Vanessa Faria de Souza

OS AVANÇOS DA MINERAÇÃO DE DADOS EDUCACIONAIS

PROCESSO, TENDÊNCIAS TEMÁTICAS E TÉCNICAS DE MINERAÇÃO

Bacharelado em
CIÊNCIA
DA COMPUTAÇÃO

INSTITUTO
FEDERAL
Rio Grande
do Sul

Bagai

ISBN 9786586734836

Este livro foi composto pela Editora Bagai.

www.editorabagai.com.br /editorabagai

/editorabagai contato@editorabagai.com.br

Vanessa Faria de Souza

OS AVANÇOS DA MINERAÇÃO DE DADOS EDUCACIONAIS

PROCESSO, TENDÊNCIAS TEMÁTICAS E TÉCNICAS DE MINERAÇÃO

Bacharelado em CIÊNCIA DA COMPUTAÇÃO

INSTITUTO FEDERAL Rio Grande do Sul

Bagai

Editor-Chefe	Cleber Bianchessi
Revisão	A autora
Projeto Gráfico	Jhonny Alves dos Reis
Diagramação	IFRS
Imagem da capa	https://bit.ly/3zdje9S
Reitor	Júlio Xandro Heck
Pró-Reitor de Pesquisa, Pós-graduação e Inovação	Eduardo Girotto

Conselho Editorial

Membros natos
Gregório Durlo Grisa
Maísa Helena Brum
Maria Cristina Caminha de Castilhos França
Cimara Valim de Mello
Greice da Silva Lorenzzetti Andreis
Aline Terra Silveira
Marilia Bonzanini Bossle
Luiz Gaspar Fensterseifer
Sílvia Schiedeck

Membros eleitos
Carine Bueira Loureiro
Daiane Romanzini
Daniela Sanfelice
Marcus André Kurtz Almança
Mariana Lima Duro
Marina Wöhlke Cyrillo
Maurício Polidoro
Paulo Roberto Janissek
Viviane Diehl

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecária responsável: Aline Terra Silveira – CRB 10/1933

S729a Souza, Vanessa Faria de
Os avanços da mineração de dados educacionais : processo, tendências temáticas e técnicas de mineração / Vanessa Faria de Souza. - 1. ed. - Curitiba, PR : Bagai, 2021.
1 arquivo em PDF (166 p.)

ISBN 9786586734836 (Livro digital)

1. Computação. 2. Mineração de dados (Computação). 3. Educação - Processamento de dados. 4. Aprendizagem. I. Título.

CDU(online): 004.6:37

https://doi.org/10.37008/978-65-86734836.20.09.21

Livro desenvolvido com recursos oriundos do IFRS, provenientes do EDITAL IFRS Nº 09/2021 – AUXÍLIO À PUBLICAÇÃO DE PRODUTOS BIBLIOGRÁFICOS.

R. Nelsi Ribas Fritsch, 1111 – Bairro Esperança
CEP: 98200-000 – Ibirubá/RS
Telefone: (54) 3324-8100
e-mail: comunicacao@ibiruba.ifrs.edu.br

Dedico este livro ao meu pai João Batista de Souza, ainda que em meio a milhares de dificuldades, nunca parou de ler...

PREFÁCIO

MINERAÇÃO DE DADOS EDUCACIONAIS – UMA APRESENTAÇÃO

"Os dados são o novo Petróleo" – Clive Humby

Na educação, o surgimento do *"Big Data"*, por meio de novas e extensas mídias educacionais, combinado com avanços na computação, significou a melhora dos processos de aprendizagem na educação formal e além. Cada vez mais, conjuntos de dados de grandes dimensões estão disponíveis a partir das interações dos alunos com softwares educacionais e em plataformas de aprendizagem on-line – entre outras fontes – com repositórios de dados públicos apoiando pesquisadores na obtenção desses dados. Nesse contexto surgiu a Mineração de Dados Educacionais (MDE).

O primeiro workshop de MDE foi realizado em 2005, em Pittsburgh, Pensilvânia. Seguiram-se workshops anuais e, em 2008, a 1ª Conferência Internacional de MDE foi realizada em Montreal, Quebec. As conferências anuais sobre MDE foram acompanhadas pelo *Journal of Educational Data Mining*, que publicou sua primeira edição em 2009, com Kalina Yacef como editora. O primeiro Manual de MDE foi publicado em 2010. No verão de 2011, a Sociedade Internacional de Mineração de Dados Educacionais (https://educationaldatamining.org/) foi formada para "promover a pesquisa científica no campo interdisciplinar de mineração de dados educacionais", organizando as conferências e periódicos e a publicação gratuita de acesso aberto de artigos de conferências e periódicos.

A MDE reúne uma comunidade interdisciplinar de cientistas da computação, cientistas da aprendizagem, psicometristas e pesquisadores de outras tradições e com o crescente interesse de pesquisa em análise de aprendizagem e mineração de dados, bem como o rápido desenvolvimento de software e métodos analíticos, é importante que pesquisadores e educadores reconheçam os aspectos únicos desta comunidade.

A MDE abrange o desenvolvimento e a aplicação de métodos para explorar os tipos únicos de dados provenientes de ambientes educacionais, podendo ser definida como a aplicação de técnicas de mineração de dados para abordar questões relativas à educação. Ela como área de pesquisa é interdisciplinar, incluindo, entre outros, recuperação de informações, sistemas de recomendação, análise de dados visuais, mineração de dados orientada por domínio, análise de redes sociais, psicopedagogia, psicologia cognitiva, psicometria, dentre outros. De fato, pode-se dizer que a MDE é a combinação de três áreas principais: Ciência da Computação, Educação e Estatística. A interseção dessas três áreas também forma subáreas intimamente relacionadas – a Educação Baseada em Computadores; Mineração de Dados; Aprendizagem de Máquina e Estatística Educacional.

Conforme uma das principais Revisões de Literatura sobre Mineração de Dados Educacionais (ROMERO; VENTURA, 2010), esta área envolve diferentes grupos de usuários ou participantes; diferentes grupos olham para informações educacionais de diferentes ângulos, de acordo com sua própria missão, visão e objetivos para usar a mineração de dados. Por exemplo, o conhecimento descoberto por técnicas de MDE pode ser usado não apenas para ajudar os professores a gerenciar suas aulas, compreender os processos de aprendizagem de seus alunos e refletir sobre seus próprios métodos de ensino, mas também para apoiar as reflexões do aluno sobre a situação e fornecer feedback.

Embora inicialmente a MDE pareça envolver apenas dois grupos principais, alunos e professores, na verdade existem mais grupos envolvidos, com muito mais objetivos, e por isso de acordo com Romero e Ventura (2010) esta área tem uma tendência de rápido desenvolvimento, sobretudo, porque há uma grande gama de pessoas envolvidas no sistema educacional que se beneficia com a sua aplicação. Segundo os autores os principais grupos beneficiados e seus objetivos na utilização da MDE são:

Usuários e Alunos – Personalizar o *e-learning*; recomendar atividades, recursos e tarefas de aprendizagem que possam melhorar sua aprendizagem; sugerir experiências de aprendizagem interessantes; sugerir links a seguir; gerar dicas adaptáveis; recomendar cursos e discussões relevantes.

Educadores, Professores, Instrutores e Tutores – Obter feedback objetivo sobre a instrução; analisar a aprendizagem e o comportamento dos alunos; detectar quais alunos precisam de suporte; prever o desempenho do aluno; classificar os alunos em grupos; encontrar padrões regulares e irregulares de alunos; encontrar erros cometidos com mais frequência pelos alunos; determinar atividades mais eficazes; melhorar a adaptação e customização de cursos.

Desenvolvedores de Cursos e Pesquisadores Educacionais – Avaliar e manter o material didático; melhorar o aprendizado do aluno; avaliar a estrutura do conteúdo do curso e sua eficácia na aprendizagem; construir automaticamente modelos de alunos e modelos de tutor; comparar técnicas de mineração de dados para poder recomendar a mais útil para cada tarefa; desenvolver ferramentas específicas de mineração de dados para fins educacionais.

Organizações, Provedores de Plataformas de Ensino e Aprendizagem, Universidades e Empresas Treinamento – Aprimorar os processos de decisão nas instituições de ensino superior; agilizar a eficiência no processo de tomada de decisão para atingir objetivos específicos; sugerir certos cursos que podem ser valiosos para cada classe de alunos; encontrar a maneira mais econômica de melhorar a retenção e as notas; selecionar os candidatos mais qualificados para a graduação; ajudar a admitir alunos que se sairão bem na universidade.

Administradores Educacionais e Gestores de Instituições de Ensino – Desenvolver a melhor forma de organizar os recursos institucionais (humanos e materiais) e sua oferta educacional; utilizar os recursos disponíveis de forma mais eficaz; aprimorar as ofertas de programas educacionais e determinar a eficácia da abordagem de ensino à distância; avaliar o professor e os currículos; definir parâmetros para melhorar a eficiência dos web sites institucionais, softwares educacionais e/ou plataformas de ensino aprendizagem e adaptá-lo aos usuários.

Como visto, são muitos os benefícios gerados pela MDE para cada grupo de atores do contexto educacional, o que a torna uma estratégia importante para melhoria contínua do ensino e aprendiza-

gem. Neste sentido, este livro é um bom primeiro passo para quem deseja entender melhor a MDE e como é seu processo de aplicação, ou ainda para pesquisadores ativos que desejam se manter atualizados sobre técnicas atuais de mineração de dados. Os capítulos são escritos com base em pesquisadores importantes da MDE e cobrem tópicos interessantes na área.

APRESENTAÇÃO

O livro *Os Avanços da Mineração de Dados Educacionais: Processo, Tendências Temáticas e Técnicas de Mineração* tem como propósito primeiramente proporcionar uma fonte com diversos estudos desenvolvidos por pesquisadores conceituados na área de MDE, muitos dos quais apoiaram o surgimento deste campo de pesquisa, e apresentar por meio da visão destes pesquisadores como este campo tem evoluído desde o seu surgimento, bem como descrever sua consolidação como área de pesquisa e suas perspectivas futuras.

Em segundo lugar este livro pretende dar suporte a pesquisadores iniciantes que desejam se aprofundar sobre quais são as principais tendências temáticas da MDE nos últimos anos, em outras palavras para quais objetivos ela tem sido empregada; quais as principais técnicas têm sido aplicadas no contexto da MDE, com explicações sobre essas técnicas de forma simples e clara. Por fim, também apresenta exemplos de aplicação do processo de MDE, de forma detalhada, para que mesmo àqueles que não tenham familiaridade com o processo de mineração de dados, se sintam capazes de desenvolver suas próprias aplicações.

Este livro é fruto da pesquisa de doutoramento da autora que fez diversos estudos de mapeamento e revisão de literatura sobre a MDE, ademais desenvolveu diversos experimentos sobre o processo de MDE, com a aplicação de técnicas de mineração sobre dados de alunos reais e gerados artificialmente, tanto no âmbito do *e-leaning*, como no presencial. Alguns desses estudos da autora, acerca do tema deste livro, podem ser consultados no Apêndice A.

O livro aborda diversos aspectos da MDE, iniciando pelas definições e caracterização do processo de aplicação; a evolução da MDE e sua consolidação como área de pesquisa; principais tendências temáticas; e diferenças entre a Mineração de Dados Educacionais e Análise de Aprendizagem. Em seguida, apresenta as técnicas de mineração de dados mais utilizadas atualmente; e estudos que fizeram a aplicação

destas técnicas de mineração no contexto educacional. Finalmente, são detalhados dois exemplos de aplicação do processo de MDE.

Desta forma, pode-se observar que uma ampla gama de tópicos sobre MDE são abordados neste livro. Visto que, muitas das informações que se têm disponíveis hoje são fruto de processos de mineração de dados cada vez mais avançados; e essas informações têm feito muitos setores da indústria, comércio, prestação de serviços, dentre outros, se desenvolver de forma sem igual; é de fundamental importância que Instituições de Ensino, também comecem a apoiar seu processo de tomada de decisão em informações reais extraídas de dados. Isso facilita e torna mais assertiva as resoluções de gestores educacionais e professores, que podem conduzir com mais certeza suas ações.

Além disso, com os avanços tecnológicos disponíveis e a digitalização da educação, causada sobretudo pela pandemia do novo Corona Vírus; que ocorreu de forma emergencial, mas que não retornará aos moldes antigos; cada vez mais a mineração de dados estará no cerne do processo de tomada de decisões nas grandes Instituições de Ensino. Por isso, este tema deve ser fundamentalmente discutido no âmbito da educação, em todos os níveis e deve-se fomentar a formação de técnicos e professores para trabalhar com MDE, que envolve muito mais do que apenas saber aplicar algoritmos de última geração.

Vanessa Faria de Souza

SUMÁRIO

CAPÍTULO 1 - MINERAÇÃO DE DADOS EDUCACIONAIS: DEFINIÇÕES E PROCESSO

A quantidade de dados disponíveis a respeito de diversos contextos aumentou expressivamente nos últimos anos devido particularmente a grande difusão das redes sociais e também a disponibilização de aplicativos e sistemas online para compras. Nesse sentido, diversas entidades organizacionais têm demonstrado eficiência na captação, organização e armazenamento de bases dados de grande volume, conseguidas por meio de transações diárias, ou em pesquisas tecnológicas e/ou científicas, todavia a chave para avançar é o domínio de como utilizar acertadamente essa grande quantidade de dados para convertê-los em conhecimentos que possam ser aproveitados em seus negócios, não importando a área a qual esteja vinculado (AGGARWAL, 2015). À vista disso, a Mineração de Dados (*Data Mining*) está cada vez mais difundida como uma estratégia para extração de conhecimentos e informações a partir de dados, que tem potencial para direcionar o processo de tomada de decisão em circunstâncias que ainda não se possua uma vasta experiência.

Para Aggarwal (2015) a Mineração de Dados (MD) é definida como:

> A mineração de dados é o estudo de coleta, limpeza, processamento, análise e obtenção de informações e ideias úteis de dados. Existe uma grande variação em termos de domínios problemáticos, aplicativos, formulações e representações de dados encontradas em aplicativos reais. Portanto, "Mineração de dados" é um termo abrangente usado para descrever esses diferentes aspectos de processamento de dados (AGGARWAL, 2015).

Com a crescente adoção de MD, essa passou a ser empregada com sucesso também no contexto educacional, auxiliando em diversos cenários, e ficou conhecida como Mineração de Dados Educacionais (*Educational Data Mining*). Este capítulo trata dos principais aspectos sobre esta área que vem se consolidando nos últimos anos, em que primeiramente é exposta sua definição, posteriormente é detalhado o processo de implementação da Mineração de Dados Educacionais (MDE).

DEFINIÇÃO DA MINERAÇÃO DE DADOS EDUCACIONAIS

Nos últimos anos a educação tem se modificado, em decorrência do avanço tecnológico disponível que direcionou a uma instrumentação do setor educacional, tanto em softwares voltados para o ensino, como na administração digital dos registros acadêmicos pelos gestores das instituições, bem como no uso da internet para a aprendizagem, em especial pela popularização do *e-learning*. Todos esses fatores impulsionaram um crescimento exponencial no volume de dados educacionais, e para se analisar uma grande quantidade de dados, é imprescindível contar com recursos computacionais, caso contrário a tarefa torna-se impraticável.

Dessa forma, as técnicas de mineração de dados estão ganhando cada vez mais importância no setor educacional, pois são uma forma de acompanhar, analisar e avaliar o processo de aprendizagem. Provavelmente, as técnicas de mineração de dados podem fornecer aos formuladores de políticas educacionais modelos para apoiar seus objetivos de aprimorar a eficiência e a qualidade do ensino e da aprendizagem. Além disso, o uso de diferentes técnicas de mineração de dados pode ser visto como base para uma mudança sistêmica, capaz de impactar de maneira positiva nas soluções de problemas específicos das Instituições de Ensino, por exemplo, viabilizando soluções que envolvem a personalização dos ambientes educacionais ou fornecendo suporte para o processo de tomada de decisão no ambiente educacional.

Nesse cenário, destaca-se a MDE que utiliza as técnicas da MD para extrair informações relevantes de conjuntos diversificados de dados educacionais. Segundo a Sociedade Internacional de Mineração de Dados Educacionais[1], a MDE pode ser definida da seguinte forma:

> É uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar dados únicos e cada vez mais em larga escala, provenientes de contextos educacionais e usa esses métodos para entender melhor os alunos e as configurações em que aprendem (EDM, 2020).

[1] http://educationaldatamining.org/.

Em outras palavras, a MD refere-se a um conjunto de técnicas computacionais para extrair informações de grandes massas de dados, e quando os dados analisados são provenientes de contextos educacionais, chama-se MDE (ROMERO; VENTURA, 2013). Corroborando essa definição (BAKHSHINATEGH *et al.*, 2018) salienta que a mineração de dados educacional (EDM) é o campo de uso de técnicas de mineração de dados em ambientes educacionais. Igualmente, De Los Reyes *et al.* (2019) define MDE como uma área voltada ao desenvolvimento de métodos para explorar dados oriundos de ambientes educacionais e utilizá-los para compreender melhor os processos de ensino e aprendizagem.

Nessa acepção, Baker, Isotani e Carvalho (2011) alegam que a MDE é definida como a área de pesquisa que tem como finalidade aperfeiçoamento e amadurecimento de técnicas para investigar conjuntos de dados obtidos em cenários educacionais. Conforme os autores, a natureza destes dados é mais diversa do que a observada nos dados tradicionalmente utilizados em tarefas de mineração, demandando adaptações e novas abordagens. Ao mesmo tempo, essa diversidade nos dados representa um potencial de implementação de recursos fundamentais para auxílio na melhoria da educação (BAKER; ISOTANI; CARVALHO, 2011; DE LOS REYES *et al.* 2019; RIGO *et al.* 2014).

Nos últimos anos, pesquisadores de uma variedade de disciplinas (incluindo ciência da computação, estatística, mineração de dados e educação) começaram a investigar como a mineração de dados pode melhorar a educação e facilitar a pesquisa educacional. A MDE é cada vez mais reconhecida como uma disciplina emergente que se concentra no desenvolvimento de métodos para explorar os tipos exclusivos de dados que vêm de um contexto educacional. Esses dados vêm de várias fontes, incluindo dados de ambientes tradicionais de sala de aula presencial, software educacional, material didático online e testes somativos / de alto risco. Essas fontes fornecem cada vez mais uma vasta quantidade de dados, que podem ser analisados para responder facilmente a questões que não eram viáveis anteriormente, envolvendo diferenças entre as populações de alunos ou envolvendo comportamentos incomuns dos alunos.

A MDE está contribuindo para a educação e a pesquisa em educação de várias maneiras, como pode ser visto na diversidade de problemas educacionais considerados nos próximos capítulos deste Livro. As contribuições da MDE influenciaram o pensamento sobre pedagogia e aprendizagem e promoveram a melhoria do software educacional, melhorando a capacidade do software de individualizar as experiências de aprendizagem dos alunos. De certa forma, o advento da MDE pode ser considerado como a educação "alcançando" outras áreas, onde a melhoria dos métodos de exploração de dados promove impactos transformadores na prática. Embora a Mineração de Dados (mais Geral) e a Mineração de Dados Educacionais possuam métodos de exploração de dados semelhantes, existem algumas diferenças importantes entre eles. Nesse sentido. Romero e Ventura (2007) elencam questões que diferenciam a MDE da MD em outros domínios:

1. *Objetivos:* que podem se relacionar à pesquisa (a) aplicada, que busca responder questões práticas, por exemplo: como melhorar o processo de aprendizagem; e (b) pura, com a finalidade de por exemplo dar sentido às observações. Na maioria das vezes esses objetivos são difíceis de quantificar e exigem seu próprio conjunto especial de técnicas de medição.
2. *Dados:* em ambientes educacionais, existem muitos tipos diferentes de dados disponíveis para mineração. Esses dados são específicos da área educacional, portanto, possuem informações semânticas intrínsecas, relacionamentos com outros dados, e vários níveis de hierarquia significativa.
3. *Técnicas:* problemas educacionais têm algumas características especiais que exigem que a questão da mineração seja tratada de uma maneira diferente. Embora a maioria das técnicas tradicionais de MD possam ser aplicadas diretamente, outras não podem e devem ser adaptadas ao problema educacional específico. Exemplo disso, é que em se tratando de cenários comuns de MD, as variáveis são em sua maioria numéricas, tratáveis diretamente por algoritmos de Aprendizagem de Máquina, enquanto que em ambientes educacionais a grande maioria é categórica, o que implica esforço em pré-processamento e

transformações, para codificar essas variáveis em numéricas, para que então possam ser interpretadas pelos algoritmos.

O PROCESSO DE MINERAÇÃO DE DADOS EDUCACIONAIS

O processo de aplicação da MDE não é trivial contendo várias fases. Nesse sentido, será explicado o processo de MD de acordo com Aggarwal (2015), depois será apresentada uma sequência de etapas considerada mais adequada, baseada em pesquisas (BAKER; ISOTANI; CARVALHO, 2011; DE LOS REYES *et al.* 2019; RIGO *et al.* 2014; ROMERO; VENTURA, 2013, 2020) específicas sobre MDE e também na experiência da autora com aplicações realizadas.

Aggarwal (2015) determina que o fluxo de trabalho de um processo típico de MD contém as seguintes fases:

1. *Coleta de dados* – A coleta de dados pode exigir o uso de hardware especializado, como uma rede de sensores, trabalho manual, como a coleta de pesquisas com usuários ou ferramentas de software como um mecanismo de rastreamento de documentos da *Web* para coletar documentos. Esta etapa é específica da plataforma e geralmente fora do domínio do cientista de dados. Após a fase de coleta, os dados geralmente são armazenados em um banco de dados ou, em geral, um *data warehouse*[2](tradução livre – armazém de dados) para processamento.
2. *Extração de recursos e limpeza de dados (Pré-Processamento e Transformação)* – Quando os dados são coletados, eles geralmente não estão em um formato adequado para processamento. Para tornar os dados adequados para processamento, é essencial transformá-los em um formato que seja interpretável aos algoritmos de mineração, como multidimensionais, séries temporais ou formato semiestruturado. O formato multidimensional é o mais comum, no qual diferentes campos de dados correspondem às diferentes propriedades medidas que são chamadas de atributos. A fase de extração de recursos geralmente é realizada em paralelo com a

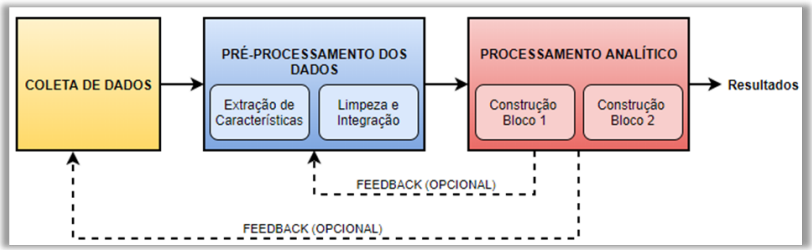
[2] É um repositório central de informações que ficam disponíveis para serem analisadas e dão suporte ao processo de tomada de decisão.

limpeza de dados, onde partes ausentes e incorretas dos dados são estimadas ou corrigidas. Em muitos casos, os dados podem ser extraídos de várias fontes e precisam ser integrados em um formato unificado para processamento. O resultado final deste procedimento é um conjunto de dados estruturados, que pode ser efetivamente usado por um programa de computador. Depois da fase de extração de recursos, os dados podem ser armazenados novamente em um banco de dados para processamento.

3. *Processamento analítico e algoritmos* – A parte final do processo de mineração é projetar métodos analíticos eficazes para extrair informações e conhecimentos relevantes a partir dos dados processados.

A sequência das etapas do processo de MD proposto por Aggarwal (2015) é apresentada na Figura 2.

Figura 1 – Processo de MD proposto por Aggarwal (2015)



Fonte: Adaptado de Aggarwal (2015)

A partir do processo de MD apresentado, pesquisas em MDE e na experiência com aplicações realizadas, algumas das etapas foram ajustadas para condizer efetivamente com o que é executado, levando em conta o fator usuário (programador ou cientista de dados), nesse sentido o processo de MDE pode ser formado pelo seguinte conjunto de etapas:

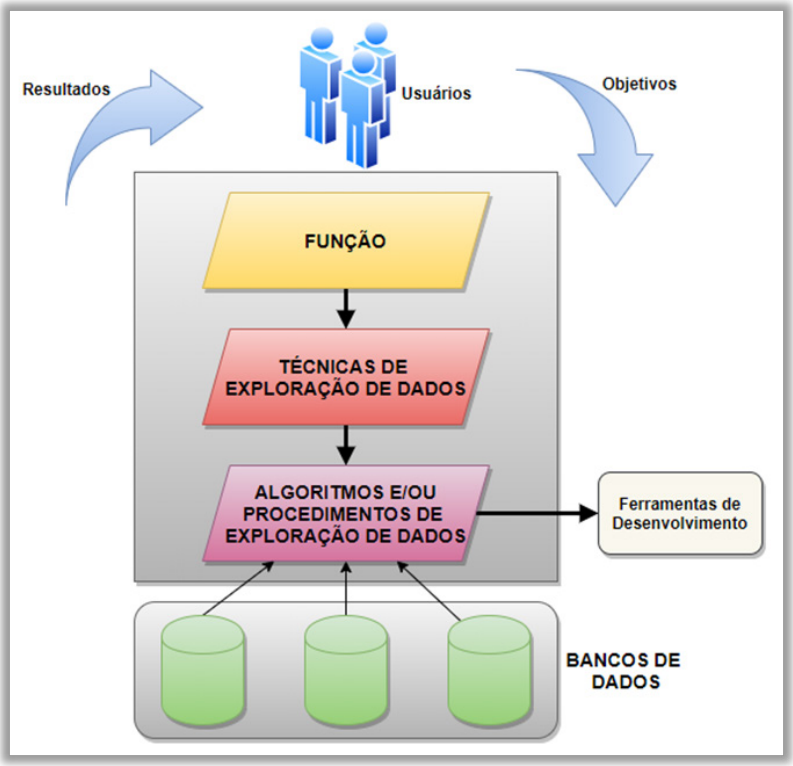
1. *Definição da função da MDE* – Nessa etapa é realizada a determinação do objetivo do processo MDE, para qual finalidade ela está sendo aplicada, como por exemplo: identificação de padrões, detecção de desvio, segmentação, sistemas de recomendação, análise de ligações e regras de associação, sumarização e visualização, mineração de textos, afinidade em grupos,

descrição de grupos, para isso é preciso especificar que tipo de conhecimento pretende-se extrair dos dados.

2. *Formatação dos dados que serão utilizados* – Diversos tipos de armazenamentos de dados e de bancos de dados podem ser manipulados no processo de mineração, cabe ao usuário definir qual formato é o mais adequado para aplicação das técnicas de mineração selecionadas, outro ponto importante é que baseado no tipo de conjunto de dados disponível para análise é que definem-se os padrões, relacionamentos ou informações que se consegue minerar. Nessa etapas todas as incoerências na base devem ser corrigidas e caso for necessário podem ser acrescentados mais atributos que sejam resultantes da combinação de outros ou que possam ser deduzidos de outros, como a idade que pode ser calculada a partir da data de nascimento, ou o total de atividades realizadas que pode ser efetuada por meio de uma soma.
3. *Definição das Técnicas de MDE* – A definição das técnicas é um passo importante, pois elas devem ser específicas para o tratamento da função de MDE estabelecida, as técnicas mais utilizadas para MDE são: a Estatística Descritiva, a Aprendizagem de Máquina e mais recentemente tem sido empregada a Aprendizagem Profunda, cabe salientar que cada uma dessas técnicas possui algoritmos, rotinas e/ou procedimentos específicos para manipulação dos dados. A literatura (AGGARWAL, 2015; BAKER; ISOTANI; CARVALHO, 2011; DE LOS REYES *et al.*, 2019; RIGO *et al.*, 2014), em muitos casos, não deixa claro as diferenças entre funcionalidades, funções e técnicas. Por exemplo, um problema que requer identificar alunos desistentes ou concluintes, para isso deve-se pensar em técnicas que possibilitem identificar esses padrões de comportamento nos alunos nessas duas categorias, isso seria a função, o propósito da MDE o de classificar esses alunos, entretanto também é possível entender como a técnica que seria aplicada, algoritmos de Aprendizagem de Máquina para classificação, nesse caso, por isso as vezes fica confusa a interpretação e diferenciação.
4. *Delineamento de como essas técnicas serão aplicadas* – Nessa etapa são selecionadas as ferramentas que vão dar suporte ao desenvolvimento de sistemas capazes de processar os dados e gerar os resultados esperados.

Na Figura 3 pode-se observar as relações entre as etapas enunciadas, tendo em vista explicar a interatividade da função da MDE com as técnicas a serem utilizadas. A MDE, levando em consideração seus objetivos, técnicas e processo de funcionamento, formam uma importante metodologia para apoio a qualquer cenário educacional, porém em cursos baseados no *e-learning*, com muitos alunos, e que (na maioria) não têm acompanhamento de professores ou tutores – em que os alunos tem uma aprendizagem mais autônoma – tais técnicas de análise de dados em massa tornam-se a solução mais viável para sanar muitos problemas identificados.

Figura 2 – Etapas do processo de MDE



Fonte: Autora

Por meio da MDE, talvez seja possível acompanhar e compreender o processo de aprendizagem, bem como outros fatores que a influenciam.

Como por exemplo, identificar que tipo de abordagem instrucional pode propiciar mais ganhos ao aluno, analisando que atributos retratam melhor seu comprometimento com o curso. Ademais, estimula-se a oportunidade de averiguar se o aluno está compreendendo ou não os conteúdos, distinguir níveis de motivação, engajamento nas tarefas on-line, descoberta de fatores ou parâmetros comportamentais de finalização e êxito em um curso, reconhecer padrões de interação, encontrar técnicas ou métodos que colaborem para a continuidade dos estudantes nos cursos até a conclusão (PURSEL *et al.*, 2016), bem como detectar possíveis fraudes, ou trapaças no sistema de aprendizagem. Estes fatores podem ajudar a personalizar o ambiente e os métodos de ensino, para oferecer melhores condições de aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

Em suma, a MDE tem se desenvolvido e pode ser considerada como uma das formas mais promissoras para extração de informações de bases de dados educacionais e suas técnicas têm se tornado cada vez mais eficientes e eficazes, graças ao número crescente de dados disponíveis e dos avanços computacionais. Para entender melhor como se deu esse desenvolvimento e como a MDE tem sido empregada no decorrer do tempo, são descritos os aspectos de sua evolução no próximo Capítulo.

CAPÍTULO 2 - A EVOLUÇÃO DA MINERAÇÃO DE DADOS EDUCACIONAIS

A disponibilidade de grandes bases de dados educacionais, fomentada pelas modernas plataformas e mídias educacionais, combinadas com avanços na computação, formam a composição ideal para o surgimento da Mineração de Dados Educacionais. Embora existam relatos sobre publicações a respeito deste tema desde 1995 (ROMERO; VENTURA, 2007) o primeiro *workshop* foi realizado em 2005, em Pittsburgh, Pensilvânia, tendo sido seguido por várias oficinas e, em 2008, ocorreu a 1.ª Conferência Internacional sobre MDE realizada em Montreal, Quebec. As conferências anuais sobre MDE impulsionaram o surgimento do *Journal of Educational Data Mining*, que publicou sua primeira edição em 2009, na sequência, o primeiro manual sobre MDE foi publicado em 2010 (ROMERO *et al.*, 2010).

Posteriormente, em 2011 a Sociedade Internacional de Mineração de Dados Educacionais foi formada com o objetivo de promover pesquisa científica na área interdisciplinar da MDE, organizando as conferências e os periódicos. No campo das publicações, uma primeira revisão de literatura foi apresentada por Romero e Ventura (2007), seguido de um modelo teórico proposto por Baker e Yacef (2009), e uma revisão bem mais abrangente sobre MDE foi desenvolvida por Romero e Ventura (2010), todos pesquisadores muito influentes na área. Na sequência, outras publicações iniciaram um amplo movimento de pesquisas nesse âmbito, e tiveram grande notoriedade.

Neste sentido, no decorrer da pesquisa bibliográfica realizada durante a elaboração deste livro, muitas publicações sobre MDE foram analisadas, e algumas delas se dedicaram a implementar revisões de literatura sobre essa área. Tais estudos auxiliaram no entendimento de como a MDE tem sido aplicada em diversos contextos educacionais, seus objetivos, as técnicas mais utilizadas, verificação de resultados alcançados e validação dos benefícios proporcionados, bem como identificação de avanços e também desafios que têm sido relatados por pesquisadores da área.

À vista disso, chamaram a atenção algumas publicações, que trouxeram grandes contribuições para pesquisadores interessados em MDE, quais sejam: Shahiri, Husain e Rashid (2015); Sukhija, Jindal e Aggarwal (2016); Schwendimann *et al.* (2017); Aldowah, Al-Samarraie e Fauzy (2019); e Romero e Ventura (2020). Estes estudos foram fundamentais para um aperfeiçoamento da compreensão da evolução da MDE no decorrer de sua consolidação como área de pesquisa e são sintetizadas na sequência.

SHAHIRI, HUSAIN E RASHID (2015)

A primeira revisão sobre MDE analisada foi desenvolvida por Shahiri, Husain e Rashid (2015), essa forneceu uma visão geral das técnicas de mineração de dados que eram usadas para prever o desempenho dos alunos, em publicações datadas entre 2002 e 2015. O estudo também se concentrou em como os algoritmos de previsão poderiam ser usados para identificar os atributos mais importantes dentre a diversidade de dados dos alunos. Nessa revisão, os autores seguiram duas questões de pesquisa para estruturar os resultados: 1) Quais são os atributos mais importantes empregados na previsão do desempenho dos alunos; e 2) Quais as técnicas/algoritmos de previsão mais eficientes. Quanto aos principais atributos, Shahiri, Husain e Rashid (2015) apontam que foram usados com frequência a média cumulativa de notas e a avaliação interna usada por 10 dos 30 artigos selecionados para a revisão.

Os autores também chegaram à conclusão, que a Aprendizagem de Máquina era a técnica mais usada e quanto à eficácia dos algoritmos as Redes Neurais tiveram a maior precisão (98%) para previsão do desempenho dos alunos, seguida das Árvores de Decisão (91%), depois as Máquinas de Vetores de Suporte e K-*Nearest Neighbors* (KNN – K-ésimo Vizinho mais Próximo) com a mesma eficácia (83%), por fim, o método menos preciso foi o Naive Bayes (76%).

Os autores afirmaram ainda, que prever o desempenho dos alunos é muito útil para ajudar educadores e alunos a melhorar o processo de ensino e aprendizagem. No entanto, é importante ressaltar que os índices de eficácia são resultado da interação entre a complexidade da questão de pesquisa com

a qualidade (e algumas vezes, a extensão da base) dos dados, não sendo uma avaliação a respeito dos métodos em si. Aponta-se como ponto negativo da revisão, que algumas das pesquisas citadas pelos autores não apontam uma diversidade de métricas de avaliação dos algoritmos – uma informação importante, que deveria ser divulgada como parte dos resultados.

SUKHIJA, JINDAL E AGGARWAL (2016)

Na sequência foi analisada a revisão sistemática desenvolvida por Sukhija, Jindal e Aggarwal (2016) que descreveram a evolução da MDE, trazendo à tona os aspectos e resultados de vários estudos divididos em 3 gerações: 1ª geração de 2001 a 2005; 2ª geração de 2006 a 2010; e 3ª geração, de 2011 até 2015.

No período de 2001 a 2005, as pesquisas se basearam no uso da MDE como uma ferramenta para antecipar os padrões que ajudam na avaliação de cursos on-line. Dessa forma, os registros de dados e registros de atividades dos alunos foram usados para analisar seu comportamento. Os autores evidenciaram, que o início da literatura relacionada à MDE, produziu pesquisas com inclinação para o ambiente de aprendizagem baseado na *Web*, devido especialmente à grande disponibilidade de dados em cursos on-line. No final deste período, as pesquisas estavam com foco no uso de algoritmos evolutivos para mineração de dados da internet.

Referente ao período de 2006 a 2010, a MDE evoluiu e os estudos passaram a buscar a aplicação de algoritmos mais eficientes. Os bancos de dados usados se tornaram provenientes de sistemas de EAD, baseados na *Web* e vinculados a grandes Instituições de Ensino, bem como o tamanho desses bancos de dados aumentou. Além disso, houve uma inclinação dos pesquisadores para análises preditivas de dados, com relação a prever os problemas e identificar os alunos em potencial, com alta probabilidade, de apresentar um desempenho acadêmico ruim, nesse sentido, sistemas de apoio à decisão para equilibrar a demanda e a oferta educacional também foram desenvolvidos. Durante esse período, a implementação da classificação baseada em Árvores de Decisão e Redes Neurais Artificiais se acentuaram no contexto educacional, com dife-

rentes objetivos. Ademais, citam que a pesquisa na área foi direcionada a dados relacionados ao ensino superior, empregando técnicas como *On-line Analytical Processing* (OLAP[3]) em combinação com DELPHI[4].

Finalmente, no que tange ao período de 2011 a 2015, a MDE evoluiu para incorporar técnicas melhores e mais eficientes, conseguindo integrar novas e mais eficientes regras de associações, ferramentas como WEKA[5] começaram a ganhar popularidade, algumas pesquisas se voltaram a utilizar dados também do ensino médio e a comunidade de pesquisa trabalhou em função de uma aceitação comercial das técnicas de MD na educação. Os pesquisadores forneceram evidências para maior correlação entre diferentes construtos no sistema educacional, levando a uma maior anuência dos resultados obtidos entre estudantes e autoridades. Além disso, extensos estudos foram realizados na busca de encontrar uma solução para as altas taxas de abandono em diferentes contextos acadêmicos, nessa perspectiva também uma grande quantidade de estudos para estimar o desempenho dos estudantes foram implementados. Sukhija, Jindal e Aggarwal (2016) ainda expuseram que os bancos de dados usados neste período ficaram consideravelmente maiores do que os anteriores e essa crescente no volume dos dados foi acompanhada do desenvolvimento de novas técnicas para MDE, como por exemplo, a retomada dos estudos com Aprendizagem Profunda *(Deep Learning)*.

Sukhija, Jindal e Aggarwal (2016) ainda, apontaram cinco lacunas na área de MDE: (1) Indisponibilidade de conjuntos de dados consistentes que sejam grandes o suficiente para refletir o sistema educacional

[3] OLAP é uma concepção de interface que propicia a capacidade de manipular e analisar uma grande dimensão de dados sob múltiplos pontos de vista. As aplicações OLAP são empregadas por gestores em qualquer nível das entidades organizacionais com o intuito de permitir análises comparativas que facilitem a sua tomada de decisões.

[4] O Delphi é muito conhecido como uma linguagem de programação. Embora, realmente ele abranja um ferramental para o desenvolvimento de software que ganhou notoriedade entre profissionais da área de desenvolvimento na implementação de aplicações de desktop, todavia atualmente é integrado e utilizado também para aplicações *Web* e *mobile*.

[5] Weka é um software livre do tipo *open source* para mineração de dados, implementado em Java, ele foi criado por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia. Desde seu desenvolvimento vem se consolidando de forma ampla como uma das principais ferramentas de MD em ambientes acadêmicos.

e seu funcionamento; (2) Necessidade de integração e versatilidade nos conjuntos de dados; (3) Grande parte das técnicas de mineração foram aplicadas isoladamente e poucos trabalhos foram realizados utilizando técnicas híbridas; (4) Havia falta de confiança das autoridades nos resultados da MDE e (5) Necessidade de comparar métodos. Pode-se dizer que embora as descobertas dos autores fossem fortemente fundamentadas, o cenário se modificou bastante deste então.

No que se refere, à primeira lacuna ressalta-se que com a evolução dos MOOCs, bases com milhões de dados estão disponíveis, como exemplo, pode-se aludir ao trabalho desenvolvido por Northcutt, Ho e Chuang (2016), onde foi utilizada uma base de dados gerada a partir de uma plataforma MOOC com 1.893.092 de usuários, que produziram em média de 200 a 1500 interações com a plataforma, cada um, por curso realizado, portanto, a indisponibilidade de conjuntos de dados já não se configura mais como um problema.

No que tange à segunda limitação salienta-se que em relação à integração das bases, ela se mantém, pois, não é possível integrar duas bases de forma simples, sem necessidade de um grande esforço de pré--processamento. No que diz respeito, à versatilidade dos dados – qualidade de não ser colinear, ou seja, dos dados não estarem relacionados – pode-se dizer que houve mudanças, pois, vários tipos diferentes de dados são usados nos modelos de MDE atuais.

A terceira lacuna, sobre uso de métodos híbridos, pode ser considerada a que menos coincide com a realidade dos experimentos realizados na área de MDE atualmente, pois muitos pesquisadores têm empregado técnicas de MDE em conjunto com outras ferramentas de pesquisa como em Gallén e Caro (2017) que utilizaram algoritmos de agrupamento e um questionário respondido pelos alunos para analisar os motivos pelos quais uma pessoa se inscreve em um MOOC. Ademais, pode-se citar como exemplo o trabalho desenvolvido por Nen-Fu *et al.* (2018) que propuseram um método hibrido para identificar grupos de alunos com diferentes perfis de motivação em MOOCS, empregando questionários e o algoritmo *K-means*, com o intuito de entender a auto-organização

de estudantes nas fases iniciais de um curso, pois os autores acreditavam haver uma ligação entre a motivação e comportamento de aprendizagem.

Com relação, à quarta limitação supõe-se que com os grandes avanços tecnológicos disponíveis e a consolidação da Inteligência Artificial (IA), presente no cotidiano das pessoas, a aceitação da MDE como aporte para tomada de decisões no setor acadêmico tenha crescido. Finalmente, quanto à quinta lacuna, é possível destacar que devido aos avanços tecnológicos muitos pesquisadores têm se dedicado a comparar novas técnicas de MDE com outras mais consolidadas, para verificação da eficácia de modelos. Nesse sentido, Gao *et al.* (2019) propuseram um novo modelo para analisar o perfil de aprendizagem e o engajamento de alunos em MOOCs e para validar sua precisão os autores o compararam com algoritmos amplamente utilizados – Regressão Linear e Máquinas de Vetores de Suporte. Além desse exemplo, aponta-se a pesquisa de Waheed *et al.* (2020), que tinha como objetivo prever o abandono em MOOCs, utilizando um Rede Neural Artificial Profunda, que foi comparada aos algoritmos de Regressão Logística e Máquinas de Vetores de Suporte.

SCHWENDIMANN *ET AL.* (2016)

A outra revisão incluída neste Capítulo foi elaborada por Schwendimann *et al.* (2016), que compreendeu, além de MDE, a Análise de Aprendizado (AA) – termo muito utilizados em inglês *Learning Analytics* – e painéis de aprendizado. Os autores cunharam o termo "painéis de aprendizado" e o definiram da seguinte forma: "uma única exibição que agrega diferentes indicadores sobre aluno(s), processo(s) de aprendizagem e/ou contexto(s) de aprendizagem em uma ou várias visualizações". Nesse sentido, os autores afirmaram que os painéis de aprendizado estão se tornando populares devido ao aumento do uso de tecnologias educacionais, como Sistemas de Gerenciamento de Aprendizagem (*Learning Management Systems* - LMS) para a execução de diversos tipos de cursos em EAD como os MOOCs, que na opinião deles constituem a base para o desenvolvimento das áreas de AA e MDE.

A revisão de Schwendimann *et al.* (2016) foi realizada em 6 bases de dados: ACM Digital Library, IEEE Xplore, SpringerLink, Science Direct, Wiley e Google Scholar, retornando 346 artigos no total, dos quais 55 foram incluídos na análise final. A revisão distinguiu entre 2 tipos de contribuições: Artigos que contribuíram com uma proposta teórica ou referencial (3 artigos; 5%); Artigos que descreveram a implementação de um painel de aprendizado (39 artigos; 71%), além de 13 artigos (5%) que apresentaram uma combinação dessas duas. A revisão finalizou delineando questões em aberto e futuras linhas sobre como trabalhar na área de painéis de aprendizado, indicando ainda que há uma necessidade longitudinal de pesquisas e de captação de dados em ambientes virtuais de aprendizado, assim como estudos que comparem sistematicamente designs de painéis diferentes.

A princípio o foco na MDE, pela revisão relatada, parece ficar difuso, entretanto no decorrer da análise percebe-se sua relação quando Schwendimann *et al.* (2016) citam os painéis de aprendizagem como subsídios para o desenvolvimento das áreas de MDE e AA. Tais painéis se constituem como uma forma sistemática de organização dos dados disponíveis dos alunos, para a finalidade de investigação. Os autores salientam ainda que os painéis apresentados nas publicações analisadas utilizaram principalmente os relatórios de Logs das atividades dos estudantes como fonte de dados, e algumas dessas pesquisas usaram Interface de Programação de Aplicativos (*Application Programming Interface* – APIs) externas, bem como atividades escritas dos estudantes, e inclusive bancos de dados institucionais. Nesse sentido, a revisão de Schwendimann *et al.* (2016) auxiliou a entender melhor, quais tipos de atributos de alunos podem servir ao propósito de investigar, por meio de MDE, o comportamento dos alunos, resultante da navegação em uma plataforma de ensino e aprendizagem.

ALDOWAH, AL-SAMARRAIE E FAUZY (2019)

Na sequência foi investigada, a revisão de Aldowah, Al-Samarraie e Fauzy (2019) que teve como foco o tema MDE e a Análise de Aprendizagem (AA) para o século XXI no ensino superior. Os autores relataram que as revisões anteriores sobre essas áreas forneceram infor-

mações substanciais sobre a base teórica correlacionada, no entanto, tais estudos não julgaram a associação entre diferentes técnicas de MDE e AA na resolução de problemas educacionais específicos e não produziram uma classificação clara das dimensões em que essas técnicas poderiam ser aplicadas com sucesso no ensino superior.

Aldowah, Al-Samarraie e Fauzy (2019) sugerem que a revisão realizada por eles, poderia ser aplicada como um guia para futuros estudos sobre o uso de técnicas de MDE e AA com o propósito de resolver problemas específicos de ensino e aprendizagem. Os autores conduziram a revisão com o intuito de responder a duas questões: "Como usar MDE e AA para resolver desafios práticos na educação?" e "Quais técnicas de mineração são mais adequadas para esses problemas?". Para responder a essas perguntas, buscaram artigos das bases de periódicos e conferências: Scopus, Web of Science, Google Scholar, ERIC, Science Direct, DBLP, ACM Digital Library, IEEEXplore e Springer, encontrando 491 artigos publicados de 2000 até 2017.

Aldowah, Al-Samarraie e Fauzy (2019) afirmaram que as técnicas podem ser agrupadas em 4 grandes dimensões: Análise de Aprendizagem Suportada por Computador (*Computer-Supported Learning Analytics* – CSLA); Análise Preditiva Suportada por Computador (*Computer-Supported Predictive Analytics* – CSPA); Análise Comportamental Suportada por Computador (*Computer-Supported Behavioral Analytics* – CSBA); e Análise de Visualização Suportada por Computador (*Computer-Supported Visualization Analytics* – CSVA). Os resultados trazidos pelos autores podem ser sintetizados da seguinte forma:

- As pesquisas sobre CSLA (120 artigos) concentraram-se principalmente no uso de análise estatística de dados para executar tarefas analíticas sofisticadas, a fim de investigar os comportamentos de aprendizagem colaborativa e de busca de informações dos alunos no contexto de um curso.
- Os estudos sobre CSPA (253 artigos) focaram, em sua grande maioria, no uso de funções preditivas ou variáveis contínuas para sugerir maneiras eficazes de melhorar o aprendizado e o desempenho dos alunos, bem como avaliar a adequação do aprendizado.

- As publicações sobre a dimensão CSBA (80 artigos), em maior parte, criaram modelos de comportamento, ações e conhecimento.
- Os estudos sobre CSVA (38 artigos) concentraram-se em métodos para explorar visualmente os dados – usando gráficos interativos por exemplo – para destacar informações úteis e produzir decisões precisas sobre as informações novas descobertas nos dados.

Os autores relataram que a mineração sequencial de padrões, a mineração de texto, a mineração de correlação, a detecção de outlier e a mineração de estimativa de densidade não são comumente usadas devido à complexidade na obtenção dos atributos necessários para regular ou adaptar-se às necessidades dos dados educacionais. Além disso, descobriram que na CSPA há uma taxa mais alta de tarefas de classificação, devido essa ser aceita como uma técnica eficaz para prever padrões de interesse e formar modelos de aprendizagem, promovendo tarefas específicas nesse sentido. Pode-se afirmar que essa revisão forneceu informações substanciais sobre a base teórica, metodológica e objetivos dessas áreas em expansão.

ROMERO E VENTURA (2020)

Para finalizar a explanação sobre as revisões focadas em MDE e temas correlatos, que deram suporte ao desenvolvimento dessa análise sobre a evolução da MDE, destaca-se Romero e Ventura (2020), que efetuaram um estudo sobre a MDE e Análise de Aprendizagem, atualizando revisões anteriores (ROMERO; VENTURA, 2007, 2010, 2013, 2017), estes autores são muito influentes quando se trata de MDE, são pesquisadores que apoiaram as primeiras iniciativas e pesquisas sobre o tema. Partindo dessa premissa, a publicação de 2020 forneceu informações sobre o estado da arte, revisando as publicações da área no sentido de elucidar: os principais marcos; o ciclo de descoberta de conhecimento; os ambientes educacionais mais utilizados; as ferramentas específicas desenvolvidas; os conjuntos de dados disponíveis gratuitamente; os métodos e técnicas mais empregados; os principais objetivos; e por fim, as tendências futuras nessa área de pesquisa. Devido à grande amplitude da revisão, são abordados os itens considerados mais relevantes: as principais alterações na MDE e na AA e suas conclusões.

No que diz respeito as principais mudanças, Romero e Ventura (2020) afirmaram que no período de 2010 a 2020 a MDE, como área de pesquisa, evoluiu enormemente e uma ampla gama de expressões relacionadas surgiram no estado da arte, os autores citam como principais as seguintes: Análise Acadêmica (*Academic Analytics)*; Análise Institucional (*Institutional Analytics)*; Análise Didática (*Teaching Analytics)*; Educação Orientada a Dados (*Data-Driven Education)*; Tomada de Decisão na Educação Orientada a Dados (*Data-Driven Decision-Making in Education)*; *Big Data* na Educação (*Big Data in Education)*; e Ciência de Dados Educacionais (*Educational Data Science)*, cujas definições são apresentadas a seguir:

- A *Academic Analytics* (AA) e a *Institutional Analytics* (IA) se preocupam com a coleta, análise e visualização de atividades do programa acadêmico como: cursos de graduação, cursos EAD, avaliação de cursos, alocação de recursos e gerenciamento para gerar *insights* institucionais. Portanto, estão focadas no desafio político/econômico.
- *Teaching Analytics* (TA) refere-se à análise das atividades de ensino, aprendizagem, dados de desempenho, bem como do design, desenvolvimento e avaliação dessas atividades, está focada no desafio educacional do ponto de vista dos instrutores.
- *Data-Driven Education* (DDE) e *Data-Driven Decision-Making in Education* (DDDM) referem-se a coletar e analisar sistematicamente vários tipos de dados educacionais, para orientar uma série de decisões com o intuito de ajudar a melhorar o sucesso de alunos e escolas.
- *Big Data in Education* (BDE) denota à aplicação de *Big Data* – conotação básica resumida em volume, variedade e valor – para dados do ambiente educacional.
- *Educational Data Science* (EDS) é definida como o uso de dados coletados de ambientes/configurações educacionais para resolução de problemas nesse contexto. A ciência de dados é um conceito para unificar estatísticas, análise de dados, Aprendizagem de Máquina, métodos e técnicas relacionados.

Quanto às conclusões, Romero e Ventura (2020), apresentaram apontamentos relacionados aos seguintes aspectos: 1) A importância e

evolução da MDE e AA; 2) Uma avaliação se as tendências encontradas na publicação de Romero e Ventura (2013) se concretizaram; 3) Os principais desafios ainda existentes, no que diz respeito a MDE e AA; e 4) As tendências para novos estudos nessas áreas.

No que se refere à importância da MDE e a AA, os autores salientaram que essas são duas comunidades interdisciplinares de cientistas da computação, cientistas de aprendizagem, psicometristas e pesquisadores de diversas áreas, mas todos com o mesmo objetivo, o de melhorar o aprendizado a partir dos dados. Os autores sugeriram que a área cresceu rapidamente nas últimas duas décadas, com duas conferências anuais[6], dois periódicos específicos[7] e com o aumento do número de livros, artigos, pesquisas e resenhas relacionados. Além disso, há uma corrente para mudar as pesquisas restritas a laboratórios para o mercado em geral, assim empregando a MDE e AA em instituições educacionais e escolas de todo o mundo. Os autores declararam, que em 2020 toda pesquisa educacional com relativa importância para o cenário envolve análise e mineração de dados. Isso indica que essas áreas se tornarão em breve maduras e amplamente utilizadas não apenas pelos pesquisadores, mas também por instrutores, administradores educacionais e negócios relacionados, em todo o mundo.

Com relação à comparação entre as tendências levantadas em sua pesquisa anterior (ROMERO; VENTURA, 2013) os autores destacaram que em 2013 dois direcionamentos foram apresentados: o primeiro dizia respeito as ferramentas para MDE e AA, o qual ainda não foi completamente alcançado; e o segundo se referia ao desenvolvimento de uma cultura baseada em dados, que continua sendo um desafio (ROMERO; VENTURA, 2020). Referente à primeira tendência, de acordo com Romero e Ventura (2013), as ferramentas para MDE e AA deveriam ser disponibilizadas gratuitamente, para que uma população mais ampla pudesse utilizá-las, o que aconteceu, pois, uma grande variedade de ferramentas de finalidade específica estão

[6] Conferência Internacional sobre Learning Analytics & Knowledge (LAK) – Conferência Internacional sobre Educational Data Mining.

[7] Journal of Edcuational Data Mining – Journal of Learning Analytics.

disponíveis de forma gratuita, porém os autores ressaltaram que ainda é necessário desenvolver ferramentas para uso geral, que possam ser aplicadas em várias tarefas e resolver diferentes problemas educacionais com a mesma interface/ferramenta, além disso, é necessário melhorar a portabilidade dos modelos obtidos por essas ferramentas (ROMERO; VENTURA, 2020). A segunda tendência pressupunha que educadores e instituições deveriam desenvolver uma cultura baseada em dados, utilizando-se deles para tomar decisões e melhorar seus processos de ensino, aprendizagem e administrativos, todavia de acordo com a pesquisa de 2020, a maioria das instituições e profissionais de ensino continuam cientes dos benefícios proporcionados pela análise de dados em larga escala, entretanto não adotaram essa cultura de forma efetiva nos seus processos gerenciais.

Romero e Ventura (2020) ainda sistematizaram os principais desafios relatados acerca de MDE e AA, quais sejam: transferibilidade e generalização; eficácia e aplicabilidade; e interpretabilidade. Transferibilidade e Generalização se referem à utilização de modelos comuns para vários contextos, o que ainda não ocorre na prática. Eficácia e Aplicabilidade dizem respeito à realização de ações de intervenção a partir das análises de dados, assim há uma grande diversidade de modelos desenvolvidos os quais não se tem informações se foram ou são aplicados na prática das Instituições de Ensino e se são efetivos na resolução de problemas dessas instituições. Interpretabilidade concerne à capacidade dos usuários compreenderem os modelos gerados para as análises de dados, acarretando um mau aproveitamento de tais modelos.

Finalmente, os autores propuseram algumas ideias, que disseram ser "visionárias e pessoais" que, em suas opiniões, podem formar tendências e direcionamentos muito promissores para a áreas de MDE e AA, para a década de 2020 resumidos no Quadro 1.

Quadro 1 – Tendências para MDE e AA

TENDÊNCIA	ANÁLISE
Levar em consideração todos os dados pessoais dos alunos durante toda a vida	Atualmente, as informações consideradas na MDE e AA baseiam-se principalmente na interação de alunos com um único ambiente educacional, mas em um futuro próximo, graças ao grande volume de dados e à Internet das Coisas (IoT), pesquisadores serão capazes de ter informações disponíveis para cada aluno desde o nascimento até o momento e em tempo real. Isso implica a integração não apenas dos dados tradicionais de desempenho coletados das Instituições de Ensino e ambientes educacionais que cada aluno utilizou, mas também as informações sobre os status de cada aluno sob diferentes pontos de vista, como médico, familiar, econômico, religioso, sexual, relacionamento emocional, psicológico e assim por diante. Todos esses dados podem ser coletados a partir de várias fontes e elas poderiam ser fundidas para serem usadas com o intuito de melhorar e personalizar o processo de aprendizagem de cada aluno em cada momento específico de sua vida, o que suscitaria um novo nível de precisão.
Aplicação e integração da MDE e AA aos futuros ambientes educacionais tecnológicos	Na última década, os grandes avanços em tecnologias inovadoras permitiram o desenvolvimento de novos sistemas educacionais a partir de dispositivos móveis, a onipresente à realidade virtual, ambientes de realidade aumentada, hologramas e nas próximas décadas, saltos quânticos serão associados à aplicação da Inteligência Artificial. Nesse contexto, não é errado pensar que os instrutores poderiam ser substituídos por máquinas sem que os alunos percebessem a mudança graças a avanços atuais em robôs humanoides inteligentes, agentes de conversação ou interfaces de voz e assistentes em ambientes educacionais. Mas esses sistemas precisam de técnicas de MDE e AA para analisar terabytes de dados e gerar modelos de análise portáteis em tempo real, a fim de enfrentar os desafios educacionais específicos desses futuros ambientes virtuais de aprendizagem.
Análise e Mineração de Dados coletados diretamente do cérebro dos alunos para uma melhor compreensão do aprendizado	O cérebro é o fator-chave para realmente entender como os alunos aprendem. Os avanços promissores na neurociência humana e neurotecnologia generalizada (interfaces cérebro-computador) estão dando origem a oportunidades sem precedentes de obter, coletar, compartilhar e manipular qualquer tipo de informação coletada do cérebro humano. Num futuro próximo, esses dados íntimos, sobre o estado psicológico e a atividade neural do aluno poderão ser analisados e minerados em tempo real, graças aos futuros dispositivos de eletroencefalografia de alta qualidade. Esses dados cerebrais, juntamente com outros dados multimodais poderiam ser integrados e usado pelos pesquisadores de MDE e AA, a fim de alcançar uma compreensão completa do processo de aprendizado.

Fonte: Romero e Ventura (2020) – *Tradução Livre*

Em conclusão, com a análise e interpretação dessas publicações, que sistematizaram boa parte do estado da arte na área, em destaque dos últimos 20 anos, foi possível compreender o início do processo de adoção de MDE, quais eram as principais técnicas, os dados utilizados para formação de bancos/bases e os resultados que foram alcançados, enfim, de entender o modo pelo qual MDE se estabeleceu como um campo consolidada de pesquisa.

Além disso, explorar como ocorreu seu processo de evolução, o qual foi constado, teve início em decorrência fundamentalmente devido a dois fatores: 1) A adoção de grandes bases de dados na educação, impulsionado sobretudo pelo surgimento de cursos *e-learning,* como os do tipo MOOCs; e 2) O avanço das tecnologias computacionais, que são indispensáveis para aplicação e evolução das técnicas vinculadas a MD. Tais avanços, proporcionaram melhorias nas técnicas/ferramentas de MD já existentes, simplificando assim as tarefas dos pesquisadores e aperfeiçoando os resultados obtidos. Dessa forma, tais técnicas/ferramentas puderam ser aplicadas e testadas ao grande volume de dados educacionais disponíveis, consolidando assim a MDE como um importante processo para exploração de dados nesse contexto.

Por fim, essas publicações propiciaram ainda, visualizar o cenário futuro para os próximos estudos com análises de dados educacionais. O que despertou interesse em descobrir quais seriam as principais vertentes temáticas de pesquisas com MDE aplicada principalmente no contexto do *e-leaning*, uma tendência educacional que vem se consolidando nos últimos anos. Para tanto, foi desenvolvido um mapeamento sistemático com o propósito de identificar essas tendências, e auxiliar no reconhecimento de tópicos de estudos promissores e ainda pouco explorados no âmbito da MDE, esse mapeamento é evidenciado no próximo Capítulo.

CAPÍTULO 3 - MINERAÇÃO DE DADOS EDUCACIONAIS: PRINCIPAIS TENDÊNCIAS TEMÁTICAS

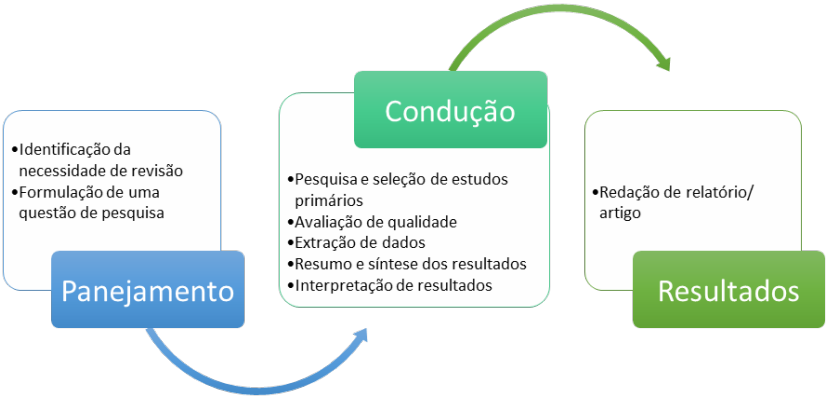
Técnicas de mineração de dados educacionais têm sido amplamente utilizadas em ambientes *e-learning* para realizar diferentes análises educacionais. Nesse contexto, foi realizado um mapeamento sistemático de literatura pela autora (SOUZA; SANTOS, 2020) em cinco bibliotecas digitais com o intuito de verificar quais são os principais temas/objetivos para os quais a Mineração de Dados Educacionais tem sido aplicada em ambientes *e-learning*. A busca abrangeu o período de 2015 a 2019, sendo encontradas 253 estudos, das quais 133 foram selecionados. Os resultados revelaram que estudos relacionados à predição e análise de Comportamento dos alunos (Desempenho, Abandono, Conclusão e Engajamento), análise em fórum de discussão e implementação de sistemas de recomendação são as temáticas mais frequentes em pesquisas com MDE no contexto do *e-leaning*. Este mapeamento é sintetizado no decorrer deste Capítulo.

DESENVOLVIMENTO DO MAPEAMENTO SISTEMÁTICO

Este mapeamento sistemático de literatura teve como objetivo selecionar publicações que aplicaram MDE para analisar os dados do *e-leaning*. O mapeamento buscou conhecer quais os principais temas existentes quando se trata dessa premissa, permeando quais são os principais objetivos da MDE aplicada nesses tipos de cursos. Um mapeamento, como o próprio nome indica, não visa avaliar como o conjunto da literatura responde a determinadas questões de pesquisa, mas sim fazer um panorama de uma determinada área, apresentando uma perspectiva que permite a identificação de oportunidades de pesquisa. Além disso, é semelhante a uma revisão sistemática, em que é preciso usar procedimentos bem definidos para encontrar, avaliar e sintetizar os resultados de pesquisas relevantes na área em estudo, no entanto com maior abrangência e menos aprofundamento em cada estudo selecionado.

Para a realização deste mapeamento foram utilizadas as diretrizes de Kitchenham e Charters (2007), que apresentam um protocolo – baseado em outros protocolos amplamente utilizados na pesquisa médica com base em evidências – que é o mais utilizado tanto na área de computação em geral, quanto em trabalhos de levantamento sistemático de literatura na área de Informática na Educação. O processo de revisão e/ou mapeamento sistemático de literatura, conforme apresentado nestas diretrizes inclui diversas atividades (Figura 3), que podem ser agrupadas em três fases principais: planejamento, condução e síntese de resultados (KITCHENHAM; CHARTERS, 2007).

Figura 3 – Fases e atividades da revisão e/ou mapeamento



Fonte: Adaptado de Kitchenham e Charters (2007)

O mapeamento então teve início com o planejamento que foi conduzido para responder a 3 questões de pesquisa:

Questão de Pesquisa 1 (QP1): Quais Conferências/Periódicos possuem mais publicações na área?

Questão de Pesquisa 2 (QP2): Em que ano houve mais publicações?

Questão de Pesquisa 3 (QP3): Quais são os tópicos mais abordados quando se trata de MDE em cursos e-leaning?

Em seguida ainda na etapa de planejamento foram realizadas as seguintes atividades: 1) Formatação da *string* de busca – a *string* de

busca foi definida e formulada da seguinte forma: ((“*educational data mining*” OR “EDM”) AND (“*e-leaning courses*” OR “*massive open online courses*”)); 2) Definição das bases de dados – foram definidas 5 bibliotecas digitais, que foram selecionadas sobretudo pela sua grande abrangência e pela qualidade das publicações indexadas, estas foram: (a) Biblioteca Digital IEEEXplore; (b) Biblioteca Digital ACM; (c) ERIC (*Education Resources Information Center*); (d) Science@Direct e (e) SciElo; e, por fim 3) Estabelecimento dos critérios de inclusão e exclusão – os critérios de inclusão e exclusão foram os seguintes: *Inclusão*: (a) Artigos completos, (b) Artigos publicados entre 2015 e 2019, (c) Artigos redigidos em inglês e (d) Artigos que descrevem a aplicação da MDE em cursos *e-learning*. *Exclusão*: (a) Artigos duplicados, (b) Uso de MDE em contextos diferentes do *e-learning*, (c) Aplicação de técnicas diferentes da MDE, (d) Revisões de literatura ou mapeamentos, e (e) Artigos em idiomas diferentes do inglês.

Após a aplicação da *string* de busca nas bases e o retorno dos artigos, foi então possível selecionar os manuscritos que atendiam aos critérios de inclusão ou excluir aqueles que atendiam aos critérios de exclusão. Para isso, em um primeiro momento foi realizada uma triagem inicial em que foram lidos apenas os resumos dos artigos, nessa fase já foi possível identificar de qual temática o manuscrito se tratava. Em seguida, os artigos selecionados foram analisados mais detalhadamente.

Para a terceira fase do mapeamento, de geração do relatório com os resultados foi utilizado o aplicativo *Planilhas* do Google, para extrair e analisar os dados, correspondendo à última etapa do mapeamento. Os registros foram formados por 8 atributos[8]: título, autores, local de publicação, ano, resumo tema/objetivo, critérios (se os critérios de inclusão/exclusão foram atendidos ou não e justificativa) e situação (incluído ou excluído). Os resultados foram analisados levando-se em consideração o número de publicações por ano, local de publicação e o tema da pesquisa.

[8] Para aceder a planilha completa com todos os manuscritos e as descrições dos 8 atributos acesse o link: https://docs.google.com/spreadsheets/d/1Ob5yWlWk1DueDYJotKI-L7sXnwUT5ku4/edit#gid=1537563813.

RESULTADOS DO MAPEAMENTO SISTEMÁTICO

O mapeamento sistemático desenvolvido selecionou artigos entre 2015 e 2019, a partir das bases de dados citadas, vale ressaltar que as publicações que não utilizaram MDE no contexto do *e-leaning* não foram selecionadas, além disso não foram incluídos artigos que eram revisões de literatura ou mapeamentos. A Tabela 1 relaciona o número de artigos devolvidos de cada base de dados.

Tabela 1 – Total de Artigos Selecionados

BIBLIOTECA DIGITAL	ESTUDOS RETORNADOS	ESTUDOS SELECIONADOS
IEEE	98	67
ACM	22	11
ERIC	47	43
SCIENCE@DIRECT	64	12
SCIELO	22	0
TOTAIS	**253**	**133**

Fonte: Souza e Santos (2020)

Conforme mostrado na Tabela 1, do total de 253 estudos encontrados, 133 estudos atenderam aos critérios de inclusão. Como se pode verificar da biblioteca SciElo nenhum artigo foi selecionado, uma vez que todos os que tratavam de MDE no contexto do *e-leaning* eram textos em espanhol, portanto, pelo critério de exclusão, nenhum pode ser selecionado. Na sequência, cada uma das questões de pesquisa norteadoras do mapeamento é respondida.

QP1: Quais Conferências/Periódicos possuem mais publicações na área?

A base de dados IEEE apresentou o maior número de publicações de acordo com os critérios de inclusão, com 67 estudos selecionadas, seguida da ERIC com 43 estudos selecionados. Science@Direct, apesar de apresentar um número significativo de trabalhos retornados, a maio-

ria não estava dentro do escopo deste mapeamento. Os 133 trabalhos selecionados e recuperados pelas bases de dados provêm de congressos e periódicos, publicados em 71 locais diferentes, dos quais 60 são congressos e 11 são periódicos. Desse total, foram classificados apenas aqueles que retornaram pelo menos 2 publicações sobre o tema, conforme mostra a Tabela 2. A Tabela 2 mostra que a Conferência Internacional de Mineração de Dados Educacionais é o evento que apresenta o maior número de pesquisas que relatam a uso de Mineração de Dados Educacionais no contexto do *e-learning*.

Tabela 2 – Total de Artigos Selecionados

Conferências/Periódicos	N° de Artigos
International Conference on Educational Data Mining (EDM)	36
IEEE Access	5
Computers in Human Behavior	4
Computers & Education	4
International Conference on Educational Innovation through Technology (EITT)	3
International Conference on Computer Science and Education (ICCSE)	3
International Conference on Learning Analytics & Knowledge Pages	3
Conference on Learning @ Scale Pages	3
IEEE 2nd International Conference on Big Data Analysis (ICBDA)	2
International Conference on Advanced Learning Technologies (ICALT)	2
International Conference on Information Reuse and Integration (IRI)	2
International Conference on Industrial Engineering and Engineering Management (IEEM)	2
Latin American Conference on Learning Technologies (LACLO)	2
Learning With MOOCS (LWMOOCS)	2
Transactions on Learning Technologies	2
Journal of Educational Data Mining	2
The Internet and Higher Education	2

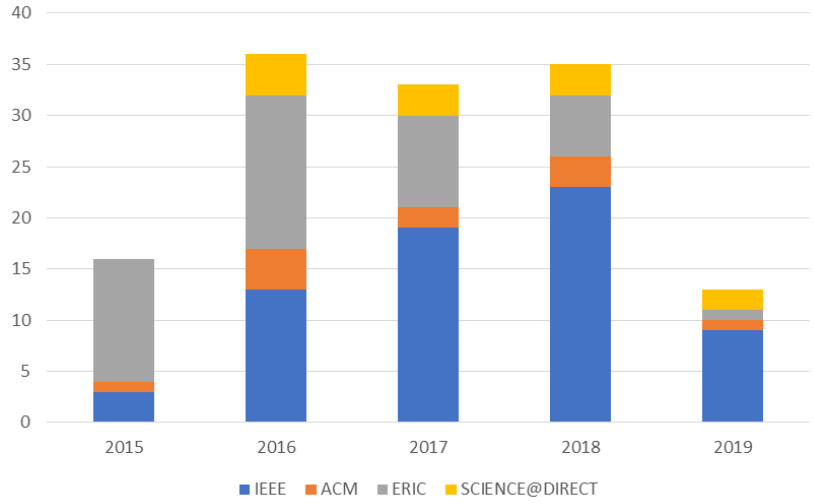
Fonte: Souza e Santos (2020)

QP2: Em que ano houve mais publicações?

A segunda questão investigada é a identificação do número de publicações por ano. Como o mapeamento foi realizado no período entre 2015 e 2019, foram identificados os seguintes números: 2015 - 16 publicações; 2016 - 36 publicações; 2017 - 33 publicações; 2018 - 35 publicações; 2019 - 13 publicações. A representação desses dados é apresentada na Figura 4.

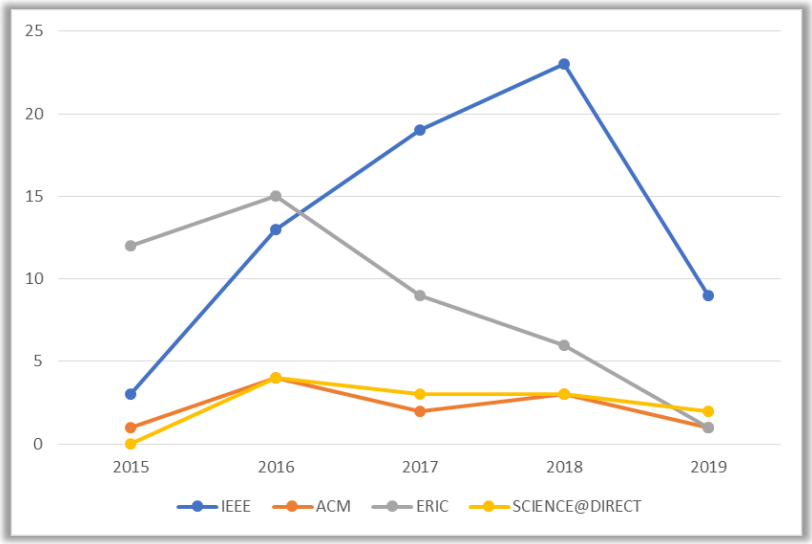
Na Figura 5, é possível visualizar melhor o número de publicações por ano em cada base de dados; analisando por base, pode-se observar nesta figura que na biblioteca IEEE há um aumento no número de publicações em 2018; e ACM (com 4 pesquisas), ERIC (15 pesquisas) e Science@Direct (4 pesquisas) apresentaram mais publicações em 2016.

Figura 4 – Número de Publicações por Ano



Fonte: Souza e Santos (2020)

Figura 5 – Número de Publicações por Ano



Fonte: Souza e Santos (2020)

QP3: Quais são os tópicos mais abordados quando se trata de MDE em cursos *e-leaning*?

A partir da leitura dos artigos, foi possível determinar sobre qual temática cada um dos estudos selecionados se tratava, e em seguida, enumerar os principais tópicos de pesquisa sobre MDE no *e-leaning*, estes podem ser listados em 14 temáticas principais: *1) Análise/predição de Comportamentos*, que aborda também – Análise/predição do desempenho (subcategoria), Análise/predição do abandono (subcategoria), Análise/predição da conclusão (subcategoria), Análise/predição do engajamento (subcategoria); estes que podem ser caracterizados como subcategorias da Análise/predição de Comportamentos, que engloba uma ampla gama de atitudes dos alunos, como por exemplo: desistir, permanecer, se engajar ou interagir nos cursos realizados*; 2) Análise de fóruns de discussão; 3) Sistemas de Recomendação; 4) Plataformas de cursos e-learning, 5) Mineração de Texto; 6) Análise de vídeo; 7) Identificação de trapaça; 8) Análise de sentimentos; 9) Análise dos currículos de cursos e-lea-*

ning; 10) Aprendizagem autorregulada; 11) Gamificação; 12) Avaliação por pares; 13) Modelo de esforço/capacidade; 14) Simulação de alunos artificiais.

A Tabela 3 mostra os temas destacados e quantidades de publicações por ano, também apresenta que investigações com ênfase na análise do comportamento dos alunos em diferentes contextos são as temáticas mais frequentes de pesquisas com MDE no âmbito do *e-learning*.

A análise/predição de Comportamento analisa como os alunos se comportam quanto realizam um curso do tipo *e-learning*, como por exemplo desenvolver uma pesquisa que relaciona o comportamento ao assistir vídeos com o desempenho em atividades avaliativas. A maioria dos trabalhos com essa tendência temática tem foco no melhoramento da experiência educacional dos cursos. A previsão de desempenho procura identificar com antecedência como será a performance do aluno no decorrer do curso, para poder intervir caso necessário e assim melhorar seu processo de aprendizagem. Artigos categorizados como predição de abandono, objetivam identificar alunos que pretendem desistir antes do encerramento do curso. Sobre predição de conclusão, é similar a predição de abandono, mas analisa o inverso. A análise do engajamento busca identificar o quanto o aluno está envolvido com a realização do curso, principalmente quantificando e qualificando as interações dos estudantes com a plataforma.

Tabela 3 – Total de Artigos por Categoria Temática

CATEGORIA	Número de Publicações por Ano					TOTAL
	2015	2016	2017	2018	2019	
Análise de Comportamento – Desempenho, Abandono, Conclusão e Engajamento	7	20	13	19	6	**65**
Análises em Fóruns de Discussão	6	2	2	5	2	**17**
Sistemas de Recomendação	0	2	8	5	1	**16**
Plataformas de Oferta	0	1	2	2	1	**6**
Mineração de Texto	1	2	2	1	0	**6**
Análises de Vídeos	1	2	1	0	1	**5**

CATEGORIA	Número de Publicações por Ano					TOTAL
	2015	2016	2017	2018	2019	
Identificação de Trapaças	0	2	3	0	0	**5**
Analises de Sentimento	0	3	0	0	0	**3**
Análise de Currículos	1	0	0	1	0	**2**
Aprendizagem Autorregulada	0	0	1	0	1	**2**
Gamificação	0	0	1	1	0	**2**
Avaliação por Pares	0	1	0	1	0	**2**
Modelo de Esforço/Capacidade	0	0	0	0	1	**1**
Simulação de Alunos Artificiais	0	1	0	0	0	**1**

Fonte: Souza e Santos (2020)

A análise de fóruns de discussão e a mineração de textos, são similares e possuem vários propósitos, dentre os quais: detecção de erros dos alunos, relevância temática, engajamento, postagens que necessitam da atenção dos professores. Geralmente, essas duas vertentes costumam possuir características interligadas, pois muitos estudos em fóruns são elaborados, por meio da mineração de texto. Além disso, foram encontrados artigos que tratam da mineração de texto em e-mails e redes sociais, em que foram realizadas análises de tópicos de discurso em redes sociais para descobrir o que alunos postam, a respeito do curso.

O tópico sistemas de recomendação também é frequente e diz respeito à implementação de sistemas que fazem alguns tipos de sugestões para usuários, como por exemplo: recomendar contatos que possuam características semelhantes; recomendar conteúdos; e também cursos dentro das plataformas. No âmbito da temática de plataformas de oferta, foi identificado um estudo inovador que propõe um novo tipo de ambiente virtual de aprendizagem com foco na análise e exploração de dados. Nas análises de vídeos os autores se concentram a detectar interações dos alunos com os vídeos, implementado maneiras de capturar essas interações, verificar suas possibilidades e impactos no desempenho.

No que se refere a identificação de trapaças em MOOCs os 5 estudos encontrados se referem ao mesmo tipo de comportamento trapaceiro, o

CAMEO *(Copying Answers using Multiple Existences Online)*. Percebeu-se que há um grande esforço em estudar esse tipo específico de irregularidade, que vem acontecendo em muitos cursos principalmente do tipo *Massive Open Online Courses* (MOOC) e está preocupando Instituições renomadas que os ofertam, como o MIT, Harvard e Stanford. O CAMEO é uma estratégia que envolve um usuário que reúne soluções para perguntas de avaliação usando uma conta de "colheita – *harvester*" e envia respostas corretas usando uma conta "mestre – *master*" separada, nesse sentido pesquisadores implementaram um método para identificação de trapaças deste tipo, com um algoritmo que utiliza MDE para rastrear os IPs *(Internet Protocol)* dos usuários e também o tempo entre postagens de atividades.

A análise de sentimentos trata de verificar as postagens dos alunos e extrair o sentimento associado, muitas vezes realizado com mineração de texto. A análise de currículo de cursos *e-leaning* busca verificar se os tipos de conteúdos apresentados aos alunos causam impactos em como esses alunos realizam os cursos. A avaliação por pares busca incentivar que os próprios alunos façam as avaliações das atividades de seus colegas, e a MDE apoia na validação destas avaliações e na verificação de sua eficácia em quanto ferramenta de ensino e aprendizagem. Quanto ao modelo de esforço/capacidade, nessa única pesquisa encontrada os autores realizaram a dosagem da complexidade das atividades e verificação do engajamento dos alunos por meio da MDE. Tais temas são muito promissores para estudos com MDE.

Por fim, cita-se como boas oportunidades de pesquisas as seguintes temáticas, no que tange a MDE no *e-leaning*: aprendizagem autorregulada – onde os estudos identificados investigaram essa teoria de análise da motivação de alunos por meio da MDE; gamificação – que é o uso de dinâmicas de jogos para engajar pessoas, resolver problemas e melhorar o aprendizado, motivando ações e comportamentos em ambientes fora do contexto de jogos, essa abordagem tem sido amplamente aplicada no *e-leaning* e há um início de análises de sua eficácia por meio da MDE; por fim, simulação de alunos artificiais – neste estudo foi realizada a geração de dados (com dimensão muito grande) criados por meio de inteligência artificial e utilizados para

treinar algoritmos, dessa forma não haverá falta de bancos de dados para treinamento dos algoritmos de Aprendizagem de Máquina ou Aprendizagem Profunda, na MDE.

CONSIDERAÇÕES SOBRE O MAPEAMENTO

Um marco importante no processo evolutivo da educação foi o surgimento dos cursos *e-leaning*, para atender às demandas advindas de um novo cenário tecnológico global. O surgimento dessa modalidade de curso contribuiu para fortalecer as mudanças nos paradigmas educacionais existentes, além de atender ao processo de democratização da educação e aos anseios do aluno com o novo perfil da era digital, cada vez mais presente nas instituições de ensino. Ademais, cursos desse tipo são capazes de gerar uma enorme quantidade de dados sobre uma grande diversidade de alunos, tais dados são mantidos pelas plataformas de oferta, fato que possibilitou explorar essa massa de dados, descobrir novos conhecimentos sobre como os indivíduos estudam, aprendem e interagem.

Foi nesse cenário que a MDE se popularizou como uma área de pesquisa interdisciplinar que trata do desenvolvimento de métodos de exploração de dados originários do campo educacional, sendo ela um dos principais mecanismos de obtenção de novos conhecimentos em grandes volumes de dados educacionais. As técnicas de MDE visam extrair informações dos dados registrados pelas plataformas durante a realização desses cursos, e podem levar à identificação de características comportamentais e indicadores relacionados à aprendizagem, as principais contribuições da MDE podem ser resumidas em: (1) a criação de modelos para melhor compreender os processos de aprendizagem; e (2) o desenvolvimento de métodos mais eficazes para apoiar a aprendizagem quando o aluno estuda utilizando um software educacional ou um Ambiente Virtual de Aprendizagem.

Nesse sentido o mapeamento sistemático exposto teve o intuito de verificar quais são os principais tópicos de pesquisa que empregaram a Mineração de Dados Educacionais em cursos *e-learning*. Os resultados

mostraram que os dados desses cursos podem ser usados em muitas temáticas de pesquisa. O mapeamento realizado apontou que a análise comportamento, análise de fóruns de discussões e sistemas de recomendação são as principais linhas de investigação, indicando ocorrências frequentes na utilização de dados de alunos matriculados em cursos *e-leaning*, para efeitos de análise, identificação e previsão desses eventos. Ademais, foram identificados tópicos ainda pouco explorados que têm potencial para serem amplamente estudados, como aprendizagem autorregulada, gamificação e simulação de alunos artificiais

CAPÍTULO 4 - MINERAÇÃO DE DADOS EDUCACIONAIS "VERSUS" ANÁLISE DE APRENDIZAGEM

A importância de analisar volumes sem precedentes de dados sobre estudantes é de grande relevância e interesse para pesquisadores da ciência de dados e da educação. Com os avanços tecnológicos atuais houve uma popularização de grandes plataformas digitais de ensino e aprendizagem, que armazenam uma grande quantidade de dados de alto valor educacional, mas é inviável analisá-los manualmente. Por esse motivo, ferramentas para analisar automaticamente esses dados são necessárias, pois esses são capazes de fornecer informações notáveis que podem ser exploradas para entender como os alunos aprendem. Verdadeiramente, um dos desafios que as instituições de ensino enfrentam na atualidade é o crescimento exponencial de dados educacionais e como transformá-los em novas ideais que podem beneficiar alunos, professores e gestores (BAKER, 2015).

À vista disso, percebeu-se potencial para novos estudos e algumas áreas de pesquisas surgiram nos últimos anos com intuito de auxiliar nessas questões. Tais áreas cresceram em conjunto com a quantidade de dados educacionais disponíveis, e podem ser exploradas para beneficiar a educação e aprimorar os processos de ensino e aprendizagem, são elas: Mineração de Dados Educacionais *(Educational Data Mining),* tema deste livro; e Análise de Aprendizagem *(Learning Analytics)* (BAKER; INVENTADO, 2014).

A Mineração de Dados Educacionais (MDE) abrange o desenvolvimento e aplicação de métodos para explorar esses tipos únicos de dados provenientes de ambientes educacionais (BAKHSHINATEGH *et al.*, 2018), podendo ser definida como a aplicação de técnicas de mineração de dados para abordar questões relativas à educação (ROMERO; VENTURA, 2013). A Análise de Aprendizagem (AA) pode ser definida como a medição, coleta, análise e relatório de dados sobre alunos e seus

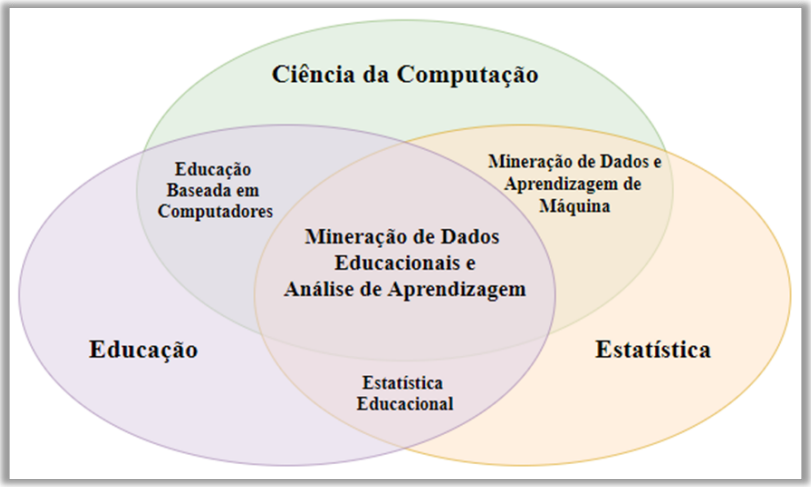
contextos, com o objetivo de entender e otimizar a aprendizagem e os ambientes em que ocorre (LANG *et al.*, 2017).

Tais abordagens compartilham um interesse em comum – métodos intensivos em dados para pesquisa educacional – e compartilham o objetivo de aprimorar a prática educacional (LIÑÁN; PÉREZ, 2015). Por um lado, AA está focada no desafio educacional e a MDE no desafio tecnológico. A AA analisa os dados para orientar a tomada de decisão e integração das dimensões técnica, social e pedagógica da aprendizagem, aplicando conhecimentos e modelos preditivos enquanto que a MDE geralmente procura novos padrões nos dados e desenvolve novos algoritmos e/ou modelos (LIÑÁN; PÉREZ, 2015).

Por fim, as diferenças entre elas são mais baseadas em foco, na investigação de forma geral, questões de pesquisa e eventual uso de modelos, do que sobre as técnicas utilizados (BAKER; INVENTADO, 2014). Independentemente das diferenças entre a AA e a MDE, as duas áreas têm sobreposição significativa, tanto nos objetivos dos pesquisadores, como nos métodos e técnicas utilizados na investigação (BAKER; INVENTADO, 2014).

Elas são áreas interdisciplinares, incluindo, entre outros, recuperação de informações, sistemas de recomendação, análise de dados visuais, mineração de dados orientada por domínio, análise de redes sociais, psicopedagogia, psicologia cognitiva, psicometria, dentre outros (ROMERO; VENTURA, 2020). De fato, elas podem ser compreendidas como a combinação de três áreas principais (Figura 6): Ciência da Computação, Educação e Estatística. A interseção dessas três áreas também forma subáreas intimamente relacionadas – a Educação Baseada em Computadores; Mineração de Dados e Aprendizagem de Máquina, e Estatística Educacional (ROMERO; VENTURA, 2020).

Figura 6 – Principais Áreas Relacionadas a MDE e AA



Fonte: Adaptado de Romero e Ventura (2020)

Devido a esta sobreposição significativa, quanto ao escopo destas duas áreas, muitos pesquisadores tem investido esforços para diferenciá-las e explicar porque existem duas áreas tão similares ao invés de uni-las para angariar mais pesquisadores e promover um avanço conjunto. Como não há uma perspectiva de que isso ocorra, pelo menos não a curto prazo, cabe indicar quais são as diferenças entre estas vertentes de análises de dados educacionais de forma mais detalhada, para apoiar pesquisadores iniciantes que ainda não sabem ao certo sobre qual concepção devem desenvolver suas investigações. Nesse sentido, são destacados os trabalhos de Siemens e Baker (2012); Moissa, Gasparini e Kemczinski (2015); e Liñán e Pérez (2015); tais contribuições são detalhadas na sequência.

SIEMENS E BAKER (2012)

Para Siemens e Baker (2012 p. 2) as comunidades de MDE e AA são definidas de maneiras relativamente semelhantes, sendo a MDE definida como: "A mineração de dados educacional é uma disciplina emergente, preocupada em desenvolver métodos para explorar

os tipos únicos de dados que vêm de ambientes educacionais e usar esses métodos para entender melhor os alunos e os ambientes que eles aprendem"; e a AA como: "[...] a medição, coleta, análise e relatórios de dados sobre os alunos e seus contextos, para fins de compreensão e otimização da aprendizagem e os ambientes em que ocorre". Para os autores as duas áreas refletem o surgimento de abordagens de educação intensivas em dados.

Extrair valor dos dados para orientar o planejamento, as intervenções e a tomada de decisões é uma mudança importante e fundamental no funcionamento dos sistemas educacionais e a Mineração de Dados Educacionais e a Análise de Aprendizagem compartilham os objetivos de melhorar a educação, melhorando a avaliação, como os problemas na educação são compreendidos e como as intervenções são planejadas e selecionadas (SIEMES; BAKER, 2012). O uso extensivo por administradores, educadores e alunos dos dados produzidos durante o processo educacional aumenta a necessidade de modelos e estratégias baseados em pesquisa. Ambas as comunidades têm o objetivo de melhorar a qualidade da análise de dados educacionais em larga escala, para apoiar tanto a pesquisa básica quanto a prática em educação (SIEMES; BAKER, 2012).

Mesmo com todas as semelhanças citadas os autores (SIEMES; BAKER, 2012) salientam que elas têm raízes diferentes e é importante observar algumas distinções, o Quadro 2 mostra algumas das principais diferenças entre as comunidades. É importante observar que essas distinções pretendem representar tendências gerais nas duas comunidades; muitos pesquisadores de MDE conduzem pesquisas que podem ser colocadas no lado da AA de cada uma dessas distinções, e muitos pesquisadores de AA conduzem pesquisas que podem ser colocadas no lado de MDE dessas distinções (SIEMES; BAKER, 2012).

Quadro 2 - Uma breve comparação da MDE e AA

CARACTE-RÍSTICAS	AA	MDE
Descoberta	Aproveitar o julgamento humano é a chave; descoberta automatizada é uma ferramenta para atingir esse objetivo	A descoberta automatizada é a chave; alavancar o julgamento humano é uma ferramenta para atingir esse objetivo
Redução e Holismo	Maior ênfase na compreensão dos sistemas como todos, em toda a sua complexidade	Maior ênfase na redução de componentes e análise de componentes individuais e relações entre eles
Origens	LAK tem origens mais fortes na web semântica, "currículo inteligente", previsão de resultados e intervenções sistêmicas	EDM tem fortes origens em software educacional e modelagem de alunos, com uma comunidade siginficiant na previsão de resultados do curso
Adaptação e Personalização	Maior foco em informar e capacitar instrutores e alunos	Maior foco na adaptação automatizada (por exemplo, pelo computador sem nenhum humano no circuito)
Técnicas e Métodos	Análise de rede social, análise de sentimento, análise de influência, análise de discurso, previsão de sucesso do aluno, análise de conceito, modelos de criação de sentido	Classificação, agrupamento, modelagem bayesiana, mineração de relacionamento, descoberta com modelos, visualização

Fonte: Siemens e Baker (2012) – *Tradução Livre*

Uma distinção importante é encontrada no tipo de descoberta que é priorizada. Em ambas as comunidades, podem ser encontradas pesquisas que usam descoberta automatizada e pesquisas que potencializam o julgamento humano por meio de visualização e outros métodos. No entanto, a MDE tem um foco consideravelmente maior na descoberta automatizada, e a AA tem um foco consideravelmente maior em alavancar o julgamento humano. Mesmo em pesquisas que combinam essas duas direções, essa preferência pode ser percebida (SIEMES; BAKER, 2012).

Esta diferença está associada a outra diferença entre as duas comunidades: o tipo de adaptação e personalização normalmente suportado pelas duas comunidades. A MDE com o maior foco na descoberta

automatizada, possui modelos mais frequentemente usados como base para a adaptação automatizada, conduzida por um sistema computacional, como um sistema de tutoria inteligente. Em contraste, os modelos de AA são mais frequentemente projetados para informar e capacitar instrutores e alunos (SIEMES; BAKER, 2012).

Uma terceira diferença importante, é a distinção entre estruturas holísticas e reducionistas. É muito mais comum na pesquisa de MDE ver pesquisas que reduzem os fenômenos a componentes e analisam os componentes individuais e as relações entre eles. O paradigma "descoberta com modelos" para pesquisa de MDE é um exemplo claro deste paradigma. Em contraste, os pesquisadores de AA normalmente colocam uma ênfase mais forte na tentativa de entender os sistemas como um todo, em toda a sua complexidade (SIEMES; BAKER, 2012).

MOISSA, GASPARINI E KEMCZINSKI (2015)

Com o objetivo de analisar as diferenças entre MDE e AA Moissa, Gasparini e Kemczinski (2015) desenvolveram um mapeamento sistemático de literatura em sete Mecanismos de Busca Acadêmica: Web of Knowledge, Engineering Village, Scopus SciVerse, IEEE Xplore, ACM Digital Library, Science Direct e Springer Link; deste processo de mapeamento foram selecionados 280 artigos, dos quais 82 estavam associados a MDE, 186 a AA e 12 não tinham uma associação evidenciada pelos autores.

De acordo Moissa, Gasparini e Kemczinski (2015), por meio do mapeamento sistemático realizado foi possível identificar que as duas áreas possuem definições e objetivos similares. Tanto a MDE como a AA tem o intuito de analisar dados educacionais para entender o processo de ensino-aprendizagem e otimizá-lo. Em muitos dos trabalhos analisados as autoras relataram que as pesquisas foram conduzidas de forma parecida e até mesmo aplicando as mesmas técnicas e concluíram que poucas diferenças foram encontradas.

Na MDE há uma predominância da utilização de técnicas típicas de Mineração de Dados como a Aprendizagem de Máquina (exemplo: agrupamento e classificação) e estatística; e na AA são empregadas na

maioria das vezes as mesmas técnicas, todavia com o intuito visualização da informação que são mais desenvolvidas na AA; nos trabalhos de AA analisados, as autoras ainda distinguiram uma técnica não utilizada nos trabalhos de MDE, a Análise de Redes Sociais (MOISSA; GASPARINI; KEMCZINSKI, 2015). Outras diferenças encontradas pelas autoras se baseiam no uso de algoritmos (maior em MDE) e de instrumentos (maior em AA). No que se refere aos dados, às intervenções, aos aspectos éticos e de privacidade nenhuma diferença significativa foi encontrada pelas autoras. No que tange aos resultados, estes também são similares em cada área (MOISSA; GASPARINI; KEMCZINSKI, 2015).

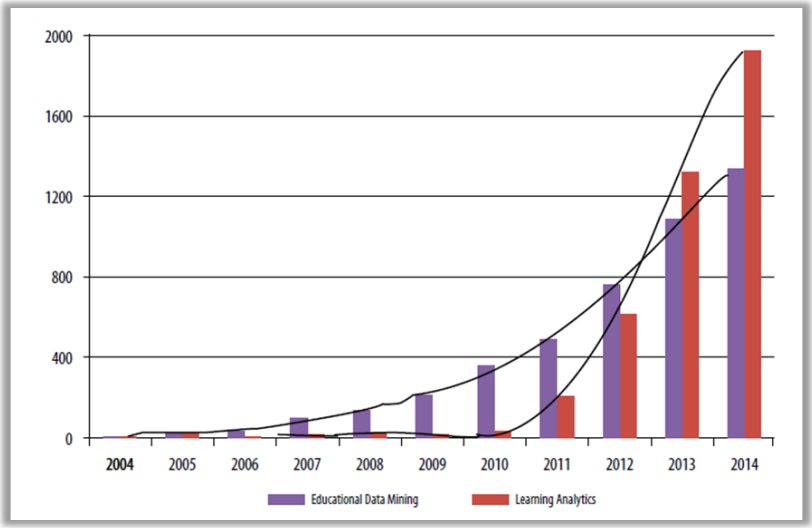
Com base neste mapeamento, Moissa, Gasparini e Kemczinski (2015) concluíram que há diferenças sutis entre a Mineração de Dados Educacionais e a Análise de Aprendizagem; sobretudo com relação ao foco de pesquisa; em que a AA tem especialmente o intuito de entender o processo de ensino-aprendizagem como um todo e desenvolver ferramentas para auxiliar neste processo, com uma abordagem mais direcionada para os fatores humanos e utiliza as técnicas de mineração como uma ferramenta. Todavia, a MDE procura empregar as técnicas de mineração mais tecnológicas como a Aprendizagem de Máquina em novos contextos e desenvolver e otimizar métodos/modelos para realizar tais análises, focando mais no fator tecnológico que no humano (MOISSA; GASPARINI; KEMCZINSKI, 2015).

LIÑÁN E PÉREZ (2015)

Liñán e Pérez (2015) destacaram que a Mineração de Dados Educacionais e a Análise de Aprendizagem são campos de pesquisa relativamente novos e promissores que visam melhorar as experiências educacionais, ajudando as partes interessadas (instrutores, alunos, administradores e pesquisadores) a tomar melhores decisões usando dados. Seu crescimento foi impulsionado pelo aumento da capacidade do computador para armazenar e analisar grandes quantidades de dados e pela disponibilidade de métodos e técnicas estatísticas, de aprendizado de máquina e de mineração de dados (LIÑÁN; PÉREZ, 2015).

Liñán e Pérez (2015) relatam haver uma sobreposição entre a MDE e a AA, relativamente considerável. Mesmo assim, algumas diferenças são destacadas na literatura, embora as duas áreas possuam basicamente o mesmo objetivo: melhorar a qualidade da educação por meio da análise de grandes quantidades de dados para extrair informações úteis para as partes interessadas. Empresas representativas em outros setores, como indústria, finanças ou saúde, já introduziram técnicas estatísticas, de aprendizado de máquina e de mineração de dados para obter um melhor desempenho por meio de decisões baseadas em dados históricos. A popularidade da MDE e da AA tem crescido desde o início da década de 2010 (Figura 7), embora a pesquisa em MDE tenha começado alguns anos antes (LIÑÁN; PÉREZ, 2015), nesse sentido os autores destacam a evolução conjunta no número de publicações nessas áreas.

Figura 7 – Evolução das referências MDE e AA no Google Scholar



Fonte: Liñán e Pérez (2015)

Quanto as diferenças Liñán e Pérez (2015) corroboram a visão de Siemens e Baker (2012), que essas dizem respeito basicamente a: Descoberta; Redução e holismo; Origens; Adaptação e personalização; e Técnicas e métodos. Os autores complementam os relatos sobre as

diferenças salientado que a AA cobre mais disciplinas do que MDE. Além da ciência da computação, estatística, psicologia e ciências da aprendizagem, AA está relacionada à ciência da informação e a sociologia. Portanto, mesmo que a fronteira entre os dois campos seja difusa e suas diferenças sejam parcialmente baseadas em suas origens e tendências, elas ainda são significativas para esses autores.

Liñán e Pérez (2015) ainda destacam que corroboram a visão de que as duas áreas não devem se unir e que a coexistência de ambas as comunidades de pesquisas leva a uma contribuição mais diversa e relevante para a sociedade, consequentemente, a comunicação e a competição entre estas áreas devem ser encorajadas.

CONSIDERAÇÕES SOBRE A MINERAÇÃO DE DADOS EDUCACIONAIS "VERSUS" ANÁLISE DE APRENDIZAGEM

Há um valor positivo em ter diferentes comunidades envolvidas em como explorar *"Big Data"* para melhorar a educação, em particular, existem diferentes padrões e valores para "boa pesquisa" e "pesquisa importante" em cada comunidade, permitindo a criatividade e o avanço que, de outra forma, não ocorreria em uma única cultura de pesquisa monolítica (SIEMENS; BAKER, 2012). Por exemplo, os pesquisadores de MDE colocaram maior foco nas questões de generalização do modelo; por outro lado, os pesquisadores de AA colocaram maior foco em atender às necessidades de várias partes interessadas com informações extraídas dos dados. Cada uma dessas questões é importante para o sucesso de longo prazo de ambos os campos, uma oportunidade importante para as duas comunidades aprenderem uma com a outra (SIEMENS; BAKER, 2012).

As comunidades MDE e AA preveem que o impacto dos dados e análises na educação será transformador nos níveis primário, secundário e pós-secundário (SIEMENS; BAKER, 2012). Um ambiente de pesquisa aberto e transparente é vital para impulsionar este importante trabalho, como disciplinas de pesquisa conectadas, mas distintas, MDE

e AA podem fornecer um consolidado campo de investigação para a excelência em pesquisa de dados educacionais, orientando formuladores de políticas, administradores, educadores e desenvolvedores de currículo, para a implantação de melhores práticas na próxima era de educação baseada em dados (SIEMENS; BAKER, 2012).

Por fim, destaca-se que existem inúmeras semelhanças entre os dois campos de pesquisa, como objetivos, metodologias e técnicas. No entanto, existem várias diferenças, atribuíveis principalmente às suas origens e tendências. A coexistência das duas respectivas comunidades científicas leva à competição com efeitos positivos na sociedade (LIÑÁN; PÉREZ, 2015; SIEMENS; BAKER, 2012). Apesar das grandes expectativas e da quantidade de trabalhos em MDE e AA, sua aplicação em ambientes educacionais ainda esbarra em algumas barreiras importantes, como a falta de uma cultura orientada a dados e de rapidez, abrangência e facilidade de uso e compreensão de ferramentas que podem ser integradas na plataforma digitais de ensino e aprendizagem.

Cabe evidenciar a importância que as técnicas associadas a mineração de dados têm no que se refere as áreas de MDE e AA. Essas áreas utilizam de técnicas intensivas para analisar dados e extrair informações relevantes, tais informações podem levar em especial gestores educacionais e professores a tomar decisões importantes, que impulsionem melhorias sobretudo no processo de ensino e aprendizagem. Nesse sentido, duas das mais relevantes técnicas aplicadas para implementação do processo de Mineração de Dados Educacionais são apresentadas de forma detalhada no Capítulo seguinte.

CAPÍTULO 5 - PRINCIPAIS TÉCNICAS DE MINERAÇÃO DE DADOS EDUCACIONAIS

Neste capítulo serão abordadas duas técnicas muito empregadas em Mineração e Ciência de Dados, por isso também amplamente utilizadas em Mineração de Dados Educacionais. Essas técnicas correspondem a Aprendizagem de Máquina (*Machine Learning*) e a Aprendizagem Profunda (*Deep Learning*). Primeiramente serão tratados os aspectos que dizem respeito a Aprendizagem de Máquina (AM), em que inicialmente ela é definida, em seguida os tipos de aprendizado de máquina existentes são apresentados, em conjunto com as funções associadas e os respectivos algoritmos e posteriormente são descritos os métodos para avaliação de modelos de AM. Em seguida é caracterizada a Aprendizagem Profunda (AP), a princípio é exposto a conceituação de AP, suas características, qual cenário propiciou sua evolução e as principais arquiteturas de AP existentes e em seguida é descrita com mais detalhes a arquitetura de AP mais empregada em problemas de classificação.

APRENDIZAGEM DE MÁQUINA

A AM consiste em extrair informações dos dados, é uma área de pesquisa formada pela interseção da Estatística, Inteligência Artificial (IA) e Ciência da Computação, muitas vezes é referenciada como análise preditiva ou aprendizado estatístico e muitos pesquisadores defendem que ela é um campo da Inteligência Artificial (BISHOP; PATTERN, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MITCHELL, 1997). Pode-se dizer que a AM basicamente investiga como os computadores podem melhorar seu desempenho com base em dados.

A primeira definição de AM foi elaborada por Samuel (1959) que a definiu como "Campo de estudo que permite que os computadores efetuem operações sem serem explicitamente programados" (SAMUEL, 1959). Uma das definições mais amplamente utilizadas para definir AM é a feita por Mitchell (1997): "Um programa de computador aprende a

partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E" (MITCHELL, 1997). Com essa afirmação o autor quer dizer que o termo especulativo de "aprendizado" empregado a essa técnica é uma forma de explicar que o sistema faz a mesma tarefa, ou tarefas, sobre um mesmo conjunto de dados de uma maneira mais eficiente a cada execução. Mitchell (1997) complementa dizendo que AM é uma área que se ocupa por investigar métodos computacionais adequados para a aquisição de novos conhecimentos, novas habilidades e novas formas de organização do conhecimento já existente. Dessa forma, o campo do aprendizado de máquina é norteado pela questão de como produzir programas que automaticamente melhoram com a sua experiência.

Para Navarro *et al.* (2017) a técnica de AM oferece soluções para automatizar a análise de *Big Data*, os autores afirmam que ela pode ser considerada como um conjunto de métodos que podem detectar automaticamente padrões nos dados, então pode-se usar esses padrões descobertos para fazer previsões ou para tomada de decisão. Nesse sentido, Alpaydin (2010) explica de forma minuciosa a AM:

> A AM tem a função de programar computadores para otimizar seu desempenho usando dados de exemplo ou experiências anteriores. Com a aplicação de algoritmos de AM gera-se um modelo definido até alguns parâmetros, e o aprendizado é o treinamento do algoritmo de AM para otimizar os parâmetros desse modelo usando os dados de treinamento ou experiência anterior. O MODELO pode ser preditivo para fazer previsões ou descritivo para obter conhecimento dos dados, ou ambos. O aprendizado de máquina usa a teoria da estatística na construção de modelos matemáticos, porque a tarefa principal é a inferência sobre os dados de uma amostra (ALPAYDIN, 2010).

Em complemento as definições apresentadas, Bishop (2011) faz a seguinte consideração sobre AM:

> Os computadores são baseados na lógica, mas precisam lidar cada vez mais com dados do mundo real, cheios

> de incerteza e ambiguidade. As abordagens modernas de aprendizagem de máquina usam a teoria da probabilidade para quantificar e calcular com essa incerteza e levaram a uma proliferação de aplicações nessa área, variando de sistemas de recomendação, pesquisa na web e de filtros de spam até reconhecimento de voz. Além disso, o advento da ampla conectividade da Internet, com armazenamento centralizado de dados, e do processamento em conjunto com algoritmos desenvolvidos para inferência probabilística computacionalmente eficientes, têm potencial para possibilitar muitas novas oportunidades para o aprendizado de máquina nos próximos anos (BISHOP, 2011).

Nota-se o entusiasmo de Bishop (2011) a respeito da AM, e como foi previsto muito se tem pesquisado nessa área, por isso muitas outras definições podem ser encontradas na literatura, algumas dessas acepções são apresentadas no Quadro 3.

Quadro 3 – Definições de Aprendizagem de Máquina

AUTORES	DEFINIÇÃO DE AM
Obermeyer e Emanuel (2016)	Em particular, AM como uma estrutura algorítmica, pode fornecer informações úteis sobre os dados, facilitar a inferência e derivar conhecimento.
Liu e Salinas (2017)	AM é um campo da ciência da computação e uma parte da IA, que se refere à ciência e engenharia pelas quais as máquinas, sistemas de computador, podem analisar e adquirir informações sobre os dados. A AM pode ajudar a desenvolver modelos de dados aprimorados usando técnicas matemáticas avançadas e manipulação de conjuntos de dados complexos e heterogêneos.
Senders *et al.* (2018)	Em um subconjunto da IA, os algoritmos de AM são capazes de assimilar padrões e se autocorrigir sem explícita programação.
Waring; Lindvall; Umeton, (2020)	A AM é uma técnica-chave que demonstra a capacidade de traduzir grandes conjuntos de dados em conhecimento acionável.
Zhou; Zheng; Zhang, (2020)	Na academia, a AM é usada para a previsão precisa de desempenho, contribuição quantificável para cada variável e análise de desempenho estocástico.

AUTORES	DEFINIÇÃO DE AM
Alqudah; Yaseen, (2020	Realizar um processo com AM significa que o computador pode descobrir uma solução sem ser especificamente programado. Ou seja, as máquinas são capazes de aprender e lidar continuamente com grandes conjuntos de dados usando algoritmos classificadores. Classificadores, que categorizam observações, são considerados a espinha dorsal da AM.
Hegde; Rokseth, (2020)	O campo de AM é um subconjunto da Inteligência Artificial e o AP é um subconjunto do AM. O termo ciência de dados é um campo que utiliza técnicas de IA, AM, AP e ciência da computação.
Lei *et al.* (2020)	Os algoritmos de AM são capazes de aprender de forma adaptativa, e adquirir conhecimento sobre diversos contextos, partir dos dados coletados, em vez de utilizar a experiência e o conhecimento de especialistas.
Weis; Jutzeler; Borgwardt, (2020)	Os métodos de AM são capazes de encontrar estatísticas e dependências nos dados, considerando também os efeitos não lineares e de interação entre esses recursos. Assim, as técnicas de aprendizado de máquina podem descobrir informações novas ou desconhecidas.
(RANA *et al.* 2020)	A AM abrange os algoritmos ou modelos estatísticos, que podem identificar padrões e construir hipóteses ou inferências baseadas na aprendizagem sobre os conjuntos de dados observados. A AM cresceu e evoluiu, à medida que a escala de informações aumentou e foi usada para identificar recursos significativos de grandes conjuntos de dados.

O contexto de AM é complexo, com propriedades vindas de várias áreas, dessa forma não é simples entender todas as terminologias utilizadas, assim para que fique mais fácil de entender do que tratam algumas questões que serão abordadas no decorrer dessa proposta de tese, é imperativo que se tenha conhecimento de alguns termos e conceitos utilizados quando se trata dessa temática, tais terminologias estão descritas no Quadro 4.

Quadro 4 – Termos relevantes para entendimento da AM

TERMO/ CONCEITO	DEFINIÇÃO
Treinamento	É a fase em que o algoritmo de AM é aplicado a uma base de dados e este deve assimilar padrões sobre os dados, e depois deve poder generalizar esses padrões.
Modelo	Modelo é o produto da submissão dos dados a um algoritmo de AM. Contextualizando, os dados são os subsídios para o algoritmo assimilar padrões, e essa assimilação é persistida por um modelo que é o resultado da aplicação de um algoritmo a uma base de dados, ou seja, o modelo é diferente de um algoritmo, embora seja também uma estrutura codificada, em outras palavras um programa. Depois da fase de treinamento de um algoritmo ele gera um modelo que pode ser salvo para ser posteriormente aplicado a uma nova base de dados, desde que essa possua os mesmos atributos. Se assemelha a um modelo gerador em probabilidade/estatística.
Teste	É a etapa que vem depois do modelo ser gerado, durante o teste o modelo deve ser aplicado a dados diferentes dos usados no decorrer do treinamento para que se possa avaliar sua eficácia, essa avaliação é realizada por meio de métodos e medida por métricas próprias da AM.
Base de Dados	Corresponde a uma matriz com várias linhas e colunas, formatada de forma que possibilite receber a aplicação de uma técnica de mineração de dados, como um algoritmo de AM ou AP.
Instância	É a referência a uma linha da base de dados, equivale a um elemento ou indivíduo do qual se tem informações, exemplo: no contexto de uma loja uma instância poderia ser o cliente, no contexto de um curso MOOC uma instância poderia ser um aluno.
Atributo	É uma coluna da base de dados, corresponde a uma informação que se tem sobre uma instância, por exemplo se a base de dados estiver relacionada a pessoas alguns atributos poderiam ser: cor dos olhos, peso ou altura.
Vetor de Atributos	Se refere a uma linha completa da base de dados, a instancia juntamente com todos os seus atributos.
Rótulos	Corresponde à classe de uma instância, por exemplo em uma base de dados de alunos, esses podem ser classificados como concluintes ou desistentes, então esses são os rótulos dos alunos, é um sinônimo para classe ou categoria.

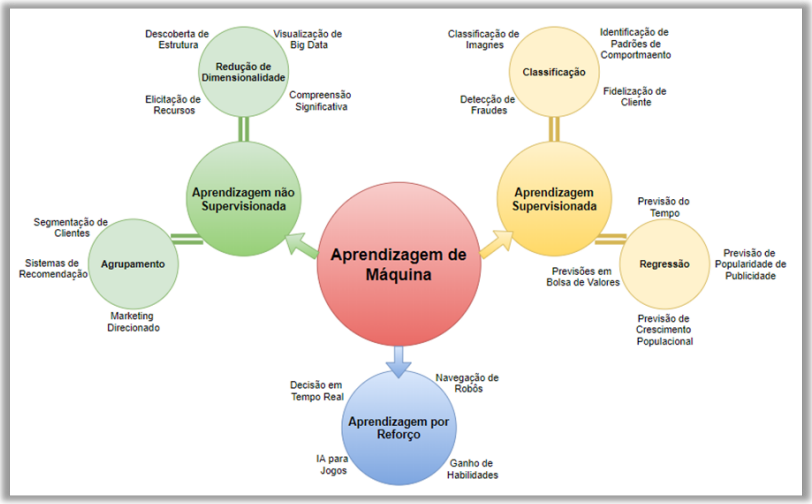
Fonte: Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Devido a essa complexidade da AM ela foi dividida, para que sua interpretação e aplicação se tornassem processos mais compreensíveis e específicos, para certificar que cada problema possuísse uma abordagem própria para sua solução. Nesse sentido, na seção a seguir são apresentados os tipos de aprendizagem de máquina que correspondem a essa divisão.

TIPOS DE APRENDIZAGEM DE MÁQUINA

A AM engloba funções próprias para cada tipo de problema levantado e algoritmos estritamente relacionados a cada uma dessas funções, para que se tornasse mais simples de entender todos os métodos que envolvem essa área ela foi dividida em tipos de aprendizagem, nesse sentido Kubat (2017) salienta que os principais tipos de AM são: Aprendizagem Supervisionada; Aprendizagem não Supervisionada e Aprendizagem por Reforço. Na Figura 8 a relação entre os tipos aprendizagem e suas funções é apresentada.

Figura 8 – Tipos de Aprendizagem de Máquina e suas funções



Fonte: Adaptado Kubat (2017)

Na aprendizagem por reforço o treinamento dos modelos acontece quanto o agente computacional pode tomar uma sequência de decisões,

então esse agente deve assimilar o que está acontecendo ao seu redor e atingir um propósito em um espaço desconhecido e eventualmente adverso, esse tipo de aprendizagem está baseado na premissa de tentativa e erro, muito utilizada em robótica e em jogos digitais, um exemplo clássico de aplicação desse tipo de aprendizagem são os robôs de exploração espacial. Como a aprendizagem por reforço está inserida em contextos diversos a MDE, não será retratada com mais detalhes. Nas seções seguintes serão caracterizadas a Aprendizagem Supervisionada e Não Supervisionada, as funções ligadas a cada uma e os respectivos algoritmos com base nos trabalhos de Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015).

APRENDIZAGEM SUPERVISIONADA

Na Aprendizagem Supervisionada, a instâncias de entrada e a categoria correspondente à qual essas instâncias pertencem são fornecidos para o algoritmo. O algoritmo de AM assimila a relação entre a entrada e a saída e, em seguida, prevê a saída para amostras de dados de entrada cujas as saídas não são fornecidas.

Por exemplo, o algoritmo de aprendizado de máquina supervisionado é alimentado com imagens de maçãs classificadas como frutas e batatas classificadas como vegetal. Após o treinamento nesses dados, o algoritmo de aprendizado de máquina supervisionado deve ser capaz de categorizar novas imagens, que não possuem uma classificação, de maçãs como frutas e batatas como vegetais. As etapas necessárias para aplicação de algoritmos de aprendizado de máquina supervisionado podem ser descritas da seguinte forma:

1. Alimente o algoritmo com os registros de entrada X e os rótulos de saída y;
2. Para cada registro de entrada, o algoritmo prevê uma saída y';
3. O erro na previsão é calculado subtraindo y de y';
4. O algoritmo se corrige removendo o erro; e
5. As etapas 1 a 4 continuam por várias iterações até o erro ser minimizado.

Em termos matemáticos, tem-se a variável de entrada X e a variável de saída y, e é necessário encontrar uma função que capture relação entre os dois, ou seja, y = f (X).

A Aprendizagem Supervisionada é utilizada para resolver dois tipos diferentes de problemas: *Classificação* e *Regressão*, também chamadas de *funções* de Aprendizagem Supervisionada. A *Classificação* refere-se ao processo de previsão de valores de saída discretos para uma entrada, por exemplo, dado uma entrada o algoritmo de classificação prevê se um e-mail é spam ou legítimo, se um tumor é benigno ou maligno, se um aluno será aprovado ou reprovado no exame. Enquanto que na *Regressão*, a tarefa do modelo de aprendizado de máquina é prever um valor contínuo, por exemplo para determinada entrada os algoritmos de regressão podem prever o preço da casa, ou as notas obtidas por um aluno em um exame.

Os principais algoritmos de Aprendizagem Supervisionada aplicados a problemas de Classificação e Regressão são: Naïve Bayes, K-Vizinhos mais Próximos (*K-Nearest Neighbors* – KNN); Árvores de Decisão, Floresta Aleatória (*Random Forest* - RF), Máquinas de Vetores de Suporte (*Suport Vector Machines),* utilizados tanto para Classificação como para Regressão; e Regressão Linear, utilizado apenas para Regressão. Esses algoritmos são detalhados na sequência.

NAÏVE BAYES

O algoritmo Naïve Bayes é um algoritmo supervisionado de aprendizado de máquina baseado no teorema de Bayes e fundamentado no princípio de independência de recurso, que afirma que os recursos de um conjunto de dados não têm relação entre si. Por exemplo, uma fruta pode ser considerada banana se tiver 15 cm ou mais de comprimento, de cor amarela e 1 cm de diâmetro. O Naïve Bayes não se preocupa se esses recursos dependem um do outro, a fruta é declarada como banana por contribuição independente desses recursos. Devido a essa suposição de independência, o algoritmo tem essa denominação de ingênuo e é o mais simples de todos os algoritmos de aprendizado de máquina e, no entanto, é muito aplicado por ser eficaz. Matematicamente, o teorema de Bayes pode ser representado como:

Equação 1 – Teorema de Bayes

$$P\,(A|B) = \frac{P\,(A|B).P\,(A)}{P\,(B)}$$

Onde os termos mencionados acima são explicados da seguinte forma: 1) *P(A | B)* é a probabilidade de o evento A ocorrer, dado o conjunto de atributos B; 2) *P(A)* é a probabilidade anterior da ocorrência do evento A; 3) *P(B|A)* é a probabilidade do conjunto de atributos se o evento A ocorrer; e 4) *P(B)* é a probabilidade prévia da ocorrência de preditores.

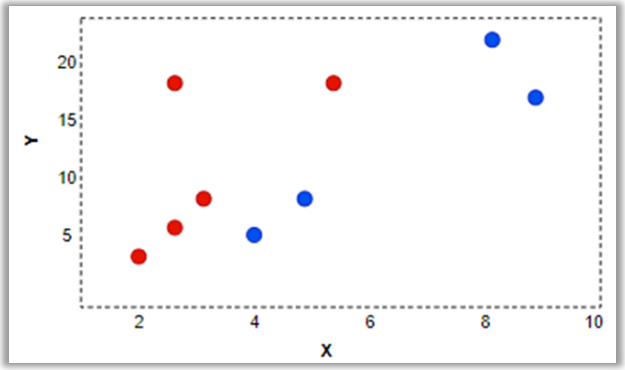
As principais aplicações do Algoritmo Naïve Bayes são: 1) Problemas de várias classes, onde é comumente empregado para classificação de texto como em problemas de análise de emoções e filtragem de spam em e-mail; 2) Combinação de algoritmos de filtragem colaborativa para construção de sistemas de recomendação, baseados em aprendizado de máquina; e 3) Como o algoritmo é rápido em comparação com outros mais avançados, é incorporado em aplicações em que tempo é um requisito crítico.

KNN (*K-NEAREST NEIGHBORS*)

KNN é um algoritmo não paramétrico, isso indica que esse algoritmo não presume que a relação entre as entradas e saídas de seus dados sigam uma função matemática em particular. O processo executado pelo KNN é simples, no treinamento o algoritmo calcula as distâncias existentes entre os atributos das instâncias da base de dados, e assimila os padrões de distância entre os dados da mesma classe, escolhendo os K atributos de dados mais próximos. Quando uma nova instância é apresentada ao algoritmo ele calcula a distância de seus atributos em comparação aos das instâncias de treinamento, dessa forma ele classifica esse novo registro com a classe que tiver a menor distância. Os cálculos utilizados podem ser a distância de Manhattan, ou a distância Euclidiana. Para utilização desse algoritmo é muito importante que os atributos da base de dados estejam escalonados, pois dimensões discrepantes nos dados ocasionam problemas nos resultados obtidos com essas fórmulas.

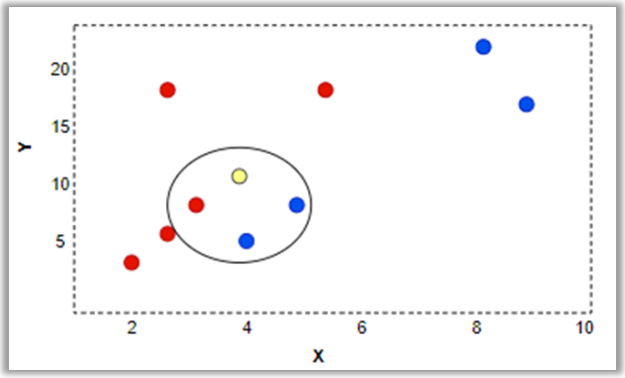
Como exemplo do funcionamento do KNN suponha-se alguns dados no espaço bidimensional, divididos em duas categorias: vermelho e azul, como mostrado na Figura 9. Considere que é necessário classificar uma nova instância, para isso deve-se calcular a distância dos atributos das instâncias vermelhas e azuis até os atributos do novo registro. Suponha que a nova instância seja o círculo amarelo e o valor de K seja três, como indicado na Figura 10, dessa forma percebe-se que dos três vizinhos mais próximos do círculo amarelo, dois são azuis e um é vermelho. Portanto, essa nova instância será classificada como azul.

Figura 9 – Exemplo do funcionamento do KNN



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Figura 10 – Classificação de uma nova instância pelo KNN



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Uma informação relevante sobre KNN é que ele não gera um modelo durante a fase de treinamento, na verdade ele usa a base de dados de treinamento como exemplo para classificar um novo registro, por isso ele é conhecido como algoritmo de aprendizado lento. Diferentemente de outros algoritmos, a lista de parâmetros exigidos pelo algoritmo KNN não é complexa, só é necessário especificar o número de vizinhos mais próximos, que corresponde ao K e o tipo de função de distância, além disso é um algoritmo rápido quando comparado por exemplo a Árvores de Decisão, uma vez que ele não gera um modelo na fase de treinamento. O KNN é simples de implementar e dados podem ser adicionados a qualquer momento, uma vez que sempre que for necessário realizar uma previsão a distância com todos os atributos dos dados é recalculada.

ÁRVORES DE DECISÃO

A Árvore de Decisão é um algoritmo de AM baseado em entropia, o princípio por trás de seu trabalho é que cada atributo no conjunto de dados é tratado como um nó na árvore de decisão. Em cada nó é tomada uma decisão sobre qual caminho escolher na árvore, dependendo do valor do atributo nesse nó específico, o processo continua até que o nó da folha seja alcançado, porque esse contém a decisão final sobre a classificação da instância.

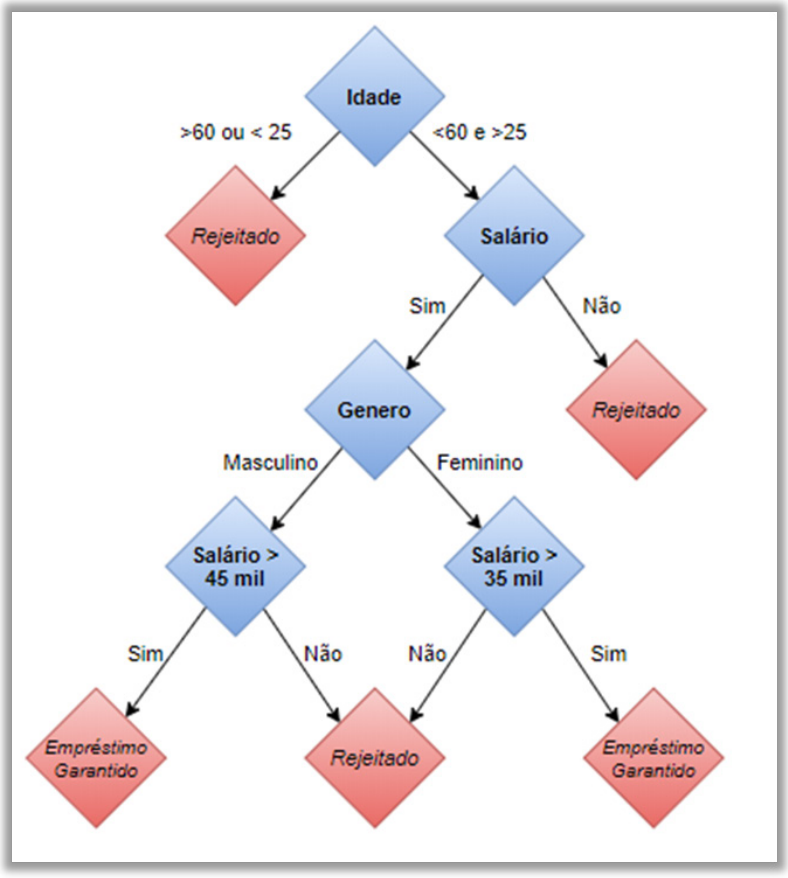
Como exemplo suponha-se que exista um banco que precisa decidir se um empréstimo deve ser concedido a um cliente específico ou não, para isso esse banco possui atributos de clientes, incluindo idade, sexo e salário em sua base de dados. Dessa forma, podem ser definidos critérios que consistem em um conjunto de regras que define se o empréstimo será concedido ou não, essas regras podem ser:

1. Se a idade do cliente for maior que 25 e menor que 60, vá para a próxima etapa. Senão, simplesmente rejeite o pedido de empréstimo.
2. Se a primeira condição for atendida, verifique se a pessoa está assalariada ou não. Se a pessoa é assalariada, vá para a etapa 3; se a pessoa estiver desempregada, rejeite o pedido de empréstimo.

3. Se a pessoa for assalariada e o gênero for masculino, vá para a etapa 4. Caso contrário, se o gênero for feminino, vá para a etapa 5.
4. Se o salário for superior a 35 mil reais por ano, conceda o empréstimo, caso contrário, rejeite o pedido.
5. Se o salário for superior a 45 mil reais por ano, conceda o empréstimo, caso contrário, rejeite o pedido.

A árvore de decisão baseada em tais regras é representada na Figura 11.

Figura 11 – Representação de uma Árvore de Decisão



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

O conjunto de regras exposto como exemplo é bastante simples e foi escolhido aleatoriamente, em problemas reais os dados são muito mais complexos e técnicas estatísticas como entropia são usadas para criar esses nós. Entropia refere-se à impureza da classificação em base de dados rotuladas. Basicamente em árvores de decisão, o recurso que resulta em entropia mínima nos rótulos de saída é definido como o nó raiz. Por exemplo, se 95% das vezes em que a idade for maior que 60 e menor que 25, o pedido de empréstimo for rejeitado, a impureza será de 5% para a idade com valores entre 60 e 25. Da mesma forma, se em 80% dos casos o empréstimo para desempregados for rejeitado, a impureza no rótulo de saída para o atributo assalariado será de 20%.

Os principais benefícios do Algoritmo de Árvore de Decisão são: 1) Funciona igualmente bem para tarefas de regressão e classificação, pode prever de forma precisa valores contínuos e discretos; 2) Pode ser usado para classificar dados lineares e não lineares; e 3) Em comparação com a maioria dos outros algoritmos de aprendizado de máquina tem processo de treinamento bastante rápido e fácil de ser interpretado.

FLORESTA ALEATÓRIA (*RANDOM FOREST*)

Uma única Árvore de Decisão pode ser enviesada, dependendo dos dados, uma abordagem que pode melhorar essa falha é utilizar várias árvores de decisão que fazem sua própria previsão e a previsão final é encontrada calculando a média de todas as previsões feitas por todas as árvores. Essa abordagem é conhecida como *ensemble learning* (aprendizado em conjunto). No aprendizado em conjunto, vários algoritmos de tipos iguais ou diferentes são unidos para criar uma maior capacidade para o modelo de AM, a Floresta Aleatória é um tipo de modelo de aprendizado em conjunto, esse algoritmo une vários algoritmos de Árvore de Decisão, criando uma floresta. O funcionamento do algoritmo Floresta Aleatória se baseia na execução das seguintes etapas:

1. Escolhe-se K atributos de dados aleatórios no conjunto de dados;
2. Cria-se um algoritmo de regressão ou classificação de árvore de decisão baseado nesses K atributos de dados;

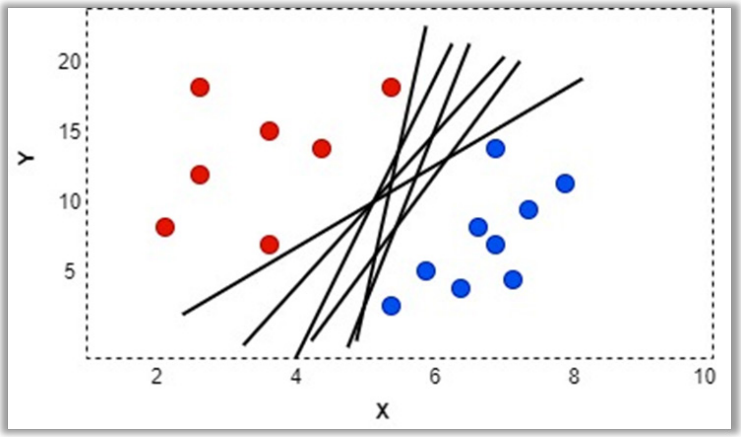
3. Seleciona-se o número de árvores para o algoritmo Floresta Aleatória e execute as etapas 1 e 2 em cada árvore;
4. Se o problema for de regressão, cada árvore prediz um valor contínuo e a saída final pode ser calculada tomando a média dos valores previstos por todas as árvores. Se o problema em questão for de classificação, cada árvore prevê um valor discreto e a classificação final pode ser selecionada por maioria de votos.

O algoritmo de Floresta Aleatória é um dos algoritmos mais estáveis independentemente da quantidade de dados, uma vez que existem várias árvores, a introdução ou remoção instâncias na base de dados pode afetar uma pequena parte dessas árvores, mas não todas, assim a estabilidade geral do algoritmo não é afetada. Além disso, ele funciona bem no caso de recursos numéricos e categóricos, e não é necessário escalonar os atributos da base de dados, pois ele não depende da distância entre os atributos das instâncias.

MÁQUINAS DE VETORES DE SUPORTE (MVS) (*SUPORTE VECTOR MACHINES*)

O algoritmo MVS se originou nos anos 60 e é um dos mais famosos algoritmos de AM, e tem sido muito utilizado desde então, antes das Redes Neurais Artificiais se popularizarem ele era considerado o algoritmo de AM mais preciso. Iniciando as explicações de como o MVS funciona indica-se que o intuito da regressão linear em um espaço de recurso bidimensional, é encontrar uma linha reta que separa com êxito os dados de diferentes classes, no entanto no mundo real, pode haver vários limites de decisão que podem classificar os dados com êxito como indicado na Figura 12.

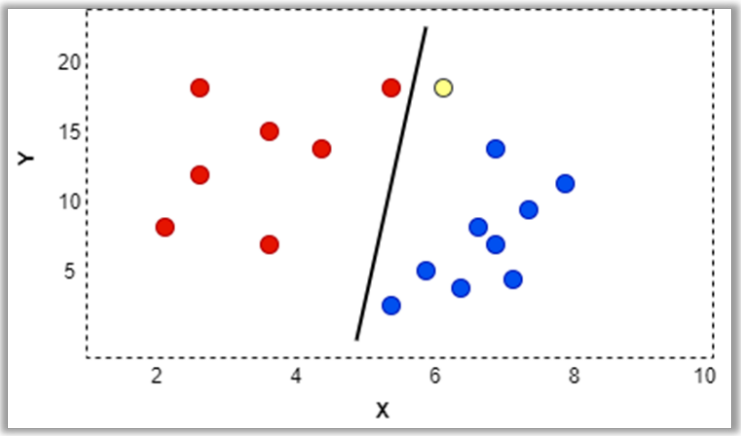
Figura 12 – Exemplo de funcionamento do MVS



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

No entanto, se uma nova instância será classificada ou não com sucesso, isso depende do limite de decisão escolhido para classificação. Como exemplo, suponha-se que seja necessário classificar uma nova instância, ou seja, o círculo amarelo, se a decisão for como na Figura 13, a nova instância será classificada como azul.

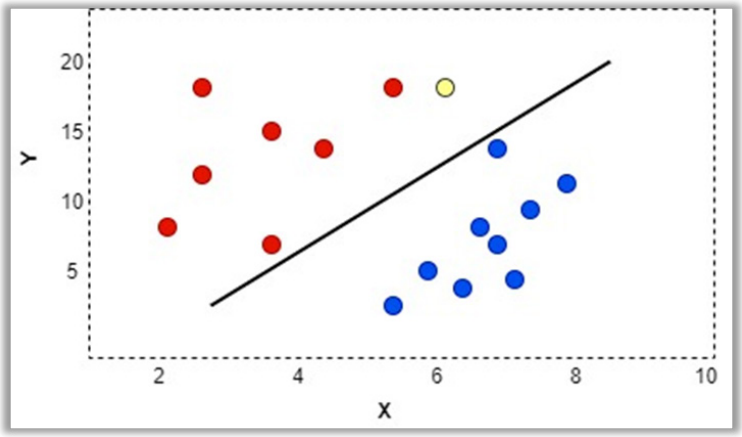
Figura 13 – Classificação de uma nova instância pelo MVS



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Por outro lado, se o limite de decisão for como na Figura 14, a nova instância será classificada como vermelha. Nas Figuras 13 e 14, observa-se que pode haver vários limites de decisão que classificam com êxito um conjunto de dados, no entanto, nem todos eles são ideais.

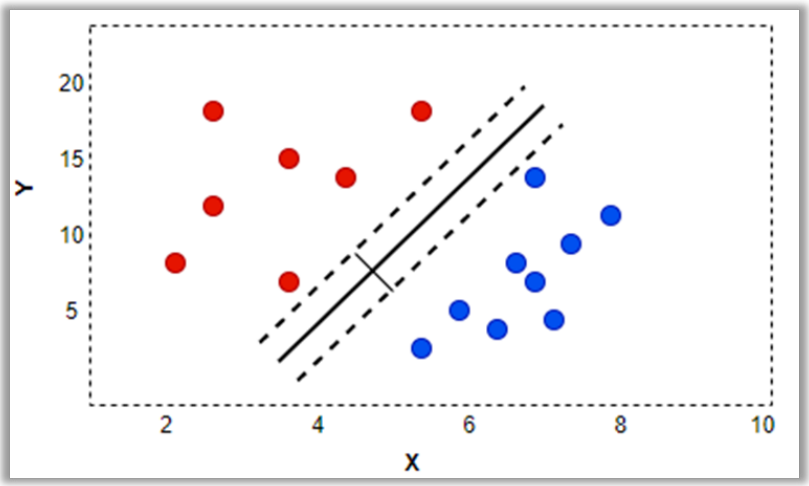
Figura 14 – Classificação de nova instância com limite diferente de decisão



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Dada uma nova instância, limites de decisão diferentes podem classificá-las diferentemente, o real objetivo do algoritmo MVS é encontrar o limite de decisão que classifica os registros de tal maneira que as chances da classificação ser incorreta seja minimizada. O algoritmo faz isso maximizando a distância entre os atributos de instâncias mais próximos de todas as classes na base dados e ele consegue encontrar esse limite com a ajuda de vetores de suporte, por isso o seu nome. Os vetores de suporte passam pelos atributos de dados mais próximos das duas classes para classificação, o trabalho do algoritmo é maximizar a distância entre esses dois vetores, traçando uma linha paralela no meio deles, esse limite de decisão é considerado o limite de decisão ideal, que pode ser ilustrado como na Figura 15.

Figura 15 – Limite de decisão do MVS



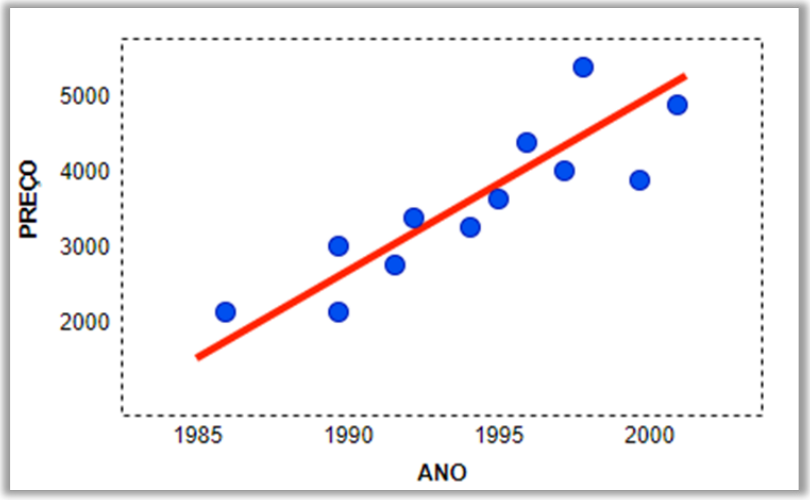
Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

Os exemplos utilizados demonstram um problema linear, entretanto uma das razões, pela qual o MVS é tão amplamente aplicável é que ele pode ser facilmente estendido para bases de dados complexas que não são linearmente separáveis. Isso é feito mapeando os registros de treinamento para um espaço de maior dimensão, onde eles se tornam um conjunto linearmente separável, essa técnica é denominada truque do *kernel* (*kernel trick*).

REGRESSÃO LINEAR

A regressão linear é uma abordagem exclusiva para regressão que identifica o relacionamento entre duas ou mais variáveis, ela é capaz de encontrar uma função linear que mapeia variáveis independentes com base nas variáveis dependentes, se esta função é plotada no espaço bidimensional, resulta em uma linha reta. Como exemplo, pode-se citar o seguinte cenário em que se deve encontrar a relação entre o preço de alguns carros e o ano de fabricação. Se o gráfico dessa relação for elaborado o ano será o eixo x e o preço o eixo y, e o algoritmo de regressão linear encontrará uma linha reta que melhor se ajusta aos dados, essa relação é mostrada na Figura 16.

Figura 16 – Representação da Regressão Linear



Fonte: Adaptado Kubat (2017); Igual e Seguí (2017); e Aggarwal (2015)

A linha é representada pela função $y = ax +$, em que *y* é a variável dependente, *a* é a inclinação da reta, *x* é a variável independente e *c* é a interceptação em *y*. Ao analisar a equação, percebe-se que que *x* e *c* são atributos dos dados, o algoritmo de regressão fornece a inclinação e a interceptação que melhor se ajusta ao conjunto de dados. Este conceito pode ser estendido para mais de uma variável independente, em que equação da função de regressão linear pode ser representada por $y = a1x1 + a2x2 + \cdots + anxn +$, na qual o *n* é o número total de variáveis independentes. Esta equação representa basicamente um hiperplano com *n* dimensões. Cabe salientar que o modelo de regressão linear de duas dimensões pode ser representado como uma linha, em três dimensões é representado na forma de plano e em mais de três dimensões é representado como hiperplano.

APRENDIZAGEM NÃO SUPERVISIONADA

Na Aprendizagem não Supervisionada, os algoritmos são alimentados com os base de dados sem rótulos, sendo assim é de responsabilidade do algoritmo identificar padrões nos registros e agrupamentos com caracterís-

ticas semelhantes. Normalmente, a maioria dos conjuntos dados do mundo real é não classificado; portanto, o aprendizado não supervisionado pode ser usado como precursor do aprendizado supervisionado. Por exemplo, a tendência dos clientes em assistir filmes em uma plataforma de *streaming* pode ser inserida em um algoritmo de aprendizado não supervisionado, e esse pode encontrar tendências, como por exemplo: sempre que cliente de perfil X assiste um filme de terror ele também assiste uma animação. Portanto, uma decisão de *marketing* seria sempre que pessoas com perfil X assistissem a um filme de terror seja oferecido em seguida uma animação.

A Aprendizagem não Supervisionada compreende duas principais funções, a *Redução de Dimensionalidade* e o *Agrupamento*. Em diversos casos as bases dados disponíveis para serem analisadas possuem muitos atributos, dos quais muitos podem não serem significativos para o processo de extração de conhecimento, nesse sentido reduzir a dimensão desses dados pode economizar recursos computacionais sem impactar na eficácia dos algoritmos, nesse contexto os algoritmos de *Redução de Dimensionalidade* são extensamente utilizados. Em relação ao *Agrupamento* é o processo de agrupar objetos semelhantes, ou seja, particionar instancias de uma base de dados não classificada, essa função pode ajudar a descobrir novas categorias de maneira não supervisionada, mesmo quando nenhuma similaridade entre os atributos é perceptível.

Os principais algoritmos para Redução de Dimensionalidade e Agrupamento são Análise dos Componentes Principais (*Principal Component Analysis*) e *K-means* respectivamente, esses algoritmos são descritos em seguida.

ANÁLISE DOS COMPONENTES PRINCIPAIS (ACP) (*PRINCIPAL COMPONENT ANALYSIS*)

A ACP é um dos algoritmos mais amplamente usados para redução de dimensionalidade, ele trabalha por seleção de recursos que causam variação máxima na saída, deixando para trás recursos que não têm efeito na saída. A intuição por trás dessa abordagem é que a variação pode ser usada como uma medida para distinguir a produção; portanto, apresenta

os responsáveis por distinguir os resultados que são mais importantes e, portanto, devem ser selecionados. O primeiro componente principal é o recurso que resulta em variação máxima, da mesma forma o segundo componente principal é o recurso que causa a segunda maior variação e assim por diante. A Análise dos Componentes Principais tem duas vantagens principais: 1) Número reduzido de recursos significa tempo de treinamento reduzido, portanto, tempo de execução mais rápido; 2) Com um número reduzido de recursos, os dados podem ser facilmente visualizados. É importante também mencionar que os dados devem ser escalonados antes da aplicação do algoritmo ACP, pois a variação pode ser grande para os atributos expressos em unidades mais altas, e o algoritmo pode ficar tendencioso em relação aos recursos não escalonados.

K-MEANS

O algoritmo *K-means* é amplamente empregado em tarefas de agrupamento de dados, em bases não classificadas e fundamenta-se especialmente na medida de similaridade. As etapas do agrupamento realizado pelo *K-means* são:

1. Escolhe-se aleatoriamente o número de centroides K – ponto central para um agrupamento – onde K corresponde ao número de grupos.
2. Encontra-se a distância entre todos os atributos das instâncias e todos os centroides. A distância pode ser Euclidiana ou Manhattan.
3. Associa-se as instâncias aos centroides de menor distância – é feita uma comparação do distanciamento dos atributos das instâncias aos centroides inicializados, para aquele centroide que a distância for menor é conferido o pertencimento da instância a esse determinado grupo.
4. Repete-se a etapa 3 para todas as instâncias formando K conjuntos.
5. Atualiza-se a localização de cada centroide, utilizando a média de todos os atributos de instâncias no grupo.
6. Repete-se as etapas 2, 3 e 4 até que as posições atualizadas de todos os centroides sejam iguais às posições anteriores.

Embora o *K-means* não ofereça garantias de precisão, sua simplicidade e velocidade são muito atraentes na prática, ao se utilizar o *K-means* alguns cuidados devem ser levados em conta, o primeiro diz respeito ao valor dos atributos, pois como esse é um algoritmo que utiliza a fórmula da distância é recomendável padronizar os dados para determinar a semelhança entre os atributos das instâncias, o segundo relaciona-se a inicialização dos centroides, pois devido sua natureza iterativa e à inicialização aleatória, os centroides podem ficar muito longe do valor ideal e não convergir, devido a isso é recomendado o uso de diferentes formas para sua inicialização.

Finalizando a explanação sobre os tipos de aprendizagem de máquina, suas funções e algoritmos, deve ser salientado que depois dos algoritmos de Aprendizagem Supervisionada serem aplicados, para descobrir se os modelos gerados são precisos, métodos e métricas de avaliação devem ser empregados, caso contrário não há como saber se esses são eficazes para o contexto da aplicação, nesse sentido esses métodos e métricas são apresentados na seção em sequência. Como os algoritmos de Aprendizagem não Supervisionada não possuem parâmetros para serem comparados, esses não são frequentemente avaliados por métodos pré-existentes, na maioria dos casos esses são avaliados por especialistas humanos.

AVALIAÇÃO DE MODELOS DE CLASSIFICAÇÃO E REGRESSÃO

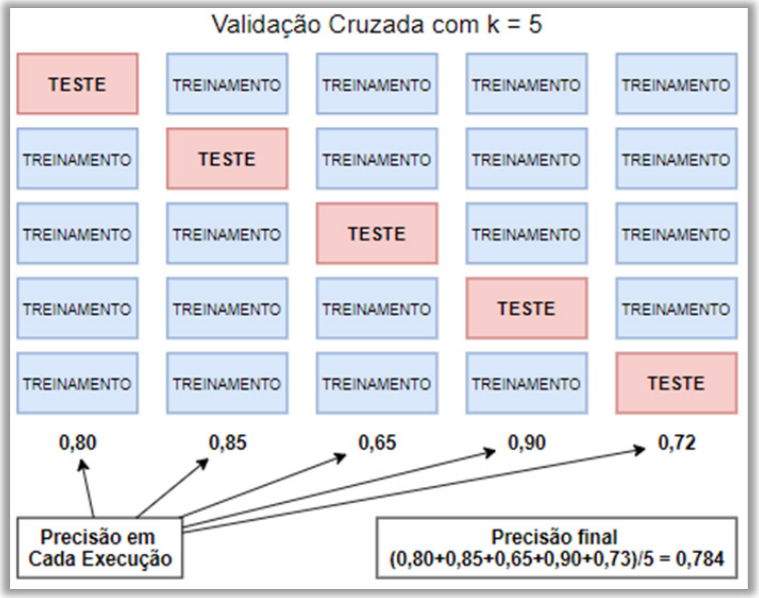
Para realizar a verificação dos resultados de um modelo de classificação ou regressão são necessários dois itens os *métodos de avaliação* e as *métricas de interpretação*, os dois devem ser aplicados em conjunto para que seja possível observar se um modelo é eficaz ou não. Os métodos indicam como esse modelo será avaliado, e as métricas traduzem os resultados da aplicação desses métodos em números que possam ser interpretados.

Os dois principais métodos de avaliação são Treinamento e Teste e Validação Cruzada. No Treinamento e Teste a base de dados é dividida de forma aleatória em duas porções, uma para treinamento e outra para teste, de acordo com Japkowicz e Shah (2014) geralmente fica 85% das instâncias para treinamento e 15% para teste. O algoritmo ao ser apli-

cado sobre a base de treinamento recolhe informações sobre os atributos das instâncias e gera um modelo de classificação ou regressão com base nesses atributos e informações, após isso esse modelo é aplicado sobre a base de teste (que contém registros diferentes da base de treinamento) e então as métricas de avaliação são calculadas sobre essa aplicação.

O método de Validação Cruzada é considerado mais detalhado e preciso (JAPKOWICZ; SHAH, 2014), nesse método o usuário escolhe em quantas partes quer dividir seus dados, geralmente usa-se entre 5 e 10 partições. Se foi escolhido, por exemplo em 5 partições o método seleciona – de forma aleatória – 4 partes para treinamento e gera o modelo, depois testa em uma dessas partes e calcula a eficácia. Em seguida ele seleciona mais 4 partes para treinamento, sendo que a partição reservada para o teste vai ser diferente da execução anterior, e também calcula sua eficácia, assim sucessivamente até que todas as 5 partições tenham sido utilizadas para teste, ao final é feita uma média de todas as execuções para encontrar a eficácia geral do modelo, a Figura 17 mostra como as etapas citadas funcionam.

Figura 17 – Etapas da Validação Cruzada



Fonte: Autora

Como observado, apenas a aplicação do método de avaliação não indica se o modelo é eficaz ou não, para isso devem ser utilizadas métricas que possibilitem interpretação do quanto o modelo foi preciso em suas classificações, em outras palavras quantificar o seu desempenho. As métricas mais utilizadas no contexto de avaliação de modelos de Aprendizagem Supervisionada estão resumidas no Quadro 5.

Os métodos de avaliação citados nessa seção são utilizados tanto para problemas de Classificação como de Regressão, no que diz respeito as métricas são utilizadas em especial no contexto da Classificação, embora a acurácia e a matriz de confusão sejam utilizadas para ambas as abordagens. Não obstante, a AM seja uma das técnicas mais aplicadas a Mineração de Dados em diversos âmbitos, há também outras técnicas que podem ser utilizadas para exploração e análise de dados, como a Aprendizagem Profunda que é uma abordagem da Aprendizagem de Máquina que trata exclusivamente de Redes Neurais Artificiais, dessa forma a seguir são apresentados os aspectos relevantes a respeito dessa técnica.

Quadro 5 – Métricas de Avaliação de Algoritmos de AM

MÉTRICA	DEFINIÇÃO
Matriz de confusão	Uma matriz de confusão fornece uma análise mais detalhada das classificações corretas e incorretas para cada classe. Uma breve explicação de como interpretar uma matriz de confusão é a seguinte: os elementos da diagonal principal representam o número de pontos para os quais o rótulo previsto é igual ao rótulo verdadeiro, enquanto qualquer coisa fora da diagonal principal foi rotulada incorretamente pelo classificador. Portanto, quanto mais altos os valores presentes na diagonal principal da matriz de confusão, melhor, indicando muitas previsões corretas.
Precisão da classificação (Acurácia)	A precisão é uma métrica de avaliação comum para problemas de classificação. É o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes. Em outras palavras, é a porcentagem de instâncias classificadas corretamente de todas as instâncias, pode ser considerada mais útil em uma classificação binária do que em problemas de classificação de várias classes, porque pode ser menos claro exatamente como a precisão se divide nessas classes.

MÉTRICA	DEFINIÇÃO
Intervalo de Confiança (IC)	Corresponde a uma métrica que indica que há uma probabilidade de 95% que a verdadeira precisão do modelo algorítmico testado esteja dentro desse intervalo.
Taxa de não informação	Essa é a precisão alcançável, sempre prevendo o rótulo da classe majoritária. Portanto, corresponde a melhor escolha, sem outras informações.
Valor de P	Consiste em um teste unilateral para verificar se a *precisão* é melhor que a *taxa de não informação*, considerando a maior porcentagem da classe dos dados.
Kappa	Corresponde a uma medida de concordância usada em escalas nominais que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas as interpretações são. É parecida com a à precisão, excetuando por ser normalizada na linha de base do acaso no conjunto de dados. É passível de considerada uma medida mais utilizada para problemas com desequilíbrio nas classes.
Área sob curva (AUC) – Taxas de Sensibilidade e Especificidade	A área sob a curva é uma métrica de desempenho para medir a capacidade de um classificador binário de discriminar entre classes positivas e negativas. Exemplos: 1) Uma área de 1,0 representa um modelo que fez todas as previsões perfeitas; 2) Uma área de 0,5 representa um modelo tão bom quanto aleatório. A AUC pode ser dividida em *Sensibilidade* e *Especificidade*. *Sensibilidade* é a verdadeira taxa positiva, são as instâncias numéricas da classe positiva que realmente foram previstas como positivas. A *Especificidade* é a verdadeira taxa negativa, ou seja, é o número de instâncias da classe negativa que foram realmente previstas como negativa.
Valores Preditivos Positivo e Negativo	*Valor preditivo positivo* – mostra o número da classe positiva prevista corretamente como uma proporção do total de previsões da classe positiva realizadas. *Valor preditivo negativo* – mostra o número da classe negativa prevista corretamente como uma proporção do total de previsões da classe negativa realizadas. Esses parâmetros descrevem o desempenho de um teste de diagnóstico.
Prevalência	Mostra com que frequência a classe positiva realmente ocorre na amostra.
Taxa de detecção	Denota o número de previsões positivas corretas da classe feitas como uma proporção de todas as previsões realizadas.
Prevalência de detecção	Apresenta o número de previsões positivas de classe feitas como uma proporção de todas as previsões realizadas.

MÉTRICA	DEFINIÇÃO
Precisão Balanceada	Atribui essencialmente a média das taxas reais positivas e negativas, isto é – (sensibilidade + especificidade) / 2.

Fonte: Adaptado Japkowicz e Shah (2014).

APRENDIZAGEM PROFUNDA

A AP é uma abordagem especial da AM, que abrange todos os tipos de "Aprendizagem", Supervisionada, não Supervisionada e por Reforço, e busca também estendê-los, para resolver problemas de Inteligência Artificial que geralmente requerem mais capacidade de processamento do que disponível na AM. A AP é muito empregada para solução de problemas relacionados à visão computacional, reconhecimento de fala, processamento de linguagem natural e reconhecimento de áudio, e se baseia na implementação de Redes Neurais Artificiais. Destaca-se, que na AP não há uma variedade de algoritmos como na AM, nesse âmbito são implementadas apenas as Redes Neurais Artificiais Multicamadas (RNAM), conhecidas também como Redes Neurais Profundas.

Para Lecun, Bengio e Hinton (2015) AP é definida da seguinte forma:

> A AP permite que modelos computacionais compostos por várias camadas de processamento aprendam representações de dados com vários níveis de abstração. Esses métodos melhoraram drasticamente o estado da arte em reconhecimento de fala, reconhecimento visual de objetos, detecção de objetos e muitos outros domínios, como descoberta de medicamentos e estudos sobre o genoma. Com a otimização do algoritmo de retropropagação para indicar como a máquina deve alterar seus parâmetros internos as Redes Neurais Profundas são capazes de descobrir estruturas complexas de informações a partir de grandes conjuntos de dados. As Redes Convolucionais trouxeram avanços no processamento de imagens, vídeo, fala e áudio, enquanto as Redes Recorrentes se sobressaíram em dados sequenciais, como texto e fala (LECUN; BENGIO; HINTON, 2015).

Para se ter uma visão geral de como os pesquisadores definem a AP no Quadro 6 são sistematizadas algumas definições disponíveis em publicações da área.

Quadro 6 – Definições de Aprendizagem Profunda

AUTORES	DEFINIÇÃO DE AP
Goodfellow, Bengio e Courville (2016)	A AP é uma subparte da AM, baseada no cálculo, que busca aprender com alto nível de abstração, usando múltiplas camadas de processamento, criadas por vários mecanismos lineares e não lineares. Com o uso da aprendizagem profunda é obtida uma ilustração para a informação de forma precisa. As principais abordagens para AP são: (1) Classificação e (2) Segmentação. Na classificação, identifica-se a classe do objeto e se fornece um rótulo para esse objeto, enquanto que na Segmentação, os dados são divididos em segmentos.
XIN *et a*l. (2018)	As etapas da AP são semelhantes às da AM. Ao contrário da AM, na AP o método de extração de recursos é automatizado em vez de manual. A AP é o único método para superar os desafios da extração de recursos no *Big Data* (BD). Os três principais tipos aprendizagem presentes na AM e AP são: Aprendizado Supervisionado, Aprendizagem não Supervisionada e Aprendizagem por Reforço.
Yang, Zhang e Su (2018)	A AP é uma nova área de AM, que se desenvolveu rapidamente. Além das Máquinas Restritas de Boltzmann (RBM), mais citadas, outras arquiteturas de AP são: codificador automático, rede neural de convolução, rede de empilhamento profundo e rede neural recorrente. As várias camadas ocultas presentes em sua estrutura, são o que lhe confere alta performance em comparação a outras técnicas.
Soffer *et al*. (2019)	A AP é um subcampo de AM em que o computador aprende sem a seleção de recursos humanos. A maioria dos modelos de AP são baseados em sistemas de Redes Neurais Artificiais, inspiradas em sistemas de redes neurais biológicas. Tais redes, são coleções de neurônios artificiais dispostos em camadas e trabalhando em uníssono, essas são estruturadas em: camada de entrada, camada intermediária – ou camadas ocultas – que realizam cálculos e uma camada de saída.
Sengupta *et al*. (2020)	Modelos de AP são arquiteturas extremamente poderosas para encontrar padrões entre diferentes combinações lineares, ou não lineares, de diferentes tipos de dados. Derivam, representações necessárias e relevantes dos dados sem a necessidade de extração manual de recursos. Nos últimos anos, algoritmos de AP estão substituindo a maioria dos algoritmos tradicionais de AM, e na maioria dos casos, superando os classificadores tradicionais.

AUTORES	DEFINIÇÃO DE AP
Murat *et al.* (2020)	Algoritmos de AP, recentemente se tornaram populares e fornecem meios para coleta de conhecimento sem a necessidade de engenharia de recursos, realizadas em sua grande maioria por humanos.
Badar, Haris e Fatima (2020)	Nos últimos anos, algoritmos de AP surgiram como métodos revolucionários que superaram a maioria dos métodos de ponta. A capacidade das redes profundas de explorar características simples, bem como características composicionais complexas de representações de dados, é referido como a razão por trás do seu sucesso.
Boulemtafe, Derhab e CHallal (2020)	A AP é uma das abordagens mais avançadas da AM e atraiu muita atenção na pesquisa, em especial por prover a capacidade de superar a dependência de recursos mapeados à mão, que são enfrentados pelos algoritmos de AM tradicionais. A AP ou, as Redes Neurais Profundas, geralmente compreendem duas fases na sua aplicação: uma etapa de treinamento para otimizar a precisão do modelo e uma fase de inferência em que o modelo é usado para análises como classificação/predição ou regressão.
Sezer, Gudelek e Ozbayoglu (2020)	AP é um tipo de Rede Neural Artificial que se constitui de diversas camadas de processamento computacional que proporciona alto nível de generalização para modelagem de dados. A principal vantagem dos modelos de AP é extrair recursos dos dados de entrada automaticamente.
Le, Torrisi e Pollastri (2020)	AP é um subcampo de AM baseado em Redes Neurais Artificiais Multicamadas, que enfatizam o uso de múltiplas camadas conectadas para transformar entradas em recursos passíveis de prever saídas correspondentes. Dado um conjunto de dados suficientemente grande de pares entrada-saída, um algoritmo de AP pode ser usado para aprender automaticamente o mapeamento de entradas com relação as saídas, ajustando um conjunto de parâmetros em cada camada da rede.

A datar de 2006, com a publicação de Hinton, Osindero e Teh (2006), na qual os autores apresentam um algoritmo (retropropagação) para o treinamento de Redes Neurais Profundas que as tornaram mais rápidas e eficientes, a AP despontou como um campo da AM, que baseia-se em um algoritmo capaz de modelar generalizações sobre dados utilizando um código composto de três tipos de camadas de processamento: 1) *Camada de entrada:* que recebe os atributos externos da base de dados; 2) *Camada oculta:* que encapsula as operações matemáticas intermediárias e dá suporte para a rede definir os resultados finais (na AP são empregadas

várias camadas ocultas); e 3) *Camada de saída:* responsável por indicar a classificação da instância analisada. Nesse sentido, o termo "profundo" refere-se à essa multiplicidade de camadas e a uma grande quantidade de camadas ocultas, por meio das quais os atributos são transformados.

Além da otimização do algoritmo para o treinamento, a emersão e a consolidação da AP só foram possíveis, por causa do aumento do volume de dados disponíveis e do avanço dos recursos computacionais (LECUN; BENGIO; HINTON, 2015). Nessa perspectiva, o principal desafio da aplicação da AP sempre foi a dependência de grandes volumes de dados para expandir seu potencial em cada contexto de aplicação, pois como as redes neurais são treinadas, por meio de exemplos, sua eficiência melhora à medida que a quantidade de dados processados aumenta. À medida que a AP evolui, ela é capaz de descobrir conhecimentos novos, antes improváveis, sobre conjuntos de dados, nesse sentido pode-se dizer que a era do *Big Data (BD)* possibilitou um amplo potencial de inovação para essa técnica, e por causa dessa disponibilidade de dados, a AP tem tido sucesso, superando o estado da arte (LECUN; BENGIO; HINTON, 2015), o que resultou em muitas aplicações em diversas áreas do conhecimento.

Em relação aos avanços dos recursos computacionais, destaca-se seu aspecto decisivo para a difusão da AP, sobretudo porque as Redes Neurais Profundas possuem várias camadas de processamento, estruturas que dependem de muitos recursos computacionais, por vezes não disponíveis em um computador comum. Por isso, sua popularização só foi possível com o surgimento de ferramentas tecnológicas que propiciaram o treinamento dessas redes em servidores on-line, essas ferramentas são popularmente conhecidas como GPUs (*Graphics Processing Unit)*, porque nos servidores onde acontecem o processamento é esse hardware que opera em conjunto com o processador, no intuito de acelerar a análise dos dados, a GPU recebe uma parte da aplicação, enquanto o processador recebe o restante, paralelizando a execução, e assim o processamento fica mais rápido. Desta forma, observou-se nos últimos anos o aumento do desempenho e disponibilidade de GPUs para cálculos matriciais paralelos em servidores on-line acessíveis, como: Pytorch (Facebook),

TensorFlow (Google), MXnet (Apache) e H2O (RStudio) que aceleraram a utilização da AP. Tudo isso, permitiu a adoção e pesquisa acentuada utilizando essa técnica, que está sendo cada vez mais aplicada em processos de tomada de decisão e se configura um campo de pesquisa ativa (LECUN; BENGIO; HINTON, 2015).

Em suma, a otimização dos algoritmos de treinamento para Redes Neurais Profundas; a disponibilidade massiva de dados na internet; e os avanços tecnológicos nas ferramentas para construção, treinamento e teste desses algoritmos, facilitaram a utilização e propiciaram a disseminação da AP. Todavia, ainda existem alguns aspectos relevantes para a implementação de Redes Neurais Profundas que influenciam diretamente no seu desempenho, esses estão listados no Quadro 7.

Quadro 7 – Aspectos da Implementação de Redes Neurais Profundas

ASPECTO	DEFINIÇÃO
Seleção de variáveis	Na maior parte das aplicações, as bases de dados possuem um grande número de atributos, os quais são inseridos com o intuito de alcançar um melhor desempenho. Porém, muitas vezes, esses dados são redundantes ou irrelevantes. Por isso, o desafio está na seleção dos atributos mais relevantes dentre todos da base de dados.
Limpeza de dados	Comumente, na etapa de obtenção da base de dados, seu pré-processamento pode não corrigir valores incorretos, ausência de dados, e/ou inconsistências. Dessa maneira, a limpeza de dados é importante no desempenho do modelo. Esse problema pode ser resolvido muitas vezes pelo conhecimento dos limites dos dados, a identificação de *outliers*, e o uso de informações estatísticas para correção de dados ausentes ou incorretos.
Representação das variáveis	Conforme o tipo de dado, os atributos podem ser codificados para obter um melhor treinamento. Dessa forma, os dados podem ser representados como discretos ou contínuos.
Normalização	A normalização é importante quando existem dados muito dispersos. Nesse cenário, valores muito altos podem sobrecarregar a função de ativação. Desse modo, uma forma fácil de normalizar os dados é somar todas as variáveis e dividir pela soma. Nesse sentido a padronização também pode ser empregada, essa consiste em subtrair da variável o valor médio e dividir pelo desvio padrão.

ASPECTO	DEFINIÇÃO
Generalização	Um fator bastante importante no decorrer do processo de treinamento é a garantia de que a rede tenha uma boa generalização. O método mais comum, para isso consiste em reservar uma parte da base de dados para validar a generalização, ou seja, para realizar os testes.

Fonte: Adaptado Lecun, Bengio e Hinton (2015)

Outra particularidade relevante na AP é que, embora não possua uma grande variedade de algoritmos, em decorrência do aumento de sua aplicação em diversos contextos, ela evoluiu e se especificou para cada problema e cenário ao qual é empregada, tais especificidades foram agrupadas em arquiteturas que descrevem os tipos de Redes Neurais Profundas mais utilizadas. Nesse sentido, Aggarwal (2018) considera como as principais arquiteturas: Redes Multilayer Perceptrons (MLP), Redes Neurais Convolucionais, Redes Neurais Recorrentes e Redes pré-treinadas – não supervisionadas, essas são descritas na sequência.

1. *Redes Multilayer Perceptrons (MLP)*: Essa arquitetura é a base para os algoritmos de AP, os modelos gerados por redes MLP são utilizados principalmente em problemas de classificação com o intuito de identificar padrões em conjuntos dados. A MLP é uma das arquiteturas mais utilizadas em processos de tomadas de decisão, pertencente à classe de redes neurais de aprendizagem supervisionada. As Redes Neurais do tipo MLP são formadas por uma camada de entrada; uma camada de saída; e entre esses dois, um número facultativo de camadas ocultas, mas que deve ser no mínimo duas, que correspondem ao efetivo meio de processamento computacional de uma rede neural artificial com essa arquitetura.
2. *Redes Neurais Convolucionais:* São Redes Neurais Profundas inspiradas na visão de seres vivos, empregadas em particular no reconhecimento de imagens e identificação de objetos. Os modelos gerados a partir de Redes Neurais Convolucionais são empregados principalmente em problemas de reconhecimento de imagem. Seu nome indica que a rede emprega uma operação matemática chamada convolução, que é um tipo especializado de operação linear. As redes convolucionais são simplesmente Redes Neurais Profundas que usam a convolução no lugar da multiplicação em pelo menos uma de suas camadas.

3. *Redes Neurais Recorrentes:* São as únicas que permitem a operação em uma sequência de vetores ao longo do tempo, elas são projetadas para tratamento de problemas que envolvam dados sequenciais, como frases de texto, séries temporais, e outras tipos de sequências. Essas redes são aplicadas em muitas tarefas de mineração de texto, pois conseguem capturar sua natureza sequencial, desta forma são utilizadas para o entendimento da linguagem e tentam prever a próxima palavra, ou conjunto de palavras, ou sentenças de alguns casos, com base nos casos anteriores. As Redes Neurais Recorrentes são redes com loops, permitindo que as informações persistam, assim elas podem se conectar com informações anteriores para uma tarefa atual. Nesse tipo de arquitetura, se enquadram as *Long Short-Term Memory* (LSTM), que é um tipo de Rede Neural Recorrente que armazena valores em intervalos arbitrários, adequada para classificar, processar e prever séries temporais com intervalos de tempo de duração desconhecida.

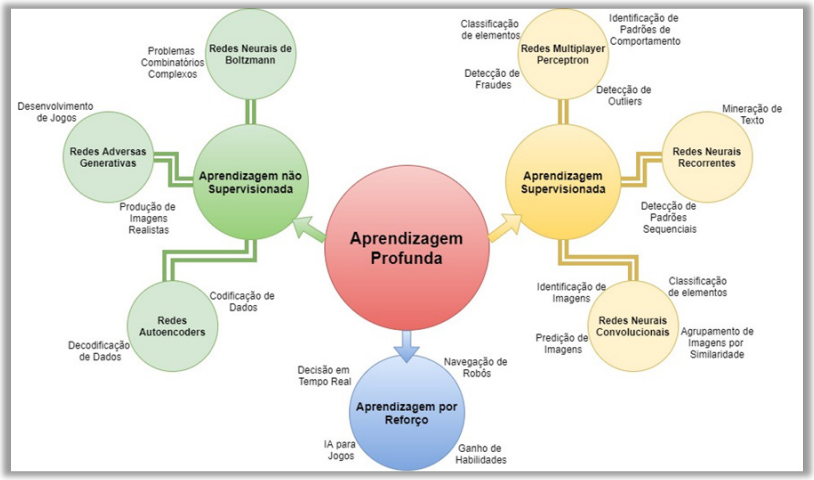
4. *Redes pré-treinadas – não supervisionadas:* O Codificador Automático (*Autoencoders*) se enquadra nesse tipo de arquitetura, ele corresponde a uma Rede Neural Profunda usada para aprendizado não supervisionado de recursos com eficiência de codificação e decodificação de dados, seu objetivo é aprender a representar os dados de entrada, para redução da dimensionalidade, compactação e fusão (possui os mesmos objetivos da Análise dos Componentes Principais). Outro exemplo dessa arquitetura são as Redes Neurais de Boltzmann, um modelo generativo não direcionado, que utiliza suas camadas ocultas para modelar a distribuição sobre as camadas visíveis (entrada e saída) e são capazes de aprender modelos internos e de representar e resolver problemas combinatórios complexos. As Redes Adversas Generativas *(Generative Adversarial Networks – GAN)*, configuram-se também como redes de aprendizagem não supervisionada em que duas redes neurais competem uma contra a outra em um jogo de soma zero[9], tais redes podem produzir imagens de fotos realistas para aplicativos, como a visualização de design de interiores ou industrial, bem como sapatos, bolsas e itens de

[9] Refere-se a um cenário em que os ganhos de um jogador sempre se traduzem em perdas para outros jogadores. Um jogo de soma zero é aquele que a pontuação final do agente ao término do jogo para os dois jogadores é igual e de sinal oposto.

vestuário, ademais são amplamente utilizadas no desenvolvimento de jogos e geração de vídeos artificiais.

Basicamente, todos os tipos de Redes Neurais Profundas se baseiam em uma ou mais (redes neurais híbridas) das arquiteturas listadas, na Figura 18. Na sequência será descrita com mais detalhes a arquitetura MLP, condizente para solução de problemas de classificação.

Figura 18 – Arquiteturas de Aprendizagem Profunda e suas funções



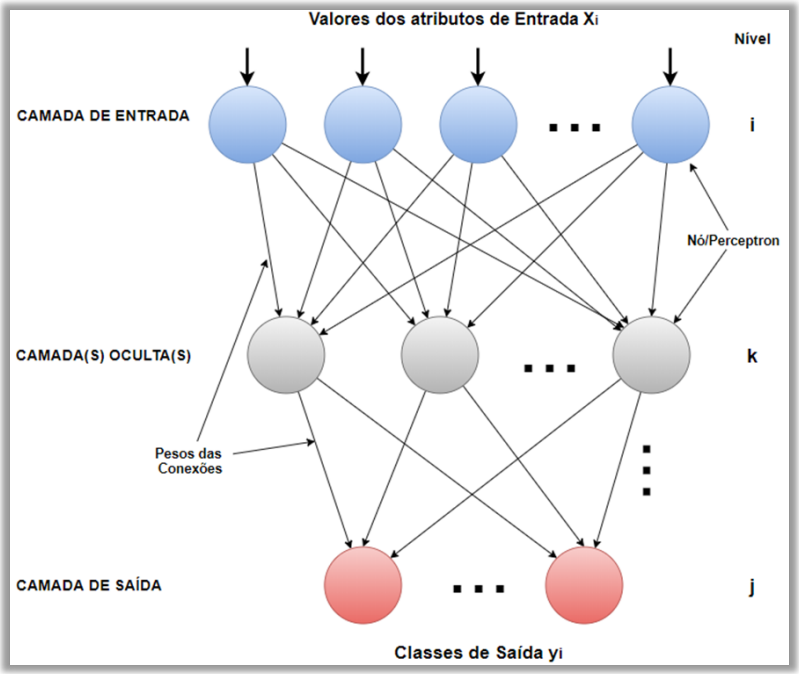
Fonte: Adaptado Aggarwal (2018)

DESCRIÇÃO DA ARQUITETURA MULTILAYER PERCEPTRONS

A arquitetura da rede neural do tipo Multilayer Perceptron é a mais simples, porém uma das mais adequadas para resolução de problemas de classificação em geral, essa tarefa está estreitamente relacionada com a identificação de padrões em dados. As redes do tipo MLP tem funcionalidade parecida a algoritmos de classificação de AM, porque pertencem à classe de Redes Neurais Profundas de Aprendizagem Supervisionada, entretanto em muitos contextos têm apresentado capacidade de processamento de dados superior aos algoritmos tradicionais (AGGARWAL, 2018; BISHOP, 1995; HAND, 1997; IGUAL; SEGUÍ, 2017; KUBAT, 2017; RIPLEY, 1996).

A arquitetura MLP consiste em uma rede de nós que são estruturas de códigos dispostos em camadas, muitos autores também denominam essa estrutura como perceptron ou neurônio artificial (AGGARWAL, 2018; BISHOP, 1995; HAND, 1997; IGUAL; SEGUÍ, 2017; KUBAT, 2017; RIPLEY, 1996). Uma rede típica de MLP consiste em três tipos de camadas de processamento: uma camada de entrada que recebe dados externos; um número arbitrário de camadas ocultas que fazem os cálculos intermediários e auxiliam a rede a encontrar os valores finais; e uma camada de saída que constrói os resultados da classificação. Destaca-se, que a diferença entre uma Rede Neural Artificial simples e uma Rede Neural Profunda é justamente a quantidade de camadas ocultas, na Rede Neural Profunda o número de camadas ocultas na maioria dos casos é superior a dois. Uma ilustração de como as camadas estão organizadas em uma rede MLP é apresentada na Figura 19.

Figura 19 – Arquitetura de Rede Multilayer Perceptron

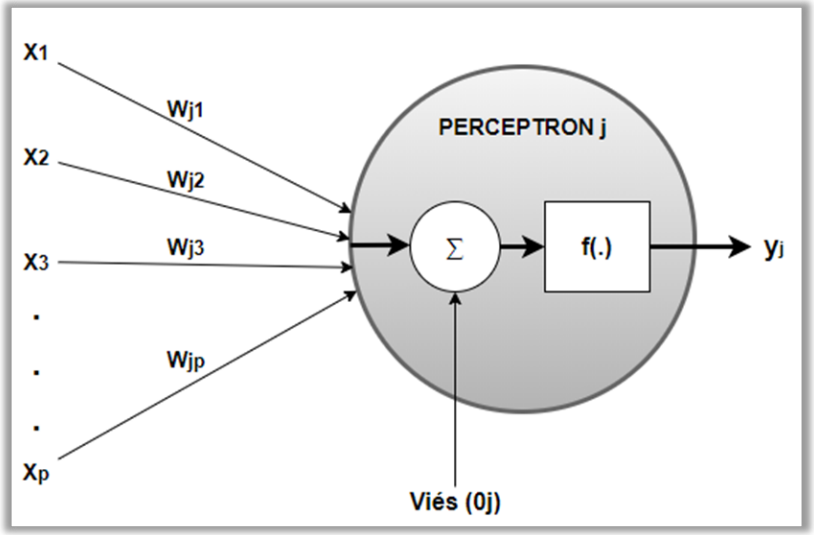


Fonte: Adaptado Aggarwal (2018)

O número nós da camada de entrada é determinado pela quantidade de atributos que há na base de dados fornecida. O número de nós na camada de saída é definido pela quantidade de classes do problema, um problema de classificação binária, por exemplo tem 2 nós na camada de saída. Para a determinação do número de camadas ocultas e a quantidade de nós em cada uma dessas camadas, existem uma variedade de metodologias para a escolha de tais aspectos, mas usualmente essa definição é realizada de forma empírica, com base em testes para cada contexto (AGGARWAL, 2018).

Ao contrário das outras camadas, nenhum cálculo está envolvido na camada de entrada. O princípio da rede é que quando os dados são apresentados na camada de entrada, os nós da rede realizam cálculos nas camadas sucessivas até que um valor de saída seja obtido para cada um dos nós, esse sinal de saída deve poder indicar a classe apropriada para as instâncias da base de dados. Um nó em uma Rede Neural Artificial pode ser representado como na Figura 20 (AGGARWAL, 2018).

Figura 20 – Um nó da Rede MLP – Perceptron



Fonte: Adaptado Aggarwal (2018)

As Redes Neurais Artificiais são algoritmos que efetuam transformações matemáticas sobre as bases de dados que recebem para processar. Em cada nó de cada camada, é realizada uma multiplicação entre o valor de entrada pelo peso do nó correlato, soma-se com o viés vinculado a esse nó e encaminha-se esse resultado adiante, aplicando a concepção de *feed forward*[10] (avançar). O viés é similar ao intercepto acrescido em uma equação linear, ele funciona como um parâmetro adicional que é aplicado para ajustar a saída junto da soma ponderada das entradas para o perceptron, é uma constante que auxilia o modelo a se adaptar melhor aos dados fornecidos (AGGARWAL, 2018).

Salienta-se que os procedimentos descritos são lineares, de maneira que, por mais complexa que seja a Rede Neural Artificial, ela unicamente manipula relações lineares entre os atributos de entrada e a classe de saída. Para modificar as Redes Neurais Profundas, de forma que essas ficassem adequadas para a modelagem de relações não-lineares, os valores de saída de cada camada passaram a ser submetidos ao processamento das denominadas *funções de ativação* (AGGARWAL, 2018). Além disso, são as funções de ativação que determinam se um perceptron deve ser ativado ou não, ou seja, se o dado que o nó está recebendo é importante, todo o processo descrito é definido nas Equações 2 e 3 (AGGARWAL, 2018):

Equação 2 – Soma ponderada realizada em cada nó

$$\vartheta j = \sum_{i=1}^{p} wji.xi + \theta j$$

Equação 3 – Combinação linear das entradas submetidas a função de ativação

$$yj = fj(\vartheta j)$$

[10] Nesse procedimento cada camada se conecta à próxima camada, dessa forma todas as conexões, têm a mesma direção, partindo da camada de entrada sentido a camada de saída.

Onde:

i ϑj é a combinação linear de entradas $x1; x2; \ldots; xp$;

ii θj é o viés;

iii wji é o peso da conexão entre a entrada xi e o neurônio j;

iv fj (.) é a função de ativação do j-ésimo neurônio; e

v yj é a saída.

A função sigmoide é uma das escolhas mais populares para função de ativação, essa é definida na Equação 4. Todavia, a função Relu (*Rectified Linear Unit*) tem se mostrado mais eficiente em tarefas de aprendizado profundo, Equação 5, por isso essa está presente em todas as bibliotecas para aplicação de AP. Ressalta-se, que existem muitas outras funções de ativação, como a função de tangente hiperbólica, a função logística e a função *softmax* (AGGARWAL, 2018). Uma questão interessante sobre o sobre o viés (θj) com relação a função de ativação, é que ele contribui para o deslocamento para a esquerda ou para direita da função de ativação, dependendo se possuiu um valor positivo ou negativo.

Equação 4 – Função de ativação Sigmoide

$$f(a) = \frac{1}{1 + e^{-a}}$$

Equação 5 – Função de ativação Relu

$$f(a) = \max(0, a)$$

O modelo gerado por uma Rede Neural Artificial corresponde a obtenção dos pesos que melhor se ajustem aos atributos de entrada fornecidos pela base de dados, esse modelo somente é alcançado mediante a atualização desses pesos no decorrer do processo de treinamento. Nesse sentido, o algoritmo de retropropagação (*backpropagation*) é um dos mais simples e eficientes métodos para o treinamento supervisionado de redes do tipo MLP (HINTON; OSINDERO; TEH, 2006), o qual é composto de duas fases: propagação e retropropagação.

Na propagação, a camada de entrada recebe um conjunto de atributos, ao qual são empregados os cálculos que foram descritos, e os resultados são propagados como entrada para a camada seguinte. Esse efeito é propagado pela rede, camada por camada, até que seja produzido um conjunto de saída como resposta atual da rede, nessa etapa os pesos são fixos.

Na fase de Retropropagação, a partir da camada de saída até a camada de entrada são ajustados todos os pesos de acordo com uma correlação do erro, isso é feito subtraindo a resposta atual da rede da resposta desejada, gerando um sinal de erro que novamente é retropropagado, através da rede no sentido contrário. Dessa forma, os pesos da rede são atualizados para fazer com que a resposta atual da rede se aproxime da resposta pretendida.

O algoritmo básico da retropropagação presente nas pesquisas de Bishop (1995) e Duda, Hart e Strok (2001) e otimizado por Hinton, Osindero e Teh (2006) funciona de forma simplificada, executando as etapas seguintes:

1. Inicializam-se todos os pesos de conexão w com pequenos valores aleatórios de um gerador de sequência;
2. Baseado nos atributos de saída (aprendizagem supervisionada) realizam-se os cálculos com os pesos e se calcula o erro;
3. Calcula as mudanças nos pesos e os atualiza, repete-se até a convergência[11] (que ocorre quando o erro E estiver abaixo de um valor predefinido ou até que o gradiente $\partial E(t)/\partial w$ seja menor que um valor predefinido).

3.1 Calcule a atualização usando a Equação 6.

Equação 6 – Cálculo da variação do peso

$$\Delta w(t) = -\eta \frac{\partial \in (t)}{\partial w}$$

[11] A Convergência é a capacidade da rede de aprender todos os padrões do conjunto de treinamento.

3.2 Atualize os pesos com a Equação 7.

Equação 7 – Cálculo da atualização dos pesos

$$w(t+1) = w(t) + \Delta w(t)$$

3.3 Calcule o erro com a Equação 8.

Equação 8 – Cálculo do erro

$$E(t+1)$$

Onde

i t é o número da iteração;

ii w é a conexão – peso; e

iii η é a taxa de aprendizado.

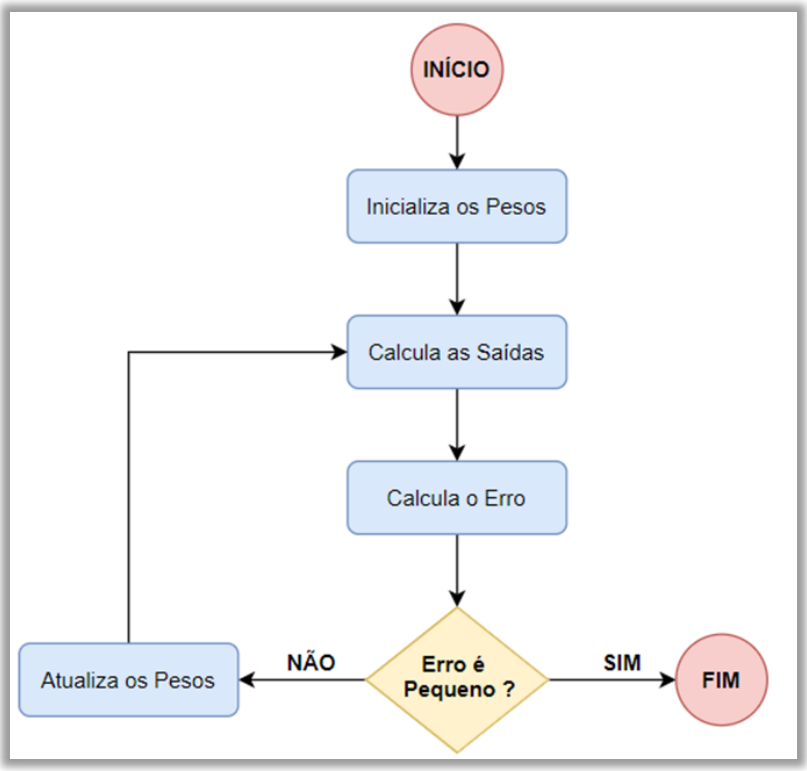
Geralmente o erro E pode ser também calculado como a função de erro quadrático médio (MSE) entre o valor real da saída da yj e a saída desejada dj, Equação 9.

Equação 9 – Cálculo do erro quadrático médio

$$E = \frac{1}{2} \sum_{j=i}^{nj} (dj - yj)^2$$

Para melhor ilustrar o processamento descrito, esse pode ser visualizado na Figura 21 que representa o fluxograma dos procedimentos que compõe funcionamento do algoritmo de retropropagação, para treinar uma rede do tipo MLP.

Figura 21 – Treinamento da uma Rede Neural Profunda



Fonte: Autora

Logo na primeira execução esse processo vai encontrar valores para as saídas, embora dificilmente sejam valores precisos, principalmente porque os pesos são atribuídos de forma aleatória, ao passo que esses devem condizer com os dados fornecidos na base, por isso devem ser atualizados até que se encontre o melhor conjunto de pesos para os dados fornecidos.

Sobre a atualização dos pesos, existem duas estratégias para esse processo de treinamento: 1) A estratégia de treinamento incremental, na qual se calcula o erro para cada registro e atualiza os pesos (SHIOTANI; FUKUDA; SHIBATA, 1995); e 2) A estratégia de treinamento em lote, na qual se define um número de registros para processar e atualizar os pesos (JANG; SUN; MIZUTANI, 2005).

Quanto ao algoritmo de retropropagação descrito, esse possui algumas deficiências que ainda não foram solucionadas, em especial quanto a taxa de aprendizagem da rede (KUBAT, 2017). A Taxa de aprendizagem é um parâmetro constante definido pelo usuário no intervalo de 0 a 1 que interfere na convergência do processo de treinamento. Se a taxa de aprendizagem for pequena o suficiente para minimizar o erro total, o processo de atualização dos pesos até encontrar um erro pequeno será muito lento. Por outro lado, uma taxa de aprendizado maior pode acelerar esse processo, contudo isso traz riscos na precisão dos resultados, levando a oscilações em torno do valor mínimo para o erro. Por isso a definição da taxa de aprendizado deve levar em conta esses aspectos e não ser nem muito pequena nem muito grande. O ideal seria utilizar uma taxa de aprendizagem grande o bastante para que o processamento fosse rápido, mas que não levasse a uma oscilação nos valores do erro (KUBAT, 2017).

A maneira mais simples de amenizar esse problema é usar o termo de *momentum,* com ele é possível evitar uma oscilação durante a busca do mínimo valor de erro. A atualização dos pesos com o algoritmo de retropropagação, incluindo o termo de *momentum* é definida pela Equação 10 (KUBAT, 2017).

Equação 10 – Cálculo para atualização dos pesos com o termo de *momentun*

$$\Delta W(t) = -\eta \frac{\partial E(t)}{\partial W} + \alpha \Delta W(t-1)$$

Onde 0 > α > 1.

A taxa de aprendizado adaptável também pode ser adotada para acelerar a convergência do algoritmo. Para estratégia de treinamento em lote, a taxa de aprendizado pode ser ajustada com a Equação 11 (KUBAT, 2017).

Equação 11 – Cálculo da taxa de aprendizado adaptável para treinamento em lote

$$\eta(t)\begin{cases}\beta\eta(t-1)\ se\ E(t) < E(t-1)\\ \theta\eta(t-1)\ se\ E(t) > kE(t-1)\\ \eta(t-1)\ outras\ formas\end{cases}$$

Onde:

i $\eta(t)$ é a taxa de aprendizado na t-iteração

ii β, θ e k são escolhidos como tal que $\beta > 1,\ 0 < \theta < 1\ e\ k > 1$

Enquanto para a estratégia de treinamento incremental, a taxa de aprendizado pode ser atualizada usando a Equação 12 (KUBAT, 2017).

Equação 12 – Cálculo da taxa de aprendizado adaptável para treinamento incremental

$$\eta(t) = \eta 0 + \lambda E(t-1)$$

Onde: η0 é a taxa de aprendizado predefinida e λ>0.

Na prática, alguns algoritmos de otimização são frequentemente utilizados para melhorar a convergência de uma Rede Neural Profunda, como o método de Descida do Gradiente, o método de Newton, o método Quase-Newton e os métodos de Gradientes Conjugados (GILL; MURRAY; WRIGHT, 1982). Esses algoritmos não foram pensados especificamente para otimização de Redes Neurais Artificiais, são antigas soluções em se tratando de Ciência da Computação, nessa área são consideradas ferramentas importantes, muito empregadas para otimizar funções complexas iterativamente dentro de diversos tipos de códigos computacionais, cujo propósito é dada alguma função arbitrária, encontrar um mínimo, dessa forma aplicados junto ao algoritmo de retropropagação tem o intuito de encontrar o menor erro.

Um dos melhores algoritmos para otimização empregados em conjunto com o de retropropagação é dos Gradientes Conjugados proposto por Polak (1971), pois apresenta baixo custo computacional e exibe bons

resultados e junto ao algoritmo da Descida do Gradiente, está disponível nas bibliotecas para AP. Considerando a otimização com o algoritmo dos Gradientes Conjugados os pesos de conexão podem ser expressos pelas Equações 13, 14 e 15 (KUBAT, 2017).

Equação 13 – Cálculo da atualização dos pesos com gradientes conjugados

$$\boldsymbol{w(t+1) = w(t) + \eta(t)d(t)}$$

Equação 14 – Cálculo do gradiente conjugado

$$\boldsymbol{d(t) = -\nabla E[w(t)] + \beta(t)d(t-1)}$$

Equação 15 – Cálculo do gradiente conjugado inicial

$$\boldsymbol{d(0) = -\nabla E[w(0)]}$$

Onde ∇E é o gradiente, $d(t)$ é gradiente conjugado, $\eta(t)$ é a taxa de aprendizagem,$\beta(t)$ é determinado na função apresentada na Equação 16 (KUBAT, 2017).

Equação 16 – Cálculo do Beta

$$\boldsymbol{\beta(t) = \frac{[\nabla E(w(t)) - \nabla E(w(t-1))]^T \nabla E[w(t)]}{\nabla E[w(t-1)]^T \nabla E[w(t-1)]}}$$

Ressalta-se que depois de aplicado o algoritmo de retropropagação, correspondente ao treinamento da Rede Neural Profunda, gera-se um modelo de classificação, formado pelos pesos assimilados pela rede em acordo com a base de dados fornecida, esse modelo pode então ser salvo e aplicado a dados novos no contexto desejado.

Para concluir essa seção algumas características sobre a arquitetura MLP são destacadas de acordo com pesquisas na área (AGGARWAL, 2018; IGUAL; SEGUÍ, 2017; KUBAT, 2017;):

1. A função de ativação de cada um dos nós é uma função não-linear – Uma característica relevante dessa função é que necessita ser suave e diferençável em todos os pontos.
2. As redes MLP são formadas de uma ou mais camadas ocultas – Essas camadas contêm nós que não são parte da camada de entrada ou saída da rede, tais nós ocultos possibilitam que a rede assimile tarefas complexas progressivamente, pelo processo de atualização dos pesos.
3. Uma rede MLP é uma rede *feed forward* – Redes desse tipo tem o fluxo de dados em uma única direção. Dessa forma, a saídas dos nós de uma camada qualquer se conectam unicamente às entradas dos nós da camada seguinte. Logo, o sinal de entrada se propaga, através da rede em um único sentido, não existe realimentação.
4. As redes MLP apresentam um forte grau de conectividade – As redes MLP ou podem ser totalmente conectadas, caso em que um neurônio de uma camada qualquer é conectado a todos os nós da camada seguinte, ou pode ser também parcialmente conectada, no caso em que alguma conexão está faltando. Na prática, a falta de algumas conexões dentro da rede é representada fazendo seu peso igual a zero.

Com o exposto foi possível entender e como é a arquitetura de uma Rede Neural Profunda utilizada no contexto de problemas de classificação, salienta-se que esse processo matemático é encapsulado pelas ferramentas utilizadas para sua implementação. Entretanto, sem o conhecimento de todos esses elementos seria difícil testar e alterar, quando necessário, os parâmetros fundamentais para processamento de dados por modelos de AP.

A AP já está consolidada em áreas como a medicina e biologia, principalmente no que diz respeito ao processamento de imagens, entretanto sua aplicação no contexto educacional para Mineração de Dados ainda não é tão consistente, embora estudos tenham relatados bons resultados com a utilização dessa técnica. Assim, para validar essa afirmação, no Capítulo seguinte são apresentados alguns estudos e constatações a respeito da aplicação dessa abordagem no âmbito educacional. Além disso, são expostos exemplo da aplicação da AM neste contexto.

CAPÍTULO 6 - APRENDIZAGEM DE MÁQUINA E APRENDIZAGEM PROFUNDA NA MINERAÇÃO DE DADOS EDUCACIONAIS

Neste Capítulo são abordados alguns estudos que aplicaram de AM e AP como técnicas de Mineração de Dados Educacionais, estes são descritos de forma geral não detalhando o processo utilizado, entretanto no Capítulo 7 serão expostos 2 exemplos completos da aplicação do processo de Mineração de Dados Educacionais com Aprendizagem Supervisionada e não Supervisionada.

APRENDIZAGEM DE MÁQUINA NA MDE

Desde a popularização do *e-learning*, e do surgimento da análise automatizada de dados educacionais muitos esforços têm sido realizados para aprimorar a experiência da aprendizagem, por esse motivo a MDE ganhou notoriedade, pois um de seus interesses é explorar a maneira como as pessoas aprendem. Avanços nessa área permitiram coletar e analisar dados sobre os alunos e seus ambientes e explorar o comportamento das pessoas enquanto aprendem e umas das principais técnicas empregadas para essa finalidade são os modelos de AM, que podem dar suporte a uma transformação da educação tradicional para uma educação otimizada e personalizada.

Como visto a AM é um campo de pesquisa que se ocupa do desenvolvimento de algoritmos que podem realizar previsões sobre grandes volumes de dados, esses algoritmos utilizados no contexto educacional podem auxiliar em inúmeras possibilidades. Nessa perspectiva, logo no início da disseminação e popularização da Inteligência Artificial e da Aprendizagem de Máquina, Baker (2000) as visualizou como um conjunto promissor de mecanismos de software e tecnologias a serem empregadas na melhoria do processo educacional, antes mesmo da consolidação da MDE como um campo de pesquisa. Nesse sentido, o autor

intensificou suas pesquisas sobre isso e agrupou as possíveis contribuições de modelos de AM ao processo educacional em três papeis principais:

1. *Modelo como ferramenta científica* – utilizado como um meio para compreender e prever algum aspecto de uma situação educacional. Por exemplo, um modelo é desenvolvido para entender porque muitos alunos têm desistido da disciplina de Cálculo, para prever o desempenho acadêmico, ou para realizar análise do comportamento de alunos.
2. *Modelo como componente* – empregado como algum aspecto do processo de ensino ou aprendizagem, sendo usado como um componente de um artefato educacional. Por exemplo, um modelo poderia ser integrado ao um ambiente de aprendizado baseado em computador, esse dessa forma se adaptar e interagir com os alunos, dando suporte em dúvidas frequentes e poderia ir refinando a forma como ajuda esses estudantes, conforme vai interagindo com eles.
3. *Modelo como base para o design* – utilizado como componente básico para projetar sistemas específicos para os processos educacionais, formando a base para o design de uma ferramenta de computador para a educação. Por exemplo, um modelo de diálogo orientado a tarefas formaria a base do design e implementação de ferramentas para comunicação mediada por computador entre alunos e professores, em um ambiente de aprendizado colaborativo suportado por computador.

Quando a Aprendizagem de Máquina é empregada como técnica de MDE, considera-se que ela está incluída na primeira possibilidade prevista por Baker (2000), com a finalidade de entender situações educacionais e com isso dar apoio no processo de tomada de decisão, nesse sentido há muitos estudos que foram e estão sendo desenvolvidos, entretanto alguns podem ser destacados, como: Chui *et al.* (2020), Zhang e Wu (2019) e Rodrigues *et al.* (2016).

CHUI *ET AL.* (2020)

No estudo desenvolvido Chui *et al.* (2020) o objetivo foi desenvolver um modelo para prever o desempenho acadêmico de alunos da graduação. Nesse contexto, os autores afirmaram que a partir de uma

perspectiva global, se um número considerável de estudantes deixar a universidade devido ao fracasso acadêmico, não apenas a reputação das próprias universidades sofre, mas as aspirações sociais dos alunos também são prejudicadas, por isso há uma necessidade de desenvolver meios precisos de prever graduandos que estejam em risco acadêmico.

Tendo esse objetivo em vista, o problema a ser tratado dizia respeito a identificação de padrões de alunos com tendência a ter um baixo desempenho. A base de dados utilizada nessa pesquisa foi a *Open University Learning Analytics* coletada entre 2013 e 2014 pela *Open University* do Reino Unido, essa base compreende 7 cursos de graduação e contém atributos de mais de 32 mil alunos sobre: atividades avaliativas, notas, perfil demográfico e uma apresentação pessoal desses estudantes.

Durante a formatação da base foram definidas duas classes para categorizar os estudantes pelos autores, Aprovação e Falha, o que configura um problema de classificação binária. Na sequência foi aplicado o algoritmo de Máquinas de Vetores de Suporte sobre a base de dados, o algoritmo foi modificado com várias otimizações para melhorar os resultados de previsões já relatados no estado da arte e diminuir o tempo de treinamento, que segundo os autores foi 60% menor que para o algoritmo tradicional. Quanto aos resultados das classificações, o modelo gerado por Chui *et al.* (2020) alcançou entre 92% e 94% de acurácia sobre os dados de teste. Concluindo, os autores relataram que com previsões precisas de alunos que tendem a ter um desempenho ruim, ações podem ser projetadas para amenizar as desistências desses alunos e motivá-los a conclusão.

ZHANG E WU (2019)

Zhang e Wu (2019), consideraram o contexto de MOOCs para o desenvolvimento de sua investigação, os autores contextualizaram o problema de sua pesquisa destacando que com o rápido desenvolvimento de cursos desse tipo, tornou-se uma questão importante na pesquisa educacional explorar as características de aprendizagem on-line e fornecer apoio à melhoria dos métodos de ensino e das atividades acadêmicas. Nesse cenário, os autores introduziram o objetivo da pesquisa que foi

prever o desempenho em MOOCs, mais especificamente a previsão de notas dos alunos. Para isso, os autores utilizaram dados dos alunos que cursaram o MOOC de Programação em Linguagem C, os atributos eram basicamente sobre: informações básicas do perfil dos alunos, pontuação nas atividades avaliativas, número de questões solucionadas, pontuação final, postagens nos fóruns de discussão.

Quanto a abordagem para a solução do problema, embora fosse esperado que os autores utilizassem Regressão, pois as notas são valores contínuos, os autores dividiram os resultados dos alunos em classes de 85 a 100 – A, de 70 a 85 – B, de 60 a 70 – C e >60 – D, portanto tornando-se um problema de classificação. Após a formatação e classificação da base de dados os autores realizaram a aplicação de 3 algoritmos de AM para geração de modelos de previsão de notas: *ID3, C4.5* e *CART* – todos baseados em Árvore de Decisão. As precisões alcançadas pelos modelos sobre a base de dados de teste foram: IDE3 – 81%, C4.4 – 75%, CART – 76%. Zhang e Wu (2019) afirmaram que os modelos baseados em árvores de decisão são consideravelmente simples de serem implementados, e têm precisão relativamente satisfatória, por isso devem ser empregados para apoiar ações que induzam a permanência de alunos em MOOCs, ou em outros cenários educacionais.

RODRIGUES *ET AL.* (2016)

Por fim, Rodrigues *et al.* (2016) desenvolveram sua pesquisa também no contexto de MOOCs, o estudo desenvolvido tinha como objetivo reconhecer perfis de engajamento de alunos nesse tipo de curso. Para esse fim, os autores utilizaram uma base de dados formada por atributos de 5 mil alunos de um curso MOOC com o tema Nova Gramática da Língua Portuguesa da plataforma *Openredu*. Os dados utilizados pelos autores diziam respeito a 15 atributos que retratavam a frequência de diferentes categorias de postagens, assiduidade e notas. Na busca pelas categorias de engajamento, os autores utilizaram o Agrupamento, por meio dos algoritmos *K-mens* e *Ward Clustering*, que possibilitaram a identificação de três perfis de engajamento no MOOC investigado:

Engajados – 16% do total de alunos; Esporádicos – 26% do total; e Desengajados – 58% do total.

- O perfil *Engajados* é descrito pelos autores como um grupo que possui uma ótima interação via fórum de discussão, mantém ritmo continuo na realização das atividades e pouca variabilidade entre as médias de notas, podendo ser considerado um grupo de alunos que realizam atividades do início ao fim do curso;
- O perfil *Esporádicos* foi detalhado como possuindo uma forte característica de mudanças, com picos repentinos durante o curso, uma das prováveis explicações é que alunos com esse perfil passam longos períodos sem entrar no ambiente e perdem alguns prazos de realização de atividades, embora possuam conhecimentos sobre o tema que faz com que tenham notas elevadas quando as realizam; e
- O perfil *Desengajados* refere-se à alunos praticamente inativos na realização do MOOC, mais de 90% só realizaram duas atividades, as notas médias e a quantidade de interações nos fóruns são inferiores aos demais grupos. Os autores salientam que esses são indícios que este grupo tem características apenas de visualização de material didático e optam por participar apenas em alguns momentos, permanecendo envolvidos, entretanto não desejam ganhar um certificado, e por este motivo não realizam as atividades.

Em conclusão Rodrigues *et al.* (2016) destacaram que a abordagem de Agrupamento pode ser usada para ajudar pesquisadores a identificar perfis comportamentais dos alunos em relação ao envolvimento em interações via fórum e durante atividades no decorrer de um MOOC, ainda relataram que entender os perfis comportamentais de alunos em cursos desse tipo pode servir como indicação para os designers atenderem a essa diversidade de padrões e orientar o desenho de estratégias adaptativas, que permitam aumentar o comprometimento dos alunos em conjunto com uma melhor experiência de aprendizado.

As pesquisas evidenciadas correspondem a exemplos de aplicação do processo de MDE, por meio da técnica de AM, todas com propósitos de entender acontecimentos no âmbito educacional e com isso ter mais

subsídios para desenvolver soluções ou tomar decisões. A partir da descrição dessas pesquisas foi possível perceber como os conceitos apresentados no Capítulo 5 podem ser desenvolvidos no contexto educacional.

TAN *ET AL*. (2018)

Outro exemplo da aplicação da técnica de AM para o mapeamento de perfis de alunos é o estudo desenvolvido por Tan *et al.* (2018), que utilizaram o algoritmo K-means no reconhecimento de perfis de aprendizagem dos alunos no MOOC "E-learning e Culturas Digitais" que foi criado pela Universidade de Edimburgo. Os autores utilizaram na sua pesquisa três tipos de atributos de 87 alunos que se matricularam neste MOOC: 1) Comportamento de aprendizagem on-line – número de postagens em fóruns, número de respostas, interações sociais, tempo de acesso a recursos; 2) Auto relatados – informações demográficas, proficiência em inglês, experiência de aprendizagem on-line, formação acadêmica e classificação das experiências de aprendizado com MOOCs; 3) Resultados de correções de questões dissertativas.

Para a aplicação do algoritmo K-means, Tan *et al.* (2018) formataram uma base de dados incluindo os atributos citados, a partir do processamento do algoritmo e análises sobre as informações geradas, os autores identificaram quatro perfis de aprendizagem entre os alunos que realizaram o MOOC: 1) Estudantes competentes e ativos; 2) Estudantes competentes e inativos; 3) Estudantes incompetentes e inativos; e 4) Espectadores. Os autores finalizaram o manuscrito afirmando que os resultados de seu estudo podem fornecer implicações úteis para projetos e implementações de MOOCs, principalmente no que tange a personalização.

APRENDIZAGEM PROFUNDA NA MDE

Desde 2006 a AP tem atraído muita atenção e sido aplicada com sucesso em muitas áreas, como reconhecimento de padrões, de fala e imagem (SCHMIDHUBER, 2015), monitoramento da integridade da máquina (ZHAO *et al.*, 2019), visão computacional, processamento de linguagem natural, detecção de intrusões e previsões médicas. No

entanto, de acordo com Yang, Zhang e Su (2019) as aplicações da AP no contexto educacional são relativamente escassas, pelo menos até o momento, em comparação a AM mais estabelecida como técnica de Mineração de Dados. A AP pode ser empregada na extração de recursos, reconhecimento e classificação de padrões, portanto é uma abordagem capaz de solucionar problemas no âmbito educacional (YANG; ZHANG; SU, 2018). A sua utilização nesse contexto pode levar a um crescente corpo de pesquisa com foco na melhoria da modelagem do comportamento e desempenho dos alunos, ampliando os horizontes de estudos na MDE, como os estudos de: Lin *et al.* (2019); Guo *et al.* (2019); Wen *et al.* (2020); e Waheed *et al.* (2020) que são brevemente sintetizados na sequência.

LIN *ET AL.* (2019)

No que tange a aplicação da AP no contexto educacional, alguns estudos começaram a despontar como o de Lin *et al.* (2019), nessa pesquisa os autores propuseram uma Rede Neural Convolucional para descoberta de conhecimento em vídeo aulas de MOOCs, com o propósito de detectar e classificar de forma automática pontos de conhecimento nesses vídeos, que de acordo com os autores, a obtenção desse tipo de informações poderia melhorar o desempenho da plataforma de aprendizado on-line.

Quanto a descoberta de conhecimento em vídeos, Lin *et al.* (2019) salientaram que tornou-se uma técnica popular usada para reconhecimento e identificação de informações produtivas em áreas como: bioinformática, assistência médica e investigação criminal, então sua aplicação no cenário educacional poderia trazer inúmeros benefícios.

No que diz respeito à Rede Neural Convolucional, sua implementação foi executada na ferramenta Tensorflow e foi treinada e testada em um conjunto de dados de 16 mil imagens e 72 horas de áudio, pertencentes a dois MOOCs diferentes. Como principal contribuição os autores relataram o desenvolvimento de uma nova abordagem de Rede Neural Convolucional para descoberta de conhecimento em vídeo, que

mostrou-se potencialmente eficaz. A partir dos resultados da pesquisa Lin *et al.* (2019) concluíram que o reconhecimento preciso da diferença entre pontos de conhecimento e pontos de não conhecimento em vídeos aulas é uma ferramenta útil para apoiar os professores e alunos.

GUO *ET AL.* (2019)

Nesse cenário, também destaca-se a investigação realizada Guo *et al.* (2019), na qual os autores implementaram uma Rede Neural Profunda Híbrida para a identificação de postagens "urgentes" que requerem atenção imediata de instrutores em fóruns de discussão em MOOCs. Para contextualizar essa problemática os autores explicam que o fórum de discussão de MOOCs é um ambiente onde alunos e professores se comunicam e que geralmente tem sobrecarga de informação, dessa forma, em postagens que os alunos expressam confusão e que necessitam de mais atenção dos professores, muitas vezes ficam sem uma rápida resposta, devido a quantidade de ruído no fórum, portanto como prestar atenção a essas mensagens urgentes a tempo, se tornou um problema a ser resolvido.

Quanto a estrutura da Rede Neural Profunda proposta no estudo, é uma Rede Neural Convolucional combinada com uma Rede Neural Recorrente, implementada para mineração de texto e funciona em 3 etapas: 1) Assimilar simultaneamente as informações semânticas e estruturais das sentenças de texto das postagens; 2) Utilizar as Redes Convolucionais em nível de caractere para capturar informações – isso foi necessário devido a muito ruído, como erros de ortografia e *emoticons* no texto das postagens; e 3) Associar as informações semânticas e estruturais com as informações de caracteres e assim chegar a representação final da frase. Guo *et al.* (2019) chegaram a resultados que superam a precisão de soluções presentes no estado da arte em até 2,4%, e concluíram que sua pesquisa pode auxiliar professores e tutores a priorizar suas respostas e gerenciar melhor várias postagens, de modo que esses profissionais da educação possam responder às perguntas dos alunos em tempo hábil e ajudar a reduzir as taxas de evasão em MOOCs.

WEN *ET AL.* (2020)

Wen *et al.* (2020) também utilizaram um modelo de AP para pesquisa no âmbito educacional, com o objetivo de identificar antecipadamente a desistência em MOOCs. Pra descrever a problemática abordada o autores salientam que as taxas de desistência nesses tipos de cursos são bastante altas e que o processo de previsão nesse cenário tem o intuito de identificar se um aluno exibirá comportamentos de aprendizagem durante vários dias consecutivos no futuro, portanto, as informações relacionadas aos comportamentos de aprendizagem de um aluno em vários dias consecutivos, devem ser consideradas.

Nesse sentido, depois que Wen *et al.* (2020) realizaram uma análise dos padrões de comportamento de aprendizagem dos alunos de um MOOC, relataram que esses estudantes geralmente exibem comportamentos de aprendizagem semelhantes em vários dias consecutivos, o *status* de aprendizagem de um aluno para o dia subsequente provavelmente será semelhante ao do dia anterior, o que os autores chamaram de *local correlation of learning behaviors* (tradução livre - correlação local de comportamentos de aprendizagem).

Embasados nessa premissa Wen *et al.* (2020) propuseram uma base de dados formada por atributos relacionados à correlação local de comportamentos de aprendizagem, sobre a qual aplicaram uma Rede Neural Convolucional, gerando um novo modelo para prever o abandono de alunos em MOOCs. Ressalta-se que, embora esse tipo de rede seja mais adequado ao processamento de imagens, pode também ser utilizado para solucionar outros tipos de problemas.

Por fim, o modelo proposto obteve uma precisão que variou de 86% a 89%, e os autores destacaram que as principais contribuições da pesquisa foram: 1) Definição do conceito de *status de aprendizagem* para encontrar a correlação local de comportamentos de aprendizagem; 2) Construção de uma base de dados formada a partir dos atributos da correlação local de comportamentos de aprendizagem; e 3) Implementação de um modelo construído a partir de uma Rede Neural Convolucional para previsão do abandono de alunos em MOOCs.

WAHEED *ET AL.* (2020)

Por fim, pode-se citar o estudo elaborado por Waheed *et al.* (2020) que desenvolveu um modelo de Rede Neural Profunda com arquitetura MLP, com o objetivo de prever o desempenho de alunos em MOOCs. Para isso os autores utilizaram relatórios de 7 MOOCs, com um total 32 mil alunos, e os atributos utilizados foram: perfil demográfico, fluxo de cliques e desempenho nas avaliações. O estudo foi conduzido com base na mineração de dois conjuntos de dados: 1) Notas das atividades avaliativas na plataforma e perfil demográfico; e 2) Atributos trimestrais do fluxo de cliques de cada aluno.

Baseado nessas extrações de atributos Waheed *et al.* (2020) construíram uma base de dados e aplicaram uma rede do tipo MLP, que resultou em um modelo para prever o risco de reprovação dos alunos. Os autores relataram que em contraste com os métodos estatísticos, as Redes Neurais Profundas facilitam a generalização, o que possibilita inferir padrões escondidos nos dos dados, dando suporte a fazer suposições sobre eles.

A precisão alcançada pelo modelo ficou entre 84% e 94% nos experimentos realizados, e Waheed *et al.* (2020) concluíram que esses resultados demonstram a efetividade do modelo implementado para a previsão precoce do desempenho de alunos em MOOCs. Os autores ainda ressaltaram que estudos como esse, orientados a dados, são necessários para auxiliar Instituições de Ensino na formulação de uma estrutura de análise de aprendizagem, contribuindo para o processo de tomada de decisão.

CONSIDERAÇÕES SOBRE A APLICAÇÃO DA APRENDIZAGEM DE MÁQUINA E APRENDIZAGEM PROFUNDA NA MDE

Tais investigações denotam o potencial da AM e da AP no contexto educacional. Cabe destacar que apesar do bom desempenho frequentemente relatado sobre modelos de AP aplicados a MDE, o grande número de parâmetros que devem ser configurados e o entendimento necessário sobre eles, (especificados no Capítulo 5), os tornam difíceis

de interpretar e implementar, por isso essa técnica não tem sido tão explorada neste contexto (YANG; ZHANG; SU, 2018).

Convém evidenciar que existem outras técnicas para MDE principalmente vinculadas a Estatística Descritiva e Inferencial, mas optou-se por abordar AM e AP pelo fato de que essas atualmente correspondem a um campo em grande expansão e tem sido amplamente utilizadas nessa área, em especial por dois motivos: 1) Facilidade de aplicação – devido as poderosas ferramentas disponíveis que encapsulam os principais conceitos matemáticos dos algoritmos aplicados; e 2) Qualidade dos resultados encontrados – devido aos avanços tecnológicos da Inteligência artificial as informações extraídas dos conjuntos de dados são na grande maioria das vezes muito relevantes.

Porém, mesmo com uma grande quantidade de estudos publicados que abordem a aplicação destas técnicas no âmbito educacional, é relativamente pequeno o número de publicações que descrevam o processo de aplicação da Mineração de Dados Educacionais em detalhes. Estes manuscritos geralmente abordam aspectos superficiais da metodologia empregada no desenvolvimento do processo de mineração. Por este motivo, há a apresentação de dois experimentos, desenvolvidos pela autora, que descrevem de forma detalhada o processo de aplicação da MDE. O processo utilizado como referência é o descrito no Capítulo 1 e ilustrado na Figura 2. Estes experimentos são descritos no Capítulo seguinte.

CAPÍTULO 7 - EXEMPLOS DE APLICAÇÃO DO PROCESSO DE MINERAÇÃO DE DADOS EDUCACIONAIS: ANÁLISE DO PERFIL DE ALUNOS E PREVISÃO DO DESEMPENHO

Neste Capítulo serão apresentados dois exemplos da aplicação do processo de MDE, exposto no Capítulo 1 (Figura 2), elaborados pela autora; esses dois exemplos tem objetivos diferentes e se utilizam das técnicas expostas no Capítulo 5. O primeiro experimento apresentado foi realizado com aplicação de um algoritmo de Aprendizagem de Máquina do tipo não Supervisionado, com o propósito mapear perfis de alunos de um *Massive Open Online Course* (MOOC) da área de química de uma plataforma brasileira, enfocando os objetivos desses alunos quando realizam um curso nesse formato. O segundo experimento foi realizado com a aplicação de algoritmos de Aprendizagem de Máquina e Aprendizagem Profunda do tipo Supervisionado, com o objetivo de realizar a previsão do desempenho de alunos em um conjunto de dados públicos e comparar as técnicas de AM e AP, ademais indicar quais os principais atributos preditores para o desempenho dos alunos. Esses dois experimentos compõe uma descrição que apresenta de maneira detalhada como realizar o processo de Mineração de Dados Educacionais e pode auxiliar pesquisadores iniciantes na área.

MAPEAMENTO DOS PERFIS COMPORTAMENTAIS DE ALUNOS EM UM MOOC BRASILEIRO

Os *Massive Open Online Courses* (MOOCs) têm recebido grande atenção desde seu surgimento em meados de 2008, sobretudo pela sua flexibilidade, acesso a materiais e aulas preparados por especialistas, e facilidade de ingresso. Segundo Shah (2018) o número e a diversidade de MOOCs continuam crescendo desde seu surgimento, o autor afirma que em 2018, mais de 900 universidades em todo o mundo lançaram 11.400 MOOCs. Todavia, apesar das vantagens dos MOOCs, muitos desafios diferentes

ainda são impostos, o que gera questões sem respostas. A esse respeito um dos problemas mais citados entre os pesquisadores são as altas taxas de desistência, que estão em torno de 90% (ALRAIMI; ZO; CIGANEK, 2015).

Na tentativa de sanar os problemas vinculados a esse modelo educacional muitas pesquisas têm sido realizadas em busca de identificar os principais perfis comportamentais de alunos que realizam MOOCs, como exemplo pode-se citar os estudos desenvolvidos por: Rodrigues *et al.* (2016); Tan *et al.* (2018); Mareca e Bordel (2019); e Shi, Peng e Wang (2017). Tais estudos se intensificam, visto que, em uma sociedade cada vez mais informatizada, oferecer serviços customizados, personalizados e adaptáveis, embora seja uma tarefa desafiadora, vale a pena, pois melhora os serviços e funcionalidades de diversos sistemas, e o conhecimento do perfil do usuário é uma fase vital neste processo. Assim, a modelagem de perfis é um campo importante que visa dar uma representação abstrata de alguns aspectos relacionados às características do usuário.

No campo educacional a modelagem do perfil do aluno pode ser um suporte para a tomada de decisões em diferentes níveis, como por exemplo: na gestão de problemas relacionados ao baixo desempenho e evasão; oferecer aos alunos a orientação mais adequada e recomendação; e definir os recursos de aprendizagem mais adaptáveis dependendo do perfil do estudante. Nesse sentido, em cursos *e-learning*, particularmente como os MOOCs, essa busca pode auxiliar na personalização dos ambientes de aprendizagem, as plataformas de oferta desses cursos, o que resultaria em mais alunos concluintes, com boas experiências de aprendizado e maior conhecimento agregado. Além disso, pode possibilitar uma maior compreensão sobre a maneira que esses estudantes se comportam na condução de seu aprendizado e apoiar na elaboração de conteúdos de forma mais adequada para diferentes públicos alvo.

O perfil do aluno representa uma estrutura que contém informações diretas e indiretas sobre seu comportamento, abordando: interesses, preferências, dados pessoais e habilidades (FERREIRA-SATLER *et al.* 2012). Muitas vezes, o conceito de perfil está relacionado ao contexto estudado, por exemplo no âmbito dos MOOCs, o foco está na interação

do aluno com a plataforma – tempo de conexão, duração da conexão, navegação na plataforma, postagens em fóruns. Diante desse contexto, este experimento teve como objetivo realizar o mapeamento dos perfis comportamentais dos alunos de um MOOC da área de química, oferecido por uma plataforma brasileira. A definição dos perfis dos alunos é importante para que se possa ter um entendimento mais amplo de como esses estudantes se comportam no decorrer da execução de um MOOC. Esse conhecimento pode auxiliar na personalização do ambiente educacional, apoiar no desenvolvimento de MOOCs mais atrativos, auxiliando no engajamento dos alunos e melhorando os índices de conclusão. Além disso, esse entendimento pode dar suporte na implementação de mecanismos e diretrizes que possam inibir práticas inadequadas na realização de um MOOC, como plágio e/ou cópias das respostas de atividades avaliativas.

Nesse contexto, este estudo tem como propósito mapear os perfis de alunos de um MOOC da área de química de uma plataforma brasileira, enfocando os objetivos desses alunos quando realizam um curso nesse formato. Para realização do mapeamento, foi utilizado o algoritmo de Agrupamento *K-means* e foram identificados 4 perfis comportamentais predominantes: *Engajados, Estratégicos, Inativos e Oportunistas*. Com os resultados apresentados por esse estudo podem ser desenvolvidas estratégias para melhorias nos MOOCs oferecidos pela plataforma, a partir do conhecimento dos perfis dos estudantes, com ações para inibição de maus comportamentos e incentivos a comportamentos desejáveis.

DESENVOLVIMENTO DO PROCESSO DE MDE PARA O MAPEAMENTO DE PERFIS DE ALUNOS

Este experimento teve como principal objetivo realizar o mapeamento dos Perfis de alunos que cursaram um MOOC da área de química de uma plataforma brasileira. Para alcançar esse objetivo foram empregados alguns procedimentos que configuram a metodologia adotada nesta pesquisa, que em termos gerais resume-se na realização do processo de MDE apresentado no Capítulo 1. Como descrito esse processo é composto basicamente por 4 fases: 1) Definição da função da MDE; 2)

Formatação dos dados que serão utilizados; 3) Definição das Técnicas de MDE; 4) Delineamento de como essas técnicas serão aplicadas. Por fim, fica subentendido, como exposto na Figura 2 do Capítulo 1, que deva ser implementada uma quinta fase que diz respeito a: 5) Análise e/ou interpretação dos resultados; que deve ser realizada conforme os objetivos do processo de mineração de dados desenvolvido. Para algoritmos de Aprendizagem Supervisionada essa análise é baseada na avaliação dos resultados com métodos e métricas quantitativas (expostas no Capítulo 5), todavia para algoritmos de Aprendizagem não Supervisionada há uma maior dependência da interpretação do pesquisador.

Essas etapas compõem basicamente os procedimentos realizados para o desenvolvimento do processo de MDE com o intuito de mapear os perfis comportamentais dos alunos, estas etapas são descritas na sequência.

DEFINIÇÃO DA FUNÇÃO DA MDE

Nessa etapa é realizada a determinação do objetivo do processo MDE, para qual finalidade ela está sendo aplicada. Neste experimento esse processo foi aplicado para realizar o Mapeamento de Perfil de alunos, desenvolvido por meio da identificação de afinidade em grupos, ou descrição de grupos.

FORMATAÇÃO DOS DADOS QUE FORAM UTILIZADOS

Primeiramente cabe realizar uma descrição do curso em que os alunos pesquisados estavam matriculados, que era do tipo MOOC, da área de química de uma plataforma de MOOCs brasileira com carga horária de 15 horas. O público do curso era diversificado, principalmente porque este não possui pré-requisitos, nem restrições para cadastramento e é gratuito. O curso tem 4 módulos de conteúdos, e um módulo inicial, no qual tem explicações sobre como o MOOC funciona, alguns materiais introdutórios e também uma pesquisa de perfil do estudante. O tipo de recurso predominante são vídeo aulas

gravadas pelo professor, mas também possui materiais em texto e todos os módulos possuem uma atividade avaliativa, no formato: questionário de múltipla escolha. Após o módulo final o aluno ainda conta com um fórum para trocar informações, realizar interações com outros participantes e opinar sobre o curso, há também um material complementar disponibilizado pelo professor. Ao concluir o aluno pode gerar o certificado de conclusão.

Este MOOC, até a extração dos dados para este experimento, contava com mais de 5.000 alunos matriculados e para este estudo foram utilizados dados de 3.540 estudantes que responderam ao questionário de perfil do aluno até a data que os dados foram extraídos (final do primeiro semestre de 2020). Com referência ao abandono no MOOC, destaca-se que não foi possível indicar de forma exata quais dos alunos já haviam desistido do curso, ou quais ainda estavam cursando, pois o MOOC não tem prazo limite de realização, bem como na base de dados formatada existem alunos que demoraram longo tempo para terminarem e emitir o certificado, enquanto outros foram mais rápidos, então não é possível afirmar certamente, mesmo em posse da média de tempo gasto por esses estudantes, quem já evadiu. Isto posto, cabe salientar que para esse estudo foram considerados concluintes os alunos que emitiram o certificado, e não concluintes aqueles que não emitiram, até a data de extração das bases de dados.

Para o desenvolvimento deste estudo foram extraídos dados de 4 relatórios disponíveis na plataforma: relatório de *Perfil do Estudante*; relatório de *Conclusão de Atividades*; relatório de *Notas;* e relatório de *Logs*; de cada um desses relatórios foram retirados dados que foram considerados relevantes para o estudo, os quais foram incorporados em uma única base de dados. Os dados constantes na base foram extraídos de interações entre junho de 2019 a agosto de 2020. Sobre esses dados foram realizados transformações e pré-processamentos para que fossem aprimoradas as informações a respeito dos alunos. Os atributos constantes na base de dados gerada para a análise e composição dos perfis podem ser observados no Quadro 8.

Quadro 8 – Atributos da Base de Dados

ID	ATRIBUTO	DEFINIÇÃO	RELATÓRIO DE ORIGEM
1	**Gênero**	Informado pelo usuário.	Extraído do Relatório de Perfil do Estudante.
2	**Idade**	Informada pelo usuário.	Extraído do Relatório de Perfil do Estudante.
3	**Escolaridade**	Informada pelo usuário – vai desde o fundamental incompleto até a pós-graduação.	Extraído do Relatório de Perfil do Estudante.
4	**Motivação**	Informada pelo usuário – que pode ser: curiosidade geral, importante para meu trabalho, ou importante para meus estudos.	Extraído do Relatório de Perfil do Estudante.
5	**Intenção**	Informada pelo usuário – diz respeito a vontade de terminar o MOOC, pode ser: Sim, Não sei, ou Não.	Extraído do Relatório de Perfil do Estudante.
6	**Nota Final**	É a média das notas dos 4 questionários realizados .	Extraído do Relatório de Notas.
7	**Tempo**	É a diferença de tempo, dada em dias, entre o primeiro acesso a alguma atividade do MOOC até o último acesso – tempo que o aluno interagiu com o curso.	Calculado a partir do Relatório de Conclusão de Atividades.
8	**Porcentagem Concluída**	É a porcentagem das atividades concluídas – exemplo se no curso há 20 recursos e/ou atividades e o aluno concluiu 10 ele tem 0,5 de porcentagem concluída.	Calculado a partir do Relatório de Conclusão de Atividades.
9	**Soma das Atividades Realizadas – Materiais em Texto**	Nesse atributo foram somados quantas vezes os alunos acessaram os materiais em texto. É o somatório apenas as atividades referentes aos conteúdos do MOOC.	Calculado a partir do Relatório de Logs.
10	**Soma das Atividades Realizadas – Vídeo Aulas**	Nesse atributo foram somados quantas vezes os alunos acessaram as vídeo aulas.	Calculado a partir do Relatório de Logs.

ID	ATRIBUTO	DEFINIÇÃO	RELATÓRIO DE ORIGEM
11	**Soma das Atividades Realizadas – Tentativas em Questionários Avaliativos**	Nesse atributo foram somados quantas vezes os alunos acessaram as tentativas de questionário.	Calculado a partir do Relatório de Logs.
13	**Soma das Atividades Realizadas**	Nesse atributo foram somados todos os acessos a recursos de conteúdos e atividades avaliativas do MOOC. Também definido como a quantidade de interações com a plataforma.	Calculado a partir do Relatório de Logs.
13	**Certificado**	Variável binária que tem valor 1 pra quem conclui e gerou o certificado e 0 para quem não requisitou o certificado.	Extraído do Relatório de Conclusão de Atividades.

Fonte: Autora

DEFINIÇÃO DAS TÉCNICAS DE MDE

Depois da formatação da base de dados foi buscado o na literatura qual seria a melhor técnica para mapear os perfis, foi então levado em consideração dois estudos similares a este experimento desenvolvido pela autora; Rodrigues *et al.* (2016) e Tan *et al.* (2018) – que foram detalhados no Capítulo 6 – estes utilizaram o algoritmo de Aprendizagem de Máquina não supervisionado *K-means* para a análise de perfis de alunos em cursos do tipo MOOC. Este algoritmo é muito utilizado para encontrar similaridades em grupos, por meio de semelhanças em atributos. Dessa forma, nesse experimento foram selecionados apenas os atributos considerados mais relevantes para a aplicação do algoritmo *K-means*, para isso foi considerado também as características do algoritmo que emprega distância entre dois pontos (distância Euclidiana) para composição dos agrupamentos por similaridade, por isso não é possível a utilização de atributos categóricos, apenas numéricos. Por isso, a base de dados na qual foi aplicada o algoritmo é composta pelos seguintes

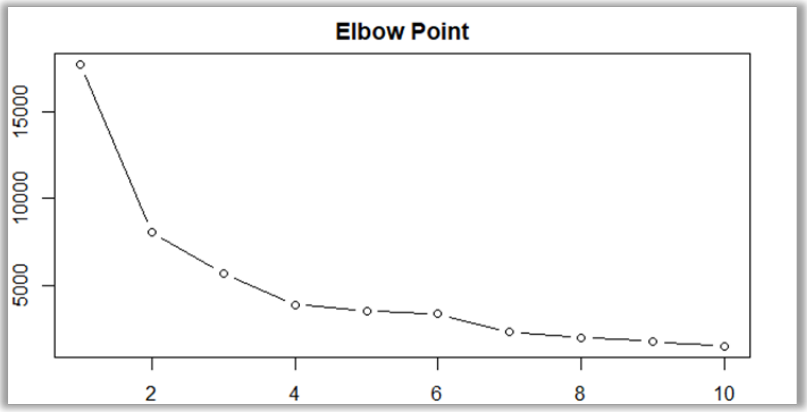
atributos: *Tempo, Porcentagem Concluída, Tentativas em Questionários Avaliativos, Soma Total da Atividades Realizadas e Certificado.*

DELINEAMENTO DE COMO AS TÉCNICAS FORAM APLICADAS

No que se refere ao Delineamento de como essa técnica foi aplicada, etapa em que são selecionadas as ferramentas que vão dar suporte ao desenvolvimento de sistemas capazes de processar os dados e gerar os resultados esperados, o algoritmo *K-means* foi aplicado, por meio da linguagem de programação e ciência de dados R.

Anteriormente a submissão da base ao algoritmo, foi efetuado um teste chamado *elbow method* (tradução livre – método do cotovelo) sobre a base de dados, a fim de descobrir qual seria a quantidade de agrupamentos que melhor se ajustaria a base. O *elbow method* é uma heurística empregada na determinação do número de grupos em um conjunto de dados, o método consiste em plotar a variação explicada em função do número de grupos e escolher o cotovelo da curva como o número de agrupamentos a serem usados. No caso da base de dados utilizada os valores mais indicados seriam entre 2 e 4, como pode ser observado na Figura 22, todavia o cotovelo está em 2, que seria o melhor valor.

Figura 22 – Método *Elbow Point*



Fonte: Autora

Entretanto, optou-se por fazer alguns teste iniciais com 4 agrupamentos, levando-se em consideração o conhecimento que se possui sobre a base de dados, porém percebeu-se que com 4 agrupamentos um grupo estava ficando muito pequeno (com apenas 47 alunos), não caracterizando um perfil singular, quanto aos objetivos no curso, diferente dos outros grupos gerados. Sendo assim foi definido a quantidade de agrupamentos como 3, e com a aplicação do *K-means* foi possível agrupar alunos que apresentaram atributos semelhantes. Foram utilizadas as configurações *default* do algoritmo em que os dados são agrupados pelo método *k-médias*, que visa particionar os pontos em k grupos de forma que a soma dos quadrados dos pontos aos centros atribuídos aos agrupamentos seja minimizada; na linguagem R o algoritmo de Hartigan e Wong (1979) é usado por padrão. Cabe destacar que para a aplicação do *K-means*, por ele utilizar como métrica a distância entre dois pontos, é importante realizar a padronização das escalas dos atributos, que nesse estudo foi feito por meio da normalização.

Dessa forma, o *K-means* foi configurado (com k=3) e aplicado sobre a base de dados, em seguida foi realizada uma investigação a respeito das características dos atributos de cada grupo identificado, sistematizando os dados de perfil, mas com ênfase nas ações realizadas na plataforma, verificando: os números de interações médias dos grupos, tempo de realização do curso, porcentagem das atividades concluídas e desempenho. Visto que, este estudo tem como intuito mapear os perfis dos alunos, com base no seu objetivo quando da realização do MOOC.

ANÁLISE E/OU INTERPRETAÇÃO DOS RESULTADOS

Nessa seção são retratadas e interpretadas as análises realizadas sobre a base de dados que teve como objetivo principal mapear os perfis dos alunos participantes deste experimento, primeiramente são apresentadas as descrições dos dados da amostra e dos agrupamentos, na sequência são definidos os perfis dos estudantes.

ANÁLISE DESCRITIVA DOS DADOS

Os estudantes participantes do curso, como já mencionado, são um público diversificado, há pessoas de diversas idades de 19 a 50 anos, sendo que há uma predominância de alunos que se caracterizam com até 19 anos, 38% dos alunos analisados; com escolarização indo desde o Ensino Fundamental Incompleto até a Pós-Graduação, nesse sentido, no que se refere à formação há mais alunos que relatam possuir Ensino Superior Incompleto, que são 36% dos integrantes do MOOC. Em relação a motivação 46% diz estar interessado no curso devido ao assunto ser "importante para seus estudos", e a grande maioria 96% diz ter intenção de terminar o MOOC; embora, apenas 24% tenha realmente finalizado. Em relação ao desempenho dos alunos, 35% tiveram notas que variaram entre 30 e 40 pontos, sendo que a avaliação desse MOOC é composta por 4 questionários com valor 10 cada um, e a nota final é somatório das notas individuais. Mais detalhes sobre a amostra pesquisada podem ser observados na Tabela 4.

Tabela 4 – Descrição da Amostra

ATRIBUTO		QUANTIDADE ABSOLUTA DE ALUNOS
Gênero	Masculino	**2.074**
	Feminino	**1.457**
	Outros	**9**
Idade	Até 19 anos	**1.375**
	20 a 24 anos	**903**
	25 a 29 anos	**417**
	30 a 34 anos	**242**
	35 a 39 anos	**209**
	40 a 49 anos	**244**
	Acima de 50 anos	**150**

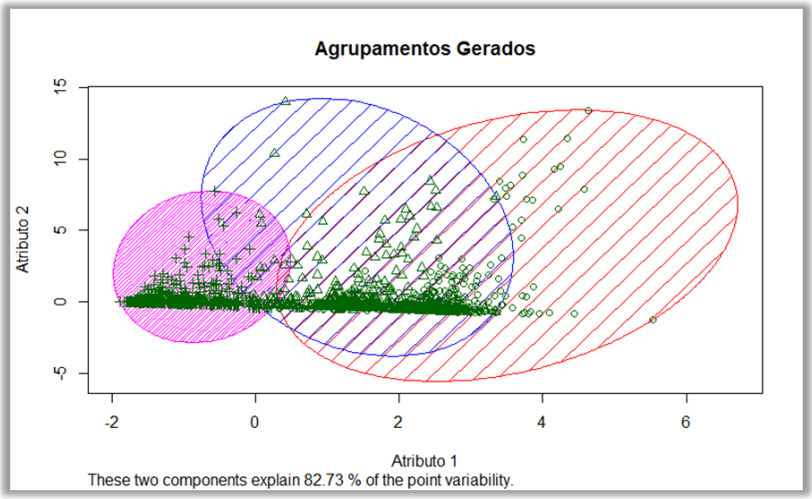
Escolaridade	Ensino fundamental incompleto	**385**
	Ensino fundamental completo	**69**
	Ensino médio incompleto	**438**
	Ensino médio completo	**373**
	Ensino superior incompleto	**1.292**
	Ensino superior completo	**435**
	Pós-Graduação	**548**
Motivação	Curiosidade geral	**1.474**
	Importante para meu trabalho	**421**
	Importante para meus estudos	**1.645**
Intenção	Não	**6**
	Não Sei	**135**
	Sim	**3.399**
Nota Final	0	**1.602**
	1-10	**387**
	11-20	**116**
	21-30	**181**
	31-40	**1.254**
Certificado	Concluintes	**876**
	Não Concluintes	**2.664**
TOTAL DE ALUNOS		**3.540**

Fonte: Autora

ANÁLISE DESCRITIVA DOS AGRUPAMENTOS

Com a aplicação do algoritmo *K-means* foram definidos os alunos pertencentes aos 3 grupos indicados, há relativa sobreposição entre os atributos dos alunos, mas com o gráfico dos agrupamentos (Figura 23) pode ser visualizado de forma mais ilustrativa a distribuição dos grupos, os atributos utilizados para formatação do gráfico são capazes de explicar 82,73% dos pontos de variação, e são definidos pela biblioteca *cluster* da linguagem R de forma automática, justamente por terem maior influência sobre os agrupamentos gerados.

Figura 23 – Gráfico dos Agrupamentos Gerados



Fonte: Autora

Agrupamento 1 (vermelho): Composto por um total de 862 alunos, dos quais 51% são homens, há alunos indo de 19 a 50 anos, mas há uma predominância de alunos mais jovens, com 36% de estudantes entre 20 e 24 anos. Quanto à escolaridade 49% dos alunos possuem superior incompleto. Quanto a motivação, 44% alega "curiosidade geral", e outros 44% diz "ser importante para seus estudos" e 12% afirma ser importante para seu trabalho, quanto a intenção de terminar 95% disseram ter intenções concluir o curso, nesse grupo todos os alunos terminaram o MOOC (100%). A média da quantidade de interações desse grupo com os recursos do curso foi de 108 interações, uma média de 3 visualizações por vídeo aula, 11 tentativas (visualizações) por questionário avaliativo e 3 acessos a cada material em texto. Os estudantes desse grupo concluíram em média 96% das atividades e passaram 13 dias realizando o curso, ou seja, 13 dias entre o primeiro acesso e a geração do certificado. Quanto ao desempenho no curso os estudantes ficaram com uma média de 36,5 pontos na nota final (indicando por volta de 9,12 pontos em cada questionário avaliativo).

Agrupamento 2 (azul): O grupo 3 possui 583 alunos, com 53% de homens, a maioria dos estudantes possui até 19 anos contabilizando 36% do grupo, e a segunda faixa etária mais populosa é de 20-24 anos correspondendo a 20% do grupo. A escolaridade predominante é o ensino superior incompleto sendo relatada por 38% dos alunos, e em seguida tem-se a pós-graduação com 20%. Com relação a motivação, 51% disseram estar motivados pelo tema "ser importante para seus estudos" e mantém-se a taxa de 95% dos participantes que possuem intenção de terminar o curso e nesse grupo a incidência de não concluintes foi bastante alta nenhum aluno concluiu. Porém, mesmo não havendo concluintes o número de interações entre os alunos e plataforma foi bastante alta, em torno de 108 interações, uma média de 3 acessos a cada vídeo aula e a cada material em texto, e 11 acessos a cada questionário. A média da porcentagem de conclusão das atividades foi de 87% e os alunos passaram em média 16 dias realizando o curso. O desempenho geral dos alunos foi de 33,4 pontos (8,35 por questionário).

Agrupamento 3 (roxo): O grupo 2 é o mais populoso com 2095 alunos, dos quais 66% são mulheres, a maioria 42% está na faixa dos 19 anos, mas há participantes em todas as faixas etárias, 30% dos alunos tem ensino superior incompleto, com relação a motivação, 46% dos estudantes desse grupo disse "ser um tema importante para seus estudos" e os alunos que dizem terem intenção de terminar são 96%. Esse grupo possui uma taxa de não conclusão alta 99% dos alunos não geraram certificados. No que se refere às interações, as médias ficaram baixas, em torno de 21 interações por aluno com a plataforma, como o curso possui 24 atividades, entre recursos de materiais em texto, vídeo aulas e questionários avaliativos, esse número não garante que os alunos acessaram pelo menos uma vez cada recurso disponibilizado. Dessa forma, os alunos tiveram uma média de porcentagem de conclusão das atividades em torno de 8%, o tempo médio que cada aluno permaneceu no curso foi em torno de 2 dias e a nota final dos alunos ficou aproximadamente em 2,8 pontos.

Mais detalhes sobre valores dos atributos dos agrupamentos gerados podem ser visualizados na Tabela 5.

Tabela 5 – Médias e Medianas dos Atributos por Grupos

ATRIBUTO	MÉTRICA	GERAL	GRUPO 1	GRUPO 2	GRUPO 3
Tempo	Mediana	0,02	2,4	1,1	0,0
	Média	7,4	13,2	16,3	2,6
Porcentagem Concluída	Mediana	19 %	100,0 %	96 %	3,8 %
	Média	42 %	96,9 %	86 %	8,5 %
Soma dos Acessos aos Materiais em Texto	Mediana	7	12,0	12,0	0,0
	Média	6,76	13,2	13,1	2,3
Soma dos Acessos às Vídeo Aulas	Mediana	25	50,0	50,0	5,0
	Média	26,31	51,5	51,3	8,9
Soma dos Acessos aos Questionários	Mediana	26	43,5	40,0	0,0
	Média	23,79	43,9	44,4	9,7
Soma das Atividades Realizadas	Mediana	60	108,0	104,0	7,0
	Média	56,86	108,7	108,9	21,1

Fonte: Autora

DEFINIÇÃO DOS PERFIS COMPORTAMENTAIS

Com a geração dos agrupamentos, por meio do algoritmo *K-means*, foi possível observar porque o *elbow method* resultou na melhor divisão em dois agrupamentos, pois ele se baseou apenas da questão de conclusão e não conclusão dos alunos. Todavia, essa é uma visão simplificada dos perfis dos alunos de um MOOC. Por isso, optou-se pela geração de 3 agrupamentos, primeiramente pelo fato da autora já possuir certa familiaridade com as informações constantes na base de dados e em segundo lugar pelos testes iniciais realizados com mais agrupamentos. Nesse sentido, a partir da exploração dos agrupamentos descritos foi possível definir[12] 4 perfis comportamentais para os alunos pesquisados, baseados nos grupos gerados e levando em consideração o que se acredita ser o *objetivo principal* dos estudantes pertencentes a cada perfil determinado:

[12] Cabe destacar que com algoritmos de Aprendizagem de Máquina não supervisionados não há métricas para avaliação, essas interpretações de resultados são na maior parte das vezes inferidas de forma subjetiva pelo pesquisador.

1. **Engajados –** correspondente a maioria dos alunos do Agrupamento 1: Destaca-se que o fator que mais indica o engajamento desses alunos é a quantidade de interações com a plataforma, em torno de 108 no total, também salienta-se que a quantidade de tempo que os alunos permaneceram realizando o curso, em média 13 dias, considera-se adequado para essa quantidade de interações. Os alunos também concluíram em média cerca de 96% das atividades do curso, evidenciando que a maioria dos recursos foi acessado, e o desempenho dos alunos comprova essa afirmativa, pois as notas por questionário se mantiveram acima de 9 pontos. Outro fator indicativo do engajamento desses alunos foi a quantidade de interações por tipo de material estudado, tendo uma média de acesso por tipo de recurso igual ou superior a 3, o que denota o interesse desses alunos em acessar os conteúdos do curso e também na obtenção do certificado, pois todos os alunos desse grupo geraram o certificado, por isso acredita-se que o objetivo predominante de alunos com esse perfil fosse a aprendizagem do tema do curso e certificar essa aprendizagem, ou seja engajados com todos os recursos do MOOC.

2. **Estratégicos –** correspondente a maioria dos alunos do Agrupamento 2: Neste grupo não há concluintes, mas, embora eles não tenham concluído (gerado certificado) a maioria desses alunos segue os mesmos padrões de comprometimento dos alunos com perfil *Engajados*, 108 interações com a plataforma, uma média de 3 acessos a cada vídeo aula e a cada material em texto, no que se refere a média da porcentagem de conclusão das atividades foi em torno de 87% e os alunos passaram aproximadamente 16 dias realizando o curso, quanto ao desempenho geral dos alunos estes obtiveram notas acima de 8 pontos por questionário, um perfil bem similar aos alunos do grupo 1. Desta forma, acredita-se que os estudantes deste perfil, principalmente aqueles que ultrapassam a média de tempo de acesso dos Engajados, tinham como principal interesse a aprendizagem dos conteúdos de um ou mais tópicos do curso, dos quais obtiveram o conhecimento e depois não realizaram os demais módulos e não geraram o certificado, possuindo assim uma estratégia de estudo individual. Ressalta-se que alguns desses indivíduos ainda podem ter gerado o certificado após a extração das bases de dados.

Devido, sobretudo, à grande quantidade de alunos do grupo 3 (2095) foi possível identificar dois perfis predominantes, um perfil entre os alunos que não concluíram e outro para os alunos que concluíram o MOOC:

1. **Inativos –** correspondente aos alunos do Agrupamento 3 que *não concluíram* o curso: São alunos que realizaram muito poucas interações com a plataforma, em torno de 21, a porcentagem concluída foi de 8% e tiveram uma média de 2 dias de realização de curso, pode-se dizer que apenas visualizaram os primeiros recursos do curso e desistiram, isso pode ocorrer por diversos motivos, um deles seria que os conteúdos disponíveis no MOOC não eram o que estavam estes estudantes estavam esperando.
2. **Oportunistas –** corresponde aos alunos dos Agrupamento 3 que *concluíram* o curso: Embora seja predominante esse perfil nos alunos do grupo 3 que concluíram (total dos alunos) também foi possível identificar alguns estudantes com algumas características de "oportunistas" no grupo 1, de forma menos frequente. Esses alunos têm um perfil de inatividade, entretanto eles terminaram o curso, porém realizando apenas as atividades necessárias para conseguir a certificação, por isso esses alunos ficaram junto aos inativos no agrupamento realizado pelo *K-means*. Foi percebido que esses alunos possuem uma quantidade razoável de interações com a plataforma, em torno de 33 interações mais precisamente, entretanto tais interações estão concentradas na realização de tentativas de questionários avaliativos, com uma média de 31 tentativas, sobrando apenas em torno de 2 interações com os outros materiais do curso (que totalizam 20). A porcentagem de conclusão de atividades também é baixa aproximadamente 22%. Porém, o que mais chama a atenção é que a média de tempo que esses alunos levaram para terminar o curso, menor que um dia, visto que o curso possui uma carga horária de 15 horas, este tempo é bem pequeno; isso significa que os alunos fizeram o primeiro acesso ao curso e geraram o certificado no mesmo dia. Outro fato que chama a atenção é o grande número de interações em tentativas nos questionários, sem acessar o restante dos materiais, como já salientado. Diante destas constatações analisa-se o perfil desses indivíduos como de "oportunistas", ou "caçadores de certificados", são indivíduos que possuem

> interesse no certificado e não acessam todos os materiais disponíveis. Especula-se, que esses indivíduos também não possuem muitos conhecimentos prévios sobre o assunto, porque necessitam fazer muitas tentativas para alcançar a nota mínima nos questionários. Dessa forma, percebeu-se que o objetivo desses indivíduos é a obtenção da certificação.

Pode-se analisar de forma mais ilustrativa, as características dos "oportunistas", na Figura 24, onde é mostrada a distribuição dos estudantes nos agrupamentos encontrados pelo *K-means*. Por exemplo, no grupo 1, onde todos emitiram certificado, pode-se verificar 295 estudantes que fizeram o curso todo em um único dia. No grupo 3 percebe-se que diferentemente dos outros agrupamentos, a maioria dos estudantes estão com índices baixos na porcentagem de atividades concluídas, bem como o tempo e a quantidade de interações com a plataforma, indicada na soma das atividades realizadas, fatores que denotam um comportamento típico de alunos desistentes (inativos), todavia há uma parcela de alunos que concluem, mesmo com tão poucas interações, característico do perfil de "oportunistas".

CONSIDERAÇÕES SOBRE O MAPEAMENTO DE PERFIS DE ALUNOS

O propósito deste estudo foi desenvolver um mapeamento de perfis comportamentais de alunos em um MOOC de uma plataforma brasileira, enfocando os objetivos e interesses desses alunos quando da realização do curso, esse mapeamento foi conduzido por meio de uma abordagem que empregou o algoritmo de aprendizagem de máquina *K-means* e uma análise dos agrupamentos gerados.

Com base na abordagem relatada foi possível observar 4 perfis comportamentais de alunos que realizaram esse curso: *1) Engajados:* estudantes que tem como principal objetivo interação com todo o conteúdo do curso, bem como sua finalização para obtenção do certificado; *2) Estratégicos:* que possuem interesse em conhecer algum conteúdo específico do curso; *3) Inativos:* alunos que realizam apenas a inscrição no curso, ou acessam apenas os primeiros materiais e não finalizaram; e *4) Oportunistas:* tem como principal intuito a geração do certificado,

sem acessar todos os materiais do curso, mesmo que aparentemente estes não possuam conhecimento prévio sobre a temática do MOOC.

Figura 24 – Distribuição dos estudantes por Agrupamento

Atributo	Intervalo	Grupo 1	Grupo 2	Grupo 3
Tempo	0 - 1	295	277	1815
	1 - 20	459	208	200
	20 - 40	52	39	41
	40 - 60	14	14	17
	>= 60	42	45	22
Porcentagem Concluída	0 - 20%	0	23	1778
	20 - 40%	17	7	276
	40 - 60	17	34	41
	60 - 80%	11	39	0
	>= 80%	817	480	0
Soma dos Acessos aos Materiais em Texto	0 - 1	4	1	1201
	1 - 10	42	53	858
	10 - 20	776	487	35
	20 - 30	37	41	0
	>= 30	3	1	1
Soma dos Acessos às Vídeo Aulas	0 - 1	1	0	785
	1 - 30	33	25	1291
	30 - 60	717	501	19
	60 - 90	97	31	0
	>= 90	12	26	0
Soma dos Acessos aos Questionários	0 - 1	0	2	1263
	1 - 30	132	124	451
	30 - 60	645	363	381
	60 - 90	83	67	0
	>= 90	1	27	0
Soma das Atividades Realizadas	0 - 1	0	0	696
	1 - 30	0	2	822
	30 - 60	16	11	222
	60 - 90	83	101	355
	>= 90	763	469	0
Certificado	Sim	862	0	14
	Não	0	583	2081

Fonte Autora

A partir da abordagem desenvolvida pode-se realizar mais testes em outros MOOCs da plataforma para que seja possível validar este experimento, verificando se perfis similares são percebidos em outros cursos, e se essa constatação pode ser generalizada. Dessa forma, com a ampliação da validação da existência desses perfis é possível realizar configurações nos

MOOCs que possam inibir os maus comportamentos apresentados pelos *Oportunistas*, como por exemplo ampliar a complexidade dos questionários avaliativos; além disso, pode-se otimizar a experiência para os *Engajados* e *Estratégicos* com o oferecimento de "planos de benefícios" com certificados diferenciados, salientando por exemplo a nota obtida no decorrer da realização do MOOC, ou oferecendo descontos em cursos não gratuitos oferecidos pela plataforma; por fim, tentar implementar dispositivos que aumentem a taxa de conclusão entre os *Inativos*, como por exemplo ao notar que esses alunos estão há alguns dias sem acessar a plataforma tentar interagir, de forma a perguntar sobre suas dificuldades, se precisam de algum tipo de ajuda e se tem interesse em continuar no curso.

PROCESSO DE MINERAÇÃO DE DADOS EDUCACIONAIS APLICADO NA PREVISÃO DO DESEMPENHO DE ALUNOS

Com a o aumento da disponibilidade de dados, sobretudo no contexto educacional, surgiram áreas específicas para extração de informações relevantes, como a MDE tema deste Livro, que integra inúmeras técnicas que dão suporte a captação, processamento e análises desses conjuntos de registros. A principal técnica associada a MDE é a Aprendizagem de Máquina, que vem sendo empregada a décadas no processamento de dados em diversos contextos, mas com a evolução tecnológica outras técnicas tem se sobressaído como Aprendizagem Profunda (tais técnicas foram descritas no Capítulo 5), baseada na aplicação de Redes Neurais Artificiais Multicamadas (RNAM).

Diante desse contexto, esse experimento teve como objetivo aplicar o processo de MDE no intuito de prever o desempenho de alunos, utilizando um conjunto de dados público do repositório *UCI Machine Learning*[13] e comparar técnicas já consolidadas no âmbito da MDE, com a técnica de AP. Dessa forma, além de verificar se os atributos que compõe a base de dados são suficientes para realizar a geração de modelos eficazes na previsão do desempenho dos alunos, foi possível avaliar se a Aprendizagem

[13] https://archive.ics.uci.edu/ml/datasets/Student+Performance.

Profunda é uma boa alternativa as técnicas mais tradicionais empregadas no âmbito da MDE. Para isso foi implementado um processo de MDE baseado nas 4 etapas descritas no Capítulo 1. Como resultado foi identificado que os modelos gerados a partir dos algoritmos tradicionais de AM têm um bom desempenho, mas inferior ao modelo AP que teve uma acurácia de 94%, bem como foi constatado que atributos relacionados as atividades escolares são mais preditores para o desempenho dos alunos do que os dados de características demográficas e socioeconômicas.

DESENVOLVIMENTO DO PROCESSO DE MDE PARA A PREVISÃO DO DESEMPENHO DE ALUNOS

Este experimento teve como principal objetivo realizar a previsão do desempenho de alunos, em duas disciplinas no ensino tradicional, utilizando técnicas de MDE. Com a realização desse processo é foi também possível identificar qual das técnicas aplicadas foi a mais eficaz na predição do desempenho desses estudantes, considerando os atributos que compõe a base de dados. À vista disso, formulou-se as questões de pesquisa que nortearam este estudo:

Questão de pesquisa 1 (QP1) – Qual a eficácia de modelos gerados a partir de algoritmos baseados em Aprendizagem de Máquina e Aprendizagem Profunda na previsão do desempenho de alunos no ensino tradicional?

Questão de pesquisa 2 (QP2) – Modelos baseados em algoritmos de Aprendizagem Profunda têm uma eficácia superior a modelos baseados em algoritmos tradicionais utilizados na Mineração de Dados Educacionais?

Questão de pesquisa 3 (QP3) – Qual o conjunto de atributos tem mais influência na previsão do desempenho de alunos?

Para responder essas questões foram empregados alguns procedimentos que configuram a metodologia adotada nesta pesquisa, que em termos gerais resume-se na realização do processo de MDE já explorado no experimento anterior: 1) Definição da função da MDE; 2) Formatação dos dados que serão utilizados; 3) Definição das Técnicas de MDE; 4) Delineamento de como essas técnicas serão aplicadas; e 5) Análise e/ou interpretação dos resultados.

DEFINIÇÃO DA FUNÇÃO DA MDE

Neste experimento o intuito do processo de MDE foi a partir dos atributos dos alunos prever o desempenho, as notas finais, identificando quais algoritmos tiveram maior eficácia nesta previsão. Esse tipo de estudo pode apoiar na realização de análises que auxiliem professores a realizar intervenções antecipadamente evitando desempenhos a baixo da média e até mesmo reprovações.

FORMATAÇÃO DOS DADOS QUE FORAM UTILIZADOS

Primeiramente os dados foram coletados do repositório de dados público o *UCI para Machine Learning*. Estes dados abordam o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos dos alunos incluem notas, frequência e características demográficas (sociais e escolares). Tais informações foram reunidas por meio de relatórios escolares e questionários. Foram fornecidos dois conjuntos de dados de 1044 alunos relativos ao desempenho em duas disciplinas distintas: Matemática e Língua Portuguesa. Os atributos constantes na base de dados extraída estão descritos na Quadro 9. Os dados sistematizados formaram um *Data Frame*[14] com 1044 linhas e 33 colunas.

Após a coleta dos dados, estes passaram pelo pré-processamento para se adequar a aplicação das técnicas de MDE – algoritmos de AM e AP – utilizadas na previsão do desempenho dos alunos, para isso foi utilizada a linguagem de programação e ciência de dados R[15] e várias tarefas foram realizadas: (1) Junção dos dados dos alunos das duas disciplinas em uma única base, para todos os alunos; (2) Transformação do atributo G3 de numérico para níveis de classificação, nesse procedimento cada faixa de notas recebeu um valor no formato de caractere, atribuindo uma categoria/classe para os registros: notas entre 20 e 16 = "A", notas entre 15 e 11 =

[14] Um Data Frame é semelhante a uma matriz, mas as suas colunas têm nomes e podem conter dados de tipos diferentes.

[15] R é uma linguagem de programação multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados.

"B", notas entre 10 e 4 = "C", notas entre 4 e 0 = "D"; (3) Formatação dos dados como Data Frame; (4) Divisão da base de dados em treinamento e teste, em que 85% dos dados da base foram definidos para treino (888 registros para treino e 156 para teste); e (5) Transformação do atributo a ser previsto (atributo meta – o desempenho final do aluno) para Factor (G3).

Quadro 9 – Atributos da Base de Dados.

ID	ATRIBUTOS	DESCRIÇÃO
1	Escola	Escola do aluno (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
2	Gênero	Gênero do aluno (binário: 'F' - feminino ou 'M' - masculino)
3	Idade	Idade do aluno (numérico: de 15 a 22)
4	Endereço	Tipo de endereço residencial do aluno (binário: 'U' - urbano ou 'R '- rural)
5	Famsize	Tamanho da família (binário:' LE3 '- menor ou igual a 3 ou' GT3 '- maior que 3)
6	Pstatus	Status de coabitação dos pais (binário:' T '- morando junto ou' A '- à parte)
7	Medu	Escolaridade da mãe (numérico: 0 - nenhum, 1 - ensino fundamental (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior)
8	Fedu	Escolaridade do pai (numérico: 0 - nenhuma, 1 - ensino primário (4º ano), 2 - 5º ao 9º ano, 3 - ensino secundário ou 4 - ensino superior)
9	Mjob	Trabalho da mãe (nominal: 'professor', 'saúde' relacionado, 'serviços' civis (por exemplo, administrativo ou policial), 'em_casa' ou 'outro')
10	Fjob	Trabalho do pai (nominal: 'professor', 'saúde' relacionado, civil 'serviços' (por exemplo, administrativo ou policial), 'em_casa' ou 'outro')
11	Razão	Razão para escolher esta escola (nominal: perto de 'casa', escola 'reputação', 'curso' preferência ou 'outro')
12	Tutor	Tutor do aluno (nominal: 'mãe', 'pai' ou 'outro')
13	Tempo de Viagem	Tempo de viagem de casa para a escola (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. A 1 hora, ou 4 -> 1 hora)

ID	ATRIBUTOS	DESCRIÇÃO
14	Horas de Estudo	Tempo de estudo semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 -> 10 horas)
15	Reprovações	Número de reprovações anteriores nas aulas (numérico: n se 1 <= n <3, senão 4)
16	Schoolup	Suporte educacional extra (binário: sim ou não)
17	Famsup	Suporte educacional familiar (binário: sim ou não)
18	Pago	Aulas extras pagas dentro da disciplina (matemática ou português) (binário: sim ou não)
19	Atividades	Atividades extracurriculares (binário: sim ou não)
20	Creche	Cursou creche (binário: sim ou não)
21	Superio	Deseja cursar o ensino superior (binário: sim ou não)
22	Internet	Acesso à internet em casa (binário: sim ou não)
23	Romântico	Com um relacionamento romântico (binário: sim ou não)
24	Famrel	Qualidade das relações familiares (numérico: de 1 - muito ruim a 5 - excelente)
25	Tempo Livre	Tempo livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto)
26	Gooout	Saindo com os amigos (numérico: de 1 - muito baixo a 5 - muito alto)
224	Dalc	Consumo de álcool durante o trabalho (numérico: de 1 - muito baixo a 5 - muito alto)
28	Walc	Consumo de álcool no fim de semana (numérico: de 1 - muito baixo a 5 - muito alto)
29	Saúde	Estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom)
30	Faltas	Número de faltas na escola (numérico: de 0 a 93)
32	G1	Nota do primeiro período (numérico: de 0 a 20)
32	G2	Nota do segundo período (numérico: de 0 a 20)
33	G3	Nota final (numérico: de 0 a 20, meta de saída)

Fonte: *UCI – Machine Leraning* (Cortez & Silva, 2008)[16]

[16] https://archive.ics.uci.edu/ml/datasets/Student+Performance.

DEFINIÇÃO DAS TÉCNICAS DE MDE

Os principais algoritmos empregados para previsão de desempenho de alunos, em muitos estudo analisados como em Chui et al. (2020); Zhang e Wu (2019); Wen *et al.* (2020) e Waheed *et al.* (2020) (citados no Capítulo 6) utilizaram algoritmos de Aprendizagem de Máquina e Aprendizagem Profunda do tipo supervisionado, podendo ser aplicadas a regressão ou a classificação. Neste experimento foram aplicados algoritmos de Aprendizagem de Máquina do tipo supervisionados de classificação e o diferencial foi compará-los com uma Rede Neural Artificial Multicamadas (RNAM) vinculada à Aprendizagem Profunda.

Em Cortez e Silva (2008), os dois conjuntos de dados, utilizados neste experimento, foram modelados com uma classificação de cinco níveis e foi utilizada a regressão, pois o atributo que os autores queriam prever era a nota (numérico). Em contraste, neste estudo as notas de desempenho dos alunos foram classificadas como A, B, C ou D, e, portanto, foram empregados algoritmos de classificação para realização das previsões. Tanto a classificação como a regressão são tarefas de Aprendizagem Supervisionada, nesse tipo de técnica a base de dados possui colunas com categorias que servem para treinar o modelo, que deve, na próxima etapa, identificar as categorias de cada linha, como explicado no Capítulo 5.

Alguns dos algoritmos de AM para classificação mais conhecidos são: Naïve Bayes; Árvore de Decisão, *Random Forest* (RF) e Suport Vector Machines (SVM), e na AP a arquitetura para Redes Neurais Artificiais Multicamadas (RNAM), Multilayer Perceptron (MLP) é uma boa opção para classificação em bases formadas por atributos simples (que não incorporam a mineração em texto, em imagem ou em vídeos). Estes algoritmos foram empregados neste experimento para previsão do desempenho dos alunos. Assim foi possível comparar os algoritmos mais tradicionais de AM com as RNAM vinculadas a AP.

DELINEAMENTO DE COMO AS TÉCNICAS FORAM APLICADAS

Para realização da aplicação desses algoritmos foi utilizada a linguagem R, em que foram empregadas bibliotecas específicas para cada algoritmo utilizado: Naïve Bayes – bliblioteca "e1071"; Arvores de Decisão – biblioteca "rpart"; *Random Forest* – biblioteca "randomForest"; Suport Vector Machine – biblioteca "e1071"; e para AP foi utilizado o framework "H2O". Além das bibliotecas outra importante questão a ser considerada na aplicação de algoritmos de AM e AP é a configuração de seus principais parâmetros, elementos que influem diretamente na eficácia dos modelos gerados. As configurações feitas nos algoritmos utilizados neste estudo estão sistematizadas na Quadro 10. Uma explanação detalhada sobre os algoritmos utilizados e também sobre a arquitetura de RNAM utilizada nesse experimento foi realizada no Capítulo 5.

Quadro 10 – Configuração dos Algoritmos.

ALGORITMO	PARÂMETROS CONFIGURADOS
Naïve Bayes	Configuração Default.
Árvore de Decisão	Configuração Default.
Random Forest	Configuração Default, e foi definido uma floresta com 30 árvores.
SVM	Foi definido o kernel "radial" e um valor de custo de 5.0.
RNAM	Foram definidas 3 camadas ocultas, com 200 neurônios cada, a quantidade de épocas de ajuste foi de 800 e a função de ativação foi a "rectifier"

Fonte: Autora

Com relação as ferramentas para a aplicação da Aprendizagem Profunda, é importante destacar que a escolha é um pouco mais complexa do que apenas a utilização das bibliotecas disponíveis na linguagem de ciência de dados a ser utilizada, como para os algoritmos de Aprendizagem de Máquina. A partir da decisão de se utilizar a AP como técnica de MDE, foi necessária a elaboração de uma pesquisa sobre as principais e melhores ferramentas disponíveis, de preferência sob a licença

de código aberto para sua aplicação. A escolha pelo código aberto, diz respeito sobretudo pela flexibilidade para configuração das ferramentas, não disponível nas proprietárias.

Para a aplicação de AP é necessário um framework[17] especializado neste tipo de algoritmo, até mesmo por questões físicas do computador – quanto a hardware – que devem ser levadas em consideração neste cenário, pois o processamento requerido pelas RNAM é mais complexo e com bases de dados de grandes dimensões, muito recurso computacional é exigido.

Neste sentido, analisou-se 3 pesquisas que realizaram comparações entre frameworks para Aprendizagem Profunda: Bahrampour *et al.*, (2015) que verificaram os desempenhos do Caffe, do Neon, do TensorFlow, do Theano e do Torch; Kovalev; Kalinovsky; Kovalev (2016) que realizaram uma avaliação do funcionamento e eficácia do Theano, do Torch, do Caffe, do TensorFlow e do DeepLearning4J; e por fim, NG *et al.* (2016) que implementaram um estudo comparativo da performance do Singa e do H2O.

Das conclusões dos autores destacam-se: O TensorFlow é suficientemente flexível, todavia seu desempenho em tempo de treinamento com uma única GPU não é satisfatória em comparação aos outros; O framework Deeplearning4J tem uma performance inferior para o treinamento, mesmo modificando a quantidade de camadas da Rede Neural; Quanto a complexidade de programação o mais complicados de se utilizar é o Torch; e O H2O obteve uma performance estável e precisa para as bases de dados testadas, possui baixa complexidade de implementação, podendo ser utilizado com as linguagens R ou Python e proporciona acesso a várias GPUs (*Graphics Processing Unit*) integradas em servidores on-line.

A partir dos itens destacados, foi possível escolher uma ferramenta flexível, de baixa complexidade e eficaz, o H2O, como ele está disponível para várias linguagens de programação optou-se por usar o R, pois essa linguagem já vem sendo utilizada pela autora em diversos estudos.

[17] De acordo com Alvim (2010) framework é uma coleção de classes que contribuem entre si propiciando melhores práticas de desenvolvimento e diminuição da repetição de tarefas. Além disso, evita variações de soluções diferentes para um mesmo tipo de problema, o que facilita a reutilização e customização dos códigos.

Além disso, o framework para linguagem R possibilita a utilização de atributos categóricos – sem precisar convertê-los para numéricos, como no Python – um diferencial decisivo para sua escolha, esse fator diminui muito a complexidade do pré-processamento e transformação a serem empregados nas bases de dados em linhas de código. Outra questão muito importante é o treinamento do modelo que é realizado em GPU, devido a grandes dimensionalidades de bases, assim como da complexidade de processamento realizado pela RNAM, sem essa possibilidade seria inviável a execução em um computador comum.

ANÁLISE E/OU INTERPRETAÇÃO DOS RESULTADOS

Por fim, realizou-se a análise e interpretação dos resultados alcançados. Essa etapa nesse experimento pretendeu avaliar a eficácia na previsão do desempenho dos alunos em cada modelo gerado pelos algoritmos. Nesse sentido, para realizar a verificação dos resultados de um modelo de classificação são necessários dois itens: os métodos de avaliação e as métricas de interpretação. Os dois devem ser aplicados em conjunto para que seja possível observar se um modelo é eficaz ou não. Os métodos indicam como esse modelo será avaliado, e as métricas traduzem os resultados da aplicação desses métodos em números que possam ser interpretados os métodos e métricas de avaliação de algoritmos mais utilizados na Aprendizagem de Máquina foram expostos no Capítulo 5.

Para este experimento o método de avaliação empregado foi o de Treinamento e Teste, em que a base de dados é dividida de forma aleatória em duas porções, uma para treinamento e outra para teste, de acordo com Japkowicz & Shah (2014) geralmente são empregados 85% das instâncias para treinamento e 15% para teste. O algoritmo ao ser aplicado sobre a base de treinamento recolhe informações sobre os atributos das instâncias e gera um modelo de classificação ou regressão com base nesses atributos e informações, após isso esse modelo é aplicado sobre a base de teste (que contém registros diferentes da base de treinamento) e então as métricas de avaliação são calculadas sobre essa aplicação.

Apenas a aplicação do método de avaliação não indica se o modelo é eficaz ou não, para isso devem ser utilizadas métricas que possibilitem interpretação do quanto o modelo foi preciso em suas classificações, em outras palavras quantificar o seu desempenho. As métricas que foram utilizadas nesse experimento foram: Precisão da classificação (Acurácia) – é o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes; Intervalo de Confiança (IC) – corresponde a uma métrica que indica que há uma probabilidade de 95% que a verdadeira precisão do modelo algorítmico testado esteja dentro desse intervalo; Taxa de não informação – é a precisão alcançável, sempre prevendo a categoria da classe majoritária; Valor de P – consiste em um teste unilateral para verificar se a precisão é melhor que a taxa de não informação, considerando a maior porcentagem da classe dos dados; Kappa – corresponde a uma medida de concordância usada em escalas nominais que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas são as interpretações.

Os resultados desse experimento foram sistematizados de acordo com as três questões de pesquisa, e são apresentados na sequência.

QP1 – QUAL A EFICÁCIA DE MODELOS GERADOS A PARTIR DE ALGORITMOS BASEADOS EM APRENDIZAGEM DE MÁQUINA E APRENDIZAGEM PROFUNDA NA PREVISÃO DO DESEMPENHO DE ALUNOS NO ENSINO TRADICIONAL?

Foi gerado um modelo para cada algoritmo utilizado nesse estudo, a partir de sua aplicação na base de dados. Os resultados referentes ao método de avaliação Treinamento/Teste e as métricas mais proeminente para a análise da eficácia do modelo estão disponíveis Tabela 6[18]. Com esses resultados percebe-se que excluindo o algoritmo Naïve Bayes, que realmente tem uma estrutura bastante simplificada, os demais algoritmos,

[18] Para conferir os resultados de todas as métricas de avaliação expostas no capítulo 5 consulte o Apêndice B.

tiveram boa performance com precisão de classificação acima de 80%, em todos os casos, sendo boas opções para a previsão do desempenho de alunos, em bases de dados compostas por registros demográficos, sociais e de rendimento escolar.

Tabela 6 – Análise dos algoritmos.

MÉTRICA	Naïve Bayes	Árvore de Decisão	*Random Forest*	SVM	RNAM
Acurácia	0,66	0,87	0,83	0,82	0,94
Intervalo de confiança de 95%	58-74%	81-92%	77-89%	75-88%	79-94%
Taxa de não informação	0,38	0,5	0,51	0,51	0,53
Valor de p	9.791e-13	2.2e-16	2.2e-16	8.754e-16	2.2e-16
Kappa	0,51	0,79	0,73	0,71	0,79

Fonte: Autora.

No que diz respeito as métricas apresentadas na Tabela 5, a primeira é a acurácia – número de previsões corretas divididas pelo número total de previsões – correspondente a 94% nas RNAM e a 87% no algoritmo de Árvore de Decisão, com um intervalo de confiança de 79-94% e 81-92% respectivamente, o que significa que há uma probabilidade de 95% que a verdadeira precisão desses modelos esteja dentro desse intervalo. Logo após, encontra-se a taxa de não informação que corresponde a 50% para Árvore de Decisão e 53% para as RNAM, essa métrica indica a precisão alcançável sempre prevendo a categoria da classe majoritária. Conforme o valor de p – que é igual para os dois – pode-se afirmar, que os modelos gerados por esses algoritmos, oferecem um desempenho significativamente melhor sobre a taxa de não informação. Na sequência a estatística Kappa, que apresentou valor de 79%, para ambos, mostra quão bem as previsões dos modelos corresponderam as categorias reais da classe; de acordo com as diretrizes propostas por Landis e Koch (1977) a Kappa nada mais é que uma concordância justa entre o modelo e as verdadeiras categorias de uma classe, uma vez que a precisão aleatória é controlada.

QP2 – MODELOS BASEADOS EM ALGORITMOS DE APRENDIZAGEM PROFUNDA TÊM UMA EFICÁCIA SUPERIOR A MODELOS BASEADOS EM ALGORITMOS TRADICIONAIS UTILIZADOS NA MINERAÇÃO DE DADOS EDUCACIONAIS?

De acordo com os resultados da avaliação o modelo de AP obteve uma acurácia superior à dos algoritmos mais tradicionais, em torno de 94%, confirmando estudos como o de Wen *et al.* (2020), Waheed *et al.* (2020) (detalhados no Capítulo 6) que encontraram valores de eficácia variando de 84% a 93%. Todavia, cabe salientar que esses valores dependem muito de como as bases de dados foram formatadas, assim como do framework e das configurações utilizadas para a aplicação das RNAM, e sendo assim esses resultados não podem ser generalizados de forma abrangente, para outras bases de dados. Ademais, as Árvore de Decisão tiveram uma precisão alta nesse conjunto de dados chegando a 87%, com um Intervalo de confiança de 81-92%, o que devido a sua simplicidade de aplicação e velocidade de processamento pode representar uma alternativa mais vantajosa em coleções de registros similares, desde que o estudo desenvolvido não necessite de acurácias muito altas.

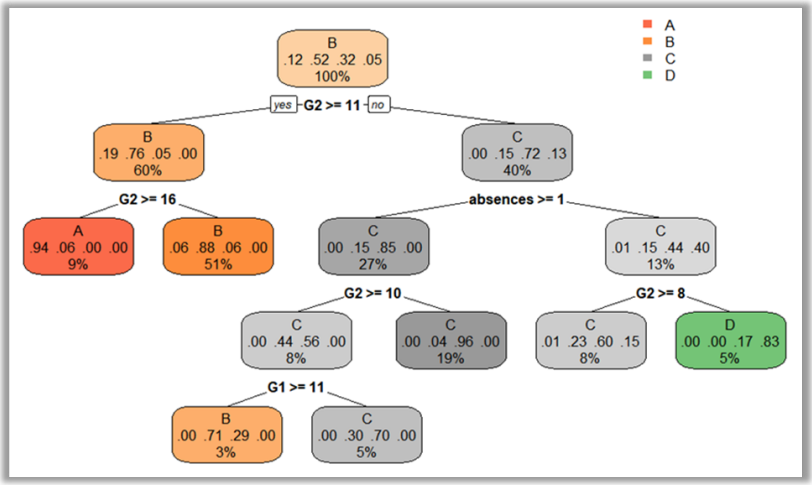
Portanto, mesmo que o modelo baseado em AP tenha apresentado uma precisão maior, não é tão superior ao algoritmo Árvore de Decisão, no contexto desse estudo, cabe então ao pesquisador decidir. Para essa decisão 4 fatores são determinantes: precisão; simplicidade de configuração; recursos computacionais requeridos; e tempo de processamento. O modelo de AP é mais preciso, mas a tarefa de configuração é mais complexa, necessita de mais recursos computacionais e tem um tempo de processamento pelo menos 2 vezes maior que os outros algoritmos utilizados. Em contra partida, o algoritmo Árvore de Decisão tem uma acurácia menor.

QP3 – QUAL O CONJUNTO DE ATRIBUTOS TEM MAIS INFLUÊNCIA NA PREVISÃO DO DESEMPENHO DE ALUNOS?

Para definir quais os elementos foram mais influentes na previsão do desempenho, foi gerado o gráfico (Figura 25) de Árvore de Decisão (com a biblioteca "rpart.plot"), este propicia visualizar quais são os atributos que estão mais no topo da árvore. Tais atributos, devido aos cálculos de entropia realizados para geração da árvore de decisão, são os mais importantes para prever o atributo meta, as notas dos alunos. Nesse sentido, não houveram grandes supressas, ou descobertas, pois de acordo com a Figura 25 os principais atributos para prever o desempenho dos alunos são as notas (G2 na raiz da árvore e nos níveis 2 e 3; e G1 no nível 4), e a quantidade de faltas (*absences* no nível 2). Dessa forma, demonstrando que os atributos vinculados as atividades escolares são mais preditores para o desempenho que dados de características demográficas e socioeconômicas.

Todavia, não se pode descartar a influência desses elementos no desempenho dos alunos, pois é de conhecimento que estudantes podem ter seu rendimento escolar prejudicado, ou abaixo do esperado, por estarem enfrentando alguma adversidade em casa, o que impacta em suas notas e pode ocasionar uma baixa frequência. Como não há registros de questionários aplicados para entender melhor esses elementos, não há uma confirmação dessas suposições, que são consideradas adequadas, mas não podem ser verificadas. Devido a essa falta de evidências, com base nos indicadores gerados pela aplicação das técnicas de MDE, os principais atributos que influenciam são os relacionados ao desempenho escolar.

Figura 25 – Gráfico de Árvore de Decisão.



Fonte: Autora.

CONSIDERAÇÕES SOBRE A PREVISÃO DO DESEMPENHO DE ALUNOS

Este experimento teve como principal objetivo aplicar o processo de MDE no intuito de realizar a previsão do desempenho de alunos, utilizando um conjunto de dados público e comparar técnicas já consolidadas no âmbito da MDE, com a técnica de AP. Ademais, foi possível identificar quais são os atributos que dão melhor suporte na previsão de desempenho dos alunos. Com relação à previsão de desempenho as técnicas de MDE aplicadas foram adequadas, em que os resultados alcançados são os seguintes: Naïve Bayes com uma acurácia de 66%; Árvore de Decisão com 87%; *Random Forest* com 83%; *Suport Vector Machine* com 82%; e RNAM com 94% de acurácia. Esses resultados confirmam que a AP aplicada no âmbito da MDE tem apresentado um bom desempenho, com uma eficácia promissora, o que confirma estudos mais amplos como os desenvolvidos por Wen *et al.* (2020) e Waheed *et al.* (2020), que foram descritos com mais detalhes no Capítulo 6.

No estudo de Wen *et al.* (2020), os autores empregaram um modelo de Aprendizagem Profunda, uma Rede Neural Convolucional, para prever a desistência em cursos do tipo MOOC. No intuito de verificar a eficácia do modelo desenvolvido os autores realizaram uma comparação com diversos algoritmos tradicionais de AM: Árvore de Decisão (tipo de implementação - CART), Naïve Bayes, *Linear Discriminant Analysis*, Regressão Logística, SVM, *Random Forest* e *Gradient Boosted Decision Tree*. Em todos os 4 experimentos realizados a acurácia da Rede Neural Convolucional foi superior aos algoritmos de AM, permanecendo entre 86% e 89%. Na pesquisa desenvolvida por Waheed *et al.,* (2020), os autores tinham como objetivo desenvolver um modelo baseado em Aprendizagem Profunda para prever o desempenho acadêmico dos alunos. Para validar seu modelo os autores o compararam com dois algoritmos tradicionais de AM: a Regressão Logística e o *Suport Vector Machine*. Os resultados apontaram que o modelo de AP obteve uma acurácia entre 84% e 93%, enquanto a Regressão Logística atingiu acurácia entre 79% e 85% e o *Suport Vector Machine* alcançou acurácia entre 79% e 89%.

Entretanto, no estudo de Wen *et al.*(2020), bem como neste experimento, outros algoritmos mais simples obtiveram também bons desempenhos, como exemplo cita-se as Árvore de Decisão que neste estudo obteve uma acuraria de 87% e em Wen *et al.*(2020) de 75%, que são resultados também relevantes para um algoritmo tecnicamente simples e que necessita de pouco esforço para configuração, e sobretudo tem um tempo de processamento bem inferior ao das RNAM. Por isso, é importante destacar que para estudos que não requeiram altos valores de precisão, algoritmos mais simples podem ser melhores opções.

Em relação conjunto de atributos com mais influência na previsão do desempenho de alunos não foi identificada uma descoberta relevante, pois de acordo com o gráfico de Árvore de Decisão gerada, os atributos referentes as notas e as faltas dos alunos são os mais preditivos para o desempenho, fato que não provocou uma surpresa. Embora, possa levantar questionamentos sobre o que ocasionou um desempenho abaixo do esperado em alguns alunos.

Por fim, este experimento apresentou como principal contribuição demostrar a aplicação do processo de MDE, em um conjunto de dados público, que pode ser replicado por outros pesquisadores, com um nível de detalhe pouco encontrado em textos dessa área. Além disso, os resultados das avaliações dos algoritmos evidenciados, podem dar suporte na escolha de métodos mais eficazes para a aplicação em conjuntos de dados educacionais, levando em consideração a Aprendizagem Profunda, que sem dúvida é uma técnica bastante relevante no contexto da MDE.

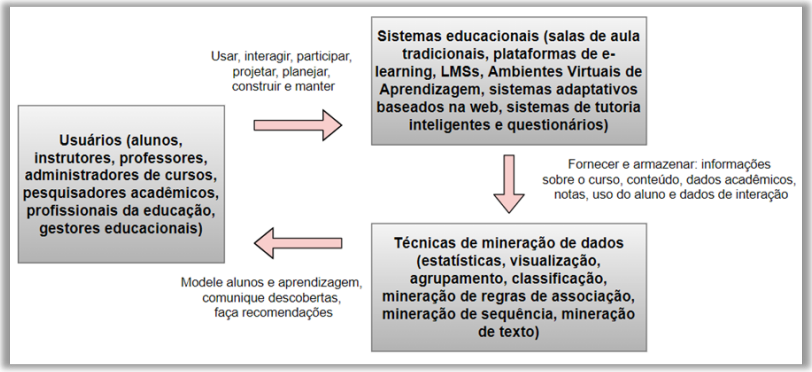
CAPÍTULO 8 - CONSIDERAÇÕES FINAIS SOBRE A MINERAÇÃO DE DADOS EDUCACIONAIS

De acordo com o primeiro Manual sobre Mineração de Dados Educacionais (ROMERO *et al.*, 2010) a aplicação de técnicas de mineração de dados a sistemas educacionais para melhorar o aprendizado pode ser vista como uma estratégia de avaliação formativa. Para estes autores a avaliação formativa de um programa educacional enquanto este ainda está em desenvolvimento, tem o propósito de melhorar continuamente este programa. Examinar como os alunos usam o sistema é uma forma de avaliar o *design* instrucional de uma maneira formativa e pode ajudar os designers educacionais a melhorar os materiais instrucionais (ROMERO *et al.*, 2010). Técnicas de mineração de dados podem ser usadas para coletar informações que podem ser usadas para auxiliar designers educacionais a estabelecer uma base pedagógica para decisões ao projetar ou modificar a abordagem pedagógica de um ambiente.

Para Romero *et al.* (2010) a aplicação da mineração de dados ao projeto de sistemas educacionais é um ciclo iterativo de formação, teste e refinamento de hipóteses (ver Figura 26). O conhecimento extraído deve entrar no ciclo de design para orientar, facilitar e aprimorar o aprendizado como um todo. Nesse processo, o objetivo não é apenas transformar dados em conhecimento, mas também filtrar o conhecimento minerado para a tomada de decisões, como destacado por Baker (2000). Como pode-se identificar na Figura 26, educadores e designers educacionais projetam, planejam, constroem e mantêm sistemas educacionais. Os alunos usam esses sistemas educacionais para aprender. Com base nas informações disponíveis sobre cursos, alunos, uso e interação, as técnicas de MDE podem ser aplicadas para descobrir conhecimentos úteis que ajudem a melhorar projetos educacionais. O conhecimento descoberto pode ser usado não apenas por designers educacionais e professores, mas também por usuários finais – sobretudo os alunos. Assim, a aplicação da mineração de dados em

sistemas educacionais pode ser orientada para apoiar as necessidades específicas de cada uma dessas categorias de interessados.

Figura 26 - Aplicação da MDE ao projeto de sistemas educacionais



Fonte: Adaptado Romero *et al.* (2010)

Em suma, conforme o primeiro Manual de Mineração de Dados Educacionais, pode-se dizer que as principais aplicações da MDE são: Comunicação com as partes interessadas; Manter e melhorar os cursos; Geração de recomendações; Previsão de desempenho dos alunos; Modelagem de alunos; e Análise da estrutura do domínio. Estas são detalhadas no Quadro 11.

Quadro 11 – Principais Aplicações da MDE

APLICAÇÕES	OBJETIVOS
Comunicação com as partes interessadas	Ajudar administradores de cursos e educadores a analisar as atividades dos alunos e as informações disponíveis nos cursos. As técnicas mais utilizadas para este tipo de objetivo são a análise exploratória de dados por meio de análises estatísticas e visualizações ou relatórios e mineração de processos.
Manter e melhorar os cursos	Apoiar os administradores de cursos e educadores a determinar como melhorar os cursos (conteúdos, atividades, links, etc.), usando informações especialmente sobre a utilização e aprendizagem dos alunos. As técnicas usadas com mais frequência para esse tipo de objetivo são algoritmos de Aprendizagem de Máquina com associação, agrupamento e classificação.

APLICAÇÕES	OBJETIVOS
Geração de recomendações	Recomendar aos alunos qual conteúdo (ou tarefas ou links) é mais adequado para eles no momento. As técnicas usadas com mais frequência para esse tipo de objetivo são algoritmos de Aprendizagem de Máquina para associação, sequenciamento, classificação e agrupamento.
Previsão de desempenho dos alunos	Prever as notas finais de um aluno ou outros tipos de resultados de aprendizagem (como retenção em um programa de graduação ou capacidade futura de aprender), com base nos dados das atividades do curso. As técnicas usadas com mais frequência para esse tipo de objetivo são algoritmos de Aprendizagem de Máquina para classificação, agrupamento e associação.
Modelagem de alunos	A modelagem de usuário no domínio educacional, como análise do perfil de alunos tem uma série de aplicações, incluindo, por exemplo, a detecção (muitas vezes em tempo real) de estados do aluno e características como satisfação, motivação, progresso de aprendizagem ou certos tipos de problemas que impactam negativamente seus resultados de aprendizagem (realizando muitos erros em atividades, mau uso ou subutilização da ajuda, enganando o sistema – trapaças, explorando recursos de aprendizagem de maneira ineficiente, etc.), detecção de sentimentos, estilos de aprendizagem e preferências. O objetivo comum é criar um modelo de aluno, uma descrição de perfil a partir de informações de uso. As técnicas frequentemente usadas para este tipo de objetivo são algoritmos de Aprendizagem de Máquina para agrupamento, todavia também podem ser aplicadas análises estatísticas e redes Bayesianas (incluindo o rastreamento de conhecimento bayesiano), modelos psicométricos e aprendizagem por reforço.
Análise da estrutura do domínio	Determinar a estrutura do domínio, usando a capacidade de prever o desempenho do aluno como uma medida da qualidade de um modelo de estrutura do domínio. O desempenho em testes ou em um ambiente de aprendizagem é utilizado para esse objetivo. As técnicas usadas com mais frequência nesse contexto são algoritmos de Aprendizagem de Máquina para regras de associação, métodos de agrupamento.

Fonte: Romero *et al.* (2010)

Destaca-se que com a evolução tecnológica e educacional, impulsionada pela emergência do ensino remoto, devido a pandemia do novo Corona Vírus, a MDE, que já era significativa desde sua origem, se

tornou uma estratégia ainda mais importante para análise do processo de ensino e aprendizagem. Como visto essa pode ser aplicada nos mais diferentes contextos e com os mais diversos objetivos para dar suporte a melhoria dos ambientes de aprendizagem – que cada vez se tornam mais digitais – muito embora a MDE também seja uma estratégia válida para âmbito de ensino tradicional (como demostrado no experimento exposto na seção 7.2).

Nesse sentido, em especial pela temática abordada, este livro é uma contribuição relevante, sobretudo para pesquisadores iniciantes na área de MDE, esta obra proporcionou um guia de fácil leitura destacando: as definições da MDE e como ocorreu sua evolução e consolidação como área de pesquisa; como é o seu processo de aplicação; as principais temáticas em que esta estratégia pode ser aplicada e como podem ser desenvolvidas, bem como oportunidades de pesquisa sobre a MDE; as principais diferenças ente MDE e Análise de Aprendizagem; suas principais técnicas que incorporam a Aprendizagem de Máquina e a Aprendizagem Profunda; ademais apresentou dois exemplos bastantes didáticos da aplicação do processo de MDE que podem ser reproduzidos por interessados em aprender mais sobre este assunto, bem como servir de subsídios para a elaboração de projetos próprios.

Em conclusão, este livro configura-se como uma fonte de diversas pesquisas na área de MDE, bem como traz algumas contribuições originais da autora como o mapeamento sistemático que buscou identificar as principais temática de pesquisa em MDE no contexto do *e-leaning*, mapeando o estado da arte entre 2015 e 2019. Ademais, ao identificar em especial uma lacuna de estudos que detalhassem a aplicação do processo de MDE de forma minuciosa, para servir de base para pesquisadores iniciantes, trouxe dois exemplos para este fim, que utilizaram dois tipos diferentes de abordagem de resolução, com Aprendizagem de Máquina Supervisionada e não Supervisionada, aplicando também a Aprendizagem Profunda, técnica ainda não consolidada no contexto da MDE. Assim, demostra-se o potencial deste compilado no avanço de pesquisas na Educação e na Ciência de Dados.

REFERÊNCIAS

AGGARWAL, Charu C. **Data Mining: The Textbook**. 1. ed. New York, USA: Springer, 2015. v. 1. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-14142-8.

AGGARWAL, Charu C. **Neural Networks and Deep Learning: A Textbook**. 1. ed. New York, USA: Springer, 2018. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-94463-0.

ALDOWAH, Hanan; AL-SAMARRAIE, Hosam; FAUZY, Wan Mohamad. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. **Telematics and Informatics**, *[S. l.]*, v. 37, p. 13–49, 2019. Disponível em: https://doi.org/10.1016/j.tele.2019.01.007.

ALPAYDIN, Ethem. **Introduction to Machine Learning**. 2. ed. Cambridge, Massachusetts: *[s. n.]*, 2010. *E-book*. Disponível em: https://doi.org/10.1007/978-1-62703-748-8_7.

ALQUDAH, Nour; YASEEN, Qussai. ScienceDirect ScienceDirect Machine Learning for Traffic Analysis : A Review Machine Learning for Traffic Analysis : A Review. **Procedia Computer Science**, *[S. l.]*, v. 170, p. 911–916, 2020. Disponível em: https://doi.org/10.1016/j.procs.2020.03.111.

ALRAIMI, K. M.; ZO, H.; CIGANEK, A. P. Understanding the MOOCs continuance: The role of openness and reputation. **Computers & Education**, v. 80, p. 28-38, 2015. Disponível em: https://doi.org/10.1016/j.compedu.2014.08.006.

ALVIM, Paulo. **Open Source com jCompany© Developer Suite.** 3a Ed. ed. Belo Horizonte: *[s. n.]*, 2010. *E-book*.

BADAR, Maryam; HARIS, Muhammad; FATIMA, Anam. Application of deep learning for retinal image analysis: A review. **Computer Science Review**, *[S. l.]*, v. 35, p. 1–18, 2020. Disponível em: https://doi.org/10.1016/j.cosrev.2019.100203.

BAHRAMPOUR, Soheil *et al.* Comparative Study of Deep Learning Software Frameworks. **Cornell Univeristy**, *[S. l.]*, v. 3, p. 1–9, 2015.

BAKER, Michael J. The roles of models in Artificial Intelligence and Education research : a prospective view. **Journal of Artificial Intelligence and Education**, *[S. l.]*, v. 11, p. 122–143, 2000.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 19, n. 02, p. 3–13, 2011. Disponível em: https://doi.org/10.5753/rbie.2011.19.02.03.

BAKER, Ryan S. J. D. **Big data and education**. 2. ed. New York, USA: A Massive Online Open Textbook (MOOT) - Teachers College, Columbia University, 2015.

BAKER, Ryan S. J. D.; YACEF, Kalina. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, *[S. l.]*, v. 1, n. 1, p. 3–17, 2009. Disponível em: https://doi.org/10.5281/zenodo.3554657.

BAKER, Ryan Shaun; INVENTADO, Paul Salvador. Educational Data Mining and Learning Analytics. *In*: J.A. LARUSSON AND B. WHITE (EDS.) (org.). **Learning Analytics: From Research to Practice**. 1. ed. New York, USA: Springer, 2014. p. 1–195. *E-book*. Disponível em: https://doi.org/10.1007/978-1-4614-3305-7.

BAKHSHINATEGH, Behdad *et al.* Educational data mining applications and tasks: A survey of the last 10 years. **Education and Information Technologies**, *[S. l.]*, v. 23, n. 1, p. 537–553, 2018. Disponível em: https://doi.org/10.1007/s10639-017-9616-z.

BISHOP, C. M. **Neural networks for pattern recognition**. 1. ed. EUA: *[s. n.]*, 1995. *E-book*.

BISHOP, Christopher M.; PATTERN. **Pattern Recognition and Machine Learning**. 1. ed. Nova York, USA: Springer, 2006. *E-book*.

BOULEMTAFES, Amine; DERHAB, Abdelouahid; CHALLAL, Yacine. A review of privacy-preserving techniques for deep learning. **Neurocomputing**, *[S. l.]*, v. 384, p. 21–45, 2020. Disponível em: https://doi.org/10.1016/j.neucom.2019.11.041.

CHUI, Kwok Tai *et al.* Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. **Computers in Human Behavior**, *[S. l.]*, v. 107, n. December 2017, p. 105584, 2020. Disponível em: https://doi.org/10.1016/j.chb.2018.06.032.

CORTEZ, P.; SILVA, A. Using Data Mining to Predict Secondary School Student Performance. **In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008),** p. 1-8, 2008.

DE LOS REYES, Daniel A. Guimarães *et al.* Predição de sucesso acadêmico de estudantes: uma análise sobre a demanda por uma abordagem baseada em transfer learning. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 27, n. 1, p. 1–25, 2019. Disponível em: https://doi.org/10.5753/rbie.2019.27.01.01.

DUDA, R. O.; HART, P. E.; STROK, D. G. **Pattern classification**. 1. ed. Wiley: *[s. n.]*, 2001. *E-book*.

EDM. **Educational Data Mining**. *[s. l.]*, 2020. Disponível em: http://educationaldatamining.org/. Acesso em: 31 maio. 2020.

FERREIRA-SATLER, Mateus et al. Fuzzy ontologies-based user profiles applied to enhance e-learning activities. **Soft Computing**, *[S. l.]*, v. 16, n. 7, p. 1129–1141, 2012. Disponível em: https://doi.org/10.1007/s00500-011-0788-y.

GALLEN, Rosa Cabedo; CARO, Edmundo Tovar. An exploratory analysis of why a person enrolls in a Massive Open Online Course within MOOCKnowledge data

collection. *In*: 2017, Athens, Greece. **Global Engineering Education Conference, (EDUCON)**. Athens, Greece: IEEE, 2017. p. 1600–1605. Disponível em: https://doi.org/10.1109/EDUCON.2017.7943062.

GAO, Lina *et al.* Modeling the effort and learning ability of students in MOOCs. **IEEE Access**, *[S. l.]*, v. 7, p. 128035–128042, 2019. Disponível em: https://doi.org/10.1109/ACCESS.2019.2937985.

GILL, by Philip E.; MURRAY, Walter; WRIGHT, Margaret H. **Practical Optimization**. 1. ed. *[S. l.: s. n.]*. *E-book*.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. Cambridge, MA, USA: *[s. n.]*, 2016. *E-book*.

GUO, Shou Xi *et al.* Attention-Based Character-Word Hybrid Neural Networks With Semantic and Structural Information for Identifying of Urgent Posts in MOOC Discussion Forums. **IEEE Access**, *[S. l.]*, v. 7, p. 120522–120532, 2019. Disponível em: https://doi.org/10.1109/ACCESS.2019.2929211.

HAND, David J. **Construction and Assessment of Classification Rules**. 1. ed. New York:: *[s. n.]*, 1997. *E-book*.

HARTIGAN, J. A; WONG, M. A. Algorithm AS 136: A K-means clustering algorithm. **Applied Statistics**, v. 28, p. 100–108. Disponível em: https://doi.org/10.2307/2346830.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. California, USA: Springer, 2009. *E-book*.

HEGDE, Jeevith; ROKSETH, Børge. Applications of machine learning methods for engineering risk assessment – A review. **Safety Science**, *[S. l.]*, v. 122, n. September 2019, p. 1–16, 2020. Disponível em: https://doi.org/10.1016/j.ssci.2019.09.015.

HINTON, Geoffrey E.; OSINDERO, Simon; TEH, Yee-Whye. A Fast Learning Algorithm for Deep Belief Nets. **Neural Computation**, *[S. l.]*, v. 18, p. 1527–1554, 2006. Disponível em: https://doi.org/10.7763/ijesd.2010.v1.67.

IGUAL, Laura; SEGUÍ, Santi. **Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications**. 1. ed. *[S. l.]*: Springer, 2017. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-50017-1.

JANG, J. S. R.; SUN, C. T.; MIZUTANI, E. Neuro-Fuzzy and Soft Computing-A Computational Approach to Learning and Machine Intelligence [Book Review]. **IEEE Transactions on Automatic Control**, *[S. l.]*, v. 42, n. 10, p. 1482–1484, 2005. Disponível em: https://doi.org/10.1109/tac.1997.633847.

JAPKOWICZ, Nathalie; SHAH, Mohak. **Evaluating Learning Algorithms: A Classification Perspective**. 1a Ed. ed. Cambridge: *[s. n.]*, 2014. *E-book*.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007. Disponível em: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.471&rep=rep1&type=pdf>. Acesso em: 13 abr. 2021.

KOVALEV, Vassili; KALINOVSKY, Alexander; KOVALEV, Sergey. Deep Learning with Theano, Torch, Caffe, TensorFlow, and Deeplearning4J: Which One Is the Best in Speed and Accuracy? **Proceedings of the 13th International Conference on Pattern Recognition and Information Processing (PRIP 2016)**, *[S. l.]*, p. 99–103, 2016.

KUBAT, Miroslav. **An Introduction to Machine Learning**. 2. ed. Coral Gables, FL, USA: Springer, 2017. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-63913-0

LANDIS, J. Richard; KOCH, Gary G. This content downloaded from 185.2.32.58 on Tue. *[S. l.]*, v. 33, n. 2, p. 363–374, 1977.

LANG, Charles *et al.* **Handbook of Learning Analytics**. 1. ed. *[S. l.]*: SOLAR - SOCIETY for LEARNING ANALYTICS RESEARCH, 2017. *E-book*. Disponível em: https://doi.org/10.18608/hla17.

LE, Quan; TORRISI, Mirko; POLLASTRI, Gianluca. Deep learning methods in protein structure prediction. **Computational and Structural Biotechnology Journal**, *[S. l.]*, v. 426, n. January, p. 1–10, 2020. Disponível em: https://doi.org/10.1016/j.csbj.2019.12.011.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, *[S. l.]*, v. 521, n. 7553, p. 436–444, 2015. Disponível em: https://doi.org/10.1038/nature14539.

LEI, Yaguo *et al.* Applications of machine learning to machine fault diagnosis: A review and roadmap. **Mechanical Systems and Signal Processing**, *[S. l.]*, v. 138, p. 1–39, 2020. Disponível em: https://doi.org/10.1016/j.ymssp. 2019.106587.

LIN, Jinjiao *et al.* Automatic Knowledge Discovery in Lecturing Videos via Deep Representation. **IEEE Access**, *[S. l.]*, v. 7, p. 33957–33963, 2019. Disponível em: https://doi.org/10.1109/ACCESS.2019.2904046.

LIÑÁN, Laura Calvet; PÉREZ, Juan Ángel Alejandro. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. **International Journal of Educational Technology in Higher Education**, *[S. l.]*, v. 12, n. 3, p. 98–112, 2015. Disponível em: https://doi.org/10.7238/rusc.v12i3.2515.

LIU, N. T.; SALINAS, J. Machine Learning for Predicting Outcomes in Trauma. **SHOCK**, *[S. l.]*, v. 48, n. 5, p. 504–510, 2017.

MARECA, Pilar; BORDEL, Borja. Students Profiles and their Behavior in MOOC Platforms, MIRIADAX Platform. *In:* 2019, Coimbra, Portugal. **14th Iberian Conference on Information Systems and Technologies (CISTI)**. Coimbra, Portugal: [s. n.], 2019. p. 1–6. Disponível em: https://doi.org/10.23919/CISTI.2019.8760696.

MITCHELL, Tom M. **Machine Learning**. 1. ed. Nova York, USA: McGraw-Hill Science/Engineering/Math, 1997. *E-book*.

MOISSA, Barbara; GASPARINI, Isabela; KEMCZINSKI, Avanilde. Educational Data Mining versus Learning Analytics: estamos reinventando a roda? Um mapeamento sistemático. *In*: 2015, Maceió, Alagoas. **XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)**. Maceió, Alagoas: *[s. n.]*, 2015. p. 1167–1176. Disponível em: https://doi.org/10.5753/cbie.sbie.2015.1167.

MURAT, Fatma *et al.* Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. **Computers in Biology and Medicine**, *[S. l.]*, v. 120, n. April, p. 1–14, 2020. Disponível em: https://doi.org/10.1016/j.compbiomed.2020.103726.

NAVARRO, Pedro J. *et al.* A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3D range data. **Sensors (Switzerland)**, *[S. l.]*, v. 17, n. 18, p. 1–20, 2017. Disponível em: https://doi.org/10.3390/s17010018.

NEN-FU, Huang *et al.* The Clustering Analysis System Based on Students' Motivation and Learning Behavior. *In*: 2018, **Proceedings of 2018 Learning With MOOCS, LWMOOCS 2018**. : IEEE, 2018. p. 117–119. Disponível em: https://doi.org/10.1109/LWMOOCS.2018.8534611.

NG, S. S. Y. *et al.* An independent study of two deep learning platforms - H2O and SINGA. *In*: 2016, Bali, Indonesia. **International Conference on Industrial Engineering and Engineering Management (IEEM 2016)**. Bali, Indonesia: IEEE, 2016. p. 1279–1283. Disponível em: https://doi.org/10.1109/IEEM.2016.7798084.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. **New England Journal of Medicine**, *[S. l.]*, v. 375, n. 13, p. 1216–1219, 2016.

POLAK, E. **Computational methods in optimization.** 2. ed. Londris: *[s. n.]*, 1971. *E-book*.

PURSEL, B. K. *et al.* Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. **Journal of Computer Assisted Learning**, *[S. l.]*, v. 32, n. 3, p. 202–217, 2016. Disponível em: https://doi.org/10.1111/jcal.12131.

RANA, Pratip *et al.* Recent advances on constraint-based models by integrating machine learning. **Current Opinion in Biotechnology**, *[S. l.]*, v. 64, p. 85–91, 2020. Disponível em: https://doi.org/10.1016/j.copbio.2019.11.007.

RIGO, Sandro José *et al.* Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 22, n. 01, p. 168–177, 2014. Disponível em: https://doi.org/10.5753/RBIE.2014.22.01.132.

RIPLEY, B. D. **Pattern recognition and neural networks**. 1. ed. Cambridge: *[s. n.]*, 1996. *E-book*.

RODRIGUES, Rodrigo Lins *et al.* Discovery engagement patterns MOOCs through cluster analysis. **IEEE Latin America Transactions**, *[S. l.]*, v. 14, n. 9, p. 4129–4135, 2016. Disponível em: https://doi.org/10.1109/TLA.2016.7785943.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, *[S. l.]*, v. 33, n. 1, p. 135–146, 2007. Disponível em: https://doi.org/10.1016/j.eswa.2006.04.005.

ROMERO, Cristbal; VENTURA, Sebastin. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews**, *[S. l.]*, v. 40, n. 6, p. 601–618, 2010. Disponível em: https://doi.org/10.1109/TSMCC.2010.2053532.

ROMERO, Cristóbal *et al.* **Handbook of Educational Data Mining**. 1. ed. Boca Raton, USA: CRC Press - Taylor & Francis, 2010. *E-book*. Disponível em: https://doi.org/10.1201/b10274.

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, *[S. l.]*, v. 3, n. 1, p. 12–27, 2013. Disponível em: https://doi.org/10.1002/widm.1075.

ROMERO, Cristobal; VENTURA, Sebastian. Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, *[S. l.]*, v. 10, n. 3, p. 1–21, 2020. Disponível em: https://doi.org/10.1002/widm.1355.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data science in massive open online courses. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, *[S. l.]*, v. 7, n. 1, 2017. Disponível em: https://doi.org/10.1002/widm.1187.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, *[S. l.]*, v. 3, n. 3, p. 210–229, 1959. Disponível em: https://doi.org/10.1147/rd.33.0210.

SHI, Yuling; PENG, Zhiyong; WANG, Hongning. Modeling student learning styles in MOOCs. *In:* 2017, Singapura. **International Conference on Information and Knowledge Management**. Singapura: [s. n.], 2017. p. 979–988. Disponível em: https://doi.org/10.1145/3132847.3132965.

SCHMIDHUBER, Jürgen. Deep Learning in neural networks: An overview. **Neural Networks**, *[S. l.]*, v. 61, p. 85–117, 2015. Disponível em: https://doi.org/10.1016/j.neunet.2014.09.003.

SCHWENDIMANN, Beat A. *et al.* Perceiving learning at a glance: A systematic literature review of learning dashboard research. **IEEE Transactions on Learning Technologies**, *[S. l.]*, v. 10, n. 1, p. 30–41, 2017. Disponível em: https://doi.org/10.1109/TLT.2016.2599522.

SENDERS, Joeky T. *et al.* Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. **World Neurosurgery**, *[S. l.]*, v. 109, n. 1, p. 476- 486., 2018. Disponível em: https://doi.org/10.1016/j.wneu.2017.09.149.

SENGUPTA, Sourya *et al.* Ophthalmic diagnosis using deep learning with fundus images – A critical review. **Artificial Intelligence in Medicine**, *[S. l.]*, v. 102, p. 1–36, 2020. Disponível em: https://doi.org/10.1016/j.artmed.2019.101758.

SEZER, Omer Berat; GUDELEK, Mehmet Ugur; OZBAYOGLU, Ahmet Murat. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. **Applied Soft Computing Journal**, *[S. l.]*, v. 90, p. 1–65, 2020. Disponível em: https://doi.org/10.1016/j.asoc.2020.106181.

SHAH, D. By the Numbers: MOOCS in 2018 Class Central (2018). Disponível em: <https://www.classcentral.com/report/mooc-stats-2018/>. Acesso em: 15 abr. 2021.

SHAHIRI, Amirah Mohamed; HUSAIN, Wahidah; RASHID, Nur'Aini Abdul. A Review on Predicting Student's Performance Using Data Mining Techniques. **Procedia Computer Science**, *[S. l.]*, v. 72, p. 414–422, 2015. Disponível em: https://doi.org/10.1016/j.procs.2015.12.157.

SHIOTANI, Shigetoshi; FUKUDA, Toshio; SHIBATA, Takanori. A neural network architecture for incremental learning. **Neurocomputing**, *[S. l.]*, v. 9, n. 2, p. 111–130, 1995. Disponível em: https://doi.org/10.1016/0925-2312(94)00061-V.

SIEMENS, George; BAKER, Ryan S. J. d. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. *In*: 2012, Vancouver, Canada. **2nd International Conference on Learning Analytics and Knowledge (LAK 2012)**. Vancouver, Canada: ACM, 2012. p. 252–254. Disponível em: https://doi.org/10.1145/2330601.2330661.

SOFFER, Shelly *et al.* Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. **Radiology**, *[S. l.]*, v. 290, n. 3, p. 590–606, 2019. Disponível em: https://doi.org/10.1148/radiol.2018180547.

SOUZA, Vanessa Faria de; SANTOS, Tony Carlos Bignardi dos . Research Topics on Educational Data Mining in MOOCS. **International Journal for Innovation Education and Research**, v. 8, p. 311-320, 2020. Disponível em: https://doi.org/10.31686/IJIER.VOL8.ISS7.2481.

SUKHIJA, Karan; JINDAL, Manish; AGGARWAL, Naveen. The recent state of educational data mining: A survey and future visions. *In*: 2015, Amritsar, India. **3rd International Conference on MOOCs, Innovation and Technology in Education (MITE)**. Amritsar, India: IEEE, 2015. p. 354–359. Disponível em: https://doi.org/10.1109/MITE.2015.7375344.

TAN, Yueying et al. Learning Profiles, Behaviors and Outcomes: Investigating International Students' Learning Experience in an English MOOC. **In: Proceedings International**

Symposium on Educational Technology (ISET 2018): IEEE, p. 214–218. Disponível em: https://doi.org/10.1109/ISET.2018.00055.

WAHEED, Hajra *et al.* Predicting academic performance of students from VLE big data using deep learning models. **Computers in Human Behavior**, *[S. l.]*, v. 104, p. 1–13, 2020. Disponível em: https://doi.org/10.1016/j.chb.2019.106189.

WARING, Jonathan; LINDVALL, Charlotta; UMETON, Renato. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. **Artificial Intelligence in Medicine**, *[S. l.]*, v. 104, p. 1–42, 2020. Disponível em: https://doi.org/10.1016/j.artmed.2020.101822.

WEIS, C. V. ..; JUTZELER, C. R. ..; BORGWARDT, K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. **Clinical Microbiology and Infection**, *[S. l.]*, 2020. Disponível em: https://doi.org/10.1016/j.cmi.2020.03.014.

WEN, Yimin *et al.* Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. **Tsinghua Science and Technology**, *[S. l.]*, v. 25, n. 3, p. 336–347, 2020. Disponível em: https://doi.org/10.26599/TST.2019.9010013.

XIN, Yang *et al.* Machine Learning and Deep Learning Methods for Cybersecurity. **IEEE Access**, *[S. l.]*, v. 20, p. 1–9, 2018. Disponível em: https://doi.org/10.1109/ACCESS.2018.2836950.

YANG, Jian; ZHANG, Xiao Ling; SU, Peng. Deep-Learning-Based Agile Teaching Framework of Software Development Courses in Computer Science Education. **Procedia Computer Science**, *[S. l.]*, v. 154, p. 137–145, 2018. Disponível em: https://doi.org/10.1016/j.procs.2019.06.021.

ZHANG, Yaling; WU, Bei. Research and application of grade prediction model based on decision tree algorithm. *In*: 2019, Chengdu, China. **Turing Celebration Conference (ACM TURC 2019)**. Chengdu, China: ACM, 2019. p. 1–6. Disponível em: https://doi.org/10.1145/3321408.3322857.

ZHAO, Rui *et al.* Deep learning and its applications to machine health monitoring. **Mechanical Systems and Signal Processing**, *[S. l.]*, v. 115, p. 213–237, 2019. Disponível em: https://doi.org/10.1016/j.ymssp. 2018.05.050.

ZHOU, Yuekuan; ZHENG, Siqian; ZHANG, Guoqiang. A review on cooling performance enhancement for phase change materials integrated systems—flexible design and smart control with machine learning applications. **Building and Environment**, *[S. l.]*, v. 174, p. 1–41, 2020. Disponível em: https://doi.org/10.1016/j.buildenv.2020.106786.

APÊNDICE A

PUBLICAÇÕES DA AUTORA SOBRE MINERAÇÃO DE DADOS EDUCACIONAIS

FARIA DE SOUZA, VANESSA. Mineração de dados educacionais com aprendizagem de máquina. **Revista Educar Mais**, v. 5, p. 766 - 787, 2021. Disponível em: https://doi.org/10.15536/reducarmais.5.2021.2417.

SOUZA, Vanessa Faria; SANTOS, TONY CARLOS BIGNARDI DOS. Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda. **Revista Brasileira de Informática na Educação (RBIE)**, v.29, p. 519 - 546, 2021. Disponível em: http://dx.doi.org/10.5753/rbie.2021.29.0.519.

SOUZA, Vanessa Faria de; PERRY, Gabriela Trindade. Proposta de melhoria dos dados de relatórios de uma plataforma de MOOCS brasileira. **Revista Brasileira de Aprendizagem Aberta e a Distância**, v. 19, p. 1-18, 2020. Disponível em: http://seer.abed.net.br/index.php/RBAAD/article/view/424.

SOUZA, Vanessa Faria; PERRY, Gabriela Trindade . Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático. **Revista Brasileira de Informática na Educação (RBIE)**, v. 28, p. 491-508, 2020. Disponível em: https://br-ie.org/pub/index.php/rbie/article/view/v28p491.

FARIA DE SOUZA, VANESSA; SANTOS, TONY CARLOS BIGNARDI DOS. Research Topics on Educational Data Mining in MOOCS. **International Journal for Innovation Education and Research**, v. 8, p. 311-320, 2020. Disponível em: https://ijier.net/ijier/article/view/2481.

SOUZA, Vanessa Faria. Mineração de dados educacionais em um MOOC brasileiro. **EAD & Tecnologias Digitais na Educação**, v. 8, p. 62-78, 2020. Disponível em: https://ojs.ufgd.edu.br/index.php/ead/article/view/11461.

DE SOUZA, VANESSA FARIA; PERRY, GABRIELA . Identifying student behavior in MOOCs using Machine Learning. **International Journal for Innovation Education and Research**, v. 7, p. 30-39, 2019. Disponível em: https://ijier.net/ijier/article/view/1318.

SOUZA, Vanessa Faria; PERRY, G. T. Mineração de Texto em Moocs: Análise da Relevância Temática de Postagens em Fóruns de Discussão. RENOTE. **Revista Novas Tecnologias na Educação**, v. 17, p. 204-213, 2019. Disponível em: https://seer.ufrgs.br/renote/article/view/99471.

APÊNDICE B

SÍNTESE DE TODAS AS MÉTRICAS GERADAS PELA BIBLIOTECA "CARET" DO R DO SEGUNDO EXPERIMENTO

1. NAÏVE BAYES

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	17	1	0	0
B	9	49	13	10
C	0	10	30	9
D	0	0	0	8

Overall Statistics

Accuracy	:	0.6667
95% CI	:	(0.5868, 0.74)
No Information Rate	:	0.3846
P-Value [Acc > NIR]	:	9.791e-13
Kappa	:	0.5138
Mcnemar's Test P-Value	:	NA Statistics by Class:

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.6538	0.8167	0.6977	0.29630
Specificity	0.9923	0.6667	0.8319	1.00000
Pos Pred Value	0.9444	0.6049	0.6122	1.00000
Neg Pred Value	0.9348	0.8533	0.8785	0.87162
Prevalence	0.1667	0.3846	0.2756	0.17308
Detection Rate	0.1090	0.3141	0.1923	0.05128
Detection Prevalence	0.1154	0.5192	0.3141	0.05128
Balanced Accuracy	0.8231	0.7417	0.7648	0.64815

2. ÁRVORE DE DECISÃO

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	14	4	0	0
B	0	72	9	0
C	0	2	47	0
D	0	0	4	4

Overall Statistics

Accuracy	:	0.8782
95% CI	:	(0.8164, 0.9251)
No Information Rate	:	0.5
P-Value [Acc > NIR]	:	< 2.2e-16
Kappa	:	0.7996
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	1.00000	0.9231	0.7833	1.00000
Specificity	0.97183	0.8846	0.9792	0.97368
Pos Pred Value	0.77778	0.8889	0.9592	0.50000
Neg Pred Value	1.00000	0.9200	0.8785	1.00000
Prevalence	0.08974	0.5000	0.3846	0.02564
Detection Rate	0.08974	0.4615	0.3013	0.02564
Detection Prevalence	0.11538	0.5192	0.3141	0.05128
Balanced Accuracy	0.98592	0.9038	0.8812	0.98684

3. RANDOM FOREST

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	12	6	0	0
B	0	71	10	0
C	0	3	46	0
D	0	0	6	2

Overall Statistics

Accuracy	:	0.8397
95% CI	:	(0.7726, 0.8935)
No Information Rate	:	0.5128
P-Value [Acc > NIR]	:	< 2.2e-16
Kappa	:	0.7326
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	1.00000	0.8875	0.7419	1.00000
Specificity	0.95833	0.8684	0.9681	0.96104
Pos Pred Value	0.66667	0.8765	0.9388	0.25000
Neg Pred Value	1.00000	0.8800	0.8505	1.00000
Prevalence	0.07692	0.5128	0.3974	0.01282
Detection Rate	0.07692	0.4551	0.2949	0.01282
Detection Prevalence	0.11538	0.5192	0.3141	0.05128
Balanced Accuracy	0.97917	0.8780	0.8550	0.98052

4. SUPORT VECTOR MACHINES

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	13	5	0	0
B	3	70	8	0
C	0	6	42	1
D	0	0	4	4

Overall Statistics

Accuracy	:	0.8269
95% CI	:	(0.7583, 0.8827)
No Information Rate	:	0.5192
P-Value [Acc > NIR]	:	8.754e-16
Kappa	:	0.7154
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.81250	0.8642	0.7778	0.80000
Specificity	0.96429	0.8533	0.9314	0.97351
Pos Pred Value	0.72222	0.8642	0.8571	0.50000
Neg Pred Value	0.97826	0.8533	0.8879	0.99324
Prevalence	0.10256	0.5192	0.3462	0.03205
Detection Rate	0.08333	0.4487	0.2692	0.02564
Detection Prevalence	0.11538	0.5192	0.3141	0.05128
Balanced Accuracy	0.88839	0.8588	0.8546	0.88675

5. REDES NEURAIS ARTIFICIAI MULTICAMADAS

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	18	0	0	0
B	3	76	2	0
C	0	3	46	0
D	0	0	0	8

Overall Statistics

Accuracy	:	0,948
95% CI	:	(0.7943, 0.9495)
No Information Rate	:	0.5321
P-Value [Acc > NIR]	:	< 2.2e-16
Kappa	:	0.7984
Mcnemar's Test P-Value	:	NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.81250	0.8554	0.8600	1.00000
Specificity	0.96429	0.8630	0.9434	0.99329
Pos Pred Value	0.72222	0.8765	0.8776	0.87500
Neg Pred Value	0.97826	0.8400	0.9346	1.00000
Prevalence	0.10256	0.5321	0.3205	0.04487
Detection Rate	0.08333	0.4551	0.2756	0.04487
Detection Prevalence	0.11538	0.5192	0.3141	0.05128
Balanced Accuracy	0.88839	0.8592	0.9017	0.99664

SOBRE A AUTORA

VANESSA FARIA DE SOUZA

Docente dedicação exclusiva do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS), *Campus* Ibirubá. Atua nos Cursos de Graduação em Ciência da Computação, Técnico em Informática Integrado do Ensino Médio, Licenciatura em Matemática e Especialização em Ensino de Linguagens e suas Tecnologias. Doutorado em andamento no Programa de Pós-Graduação em Informática na Educação (PPGIE) da Universidade Federal do Rio Grande do Sul (UFRGS). Mestre em Informática pelo Programa de Pós-Graduação em Informática (PPGI) da Universidade Tecnológica Federal do Paraná (UTFPR), na subárea de Computação Aplicada com ênfase em Engenharia de Software. Especialista em Educação Especial Inclusiva, com ênfase em Tecnologia Assistiva pela Universidade Estadual do Norte do Paraná (UENP). Graduada em Sistemas de Informação pela UENP com Bacharelado em Sistemas de Informação e Licenciatura em Computação. Licenciada em Matemática pela UTFPR.

CV: http://lattes.cnpq.br/0052710619133749

ÍNDICE REMISSIVO

Este livro foi composto pela Editora Bagai.

www.editorabagai.com.br /editorabagai

/editorabagai contato@editorabagai.com.br