (modelagens lineares e não lineares, testes estatísticos clássicos, análises de séries temporais, classificação, agrupamento...) e técnicas gráficas.

É amplamente extensível com o uso dos pacotes, que são bibliotecas para funções específicas ou áreas de estudo específicas. Um conjunto de pacotes é incluído com a instalação padrão do R, mas muito outros estão disponíveis na rede de distribuição do R (em inglês CRAN).

O código fonte do R está disponível sob a licença GPL[2] e as versões binárias pré-compiladas são fornecidas para Windows, Macintosh, e muitos sistemas operacionais Unix/Linux.

A linguagem é amplamente utilizada atualmente por estatísticos e analistas de dados para desenvolvimento de softwares estatísticos e processos de análise de dados.

## C.2 O WORKBENCH WEKA

Weka[3] acrônimo que significa *Waikato Environment for Knowledge Analysis* (Ambiente Waikato para Análise do Conhecimento), foi desenvolvido pela Universidade de Waikato, na Nova Zelândia, com o objetivo de prover a rápida experimentação de métodos de análise em novos conjuntos de dados, de forma rápida e flexível.

O sistema foi escrito em Java e é distribuído de acordo com os termos da licença GPL. O Weka funciona nos principais sistemas operacionais: Linux, Windows e Macintosh, e é

[1] Disponível para download em: https://cran.r-project.org/mirrors.html

[2] GPL, sigla em Inglês para: General PublicLicense (Licença Pública Geral), que é uma licença de uso de software que garante ao usuário final o direito de estudar, compartilhar e modificar um software sob essa licença (GNU.ORG).

[3] Disponível para download em: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

classificado como uma "bancada de trabalho"[4] contendo uma coleção de ferramentas de pré-processamento de dados e das mais atuais implementações de algoritmos de aprendizado de máquina. O ambiente oferece suporte completo ao processo de experimentação em mineração de dados, ou seja: suporte ao preparo dos dados de entrada (pré-processamento dos dados), avaliação estatística de esquemas de aprendizado, bem como métodos para o processamento dos resultados (pós-processamento) e integração com aplicações externas através de uma API Java (WITTEN; FRANK; HALL, 2011).

[4] Em inglês: workbench.

# APÊNDICE D – EXECUÇÃO DO SCRIPT R - SELEÇÃO INICIAL DE COVARIADAS - APRENDER

```
## Aqui vamos demostrar o processo utilizando o stepAIC via modelo de regressão linear normal
## para fornecer um modelo inicial a ser testado e melhorado com base no modelo de regressão
## beta.

> install.packages("betareg") # Instala o pacote betareg no R
> library(MASS)  # Carrega programas necessários para realiza o stepAIC
> library(betareg) #Carrega o betareg.
>
> dados1<-read.table("C:\\Users\\Cheops\\Moodle.txt", header=TRUE) ## Carregando os dados
> attach(dados1)
> dados1<-data.frame(dados1)

####### Abaixo estão descritas todas as variáveis originais contidas no dataset #######

> Variates<-cbind(id,periodo,diasinteracao,idade,genero,user_login,course_view,resource_view,
wiki_view,glossary_view,forum_view_discussion,forum_view_forum,forum_add_post,
forum_update_post,wiki_comments,message_write,glossary_add_entry,assign,blog,chat,choice,
course,data,forum,glossary,message,quiz,IDscorm,scorm,user,wiki,per_aprender,per_aprox,per_evit,
per_t,IDquiz,IDblog)

> dados2<-read.table("C:\\Users\\Cheops\\Moodle_AIC.txt", header=TRUE) ## Carregando os dados
> attach(dados2)
> dados2<-data.frame(dados2)

### Abaixo estão descritas as covariadas que irão entrar no processo de seleção stepAIC  ###

>Covariates_Model<-cbind(periodo,diasinteracao,idade,genero,user_login,course_view,
resource_view,wiki_view,glossary_view,forum_view_discussion,forum_view_forum,forum_add_post,
forum_update_post,wiki_comments,message_write,glossary_add_entry,assign,blog,chat,choice,
course,data,forum,glossary,message,quiz,IDscorm,scorm,user,wiki,IDquiz,IDblog)

######################################## APRENDER ########################################

> Aprender_logit<-log((per_aprender)/(1 - per_aprender)) ## Essa transformação da resposta é
## chamada de logito, a qual transforma dados do intervalo unitário para os reais. Para usar
## o stepAIC da distribuição normal é indicado fazer essa transformação.
>
> Aprender.lm <- lm(Aprender_logit ~ ., data = dados2) ## Regressão linear normal
## considerando todas as variaáveis de Covariates_Model.
> Aprender.lm2 <- stepAIC(Aprender.lm, trace = FALSE) ## O método stepAIC
>
> summary(Aprender.lm2) ## Apresenta o modelo final, os coeficientes estimados e a
## significância de cada covariada.

Call:
lm(formula = Aprender_logit ~ diasinteracao + idade + genero +
    user_login + forum_view_discussion + forum_view_forum + forum_add_post +
    forum_update_post + wiki_comments + glossary_add_entry +
    course + forum + glossary + message + user + wiki, data = dados2)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.49207 -0.12379  0.00051  0.13165  0.37802

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.569e-01  8.165e-02  -3.146 0.002020 **
diasinteracao          2.836e-04  1.667e-04   1.701 0.091055 .
idade                  3.875e-03  2.086e-03   1.858 0.065305 .
genero                 7.072e-02  3.642e-02   1.942 0.054148 .
user_login            -5.815e-04  3.754e-04  -1.549 0.123618
forum_view_discussion  5.779e-03  1.502e-03   3.847 0.000180 ***
forum_view_forum       6.167e-03  1.574e-03   3.918 0.000139 ***
forum_add_post         4.230e-03  2.498e-03   1.693 0.092599 .
forum_update_post      1.380e-02  5.894e-03   2.342 0.020575 *
wiki_comments          3.154e-02  1.836e-02   1.718 0.088028 .
glossary_add_entry    -3.711e-02  1.227e-02  -3.025 0.002952 **
course                -1.686e-04  7.078e-05  -2.382 0.018570 *
forum                 -5.722e-03  1.450e-03  -3.947 0.000125 ***
glossary               4.547e-03  1.342e-03   3.388 0.000913 ***
message               -5.846e-04  4.109e-04  -1.423 0.157027
user                   3.429e-04  1.967e-04   1.744 0.083384 .
wiki                  -1.223e-02  5.290e-03  -2.312 0.022250 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2043 on 141 degrees of freedom
Multiple R-squared:  0.3393,Adjusted R-squared:  0.2643
F-statistic: 4.525 on 16 and 141 DF,  p-value: 3.276e-07
>

############ Ajustando o modelo beta com as covariadas selecionadas pelo stepAIC #############

> fitAprender<-betareg(per_aprender~ diasinteracao + idade + genero + user_login +
forum_view_discussion + forum_view_forum + forum_add_post + forum_update_post +
wiki_comments + glossary_add_entry + course + forum + glossary + message + user + wiki)
>
> summary(fitAprender)

Call:
betareg(formula = per_aprender ~ diasinteracao + idade + genero + user_login +
    forum_view_discussion + forum_view_forum + forum_add_post + forum_update_post +
    wiki_comments + glossary_add_entry + course + forum + glossary + message +
    user + wiki)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.4736 -0.6817  0.0066  0.6995  2.1646

Coefficients (mean model with logit link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.558e-01  7.626e-02  -3.354 0.000795 ***
diasinteracao          2.862e-04  1.557e-04   1.838 0.066064 .
idade                  3.828e-03  1.948e-03   1.966 0.049332 *
genero                 6.929e-02  3.402e-02   2.037 0.041689 *
user_login            -5.838e-04  3.509e-04  -1.664 0.096182 .
```

```
forum_view_discussion  5.360e-03  1.465e-03   3.658 0.000254 ***
forum_view_forum       5.727e-03  1.533e-03   3.736 0.000187 ***
forum_add_post         3.881e-03  2.363e-03   1.642 0.100530
forum_update_post      1.334e-02  5.537e-03   2.409 0.016013 *
wiki_comments          3.144e-02  1.722e-02   1.826 0.067782 .
glossary_add_entry    -3.728e-02  1.155e-02  -3.226 0.001254 **
course                -1.654e-04  6.617e-05  -2.500 0.012414 *
forum                 -5.301e-03  1.418e-03  -3.738 0.000185 ***
glossary               4.550e-03  1.264e-03   3.598 0.000320 ***
message               -5.751e-04  3.841e-04  -1.497 0.134331
user                   3.416e-04  1.838e-04   1.859 0.063076 .
wiki                  -1.220e-02  4.964e-03  -2.458 0.013985 *

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   109.20      12.23   8.929   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 257.2 on 18 Df
Pseudo R-squared: 0.339
Number of iterations: 30 (BFGS) + 2 (Fisher scoring)
>
```

# APÊNDICE E – EXECUÇÃO DO SCRIPT R - SELEÇÃO INICIAL DE COVARIADAS - PERFORMANCE-APROXIMAÇÃO

```
## Aqui vamos assumir que os dados e bibliotecas já foram carregados, vide código anterior.
## Também não entraremos em detalhes nos comentários pelo fato de o código ser bastante
## semelhante ao anterior, apenas alterando a variável-meta,

################################# PERFORMANCE-APROXIMAÇÃO ################################
> Aproximacao_logit<-log((per_aprox)/(1 - per_aprox))
> Aproximacao.lm <- lm(Aproximacao_logit ~ ., data = dados2)
> Aproximacao.lm2 <- stepAIC(Aproximacao.lm, trace = FALSE)
> summary(Aproximacao.lm2)

Call:
lm(formula = Aproximacao_logit ~ idade + genero + course_view +
    glossary_view + forum_add_post + forum_update_post + wiki_comments +
    glossary_add_entry + assign + blog + glossary + message +
    IDscorm + user + wiki + IDblog, data = dados2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65400 -0.16150  0.01751  0.16565  0.67120

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -8.707e-01  1.056e-01  -8.249 1.02e-13 ***
idade              -4.232e-03  2.719e-03  -1.556  0.12188
genero             -1.404e-01  5.150e-02  -2.726  0.00722 **
course_view         2.315e-04  1.022e-04   2.265  0.02506 *
glossary_view       9.979e-02  3.980e-02   2.507  0.01330 *
forum_add_post     -3.060e-03  2.169e-03  -1.410  0.16063
forum_update_post  -1.418e-02  7.239e-03  -1.959  0.05203 .
wiki_comments      -4.608e-02  2.286e-02  -2.016  0.04574 *
glossary_add_entry  1.483e-01  4.634e-02   3.201  0.00169 **
assign              3.933e-04  2.511e-04   1.567  0.11940
blog                1.629e-03  7.735e-04   2.106  0.03701 *
glossary           -1.044e-01  3.942e-02  -2.649  0.00900 **
message             1.164e-03  5.448e-04   2.137  0.03434 *
IDscorm             1.932e-01  7.462e-02   2.589  0.01064 *
user               -1.147e-04  6.726e-05  -1.705  0.09038 .
wiki                1.526e-02  6.833e-03   2.233  0.02712 *
IDblog             -7.767e-02  5.512e-02  -1.409  0.16096
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2682 on 141 degrees of freedom
Multiple R-squared:  0.334,Adjusted R-squared:  0.2585
F-statistic:  4.42 on 16 and 141 DF,  p-value: 5.159e-07

> fitAproximacao<-betareg(per_aprox~idade + genero + course_view + glossary_view +
```

```
forum_add_post + forum_update_post + wiki_comments + glossary_add_entry + assign + blog +
glossary + message + IDscorm + user + wiki + IDblog)
> summary(fitAproximacao)
Call:
betareg(formula = per_aprox ~ idade + genero + course_view + glossary_view +
    forum_add_post + forum_update_post + wiki_comments + glossary_add_entry +
    assign + blog + glossary + message + IDscorm + user + wiki + IDblog)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-2.7197 -0.7129  0.0961  0.7097  2.9244

Coefficients (mean model with logit link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -8.709e-01  9.829e-02  -8.860  < 2e-16 ***
idade              -3.711e-03  2.539e-03  -1.462 0.143777
genero             -1.317e-01  4.802e-02  -2.741 0.006117 **
course_view         2.294e-04  9.449e-05   2.428 0.015174 *
glossary_view       9.922e-02  3.860e-02   2.571 0.010144 *
forum_add_post     -3.173e-03  2.026e-03  -1.566 0.117400
forum_update_post  -1.361e-02  6.736e-03  -2.021 0.043262 *
wiki_comments      -4.647e-02  2.145e-02  -2.166 0.030292 *
glossary_add_entry  1.478e-01  4.433e-02   3.335 0.000853 ***
assign              3.993e-04  2.325e-04   1.718 0.085847 .
blog                1.679e-03  6.936e-04   2.421 0.015488 *
glossary           -1.038e-01  3.820e-02  -2.717 0.006585 **
message             1.194e-03  4.942e-04   2.416 0.015709 *
IDscorm             1.937e-01  6.906e-02   2.806 0.005022 **
user               -1.246e-04  6.201e-05  -2.009 0.044497 *
wiki                1.523e-02  6.258e-03   2.433 0.014974 *
IDblog             -7.525e-02  5.139e-02  -1.464 0.143145

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   79.346      8.882   8.934   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 250.8 on 18 Df
Pseudo R-squared: 0.3336
Number of iterations: 29 (BFGS) + 2 (Fisher scoring)
```

# APÊNDICE F – EXECUÇÃO DO SCRIPT R - SELEÇÃO INICIAL DE COVARIADAS - PERFORMANCE-EVITAÇÃO

```
## Aqui vamos assumir que os dados e bibliotecas já foram carregados, vide código de Aprender.

################################## PERFORMANCE-EVITAÇÃO ##################################
> Evitacao_logit<-log((per_evit)/(1 - per_evit))
> Evitacao.lm <- lm(Evitacao_logit ~ ., data = dados2)
> Evitacao.lm2 <- stepAIC(Evitacao.lm, trace = FALSE)
> summary(Evitacao.lm2)

Call:
lm(formula = Evitacao_logit ~ genero + resource_view + forum_view_discussion +
    forum_view_forum + assign + blog + forum + message + scorm,
    data = dados2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45150 -0.20413 -0.05753  0.17390  0.86794

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.1943201  0.0485393 -24.605  < 2e-16 ***
genero                 0.0887410  0.0515816   1.720   0.0874 .
resource_view          0.0008680  0.0005715   1.519   0.1309
forum_view_discussion -0.0072736  0.0016959  -4.289 3.22e-05 ***
forum_view_forum      -0.0072120  0.0017688  -4.077 7.41e-05 ***
assign                -0.0006481  0.0002733  -2.371   0.0190 *
blog                  -0.0012313  0.0006221  -1.979   0.0496 *
forum                  0.0068256  0.0015737   4.337 2.66e-05 ***
message               -0.0008421  0.0005273  -1.597   0.1124
scorm                 -0.0014018  0.0006985  -2.007   0.0466 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2802 on 148 degrees of freedom
Multiple R-squared:  0.1749,Adjusted R-squared:  0.1247
F-statistic: 3.485 on 9 and 148 DF,  p-value: 0.0006226

> fitEvitar<-betareg(per_evit~genero + resource_view + forum_view_discussion +
forum_view_forum + assign + blog + forum + message + scorm)
> summary(fitEvitar)

Call:
betareg(formula = per_evit ~ genero + resource_view + forum_view_discussion +
    forum_view_forum + assign + blog + forum + message + scorm)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-1.6820 -0.7392 -0.2117  0.6440  3.1989
```

```
Coefficients (mean model with logit link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.1754021  0.0471811 -24.913  < 2e-16 ***
genero                 0.0869389  0.0498612   1.744   0.0812 .
resource_view          0.0008928  0.0005564   1.605   0.1086
forum_view_discussion -0.0072104  0.0015603  -4.621 3.82e-06 ***
forum_view_forum      -0.0071751  0.0016177  -4.435 9.19e-06 ***
assign                -0.0006481  0.0002640  -2.455   0.0141 *
blog                  -0.0011942  0.0006269  -1.905   0.0568 .
forum                  0.0067829  0.0014410   4.707 2.51e-06 ***
message               -0.0008563  0.0005251  -1.631   0.1029
scorm                 -0.0013743  0.0006932  -1.982   0.0474 *

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   74.666      8.361    8.93   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 255.7 on 11 Df
Pseudo R-squared: 0.1747
Number of iterations: 22 (BFGS) + 2 (Fisher scoring)
```

# APÊNDICE G – EXECUÇÃO DO SCRIPT R - HISTOGRAMA VARIÁVEIS ALEATÓRIAS

```
> install.packages("betareg") # Instala o pacote betareg no R
> library(betareg) #Carrega o betareg.
> dados1<-read.table("C:\\Users\\Cheops\\Dados.txt", header=TRUE) ## Carregando os dados
> attach(dados1)
> dados1<-data.frame(dados1)
> par(mfrow=c(1,3))
> par(pty="s")
> ajuste=betareg(per_aprender~1) # Para criar a curva estimada da beta
> phi=ajuste$coefficients$precision # parâmetro phi gerado pelo betareg
> p=mean(ajuste$fitted)*phi  ## mean(ajuste$fitted) é a média estimada pelo betareg
## "p" parâmetro original da distribuição beta no R
> q=phi-mean(ajuste$fitted)*phi ## "q", parâmetro original da distribuição beta no R
## p=mu*phi
## p+q=phi
> a=mean(per_aprender) # média da variável para estimar a densidade normal
> b=sd(per_aprender)   # desvio padrão da variável para estima a densidade normal
> hist(per_aprender, xlab="", ylab="Densidade", main="Aprender", probability = TRUE,
cex.lab=1.4,cex.main=1.6,cex.axis=1.4, xlim=c(0.2,0.8),ylim=c(0,9))
> x<-per_aprender
> curve(dnorm(x, mean = a, sd = b, log = FALSE), col = 2, lty = 1, lwd = 1, add = TRUE, xlim=c(0.2,0.8))
> curve(dbeta(x, p, q, ncp = 0, log = FALSE), col = 4, lty = 2, lwd = 2, add = TRUE, xlim=c(0.2,0.8))
> ajuste= betareg(per_aprox~1)
> phi=ajuste$coefficients$precision
> p=mean(ajuste$fitted)*phi
> q=phi-mean(ajuste$fitted)*phi
> a=mean(per_aprox)
> b=sd(per_aprox)
> hist(per_aprox, xlab="", ylab="Densidade", main="Aproximação", probability = TRUE,
cex.lab=1.4,cex.main=1.6,xlim=c(0.05,0.6),cex.axis=1.4, ylim=c(0,9))
> x<-per_aprox
> curve(dnorm(x, mean = a, sd = b, log = FALSE), col = 2, lty = 1, lwd = 1, add = TRUE, xlim=c(0.05,0.6))
> curve(dbeta(x, p, q, ncp = 0, log = FALSE), col = 4, lty = 2, lwd = 2, add = TRUE, xlim=c(0.05,0.6))
> legend(0.10,8.7, c("Curva estimada normal","Curva estimada beta"),bty = "n",col=c(2,4),lty=c(1,2),cex=1.4)
> ajuste= betareg(per_evit~1)
> phi=ajuste$coefficients$precision
> p=mean(ajuste$fitted)*phi
> q=phi-mean(ajuste$fitted)*phi
> a=mean(per_evit)
> b=sd(per_evit)
> hist(per_evit, xlab="", ylab="Densidade", main="Evitação", probability = TRUE,
cex.lab=1.4,cex.main=1.6,xlim=c(0,0.5),cex.axis=1.4, ylim=c(0,9))
> x<-per_evit
> curve(dnorm(x, mean = a, sd = b, log = FALSE), col = 2, lty = 1, lwd = 1, add = TRUE, xlim=c(0,0.5))
> curve(dbeta(x, p, q, ncp = 0, log = FALSE), col = 4, lty = 2, lwd = 2, add = TRUE, xlim=c(0,0.5))
> dev.off()
```

# APÊNDICE H – EXECUÇÃO DO SCRIPT R - MODELOS FINAIS

## Modelo final para Aprender

```
> fitAprender<-betareg(per_aprender~idade:genero + log(forum_update_post + 1.0):genero
+ log(course_view):idade|exp(per_evit):IDquiz+log(per_evit) + forum_update_post:genero
+ log(forum_update_post+1.0):scorm)
> summary(fitAprender)

Call:
betareg(formula = per_aprender ~ idade:genero + log(forum_update_post +
1):genero + log(course_view):idade | exp(per_evit):IDquiz + log(per_evit) +
forum_update_post:genero + log(forum_update_post + 1):scorm)

Standardized weighted residuals 2:
Min      1Q  Median      3Q     Max
-3.0019 -0.8055 -0.0056  0.6245  2.3387

Coefficients (mean model with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -0.1443445  0.0417359  -3.459 0.000543 ***
idade:genero                       0.0029239  0.0011263   2.596 0.009429 **
genero:log(forum_update_post + 1)  0.0618258  0.0094861   6.517 7.15e-11 ***
idade:log(course_view)             0.0004878  0.0002489   1.960 0.049987 *

Phi coefficients (precision model with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.720147   0.780245   2.205 0.027481 *
log(per_evit)                    -1.975947   0.510449  -3.871 0.000108 ***
exp(per_evit):IDquiz             -0.380347   0.197701  -1.924 0.054373 .
forum_update_post:genero          0.117064   0.030211   3.875 0.000107 ***
log(forum_update_post + 1):scorm -0.005126   0.002608  -1.965 0.049357 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 249.3 on 9 Df
Pseudo R-squared: 0.08111
Number of iterations: 31 (BFGS) + 1 (Fisher scoring)
>
```

## Modelo final para Performance-Aproximação

```
> fitAproximacao<-betareg(per_aprox~idade:genero+log(scorm + 1.0):IDscorm
+ log(course_view):genero + blog:IDblog|scorm:idade
+ log(course_view):genero + scorm:IDscorm + genero:idade)
> summary(fitAproximacao)

Call:
```

```
betareg(formula = per_aprox ~ idade:genero + log(scorm + 1):IDscorm + log(course_view):genero +
blog:IDblog | scorm:idade + log(course_view):genero + scorm:IDscorm + genero:idade)

Standardized weighted residuals 2:
Min      1Q  Median      3Q     Max
-2.1293 -0.7654  0.0574  0.7832  2.5895

Coefficients (mean model with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.9692220  0.0398433 -24.326  < 2e-16 ***
idade:genero            -0.0098561  0.0022882  -4.307 1.65e-05 ***
log(scorm + 1):IDscorm   0.0164387  0.0074584   2.204   0.0275 *
genero:log(course_view)  0.0314078  0.0161569   1.944   0.0519 .
blog:IDblog              0.0013571  0.0003223   4.211 2.54e-05 ***

Phi coefficients (precision model with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.9905086  0.1882116  21.202  < 2e-16 ***
scorm:idade              0.0011997  0.0002212   5.424 5.83e-08 ***
log(course_view):genero -0.1885166  0.0925673  -2.037 0.041697 *
scorm:IDscorm           -0.0281645  0.0073452  -3.834 0.000126 ***
idade:genero             0.0481706  0.0181243   2.658 0.007866 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 239.8 on 10 Df
Pseudo R-squared: 0.1592
Number of iterations: 24 (BFGS) + 5 (Fisher scoring)
>
```

## Modelo final para Performance-Evitação

```
> IDblogIDquiz<-IDblog+IDquiz
> fitEvitacao<-betareg(per_evit~quiz:IDquiz + log(message_write +1.0) + blog:IDblog
+ log(forum_add_post+1.0):scorm + log(forum_view_forum):scorm|forum_add_post:scorm
+ quiz:IDblogIDquiz + message + log(per_aprox) + forum_view_forum:genero)
> summary(fitEvitacao)

Call:
betareg(formula = per_evit ~ quiz:IDquiz + log(message_write + 1) + blog:IDblog +
log(forum_add_post + 1):scorm + log(forum_view_forum):scorm | forum_add_post:scorm +
quiz:IDblogIDquiz + message + log(per_aprox) + forum_view_forum:genero)

Standardized weighted residuals 2:
Min      1Q  Median      3Q     Max
-2.1341 -0.6090  0.0024  0.8174  2.4154

Coefficients (mean model with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.2048027  0.0441444 -27.292  < 2e-16 ***
log(message_write + 1)       -0.0169066  0.0088751  -1.905  0.05679 .
quiz:IDquiz                   0.0017588  0.0006551   2.685  0.00726 **
blog:IDblog                  -0.0007141  0.0004004  -1.783  0.07454 .
log(forum_add_post + 1):scorm -0.0014897  0.0007086  -2.102  0.03551 *
```

```
scorm:log(forum_view_forum)    0.0006929  0.0004083   1.697  0.08967 .

Phi coefficients (precision model with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept)              5.3477818  0.7105849   7.526 5.24e-14 ***
message                  0.0078429  0.0025086   3.126  0.00177 **
log(per_aprox)           1.0695581  0.5244892   2.039  0.04143 *
forum_add_post:scorm     0.0005531  0.0002732   2.025  0.04289 *
quiz:IDblogIDquiz       -0.0043528  0.0017851  -2.438  0.01475 *
forum_view_forum:genero  0.0018983  0.0008356   2.272  0.02310 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 261.2 on 12 Df
Pseudo R-squared: 0.07306
Number of iterations: 25 (BFGS) + 4 (Fisher scoring)
>
```

# Anexos

# ANEXO A – AUTORIZAÇÃO INSTITUCIONAL

**Autorização Institucional**

Eu, Fernando Silvio Cavalcante Pimentel, responsável pela Coordenadoria Institucional de Educação a Distância da Universidade Federal de Alagoas, declaro que fui informado dos objetivos da pesquisa "Um modelo computacional para classificação da motivação de estudantes em educação online", a ser executada pelo Sr. Cheops Araújo Malta, Prof. Dr. Alan Pedro da Silva e Prof. Dr. Ig Ibert Bittencourt Santana Pinto e concordo em autorizar a execução da mesma nesta instituição. Caso necessário, a qualquer momento como instituição CO-PARTICIPNATE desta pesquisa poderemos revogar esta autorização, se comprovada atividades que causem algum prejuízo à esta instituição ou ainda, a qualquer dado que comprometa o sigilo da participação dos integrantes desta instituição. Declaro também, que não recebemos qualquer pagamento por esta autorização bem como os participantes também não receberão qualquer tipo de pagamento.

| | |
|---|---|
| Cheops Araujo Malta<br>Pesquisador | Fernando Silvio Cavalcante Pimentel<br>Responsável pela Instituição<br>Prof. Fernando Silvio C. Pimentel<br>Coordenador Adjunto<br>UAB/UFAL/CIED |
| Alan Pedro da Silva<br>Orientador | Ig Ibert Bittencourt Santana Pinto<br>Orientador |

# ANEXO B – FOLHA DE ROSTO DA SUBMISSÃO DO PROJETO AO CEP

Plataforma Brasil — MINISTÉRIO DA SAÚDE - Conselho Nacional de Saúde - Comissão Nacional de Ética em Pesquisa – CONEP

**FOLHA DE ROSTO PARA PESQUISA ENVOLVENDO SERES HUMANOS**

| | | | |
|---|---|---|---|
| 1. Projeto de Pesquisa:<br>Um modelo computacional para classificação da motivação de estudantes em educação online | | | 2. Número de Participantes da Pesquisa:<br>500 |
| 3. Área Temática: | | | |
| 4. Área do Conhecimento:<br>Grande Área 1. Ciências Exatas e da Terra | | | |
| **PESQUISADOR RESPONSÁVEL** | | | |
| 5. Nome:<br>CHEOPS ARAUJO MALTA | | | |
| 6. CPF:<br>054.072.554-48 | 7. Endereço (Rua, n.º):<br>RUA DELMIRO GOUVEIA, 532 CAMOXINGA SANTANA DO IPANEMA ALAGOAS 57500000 | | |
| 8. Nacionalidade:<br>BRASILEIRO | 9. Telefone:<br>(82) 9945-0595 | 10. Outro Telefone: | 11. Email:<br>cheopsmalta@gmail.com |
| 12. Cargo: | | | |

Termo de Compromisso: Declaro que conheço e cumprirei os requisitos da Resolução CNS 466/12 e suas complementares. Comprometo-me a utilizar os materiais e dados coletados exclusivamente para os fins previstos no protocolo e a publicar os resultados sejam eles favoráveis ou não. Aceito as responsabilidades pela condução científica do projeto acima. Tenho ciência que essa folha será anexada ao projeto devidamente assinada por todos os responsáveis e fará parte integrante da documentação do mesmo.

Data: 30 / 10 / 14

Cheops Araujo Malta

Assinatura

| | | |
|---|---|---|
| **INSTITUIÇÃO PROPONENTE** | | |
| 13. Nome:<br>Universidade Federal de Alagoas | 14. CNPJ:<br>24.464.109/0001-48 | 15. Unidade/Órgão:<br>PPGMCC / IC |
| 16. Telefone:<br>(82) 3214-1051 | 17. Outro Telefone: | |

Termo de Compromisso (do responsável pela instituição ): Declaro que conheço e cumprirei os requisitos da Resolução CNS 466/12 e suas Complementares e como esta instituição tem condições para o desenvolvimento deste projeto, autorizo sua execução.

Responsável: AYDANO PAMPONET MACHADO CPF: 031.776.634-16

Cargo/Função: COORDENADOR DO PPGMCC / IC / UFAL

Data: 17 / 10 / 14

Assinatura

Prof. Dr. Aydano Pamponet Machado
Coordenador do PPGMCC
Mat. SIAPE 2351211
IC/UFAL

**PATROCINADOR PRINCIPAL**

Não se aplica.

Alan Pedro da Silva
ORIENTADOR
SIAPE: 2494583

# ANEXO C – PARECER CONSUBSTANCIADO DO CEP

UNIVERSIDADE FEDERAL DE ALAGOAS

## PARECER CONSUBSTANCIADO DO CEP

**DADOS DO PROJETO DE PESQUISA**

**Título da Pesquisa:** Um modelo computacional para classificação da motivação de estudantes em educação online

**Pesquisador:** CHEOPS ARAUJO MALTA

**Área Temática:**

**Versão:** 2

**CAAE:** 38012314.3.0000.5013

**Instituição Proponente:** Universidade Federal de Alagoas

**Patrocinador Principal:** Financiamento Próprio

**DADOS DO PARECER**

**Número do Parecer:** 922.559

**Data da Relatoria:** 17/12/2014

**Apresentação do Projeto:**

A Educação a Distância atrelada às ferramentas tecnológicas, possibilita cada vez mais acessibilidade ao ensino de qualidade. Esta modalidade oferece diversas ferramentas para que tanto alunos quanto professores possam desenvolver suas atividades. Na EAD o aluno deixa de ser um sujeito passivo, aquele que simplesmente escuta e aplica o que o professor apresenta e passa a ser um sujeito ativo dentro do processo de aprendizagem, ele torna-se o centro do processo e o principal responsável pela busca do conhecimento definindo seu ritmo de estudos e etc. Apesar dos grandes avanços no que se refere à inserção desta modalidade de ensino no âmbito das universidades e no Brasil como um todo, muitas ainda são as dificuldades encontradas, uma destas está relacionada com a motivação dos alunos para aprender. Este estudo busca, por sua vez, identificar os padrões motivacionais dos alunos da educação a distância da UFAL sob a perspectiva da Teoria de Metas de Realização, permitindo assim a proposição de um modelo computacional que permita a inferência da classificação motivacional dos alunos da EAD de forma automática.

A natureza deste estudo se caracteriza por uma pesquisa do tipo descritiva exploratória.

O projeto em questão será realizado dentro do modelo hipotético-dedutivo de caráter transversal. Toma o modelo survey para a coleta de dados quantitativos e para análise dados serão utilizados

**Endereço:** Campus A . C Simões Cidade Universitária
**Bairro:** Tabuleiro dos Martins **CEP:** 57.072-900
**UF:** AL **Município:** MACEIO
**Telefone:** (82)3214-1041 **Fax:** (82)3214-1700 **E-mail:** comitedeeticaufal@gmail.com

UNIVERSIDADE FEDERAL DE ALAGOAS




instrumentos estatísticos. Também serão usados os dados das interações dos sujeitos no ambiente virtual de aprendizagem para correlação com fatores identificados no survey. Farão parte deste estudo como respondentes alunos da educação a distância da UFAL, sendo estes de ambos os sexos e dos diversos cursos de graduação ofertados, que comporão uma amostra não-probabilística.

**Objetivo da Pesquisa:**
Objetivo Primário:
Identificar os fatores que motivam os alunos na educação online, de modo que a motivação dos mesmos possa ser classificada utilizando como arcabouço a Teoria de Metas de Realização.

Objetivo Secundário:
Avaliar os padrões motivacionais dos alunos da educação a distância da UFAL e Propor um modelo computacional que permita a inferência da classificação motivacional dos alunos de forma automática;

**Avaliação dos Riscos e Benefícios:**
Riscos:
Os riscos advindos desta pesquisa estão associados ao desconforto emocional e psicológico resultantes do conhecimento que terá o sujeito quando ao responder o questionário e tomar consciência de que sua atuação está em processo de análise. Não há agravamentos àqueles.
Benefícios:
Os benefícios se fundamentam na contribuição para a construção de um modelo computacional que permita a classificação da motivação dos alunos
da educação online.

**Comentários e Considerações sobre a Pesquisa:**
Pesquisa relevante para o campo da Educação a Distância.
A lista de pendências:
NO TCLE: revisar cláusula "inibição dentro do olhar do pesquisador", pois não se entende o que é dito;deixar claro como ser dará a análise do ambiente virtual de aprendizagem.
Os alunos são pessoas em situação de vulnerabilidade, sendo assim, é necessário esclarecer no TCLE que a pesquisa não terá qualquer impacto na avaliação das disciplinas.
Verificar cronograma de pesquisa, tendo em vista que o mesmo pode ficar desatualizado, tendo

**Endereço:** Campus A . C Simões Cidade Universitária
**Bairro:** Tabuleiro dos Martins **CEP:** 57.072-900
**UF:** AL **Município:** MACEIO
**Telefone:** (82)3214-1041 **Fax:** (82)3214-1700 **E-mail:** comitedeeticaufal@gmail.com

UNIVERSIDADE FEDERAL DE ALAGOAS




em vista a pendência foi sanada.

**Considerações sobre os Termos de apresentação obrigatória:**

Foram analisados os documentos: Folha de Rosto folha_de_rosto; TCLE - Modelo de Termo de Consentimento Livre e Esclarecido; Declarações Diversas; Vínculo Instituições Participantes vinculo; Projeto Detalhado.

**Recomendações:**

**Conclusões ou Pendências e Lista de Inadequações:**

Protocolo atende as recomendações éticas da resolução 466/12.

**Situação do Parecer:**

Aprovado

**Necessita Apreciação da CONEP:**

Não

**Considerações Finais a critério do CEP:**

MACEIO, 18 de Dezembro de 2014

**Assinado por:**
**Deise Juliana Francisco**
**(Coordenador)**

**Endereço:** Campus A . C Simões Cidade Universitária
**Bairro:** Tabuleiro dos Martins **CEP:** 57.072-900
**UF:** AL **Município:** MACEIO
**Telefone:** (82)3214-1041 **Fax:** (82)3214-1700 **E-mail:** comitedeeticaufal@gmail.com