CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
UNIDADE DE PÓS-GRADUAÇÃO, EXTENSÃO E PESQUISA
MESTRADO PROFISSIONAL EM GESTÃO E DESENVOLVIMENTO DA EDUCAÇÃO PROFISSIONAL

HENRIQUE RUIZ POYATOS NETO

AVALIAÇÃO DE DISCENTES NA MODALIDADE DE ENSINO A DISTÂNCIA EM CURSOS DE EDUCAÇÃO PROFISSIONAL EM NÍVEL SUPERIOR, POR MEIO DE ALGORITMOS PREDITIVOS

São Paulo
Abril/2019

HENRIQUE RUIZ POYATOS NETO

AVALIAÇÃO DE DISCENTES NA MODALIDADE DE ENSINO A DISTÂNCIA EM CURSOS DE EDUCAÇÃO PROFISSIONAL EM NÍVEL SUPERIOR, POR MEIO DE ALGORITMOS PREDITIVOS

Dissertação apresentada como exigência parcial para a obtenção do título de Mestre em Gestão e Desenvolvimento da Educação Profissional do Centro Estadual de Educação Tecnológica Paula Souza, no Programa de Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional, sob a orientação do Profa. Dra. Celi Langhi.

São Paulo
Abril/2019

FICHA ELABORADA PELA BIBLIOTECA NELSON ALVES VIANA
FATEC-SP / CPS – CRB8-8281

P891a

Poyatos Neto, Henrique Ruiz

Avaliação de discentes na modalidade de ensino a distância em cursos de educação profissional em nível superior, por meio de algoritmos preditivos / Henrique Ruiz Poyatos Neto. – São Paulo: CPS, 2019.

96 f. : il.

Orientadora: Profa. Dra. Celi Langhi

Dissertação (Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional) - Centro Estadual de Educação Tecnológica Paula Souza, 2019.

1. Educação profissional. 2. EaD. 3. Ensino a Distancia. 4. *Big Data*. 5. *Learning Analytics*. I. Langhi, Celi. II. Centro Estadual de Educação Tecnológica Paula Souza. III. Título.

HENRIQUE RUIZ POYATOS NETO

AVALIAÇÃO DE DISCENTES NA MODALIDADE DE ENSINO A DISTÂNCIA EM CURSOS DE EDUCAÇÃO PROFISSIONAL EM NÍVEL SUPERIOR, POR MEIO DE ALGORITMOS PREDITIVOS

_____

Profa. Dra. Celi Langhi

_____

Prof. Dr. Ricardo Sartorello

_____

Profa. Dra. Marilia Macorin de Azevedo

São Paulo, 16 de abril de 2019

Dedico este trabalho a minha esposa Patricia e
meus filhos, que a cada dia que passa me
tornam um ser humano melhor.

AGRADECIMENTOS

A minha esposa Patricia Poyatos e filhos Giuliano e Theo, que me apoiaram com seu amor nos momentos mais difíceis desta jornada.

A minha ex-aluna e agora colega de trabalho Ana Raquel Fernandes Cunha, pelas dicas sobre a linguagem R e consultas aos objetos de aprendizado que tornaram minha tarefa de consolidação das consultas muito mais fácil. A aluna se torna a professora e o professor, o aluno.

Ao amigo e professor de Inteligência Artificial Felipe Teodoro, que apontou os melhores caminhos quando as dúvidas sobre predição surgiram.

A amiga Cristiane Maria Paiva de Melo, que revisou este trabalho e me encorajou durante todo o período de sua realização.

Ao meu chefe e amigo, Leandro Rubim de Freitas, que me apoiou e acreditou em meu potencial nos momentos em que eu duvidava de minhas capacidades.

Aos professores Dra. Marilia Macorin de Azevedo e Dr. Ricardo Sartorello que participaram de minha banca examinadora e contribuíram muito para o enriquecimento deste estudo.

E um especial agradecimento à minha orientadora Celi Langhi que me orientou da melhor forma possível na concepção e realização deste trabalho.

O homem sábio é forte, e o homem de
conhecimento consolida a força
(Livro dos Provérbios)

RESUMO

POYATOS NETO, H. R. **Avaliação de discentes na modalidade de ensino a distância em cursos de educação profissional em nível superior, por meio de algoritmos preditivos.** 96f. Dissertação (Mestrado Profissional em Gestão e Desenvolvimento da Educação Profissional). Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2019.

A oferta de cursos a distância no ensino superior se amplia cada vez mais, assim como o número de alunos que evadem estes cursos. Quarenta por cento das instituições de ensino alegam desconhecer as razões da evasão de seus alunos, o que faz com que sejam necessários processos adequados para uma avaliação precoce da aprendizagem dos alunos, de forma a priorizar aqueles que passam por maior dificuldade, mesmo que não peça ajuda. Assim sendo, o objetivo deste trabalho é empregar algoritmos preditivos utilizados em aprendizagem de máquina assistida e aplicá-los na análise de dados acadêmicos dos alunos, prevendo da forma mais eficaz possível seus desempenhos nas avaliações somativas e possibilitando à instituição de ensino ações preventivas em situações desfavoráveis. A metodologia usada foi a netnografia com uma amostra composta pelas interações realizadas no ambiente virtual de aprendizagem dos alunos de sete cursos de ensino superior totalmente a distância da Faculdade de Informática e Administração Paulista (FIAP). Como resultados observou-se que algoritmos de predição apresentam taxas de acerto diferentes, as amostragens colhidas nos períodos iniciais de curso podem ser mais significativas do que em períodos maiores e indicadores como a exposição do aluno aos objetos de aprendizagem podem ser menos relevantes do que se presume. Sugere-se que as instituições de ensino que façam uso do mesmo ambiente virtual de aprendizagem possam adaptar os algoritmos preditivos utilizados neste estudo e realizem suas próprias predições, auxiliando no processo de ensino-aprendizagem de seus alunos e, por consequência, diminuir seus índices de evasão. O produto gerado por este trabalho é um algoritmo preditivo escrito em Linguagem R para fins de avaliação dos alunos em cursos superiores na modalidade de ensino a distância que utilizam a plataforma Moodle (software intitulado "Sistema para predição de avaliação formativa no ensino a distancia em curso superior" registrado no Instituto Nacional da Propriedade Industrial sob o número BR512019000915-0).

Palavras-chave: Educação Profissional; EaD; Ensino a distância; *Big Data; Learning Analytics*; Avaliação; EDM; Análise preditiva

ABSTRACT

POYATOS NETO, H. R. **Evaluation of distance education students in professional education courses at the higher level, through predictive algorithms**. 96f. Dissertation (Professional Master in Management and Development of Professional Education). Centro Estadual de Educação Tecnológica Paula Souza, São Paulo, 2019.

The supply of distance courses in higher education is increasing, as is the number of students who avoid these courses. Forty percent of the educational institutions claim to be unaware of the reasons for their students' evasion, which means that adequate processes are necessary for an early assessment of student learning to prioritize those who are more difficult, even if they do not ask help. Therefore, the objective of this work is to use predictive algorithms used in assisted machine learning and to apply them in the analysis of the students' academic data, providing the most effective possible performance in summative evaluations and enabling the institution of preventive actions in situations unfavorable. The methodology used was the netnography with a sample made up of the interactions carried out in the virtual learning environment of the students of seven higher education courses totally at a distance from the Faculdade de Informática e Administração Paulista (FIAP). As results, it was observed that prediction algorithms have different hit rates, the samplings collected in the initial periods of the course may be more significant than in larger periods and indicators such as the student's exposure to the learning objects may be less relevant than if presumed It is suggested that educational institutions that make use of the same virtual learning environment can adapt the predictive algorithms used in this study and make their own predictions, helping in the teaching-learning process of their students and, consequently, evasion. The product generated by this work is a predictive algorithm written in R language for the purpose of evaluating students in distance learning courses using the Moodle platform (software titled " Sistema para predição de avaliação formativa no ensino a distancia em curso superior" registered in the Instituto Nacional da Propriedade Industrial, number BR512019000915-0).

Key words: Professional Education. EaD. Distance education. Big data. Learning Analytics. Evaluation. EDM. Predictive analysis.

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS

ABED	Associação Brasileira de Ensino a Distância
AVA	Ambiente virtual de aprendizagem
BI	*Business Intelligence*
EaD	Ensino a distância
EDM	*Educational Data Mining*, ou mineração de dados educacionais
ETL	*Extract, Transform* e *Load* ou extração, transformação e carga
FN	Falso negativo
FP	Falso positivo
HDFS	*Hadoop Filesystem* ou sistema de arquivos Hadoop
INPE	Instituto Nacional de Pesquisas Espaciais
KNN	*K-Nearest Neighbor*
LA	*Learning Analytics*, ou análise de aprendizado
LDB	Diretrizes e Bases da Educação Nacional
SGBD	Sistema gerenciador de banco de dados
SQL	*Structured Query Language* ou linguagem de consulta estruturada
SVM	*Support Vector Machine*
SV	*Support Vector* ou vetor de suporte
TIC	Tecnologias de informação e comunicação

SUMÁRIO

INTRODUÇÃO

Poucas inovações tecnológicas provocaram tantas mudanças na sociedade em tão curto espaço de tempo como as Tecnologias de Informação e Comunicação (TICs), especialmente a Internet e, em decorrência disso, surgiram novas maneiras de se conviver no mundo moderno, renovadas constantemente pela revolução nas mídias e nas suas possibilidades (LÉVY, 2004).

A Internet possibilita vários tipos de aplicações educacionais: de divulgação, de pesquisa, de apoio ao ensino, de comunicação e mais recentemente, de ensino. A divulgação pode ser da instituição de ensino ou de seus participantes, como alunos e professores. Estas partes envolvidas podem realizar pesquisas individualmente ou em grupo, durante a aula ou extraclasse, de forma obrigatória ou livre. Nas atividades de apoio ao ensino, o processo de ensino pode ser enriquecido com textos, imagens, sons do tema específico do programa, acompanhado de livros, revistas e vídeos. A comunicação ocorre entre professores e alunos, entre professores e professores, entre alunos e outros colegas da mesma ou de outras cidades e países, ou seja, as possibilidades e democratização da comunicação é enorme (MORAN, 1997).

Se, no passado, os processos de ensino e aprendizagem se davam majoritariamente de forma presencial, as TICs possibilitaram que estes ocorram parcialmente ou de forma totalmente remota, em uma modalidade conhecida como ensino a distância (EaD). A Lei de Diretrizes e Bases da Educação Nacional (LDB) legalizou o ensino a distância e as tecnologias de informação e comunicação (TIC) modernizaram esta modalidade de ensino, tida anteriormente como atrasada e de segunda classe. A interconectividade que a Internet e as redes desenvolveram começam a revolucionar a forma de ensinar e aprender (MORAN, 2013).

Esta modalidade de ensino está em grande expansão no mundo todo e o Brasil considera o ensino a distância um componente estratégico na busca por diminuir sua defasagem educacional, se comparado a outros países no mundo. A portaria normativa nº 11 de 20 de junho de 2017 e a portaria nº 132 de 5 de junho de 2018 promovem uma flexibilização nos processos de credenciamento de instituições e cursos para a modalidade EaD, maior oferta de cursos lato sensu e a regulamentação da modalidade para o ensino fundamental, outrora restrito ao ensino superior.

Entretanto, seja na modalidade presencial ou a distância, altos índices de evasão escolar são registrados no mundo todo. Segundo o censo 2016 da ABED (2017) realizado com instituições de ensino brasileiras, 32% das instituições que oferecem cursos regulamentados totalmente a distância informam taxas de evasão entre 11% a 25%. É importante salientar que nem todas as instituições conhecem os motivos pela qual seus alunos evadem. Segundo o mesmo censo, apenas 60% das instituições privadas com fins lucrativos e 41% das instituições educacionais públicas federais alegam conhecer os motivos. Presume-se que, por desconhecer as causas, poucas ações efetivas estejam sendo tomadas para mitigar esses números.

O ensino a distância não oferece um ensino centrado no professor e sim pautado na capacidade do aluno em superar-se e buscar/construir seu próprio conhecimento e essa autonomia requerida pelo EaD é um dos maiores desafios dessa modalidade educacional (IVASHITA; COELHO, 2009). Em um estudo realizado por Gilberto (2014), os sujeitos apontam dificuldades como um aprendizado solitário e a complexidade da comunicação escrita nos processos de interação. Aliado à uma comunicação com o tutor assíncrona e não imediata, presume-se que o discente não se sente estimulado a buscar auxílio em momentos de dificuldade e frustração. Ao observar os números de evasão da pesquisa da ABED (2017), o desconhecimento de parte das instituições nas causas da evasão e os desafios e dificuldades apontadas pelos alunos no estudo de Gilberto (2014), pode-se concluir que a falta de visibilidade do aluno na modalidade EaD é um problema a ser enfrentado.

No entanto, o aumento de softwares educativos instrumentados como ambientes virtuais de aprendizagem (AVA) e bancos de dados com pontuações de testes dos alunos criaram grandes repositórios de dados que podem ser utilizados para identificar desinteresse e dificuldades no aprendizado por parte do aluno, resultando em uma eventual evasão. A Mineração de Dados Educacionais (EDM) é um campo interdisciplinar que reúne pesquisadores da ciência da computação, educação, psicologia, psicometria e estatística para analisar grandes conjuntos de dados e para abordar essas importantes questões educacionais (NASIRI; MINAEI; VAFAEI, 2012).

Usando recursos extraídos de dados de registro e marcas obtidas nas avaliações formativas, alguns pesquisadores (MERCERON; YACEF, 2005) usam técnicas de classificação para prever o desempenho do aluno com bastante precisão, permitindo a tutores identificar os alunos em risco e oferecer suporte antes de avaliação somativa.

Segundo Picciano (2012), é possível prever, logo após a primeira semana de um curso e com 70% de precisão, se algum aluno concluirá o curso com sucesso. Um segundo modelo poderia ser gerado com atualizações semanais usando fatores preditivos semelhantes. Um instrutor pode revisar o envolvimento do aluno a qualquer momento durante o curso e os dados no PACE (a performance do aluno, semelhante à medida de um corredor em uma maratona) são mantidos em tempo real.

É a partir dessa constatação que surge a seguinte questão de pesquisa: é possível empregar algoritmos preditivos para prever o desempenho dos alunos na avaliação somativa, a partir de seu comportamento no AVA de uma determinada instituição de ensino logo nos primeiros meses de curso? Quais dados são relevantes para a predição e quais procedimentos trarão o melhor resultado para esta aplicação em específico?

Uma das hipóteses é que algoritmos de classificação mais robustos trarão percentuais de acuidade maiores que técnicas mais simples. Outra hipótese é que características como a idade e gênero são relevantes na predição, pois é possível observar comportamentos de estudo distintos entre gêneros e faixas etárias diferentes. Além disso, presume-se que a acuidade do algoritmo de predição aumentará na medida em que novos meses de análise são acrescidos no algoritmo de aprendizagem de máquina.

Assim sendo, o objetivo geral deste trabalho é propor um modelo de predição precoce que mapeie o desempenho dos alunos em cursos de nível superior, na modalidade de ensino a distância, prevendo seu desempenho na avaliação somativa a partir de seus dados acadêmicos logo nos primeiros meses de curso. Os objetivos específicos são: analisar quais dados são relevantes para a predição, e qual algoritmo de classificação trará os melhores resultados na previsão do resultado da avaliação somativa dos alunos.

Para atingir este objetivo, o trabalho foi organizado em quatro capítulos, além da introdução e considerações finais.

No primeiro capítulo de fundamentação teórica é realizado um breve histórico sobre o Ensino a Distância (EaD) no Brasil e como esta modalidade de ensino e aprendizagem foi potencializada com o advento das tecnologias de informação e comunicação. Adicionalmente, são apresentadas as suas características, oferta de cursos e demanda pelos alunos em território nacional, vantagens, desvantagens e o porquê da implementação por si só não representar uma revolução tecnológica, devendo ser bem planejada pelas instituições e o quanto seu sucesso

depende de um corpo docente qualificado e treinado para lidar com o novo paradigma educacional que se apresenta.

O subitem “Ambientes virtuais de aprendizagem” apresenta informações sobre esses sistemas computacionais utilizados para a mediação das comunicações entre instituições de ensino, docentes e discentes, que permitem a transmissão de informação de forma multimídia e possibilita aos alunos a construção de um conhecimento de forma assíncrona. Apresenta-se quais ferramentas podem compor estes AVAs, suas possibilidades e limitações.

No subitem seguinte, “Materiais didáticos para EaD” destaca-se a importância do processo de criação de um material didático específico para ensino a distância que permita ao educando percorrer diferentes trilhas de aprendizagem, explore as múltiplas mídias disponíveis e que esteja alinhado com a flexibilidade que a modalidade proporciona. O material deve promover o desenvolvimento das habilidades metacognitivas e facilitar o aprendizado cooperativo. No entanto, cabe à avaliação mensurar se o material didático está cumprindo seu objetivo.

No subitem intitulado “Avaliações e sua importância no EaD” é apresentado um breve histórico com o intuito de definir o que é avaliação, seus objetivos e pontuar as diferenças entre o que é uma avaliação e o que é um exame. Além disso, apresenta-se como as avaliações podem acontecer em um ambiente virtual de aprendizagem e sua importância como instrumento de diagnóstico e acompanhamento da performance de alunos, aderência de currículos e eficácia das instituições.

O segundo capítulo aborda **o método da pesquisa**, começando pelo subitem “Netnografia” que se destina a introduzir no estudo este neologismo recente, a ramificação da etnologia que se apresenta como um método de estudo usado para descrever costumes e tradições de um determinado grupo de pessoas, adicionando ou restringindo a coleta de dados as comunicações e ações dos indivíduos que aconteçam pela Internet, mediadas por um computador ou dispositivo eletrônico contemporâneo, como um *tablet* ou *smartphone*. Apresenta-se nesse subitem as limitações inerentes a este método de estudo que explicam, em parte, a relutância de pesquisadores e antropólogos ao considerar a netnografia em seus estudos. Assim como existem limitações, apresentam-se também as vantagens que este tipo de estudo pode trazer e o porquê a coleta de dados a partir de comunidades digitais é altamente recomendada, a depender do estudo que precise ser realizado.

O subitem seguinte, intitulado "*Business Intelligence e Big Data*" tem como objetivo apresentar como vários dos conceitos presentes da netnografia já são usados há décadas pela iniciativa privada e como os dados têm sido usados para impulsionar a venda de produtos ou serviços e melhorar a tomada de decisão de grandes empresas. Quanto maior for a capacidade de coleta, maior deverá ser a capacidade de análise destes dados e melhores decisões de negócio são tomadas, justificando assim o alto investimento.

No entanto, não são apenas setores como indústria e comércio que podem se beneficiar de uma massiva coleta e análise de dados para melhorar suas vendas e prestações de serviços. O subitem "*Education Data Mining* (EDM) e *Learning Analytics* (LA)" apresenta dois ramos de pesquisa recentes oriundos de *Business Intelligence* e *Big Data* que aplicam de maneira adequada as ferramentas, processo de coleta e análise para a realidade acadêmica. Adicionalmente, descreve-se neste subitem as vantagens que podem ser alcançadas e cuidados a serem tomados.

Posteriormente, apresenta-se uma breve explicação sobre aprendizagem de máquina assistida e os algoritmos de classificação utilizados, que são introduzidos dos mais simples aos mais robustos: árvore de decisão, *Naïve Bayes, k-Nearest Neighbor e Support Vector Machine*. Os resultados serão apresentados em matrizes de confusão que foram brevemente explicadas, concluindo assim a fundamentação teórica deste trabalho.

CAPÍTULO 1 – ENSINO A DISTÂNCIA

Ensino a distância (EaD) é um dos termos utilizados para definir a modalidade educacional em que discentes e docentes interagem com o processo de ensino e aprendizagem geograficamente distantes e em tempos diferentes. Esta modalidade educacional começou a ser utilizada no Brasil no final do século XIX por meio da correspondência postal, da qual o material didático era enviado pelo correio de uma instituição de ensino ao aluno. O primeiro curso de que se tem notícia foi um curso profissionalizante de datilografia divulgado na primeira edição do Jornal do Brasil em 1891 (LANGHI, 2005).

Nas décadas iniciais do século XX, a difusão da mídia radiofônica prolifera os cursos de eletrônica por correspondência e o meio passa a ser utilizado para interação entre as partes envolvidas no EaD especialmente entre 1940 e 1960, possibilitando a instituição de ensino transmitir através de ondas eletromagnéticas de longa distância o mesmo conhecimento a vários receptores simultaneamente. Projetos como o Minerva, transmitido pela Rádio MEC e o Sistema Avançado de Comunicações Interdisciplinares (Projeto Saci), iniciativa do Instituto Nacional de Pesquisas Espaciais (INPE), permitiu a milhares de pessoas realizarem seus estudos (SARAIVA, 1996).

Na década de 1960 o avanço tecnológico e sua adoção massiva possibilita a inclusão de uma nova mídia a esta modalidade educacional: o audiovisual, transmitido através de ondas eletromagnéticas de longa distância assim como o rádio, a mídia televisiva enriquece o material didático transmitido aos discentes não apenas o áudio presente nas transmissões radiofônicas, mas imagens em movimento (LANGHI, 2005).

O avanço e massificação das tecnologias de informação e comunicação (TIC), e em especial a Internet, tem transformado o modo em que a comunicação entre as pessoas acontece, provocando uma revolução silenciosa na sociedade, modificando relações de trabalho e a forma que fazemos negócio. Além disso, as TIC têm promovido transformações na área educacional com novas possibilidades no ensino e aprendizagem ao propiciar à modalidade de Ensino a Distância, que combinam os já conhecidos recursos educacionais, à agregação de modernas ferramentas tecnológicas de informação e comunicação. É neste contexto que o ensino a distância surge como uma das mais importantes formas de difundir a educação e conhecimento (MAIA, 2004).

Essa estrutura provida pelas tecnologias de informação e comunicação reavivou as práticas de EaD, pois a modalidade proporciona maior flexibilidade do tempo, a quebra de

barreiras espaciais e o envio e recebimento praticamente instantâneo de materiais didáticos e atividades. O Ensino a Distância suportado pelas TICs permite não apenas transmitir conteúdos digitalizados e de forma multimídia, mas tem como potencial a interatividade das partes interessadas ao desenvolver atividades a distância com base na interação e na produção de conhecimento (ALMEIDA, 2003).

Entretanto, Mendonça (2013) observa que o Ensino a Distância depende das TICs para encurtar as diferenças de espaço e tempo (CORREIA; SANTOS, 2013). Isso pode ser observado na definição atual de Ensino a distância estabelecida pelo Estado brasileiro, que o define como:

> [...] modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorra com a utilização de meios e tecnologias de informação e comunicação, com pessoal qualificado, com políticas de acesso, com acompanhamento e avaliação compatíveis, entre outros, e desenvolva atividades educativas por estudantes e profissionais da educação que estejam em lugares e tempos diversos (Decreto Federal nº 9.057, de 25 de maio de 2017, que regulamenta o artigo 80 da Lei nº 9.394, de 20 de dezembro de 1996, artigo 1º2017).

Assim sendo, conforme salientado por Maia e Meirelles (2003), o EaD consiste nos seguintes elementos centrais: a separação física entre o professor e o aluno; na utilização das TICs na transmissão de conteúdos didáticos de professores aos seus alunos; em uma interação professor aluno bidirecional, que possibilita ao aluno iniciar o processo de comunicação e na realização de encontros presenciais que possam consolidar conhecimentos e habilidades desenvolvidas a distância e que promovam a socialização das partes envolvidas.

As tecnologias da informação e comunicação (TICs) possuem três elementos a se destacar: o computador ou equipamento similar, que realiza os cálculos e operações lógicas com rapidez e confiabilidade; a comunicação, que é essencial à condição humana e, por meio de, ocorrem os processos de transmissão e recepção da informação; e a automação promovida pelos sistemas da informação, e é por esta razão que as TICs se tornaram o elo fundamental do Ensino a Distância pois, em sua ausência, não é possível acontecer a interação entre professor e aluno (RAMOS, 2008; CORREIA; SANTOS, 2013).

Ao observar o EaD sob uma perspectiva de interação e construção colaborativa de conhecimento, seus ambientes digitais favorecem o desenvolvimento de competências e habilidades relacionadas com a escrita com o intuito de expressar o próprio pensamento, interpretação de textos, sejam estes elaborados por um conteudista ou na representação do

pensamento do aprendiz e a comunicação de suas ideias, contribuindo para a produção individual e coletiva de conhecimentos. Além disso, deve-se destacar o potencial da modalidade ao possibilitar ao discente a oportunidade de avaliar continuamente o próprio trabalho individualmente, efetuando reformulações praticamente instantâneas e produzindo assim novos saberes, assim como participar de produções coletivamente ao analisar criticamente, emitir *feedback* ou espelhar-se nas produções de outros discentes. Torna-se fundamental favorecer a interação entre os participantes e a representação do pensamento do aprendiz. Destaca-se como vantagem adicional no uso das TICs em EaD a sua contribuição no registro contínuo das produções, interações e caminhos percorridos pelos discentes, o que possibilita uma avaliação processual que permite a rápida recuperação de qualquer etapa do processo, analisá-la e promover mudanças contínuas, visando ao melhor desenvolvimento do aprendiz. Assim sendo, diferentemente da modalidade presencial, mesmo após a conclusão das interações, é possível recuperar todas elas, rever todo o processo e refazer as análises mais pertinentes em termos de avaliação (ALMEIDA, 2003).

As tecnologias da informação e comunicação estão criando maneiras diferentes de se distribuir socialmente o conhecimento e é por esta razão que se torna necessário discutir novas formas de alfabetização: literária, gráfica, informática, científica, entre outras (POZO, 2004).

O Ensino a Distância se torna um sistema tecnológico de comunicação em massa de mão-dupla, composto de diversos recursos didáticos e de apoio tutorial que propiciam a aprendizagem autônoma do aluno e substituem a integração presencial entre professor e aluno (ARETIO, 1990). Suas vantagens ao quebrar as barreiras do espaço e tempo e sua capacidade ao democratizar o conhecimento ao ofertar cursos em regiões do país em que tais cursos não seriam oferecidos no modelo tradicional, torna esta modalidade educacional uma possibilidade atraente, seja para a educação formal, como na capacitação de profissionais no mercado de trabalho.

Segundo a ABED (2017) eram 3.734.887 alunos estudando na modalidade EAD em 2016, sendo 561.667 alunos em cursos regulares totalmente a distância, 217.175 em cursos regulamentados semipresenciais, 1.675.131 em cursos livres não corporativos e 1.280.914 em cursos livres corporativos.

Há uma grande oferta de cursos na modalidade EAD. De acordo com a ABED (2017), foram 2369 cursos disponíveis nessa modalidade no ano de 2016, destacando-se os 1098 cursos superiores de pós-graduação do tipo *lato sensu* ou especialização, 219 cursos técnicos

profissionalizantes, 235 cursos superiores de tecnologia, 210 cursos superiores de graduação em licenciatura e 142 cursos superiores de bacharelado.

Esta demanda por ser compreendida pelas vantagens que o EaD traz para o aprendiz. Trata-se de uma modalidade educacional que permite ao aluno a administração de seu próprio tempo, dando-lhe autonomia para realizar as atividades propostas no momento em que considera mais apropriado desde que, no entanto, sejam respeitados os prazos constituídos pelo próprio andamento do curso (ALMEIDA, 2003).

Em um tempo presente permeado por intervenções tecnológicas, as TICs têm sido ferramentas de transformação social, promovendo novos conceitos de interação social e maior liberdade à organização social (LEVY, 1999; CORREIA; SANTOS, 2013). Entretanto, indivíduos que não se atualizarem e não se tornarem parte deste processo inovador se tornarão subitamente "analfabetos tecnológicos", tornando-se inaptos às suas ocupações profissionais originais ou novas ocupações que surgirão, pois ambas irão requerer maior qualificação do que as exigidas anteriormente (MAIA, 2004). Se em um primeiro momento as interações tecnológicas exigidas pela modalidade EaD, por meio das TICs, possam parecer uma desvantagem desta modalidade educacional, ao exercitá-las os indivíduos ficarão cada vez mais distantes do "analfabetismo tecnológico".

Além disso, o grande trunfo da tecnologia não é fazer antigos processos funcionarem mais rápido ou melhor, mas ao permitir que as organizações desfaçam antigos processos e regras, possibilita a elas criar novas formas de trabalho, baseados no desenvolvimento de processos orientados a cooperação, sustentados pela convivência-consciência-transcendência (FRANCIOSI; MEDEIROS; COLLA, 2003). Se for bem planejado, um curso ministrado na modalidade EaD não apenas prepara seus discentes para esta nova lógica de trabalho como representa por si só o rompimento dos antigos processos e regras, pois um curso planejado nesta modalidade não pode ser simplesmente uma transposição do material didático e processos educacionais tradicionais para o meio digital.

No entanto, o emprego de uma determinada tecnologia como suporte à modalidade EaD não constitui em uma revolução metodológica, mas a torna possível. Desta forma, pode tanto ser usada na tentativa de simular a educação presencial quanto fazer uso de uma nova mídia e oferecer novas possibilidades de aprendizagem, ao explorar características inerentes à tecnologia (PERAYA, 2002; ALMEIDA, 2003). A utilização de TICs não garante a constituição de ambientes de aprendizagem, sendo necessário o planejamento de um ambiente idealizado que busque uma sinergia entre visão-e-ação, teoria-e-prática, sonho-e-realidade,

todo-e-parte, individual-e-coletivo, cada um-e-consigo mesmo (FRANCIOSI; MEDEIROS; COLLA, 2003).

A instituição que possui cursos na modalidade EaD, que se restringe a distribuir informações de maneira uniforme a todos seus alunos que estudam os conceitos recebidos, realizam os exercícios e os submetem de volta à instituição que, por sua vez, libera novos módulos do curso pode enfrentar altos índices de desistência. Aliando-se à dispersão geográfica dos outros alunos e a separação física do professor, pode-se gerar uma sensação de isolamento que desmotiva o discente que, consequentemente, deixa de participar das atividades on-line (ALMEIDA, 2003; TORRES; MARRIOTT, 2006).

Assim sendo, faz-se necessário desenvolver uma linguagem pedagógica apropriada para o processo de ensino/aprendizagem por meio das TICs. Deve-se manter os alunos motivados, incentivar sua interação com seus pares e tutores e as avaliações devem ser constantes, apropriadas ao conteúdo abordado e dentro das capacidades do aluno, e devem ser repensadas sempre que verificando que não estão atingindo os objetivos estabelecidos previamente (MAIA; MEIRELLES, 2003; CORREIA; SANTOS, 2013).

No entanto, estes não são os únicos desafios que a modalidade de ensino enfrenta. Para Pozo (2008), um dos principais desafios é a capacidade em que uma instituição de ensino é capaz de converter informações em conhecimento e, para tal, não basta transpor a maneira em que o aprendizado presencial acontece para a o aprendizado a distância, pois as TICs estão incontestavelmente mais presente na vida das pessoas. Sob esta perspectiva, faz-se necessário repensar até mesmo a forma em que a aprendizagem presencial acontece, uma vez que as novas tecnologias de informação e comunicação transformaram profundamente a forma e velocidade em que as pessoas têm acesso à informação.

Além da apropriação das TICs, Perry e cols (2006) defendem o papel dos tutores fornecendo *feedbacks* constantes proporcionando à IES a oportunidade de compreender adequadamente como pode transformar o processo de transmissão de informações em construção de conhecimento, desenvolvendo assim estratégicas didáticas e pedagógicas condizentes com o ensino a distância. Azzolino e Nabarretti (2008) corroboram a defesa ao afirmar que as instituições precisam desenvolver métodos, técnicas e educadores capazes de preparar seus discentes a absorver e filtrar o grande volume de informações as quais são nos dias de hoje expostos. Além disso, os autores ressaltam a importância das instituições de ensino acompanharem o surgimento de novas tecnologias de informação e comunicação que sejam aplicáveis ao ensino a distância aprimorando e readequando seus recursos humanos de

forma a estarem sempre preparados a cumprir seu papel neste novo paradigma educacional. Mercado e cols (2008) afirmam que o professor no EaD é o especialista que planeja o curso, produz e garante a qualidade do material didático que será utilizado e, por sua vez e de forma síncrona ou assíncrona, presencial ou a distância, cabe ao tutor facilitar a comunicação de forma a efetivar o referido material, conduzindo, acompanhando e avaliando a aprendizagem dos alunos.

Por fim, é possível concluir que os desafios na gestão das TICs no EaD são: (a) potencializar o uso das TICs de forma que possam enriquecer a experiência e facilitar o processo de ensino e aprendizagem, ao disponibilizar ambientes atrativos com interfaces de fácil manuseio, repletas de recursos para aprendizagem individual e em grupo e que, assim, possam prender a atenção do aluno e (b) capacitar professores e alunos na utilização das TICs, aproveitando assim todo o seu potencial (VIEIRA, 2011; FERNANDES; FERNANDES, 2012; CORREIA; SANTOS, 2013) sendo o foro desta integração os ambientes virtuais de aprendizagem.

1.1 Ambientes virtuais de aprendizagem

Tratam-se de sistemas computacionais disponíveis na internet, destinados ao suporte de atividades de ensino e aprendizagem mediadas pelas tecnologias de informação e comunicação que permitem transmitir informações por meio de múltiplas mídias (hipertexto, imagens, áudio e vídeo) de uma maneira organizada, permitindo a construção de novos conhecimentos de maneira assíncrona, ou seja, sem que professor e aluno estejam disponíveis no mesmo espaço e tempo (ALMEIDA, 2003).

Tais ambientes digitais de aprendizagem permitem desenvolver interações entre pessoas e objetos de conhecimento, elaborar e socializar produções com o objetivo de atingir determinados objetivos educacionais. Os recursos à disposição são tradicionalmente ofertados pela internet, como correio eletrônico, fórum, bate-papo, conferência, banco de recursos e outros, porém propiciam à instituição uma gestão de informação seguindo critérios previamente estabelecidos de acordo com suas necessidades e características de cada software (ALMEIDA, 2003). Franciosi e colaboradores (2003) destacam os mecanismos como uma Biblioteca Digital, mural de avisos, publicação de tarefas, pré-testes com feedback imediato, vários deles atuando como meios para uma autoria coletiva. Essas ferramentas permitem mediatizar a comunicação entre emissor-receptor permitindo trocas sociais mais efetivas ao utilizar como meios as linguagens escrita, oral, imagens ou vídeos (ALMEIDA, 2003). Assim

sendo, Kenski (2003, p. 101) alerta que "para se realizar ensino a distância mediado por novas tecnologias, é preciso contar com uma infraestrutura organizacional (técnica, pedagógica e administrativa) complexa, na qual o ensino será desenvolvido."

AVAs são como salas de aula online, compostas de interfaces e ferramentas disponibilizadas com o intuito de promover aprendizagem ao construir novos conhecimentos e habilidades, ao disponibilizar materiais didáticos e atividades propostas pelo professor, permitindo interações entre alunos e professor sejam estas individualmente ou colaborativamente (SILVA, 2011; BARBOSA, 2012).

Ao oferecer ao aluno possibilidades de acessar em tempo e espaço diferenciados as atividades do curso, fazer contatos, interagir, estudar de forma autônoma ou com a orientação do tutor por métodos e estratégias que dimensionem a busca do conhecimento, o AVA proporciona uma socialização entre as partes envolvidas, combinando a flexibilidade da interação humana, a independência no tempo e no espaço sem, no entanto, perder velocidade. As atividades se desenvolvem no tempo, ritmo de trabalho e espaço em que cada participante se localiza, de acordo com um planejamento prévio denominado design educacional (CAMPOS; ROCHA, 1998; BELLONI, 2009; BARBOSA, 2012).

Além do ensino a distância promovido em ambientes digitais e interativos de aprendizagem romper as distâncias espaço-temporais, ela viabiliza a recursividade, múltiplas conexões, trajetórias e interferências, não se restringindo aos materiais didáticos e atividades definidas previamente. Para Moraes (1997, p. 68), o EaD pode ser entendido como um sistema aberto "com mecanismos de participação e descentralização flexíveis, com regras de controle discutidas pela comunidade e decisões tomadas por grupos interdisciplinares" (ALMEIDA, 2003).

Sendo assim, os ambientes digitais de aprendizagem podem ser utilizados como suporte para sistemas de ensino parcialmente ou exclusivamente *on-line*. No caso de sistemas semipresenciais, os AVAs podem expandir as interações da sala de aula para além do espaço-tempo disponível para o encontro face a face, provendo suporte tanto para ações presenciais como atividades a distância (ALMEIDA, 2003). O principal objetivo ao se utilizar os AVAs é maximizar as possibilidades de interação das partes interessadas e desenvolver ações compartilhadas, em que todos exercem os papéis de professor e aluno (FRANCIOSI; MEDEIROS; COLLA, 2003).

No entanto, para que um ambiente virtual de aprendizagem realize uma intermediação eficiente na comunicação síncrona ou assíncrona entre as partes interessadas neste processo educativo *on-line*, é necessário que estas se disponham a realizar comunicações mais ricas e socialmente compartilhadas, afinal, o conhecimento elaborado de forma solidária pode resultar em uma aprendizagem mais significativa e um maior desenvolvimento cognitivo (VALENTE, 2003; BARBOSA, 2012).

Ao promover uma interação maior de seus participantes por meio de um AVA, incorporam-se relações socioculturais nos processos de construção do conhecimento, alinhando-se assim a teoria do socioconstrutivismo. Desta maneira, provê-se ao aprendiz um arcabouço teórico que o permita resolver problemas de maneira mais eficiente, em um modelo educacional mais aderente às reivindicações da sociedade contemporânea. A comunicação, o diálogo e a colaboração se tornam aspectos importantes nas estratégias pedagógicas para desenvolver o ensino/aprendizagem (D'ÁVILA, 2006; BARBOSA, 2012).

A interação entre as partes interessadas acontece no ambiente virtual de aprendizagem por meio dos contatos via e-mail, participações nos *chats* e fóruns, quando algum tema é enviado para discussão. Os *chats*, sejam estes por texto ou videoconferências, promovem uma maior integração entre docentes e discentes e são as ferramentas que mais favorecem "o estar junto virtual" (ANDRADE; VICARI, 2011; BARBOSA, 2012).

O potencial interativo do uso das TIC permite desenvolver contatos e atividades por meio do diálogo participativo que possibilitam o "estar junto virtual", que pode ser estabelecido por meio de múltiplas interações que visam o constante acompanhamento e assessoria desse aprendiz, determinando seu potencial e, assim, propor desafios acadêmicos que o motivem e atribuam significado ao seu aprendizado. Tais desafios levam o aprendiz a aplicar seu conhecimento, a transformá-lo e a buscar outras informações, construindo assim novos conhecimentos (VALENTE, 2003; ANDRADE; VICARI, 2011; BARBOSA, 2012).

Cada recurso mediático utilizado no EaD contém suas próprias características estruturais e diferentes níveis de diálogos limitados pela própria mídia, assim como em um ambiente de sala de aula o nível de diálogo e participação dos alunos depende da abordagem pedagógica, estratégicas e mediações promovidas pelo professor (ALMEIDA, 2003). Seu papel deve ser centrado em usar suas habilidades para motivar, despertar o interesse e apoiar seus alunos, bem como preparar o ambiente e organizar os materiais didáticos envolvidos no processo. Os alunos deixam de ser receptores passivos de informação e se tornam construtores

e socializadores do conhecimento (VALENTE, 2003; ANDRADE; VICARI, 2011; BARBOSA, 2012).

Como um orientador do aluno que o acompanha em seu desenvolvimento no curso, provocando-o para fazê-lo refletir, compreendendo seus equívocos e depurando suas produções, o tutor não deve, por sua vez, realizar um plantão integral, devendo apenas se fazer presente em determinados momentos para acompanhar o aluno. O tutor não deve entrar em embate ou controlar seu desempenho, caso contrário criará uma dependência que não o permitirá a autonomia desejada, perpetuando a hierarquia das relações aluno-professor do ensino tradicional, que já se mostraram inadequadas e ineficientes (ALMEIDA, 2003).

O professor deixa de ser um transferidor de conhecimentos prontos e se torna um formulador de problemas e um provocador de situações, sistematizando a experiência do aluno, arquitetando sua trilha de conhecimentos, um agente da construção de seu conhecimento (BARBOSA, 2012).

Atuar neste ambiente virtual significa expressar pensamentos, dialogar, trocar informações e experiência, tomar decisões e, assim, produzir conhecimento. Por estão razão, o EaD não deve ser considerado uma solução paliativa para atender alunos que estejam geograficamente longe das instituições educacionais ou uma simples transposição de conteúdos e métodos do ensino tradicional para um outro meio, pois curso ministrados na modalidade EAD podem ter um nível de interação entre os participantes priorizado. Contudo, em alguns ambientes virtuais, a dinâmica de atividades pedagógicas desperdiça o potencial de comunicação e interatividade que as TIC são capazes de possibilitar ao aluno para que este se sinta socialmente integrado e assim, desenvolver novas competências cognitivas (BARBOSA, 2012).

Com base na teoria socioconstrutivista, é possível afirmar que a medição entre as partes envolvidas é essencial no processo de ensino/aprendizagem e permitem compreender o papel das tecnologias digitais como artefatos estruturados que direcionam a prática educativa do docente, que devem incentivar a participação dos discentes nos espaços virtuais promovidos pelo AVA, buscando-se assim uma aprendizagem afetiva e colaborativa (BARBOSA, 2012).

Conforme defendido por Maturana (1990), a cooperação é essencial na vida humana cotidiana por estar fundamentada na confiança e no respeito mútuo; sempre que se faz presente, a cooperação recupera o senso de participação e responsabilidade, provê consciência

de ser parte-e-todo e contribui para a transformação de barreiras em pontes e de adversários em aliados, características fundamentais nas relações sociais e, por consequência, relações profissionais. Assim sendo, oportunidades pela qual a cooperação possa ser exercitada serão positivas, cabendo aos professores e instituição elaborar atividades com este objetivo, uma vez que os ambientes virtuais de aprendizagem possuem ferramentas para suportá-las. O material didático, por sua vez, deve fornecer o suporte ao processo de conhecimento que levará os alunos a realizá-las.

1.2 Materiais didáticos para EaD

O ensino a distância (EaD) constitui "parte de um processo de inovação educacional mais amplo que é a integração das novas tecnologias de informação e comunicação nos processos educacionais" (BELLONI, 2002, p. 123), uma modalidade de aprendizagem mais flexível, apoiada nos pressupostos de autonomia individual e liberdade intelectual (LAGUARDIA; PORTELA; VASCONCELLOS, 2007).

É importante destacar a possibilidade provida pelo ensino a distância em que cada estudante pode percorrer caminhos distintos de aprendizagem, ao ter a possibilidade de traçar caminhos diferentes, nas diversas conexões existentes entre os hipertextos e imagens e ao vincular mídias e recursos ao seu repertório; pode criar nós e conexões. O aprendiz se torna receptor, mas também emissor de informações, é leitor, escritor e comunicador (ALMEIDA, 2003).

A integração entre a tecnologia digital com os recursos da telecomunicação evidenciou possibilidades de ampliar o acesso à educação, embora esse uso por si só não implique práticas mais inovadoras e não represente mudanças nas concepções de conhecimento, ensino e aprendizagem (ALMEIDA, 2003). Assim sendo, utilizar as tecnologias de informação e comunicação como suporte ao ensino a distância apenas para colocar o aluno diante de informações, problemas e objetos de conhecimento podem não ser suficientes para envolvê-lo e engajá-lo no processo de aprendizagem. Faz-se necessário criar um ambiente que favoreça a aprendizagem significativa ao aluno, despertando sua disposição para o aprendizado ao disponibilizar as informações pertinentes de maneira organizada e no momento apropriado, promovendo a interiorização de conceitos construídos (ALMEIDA, 2003; POZO, 2004).

Com o advento das novas tecnologias da informação promoveu-se uma democratização não apenas em como a informação é recepcionada, mas também na geração

de novas informações. Os meios de comunicação e instituições de ensino deixaram de ser os únicos emissores, papel que hoje também é realizado por indivíduos e, por esta razão, as informações são muito mais voláteis e sua veracidade é variada. Por esta razão, cabe à escola o papel de formar os alunos para terem acesso e darem sentido à informação, proporcionando-lhes capacidades de aprendizagem que lhes permitam uma assimilação crítica da informação (POZO; POSTIGO, 2000).

Portanto uma das metas essenciais da educação, para poder atender às exigências dessa nova sociedade da aprendizagem é fomentar nos alunos capacidades de gestão do conhecimento, já que, para além da aquisição de conhecimentos pontuais concretos, pois esta habilidade é essencial para ajudá-los a enfrentar as tarefas e os desafios que os aguardam na sociedade do conhecimento. Faz-se necessário capacitar os alunos para a aquisição de informação, sua interpretação e análise, a compreensão e comunicação da mesma.

Para que seja possível averiguar se o ambiente virtual de aprendizagem está cumprindo o seu objetivo, emprega-se como ferramenta a avaliação.

1.3 Avaliações e sua importância no EaD

Avaliação é uma tarefa didática necessária do trabalho docente que cumpre funções pedagógico-didáticas, realizada ao decorrer do trabalho conjunto do professor e dos alunos cujos resultados obtidos são comparados com os objetivos inicialmente propostos, cumprindo assim uma função diagnóstica. Além disso, é possível constatar progressos, dificuldades, e reorientar o trabalho para as correções necessárias, coletados a partir de várias manifestações de situações didáticas, como análise de provas, exercícios, respostas dadas pelos alunos, realização de tarefas, entre outros, proporcionando dados que podem auxiliar o professor a tomar decisões sobre seu trabalho docente (LIBANEO, 2017).

Primeiramente, faz-se necessário compreender a diferença entre exame e avaliação. Segundo Luckesi (2014), o exame se restringe a classificar o educando em uma escala de notas que varia de 0 (zero) a 10 (dez) ou atribuição semelhante sem entrar no mérito da qualidade do aprendizado atingido pelo discente, restringindo-se a separá-los entre aqueles que aprenderam dos que não aprenderam.

Desta forma, o ato de examinar averigua o passado e o que foi aprendido ignorando assim o que não foi aprendido. Por esta razão, em geral o educando se concentra em “tirar nota” e não no seu aprendizado em si. Examinar se restringir em classificar o aluno por sua

capacidade de responder e demonstrar conhecimento e condicioná-lo a responder mediante a uma recompensa representada pela nota alta em uma escala, e não no ato de aprender em si (POYATOS et al., 2018).

Por outro lado, Luckesi (2014) afirma que a avaliação promove uma investigação sobre o desempenho do estudante que, em um caso de não-aprendizado, seja reorientado até que adquira o conhecimento ou habilidade considerada essencial, tendo portanto um caráter diagnóstico; se o ato de examinar está voltado para o passado e na medição do desempenho escolar do aluno, o ato de avaliar se apresenta como uma investigação sobre o seu desempenho escolar, ao averiguar o que o aluno aprendeu ou não, permitindo ao professor e o estudante perspectivas futuras e o planejamento dos próximos passos na direção de abordar o que não foi aprendido.

Além disso, a avaliação pode ser utilizada para diagnosticar e medir outros desempenhos além do aluno, como a investigação do desempenho de uma turma de alunos e até mesmo das instituições de ensino, além do cumprimento de currículos escolares. Conforme afirma Depresbiteris e Tavares (2017), Ralph W. Tyler provocou um grande impacto na literatura das avaliações ao propor atividades dotadas de diversos instrumentos, como escalas de atitude, inventários, questionários, fichas de registro de observação e outros com o intuito de coletar evidências para a avaliação a qualidade dos currículos.

É importante destacar a contribuição de Michael Scriven (2007) à literatura das avaliações ao dividir o ato de avaliar em dois instrumentos diferentes, a **avaliação formativa** e **avaliação somativa**.

A avaliação formativa deve ocorrer continuamente, ao aplicar várias avaliações ao longo do desenvolvimento de todo o programa educacional com o intuito de proporcionar informações úteis ao avaliado identificar e conscientizar-se de seu nível de desempenho e, assim, poder evoluir para um nível que lhe assegure uma avaliação final positiva, além de proporcionar a discentes e instituição dados que permitam o aprimoramento das ações educacionais, como auxiliar os alunos a evoluir seu aprendizado ou pela continuidade ou não de um programa educacional (PARREIRA; SILVA, 2015; DEPRESBITERIS; TAVARES, 2017). Por esta razão, Depresbiteris e Tavares (2017) defendem que o professor deve evitar qualificar o aluno em um único momento, pois a avaliação formativa permite o acompanhamento do processo educacional e estimula a aprendizagem por meio de contínua retroalimentação oral e escrita, questionamentos, reflexões, palavras de alento e de reconhecimento.

Por outro lado, a avaliação somativa é um produto de um processo responsável por determinar mérito, valor ou significado, capaz de situar o aluno em uma escala de capacidade, competências ou conhecimento, averiguando seu mérito relativo. A lógica geral da avaliação estabelecida por Scriven (2007) integra quatro passos fundamentais, sendo: estabelecer critérios de mérito; construir padrões de comparação; medir o desempenho e compará-lo com os padrões; e integrar os dados num juízo sobre o mérito ou valor (SCRIVEN, 2007; PARREIRA; SILVA, 2015).

Para Benson (2003), os métodos utilizados em uma avaliação em AVAs podem estar relacionados à modalidade de avaliação. A recomendação de Laguardia e colaboradores (2007) para uma avaliação diagnóstica que anteceda a formação e permita um ajuste no programa nos conhecimentos ministradores e habilidades a serem desenvolvidas pode ser realizada por meio de inquéritos eletrônicos (*web surveys*) que abordem os aspectos relacionados às expectativas dos alunos, seus estilos de aprendizagem, abordagens de estudo e medidas de autoeficácia computacional.

A avaliação formativa pode ser feita em um AVA não apenas por meio das tradicionais análises de respostas selecionadas (múltipla escolha, verdadeira-ou-falsa e pareamento de questões) e respostas construídas (preenchimento de lacunas) mas também usando ferramentas síncronas e assíncronas, atividades reflexivas, mapeamento conceitual e criação de portfólios, visando adequação entre os objetivos e os processos de aprendizagem, o grau de assimilação dos conceitos pelos alunos, sua capacidade de avaliação crítica e aplicação à realidade (LAGUARDIA; PORTELA; VASCONCELLOS, 2007).

A avaliação formativa pode ser aplicada com o intuito de formar indivíduos que sejam capazes de construir novos conhecimentos, realizar tarefas complexas e resolver problemas, planejada de forma a ser significativa e relevante ao aluno e promover aprendizagem ativa, construindo novos conhecimentos com o acompanhamento e orientação de um tutor, tornando-se o que é chamado por alguns autores como avaliação baseada em performance (HAERTEL, 1999; OTSUKA; ROCHA, 2002).

Pozo e Postigo (1994) realizaram algumas considerações acerca da melhor forma de se elaborar problemas a serem resolvidos pelos alunos, ao apresentar critérios para orientar a elaboração de problemas de tal forma que esses não sejam considerados exercícios pelos alunos. Segundo estes autores, deve-se: a) priorizar tarefas com escopo aberto em detrimento de um escopo fechado, pois estes permitem diferentes possibilidades de solução; b) evitar que a forma que o problema é apresentado se confunda com tipos de problemas; c) diversificar os

contextos das estratégias de ensino para que os alunos trabalhem os mesmos tipos de problemas em distintos momentos do currículo; d) propor problemas situados em um cenário cotidiano, seja este pessoal ou profissional, porém com o objetivo acadêmico e desenvolver novos conhecimentos ou habilidades; e) adequar a definição do problema aos objetivos da tarefa a ser realizada e f) utilizar os problemas para diversos fins de forma a evitar que as tarefas relacionadas à solução de problemas se tornem ilustrações, demonstrações ou exemplificação de algum conteúdo. Resumidamente, esses critérios envolvem a formulação do problema, o processo de solução pelos alunos e a avaliação que se faz sobre os mesmos (POZO; POSTIGO, 2000; LANGHI, 2005).

Avaliações baseadas em resolução de problemas promovem um aprendizado mais significativo quando analisados de forma cognitiva. Para Sternberg (2003, p. 9), problemas são mentalmente organizados em representações mentais compostas por quatro partes: a) descrição do estado inicial do problema; b) descrição do estado do objetivo; c) um conjunto de operadores permitidos e d) um conjunto de restrições. Esta representação mental permite ao aluno lembrar do problema juntamente com a informação de resolução associada, organizando as condições e regras desse problema de forma e, por fim, determinar quais estratégias serão úteis em sua resolução. Sternberg (2003) apresenta um ciclo para a resolução de problemas, composto pelos sete estágios: a) o reconhecimento e identificação do problema; b) definição e representação mental do problema; c) desenvolvimento da estratégia a ser usada ao solucionar o problema; d) a organização do conhecimento acerca do problema; e) alocação dos recursos físicos e psicológicos a serem utilizados na solução do problema; f) monitoramento do processo em relação ao objetivo e g) avaliação da exatidão da solução implementada (POYATOS et al., 2018).

Por sua vez, a avaliação somativa em um AVA pode utilizar como técnica a análise das respostas aos questionários de satisfação do usuário (*smile sheets*), a comparação entre as respostas aos testes feitos antes e após o curso, análise de artigos, documentos ou relatórios de trabalhos de campo escritos individualmente ou em grupo, além da autoavaliação, que é uma técnica recorrente em avaliações de ambientes colaborativos e participativos. A avaliação somativa tem função classificatória e é realizada para averiguar se os alunos alcançaram os níveis de aprendizagem previamente estabelecidos (LAGUARDIA; PORTELA; VASCONCELLOS, 2007).

A avaliação formativa pode ser aplicada com o intuito de formar indivíduos que sejam capazes de construir novos conhecimentos, realizar tarefas complexas e resolver problemas.

No entanto, a avaliação baseada em performance (WIGGINS, 1990; HAERTEL, 1999) é uma forma de avaliação formativa baseada no acompanhamento e orientação do discente durante o desenvolvimento de tarefas que sejam significativas e relevantes para o mesmo, planejadas de forma a promover a aprendizagem ativa e, assim, construir novos conhecimentos (OTSUKA; ROCHA, 2002).

Laguardia e colaboradores (2007) afirmam que embora a avaliação somativa em ambientes virtuais restrinja-se geralmente à análise da pontuação obtida para aprovação, a disponibilidade de registros digitais das atividades dos alunos realizadas ao longo do curso, juntamente com as avaliações realizadas pelos dos tutores, abrem possibilidades para a aplicação de métodos de análise qualitativos.

CAPÍTULO 2 – METODOLOGIA

No método de estudo conhecido como Netnografia é possível estabelecer um paralelo entre os conceitos básicos deste método e aqueles aplicados por grandes empresas de mercado, ao observar os comportamentos digitais de potenciais e atuais clientes nas áreas de estudo conhecidas como *Business Intelligence* (BI) e *Big Data*, apresentadas no subitem posterior.

Posteriormente, apresentam-se duas áreas de estudo conhecidas como *Education Data Mining (EDM)* e *Learning Analytics* (LA) que são oriundas do Big Data e se propõem a pensar os processos de Big Data para a área acadêmica e, após apresentá-las, os demais subitens apresentam as especificidades deste estudo.

2.1 Netnografia

No século XIX, antropólogos julgaram necessário a criação da etnografia, um método de estudo usado para descrever costumes e tradições ao coletar dados de um determinado grupo baseando-se na presença física do pesquisador que realizava uma observação *in loco*, possibilitando ao mesmo confirmar ou refutar hipóteses previamente levantadas.

Quase um século depois, a criação da Internet na década de 1970 e a massificação de seu uso nas décadas de 1990 e seguintes causaram uma grande transformação na forma como as pessoas se expressam e se comunicam umas com as outras, permeando tantas áreas da vida social contemporânea de forma abrangente e chega-se à conclusão que, para se compreender adequadamente muitas das facetas mais importantes da vida social e cultural, é impossível realizar um estudo adequado sem incorporar a Internet e as comunicações mediadas por computador (KOZINETS, 2014).

Netnografia é uma ramificação da etnografia que analisa o comportamento das pessoas e grupos sociais na internet. Kozinets (2014) afirma que a própria etnografia é um neologismo criado no século XIX para definir uma prática que não existia exatamente da forma ou com os objetivos que os criadores do termo estavam tentando comunicar, tornando-se uma “filosofia moral comparativa”. O autor defende que o mundo da pesquisa e inovação intelectual estão permeados por neologismos que causam estranhamento em seu surgimento, como “cibernética”, “psicolinguística” ou “software” e que novos mapeamentos da realidade diferem em seus métodos, ferramentas ou procedimentos ao ponto que exigem novos nomes para defini-los e que esses levam tempo para se estabelecer. Para este autor, portanto,

Netnografia conduz uma pesquisa cultural no mundo contemporâneo da internet e outras TICs que diferem consideravelmente dos procedimentos da etnografia tradicional.

A relutância no emprego do termo por pesquisadores é, em parte, um reflexo da própria relutância e lentidão que antropólogos em geral demonstram para observar comunidades on-line. Conforme levantado por Beaulieu (2004), dois aspectos são considerados problemáticos por pesquisadores ao observar comportamentos de grupos sociais on-line: a) a ausência da comunicação "face-a-face" e b) as comunicações se dão unicamente em modo texto, reforçados por Thompsen e colaboradores:

> As comunidades on-line apresentam ao pesquisador nada além de texto. O etnógrafo não pode observar pessoas, a não ser através de suas contribuições textuais para um fórum. Todo comportamento é verbal na forma de texto. Não há outros artefatos para analisar além do texto. [...] Essa ênfase necessária no texto apresenta oportunidades e limites severos. Em certo sentido, há menos para o etnógrafo perder em um mundo de interação baseado em texto. Todo discurso, comportamento, regras comunitárias e história da comunidade, em princípio, provavelmente estarão disponíveis on-line para a inspeção do pesquisador. Isso pode fazer com que a tarefa pareça enganosamente fácil. Um pesquisador pode se sentar em seu próprio computador, navegar pelo arquivo de uma comunidade, monitorar postagens atuais e ter as condições de trabalho de campo mais fáceis do mundo (THOMPSEN; STRAUBHAAR; BOLYARD, 1998, p. 7, tradução do autor).

Os aspectos levantados pelos pesquisadores são pertinentes e retratam de forma fidedigna o momento em que foram realizadas: as tecnologias de informação e comunicação na década de 1990 limitavam a velocidade de comunicação, restringindo as comunicações instantâneas em um modo texto. No entanto, é importante ressaltar o avanço tecnológico ocorrido nas últimas décadas; o aumento crescente da velocidade em que os dados são transmitidos pela Internet foi acompanhado pela proliferação de uma comunicação mais rica, acompanhada de imagens, sejam estas estáticas ou em movimento.

As tecnologias de informação e comunicação (TICs) têm permeado tantas áreas da vida social contemporânea de forma tão abrangente e presente que cientistas sociais chegam cada vez mais à conclusão de que não é possível compreender adequadamente as diversas facetas que compõem a vida social e cultural moderna sem incorporar as comunidades digitais (KOZINETS, 2014).

Conforme observado por Paccagnella (1997), a netnografia possui algumas vantagens inerentes ao cenário técnico que se propõe a observar. Em primeiro lugar, o número de

indivíduos a serem observados na etnografia presencial pode ser drasticamente reduzido por questões orçamentárias ou prazo, enquanto na netnografia pesquisas profundas e interpretativas sobre comunidades virtuais são beneficiadas pelo uso preciso de novas ferramentas analíticas, poderosas e flexíveis, possibilitando a coleta, organização e exploração dos dados digitais com custos menores. Além disso, se na etnografia presencial os dados e comportamentos dos indivíduos observados podem ser distorcidos pela simples presença do observador, este risco é mitigado na netnografia, já que a presença do observador digital em geral passa desapercebida por seus observados.

É importante ressaltar que uma pesquisa de campo conduzida com técnicas discretas está inevitavelmente fadada a criar grandes problemas éticos cujas diretrizes não se possui um consenso. Por exemplo, alguns acadêmicos como Marvin (2006) acreditam ter uma obrigação ética de obter permissão explícita dos observados ao publicar seus registros em trabalhos acadêmicos; outros como Danowski e Edison-Swift (1985) coletam registros que são processados apenas por software estatístico e não lidos por humanos e, por esta razão, não pedem permissão enquanto outros como Reid (1991) simplesmente não declaram de forma explícita se a permissão para a publicação dos registros foi obtida. Seja como for, todos estão preocupados com a privacidade dos observados ao tomarem preocupações como alterar nomes, pseudônimos e outras informações pessoais mais sensíveis, denotando respeito pela realidade social do ciberespaço (PACCAGNELLA, 1997).

Cabe, no entanto, a definição da palavra “privacidade”. Se pensarmos na privacidade como a quantidade de informação que podemos manter em segredo ou desconhecida, esse tipo de privacidade certamente está diminuindo. Estamos vivendo uma revolução da informação e a coleta, o uso e a análise de dados pessoais são inevitáveis. No entanto, a privacidade como a questão de quais regras devem reger o uso de informações pessoais, então a privacidade nunca foi tão presente, pois é uma das questões mais importantes e vitais discutida pela sociedade moderna. Temos algumas regras de privacidade para controlar os fluxos existentes de informações pessoais, mas não temos regras para governar novos fluxos, usos e decisões derivadas destes dados. Faz-se necessário a criação de novas leis que regulem os custos sociais das novas ferramentas sem sacrificar seus benefícios inegáveis (RICHARDS; KING, 2014).

2.2 *Business Intelligence* e *Big Data*

Herbert Simon recebeu o Prêmio Nobel de Economia em 1978 por sua pesquisa sobre tomada de decisões nas organizações. Sua teoria é amplamente reconhecida como uma suposição fundamental na compreensão dos processos organizacionais, como a tomada de decisões e o planejamento, tendo como princípio o fato de que as organizações operam baseadas em comportamentos racionais e sociais continuamente (PICCIANO, 2012).

Nas décadas de 1980 e 1990, a crescente evolução no poder computacional de processamento de dados e a redução dos custos no armazenamento de informações popularizou uma área de estudos que ficou conhecida como *Business Intelligence* (BI), da qual os dados oriundos dos sistemas operativos das empresas (em geral sistemas integrados conhecidos como ERPs) eram replicados em armazéns de dados (*data warehouses*) e combinados em dimensões diferentes, provendo às suas empresas conhecimento de negócio que as permitia tomar decisões melhores, processo que ficou conhecido como tomada de decisão baseada em dados.

Com o poder de processamento crescendo continuamente e a redução de custos no armazenamento de dados caindo drasticamente nas décadas seguintes, além da popularização da Internet e das redes sociais, o *Business Intelligence* (BI) está evoluindo para um conceito mais sofisticado conhecido como *Big Data*, que visa a análise de dados estruturados (providos pelos sistemas tradicionais da empresa) e dados não-estruturados (como e-mails, imagens ou postagens em redes sociais) de maneira integrada e em grande volume, visando conhecimento mais rico sobre o negócio e tomadas de decisões mais acertivas.

Big Data é um termo genérico que pressupõe que a informação ou o(s) sistema(s) de bases de dados utilizado(s) como armazenamento principal é capaz de armazenar grandes quantidades de dados longitudinalmente e até transações muito específicas (PICCIANO, 2012).

As redes sociais e a tecnologia móvel tornaram a presença on-line de seus usuários constante; desta forma, a tomada de decisão administrativa também evoluiu à medida que mais dados foram disponibilizados a partir de sistemas de informação integrados (dados estruturados) que poderiam ser integrados aos dados não-estruturados de diversas fontes na Internet, permitindo descobrir respostas a perguntas do tipo “e se” usando linguagens de consulta de banco de dados e sistemas de suporte à decisão (PICCIANO, 2012).

Os modernos sistemas informatizados de informação estão facilitando e instilando um maior grau de racionalidade na tomada de decisões em todas as organizações e as ajudam a

ajustar, adaptar e aprender para desempenhar suas funções administrativas. As ferramentas providas pela *Business Intelligence* e posteriormente *Big Data* não substituem o tomador de decisões; elas o ajudam a refinar o processo de tomada de decisão (PICCIANO, 2012).

Assim sendo, é essencial a análise de dados para tomadas de decisão mais acertivas, especialmente nos casos em que a bases de dados possui milhões de registros; um caso de uso comum é a empresa de comércio eletrônico Amazon.com que examina o tráfego, as compras ou os padrões de navegação do site para determinar quais de seus clientes estão mais ou menos propensos a comprar produtos específicos. Em posse de tais informações, as empresas podem realizar várias ações, como enviar notificações aos clientes de novos produtos assim que se tornam disponíveis ou oferecer descontos personalizados (PICCIANO, 2012). Ao cruzar os dados de seus sistemas operativos com outras fontes de dados como a Bolsa de Valores de Nova York, Facebook e Ancestry.com, as empresas são capazes de prever o comportamento do consumidor, aprimorar produtos e serviços e tomar decisões de negócios mais bem informadas (REYES, 2015).

Entre os termos relacionados à *Big Data* e a tomada de decisão baseada em dados incluem: *data warehousing*, *data mining* e desagregação de dados. O *Data Warehouse*, ou armazém de dados, é essencialmente um sistema de informações de bancos de dados projetado para a integração de grandes volumes de dados oriundos de diversas fontes estruturadas de informação. No entanto, a integração de dados não-estruturados para análise exigiu estruturas menos rígidas de armazenamento de dados, capazes de armazenar e manter grandes volumes de dados nos mais diferentes formatos e a estas estruturadas que representam uma evolução do *Data Warehouse* dá-se o nome de *Data Lakes* (ou lagos de dados).

A mineração de dados (*Data Mining*) é um termo usado em pesquisa e estatística que se refere à busca ou "escavação" de um arquivo de dados, realizando cruzamentos de dados para obter informações que compreendam melhor um determinado fenômeno e é uma área de pesquisa que permite a descoberta de informações novas e potencialmente úteis a partir de grandes volumes de informação (WITTEN; FRANK, 1999; PICCIANO, 2012).

A desagregação de dados, por sua vez, refere-se ao uso de ferramentas de software para dividir os arquivos de dados em várias características como, por exemplo, obter o resultado das vendas de um produto por gênero, faixa etária, etnia ou outras características (PICCIANO, 2012).

No entanto, ao aplicar os métodos de mineração de dados que são utilizados convencionalmente, verificou-se a necessidade de desenvolver métodos específicos para a área acadêmica, devido à necessidade de explicar explicitamente (e as oportunidades de explorar) a hierarquia multinível e a não-independência nos dados educacionais e, por esta razão, é cada vez mais comum o desenvolvimento de modelos específicos para mineração de dados educacionais (BAKER; YACEF, 2009).

2.3 *Education Data Mining* (EDM) e *Learning Analytics* (LA)

Education Data Mining (EDM) ou mineração de dados educacionais, de acordo com a sua própria comunidade de pesquisadores (EDM, 2018), é um campo de estudo recente que tem como foco o desenvolvimento de métodos para explorar os tipos únicos de dados providos pelo ramo acadêmico, seja ele presencial ou a distância, e com o uso destes métodos compreender melhor os alunos e como o aprendizado acontece (BAKER; YACEF, 2009).

Os métodos utilizados para mineração de dados educacionais combinam conhecimentos das áreas de mineração de dados convencional (*data mining*), aprendizado de máquina (*machine learning*), psicometria, estatística, visualização de informações e modelos computacionais (BAKER; YACEF, 2009). No ensino superior, a análise está começando a ser usada em vários aplicativos que abordam o desempenho, os resultados e a persistência do aluno. Em um cenário de *Big Data*, os dados são coletados para cada transação de aluno em um curso, especialmente em um curso na modalidade de ensino a distância: cada entrada de aluno em uma avaliação de curso, fórum de discussão, atividade de wiki pode ser registrada, conteúdos visualizados e vídeos assistidos, gerando milhares de transações por aluno por curso. Além disso, esses dados podem ser coletados em tempo real ou quase em tempo real e, em seguida, analisados para sugerir ações corretivas (PICCIANO, 2012).

Os pesquisadores de Mineração de Dados Educacionais estudam várias áreas, incluindo aprendizado individual a partir de software educacional, aprendizado colaborativo auxiliado por computador, testes adaptativos por computador (e testes mais amplos) e os fatores associados à falha ou não-retenção de alunos em cursos. Em todos esses domínios, uma das principais áreas de aplicação tem sido a melhoria dos modelos de alunos que representam informações sobre as características ou sua situação atual, como seu conhecimento atual, motivação, meta-cognição e atitudes (BAKER; YACEF, 2009).

Sistemas de manutenção de registros de estudantes universitários mantiveram informações de resultados sobre os alunos, como notas em cada curso. Essas informações poderiam ser usadas por pesquisadores institucionais para estudar padrões de desempenho dos alunos ao longo do tempo, geralmente de um semestre a outro ou de um ano para outro (PICCIANO, 2012).

Assim como um professor possui a capacidade de se adaptar para viabilizar um ensino que favoreça o aprendizado de um único aluno, este mesmo professor pode aprender mais sobre como estudantes aprendem, refletir e aprimorar suas práticas ao observar um grupo de estudantes (MERCERON; YACEF, 2005)

Segundo Heiner, Heffernan e Barnes (2007), pesquisadores têm usado mineração de dados educacionais para detectar afeição ou desmotivação do aluno, detectar tentativas do estudante burlar o aprendizado ao explorar brechas de sistema, conduzir os esforços de estudo dos alunos, desenvolver ou refinar modelos de estudo, mensurar os efeitos de intervenções individuais, aprimorar o suporte ao ensino e prever comportamento e performance dos estudantes.

Em um artigo publicado pela IBM Software Group (2011) intitulado *Analytics for Achievement*, oito categorias de aplicações foram apresentadas e são resumidas no Quadro 1:

Quadro 1 – Oito categorias de aplicação para *Learning Analytics*

Categoria	**Descrição**
Medir e monitorar o desempenho dos alunos	O primeiro passo lógico é obter uma imagem clara de como os alunos estão se saindo em relação a outros alunos, seja em uma esfera local ou global. Os sistemas de análise permitem que as escolas visualizem os dados de desempenho de alunos e grupos em vários eventos avaliativos, permitindo comparar as notas de estudante, escola, distrito ou diretoria e nacional e visualizar e acompanhar o progresso por turma, professor, curso ou programa. Uma melhor percepção do desempenho dos alunos é o primeiro passo para desvendar o potencial dos alunos.
Relatório de resultados	Relatar o desempenho dos alunos é uma necessidade comum para organizações educacionais de todos os níveis, seja uma função de lei, financiamento ou relações públicas. As ferramentas de análise permitem que educadores investiguem segmentos especiais da população ou por atributos individuais, a fim de compreender melhor fatores no sucesso ou fracasso de uma iniciativa educacional.
Detectando os "pontos fora	Uma das ferramentas mais eficazes para que as escolas melhorem o

da curva" para intervenção precoce	desempenho geral é a intervenção precoce dos alunos "fora da média", sejam estes alunos com baixíssimo rendimento (em risco) e alunos de alto desempenho. O objetivo é detectar os alunos de baixo desempenho cedo o bastante para que medidas possam ser tomadas. Geralmente a instituição escolar não têm os recursos para acompanhar o progresso de um aluno em todos os assuntos e avaliações de ano para ano. Uma mensuração mais intensa do desempenho dos alunos aliada a análises pode ajudar a destacar os fatores que indicam uma queda no desempenho. Por outro lado, os alunos de alta performance estão sempre em risco de ficarem entediados e, portanto, desmotivados a aprender. A identificação igualmente precoce destes estudantes permite a instituição realizar um trabalho diferenciado, seja ofertando a eles conteúdos didáticos complementares ou atividades mais desafiadoras, de forma a mantê-los sempre motivados a aprender.
Prever o potencial dos alunos	Embora instituições educacionais concentrem esforços no auxílio de alunos de baixo desempenho, não podem deixar de acompanhar o aluno em performance dentro de média, afinal, educadores desejam que seus alunos atingiam seu potencial. Com análise preditiva a partir dos dados dos alunos, é possível estabelecer padrões que sugiram o desempenho individual do aluno, sendo possível descobrir causas e oferecer auxílio proativamente em casos de alunos cujo desempenho esteja abaixo do previsto.
Prevenir a evasão escolar	Teóricos identificaram uma série de sinais de alerta que podem apontar para a necessidade de intervenção, sendo a frequência às aulas e envolvimento em atividades escolares os indicadores mais comuns na modalidade presencial. Estes indicadores possuem equivalência no ensino a distância, como o número de vezes da qual o aluno acessa o ambiente virtual de aprendizagem, tempo de permanência em cada um dos objetos de aprendizagem, frequência e performance em avaliações somativas, entre vários outros indicadores que podem ser determinados para mensurar o interesse do aluno no curso. Se ofertados em tempo real ou quase em tempo real, medidas proativas podem ser tomadas evitando a evasão dos alunos que apresentaram os índices mais baixos.
Identificar e desenvolver os atributos-chave dos bons professores	A análise de dados pode auxiliar as instituições de ensino a identificar, recrutar e reter os melhores professores, assim como melhorar as práticas de ensino de todos eles. A plotagem de notas com dados demográficos pode destacar os professores que estão obtendo o melhor desempenho dos alunos mais desfavorecidos. A satisfação do aluno com os professores também se correlaciona

	diretamente com o sucesso do aluno. Assim sendo, dados qualitativos de feedback dos alunos podem revelar *insights* subjetivos sobre a experiência em sala de aula. Entender os principais atributos de um grande professor pode ajudar um conselho escolar a estabelecer melhores padrões para o desempenho do professor e usar essa informação para uma vez identificados, os professores excepcionais poderem compartilhar técnicas práticas para gestão de sala de aula, disciplina, motivação, incentivo e práticas de ensino com seus colegas e sua gestão para ajudar todos os professores a obter os melhores resultados.
Analisando testes padronizados para uniformizar o desempenho	Testes padronizados como o ENEM e o ENADE são necessários como um indicador da saúde do sistema educacional e podem ser úteis como uma excelente linha de base do desempenho institucional e podem revelar oportunidades de melhoria. A classificação pode ajudar os educadores a examinar os fatores por trás do desempenho e trabalhar mais para direcionar mais recursos para escolas de baixo desempenho em uma tentativa de nivelar o sistema educacional como um todo, como aconteceu com a Finlândia.
Testar e desenvolver currículos	Os currículos sempre serão atualizados em um esforço contínuo de se desenvolver da melhor maneira possível as habilidades e conhecimentos mais adequados à sociedade e ao mercado de trabalho. A análise de dados pode ser utilizada com o intuito de monitorar o desempenho dos alunos assim que forem expostos a novos elementos do currículo, analisando sua aceitação entre docentes e discentes, adequando materiais didáticos e atividades de forma a atingir este objetivo. Mudanças podem ser implementadas em um escopo menor, como serem disponibilizadas para uma população menor de alunos, podendo ser ampliadas gradativamente à medida que apresentam bons resultados. À medida em que educadores descubram que diferentes métodos funcionam para diferentes grupos de alunos, estes registros históricos podem alimentar uma análise preditiva capaz de prever, por exemplo, que tipo de programa de matemática será ideal para determinados grupos de alunos e quanto tempo levará para adotar, economizando tempo, recursos financeiros e garantindo o sucesso de currículos personalizados.

Fonte: IBM Software Group (2011), adaptado pelo autor (2019)

Estas oito categorias indicam o desenvolvimento de recursos de informações confiáveis e oportunos para coletar, classificar e analisar os dados usados no processo de tomada de decisão. É importante observar que a análise de dados é usada para informar e não para substituir a experiência, a perícia, a intuição, o discernimento e a perspicácia de

educadores competentes. Em uma organização educacional moderna, a tomada de decisões é um componente integral de processos complexos de gerenciamento, como planejamento acadêmico, elaboração de políticas e orçamento, exigindo a participação das partes interessadas e, o mais importante, procuram incluir informações que ajudem todos os envolvidos no processo de decisão (PICCIANO, 2012).

Como a análise de aprendizado exige grandes quantidades de dados coletados dos alunos e integrada a outros bancos de dados, as instituições de ensino precisam ter cuidado com a privacidade, o perfil dos dados e os direitos dos alunos em registrar seus dados individuais e comportamentos. Estas instituições sempre realizaram registros acerca da avaliação do desempenho do aluno e do comportamento acadêmico, mas a análise da aprendizagem a partir de grandes massas de dados leva o registro do comportamento a outro nível e escopo. Por mais bem-intencionada que a análise de aprendizado possa ser em termos de ajudar os alunos a terem sucesso, essa abordagem de *Big Data* também pode ser vista como um grande "*Big Brother*" e como uma invasão de privacidade que alguns alunos prefeririam não ter imposto. Devem ser tomadas precauções para garantir que as extensas coletas de dados das transações instrucionais dos alunos não sejam utilizadas de maneira a prejudicar os indivíduos (PICCIANO, 2012). São novas questões éticas a serem abordadas. Ao evoluir a tecnologia, como o rastreamento de localização e biometria, é possível coletar uma variedade de dados que vão além do desempenho acadêmico. Se os alunos tiverem a percepção que sua privacidade está sendo invadida, eles podem relutar em permitir que seus dados sejam usados para pesquisa e análise (REYES, 2015).

A privacidade não deve ser pensada apenas como o quanto é secreto, mas sim sobre quais regras existem (legais, sociais ou outras) para governar o uso da informação, bem como sua divulgação. Embora as tecnologias digitais e as práticas governamentais e corporativas estejam colocando muitas noções existentes de privacidade sob ameaça, a privacidade em geral não está acabando, porque a privacidade é mais do que apenas segredo, é uma forma abreviada que usamos para identificar regras de informação (RICHARDS; KING, 2014).

A utilidade da mineração desses dados é promissora, mas a análise estereotipada deve ser mitigada e alguns pesquisadores já começam a estabelecer algumas diretrizes para garantir que os dados possam ser utilmente removíveis de sua experiência de dados de mineração no projeto (MERCERON; YACEF, 2005). Além do caso de usuários de teste que podem "sujar" os dados e comprometer os resultados de uma análise (e devem ser devidamente removidos da

amostragem), outros modelos podem ser gerados a partir de comportamento atual de uma etnia ou faixa etária que não será necessariamente perpetuado no futuro.

Para aproveitar todo o potencial de *Big Data* e a análise de aprendizado (ou *Learning Analytics* - LA), é praticamente um requisito que o processamento de transações seja eletrônico e automatizado em vez de manual, possibilitando a escalabilidade da análise. Além disso, é importante que as transações instrucionais sejam coletadas conforme elas ocorram, e um ambiente virtual de aprendizagem (AVA) é fundamental neste cenário. A maioria destes ambientes fornece monitoramento constante da atividade dos alunos, sejam eles: respostas, postagens em um painel de discussão, acessos a material de leitura, conclusões de um teste ou alguma outra avaliação. Mesmo ao usar um AVA básico, ao usar todos seus recursos em um curso *on-line* robusto de quinze semanas, é possível gerar milhares de transações por aluno. A gravação em tempo real e a análise dessas transações podem ser usadas para alimentar um aplicativo de análise de aprendizado. As transações instrucionais também devem ser integradas a outros recursos, como dados dos sistemas de informações da faculdade (estudante, curso, faculdade) e redes sociais, provendo um cruzamento de informações mais completo (PICCIANO, 2012).

Uma das principais vantagens em análise de aprendizado (LA) é a possibilidade de mensurar o progresso de um aluno em qualquer estágio e durante qualquer atividade em um curso; se em uma avaliação tradicional do curso algumas informações sobre as motivações e opiniões do aluno são descobertas apenas quando o curso termina, a LA permite entender como os alunos estão usando o conteúdo, interagindo e participando de um curso, para que a intervenção precoce seja realizada (REYES, 2015).

No geral, a análise de aprendizado (LA) pode ser usada por todos os envolvidos na tomada de decisões educacionais de várias maneiras para alcançar o objetivo de melhorar os resultados de aprendizagem, não apenas para prever o desempenho de um aluno através de técnicas de análise preditiva, como também pode reconhecer padrões que forneçam sugestões de recursos de aprendizagem relevantes para as necessidades do aluno, uma verdadeira experiência de aprendizagem personalizada adaptada em tempo real (SIEMENS, 2012; VERBERT et al., 2012). Em um estudo, Tang e McCalla (2004) relatam uma instanciação de tal sistema que integra o agrupamento e a filtragem colaborativa para recomendar conteúdo aos alunos. Os autores apresentam um estudo realizado com alunos simulados e uma posterior avaliação bem-sucedida do sistema com alunos reais.

Além disso, a análise de aprendizado (LA) também pode aumentar a conscientização do aluno, pois os educadores compartilham e discutem a análise a partir dos dados com o aluno e oferecem a capacidade de refletir sobre o processo de aprendizagem, podem melhorar o aprendizado social e identificar comportamentos indesejáveis nos alunos e permitir a detecção de sentimentos como frustração, confusão ou tédio (REYES, 2015).

Dada a possibilidade de obtenção e análise de dados, pode-se desenvolver processos de avaliação mais precisos e específicos sobre a aprendizagem dos alunos.

2.4 Aprendizado de máquina assistida: algoritmos de classificação

Aprendizado de máquina assistido ou indutivo é o processo de aprender um conjunto de regras a partir de instâncias (exemplos em um conjunto de treinamento) ou, de maneira mais geral, criar um classificador que pode ser usado para generalizar a partir de novas instâncias (KOTSIANTIS, 2007).

O primeiro passo é coletar o conjunto de dados e, se um especialista estiver disponível, ele poderá sugerir quais atributos (que também são chamados de dimensões) são os mais relevantes. Entretanto, na falta de um especialista, o método mais simples é o de "força bruta", que significa medir tudo o que está disponível, na esperança de que as características mais relevantes possam ser isoladas. Geralmente dados coletados por "força bruta" não são os mais adequados pois eles contém, na maioria dos casos, valores de ruído e falta de recursos, necessitando de um pré-processamento significativo, cerca de 80% do tempo no processo de engenharia de dados (ZHANG, S., ZHANG, C., YANG, 2003; KOTSIANTIS, 2007).

É geralmente aceito que medidas robustas de sucesso de previsão façam uso de dados independentes, isto é, dados não utilizados para desenvolver o modelo de previsão. O Quadro 2 descreve algumas estratégias usadas para obter dados de teste; embora os dois conjuntos de dados necessários para desenvolver e testar previsões sejam chamados no quadro de 'treinamento' e 'teste', em algumas literaturas podem, por exemplo, podem ser chamados de dados de aprendizagem e validação e até outras variações. Trata-se de uma prática comum dividir ou particionar os dados disponíveis para fornecer os dados de treinamento e teste (FIELDING; BELL, 2017).

Quadro 2 – Métodos de particionamento de dados para a atribuição de casos ao treinamento e teste

Método	**Exemplos**	**Descrição**
Resubstituição	Stockwell (1992) Osborne e Tigar (1992)	Nenhum particionamento é realizado: os mesmos dados são

		usados para treinamento e teste. Sua tendência é fornecer perspectivas otimistas do sucesso da predição.
Boostrapping	Buckland & Elston (1993) Verbyla & Litaitis (1989)	Amostras de *bootstrap* (amostragem com substituição) são usadas para avaliar o sucesso da previsão. A precisão é geralmente relatada como uma média e confiança limites.
Randomização	Capen et al (1986)	Amostras aleatórias são obtidas por amostragem sem reposição. Precisão é geralmente relatada como uma média e limites de confiança.
Amostragem prospectiva	Capen et al (1986) Fielding & Haworth (1995) Morrison et al. (1987)	Uma nova amostra de casos é obtida após o modelo ter sido desenvolvido, que podem ser de uma região ou tempo diferente.
Particionamento *k-fold*	Stockwell (1992)	Os dados são divididos em conjuntos k (k> 2), dos quais apenas um é usado para treinamento. Os conjuntos restantes de k - 1 são agrupados para fins de teste. A precisão é geralmente relatada como uma média e limites de confiança.
Particionamento *k-fold Leave-One-Out (L-O-O)*	Capen et al (1986) Osborne e Tigar (1992)	Também conhecida como amostragem *jack knife*, n amostras de 1 caso são testadas sequencialmente, sendo que os restantes n - 1 casos formam o conjunto de treino.
Particionamento *k-fold K=2*	Smith (1994)	Os dados são divididos em um conjunto de treinamento e um conjunto de testes. Uma variedade de estratégias pode ser empregada para determinar a divisão.

Fonte: FIELDING e BELL (2017), tradução livre.

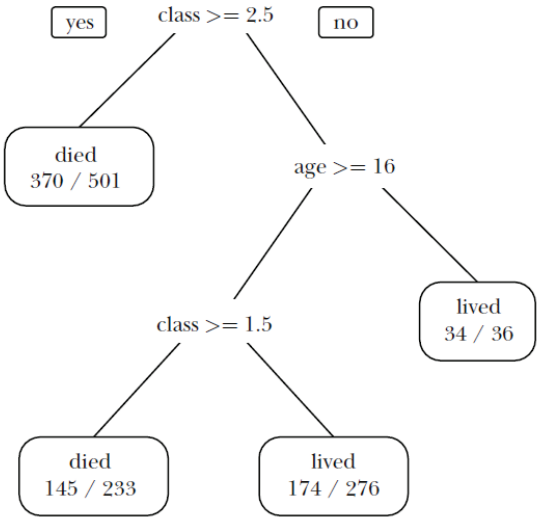
Com as amostras devidamente separadas, cabe a aplicação de um algoritmo da classificação e algumas possibilidades são apresentadas neste estudo.

2.4.1 Árvore de decisão

As árvores de decisão classificam as instâncias (ou ocorrências) tomando como base os valores de atributos (ou dimensões), já que cada nó em uma árvore de decisão representa um atributo em uma instância a ser classificada, e cada ramo representa um valor que o nó pode assumir (KOTSIANTIS, 2007).

VARIAN (2014) ilustra o uso de árvore de decisão ao utilizá-la para classificar os sobreviventes do Titanic, baseado em dois atributos: a classe em que os sobreviventes viajavam (primeira, segunda e terceira classe) e suas idades, como pode ser visto na Figura 1.

Figura 1 – Classificação dos sobreviventes do Titanic usando árvore de decisão



Fonte: VARIAN (2014)

No entanto, a representação gráfica de árvore de decisão não é a única maneira de se representar as regras a serem seguidas pelo modelo. A Tabela 1 é a representação do mesmo exemplo de VARIAN (2014) no formato de formulário de regras.

Tabela 1 - Classificação dos sobreviventes do Titanic no formato formulário de regras

Características	**Predição**	**Atual/Total**
Terceira classe	Faleceu	370/501
Primeira e segunda classes, idade baixo de 16 anos	Viveu	34/36
Segunda classe, idade igual ou maior a 16 anos	Faleceu	145/233
Primeira classe, idade igual ou maior a 16 anos	Viveu	174/276

Fonte: VARIAN (2014, tradução nossa)

2.4.2 Naïve Bayes

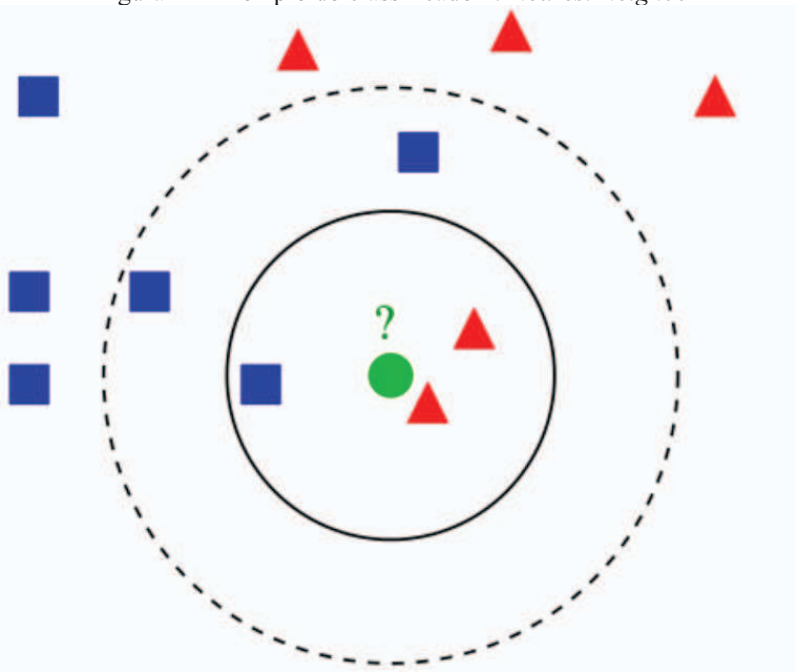
O algoritmo Naïve Bayes é um classificador probabilístico simples que calcula um conjunto de probabilidades contando a frequência e as combinações de valores em um determinado conjunto de dados e, para tal, faz uso do teorema de Bayes. Por assumir que todos os atributos são independentes, tende a ter um bom desempenho e aprender rapidamente em vários problemas de classificação supervisionados, embora a independência condicional raramente é verdadeira em aplicações no mundo real e por isso é considerado um classificador simples (PATIL; SHEREKAR, 2013).

2.4.3 k-Nearest Neighbor (KNN)

Conjuntos de dados que possuam um volume pequeno de instâncias (ou registros) e poucas dimensões são tradicionalmente analisados com algoritmos presentes em sistemas gerenciadores de bancos de dados, como *B+tree* ou *R-trees*. Entretanto, quando o conjunto de dados possui alto volume de instâncias e muitas dimensões, a técnica do *Nearest Neighbor* (ou vizinho mais próximo, em uma tradução livre) é vital nestas situações. KNN é uma técnica de classificação de dados que encontra o vizinho mais próximo baseado no valor de *k* e é simples e fácil de aprender, rápida para o treinamento de modelos, altamente eficiente em grandes volumes de dados e eficaz no campo de reconhecimento de padrões, mesmo para dados de treinamento ruidosos (DHANABAL; CHANDRAMATHI, 2011).

Ajanki (2010) ilustra o funcionamento de KNN conforme mostrado na Figura 2; o ponto representado em verde é classificado como um triângulo vermelho quando o valor de *k* for menor (círculo sólido), pois existem dois triângulos e apenas um quadrado azul. Entretanto, se o valor de k for maior (representado pelo círculo pontilhado) o ponto verde se torna um quadrado azul, pois o conjunto possui agora três quadrados azuis (maioria) e apenas dois triângulos vermelhos.

Figura 2 – Exemplo do classificador *k-Nearest Neighbor*



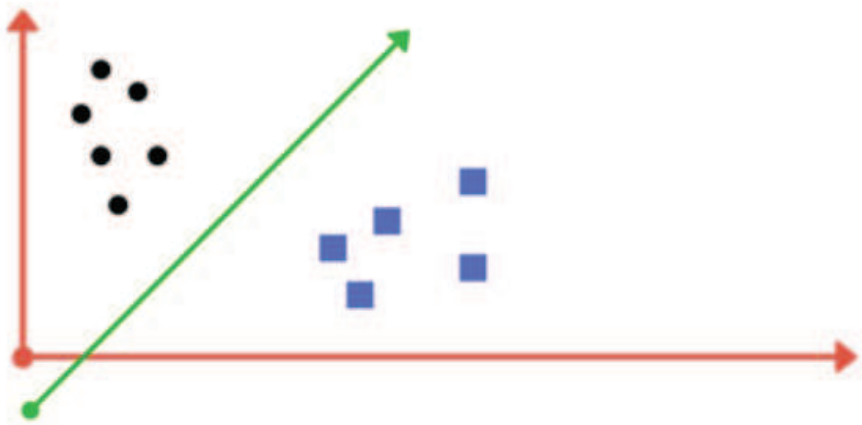
Fonte: AJANKI (2010)

2.4.4 Support Vector Machine (SVM)

Support Vector Machine é uma ferramenta universal para resolver problemas de estimativa de funções multidimensionais. Inicialmente, foi projetado para resolver problemas de reconhecimento de padrões em situações em que para encontrar uma regra de decisão com boa capacidade de generalização são selecionados alguns pequenos subconjuntos dos dados de treinamento que são chamados de Vetores de Suporte (SVs). A separação ideal dos vetores de suporte é equivalente à separação ideal de todos os dados (VAPNIK; GOLOWICH; SMOLA, 1998).

Trata-se de um classificador discriminativo formalmente definido por um hiperplano de separação, pois os dados de treinamento rotulados geram um hiperplano ideal que é utilizado para categorizar novos exemplos. Em situações bidimensionais esse hiperplano é uma linha que divide um plano em duas partes e cada classe fica em cada lado (PATEL, 2017).

Figura 3 – Exemplo do classificador *SVM* em duas dimensões



Fonte: PATEL (2017)

2.4.5 Matriz de confusão

Uma matriz de confusão ilustra a precisão da solução para um problema de classificação e contém informações reais (presentes na amostra de teste) e classificações previstas feitas por uma classificação sistema. Há dois possíveis erros de predição: falsos positivos (FP) e falsos negativos (FN). O desempenho de algoritmos de classificação é normalmente sumarizado em uma matriz de confusão ou erro que cruza os padrões de presença / ausência observados e previstos. Um exemplo pode ser observado na Figura 4, sendo “a” o número de previsões corretas que uma instância é positiva, “b” o número de falsos positivos (predições incorretas), “c” o número de falsos negativos (predições

incorretas) e “d” o número de previsões corretas que uma instância é negativa. (PATIL; SHEREKAR, 2013; FIELDING; BELL, 2017)

Figura 4 – Matriz de confusão ou erro

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Fonte: FIELDING e BELL (2017)

Os dados na matriz de confusão são geralmente apresentados como contagens, mas podem ser apresentados como porcentagens e pode possuir mais linhas e colunas quando houver mais do que dois valores possíveis.

2.5 Análise utilizando Linguagem R

R é uma linguagem de computação estatística criada por Ross Ihaka and Robert Gentleman que combina recursos úteis de duas linguagens de computador existentes, a linguagem S de Becker, Chambers e Wilks (1988) e a Scheme de Steel e Sussman (1975). A nova linguagem oferecia vantagens nas áreas de portabilidade, eficiência computacional, gerenciamento de memória e escopo (IHAKA; GENTLEMAN, 1996).

CAPÍTULO 3 – MÉTODO

Neste capítulo é apresentado o método utilizado nesta pesquisa, a natureza de seu estudo, materiais utilizados e procedimentos adotados.

3.1 Natureza do estudo

A Faculdade de Informática e Administração Paulista (FIAP) é uma instituição de ensino particular fundada em 1993 na cidade de São Paulo com o objetivo de oferecer cursos em ensino superior nas áreas de Informática e Administração. Posteriormente, passou a ofertar cursos de pós-graduação do tipo *lato sensu* (nomeados atualmente pela instituição como MBA, *Master of Business Administration*) e cursos livres de curta duração sob a marca SHIFT.

A instituição é reconhecida pelo mercado de trabalho que atua por formar profissionais com qualidade e, se comparada a outras instituições em território nacional, optou tardiamente por ofertar seus cursos na modalidade de ensino a distância (EaD), por temer oferecer uma formação não-presencial que não estivesse à altura de seus cursos presenciais e da expectativa do mercado.

Embora já utilizasse ambientes virtuais de aprendizagem (AVA) há mais de uma década, tais ferramentas tinham como objetivo estender a sala de aula, ao permitir entrega de atividades extraclasse previstas em seus cursos.

No ano de 2014 a instituição lançou suas duas primeiras disciplinas totalmente a distância que compunham vinte por cento da carga horária de alguns de seus cursos presenciais em ensino superior, conforme permitido em lei na época, assim como alguns cursos livres de curta duração do selo SHIFT, iniciativas que serviram como um debute e ambiente de experimentação na criação e experimentação do material didático, recursos audiovisuais (como áudio e vídeo pré-produzidos e transmissões ao vivo), AVA e avaliações a serem realizadas.

Após a publicação da autorização de cursos totalmente a distância em Diário Oficial no final de 2015, no ano seguinte os dois primeiros cursos em ensino superior na modalidade foram oferecidos, e os escolhidos foram “Tecnologia em Análise e Desenvolvimento de

Sistemas" e "Tecnologia em Gestão de Tecnologia da Informação", sendo o primeiro curso mencionado o escolhido para este estudo.

O ambiente virtual de aprendizagem (AVA) adotado pela FIAP desde 2015 é o Moodle, software que foi desenvolvido com princípios pedagógicos e utilizado em ensino a distância, ensino híbrido, sala de aula invertida e outras iniciativas educacionais em ambientes de trabalho, escolas e universidades em todas as partes do mundo. Por ser um *software* livre e de código-fonte aberto (distribuído sob a licença livre *General Public License* do projeto GNU), o Moodle pode ser utilizado livremente, seja para propósitos pessoais ou comerciais, e suas funcionalidades podem ser modificadas e novas podem ser implementadas, estendendo o código-fonte por meio de *plugins*.

A Faculdade de Informática e Administração Paulista (FIAP) personalizou seu ambiente virtual de aprendizagem Moodle para suas necessidades acadêmicas, tanto esteticamente quanto adicionando ao ambiente novas funcionalidades, por meio de *plugins* de personalização disponíveis no mercado e desenvolvendo seus próprios.

Os dados gerados por meio da interação dos alunos em um ambiente virtual de aprendizagem como o Moodle precisam ser armazenados em um sistema de gerenciamento de banco de dados (SGBD, ou *Data Base Management System* (DBMS)), responsável por gerenciar, persistir, manipular e organizar os dados nele armazenados, de forma eficiente e segura.

Embora o ambiente virtual de aprendizagem Moodle permita atualmente o uso de outros SGBDs como PostgreSQL, Microsoft SQL Server e Oracle Server, a implementação do AVA no ambiente da FIAP utiliza o SGBD original do Moodle, o banco de dados MySQL. Por possuir uma licença híbrida (parcialmente livre), a infraestrutura da faculdade e o ambiente criado para este estudo fazem uso da versão livre deste SGBD, conhecido como MariaDB.

Foram sujeitos da pesquisa 483 alunos ingressantes no ano letivo de 2018 de sete cursos superiores ministrados totalmente a distância, oferecidos pela Faculdade de Informática e Administração Paulista (FIAP).

3.3 Material

Para a realização deste estudo foi necessária a implementação de uma infraestrutura tecnológica de Big Data, de forma que houvesse maior liberdade no uso de ferramentas e no

tratamento dos dados. Diferentemente de ambientes tecnológicos de *Business Intelligence* que, em sua maioria, requerem a aquisição de licenças de software que poderia encarecer o estudo, a infraestrutura necessária para um ambiente *Big Data* pode ser criada com a composição de dezenas de *softwares* livres e de código-fonte aberto, sendo a maioria deles mantido pela fundação Apache (http://apache.org/).

A escolha de quais ferramentas *Big Data* utilizar é orientada pela necessidade de armazenamento e tratamento de dados que grandes corporações possuem, resultando em ambientes muito distintos entre si. Assim sendo, o ambiente resultante deste estudo é fruto da necessidade analítica aqui proposta.

3.3.1 Sistema operacional do ambiente *Big Data*

O sistema operacional escolhido para abrigar o ambiente Big Data deste estudo é uma versão do sistema operacional livre e de código-fonte aberto Linux conhecida como CentOS. Esta versão, por sua vez, é uma derivação de fontes de outra distribuição Linux conhecida como *Red Hat Enterprise Linux* (RHEL), a mais bem-sucedida versão comercial do sistema. Conforme observado em pesquisas e experimentações prévias, os *softwares* de Big Data operam de maneira mais estável nestas versões do sistema operacional, e é por esta razão que a infraestrutura do estudo fez uso do CentOS Linux na versão 7.5.1804.

Por não haver grande necessidade de performance na análise de dados e se tratar de um pequeno volume de dados (medidos em dezenas de gigabytes), optou-se por uma única instância do sistema operacional, não se tratando, portanto, de um *cluster*, situação pela qual várias instâncias de um sistema operacional são ligadas entre si trabalhando de forma colaborativa. Esta instância única foi instalada em um ambiente virtualizado, conceito que a computação em nuvem (*Cloud computing*) se baseia, da qual há uma abstração do hardware utilizado. O software de virtualização escolhido foi o *Oracle VirtualBox* na versão 5.2.20 r125813, que pode ser baixado gratuitamente em (http://www.virtualbox.org/).

3.3.2 O lago de dados do ambiente *Big Data*

Um lago de dados é um repositório de armazenamento massivamente escalável que contém uma grande quantidade de dados brutos em seu formato nativo, ou seja, exatamente como são, e servem para alimentar sistemas de análise que devem pegar os dados brutos, tratá-los e transformá-los em informações úteis. Os lagos de dados são tipicamente

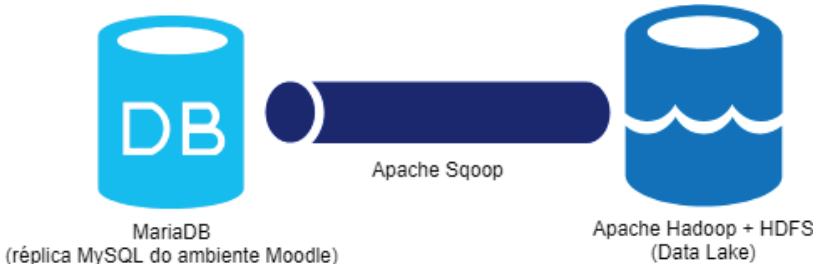
construídos para lidar com volumes grandes e rápidos de dados não estruturados, em contraste com os *data warehouses* clássicos de *Business Intelligence* que armazenam dados já estruturados (MILOSLAVSKAYA; TOLSTOY, 2016). Tratam-se de repositórios integradores que recebem dados das mais variadas origens, como sistemas empresariais integrados (*Enterprise Resource Planning* – ERP), sistemas de gerenciamento de relacionamento com clientes (*Customer Relationship Management* – CRM), portais de notícias, e-mails, planilhas eletrônicas, rede sociais, entre outros.

O lago de dados (*Data Lake*) necessário para abrigar os dados estruturados e não-estruturados desse estudo foi criado com o software Apache Hadoop na versão 2.7.6, um software livre e de código-fonte aberto mantido pela fundação Apache. O *Apache Hadoop* possui um componente conhecido como *Hadoop Filesystem* (HDFS), que permite abrigar arquivos de dados de grandes proporções e de maneira distribuída, possibilitando o armazenamento e análise posteriores de grandes volumes de dados, como a área de pesquisa de *Big Data* se propõe a realizar.

Os dados oriundos do ambiente virtual de aprendizagem foram carregados na infraestrutura da seguinte maneira: uma implementação do sistema de gerenciamento de banco de dados MySQL, um MariaDB versão 5.5.60, foi instalado no ambiente virtualizado do estudo e recebeu a carga de dados a partir de cópias de segurança do servidor MySQL da faculdade, tornando-se uma réplica idêntica do banco de dados utilizado pelo ambiente virtual de aprendizagem. Trata-se de uma boa prática de mercado, pois em uma eventual análise de dados massiva que exija muito poder computacional ou em uma eventual perda de dados, os serviços prestados em ensino a distância pela faculdade aos seus alunos no AVA são absolutamente preservados.

Os dados estruturados contidos na réplica MySQL do ambiente virtualizado devem, por sua vez, ser transferidos para o lago de dados criado pelo Apache Hadoop em seu HDFS. Para a realização desta tarefa, foi utilizado o software livre e de código-aberto Apache Sqoop na versão 1.4.7 que, aliado a um conector MySQL escrito na linguagem Java (8.0.13), são específicos para este tipo de necessidade, atuando como uma espécie de "túnel" de dados, transferindo os dados do MySQL para o HDFS, conforme mostrado na Figura 6:

Figura 5 – Representação da estrutura de dados do estudo



Fonte: Elaborado pelo autor (2018)

3.3.3 Ferramentas de análise no ambiente *Big Data*

O ambiente de *Big Data* deste estudo também foi equipado com ferramentas que possibilitariam a análise dos dados contidos no lago de dados. Uma das ferramentas escolhidas foi o Apache Hive, que facilita a leitura, escrita e gerenciamento de grandes massas de dados presentes no *Apache Hadoop* e seu HDFS, ao criar pseudo-tabelas estruturadas e permitindo a interação com os dados utilizando a linguagem SQL (*Structured Query Language*) que é um padrão de mercado e suportada por qualquer ferramenta SGBD. A versão que foi instalada no ambiente foi a 1.2.2 e, por se tratar de um software livre e de código-fonte aberto, pode ser baixada gratuitamente em http://hive.apache.org/.

Adicionalmente, o suporte à linguagem R em sua versão 3.5.1 foi adicionado ao ambiente de estudo. O R é uma linguagem criada especificamente para computação estatística amplamente usada por estatísticos e mineradores de dados. Possui pacotes integrados de software que facilitam o trabalho com dados, cálculos e a exibição de gráficos.

3.4 Procedimento

Para a conveniência do estudo os dados foram consolidados, transformados e sintetizados no sistema gerenciador de banco de dados MySQL utilizando procedimentos armazenados (*stored procedures*) criados pelo pesquisador, assemelhando-se a uma abordagem clássica de *Business Intelligence*, cuja etapa conhecida como ETL (acrônimo na língua inglesa para *Extract, Transform* e *Load*, ou extração, transformação e carga) acontece antes dos dados serem carregados em um *Data Warehouse* diferentemente do lago de dados que, segundo Miloslavskaya e Tolstoy (2016), recebe seus dados em forma bruta no lago de dados para transformação posterior. A abordagem da transformação prévia foi escolhida com o intuito de aproveitar conhecimentos prévios do pesquisador na linguagem de programação vigente no SGBD MySQL.

A consolidação foi iniciada a partir dos dados presentes no sistema de provas, eliminando assim eventuais sujeiras nos dados causadas por usuários de teste e mantenedores do ambiente virtual de aprendizagem, conforme alertado por Merceron e Yacef (2005). Optou-se, portanto, em analisar apenas os alunos que realizaram a avaliação somativa ao final do ano letivo. Embora fosse relevante para o estudo analisar eventuais alunos que evadiram durante o ano letivo e por esta razão não realizaram a avaliação somativa, não era possível afirmar que os dados oriundos do sistema de provas foram fornecidos integralmente pela instituição, tornando-se impossível de determinar a diferença entre um aluno evadido e um aluno cujo resultado na avaliação somativa era desconhecido por este trabalho, o que poderia causar uma distorção nos resultados obtidos.

O procedimento técnico pode ser observado no Quadro 13 presente no Apêndice B deste trabalho e é possível observar que dados como a data de nascimento e gênero dos alunos foram obtidos pela carga de dados do sistema de provas e não estavam presentes no AVA. Já que uma das hipóteses a serem testadas é a relevância destes dois atributos (doravante chamados de dimensões) na predição do resultado da avaliação somativa, reforça-se assim a decisão da amostragem ser contida apenas por alunos que realizaram a prova ao final do ano letivo.

Quanto à amostragem utilizada, no início dos trabalhos a premissa era a utilização de dados dos alunos do curso de "Análise e Desenvolvimento de Sistemas" no ano letivo de 2017. Entretanto, observou-se dois fatores: a) as dimensões determinadas para o trabalho não eram aplicáveis apenas ao curso em específico, mas a todos os sete cursos de graduação ministrados totalmente a distância pela instituição e b) o aumento da amostragem de 300 alunos de um curso em específico em 2017 e 2018 para 483 alunos de sete cursos distintos em 2018 alcançaram resultados numericamente melhores e mais relevantes à instituição.

As avaliações formativas dos cursos analisados não fazem parte da carga proveniente do Moodle, pois tais avaliações ocorrem em uma plataforma de avaliação proprietária da instituição e, portanto, apartada. As avaliações são compostas por quarenta questões de múltipla escolha composta por todas as disciplinas do curso. Desta primeira carga oriunda do sistema de provas, definiu-se as seguintes dimensões:

- **Idade:** gerada a partir da data de nascimento do aluno, corresponde à idade do aluno no início do aluno letivo de 2018. O cálculo pode ser observado no início

do Quadro 17 presente no Apêndice B e se trata de um número real, não inteiro (sendo 29,652472 anos uma idade válida para o estudo);

- **Gênero:** gerado a partir da coluna "sexo" com os valores válidos "M" para gênero masculino e "F" para gênero feminino, seus valores válidos foram transformados para números inteiros, sendo "1" para masculino e "2" para feminino, transformação exigida para a linguagem R que pode ser observada no início do Quadro 17 presente no Apêndice B;
- **Índice de Nota:** índice gerado a partir da nota obtida pelo aluno na avaliação somativa, calculado a partir da nota obtida dividida pela nota máxima (atualmente 30). O cálculo foi omitido do Quadro 13 no Apêndice B para preservar as particularidades da infraestrutura da instituição;
- **Resultado:** trata-se da informação a ser prevista pelos algoritmos de classificação e foi calculada a partir da dimensão índice de nota na linguagem R, como pode ser observado em qualquer quadro presente no Apêndice C. Caso o aluno possua um **índice de nota** igual ou superior a 0,6 (o que corresponderia a 60% de aproveitamento na avaliação somativa) o resultado foi considerado "bom", caso contrário, "ruim". É importante observar que a nota final do discente é composta por outras avaliações além desta, sendo assim, um resultado considerado "ruim" não corresponde necessariamente à reprovação do aluno, significando apenas que obteve um resultado abaixo do esperado.

A carga de dados oriunda do Moodle da instituição contém o que o AVA tradicionalmente registra em suas funcionalidades originais além de alguns dados adicionais gerados pelas funcionalidades extras providas pela sua customização e a partir destes dados foram definidas as seguintes dimensões:

- **Índice de pontos:** índice gerado a partir dos pontos obtidos pelo discente em avaliações formativas no período da amostra dividido pelos pontos máximos que poderiam ser obtidos, cálculo que pode ser observado no Quadro 14 do Apêndice B. Por se tratar de alunos de cursos distintos, os pontos máximos no período de amostragem podem variar de um curso para outro, sendo assim, transformar os pontos em um índice real que varia de 0 até 1 foi necessário para equalizar a amostragem;

- **Índice de Entregas:** índice gerado a partir do número de entregas de avaliações formativas efetuadas pelo discente dividido pelo número máximo de entregas possíveis no mesmo período. O cálculo que pode ser observado no Quadro 14 do Apêndice B e, a exemplo do índice de pontos, o número máximo de entregas é distinto de um curso para outro, cabendo à transformação de um índice para equalização da amostra;
- **Índice de Entregas efetuadas com atraso:** índice gerado a partir do número de entregas que o discente efetuou com atraso no período da amostragem dividido pelo número de entregas possíveis, conforme cálculo realizado no Quadro 15 do Apêndice B. Alunos cujo índice tende a zero efetuaram as entregas das avaliações formativas pontualmente;
- **Número de interações realizadas no fórum:** a dimensão corresponde ao número de vezes que o discente realizou alguma postagem no fórum, mostrando sua interação na plataforma, conforme consolidado no Quadro 16 do Apêndice B. Embora não tenha sido transformado em um índice, a função scale() da linguagem R, observada em qualquer quadro do Apêndice C, transformou o dado em um índice entre -1 e 1.
- **Tempo Total em HTML, Vídeos, Áudios e PDF:** as quatro dimensões correspondem ao número de segundos de exposição do discente nos distintos objetos de aprendizagem. O cálculo destas dimensões foi omitido deste trabalho de forma a não expor particularidades do ambiente da instituição, já que este controle não é nativo do ambiente virtual de aprendizagem Moodle. Embora não tenha sido transformado em um índice, a função scale() da linguagem R, observada em qualquer quadro do Apêndice C, transformou os dados em índices entre -1 e 1.

É importante observar que, embora os procedimentos de consolidação dos dados apresentados no Apêndice B constem a identificação única dos alunos, estes por sua vez foram omitidos no procedimento de exportação para a linguagem R, mostrada no Quadro 17 do Apêndice B. Desta forma, os dados foram anonimizados, preservando a privacidade dos alunos cuja identificação única é irrelevante para o estudo. A identificação dos alunos foi mantida na etapa de consolidação para mera conferência, de forma a garantir que procedimentos e cálculos foram realizados corretamente.

Os dados contendo as dez dimensões a serem analisadas e o índice de notas que foi utilizado para determinar o indicador de resultado “bom” e “ruim” foram importados para a linguagem R. Como método de particionamento de dados, foi escolhida a Randomização utilizada por Capen et al (1986) na seguinte proporção: setenta por cento dos dados selecionados aleatoriamente compuseram a amostragem de treinamento para aprendizagem de máquina assistida e os trinta por cento restantes foram designados como amostragem de teste; tais proporções já foram usadas em estudos anteriores, como pode ser visto no trabalho de Cha et al (2006). O particionamento ocorreu na linguagem R e pode ser observado em qualquer quadro presente no Apêndice C.

Sendo assim, no começo de cada procedimento na linguagem R a amostragem de 483 alunos foi particionada em 70/30. A partição de 70% da amostragem, contendo o indicador “bom” e “ruim” é submetido a um algoritmo de classificação que aprende a classificar os alunos com esses indicadores, o que é chamado de aprendizagem de máquina assistida. Na segunda parte dos procedimentos em R, os 30% restantes das amostras (128 alunos) foram submetidos ao modelo gerado pelo classificador sem o indicador “bom” ou “ruim”, pois o objetivo é justamente prever este indicador. Por fim, o indicador previsto é confrontado com o indicador real obtido por esses alunos, apresentado em uma matriz de confusão e um percentual de acuidade do algoritmo. Para viabilizar os testes comparativos, a amostragem randomizada é a mesma em todas as execuções dos procedimentos.

Uma vez que o indicador a ser previsto possui apenas dois estados (“bom” e “ruim”), uma escolha arbitrária por um destes resultados, semelhante ao lançar uma moeda de duas faces, determina uma taxa de acerto mínima de 50%, assim sendo, quaisquer resultados abaixo deste percentual são considerados insatisfatórios. Em estudos semelhantes a este, como Santana e colaboradores (2015) e Silva e colaboradores (2015), foram obtidas taxas de acerto na ordem de 74%, o que foi considerado por seus pesquisadores uma taxa satisfatória, assim sendo, estipula-se este número como patamar a ser alcançado por este estudo.

Na primeira parte do estudo, a amostragem corresponde ao mês de fevereiro de 2018, utilizando as dez dimensões definidas é submetida a quatro algoritmos de classificação diferentes: Árvore de Decisão, *Naïve Bayes*, *K-nearest neighbor* (KNN) e *Support Vector Machine* (SVM) de modo a determinar o classificador mais indicado para o problema proposto. A hipótese é que o percentual de acuidade aumente na ordem em que os classificadores foram apresentados, dos algoritmos mais simples (começando pelo Árvore de Decisão) aos mais complexos (finalizando pelo SVM).

Na segunda parte do estudo, as dez dimensões foram submetidas a amostragens de períodos distintos, variando de fevereiro de 2018 até junho de 2018, de forma cumulativa (ou seja, a amostragem de junho possui um acumulado de primeiro semestre letivo completo, conforme pode ser observado no cálculo do Quadro 17 no Apêndice B). A hipótese a ser testada é se a acuidade do algoritmo de classificação aumenta conforme a amostragem se aproxima da avaliação somativa.

Na terceira e última parte do estudo, a amostragem de fevereiro de 2018 é submetida ao classificador mais indicado, variando a presença dos dez indicadores. Esses foram removidos um a um, de forma a determinar quais são relevantes e quais não são para a acuidade do algoritmo de classificação. A hipótese levantada é que todas as dimensões definidas são relevantes ao cálculo do algoritmo.

CAPÍTULO 4 – ANÁLISE E DISCUSSÕES

Na primeira parte do estudo, a mesma amostragem de fevereiro de 2018 foi submetida aos quatro algoritmos de classificação apresentados neste estudo. O primeiro a ser testado foi a técnica da árvore de decisão, cujo procedimento técnico é apresentado na íntegra no Quadro 18 do Apêndice C. O Quadro 3 é a representação do modelo obtido pelo treinamento realizado pelos 70% da amostragem (partição de treinamento).

Quadro 3 – Árvore de decisão representada em modo texto

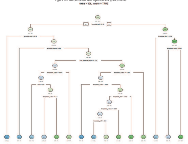
```
n= 355

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 355 158 ruim (0.4450704 0.5549296)
    2) tempototal_pdf< -0.1530774 262 129 ruim (0.4923664 0.5076336)
      4) tempototal_pdf>=-0.1531246 7   0 bom (1.0000000 0.0000000)
*
      5) tempototal_pdf< -0.1531246 255 122 ruim (0.4784314
0.5215686)
       10) tempototal_audios>=0.3014984 12   2 bom (0.8333333
0.1666667) *
       11) tempototal_audios< 0.3014984 243 112 ruim (0.4609053
0.5390947)
         22) num_interacoes_forum>=-0.2847319 107  49 bom (0.5420561
0.4579439)
           44) tempototal_videos< -0.07689081 68  23 bom (0.6617647
0.3382353)
             88) idade< 0.3595957 48  12 bom (0.7500000 0.2500000) *
             89) idade>=0.3595957 20   9 ruim (0.4500000 0.5500000)
              178) tempototal_audios>=-0.1969168 7   2 bom
(0.7142857 0.2857143) *
              179) tempototal_audios< -0.1969168 13   4 ruim
(0.3076923 0.6923077) *
           45) tempototal_videos>=-0.07689081 39  13 ruim (0.3333333
0.6666667) *
         23) num_interacoes_forum< -0.2847319 136  54 ruim
(0.3970588 0.6029412)
           46) tempototal_videos>=-0.08091537 54  26 bom (0.5185185
0.4814815)
             92) tempototal_videos< -0.07890213 10   1 bom
(0.9000000 0.1000000) *
             93) tempototal_videos>=-0.07890213 44  19 ruim
(0.4318182 0.5681818)
              186) tempototal_html< -0.07200723 34  17 bom
(0.5000000 0.5000000)
                372) tempototal_videos>=-0.07551272 26  11 bom
(0.5769231 0.4230769)
                  744) tempototal_pdf>=-0.1532353 8   1 bom
(0.8750000 0.1250000) *
                  745) tempototal_pdf< -0.1532353 18   8 ruim
(0.4444444 0.5555556) *
                373) tempototal_videos< -0.07551272 8   2 ruim
```

```
(0.2500000 0.7500000) *
             187) tempototal_html>=-0.07200723 10   2 ruim
(0.2000000 0.8000000) *
          47) tempototal_videos< -0.08091537 82  26 ruim (0.3170732
0.6829268) *
    3) tempototal_pdf>=-0.1530774 93  29 ruim (0.3118280 0.6881720)
      6) tempototal_html< -0.07416276 12   3 bom (0.7500000
0.2500000) *
      7) tempototal_html>=-0.07416276 81  20 ruim (0.2469136
0.7530864)
       14) tempototal_audios>=0.4148729 8   3 bom (0.6250000
0.3750000) *
       15) tempototal_audios< 0.4148729 73  15 ruim (0.2054795
0.7945205) *
```

Fonte: Elaborado pelo autor (2019)

O mesmo modelo obtido foi representado na Figura 6 utilizando uma das bibliotecas de plotagem da linguagem R.

Figura 6 – Árvore de decisão representada graficamente



Fonte: Elaborado pelo autor (2019)

Por fim, a predição foi realizada utilizando na partição de teste utilizando o modelo gerado e, para efeito comparativo foi gerada a matriz de confusão utilizando a função confusionMatrix() de uma das bibliotecas da linguagem R. Além da matriz de confusão, a função em questão gera uma série de indicativos, e o que será comparado entre os classificadores é o *Accuracy*, que indica o percentual de acuidade alcançado. O Quadro 4 apresenta os resultados obtidos, e é possível constatar uma taxa de acerto de 60,94% pela técnica de árvore de decisão. Ao observar a matriz de confusão, é possível observar que o número de falsos positivos (28) é ligeiramente maior que o número de falsos negativos (22).

Quadro 4 – Matriz de confusão e taxa de acerto resultantes da árvore de decisão

```
Confusion Matrix and Statistics

          Reference
Prediction bom ruim
      bom   28   28
      ruim  22   50

               Accuracy : 0.6094
                 95% CI : (0.5192, 0.6944)
    No Information Rate : 0.6094
    P-Value [Acc > NIR] : 0.5387

                  Kappa : 0.1968
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.5600
            Specificity : 0.6410
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.6944
             Prevalence : 0.3906
         Detection Rate : 0.2188
   Detection Prevalence : 0.4375
      Balanced Accuracy : 0.6005

       'Positive' Class : bom
```

Fonte: Elaborado pelo autor (2019)

O segundo algoritmo de classificação testado foi o Naïve Bayes, cujo procedimento técnico em linguagem R é apresentado na íntegra no Quadro 19 do Apêndice C. No Quadro 5 apresentado é possível observar o modelo do classificador Naïve Bayes sumarizado em modo texto.

Quadro 5 – Modelo do classificador Naïve Bayes representado em modo texto

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, type = "class")

A-priori probabilities:
```

```
Y
      bom      ruim
0.4450704 0.5549296

Conditional probabilities:
      idade
Y             [,1]      [,2]
  bom   0.02315861 1.0060074
  ruim -0.01857391 0.9973314

      genero
Y            [,1]      [,2]
  bom  -0.1329911 0.8718269
  ruim  0.1066629 1.0823994

      entregas_index
Y             [,1]     [,2]
  bom  -0.01666513 0.988908
  ruim  0.01336594 1.011123

      pontos_index
Y             [,1]      [,2]
  bom  -0.02087328 0.9799557
  ruim  0.01674101 1.0179681

      tempototal_html
Y             [,1]      [,2]
  bom  -0.04211561 0.2658836
  ruim  0.03377800 1.3217104

      tempototal_videos
Y             [,1]       [,2]
  bom  -0.06531528 0.03724427
  ruim  0.05238485 1.34119878

      tempototal_audios
Y             [,1]      [,2]
  bom   0.04501000 1.1121497
  ruim -0.03609939 0.9013411

      tempototal_pdf
Y             [,1]      [,2]
  bom  -0.06255264 0.3936731
  ruim  0.05016912 1.2947191

      entregas_com_atraso
Y             [,1]      [,2]
  bom  -0.04043331 0.8697704
  ruim  0.03242874 1.0944292

      num_interacoes_forum
Y             [,1]      [,2]
  bom   0.01534427 0.8220804
  ruim -0.01230657 1.1244723
```

Fonte: Elaborado pelo autor (2019)

No Quadro 6 observamos a matriz de confusão e taxa de acerto deste classificador: além de uma taxa de acerto consideravelmente inferior (51,56%), a matriz de confusão mostra que o classificador foi ineficaz na identificação dos alunos cujo indicador na avaliação somativa foi considerado “bom”, um único acerto e um elevado número de falsos positivos (13) e falsos negativos (49).

Quadro 6 – Matriz de confusão e taxa de acerto resultantes do classificador Naïve Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction bom ruim
      bom    1   13
      ruim  49   65

               Accuracy : 0.5156
                 95% CI : (0.4257, 0.6048)
    No Information Rate : 0.6094
    P-Value [Acc > NIR] : 0.9876

                  Kappa : -0.1684
 Mcnemar's Test P-Value : 8.789e-06

            Sensitivity : 0.020000
            Specificity : 0.833333
         Pos Pred Value : 0.071429
         Neg Pred Value : 0.570175
             Prevalence : 0.390625
         Detection Rate : 0.007812
   Detection Prevalence : 0.109375
      Balanced Accuracy : 0.426667

       'Positive' Class : bom
```

Fonte: Elaborado pelo autor (2019)

O terceiro algoritmo de classificação testado foi o *K-Nearest Neighbor (Knn)*, cujo procedimento técnico na linguagem R é apresentado na íntegra no Quadro 20 do Apêndice C. O procedimento foi realizado algumas dezenas de vezes para determinar o valor de k que apresentava o melhor resultado, e o melhor valor testado é k=71. O Quadro 7 mostra o resultado obtido quando k=71, atingindo uma taxa de acerto de 67,97%, falsos positivos em 9 ocorrências e falsos negativos em 32 ocorrências.

Quadro 7 – Matriz de confusão e taxa de acerto resultantes do classificador *K-Nearest Neighbor (Knn)*

```
Confusion Matrix and Statistics

          Reference
Prediction bom ruim
      bom   18    9
      ruim  32   69
```

```
               Accuracy : 0.6797
                 95% CI : (0.5915, 0.7594)
    No Information Rate : 0.6094
    P-Value [Acc > NIR] : 0.0604784

                  Kappa : 0.2666
 Mcnemar's Test P-Value : 0.0005908

            Sensitivity : 0.3600
            Specificity : 0.8846
         Pos Pred Value : 0.6667
         Neg Pred Value : 0.6832
             Prevalence : 0.3906
         Detection Rate : 0.1406
   Detection Prevalence : 0.2109
      Balanced Accuracy : 0.6223

       'Positive' Class : bom
```

Fonte: Elaborado pelo autor (2019)

Por fim, o último algoritmo de classificação testado foi o *Support Vector Machine* que, assim como os três primeiros, é apresentado na íntegra no Quadro 21 do Apêndice C. O procedimento foi repetido algumas dezenas de vezes e fez uso de uma função de tune() na linguagem R, de modo a determinar os melhores valores para os parâmetros "*gamma*" e "*cost*". O Quadro 8 exibe os resultados quando *cost*=1 e *gamma*=2, parâmetros que apresentaram os melhores resultados: SVM obteve uma taxa de acerto de 64,06%, 14 ocorrências nos falsos positivos e 32 ocorrências em falsos negativos.

Quadro 8 – Matriz de confusão e taxa de acerto resultantes do classificador *Support Vector Machine (SVM)*

```
Confusion Matrix and Statistics

          Reference
Prediction bom ruim
      bom   18   14
      ruim  32   64

               Accuracy : 0.6406
                 95% CI : (0.5511, 0.7235)
    No Information Rate : 0.6094
    P-Value [Acc > NIR] : 0.26440

                  Kappa : 0.193
 Mcnemar's Test P-Value : 0.01219

            Sensitivity : 0.3600
            Specificity : 0.8205
         Pos Pred Value : 0.5625
         Neg Pred Value : 0.6667
             Prevalence : 0.3906
         Detection Rate : 0.1406
   Detection Prevalence : 0.2500
```

```
       Balanced Accuracy : 0.5903

        'Positive' Class : bom
```

Fonte: Elaborado pelo autor (2019)

A Tabela 2 sumariza os resultados da primeira parte do estudo e é possível concluir que o algoritmo de classificação *K-nearest neighbor* obteve uma ligeira vantagem em relação ao *Support Vector Machine,* obtendo os melhores resultados. A primeira hipótese que afirma que classificadores robustos obteriam os melhores resultados foi confirmada, embora a expectativa que SVM superasse KNN não foi confirmada.

Tabela 2 - Resultados obtidos utilizando diferentes algoritmos de classificação

Algoritmo preditivo	**Taxa de acerto**	**Falsos positivos**	**Falsos negativos**
Árvore de decisão	60,94%	28	22
Naïve Bayes	51,56%	13	49
K-nearest neighbor (KNN)	67,97%	9	32
Support Vector Machine (SVM)	64,06%	14	32

Fonte: Elaborado pelo autor (2019)

Na segunda parte do estudo, as dez dimensões foram submetidas aos algoritmos de classificação *K-nearest neighbor* (com k=71) e *Support Vector Machine* (com gamma=2 e cost=1) aumentando o período da amostragem, começando em fevereiro de 2018 indo até junho de 2018, de forma cumulativas. As taxas de acerto obtidas podem ser observadas na Tabela 3.

Tabela 3 - Resultados obtidos em períodos de amostragem diferentes

Algoritmo preditivo	**Taxas de acerto**				
	Fev/2018	**Mar/2018**	**Abr/2018**	**Mai/2018**	**Jun/2018**
K-Nearest Neighbor	67,97%	58,59%	57,03%	56,25%	53,91%
Support Vector Machine	64,06%	53,91%	55,47%	57,81%	61,72%

Fonte: Elaborado pelo autor (2019)

Observa-se uma drástica queda na taxa de acerto dos dois algoritmos, refutando a hipótese que quanto mais perto da avaliação somativa, maior seria a acuidade da predição. Há um fator positivo neste resultado, pois observa-se ser possível classificar com uma taxa de acerto próxima aos 70% logo nos primeiros 30 dias do curso, indo ao encontro do objetivo de

identificar os alunos cuja tendência de baixo rendimento é maior, oferecendo a estes um auxílio de maneira proativa.

O alto desempenho nos primeiros trinta dias de curso é resultado da maior variância entre as dez dimensões selecionadas: o número de interações no fórum, assim como o número de entregas não efetuadas ou feitas com atraso é maior no início dos cursos. Nos meses seguintes, as variações entre estas dimensões são menores e a amostragem se torna mais uniforme, o que explicaria a maior dificuldade dos algoritmos em classificar os alunos.

Na terceira parte do estudo, as dimensões foram removidas uma a uma, observando-se a taxa de acerto obtida utilizando o algoritmo de classificação melhor sucedido, o *K-Nearest Neighbor*, utilizando k=71. A Tabela 4 mostra um comparativo entre a relevância de cada uma das dimensões definidas no estudo.

Tabela 4 - Resultados obtidos pelo KNN utilizando dimensões diferentes

Dimensões utilizadas	**Taxa de acerto (Fev/2018)**
Todas as dez dimensões	67,97%
Excetuando-se gênero	62,50%
Excetuando-se idade	55,47%
Excetuando-se gênero e idade	51,56%
Excetuando-se número de interações no fórum	58,59%
Excetuando-se índice de entregas em atraso	67,19%
Excetuando-se índice de pontos	62,50%
Excetuando-se índice de entregas	57,81%
Excetuando-se exposição ao HTML	69,53%
Excetuando-se exposição aos vídeos	67,97%
Excetuando-se exposição aos áudios	64,84%
Excetuando-se exposição aos PDFs	64,84%

Fonte: Elaborado pelo autor (2019)

Observa-se que as dimensões mais relevantes para a predição são: a) a idade do estudante, b) o índice de entregas e c) o número de interações no fórum, ordenados por ordem

de relevância. Os itens “b” e “c”, conforme discutidos anteriormente, são os que apresentam a maior variação no primeiro mês de curso, tornando-se mais uniformes nos meses seguintes.

Além disso, a exposição do discente aos objetos de aprendizagem foram as dimensões que menos contribuíram para a taxa de acerto: enquanto a exposição aos áudios e PDFs foram pouco significativas, o índice de exposição aos vídeos foi irrelevante e o índice de exposição ao HTML “sujou” o algoritmo preditivo, pois sua ausência fez com que a taxa de acerto aumentasse, chegando próxima aos 70%. A Tabela 5 apresenta um comparativo do algoritmo de classificação *K-Nearest Neighbor* sendo k=71 utilizando todas as dimensões no primeiro semestre letivo, e sem índices de exposição ao HTML e vídeos.

Tabela 5 - Resultados obtidos pelo KNN utilizando dimensões e períodos de amostragem diferentes

Dimensões utilizadas	**Taxas de acerto**				
	Fev/2018	**Mar/2018**	**Abr/2018**	**Mai/2018**	**Jun/2018**
Todas as dimensões	67,97%	58,59%	57,03%	56,25%	53,91%
Excetuando-se exposição ao HTML	69,53%	60,16%	58,59%	56,25%	54,69%
Excetuando-se exposição ao HTML e vídeos	69,53%	60,16%	61,72%	57,03%	57,81%

Fonte: Elaborado pelo autor (2019)

Sendo assim, é possível refutar a última hipótese que afirmava que todas as dez dimensões eram relevantes ao algoritmo preditivo pois, como pode ser observado na última parte do estudo, a ausência dos índices de exposição aos objetos de aprendizagem em HTML e vídeos torna o algoritmo preditivo mais eficiente inclusive a longo prazo.

Estes resultados dos objetos de aprendizagem, somando aos altos números de falsos negativos apresentados por todos os algoritmos de classificação, indicam que a predição falha em identificar discentes que já possuem conhecimento prévio e discentes que são expostos a outros objetos de aprendizagem externos ao ambiente virtual de aprendizagem pois, nestes casos, os discentes apresentam bons resultados nas avaliações somativas mesmo com baixos índices de exposição aos objetos de aprendizagem.

CONSIDERAÇÕES FINAIS

O avanço tecnológico provê ferramentas importantes que aumentam a eficiência de empresas bilionárias ao transformar dados em informação, com investimentos que chegam a dezenas de milhões de dólares. Este estudo visa contribuir para a conscientização do fato de que instituições de ensino podem se beneficiar muito das mesmas ferramentas utilizadas pelas grandes corporações, tornando-se mais eficiente na missão de desenvolver habilidades e conhecimentos de seus alunos e, por consequência, diminuir a evasão e melhorando sua saúde financeira.

Entretanto, algumas ponderações devem ser realizadas. Este estudo concluiu que a idade é uma dimensão importante na predição dos dados, pois o comportamento de estudo é distinto em diferentes faixas etárias. Isso não significa, no entanto, que estes comportamentos não possam sofrer mudanças em um futuro próximo. Independentemente do algoritmo preditivo escolhido o aprendizado de máquina deve ser constante, permitindo que o modelo preditivo seja aperfeiçoado e com a inserção de dados atualizados, mitigando uma análise estereotipada.

Outra conclusão importante deste estudo é que quanto maior for a amostragem, melhores serão os resultados obtidos. A mudança do número de alunos acompanhados, de 300 alunos de um único curso para 483 alunos em cursos distintos foi determinante para a obtenção de melhores resultados.

O estudo não atingiu o patamar de taxa de acerto de 73% estudos semelhantes e uma das razões é o fato da amostragem utilizada ser pequena, fazendo a taxa de acerto oscilar em até dez pontos percentuais, a depender da partição de treinamento utilizada. Assim sendo, um número maior de alunos acompanhados resultará em taxas de acerto mais estáveis.

Além disso, o alto número de falsos negativos obtido neste estudo indica pontos de melhoria a serem almejados. A criação de um índice de conhecimento prévio do aluno, a ser determinado a partir de uma avaliação diagnóstica ou ainda que uma entrevista prévia com o discente pode ser determinante na redução de falsos negativos e obtenção de taxas de acerto mais expressivas.

Sugere-se que as instituições de ensino que façam uso do mesmo ambiente virtual de aprendizagem possam adaptar os algoritmos preditivos utilizados neste estudo e realizem suas próprias predições, auxiliando no processo de ensino-aprendizagem de seus alunos e, por

consequência, diminuir seus índices de evasão. A instituição de ensino sujeita ao estudo manifestou interesse em adotar e evoluir o algoritmo apresentado neste estudo.

O produto gerado por este trabalho é um algoritmo preditivo escrito em Linguagem R para fins de avaliação dos alunos em cursos a distância que utilizam a plataforma Moodle. Isso gerou um software intitulado “Sistema para predição de avaliação formativa no ensino a distancia em curso superior”, devidamente registrado no Instituto Nacional da Propriedade Industrial (INPI) sob o número BR512019000915-0 conforme apresentado no Anexo B.

Por fim, conclui-se que há um grande potencial na análise dos dados presentes em um ambiente virtual de aprendizagem, pois a identificação precoce de alunos em dificuldade pode resultar em ações proativas por parte de tutores da instituição de ensino que podem aperfeiçoar o processo de ensino e aprendizagem. Embora o Moodle seja um excelente ambiente virtual de aprendizagem, não possui ferramentas analíticas em sua instalação padrão, cabendo às instituições de ensino que optem por seu uso a personalização da ferramenta para alcançar este objetivo. Para tal, os *softwares* a serem utilizados são gratuitos e de código-fonte aberto, assim sendo, as instituições de ensino devem realizar seus investimentos em infraestrutura e mão-de-obra qualificada, que podem começar baixos em seu início para a criação de uma prova de conceito com centenas de alunos, e podem ser ampliados na medida em que os algoritmos de predição mostrem resultados. Por esta razão, espera-se que este estudo traga contribuições para que outras instituições de ensino busquem soluções semelhantes a essa, melhorando a educação no Brasil como um todo.

REFERÊNCIAS

ABED. **Censo EAD Brasil 2016 - Relatório Analítico de Aprendizagem a Distância no Brasil**, 2017. Disponível em: <http://abed.org.br/censoead2016/Censo_EAD_2016_portugues.pdf>. Acesso em: 2 set. 2018.

AJANKI, A. **Knn Classification**. Disponível em: <https://en.wikipedia.org/wiki/File:KnnClassification.svg>. Acesso em: 17 mar. 2019.

ALMEIDA, M. E. B. de. Educação a distância na internet: abordagens e contribuições dos ambientes digitais de aprendizagem. **Educação e Pesquisa**, v. 29, n. 2, p. 327–340, 2003. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-97022003000200010&lng=pt&tlng=pt>. Acesso em: 5 set. 2018.

ANDRADE, A. F. de; VICARI, R. M. Construindo um ambiente de apren- dizagem a distância inspirado na concepção sociointeracionista de Vygotsky. In: **SILVA, Marco (Org.). Educação online: teorias, práticas, legislação, formação corporativa.** 3. ed. ed. São Paulo: Ed. Loyola, 2011. p. 259-260–261.

ARETIO, J. **Un concepto integrador de ensenansa a distância. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 8º, 1990. Caracas. Anais.**Caracas:ICDE, 1990.

AZZOLINO, A. P.; NABARRETTI, C. P. Gestão nas IES privadas: capacitação do corpo docente e os paradigmas das tecnologias da informação e comunicação no ensino a distância. **Revista de Ciências Gerenciais**, v. XII, n. nº. 16, 2008.

BAKER, R. S. J. D.; YACEF, K. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–16, 2009.

BARBOSA, C. M. A. M. A aprendizagem mediada por TIC: interação e cognição em perspectiva. **RBAAD - Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 11, 2012. Disponível em: <http://www.abed.org.br/revistacientifica/Revista_PDF_Doc/2012/artigo_07_v112012.pdf>. Acesso em: 5 set. 2018.

BEAULIEU, A. Mediating ethnography: Objectivity and the making of ethnographies of the internet. **Social Epistemology**, v. 18, n. 2–3, p. 139–163, 2004.

BECKER, R. A.; CHAMBERS, J. M.; WILKS, A. R. **The New S Language**. Pacific Grove: Wadsworth, 1988.

BELLONI, M. L. Ensaio sobre a educação a distância no Brasil. **Educação & Sociedade**, v. 78, p. 117–42, 2002.

BELLONI, M. L. **Educação a Distância**. 5a. edição ed. Campinas: Autores Associados, 2009.

BENSON, A. D. Assessing participant learning in online environments. **New Directions for Adult and Continuing Education**, n. 100, p. 69–78, 2003.

BRASIL. **Decreto nº 9.057, de 25 de maio de 2017. Regulamenta o art. 80 da Lei nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional. Diário Oficial, Brasília, DF, 26 mai 2017. Seção 1, p. 3.**

BUCKLAND, S.T. & ELSTON, D. A. Empirical models for the spa- tial distribution of wildlife. **Journal of Applied Ecology**, n. 30, p. 478–495, 1993.

C. SANTANA, L.; M. A. MACIEL, A.; L. RODRIGUES, R. Avaliação do Perfil de Uso no Ambiente Moodle Utilizando Técnicas de Mineração de Dados. **Anais do XXV Simpósio Brasileiro de Informática na Educação (SBIE 2014)**, v. 1, n. Cbie, p. 269, 2015.

CAMPOS, F.; ROCHA, A. R. **Design instrucional econstrutivismo: em busca de modelos para o desenvolvimento de software**IV Congressso RIBIE, Brasilia, 1998. Disponível em: <http://www.niee.ufrgs.br/eventos/RIBIE/1998/pdf/com_pos_dem/250M.pdf>. Acesso em: 12 set. 2018.

CAPEN, D.E., FENWICK, J.W., INKLEY, D.B. & BOYNTON, A. C. Multivariate models of songbird habitat in New England forests. In: J.A. VERNER, M. L. M. AND C. J. R. (Ed.). **Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates**. Madison: University of Wisconsin Press, 1986. p. 171–175.

CHA, H. J.; KIM, Y. S.; PARK, S. H.; YOON, T. B.; JUNG, Y. M.; LEE, J. H. Learning styles diagnosis based on user interface behaviors for the customization of learning interfaces in an intelligent tutoring system. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 4053 LNCS, p. 513–524, 2006.

CORREIA, R. L.; SANTOS, J. G. dos. A Importância da Tecnologia da Informação e Comunicação (TIC) na Educação a Distância (EAD) do Ensino Superior (IES). **Revista Aprendizagem em EAD**, v. 2, n. 1996, p. 1–16, 2013.

D'ÁVILA, C. Por uma didática colabo- rativa no contexto das comunidades virtuais de aprendizagem. In: **SANTOS, Edméa; ALVES, Lynn (Org.). Práticas pedagógicas e tecnologias digitais.** Rio de Janeiro: E-papers, 2006. p. 91-98–100.

DANOWSKI, J. A.; EDISON-SWIFT, P. Crisis effects on intraorganizational computer-based communication. **Communication Research**, v. 12, n. 2, p. 251–270, 30 abr. 1985. Disponível em: <http://journals.sagepub.com/doi/10.1177/009365085012002005>. Acesso em: 3 out. 2018.

DEPRESBITERIS, L.; TAVARES, M. R. **Diversificar é preciso...: Instrumentos e técnicas de avaliação de aprendizagem**. 1a. edição ed. São Paulo: Senac São Paulo, 2017.

DHANABAL, S.; CHANDRAMATHI, S. A Review of various k-Nearest Neighbor Query Processing Techniques. **International Journal of Computer Applications**, v. 31, n. 7, p. 975–8887, 2011. Disponível em: <https://pdfs.semanticscholar.org/ea1c/116ff0700b9250ef01315a94366ce8bf753c.pdf>. Acesso em: 17 mar. 2019.

EDM. **Education Data Mining Website**. Disponível em: <www.educationaldatamining.org>. Acesso em: 30 set. 2018.

FERNANDES, A. P. L. M.; FERNANDES, R. R. **A Importância das TICs como Recurso Didático no Ensino da Matemática Financeira.**Simpósio de Excelência em Gestão e Tecnologia: gestão, inovação e tecnologia para sustentabilidade. IX SEGeT 2012, pp.1-10, 2012.

FIELDING, A.H. & HAWORTH, P. F. Testing the generality of bird-habitat models. **Conservation Biology**, n. 9, p. 1466–1481, 1995.

FIELDING, A. H.; BELL, J. F. !!!!A review of methods for the assessment of prediction errors in PB models - Copy.pdf. v. 24, n. 1, p. 38–49, 2017.

FRANCIOSI, B.; MEDEIROS, M.; COLLA, A. Caos, criatividade e ambientes de aprendizagem. **Educação a Distância**, p. 2002–2003, 2003. Disponível em: <http://www.ead.pucrs.br/biblioteca/artigo/RESUMO_CAOS.PDF>. Acesso em: 5 set. 2018.

GILBERTO, I. J. L. A função docente e a prática pedagógica na Educação a Distância. **Educação & Linguagem**, v. 17, n. 2, p. 89–108, 2014.

HAERTEL, E. H. Performance Assessment and Educational Reform. **Phi Delta Kappan**, v. 80, n. 9, p. 662–666, 1999.

HEINER, C.; HEFFERNAN, N.; BARNES, T. Educational data mining Workshop. **Workshop of Educational Data Mining**, n. July, 2007. Disponível em: <http://web.cs.wpi.edu/~nth/pubs_and_grants/papers/2007/AIED-EDM/AIED-EDM_proceeding_full2.pdf>. Acesso em: 20 nov. 2018.

IBM SOFTWARE GROUP. **Analytics for Achievement: Understand success and boost performance in primary and secondary education**. [s.l: s.n.]. Disponível em: <ftp://public.dhe.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_analytics_for_achievement_understand_success_and_boost_performance_in_education.pdf>. Acesso em: 20 nov. 2018.

IHAKA, R.; GENTLEMAN, R. R : A Language for Data Analysis and Graphics. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 299–314, 1996. Disponível em: <http://www.jstor.org/stable/pdf/1390807.pdf?acceptTC=true>. Acesso em: 12 mar. 2019.

IVASHITA, S. B.; COELHO, M. P. EaD: O importante papel do professor-tutor. In: IX Congresso Nacional de Educação - EDUCERE III Encontro Sul Brasileiro de Psicopedagogia, **Anais**. 2009.

KENSKI, V. M. Novas tecnologias na educação presencial e a distância I. In: **In: BARBOSA, Raquel Lazzari Leite (Org.). Formação de Educadores: desafios e perspectivas.** São Paulo: UNESP, 2003.

KOTSIANTIS, S. B. Supervised Machine Learning: A Review of Classification Techniques. **Informatica**, v. 31, p. 249–268, 2007.

KOZINETS, R. V. **Netnografia: Realizando Pesquisa Etnográfica Online**. 1a. edição ed. Porto Alegre: Penso, 2014.

LAGUARDIA, J.; PORTELA, M. C.; VASCONCELLOS, M. M. Avaliação em ambientes virtuais de aprendizagem. **Educação e Pesquisa**, v. 33, n. 3, p. 513–530, 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-97022007000300009&lng=pt&tlng=pt>. Acesso em: 20 nov. 2018.

LANGHI, C. **Materiais instrucionais para o ensino à distância: estudo sobre a aplicação da teoria de aprendizagem significativa de Ausubel na produção de conteúdos para cursos via Internet**. 2005. Tese (Doutoramento em Psicologia). Instituto de Psicologia da USP, Universidade de São Paulo, São Paulo, 2005.

LEVY, P. **Cibercultura.** Rio de Janeiro: Editora 34, 1999.

LÉVY, P. **As Tecnologias da Inteligência – O Futuro do Pensamento na Era da Informática**. 10a. ediçã ed. São Paulo: Editora 34, 2004.

LIBANEO, J. C. **Didática**. São Paulo: Cortez Editora, 2017.

LUCKESI, C. C. **Avaliação da aprendizagem escolar: Estudos e proposições**. São Paulo: Cortez Editora, 2014.

MAIA, M. de C. O uso da tecnologia de informação para a educação a distância no ensino superior. **Higher Education**, p. 294, 2004. Disponível em: <http://bibliotecadigital.fgv.br/dspace/handle/10438/2463>. Acesso em: 5 set. 2018.

MAIA, M. de C.; MEIRELLES, F. de S. Educação a distância e o Ensino Superior no Brasil. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 2, p. 1–8, 2003. Disponível em: <http://seer.abed.net.br/index.php/RBAAD/article/view/131>. Acesso em: 5 set. 2018.

MARVIN, L.-E. Spoof, Spam, Lurk, and Lag: the Aesthetics of Text-based Virtual Realities. **Journal of Computer-Mediated Communication**, v. 1, n. 2, p. 0–0, 23 jun. 2006. Disponível em: <https://academic.oup.com/jcmc/article/4584321>. Acesso em: 3 out. 2018.

MATURANA, H. **Emociones y Lenguaje en Educacion y Politica**. Santiago: Hachete, 1990.

MENDONÇA, J. R. C. et al. **Competências Eletrônicas de Professores para Educação a Distância no Ensino Superior no Brasil: discussão e proposição de modelo de análise.** 2013. Pernambuco: Universidade Federal de Pernambuco, 2013.

MERCADO, L. P. L.; FIGUEREDO, L. K. de A.; JOBIM, D. R. de B. Formação de tutores do curso piloto de administração a distância da universidade aberta do Brasil. In: **MERCADO, Luís Paulo Leopoldo (Org.). Práticas de formação de professores na educação a distância**. Maceió: EDUFAL, 2008. p. 98.

MERCERON, A.; YACEF, K. Educational data mining: A case study. **Artificial intelligence in education: Supporting learning through intelligent and sociall informed technology**, p. 467–474, 2005. Disponível em: <http://sydney.edu.au/engineering/it/~kalina/publis/merceron_yacef_aied05.pdf>. Acesso em: 20 nov. 2018.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big Data, Fast Data and Data Lake Concepts. **Procedia Computer Science**, v. 88, p. 300–305, 2016. Disponível em: <http://dx.doi.org/10.1016/j.procs.2016.07.439>. Acesso em: 12 mar. 2019.

MORAES, M. C. **O paradigma educacional emergente**. Campinas: Papirus Editora, 1997.

MORAN, J. A integração das tecnologias na educação. In: **A Educação que desejamos: novos desafios e como chegar lá**. 5a. edição ed. Campinas: Papirus, 2013. p. 89–90.

MORAN, J. M. Como utilizar a Internet na educação. **Ciência da Informação**, v. 26, n. 2, p. 146–153, maio 1997. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651997000200006&lng=pt&tlng=pt>. Acesso em: 19 mar. 2019.

MORRISON, M.L., TIMOSS, I.C. & WITH, K. A. Development and testing linear regression models predicting bird-habitat re- lationships. **Journal of Wildlife Management**, n. 51, p. 247–253, 1987.

NASIRI, M.; MINAEI, B.; VAFAEI, F. Predicting GPA and academic dismissal in LMS using educational data mining: A case mining. **3rd International Conference on eLearning and eTeaching, ICeLeT 2012**, n. January 2015, p. 53–58, 2012.

OSBORNE, P.E. & TIGAR, B. J. Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. **Journal of Applied Ecology**, n. 29, p. 55–62, 1992.

OTSUKA, J. L.; ROCHA, H. V. da. **Avaliação formativa em ambientes de EaD.** XIII Simpósio Brasileiro de Informática na Educação – SBIE – UNISINOS, 2002. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/174/160>. Acesso em: 2 set. 2018.

PACCAGNELLA, L. Getting the Seat of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities. **Journal of Computer-Mediated Communication**, v. 3, n. 1, p. 0, 1997. Disponível em: <http://jcmc.indiana.edu/vol3/issue1/paccagnella.html>. Acesso em: 20 nov. 2018.

PARREIRA, A.; SILVA, A. L. da. A lógica complexa da avaliação. **Ensaio: aval. pol. públ. Educ., Rio de Janeiro, p.367-388**, v. 23, n. 87, 2015.

PATEL, S. **Chapter 2 : SVM (Support Vector Machine) — Theory**. Disponível em: <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>. Acesso em: 17 mar. 2019.

PATIL, T. R.; SHEREKAR, M. . Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. **International Journal Of Computer Science And Applications**, v. 6, n. 2, p. 256–261, 2013.

PERAYA, D. O ciberespaço: um dispositivo de comunicação e de formação midiatizada. In: **In: ALAVA, S. Ciberespaço e formações abertas: rumo a novas práticas educacionais.** Porto Alegre: Artmed, 2002.

PERRY, G. T.; TIMM, M. I.; FERREIRA, R. C. F.; SCHNAID, F.; ZARO, M. A. Desafios da gestão de EAD: necessidades específicas para o ensino científico e tecnológico. **RENOTE: Revista Novas Tecnologias na Educação. CINTED. Universidade Federal do Rio Grande do Sul.**, v. 4, n. Nº1, Julho, 2006, 2006.

PICCIANO, A. G. The Evolution of Big Data and Learning Analytics in American Higher Education. **Journal of Asynchronous Learning Networks**, v. 16, n. 3, p. 9–20, 2012. Disponível em: <http://files.eric.ed.gov/fulltext/EJ982669.pdf>. Acesso em: 20 nov. 2018.

POYATOS, H.; MENDES, M. H.; RUBIM, L.; LANGUI, C. **Avaliações em Educação a Distância: desafios e oportunidades.**São Paulo:XIII Workshop de pós-graduação e pesquisa do Centro Paula Souza, 2018.

POZO, J. I. a Sociedade Da Aprendizagem E O Desafio De Converter InformaçãO Em Conhecimento. **Revista PáTio**, v. Ano 8, n. Agosto/Outubro, p. 34–36, 2004. Disponível em: <http://www.udemo.org.br/A sociedade.pdf>. Acesso em: 5 set. 2018.

POZO, J. I. A sociedade da aprendizagem e o desafio de converter informação em conhecimento. In: **SALGADO, Maria.Tecnologias na Educação: ensinando e aprendendo com as TIC: guia do cursista.** Brasília: Ministério da Educação, Secretária de Educação à Distância, 2008.

POZO, J. I.; POSTIGO, Y. La solución de problemas como contenido procedimental em la educación obligatoria. In: **Solución de problemas**. Madrid: Santillana, 1994. p. 322–347.

POZO, J. I.; POSTIGO, Y. **Los procedimientos como contenidos escolares: uso estratégico de la información**. Barcelona: Edebé, 2000.

RAMOS, S. R. **Tecnologias da Informação e Comunicação: conceitos básicos**. Disponível em: <http://esms.edu.pt/>. Acesso em: 23 ago. 2013.

REID, E. M. **Electropolis: Communication and Community On Internet Relay Chat**. Disponível em: <http://www.aluluei.com/electropolis.htm>. Acesso em: 3 out. 2018.

REYES, J. The skinny on big data in education: Learning analytics simplified. **TechTrends**, v. 59, n. 2, p. 75–80, 2015. Disponível em: <http://ezproxy.deakin.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=100711546&site=eds-live&scope=site>. Acesso em: 20 nov. 2018.

RICHARDS, N. M.; KING, J. H. Big Data Ethics. **Wake Forest Law Review**, v. 1, p. 40, 2014.

SARAIVA, T. Educação a distância no Brasil: lições da história. **Em Aberto, Brasília**, n. 70, 1996.

SCRIVEN, M. **The logic of evaluation**ClaremontClaremont Graduate University, 2007. Disponível em: <https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1390&context=ossaarchive>. Acesso em: 2 set. 2018.

SIEMENS, G. **Learning analytics: envisioning a research discipline and a domain of practice. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 4-8)**ACM. SoLAR, 2012.

SILVA, F.; SILVA, J. Da; SILVA, R.; FONSECA, L. C. Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. **Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)**, v. 1, n. Sbie, p. 1187, 2015.

SILVA, M. Criar e professorar um curso online: relato de experiência. In: **Educação online: teorias, práticas, legislação, formação corporativa.** 3. edição ed. São Paulo: Ed. Loyola, 2011. p. 55–64.

SMITH, P. A. Autocorrelation in logistic regression modelling of species' distributions. **Global Ecology and Biogeography Letters**, n. 4, p. 47–61, 1994.

STEEL, G. L.; SUSSMAN, G. J. **Scheme: An Interpreter for the Extended Lambda Calculus**. 1975. Massachusetts Institute of Technology, 1975.

STERNBERG, R. J. **The psychology of problem solving**. 1a. edição ed. Cambridge: Cambridge University Press, 2003.

STOCKWELL, D. R. B. **Machine learning and the problem of prediction and explanation in ecological modelling.** 1992. Australian National University, 1992.

TANG, T.; MCCALLA, G. Utilizing Artificial Learners to Help Overcome the Cold-Start Problem in a Pedagogically-Oriented Paper Recommendation System. **Proceedings of the International Conference on Adaptive Hypermedia**, p. 245–254, 2004.

THOMPSEN, S. R.; STRAUBHAAR, J. D.; BOLYARD, D. M. Ethnomethodology and the study of online communities. **Information Research**, v. 4, n. 1, p. 2010, 1998. Disponível em: <http://informationr.net/ir/4-1/paper50.html>. Acesso em: 20 nov. 2018.

TORRES, P. L.; MARRIOTT, R. de C. V. A aprendizagem colaborativa no LOLA. In: **SANTOS, Edméa; ALVES, Lynn (Org.). Práticas pedagógicas e tecnolo- gias digitais**. Rio de Janeiro: E-papers, 2006. p. 161–162.

VALENTE, J. A. Curso de especialização em desenvolvimento de projetos pedagógicos com o uso das novas tecnologias: descrição e fundamentos. In: **VALENTE, José Armando; PRADO, Maria Elisabette B. Brito; ALMEIDA, Maria Elizabeth Bianconcini de. Educação a distância via Internet.** São Paulo: Avercamp, 2003. p. 31.

VAPNIK, V.; GOLOWICH, S. E.; SMOLA, A. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In: SUYKENS J.A.K.; KANDEWALLE J. (Ed.). **Nonlinear Modeling: advanced black-box techniques**. Boston: Kluwer Academic Publishers, 1998. p. 55–85.

VARIAN, H. R. Big Data: New Tricks for Econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3–28, 2014.

VERBERT, K.; MANOUSELIS, N.; DRACHSLER, H.; DUVAL, E. Dataset-driven research to support learning and knowledge analytics. **Educational Technology & Society**, v. 15, p. 133–148, 2012.

VERBYLA, D.L. & LITAITIS, J. A. Resampling methods for evaluating classification accuracy of wildlife habitat models. **Environmental Management**, n. 13, p. 783–787, 1989.

VIEIRA, R. S. O Papel das tecnologias da informação e comunicação na educação a distância: um estudo sobre a percepção do professor/tutor. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 10, p. 65–70, 2011. Disponível em: <http://www.abed.org.br/revistacientifica/Revista_PDF_Doc/2011/Artigo_05.pdf>. Acesso em: 20 nov. 2018.

WIGGINS, G. The Case for Authentic Assessment. **o Practical Assessment, Research & Evaluation**, v. 2, n. 2, 1990. Disponível em: <http://pareonline.net/getvn.asp?v=2&n=2>. Acesso em: 2 set. 2018.

WITTEN, I. H.; FRANK, E. **Data mining: Practical Machine Learning Tools and Techniques with Java Implementations.** São Franscisco: Morgan Kaufmann, 1999.

ZHANG, S., ZHANG, C., YANG, Q. Data Preparation for Data Mining. **Applied Artificial Intelligence**, v. 17, p. 375–381, 2003.

APÊNDICE A – ARQUIVOS DE CONFIGURAÇÃO PARA BIG DATA

Neste **Apêndice A** são apresentados os conteúdos dos arquivos de configuração utilizados na infraestrutura de *Big Data*, para fins de replicação do experimentado ou eventual adequação do algoritmo preditivo para outras instituições de ensino. É importante frisar novamente que, para a conveniência deste estudo, o **Apache Hadoop** foi configurado para funcionar com um único nó em seu *cluster*, o que não condiz com a maioria das aplicações reais que possuem diversos nós em um *cluster* para lidar com o alto volume de dados.

Quadro 9 – O arquivo de configuração /opt/hadoop/etc/hadoop/core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing,
software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
  See the License for the specific language governing permissions
and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://0.0.0.0:9000</value>
</property>
<!--
<property>
    <name>dfs.permissions</name>
       <value>false</value>
</property>
-->
</configuration>
```

Fonte: Elaborado pelo autor (2019)

Quadro 10 – O arquivo de configuração /opt/hadoop/etc/hadoop/hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at
```

```
    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing,
software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
  See the License for the specific language governing permissions
and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
    <name>dfs.replication</name>
    <value>1</value>
</property>
</configuration>
```

Fonte: Elaborado pelo autor (2019)

Quadro 11 – O arquivo de configuração /opt/hadoop/etc/hadoop/mapred-site.xml

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing,
software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
  See the License for the specific language governing permissions
and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

Fonte: Elaborado pelo autor (2019)

Quadro 12 – O arquivo de configuração /opt/hadoop/etc/hadoop/yarn-site.xml

```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing,
software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
  See the License for the specific language governing permissions
and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->

<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.resourcemanager.address</name>
<value>127.0.0.1:8032</value>
</property>
<property>
<name>yarn.resourcemanager.scheduler.address</name>
<value>127.0.0.1:8030</value>
</property>
<property>
<name>yarn.resourcemanager.resource-tracker.address</name>
<value>127.0.0.1:8031</value>
</property>

</configuration>
```

Fonte: Elaborado pelo autor (2019)

APÊNDICE B – PROCEDIMENTOS PARA CONSOLIDAÇÃO DOS DADOS

Neste **Apêndice B** são apresentados os procedimentos utilizados para a consolidação e cálculos dos dados utilizados na pesquisa. Para a conveniência do estudo, os dados foram consolidados, transformados e sintetizados previamente no sistema gerenciador de banco de dados MySQL, utilizando procedimentos armazenados (*stored procedures*). Em casos reais, os dados são carregados em sua forma bruta no lago de dados, e os procedimentos relatados são feitos posteriormente.

A abordagem adotada na pesquisa aproveitou conhecimentos prévios de seu pesquisador e trouxe como vantagem adicional a exportação da massa de dados previamente tratada em arquivos CSV que podem ser carregados diretamente no software **R Studio**, eliminando a necessidade a infraestrutura de *Big Data* para pequenos volumes de dados. Um exemplo da exportação para CSV é apresentada no Quadro 17.

Os procedimentos armazenados que são apresentados nos quadros deste apêndice são funcionais, no entanto, foram ligeiramente alterados para preservar eventuais particularidades da infraestrutura do AVA Moodle utilizado pela instituição de ensino da pesquisa e é por esta mesma razão que a consolidação do tempo de permanência nos objetos de aprendizagem foi omitida deste Apêndice B. Por outro lado, os procedimentos mostrados podem ser utilizados em qualquer implementação do Moodle versão 2.x ou superior com adequações mínimas.

Quadro 13 – Estrutura e cargas de dados iniciais do consolidado de alunos

```
-- Criação de uma estrutura para receber dados pessoais e notas da
avaliação formativa
CREATE TABLE Alunos_sistema_prova
(RM NUMERIC(6) NOT NULL PRIMARY KEY,
nome VARCHAR(100) NOT NULL,
cidade VARCHAR(100) NOT NULL,
dt_nascimento DATETIME NOT NULL,
sexo CHAR (1) NOT NULL,
turma CHAR(11) NOT NULL,
ano CHAR(4) NOT NULL,
curso VARCHAR(100) NOT NULL);

-- Carga de dados a partir do arquivo CSV
TRUNCATE Alunos_sistema_prova;
LOAD DATA LOCAL INFILE '/root/Alunos.csv' INTO TABLE
Alunos_sistema_prova
CHARACTER SET UTF8
FIELDS TERMINATED BY ';';

-- Estrutura de dados consolidados dos alunos
CREATE TABLE consolidado_alunos
(userId NUMERIC(5) NOT NULL PRIMARY KEY,
 RM NUMERIC(6) NOT NULL UNIQUE,
```

```
 nome VARCHAR(100) NOT NULL,
 cidade VARCHAR(100) NOT NULL,
 dt_nascimento DATETIME NOT NULL,
 sexo CHAR (1) NOT NULL);

-- Carga a partir do sistema de prova
INSERT INTO consolidado_alunos
SELECT u.id, asp.RM, asp.nome, asp.cidade, asp.dt_nascimento,
asp.sexo
FROM Alunos_sistema_prova asp
INNER JOIN moodle_user u
ON (asp.RM = u.username);

-- Estrutura para as turmas de alunos
CREATE TABLE consolidado_alunos_turma(
  userId NUMERIC(5) NOT NULL,
  RM NUMERIC(6) NOT NULL UNIQUE,
  categoryId BIGINT(10) NOT NULL,
  turma CHAR(11) NOT NULL,
  ano CHAR(4) NOT NULL,
  curso VARCHAR(100) NOT NULL,
  PRIMARY KEY (userId, categoryId)
);

-- Carregando apenas as turmas que o aluno se matriculou em 2018
-- removendo as cargas de 2017 das turmas pré-existentes
INSERT INTO consolidado_alunos_turma
SELECT ca.userId, ca.RM, fcc.id, fcc.name, asp.ano, asp.curso
FROM consolidado_alunos ca
LEFT JOIN Alunos_sistema_prova asp
ON (asp.RM = ca.RM)
INNER JOIN moodle_course_categories fcc
ON (fcc.name = asp.turma)
WHERE FROM_UNIXTIME(fcc.timemodified) > '2017-11-01 00:00:00'
ORDER BY name, asp.nome, name;

-- Estrutura para cruzamento com os cursos (presentes no Moodle)
CREATE TABLE consolidado_alunos_cursos
(userId NUMERIC(5) NOT NULL,
 RM NUMERIC(6) NOT NULL,
 courseId BIGINT(10) NOT NULL,
 PRIMARY KEY(userId, courseId));

-- Carga de dados dos cursos em que o aluno está matriculado.
INSERT INTO consolidado_alunos_cursos
SELECT ca.userId, ca.rm, c.id
FROM consolidado_alunos ca
INNER JOIN consolidado_alunos_turma cat
ON (ca.userId = cat.userId)
INNER JOIN moodle_course_categories fcc
ON (cat.categoryId = fcc.id)
INNER JOIN moodle_course c
ON (c.category = fcc.id)
ORDER BY ca.userId, ca.rm, c.id;

-- Alteração para dados da avaliação somativa
```

```
ALTER TABLE consolidado_alunos_turma
ADD COLUMN nota NUMERIC(9,4) NOT NULL DEFAULT 0,
ADD COLUMN nota_index NUMERIC(7,6) NOT NULL DEFAULT 0,
ADD COLUMN QtdMaximoPonto INT NOT NULL DEFAULT 0,
ADD COLUMN QtdQuestao INT NOT NULL DEFAULT 0,
ADD COLUMN datahorainicio TIMESTAMP,
ADD COLUMN datahorafim TIMESTAMP,
ADD COLUMN tempo_de_prova TIME;

-- Procedimento de carga da avaliação somativa omitido.
```

Fonte: Elaborado pelo autor (2019)

Quadro 14 – Procedimento para consolidação dos dados das atividades formativas

```
-- 1 - Script para consolidado dos dados das entregas remotas
DELIMITER $$
CREATE PROCEDURE sp_consolidado_entregas_remotas_v2 (IN _ano
CHAR(4), IN _mes CHAR(2))
    BEGIN
      SET @dt_inicio = CONCAT(_ano,'-',_mes,'-01 00:00:00');
      SET @dt_fim = CONCAT(LAST_DAY(@dt_inicio), ' 23:59:59');

      UPDATE consolidado_alunos_indices cai,
      (SELECT u.id AS userId,
        cat.turma,
      TRUNCATE(COUNT(a.id), 0) as tarefas,
      COUNT(ass.id) as entregas,
      IFNULL(COUNT(ass.id)/TRUNCATE(COUNT(a.id), 0),0) as
entregas_index,
      TRUNCATE(SUM(a.grade), 0) as pontos_maximos,
      IFNULL(TRUNCATE(SUM(ag.grade),0),0) as pontos_conquistados,
      IFNULL(IFNULL(TRUNCATE(SUM(ag.grade),0),0) /
TRUNCATE(SUM(a.grade), 0), 0) as pontos_index
      FROM moodle_assign a
      JOIN moodle_user u
      LEFT JOIN moodle_assign_submission ass
      ON (a.id = ass.assignment) AND (ass.userid = u.id)
      LEFT JOIN moodle_assign_grades ag
      ON (ag.assignment = a.id) AND (ag.userid = u.id)
      INNER JOIN consolidado_alunos_cursos cac
      ON (a.course = cac.courseId AND cac.userId = u.id)
      INNER JOIN consolidado_alunos_turma cat
      ON (cac.userId = cat.userId)
      AND FROM_UNIXTIME(a.duedate) >= @dt_inicio AND
FROM_UNIXTIME(duedate) <= @dt_fim
      GROUP BY userId, cat.turma) src
      SET cai.tarefas = src.tarefas,
          cai.entregas = src.entregas,
          cai.entregas_index = src.entregas_index,
          cai.pontos_maximos = src.pontos_maximos,
          cai.pontos_conquistados = src.pontos_conquistados,
          cai.pontos_index = src.pontos_index
      WHERE cai.userId = src.userId
      AND cai.turma = src.turma
      AND ano = _ano
      AND mes = _mes;
    END$$
```

```
DELIMITER ;

-- 2 - Utilização do script para geração de dados consolidados
-- atividades remotas
CALL sp_consolidado_entregas_remotas_v2('2018','02');
CALL sp_consolidado_entregas_remotas_v2('2018','03');
CALL sp_consolidado_entregas_remotas_v2('2018','04');
CALL sp_consolidado_entregas_remotas_v2('2018','05');
CALL sp_consolidado_entregas_remotas_v2('2018','06');
```

Fonte: Elaborado pelo autor (2019)

Quadro 15 – Procedimento para consolidação dos dados de entregas com atraso

```
-- 1 - Script para consolidado dos dados de entregas com atraso
DELIMITER $$
CREATE PROCEDURE sp_consolidado_entregas_com_atraso_v2 (IN _ano
CHAR(4), IN _mes CHAR(2))
    BEGIN
      SET @dt_inicio = CONCAT(_ano,'-',_mes,'-01 00:00:00');
      SET @dt_fim = CONCAT(LAST_DAY(@dt_inicio), ' 23:59:59');

      UPDATE consolidado_alunos_indices cai,
             (SELECT u.id AS userId,
               cat.turma,
               COUNT(*) entregas_com_atraso
             FROM moodle_assign_submission ass
             INNER JOIN moodle_assign a
             ON (ass.assignment = a.id)
             INNER JOIN moodle_user u
             ON (ass.userid = u.id)
             INNER JOIN consolidado_alunos_cursos cac
             ON (a.course = cac.courseId AND cac.userId = u.id)
             INNER JOIN consolidado_alunos_turma cat
             ON (cac.userId = cat.userId)
              AND a.duedate < ass.timecreated
              AND FROM_UNIXTIME(a.duedate) >= @dt_inicio AND
FROM_UNIXTIME(duedate) <= @dt_fim
              GROUP BY u.id, cat.turma
            ) AS src
      SET cai.entregas_com_atraso = src.entregas_com_atraso,
          cai.entregas_com_atraso_index = src.entregas_com_atraso /
cai.entregas
      WHERE cai.ano = _ano
      AND cai.mes = _mes
      AND cai.turma = src.turma
      AND cai.userid = src.userid;

    END$$
DELIMITER ;

-- 2 - Utilização do script para geração de dados consolidados
-- de entregas com atraso
CALL sp_consolidado_entregas_com_atraso_v2('2018', '02');
CALL sp_consolidado_entregas_com_atraso_v2('2018', '03');
CALL sp_consolidado_entregas_com_atraso_v2('2018', '04');
CALL sp_consolidado_entregas_com_atraso_v2('2018', '05');
CALL sp_consolidado_entregas_com_atraso_v2('2018', '06');
```

Fonte: Elaborado pelo autor (2019)

Quadro 16 – Procedimento para consolidação dos dados de interação no fórum

```
-- 1 - Script para consolidado dos dados de interação no fórum
DELIMITER $$
CREATE PROCEDURE sp_consolidado_forum_v2 (IN _ano CHAR(4), IN _mes
CHAR(2))
  BEGIN
      SET @dt_inicio = CONCAT(_ano,'-',_mes,'-01 00:00:00');
      SET @dt_fim = CONCAT(LAST_DAY(@dt_inicio), ' 23:59:59');

      UPDATE consolidado_alunos_indices cai,
         (SELECT u.id AS userId,
                 cat.turma,
                 COUNT(*) AS num_interacoes_forum
          FROM moodle_forum_posts fp
          INNER JOIN moodle_forum_discussions md
          ON (fp.discussion = md.id)
          INNER JOIN moodle_forum f
          ON (md.forum = f.id) AND (md.course = f.course)
          INNER JOIN moodle_user u
          ON (fp.userid = u.id)
          INNER JOIN consolidado_alunos_cursos cac
          ON (md.course = cac.courseId AND cac.userId = u.id)
          INNER JOIN consolidado_alunos_turma cat
          ON (cac.userId = cat.userId)
          AND FROM_UNIXTIME(fp.created) >= @dt_inicio
          AND FROM_UNIXTIME(fp.created) <= @dt_fim
          GROUP BY u.id) AS src
        SET cai.num_interacoes_forum = src.num_interacoes_forum
        WHERE cai.ano = _ano
        AND cai.mes = _mes
        AND cai.turma = src.turma
        AND cai.userid = src.userid;
  END$$
DELIMITER ;

-- 2 - Utilização do script para geração de dados consolidados
-- de interação no fórum
CALL sp_consolidado_forum_v2('2018', '02');
CALL sp_consolidado_forum_v2('2018', '03');
CALL sp_consolidado_forum_v2('2018', '04');
CALL sp_consolidado_forum_v2('2018', '05');
CALL sp_consolidado_forum_v2('2018', '06');
```

Fonte: Elaborado pelo autor (2019)

Quadro 17 – Procedimento de exportação de consolidado para fevereiro de 2018

```
mysql -B -u root tableaumoodle -h localhost -e "
SELECT DATEDIFF('2018-01-01 00:00:00', ca.dt_nascimento)/365.25 AS
idade,
       IF(ca.sexo='M', 1, 2) AS genero,
       TRUNCATE(AVG(cai.entregas_index),4) AS entregas_index,
       TRUNCATE(AVG(cai.pontos_index),4) AS pontos_index,
       SUM(cai.tempototal_html) AS tempototal_html,
       SUM(cai.tempototal_videos) AS tempototal_videos,
       SUM(cai.tempototal_audios) AS tempototal_audios,
```

```
       SUM(cai.tempototal_pdf) AS tempototal_pdf,
       SUM(cai.entregas_com_atraso) AS entregas_com_atraso,
       SUM(cai.num_interacoes_forum) AS num_interacoes_forum,
       cat.nota_index
FROM consolidado_alunos ca
INNER JOIN consolidado_alunos_turma cat
ON (ca.userId = cat.userId AND ca.RM = cat.RM)
INNER JOIN consolidado_alunos_indices cai
ON (cat.userId = cai.userId AND cat.turma = cai.turma AND cat.ano =
cai.ano)
WHERE cat.nota_index <> 0
AND cai.ano = 2018
AND cai.mes <= 2
GROUP BY idade,
       genero,
       cat.nota_index
" | sed "s/\t/;/g" > consolidado_2018_02.csv
```

Fonte: Elaborado pelo autor (2019)

APÊNDICE C – PROCEDIMENTOS NA LINGUAGEM "R" PARA PREDIÇÃO DE DADOS

Neste **Apêndice C** são apresentados os procedimentos escritos em linguagem R que foram utilizados para a predição de dados. Os códigos-fonte foram comentados passo-a-passo para melhor compreensão.

Quadro 18 – Procedimento para predição utilizando árvore de decisão

```
#Instalação dos pacotes necessários ao procedimento
#install.packages("rpart", dependencies=T)
#install.packages("rpart.plot", dependencies = T)
#install.packages("caret", dependencies=T)

#Carregamento das bibliotecas para árvore de decisão
library("rpart")
library("rpart.plot")
library("caret")

#Carga de dados a partir do arquivo CSV (pode ser substituído pelo
RHive)
alunos2018 = read.csv(file.choose(), sep=";", header=T)


#Definição da semente para amostragem fixa
set.seed(10100)

#Definição do fator a ser previsto
alunos2018['resultado'] = as.factor(ifelse(alunos2018$nota_index >=
0.6, "bom", "ruim"))

#Geração de Matriz para divisão 70% (treino) e 30% (teste)
amostra = sample(2,nrow(alunos2018),replace=T,prob=c(0.7,0.3))

#Separação da Amostragem de treino (70%)
alunos2018_treino = alunos2018[amostra==1,]
alunos2018_treino_norm = scale(alunos2018_treino[,1:10]);
alunos2018_treino_norm = as.data.frame(alunos2018_treino_norm)
alunos2018_treino_norm['resultado'] = alunos2018_treino$resultado;

#Separação da Amostragem de teste (30%)
alunos2018_teste = scale(alunos2018[amostra==2,1:10])
alunos2018_teste = as.data.frame(alunos2018_teste)

#Criação do modelo utilizando árvore de decisão (rpart)
arvore = rpart(resultado ~
idade+genero+entregas_index+pontos_index+tempototal_html+tempototal_
videos+tempototal_audios+tempototal_pdf+entregas_com_atraso+num_inte
racoes_forum, data=alunos2018_treino_norm, method="class")

#Impressão da árvore de decisão exibida no Quadro 3
print(arvore)

#Plotagem da árvore de decisão gráfica exibida na Figura 7
```

```
rpart.plot(arvore, main="extra = 106,  under = TRUE", extra=106,
under=TRUE, faclen=0)

#Predição dos 30% de treino utilizando árvore de decisão
predicao = predict(arvore, newdata = alunos2018_teste, type="class")

#Geração da Matriz de Confusão
classes_teste = ifelse(alunos2018[amostra==2,11] >= 0.6, "bom",
"ruim")
classes_teste = factor(classes_teste, levels = c("bom", "ruim"))
confusionMatrix(predicao, classes_teste )
```

Fonte: Elaborado pelo autor (2019)

Quadro 19 – Procedimento para predição utilizando Naïve Bayes

```
#Instalação dos pacotes necessários ao procedimento
#install.packages("e1071", dependencies=T)
#install.packages("caret", dependencies=T)

#Carregamento das bibliotecas para árvore de decisão
library("e1071")
library("caret")

#Carga de dados a partir do arquivo CSV (pode ser substituído pelo
RHive)
alunos2018 = read.csv(file.choose(), sep=";", header=T)

#Definição da semente para amostragem fixa
set.seed(10100)

#Definição do fator a ser previsto
alunos2018['resultado'] = as.factor(ifelse(alunos2018$nota_index >=
0.6, "bom", "ruim"))

#Geração de Matriz para divisão 70% (treino) e 30% (teste)
amostra = sample(2,nrow(alunos2018),replace=T,prob=c(0.7,0.3))

#Separação da Amostragem de treino (70%)
alunos2018_treino = alunos2018[amostra==1,]
alunos2018_treino_norm = scale(alunos2018_treino[,1:10]);
alunos2018_treino_norm = as.data.frame(alunos2018_treino_norm)
alunos2018_treino_norm['resultado'] = alunos2018_treino$resultado;

#Separação da Amostragem de teste (30%)
alunos2018_teste = scale(alunos2018[amostra==2,1:10])
alunos2018_teste = as.data.frame(alunos2018_teste)

#Criação do modelo utilizando Naive Bayes
modelo = naiveBayes(resultado ~
idade+genero+entregas_index+pontos_index+tempototal_html+tempototal_
videos+tempototal_audios+tempototal_pdf+entregas_com_atraso+num_inte
racoes_forum, data=alunos2018_treino_norm, type="class")

#Impressão da árvore de decisão exibida no Quadro 5
print(modelo)

#Predição dos 30% de treino utilizando Naive Bayes
```

```
predicao = predict(modelo, newdata = alunos2018_teste, type="class")
predicao

#Geração da Matriz de Confusão
classes_teste = ifelse(alunos2018[amostra==2,11] >= 0.6, "bom",
"ruim")
classes_teste = factor(classes_teste, levels = c("bom", "ruim"))
confusionMatrix(predicao, classes_teste )
```

Fonte: Elaborado pelo autor (2019)

Quadro 20 – Procedimento para predição utilizando K-Nearest Neighbor

```
#Instalação dos pacotes necessários ao procedimento
#install.packages("class", dependencies=T)
#install.packages("caret", dependencies=T)

#Carregamento das bibliotecas para árvore de decisão
library(caret)
library(class)

#Carga de dados a partir do arquivo CSV (pode ser substituído pelo
RHive)
alunos2018 = read.csv(file.choose(), sep=";", header=T)

#Definição da semente para amostragem fixa
set.seed(10100)

#Definição do fator a ser previsto
alunos2018['resultado'] = as.factor(ifelse(alunos2018$nota_index >=
0.6, "bom", "ruim"))

#Geração de Matriz para divisão 70% (treino) e 30% (teste)
amostra = sample(2,nrow(alunos2018),replace=T,prob=c(0.7,0.3))

#Separação dos fatores em classes treino e classes teste
classes_treino = alunos2018[amostra==1,12]
classes_teste = alunos2018[amostra==2,12]

#Separação da Amostragem de treino (70%)
alunos2018_treino =
as.data.frame(scale(alunos2018[amostra==1,1:10]))

#Separação da Amostragem de teste (30%)
alunos2018_teste = as.data.frame(scale(alunos2018[amostra==2,1:10]))

#Predição dos 30% de treino utilizando k-nearest neighbor
predicao = knn(alunos2018_treino, alunos2018_teste, classes_treino,
k=71)

#Geração da Matriz de Confusão
confusionMatrix(predicao, classes_teste )
```

Fonte: Elaborado pelo autor (2019)

Quadro 21 – Procedimento para predição utilizando *Support Vector Machine* (SVM)

```
#Instalação dos pacotes necessários ao procedimento
#install.packages("e1071", dependencies=T)

#Carregamento das bibliotecas para SVM
library(e1071)

#Carga de dados a partir do arquivo CSV (pode ser substituído pelo
RHive)
alunos2018 = read.csv(file.choose(), sep=";", header=T)

#Definição da semente para amostragem fixa
set.seed(10100)

#Definição do fator a ser previsto
alunos2018['resultado'] = as.factor(ifelse(alunos2018$nota_index >=
0.6, "bom", "ruim"))

#Geração de Matriz para divisão 70% (treino) e 30% (teste)
amostra = sample(2,nrow(alunos2018),replace=T,prob=c(0.7,0.3))

#Separação dos fatores em classes treino e classes teste
classes_treino = alunos2018[amostra==1,12]
classes_teste = alunos2018[amostra==2,12]

#Separação da Amostragem de treino (70%)
alunos2018_treino =
as.data.frame(scale(alunos2018[amostra==1,1:10]))
alunos2018_treino_com_resultado = alunos2018_treino
alunos2018_treino_com_resultado['resultado'] =
alunos2018[amostra==1,12]

#Separação da Amostragem de teste (30%)
alunos2018_teste = as.data.frame(scale(alunos2018[amostra==2,1:10]))

#Geração do modelo utilizando Support Vector Machine (SVM)
modelo = svm(resultado ~ ., data = alunos2018_treino_com_resultado,
type = 'C-classification',  kernel="radial", cost=1, gamma=2)

#Processo de tunning para chegar ao melhor "cost" e "gamma"
Stuned = tune(svm,
              resultado ~ .,
              data=alunos2018_treino_com_resultado,
              type = "C-classification",
              kernel  = "radial",
              ranges=list(gamma = 2^(-9:10),
                          cost = 2^(-9:10))          )
#Resultado de tune
summary(Stuned)

#Predição dos 30% de treino utilizando SVM
predicao = predict(modelo, alunos2018_teste)

#Geração da Matriz de Confusão
confusionMatrix(predicao, classes_teste )
```

Fonte: Elaborado pelo autor (2019)

ANEXO A – AUTORIZAÇÃO PARA USO DOS DADOS

São Paulo, 20 de novembro de 2018

AUTORIZAÇÃO

As partes

Leandro Rubim de Freitas, FIAP ON Leader, casado, CPF 288.984.638-55, trabalha na VSTP Educação LTDA (FIAP) no endereço Av. Lins de Vasconcelos, 1222 – Cambuci – São Paulo, SP, CEP 01538-001, representando a doravante chamada de instituição

E

Henrique Ruiz Poyatos Neto, coordenador e professor, casado, CPF 218.987.458-03, residente na Av. Lacerda Franco, 1313 – Cambuci – São Paulo, SP, CEP 01536-001, doravante chamado de pesquisador.

I – A instituição se compromete a ceder os dados de seu ambiente virtual de aprendizagem (AVA) e sistema de provas estritamente para fins de pesquisa acadêmica e trabalho de Mestrado do pesquisador;

II – O pesquisador se compromete a não expor os dados cedidos pela instituição a terceiros de qualquer natureza sob qualquer pretexto.

III – O pesquisador se compromete a divulgar apenas resultados referentes ao algoritmo de predição, não expondo os dados de nenhum aluno de instituição unicamente.

IV – O pesquisador se compromete a omitir procedimentos técnicos que exponham particularidades referentes à infraestrutura de dados que possam ser exploradas por uma eventual invasão.

V – O pesquisador se compromete a destruir os dados utilizados na pesquisa assim que o trabalho for concluído, assim como ceder os procedimentos realizados a contribuir para a melhoria no atendimento dos alunos da instituição.

_____

Henrique Ruiz Poyatos Neto
Pesquisador
CPF 218.987.458-03

_____

Leandro Rubim de Freitas
FIAP ON Leader, VSTP Educação LTDA
CPF 288.984.638-55

REPÚBLICA FEDERATIVA DO BRASIL

MINISTÉRIO DA ECONOMIA

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512019000915-0**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 30/04/2019, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Sistema para predição de avaliação formativa no ensino a distancia em curso superior

Data de publicação: 30/04/2019

Data de criação: 01/10/2018

Titular(es): MARCELO TSUGUIO OKANO

Autor(es): MARCELO TSUGUIO OKANO; CELI LANGHI; HELENA GEMIGNANI PETEROSSI; SERGIO EUGENIO MENINO; HENRIQUE POYATOS NETO

Linguagem: R

Campo de aplicação: ED-06

Tipo de programa: AP-01

Algoritmo hash: SHA-512

Resumo digital hash:
de23e5d198e7f1f95337be5ab02757609e7d1ecf983e259aa706a378e47e0d916ef0fd8145b474c08804ca1dd4f538f831c6526f2cd5d874cb73c556a6c220a8

Expedido em: 21/05/2019

Aprovado por:

Liane Elizabeth Caldeira Lage

Diretora de Patentes, Programas de Computador e Topografias de Circuitos Integrados