UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LAERTES PEREIRA JUNIOR

ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA: UM ESTUDO DE CASO NA ÁREA EDUCACIONAL

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA

2020

LAERTES PEREIRA JUNIOR

ANÁLSE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA: UM ESTUDO DE CASO NA ÁREA EDUCACIONAL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação do Departamento Acadêmico de Informática, da Universidade Tecnológica Federal do Paraná.

Orientadora: Profa. Dra. Simone Nasser Matos

Coorientadora: Profa. Dra. Helyane Bronoski Borges

PONTA GROSSA

2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Ponta Grossa

Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Bacharelado em Ciência da Computação

TERMO DE APROVAÇÃO

ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA: UM ESTUDO DE CASO NA ÁREA EDUCACIONAL

por

LAERTES PEREIRA JUNIOR

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 24 de setembro de 2020 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O aluno Laertes Pereira Junior foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof(a). Dra. Simone Nasser Matos
Orientador(a)

_____
Prof(a). Dra. Helyane Bronoski Borges
Coorientadora

_____
Prof(a). Dra. Eliana Cláudia Mayumi Ishikawa
Membro titular

_____
Prof. MSc. Geraldo Ranthum
Membro titular

_____
Prof. MSc. Geraldo Ranthum
Responsável pelo Trabalho de Conclusão de Curso

_____
Prof(a). Dra. Mauren Louise Sguario
Coordenador do curso

AGRADECIMENTOS

Agradeço a Deus pela saúde, força e oportunidades que me ajudaram a superar todos os momentos difíceis.

Aos meus pais por todo incentivo e dedicação comigo durante toda esta jornada, sem o apoio deles com certeza nada disso seria possível.

À minha orientadora profa. Dra. Simone Nasser Matos e coorientadora profa. Dra. Helyane Bronoski Borges pela excepcional orientação e ajuda, por todas as dicas dadas, pelas reuniões que promoveram excelentes ideias e pela experiência transmitida por elas, muito obrigado.

Aos meus amigos pela ajuda e paciência em trabalhos, provas, grupos de estudo, etc., cada minuto gasto de aprendizado com vocês foi essencial.

Pelo financiamento desta pesquisa promovido pelo “Edital 02/2019 - PROGRAD / PROREC - Apoio à execução de trabalhos de conclusão de curso - TCC” da Universidade Tecnológica Federal do Paraná.

RESUMO

PEREIRA, L. **Análise comparativa de algoritmos de aprendizagem de máquina: um estudo de caso na área educacional**. 2020. 108 f. Trabalho de Conclusão de Curso - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2020.

Na área educacional a aplicação de algoritmos de máquina permite definir estratégias que ajudam os estudantes a melhorar seu desempenho para progredir no aprendizado, auxiliar professores e pesquisadores a descobrirem novas maneiras de se aprimorarem, prevê risco de evasão de alunos, avaliam o desempenho de estudantes atuando no ambiente educacional, entre outros. Este trabalho aplicou ferramentas de mineração de dados utilizando algoritmos de aprendizagem de máquina em uma base referente a opinião de alunos do ensino superior sobre o ensino remoto no período de pandemia. A etapa de mineração de dados teve como objetivo inferir regras e padrões que possam modelar o perfil dos estudantes em relação a metodologia EAD. Os resultados encontrados mostram características diferentes para grupos de alunos contendo variadas opiniões sobre a aplicação do ensino a distância. Os algoritmos mostraram resultados satisfatórios em relação a acurácia e foi possível analisar os perfis destes alunos por meio das regras definidas por eles.

Palavras-chave: Educação. Mineração de Dados. Aprendizagem de Máquina.

ABSTRACT

PEREIRA, L. **Comparative analysis of machine learning algorithms: a case study in educational field.** 2020. 108 p. Course Conclusion Work - Federal University of Technology - Paraná. Ponta Grossa, 2020.

In the educational area, the application of machine algorithms allows define strategies that help students to improve their performance to progress in learning, assist teachers and researchers to discover new ways to improve themselves, predict the risk of student evasion, evaluate the performance of students acting in the educational environment, among others. This work applied data mining tools using machine learning algorithms on a database referring to the opinion of higher education students about remote education in the pandemic period. The data mining step aimed to infer rules and patterns that can model students' profile in relation to distance learning methodology. The results found show different characteristics for groups of students with varying opinions on the application of distance learning. The algorithms showed satisfactory results in relation to accuracy and it was possible to analyze the profiles of these students through the rules defined by them.

Keywords: Education. Data Mining. Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE SIGLAS E ACRÔNIMOS

AM	Aprendizagem de Máquina
APPs	Aplicativos
ASSARTE	Associação Artesanal do Excepcional de Ponta Grossa
AVAs	Ambientes Virtuais de Aprendizagem
C0-k3	*cluster* 0 *k-means*-3
C1-k3	*cluster* 1 *k-means*-3
C2-k3	*cluster* 2 *k-means*-3
C0-k5	*cluster* 0 *k-means*-5
C1-k5	*cluster* 1 *k-means*-5
C2-k5	*cluster* 2 *k-means*-5
C3-k5	*cluster* 3 *k-means*-5
C4-k5	*cluster* 4 *k-means*-5
C0-SOM	*cluster* 0 SOM
C1-SOM	*cluster* 1 SOM
C2-SOM	*cluster* 2 SOM
C3-SOM	*cluster* 3 SOM
CTA	Capacidade Total de Acerto do Modelo
DI	Deficiência Intelectual
EAD	Educação a Distância
FP	*False Positives*
FN	*False Negatives*
IA	Inteligência Artificial
IAED	Inteligência Artificial na Educação
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IoT	*Internet of Things*
KDD	*Knowledge Discovery in Databases*
LA	*Learning Analytics*
LMSs	*Learning Management Systems*
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MOOCs	*Massive Open Online Courses*
PART	*Partial Decision Trees*
PLN	Processamento de Linguagem Natural
ROC	*Receiver Operating Characteristic*
SEs	*Softwares* Educacionais
STIs	Sistemas Tutores Inteligentes
SOM	*Sel-organizing maps*
TICs	Tecnologias da Informação e Comunicação
TN	*True Negatives*
TP	*True Positives*
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Estudos na área da informática vêm ajudando a humanidade em vários segmentos da sociedade ao longo dos anos, dentre estes, a aplicação em ambientes educacionais. Segundo Vygotsky (2005), a socialização e o ambiente computacional permitem mudanças no desenvolvimento dos alunos, os quais não ocorrem somente em salas de aula tradicionais.

Estas melhorias abordam tanto a aprendizagem direta dos alunos como por exemplo, uso de softwares educativos em salas de aula para avaliar o desempenho dos alunos após a aplicação dessas ferramentas, como utilizando pesquisas na área de mineração de dados educacionais, para identificar melhorias no ambiente educacional. Para Baker, Isotani e Carvalho (2011), este campo é definido como a área de pesquisa que possui como foco principal o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educativos.

A Aprendizagem de Máquina (AM) a partir de um conjunto dados provenientes de softwares educativos e bases educacionais, pode ser aplicada, por exemplo, para compreender o aluno em relação a padrões de aprendizagem, entender o contexto na qual a aprendizagem ocorre, além de outros fatores (BAKER; ISOTANI; CARVALHO, 2011).

Segundo Rich e Knight (1991), um computador pode servir como interpretador de informações recebidas de maneira a melhorar seu desempenho, ou seja, a aprendizagem de máquina é fundamentada em coletar informações e analisá-las em busca de gerar resultados melhores. Existem diversas definições na literatura que tangem o conceito de aprendizagem de máquina, porém, no que diz respeito a aplicações computacionais, AM trata a questão de como construir programas de computador que venham a “aprender” com a experiência, ou seja, o quão melhorado foi o desempenho em determinada tarefa por meio da experiência (MITCHELL, 1997).

Costa *et al*. (2012) descrevem conceitos, técnicas, ferramentas e aplicações da mineração de dados em busca de melhor compreender o comportamento dos estudantes e a forma como eles aprendem. Nascimento, Junior e Fagundes (2018) apresentam um estudo sobre indicadores da educação em bases de dados do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) utilizando as técnicas de mineração de dados educacionais, tal como a aprendizagem de máquina (AM).

Para Monard e Baranauskas (2003) a AM é uma ferramenta poderosa, entretanto, não há indícios de que um único algoritmo traga o melhor desempenho para todos os problemas. Trabalhos referentes a comparação de algoritmos de AM podem ser encontrados na literatura tais como Oliveira (2017) que realizou um estudo comparativo de algoritmos de aprendizagem de máquina utilizados em técnicas de *fingerprinting* para localização (GPS) *indoor* em diferentes tipos de *smartphone*, os algoritmos utilizados foram escolhidos com base nas suas aplicações prévias em relação ao tema, por fim os algoritmos foram avaliados pela acurácia e pelas métricas retiradas da matriz de confusão como *precision* e *recall*, além da capacidade de generalização de cada um. Filho (2018) promoveu uma análise comparativa entre algoritmos de aprendizado profundo para detecção de distração de motoristas enquanto dirigem por meio de imagens. A avaliação dos algoritmos foi realizada por meio da acurácia e as métricas da matriz de confusão como *precision*, *recall* e *F1-score*.

Este trabalho propõe a aplicação e comparação de algoritmos de aprendizagem em um problema na área educacional. O estudo foca na mineração de dados a partir do uso dos algoritmos aplicados a uma base sobre a opinião de estudantes do ensino superior na modalidade presencial a respeito do ensino a distância (EAD) no período de pandemia da Covid-19. A base de dados utilizada no experimento foi criada a partir das respostas dos alunos ao responder um formulário a eles disponibilizados pelas redes sociais. O formulário foi composto por questões sobre informações pessoais do aluno, condições de acesso, condições psicológicas, infraestrutura familiar e social, situação econômica, entre outras.

O experimento foi realizado com base nas etapas do processo de descoberta de conhecimento, as fases compreendem a coleta dos dados, análise para pré-processamento, aplicação algoritmos de aprendizagem de máquina e análise dos resultados. A análise comparativa dos algoritmos: SOM, *k-means,* árvore de decisão (J48), JRip e PART foram feitas usando taxas de acerto dos classificadores e nas métricas da matriz de confusão como Precisão*, Recall* e curva ROC.

1.1 JUSTIFICATIVA

Segundo o INEP, órgão responsável pelo levantamento e divulgação de informações sobre a educação no país, os índices de repetência nos ensinos médio e fundamental chegou a quase 16% nos anos entre 2014 e 2015 (INEP, 2017). Assim, aplicar algoritmos de aprendizagem de máquina e avaliar esses modelos educacionais possibilita maior compreensão dos fatores que permitam uma melhora na qualidade das técnicas de ensino e desempenho dos estudantes, em relação a evasão de alunos no ensino médio aqui apresentada, por exemplo.

Para os autores da Nota Técnica nº 16 Inteligência Artificial na Educação, Isotani e Pinto (CIEB, 2019) o contexto da Inteligência Artificial aplicado à educação traz dois objetivos fundamentais: o objetivo educacional que busca compreender de maneira mais profunda e com mais detalhes sobre como e quando a aprendizagem ocorre, elaborando subsídios para melhorar as práticas educacionais/instrucionais, e o objetivo tecnológico, que visa promover o desenvolvimento de ambientes adaptativos de aprendizagem mais flexíveis, inclusivos, personalizados, envolventes e eficazes. Ainda, segundo eles, o espectro do uso da IA na educação compreende (CIEB, 2019):

- Entender o fenômeno para apoiar a tomada de decisão
- Atuar no ambiente para promover os objetivos Educacionais
- Retroalimentar o sistema com vistas a melhorar as duas ações anteriores

Portanto, na tentativa de detectar possíveis melhorias na área educacional, busca-se aplicar algoritmos de aprendizagem de máquina para aprimorar métodos educacionais e de aprendizagem dos alunos.

1.2 OBJETIVO

Aplicar algoritmos de aprendizagem de máquina para analisar e identificar padrões e informações relevantes de perfis de alunos do ensino superior, na modalidade presencial, sobre as aulas em ensino remoto no período de pandemia da COVID-19. Os objetivos específicos são:

- Criar uma base de dados com informações e características de alunos do ensino superior em situação de ensino remoto ou não.
- Identificar critérios para comparação e avaliação dos algoritmos de AM.
- Analisar os resultados das informações extraídas.

1.3 ORGANIZAÇÃO DO TRABALHO

Esse trabalho está organizado em seis capítulos. O Capítulo 2 aborda sobre educação e inteligência artificial, a importância da tecnologia no ensino e ferramentas desenvolvidas. São descritas técnicas da IA aplicadas ao ensino, trabalhos desenvolvidos e avanços da tecnologia no futuro para educação.

O Capítulo 3 discorre sobre os conceitos de mineração de dados e aprendizagem de máquina, são abordados a importância da AM e algumas das suas variadas aplicações, além de uma mostra geral dos algoritmos desenvolvidos nesta área. Por fim, trabalhos relacionados a comparação de algoritmos de aprendizagem de máquina e suas técnicas são retratados.

O capítulo 4 fala da metodologia aplicada para execução dos experimentos, são descritas as fases e o que foi realizado em cada uma delas.

O capítulo 5 apresenta os resultados obtidos durante os experimentos, as saídas geradas pelos algoritmos são descritas e detalhadas, além de suas métricas de avaliação e desempenho. Ainda, é realizada uma discussão comparativa dos resultados, analisando as informações encontradas de maneira geral e o desempenho dos algoritmos utilizados, com o objetivo de compreender melhor estes dados.

Por fim, o capítulo 6 refere-se a conclusão desta pesquisa e debate sobre possíveis trabalhos futuros.

2 EDUCAÇÃO E INTELIGÊNCIA ARTIFICIAL

Este Capítulo discorre sobre a importância do emprego da inteligência artificial, no processo educacional, como uma ferramenta transformadora destes processos. A Seção 2.1 aborda a importância do uso das tecnologias na área de educação. A Seção 2.2 trata sobre a inteligência artificial aplicada a educação. Por fim, a Seção 2.3 trata as considerações finais do capítulo.

2.1 IMPORTÂNCIA DA TECNOLOGIA DA INFORMAÇÃO NA EDUCAÇÃO

A educação vem sofrendo gradativas transformações dentro da sociedade. Estas mudanças acontecem devido ao avanço das tecnologias de informação e comunicação (TICs) que também sofrem tais ações e evoluem com o passar dos anos. De acordo com Cysneiros (2000), a inserção da informática no ambiente educacional faz-se necessária, consentindo o acesso dos indivíduos a um bem cultural que deve estar ao alcance de todos. Projetos de Inclusão Digital continuam sendo desenvolvidos ao longo dos anos em diversas escolas da educação básica, tendo estes o objetivo de prover uma melhoria na educação para formar alunos ainda mais competitivos para o mercado de trabalho e prepara-los para uma sociedade globalizada.

Para Marques e Caetano (2002), a informática quando aplicada no ensino, proporciona flexibilidade na aprendizagem, unifica as teorias e as práticas em que os alunos aprendem e entendem como, por que, onde e quando eles aprendem. Entretanto, a informática não deve ser atrelada como redentora da educação e sim vista como um elemento adicional na contribuição e construção de uma escola que possa criar mecanismos que auxiliem alunos na superação de suas barreiras.

Lopes (2004) realizou um estudo com base na experiência da introdução da informática em uma escola do estado de São Paulo. Ele observou os processos, tanto dos professores quanto dos alunos e concluiu que projetos de informática nas escolas devem fazer parte do programa político pedagógico. Além disso, alguns fatores resultantes foram observados como a definição de um momento para a prática do projeto, a existência da figura de um coordenador de informática para gerenciamento

e mobilização dos professores e o engajamento em um projeto pedagógico para apoio da direção que oferece os recursos necessários.

Outra pesquisa realizada por Borges (2008) apresenta um estudo de caso para implementação de um projeto de inclusão digital em uma escola pública de Ipatinga/MG. Apesar das dificuldades encontradas durante o projeto, muitas contribuições foram alcançadas por eles, dentre estas: maior participação dos alunos na escola, melhorias nas disciplinas e frequência dos alunos; maior trabalho em equipe possibilitando a socialização do conhecimento e maior interação entre os estudantes; aprendizagem mais prazerosa e possibilitando o acesso de alunos carentes ao computador e tecnologia.

Para Folque (2011) analisando a escola como um ambiente de criação da cultura e transformação social, é papel das instituições agregar os mais diversificados e avançados produtos culturais e práticas sociais, pois espera-se que a escola contribua no sentido de ajudar crianças e jovens a viver em um ambiente cada vez mais “automatizado”, por meio do uso da eletrônica e das telecomunicações. Desse modo, cada vez mais a tecnologia se mostra vigente na escola e no aprendizado do aluno, seja pelo uso de ferramentas tecnológicas em sala ou por meio de projetos que envolvem educação e tecnologia (DE OLIVEIRA, 2015).

De acordo com Bacich, Neto e Trevisani (2015), estruturalmente, a escola dos dias atuais não se distingue das do início do século passado, contudo, os estudantes de hoje não mais aprendem da mesma maneira que os do século anterior. Para esses autores, por meio da facilidade de acesso à informação, novas formas de aprendizagem emergiram trazendo a construção coletiva e compartilhada do conhecimento com toda a sociedade a partir de um único clique no mouse. Assim, nessa construção onde muitos estão inseridos, é possível identificar que não há um conhecimento pronto e concluso, mas reorganizações de conceitos que levam em consideração diferentes cenários.

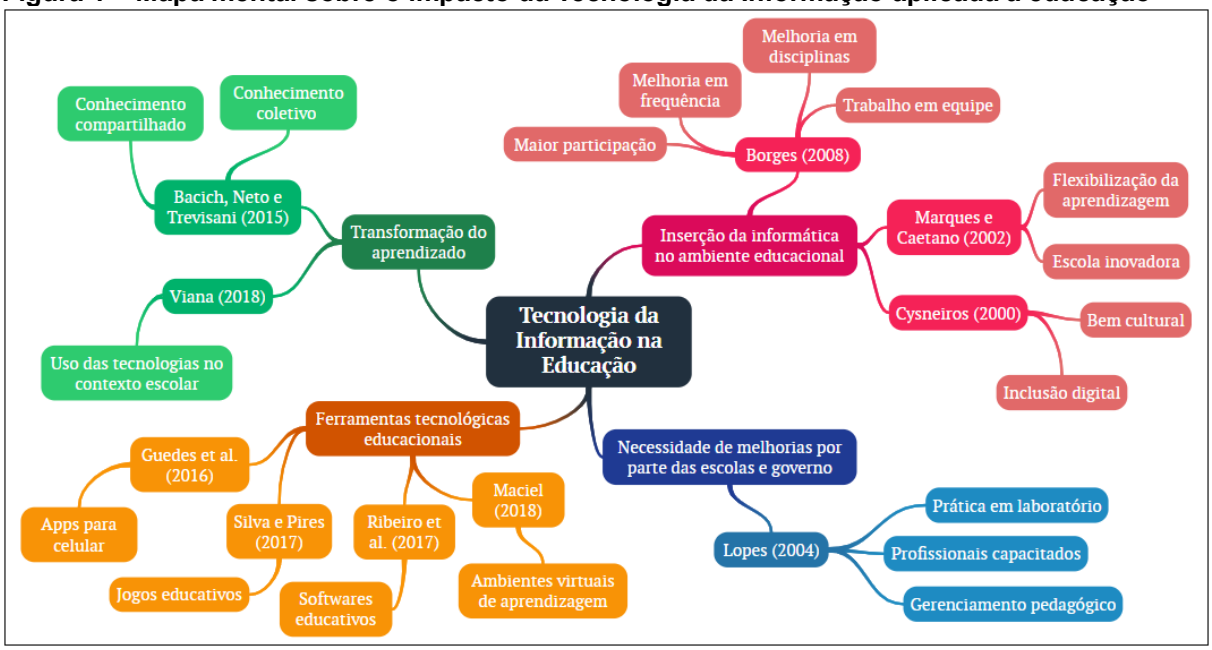
Diante disto, entende-se que por meio do uso das tecnologias há possibilidade de aproximar a convivência das pessoas e aumentar as chances de inclusão dos indivíduos por meio da informação, proporcionando a estes uma nova experiência. Almeida (2015) expressa que:

> O professor que associa a TIC aos métodos ativos de aprendizagem desenvolve a habilidade técnica relacionada ao domínio da tecnologia e,

> sobretudo, articula esse domínio com a prática pedagógica e com as teorias educacionais que o auxiliem a refletir sobre a própria prática e a transformá-la, visando explorar as potencialidades pedagógicas da TIC em relação à aprendizagem e à consequente constituição de redes de conhecimentos (ALMEIDA 2015. p. 72).

Outra pesquisa elaborada por Viana (2018), mostrou a utilização das novas tecnologias no contexto escolar, levando em consideração o docente como motivador intermediador nesse processo onde é indicado que o professor precisa estar aberto as mudanças de paradigmas, uma vez que o uso das mídias no contexto educacional vieram para inovar os métodos tradicionais de ensino e aprendizagem. É necessário que todos os membros inseridos no contexto escolar dos alunos, inclusive os pais, tenham seu papel redesenhado. A Figura 1 apresenta um mapa mental sobre o impacto da tecnologia da informação no ambiente educacional de acordo com as informações aqui apresentadas. A contribuição que as TICs oferecem é explícita, pois agregam fatores pedagógicos e sociais para a melhoria do processo educacional e do ambiente escolar.

Figura 1 – Mapa mental sobre o impacto da Tecnologia da Informação aplicada à educação



Fonte: Autoria própria

A tecnologia não está inserida apenas na educação em sala de aula, e é por causa da evolução das TICs que ela também se traduz no surgimento da Educação a Distância (EAD). Segundo Da Costa (2017) a EAD é uma modalidade de ensino-aprendizagem mediada pelas Tecnologias da Informação e Comunicação (TICs) que

permitem ao professor e o aluno estarem fisicamente em ambientes distantes ou diferentes. A EAD permite que o ensino-aprendizagem sejam proporcionados por meio de tecnologias como os correios, rádio, televisão, vídeos e, principalmente, a internet. Com isso, a informática representa papel fundamental neste processo por meio do desenvolvimento de softwares, plataformas *online* de ensino, tecnologias *streaming*, criação de canais de comunicação como *e-mails* e *chats*, entre outros (PIRES; ARSAND, 2017).

Além destas, outras podem ser as ferramentas educacionais providas das tecnologias tais *apps* para celulares (GUEDES *et al.*, 2016), jogos (SILVA; PIRES, 2017) e softwares educativos (SEs) (RIBEIRO *et al.*, (2017), ambientes virtuais de aprendizagem (AVAs) (RIBEIRO *et al.*, 2007).

2.2 INTELIGÊNCIA ARTIFICIAL APLICADA À EDUCAÇÃO

A Inteligência Artificial é uma área em que os estudos já são produzidos a um certo tempo. Esta ciência possibilita a simulação do pensamento humano nas mais diferentes áreas, e com isso, a possibilidade de descobrir meios de solucionar problemas (SANTOS, 2005).

Os questionamentos acerca da inteligência artificial sempre foram observados por estudiosos e levam as discussões sobre os impactos que a Inteligência Artificial (IA) e a informática podem trazer a sociedade. A IA aplicada a educação é uma área de pesquisa multi e interdisciplinar, pois considera o uso das tecnologias da IA em softwares que possuem como objetivo o ensino e a aprendizagem (VICARI, 2018). Desta forma, a IA educacional reúne a área da inteligência artificial, que é interdisciplinar, com as ciências de aprendizagem (educação, psicologia, neurociência, linguística, sociologia e antropologia) para proporcionar o desenvolvimento de ambientes de aprendizagem adaptáveis e outras ferramentas da IA na educação flexíveis, inclusivas, personalizadas, envolventes e eficazes (LUCKIN *et al.*, 2016).

Segundo Luckin *et al.* (2016, p. 18), a IA educacional possuí como pilar principal o objetivo científico de “tornar formas computacionalmente precisas e explícitas de conhecimento educacional, psicológico e social que muitas vezes são deixadas implícitas”. Ou seja, além de ser o mecanismo de muitas tecnologias ditas

“inteligentes”, a IA aplicada à educação também é uma ferramenta para abrir o que as vezes é chamado de “caixa preta da aprendizagem”, provendo entendimentos mais profundos e com maior riqueza de detalhes de como a aprendizagem realmente acontece. Um exemplo é entender como o ensino pode vir a ser influenciado pelo contexto socioeconômico e físico do aluno ou então pela tecnologia.

A inteligência artificial trabalha de certa forma com a investigação, pois ela examina a forma que o ser humano raciocina. Sendo assim, a IA visa transformar o pensamento em tecnologia, dedicando seus esforços com o objetivo de construir esclarecimentos algorítmicos dos processos mentais humanos. Tais processos dividem-se em quatro grupos: a) área conexionista, capacidade dos computadores de identificarem e aprenderem modelos; b) segmento ligado à biologia molecular na tentativa de gerar vida artificial; c) campo referente à robótica em conjunto com a biologia, com o intuito de criar máquinas que hospedem IA; d) área compatível com a psicologia, epistemologia e sociologia que procura mostrar à máquina formas de raciocínio e procura (FAVA, 2016).

De acordo com Vicari (2018), os sistemas educacionais com maiores aplicações que utilizam da tecnologia da IA são os: Sistemas Tutores Inteligentes Afetivos (STIs), os *Learning Management Systems* (LMSs), a Robótica Educacional Inteligente e os *Massive Open Online Courses* (MOOCs), sendo estes todos referentes a *Learning Analytics* (LA). Além disso, cada uma dessas aplicações utilizam as técnicas da IA de forma diferente, uma vez que se tratam de aplicações distintas. O Quadro 1 descreve brevemente cada um destes segmentos com suas aplicações voltadas a educação.

Quadro 1 - Sistemas educacionais que utilizam tecnologia da IA

Sistema	Descrição
Tutores Inteligentes Afetivos (STIs)	Sistemas que reconhecem as emoções dos alunos ou geram emoções para o tutor interagir de forma afetiva com o aluno.
Robótica Inteligente Educacional	A robótica inteligente educacional recupera, principalmente, robôs e suas plataformas de programação que podem ser utilizados na educação.
Processamento de Linguagem Natural (PLN)	Comtempla a geração e compreensão automática de línguas humanas naturais. O PLN na educação trata, basicamente, da aplicação desse processamento em interfaces educacionais que permitem a tradução simultânea.
MOOCs	É a sigla em inglês para *Massive Open Online Courses,* ou seja, cursos *on-line* abertos cujo objetivo é atingir um grande público
Baseado em jogos e Aprendizagem	Termo da área de Educação que surge da necessidade de inserir metodologias interativas entre os alunos, ou entre alunos e professor.

Fonte: Adaptado de Vicari (2018)

Quadro 1 - Sistemas educacionais que utilizam tecnologia da IA

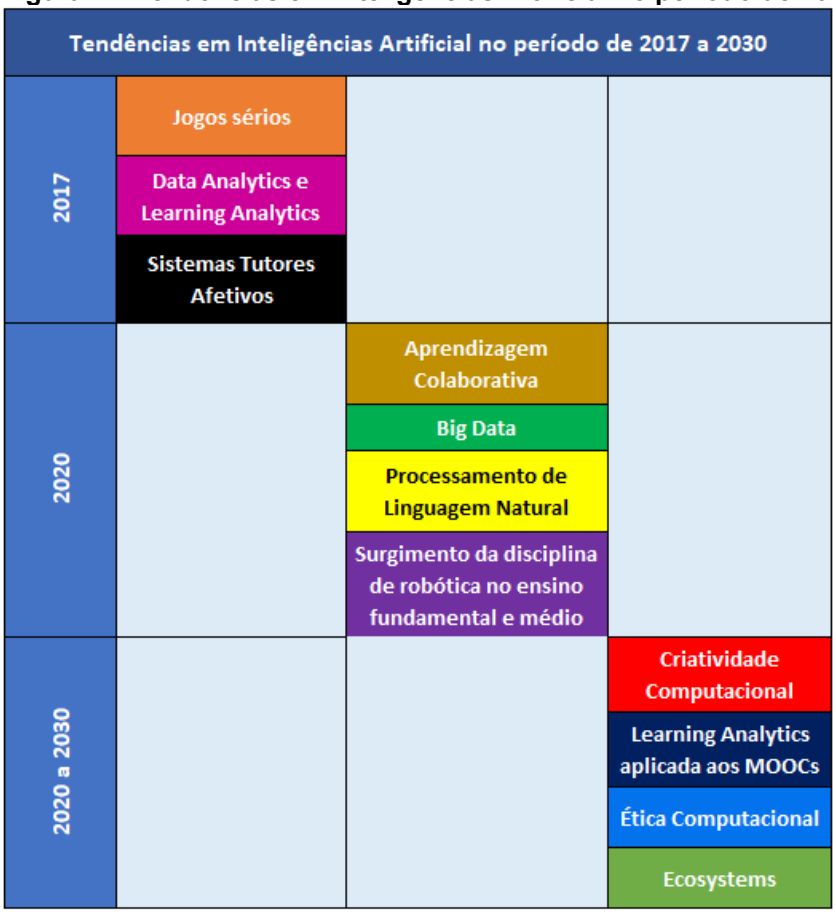
Sistema	Descrição
Aprendizagem Colaborativa	Termo da área de Educação que surge da necessidade de inserir metodologias interativas entre os alunos, ou entre alunos e professor.

Fonte: Adaptado de Vicari (2018)

Em seu estudo, Vicari (2018) tenta mostrar tendências em Inteligência Artificial na Educação no período de 2017 a 2030, e realizou uma busca a partir dos termos apresentados destas tecnologias, em conferências e periódicos relevantes destas áreas. A partir disso, a autora busca realizar uma prospecção das áreas para o futuro, por meio de uma análise qualitativa das publicações selecionadas previamente. Com o uso de um *Roadmap*, a autora cita sistemas como Jogos Sérios e Sistemas Tutores Afetivos como produtos já existentes no mercado, e suas pesquisas já advém de um período anterior a 2017. A Aprendizagem Colaborativa e o Processamento de Linguagem Natural são áreas que foram prospectadas até 2020, por já possuírem uma oferta considerável destes produtos no mercado. Já o intervalo entre 2020 e 2030 a autora cita áreas consideradas tecnologias embrionárias como a Criatividade Computacional, Ética Computacional e os *Ecosystems*, pois estas tecnologias são apresentadas como temas de pesquisa atual e por este motivo foram classificadas a longo prazo.

Vicari (2018) ainda menciona que o período de 2017 foi diretamente vinculado ao uso de três diferentes realidades tecnológicas subjacentes à IA, onde juntas, mudaram o perfil do uso das tecnologias educacionais, sendo estas: redes sem fio (*Wi-Fi*), tecnologias móveis (celular e *tablet*) e armazenamento de conteúdos em nuvens. Todas estas tecnologias influenciam a IA e são responsáveis pelo surgimento das novas tecnologias como *Learning Analytics*, *Big Data*, a possibilidade do treinamento de algoritmos de Aprendizagem de Máquina (AM) com grandes quantidades de dados, etc. A Figura 2 denota as áreas com aplicação e uso da inteligência artificial no período entre 2017 a 2030, com base nas informações produzidas pelo estudo de Vicari (2018), as diferentes cores representam as tecnologias produzidas.

Figura 2 – Tendências em Inteligências Artificial no período de 2017 a 2030



Fonte: Adaptado de Vicari (2018)

Como mencionado pela autora sobre a aplicação destes na educação, os algoritmos de aprendizagem de máquina são utilizados neste estudo, sendo possível a coleta das informações para o trabalho com o uso de ferramentas tecnológicas, sem a necessidade de uma pesquisa de campo. Os algoritmos de aprendizagem de máquina, tem como objetivo encontrar informações relevantes a partir de grandes quantidades de dados, visto que estas técnicas são bastante promissoras e de grande ajuda para novas descobertas e aplicações educacionais.

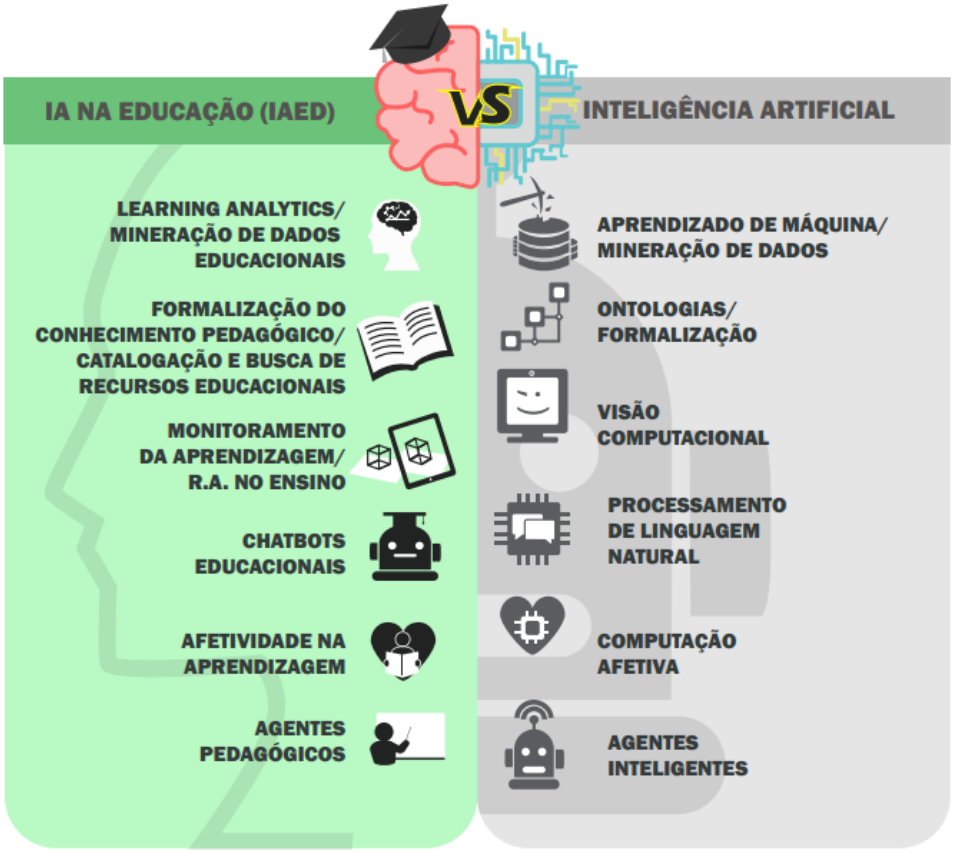
Isotani e Pinto (CIEB, 2019) autores da nota técnica nº 16 Inteligência Artificial na Educação (IAED), apresentam a IA como uma tecnologia emergente, e descrevem suas contribuições para a educação e referenciais para implementação de IA no setor. Assim como define-se a IA, eles definem a IAED como:

> Um sistema de computador projetado para interagir com o ecossistema educacional (atores, recursos, visões pedagógicas etc.), por meio de capacidades e comportamentos inteligentes (utilizando algoritmos ou técnicas provindas da área de IA), para entender e encontrar soluções de

> problemas educacionais complexos que eram então compreendidos e resolvidos essencialmente por humanos.

Também, os autores citam que a IAED amplia as capacidades do professor, permitindo que ele foque na tarefa mais importante de acompanhar os estudantes individualmente apoiando com mais efetividade o processo de ensino e aprendizagem. A Figura 3 mostra algumas oportunidades que a IA oferece no contexto educacional com base nas áreas da computação que utilizam IA.

Figura 3 – Relação entre áreas da IAED com as técnicas da IA



Fonte: CIEB (2019)

O aprendizado de máquina e mineração de dados estão ligados a área de *Learning Analytics* e Mineração de Dados Educacionais, como também visto em Vicari (2018). Para os autores Isotani e Pinto (CIEB, 2019), as duas áreas, apesar de distintas são fortemente correlatas, possuem como objetivo melhorar a educação a partir da análise e do uso de grandes quantidades de dados, provendo ferramentas para entender melhor os problemas educacionais, tanto do ponto de vista dos estudantes quanto dos professores e gestores, facilitando o desenvolvimento de modelos e estratégias para entender melhor o processo de ensino e aprendizagem.

Em relação a outras aplicações envolvendo IA na educação, Pinto (2014) promove a inserção de um robô humanoide em salas de aula. Este robô tem características físicas parecidas como a dos seres humanos, em que a máquina é capaz de reconhecer figuras geométricas planas, podendo ser estendido para outros tipos de conteúdo e assim utilizado como ferramenta de ensino.

Semensato *et al.* (2015) descrevem o uso da inteligência artificial dentro da educação a distância, descrevendo a importância da IA na organização da EAD como um aumento da facilidade de uso, interação com o usuário, *feedback* de tutor *online* e diminuição da necessidade de encontros presenciais.

Outra aplicação em destaque é atribuída a escola de negócios *Saint Paul*, que informou em 2017 a disponibilização de uma plataforma de educação corporativa que engloba um assistente de inteligência cognitiva chamado Watson, com recursos de *e-learning*, vídeos e biblioteca. A plataforma recebeu o nome de LIT, e pode ser acessada a qualquer momento por meio de aplicativo celular, *tablet* ou *desktop* (SORAIA, 2017).

Também, tem-se a ferramenta *Mr Turing*, uma plataforma que utiliza a inteligência artificial para ensinar inglês. O programa consiste de um *chatbot* que interage com o usuário reorganizando o conteúdo de acordo com as necessidades e baseado no desempenho do usuário. A aplicação é disponibilizada no *Facebook* (DATAH, 2017).

2.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

A inteligência artificial é uma ferramenta de colaboração importante porque pode ser aplicada em áreas de forma multi e interdisciplinar. Ela se mostra como uma ferramenta transformadora com o propósito de desenvolver, analisar e descobrir recursos. Por isto, é promissor o uso deste mecanismo nos segmentos da educação para encontrar formas de tornar este ambiente mais prazeroso e um meio facilitador que busque impulsionar a área educacional, não só no Brasil, mas também no mundo.

Outro ponto importante refere-se a tecnologia inserida no ambiente escolar. As escolas, em muitos casos, já não são o ponto de partida da integração da tecnologia com os alunos, isto é, muitos estudantes já estão ambientados com produtos tecnológicos (como celulares, tablets, videogames, etc.), assim as escolas

precisam estar equipadas com esses dispositivos, reformulando o ambiente escolar tornando-o mais inovador, atrativo e desenvolvedor.

As ferramentas tecnológicas de ensino passam por estudos, assim elas apresentam um quadro de constante evolução, e desta forma é preciso preparar os alunos para as mudanças que possam vir a ocorrer dentro das salas de aula, pois em um período próximo, acredita-se que novas ferramentas, como até mesmo robôs, possam estar auxiliando o processo de aprendizagem destes estudantes.

Entretanto, como nem todos estão inseridos na mesma classe social, como seria para um aluno chegar em uma sala onde não consiga sequer utilizar das ferramentas tecnológicas disponíveis? Por esses e outros motivos, que não somente professores, mas toda a equipe escolar precisa estar atenta a estas evoluções. Ademais, o uso de ferramentas tecnológicas promove o avanço do ambiente escolar em um contexto geral, contribuindo para melhores análises e pode fornecer novas descobertas ao cenário educacional.

Também, observando a prospecção das tendências em IA desenvolvida por Vicari (2018) e as pesquisas apresentadas pelos autores abordadas nesta seção, é possível notar a relação entre os estudos envolvendo robótica, sistemas tutores inteligentes, análise de dados e IA aplicada ao ensino e ferramentas EAD são pesquisas ocorrendo neste período e que estão prospecção para um maior avanço e desenvolvimento nos próximos anos.

3 APRENDIZAGEM DE MÁQUINA E MINERAÇÃO DE DADOS EDUCACIONAIS

Este Capítulo apresenta definições sobre a Aprendizagem de Máquina, Mineração de Dados e Mineração de Dados Educacionais. Para melhor entendimento destas áreas, o conteúdo divide-se em seções. A Seção 3.1 trata as etapas presentes no processo de mineração de dados e descoberta de conhecimento. A Seção 3.2 aborda os conceitos da aprendizagem de máquina, são conceituadas as bases de dados, alguns dos algoritmos de aprendizagem de máquina mais utilizados em modelos educacionais além das técnicas de comparação entre algoritmos de AM. A Seção 3.3 discorre sobre a mineração de dados educacionais e AM. Por fim, a Seção 3.4 aborda as considerações finais do capítulo.

3.1 MINERAÇÃO DE DADOS E DESCOBERTA DE CONHECIMENTO

De acordo com Fayyad (1996) o modelo usual para transformação dos dados em informação, consiste de um processamento manual de todas as informações por especialistas, de forma a produzir relatórios que deverão ser analisados. Porém, para a grande maioria dessas aplicações, esse processo manual se torna impraticável devido ao grande volume de dados existentes. Também, o processo de descoberta de conhecimento em bases de dados (KDD – *Knowledge Discovery in Databases*) é descrito pelo autor como uma tentativa de solucionar o problema causado pela chamada “era da informação”: a sobrecarga de dados.

A Mineração de Dados (MD), segundo Carvalho (2005), é descrita como o uso de técnicas automáticas de exploração de grandes quantidades de dados com o objetivo de descobrir novos padrões e relações, que, devido a este grande volume de dados, não seriam facilmente descobertas pelo ser humano. Também, para Han e Kamber (2001) mineração de dados, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados.

A aprendizagem de máquina é considerada um subcampo da inteligência artificial (IA), e estuda o desenvolvimento de métodos capazes de extrair informações e conhecimento a partir de uma determinada amostra de dados (TAKAKURA *et al*., 2017).

Sendo assim, a utilização das técnicas de mineração de dados e aprendizagem de máquina estão relacionadas a manipulação da grande massa de dados que cresce de maneira constante, encontrando informações que são de difícil extração, e que quando reveladas, se mostram de grande valia na tomada de decisão.

A definição acerca dos termos Descoberta de Conhecimento e Mineração de Dados ainda não é consenso entre autores. Em Rezende (2005), Wang (2005) e Han e Kamber (2006) os termos possuem mesmo significado. Para Fayyad (1996) o KDD refere-se a todo o processo de descoberta de conhecimento, enquanto a mineração de dados a uma das atividades do processo. Neste trabalho o foco está mais voltado à aplicação da Mineração de Dados, as etapas que são posteriormente aplicadas e discutidas são: Entendimento do problema e coleta dos dados, Pré-processamento, Mineração de dados e Pós-processamento.

3.1.1 ENTENDIMENTO DO PROBLEMA E COLETA DOS DADOS

As fontes de dados podem vir de diversos locais e possuírem diversos formatos. Nesta fase, é definido o entendimento do problema e o que se deseja analisar. Segundo Olson (2008) é necessário conhecer os dados buscando descrever o problema de forma clara e identificar os dados relevantes para o problema.

3.1.2 PRÉ-PROCESSAMENTO

Geralmente, existem diversos passos que procedem o processo de aprendizagem. Por exemplo, um passo para realizar uma limpeza pode ser usado para melhorar a qualidade dos dados, modificando seu formato ou conteúdo, como remover ou corrigir os valores de dados incorretos (MONARD; BARANAUSKAS, 2000). Este processo deve realizar limpeza e transformação dos dados de um conjunto para promover uma melhor descoberta de conhecimento, em geral, é denominada pré-processamento de dados.

Para Alasadi e Bhaya (2017) o pré-processamento é uma das principais tarefas de mineração de dados que inclui a preparação e transformação dos dados em uma forma adequada para os procedimentos de mineração. Esta técnica visa

reduzir o tamanho dos dados, encontrar relações entre eles, normalizar, remover discrepâncias e extrair recursos para os dados. Neste trabalho são abordados de forma breve alguns dos principais: limpeza e transformação de dados.

A limpeza é o primeiro passo nas técnicas de pré-processamento de dados, são usadas para encontrar os valores faltantes, suavizar os dados de ruído, reconhecer *outliers* e corrigir inconsistentes.

Os dados faltantes podem ser preenchidos da seguinte forma:

- Ignorar a tupla: essa opção é escolhida quando o valor do atributo classe não existe. Não é considerado um método eficaz, porém, é usado quando uma tupla possuí vários atributos com valores vazios.
- Usar valor média do atributo: esse método funciona substituindo o valor faltante de um atributo específico pelo valor médio desse atributo.
- Usar o valor mais provável: essa abordagem é usada com técnicas como regressão baseada em inferência usando uma indução de árvore de decisão ou formalismo Bayesiano.

A transformação de dados inclui a modificação dos dados em valores adequados para o processo de mineração. Para Baskar *et al.* (2013) a transformação de dados envolve o seguinte:

- Normalização: este método ajusta os valores dos dados em um intervalo específico, como entre 0-1 ou -1-1. Considerado útil para as técnicas de classificação, redes neurais artificiais e algoritmos de agrupamento. O uso das escalas Mínimo-máximo, z-*score* e escala decimal são algumas das formas mais populares de normalização.
- Seleção de subconjunto de atributos: possuí o objetivo de reduzir o tamanho do conjunto de dados removendo atributos ou dimensões redundantes ou atributos irrelevantes.

3.1.3 MINERAÇÃO DE DADOS

A etapa de mineração de dados é considerada a principal dentre as etapas do processo de mineração e descoberta de conhecimento em bases de dados, seu objetivo é extrair padrões e informações dos dados. Esta etapa é considerada o centro do processo buscando ajustar adequadamente os modelos ou determinar padrões a

partir dos dados observados. É também vista como uma maneira de selecionar, explorar e modelar grandes quantidades de dados para identificar padrões de comportamento (FAYYAD, 1996).

É nesta fase que os algoritmos de mineração são aplicados, sendo a escolha destes dependentes dos objetivos desejados (MCCUE, 2007).

3.1.4 PÓS-PROCESSAMENTO

Nesta etapa, algumas das operações principais são a avaliação de padrões, onde identifica-se e interpreta padrões relevantes, ou seja, aquisição de alguma informação relevante obtida a partir da análise dos dados na fase de mineração, e apresentação dos resultados, que consiste representação e visualização dos resultados da mineração de dados (HAN, KAMBER, 2006).

3.1.5 TAREFAS DO PROCESSO DE MINERAÇÃO DE DADOS

A mineração de dados é em geral classificada pela sua capacidade em realizar determinadas tarefas, as principais são (LAROSE, 2005): Classificação, Agrupamento e Associação.

A Classificação visa identificar a qual classe um determinado exemplo ou registro pertence. O modelo analisa o conjunto de registros fornecidos, com cada registro contendo a sua respectiva classe, de forma a “aprender” como classificar um novo registro (LAROSE, 2005).

O Agrupamento visa identificar e aproximar registros similares. Um agrupamento pode ser descrito como uma coleção de registros similares entre si, mas diferente dos outros registros nos demais agrupamentos (LAROSE, 2005). Nesta tarefa, não é necessário que os registros sejam previamente categorizados como na classificação. Ainda, o agrupamento não possui intenção de classificar ou prever o valor de uma variável, apenas de identificar grupos de registros com características comuns.

A Associação consiste em identificar quais atributos estão relacionados. Possuem o formato SE atributo X ENTÃO atributo Y. Considerada uma das tarefas

mais conhecidas por conta dos seus bons resultados obtidos. Assim como no agrupamento, a associação também não exige que os registros sejam previamente categorizados (LAROSE, 2005).

As tarefas do processo de mineração encontram-se representadas na aprendizagem de máquina por meio dos algoritmos desenvolvidos nesta área. Existem diversos algoritmos criados para um mesmo segmento entre as tarefas de mineração. Estas tarefas são novamente filtradas dentro da Aprendizagem de Máquina tais como a classificação atribuída a aprendizagem supervisionada, e o agrupamento e a associação à aprendizagem não-supervisionada, temas discutidos posteriormente na seção 3.2.

3.2 APRENDIZAGEM DE MÁQUINA

Segundo Horst (1999), a habilidade de aprender é um dos atributos mais importantes do comportamento inteligente, logo, o estudo e a modelagem computacional dos processos de aprendizado em suas múltiplas manifestações tem se constituído, basicamente, como o objetivo maior das inúmeras pesquisas em aprendizado de máquina.

As pesquisas na área da AM que possuíam certa limitação a estudos teóricos e posterior estudos experimentais, agora recebem maior atenção devido a sua aplicação prática, e a medida em que aumentam o número de aplicações bem-sucedidas de AM, o grau de aceitação em relação ao uso destas pesquisas consequentemente aumenta (HORST, 1999).

Existem inúmeras atividades associadas à noção de aprendizagem, dificultando a definição exata do termo e tornando-o dependente de contexto (HENKE, *et al*., 2011). Porém, no contexto de aplicações computacionais, Alpaydin (2010, p. 3) traz uma definição de aprendizagem de máquina como “programas de computador utilizados para otimizar um critério de desempenho, usando dados de exemplo ou experiência do passado”.

Ben-David e Shalev-Shwartz (2014) levantam a seguinte questão: quando se precisa da aprendizagem de máquina? Ela não somente serve de amparo para realizar uma tarefa cuja execução e desenvolvimento são lentos. Deve-se ser analisado a complexidade e a necessidade de adaptabilidade do problema. Tarefas

que são muito complexas para programar (reconhecimento de imagens), ou que estão muito além da capacidade humana (análise de dados genômicos), apresentam resultados satisfatórios a partir da utilização da AM, uma vez que são expostas a muitos exemplos de treinamento e teste.

Outra questão importante diz respeito sobre o que distingue os mecanismos de aprendizado que resultam em apenas um palpite de um aprendizado útil? Questão crucial para o desenvolvimento do aprendizado automatizado, pois humanos possuem a capacidade de filtrar conclusões aleatórias sem sentido de aprendizado. A partir do momento que se exporta a tarefa de aprender para uma máquina, é necessário oferecer princípios bem definidos que não irão permitir que um programa chegue as conclusões sem sentido ou inúteis. Desenvolver estes princípios é um dos objetivos centrais da teoria do aprendizado de máquinas (BEN-DAVID; SHALEV-SHWARTZ, 2014).

As próximas seções apresentam informações importantes sobre aprendizado de máquina, são abordados assuntos como: tipos de aprendizagem utilizados pela AM, os processos que envolvem a descoberta de conhecimento baseado na concepção do pensamento humano e suas variações, configurações das bases de dados necessárias para execução dos algoritmos, e os métodos de avaliação e comparação destes para a escolha e uso do melhor modelo. Por fim, um apanhado sobre a aplicação da aprendizagem de máquina na educação detalha algumas das contribuições promovidas pelo seu uso no ambiente escolar.

3.2.1 TIPOS DE APRENDIZAGEM

Em geral, os algoritmos de AM são utilizados de forma a produzir classificadores para um conjunto de exemplos. Por classificação, entende-se o processo de atribuir, a um determinado dado, o rótulo de uma (ou mais) classe(s) a qual ele pertence. Neste sentido, as técnicas de AM são empregadas na indução, a partir de um conjunto de exemplos, de um classificador, que deve ser capaz de prever a classe de novos dados do domínio em que ele foi treinado (ESTEVES *et al*., 2009). Como exemplo, pode-se descrever uma classificação de e-mails, onde estes possuem duas classes: *spam* e não *spam*. Neste problema tanto a entrada (um documento e-mail, seja qual for seu tipo) quanto a saída (spam ou não spam) são conhecidas,

entretanto, a maneira como a entrada deve ser convertida na saída é que se apresenta desconhecida (HENKE, *et al.*, 2011).

Takakura *et al.* (2017) descreve a indução, como um princípio de inferência utilizado nas técnicas de AM, que se trata de uma forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos. Ela é caracterizada como o raciocínio que se origina em um conceito específico e o generaliza, ou seja, da parte para o todo. Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. Portanto, as hipóteses geradas por meio da inferência indutiva podem ou não preservar a verdade (MONARD; BARANAUSKAS, 2003). Um exemplo de indução é:

- Premissa 1: Pedro é homem e mortal;
- Premissa 2: João é homem e mortal;
- Premissa 3: Antônio é homem e mortal;
- Conclusão: Todos os homens são mortais.

Segundo Monard e Barnauskas (2003), apesar da indução ser o recurso de maior uso pelo cérebro humano para formar novo conhecimento, ela precisa ser utilizada com cautela, pois se o número de exemplos for baixo ou se os exemplos não forem bem definidos, as hipóteses obtidas podem ser de pouco valor. Portanto, os algoritmos de aprendizagem de máquina podem ser fatores chave para solucionar problemas de natureza indutiva, pois tratam-se de algoritmos que “aprender” a definir padrões das classes envolvidas no problema, a partir de exemplos reais obtidos do ambiente (HENKE *et al.*, 2011).

A obtenção do conhecimento proveniente da aprendizagem indutiva pode ocorrer de duas formas: Aprendizado Supervisionado e Não Supervisionado (NUNAN, 2012). Na aprendizagem supervisionada, o sistema precisa conhecer o ambiente, em que esse conhecimento é representado por um conjunto de exemplos de pares de entrada-saída que são transmitidos em uma sequência de instruções que o computador deve seguir para atingir o efeito esperado (HAYKIN, 2008). É comum que este modelo seja chamado também de classificador. Portanto, é importante que exista um conjunto de dados de treinamento de qualidade para que o modelo criado possa ser capaz de predizer novas instâncias de forma eficiente (CARVALHO, 2014).

Na aprendizagem não-supervisionada, não há exemplos para serem aprendidos e são descritas duas subdivisões: aprendizagem por reforço e a aprendizagem não-supervisionada. A aprendizagem por reforço é definida a partir de

um mapeamento de entrada-saída atingida por meio de interação com o ambiente de modo a minimizar um índice escalar de desempenho (HAYKIN, 2008). Na aprendizagem não-supervisionada não há exemplos rotulados, ou seja, não há valores de saída desejados. A tarefa de aprendizagem envolve obter alguma compreensão dos dados de entrada e desenvolve habilidades para formar representações internas de modo que seja possível codificar características da entrada e, assim, criar novas classes automaticamente (HAYKIN, 2008). A abordagem do aprendizado não-supervisionado é indicada quando o propósito do sistema não é desenvolver um modelo de predição, e sim um modelo cuja função seja descobrir regularidades nos dados que possam ser utilizáveis (CARVALHO, 2014).

Portanto, para que os modelos cumpram o seu papel na descoberta do conhecimento, é necessário que as bases de dados sigam os critérios necessários na fase de Pré-processamento, para que sejam tratados e limpos de forma a produzirem informações concretas. A próxima seção detalha um pouco sobre as bases de dados e suas características para um melhor entendimento.

3.2.2 BASES DE DADOS

As bases de dados são elementos essenciais dentro da aprendizagem de máquina, e não somente dentro da IA, elas servem para armazenamento e consulta de informações em inúmeras áreas, como uma base de clientes de um supermercado, por exemplo. Bases de dados podem ser descritas como fontes de informação eletrônicas, pesquisáveis de modo interativo ou conversacional por meio de um computador (POBLACIÓN; WITTER; SILVA, 2006). Como principais objetivos, as bases procuram promover o acesso à informação, fornecer informações atualizadas, precisas e confiáveis, atender às necessidades do público alvo e fornecer mecanismos de recuperação.

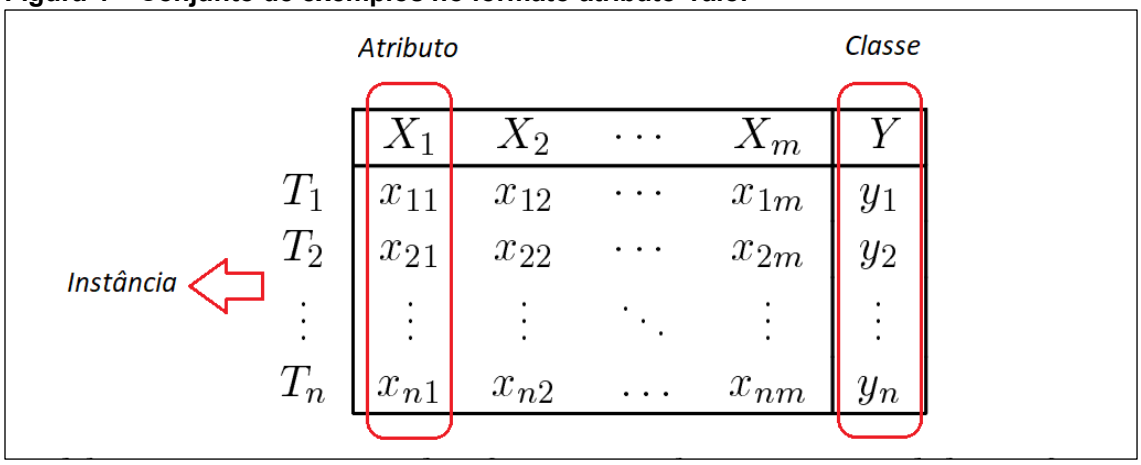
Uma instância (também descrita como exemplo) pode ser representada como um vetor de valores de atributos de uma base de dados. A instância descreve a entidade básica com a qual se está lidando, como um paciente, uma sequência de DNA ou dados médicos sobre alguma doença.

O atributo descreve as características de uma instância. No geral, existem dois tipos de atributos: nominal/categórico (e.g., cor: vermelho, verde, azul) e contínuo/numérico (e.g., o peso de uma pessoa $\in \mathbb{R}$, um número real qualquer).

No aprendizado supervisionado, toda instância possuí um atributo em particular, a classe (também chamada de saída ou rótulo), que descreve o elemento ao qual se gostaria de prever. Os valores das classes são tipicamente pertencentes a um conjunto discreto (nominal) de classes {C_1, C_2, ..., C_k} no caso da classificação, ou de valores reais no caso da regressão.

Em computação uma base de dados pode ser chamada de conjunto de dados ou *dataset*. Um conjunto de dados é um conjunto de instâncias contendo valores de atributos, bem como a classe associada. A Figura 4 mostra o formato padrão de um conjunto de dados *T* com *n* instâncias e *m* atributos. No conjunto, a linha *i* se refere a *i*-ésima instância (*i* = 1, 2, ..., *n*) e a entrada da coluna X_{ij} se refere ao valor do *j*-ésimo (*j* = 1, 2, ..., *m*) atributo X_j da instância *i* (MONARD; BARANAUSKAS, 2000).

Figura 4 – Conjunto de exemplos no formato atributo-valor



Fonte: Adaptado de Monard e Baranauskas (2000)

Dentro da aprendizagem de máquina os modelos precisam ser testados, para isso os conjuntos de dados passam por configurações para as corretas validações das técnicas. Monard e Baranauskas (2000) definem o conceito de conjunto de treinamento e teste como se segue:

> Usualmente um conjunto de dados é dividido em dois subconjuntos disjuntos: o conjunto de treinamento usado para o aprendizado do conceito e o conjunto de teste usado para medir o grau de afetividade do conceito aprendido. Os subconjuntos são normalmente disjuntos para assegurar que as medidas obtidas, utilizando o conjunto de testes, sejam de um conjunto diferente do

> usado para realizar o aprendizado, tornando a medida estatisticamente válida.

É compreensível a necessidade desta distinção entre os conjuntos, uma vez que os exemplos apresentados para o teste não participam na formação do classificador. O Quadro 2 detalha um exemplo de conjunto de dados com seus respectivos exemplos, atributos e classes.

Quadro 2 - Exemplo de um conjunto de dados com seus componentes e respectivos valores

ID	Dor de Cabeça	Dor nas articulações	Febre alta	Moleza e dor no corpo	Está com dengue?
1	Sim	Sim	Sim	Não	Sim
2	Sim	Não	Sim	Sim	Sim
3	Não	Sim	Não	Sim	Sim
4	Sim	Sim	Sim	Não	Não
5	Não	Sim	Sim	Não	Não
6	Sim	Sim	Não	Sim	Sim

Fonte: Autoria própria

A base é composta pelos atributos "*Id", "Dor de cabeça", "Dor nas articulações", "Febre alta", "Moleza e dor no corpo"* e "*Está com dengue?"*. O atributo *"Id"* é classificado como um atributo numérico, visto que seus valores correspondem a uma sequência contínua de números, já os demais atributos são descritos como categóricos ou nominais, em que seus valores são representados pelas palavras *'Sim'* e *'Não'*.

Encontra-se nesta base algumas possibilidades de análises explorando o último atributo *"Está com dengue?"* em que se pode classificá-lo como atributo meta ou classe, de forma a tentar prever seu resultado (*"Sim"* ou *"Não"*) a partir de novas instâncias apresentadas. É possível também aplicar técnicas para encontrar regras que "construam" um caminho a uma determinada resposta, ou ainda criar grupos entre as instâncias cujas características são comuns entre si para inferir o resultado.

3.2.3 ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Existem diversos algoritmos de aprendizagem de máquina, tanto para aprendizagem supervisionada como não-supervisionada. Entre os algoritmos supervisionados pode-se citar as Árvores de Decisão (*Decision Trees*) (CARVALHO,

2005), JRip (COHEN, 1995), PART (FRANK; WITTEN, 1998), *k-NN* (DUDA *et al*., 2000) e Floresta Aleatória (*Random Forest*) (BREIMAN, 2001). Para os algoritmos não-supervisionados temos os Mapas Auto-organizáveis (SOM) (KOHONEN, 1990), *k-means* (PIMENTEL *et al.,* 2003) e *Apriori* (RAO; GUPTA, 2012). Todos estes são apresentados nas próximas seções.

3.2.3.1 Árvores de Decisão

Os algoritmos de árvore de decisão baseiam-se na estratégia de dividir para conquistar (FONSECA, 1994). Para Crepaldi *et al*. (2011), as árvores de decisão são descritas como representações simples do conhecimento e trazem uma maneira diferente de se construir classificadores que predizem classes ou informações úteis baseadas em valores de atributos de um conjunto de dados.

Segundo Carvalho (2005), uma árvore de decisão é induzida por meio de um conjunto de exemplos de treinamento em que as classes são conhecidas de maneira prévia, sua estrutura é organizada de forma que os nós internos (não-folha) são rotulados com o nome dos atributos previsores, as arestas partem de um nó interno e carregam consigo os valores do atributo naquele nó, por fim, os nós folhas (os últimos nós da árvore) carregam uma classe, sendo está a classe do nó em questão além de ser a classe prevista para os exemplos que venham a pertencer àquele nó folha.

A regra de classificação de um exemplo ocorre fazendo o mesmo “caminhar” pela árvore, saindo do nó inicial (ou nó raiz), percorrendo as arestas que unem os nós por meio das condições que os mesmos apresentam. Ao chegar em um nó folha, a classe pertencente a este nó é atribuída àquele exemplo (CARVALHO, 2005).

Por meio dos caminhos retratados pela árvore de decisão é possível derivar regras, ou seja, o caminho do nó raiz até algum nó folha da árvore. Em geral, esses dois processos ocorrem juntos porque as árvores tendem a crescer muito de acordo com certas aplicações. Muitas vezes elas são substituídas pelas regras e isso acontece porque as mesmas são facilmente modularizadas (INGARGIOLA, 1996).

3.2.3.2 JRip

O algoritmo JRIP implementa um aprendizado de regras proposicional, utilizando Poda Incremental Repetida para Produzir Redução de Erros (RIPPER), proposto por William W. Cohen (1995) como uma versão otimizada do IREP. Este classificador integra firmemente a redução de poda de erros com um algoritmo de aprendizado de regras baseado na estratégia de dividir para conquistar, assim como nas árvores de decisão.

Basicamente, o algoritmo constrói um conjunto de regras de forma gulosa, uma regra de cada vez. Depois que uma regra é encontrada, todos os exemplos cobertos pela regra (positivos e negativos) são excluídos. Esse processo é repetido até que não se tenha mais exemplos positivos ou até que a regra encontrada pelo algoritmo possua um valor de erro significativamente alto que não possa ser aceitável.

3.2.3.3 PART

Assim como no JRIP, o algoritmo PART (*Partial Decision Trees*) também é baseado nas árvores de decisão. Adota a estratégia de dividir para conquistar para criar uma árvore de decisão parcial em cada interação, transformando a “melhor” folha em uma regra. Segundo Frank e Witten (1998) criadores do algoritmo, a ideia principal se dá pela construção de uma árvore de decisão “parcial” em vez de uma totalmente explorada.

Os autores descrevem uma árvore de decisão parcial como uma árvore de decisão comum que contém ramificações para subárvores indefinidas. Para gerar essa árvore, operações de construção e remoção são criadas para encontrar uma subárvore “estável” que não possa ser mais simplificada. Depois que essa subárvore é encontrada, a construção de árvores finaliza e uma única regra é lida.

3.2.3.4 Mapas auto-organizáveis (SOM)

O algoritmo de mapas auto-organizáveis ou *Self-organizing maps* (SOM) ou mapas de Kohonen (1990), trata-se de um modelo que utiliza redes neurais artificiais para aprendizado não-supervisionado. Formado basicamente por duas camadas, a

camada de entrada e a camada de saída, para cada neurônio da camada de entrada, estimula-se todos os neurônios da camada de saída, os quais estão dispostos no mapa e interligados entre si em uma dada topologia. O mapa se auto-organiza por meio de vetores de entrada que ativam o mesmo neurônio, definindo conjuntos de vetores de entrada associados a cada neurônio. Ainda, conjuntos de vetores que possuam similaridades entre si devem ativar neurônios próximos no mapa de saída, definindo assim regiões de similaridade.

Esta organização acontece por meio dos ajustes dos pesos das conexões dos neurônios. Cada posição *i* está relacionada a um padrão *i* correspondente, determinado por um vetor de pesos com a camada de entrada. O vetor de pesos denota um vetor característico, no caso o *cluster*, o qual classifica e representa uma região do espaço de entrada. Cada entrada apresentada é comparada com os vetores característicos (pesos), de modo que a entrada ativa uma única posição *i*, sendo esta a que possuir vetor característico de maior similaridade com a entrada. Para o cálculo da similaridade é utilizado a distância euclidiana, porém outros cálculos também podem ser empregados.

Assim, os pesos de cada posição do mapa são ajustados pelo processo de aprendizagem, onde torna-se o vetor característico da posição ativa mais similar ao padrão ora apresentado, este ajuste, propagado às posições vizinhas, define as regiões de maior similaridade, ordenando espacialmente as entradas (KOHONEN, 1990).

3.2.3.5 *k-means*

O algoritmo *k-means* é um algoritmo de agrupamento (*clustering*), que pode ser chamado de *k*-médias, muito famoso devido a sua facilidade de implementação e sua ordem de complexidade O(*n*), onde *n* é o número de padrões (JAIN *et al.*, 1999). Fontana e Naldi (2009), descrevem a utilização do *k-means* a partir do conceito de centroides como protótipos representativos dos grupos, onde o centroide simboliza o centro de um grupo, sendo este calculado por meio da média de todos os objetos de um grupo.

O algoritmo beneficia-se da estratégia que utiliza o algoritmo de agrupamento de dados por *k-médias*. O propósito deste algoritmo está em encontrar a melhor

divisão de *P* dados em *K* grupos *Ci, i* = 1, ... *K*, de modo que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja a menor distância possível (PIMENTEL *et al.,* 2003). Jain *et al*. (1999) afirmam que um dos principais problemas deste algoritmo é a sensibilidade para a seleção da partição inicial, fazendo com que o algoritmo venha a convergir a um mínimo local do valor da função critério, nos casos onde a partição inicial, não seja a melhor partição escolhida.

3.2.3.6 *k-NN*

Para Cunningham e Delany (2007) a intuição implícita à classificação pelo algoritmo *k-NN* é bastante direta, os exemplos são classificados baseados nos rótulos das classes dos vizinhos mais próximos (*nearest neighbor*). No geral, é útil considerar mais de um vizinho, uma vez que a técnica é mais conhecida como classificação pelos k-vizinhos mais próximos, onde *k* vizinhos são usados na determinação da classe.

O treinamento de um classificador *k-NN* consiste apenas em armazenar os padrões apresentados durante essa fase. Para cada padrão classificado, é preciso realizar o cálculo da distância deste padrão para todos os padrões de treinamento, assim o algoritmo classifica um determinado padrão de entrada como a classe que corresponde à maioria dos *k* padrões de treinamento mais próximos daquele padrão (entrada), onde *k* é um valor natural normalmente pequeno (DUDA *et al*., 2000). Se tratando de classificações binárias, isto é, onde existem apenas duas classes, é aconselhável escolher um valor ímpar para *k*, removendo a possibilidade de empates (MADEIRO *et al.*, 2009).

3.2.3.7 Floresta Aleatória

O método Floresta Aleatória ou *Random Forest* trata-se de um algoritmo classificador que utiliza a técnica de árvores de decisão criado por Breiman (2001). Enquanto uma árvore possui o objetivo de construção total de uma estrutura a partir de uma base de dados, o *Random Forest* tem a finalidade de criação de diversas árvores de decisão utilizando apenas um subconjunto de atributos, selecionados de forma aleatória, a partir do conjunto original, contendo todos os atributos e que estes possuem um tipo de amostragem chamado de *bootstrap*, cujo tipo é com reposição,

ou seja, os atributos da base podem se repetir em mais de uma árvore dentre as diversas criadas pelo algoritmo, possibilitando uma melhor análise dos dados (NETO, 2014).

A partir de cada subconjunto criado, uma árvore de decisão é gerada. A construção dessas árvores acontece por meio de uma seleção de atributos aleatória dos subconjuntos, os quais são utilizados nos nós de cada uma das árvores geradas. Assim, uma floresta aleatória é um conjunto dessas árvores de decisão. Após a formação da floresta, há uma grande quantidade de árvores a serem testadas e todas contribuem para a classificação do objeto em questão, sendo esta, por meio de um voto sobre qual classe o atributo meta deve pertencer (NETO, 2014). Finalmente, a previsão da floresta aleatória é obtida por uma votação majoritária sobre as previsões de cada uma das árvores construídas (BEN-DAVID; SHALEV-SHWARTZ, 2014).

3.2.3.8 *Apriori*

O algoritmo *Apriori* é de fácil execução e muito simples, usado para extrair todos os conjuntos de itens frequentes em bancos de dados. O algoritmo (RAO; GUPTA, 2012) realiza diversas pesquisas em um banco de dados para encontrar conjuntos de itens frequentes onde *k*-conjuntos de dados são usados para gerar *k*+1-conjuntos de dados. Cada conjunto de itens *k* deve ser maior ou igual ao limite mínimo de suporte para ser a frequência. Caso contrário, o conjunto é chamado de conjunto de itens candidatos.

Primeiro, o algoritmo verifica o banco de dados para encontrar a frequência de conjuntos de 1 item, ou seja, que contém apenas um item, contando cada item no banco de dados. A frequência de conjuntos de 1 item é usada para encontrar os conjuntos de itens com 2 itens, que por sua vez é usada para encontrar conjuntos de 3 itens e assim por diante até que não haja mais *k* conjuntos de itens.

3.2.4 AVALIAÇÃO E COMPARAÇÃO DE ALGORITMOS DE AM

Segundo Oliveira (2016), modelos de aprendizagem não garantem o acerto total de suas previsões, quanto maior a dificuldade do problema mais complicado realizar as previsões de forma correta. Logo, é importante que se possa estimar

acertadamente a capacidade preditiva de um modelo, para saber qual o seu grau de confiabilidade quando aplicado aos cenários reais. O principal objetivo da avaliação de modelos preditivos busca mensurar a probabilidade de o modelo realizar previsões corretas quando uma nova entrada é apresentada. As próximas seções definem estes critérios e apresentam tópicos significativos para a realização de avaliação de modelos.

3.2.4.1 MÉTRICAS E MÉTODOS DE AVALIAÇÃO

A eficácia preditiva de um modelo em um conjunto de dados (no caso dos rotulados) consegue ser avaliada por meio de inúmeras métricas. Se o objetivo de uma aplicação é voltado a prever uma determinada classe, por exemplo, estimar se um aluno irá reprovar de ano, considera-se esta classe como positiva, enquanto se o aluno irá passar de ano considera a classe como negativa (classificador binário).

A partir disso, quatro medidas elementares foram estipuladas para servir como base para as métricas de avaliação dos modelos: verdadeiros positivos (TP) que são as observações classificadas como positivas e que possuem em sua classe o rótulo positivo; verdadeiros negativos (TN) que são as observações negativas classificadas de forma correta; falsos positivos (FP) que tratam exemplos classificados incorretamente como classes positivas; e, falsos negativos (FN) que descrevem exemplos classificados incorretamente com classe negativa. Estas medidas elementares possuem uma disposição visual expostas em uma tabela de contingência também chamada de matriz de confusão (OLIVEIRA, 2016).

Os valores encontrados pela matriz de confusão servem como parâmetros para derivar uma outra série de métricas para avaliação de modelos. O quadro 3 exibe um exemplo de uma matriz de confusão com os seus respectivos parâmetros.

Quadro 3 - Matriz de Confusão

Resultado do Modelo	**Real**		
	Positivo (D_+)	**Negativo** (D_-)	**Total**
Positivo (T_+)	$a\ (TP)$	$b\ (FP)$	$a + b$
Negativo (T_-)	$c\ (FN)$	$d\ (TN)$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Fonte: Mazucheli *et al.* (2008)

Segundo Mazucheli *et al*. (2008), aplicando um exemplo aos parâmetros apresentados no quadro 3 tem-se as seguintes métricas:

- (a): representa a quantidade de instâncias positivas classificadas corretamente como positivas, isto é, os verdadeiros positivos (TP);
- (b): representa a quantidade de instâncias negativas classificadas incorretamente como positivas, ou seja, revela o total de falsos-positivos (FP);
- (c): representa a quantidade de instâncias positivas classificadas incorretamente como negativas, ou seja, a quantidade total de falsos-negativos (FN);
- (d): representa a quantidade de instâncias negativas classificadas corretamente como negativas, isto é, os verdadeiros negativos (TN);
- $(a + c)$: representa a quantidade total de instâncias positivas;
- $(b + d)$: representa a quantidade total de instâncias negativas;
- $(a + b)$: representa a quantidade de instâncias identificadas pelo modelo como positivas;
- $(c + d)$: representa a quantidade de instâncias identificadas pelo modelo como negativas.

A capacidade de previsão de um modelo está diretamente relacionada com suas medidas de desempenho, obtidas por meio de cálculos utilizando os parâmetros visualizados no quadro 3. Dentre estes, tem-se métricas como a sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo e acurácia.

De acordo com Oliveira (2016) a sensibilidade é a razão entre os verdadeiros positivos e todos os positivos, isto é, quantos dos casos positivos o modelo conseguiu prever corretamente. Mazucheli *et al.* (2008) identifica a sensibilidade, de acordo com o quadro 3, por meio da equação (1):

$$S = P(T_+|D_+) = \frac{TP}{TP + FN} = \frac{a}{a + c} \quad (1)$$

A especificidade trata a razão entre as instâncias classificadas corretamente como negativas e todos os casos negativos (OLIVEIRA, 2016). Essa métrica expressa a probabilidade de o modelo retornar um resultado negativo dado que a instância seja

livre da característica em questão. Considerando o quadro 3, tem-se a especificidade dada em (2) por (MAZUCHELI *et al.*, 2008):

$$E = P(T_-|D_-) = \frac{TN}{TN + FP} = \frac{b}{b + d} \tag{2}$$

O valor preditivo positivo se dá pela taxa de instâncias verdadeiras positivas com relação a todas as classificadas como positivas (OLIVEIRA, 2016). Considerando a quadro 3, o valor preditivo positivo é dado em (3) por (MAZUCHELI *et al.,* 2008):

$$VPP = P(D_+|T_+) = \frac{TP}{TP + FP} = \frac{a}{a + b} \tag{3}$$

O valor preditivo negativo se dá pela taxa de instâncias verdadeiras negativas com relação as classificadas como negativas (OLIVEIRA, 2016). Considerando o quadro 3, o valor preditivo negativo é dado (4) por (MAZUCHELI *et al.,* 2008):

$$VPN = P(D_-|T_-) = \frac{TN}{TN + FN} = \frac{d}{c + d} \tag{4}$$

A acurácia é estabelecida como a proporção de acertos de um modelo, tanto as instâncias positivas quanto as negativas, isto é, trata a proporção dos verdadeiros positivos e verdadeiros negativos em relação a todos os resultados possíveis. Ela também é chamada de capacidade total de acerto do modelo (CTA) (MAZUCHELI *et al.,* 2008). Para Oliveira (2016), a acurácia não é uma métrica ideal para orientar a escolha de um modelo uma vez que elas se tornam métricas inadequadas quando se trabalha com uma distribuição desbalanceada das classes nos dados. Considerando o quadro 3, Mazucheli *et al.* (2008) calcula a acurácia como em (5):

$$CTA = \frac{TP + TN}{TP + FP + TN + FN} = \frac{a + d}{a + b + c + d} \tag{5}$$

De acordo com Oliveira (2016), quando determinado classificador possui uma resposta contínua ou uma estimativa de probabilidade, consegue-se executar uma variação em um limiar de decisão em que as instâncias em lados opostos a esse limiar apresentam classes diferentes. Por meio dessa variação de limar é possível desenhar

uma curva identificando as taxas de verdadeiros positivos (sensibilidade) e de falsos positivos (inverso da especificidade) em cada limiar. Essa curva é chamada de ROC (*Receiver Operating Characteristic*), que descreve uma comparação de duas características receptoras (sensibilidade e especificidade) conforme mudança de critérios, indicando os *trade-offs* entre os verdadeiros positivos (benefícios) e os falsos positivos (custos).

3.2.4.2 MÉTODOS DE VALIDAÇÃO DE MODELOS: REAMOSTRAGEM

Quando se trata sobre os métodos de avaliação de modelos, o cálculo das métricas são fatores importantes, pois trabalham com um conjunto de dados provenientes da base.

O emprego dos cálculos das métricas nunca devem ser aplicados em conjuntos de dados que foram utilizados para o treinamento de um modelo, uma vez que as dificuldades encontradas na geração de modelos preditivos relatam a tendência de muitos algoritmos se sobreajustarem aos dados de treinamento. Sendo assim, quando utilizado esses conjuntos eles podem apresentar bom desempenho, porém sem garantias que tal desempenho seja o mesmo para dados ainda não classificados e o algoritmo pode não conseguir generalizar para novos dados de entrada (OLIVEIRA, 2016).

As medidas correspondentes a erros que sejam adquiridos por meio do conjunto de teste podem ser consideradas como o erro verdadeiro, uma vez que se aproximam do erro populacional se o tamanho do conjunto de teste for suficientemente grande. Quando uma amostra de teste alcança um tamanho de 1000, as estimativas são acuradas e com 5000 instâncias a estimativa da amostra é quase idêntica ao erro verdadeiro da população (WEISS, KULIKOWSKI, 1991).

É importante estimar o erro verdadeiro de uma amostra para ser aleatória. Em problemas reais, é fornecida uma amostra única de uma população de tamanho *n* e a partir desta única amostra têm-se a tarefa de estimar o erro verdadeiro para esta população e não para todas as populações (BARANAUSKAS, MONARD, 2000).

De acordo com Baranauskas e Monard (2000), existem diversos paradigmas para estimar o erro verdadeiro de uma população entre eles o *Holdout* é uma das principais técnicas. Esse estimador divide o conjunto de dados em uma porcentagem

fixa de instâncias *p* para treinamento e (1 – *p*) para teste, geralmente *p* > 1/2 do conjunto. Alguns valores usados para *p* são *p* = 2/3 e (1 – *p*) = 1/3, entretanto, não há fundamentação teórica sobre esses valores.

Outro estimador é o *k-fold-cross-validation* (validação cruzada), nesta validação o conjunto de dados é dividido aleatoriamente em *k* partições mutuamente exclusivas (*folds*). Após isso, utiliza-se uma das partições como conjunto de teste e os restantes para treinamento. O processo é repetido variando a partição de teste de forma que no fim do processo todas as partições são utilizadas como teste exatamente uma vez. O erro gerado a partir da validação cruzada é a média dos erros sobre todas as *k* partições.

Um caso especial de validação cruzada é conhecido como *Leave-one-out*. Procedimento computacionalmente caro, frequentemente em pequenos conjuntos de dados. Para uma amostra de tamanho *n* são utilizadas *n* – 1 instâncias para treinar o modelo (sendo *n* o total de instâncias) e 1 instância para teste do modelo a cada iteração. Esse processo é repetido *n* vezes sem criar um preditor e deixa uma instância de fora do conjunto. O erro é descrito como a soma dos erros nas instâncias únicas de teste divididas por *n* (BARANAUSKAS, MONARD, 2000).

3.3 MINERAÇÃO DE DADOS EDUCACIONAIS

Autores fazem uso e recomendam a aplicação das técnicas da aprendizagem de máquina para o estudo e avaliação de problemas na área educacional. A busca pelos trabalhos foi embasada em pesquisas científicas presentes na literatura, os repositórios de busca foram *sites* como Google Acadêmico, IEEE *Xplore*, *ResearchGate*, *SciELO*, *Journal of the Brazilian Computer Society* (JBCS), repositórios de universidades, entre outros.

As pesquisas focaram em trabalhos acadêmicos (dissertações, teses), relatórios técnicos e científicos, livros, artigos em conferências e *journals*. As palavras-chave de maior uso para a busca se deram por “Educação”, “Inteligência Artificial”, “Mineração de Dados”, “Mineração de dados Educacionais”, “Aprendizagem de Máquina”, “Aprendizagem de Máquina aplicada à Educação”, e “Inteligência Artificial na Educação”. Além disso, estas palavras foram traduzidas para o inglês também para uso na busca por trabalhos, a tradução se deu a partir de pesquisas encontradas na

língua inglesa para que não houvessem erros de tradução, como “*Education*”, “*Artificial Intelligence*”, “*Data Mining*”, “*Educational Data Mining*”, “*Machine Learning*”.

Alguns critérios foram adotados para a escolha dos trabalhos a serem lidos e analisados como pesquisas consolidadas na área de inteligência artificial e aprendizagem de máquina aplicada à educação, artigos e trabalhos de congressos e eventos da área de mineração de dados e AM educacional e relevância dos artigos como locais de publicação (periódicos, revistas, universidades). Os trabalhos selecionados foram: Hijazi e Naqvi (2006), Baker, Barnes e Beck (2008), Kampff *et al*. (2008), Shih, Koedinger e Scheines (2010), Lopez *et al*. (2012), Bazaldua, Baker e San Pedro (2014), Rau, Mason e Nowak (2016), Fang *et al*. (2018) e Yang *et al.* (2019).

Hijazi e Naqvi (2006) conduziram um estudo sobre o desempenho de estudantes, com base numa hipótese, enquadrada por eles, declarada como “Ações do aluno em relação à frequência às aulas, horas gastas no estudo diariamente após a faculdade, renda familiar dos alunos, idade da mãe e educação da mãe está significativamente relacionada ao desempenho do aluno”. A partir do uso do modelo de regressão linear simples, verificou-se que fatores como a educação da mãe e a renda familiar do aluno estão altamente correlacionados com o desempenho acadêmico do aluno.

Baker, Barnes e Beck (2008) descrevem o conceito da mineração de dados educacionais como o processo de conversão de dados de sistemas educacionais em informações que podem ser usadas por desenvolvedores de softwares educativos, estudantes, professores, pais e outros pesquisadores educacionais.

A Mineração de Dados Educacionais (MDE) emergiu-se como uma área de pesquisa independente a alguns anos, mais especificamente em 2008, junto a criação da Conferência Internacional em Mineração de Dados Educacionais, além do jornal de Mineração de Dados Educacionais (BAKER, 2010). Com o surgimento destes órgãos, novas pesquisas estão surgindo e sendo disponibilizadas pela conferência internacional de Mineração de Dados Educacionais, trazendo os mais recentes avanços e descobertas desta área.

Kampff *et al*. (2008) propuseram aplicar mineração de dados na construção de alertas em ambientes virtuais de aprendizagem para identificar perfis de alunos com risco de evasão ou reprovação, de forma a promover alterações nos AVA que facilitem a implementação de alertas como apoio a prática docente.

Shih, Koedinger e Scheines (2010) desenvolveram uma pesquisa para descobrir táticas de estudos de alunos utilizando descoberta não-supervisionada, em que o algoritmo incorpora medidas educacionais no nível do aluno diretamente no processo de aprendizagem. Os dados para este estudo derivam da plataforma tutora de ensino *Geometry Cognitive Tutor*. Os algoritmos de comportamento do aluno, além de prever o ganho de aprendizagem, sugerem que os alunos que aprendem melhor tendem a fazer tentativas mais persistentes ao invés de utilizar a ajuda do software.

Lopez *et al*. (2012) compararam técnicas de agrupamento para prever notas finais baseadas nas respostas de participação de alunos em fóruns, buscando a existência de uma correlação entre as notas finais dos estudantes com a participação dos mesmos nos fóruns educativos.

Bazaldua, Baker e San Pedro (2014) utilizaram dados de respostas de atividades de alunos, submetidos a um ambiente de aprendizagem online na área da matemática, com o objetivo de encontrar regras de associação para avaliar as relações entre os comportamentos afetados e desmembrados dos alunos durante o uso do sistema. Após a aplicação do modelo de mineração, foram utilizadas diferentes métricas para selecionar as melhores regras e por fim compará-las com a opinião de avaliação de especialistas da área.

Rau, Mason e Nowak (2016) utilizaram algoritmos de AM para analisar a similaridade de representações visuais de moléculas químicas apresentadas a alunos de graduação em Química de uma universidade dos EUA. Dada a representação visual de uma molécula os alunos julgavam sua similaridade comparada a outras. A partir das respostas, eles aplicaram duas abordagens para medir a similaridade descritas como: aprendizado de similaridade por *ranking*, e escala multidimensional não métrica. Por fim, eles não conseguiram identificar uma grande diferença entre os dois modelos, porém recomendaram o uso de uma das abordagens por otimizar o erro de previsão, cuja característica mede objetivamente a qualidade entre os modelos.

Fang *et al*. (2018) utilizaram as técnicas de agrupamento *k-Means* e Análise Hierárquica para identificar padrões na aprendizagem de adultos com baixa alfabetização a partir da interação com um sistema tutor inteligente. Eles conseguiram separar os alunos em grupos, entre: leitores proficientes, leitores com dificuldades, leitores conscientes e leitores desengajados, baseando-se nos seus padrões de comportamento em relação as atividades praticadas, além de identificarem pontos

fortes e fracos dos alunos podendo melhorar materiais específicos da aprendizagem para cada indivíduo.

Yang *et al.* (2019) utilizaram sistemas tutores inteligentes para alunos em salas de aula coletando dados afetivos e aplicando estes a detectores afetivos para prever estados afetivos dos alunos a partir de um conjunto de características das atividades executadas. Isso possibilita identificar pontos em que o aluno possa se sentir confuso ou entediado de forma a aprimorar a abordagem educativa do STI. O Quadro 4 apresenta um comparativo entre as pesquisas discutidas nesta seção.

Quadro 4 - Comparativo entre trabalhos relacionados

Autor (es)	Ano	Base de Dados	Algoritmo de Mineração	Resultados do Trabalho
Hijazi e Naqvi	2006	Dados de alunos coletados de diferentes universidades particulares	Regressão Linear Simples	Identificaram atributos relevantes correlacionados ao bom/mau desempenho acadêmico de alunos.
Kampff *et al.*	2008	Dados coletados de uma Universidade atuante na modalidade EaD	Árvore de decisão	Identificaram atributos referentes a perfis de alunos para gerar alertas de forma a auxiliar a tomada de decisão docente.
Shih, Koedinger e Scheines	2010	Respostas da plataforma de ensino *Geometry Cognitive Tutor*	*Bottom-out hint* (Modelo de formulação de dicas de baixo pra cima)	Identificaram ganho de aprendizagem de alunos por meio de características específicas dos mesmos.
Lopez *et al*.	2012	Dados coletados de um fórum da plataforma *Moodle* do curso de Engenharia da Computação de uma Universidade	Meta-Classificador utilizando técnica de agrupamento para classificação	Identificaram que a participação dos alunos em fóruns pode impactar no seu desempenho final.
Bazaldua, Baker e San Pedro	2014	Respostas de atividades dentro do sistema educacional online ASSISTments	*Apriori* utilizando o pacote *arules* para descobrir regras de associação	Identificaram a acurácia das métricas de avaliação de modelos de regras de associação na geração das mesmas.
Rau, Mason e Nowak	2016	Respostas de atividades propostas a alunos do curso de química de uma universidade dos EUA	Aprendizado de similaridade por *ranking*; Escala multidimensional não métrica utilizando *k-fold cross validation*	Obtiveram resultados importantes fornecendo caminhos mais formais para designers construírem aprendizagens perceptivas, contribuindo assim para a pesquisa em modelagem cognitiva.

Fonte: Autoria própria

Quadro 4 - Comparativo entre trabalhos relacionados

Autor (es)	Ano	Base de Dados	Algoritmo de Mineração	Resultados do Trabalho
Fang *et al.*	2018	Respostas de alunos providos de atividades de diferentes áreas conduzidas por professores e sistema *AutoTutor*	*k-means* e Análise Hierárquica	Identificaram agrupamentos de alunos com base nos seus padrões de comportamento em relação a prática de atividades, além de identificar pontos fortes e fracos podendo melhorar pontos específicos da aprendizagem.
Yang *et al.*	2019	Registros de estados afetivos, de efeito e expressões de alunos durante aprendizagem com auxílio de um sistema tutor inteligente	Detectores *sensor-free* com uso de classificadores baseados em Aprendizado de Máquina	Utilizaram STIs para coletar dados de estados afetivos de alunos e aplicaram detectores *sensor-free* para prever esses estados a partir de um conjunto de características de atividades.

Fonte: Autoria própria

Observa-se que os trabalhos apresentados buscam seus dados para análise, no geral, a partir de informações e respostas de alunos. Os autores utilizaram métodos como pesquisa ao público (desde que o público esteja inserido no ambiente educacional), dados de alunos de universidades, e a maioria das pesquisas extraíram as informações de ferramentas educacionais, pois essas plataformas produzem grande quantidade de dados e fornecem a possibilidade de encontrar informações importantes.

Também se pode concluir que praticamente todos os trabalhos implementam diferentes algoritmos, pois existem diversas implementações para um mesmo segmento da AM (classificação, agrupamento, associação), o que possibilita diversificar a escolha dos algoritmos para utilização. Os algoritmos de classificação e agrupamento aparecem em maior quantidade de aplicações e alguns fazem uso de algoritmos de associação. Ainda, como resultados, as pesquisas buscam identificar critérios para avaliação e melhorias no ensino, recuperação de informações e obtenção de critérios para aprimoramento de ferramentas.

3.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este Capítulo apresentou informações sobre a aprendizagem de máquina. O desenvolvimento desta área mostra a motivação buscada pelos humanos na realização de tarefas complexas, e por meio desta, descreve o potencial da AM na

resolução destes problemas. Ainda, a preocupação de se transmitir adequadamente a capacidade de inteligência humana para as máquinas, caracteriza outro fomento da sociedade cujo propósito visa a criação de algoritmos de AM que retornem os resultados precisos.

Também foram listados alguns dos principais algoritmos de AM e discutido sobre seu funcionamento e empregabilidade. Além destes, as abordagens referentes aos métodos de comparação e avaliação das técnicas de aprendizagem de máquina foram retratadas, pois representam elementos chaves nas análises que foram realizadas neste trabalho.

Diversos segmentos da área educacional salientam as grandes vantagens da utilização da AM em busca de melhorias na educação. Estas melhorias visam não apenas desenvolver novas técnicas de aprendizado, elas buscam melhorar o ensino como um todo, desenvolvendo meios que possam avaliar o sistema educacional, os alunos e professores. Logo, o próximo capítulo traz a apresentação à metodologia deste trabalho, sendo está sobre a aplicação da Aprendizagem de Máquina na tentativa pela descoberta de padrões e informações úteis em grupos de alunos do ensino superior na pandemia em relação à prática dos estudos na modalidade EAD.

4 APLICAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA: UM ESTUDO NA ÁREA EDUCACIONAL

Este Capítulo apresenta a execução dos passos para aplicação dos algoritmos de aprendizagem de máquina em bases educacionais. As etapas de mineração de dados discutidas no capítulo 3 são aqui empregadas e detalhadas.

Muitas são as informações providas dos dados educacionais que podem ser recuperadas e utilizadas para obter novas informações, gerando maior aproveitamento destes dados. O estudo de caso abordou a questão da educação no período de pandemia da COVID-19, pois houve a necessidade da suspensão do método de ensino presencial em todos os níveis de aprendizado, e muitas instituições de ensino não estavam preparadas para lidar com a situação deste cenário.

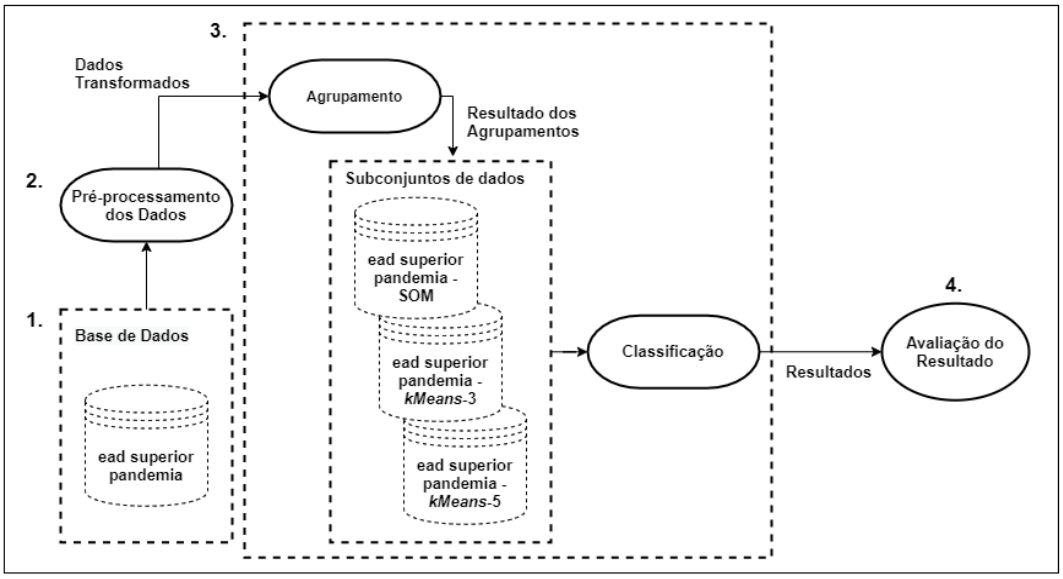
Instituições implantaram o ensino a distância objetivando a continuidade das aulas para evitar a perda do ano letivo e a evasão/reprovação de estudantes, entretanto, muitos inseridos nesse contexto podem não ter o devido acesso, seja por fatores estruturais, econômicos ou psicológicos. Assim, a pesquisa buscou identificar esses fatores que influenciam alunos a optarem pelo ensino remoto ou não. O público alvo foram alunos do ensino superior que tiveram suas aulas suspensas submetidos a uma pesquisa de opinião sobre o ensino remoto. As respostas da pesquisa puderam ser posteriormente transformadas em dados para utilização dos algoritmos de aprendizagem de máquina.

A Seção 4.1 aborda a metodologia utilizada para a realização do experimento. A Seção 4.2 descreve sobre a coleta dos dados e criação da base. A Seção 4.3 relata sobre as tarefas do pré-processamentos dos dados. A Seção 4.4 trata da aplicação dos algoritmos de AM na base desenvolvida. A Seção 4.5 comenta as tarefas do pós-processamento. Por fim, a Seção 4.6 apresenta as considerações finais do capítulo.

4.1 METODOLOGIA

A metodologia utilizada para a realização dos experimentos é formada por quatro etapas principais e apresentada na Figura 5, sendo: 1. Coleta dos Dados e Criação da Base de Dados, 2. Pré-Processamento dos Dados, 3. Aplicação de Algoritmos de Mineração de Dados e 4. Avaliação dos Resultados.

Figura 5 – Uma visão geral da metodologia



Fonte: Autoria própria

A etapa 1 consiste do entendimento do problema para criação do conjunto de dados. A etapa 2 discorre sobre o tratamento dos dados no pré-processamento. A etapa 3 apresenta a execução dos experimentos utilizando os algoritmos de AM na base criada. Por fim, a etapa 4 discute os resultados. As próximas seções detalham cada uma dessas etapas.

4.2 COLETA DOS DADOS E CRIAÇÃO DA BASE DE DADOS

O conjunto de dados para este trabalho foi obtido por meio de uma pesquisa com alunos do ensino superior. Os dados foram coletados a partir de um formulário desenvolvido na plataforma *Google Forms*, o mesmo foi divulgado via redes sociais em páginas de universidades.

A base foi nomeada como *ead_superior_pandemia*, formada por 30 questões, sendo estas os respectivos atributos. O modelo das questões foi de múltipla escolha com algumas sendo de resposta única e outras de resposta múltipla. Algumas questões foram oportunizadas ao aluno o preenchimento do campo “Outros”, por exemplo, a pergunta indagava se o aluno possui alguma doença psicológica que afetasse seu desempenho, e apresentava algumas alternativas, além de permitir ao aluno informar alguma outra doença caso a mesma não estivesse disponível nas opções. Um total de 483 respostas foram enviadas pelos alunos do ensino superior presencial (verificar Apêndice A) e cada resposta constitui os registros da base

ead_superior_pandemia. Um exemplo de pergunta formulada para posterior transformação no formato atributo-valor é apresentado no Quadro 5 abaixo.

Quadro 5 - Exemplo de uma pergunta desenvolvida para o formulário da pesquisa de opinião

Pergunta	*Qual seu grau de adaptabilidade e compreensão das ferramentas tecnológicas EAD?*
Alternativas	• Fácil adaptação, as ferramentas são claras e fáceis de manusear
	• Adaptação moderada, poucas dificuldades em relação as seções e conteúdos
	• Difícil adaptação, na maioria das vezes não encontro o que desejo e não consigo usar a ferramenta

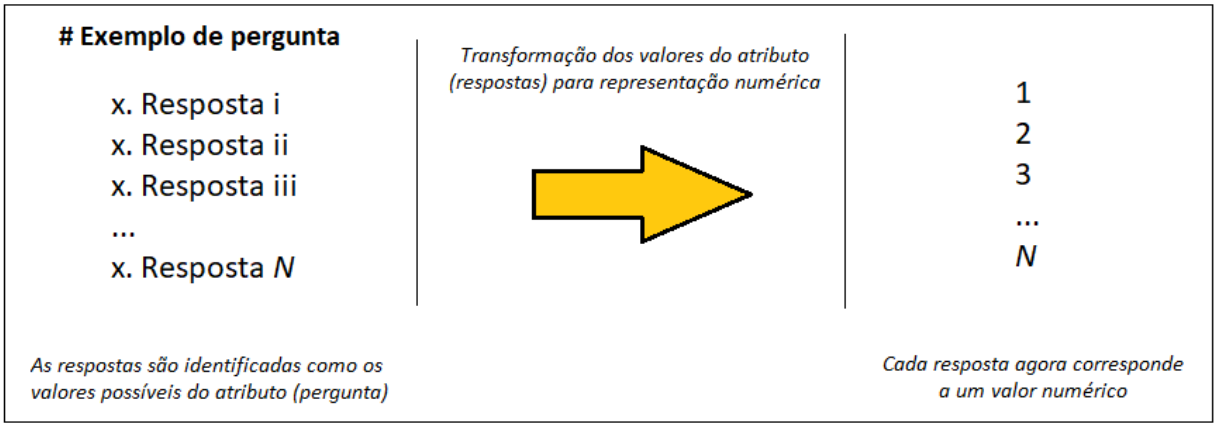
Fonte: Autoria própria

Nesse caso apenas uma resposta foi solicitada ao aluno, que se torna um valor para o atributo “Adaptabilidade as ferramentas EAD” derivado desta questão, com as opções “Fácil adaptação”, “Média adaptação” ou “Difícil adaptação”, por exemplo.

4.3 PRÉ-PROCESSAMENTO

A base é formada, quase que em sua totalidade, por atributos nominais. Assim, fez-se necessário transformar os valores dos atributos para posterior aplicação dos algoritmos de Aprendizado de Máquina. A Figura 6 expressa como foi realizado esse processo.

Figura 6 – Processo utilizado para codificação das respostas do formulário



Fonte: Autoria própria

Os valores foram codificados criando-se uma representação numérica para cada valor do atributo, isso não altera de fato os valores reais do atributo, apenas modifica-os para utilização dos algoritmos.

Após este processo, foi aplicado ao conjunto a normalização min-max (HAN, KAMBER, 2006) para normalizar os valores e submetê-los a execução dos algoritmos. Segundo os autores, normalizar os dados tenta dar a todos os atributos do conjunto de dados um peso igual. A normalização é particularmente útil para algoritmos de classificação que envolvem redes neurais ou medições de distância como classificação de vizinho mais próximo e agrupamento. Ainda, para os métodos baseados em distância, a normalização ajuda a evitar que atributos com faixas inicialmente grandes (como valores de renda, por exemplo) superem atributos de faixas menores (como atributos binários), além de ser útil quando não há conhecimento prévio dos dados.

O processo da normalização min-max executa uma transformação linear nos valores de dados originais. Suponha que min_A e max_A sejam os valores mínimo e máximo de um atributo, *A*. A normalização mapeia um valor, v_i de *A* para v'_i, no intervalo [*novo_min, novo_max*], que corresponde ao valor de intervalo que se deseja normalizar, por meio da equação 6 (HAN, KAMBER, 2006).

$$v_i^1 = \frac{v_i - min_A}{max_A - min_A}(new\_min_A - new\_min_A) + new\_min_A \quad (6)$$

Aplicando a fórmula a um exemplo da base criada, temos os valores de mínimo e máximo para o atributo denominado ‘classe_social’ sendo 0 e 3 respectivamente. Se queremos mapear o atributo ao intervalo de [0, 1], aplicando a normalização min-max, o valor 1, que corresponde a uma das opções de resposta do atributo, é transformado para $\frac{1-0}{3-0}(1-0)+0 = 0{,}3333333$.

4.4 MINERAÇÃO DE DADOS

Os experimentos foram realizados no *software Weka* versão 3.8.4. O *Weka* (*Waikato Environment for Knowledge Analysis*) trata-se de um pacote composto por um conjunto de implementações de diversos algoritmos de Mineração de Dados (BOUCKAERT *et al*., 2018). Os algoritmos utilizados nos experimentos foram executados utilizando os parâmetros de configuração padrão do *Weka*.

Após a etapa de pré-processamento os dados foram submetidos a aplicação dos algoritmos de aprendizagem de máquina. Os algoritmos *Self-organizing maps* (SOM), e *k-means* foram utilizados para geração de agrupamentos. O algoritmo SOM foi usado para determinar automaticamente uma quantidade de grupos, sendo gerado 4 agrupamentos. Para o *k-means*, as escolhas dos valores de *k* consideraram a quantidade de agrupamentos gerados anteriormente pelo SOM, sendo estes *k*=3 e *k*=5, a fim de trabalhar em proporções próximas um do outro.

Como resultado dos agrupamentos, três subconjuntos foram gerados, cada um com o seu respectivo atributo meta relacionado aos *clusters* gerados por cada algoritmo. Cada subconjunto foi submetido a três algoritmos de classificação sendo estes: JRip, PART e J48. Estes algoritmos foram utilizados para gerar regras referentes aos grupos formados na fase de agrupamento. Os resultados foram avaliados por meio das estimativas de acurácia dos modelos, utilizando o método de validação cruzada estratificada com 10 partições (10 *fold-cross-validation*) em cada teste com os classificadores.

4.5 PÓS-PROCESSAMENTO

A etapa de pós-processamento aborda a avaliação, análise e discussão dos resultados obtidos. Para avaliação dos algoritmos foi observado as métricas de acuracidade geradas durante realização dos experimentos em cada um dos testes. As regras geradas como resultado são então analisadas e posteriormente discutidas, de forma a auxiliar na compreensão e unificar os resultados para identificação das informações relevantes encontradas.

4.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este Capítulo abordou sobre o processo de aplicação de algoritmos de aprendizagem de máquina em bases educacionais. As etapas necessárias para execução dessas ferramentas foram detalhadas como o tratamento dos dados, o formato dos experimentos com os algoritmos de AM e como a fase de resultados trabalha com a discussão destes para apresentação clara de padrões e informações úteis extraídas por meio da mineração desses dados.

5 RESULTADOS

A etapa de avaliação e discussão dos resultados é importante porque além de encontrar padrões relevantes nos dados de entrada, é preciso entender e julgar a utilidade do conhecimento extraído. Esta análise é realizada por métodos, geralmente estatísticos, que observa critérios de desempenho do processo e considera fatores como a precisão e a representação do conhecimento extraído (DA ROCHA, 2000).

Este Capítulo apresenta os resultados desta pesquisa. A Seção 5.1 lista os resultados do agrupamento e classificação representados pelas suas saídas, sendo para o agrupamento seus grupos e quantidade de instâncias e para a classificação suas taxas de acerto. A Seção 5.2 detalha os resultados da aplicação dos classificadores na base resultante do algoritmo SOM, seguida pela Seção 5.3 que detalha os resultados desta aplicação nas bases resultantes do algoritmo *k-means* com *k*=3 e *k*=5. A Seção 5.4 aborda uma discussão dos resultados encontrados. A Seção 5.5 apresenta como estes resultados podem ser replicados para outros domínios da área educacional, isto é a exigência do edital em que esta pesquisa foi contemplada com bolsa. Por fim, a Seção 5.6 trata das considerações finais do capítulo.

5.1 RESULTADO DOS ALGORITMOS DE AGRUPAMENTO E CLASSIFICAÇÃO

Para representar os resultados dos experimentos utilizará a sigla C*i*-ALG em que *i* corresponde a variação da quantidade de *cluster* gerados pelos algoritmos e ALG é o algoritmo de agrupamento utilizado. O algoritmo SOM gerou 4 grupos, logo têm-se as seguintes representações: C0-SOM, C1-SOM, C2-SOM e C3-SOM. Para o algoritmo *k-means*, têm-se: C0-k3, C1-k3 e C2-k3 usando *k*=3 e C0-k5, C1-k5, C2-k5, C3-k5 e C4-k5 para *k*=5. A Tabela 1 apresenta a quantidade de instâncias atribuídas, para cada grupo gerado, pelos algoritmos de agrupamento.

Tabela 1 - Resultado dos agrupamentos com respectivos *clusters* e quantidade de instâncias

Algoritmos	**SOM**				***k*=3**			***k*=5**				
Grupos	**C0-SOM**	**C1-SOM**	**C2-SOM**	**C3-SOM**	**C0-k3**	**C1-k3**	**C2-k3**	**C0-k5**	**C1-k5**	**C2-k5**	**C3-k5**	**C4-k5**
Quantidade de Instâncias	138	150	93	102	125	154	204	91	67	119	165	41

Fonte: Autoria própria

A Tabela 2 apresenta a taxa de acerto dos algoritmos JRip, PART e J48, logo após, a Figura 7 apresenta um gráfico comparando essas medidas.

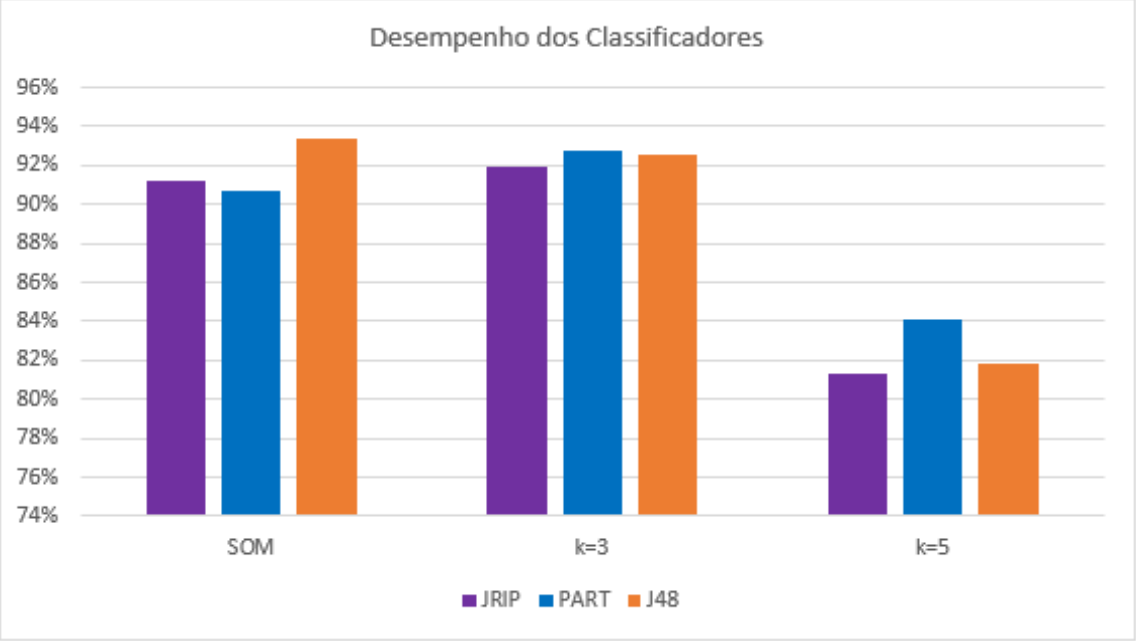
Tabela 2 - Desempenho dos classificadores baseado em suas taxas de acerto

Classificador	Taxa de Acerto		
	SOM	***k=3***	***k=5***
JRIP	91.25%	91.92%	81.36%
PART	90.68%	92.75%	84.05%
J48	93.37%	92.54%	81.78%

Fonte: Autoria própria

Os classificadores foram aplicados para cada um dos resultados dos algoritmos de agrupamento, ou seja, os algoritmos JRIP, PART e J48 foram executados para o SOM e para o *k-means* com *k*=3 e com *k*=5, respectivamente.

Figura 7 – Comparação dos classificadores em relação a suas Taxas de Acerto



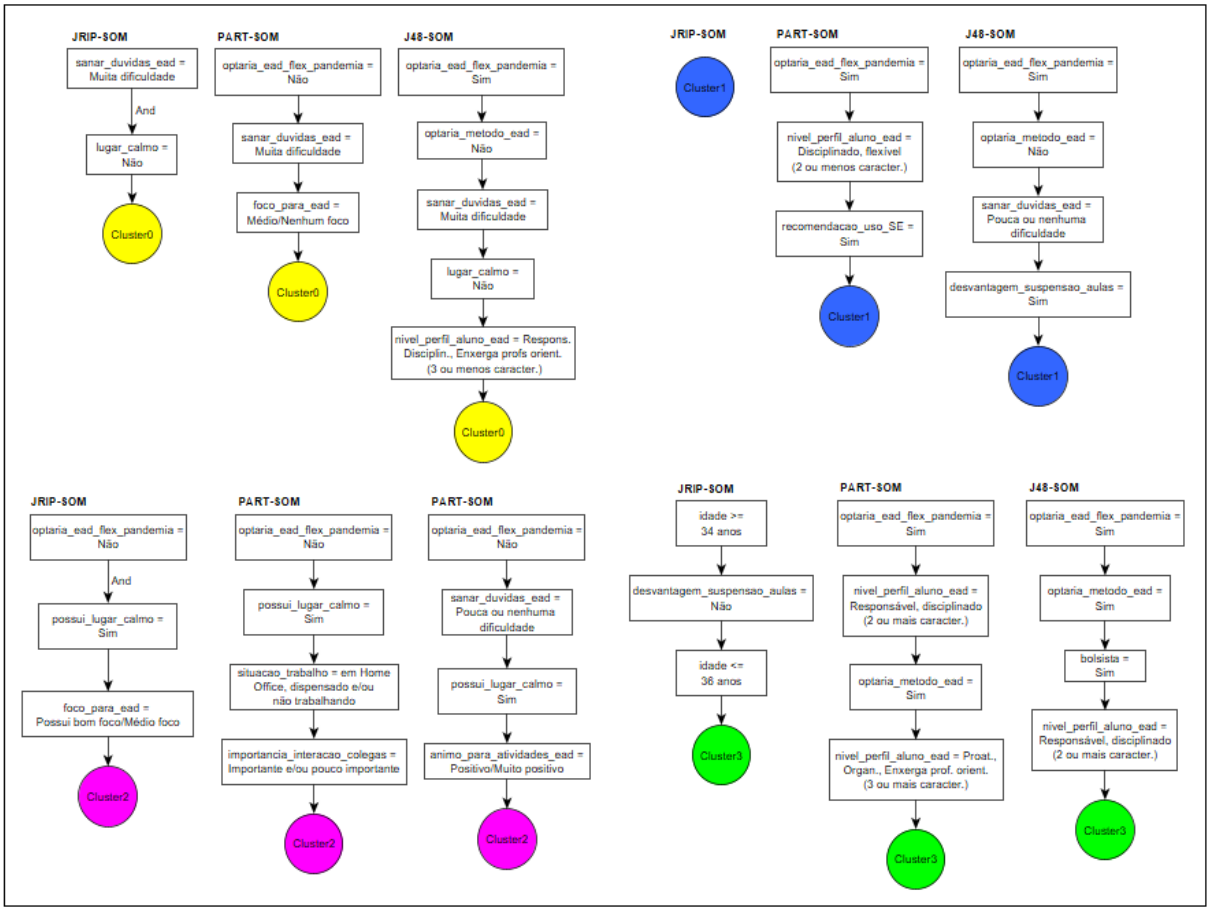
Fonte: Autoria própria

Nota-se que os algoritmos mostraram comportamento comum em relação as suas taxas de acerto. Entretanto para os grupos gerados pelo *k-means* com *k* = 5, a métrica obteve uma significante queda, podendo inferir que para resultados relevantes a quantidade de 3 a 4 grupos se mostra mais ideal.

Como descrito na metodologia, após a execução dos algoritmos de agrupamento, criaram-se atributos classe para a base. Dessa forma, foi possível executar os algoritmos de classificação para gerar as regras destes grupos. Para uma melhor análise dos resultados foi realizado uma visualização gráfica das regras utilizando diferentes cores para cada agrupamento. Como foram geradas diferentes

quantidades de grupos em cada um dos algoritmos de agrupamento a ordem das cores se manteve a mesma em todos até que um destes diferenciasse a quantidade, em que uma nova cor foi adicionada para um algoritmo que apresentou uma quantidade maior de grupos que outro. Este esquema foi útil para visualizar melhor as correlações existentes entre os grupos e as regras geradas por estes. A Figura 8 apresenta este esquema de cores considerando como exemplo os grupos gerados pelo SOM.

Figura 8 – Esquema de representação de cores para cada grupo e suas regras



Fonte: Autoria própria

Neste exemplo, os agrupamentos pertencem ao algoritmo SOM, o mesmo foi realizado para os outros algoritmos de agrupamento. Cada grupo corresponde a uma cor, nota-se ainda que cada regra pertence a um algoritmo classificador diferente, possibilitando identificar informações e características comuns entre as regras geradas em cada classificador

5.2 RESULTADOS COM O ALGORITMO SOM

O algoritmo SOM gerou quatro agrupamentos: i) Grupos C0-SOM e C2-SOM se mostraram contrários a aplicação do ensino a distância tanto durante a época de pandemia quanto fora dela; ii) C1-SOM e C3-SOM se mostraram um pouco mais favor, sendo o C1-SOM a favor do EAD apenas durante a época de pandemia e o C3-SOM a favor em ambas as situações. As Figuras 9, 10 e 11 apresentam os resultados da aplicação dos algoritmos JRip, PART e J48 respectivamente.

Figura 9 – Resultados do classificador JRip para o SOM

```
=== Summary ===

Correctly Classified Instances         444               91.9255 %
Incorrectly Classified Instances        39                8.0745 %
Kappa statistic                          0.8911
Mean absolute error                      0.062
Root mean squared error                  0.1978
Relative absolute error                 16.7617 %
Root relative squared error             45.9709 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,899    0,029    0,925      0,899    0,912      0,877    0,936     0,859     cluster0
                0,940    0,018    0,959      0,940    0,949      0,927    0,967     0,919     cluster1
                0,903    0,028    0,884      0,903    0,894      0,868    0,959     0,789     cluster2
                0,931    0,031    0,888      0,931    0,909      0,884    0,957     0,832     cluster3
Weighted Avg.   0,919    0,026    0,920      0,919    0,919      0,893    0,955     0,859

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 124   1   9   4 |   a = cluster0
   2 141   0   7 |   b = cluster1
   8   0  84   1 |   c = cluster2
   0   5   2  95 |   d = cluster3
```

Fonte: Autoria própria

A acuracidade neste teste foi de 91,9255%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 10 – Resultados do classificador PART para o SOM

```
=== Summary ===

Correctly Classified Instances         438               90.6832 %
Incorrectly Classified Instances        45                9.3168 %
Kappa statistic                          0.8744
Mean absolute error                      0.0515
Root mean squared error                  0.2078
Relative absolute error                 13.9128 %
Root relative squared error             48.3039 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,884    0,043    0,891      0,884    0,887      0,842    0,937     0,877     cluster0
                0,927    0,018    0,959      0,927    0,942      0,917    0,970     0,936     cluster1
                0,871    0,038    0,844      0,871    0,857      0,823    0,963     0,884     cluster2
                0,941    0,024    0,914      0,941    0,928      0,908    0,977     0,929     cluster3
Weighted Avg.   0,907    0,030    0,908      0,907    0,907      0,876    0,961     0,908

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 122   3  12   1 |   a = cluster0
   3 139   0   8 |   b = cluster1
  12   0  81   0 |   c = cluster2
   0   3   3  96 |   d = cluster3
```

Fonte: Autoria própria

A acuracidade neste teste foi de 90,6832%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 11 – Resultados do classificador J48 para o SOM

```
=== Summary ===

Correctly Classified Instances         451               93.3747 %
Incorrectly Classified Instances        32                6.6253 %
Kappa statistic                          0.9104
Mean absolute error                      0.0405
Root mean squared error                  0.1797
Relative absolute error                 10.9509 %
Root relative squared error             41.7722 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,942    0,041    0,903      0,942    0,922      0,890    0,956     0,841     cluster0
                 0,960    0,009    0,980      0,960    0,970      0,956    0,977     0,961     cluster1
                 0,849    0,023    0,898      0,849    0,873      0,844    0,934     0,857     cluster2
                 0,961    0,016    0,942      0,961    0,951      0,938    0,990     0,936     cluster3
Weighted Avg.    0,934    0,022    0,934      0,934    0,934      0,912    0,965     0,902

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 130   1   7   0 |   a = cluster0
   2 144   0   4 |   b = cluster1
  12   0  79   2 |   c = cluster2
   0   2   2  98 |   d = cluster3
```

Fonte: Autoria própria

A acuracidade neste teste foi de 93,3747%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Para os grupos C0-SOM e C2-SOM características como ‘foco_para_ead’, ‘sanar_duvidas_ead’, ‘nivel_perfil_aluno_ead’, ‘uso_software_educativo’. ‘lugar_calmo’ e ‘nivel_estimulo_negativo_estresse’, foram alguns dos divisores entre os grupos. Estudantes inseridos no C0-SOM apresentaram estes índices de forma mais negativa, não se adaptam as ferramentas EAD, possuem doença psicológica que afeta o desempenho e são alunos totalmente dependentes, comparado aos indivíduos do C2-SOM.

Nos agrupamentos C1-SOM e C3-SOM, as características que dividiram os indivíduos foram ‘nivel_perfil_aluno_ead’, ‘nivel_estimulo_negativo_estresse’ e ‘desvantagem_suspensao_aulas’. Indivíduos com maiores características de adaptação compuseram o C3-SOM, no geral, foram identificados indivíduos em nível de pós-graduação, com maiores níveis de perfil do aluno EAD, mais independentes e com menos estímulos negativos. O C1-SOM selecionou indivíduos com maiores dificuldades, apesar de serem poucas as diferenças, mostraram ânimo negativo para as atividades, maiores estímulos negativos ao estresse e são mais dependentes dos professores, além de a maior parte ter apresentado alguma desvantagem por conta

da suspensão das aulas. O Quadro 6 apresenta regras geradas para o *cluster* C0-SOM, regras dos demais *clusters* estão disponíveis no Apêndice B.

Quadro 6 - Regras geradas pela base SOM para o *cluster* C0-SOM

CLUSTER 0
Regra 1 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** sanar_duvidas_ead = "Muita dificuldade" **E** foco_para_ead = "Nenhum foco" **Então: C0-SOM**
Regra 2 **SE** sanar_duvidas_ead = "Muita dificuldade" **E** lugar_calmo = "Não" **Então: C0-SOM**
Regra 3 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** sanar_duvidas_ead = "Pouca dificuldade, Muita dificuldade" **E** optaria_metodo_ead = "Não" **E** idade = "Menor ou igual a 31 anos" **Então: C0-SOM**
Regra 4 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** optaria_metodo_ead = "Não" **E** sanar_duvidas_ead = "Muita dificuldade" **E** lugar_calmo = "Não" **E** nivel_perfil_aluno_ead = "Responsável, Disciplinado, Enxerga profs. como orient. (3 ou menos características)" **Então: C0-SOM**
Regra 5 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** sanar_duvidas_ead = "Muita dificuldade" **E** sexo = "Masculino" **Então: C0-SOM**

Fonte: Autoria própria

Quadro 6 - Regras geradas pela base SOM para o *cluster* C0-SOM

CLUSTER 0
Regra 6
SE
optaria_metodo_ead_flex_pandemia = "Não" **E**
sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E**
foco_para_ead = "Nenhuma/Pouca dificuldade" **E**
lugar_calmo = "Não" **E**
animo_para_atividades_ead = "Negativo, Muito negativo" **E**
bolsista = "Não"
Então: C0-SOM

Fonte: Autoria própria

5.3 RESULTADOS COM ALGORITMO *K-MEANS*

Para o *k-means* com *k*=3, os grupos C0-k3 e C1-k3 selecionaram indivíduos que optariam pelo EAD durante pandemia, mas não fora dela, e o C2-k3 selecionou indivíduos que não optariam em nenhuma das ocasiões. As Figuras 12, 13 e 14 apresentam os resultados da aplicação dos algoritmos JRip, PART e J48 respectivamente.

Figura 12 – Resultados do classificador JRip para o *k-means* com *k* = 3

```
=== Summary ===

Correctly Classified Instances         444               91.9255 %
Incorrectly Classified Instances        39                8.0745 %
Kappa statistic                          0.8769
Mean absolute error                      0.0623
Root mean squared error                  0.2135
Relative absolute error                 14.3197 %
Root relative squared error             45.7658 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,968    0,028    0,924      0,968    0,945      0,926    0,980     0,970     cluster0
                 0,929    0,046    0,905      0,929    0,917      0,877    0,969     0,911     cluster1
                 0,882    0,050    0,928      0,882    0,905      0,838    0,956     0,928     cluster2
Weighted Avg.    0,919    0,043    0,919      0,919    0,919      0,873    0,966     0,933

=== Confusion Matrix ===

   a   b   c   <-- classified as
 121   1   3 |   a = cluster0
   0 143  11 |   b = cluster1
  10  14 180 |   c = cluster2
```

Fonte: Autoria própria

A acuracidade neste teste foi de 91,9255%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 13 – Resultados do classificador PART para o *k-means* com *k* = 3

```
=== Summary ===

Correctly Classified Instances         448               92.7536 %
Incorrectly Classified Instances        35                7.2464 %
Kappa statistic                          0.8891
Mean absolute error                      0.0513
Root mean squared error                  0.2117
Relative absolute error                 11.7743 %
Root relative squared error             45.3695 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,944    0,011    0,967      0,944    0,955      0,940    0,985     0,968     cluster0
                 0,942    0,046    0,906      0,942    0,924      0,887    0,974     0,928     cluster1
                 0,907    0,057    0,920      0,907    0,914      0,851    0,960     0,939     cluster2
Weighted Avg.    0,928    0,042    0,928      0,928    0,928      0,886    0,971     0,943

=== Confusion Matrix ===

   a   b   c   <-- classified as
 118   0   7 |   a = cluster0
   0 145   9 |   b = cluster1
   4  15 185 |   c = cluster2
```

Fonte: Autoria própria

A acuracidade neste teste foi de 92,7536%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 14 – Resultados do classificador J48 para o *k-means* com *k* = 3

```
=== Summary ===

Correctly Classified Instances         447               92.5466 %
Incorrectly Classified Instances        36                7.4534 %
Kappa statistic                          0.8855
Mean absolute error                      0.0538
Root mean squared error                  0.2076
Relative absolute error                 12.3622 %
Root relative squared error             44.4964 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,912    0,020    0,942      0,912    0,927      0,902    0,977     0,956     cluster0
                 0,935    0,024    0,947      0,935    0,941      0,914    0,971     0,905     cluster1
                 0,926    0,075    0,900      0,926    0,913      0,848    0,949     0,918     cluster2
Weighted Avg.    0,925    0,045    0,926      0,925    0,926      0,883    0,963     0,924

=== Confusion Matrix ===

   a   b   c   <-- classified as
 114   0  11 |   a = cluster0
   0 144  10 |   b = cluster1
   7   8 189 |   c = cluster2
```

Fonte: Autoria própria

A acuracidade neste teste foi de 92,5466%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Analisando algumas regras desses agrupamentos o atributo ‘sexo’ e ‘animo_para_atividades_ead’ foram caracterizados como principais separadores entre os três clusters. Para os grupos C0-k3 e C1-k3 poucas foram as diferenças encontradas, o atributo ‘nivel_estimulo_negativo_estresse’ mostrou mais estímulos negativos nos indivíduos do C0-k3.

Em relação a C2-k3, características como ‘foco_para_ead’, ‘sanar_duvidas_ead’, ‘doenca_psicologica’, ‘nivel_perfil_aluno_ead’ e ‘lugar_calmo’ foram fatores negativos mais presentes. O Quadro 7 apresenta regras geradas para o *cluster* C0-k3, regras dos demais *clusters* estão disponíveis no Apêndice C.

Quadro 7- Regras geradas pela base *kMeans*3 para o *cluster* C0-k3

CLUSTER 0
Regra 1 **SE** sexo = “Feminino” **E** optaria_metodo_ead_flex_pandemia = “Sim” **E** animo_para_atividades_ead = “Muito positivo, Positivo e/ou Negativo” **Então: C0-k3**
Regra 2 **SE** sexo = "Feminino" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** foco_para_ead = "Bom/Médio foco" **Então: C0-k3**
Regra 3 **SE** sexo = "Feminino" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** foco_para_ead = "Bom/Médio foco" **E** nivel_estimulo_negativo_estresse = "Dificuldade de absorção do conteúdo (1 estímulo apenas)" **Então: C0-k3**
Regra 4 **SE** sexo = "Feminino" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** foco_para_ead = "Bom/Médio foco" **Então: C0-k3**

Fonte: Autoria própria

Quadro 7 - Regras geradas pela base *kMeans*3 para o *cluster* C0-k3

CLUSTER 0
Regra 5
SE desvantagem_suspensao_aulas = "Sim" **E** lugar_calmo = "Sim" **Então: C0-k3**
Regra 6
SE sexo = "Feminino" **E** adaptabilidade_ead = "Fácil adaptação" **E** moradia = "Sozinho, Com família" **Então: C0-k3**

Fonte: Autoria própria

Para *k*=5, os agrupamentos C0-k5 e C4-k5 foram compostos por indivíduos que optariam apenas pelo EAD durante pandemia. Nos grupos C1-k5 e C2-k5, os indivíduos não optariam pelo EAD em ambas as situações, e no grupo C3-k5, separados os indivíduos que optariam pelo EAD durante pandemia e fora dela. As Figuras 15, 16 e 17 apresentam os resultados da aplicação dos algoritmos JRip, PART e J48 respectivamente.

Figura 15 – Resultados do classificador JRip para o *k-means* com *k* = 5

```
=== Summary ===

Correctly Classified Instances         393               81.3665 %
Incorrectly Classified Instances        90               18.6335 %
Kappa statistic                          0.7546
Mean absolute error                      0.0954
Root mean squared error                  0.2603
Relative absolute error                 31.3361 %
Root relative squared error             66.7246 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,769    0,038    0,824      0,769    0,795      0,751    0,896     0,751     cluster0
                 0,851    0,017    0,891      0,851    0,870      0,850    0,915     0,762     cluster1
                 0,832    0,052    0,839      0,832    0,835      0,782    0,911     0,837     cluster2
                 0,824    0,110    0,795      0,824    0,810      0,708    0,870     0,757     cluster3
                 0,756    0,032    0,689      0,756    0,721      0,695    0,896     0,526     cluster4
Weighted Avg.    0,814    0,063    0,816      0,814    0,814      0,753    0,893     0,757

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  70   5   4  11   1 |   a = cluster0
   1  57   2   4   3 |   b = cluster1
   2   0  99  14   4 |   c = cluster2
  10   2  11 136   6 |   d = cluster3
   2   0   2   6  31 |   e = cluster4
```

Fonte: Autoria própria

A acuracidade neste teste foi de 81,3665%. As principais medidas de avaliação como *Precision, Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 16 – Resultados do classificador PART para o *k-means* com *k* = 5

```
=== Summary ===

Correctly Classified Instances         406               84.058  %
Incorrectly Classified Instances        77               15.942  %
Kappa statistic                          0.7902
Mean absolute error                      0.0681
Root mean squared error                  0.2407
Relative absolute error                 22.3768 %
Root relative squared error             61.7202 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,780    0,046    0,798      0,780    0,789      0,741    0,918     0,754     cluster0
                 0,910    0,034    0,813      0,910    0,859      0,837    0,949     0,816     cluster1
                 0,924    0,033    0,902      0,924    0,913      0,884    0,962     0,927     cluster2
                 0,855    0,072    0,860      0,855    0,857      0,783    0,910     0,822     cluster3
                 0,561    0,023    0,697      0,561    0,622      0,595    0,806     0,523     cluster4
Weighted Avg.    0,841    0,048    0,838      0,841    0,838      0,792    0,921     0,809

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  71   5   2  12   1 |   a = cluster0
   3  61   0   3   0 |   b = cluster1
   4   0 110   4   1 |   c = cluster2
   5   6   5 141   8 |   d = cluster3
   6   3   5   4  23 |   e = cluster4
```

Fonte: Autoria própria

A acuracidade neste teste foi de 84,058%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Figura 17 – Resultados do classificador J48 para o *k-means* com *k* = 5

```
=== Summary ===

Correctly Classified Instances         395               81.7805 %
Incorrectly Classified Instances        88               18.2195 %
Kappa statistic                          0.7594
Mean absolute error                      0.0827
Root mean squared error                  0.2565
Relative absolute error                 27.1516 %
Root relative squared error             65.7662 %
Total Number of Instances              483

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,736    0,054    0,761      0,736    0,749      0,692    0,879     0,708     cluster0
                 0,910    0,019    0,884      0,910    0,897      0,880    0,954     0,765     cluster1
                 0,908    0,027    0,915      0,908    0,911      0,883    0,964     0,916     cluster2
                 0,830    0,113    0,792      0,830    0,811      0,709    0,893     0,764     cluster3
                 0,537    0,029    0,629      0,537    0,579      0,545    0,772     0,492     cluster4
Weighted Avg.    0,818    0,061    0,815      0,818    0,816      0,758    0,906     0,768

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  67   2   2  16   4 |   a = cluster0
   1  61   0   5   0 |   b = cluster1
   4   0 108   7   0 |   c = cluster2
   9   4   6 137   9 |   d = cluster3
   7   2   2   8  22 |   e = cluster4
```

Fonte: Autoria própria

A acuracidade neste teste foi de 81,7805%. As principais medidas de avaliação como *Precision*, *Recall* e ROC *Area* são apresentados para cada classe. A matriz de confusão denota de forma geral quão bom foi o emprego do classificador.

Entre os grupos C0-k5 e C4-k5 observou-se que os atributos ‘nivel_academico’, ‘situacao_emprego’ e ‘nivel_estimulo_negativo_estresse’ influenciou a divisão dos indivíduos, sendo o grupo C0-k5 formado por pessoas a nível de graduação e menos adaptáveis à metodologia e o grupo C4-k5 indivíduos mais velhos, na maioria da pós-graduação e com características mais positivas em relação

ao EAD. Nos grupos C1-k5 e C2-k5 poucas discrepâncias foram encontradas entre os indivíduos, pois possuem quase as mesmas características, os atributos ‘grupo_de_risco_covid19’ e ‘doenca_psicologica’ foram os principais divisores entre esses agrupamentos.

Por fim, o grupo C3-k5 é composto por indivíduos a favor dos métodos EAD, as características ‘foco_para_ead’, ‘nivel_perfil_aluno_ead’, ‘adaptabilidade_ferramentas_ead’ e ‘nivel_estimulo_negativo_estresse’ foram fatores mais positivos em relação aos outros quatro agrupamentos. O Quadro 8 apresenta regras geradas para o *clusters* C0-k5, regras dos demais *clusters* estão disponíveis no Apêndice D.

Quadro 8 - Regras geradas pela base *kMeans*5 para o *cluster* C0-k5

CLUSTER 0
Regra 1
SE sexo = "Feminino" **E** nivel_estimulo_negativo_estresse = "Dificuldade de absorção do conteúdo, Preocupação com o futuro (2 ou mais estímulos)" **E** nivel_perfil_aluno_ead = "Organizado, Responsável, Enxerga profs. como orient. (3 ou menos características)" **Então: C0-k5**
Regra 2
SE grupo_de_risco_covid19 = "Não" **E** sexo = "Feminino" **E** filhos = "Não possui filhos" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** sanar_duvidas_ead = "Pouca/Muita dificuldade" **E** optaria_metodo_ead = "Não" **E** desvantagem_suspensao_aulas = "Sim" **Então: C0-k5**

Fonte: Autoria própria

Quadro 8 - Regras geradas pela base *kMeans*5 para o *cluster* C0-k5

CLUSTER 0
Regra 3
SE grupo_de_risco_covid19 = "Não" **E** sexo = "Feminino" **E** filhos = "Não possui filhos" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** lugar_calmo = "Sim" **E** importancia_interacao_professor = "Muito importante" **E** optaria_metodo_ead = "Não" **E** idade = "Menor ou igual a 26 anos" **Então: C0-k5**
Regra 4
SE sexo = “Feminino” **E** grupo_de_risco_covid19 = “Não” **E** foco_para_ead = “Médio foco ou Nenhum foco” **E** nivel_academico = “Graduação, Especialização e/ou Mestrado” **E** optaria_metodo_ead = “Não” **E** desvantagem_suspensao_aulas = “Sim” **Então: C0-k5**
Regra 5
SE situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado do trabalho" **E** nivel_academico = "Graduação" **E** nivel_estimulo_negativo_estresse = "Preocupação com o futuro (1 ou mais estímulos)" **Então: C0-k5**
Regra 6
SE sexo = "Feminino" **E** importancia_interacao_professor = "Muito importante" **E** tecnologias_ead_conhece = "Hangout Meet Google Education, Moodle, Skype (3 ou mais tecnologias)" **Então: C0-k5**

Fonte: Autoria própria

5.4 DISCUSSÃO DOS RESULTADOS

A experimentação com diferentes algoritmos de mineração possibilitou gerar diversos resultados para comparação das regras geradas e das informações descobertas. Esses resultados classificaram indivíduos de acordo com características positivas e negativas em relação a fatores de adaptação, condições psicológicas e condições estruturais.

Analisou-se que indivíduos que não optariam pelo EAD, em sua maioria, possuíam maiores estímulos negativos do estresse, média adaptação às ferramentas, relataram dúvidas quanto a aplicação de softwares educativos para auxílio das atividades, mostraram-se mais dependentes por considerarem a interação presencial com os professores fundamental, além de apresentarem baixos níveis de perfil do aluno EAD. Características como doença psicológica e se possuem lugar calmo também se mostraram relevantes.

Entre os indivíduos que optariam pelo ensino remoto, estes exibiram maiores características de perfil do aluno EAD, fácil adaptação para as ferramentas tecnológicas, optam pelo uso do software educativo, relataram poucos estímulos negativos do estresse e mostram ser mais independentes por conta de não acharem a interação presencial entre aluno e professor e aluno e colegas fundamental, são mais autogerenciáveis.

Regras mostraram casos em que os atributos como idade, grau de instrução, aluno bolsista e conhecimento das ferramentas tecnológicas EAD disponíveis se mostraram relevantes. Os algoritmos classificaram estudantes mais jovens, em nível de graduação e que relataram conhecer poucas ferramentas EAD como menos adeptos à metodologia. Estudantes de maior faixa etária, de maior grau de instrução (pós-graduação) e alunos bolsistas apareceram em muitas regras que classificaram estes como mais adeptos, mostrando grupos de alunos que possuem um maior vínculo com as instituições apresentam maior dedicação. Também, foi possível inferir que muitos alunos possam ainda não estar preparados para aplicação da metodologia EAD, e apenas optariam por esta por conta de alguma desvantagem ocasionada pela suspensão das aulas presenciais.

Assim, a identificação destes elementos possibilita uma visão geral dos indivíduos e facilita o entendimento desses pontos, auxiliando na criação de

estratégias pedagógicas que visem aprimorar a integração de alunos na modalidade EAD.

Em relação aos classificadores, o algoritmo PART obteve maior taxa de acuracidade, porém o modelo J48 também apresentou uma performance semelhante e com bons resultados. Para o *k-means* com *k*=5, houve uma queda significativa de desempenho, podendo inferir que a quantidade ideal de grupos para este caso seja de 3 ou 4 agrupamentos.

Por fim, foi possível realizar uma projeção de divisão dos grupos estipulada por meio dos resultados, projetando ao fim 3 grupos, sendo estes, um formado por pessoas que optariam pelo método EAD durante e fora de pandemia, outro composto por pessoas que optariam pelo EAD apenas durante a pandemia e o último composto por pessoas que não optariam pelo EAD em ambas as situações.

5.5 REUSO DO EXPERIMENTO PARA OUTROS DOMÍNIOS DA ÁREA EDUCACIONAL

A pesquisa apresentada neste trabalho foi financiada pelo “Edital 02/2019 - PROGRAD / PROREC - Apoio à execução de trabalhos de conclusão de curso - TCC” da Universidade Tecnológica Federal do Paraná, por meio da colaboração com a instituição de ensino dos alunos com DI em que iriam ser extraídas as informações destes para a utilização dos algoritmos de aprendizagem de máquina.

Como as adversidades que surgiram devido a pandemia da COVID-19 cancelaram as aulas presenciais na instituição parceira, não mais foi possível manter o contato com esses estudantes, o que dificultou a continuidade da pesquisa nesse com este público. A alternativa foi utilizar os algoritmos de AM para extrair informações sobre a opinião de alunos referente ao ensino remoto pois contempla questões educacionais, foco deste trabalho.

É importante ressaltar que os alunos com deficiência intelectual da instituição Dra. Zilda Arns, que hoje somam 120 alunos, em relação ao ensino remoto, somente 17 encontram-se desenvolvendo atividades via *Whatsapp* o que é um número pequeno para realizar a mineração de dados. Mas, para cumprir com o objetivo do edital, foi formulado uma ideia de como a pesquisa apresentada pode ser replicada a este público.

A primeira etapa é a adaptação do formulário em relação as questões, elaborando perguntas mais diretas, simples e com pouco texto, inserindo imagens nas questões tornando tanto as perguntas quanto as alternativas mais intuitivas. Por exemplo, uma pergunta referente ao ânimo do aluno durante a prática da atividade EAD, poderia ter nas suas alternativas imagens representando emoções como expressões negativas, medianas ou neutras e negativas. Isto pode ser aplicado a todas as perguntas deste mesmo tipo, facilitando a compreensão e preenchimento por estes alunos.

A segunda etapa é a ajuda de uma pessoa para auxiliar o aluno durante o preenchimento do formulário, é importante para que o mesmo fosse respondido corretamente. Assim, pode-se identificar que a maior dificuldade na replicação desta pesquisa estaria na adaptação das perguntas do formulário para os alunos com DI, pois a experimentação não mudaria, apenas os resultados da análise, visto que o público alvo é outro.

5.6 COMPARATIVO DESTA PESQUISA COM TRABALHOS DA LITERATURA

Esta seção realiza uma análise entre a pesquisa desenvolvida neste trabalho com as relacionadas na literatura. Em Hijazi e Naqvi (2006) os autores desenvolveram uma base de dados a partir de um questionário que tratou o perfil dos alunos com base no desempenho escolar e algumas características pessoais. O público alvo da pesquisa foi aplicada a estudantes de graduação. Os autores aplicaram o algoritmo de regressão linear simples.

Em Kampff *et al*. (2008) os autores trabalharam com dados de alunos em AVAs. A base de dados foi criada com atributos referentes a informações de acesso e interação dos alunos no AVA. O público alvo foram estudantes de graduação presencial que possuíam matérias ofertadas via EAD. Foram utilizados algoritmos de árvore de decisão para descoberta de regras.

Em Shih, Koedinger e Scheines (2010) os autores obtiveram uma base de dados a partir da aplicação de um tutor inteligente para ensino da matemática. Foram utilizados algoritmos de aprendizado não-supervisionado para identificar comportamentos do aluno ao utilizar a ferramenta. Os autores não especificaram o público alvo.

Em Lopez *et al.* (2012) os autores coletaram os dados da plataforma de ensino *Moodle*, o trabalho se deu pela coleta de informações sobre estatísticas de uso na plataforma. O público alvo foram estudantes de graduação. Os autores aplicaram diversos algoritmos de agrupamento e classificação e realizaram uma análise comparativa destes algoritmos a partir das métricas de acuracidade.

Em Bazaldua, Baker e San Pedro (2014) os autores obtiveram uma base de dados de respostas de atividades de alunos realizadas em um ambiente de aprendizagem *online*. O público alvo foram estudantes do ensino fundamental. Utilizaram o algoritmo *Apriori* para gerar regras referentes ao comportamento dos alunos durante o uso do sistema. Por fim, os autores selecionaram as melhores regras comparando diferentes métricas desses algoritmos.

Em Rau, Mason e Nowak (2016) a base de dados foi criada a partir da aplicação de uma pesquisa para alunos de graduação em química, referente a similaridade de estruturas químicas, em que os alunos responderam 50 questões julgando essas similaridades. Eles utilizaram *k-fold cross validation* para avaliar a precisão da previsão dos modelos de similaridade. Ao fim, os autores comparam ambas as abordagens de similaridade utilizadas.

Em Fang *et al*. (2018) os autores obtiveram uma base de dados por meio da aplicação de um sistema tutor inteligente para estudantes adultos com baixa alfabetização. Os algoritmos *k-means* e *Hierarchical Clustering* foram utilizados para descoberta de padrões de aprendizagem no uso da ferramenta. A análise de performance dos algoritmos foi comparada ao fim para a seleção da solução final.

Em Yang *et al.* (2019) os autores usaram uma base de dados com registros afetivos dos alunos por meio do uso de um STI. O público alvo não é informado, a base é disponibilizada na *internet*. Eles aplicaram detectores afetivos que utilizam classificadores baseados em aprendizado de máquina para prever os estados afetivos dos alunos. Ao fim, os diferentes métodos de aprendizagem ativa aplicados por eles foram comparados com base nos valores de avaliação da curva ROC. O Quadro 9 apresenta uma comparação das informações dos trabalhos apresentados e da presente pesquisa.

Quadro 9 - Comparativo dos trabalhos apresentados em relação ao trabalho desenvolvido

Autor Trabalho	Público Alvo	Obtenção da Base	Algoritmo de Mineração	Metodologia
Hijazi e Naqvi	Estudantes de Graduação presencial	Base desenvolvida pelo autor	Regressão Linear Simples	• Aplicação dos algoritmos • Análise dos resultados
Kampff *et al.*	Estudantes de Graduação presencial	Base desenvolvida pelo autor	Árvore de decisão	• Aplicação dos algoritmos • Análise dos resultados
Shih, Koedinger e Scheines	Estudantes com nível escolar não informado	Base obtida por aplicação de ferramenta tecnológica	Modelo de *Bottom-out hint*	• Aplicação dos algoritmos • Análise dos resultados
Lopez *et al*.	Estudantes de Graduação presencial	Base desenvolvida pelo autor	*Simple k-means*, *Hierarchical Clusterer*, *Xmeans*, EM, outros. JRip, J48, *Random Forest*, *Naïve Bayes*, outros.	• Aplicação dos algoritmos • Análise dos resultados • Avaliação e Comparação dos algoritmos
Bazaldua, Baker e San Pedro	Estudantes do Ensino Fundamental	Base obtida por aplicação de ferramenta tecnológica	*Apriori*	• Aplicação dos algoritmos • Análise dos resultados
Rau, Mason e Nowak	Estudantes de Graduação presencial	Base desenvolvida pelo autor	*k-fold cross validation*	• Aplicação dos algoritmos • Análise dos resultados • Avaliação e Comparação dos algoritmos
Fang *et al.*	Estudantes adultos com baixa alfabetização	Base obtida por aplicação de ferramenta tecnológica	*k-means* e Análise Hierárquica	• Aplicação dos algoritmos • Análise dos resultados • Avaliação e Comparação dos algoritmos
Yang *et al.*	Estudantes com nível escolar não informado	Base disponibilizada na *internet*	Regressão Logística aplicada a detectores afetivos	• Aplicação dos algoritmos • Análise dos resultados • Avaliação e Comparação dos algoritmos
Pesquisa proposta	Estudantes de Graduação presencial	Base desenvolvida pelo autor	*Self-organized maps*, *k*-means JRip, PART e J48	• Aplicação dos algoritmos • Análise dos resultados • Avaliação e Comparação dos algoritmos

Fonte: Autoria própria

As informações exibidas comparam o público alvo escolhido, como os autores obtiveram a base de dados, quais algoritmos foram usados, e se ao fim houve apenas a análise dos resultados ou análise, avaliação e comparação dos resultados.

Nota-se que a maioria dos trabalhos tiveram como público alvo alunos de graduação na modalidade presencial. A obtenção da base de dados mostra que a maioria dos autores desenvolveu a base utilizada. Ainda, a maior parte das pesquisas adotou a metodologia de aplicação dos algoritmos, análise dos resultados, avaliação e comparação dos resultados. Observa-se que novas pesquisas podem ser desenvolvidas no domínio fundamental e médio.

5.7 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este Capítulo mostrou como o experimento foi realizado, bem como os resultados obtidos pela aplicação dos algoritmos de agrupamento e classificação. A análise e interpretação das regras encontradas também foram apresentadas como o esquema de cores gerado para os grupos e suas respectivas regras, além das medidas de avaliação dos classificadores como matriz de confusão e acuracidade para cada execução dos algoritmos. Na discussão dos resultados, uma das principais fases da mineração de dados, foi explicada a análise das regras para extrair informações e pontos positivos e negativos sobre os dados trabalhados.

Foi descrita como a pesquisa pode ser replicada para a área de ensino na modalidade especial. Por este público possuir uma maior dependência e dificuldade na execução das atividades educacionais, foi sugerido a adaptação do formulário usado na coleta de dados. Por fim, a pesquisa desenvolvida é comparada com outros trabalhos na área educacional aqui apresentados, foram levantados atributos como problemática, obtenção da base de dados, algoritmos utilizados e metodologia, sendo estas fases que foram implementadas neste estudo.

6 CONCLUSÃO

Este trabalho teve como finalidade o emprego e avaliação de algoritmos de aprendizagem de máquina aplicados à área educacional. O problema em questão foi extrair informações e identificar perfis de alunos do ensino superior presencial sobre a aplicação do ensino a distância no período de pandemia.

A coleta dos dados se deu pelo desenvolvimento de um conjunto de questões sobre os estudantes e sobre a opinião pela adesão do ensino à distância. Algumas dificuldades foram encontradas durante a execução da experimentação, como a coleta das respostas e interpretação das regras. Em relação a coleta, para a criação do conjunto de dados foi preciso alcançar uma quantidade *x* de respostas a fim de que o experimento pudesse ser realizado, e que demandou um maior tempo na fase de coletas dos dados e entendimento do problema. A interpretação das regras dos grupos, após a fase de mineração passou por diversos passos, além da representação dos *clusters* por cores apresentado na seção 5.1, os atributos passaram por uma análise individual, identificando os valores mais predominantes em cada atributo na tentativa de compreender melhor um dado agrupamento, por exemplo.

Foram utilizados os algoritmos de agrupamento SOM e *k-means* para gerar grupos de estudantes com características comuns, após esta fase foi possível gerar regras sobre esses grupos por meio de algoritmos de classificação PART, JRip e J48, baseados na árvore de decisão, para extrair as informações. Os algoritmos foram comparados com base nas taxas de acerto dos classificadores e nas métricas da matriz de confusão como Precisão*, Recall* e curva ROC.

Os resultados mostraram que os alunos em condições mais adversas de adaptação, econômica, psicológica ou de acesso possuem um menor ânimo para desenvolver as atividades, ao contrário de alunos com boas condições de acesso, infraestrutura, vínculo com a instituição de ensino e melhor adaptabilidade que se mostraram engajados com a implementação do EAD.

Apesar da resposta positiva de diversos alunos, muitos escolheram por não optar pela metodologia EAD, mesmo apresentando grau positivo para as questões de acesso e adaptação, o que leva uma busca por maiores informações sobre estes estudantes. Com base nas métricas de acuracidade e no resultado das regras descobertas, os classificadores PART e J48 mostraram ser adequados para o conjunto de dados.

6.1 TRABALHOS FUTUROS

O experimento pode servir para novas pesquisas, é possível aumentar as questões em relação as atividades práticas e teóricas dos cursos, buscando entender o nível de ocorrências destas aulas. Por exemplo, um curso de química onde os alunos possuem aulas de laboratório diferente de um curso de direito onde não há tantas práticas em laboratórios, mas há aulas externas como ida a tribunais, sessões jurídicas, etc., para saber o impacto em cada indivíduo pela escolha do EAD.

A aplicação de algoritmos de redução de dimensionalidade pode ser uma nova pesquisa, pois ajuda a filtrar ainda mais as regras e gera novos resultados. Ainda, pode-se aplicar os algoritmos para uma área acadêmica específica, com uma população de indivíduos com características em comum.

REFERÊNCIAS

ALASADI, S. A.; BHAYA, W. S. Review of data preprocessing techniques in data mining. **Journal of Engineering and Applied Sciences**, v. 12, n. 16, p. 4102-4107, 2017.

ALMEIDA, M. E. B. **Tecnologia na escola**. p. 69-73. Disponível em < http://portal.mec.gov.br/seed/arquivos/pdf/2sf.pdf>. Acesso em: 21 de mar. 2020.

ALPAYDIN, E. **Introduction to Machine Learning.** 2. ed. Massachusetts: MIT Press, 2014.

BACICH, L.; NETO, A. T.; TREVISANI, F. D. M. **Ensino híbrido**: personalização e tecnologia na educação. Porto Alegre: Penso, 2015.

BAKER, R. S. J. D.; ISOTANI, S.; DE CARVALHO, A. M. J. B. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação.** v. 19, n. 2, 2011.

BAKER, R. S. J.; BARNES, T.; BECK, J. E. Educational Data Mining. In: INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2008, Montreal. **Proceedings...** Montreal, 2008, p. 2.

BAKER, R. S. J. D. Data mining for education. **International encyclopedia of education**. v. 7, n. 3, p. 112–118, 2010.

BARANAUSKAS, J. A.; MONARD, M. C. **Reviewing Some Machine Learning Concepts and Methods**: Relatórios Técnicos do ICMC. São Carlos, 2000.

BARARDO, D. G. *et al.* Machine Learning for predicting lifespan-extending chemical compounds. **Aging.** Nova York, v. 9, n. 7, 2017.

BASKAR, S. S.; AROCKIAM, L.; CHARLES, S. A systematic approach on data pre-processing in data mining. **Compusoft**, v. 2, n. 11, 2013, p. 335.

BAZALDUA, D. A. L.; BAKER, R. S.; SAN PEDRO, M. O. Z. Comparing Expert and Metric-Based Assessments of Association Rule Interestingness. In: 7 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2014, London, **Proceedings…** London, jul. 2014, p. 44 – 51.

BORGES, M. F. V. Inserção da Informática no Ambiente Escolar: inclusão digital e laboratórios de informática numa rede municipal de ensino. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 2008, Belém do Pará. **Anais...** Belém do Pará, 2008.

BOUCKAERT, R. R. *et al*. Weka Manual for Version 3-8-3. Sep 4, 2018. Disponível em: <https://www.cs.waikato. ac.nz/ml/ weka/> Acesso em: 16 de jun de 2020.

BREIMAN, L. Random forests. **Machine Learning Journal**. Hingham, v.45, 2001, p.5–32

BULEGON, H.; MORO, C. M. C. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. **Journal of Health Informatics.** São Paulo, v. 2, n. 2, 2010.

CARVALHO, D. R. **Árvore de Decisão / Algoritmo Genético para Tratar o Problema de Pequenos Disjuntos em Classificação de Dados.** 162f. 2005. Tese (Doutorado em Ciência em Engenharia Civil). Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2005.

CARVALHO, H. M. **Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão.** 98 f. 2014. Monografia (Bacharelado em Engenharia de Software). Universidade de Brasília. Brasília, 2014.

CARVALHO, L. A. V. ***Datamining* – A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2005.

CASTELLS, M. **A sociedade em rede. A era da informação: economia, sociedade e cultura.** São Paulo: Paz e Terra, 2003.

CIEB. **Nota Técnica nº 16 de 30 de outubro de 2019.** Inteligência Artificial na educação. Disponível em: <https://cieb.net.br/wp-content/uploads/2019/11/CIEB _Nota_Tecnica16_nov_2019_digital.pdf>. Acesso em: 04 set. 2020.

COHEN, W. W. **Fast effective rule induction.** In: Machine learning proceedings 1995. Morgan Kaufmann, 1995. p. 115-123.

COSTA, B. D. S.; RIBEIRO, G. C.; GUEDES, A. M. A. Uso do Software Educacional Duolingo no Ensino da Língua Inglesa. **Nuevas Ideas en Informática Educativa.** Santiago do Chile, v. 12, 2016, p. 501-504.

COSTA, E. *et al*. JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO, n. 1, 2012. Rio de Janeiro. **Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações.** Rio de Janeiro: UFRJ, 2012, 29 p.

CREPALDI, P. G. *et al*. Um estudo sobre a Árvore de Decisão e sua Importância na Habilidade de Aprendizado. **Revista Eletrônica Múltiplo Saber.** Londrina, v. 14, n. 1, dez. 2011.

CUNNINGHAM, P.; DELANY, S. k-nearest Neighbour classifiers. Technical report, UCD School of Computer Science and Informatics, 2007.

CYSNEIROS, P. G. A gestão da Informática na Escola Pública. In: XI SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2000, Maceió. **Anais… do SBIE.** Maceió, 2000.

DATAH. Disponível em: <https://www.datah.ai/single-post/2017/06/09/Bot-Mr-Turing-ensina-ingl%C3%AAs-pelo-Facebook-Messenger> Acesso em: 10 de mar de 2020.

DA COSTA, A. R. A EDUCAÇÃO A DISTÂNCIA NO BRASIL: Concepções, histórico e bases legais. **Revista Científica da FASETE**, p. 69-74, 2017.

DA ROCHA, C. A. J. PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS - ELEMENTOS DE APOIO E PRINCIPAIS PROBLEMAS. **Revista Traços**, v. 3, n. 5, p. 5 - 14, 2000.

DE OLIVEIRA, C. TIC'S na educação: a utilização das tecnologias da informação e comunicação na aprendizagem do aluno. **Pedagogia em Ação**, v. 7, n. 1, 2015.

DIETTERICH, T. G. Machine-Learning Research. **AI Magazine**, v. 18, n. 4, p. 97, 15 Dec. 1997.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification.** 2 ed. New York: John Wiley & Sons, 2001.

ESTEVES, R. S.; LORENA, A. C.; NASCIMENTO, M. Z. Aplicação de técnicas Aprendizado de Máquina na Classificação de Imagens Mamográficas. In: 2º SIMPÓSIO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DO ABC, 2009, Santo André. **Anais...** Santo André: UFABC, nov. 2009.

FANG, Y. *et al*. Clustering the Learning Patterns of Adults with Low Literacy Skills Interacting with na Intelligent Tutoring System. In: 11 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2018, Buffalo. **Proceedings...** Buffalo, jul. 2018, p. 348 – 354.

FAVA, R. **Educação para o século 21: a era do indivíduo digital.** São Paulo: Saraiva, 2016.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** American Association for Artificial Intelligence, 1996.

FILHO, E. S. L. **Comparando algoritmos de aprendizado profundo para o problema de detecção de distração de motoristas a partir de imagens**. Quixadá, 2018. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Software). Coordenadoria de Graduação, Universidade Federal do Ceará.

FOLQUE, M. A. Infância e cibercultura: como educar a geração que já nasceu no mundo digital. **Revista pátio educação infantil.** n. 28, p. 4-11, 2011.

FONSECA, J. **Indução de Árvores de Decisão**: HistClass – proposta de um algoritmo não paramétrico. 1994. 140 p. Dissertação (Mestrado em Engenharia Informática). Universidade Nova Lisboa, Lisboa, 1994.

FONTANA, A.; NALDI, M. C. **Estudo e comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados.** Universidade de São Paulo. Brasil, 2009.

FRANK, E.; WITTEN, I. H. **Generating accurate rule sets without global optimization.** 1998.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** The Morgan Kaufmann series in data management systems. Morgan Kaufmann Publishers, San Francisco, USA, August 2001.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques.** Elsevier, 2006.

HAYKIN, S. **Neural Networks and Learning Machines.** 3. ed. New Jersey: Prentice Hall.

HENKE, M. *et al*. **Aprendizagem de Máquina para Segurança em Redes de Computadores: Métodos e Aplicações.** In: Minicursos / SBSeg 2011. Porto Alegre: Sociedade Brasileira de Computação (SBC), 2011, v. 1, p. 55-105.

HIJAZI, S.T.; NAQVI, S. Factors Affecting Students' Performance: A Case of Private Colleges. **Bangladesh e-Journal of Sociology**, v. 3, n. 1, p. 1-10, jan. 2006.

HORST, P. S. Avaliação do Conhecimento Adquirido por Algoritmos de Aprendizado de Máquina Utilizando Exemplos. In: SIMPÓSIO DE PÓS-GRADUAÇÃO DO ICMC-USP, 1999, São Carlos. **Anais...** São Carlos: USP, ago. 1999.

INGARGIOLA, Giorgio. **Building classification models: ID3 and C4. 5**. Disponível por WWW em: http://www.cis.temple. edu/~ ingargio/cis587/readings/id3-c45. html, 1996.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: < http://portal.inep.gov.br/web/guest/inicio>. Acesso em: 11 out. 2019.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. **Data clustering**: a review. ACM computing surveys (CSUR), v. 31, n. 3, p. 264-323, 1999.

KALOGIROU, S. A. Artificial Neural Networks in Energy Applications in Buildings. **International Journal of Low-Carbon Technologies.** Oxford, v. 1, n. 3, jul. 2006.

KAMPFF, A. J. C.; REATEGUI, E. B.; LIMA, J. V. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à pratica docente. **Novas Tecnologia na Educação.** Rio Grande do Sul, v. 6, n. 2, dez. 2008.

KOHONEN, T. **The self-organizing map.** Proceedings of the IEEE, v. 78, n. 9, p. 1464-1480, 1990.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley and Sons, Inc, 2005.

LITTLEWORT, G. *et al*. Analysis of Machine Learning Methods for Real-Time Recognition of Facial Expressions from Video. In: IEEE COMPUTER SCIENCE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2004, Washington. **Proceedings...** Washington: IEEE Computer Society, jul. 2003.

LOPES, J. J. **Introdução da Informática no Ambiente Escola.** 2004. Artigo (Disciplina). Programa de Pós-Graduação em Educação Matemática, Universidade Estadual Paulista. Rio Claro, 2004.

LOPEZ, M. I.; LUNA, J. M.; ROMERO, C.; VENTURA, S. Classification via Clustering for predicting final marks based on student participation in forums. In: 5 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2012, Pittsburgh. **Proceedings...** Chania, jun. 2012, p. 148 – 151.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às *Support Vector Machines*. **Revista de Informática Teórica e Aplicada.** Porto Alegre, v. 14, n. 2, 2007.

LUCKIN, R.; HOLMES, W.; GRFFITHS, M.; FORCIER, L.B. Intelligence Unleashed: An Argument for AI in Education. London: Pearson. Disponível em: static.googleusercontent.com/media/edu.google.com/pt-BR//pdfs/Intelligence-Unleashed-Publication.pdf. Acesso em: 21 set de 2017.

MADEIRO, S. S. *et al*. Uma Abordagem Multi-Objetiva Híbrida para Seleção e Atribuição de Pesos a Caracteristicas para Classificadores k-NN. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL (ENIA), 2009, Recife. **Anais...** Recife: UPE, 2009.

MARQUES, A. C.; CAETANO, J. S. Utilização da Informática na Escola. In: MERCADO, L. P. L. **Novas Tecnologias na Educação: Reflexões Sobre a Prática.** Maceió, edUFAL, 2002. p. 131 – 137.

MAZUCHELI, J. A., LOUZADA-NETO, F., MATINEZ, E. Z. **Algumas medidas do valor preditivo de um modelo de classificação.** Outubro 2006. Nº. 162.

MCCUE, C. **Data Mining and Predictive Analysis - Intelligence Gathering and Crime Analysis.** Elsevier, 2007.

MITCHELL, T. **Machine Learning**. McGraw-Hill: New York, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In: REZENDE, S. O. (Org.). ***Sistemas Inteligentes***. Barueri: Manole, 2003. p. 40-56.

NASCIMENTO, E. G. A. **Avaliação do uso do software GeoGebra no ensino de geometria: reflexão da prática na escola.** Actas de la Conferencia Latinoamericana de GeoGebra ISSN 2301-0185. Uruguay, v. 1, p. 125-132, jan. 2012.

NASCIMENTO, R. L. S. D.; JUNIOR, G. G. D. C.; FAGUNDES, R. A. D. A. Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de Dados do INEP. **Novas tecnologias na Educação.** Porto Alegre, v. 16, n. 1, jul. 2018.

NETO, C. D. G. **Potencial de Técnicas de Mineração de Dados para o Mapeamento de Áreas Cafeeiras.** 2014. Monografia (Apresentação de disciplina). Instituto Nacional de Pesquisas Espaciais – INPE. São José dos Campos, 2014.

NUNAN, A. E. **Detecção de Cross-Site Scripting em Páginas Web.** 2012. Dissertação (Mestrado em Informática). Universidade Federal do Amazonas. Manaus, 2012.

OLIVEIRA, A. R. **Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado.** 2016. Dissertação (Mestrado em Ciência da Computação). Universidade Federal do Rio Grande do Sul. Porto Alegre, 2016.

OLIVEIRA, L. L.; FREITAS, A. A.; TINÓS, R. Multi-objective genetic algorithms in the study of the genetic code's adaptability. **Information Sciences.** v. 425, p. 48 – 61, jan. 2018.

OLIVEIRA, L. L **Uma análise de algoritmos de aprendizagem de máquina aplicados em técnicas de localização indoor para diferentes tipos de smartphones**. Recife, 2017. Trabalho de Graduação (Bacharel em Engenharia da Computação). Coordenadoria de Graduação, Universidade Federal de Pernambuco.

OLSON, D. L; DELEN, D. **Advanced Data Mining Techniques**. Springer, 2008

PIMENTEL, E. P.; DE FRANÇA, V. F.; OMAR, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO – SBIE, 2003, Rio de Janeiro. **Anais... do SBIE.** Rio de Janeiro, 2003, p. 495-504.

PINTO, A. H. M. **Um sistema de reconhecimento de objetos incorporado a um robô humanoide com aplicação na educação.** 2014. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional). Universidade de São Paulo. São Carlos, 2014.

PIRES, C. S.; ARSAND, D. R. Análise da utilização das tecnologias da informação e comunicação na educação a distância (EaD). **Revista Thema**, v. 14, n. 1, p. 182-198, 2017.

POBLACIÓN, D. A.; WITTER, G. P.; SILVA, J. F. M. D. Comunicação & Produção Científica: contexto, indicadores e avaliação. São Paulo: Angellara, 2006. 426 p. (Comunicação & Pesquisa.).

RAMOS, E. M. F. **Educação e informática - reflexões básicas.** Graf & Tec, Florianópolis, v. 0, n. 0, p. 11-26, 1996.

RAMOS, J. S. **A Utilização de Ferramentas Tecnológicas no Processo de Ensino-Aprendizagem no Ensino Médio de Colégios Públicos Estaduais de Joinville/SC.** Trabalho de Conclusão de Curso (Pós-graduação *Latu Sensu* em Ciências e Tecnologia). Universidade Federal de Santa Catarina. Florianópolis.

RAO, S.; GUPTA, P. **Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm.** 2012.

RAU, M. A.; MASON, B.; NOWAK, R. How to Model Implicit Knowledge? Similarity Learning Methods to Assess Perceptions of Visual Representations. In: 9 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2016, North Carolina, **Proceedings…** North Carolina, jun. 2016, p. 199 – 206.

RIBEIRO, E. N.; MENDONÇA, G. A. A.; MENDONÇA, A. F. A Importância dos Ambientes Virtuais de Aprendizagem na Busca de Novos Domínios da EAD. In: Congresso Nacional de Educação a Distância, n. 13, 2007, Curitiba. **Anais...** Curitiba.

RIBEIRO, G. C. *et al.* Software Livre como Ferramenta no Processo do Ensino Aprendizado: Uma experiência com Turmas do EJA. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2017, Recife. **Anais... do XXVIII Workshop de Informática na Escola.** Recife, 2017.

REZENDE, S. O. **Mineração de Dados.** XXV Congresso da Sociedade Brasileira de Computação, 2005.

RICH, E.; KNIGHT, K. **Artificial Intelligence**. McGraw-Hill: New York, 1991.

SANTOS, A. V. **Ambiente Virtual Inteligente Aplicado em Controle Estatístico da Qualidade STCEQ.NET.** 2005. Dissertação (Mestrado em Ciência da Computação). Universidade Federal de Santa Catarina. Florianópolis, 2005.

SEMENSATO, M. R.; FRANCELINO, L. A.; MALTA, L. S. O uso da Inteligência Artificial na Educação à Distância. **Revista Cesuca Virtual: Conhecimento sem Fronteiras.** Cachoeirinha, v. 2, n. 4, ago. 2015.

SILVA, E. PIRES, F. G. S. O uso do jogo educacional “Eu sei Contar” como auxílio da matemática no ensino infantil. In: WORKSHOP DE INFORMÁTICA NA ESCOLA. 2017. **Anais... do Workshop de Informática na Escola.** 2017.

SILVEIRA, S.R. **Estudo e Construção de uma ferramenta de autoria multimídia para a elaboração de jogos educativos.** 1999. Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul, 1999.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning: From Theory to Algorithms.** Cambridge University Press, New York, 2014.

SHIH, B.; KOEDINGER, K. R.; SCHEINES, R. A Response Time Model for Bottom-Out Hints as Worked Examples. In: 1 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2008, Montreal. **Proceedings...** Montreal, jun. 2008, p. 117 - 126.

TAKAKURA, A. M. *et al*. Uso do aprendizado de máquina no diagnóstico médico de patologias. **Colloquium Exactarum.** Presidente Prudente, v. 10, n. 1, p. 78 – 90, out. 2017.

TRINDADE, V. C. **Tecnologia como facilitadora da aprendizagem do aluno com deficiência intelectual**. Florianópolis, 2016. Monografia (Especialização em Educação na Cultura Digital). Coordenadoria de Pós-Graduação, Universidade Federal de Santa Catarina.

UNFER, E. **A contribuição de softwares livres na alfabetização de criança com dificuldades de aprendizagem.** Agudo, 2014. Artigo (Especialização em Tecnologias da Informação e da Comunicação Aplicada à Educação). Universidade Federal de Santa Maria.

WEISS, S. M.; KULIKOWSKI, C. A. Computer Systems that Learn. Morgan Kaufmann, 1991, San Mateo, CA.

VALENTE, J. A. Informática na educação no Brasil: Análise e contextualização histórica. In VALENTE, J. A. (org). **O computador na sociedade do conhecimento.** Campinas: Unicamp/NIED, 1999.

VIANA, L. C. **Informática na Educação: O Uso das Tecnologias na Prática Docente.** 2018. Artigo (Especialização em Mídias da Educação). Universidade Federal de Santa Maria. Santa Maria, 2018.

VICARI, R. M. **Tendências em inteligência artificial na educação no período de 2017 a 2030: sumário executivo**. 2018.

VYGOTSKY, L. S. A Construção do Pensamento e da Linguagem. São Paulo: **Martins Fontes**, 2005.

WANG, J. **Encyclopedia of Data Warehousing and Mining.** Idea Group Reference, 2005.

YANG, T. Y.; BAKER, R. S.; STUDER, C.; HEFFERNAN, N.; LAN, A. S. Active Learning for Student Affect Detection. In: 12 INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2019, Montréal. **Proceedings...** Montréal, jul. 2019, p. 208 – 217.

YOSHIDA, Soraia. Saint Paul lança plataforma de ensino com Inteligência artificial IBM Watson. **Revista Época Negócios**, 14 de novembro de 2017. Disponível em: < https://epocanegocios.globo.com/Carreira/noticia/2017/11/saint-paul-lanca-plataformade-ensino-com-inteligencia-artificial-ibm-watson.html >. Acesso em: 10 de mar de 2020.

APÊNDICE A – DESCRIÇÃO DO CONJUNTO DE DADOS *EAD_SUPERIOR_PANDEMIA*

Quadro 10 - Descrição do Conjunto de Dados *ead_superior_pandemia*

Categoria	Atributo	Descrição	Valores Possíveis
Informações Pessoais	idade	Idade dos estudantes	Entre 16 e 61 anos
	sexo	Sexo/gênero dos estudantes	Masculino, Feminino
Informações acadêmicas	nivel_academico	Grau de instrução	Graduação, Especialização, Mestrado, Doutorado
	curso	Curso que está matriculado	Curso ficou para preenchimento pelo aluno
	bolsista	Se é aluno bolsista	Sim, Não
Informações socioeconômicas	classe_social	Padrão de vida que melhor se encaixa	Classe baixa, Classe média baixa, Classe média, Classe média alta
	situacao_moradia	Situação atual de moradia	Sozinho, Em Família, Em República, Divide Apartamento
	lugar_calmo	Se possui lugar calmo para as atividades ou não	Sim, Não
	situacao_emprego	Se está trabalhando	Trabalhando dirigindo-se ao trabalho, Trabalhando em *Home Office*, Dispensado das Atividades, Não trabalhando
	filhos	Se possui filhos	Possuo filhos e teria as atividades EAD afetadas, Possuo filhos mas não seria afetado pelo EAD, Não possuo filhos
	situacao_internet	Se possui internet e qual o tipo de provedor	Banda larga fixa, Banda larga móvel, Ambas, Não possuo internet
	possui_computador_notebook	Se possui computador ou *notebook*	Sim, Não
	possui_dispositivo_virtual	Se possui dispositivo virtual	Celular, Tablet, Outros
	situacao_disp_virtual	Situação do dispositivo virtual/computador	Computador/Notebook pessoal e/ou Emprestado, Celular/Notebook pessoal e/ou Emprestado, Uso de dispositivo dentro/Emprestado do campus,
Informações sobre condições psicológicas	doença_psicologica	Se possui alguma doença psicológica que possa afetar o desempenho	Depressão, Ansiedade, Bipolaridade, Síndrome do Pânico, Não possuo doença deste tipo
	nivel_estimulo_negativo_estresse	Situações que posam desencadear estímulos negativos em estudantes	Necessidade de estudar várias matérias, Acúmulo de atividades, Dificuldade de absorção do conteúdo, Preocupação com o futuro, Não possuo estímulos

Fonte: Autoria própria

Quadro 10 - Descrição do Conjunto de Dados *ead_superior_pandemia*

Categoria	Atributo	Descrição	Valores Possíveis
Informações referentes a adaptabilidade e conhecimento do aluno sobre EAD	foco_para_ead	Nível de foco para realização das atividades	Bom foco, Médio foco, Nenhum foco
	sanar_duvidas_ead	Nível de dificuldade em sanar dúvidas via EAD	Nenhuma dificuldade, Pouca dificuldade, Muita dificuldade
	animo_para _atividades_ead	Possui ânimo para execução de atividades EAD	Muito positivo, Positivo, Negativo, Muito negativo
	adaptabilidade _ferramentas_ead	Nível de adaptabilidade às ferramentas EAD	Fácil, Média, Difícil adaptação
	nivel_perfil_aluno _ead	Características sobre o perfil do aluno EAD mais em comum com as que possui	Proativo, Organizado, Responsável, Disciplinado, Flexível, Possui Autonomia, Enxerga professores como orientadores
	tecnologias_ead _conhece	Algumas tecnologias que os entrevistados conhecem ou não	Hangout Meet Google Education, Microsoft Teams, Moodle, Skype, Slack, Nenhuma das alternativas
	tecnologias _fundamentnais_ead	Opinião sobre quais das tecnologias listadas precisam ser fundamentais em plataformas EAD	Videoaulas, Áudio e Videoconferência, Chats, Fóruns, Bibliotecas Virtuais, Outros
Informações referentes ao engajamento do aluno	uso_software _educativo	Opinião sobre sugestão de aplicação de softwares educacionais para auxílio das atividades EAD	Sim, Talvez, Não
	importancia _interacao_professor	Nível de importância da interação presencial entre aluno e professor	Muito importante, Importante, Pouco importante
	importancia _interacao_colegas	Nível de importância da interação presencial entre aluno e colegas de classe	Muito importante, Importante, Pouco importante
Informações sobre situação atual do aluno em relação as aulas e a pandemia	desvantagem _suspensao_aulas	Se foi prejudicado pela suspensão das aulas	Sim, Não
	grupo_risco_Covid_19	Se está no grupo de risco da Covid-19	Sim, Não
Informações referente a opinião dos alunos sobre aplicação do EAD	optaria_metodo_ead	Se optaria pelo método EAD fora de pandemia	Sim, Não
	optaria_metodo _ead_flexibilizado _pandemia	Se optaria pelo método EAD durante pandemia	Sim, Não

Fonte: Autoria própria

APÊNDICE B - REGRAS GERADAS PELO ALGORITMO SOM

Quadro 11 - Regras geradas pela base SOM para o *cluster* C1-SOM

CLUSTER 1
Regra 1 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** optaria_metodo_ead = "Não" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E** desvantagem_suspensao_aulas = "Sim" **Então: C1-SOM**
Regra 2 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** nivel_perfil_aluno_ead = "Disciplinado, Flexível (2 ou menos características)" **E** uso_software_educativo = "Sim" **Então: C1-SOM**
Regra 3 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** optaria_metodo_ead = "Não" **E** sanar_duvidas_ead = "Pouca/Muita dificuldade" **E** lugar_calmo = "Sim" **Então: C1-SOM**
Regra 4 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** optaria_metodo_ead = "Sim" **E** bolsista = "Sim" **E** nivel_perfil_aluno_ead = "Proativo, Organizado, Enxerga prof. como orient. (3 ou menos características)" **Então: C1-SOM**

Fonte: Autoria própria

Quadro 12 - Regras geradas pela base SOM para o *cluster* C2-SOM

CLUSTER 2
Regra 1 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** lugar_calmo = "Sim" **E** situacao_emprego = "Trabalhando em Home Office, Dispensado das funções, Não trabalhando" **E** importancia_interacao_colegas = "Importante, Pouco importante" **Então: C2-SOM**
Regra 2 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** lugar_calmo = "Sim" **E** uso_software_educativo = "Sim/Talvez optaria pelo uso" **E** optaria_metodo_ead = "Não" **E** tecnologias_ead_conhece = "Hangout Meet Google Education, Skype (2 ou mais tecnologias)" **Então: C2-SOM**
Regra 3 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** sexo = "Masculino" **E** uso_software_educativo = "Sim/Talvez optaria pelo uso" **Então: C2-SOM**
Regra 4 **SE** optaria_metodo_ead_flex_pandemia = "Não" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E** lugar_calmo = "Sim" **Então: C2-SOM**

Fonte: Autoria própria

Quadro 12 – Regras geradas pela base SOM para o *cluster* C2-SOM

CLUSTER 2
Regra 5
SE optaria_metodo_ead_flex_pandemia = "Não" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E** lugar_calmo = "Sim" **E** situacao_emprego = "Trabalhando, dirigindo-se ao local de trabalho" **E** desvantagem_suspensao_aulas = "Sim" **Então: C2-SOM**
Regra 6
SE optaria_metodo_ead_flex_pandemia = "Não" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E** lugar_calmo = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** bolsista = "Sim" **E** foco_para_ead = "Bom/Médio foco" **Então: C2-SOM**

Fonte: Autoria própria

Quadro 13 - Regras geradas pela base SOM para o *cluster* C3-SOM

CLUSTER 3
Regra 1 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** nivel_perfil_aluno_ead = "Responsável, Disciplinado (2 ou mais caracteristicas)" **E** optaria_metodo_ead = "Sim" **E** nivel_perfil_aluno_ead = "Proativo, Organizado, Enxerga profs. como orient. (3 ou mais caracter.)" **Então: C3-SOM**
Regra 2 **SE** optaria_metodo_ead_flex_pandemia = "Sim" **E** optaria_metodo_ead = "Sim" **E** bolsista = "Sim" **E** nivel_perfil_aluno_ead = "Proativo, Organizado, Enxerga profs. como orient. (3 ou mais caracter.)" **Então: C3-SOM**
Regra 3 **SE** idade = "Maior ou igual a 34 anos" **E** desvantagem_suspensao_aulas = "Não" **E** idade = "Menor ou igual a 36 anos" **Então: C3-SOM**

Fonte: Autoria própria

APÊNDICE C - REGRAS GERADAS PELO ALGORITMO K-MEANS COM K = 3

Quadro 14 - Regras geradas pela base *kMeans*3 para o *cluster* C1-k3

CLUSTER 1
Regra 1 **SE** sexo = “Masculino” **E** optaria_metodo_ead_flex_pandemia = “Sim” **E** foco_para_ead = “Nenhum foco“ **E** sanar_duvidas_ead = “Pouca ou nenhuma dificuldade” **Então: C1-k3**
Regra 2 **SE** sexo = "Masculino" **E** doenca_psicologica = "Síndrome do Pânico (1 doença apenas)" **E** animo_para_atividades_ead = "Muito positivo, Positivo, Negativo" **E** classe_social = "Classe média baixa, Classe média, Classe média alta" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **Então: C1-k3**
Regra 3 **SE** nivel_perfil_aluno_ead = "Organizado, Responsável, Disciplinado, Flexível, Enxerga profs. como orient. (5 ou mais caracter.)" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **Então: C1-k3**
Regra 4 **SE** sanar_duvidas_ead = "Nenhuma dificuldade" **E** foco_para_ead = "Bom/Médio foco" **E** tecnologias_ead_conhece = "Moodle, Skype (2 ou mais tecnologias)" **Então: C1-k3**
Regra 5 **SE** sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Sim" **E** foco_para_ead = "Bom/Médio foco" **Então: C1-k3**

Fonte: Autoria própria

Quadro 14 – Regras geradas pela base *kMeans*3 para o *cluster* C1-k3

CLUSTER 1
Regra 6
SE sexo = "Masculino" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** foco_para_ead = "Bom foco" **Então: C1-k3**

Fonte: Autoria própria

Quadro 15 - Regras geradas pela base *kMeans*3 para o *cluster* C2-k3

CLUSTER 2
Regra 1
SE desvantagem_suspensao_aulas = “Sim” **E** optaria_metodo_ead_flex_pandemia = “Não” **E** moradia = “Com família, Divide apartamento ou República” **E** importancia_interacao_colegas = “Muito importante” **Então: C2-k3**
Regra 2
SE sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Não" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** foco_para_ead = "Pouco/Nenhum foco" **E** bolsista = "Não" **E** nivel_academico = "Graduação, Especialização" **Então: C2-k3**
Regra 3
SE foco_para_ead = "Pouco/Nenhum foco" **E** sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Não" **E** sanar_duvidas_ead = "Pouca/Muita dificuldade" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **Então: C2-k3**
Regra 4
SE sexo = "Masculino" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** sanar_duvidas_ead = "Pouca/Muita dificuldade" **Então: C2-k3**

Fonte: Autoria própria

Quadro 15 – Regras geradas pela base *kMeans*3 para o *cluster* C2-k3

CLUSTER 2
Regra 5
SE sexo = "Feminino" **E** optaria_metodo_ead_flex_pandemia = "Não" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** foco_para_ead = "Pouco/Nenhum foco" **Então: C2-k3**
Regra 6
SE sexo = "Feminino" **E** optaria_metodo_ead = "Não" **E** adaptabilidade_ead = "Média/Difícil adaptação" **Então: C2-k3**

Fonte: Autoria própria

APÊNDICE D - REGRAS GERADAS PELO ALGORITMO K-MEANS COM K = 5

Quadro 16 - Regras geradas pela base *kMeans*5 para o *cluster* C1-k5

CLUSTER 1
Regra 1 **SE** grupo_de_risco_covid19 = “Sim” **E** optaria_metodo_ead_flex_pandemia = “Sim” **E** filhos = “Não possui filhos” **E** sexo = “Masculino” **E** animo_para_atividades_ead = “Negativo e/ou muito negativo” **E** classe_social = “Classe baixa e/ou média baixa” **Então: C1-k5**
Regra 2 **SE** grupo_de_risco_covid19 = "Sim" **E** tecnologias_fundamentais_ead = "Videoaulas, Chats, Bibliotecas Virtuais (3 ou menos tecnologias)" **E** doenca_psicologica = "Transtorno de Ansiedade (1 ou mais doenças) **E** sexo = "Masculino" **Então: C1-k5**
Regra 3 **SE** grupo_de_risco_covid19 = "Sim" **E** nivel_academico = "Graduação, Especialização" **E** uso_software_educativo = "Talvez/Não optaria pelo uso" **Então: C1-k5**
Regra 4 **SE** grupo_de_risco_covid19 = "Sim" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **E** adaptabilidade_ead = "Média/Difícil adaptação" **Então: C1-k5**
Regra 5 **SE** foco_para_ead = "Nenhum foco" **E** idade = "Igual ou menor que 25 anos" **Então: C1-k5**

Fonte: Autoria própria

Quadro 16 – Regras geradas pela base *kMeans*5 para o *cluster* C1-k5

CLUSTER 1
Regra 6
SE grupo_de_risco_covid19 = "Sim" **E** optaria_metodo_ead_flexibilizado_pandemia = "Não" **Então: C1-k5**

Fonte: Autoria própria

Quadro 17 - Regras geradas pela base *kMeans*5 para o *cluster* C2-k5

CLUSTER 2
Regra 1 **SE** foco_para_ead = “Nenhum foco, não se adapta” **E** grupo_de_risco_covid19 = “Não” **E** uso_software_educativo = “Sim e/ou talvez” **E** optaria_metodo_ead = “Não” **Então: C2-k5**
Regra 2 **SE** sexo = "Masculino " **E** optaria_metodo_ead_flex_pandemia = "Não" **E** grupo_de_risco_covid19 = "Não" **E** nivel_academico = "Graduação" **Então: C2-k5**
Regra 3 **SE** animo_para_atividades_ead = "Negativo, Muito negativo" **E** adaptabilidade_ead = "Média/Difícil adaptação" **E** grupo_de_risco_covid19 = "Não" **Então: C2-k5**
Regra 4 **SE** sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Não" **E** nivel_academico = "Graduação" **Então: C2-k5**

Fonte: Autoria própria

Quadro 18 - Regras geradas pela base *kMeans*5 para o *cluster* C3-k5

CLUSTER 3
Regra 1
SE importancia_interacao_professor = “Importante e/ou pouco importante” **E** idade = “34 anos ou menos” **E** classe_social = “Classe média baixa, média e/ou média alta” **E** nivel_perfil_aluno_ead = “Disciplinado (1 ou mais características)” **Então: C3-k5**
Regra 2
SE grupo_de_risco_covid19 = “Não” **E** sexo = “Masculino” **E** optaria_metodo_ead_flex_pandemia = “Sim” **E** foco_para_ead = “Bom e/ou médio foco” **E** filhos = “Não possui filhos” **E** adaptabilidade_ferramentas_ead = “Fácil adaptação” **Então: C3-k5**
Regra 3
SE animo_para_atividades_ead = "Muito positivo, Positivo" **E** filhos = "Não possui filhos, Possui filhos mas não seria afetado pelo EAD" **E** optaria_metodo_ead_flex_pandemia = "Sim" **E** sexo = "Masculino" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **Então: C3-k5**
Regra 4
SE sexo = "Feminino" **E** optaria_metodo_ead = "Sim" **E** lugar_calmo = "Sim” **E** foco_para_ead = "Bom/Médio foco" **E** filhos = "Não possui filhos, Possui filhos mas não seria afetado pelo EAD" **E** sanar_duvidas_ead = "Nenhuma/Pouca dificuldade" **E** importancia_interacao_colegas = "Importante/Pouco importante" **Então: C3-k5**

Fonte: Autoria própria

Quadro 18 – Regras geradas pela base *kMeans*5 para o *cluster* C3-k5

CLUSTER 3
Regra 5
SE grupo_de_risco_covid19 = "Sim" **E** optaria_metodo_ead_flex_pandemia = "Sim" **E** filhos = "Não possui filhos" **E** sexo = "Masculino" **E** animo_para_atividades_ead = "Negativo, Muito negativo" **E** classe_social = "Classe média / Média alta" **Então: C3-k5**
Regra 6
SE situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado das funções" **E** classe_social = "Clsse média, Classe média alta" **E** adaptabilidade_ead = "Fácil adaptação" **Então: C3-k5**
Regra 7
SE sexo = "Feminino" **E** filhos = "Não possui filhos" **E** animo_para_atividades_ead = "Muito positivo, Positivo" **E** lugar_calmo = "Sim" **E** importancia_interacao_professor = "Importante, Pouco importante" **E** nivel_estimulo_negativo_estresse = "Necessidade estudar várias matérias, Acúmulo de atividades, Preocupação com o futuro (3 ou menos estímulos)" **Então: C3-k5**

Fonte: Autoria própria

Quadro 19 - Regras geradas pela base *kMeans*5 para o *cluster* C4-k5

CLUSTER 4
Regra 1 **SE** nivel_academico = “Especialização, Mestrado e/ou Doutorado” **E** optaria_metodo_ead_flex_pandemia = “Não” **E** situacao_emprego = “Trabalhando dirigindo-se ao local de trabalho e/ou em Home Office” **Então: C4-k5**
Regra 2 **SE** filhos = "Possui filhos com ou sem atividades afetadas" **E** sanar_duvidas_ead = "Pouca/Muita dificuldade" **E** situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado das funções" **E** optaria_metodo_ead = "Não" **Então: C4-k5**
Regra 3 **SE** situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado das funções" **E** lugar_calmo = "Não" **Então: C4-k5**
Regra 4 **SE** nivel_academico = "Especialização, Mestrado, Doutorado" **E** optaria_metodo_ead_flex_pandemia = "Não" **E** situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando Home Office" **Então: C4-k5**
Regra 5 **SE** idade = "Maior ou igual a 27 anos" **E** importancia_interacao_professor = "Muito importante" **E** nivel_perfil_aluno_ead = "Organizado, Responsável, Disciplinado (3 ou menos caracter.)" **E** situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado das funções" **Então: C4-k5**

Fonte: Autoria própria

Quadro 19 – Regras geradas pela base *kMeans*5 para o *cluster* C4-k5

CLUSTER 4
Regra 6
SE grupo_de_risco_covid19 = "Não" **E** sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Sim" **E** nivel_academico = "Especialização, Mestrado, Doutorado" **E** situacao_emprego = "Dirigindo-se ao trabalho, Trabalhando em Home Office, Dispensado das funções" **Então: C4-k5**
Regra 7
SE grupo_de_risco_covid19 = "Não" **E** sexo = "Masculino" **E** optaria_metodo_ead_flex_pandemia = "Sim" **E** foco_para_ead = "Bom/Médio foco" **E** filhos = "Possui filhos com ou sem atividades afetadas" **E** importancia_interacao_professor = "Muito importante" **Então: C4-k5**
Regra 8
SE grupo_de_risco_covid19 = "Não" **E** sexo = "Feminino" **E** filhos = "Possui filhos com ou sem atividades afetadas" **E** doenca_psicologica = "Não possui doença psicológica" **Então: C4-k5**

Fonte: Autoria própria