UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

ESRON DTAMAR DA SILVA

AVALIAÇÃO DE MODELOS PREDITIVOS DE EVASÃO NA EAD DA UNIVASF, A PARTIR DA TEORIA DA DISTÂNCIA TRANSACIONAL

JUAZEIRO - BA
2019

UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

ESRON DTAMAR DA SILVA

AVALIAÇÃO DE MODELOS PREDITIVOS DE EVASÃO NA EAD DA UNIVASF, A PARTIR DA TEORIA DA DISTÂNCIA TRANSACIONAL

Trabalho apresentado à Universidade Federal do Vale do São Francisco - Univasf, *Campus* Juazeiro, como requisito da obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Jorge Luis Cavalcanti Ramos

JUAZEIRO - BA
2019

RESUMO

Os avanços do acesso às tecnologias da informação criou um ambiente fértil para a pesquisa na área de educação a distância (EAD). Porém, apesar da grande disponibilidade e flexibilidade, os cursos da modalidade EAD, no Brasil, ainda sofrem com o problema da evasão de estudantes. Acompanhando o crescimento da EAD, se desenvolve também a área de Mineração de Dados Educacionais (MDE). Este trabalho propõe a utilização de uma metodologia fundamentada em técnicas de MDE com o objetivo de construir e avaliar modelos de classificação da situação final de estudantes da EAD entre duas classes, evadidos e não evadidos. Na construção dos modelos, serão utilizados dados obtidos no contexto da Universidade Federal do Vale do São Francisco (UNIVASF), e, alem disso, as variáveis preditoras serão concebidas a partir dos construtos da Teoria da Distância Transacional (TDT).

Palavras-chave: Educação a distância. Mineração de dados educacionais. Evasão. Aprendizagem de máquina

ABSTRACT

The advances in information technologies created a very fertile research environment in the distance education field. However, despite of the great disponibility and flexibility, the brazilian DE course genre still suffers from the students evasion problem, as shown by the EAD.BR census in the last few years. Along with the growing in DE, the Educational Data Mining is developing as well. Inspired by the Transactional Distance Theory, this works proposes the using of knowledge discovery in databases to construct and compare classification models on the final situation of a DE student between two classes, evasor and not-evasor. This work applies the methodology proposed by Ramos et al (2016), adapted to the EAD context in UNIVASF, where the variables utilized in the predictive models are constructed based on the TDT attributes, comparing the results found with a new educational scenario.

Key-words: Distance education. Educational data mining. School evasion.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABED	Associação Brasileira de Educação a Distância
AVA	Ambiente Virtual de Aprendizagem
CFA	Análise Fatorial Confirmatória
DM	*Data Mining*
DT	Distância Transacional
EAD	Educação a Distância
EDM	*Educational Data Mining*
IES	Instituição de ensino superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	*Knowledge Discovery in Databases*
KNN	*K-nearest Neighbors*
LA	*Learning Analytics*
LMS	*Learning Management System*
ML	*Machine Learning*
RL	Regressão Logistica
SVM	*Support Vector Machine*
TCC II	Trabalho de Conclusão de Curso II
TDT	Teoria da Distância Transacional
UNIVASF	Universidade Federal do Vale do São Francisco

SUMÁRIO

1 INTRODUÇÃO

O desafio de levar educação e formação profissional a lugares remotos, onde, dificilmente, a formação presencial tradicional conseguiria alcançar de maneira efetiva, é uma das principais bandeiras da Educação a Distância (EAD). Entretanto, outros desafios surgem em decorrência da expansão da modalidade: como garantir a qualidade dessa formação? como transpor modelos de educação presencial para a distância? como atuar de maneira a prevenir e reduzir os altos índices de evasão, ainda, verificados na modalidade? São questões como essas que devem ser respondidas a partir do desenvolvimento de pesquisas nessa modalidade. O uso de novas tecnologias e de novos processos pode contribuir com essas pesquisas.

Diversas iniciativas reforçam o crescimento da EAD e exigem uma atenção maior nos aspectos importantes para a consolidação e manutenção das atividades dessa modalidade nas instituições. Dentre esses aspectos, está a necessidade de pesquisas voltadas para a modalidade, como forma de se agregar procedimentos validados cientificamente, ferramentas de gestão mais eficientes e metodologias inovadoras, capazes de superar grandes desafios impostos pela EAD.

O avanço da modalidade de EAD requer o desenvolvimento de recursos que permitam o acompanhamento de cursos oferecidos em um ambiente virtual de aprendizagem. Esses recursos podem ser obtidos a partir de metodologias de análise que envolvem o conhecimento das estratégias pedagógicas dos cursos EAD, o levantamento das necessidades apontadas pelos profissionais que atuam na área e na elicitação dos requisitos para implementação de ferramentas de visualização de dados de diversas atividades dentro de um contexto educacional. (RAMOS, 2016)

Com a expansão do EAD de maneira responsável e planejada, com infraestrutura compatível e recursos humanos qualificados, será possível a oferta de novos cursos pelas instituições, disseminando conhecimento e possibilitando mais oportunidades para o desenvolvimento regional.

Para RAMOS (2016), os altos índices de evasão dos alunos em cursos EAD representam um grande desafio para todos os que atuam na modalidade. Além desses índices estarem em níveis elevados, observou-se também que estão em crescimento. Com isso há uma necessidade contínua de desenvolvimento de pesquisas que apontem caminhos, métodos e ferramentas que os auxiliem a enfrentar melhor esse problema. O uso de técnicas estatísticas e de mineração de dados, em conjunto com teorias consolidadas na modalidade, pode fundamentar modelos eficientes de detecção precoce do risco de evasão pelos alunos.

No estudo apresentado por RAMOS (2016), foram desenvolvidos, testados e vali-

dados, modelos preditivos da evasão de estudantes de graduação em cursos ofertados na modalidade EAD, tomando como base as variáveis que compõem cada um dos construtos da Teoria da Distância Transacional (MOORE, 2008). Essa pesquisa ocorreu a partir dos dados de cursos de licenciatura em Biologia e Pedagogia, ambos ofertados por EAD, na Universidade de Pernambuco (UPE).

A citada pesquisa testou cinco algoritmos de classificação para definição dos modelos preditivos: Árvore de Decisão, Máquina de Vetor de Suporte (SVM, do inglês, *Support Vector Machine*), Rede Neural Artificial, K-Vizinhos Mais Próximos (KNN, do inglês, *K-nearest Neighbors*) e Regressão Logística, sendo este último o que apresentou resultados mais relevantes, embora os demais não ficaram muito distantes, nas métricas analisadas.

A partir dessa referência, este estudo foi desenvolvido no sentido de verificar se o mesmo conjunto de variáveis usadas e os algoritmos de classificação aplicados, podem também ser replicados e validados em outro cenário educacional. Desta vez nos cursos de graduação em Administração Pública e na Licenciatura em Biologia, ofertados também por EAD, mas pela Universidade Federal do Vale do São Francisco (UNIVASF).

Algumas adpatações no processo de replicação do estudo foram necessários, tais como: mudança de tecnologia, ajuste nos scripts de coleta de dados e redução de cinco para três algoritmos de classificação. Essas alterações não alteraram os objetivos do trabalho, apenas forneceram novas e adequadas condições para o seu desenvolvimento.

Assim, a principal questão a ser esclarecida neste trabalho é se um conjunto de variáveis representativas da Teoria da Distância Transacional (TDT) e os alguns dos algoritmos classificadores, também podem ser usados em modelos preditivos de evasão na EAD, em um cenário diferente do originalmente apresentado por RAMOS (2016).

Espera-se com este trabalho contribuir para o fortalecimento da EAD, além de fomentar a linha de pesquisa voltada para o estudo das tecnologias educacionais, tão evidenciadas e diversificadas, a partir do uso cada vez maior das tecnologias de informação e comunicação no processo de ensino/aprendizagem, particularmente aquelas destinadas a reduzir os atuais índices de evasão verificados na modalidade.

1.1 OBJETIVOS

Esta pesquisa será desenvolvida com o propósito de atingir os seguintes objetivos geral e específicos:

1.1.1 Objetivo geral

Avaliar se um conjunto de modelos de predição desenvolvidos para outro cenário de EAD, pode também ser usado para prever quais alunos têm tendência a evasão em

cursos nessa modalidade na UNIVASF, mantendo os resultados em níveis satisfatórios, comparados aos originais.

1.1.2 Objetivos específicos

- Adaptar os modelos preditivos já desenvolvidos para uma outra ferramenta tecnológica;
- Aplicar os classificadores em bases de dados de cursos EAD da UNIVASF;
- Avaliar os resultados dos classificadores segundo métricas consolidadas.

1.2 ORGANIZAÇÃO DO TEXTO

Esse trabalho está organizado em 5 capítulos. No primeiro capítulo apresenta-se o projeto, uma contextualização sobre o problema abordado, assim como os objetivos gerais e específicos.

No segundo capítulo, é realiada uma revisão sobre a TDT, MDE e Aprendizagem Supervisionada, com objetivo de promover um maior detalhamento sobre os conceitos utilizados ao longo do texto. Também neste capítulo são apresentados resumos de trabalhos relacionados com esta pesquisa.

O terceiro capítulo explora os detalhes da caracterização da pesquisa e a metodologia aplicada, Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge Discovery in Databases*).

O quarto capítulo contém o cronograma de atividades para a disciplina Trabalho de Conclusão de Curso II (TCC II).

E por fim, o quinto capítulo contém as considerações finais, os resultados esperados e a contribuição da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

A necessidade de apoiar o desenvolvimento da EAD tem feito surgir novas teorias, métodos, abordagens de ensino e tratamento das informações, geradas nas diversas tecnologias usadas nessa modalidade. Nas seções seguintes, serão abordadas as principais temáticas envolvidas neste estudo, o que oferecerá as bases teóricas necessárias para a fundamentação da pesquisa.

2.1 TEORIA DA DISTÂNCIA TRANSACIONAL

Em 1972, Michael Grahame Moore propôs uma teoria pioneira para a EAD. Essa teoria seria, posteriormente, denominada de Teoria da Distância Transacional (TDT). Ao longo de mais de 40 anos, desde da proposição da TDT, o próprio autor e outros pesquisadores trataram de atualizá-la, principalmente, em razão da evolução tecnológica. Os textos originais do autor estão nas suas obras de 1973, 1993 e 2013 (MOORE, 1973, 1993, 2013).

Em seus estudos, Moore (2008) afirmou que na EAD não existe apenas uma distância física entre professores e alunos, mas, também, uma distância psicológica. Na TDT, as interações do estudante, com o professor, com o conteúdo e com os estudantes, podem ser estudadas com base em construtos elementares, sendo eles, a estrutura dos programas ou cursos, o diálogo entre alunos e professores e o grau de autonomia do discente. De acordo com a TDT a EAD tem a sua própria identidade e características pedagógicas distintivas. Como outras teorias, a TDT pode ser usada no estabelecimento de uma heurística, para a tomada de decisões em projetos de cursos EAD (MOORE, 2008).

Dewey e Bentley (1960) elaboraram o conceito de transação, que, conforme foi exposto posteriormente por Boyd e Apps (1980), denota a interação entre o ambiente, os indivíduos e os padrões de comportamento numa dada situação. Uma transação, em EAD, é a interação entre professores e alunos que estão espacialmente separados. Como foi definido por Moore (2008), essa separação cria padrões especiais de comportamento que afetam tanto o ensino quanto o aprendizado. Derivado da separação, surge um espaço psicológico e comunicacional propício a mal-entendidos nas interações instrutor-aluno. A esta separação é dado o nome de Distância Transacional (DT).

Faz-se necessário lembrar que, segundo Moore (2008) a distância transacional não é um valor fixo ou dicotômico, na verdade, é um valor relativo e contínuo. Além disso, essa distância é diferente para cada estudante, mesmo entre os que compartilham o mesmo curso. Foi apontado por Rumble (1986) que existe uma DT mesmo em cursos presenciais. Com base nisso, podemos dizer que a EAD é um subconjunto da educação e os estudos

realizados em EAD podem auxiliar a teoria e a prática da educação tradicional. Porém, em uma situação classificada como EAD, a distância entre os participantes — professores e alunos — é grande o suficiente para justificar a investigação de técnicas próprias de ensino-aprendizagem.

Os procedimentos de ensino se dividem em dois grupos, e acontece também um terceiro grupo de variáveis que descreve o comportamento dos alunos. A DT é uma função desses três grupos de variáveis. Na TDT, estes grupos de variáveis recebem o nome de Diálogo, Estrutura e Autonomia do Aluno (MOORE, 2008).

2.1.1 Diálogo

O diálogo foi originalmente definido por Moore (1973, 1993, 2013) como sendo interações focadas, positivas e propositais entre o professor e os alunos. Ainda segundo Moore, o diálogo ocorre entre professores e alunos quando alguém ensina e os demais reagem. Interações negativas ou neutras não são classificadas como diálogo. O diálogo deve ser direcionado para o aperfeiçoamento da compreensão por parte do aluno.

"A extensão e natureza do diálogo são determinadas pela filosofia educacional da instituição responsável pelo projeto do curso, pelas personalidades do professor e do aluno, pelo tema do curso e por fatores ambientais." (CABAU; COSTA, 2018, p. 438).

Moore (2008) cita meios de comunicação como um importante fator ambiental na EAD, no entanto, relata ser importante que outras variáveis sejam atendidas à medida que a EAD amadurece, as variáveis destacadas por Moore foram: projeto de curso, seleção e treinamento de instrutores e o estilo de aprendizagem dos alunos.

O diálogo é o mediador central da DT e referenciado como medida de aprendizado ao passo que a DT seria uma medida de não-aprendizado. No entanto, já que o diálogo não se limita apenas à interação professor-aluno, especialmente com os avanços da EAD provendo novas formas de interações entre estudantes, diversos pesquisadores vêm propondo a inclusão de interações entre alunos no conceito de diálogo (BENSON; SAMARAWICKREMA, 2009; CHEN; WILLITS, 1999; HUANG *et al.*, 2016).

2.1.2 Estrutura do curso

A estrutura do curso diz respeito aos elementos do projeto, bem como, divisão do curso em unidades, objetivos, estratégias institucionais e métodos de avaliação. A estrutura transmite a flexibilidade ou rigidez dos elementos do curso. É, também, responsável pela facilitação ou não-facilitação do diálogo (MOORE, 2008).

Como o diálogo, a estrutura do curso é uma variável qualitativa, e a medida da estrutura em um programa EAD é, normalmente, determinada pela natureza dos meios de comunicação empregados, e também pela filosofia e personalidade dos professores,

pelas personalidades dos alunos e pelas restrições impostas pelas instituições educacionais (MOORE, 2008).

Embora Moore atribua como qualitativa o tipo de variável relacionada ao diálogo e à estrutura, diversos estudos recentes mostraram que é possível quantificar e mensurar esses componentes da TDT (ZHANG, 2003; HORZUM, 2011; PAUL *et al.*, 2015; RAMOS, 2016).

Em cursos gravados em fitas, discos, ou mesmo cursos televisionados a estrutura é rígida e o diálogo não existe, pois não é possível reorganizar o conteúdo para levar em consideração as interações de um aluno. Em contrapartida cursos por teleconferências, permitem ampla variedade de respostas alternativas do instrutor às perguntas dos participantes. Um curso altamente estruturado não possibilita o diálogo professor-aluno, consequentemente, a DT entre alunos e professores aumenta. No entanto, o contrário não pode ser generalizado. "...a extensão do diálogo e a flexibilidade da estrutura variam de programa para programa. É essa variação que dá a um programa maior ou menor distância transacional que outro" (MOORE, 2008).

Em um programa com pequena DT os alunos recebem instruções e orientações por meio do diálogo com o instrutor, nesse caso é possível ter uma estrutura aberta, que dê respaldo para tais interações. Em programas com maior DT é necessário uma estrutura robusta, materiais didáticos que forneçam todas as orientações, instruções e aconselhamentos que o instrutor puder prever, mas sem a possibilidade de alterações por meio de diálogo aluno-professor (MOORE, 2008).

Temos então que, em programas com maior DT, os alunos precisam se responsabilizar em escolher quais atividades e avaliações serão feitas e quando serão feitas. Mesmo que o curso seja bem estruturado, o estudante, na falta de diálogo, decidirá quais atividades serão realizadas, quando, e qual a importância de cada uma. Sendo assim, quanto maior a DT mais é exigido uma autonomia do aluno (MOORE, 2008).

2.1.3 Autonomia do aluno

No período do surgimento da TDT, década de 1970, ela representava a fusão de duas tradições pedagógicas que pareciam contraditórias. Uma, a tradição humanística, que valorizava o diálogo aberto, não-estruturado e interpessoal, tanto na educação quanto no aconselhamento. A outra, a tradição behaviorista, que valorizava o projeto sistemático da instrução, baseado em objetivos comportamentais com o máximo de controle do processo de aprendizagem por parte do professor. No início dos anos 1970, a EAD era dominada pela tradição behaviorista. Tanto que, o título do primeiro trabalho sobre a TDT de Moore (1972) foi: "A autonomia do aluno — a segunda dimensão da aprendizagem independente". Nesse trabalho Moore afirmou que: "educadores por correspondência limitavam o potencial do seu método ao negligenciar a habilidade dos alunos em compartilharem a responsabilidade

por seus próprios processos de aprendizagem" (MOORE, 2008).

O termo "autonomia do aluno" foi escolhido para descrever os padrões de comportamento de alunos que usavam materiais didáticos e programas de ensino para atingir seus próprios objetivos, à sua maneira e sob seu próprio controle (MOORE, 2008).

Autonomia do aluno se refere a capacidade de se auto-direcionar. Moore (2008) definiu o estudante autônomo ideal como "a pessoa emocionalmente independente de um professor" e quem "tem capacidade de abordar o assunto estudado diretamente sem a ajuda de um instrutor". Diferente da estrutura do curso e do diálogo, é um fator que depende apenas do aluno. Um aprendiz pouco autônomo pode precisar de um direcionamento maior e uma estrutura mais rígida (HUANG *et al.*, 2016).

2.2 EVASÃO DE ALUNOS NA EAD

Segundo o censo, realizado pela Associação Brasileira de Educação a Distância (ABED), com dados de 2016, consultou 340 instituições em todo o país, formadoras e fornecedoras de de produtos e serviços para EAD (ABED, 2017).

De acordo com o Censo EAD.BR 2016, as taxas de evasão informadas pelos respondentes recaíram, principalmente, entre 11% e 25%. O censo, também, revelou que, entre os respondentes, cursos semipresenciais tem taxa de evasão menor que cursos totalmente a distância. A Tabela 1 compara os índices dos censos realizados pela ABED entre 2014 e 2017 (ABED, 2014; ABED, 2015; ABED, 2016; ABED, 2017).

Tabela 1 – Taxas de evasão ao longo dos anos segundo o censo realizado pela ABED

Taxas de evasão declaradas	**Percentuais de instituições declarantes, por faixa**			
	2013	**2014**	**2015**	**2016**
Até 25%	65%	50%	53%	58%
Entre 26 e 50%	24%	38%	40%	19%
Acima de 50%	2%	2%	7%	1%
Não declararam	9%	10%	-	22%

Fonte: ABED (2014), ABED (2015), ABED (2016), ABED (2017).

Entre os motivos para a evasão investigados e declarados no censo, questões financeiras e falta de tempo foram os citados como os que geram maior evasão. Houve uma parcela considerável de respondentes que acredita que a evasão não é um problema em cursos totalmente a distância, pois os participantes podem sempre retornar.

Em cursos livres, o motivo mais citado foi a falta de tempo, e, também, grande parte dos respondentes acredita que os alunos desses cursos sempre podem retornar.

O Censo EAD.BR 2016 apontou que cursos presenciais, semipresenciais e corporativos possuem mecanismos que vão além do conteúdo e da interação online com professores para manter seus alunos engajados. Já os cursos totalmente a distância e cursos livres não corporativos dependem apenas da experiência do aluno com o conteúdo e com seus professores e tutores.

A Tabela 2 apresenta dados dos indicadores da evasão, em cursos superiores a distância, segundo o Mapa do Ensino Superior no Brasil Edições 2015 e 2016, que foram publicados pelo Sindicato das Empresas Mantenedoras do Ensino Superior (SEMESP) feito com base nos dados do INEP dos anos 2013 e 2014.

Tabela 2 – Taxas de evasão em cursos superiores presenciais e a distância

Ano	**Cursos presenciais**		**Cursos a Distância**	
	IES públicas	**IES privadas**	**IES públicas**	**IES privadas**
2013	17,8%	27,4%	25,6%	29,2%
2014	18,3%	27,9%	26,8%	32,5%

Fonte: SEMESP (2015), SEMESP (2016)

Segundo o trabalho de Paz e Cazella (2017) a evasão em instituições de ensino superior (IES) é um tema complexo na gestão universitária no Brasil. Um grave problemas das universidades brasileiras é o aumento das taxas de evasão escolar.

Manhães *et al.* (2012) identificaram que a descoberta precoce de grupos de estudantes com risco de evasão é condição importante para reduzir tal problema, pois possibilita proporcionar algum tipo de atendimento personalizado para a situação de cada aluno. Ainda segundo Manhães *et al.* (2012), os processos de identificação desses grupos à época eram manuais e sujeitos a falhas e dependiam, primordialmente, da experiência do docente.

2.3 RELAÇÃO ENTRE A DISTÂNCIA TRANSACIONAL E A EVASÃO EM CURSOS A DISTÂNCIA

Pela sua definição, a Distância Transacional é um dos fatores que pode gerar maior dificuldade no engajamento e na comunicação do estudante no ambiente de aprendizagem (GOEL *et al.*, 2012). Além de Moore (2008), outros autores afirmaram que quanto maior for a DT, maior a possibilidade de ocorrência de problemas como atritos, insatisfações e abandono de cursos (ZHANG, 2003; STEINMAN, 2007; HORZUM, 2011; MBWESA, 2014; PAUL *et al.*, 2015).

Segundo Zhang (2003), em sua tese de doutorado, demonstrou a existência de uma correlação negativa entre a distância transacional e o envolvimento dos alunos com a sua

aprendizagem, assim como com a sensação de satisfação e a intenção do aluno em persistir no seu curso *on-line*.

Para Steinman (2007), as percepções dos alunos de cursos *on-line* podem ser negativas se eles experimentam grande DT com o instrutor e com outros alunos, podendo ainda influenciar sua decisão de permanecer ou abandonar o curso. Assim, uma vez que a DT afeta a satisfação e retenção dos alunos, esse conceito é visto como um importante tópico de discussão sobre evasão em cursos *on-line*.

A obtenção dos construtos da DT pode refletir uma condição ou um estado de um curso no tempo de sua execução, permitindo, por exemplo, que professores e tutores notem um distanciamento exagerado de determinados alunos e consigam intervir no sentido de prevenir ou reverter situações de evasão de alunos do curso (HORZUM, 2011).

2.4 DESCOBERTA DE CONHECIMENTO

Segundo Costa *et al.* (2012) Mineração de Dados (DM, do inglês, *Data Mining*), pode ser interpretada como uma etapa de um processo mais amplo denominado como Descoberta de Conhecimento em Bases de Dados (KDD, do inglês, *Knowledge Discovery in Databases*). No KDD são identificadas duas grandes etapas: a de pré-processamento de dados, na qual os dados são captados, tratados e organizados, e a de pós-processamento dos resultados obtidos da etapa de mineração (FAYYAD *et al.*, 1996).

Para a obtenção de conhecimentos relevantes, no KDD, é necessário estabelecer metas bem definidas. Segundo Fayyad *et al.* (1996), as metas são definidas em função do objetivo na utilização da metodologia, sendo dois tipos básicos de metas: verificação e descoberta. No caso de verificação, o sistema está limitado a testar hipóteses definidas pelo usuário, enquanto que em descoberta o sistema encontra novos padrões de forma autônoma. Quando a meta é do tipo descoberta, em geral, o objetivo está relacionado com as seguintes tarefas de mineração de dados: predição e descrição.

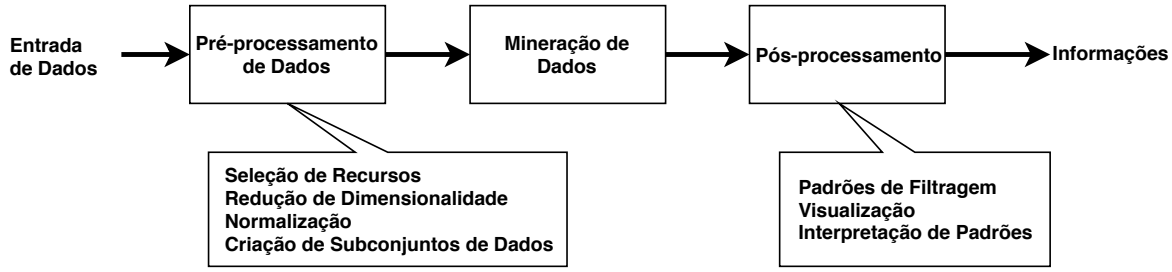
As tarefas preditivas buscam descobrir o valor de um determinado atributo com base nos valores de outros atributos. O atributo a ser predito pode ser chamado de variável preditiva, dependente ou alvo, já os atributos utilizados na predição são chamados de variáveis preditoras, independentes ou explicativas. Sendo generalista, a predição utiliza um conjunto de variáveis para predizer o valor de outras (FAYYAD *et al.*, 1996).

Tarefas descritivas objetivam encontrar padrões — correlações, tendências, grupos, trajetórias e anomalias — que representem os dados (FAYYAD *et al.*, 1996).

Para realizar tarefas de predição e descrição são utilizados alguma das seguintes tarefas e métodos de mineração de dados: classificação, regressão, agrupamento, sumarização, modelagem de dependência e identificação de mudanças e desvios.

Segundo Tan *et al.* (2009), DM é uma parte do KDD, um processo geral de conversão de dados brutos em informações úteis, sendo este composto de uma séries de passos de transformação, do pré-processamentos dos dados até o pós-processamento dos resultados da mineração de dados. A Figura 1 ilustra uma visão geral do KDD segundo Tan et al.

Figura 1 – Processo de descoberta de conhecimento em bases de dados



Fonte: Tan *et al.* (2009).

Ainda segundo Tan *et al.* (2009), os dados de entrada podem estar armazenados nos mais diversos formatos (tabelas eletrônicas, bases de dados estruturadas, arquivos simples), e podem estar em um único repositório ou distribuídos por diversas fontes. A etapa de pré-processamento é responsável por transformar os dados brutos em dados apropriados para as análises seguintes. Fusão de dados de múltiplas fontes, limpeza para remoção de ruídos, e seleção de características relevantes à DM, são passos importantes realizados na etapa de pré-processamento. Como existem diversas formas de se coletar e armazenar os dados, o pré-processamento se torna, muitas vezes, a etapa mais demorada e trabalhosa do KDD.

De acordo com Tan *et al.* (2009), o pós-processamento é a etapa do KDD na qual os dados válidos e úteis gerados na etapa de mineração são integrados a ferramentas de auxílio na tomada de decisões. Um exemplo de pós-processamento é a visualização de dados, que permite por meio de gráficos, auxiliar na interpretação de comportamentos e características dos dados. Também podem ser utilizados testes estatísticos para eliminar resultados não legítimos da mineração de dados.

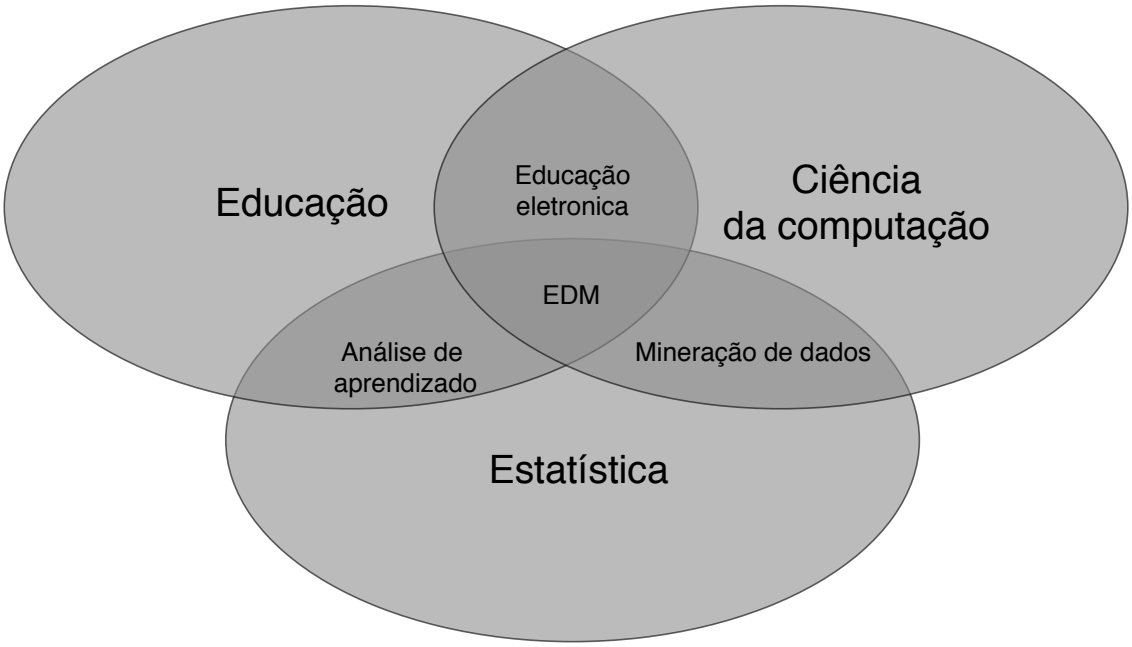
2.4.1 Mineração de Dados Educacionais

Segundo Costa *et al.* (2012), a área emergente de Mineração de Dados Educacionais (EDM, do inglês, *Educational Data Mining*) procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como AVAs, Sistemas Tutores Inteligentes (STIs), entre outros.

Muitos métodos utilizados em EDM são, originalmente, da área de mineração de dados. No entanto, segundo Baker *et al.* (2010), muitas vezes estes métodos devem ser modificados, por se fazer necessário considerar a hierarquia da informação. Existe também, falta de independência estatística nos tipos de dados encontrados ao coletar informações em contextos educacionais. Logo, diversos algoritmos e ferramentas utilizadas na área de DM não podem ser aplicados para análise de dados educacionais sem sofrerem os devidos ajustes (BAKER *et al.*, 2011; COSTA *et al.*, 2012).

A EDM pode ser descrita como a combinação de três áreas principais (Figura 2): ciência da computação, educação e estatística. As interseções dessas três áreas forma subáreas próximas da EDM, como sendo análise de aprendizado (LA, do inglês, *Learning Analytics*), ambientes de aprendizado baseados em computador e aprendizado de máquina (ROMERO; VENTURA, 2013).

Figura 2 – Principais áreas relacionadas com EDM



Fonte: Romero e Ventura (2013).

2.5 APRENDIZAGEM SUPERVISIONADA

O campo do Aprendizado de Máquina (ML, do inglês, *Machine Learning*) fornece uma ampla área para cientistas explorarem modelos e algoritmos de aprendizado que podem ajudar "máquinas" (computadores) a aprender sobre um sistema com base em dados. Em outras palavras, o objetivo do ML é construir sistemas inteligentes. Algoritmos de aprendizado são ferramentas de reconhecimento de padrão. A seguir é apresentado, de uma forma geral, a descrição de um problema de ML. Suponha que são dados um

conjunto de dados e sua respectiva resposta para um sistema. Então, o problema de ML pode ser definido como ajustar um modelo entre eles, os dados e sua resposta, e como treinar e validar o modelo para aprender as características do sistema por meio dos dados (SUTHAHARAN, 2016).

A tarefa de Aprendizagem Supervisionada é a seguinte:

Dados um conjunto de treinamento de N exemplos de pares entrada/saída

$$(x_1, y_1), (x_2, y_2) \dots (x_n, y_n),$$

onde cada y_j foi gerado por uma função desconhecida $y = f(x)$, descobrir uma função h que aproxime a verdadeira função f (RUSSELL; NORVIG, 2011).

Na definição anterior, x e y podem ser qualquer valor, não necessariamente numérico. A função h é uma hipótese. Aprender é procurar em um espaço de hipóteses possíveis por uma que tenha alto desempenho, mesmo em exemplos não contidos no conjunto de treinamento. Para mensurar a acurácia de uma hipótese se utiliza um conjunto de teste, exemplos que são distintos do conjunto de treinamento. É dito que uma hipótese generaliza bem se prediz corretamente os valores y para exemplos novos (RUSSELL; NORVIG, 2011).

Quando a saída y é uma em um conjunto finito de valores, o problema de aprendizado é denominado classificação, e é chamado classificação booleana ou classificação binária quando existem apenas dois valores possíveis(RUSSELL; NORVIG, 2011).

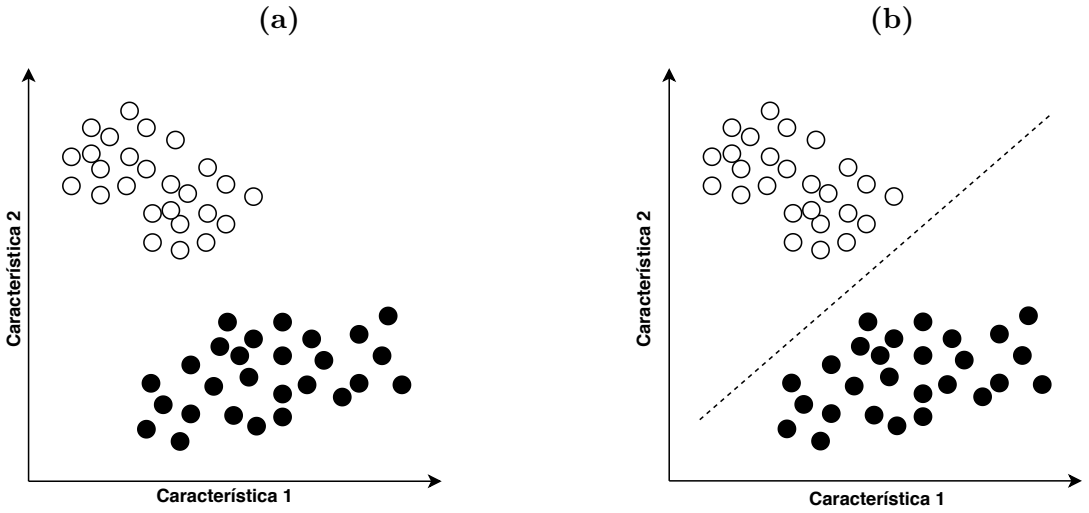
2.5.1 Classificação

De acordo com Tan *et al.* (2009) a Classificação é a tarefa de aprender uma função alvo f que mapeie cada conjunto de atributos x para um dos rótulos de classes y pré-determinados. A função alvo também pode ser chamada de modelo de classificação.

Em problemas de classificação, assumimos que são disponibilizados dados etiquetados (classes) para gerar regras que podem ajudar a atribuir uma etiqueta a novos dados que não possuem classes. Nesse caso, podemos derivar uma regra exata pela disponibilidade das classes. A Figura 3 ilustra um exemplo com duas classes, etiquetadas com pontos brancos e pontos pretos, e uma reta (Figura 3b) representado a regra que nos ajuda a estabelecer uma classe para cada novo ponto (SUTHAHARAN, 2016).

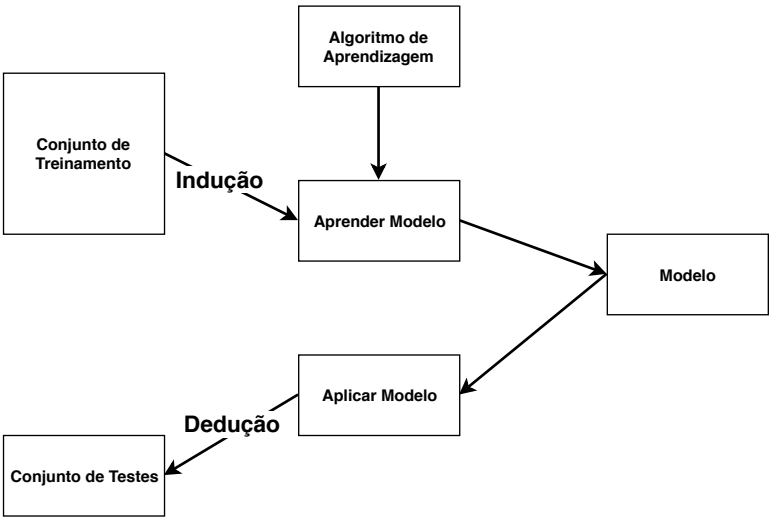
Em uma abordagem geral, para a construção de um modelo de classificação, primeiro, um conjunto de treinamento consistindo de registros com rótulos de classe conhecido é fornecido. O conjunto de treinamento é usado para construir um modelo de classificação, que é então aplicado a um conjunto de testes, constituído por registros com rótulos desconhecidos para o modelo. A Figura 4 ilustra essa abordagem geral (TAN *et al.*, 2009).

Figura 3 – Exemplo de classificação



Fonte: Suthaharan (2016).

Figura 4 – Abordagem geral para a construção de um modelo de classificação



Fonte: Tan *et al.* (2009).

Segundo Tan *et al.* (2009) a avaliação do desempenho de um modelo de classificação é baseada na contagem de registros do conjunto de teste que foram classificados correta e incorretamente. Estas contagens são organizadas em uma tabela denominada matriz de confusão. O Quadro 1 apresenta uma matriz de confusão para um problema de classificação binária. A partir das entradas da matriz de confusão, o número de previsões corretas realizadas pelo modelo é $(f_{11} + f_{00})$ e o número de previsões incorretas é $(f_{10} + f_{01})$.

A matriz de confusão mostra informações importantes para determinar o desem-

Quadro 1 – Exemplo de matriz de confusão

		Classe prevista	
		Classe = 1	Classe = 0
Classe real	Classe = 1	f_{11}	f_{10}
	Classe = 0	f_{01}	f_{00}

Fonte: O autor.

penho do modelo, no entanto, resumir essas informações em um único número é mais conveniente quando queremos comparar o desempenho entre diferentes modelos. Isso pode ser feito usando uma métrica de desempenho como a precisão, que é definida da seguinte maneira (TAN *et al.*, 2009):

$$\text{Precisão} = \frac{\text{Número de previsões corretas}}{\text{Número total de previsões}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

2.5.1.1 Árvore de decisão

Em ML, existem dois tipos de árvores de decisão: árvores de regressão e árvores de classificação. Uma árvore de decisão utiliza uma abordagem baseada em regras para dividir o domínio dos dados em múltiplos espaços lineares e prever respostas. Se as respostas previstas forem contínuas, então a árvore de decisão é chamada de árvore de regressão, e se as previsões são discretas, ou seja, pertencem a uma classe, então a árvore de decisão é chamada de árvore de classificação (SUTHAHARAN, 2016).

Segundo Suthaharan (2016), árvores de decisão são modelos de aprendizado supervisionado, que mapeiam o domínio dos dados hierarquicamente em um conjunto de respostas. Dividindo o domínio dos dados (também chamado de nó), recursivamente, em dois subdomínios, de forma que os subdomínios tenham um maior ganho de informação que o nó que foi dividido. Já que o objetivo do aprendizado supervisionado é a classificação dos dados, portanto, o aumento do ganho de informação influencia na eficiência da classificação nos subdomínios criados pela divisão. Encontrar a divisão que traga o máximo de ganho de informação, ou seja, eficiência na classificação é o objetivo do dos algoritmos de otimização no aprendizado supervisionado baseado em árvores de decisão.

Suthaharan (2016) traz um exemplo de classificação usando árvore de decisão. Suponha que temos um sistema que produz eventos (observações) que podem pertencer a uma de duas classes, 0 ou 1, e estes eventos dependem apenas de uma variável. Consequentemente, definimos o domínio como: $D = \{e_1, e_2, \ldots, e_n\}$ (assumimos que isto é um conjunto ordenado), e seus rótulos de classe correspondentes $L = r_1, r_2, r_3, \ldots, r_n$, onde r_i pertence $\{0, 1\}$, e $i = 1 \ldots n$. A propagação dos rótulos das classes sobre o domínio dos dados determina a facilidade na classificação. Representamos o ganho de informação do domínio D em relação a L por I_i e dividimos o conjunto ordenado na localização m

para formar dois subdomínios $D_1 = \{e_1, e_2, \ldots, e_m\}$ e $D_2 = \{e_{m+1}, e_{m+2}, \ldots, e_n\}$ com os conjuntos de respostas correspondentes $L_1 = \{r_1, r_2, \ldots, r_m\}$ e $L_2 = \{r_{m+1}, r_{m+2}, \ldots, r_n\}$. Se os respectivos ganhos de informação são I_{i1} e I_{i2}, então m será considerado a melhor divisão se a média$(I_{i1}, I_{i2}) > I_i$. Não obstante, precisamos de uma boa medida quantitativa para mensurar o ganho de informação obtido após a divisão dos dados.

Vamos supor que p_0 e p_1 representem as probabilidades de que as classes 0 e 1 possam ser extraídas do domínio D, respectivamente. Se, por exemplo, $|p_0 - p_1| \to 1$; então podemos observar que uma classe em particular tem grande predominância neste domínio, portanto, não é mais necessário dividir os dados. Similarmente, se $|p_0 - p_1| \to 0$, então as classes tem predominância igual no domínio; logo, uma divisão é necessária. Neste caso geramos dois subdomínios D_1 e D_2. Digamos que, q_0 e q_1 são as probabilidades de que a classe 0 e a classe 1 sejam derivadas do subdomínio D_1, respectivamente. Se a divisão for eficiente, $q_0 > p_0$ ou $q_1 > p_1$. Assumindo $q_0 > p_0$, então $q_0 = p_0 + \epsilon$, onde $\epsilon > 0$.

$$|q_0 - q_1| = |2q_0 - 1| = |2(p_0 + \epsilon) - 1| = |2p_0 + 2\epsilon - 1|$$

$$|q_0 - q_1| = |p_0 + 1 - p_1 + 2\epsilon - 1| = |p_0 - p_1 + 2\epsilon|$$

Esta equação enfatiza a seguinte inequação, (quando $q_0 > p_0$):

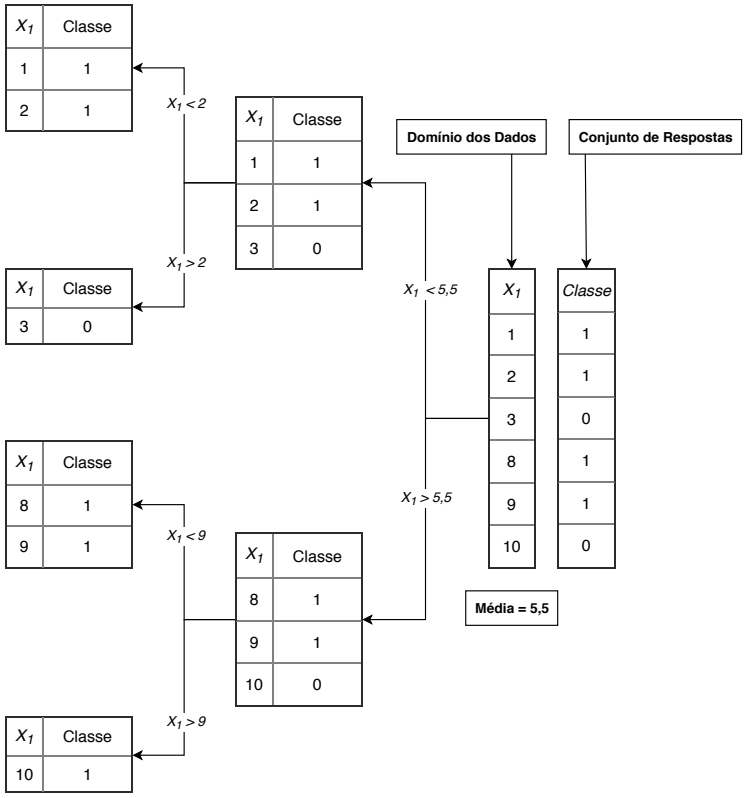
$$|q_0 - q_1| > |p_0 - p_1|$$

As diferenças absolutas na inequação acima são as medidas quantitativas de proporcionalidade entre as classes em seus respectivos subdomínios. Essas medida probabilística é uma boa métrica para abordagem de otimização de árvores de decisão. A Figura 5 vemos um exemplo de árvore de decisão em termos de divisão de domínios focado no ganho de informação.

2.5.1.2 K-ésimo vizinho mais próximo

Peterson (2009) introduziram um método não paramétrico de reconhecimento de padrões que ficou conhecido como Regra do K-ésimo Vizinho Mais Próximo (do inglês, *K-nearest-neighbor*, KNN). O KNN é um dos algoritmos de classificação mais simples e mais fundamentais e deveria ser a primeira escolha para um estudo de classificação quando se tem pouco ou nenhum conhecimento sobre a distribuição dos dados. A classificação com KNN foi desenvolvida a partir da necessidade de realizar análises discriminatórias quando estimativas confiáveis de densidade de probabilidade dos dados não são conhecidas ou difíceis de determinar. Peterson (2009) descreveram as propriedades formais do KNN, por exemplo, foi demonstrado que para $k = 1$ e $n \to \infty$ o erro de classificação do KNN é limitado pelo dobro da taxa de erro de Bayes. Desde que essas propriedades formais foram estabelecidas seguiu-se uma longa linha de investigações incluindo uma abordagem de

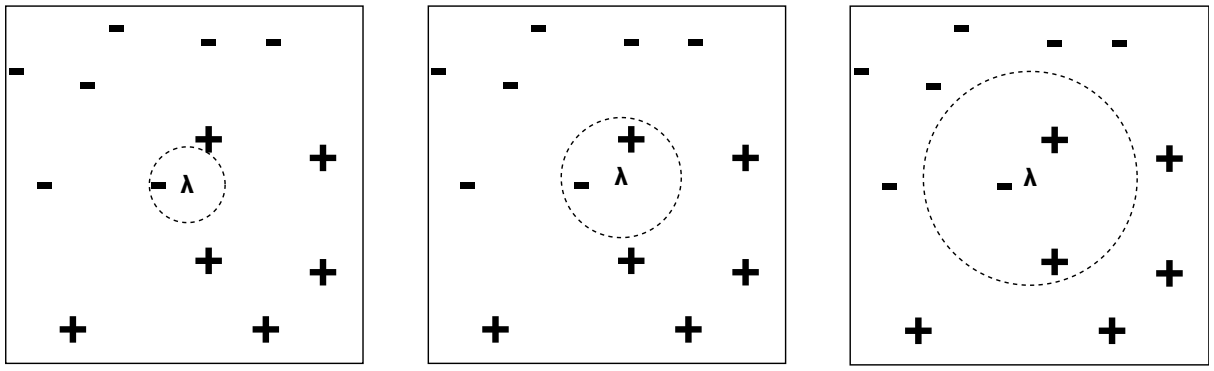
Figura 5 – Exemplo de árvore de decisão usada para classificação construída com um domínio de dados uni-dimensional



Fonte: Suthaharan (2016).

Figura 6 – Os 1, 2 e 3 vizinhos mais próximos de um ponto dado

(a) 1-vizinho mais próximo **(b)** 2-vizinhos mais próximo **(c)** 3-vizinhos mais próximo



Fonte: Tan *et al.* (2009).

rejeição , melhoramento em relação com a taxa de erro de Bayes , abordagem com pesos nas distâncias (PETERSON, 2009).

Segundo Tan *et al.* (2009) um classificador que utiliza KNN representa cada exemplo, de treinamento ou de teste, como um ponto de dado em um espaço d-dimensional, onde d é a quantidade de atributos. Dado um exemplo de teste, calcula-se a sua proximidade com o resto dos pontos de dados do conjunto de treinamento, usando alguma medida de distância, geralmente a distância euclidiana. Os k vizinhos mais próximos de um determinado ponto de teste z se referem aos k pontos com menor distância de z. Então, z é classificado com base nos rótulos de classe do seus vizinhos mais próximos e lhe é atribuída a classe majoritária dos seus vizinhos mais próximos. No exemplo da Figura 6 determinamos que o símbolo **-** representa a classe negativo e o símbolo $+$ representa a classe positivo e λ representa um ponto dado a ser classificado. Na Figura 6a, onde $k = 1$, foi atribuído ao ponto dado a classe negativo. Na Figura 6b, com $k = 2$, os dois vizinhos mais próximos do ponto dado tem classes distintas, portanto, podemos atribuir aleatoriamente qualquer uma das duas classes. Na Figura 6c, com $k = 3$, dois dos vizinhos mais próximos do ponto dado são da classe positivo e apenas um é da classe negativo, logo, atribuímos ao ponto dado a classe positivo.

O desempenho de um classificador usando KNN pode ser melhorado quando os atributos são transformados antes da análise de classificação. A forma mais comum de transformação é a normalização ou padronização. A normalização remove efeitos provocados por atributos com escalas diferentes, exemplo, o atributo peso de um paciente pode ser baseado na unidade quilograma enquanto os valores de proteína no sangue são baseados em nanograma por decilitro variando entre -3 e 3, logo, o peso do paciente teria maior influência no cálculo das distâncias entre os pontos de exemplo e, por consequência, na classificação (PETERSON, 2009).

2.5.1.3 Regressão logística

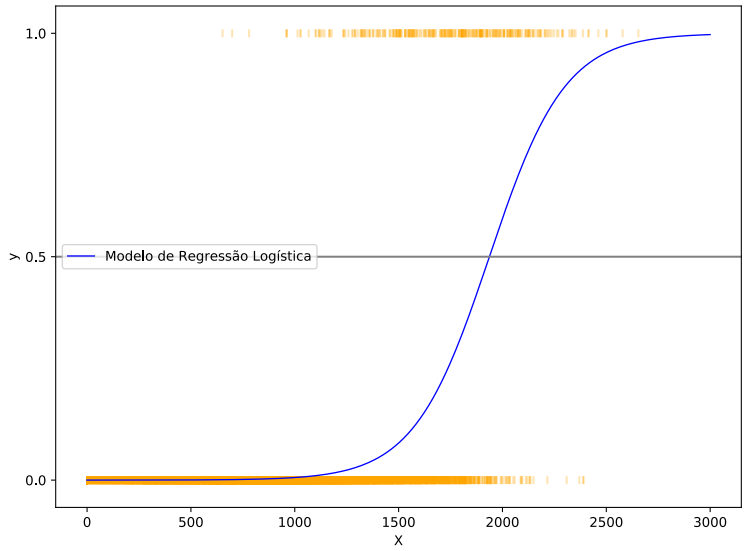
A Regressão Logística (RL) é uma generalização da regressão linear. É usada principalmente para prever variáveis dependentes binárias ou de múltiplas classes. Como a variável de resposta é discreta, ela não pode ser modelada diretamente por regressão linear. Portanto, em vez de prever uma estimativa de ponto do evento em si, o modelo baseia-se para prever a probabilidade de sua ocorrência (ŞEN *et al.*, 2012).

O modelos de RL surge do desejo de modelar as probabilidades posteriores de de K classes através de funções lineares em x, ao mesmo tempo, garantir que a soma dessas probabilidades seja um e elas permaneçam no intervalo entre 0 e 1 (JAMES *et al.*, 2013).

Uma vantagem da RL no processo de classificação de uma variável dependente binária (binomial), é que, nela, pode ser usado um conjunto de variáveis independentes numéricas ou categóricas (KLEINBAUM; KLEIN, 2002).

Dada uma variável ou conjunto de variáveis X, podemos utilizar um modelo de RL para calcular a probabilidade de pertencimento à classe y. Para cada valor ou valores

Figura 7 – Previsões utilizando regressão logística. As probabilidades se encontram no intervalo entre 0 e 1



Fonte: James *et al.* (2013).

de X pode ser feita uma previsão para a classe y. Por exemplo, pode-se dizer que o item sendo testado pertence a classe y sempre que o modelo RL retornar uma probabilidade maior que 50%, sendo que este limiar pode ser ajustado de acordo com a necessidade do problema abordado (JAMES *et al.*, 2013).

Os modelos de RL para diversas variáveis é descrito pela seguinte fórmula:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i}$$

Onde, β_0, β_1, ..., β_k são os coeficientes das variáveis que explicam a ocorrência de determinado evento p_i é a probabilidade de um evento ocorrer dado o conjunto de variáveis i.

O resultado do modelo logit é uma curva em forma de S (Figura 7). Para estimar um modelo de regressão logística, essa curva de valores previstos é ajustada aos dados reais, analogamente como é feito com uma relação linear em regressão múltipla.

Em níveis muito baixos da variável independente, a probabilidade se aproxima de 0%, mas nunca alcança tal valor. Da mesma forma, quando o valor da variável independente aumenta, os valores previstos crescem para acima da curva, mas, em seguida, a inclinação começa a diminuir, aproximando a probabilidade de 100%, sem, entretanto, exceder esse valor (HAIR *et al.*, 2009).

2.6 TRABALHOS RELACIONADOS

Ramos *et al.* (2016), propôs o mapeamento do comportamento de usuários de um *Learning Management System* (LMS), em variáveis que representam os contrutos da TDT. O objetivo foi descrever e validar um conjunto de variáveis com as quais esses construtos podem ser medidos, permitindo o desenvolvimento de pesquisas na área, assim como a obtenção destas medidas a qualquer momento do curso e sem a necessidade de questionários. A criação e validação de um conjunto final de variáveis foi feita a partir da Análise Fatorial Confirmatória (CFA), que apontou como cada construto pode ser representado por um conjunto de atributos obtidos a partir do banco de dados do LMS.

Ramos *et al.* (2018), analisaram a performance de diferentes algoritmos na previsão da evasão em alunos EAD. Neste trabalho foram utilizados dados de turmas de dois cursos de graduação na Universidade de Pernambuco (UPE). Os algoritmos testados foram: Árvore de Decisão, Máquina de Vetor de Suporte (SVM), Rede Neural Artificial, K-Vizinhos Mais Próximos (KNN) e Regressão Logística. As variáveis foram construídas com base na TDT. O algoritmo com maior acurácia foi o KNN, com maior precisão foi o SVM e a Regressão Logística teve os maiores valores de Recall e Área Sob a Curva ROC (AUC).

Queiroga *et al.* (2018), elaboraram um modelo de predição da evasão de estudantes em cursos técnicos a distância, através de mineração de dados, utilizando dados de turmas EAD do Câmpus Visconde da Graça (CaVG) do Instituto Federal Sul-rio-grandense (IFSul). Os algoritmos utilizados para gerar os modelos testados foram: *Bayes Net*, *Simple Logistic*, *Multilayer Perceptron*, *Random Forest* e J48, implementados na biblioteca WEKA. Todos os algoritmos selecionados previram com exatidão de 95% a evasão de um aluno antes do final do primeiro ano. O algoritmo que mais se destacou no quesito acurácia foi o Random Forest com 85%.

Manhães *et al.* (2011), utilizaram mineração de dados para identificar antecipadamente alunos com risco de evasão. Foram utilizados dados de cursos de graduação da Universidade Federal do Rio de Janeiro (UFRJ). Os resultados mostraram que utilizando as primeiras notas semestrais dos calouros é possível identificar com precisão de 80% a situação final do aluno no curso.

Paz e Cazella (2017), Aplicaram KDD em dados coletados em uma IES, e, através da tarefa de classificação, utilizando a técnica de árvores de decisão, atingiram acurácia de 90% na identificação de alunos evasores.

RAMOS (2016) desenvolveu e testou modelos preditivos baseados nos algoritmos Árvore de Decisão (TreeDecision), Máquina de Vetor de Suporte (SVM), Rede Neural Artificial (NeuralNet), k-Nearest Neighbors (KNN) e Regressão Logística (RegLog), usando com base as variáveis representativas dos construtos da distância transacional, obtidas em

trabalho anterior (RAMOS *et al.*, 2016). Esse trabalho serviu como principal referência para o desenvolvimento deste estudo, a fim de verificar a aplicabilidade do método em um outro cenário educacional.

3 METODOLOGIA PROPOSTA

3.1 CARACTERIZAÇÃO DA PESQUISA

Segundo Marconi e Lakatos (2003), a pesquisa é um procedimento formal, com método de pensamento reflexivo, que requer um tratamento científico e se constitui no caminho para conhecer a realidade ou para descobrir verdades parciais. A pesquisa é um procedimento sistemático e crítico, que permite descobrir novos fatos, relações ou leis acerca de qualquer campo do conhecimento.

Uma pesquisa pode ser caracterizada segundo os seguintes critérios (GIL, 2008):

a) Quanto à natureza: básica ou aplicada;

b) Quanto aos objetivos: exploratória, descritiva ou explicativa;

c) Quanto à abordagem: qualitativa ou quantitativa;

d) Quanto aos procedimentos: documental, bibliográfica, experimental, levantamento, estudo de caso, entre outros.

Este trabalho pode ser classificado como de natureza aplicada, já que será aplicada uma metodologia de busca de conhecimentos em bancos de dados e métodos de classificação para prever a evasão de cursos EAD.

Em relação aos objetivos podemos classificar este trabalho como pesquisa exploratória e descritiva. Tendo como base Gil (2002), a pesquisa exploratória busca ampliar o conhecimento sobre o problema, procurando torná-lo mais explícito ou a construção de hipóteses, tendo como objetivo central o aperfeiçoamento de ideias ou a revelação de intuições. E a pesquisa descritiva objetiva descrever características de determinado fenômeno ou população. Este trabalho utiliza uma metodologia de exploração de conhecimento para tentar prever um comportamento em um conjunto de uma população.

Quanto à abordagem, este trabalho é classificado como quantitativo, em razão da utilização de abordagens algorítmicas de Mineração de Dados, a partir das quais serão extraídas as características dos estudantes de EAD e aplicadas modelos de classificação que farão a devida categorização.

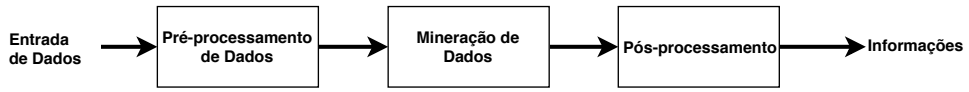
No quesito procedimentos classificamos este trabalho como pesquisa experimental. De acordo com Gil (2002), a pesquisa experimental consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos qua variável produz no objeto.

No caso deste trabalho, o objeto de estudo é a evasão na EAD da UNIVASF e as variáveis foram definidas pela TDT.

3.2 MÉTODO

Para tratamento e preparação dos dados para os diferentes modelos de classificação que serão avaliados, será utilizado o processo de Descobrimento de Conhecimento em Banco de Dados (do inglês, *Knowledge Discovery in Databases*, KDD) como descrito por Tan *et al.* (2009) e ilustrado na Figura 8. Este processo envolve uma série de passos com o objetivo de transformar dados brutos em informações úteis.

Figura 8 – Fluxo básico do processo KDD



Fonte: Tan *et al.* (2009).

A fase de entrada de dados será desenvolvida, baseado no trabalho de RAMOS (2016), coletando as variáveis mais relevantes que poderiam representar cada um dos três construtos da TDT. Os dados de onde serão retiradas as variáveis estão armazenados nas bases de dados do Moodle, atualmente em uso pelos cursos de graduação oferecidos pela UNIVASF na modalidade EAD.

Na etapa de pré-processamento de dados ocorrem as transformações e adaptações dos dados para os algoritmos de Mineração de Dados. Entre essas transformações podemos citar: normalização, limpeza de valores faltantes, identificação de outliers, entre outros. Esta etapa, geralmente, exige muito tempo e esforço. A correta execução deste passo resultará em melhores resultados nas etapas posteriores.

No contexto deste trabalho, utilizaremos ferramentas de análises exploratórias e *scripts* de buscas em bancos de dados para construção da base de dados a ser utilizada na etapa posterior.

Em seguida, ocorre a etapa de mineração de dados, onde são buscados padrões de interesse ou características que representem as tendências dos dados, entre os métodos de busca de padrões podemos citar: clusterização, classificação, regressão, entre outros.

Para este trabalho, utilizaremos os algoritmos de classificação a seguir: Árvore de Decisão, KNN, e Regressão Logística. Nesta etapa, os parâmetros dos algoritmos de classificação serão ajustados para que a performance dos mesmos seja melhorada.

Na última etapa, pós-processamento, serão avaliados e interpretados os padrões extraídos na etapa de mineração, podem ocorrer retornos a qualquer etapa anterior

para mais iterações. Esta etapa pode envolver a visualização dos padrões e modelos gerados, ou visualização dos dados fornecidos. Neste passo, o conhecimento descoberto será documentado para possível uso posterior, em uma ferramenta de geração de relatórios ou de visualização, tipo dashboard.

3.3 MATERIAIS

3.3.1 Moodle

O Moodle[1] é uma plataforma de ensino projetada para oferecer a educadores, administradores e estudantes, com uma sistema integrado, simple e robusto, a criação de ambientes de aprendizado personalizados. É financiado por uma rede de mais de 80 empresas ao redor do mundo.

Moodle é um software de código aberto sob a licença GNU General Public License. Qualquer pessoa pode adaptá-lo, estendê-lo ou modificá-lo, tanto para uso comercial ou não-comercial, sem nenhum tipo taxa de licenciamento e se beneficiando de sua eficiência e custo, flexibilidade e outras vantagens de usar o Moodle.

3.3.2 MySQL

MySQL[2] é a base de dados mais popular no mundo. Provê performance, confiabilidade e facilidade de uso, MySQL vem liderando a escolha de aplicações web, usado por grandes empresas na internet como: Facebook, Twitter, YouTube, Yahoo! e muitas outras.

MySQL é sistema de gerenciamento de banco de dados (SGDB), baseado na linguagem SQL (do inglês, Structured Query Language). Entre as vantagens suas vantagens podemos listar: portabilidade, compatibilidade, excelente desempenho e estabilidade, facilidade de manuseio e é um software livre sob a licença GPL.

3.3.3 Python

Python[3] é uma linguagem de programação de código aberto classificada como linguagem de alto nível de abstração. Considerada de fácil manuseio mesmo por usuários iniciantes. É mantida e desenvolvida pela Python Software Foundation

Graças a sua enorme comunidade, existem diversos pacotes e bibliotecas desenvolvidas em Python para as mais variadas tarefas, desde servidores HTTP, desenvolvimento de aplicações desktop até mineração de dados, inteligência artificial e estatística.

[1] <https://moodle.org/> Acesso em: 06 de mar. 2019
[2] <https://www.mysql.com/> Acesso em: 06 de mar. 2019
[3] <https://www.python.org/> Acesso em: 06 de mar. 2019

3.3.4 Anaconda Python Distribution

A distribuição de código aberto Anaconda[4] é uma maneira fácil de realizar tarefas de mineração de dados e aprendizado de máquina em ambientes Linux, Windows ou Mac OS X. Anaconda é um gerenciador de pacotes e ambientes e uma distribuição Python especializada em data science com mais de 1500 pacotes de código aberto.

3.3.5 Jupyter Notebook

Jupyter Notebook[5] é uma aplicação web de código aberto que permite a criação e compartilhamento de documentos que contém código em tempo de execução, equações, visualizações e textos narrativos. Funciona como uma IDE (do inglês, Integrated Development Environment) e foi desenvolvido para tarefas de limpeza e transformação de dados, simulações numéricas, modelagem estatística, visualização de dados, aprendizado de máquina e mais.

Jupyter Notebook suporta mais de 40 linguagens de programação incluindo Python e já vem pré configurado na distribuição Anaconda.

3.3.6 Python Data Analysis Library

Python Data Analysis Library[6], ou simplesmente pandas, é uma biblioteca de código aberto sob a licença BSD que provê estruturas de dados e ferramentas de análise de dados de alta performance e fácil uso para a linguagem de programação Python. Pandas proporciona estruturas de dados rápidas, flexíveis e expressivas desenvolvidas para uso com dados relacionais ou etiquetados.

A biblioteca pandas já vem configurada para uso na distribuição Anaconda.

3.3.7 Numpy

NumPy[7] é o pacote fundamental para computação científica em Python. Contento, além de outra funcionalidades, um poderoso vetor n-dimensional, funções de broadcast sofisticadas, ferramentas de integração com códigos C/C++ e Fortran, ferramentas de algebra linear, transformadas de Fourier e números aleatórios.

Além dos óbvios usos científicos, NumPy também pode ser usado como um invólucro para dados genéricos. Tipos de dados arbitrários podem ser definidos, isso permite que seja integrado de forma rápida com uma miríade de bases de dados.

[4] <https://www.anaconda.com/> Acesso em: 06 de mar. 2019
[5] <https://jupyter.org/> Acesso em: 06 de mar. 2019
[6] <https://pandas.pydata.org/> Acesso em: 06 de mar. 2019
[7] <http://www.numpy.org/> Acesso em: 06 de mar. 2019

NumPy é uma biblioteca de código aberto sob a licença BSD e é presente na distribuição Anaconda.

3.3.8 Scikit-learn

Scikit-learn[8] é um módulo Python para aprendizado de máquina de código aberto sob a licença BSD. Além das principais tarefas de mineração, como: classificação, regressão e clusterização a biblioteca proporciona as visualizações mais básicas para análise exploratória. Scikit-learn é compatível com pandas e NumPy e pode ser encontrado na distribuição Anaconda.

[8] <https://scikit-learn.org/> Acesso em: 06 de mar. 2019

4 CONSIDERAÇÕES FINAIS

A modalidade EAD ajuda a democratizar o ensino, levando-o às regiões de difícil acesso aos professores ou dando a oportunidade ao estudante de criar sua própria rotina de estudos. A evasão desta modalidade de ensino ainda é um grande problema a ser resolvido, logo, existe a necessidade de pesquisa científica nesta área.

Com o uso crescente de ferramentas de tecnologia da informação em EAD fica evidente que o uso de aprendizagem de máquina pode ser utilizado para modelar e prever os fenômenos que causam a evasão.

O fluxo de descoberta de conhecimento em bases de dados descrito na metodologia deste trabalho será utilizado como arcabouço para o desenvolvimento dos modelos preditivos que serão comparados no decorrer da disciplina Trabalho de Conclusão de Curso II (TCC II). Espera-se obter resultados satisfatórios, aproximados aos encontrados na literatura.

5 CRONOGRAMA

A tabela 3 mostra o cronograma de atividades a serem executadas para o TCC II, com base no calendário de 2019.1 da UNIVASF.

Tabela 3 – Cronograma de atividades para o TCC II

Atividade	**Mai.**	**Jun.**	**Jul.**	**Ago.**
Pré-processamento de dados	X	X	X	
Mineração de dados			X	
Pós-processamento de dados				X
Escrita do TCC II	X	X	X	X
Defesa do TCC II				X

Fonte: O autor.

REFERÊNCIAS

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA. **CENSO EAD.BR 2013**. 2014. Disponível em: <http://www.abed.org.br/censoead2013/CENSO_EAD_2013_PORTUGUES.pdf>. Acesso em: 13 jan. 2019. Citado na página 16.

_____. **CENSO EAD.BR 2014**: Relatório analítico da aprendizagem a distância no brasil. 2015. Disponível em: <http://www.abed.org.br/censoead2014/CensoEAD2014_portugues.pdf>. Acesso em: 13 jan. 2019. Citado na página 16.

_____. **CENSO EAD.BR 2015**: Relatório analítico da aprendizagem a distância no brasil. 2016. Disponível em: <http://abed.org.br/arquivos/Censo_EAD_2015_POR.pdf>. Acesso em: 13 jan. 2019. Citado na página 16.

_____. **CENSO EAD.BR 2016**: Relatório analítico da aprendizagem a distância no brasil. 2017. Disponível em: <http://abed.org.br/censoead2016/Censo_EAD_2016_portugues.pdf>. Acesso em: 13 jan. 2019. Citado na página 16.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011. Citado na página 20.

BAKER, R. *et al.* Data mining for education. **International encyclopedia of education**, Elsevier Oxford, UK, v. 7, n. 3, p. 112–118, 2010. Citado na página 20.

BENSON, R.; SAMARAWICKREMA, G. Addressing the context of e-learning: using transactional distance theory to inform design. **Distance Education**, Taylor & Francis, v. 30, n. 1, p. 5–21, 2009. Citado na página 14.

BOYD, R. D.; APPS, J. W. Redefining the discipline of adult education. **The AEA Handbook Series in Adult Education**, Jossey-Bass„ 1980. Citado na página 13.

CABAU, N. C. F.; COSTA, M. L. F. A teoria da distância transacional: um mapeamento de teses e dissertações brasileiras (the theory of transactional distance: a mapping of brazilian theses and dissertations). **Revista Eletrônica de Educação**, v. 12, n. 2, p. 431–447, 2018. Citado na página 14.

CHEN, Y.-J.; WILLITS, F. K. Dimensions of educational transactions in a videoconferencing learning environment. **American Journal of Distance Education**, Taylor & Francis, v. 13, n. 1, p. 45–59, 1999. Citado na página 14.

COSTA, E. *et al.* Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1–29, 2012. Citado 3 vezes nas páginas 18, 19 e 20.

DEWEY, J.; BENTLEY, A. F. **Knowing and the known**. [S.l.]: Beacon Press Boston, 1960. Citado na página 13.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996. Citado na página 18.

GIL, A. C. Como elaborar projetos de pesquisa. **São Paulo**, v. 5, n. 61, p. 16–17, 2002. Citado na página 30.

_____. **Métodos e técnicas de pesquisa social**. 6. ed. [S.l.]: Editora Atlas SA SA, 2008. Citado na página 30.

GOEL, L.; ZHANG, P.; TEMPLETON, M. Transactional distance revisited: Bridging face and empirical validity. **Computers in Human Behavior**, Elsevier, v. 28, n. 4, p. 1122–1129, 2012. Citado na página 17.

HAIR, J. F. *et al.* **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009. Citado na página 27.

HORZUM, M. B. Developing transactional distance scale and examining transactional distance perception of blended learning students in terms of different variables. **Educational sciences: Theory and practice**, ERIC, v. 11, n. 3, p. 1582–1587, 2011. Citado 3 vezes nas páginas 15, 17 e 18.

HUANG, X. *et al.* Understanding transactional distance in web-based learning environments: An empirical study. **British Journal of Educational Technology**, Wiley Online Library, v. 47, n. 4, p. 734–747, 2016. Citado 2 vezes nas páginas 14 e 16.

JAMES, G. *et al.* **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado 2 vezes nas páginas 26 e 27.

KLEINBAUM, D. G.; KLEIN, M. Analysis of matched data using logistic regression. **Logistic regression: A self-learning text**, Springer, p. 227–265, 2002. Citado na página 26.

MANHÃES, L. *et al.* Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. **Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo**, 2012. Citado na página 17.

MANHÃES, L. M. B. *et al.* Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. **Simpósio Brasileiro de Informática na Educação-SBIE**, 2011. Citado na página 28.

MARCONI, M. d. A.; LAKATOS, E. M. **Fundamentos de Metodologia Científica**. 5. ed. São Paulo: Editora Atlas SA, 2003. Citado na página 30.

MBWESA, J. K. Transactional distance as a predictor of perceived learner satisfaction in distance learning courses: A case study of bachelor of education arts program, university of nairobi, kenya. **Journal of Education and Training Studies**, v. 2, n. 2, p. 176–188, 2014. Citado na página 17.

MOORE, M. G. Learner autonomy: The second dimension of independent learning. **Convergence**, International Council for Adult Education, v. 5, n. 2, p. 76, 1972. Citado na página 15.

_____. The theory of transactional distance. In: **Handbook of distance education**. [S.l.]: Routledge, 1973, 1993, 2013. Citado 2 vezes nas páginas 13 e 14.

_____. Teoria da distância transacional. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 1, n. 0, 2008. ISSN 1806-1362. Disponível em: <http://seer.abed.net.br/index.php/RBAAD/article/view/111>. Citado 6 vezes nas páginas 11, 13, 14, 15, 16 e 17.

PAUL, R. C. *et al.* Revisiting zhang's scale of transactional distance: Refinement and validation using structural equation modeling. **Distance Education**, Taylor & Francis, v. 36, n. 3, p. 364–382, 2015. Citado 2 vezes nas páginas 15 e 17.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, n. 1, p. 624. Citado 2 vezes nas páginas 17 e 28.

PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009. Revision #137311. Citado 3 vezes nas páginas 24, 25 e 26.

QUEIROGA, E. M. *et al.* Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações. **Revista Thema**, v. 15, n. 2, p. 425–438, 2018. Citado na página 28.

RAMOS, J. L. C. **Uma abordagem preditiva da evasão na educação a distância a partir dos construtos da distância transacional**. Tese (Doutorado) — Universidade Federal de Pernambuco, 2016. Citado 5 vezes nas páginas 10, 11, 15, 28 e 31.

RAMOS, J. L. C. *et al.* Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1463. Citado na página 28.

_____. Mapeamento de dados de um lms para medida de construtos da distância transacional. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2016. v. 27, n. 1, p. 1056. Citado 2 vezes nas páginas 28 e 29.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013. Citado na página 20.

RUMBLE, G. **The planning and management of distance education**. [S.l.]: Croom Helm, 1986. Citado na página 13.

RUSSELL, S.; NORVIG, P. Artificial intelligence a modern approach 3rd edition pdf. **Hong Kong: Pearson Education Asia**, 2011. Citado na página 21.

ŞEN, B.; UÇAR, E.; DELEN, D. Predicting and analyzing secondary education placement-test scores: A data mining approach. **Expert Systems with Applications**, Elsevier, v. 39, n. 10, p. 9468–9476, 2012. Citado na página 26.

SINDICATO DAS MANTENEDORAS DO ENSINO SUPERIOR. **Mapa do Ensino Superior no Brasil — 2015**. 2015. Disponível em: <http://convergenciacom.net/pdf/mapa-ensino-superior-brasil-2015.pdf>. Acesso em: 21 jan. 2019. Citado na página 17.

_____. **Mapa do Ensino Superior no Brasil — 2016**. 2016. Disponível em: <http://convergenciacom.net/pdf/mapa_ensino_superior_2016.pdf>. Acesso em: 21 jan. 2019. Citado na página 17.

STEINMAN, D. Educational experiences and the online student. **TechTrends**, Springer, v. 51, n. 5, p. 46–52, 2007. Citado 2 vezes nas páginas 17 e 18.

SUTHAHARAN, S. Machine learning models and algorithms for big data classification. In: **Integrated Series in Information Systems**. [S.l.]: Springer, 2016. v. 36. Citado 4 vezes nas páginas 21, 22, 23 e 25.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao datamining: mineração de dados**. [S.l.]: Ciência Moderna, 2009. Citado 7 vezes nas páginas 19, 21, 22, 23, 25, 26 e 31.

ZHANG, A. M. **Transactional distance in web-based college learning environments: Toward measurement and theory construction**. [S.l.]: Virginia Commonwealth University, 2003. Citado 2 vezes nas páginas 15 e 17.