Universidade Federal do Amazonas

Instituto de Computação

Programa de Pós-Graduação em Informática

Predição de Evasão de Cursos Técnicos em EaD através de Técnicas de Aprendizado de Máquina em Duas Etapas

MARIELA MIZOTA TAMADA

Manaus - Amazonas

Junho de 2022

Universidade Federal do Amazonas
Instituto de Computação
Programa de Pós-Graduação em Informática

Predição de Evasão de Cursos Técnicos em EaD através de Técnicas de Aprendizado de Máquina em Duas Etapas

Tese apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas (PPGI-UFAM), como requisito parcial para a obtenção do título de Doutora em Informática.

Orientador: Prof. Dr. Rafael Giusti
Coorientador: Prof. Dr. José Francisco de Magalhães Netto

Manaus - Amazonas

Junho, 2022

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Tamada, Mariela Mizota
T153p Predição de evasão de cursos técnicos em EaD através de técnicas de aprendizado de máquina em duas etapas / Mariela Mizota Tamada . 2022
155 f.: il. color; 31 cm.

Orientador: Rafael Giusti
Coorientador: José Francisco de Magalhães Netto
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Educação a distância. 2. Predição de evasão. 3. Modelos de agrupamento e classificação. 4. Logs do Moodle. 5. Ensino técnico. I. Giusti, Rafael. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Predição de Evasão de Cursos Técnicos em EaD através de Técnicas de Aprendizado de Máquina em Duas Etapas"

MARIELA MIZOTA TAMADA

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Rafael Giusti - PRESIDENTE

Prof. David Braga Fernandes de Oliveira - MEMBRO INTERNO

Fabíola Guerra Nakamura

Profa. Fabíola Guerra Nakamura - MEMBRO INTERNO

Prof. Crediné Silva de Menezes - MEMBRO EXTERNO

Prof. José Luiz de Souza Pio - MEMBRO EXTERNO

Manaus, 27 de Junho de 2022

Dedico esta Tese aos meus pais e a minha família por todo o apoio recebido ao longo da minha caminhada.

AGRADECIMENTOS

Agradeço à sociedade como um todo porque tenho uma história de lutas, alegrias, frustrações e conquistas, e todos aqueles que, de alguma forma, cruzei nesta caminhada influenciaram na minha formação e realização de mais um projeto de vida.

A Horácio, meu esposo, e aos meus filhos, Márcio e Luciana, razão da minha vida e minha luta, por todo o apoio e paciência nessa trajetória.

Aos meus pais Akito (*in memoriam*) e Yoko, ambos imigrantes japoneses pós-guerra, que tiveram que sair do país de origem atrás de novas oportunidades e me ensinaram a importância dos estudos e do respeito.

Ao meu orientador Prof. Dr. Rafael Giusti, um excelente professor que aceitou a minha proposta de monografia e me incentivou até o fim. Agradeço ainda todo o ensino, apoio, paciência e contribuições fundamentais na escrita dos trabalhos que publicamos. Foram longos 18 meses ininterruptos de reuniões online semanais com conversas que deram resultados práticos importantes.

Agradeço, de igual forma, ao meu coorientador Prof. Dr. José Francisco de Magalhães Netto por poder fazer parte do seu grupo de pesquisa e ter acreditado nas minhas capacidades, contando com o seu apoio desde o início do curso.

À Profa. Dra. Fabíola Guerra Nakamura, quero expressar a minha gratidão pela forma amiga e generosa que me atendeu toda vez que precisei.

Aos membros da banca examinadora da qualificação e da defesa pelas valiosas contribuições para concluir a jornada do doutorado.

Aos meus colegas do grupo de IA na Educação (Arcanjo, Joethe, Andreza, Dhanielly, Thais e Márcio) pela amizade, apoio e parceria no decorrer da jornada.

Ao meu colega Albert e esposa Elda, que me receberam na sua casa nas tantas vezes que tive que me deslocar até Manaus. O Albert me ajudou muito nas disciplinas e me deu apoio emocional quando eu pensava que não conseguiria concluir a pesquisa.

Aos meus colegas da turma ‘Fora da sede’, que cursamos as disciplinas em Rio Branco (AC) compartilhando muitas horas de estudos.

À Universidade Federal do Amazonas (UFAM) pela oportunidade e suporte concedido para viabilizar esta formação. Em especial aos professores e funcionários da secretaria por todo apoio, disponibilidade e atenção concedida durante este período.

Ao Instituto Federal de Rondônia (IFRO) pela confiança depositada em concluir este curso tão importante para melhorar a formação do seu corpo docente. Ao meu colega Celso, quem consultei várias vezes sobre o banco de dados do nosso AVA. À minha colega e grande amiga Maria Ivanilse, que também estamos juntas neste curso, pelo apoio e parceria, e estou torcendo por você.

À Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) pelo projeto RH-Ti Capacitação FAPEAM, da UFAM em parceria com a Universidade Federal do Acre (UFAC) e a Universidade Federal de Rondônia (UNIR), que possibilitou a formação de vários mestres e uma doutora, eu.

Ao Brasil por oferecer uma infraestrutura de instituições públicas de ensino de qualidade e possibilitar o meu mestrado e doutorado. Espero retribuir à comunidade.

Finalizando, faço uma reflexão sobre a difícil situação que estamos passando com a pandemia de Covid-19. Foi necessária muita persistência e determinação para conciliar a conclusão do doutorado com as atividades do trabalho docente, que não foi interrompida, a família e a situação pandêmica, e que por algumas vezes senti a necessidade de parar, refletir e orar.

Predição de Evasão de Cursos Técnicos em EaD
através de Técnicas de Aprendizado de Máquina em Duas Etapas

RESUMO

A Mineração de Dados Educacionais integra inúmeras técnicas que dão suporte a processamento e análises de dados gerados e coletados nos Ambientes Virtuais de Aprendizagem (AVA). Ela tem o intuito de extrair informações relevantes no ambiente escolar e emprega como principal técnica o Aprendizado de Máquina. O objetivo geral desta tese é definir uma metodologia que auxilie os gestores educacionais na detecção do risco de evasão de estudantes em EaD, com base nas mudanças das características de comportamento extraídas por técnicas de Aprendizado de Máquina. Devido ao contexto dinâmico do ambiente educacional, os modelos propostos são construídos em diferentes momentos do curso, com dados coletados das interações dos estudantes com o AVA aos 10%, 25%, 50% e 75% dos 2 anos de duração de cursos técnicos. Para uma melhor predição, apresenta-se uma técnica em duas etapas, com abordagem não supervisionada para agrupar estudantes sem definir um número de grupos a priori, e então uma abordagem supervisionada na qual o *cluster* atribuído a cada estudante é um novo atributo de entrada para um modelo de classificação. As técnicas não supervisionadas também são utilizadas como ferramenta para estudar o domínio dos dados. Como resultados, o algoritmo de agrupamento *K-means* revelou a presença de quatro grupos coerentes de estudantes pelas suas características de comportamento no AVA aos 10% do andamento do curso, e contribui na melhora de predição dos alunos em risco de evasão, com retorno da métrica F1 acima de 80% nos diferentes classificadores testados. Os resultados mostram alta correlação com a conclusão ou não conclusão do curso e traz *insights* e novos conhecimentos sobre os estudantes. Este trabalho aborda essas duas técnicas em cascata ou em duas etapas, e não tem se encontrado pesquisas com essa abordagem em Educação a Distância. Além desses resultados, este trabalho é relevante por pesquisar os cursos técnicos, que são considerados de grande importância para o desenvolvimento social e econômico do país, embora sejam escassos pesquisas e estudos que foquem nesse nível de ensino.

Palavras-chave: Educação a Distância; Predição de Evasão; Modelos de agrupamento e classificação; Logs do Moodle; Ensino Técnico.

Predicting Dropout of Technical Courses at Virtual Learning with Machine Learning in Two Steps

ABSTRACT

Educational Data Mining integrates numerous techniques that support the processing and analysis of data generated and collected from Learning Management Systems (LMS). It aims to extract relevant information in the school environment and uses Machine Learning as its main technique. The general objective of this thesis is to define a methodology that helps educational managers in detecting the risk of dropout of students in distance education, based on changes in behavior characteristics extracted by Machine Learning techniques. Due to the dynamic context of the educational environment, the proposed models are built at different times of the course, with data collected from the students' interactions with the LMS at 10%, 25%, 50% and 75% of the two-year duration of technical courses. To provide better predictors, a two-step technique is presented, with an unsupervised approach to group students without defining a number of groups a priori, and then a supervised approach in which the cluster assigned to each student is a new input attribute to a classification model. Unsupervised techniques are also employed as a tool to study the data domain. As a result, the K-means clustering algorithm revealed the presence of four coherent groups of students according to their behavior toward the LMS, and contributes to the improvement of the prediction of students at risk of dropping out, with a return of the F1 metric above 80% in the different classifiers tested. The results show a high correlation with the completion or non-completion of the course and bring insights and new knowledge about the students. This work addresses these two techniques in cascade or in two stages, and no research with that approach has been found in Distance Education. In addition to these results, this work is relevant due to being focused on technical courses, which are considered of great importance for the social and economic development of the country, despite research and studies focusing on this level of education being scarce.

Keywords: Distance education; Dropout prediction; Classification and clustering model; Moodle logs; Technical course.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABED	Associação Brasileira de Educação a distância
AM	Aprendizado de Máquina
AVA	Ambiente Virtual de Aprendizagem
AVEA	Ambiente Virtual de Ensino-Aprendizagem
BD	Banco de Dados
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBIE	Congresso Brasileiro de Informática na Educação
CDVAA	Coordenação de Design Visual e Ambientes de Aprendizagem
CDMI	Coordenação de Material e Designer Instrucional
CEFET	Centro Federal de Educação Tecnológica
CEP	Comitê de Ética em Pesquisa
CONIF	Conselho Nacional dos Institutos Federais
CPF	Cadastro de Pessoa Física
CPGA	Coordenação de Produção e Geração Audiovisual
CSV	*Comma-separated Values*
DT	*Decision Tree*
EAD	Educação a distância
EDM	Educational Data Mining
Encceja	Exame Nacional para Certificação de Competências de Jovens e Adultos
EPT	Ensino Técnico Profissionalizante
EUA	Estados Unidos de América
FIC	Formação Inicial e Continuada
GBT	Gradient Boost Tree
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDE	*Integrated Development Environment*
IEEE	*Institute of Electrical and Electronics Engineers*
IES	Instituição de Ensino Superior
IF	Instituto Federal
IFRO	Instituto Federal de Educação, Ciência e Tecnologia de Rondônia
INEP	Instituto de Pesquisas Educacionais Anísio Teixeira
ITS	*Intelligent Tutorial System*

KDD	*Knowledge discovery from data*
KDDB	*Knowledge discovery in database*
KNN	*K Nearest-Neighbors*
LA	*Learning Analytics*
LMS	*Learning Management System*
LR	*Logistic Regression*
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação e Cultura do Brasil
MOOC	*Massive Open Online Courses*
Moodle	*Modular Object-Oriented Dynamic Learning Environment*
ML	*Machine Learning*
MSL	Mapeamento Sistemático da Literatura
NB	Naive Bayes
NN	Neural Network
NTIC	Novas Tecnologias da Informação e da Comunicação
OCDE	Organização para Cooperação e Desenvolvimento Econômico
PnadC	Pesquisa Nacional por Amostra de Domicílios Contínua
PPC	Projeto Pedagógico do Curso
Pro Funcionário	Programa de Formação Inicial em Serviço dos Profissionais da Educação Básica dos Sistemas Públicos de Ensino
Pronatec	Programa Nacional de Acesso ao Ensino Técnico e Emprego
PVZN	Porto Velho Zona Norte
RF	*Random Forest*
RO	Rondônia
RS	Revisão Sistemática
RSL	Revisão Sistemática da Literatura
SCORM	*Sharable Content Object Reference Model*
SISTEC	Sistema Nacional de Informações da Educação Profissional e Tecnológica
StArt	*State of Art throught Systematic Review*
STI	Sistema Tutorial Inteligente
SIGA	Sistema Integrado de Gestão Acadêmica
SUAP	Sistema Unificado de Administração Pública
SVM	*Support Vector Machine*
TIC	Tecnologias da Informação e da Comunicação

VLE	*Virtual Learning System*
WEKA	*Waikato Environment for Knowledge Analysis*
YALE	*Yet Another Learning Environment*

SUMÁRIO

Capítulo 1
Introdução

O presente trabalho apresenta um modelo metodológico que está inserido na área de pesquisa de Informática na Educação. Este capítulo aborda a caracterização do problema e as estratégias para solução, hipóteses, objetivos, motivação e escopo desta pesquisa, seguida da justificativa e método que norteou o desenvolvimento deste trabalho. Nas últimas seções são descritas as potenciais contribuições, e o capítulo se encerra com a apresentação da organização dos demais capítulos que compõem esta tese. Consideram-se as expressões evasão, desistência, reprovação, insucesso ou fracasso no tempo de integralização do curso para referenciar, de um modo geral, a evasão escolar.

1.1 Considerações Iniciais

A Educação a Distância (EaD) traz desafios para docentes, técnicos, administradores e estudantes para gerenciar o processo de ensino-aprendizado por meio de uma plataforma virtual e, com isso, acentua-se o problema da evasão de estudantes. Para Seidman (1996), a chave para reduzir os níveis de evasão consiste na identificação precoce de estudantes em risco, permitindo intervenção intensiva e contínua. A massificação da EaD e o grande número de estudantes leva à necessidade de que essa identificação seja apoiada ou até mesmo realizada por sistemas inteligentes, capazes de diagnosticar com antecedência aqueles estudantes com alta probabilidade de desistir e que também possam fornecer aos gestores uma visualização com informações estratégicas para identificar os casos e as causas de prováveis abandonos e desistências do curso.

O Ambiente Virtual de Aprendizagem (AVA) apoia todo o processo de comunicação entre os estudantes, professores, materiais didáticos e a comunidade, fazendo com que todos participem de modo interativo (KEMCZINSKI, 2005). Os AVAs armazenam, além de registros acadêmicos, como frequência e nota das atividades, todo o histórico de ações dos estudantes, tais como visualização e envio de tarefas, tentativas de resolução de questionários, visualização de aulas pré-gravadas, dentre outras interações na plataforma virtual. À medida que esses dados se tornam mais completos e complexos, é possível desenvolver algoritmos capazes de realizar análises e predições igualmente mais complexas e também mais úteis.

A Mineração de Dados Educacionais (MDE) é a aplicação de técnicas de Mineração de Dados (MD) a dados educacionais e, portanto, seu objetivo é analisar esses tipos de dados para resolver problemas de pesquisa educacional (BARNES *et al*., 2009). A MDE é um análogo do processo de extração de conhecimento de base de dados no contexto educacional, e emprega Aprendizado de Máquina (AM) como sua principal ferramenta. Para Mitchell (1997), o AM é um campo multidisciplinar voltado à construção de programas que aprendem automaticamente com a experiência, em contraste com programas que são projetados manualmente para resolver um problema.

Os algoritmos de AM podem ser usados para extrair modelos preditivos ou descritivos em diferentes tarefas, tais como classificação, regressão e agrupamento, dentre outras. No contexto educacional, técnicas de classificação podem ser usadas, por exemplo, para predizer casos de abandono de um curso, e técnicas de agrupamento podem ser usadas para encontrar grupos de estudantes com comportamentos semelhantes com respeito ao curso e, dessa forma, obter uma descrição dos grupos.

Um recurso importante dos modelos preditivos construídos a partir de dados do AVA é sua capacidade de predição antecipada. Se um modelo é capaz de identificar estudantes em risco nos estágios iniciais do curso, os professores podem conceber maneiras de ajudar sua aprendizagem (MACFADYEN e DAWSON, 2010). No entanto, a maioria dos trabalhos tipicamente constrói modelos preditivos a partir de dados compilados ao final do curso, negligenciando o valor prático de uma predição antecipada enquanto o curso ainda está em andamento (HU *et al*., 2014).

Esse panorama desencadeou a motivação pela busca e desenvolvimento de uma solução tecnológica que pudesse contribuir na redução da taxa de evasão escolar identificada em cursos EaD.

A identificação e viabilidade para o desenvolvimento desse método foram possíveis mediante acesso ao banco de dados da plataforma de aprendizagem virtual do IFRO (Instituto Federal de Educação, Ciência e Tecnologia de Rondônia), e com estudos das técnicas supervisionadas e não supervisionadas do AM para caracterização dos grupos de estudantes e posterior predição de risco de evasão.

Esta Tese de Doutorado colabora e contribui com o contexto educacional em EaD, cenário este em contínuo crescimento e evolução, e tem como objetivo definir uma metodologia que considere o dinamismo das mudanças de comportamento do estudante no AVA e auxilie

os gestores educacionais a fazer intervenções precoces para reduzir o risco de evasão, com base em informações extraídas por processos de MD e técnicas de AM, de cursos ofertados na modalidade a distância.

1.2 Descrição do Problema e Estratégia de Solução

O grande marco de 2017 para a EaD foi a flexibilização da regulamentação e oferta de cursos a distância em níveis superiores por meio da Portaria Normativa no. 11, de 20 de junho de 2017 (BRASIL, 2017). Com essa mudança, passou a ser possível criar instituições que oferecessem EaD sem a contrapartida presencial, e facilitou-se a criação de polos (ABED, 2018, p.17). Mais recentemente, o interesse por esse assunto aumentou à medida que escolas de todos os níveis buscavam opções para continuar ministrando cursos durante a pandemia de Covid-19 (MAREK, 2021; ALTURKI, 2021).

O aprendizado formal dentro do âmbito escolar é importante para o desenvolvimento humano no campo pessoal, intelectual, profissional, social e financeiro. O estudante almeja a obtenção de um diploma, que pode se transformar em ascensão nesses aspectos, mas depara-se com inúmeros desafios que o levam a abandonar o curso. Os cursos avaliados neste trabalho tiveram taxa média de insucesso acima de 50%, incluindo casos de reprovação, evasão e desistência.

Um ponto importante na identificação precoce dos estudantes sujeitos a evasão e/ou retenção é o tempo que o estudante leva para tomar uma decisão. Não é uma decisão repentina, pelo que devemos detectar o desempenho e as dificuldades com antecedência a saída definitiva do estudante (BARDAGI, 2009), e melhorar a eficiência na redução de evasão.

A fim de caracterizar e confirmar essa tendência na comunidade acadêmica, foi executada uma Revisão Sistemática da Literatura (RSL) sobre trabalhos internacionais, publicados até 2019, que abordam técnicas de predição de evasão em EaD. Essa RSL foi posteriormente complementada com outras RSLs internacionais e nacionais dos últimos 3 anos. Em publicações internacionais, a maioria dos trabalhos aborda os *Massive Open Online Courses* (MOOC), que abrangem cursos curtos com um público massivo de milhares de estudantes *online* simultâneos em períodos com calendário livre sem uma data de início e de conclusão. Também é prevalente o uso de técnicas supervisionadas que comparam os resultados dos algoritmos de AM.

No caso de publicações nacionais recentes, este trabalho baseou-se principalmente nos resultados de Colpo *et al.* (2020), que analisaram publicações que usam técnicas de MDE para predição de evasão escolar no Brasil e publicações realizadas no Congresso Brasileiro de Informática na Educação (CBIE), principal evento da área no país. Os autores concluíram que essas pesquisas demonstram maior interesse da evasão no nível de graduação, na modalidade presencial e na rede pública de ensino, unanimidade no emprego de técnicas de classificação, com predominância de algoritmos de árvores de decisão, geralmente usando a ferramenta Weka. A pesquisa ressalta a carência de trabalhos na esfera pública nos níveis de ensino fundamental e técnico.

Lüsche e Dore (2011), Meira (2015), Linke e Nogueira (2017) já vêm relatando na última década sobre essa falta de pesquisas na área com foco nos estudantes do ensino técnico brasileiro, tanto público quanto privado, que enfrenta altas taxas de desistência escolar. Isso pode ser um indicativo de uma área de pesquisa que pouco tem avançado nesses últimos anos.

Essas considerações são relevantes para o presente trabalho, que combina técnicas de AM supervisionado e não supervisionado na análise de cursos em EaD.

Uma questão importante é que os docentes dos cursos analisados têm acesso apenas às notas e frequências dos discentes de suas próprias disciplinas, sem acesso ao importante histórico do estudante, dados de apoios, bolsas, participações em atividades extracurriculares, extensão, estágio etc. Esses dados estão disponíveis apenas às equipes de gestão, que frequentemente precisam do uso de sistemas distintos. Nota e frequência são as únicas informações disponibilizadas pelo AVA para tomadas de decisões pelo docente. Essas informações são relevantes, porém fornecem uma visão incompleta do estudante e não permitem ao docente identificar adequadamente os casos de risco. Entretanto, existe informação valiosa “oculta” nos dados dos AVAs, que poderiam ser melhor explorados por docentes e/ou coordenadores de curso. Tais dados podem ser analisados com técnicas de MDE, a fim de extrair novas informações, inicialmente implícitos.

Para melhorar o acompanhamento do comportamento dos estudantes nos AVAs, Sha *et al.* (2012) mostram que dados do fluxo de cliques (acessos) e/ou visualizações são interessantes, pois são coletados enquanto o comportamento de aprendizagem está acontecendo. Esse tipo de dado se contrasta a outros, como frequência e nota, que podem ser influenciados pelo fato de que o estudante sabe que será avaliado por critérios baseados nesses dados.

A carência de trabalhos no ensino técnico mencionada por Colpo et. al. (2020) é, em parte, suprida neste trabalho, cuja abordagem é validada com um caso descritivo para instanciar

o arcabouço conceitual no IFRO, uma instituição pública federal no Brasil, que oferece cursos em formatos totalmente online, híbrido ou presencial em diferentes níveis, incluindo ensino superior, técnico concomitante (dupla matrícula), técnico subsequente (após o ensino médio), técnico integrado (ensino médio de 4 anos), graduação e pós-graduação. Esta pesquisa tem como foco os estudantes dos cursos técnicos concomitantes ao ensino médio, com duração de 2 anos.

O problema está delimitado aos cursos na modalidade EaD, com análise de 9 turmas, entre 2016 e 2019, dos cursos técnicos concomitantes de Informática para Internet, de Finanças e de Administração. Os dados são analisados, limpos, selecionados e divididos para que as partes sirvam para treinamento e testes dos algoritmos de AM, com técnicas supervisionadas e não supervisionadas, considerando o contexto e dinamismo do ambiente educacional para descrever o comportamento dos estudantes no AVA que identifique um estudante em risco de evasão.

1.3 Hipótese

Este trabalho aborda modelos de predição de evasão que propiciem intervenção para reduzir a evasão de estudantes de cursos na educação a distância. Esta pesquisa testa a Hipótese Inicial definida a seguir:

> Hipótese Inicial: *A análise de agrupamentos dos estudantes pela forma de interação com a interface da plataforma de aprendizagem virtual permite predizer sucesso ou não na conclusão do curso em até 10% do início do curso.*

1.4 Objetivos

O objetivo geral desta tese é definir uma metodologia que auxilie os gestores educacionais na detecção do risco de evasão de estudantes em EaD durante o andamento do curso, com base nas mudanças das características de comportamento extraídas por processos de Mineração de Dados Educacionais e técnicas de Aprendizado de Máquina.

Para alcançar esse objetivo, definem-se os seguintes objetivos específicos[1].

[1] "definir os objetivos específicos significa aprofundar as intenções expressas nos objetivos gerais" (CERVO, BERVIAN e DA SILVA, 2007). Portanto, nesta parte, o autor deve expor suas metas para se chegar ao objetivo geral da pesquisa. As metas consistem em várias etapas que devem ser realizadas para que se consiga alcançar o resultado desejado.

1. Caracterizar as mudanças de comportamento dos estudantes durante o andamento do curso após 10%, 25%, 50% e 75% da duração do curso.
2. Agrupar os estudantes pelos seus comportamentos com relação ao AVA e correlacionar os grupos de estudantes com a conclusão ou não conclusão do curso.
3. Promover uma predição confiável e antecipada do risco de evasão usando dados de logs do AVA obtidos nos primeiros meses do curso.

1.5 Justificativa

Com base na revisão da literatura que foi realizada durante a execução deste trabalho, pode-se apontar os seguintes fatos como justificativas para investir em pesquisas para predizer e mitigar a evasão escolar:

- O custo de uma evasão é seis vezes maior que a prevenção para evitar o abandono escolar (BAAS, 1991).
- As metodologias têm sido ineficientes no uso dos dados, perdendo oportunidades de intervenção a tempo (SIEMENS e LONG, 2011).
- A Associação Brasileira de Educação a distância-ABED, no censo de 2019, mostra que 32% de instituições que oferecem cursos regulamentados totalmente a distância têm uma evasão entre 11% e 25% e que 19% das instituições têm entre 26% e 50%.
- A evasão nos cursos de EaD tem causado perdas que vão desde a ociosidade de recursos pessoais e materiais das instituições até o fechamento de cursos com muitos estudantes evadidos. O problema é agravado devido aos poucos trabalhos de combate à evasão de estudantes em cursos dessa modalidade de ensino somado à inexistência de uma política efetiva de combate à evasão nos cursos de EaD (BITTENCOURT e MERCADO, 2014).
- Atualmente, o processo de intervenção em caso de detecção de risco de evasão é manual, subjetivo, empírico e sujeito a falhas, pois depende da experiência acadêmica e da participação dos docentes, considerando que os docentes têm desempenhado diversas atividades, assim como também existe uma grande quantidade de alunos em cursos EaD (MANHÃES *et al*, 2011). Portanto, mecanismos que automatizam ou auxiliam a predição de grupos de estudantes com risco de evasão é importante para diminuir o problema da evasão.
- Os trabalhos publicados no Brasil que abordam a evasão, na sua maioria, são focados em entidades públicas de ensino superior, na modalidade presencial, e ressalta-se a

carência de trabalhos na esfera pública nos níveis de ensino fundamental e técnico (COLPO *et. al*, 2020).

- Encontraram-se na literatura diversos trabalhos que apresentam um método de predição de estudantes em risco de não concluir o curso e, em grande parte, usam técnicas supervisionadas para classificação e verificam a taxa de acerto em turmas já encerradas. Poucos trabalhos abordam esse problema com métodos para analisar o dinamismo de turmas em andamento para uma predição antecipada.

Diante desse panorama, que apresenta uma baixa taxa de formados, faz-se necessário pensar em soluções para aproveitamento ou que evitem uma perda maior dos recursos dessas instituições. Para isso, docentes e gestores precisam de ferramentas automatizadas que auxiliem na tomada de decisões para reduzir a evasão escolar, uma vez que o orçamento que recebem as instituições de ensino está relacionado à quantidade de matrículas ativas que possuem.

1.6 Terminologia

Estabeleceremos algumas definições que abrangem o escopo deste trabalho.

- AVA (Ambiente Virtual de Aprendizagem) é o conceito sobre um sistema computacional de gerenciamento de cursos que propicia o espaço online no qual ocorre a interação entre atores no processo ensino-aprendizagem.
- Moodle: é um software para apoiar a aprendizagem dentro do AVA.
- *Clustering*: técnica de agrupamento em AM e referenciado neste trabalho também como agrupamento.
- EaD/ e-Learning: Educação a distância atrelada às TICs (Tecnologia da Informação e Comunicação), por meio de uma plataforma virtual de aprendizagem.
- Evadido: O estudante que não consegue se formar após três anos, tempo de integralização do curso concomitante, considerando aqueles que desistem ou reprovam. Consideram-se as expressões evasão, desistência, insucesso ou fracasso para referenciar, de um modo geral, a evasão escolar.
- Sucesso: estudante que finalizou o curso, após ser aprovado em todas as disciplinas.
- Turma: é um grupo de estudantes matriculados em um curso que (idealmente) iniciou seus estudos simultaneamente e dos quais espera-se que se formem simultaneamente com cumprimento da carga horária, ou seja, aqueles que iniciaram e finalizaram na mesma turma e desconsiderou-se alunos retidos, que pagam disciplina.

Com respeito às tecnologias adotadas para a execução deste projeto, é oportuno expor algumas ponderações sobre as técnicas escolhidas nesta proposta:

- Consultas e análise exaustiva, por meio da linguagem SQL, das bases de dados do Moodle.
- As técnicas de Mineração de Dados Educacionais (MDE) são utilizadas para análise e extração de conhecimento da base de dados da plataforma Moodle.
- As técnicas de Aprendizado de Máquina (AM) são para predição de evasão, cujo processo requer seleção de atributos relevantes que influenciam na predição, treinamento, testes e validação de dados, comparativos entre os algoritmos que levam a uma análise descritiva e explicativa.
- Para desenvolver a proposta foram utilizadas ferramentas como:
    - StArt para organizar a revisão sistemática da literatura;
    - Postgres para restaurar as cópias dos bancos de dados do Moodle;
    - DBeaver Enterprise para executar as queries criadas para análise por períodos do comportamento do estudante;
    - Rapid Miner para executar os algoritmos de AM para predição, testes, validação, avaliar desempenho dos algoritmos e seleção de características relevantes, comparativos nos resultados dos algoritmos e gerar tabelas e figuras;
    - Planilha eletrônica para integrar os resultados do Rapid Miner com a base Postgres, tanto importação como exportação dos dados, para análises mais detalhadas em ambas ferramentas e gerar tabelas e gráficos.

1.7 Indicação do Método da Pesquisa

Wazlawick (2014, p.19) explica que “O método científico é particularmente importante em computação porque, como ciência, ela não pode se ocupar apenas da coleta de dados. A explicação dos dados é muito mais importante”. Ainda, o autor defende que a metodologia seria o estudo dos métodos e que “[...] apesar do uso corrente, linguisticamente seria mais correto afirmar que um trabalho científico individualmente tem um método de pesquisa e não uma metodologia. ” (WAZLAWICK, 2014, p.64). E neste trabalho adotou-se o termo método de pesquisa.

Esta tese se caracteriza como um estudo descritivo, com fins de pesquisa explicativa, e para atingir os objetivos propostos a abordagem utilizada neste trabalho é tanto quantitativa

quanto qualitativa, tem natureza aplicada[2] e os principais tipos de instrumentos utilizados são fontes bibliográficas e observações, com uso exclusivamente de fontes secundárias.

Para definir o delineamento metodológico, deve-se considerar que os cursos concomitantes ao ensino médio do IFRO, campus Porto Velho Zona Norte (PVZN), objeto desta pesquisa, já finalizaram.

O Fonseca (2002, p.32) define que:

> A pesquisa *ex-post-facto* tem por objetivo investigar possíveis relações de causa e efeito entre um determinado fato identificado pelo pesquisador e um fenômeno que ocorre posteriormente. A principal característica deste tipo de pesquisa é o fato de os dados serem coletados após a ocorrência dos eventos.

Como exemplo desse tipo de pesquisa, pode-se citar uma pesquisa que intenta analisar os indicadores de evasão. Do outro lado, existe um estudo experimental, que pode ser exemplificado com um determinado tratamento em grupo de alunos e depois observar o índice de evasão.

A pesquisa descritiva pretende descrever os fatos e fenômenos de determinada realidade e exige do pesquisador uma série de informações sobre o que deseja pesquisar (TRIVIÑOS, 1987). São exemplos de pesquisa descritiva: estudos de caso, análise documental, pesquisa *ex-post-facto*.

Segundo Gil (2007, p. 43), uma pesquisa explicativa pode ser a continuação de outra descritiva, posto que a identificação de fatores que determinam um fenômeno exige que este esteja suficientemente descrito e detalhado. Para Wazlawick (2014), que analisa metodologias em ciência da computação, a pesquisa explicativa é a mais complexa e completa. É considerada a pesquisa científica por excelência porque, além de analisar os dados observados, busca suas causas e explicações, ou seja, os fatores determinantes desses dados, explicando o porquê das coisas através dos resultados oferecidos.

Quanto à abordagem do problema, trata-se de uma pesquisa quantitativa. Uma vez que pretende-se medir as distâncias entre as instâncias para determinar os grupos, usar dados normalizados e análises estatísticas na classificação. No entanto, envolve também aspectos qualitativos, uma vez que pretende-se uma pesquisa que procura descrever as características dos grupos formados e do fenômeno de evasão.

---

[2] Objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos. Envolve verdades e interesses locais (GERHARDT e SILVEIRA, 2009)

Nos cursos estudados, devido ao grande volume de dados que é gerado, fica inviável o uso de métodos manuais para produzir o conhecimento necessário para atingir os objetivos. Por esse motivo, justifica-se o uso de técnicas de MD.

Segundo Wazlawick (2014), o método da pesquisa permite refazer uma sequência de passos a serem seguidos para chegar ao objetivo principal do trabalho. Dessa forma, levam a validar a hipótese.

Considerando que,

> "O método propriamente dito de um trabalho científico só pode ser estabelecido depois que o objetivo tiver sido definido. Por esse motivo, no caso da computação, normalmente a revisão bibliográfica não deveria nem fazer parte do método.
>
> A revisão bibliográfica consiste em um passo do trabalho no qual o aluno vai iniciar ou aprofundar seus conhecimentos em um campo do saber para que possa então propor um objetivo que seja coerente com o grau que deseja obter. Ou seja, **a etapa de revisão bibliográfica não seria parte do método** (grifo nosso), mas um pré-requisito para a realização do trabalho de pesquisa, pois quem não estudou o assunto não tem como propor um objetivo válido" (WAZLAWICK, 2014, p.65).

As etapas do método de pesquisa são:

(a) Caracterização de um arcabouço conceitual genérico a partir dos resultados da pesquisa teórica que define amostra, técnicas, etapas e processos.
(b) Engenharia de atributos, para pré-processamento, que inclui limpeza, transformação e seleção de atributos que servem como dados de entrada para as técnicas de AM.
(c) Análise descritiva dos grupos de estudantes descobertos pelas características de comportamento no AVA.
(d) Execução dos algoritmos de classificação para predição de evasão e comparação de resultados.
(e) Análises e conclusões dos resultados parciais obtidos.
(f) Ajustes dos dados e das técnicas em AM, repetir as etapas (b) a (e) até obter um método que forneça um instrumento de predição que contemple o comportamento dinâmico do contexto educacional.
(g) Apresentação dos resultados de predição de fracasso na conclusão do curso e elaboração dos gráficos mais adequados que compõem a análise descritiva.

1.8 Principais Contribuições

Após a contextualização dos problemas e estratégias para uma solução, as potenciais contribuições deste trabalho são, a seguir:

- Aplicar a metodologia proposta em outros cursos EaD e observar e descrever o comportamento dos estudantes em cenários reais, com o curso em andamento, quando as informações de registro disponíveis são apenas as geradas até o momento. Os resultados deste trabalho podem ser usados e complementados no âmbito nacional e internacional com cursos em diferentes níveis de ensino.
- Trazer a importância da análise do dinamismo do comportamento do estudante com as descobertas de *insights* com os agrupamentos resultantes das técnicas não supervisionadas e esse conhecimento ser utilizado para classificadores de risco de evasão do estudante, usando técnicas supervisionadas. Geralmente, as pesquisas focam só as técnicas supervisionadas, que encontram padrões nos modelos treinados de turmas encerradas para propor modelo de predição de evasão em outras turmas, e neste trabalho é precedido por uma etapa para interpretar as características de comportamento dos grupos formados.
- Em termos de publicações geradas pela tese, foram produzidos 3 artigos internacionais, com 2 artigos em conferências da IEEE e um artigo publicado em periódico (*journal*) classificado no estrato superior *Qualis*.

1.9 Organização da Tese

Assim, a partir das explicações sobre o problema, dos objetivos, da justificativa e da indicação metodológica em função dos estudos desenvolvidos, o trabalho é apresentado em 6 capítulos da seguinte forma:

- Capítulo 2 traz um referencial teórico que detalha os principais tópicos do documento: Ensino Profissional Técnico no Brasil, Educação a Distância, Evasão, Mineração de Dados Educacional e Aprendizado de Máquina, mencionando as principais técnicas e os trabalhos correlatos ao tema.
- Capítulo 3 discutem-se os trabalhos relacionados referentes ao problema da evasão escolar utilizando MD e AM.
- Capítulo 4 apresenta a metodologia proposta com a descrição dos passos na base de dados, a complexa tarefa de engenharia de atributos, detalha as técnicas de agrupamento e posterior predição de evasão, e as métricas analisadas.
- Capítulo 5 é a apresentação dos experimentos em duas etapas: *clustering* e classificação, com análise e discussão dos resultados, com as descobertas em cada etapa.
- Por fim, o Capítulo 6 apresenta as considerações finais desta tese, limitações, dificuldades no percurso da pesquisa e sugestões de trabalhos futuros.

Após as Referências, os apêndices e anexos trazem a matriz curricular dos 3 cursos da amostra, gráficos complementares, os artigos publicados e parte relevante das consultas elaboradas para extração de dados do Moodle.

Capítulo 2

Referencial Teórico

Neste capítulo são apresentados os conceitos prévios necessários para o entendimento e desenvolvimento do trabalho. Os fundamentos teóricos são divididos em 5 seções, em que são estabelecidas algumas convenções e definições para os termos a serem usados durante todo o texto deste documento, com a finalidade de orientar o entendimento da pesquisa. Esta pesquisa utiliza 5 termos principais: Ensino Profissional Técnico (EPT) no Brasil, Educação a distância (EaD), Evasão, Mineração de Dados Educacionais (MDE) e Aprendizado de Máquina (AM), com uma seção dedicada amplamente para cada termo e assuntos correlatos.

2.1 Ensino Profissional Técnico (EPT)

De acordo com o portal do MEC (Ministério da Educação) [3], a educação profissional técnica de nível médio inclui desde as qualificações profissionais técnicas de nível médio (EPTNM), como saídas intermediárias, até a correspondente habilitação profissional do técnico de nível médio.

A Resolução do Conselho Nacional de Educação CNE/CP Nº 1, de 5 de janeiro de 2021 (BRASIL, 2021), no seu Capítulo VI - Da Estrutura E Organização Da Educação Profissional Técnica De Nível Médio define que:

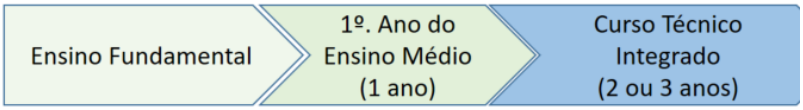
> Art. 16. Os cursos técnicos serão desenvolvidos nas formas integrada, concomitante ou subsequente ao Ensino Médio, assim caracterizadas:
>
> I - integrada, ofertada somente a quem já tenha concluído o Ensino Fundamental, com matrícula única na mesma instituição, de modo a conduzir o estudante à habilitação profissional técnica ao mesmo tempo em que conclui a última etapa da Educação Básica;
>
> II - concomitante, ofertada a quem ingressa no Ensino Médio ou já o esteja cursando, efetuando-se matrículas distintas para cada curso, aproveitando oportunidades educacionais disponíveis, seja em unidades de ensino da mesma instituição ou em distintas instituições e redes de ensino;
>
> III - concomitante intercomplementar, desenvolvida simultaneamente em distintas instituições ou redes de ensino, mas integrada no conteúdo, mediante a ação de convênio ou acordo de intercomplementaridade, para a execução de projeto pedagógico unificado; e
>
> IV - subsequente, desenvolvida em cursos destinados exclusivamente a quem já tenha concluído o Ensino Médio.

[3] http://portal.mec.gov.br/cursos-da-ept/cursos-da-educacao-profissional-tecnica-de-nivel-medio

Ou seja, no Brasil, *o ensino técnico é um nível de ensino enquadrado no nível médio dos sistemas educativos* e pode ser ofertado como:

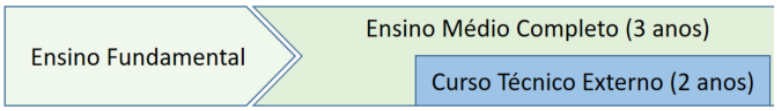
1) Curso Técnico Integrado: substitui parcialmente o ensino médio e pode ser iniciado logo após o estudante concluir o primeiro ano do ensino médio (Figura 1). Ao concluir um curso técnico integrado, o estudante recebe dois certificados: o de conclusão do ensino médio e o de conclusão do curso técnico escolhido.

Figura 1- Curso Técnico Integrado



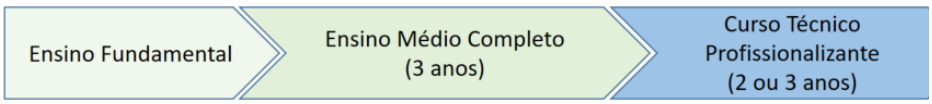
2) Curso Técnico Externo ou Concomitante: acontece em paralelo ao ensino médio (Figura 2) e exige que o estudante tenha concluído o primeiro ano do ensino médio. No caso deste tipo de formação o estudante pode cursar no contraturno. Após concluir o ensino técnico externo, o aluno recebe o certificado, se tiver concluído o ensino médio regular.

Figura 2- Curso Técnico Externo ou Concomitante.



3) Curso Técnico Profissionalizante ou Subsequente: é uma opção para quem já concluiu o ensino médio e quer realizar um curso técnico (Figura 3).

Figura 3 - Curso Técnico Profissionalizante ou Subsequente.



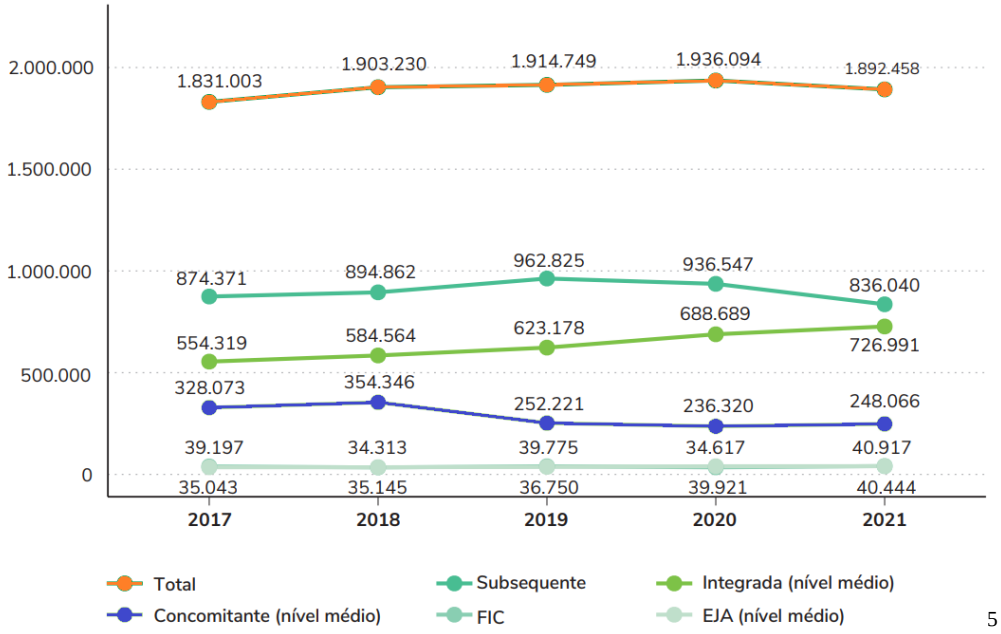
Os cursos técnicos podem ser estruturados com carga horária variando entre 800, 1.000 e 1.200 horas, dependendo da respectiva habilitação profissional técnica e destinam-se a pessoas que tenham concluído o ensino fundamental, estejam cursando ou tenham concluído o ensino médio. É importante ressaltar que para a obtenção do diploma de técnico é necessário a conclusão do ensino médio (portal MEC[4]).

[4] http://portal.mec.gov.br/cursos-da-ept/cursos-da-educacao-profissional-tecnica-de-nivel-medio

Para entender melhor a evolução da Educação no Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) publica anualmente o Censo Escolar em todos os níveis de ensino.

Segundo o último relatório do INEP, de 2021, foram registradas 7,8 milhões de matrículas no ensino médio desse ano, o que significa um aumento de 2,9% em relação a 2020. Por outro lado, o número de matrículas da educação profissional divergiu dessa tendência de crescimento e, na Figura 4, observa-se que o total de matrículas de 2021, em relação ano anterior, teve uma redução de 2,3%. Já as modalidades ofertadas no ensino médio (integrado e concomitante) tiveram alta, com a diferença que a modalidade integrada está em alta nos últimos 5 anos, enquanto a modalidade concomitante teve uma queda de 43%, se comparado a 2018.

Figura 4 – Número de Matrículas na Educação Profissional (2017-2021).



Fonte: adaptado de INEP (2021).

O relatório do INEP destaca que houve uma melhora significativa da taxa de aprovação no período de 2020 a 2021, influenciada pela adoção de ajustes no planejamento curricular das escolas diante da pandemia de Covid-19 e alinhada às recomendações do Conselho Nacional de Educação (CNE).

[5] FIC (Formação Inicial e Continuada); EJA (Educação de Jovens e Adultos).

2.1.1 Importância do ensino técnico

O ensino técnico foi promovido pelo Governo Federal por meio do Pronatec (Programa Nacional de Acesso ao Ensino Técnico e Emprego), entre 2011 e 2014, por meio de programas, projetos e ações de assistência técnica e financeira. Essa iniciativa incluiu a "expansão física das redes públicas federal, distrital e estaduais, com a construção e ampliação de escolas de educação profissional em todo o País, redução da capacidade ociosa das instituições, ampliação da oferta de educação profissional a distância e oferta de Bolsa Formação Estudante e Trabalhador" (MEC, 2015, p.36).

Assim, em pesquisas recentes, Rego *et al*. (2021) destacam que os estudantes têm como alternativa os cursos EPT para acessar ao mercado de trabalho e confirmam a importância dessa modalidade educacional no atual cenário brasileiro e frente às exigências tecnológicas, seja na melhor qualificação ou na requalificação para uma reinserção profissional no mercado de trabalho, baseados nos 76,2% dos entrevistados na PNAD, da sua última pesquisa em 2014, que aponta que o conteúdo aprendido, junto com o diploma, foram os principais fatores para obtenção de emprego na área de formação.

Nesse contexto de avanços do EPT, o Decreto n° 9.570, de 22 de novembro de 2018 (BRASIL, 2018), para regulamentar a Lei do Aprendiz, a legislação brasileira estabelece a obrigatoriedade para que empresas de qualquer natureza empreguem ou matriculem no mínimo 5% de aprendizes jovens de 14 a 24 anos em cada estabelecimento, cujas funções demandem formação profissional. Isso permite abrir oportunidades concretas de inserção no mercado de trabalho para quem tem habilitação profissional de técnico de nível médio.

O relatório da OCDE (Organização para a Cooperação e o Desenvolvimento Econômico) de 2021, com o tema "Perspectivas da política educacional no Brasil", apresenta uma análise detalhada do desempenho do sistema educacional brasileiro em relação a países comparativamente relevantes, e aponta Letônia, Canadá e Reino Unido, como experiências internacionais para resolver problemas de evasão escolar em diferentes níveis educacionais, além de relacionar passos para que o País melhore a qualidade e a equidade dos resultados educacionais.

Esse relatório da OCDE aponta que um dos principais desafios no Brasil é aumentar as matrículas de qualidade em toda a oferta de EPT (OCDE, 2021, p.16):

> "[...] A educação profissional e técnica (EPT) pode facilitar a entrada no mercado de trabalho, mas muitos países da OCDE ainda não possuem programas de formação no local de trabalho suficientemente desenvolvidos. [...]. No entanto, a taxa de matrícula

> nessa modalidade educacional no Brasil é baixa**: apenas 11% dos alunos do ensino médio em 2018 participavam de cursos técnicos, bem abaixo da média da OCDE de 42% e da meta do PNE de 25% até 2024** (grifo nosso). As taxas de conclusão também são baixas: em 2018, apenas 58% dos alunos da educação profissional e técnica do ensino médio tinham se formado 2 anos após a duração teórica. As reformas atuais visam o modelo integrado, que registra taxas de evasão mais baixas e tem como objetivo aumentar as matrículas".

É importante destacar desse parágrafo que no Brasil somente 11% dos alunos do ensino médio optam por este modelo de formação, muito abaixo dos 42% da média da OCDE e discrepante se comparado a países como Alemanha (45%), Reino Unido (63%), Coreia do Sul (65%) ou Chile (31%), do OCDE 2017.

Pode ser que esse cenário melhore com várias ações e uma delas seja o Novo Ensino Médio, iniciado em 2022, com proposta de um novo modelo de aprendizagem por áreas de conhecimento, que permite ao estudante optar por uma formação técnica e profissionalizante com, no mínimo, 1.200 horas reservadas para a Formação Técnica e Profissional.

2.2 Educação a Distância (EaD) no Brasil

É comum ver o termo EaD e ficar na dúvida se a leitura seria 'Ensino a Distância' ou 'Educação a Distância'. Moran (2002, p.1) explica que "...na expressão 'ensino a distância' a ênfase é dada ao papel do professor (como alguém que ensina a distância). Preferimos a palavra 'educação' que é mais abrangente, embora nenhuma das expressões seja perfeitamente adequada". Dessa forma, neste trabalho adotou-se a expressão 'Educação a Distância' para a abreviatura EaD.

Segundo Moran (2009), a educação se apresenta em três formatos, sendo elas a presencial, semipresencial e educação a distância. A educação presencial ocorre em cursos regulares, na qual professores e estudantes são alocados em uma sala de aula, geralmente no mesmo espaço físico e no mesmo horário. No semipresencial as aulas são divididas em uma parte presencial e a outra parte a distância. Nos momentos das aulas a distância, os estudantes têm encontros virtuais e interagem por meio das TICs.

Atualmente, a modalidade semipresencial recebe diferentes terminologias: ensino híbrido, educação bimodal, *b-learning ou blended learning*. Segundo Bacich *et al*. (2015, p.3) "o ensino híbrido é uma abordagem pedagógica que combina atividades presenciais e atividades realizadas por meio das tecnologias digitais de informação e comunicação (TDICs). Existem diferentes propostas de como combinar essas atividades, porém, na essência, a estratégia consiste em colocar o foco do processo de aprendizagem no aluno" e o aluno estuda o material em diferentes situações e ambientes, para resolver projetos ou problemas propostos, realizando

atividades, discussões ou laboratórios, entre outros, com o apoio do professor e colaboração dos colegas.

No caso da modalidade a distância, são utilizados termos como *e-learning* ou *electronic learning* ou educação *online*. Moran (2009) define a EaD como um processo de ensino-aprendizagem, mediado por tecnologias, no qual professores e alunos estão separados espacial e/ou temporariamente. Apesar de não estarem juntos, de maneira presencial, eles podem estar conectados, interligados por tecnologias, principalmente as telemáticas, como a Internet.

Pelas limitações temporais ao processo de aprendizagem, a EaD pode ser classificada em duas categorias: síncrona e assíncrona (MANDALA, 2013). Na forma síncrona os estudantes participam de uma interação ao vivo com o professor ou tutor por meio de videoconferência, bate-papo e outras ferramentas de comunicação disponíveis no AVA, que permite aos usuários desenvolver comunidades de aprendizagem *online*. Na forma assíncrona, os alunos podem acessar o AVA para baixar o conteúdo ou enviar mensagens para tutores ou colegas a qualquer momento, e a interação *off-line* também permite apoiar a colaboração entre os participantes.

Assim, a EaD permite a definição de uma metodologia pedagógica para um modelo de aprendizado baseado em tecnologia e a utilização de recursos de áudio, de vídeo e de ferramentas de interatividade por meio de uma plataforma virtual de aprendizagem.

Para o Ministério da Educação e Cultura no Brasil (MEC), no seu art. 1.º do Decreto 9.057, de 25 de maio de 2017 (BRASIL, 2017), considera-se:

> "[...] educação a distância a modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorra com a utilização de meios e tecnologias de informação e comunicação, com pessoal qualificado, com políticas de acesso, com acompanhamento e avaliação compatíveis, entre outros, e desenvolva atividades educativas por estudantes e profissionais da educação que estejam em lugares e tempos diversos"

Para conceituar EaD em cursos profissionalizantes a Resolução CNE/CP Nº 1, de 5 de janeiro de 2021 (BRASIL, 2021), que define as Diretrizes Curriculares Nacionais Gerais para a Educação Profissional e Tecnológica, considera:

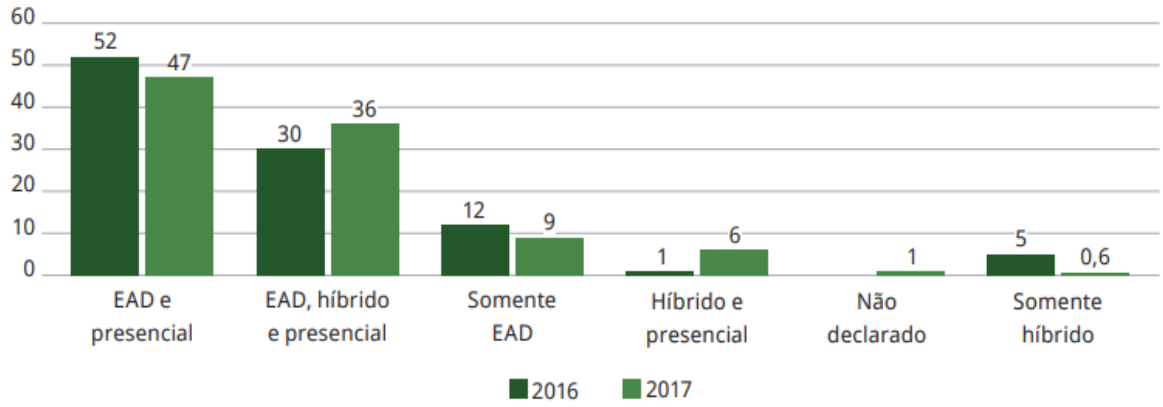
> Art. 40. A modalidade EaD é aqui entendida como uma forma de desenvolvimento do processo de ensino-aprendizagem que permite a atuação direta do docente e do estudante em ambientes físicos diferentes, em consonância com o disposto no art. 80 da Lei nº 9.394/1996 e sua regulamentação.
>
> Art. 41. A oferta de cursos de qualificação profissional na modalidade a distância deve observar as condições necessárias para o desenvolvimento das competências requeridas pelo respectivo perfil profissional, resguardada a indissociabilidade entre teoria e prática.

Segundo Bienkowski *et al.* (2012) no modelo tradicional todos os estudantes vão ouvir as mesmas aulas e completar as mesmas tarefas na mesma sequência e ao mesmo ritmo. E no modelo EaD promove-se um processo de aprendizagem mais personalizado, em que o estudante tem um papel ativo. Em qualquer forma de ensino, o papel do instrutor é conceber, organizar e apoiar experiências de aprendizagem.

No entanto, para Liñán e Pérez (2015), os cursos *e-learning*, ou a distância, também apresentam taxas de abandono mais elevados que o presencial devido ao fato de que a educação a distância pode criar uma sensação de isolamento dos estudantes, que podem se sentir desconectados de outros estudantes, dos professores, dos tutores e da instituição de ensino.

Segundo o relatório da ABED (2018) referente ao censo em 2017, em geral, as instituições oferecem mais de um tipo de curso, sendo a combinação de EAD e presencial a mais frequente, com 47% das instituições que oferecem essa modalidade associada, seguida de 36% de instituições com cursos a distância, híbridos e presenciais (Figura 5). De acordo com (ABED, 2018, p.52), houve "um aumento no percentual de instituições que oferecem diferentes modalidades e uma redução nas que oferecem somente uma modalidade, o que revela que a diversidade de oferta parece ser a tendência atual".

Figura 5 - Evolução do número de cursos oferecidos por uma mesma instituição, por tipo de curso, em percentual.



Fonte: ABED (2018, p.52).

Os números da Figura 5 indicam uma redução de instituições que ofertam somente EaD e um crescimento de instituições que ofertam híbrido e presencial, ou estes dois com a EaD. O aumento do ensino híbrido, em parte, é pela revogação do Decreto nº 5.622, de 19 de dezembro de 2005, substituído pelo Decreto nº9.057, de 25 de maio de 2017 (BRASIL, 2017), e a Resolução n° 03, de 21 de novembro de 2018 (BRASIL, 2018), que, segundo Da Silva *et al.*

(2020), essas diretrizes trouxeram mudanças significativas na organização didático-pedagógica e de gestão da EaD na educação básica e profissional.

Em referência ao ensino médio, incluindo a Educação Técnica Integrada e/ou Concomitante ao Ensino Médio, a Resolução 03/2018 trouxe a novidade da regulamentação sobre a carga horária semipresencial nos cursos de nível médio, onde no art. 17, § 15 menciona que as atividades realizadas a distância podem contemplar até 20% da carga horária total, necessariamente com acompanhamento docente da unidade escolar onde o estudante está matriculado, e na educação de jovens e adultos (EJA).

Por exemplo, no art. 17, § 5°, na modalidade de educação de jovens e adultos é possível oferecer até 80% (oitenta por cento) da carga horária total dos cursos a distância, tanto na formação geral básica, quanto nos itinerários formativos do currículo, incluindo a educação técnica e profissional, desde que haja suporte tecnológico e pedagógico apropriado (DA SILVA *et al.,* 2020). Por outro lado, permanece vigente a determinação do art. 33 da resolução CNE/CEB n° 06, de 20 de setembro de 2012 (BRASIL, 2012), segundo o qual as instituições são obrigadas a adotar um mínimo de 20% de carga horária presencial no âmbito dos cursos técnicos na modalidade a distância, com exceção da área da saúde, que deve cumprir 50%, no mínimo.

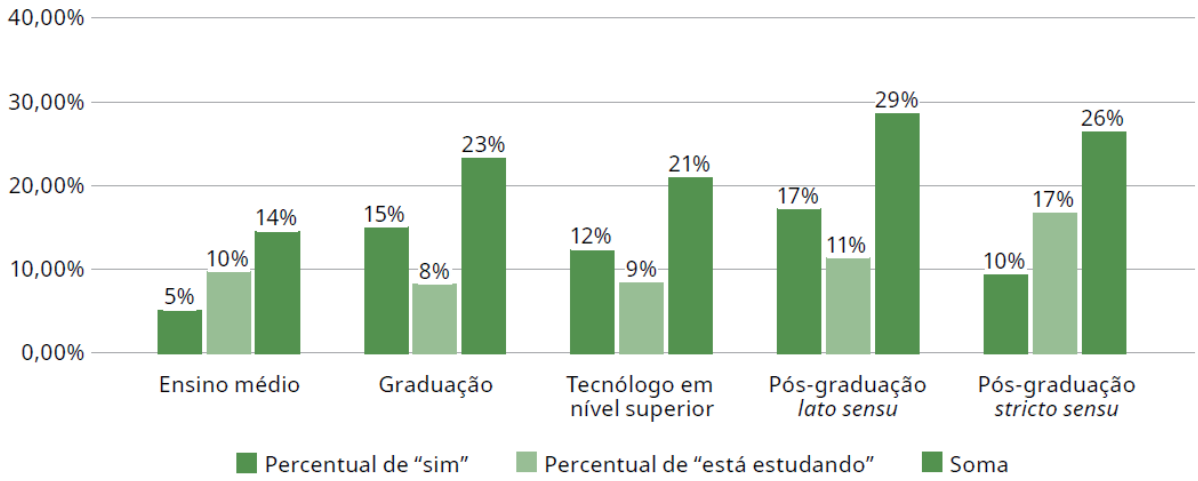
Com essa liberação de EaD no ensino médio e, também, na pós-graduação stricto sensu a distância, surgiu uma nova questão no Censo EAD.BR (2019-2020) relacionado a quais níveis de ensino as instituições pretendiam se expandir na modalidade EaD. Participaram desse censo 357 instituições (208 públicas[6] e 149 privadas, com 11 institutos federais), que, em sua maioria, são instituições de nível superior, e responderam à questão: “se ainda não oferece um curso neste nível, pretende oferecer?” (ABED, 2020, p.42), referindo-se a cada nível de ensino que a legislação permite ofertar na modalidade a distância.

A Figura 6 mostra que a prioridade das instituições de ensino está na pós-graduação lato sensu, com 17,2% do IES para começar nessa modalidade e 11,5% pensando no assunto, o que revela uma preferência por estudantes formados e com suposta renda própria para uma modalidade on-line de ensino. Isso confirma que as IES focam o público que conhece cursos on-line, e “a abertura do mercado de EAD no ensino médio, que terá demandas mais urgentes

[6] Onze IFs (Institutos Federais), correspondentes a 11 estados, fazem parte das 208 instituições que participaram da pesquisa. O IFRO não participou.

a partir de 2021, provavelmente ficará a cargo de outros *players* do mercado educacional" (ABED, 2020 p.43). É importante ressaltar que os dados coletados para elaboração desse último relatório são anteriores à pandemia de Covid-19.

Figura 6 - Em quais níveis de ensino as IES pretendem começar a oferecer cursos EAD.



Fonte: ABED (2020, p.42).

Neste trabalho os cursos a distância acontecem em Ambiente Virtual de Aprendizagem (AVA) na plataforma do Moodle.

2.2.1 Plataforma Virtual de Aprendizagem Moodle

O MOODLE[7] (*Modular Object-Oriented Dynamic Learning Environment*) é um software livre de apoio à aprendizagem executado num ambiente virtual, criado em 1999 pelo educador e cientista computacional Martin Dougiamas (ALVES *et al*, 2009), sob a forma de comunidade virtual. O Moodle é o nome do software de um AVA utilizado no mundo inteiro e, também, representa uma comunidade global de educadores, desenvolvedores e outros relacionados na construção da plataforma para aprendizado online.

Esse conceito envolve uma filosofia educacional baseado na "pedagogia social construcionista"[8], tema que considera conceitos principais relacionados como Construtivismo e Construcionismo. O Construtivismo é umas das teorias mais importantes da educação, baseado nos conceitos de Jean Piaget (1976), e pressupõe que a cognição se desenvolve e constrói seu próprio conhecimento a partir da interação com o meio físico e social. Já o

[7] http://moodle.org

[8] https://docs.moodle.org/32/en/Philosophy#Social_constructivism

Construcionismo é baseado nos estudos de Saymour Papert (1997), que propõe a construção do conhecimento por meio do uso de tecnologias. Essa teoria se concentra no modo de aprendizagem e afirma que a aprendizagem é particularmente eficaz quando se constrói algo que os outros possam experimentar, como uma frase falada, uma postagem na Internet ou desenvolvimento de software, ao invés de ser transmitido sem mudanças.

Ou seja, o AVA não é meramente um repositório de materiais e recursos prontos, mas um ambiente que promove a interação e auxilia na construção do conhecimento com base nas habilidades, capacidades e conhecimentos próprios do estudante, que vai depender da proposta pedagógica do curso e do contexto educacional. Para auxiliar na construção do aprendizado, o Moodle proporciona ferramentas de interação com e entre os participantes do curso, como *wikis*, e-livros, fórum de discussão, *chat*, glossários, banco de dados etc., mas depende do método de ensino aplicado para conseguir o aprendizado como pretendido na concepção do AVA.

O Moodle é disponibilizado livremente na forma de software livre (sob a licença de software livre *GNU Public License*) e pode ser instalado em diversos ambientes (Unix, Linux, Windows, Mac OS) desde que os mesmos consigam ter um servidor web com suporte à linguagem PHP, como Apache e IIS. Assim, as instituições de ensino podem adaptar, estender ou personalizar o Moodle. Também aceita plugins para adicionar funções e outros programas de terceiros.

2.2.2 EaD no IFRO

No crescimento de ofertas nas diversas modalidades educacionais, a Educação Profissional Tecnológica (EPT) também ganha espaço de relevância, sobretudo a partir da criação dos Institutos Federais de Educação, Ciência e Tecnologia, criado através da Lei Nº. 11.892 de 29 de dezembro de 2008 (BRASIL, 2008). Essa lei reorganizou a Rede Federal de Educação Profissional, Científica e Tecnológica composta pelas Escolas Técnicas, Agrotécnicas e CEFET's (Centro Federal de Educação Tecnológica), transformando-os em Institutos Federais de Educação, Ciência e Tecnologia.

Os Institutos Federais[9] são instituições pluricurriculares e multicampi (reitoria, campus, campus avançado, polos de inovação e polos de educação a distância), especializados na oferta de educação profissional e tecnológica (EPT) em todos os seus níveis e formas de articulação

[9] http://portal.mec.gov.br/rede-federal-inicial/instituicoes

com os demais níveis e modalidades da Educação Nacional, oferta os diferentes tipos de cursos de EPT, além de licenciaturas, bacharelados e pós-graduação stricto sensu.

Instituídos no momento de constituição da Rede Federal, os institutos têm como obrigatoriedade legal garantir um mínimo de 50% de suas vagas para a oferta de cursos técnicos de nível médio, prioritariamente na forma integrada. Devem, ainda, garantir o mínimo de 20% de suas vagas para atender a oferta de cursos de licenciatura, bem como programas especiais de formação pedagógica, com vistas à formação de professores para a educação básica, sobretudo nas áreas de ciências e matemática, e para a educação profissional.

Destaca-se também sua atribuição no desenvolvimento de soluções técnicas e tecnológicas por meio de pesquisas aplicadas e as ações de extensão junto à comunidade com vistas ao avanço econômico e social local e regional.

Desse modo, programas para a Rede Federal de Educação Profissional buscam na EaD uma forma de ampliar o alcance à formação pessoal. Isso não apenas no que se refere à quantidade de formados, mas também alcance da população em áreas rurais, ribeirinhas, urbanas etc.

> "A EaD nos Institutos Federais pode ocorrer de formas diferentes em cada instituto, pois desde o início foi dada a eles a **opção de escolher o modelo de gestão da modalidade a distância (grifo nosso)**. Essa escolha foi orientada pelo documento que regulamenta a EaD nos Institutos Federais, elaborado pelo Conselho Nacional dos Institutos Federais (CONIF). Com base nas orientações previstas nesse documento, cada Instituto Federal reconheceu a EaD e sua gestão de forma diferente, sendo possível observar a existência ou não do processo de institucionalização, uma vez que boa parte da oferta de EaD é fomentada por programas que contam com o recebimento de bolsas para quem trabalha. Essas bolsas exigem que o trabalho na EaD ocorra fora do horário das demais atividades existentes no Instituto" (MEDEIROS, 2019, p.2).

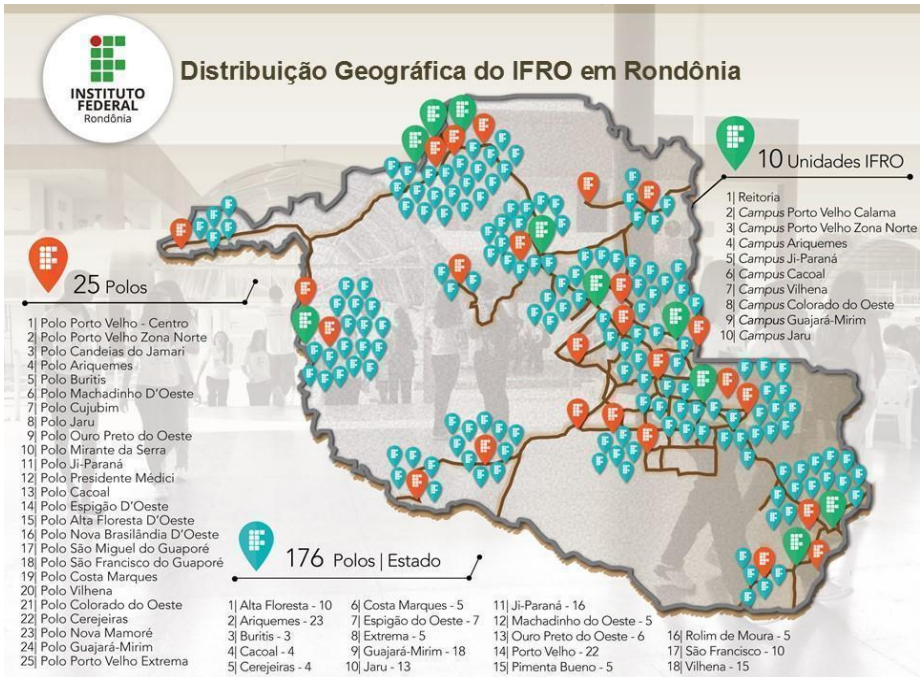
No total, em 2019, considerando todos os campi de instituições públicas vinculadas à administração federal do Brasil, os Institutos Federais abrangem 661 unidades educacionais, presente nos 26 estados mais o Distrito Federal, vinculados a 38 institutos federais, 2 Centros Federais de Educação Tecnológica (CEFET), a Universidade Tecnológica Federal do Paraná (UTFPR), a 22 escolas técnicas vinculadas às universidades federais e ao Colégio Pedro II.

O Estado de Rondônia está representado pelo Instituto Federal de Rondônia[10] (IFRO), que por meio de programas próprios e do Ministério da Educação, a exemplo da Rede e-Tec Brasil e Pronatec, que fomentam a expansão da educação profissional e tecnológica, tem expandido a oferta de cursos técnicos em todas as regiões do Estado.

[10] https://portal.ifro.edu.br/sobre-o-ifro

Territorialmente, o IFRO está presente em 33 dos 52 municípios de RO com 10 *campi* presenciais (Ariquemes, Cacoal, Colorado do Oeste, Guajará-Mirim, Jaru, Ji-Paraná, Porto Velho Calama, Porto Velho Zona Norte, São Miguel do Guaporé e Vilhena). Na EaD, além dos 25 polos[11] já atendidos, em 2016 o IFRO formalizou um termo de cooperação com o Governo do Estado para o atendimento de mais 176 polos. O governo do Estado e as prefeituras são parceiros diretos no apoio à infraestrutura dos polos nos municípios onde o Instituto ainda não possui campus instalado (Figura 7).

Figura 7 - Distribuição geográfica dos polos EaD e campi em RO.



Fonte: Portal do IFRO[12].

Dentre os *campi* presenciais, o campus PVZN, criado em 2010, está localizado na capital do estado e local desta pesquisa, oferta atualmente: superior de tecnologia (presencial), técnico subsequente (presencial e EaD), técnico concomitante (EaD), licenciatura (EaD) e pós-graduação (EaD).

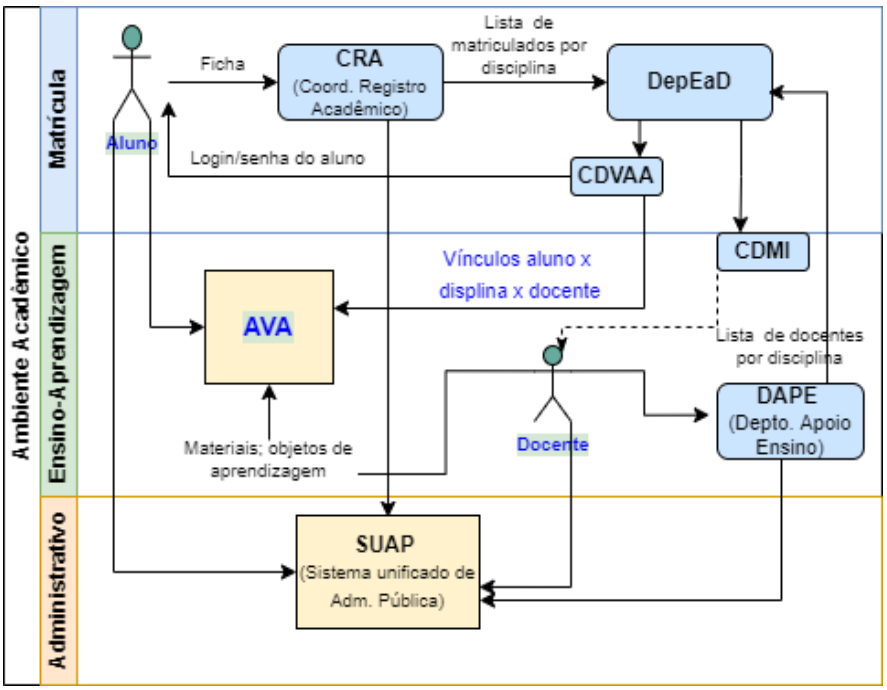
Todos os estudantes matriculados recebem *login* e senha com acesso restrito ao seu curso no Moodle. Para fazer o vínculo dos estudantes a cada disciplina de cada curso a Coordenação de Registros Acadêmicos (CRA), que recebe as matrículas, envia a lista ao

[11] Local externo ao instituto que oferece estrutura física de salas de aula, e pode contar com laboratórios, biblioteca e secretaria.

[12] Disponível em <https://portal.ifro.edu.br/images/imagens_menu/O_INSTITUTO/IFRO-em-RO-2016.jpg> Accesso em ago.2020

Departamento de Produção EaD (DepEaD), e a Coordenação de Design Virtual e Ambientes de Aprendizagem (CDVAA) faz o vínculo e configurações de acesso dos estudantes ao Moodle e, também, vincula os docentes em cada disciplina (Figura 8).

Figura 8 – Fluxograma para acesso ao Moodle.



Para ter uma dimensão do público atendido no campus PVZN, passaram pela CRA, em 2021: 57 certificados Encceja (Exame Nacional para Certificação de Competências de Jovens e Adultos), 276 certificados FIC (Formação Inicial e Continuada), 3.093 matrículas de alunos, 1.580 certificados e diplomas emitidos, 162 cancelamentos de matrículas e 1.715 alunos evadidos[13].

Existem produções de cursos EaD em outros estados com aulas pré-gravadas, entretanto o campus PVZN é um dos 3 campi dos IFs no Brasil que contam com estúdio de gravação com diversos equipamentos de alta tecnologia (Figura 9a), profissionais especializados em produção de conteúdo, formação de professores para enfrentar às câmeras, transmissão, gravação e armazenagem de audiovisuais produzidos no estúdio, contratações de professores e tutores bolsistas para disciplinas específicas e outros recursos para o preparo e as transmissões das aulas online.

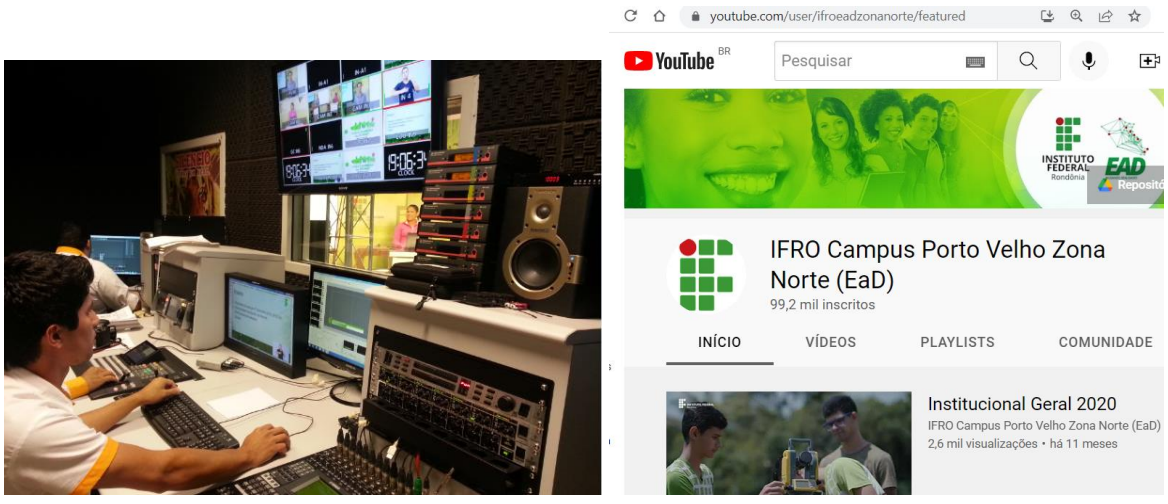
[13] Considera-se evadido o discente que não renovar a matrícula dentro do prazo semestral após 2 períodos consecutivos.

Essa moderna infraestrutura para uma transmissão via satélite, com interações ao vivo entre estudantes e docentes, que permitem chegar a vários polos, mesmo onde não há cobertura ou rede fixa de Internet, também tem recebido o nome de ensino telepresencial ou presencial conectado. Com a pandemia do Covid-19 as atividades com uso de satélites foram suspensas pelo alto custo fixo de manutenção

As aulas ficam salvas, editadas e organizadas em repositório, que o campus armazena em *datacenter* e disponibiliza por meio de um canal no YouTube[14] (Figura 9b). Atualmente, esse canal conta com mais de 100 mil inscritos e mais de 5.000 videoaulas disponibilizadas gratuitamente, gravadas durante as aulas EaD online, realizadas desde 2013. Essas mesmas aulas estão no Moodle, organizadas por tópicos em cada disciplina e turma. Nele também foi disponibilizado um menu para acesso a todo o material didático por curso elaborado pela Rede e-Tec, que constitui uma das ações do Pronatec.

A Figura 10 mostra uma gravação no estúdio da disciplina Programação Orientada a Objeto do curso técnico subsequente de Informática para Internet. Esse tipo de aula conta com professor formador, professor assistente e um intérprete de libras (canto inferior direito da tela), que fica em outra sala do estúdio. Cada polo tem um tutor bolsista para auxiliar os alunos, mas não é necessário ter formação na área do curso. O professor assistente recebe as dúvidas dos estudantes enviadas pelos tutores dos seus respectivos polos, seleciona as dúvidas e comenta as respostas com o professor formador. Portanto, são aulas que têm interações ao vivo com o público estudantil.

Figura 9 - Aula EaD: (a) Centro de controle, gravação e edição das aulas;
(b) Repositório das videoaulas no YouTube



[14] https://www.youtube.com/user/ifroeadzonanorte

Figura 10 - Aula online EaD no estúdio do campus IFRO, Porto Velho Zona Norte.

2.2.3 Curso Técnico Concomitante ao Ensino Médio pode ser EaD ?

Sim, este trabalho analisou 3 cursos técnicos concomitantes ao ensino médio na modalidade EaD do campus PVZN do IFRO, porém não foram ofertados no formato telepresencial, como visto na seção anterior.

Considerando o art. 33 da resolução CNE/CEB n° 06/2012, no qual as instituições são obrigadas a adotar um mínimo de 20% de carga horária presencial no âmbito dos cursos técnicos na modalidade a distância, o Projeto Pedagógico do Curso (PPC) contempla esses cursos EaD com encontros semanais presenciais nos polos e a forma de interagir no Moodle é igual ao formato totalmente EaD, com envio de roteiro de aprendizagem, prazos, materiais padronizados, verificações pelo DepEaD e pelo Departamento de Apoio ao Ensino (DAPE). No caso dos cursos analisados, os alunos do ensino médio regular precisam ter o encontro presencial no campus e direto com o docente. Para isso, é necessário disponibilizar uma sala de aula física e/ou laboratório, o que limita a 40 vagas.

As matrizes curriculares dos 3 cursos técnicos concomitantes analisados neste trabalho encontram-se no ANEXO A.

Algumas informações para contextualizar esse tipo de cursos são:

a) Processo seletivo de estudantes. Estar matriculado no 1º ou 2º do ensino médio em escola regular. Nesse nível, o estudante cursa as disciplinas de ensino médio em sua escola de origem e cursa apenas as disciplinas do curso técnico no IFRO.
b) Critério de seleção: classificação do candidato pelo seu desempenho (notas/conceitos) nas disciplinas de Língua Portuguesa, Matemática, Ciências, História e Geografia do ensino fundamental (7º ao 9ª ano), abrange a conclusão via ensino fundamental

regular, provão ou Encceja (Exame Nacional para Certificação de Competências de Jovens e Adultos).

c) Integralização do curso:

> RESOLUÇÃO Nº 88/CONSUP/IFRO/2016, de 26 de dezembro de 2016 (Rondônia, 2016), no seu Art. 36. O prazo máximo de integralização de matrizes curriculares, incluindo--se todos os componentes curriculares, como as disciplinas, estágios, trabalhos de conclusão de curso e atividades complementares, se houver, limita-se ao dobro do tempo mínimo estabelecido nos projetos pedagógicos correspondentes.

No caso dos cursos concomitantes, cuja duração é de 2 anos, o prazo máximo de integralização é mais 1 ano, ou seja, o estudante deve concluir com sucesso o curso em até 3 anos.

d) Critério de aprovação:

> RESOLUÇÃO Nº 88/CONSUP/IFRO/2016, de 26 de dezembro de 2016 (Rondônia, 2016). CAPÍTULO III. DAS CONDIÇÕES DE PROMOÇÃO
> Art. 94. Para ser considerado promovido, o estudante deve atingir pelo menos 60 pontos por disciplina na média por período ou 50 pontos após exame final, e cumprir a frequência mínima estabelecida em Lei.
>
> Art. 95. Ao longo do período letivo, o estudante que apresentar dificuldades e resultados que o impeçam de atingir a nota mínima estabelecida terá o direito de participar de estudos de recuperação.
>
> Art. 96. O estudante que, após estudos de recuperação, não obtiver média suficiente para sua promoção em cada componente e período, terá direito a realizar exame final.

No caso do curso concomitante, a taxa de frequência mínima é de 75%.

2.3 Evasão Escolar

De acordo com o Ministério da Educação, são identificados como “evadidos do curso os alunos que não se diplomaram no tempo máximo de integralização curricular e que não estão mais vinculados ao curso em questão” (BRASIL, 1997, p. 22). Favero (2006) inclui no conceito de evasão aqueles que nunca estiveram ou se manifestaram no decorrer do curso para seus professores, tutores e colegas. Para Santos et al. (2008, p.2), a evasão é “[...] a desistência definitiva do estudante em qualquer etapa do curso e a mesma pode ser considerada como um fator frequente em cursos a distância”.

Márquez-Vera *et al.* (2016) sugerem que a identificação precoce dos estudantes vulneráveis que são propensos a deixar os seus cursos é uma estratégia fundamental para o sucesso na intervenção escolar. É necessário detectar o mais cedo possível os estudantes que estão em risco e, assim, intervir precocemente para facilitar a permanência dos estudantes no sistema e, em consequência, evitar que esses estudantes abandonem seus estudos

No Brasil, existem diversos termos utilizados para categorizar o sucesso ou insucesso dos estudantes. De uma maneira geral, o sucesso dos estudantes está relacionado à obtenção do diploma de finalização do curso, chamado de diplomação. Outros termos são utilizados para descrever o insucesso escolar e os mais empregados são: evasão, abandono ou atraso (MANHÃES, 2014).

O Ministério da Educação e Cultura do Brasil -MEC- (1997) define que a evasão pode ser entendida em três eixos:

a) Evasão de curso - quando o estudante se desliga do curso em situações diversas: abandono (deixa de matricular-se), desistência (oficial), transferência (mudança de curso), exclusão por norma institucional;
b) Evasão da instituição - quando o estudante se desliga da instituição na qual está matriculado;
c) Evasão do sistema - quanto o estudante abandona de forma definitiva ou temporária o ensino.

O termo retenção é outro termo que costuma ser comparado com a evasão e trata-se do estudante que, apesar de esgotado o prazo máximo de integralização curricular fixado pelas normas da instituição, ainda não concluiu o curso, mantendo-se matriculado, sendo mais fácil de encontrá-lo. Por outro lado, o estudante evadido quebra o vínculo e perde-se o contato.

No contexto deste trabalho refere-se à evasão como '**evasão do curso**', que são os casos de estudantes que, uma vez matriculados, não concluem o curso por qualquer motivo.

No registro acadêmico do estudante existe uma diferenciação entre evadido e desistente, considerando-se evadido aquele que abandonou o curso sem comunicar e desistente aquele que confirmou a sua saída do curso. A esses casos juntam-se os casos de estudantes reprovados no curso, ou seja, com pelo menos uma disciplina reprovada no tempo previsto nas normas. Neste trabalho todos esses casos foram rotulados como "evadidos", como uma forma de abranger os não egressos do curso.

A Figura 11 corresponde à Tabela 3.5 do relatório do CENSO EAD.BR 2019/2020 (ABED, 2021) e mostra o percentual a quantidade de instituições por percentual de evasão até 2019.

Figura 11- Percentual de evasão observado nas IES públicas.

Tipos de oferta	Número de IES públicas respondentes por percentual de evasão										
	Entre 0% e 5%	Entre 6% e 10%	Entre 11% e 15%	Entre 16% e 20%	Entre 21% e 25%	Entre 26% e 50%	Entre 51% e 75%	Total de IES com evasão	Inf. indisponível	Vazia	Não se aplica
Graduação EAD	5	2	8	9	11	18	2	55	9	143	1
Pós-graduação	1	1	0	1	0	5	0	8	3	197	0
Livre não corporativo	13	13	5	9	12	14	2	68	14	122	4
Livre corporativo	4	6	2	2	7	8	0	29	9	169	1
Graduação presencial	22	21	20	18	6	9	2	98	29	78	3

Fonte: ABED (2021).

A maior quantidade de casos se concentra entre os 26% e 50% de evadidos em cursos de graduação EaD. Esses números mostram que existe menos evasão nos cursos presenciais e maior evasão nos cursos EaD, porém não há informações sobre cursos técnicos, pois trata-se de um censo para IES e, atualmente, não existe esse tipo de levantamento.

2.4 Mineração de Dados

Para Fayyad *et. al* (1996), MD pode ser definida como o processo envolvido na extração de informações interessantes, interpretáveis, úteis e novas de dados. Amaral (2016) destaca que as técnicas de MD permitem extrair, analisar e explorar conhecimento de uma massa de dados em busca de padrões, erros e associações que, de outra maneira, permaneceriam escondidos nas grandes bases. Já o AM é um método de análise de dados que automatiza a construção de modelos analíticos (CORCOVIA e ALVES, 2019). Técnicas de AM são frequentemente empregadas em processos de MD.

Os sistemas de gerenciamento acadêmicos armazenam grandes volumes de dados sobre os estudantes e o processo para converter esses dados em informações úteis é conhecido como KDD (*Knowledge Discovery from Data* ou Descoberta de Conhecimento em Base de Dados). O KDD faz uso de algoritmos específicos para a extração de padrões dos dados (WITTEN *et al*., 2016) e a MD é uma de suas partes fundamentais (Figura 12). Segundo Goldschmidt *et al*. (2015) e Han e Kamber (2011), as etapas do KDD são:

i. Seleção de dados: identificar e selecionar quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo;

ii. Limpeza de dados: Realizar tratamento sobre os dados, a fim de assegurar a qualidade relacionada à completude, veracidade e integridade, ou seja, dados inconsistentes ou fora dos padrões são removidos;

iii. Integração de dados: Reunir várias fontes de dados, mantendo a consistência e a coerência dos dados integrados;

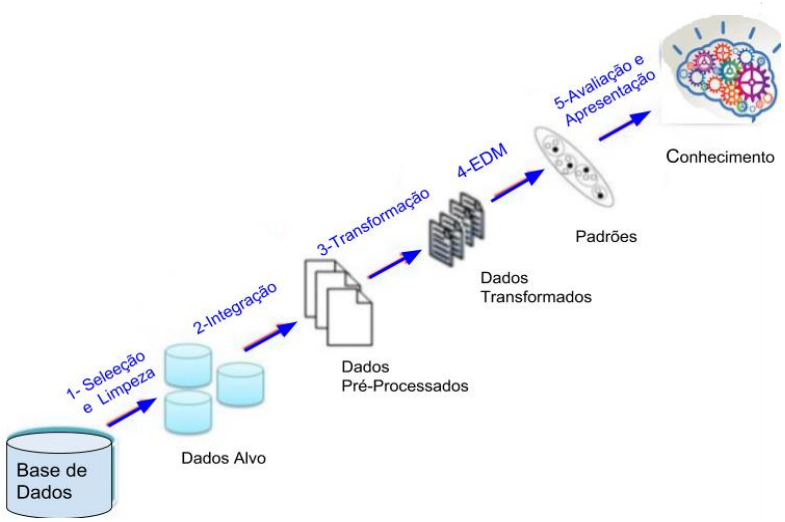
iv. Transformação de dados: Codificar os dados para o formato apropriado para a próxima etapa. Esta fase é realizada dependendo dos algoritmos que será aplicado na mineração de dados;

v. Mineração de Dados: Aplicar métodos com o propósito de extrair os padrões de interesse;

vi. Avaliação de padrões: Identificar os padrões de interesse de acordo com algum critério do usuário;

vii. Apresentação de conhecimento: Tornar o conhecimento extraído compreensível ao homem através de gráficos, diagramas ou relatórios demonstrativos.

Figura 12- Etapas do Processo KDD de Descoberta de Conhecimento em Banco de Dados.



Fonte: Adaptado de Goldschmidt et al. (2015).

De modo geral, os trabalhos que utilizam mineração de dados para estudos de evasão escolar seguem as etapas definidas pelo KDD (JIMÉNEZ-GÓMEZ *et al.*, 2015; MEHTA e BUCH, 2016).

2.4.2 Mineração de Dados no Contexto Educacional

Neste trabalho, o conhecimento de mineração de dados é combinado com estratégias educacionais para identificar as características que delineiam o comportamento dos estudantes. E, segundo Barnes *et al.* (2009), a Mineração de Dados Educacionais (MDE) é a aplicação de

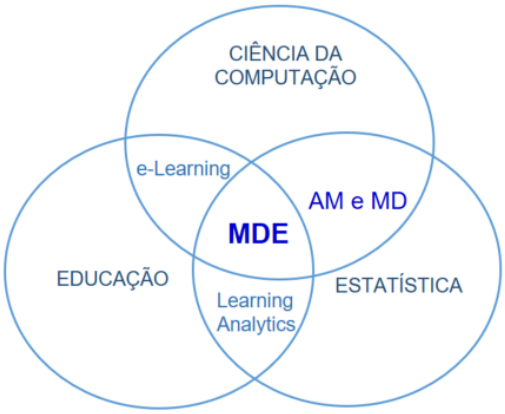
técnicas de Mineração de Dados a dados educacionais e, portanto, seu objetivo é analisar esses tipos de dados para resolver problemas de pesquisa educacional.

Cambruzzi *et al.* (2014) entendem que as instituições de ensino, em geral, estão preocupadas com ações preventivas para minimizar a evasão e que, para isso, precisam identificar variáveis associadas com características dinâmicas, como interações com o ambiente de aprendizado, pessoas ou materiais instrucionais e ferramentas de apoio utilizadas nas aulas. E essa grande disponibilidade de conjuntos de dados, que em boa parte são dados não estruturados, tem fomentado o interesse nas técnicas de MD na busca de respostas específicas da Educação.

A primeira Conferência Internacional sobre MDE foi realizada em 2008, em Montreal, e desde então são organizadas conferências anuais. As sociedades mais populares são a Sociedade Internacional de *Educational Data Mining*, criada em 2011, e da IEEE *Task Force of Educational Data Mining*[15], formada em 2012. Isto indica que o uso de mineração de dados em ambiente educacional é relativamente recente.

Em MDE adaptam-se métodos da Estatística, AM e de MD para estudar dados educacionais gerados basicamente por estudantes e instrutores. Sua aplicação pode ajudar a analisar os processos de aprendizagem dos estudantes, considerando sua interação com o meio ambiente (SIEMENS e BAKER, 2012). A partir do seu relacionamento com as diversas áreas de conhecimento, a interseção de Ciência da Computação com Estatística é a subárea de MD e AMe a subárea *e-Learning*, aprendizado on-line, é a interseção da Ciência da Computação com a Educação (Figura 13).

Figura 13 - Principais áreas relacionadas a MDE.



Fonte: Adaptado de Romero e Ventura (2013).

[15] http://datamining.it.uts.edu.au / EDD

Inicialmente, algumas oficinas foram realizadas em conferências sobre Inteligência Artificial na Educação e Sistemas Tutores Inteligentes (STI). O aumento do desenvolvimento de cursos on-line abertos massivos (MOOCs) e a nova disponibilidade de STI permitem que os pesquisadores coletem grandes quantidades de dados, e esses grandes conjuntos de dados facilitam a aplicação de abordagens de Aprendizado de Máquina e ciência de dados, por exemplo, para abordar tarefas como predição e prevenção de evasão ou personalização para melhorar o aprendizado (TERUEL e ALAMY, 2018).

Para Manhães (2014) a área da MDpossui recursos e técnicas que podem ser utilizados para resolver problemas de predição em diversas áreas e, atualmente, está sendo utilizada, ampla e consistentemente, para resolver problemas envolvendo dados educacionais. Por exemplo, Cunha *et al.* (2016) utilizam as técnicas de MDpara detectar comportamentos relacionados a evasão escolar e a reprovação através de uma base de dados acadêmica.

2.4.3 Técnicas Aplicadas na MDE

No Brasil, o primeiro evento com temática relacionada à MDE aconteceu em 2004 com análise de dados educacionais na 7ª. Conferência Internacional de Sistemas Tutores Inteligentes. Posteriormente, Baker *et al.* (2011) fizeram uma análise dos principais trabalhos de pesquisa na área de MDE e apresentaram várias linhas de pesquisa com uma classificação das principais sub-áreas de pesquisa empregadas em MDE em formato de taxonomia. Dentre as técnicas destacam-se predição, agrupamento e mineração de relações.

a) Na predição o objetivo é inferir o valor de um conjunto de variáveis-alvo, denominadas variáveis preditivas (*predicted variables*), através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras (*predictor variables*). Em predição tem as sub-áreas de classificação, regressão e estimação de densidade (BAKER *et al*., 2011).
b) No agrupamento, o objetivo principal é encontrar uma estrutura que permita separar os dados em grupos ou categorias coesas, de maneira que os elementos de um grupo tenham características semelhantes entre si e distintas dos elementos dos outros grupos (AMERSHI e CONATI, 2009). Normalmente, utiliza-se de algum tipo de medida de distância ou similaridade para determinar quais objetos possuem características próximas ou não. Vellido *et al.* (2007) sugerem formar grupos de alunos com base em seus padrões de aprendizagem e de intervenção. Amershi e Conati (2009) também dão exemplos como achar grupos de escolas (para investigar as diferenças e similaridades

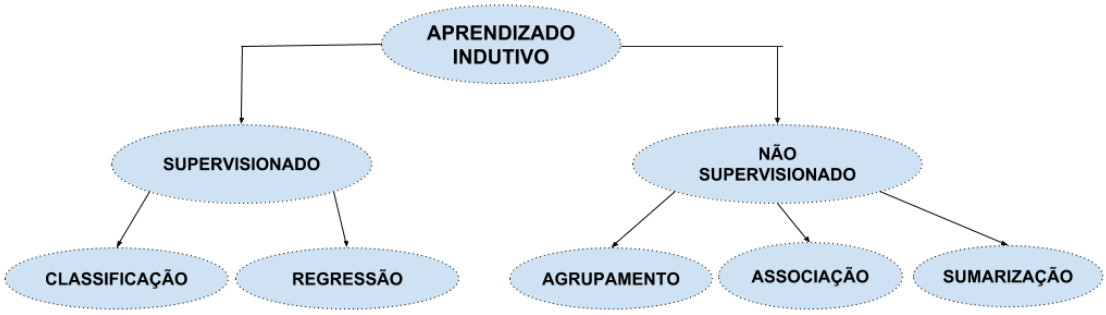
entre escolas), ou achar grupos de alunos (para investigar as diferenças e similaridades entre alunos), ou até grupos de atos (para investigar padrões de comportamento).

c) Em mineração de relações, usa regras de associação, correlações, padrões sequenciais e mineração de causas entre variáveis. Essas técnicas podem ser usadas para gerar padrões descritivos através da avaliação do comportamento dos dados (FAYYAD *et al.*, 1996). Segundo Baker *et. al* (2011, p.3), "...a meta é descobrir possíveis relações entre variáveis em bancos de dados. Esta tarefa pode envolver a tentativa de aprender quais variáveis são mais fortemente associadas com uma variável específica, previamente conhecida e importante, ou pode envolver as relações entre quaisquer variáveis presentes".

2.5 Técnicas de Aprendizado de Máquina aplicadas em MDE

Os modelos de AM são geralmente obtidos por um processo indutivo, no qual podemos dizer que o objetivo é "generalizar os dados". O Aprendizado Indutivo pode ser sub-categorizado em dois tipos: supervisionado e não-supervisionado (HEYKIN, 2009; FRIEDMAN *et al*., 2001; JAMES *et al*., 2013), como mostrado na Figura 14.

Figura 14 - Hierarquia das técnicas de Aprendizado de Máquina.



Fonte: Faceli *et al.* (2017)

Classificação e regressão são as principais tarefas de AM supervisionado. Na classificação, o rótulo indica a categoria em que o exemplo correspondente se enquadra; na regressão, o rótulo é uma saída de valor numérico, como temperatura, altura, preço etc. (ZHOU e LI, 2010). Como exemplo de classificação pode-se considerar a tarefa de predizer se um estudante irá evadir-se ou não de um curso EaD, enquanto um exemplo de regressão pode ser a predição da nota obtida por um estudante (por exemplo, em uma escala de 0 a 10).

Alguns bons exemplos de algoritmos para aprendizagem supervisionada são: árvores de decisão (KOTSIANTIS, 2011; NEVILLE 1999), florestas aleatórias (LOUPPE 2014; CUTLER 2010), k-vizinhos mais próximos ou k-NN (YADAV e SHUKLA, 2016), regressão (HARREL, 2011) e regressão logística (SHEATHER, 2009; PARK, 2013).

Por sua vez, no aprendizado não supervisionado, o objetivo é explorar o conjunto de dados a partir de sua regularidade, e não há um rótulo a dar na saída do algoritmo. O agrupamento, abordado na seção anterior, é uma das tarefas mais importantes de AM supervisionado. Outras tarefas importantes são associação, na qual procura-se padrões de associações entre os elementos de um conjunto de dados, e sumarização, na qual o objetivo é encontrar uma descrição simples para um conjunto de dados.

Um exemplo de técnicas de agrupamento para análise de evasão é do Oeda e Hashimoto (2017), que analisam as edições nos códigos fonte de aulas de programação armazenadas como dados de log para detectar os estudantes que não conseguem acompanhar as aulas. Para a detecção de *outliers* (anomalias) para agrupar dados usando algoritmos K-means em *clustering* do aprendizado não supervisionado e sistemas de tutores inteligentes.

Para problemas de evasão de estudantes são encontradas, na sua grande maioria, técnicas de classificação. Algumas pesquisas com essa abordagem são as citadas no resultado da revisão de literatura, no Capítulo 3. Várias dessas pesquisas propõem uma extensão ou adaptação de resultados publicados, outros fazem comparativos das acurácias dos diferentes algoritmos, e no caso do uso de técnicas não supervisionadas, que são minoria, propõem soluções alinhadas às técnicas de LA (*Learning Analytics*).

Como resultado dos conceitos estudados nesta seção, segue o Quadro 1 com o comparativo das técnicas supervisionadas e não supervisionadas e as suas principais diferenças.

Quadro 1 - Comparativo de técnicas supervisionadas e não supervisionadas em AM.

Supervisionada	**Não supervisionada**
A amostra é composta por dados rotulados (variáveis preditivas)	Dados não rotulados (não há variável preditiva)
Procura um mapeamento entre as características de entrada e o alvo	Encontra uma estrutura nos dados
Usado principalmente para tarefas preditivas	Usado principalmente em análise

As principais tarefas são classificação e regressão	As principais tarefas são agrupamento e identificação de anomalias
Geralmente, o objetivo é maximizar uma métrica de desempenho (ex.: acurácia)	Geralmente, o objetivo é encontrar uma representação útil e potencialmente nova dos dados

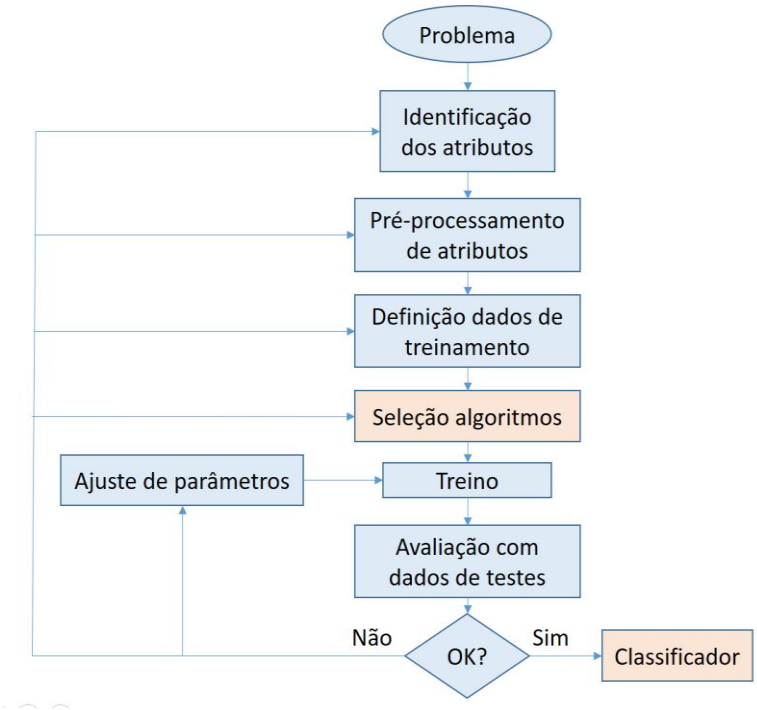
Para definir os atributos do modelo, segundo Faceli *et al.* (2017, p.47), “as diversas técnicas existentes podem ser divididas em duas grandes abordagens: agregação e seleção de atributos. Enquanto as técnicas de agregação substituem os atributos originais por novos atributos formados pela combinação de grupos de atributos, as técnicas de seleção mantêm uma parte dos atributos originais e descartam os demais atributos”. Neste trabalho aplicaram-se ambas técnicas nos experimentos.

As técnicas de agregação de atributos também são conhecidas como extração de características, ao transformar ou combinar um conjunto de características originais para extrair ou criar novas características. É comum que essa técnica de extração preceda a seleção das características ou atributos, que elimina os atributos mais irrelevantes. Deve ser seguido um determinado critério para não reduzir de forma a perder o poder de discriminação, e estimar a dimensão ideal para o conjunto de dados.

Na etapa de pré-processamento deste trabalho foi utilizada a técnica de agregação para transformar 9 tipos de ações registradas no log em 4 atributos e, posteriormente, foi utilizada a técnica de seleção de atributos baseado nos critérios da ferramenta de AM, que reduziu os 13 atributos iniciais para 6 atributos, utilizados na construção dos modelos.

Após o pré-processamento, as técnicas supervisionadas consistem em usar o conjunto de dados inicial e comparar uma série de algoritmos de classificação, definindo o melhor resultado obtido com o algoritmo e taxa de acerto em relação à média dos demais (KOTSIANTIS, 2016). Com base nessa definição, o processo de aplicação de ML supervisionado é descrito na Figura 15, para escolha de um classificador com base na amostra dos dados de treinamento e aplicada nos dados de testes. É um modelo que requer de vários ciclos de comparações e ajustes, com situações que pode ser que precise reiniciar o processo desde a identificação dos atributos.

Figura 15- Etapas da Classificação em Aprendizado de Máquina.



Witten *et al.* (2009) destaca os cinco principais métodos classificadores: árvore de decisão (DT–*Decision Tree*), incluindo florestas aleatórias (RF–*Random Forests*) e árvores induzidas com *Gradient Boosting* (GB), Naive Bayes (NB), rede neural rasa (MLP–*Multi Layer Perceptron*), k-vizinhos mais próximos (kNN–*k-Nearest Neighors*) e máquina de vetores de suporte (SVM–*Support Vector Machine*), e acrescentou-se regressão. Baseado nesses métodos, associaram-se os algoritmos mais importantes a cada categoria (Quadro 2).

Cada uma destas técnicas possui vantagens e desvantagens, assim como cenários de aplicação mais propícios.

Quadro 2 - Algoritmos ou modelo para os principais métodos de classificação.

Categoria	Algoritmo ou Modelo
Árvore de Decisão	ID3, C4.5, CART, *Random Forest* (RF) *Gradient Boosting* (GB)
Bayesiano	Naive Bayes (NB)
Rede Neural	*Multilayer Perceptron* (MLP), *Backpropagation, Deep Learning*(DL)
kNN (*K-Nearest Neighbor*)	kNN
SVM (*Support Vector Machine*)	*Sequential Minimal Optimization* (SMO)
Regressão	Regressão gradiente descendente

Por outro lado, as técnicas do aprendizado não supervisionado, ou descritivo, de acordo com Faceli *et al.* (2017, p.178), "se referem à identificação de informações relevantes nos dados sem a presença de um elemento externo para guiar o aprendizado". Uma técnica com uma trajetória de mais de 50 anos é o K-means (JAIN, 2010), que, apesar de ser tão antigo e vários outros algoritmos de agrupamento terem sido publicados desde então, ainda é amplamente utilizado.

Faceli *et al*. (2017) definem o *cluster* como uma coleção de objetos próximos ou que satisfazem alguma relação espacial, considerando os objetos pontos em um espaço de dimensão *d*, e ressaltam que não existe uma definição formal única e precisa para *cluster*, embora a ideia do que constitui seja intuitiva (grupos de objetos similares).

Estivil-Castro (2002) e Kleinberg (2002) afirmam que não existe um único algoritmo de agrupamento capaz de encontrar todos os tipos de agrupamentos que podem estar presentes em um conjunto de dados. Segundo Handl e Knowles (2007), existe a possibilidade de que mais de uma estrutura relevante esteja presente em um conjunto de dados. O agrupamento é uma tarefa inerentemente subjetiva.

Como exposto, a técnica de agrupamento parece intuitivamente fácil e ao mesmo tempo complexa para ser resolvida em algoritmos e existe uma certa dificuldade em projetar um algoritmo de agrupamento de propósito geral.

Normalmente, em um agrupamento K-means é informando o valor de k, que define o número de *clusters* ou grupos, como Márquez-Vera *et al*. (2016) ou Manhães *et al.* (2014) que separam em 3 grupos, prevendo estudantes excelentes, regulares e ruins. Outros grupos foram descobertos e avaliados por Riestra-González *et al.* (2020) que definem 6 grupos como o adequado para aplicar em qualquer tipo de curso e duração. Com a identificação dos grupos pode-se descrever as características de cada um deles, permitindo achados nos dados originais e descobrir correlações entre os atributos dos dados que não seriam facilmente percebidas sem o emprego de agrupamento.

Abordagens supervisionadas são utilizadas na maior parte das aplicações de IA e trabalham com elementos rotulados, e no mundo real esse rótulo será conhecido só no final do curso, ou seja, a técnica de aprendizagem poderá ser aplicada depois que o curso acabar, quando os rótulos estão disponíveis e confirmados. Dessa forma, seria possível fazer uma predição em

AM com uma base de treinamento com instâncias rotuladas, usando dados passados para aprender e classificar situações futuras.

É importante não generalizar uma predição inicial porque as informações de log reais são diferentes daquelas usadas para treinar os modelos. Esse cuidado é alertado por López-Zambrano *et al.* (2020), que estudaram a portabilidade de modelos preditivos de desempenho do estudante em cursos com o mesmo grau e nível semelhante de uso de AVA. Ao transferir modelos entre cursos do mesmo nível, os valores de AUC[16] caem de 0,09 a 0,28. Essas perdas variam ao portar modelos entre cursos com um nível semelhante de uso do Moodle.

Essa baixa no desempenho sugere que o cenário ideal seria que o modelo fosse treinado com dados do curso atual, ao invés de replicar o modelo gerado em outro contexto, pois cada grupo possui características próprias. O ambiente educacional tem características de um ambiente dinâmico, na qual os cursos, colegas, professores e outros fatores podem influenciar sobre o comportamento e o sucesso na obtenção do diploma.

Encontrar uma solução para predição de um contexto dinâmico é um grande desafio, pois ir rotulando ou ajustando de forma online pode custar "caro". De maneira geral, Faceli *et al.* (2017) afirmam que não existe técnica universal, ou seja, não é possível estabelecer a priori que uma técnica de AM em particular se sairá melhor na resolução de qualquer tipo de problema. Conforme Wolpert (1996), o desempenho de um algoritmo depende de como seu viés indutivo se adapta a propriedades presentes no conjunto de dados a ser utilizado.

2.6 Considerações Finais sobre o Capítulo

O capítulo apresentou os principais conceitos relacionados à MDE de Dados Educacionais, que neste trabalho usa amostra de estudantes de cursos técnicos concomitantes ao ensino médio e destaca a importância social e econômica do ensino técnico no Brasil, que ainda precisa ser melhor explorada nas pesquisas. Também, foram apresentados os conceitos de AM para aplicação na predição de desempenho e de risco de evasão escolar. A continuação, o Capítulo 3 apresenta os trabalhos relacionados a esta tese.

[16] área sob o valor da curva ROC está entre 0,5 e 1, onde 0,5 denota um classificador errado e 1 denota um classificador excelente.

Capítulo 3
Estado da Arte

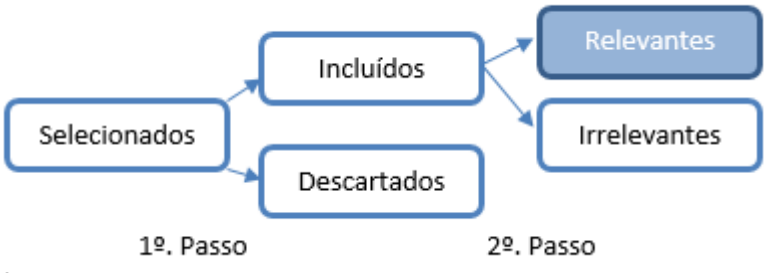
Neste capítulo são apresentados os trabalhos relacionados ao tema de pesquisa por meio de uma revisão sistemática da literatura (RSL), que abrange os conceitos considerados no Capítulo 2 e que permite uma visão ampla dos métodos propostos ou implementados no Brasil e no mundo para reduzir a evasão de estudantes na modalidade EaD usando as técnicas de AM e, dessa forma, agrupar e/ou classificar os trabalhos encontrados

3.1 Revisão Sistemática da Literatura

Esta pesquisa foi direcionada por uma Revisão Sistemática da Literatura (RSL) sobre soluções ou propostas para predição de evasão de estudantes usando técnicas de AM (TAMADA et al., 2019). A RSL foi executada em três fases, de acordo com Kitchenham e Charters (2007): Planejamento, Condução e Relatórios.

As etapas de planejamento (Figura 16) envolveram a identificação da necessidade de uma revisão e o desenvolvimento de um protocolo de revisão. Os artigos recuperados na busca automatizada foram armazenados em uma lista que inclui: título, autores, ano, resumo e palavras-chave.

Figura 16 - Etapas da Revisão Sistemática.



A condução das etapas envolveu a identificação de pesquisas, a seleção de estudos primários, o estudo de avaliação da qualidade, extração e monitoramento de dados e síntese de dados. Para esta etapa, o título, as palavras-chave e os resumos foram analisados pelo pesquisador, que determinou se o artigo era relevante ou irrelevante para a revisão. Todos os trabalhos selecionados foram reavaliados com base na leitura do texto na íntegra. No final, os trabalhos selecionados nesta etapa foram utilizados para extração de dados.

Dessa forma, o objetivo desta seção é conhecer os trabalhos relacionados a esta pesquisa, encontrados através da revisão sistemática da literatura, e os resultados obtidos servem de base teórica para a escolha de técnicas mais adequadas ao contexto educacional e, especificamente, as técnicas de predição de evasão que permitam definir um método eficiente de intervenções pedagógicas para que o estudante obtenha o diploma.

3.1.1 Processo de Busca em RSL

O objetivo foi pesquisar propostas para reduzir as taxas de evasão de estudantes na educação a distância usando técnicas de AM para promover ações proativas.

Com o objetivo de orientar o processo de construção da *string* de busca, a sequência de pesquisa genérica usada em nossa pesquisa é mostrada no Quadro 3. A *string* de busca genérica foi modificada para se ajustar à pesquisa avançada de cada uma das principais bibliotecas acadêmicas digitais on-line em Tecnologia: IEEE Xplore[17], ACM Digital Library[18], *Scopus*[19],*SpringerLink*[20] *e Research Gate*[21]. A *Scopus* armazena publicações de diversas fontes como *Springer*, *ACM*, *ScienceDirect/Elsevier*[22] e *British Computer Society* e algumas do IEEE Xplore, e eventuais duplicidades são descartadas no primeiro passo.

Quadro 3 - Termos utilizados na *string* de busca.

Termo Principal	Termos derivados
Evasão	(Dropout OR “dropping out”)
Educação	(“distance learning” OR “distance education” OR e-Learning OR VLE OR “virtual learning environment“ OR LMS OR EDM OR “educational data mining”)
Aprendizado de Máquina	“Machine learning”
Predição	(predict OR predicting OR prediction OR predictive)

[17] IEEE Xplore: https://ieeexplore.ieee.org/Xplore/home.jsp

[18] ACM Digital Library: https://dl.acm.org

[19] Scopus: https://www.scopus.com

[20] SpringerLink: http://link.springer.com

[21] Research Gate: https://www.researchgate.net/

[22] Science Direct: http://www.sciencedirect.com

3.1.2 Seleção de Dados

Na seleção primária, foram lidos o título, resumo e palavras-chave, enquanto na seleção secundária foram lidas a introdução e a conclusão. As publicações recuperadas pelos mecanismos de busca foram organizadas com auxílio de uma ferramenta específica para organização de pesquisa denominada StArt (*State of Art throught Systematic Review*)[23].

Durante a seleção na primeira etapa é usado o *checklist* de inclusão para organizar a inclusão de estudos que estão dentro do escopo da pesquisa e a exclusão dos que estão fora do escopo. As publicações até 2019 que foram selecionadas na etapa inicial até chegar nas mais relevantes estão disponíveis em https://goo.gl/1U3AKS.

3.1.3 Análise dos Resultados

Os 13 artigos de pesquisa selecionados na etapa final (Quadro 4) mostram técnicas ou conjuntos de técnicas que predizem com eficiência os estudantes que correm o risco de desistir e propõem uma solução.

Foram selecionados para a etapa final da RSL 13 trabalhos (Quadro 4), todos os quais têm como objetivo principal a predição e redução de evasão. Os trabalhos abordam diferentes técnicas de predição de evasão no ambiente educacional e apontam possíveis causas de evasão.

Seis artigos (UDDIN e LEE, 2017; CHEN *et al.*, 2016; FEI e YEUNG, 2015; HONG e WEI, 2017; QIU e LIU, 2018; LAVETI *et al.*, 2017) combinam técnicas de predição mostrando novas abordagens e resultados mais eficientes do que métodos únicos. Uddin e Lee (2017) combinaram árvores de decisão e rede Bayesiana para projetar seu modelo.

Na sua maioria, os trabalhos em predição de evasão utilizaram técnicas supervisionadas e foram encontrados só dois trabalhos com métodos de aprendizado não supervisionado. Oeda e Hashimoto (2017) empregaram K-Means e os *outliers* dos grupos são estudantes que se destacam por terem desempenho muito acima ou muito abaixo do esperado. Teruel e Alonso (2018) usaram redes LSTM como etapa inicial de uma solução não supervisionada para predição de evasão em um MOOC.

Quadro 4 - Demonstrativo dos artigos selecionados na RSL.

[23] http://lapes.dc.ufscar.br/ , desenvolvida na Universidade de Federal de São Carlos, pelo Laboratório de Pesquisa em Engenharia de Software (LaPES).

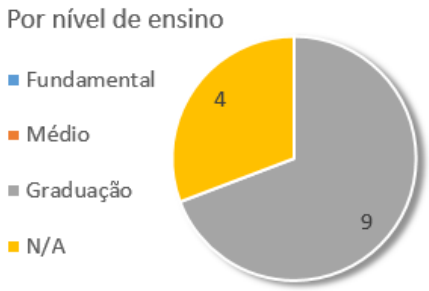
Título	Autor(es)	Ano	Algoritmo[24]	Visão Geral
Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout.	Burgos *et al.*	2018	LR	12 variáveis preditivas com base nos dados acadêmico: Redução de 14% na taxa de evasão.
Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention	Xing, W.; Du, D	2018	DL	· Pontuações de *quiz* · Dados de rastreamento fornecidos diretamente pela API contendo informações sobre as páginas visitadas
Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction	Uddin, M.F.; Lee, J.	2017	DT NB	· 20 atributos de personalidade e formação acadêmica · 15 atributos de redes sociais · Uso de dados acadêmicos. · Uma pesquisa online foi projetada e usada para coletar dados dos estudantes.
DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction	Chen, Y. *et al.*	2016	LR NN RF	· Marcas de atribuição ·quantidade de postagens ·quantidade de registros de pausa ·quantidade de dias ativos
Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes	Oeda, S.; Hashimoto, G.	2017	K-means	- Histórico da digitação nos comandos dos algoritmos - Formação de clusters por anomalias
What decides the dropout in MOOCS?	Lu, X. *et al*.	2017	SVM	19 principais atributos depois de analisar a estrutura de log
Temporal Models for Predicting Student Dropout in Massive Open Online Courses	Fei, M.; Yeung, D.Y.	2015	LSTM RNN	· Nº de vezes que um estudante navega pela página do curso · Nº de atividades-problema do curso · Nº de atividades com outros objetos do curso
A Big Data Framework for Early Identification of Dropout Students in MOOC	Tang, J.K.T. *et al*.	2017	DT	· Informação demográfica, · Informações de inscrição · Informações de atividade, como o intervalo de tempo médio entre os vídeos
Discovering learning behavior patterns to predict dropout in MOOC	Hong, B. *et a*l.	2017	MLR RF SVM	13 características de comportamento de aprendizagem: número ativo, taxa ativa, visita de vídeo, visita wiki etc.
An Integrated Framework with Feature Selection for Dropout Prediction in Massive Open Online Courses	Qiu, L.; Liu, Y.	2018	LR RBF SVM	· lição · Assistir o vídeo · acessar ou navegar em outros objetos · ler o wiki

[24] Siglas dos algoritmos: DL- Deep Learning; DT- Decision Tree; GB- Gradient Boost; K- K means; kNN- K Nearest Neighbors; LR- Logistic Regression; LSTM- Long Short Term Memory; MLP- Multilayer Preceptron; MLR- Multinomial Logistic Regression; NB- Naive Bayes; NN- Neural Networks; RBF- Radial Basis Function Neural Network; RF- Random Forest; RNN- Recurrent Neural Network; SVM- Support Vector Machine.

Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing	Laveti, R. *et al.*	2017	GBT LR RF	·16 funcionalidades: quantidade de eventos na última semana, quantidade dias entre o final do curso e o último dia de acesso ao material do curso etc.
Dropout detection in MOOCs: An exploratory analysis	Isidro, C. *et al.*	2018	DL	· Nº de exercícios resolvidos corretamente · Nº de vezes que os vídeos são rebobinados · Nº de tentativas para resolver cada exercício · Nº de problemas realizados
Co-embeddings for Student Modeling in Virtual Learning Environments	Teruel, M.; Alonso A.	2018	LSTM RNN	logs de eventos como acesso ao conteúdo de vídeo, resolução de um problema etc.

A maioria dos trabalhos (11 dos 13 no total) pesquisaram cursos curtos com método MOOC de ensino massivo, uma metodologia que foca na autonomia do estudante, sem tutor e não exige ligação a uma instituição de ensino. Nos dois trabalhos restantes, um utilizou o Moodle em curso de graduação (BURGOS *et al.*, 2018) e uma publicação (UDDIN e LEE, 2017) não especificou o tipo de ambiente pesquisado. Em relação ao nível de ensino (Figura 17), nove publicações focaram pesquisas sobre estudantes do ensino superior e os quatro restantes usaram bases de dados públicas e não informam qual o nível de ensino da amostra de dados analisada.

Figura 17 - Nível de ensino nos trabalhos selecionados.



A maioria das pesquisas utilizou predição via mineração de registros de fluxo de cliques (UDDIN e LEE, 2017; CHEN *et al.*, 2016; LU *et al.*, 2017; FEI e YEUNG, 2015; HONG e WEI, 2017; QIU e LIU, 2018) em atividades de vídeo, extraindo o comportamento do estudante, que contém características fundamentais como refletir (reproduzir e pausar) e revisar (reproduzir e pular para trás, rever novamente).

Além disso, foram testados fóruns de discussão (XING e DU, 2018; CHEN *et al*., 2016; LU *et al.*, 2017; FEI e YEUNG, 2015; HONG e WEI, 2017; QIU e LIU, 2018), notas ou tentativas de questionário (XING e DU, 2018), registros de notas, horários de ensino (BURGOS *et al*, 2018), redes sociais (UDDIN e LEE, 2017) e outros (Quadro 5).

Vários autores fazem a predição de abandono construída semanalmente (BURGOS *et al*., 2018; XING e DU, 2017; CHEN *et al*., 2016; LU *et al.*, 2017; FEI e YEUNG, 2015; ISIDRO *et al*., 2018; TERUEL, 2018). Dentre eles, Xing e Du (2018) exploram a técnica de Aprendizagem Profunda para construir um modelo semanal de predição. Para Fei e Yeung (2015) os recursos registrados na segunda semana são diretamente concatenados ou adicionados aos recursos registrados na primeira semana e da mesma forma para as semanas posteriores. Isidro *et al*. (2018) geram um arquivo diariamente para registrar todos os eventos, e modelos são gerados semanalmente para cada um dos estudantes, de acordo com a data em que se matriculam no curso.

3.2 Trabalhos de pesquisa sobre agrupamento de estudantes em AVA

A RSL demonstrou que a maioria das pesquisas sobre o tema de evasão de estudantes utiliza técnicas supervisionadas. Essa RSL foi complementada por um mapeamento não sistemático da literatura, no qual a intenção era encontrar trabalhos que usem técnicas não supervisionadas na MDE. Os trabalhos de pesquisa a seguir visam agrupar os estudantes no que diz respeito à sua interação com AVAs.

O conjunto de dados de Talavera e Gaudioso (2004) inclui dados de log do AVA, conhecimento prévio, dados demográficos e interesses dos estudantes (obtidos de pesquisas realizadas com estudantes) para descobrir padrões de interação do AVA de um curso. O algoritmo Expectation-Maximization (EM) encontrou seis grupos apresentando boa correlação com o desempenho dos estudantes.

Hung e Zhang (2008) trabalharam entradas de log para 98 estudantes de graduação para descobrir os comportamentos de aprendizagem online dos estudantes. Foram encontrados três grupos, sendo que dois correlacionaram-se com estudantes com desempenho acima da média e alta interação com o AVA, ao passo que o terceiro se correlacionou com padrões de baixa interação e baixo desempenho.

Romero *et al*. (2013) analisaram dados da interação dos estudantes com o AVA, incluindo engajamento nos fóruns de discussão, além da interação com as atividades e as notas

obtidas. Eles compararam técnicas puramente supervisionadas (classificação) com estratégias em dois passos nas quais os estudantes são agrupados por um algoritmo EM e depois regras de associação são extraídas para prever casos futuros de desempenho bom ou ruim.

Cerezo *et al*. (2016) utilizaram dados extraídos de logs do Moodle para agrupar estudantes de um curso a distância para estudar seu processo de aprendizagem assíncrona. Os algoritmos K-means e Expectation-Maximization (EM) descobrem que as características de procrastinação e socialização são as mais determinantes para os grupos. Dos quatro grupos, três grupos apresentam diferenças nas notas finais dos estudantes.

Francis e Babu (2019) propõem abordagens híbridas para avaliar o desempenho acadêmico. Na fase 1, o estudo realizou os experimentos utilizando SVM, NB, Árvore de decisão e Rede Neural para identificar quais das características apresentam melhores resultados. Na fase 2, as características obtidas são passadas para o algoritmo de agrupamento K-Means para adquirir características de agrupamentos indicando alunos de alto, médio e baixo desempenho. Nesta abordagem, onde o classificador é colocado antes do agrupamento, pode existir um viés, pois informações de dados rotulados, indiretamente, são utilizadas no agrupamento.

Riestra-González *et al*. (2020) utilizaram classificadores para criar modelos para a predição inicial do desempenho dos estudantes na resolução de tarefas em AVA, apenas analisando os arquivos de log do Moodle gerados até o momento da predição. Eles analisaram 699 cursos de ensino superior em diferentes áreas, cursos curtos ou de graduação, e com 10%, 25%, 33% e 50% da duração do curso. Como resultado, eles conseguiram identificar quatro padrões de comportamento no Moodle. Eles criaram um modelo que consegue generalizar para qualquer tipo de curso e duração e define que seis grupos é o padrão encontrado. Por outro lado, criaram um modelo de predição de desempenho e as redes neurais MLP (Multilayer Perceptron) obtiveram 80,10% de acerto quando 10% do curso foi ministrado.

No trabalho de Iatrellis (2021), que aplicou um estudo de caso com os estudantes do ensino superior, foram inicialmente agrupados com base na semelhança de fatores e métricas específicas relacionadas à educação e, posteriormente, aplicou um método de classificação e regressão baseado em árvores, com o *Random Forest*. Usou em dados com histórico acadêmico o algoritmo K-*means*, que revelou a presença de três grupos que melhor separavam o conjunto de estudantes. Posteriormente, abordou cada grupo específico e fez uma predição de estudantes individualmente a fim de predizer o tempo de conclusão do curso e matrícula em programas

educacionais oferecidos pela própria instituição, como uma pós-graduação, e, consequentemente, programar e alocar melhor seus recursos, bolsas e auxílio financeiro.

As pesquisas de Riestra-González (2020) e Iatrellis (2021) são os trabalhos encontrados mais recentes, abrangentes e relevantes sobre análise de dados no AVA para predição em cursos EaD, que utilizam tanto técnicas supervisionadas como não supervisionadas na mesma pesquisa.

Um resumo dos trabalhos encontrados nesta revisão de literatura está no Quadro 5, na qual o primeiro da lista é o trabalho encontrado na RSL deste trabalho e os demais foram encontrados após pesquisa específica de técnicas de agrupamento na EaD utilizando as mesmas máquinas de busca. Todos os trabalhos encontrados têm como alvo cursos de nível superior.

Quadro 5 - Síntese da revisão da literatura sobre técnicas não supervisionadas em estudantes no AVA.

Autor(es)	Técnica	Ano	Algoritmos	Alvo	Visão geral
Oeda, S.; Hashimoto, G.	Cluster	2017	K-mean	anomalias	Agrupou estudantes de acordo as habilidades em resolver algoritmos de programação.
Talavera e Gaudioso	Cluster	2004	Expectation-Maximization	N/D	Encontrou seis clusters com interações no AVA de somente um curso
Hung e Zhang	Cluster	2008	K-mean	N/D	Encontrou três clusters com padrões de comportamento de aprendizagem.
Romero *et al*.	Cluster + Associação	2013	Expectativa-Maximização	notas finais	Uso de diferentes abordagens de mineração de dados para melhorar a predição do desempenho final dos estudantes usando dados do fórum.
Cerezo *et al*.	Cluster	2016	K-mean; Expectativa-Maximização	N/D	Encontrou quatro clusters estudando o processo de aprendizagem assíncrona dos estudantes.
Francis e Babu	Classificação e Cluster	2019	SVM, NB, DT, K-mean	Desempenho acadêmico	Três grupos separados em alto, médio e baixo desempenho.
Riestra-González *et al*.	Classificação e Cluster	2020	Perceptron multicamadas K-mean	resolver tarefas	Predição no desempenho das tarefas usando os arquivos de log do Moodle gerados até o momento da predição. Também identificou 6 clusters para 699 cursos em diferentes áreas, tipos e períodos em 10%, 25%, 33% e 50% da duração do curso
Iatrellis	Cluster e Classificação	2021	K-mean Random Forest	Otimizar recursos financeiros da instituição	Encontrou três grupos em cursos do ensino superior para uma melhor predição de tempo para conclusão do curso e matrícula na pós-graduação.

Outros trabalhos de pesquisa usam informações recuperadas de AVAs para predizer o desempenho dos estudantes. No entanto, eles usam vários cursos, mas não fornecem predições antecipadas enquanto o curso está em andamento. Em vez disso, eles usam todos os dados do AVA gerados ao longo do curso, reduzindo sua capacidade de predição antecipada.

3.3 RSLs Complementares

Como parte do trabalho constante de revisão bibliográfica, outros trabalhos foram estudados após a conclusão da RSL apresentada na seção anterior. Destacam-se aqui dois *surveys* recentes que trazem avanços na área de MDE.

Em "*A Literature Review on Intelligent Services Applied to Distance Learning*" (DA SILVA *et al*., 2021), a descoberta dos comportamentos dos estudantes possibilita uma ampla variedade de serviços inteligentes para auxiliar no processo de aprendizagem. Nessa revisão da literatura foram analisados trabalhos publicados entre 2010 a maio de 2021. Dos 51 artigos, a maioria refere-se a sistemas de recomendação (18) e sistemas de ensino (17). São abordados 13 modelos preditivos e apenas 3 trabalhos apresentam instrumentos de avaliação. Dentre os 13 modelos preditivos, todos comparam o desempenho de algoritmos classificatórios e propõem sistemas de aplicação.

A outra RSL é de Albreiki *et al*. (2021), que apresentam publicações entre 2009 e 2021 sobre MDE, principalmente relacionada à identificação de estudantes desistentes e em risco de evasão. Os resultados da revisão indicaram que várias técnicas de AM são usadas para entender e superar os desafios subjacentes; predição de estudantes em risco e predição de desistência dos estudantes, usando dois tipos de conjuntos de dados: bancos de dados de estudantes e das plataformas de aprendizado online.

Novamente, todos os trabalhos encontrados nessas duas revisões são sobre estudos realizados em cursos do ensino superior. A maioria usa técnicas supervisionadas e se referem a modelos de turmas encerradas, mas não explica a eficiência que poderia levar a outras turmas. Dentre as principais tecnologias para análise no AVA, os algoritmos de *Deep Learning* se destacam com três artigos.

No geral, os algoritmos DT, NB e SVM foram aplicados para análise de desempenho e nas predições de evasão, usando dados estáticos e dinâmicos. Da Silva *et al.* (2021) selecionam 51 artigos finais e concluem que:

a) A natureza temporal dos recursos usados para predições de estudantes em risco de evasão não foi estudada em seu potencial. Os valores dessas características mudam com o tempo devido à sua natureza dinâmica. A incorporação de características temporais para classificação pode melhorar o desempenho do preditor.
b) A maioria das pesquisas abordou o problema como uma tarefa de classificação.
c) Os problemas mencionados acima são tratados como classificação binária, enquanto várias outras classes poderiam ser introduzidas para ajudar a gestão a desenvolver planos de intervenção mais eficazes.
d) Foi dado menos atenção às tarefas de engenharia de atributos, onde os tipos de atributos podem influenciar o desempenho do preditor. Dados demográficos, acadêmicos e registros de sessões de interação no AVA foram os recursos mais utilizados nos estudos.
e) Os modelos preditivos usam as informações de registro recuperadas nos estágios iniciais dos cursos, mas eles se concentram apenas em um algoritmo específico.

3.4 Evasão escolar no Ensino Técnico no Brasil

Além das pesquisas abordadas nas seções anteriores, foram pesquisados trabalhos, a partir de 2017, focados especificamente no uso de técnicas para predição de evasão, retenção ou desempenho, em cursos técnicos EaD no Brasil. Esse último filtro foi adicionado só no contexto brasileiro por 2 motivos: as RSL internacionais já mostram o panorama geral e precisou-se de buscas específicas no Brasil, incluindo dissertações e teses; cursos técnicos em outros países tem outras denominações e ofertas diferentes ao sistema educacional brasileiro para poder comparar. Encontraram-se 8 trabalhos, dos quais 6 trabalhos foram descartados e as 2 dissertações abaixo (Quadro 6) foram selecionadas como relevantes e não foram encontradas nas RSL nacionais.

Quadro 6 – Trabalhos sobre predição de evasão escolar em curso técnico.

Autor	Técnica	Ano	Algoritmos	Alvo	Visão geral
Queiroga	Classificação	2017	Floresta Aleatória	evasão	Predição usando apenas contagem de interações no AVA partir da 10ª. semana usando log em curso técnico
Barbosa	Classificação	2020	J48 no Weka	evasão	Análise de perfil de estudantes evadidos em curso técnico.

Queiroga (2017) usou a quantidade de interações dos estudantes para treinar classificadores de predição de evasão em quatro cursos técnicos. Os atributos são a quantidade

diária de interações (coletadas em 721 dias), a quantidade semanal de interações (103 semanas) e estatísticas das interações semanais (média, mediana e desvio padrão), usados em 5 algoritmos de classificação na ferramenta Weka em 2 cenários: a) treinamento e teste do mesmo curso; b) treinamento com 3 cursos e o curso restante para testes, e conclui que a Floresta Aleatória obtém 84% de acurácia na 10ª semana no primeiro cenário e no segundo cenário alcança 80%. Observou que os acadêmicos que acessavam o AVA muito além da média da turma também evadiam. Por outro lado, Barbosa (2020), analisa o perfil do estudante evadido só em 1 curso técnico de Informática com a nota de avaliação e dados da ficha do estudante de curso técnico e conclui que o J48 (árvore de decisão no Weka) tem melhor acurácia.

3.5 Considerações Finais sobre o Capítulo

Este capítulo apresentou uma revisão sistemática da literatura (RSL) que, após uma pesquisa inicial com 199 estudos publicados até 2019, resume as informações de 13 soluções para redução de evasão de estudantes usando técnicas de AM. Com isso, conseguiu-se um levantamento inicial e visão geral das publicações.

Para verificar trabalhos mais recentes, essa revisão inicial foi complementada com duas RSLs com pesquisas a partir de 2019. Também foram feitas outras pesquisas específicas sobre o uso de técnicas não supervisionadas, ainda pouco exploradas para problemas no contexto estudado, para poder comparar com a proposta desta tese. Para RSL em português foi referenciada a pesquisa de Colpo et al. (2020). Por último, foi feita uma pesquisa de trabalhos no Brasil com cursos técnicos EaD.

Pode-se concluir que os métodos supervisionados sugerem ser mais apropriados para analisar os cursos finalizados, por ter o rótulo de classificação, e os métodos não supervisionados são mais apropriados para um curso em andamento.

Encontraram-se na literatura diversos trabalhos que apresentam um método de predição de estudantes em risco de não concluir o curso e, em grande parte, usam técnicas supervisionadas para classificação e verificam a taxa de acerto em turmas encerradas e não fornecem predições antecipadas quando o curso está em andamento. Esta tese define uma metodologia com duas etapas, sendo a primeira com uso de técnicas de agrupamento (não supervisionadas) para análise de comportamento e correlação com o desempenho dos estudantes de turmas em andamento ou encerradas, e os grupos resultantes são utilizados como atributos qualitativos de entrada para métodos de classificação, (técnicas supervisionadas) na

predição da evasão na EaD. Nesse contexto, considera-se que esta tese seja uma pesquisa inédita.

Capítulo 4
Métodos e Experimentos

Este capítulo apresenta a caracterização e definição da metodologia proposta, informando as ferramentas, coleta de dados, pré-processamento, transformação dos atributos e as técnicas de AM. O problema apresentado e seus desdobramentos são analisados sob a ótica das metodologias para predição de estudantes em risco de não concluir o curso, e envolvem conhecimentos de gerenciamento e linguagem de manipulação no banco de dados, conhecimento profundo dos dados no contexto educacional, do uso de técnicas supervisionadas e não supervisionadas de AM, da interpretação dos resultados, ajustes e melhorias.

4.1 Comitê de Ética em Pesquisa (CEP)

As questões de ética da pesquisa com humanos devem necessariamente ser avaliadas e aprovadas por uma comissão específica para esse fim.

A pesquisa nesta tese foi realizada seguindo o protocolo aprovado pelo Comitê de Ética em Pesquisa (CEP) pelo Número do Parecer 2.757.851 (2018). As condições descritas nesse protocolo são o uso de dados secundários provenientes de *backups* de banco de dados do Moodle e de outras bases de dados do campus de Porto Velho Zona Norte do IFRO. Esses dados são trabalhados e divulgados de forma anonimizada.

Esse protocolo foi obtido através da Plataforma Brasil, que é um sistema eletrônico criado pelo Governo Federal para sistematizar o recebimento de projetos de pesquisa que envolvam seres humanos, em todo o país, cujo parecer foi aprovado, conforme ANEXO B.

4.2 Configuração de Ambiente

O ambiente de execução descrito a seguir exige recursos computacionais acessíveis e software livre ou de acesso gratuito. A exceção é o pacote MS Office, que foi utilizado neste trabalho por sua facilidade de uso, mas para o qual existem alternativas livres ou gratuitas, como LibreOffice e Google Documentos. Todos os programas possuem versões para Windows e Linux.

- Intel® Core™ i7-7500 CPU @2.70GHz 2.90GHz, 12.0GB RAM;
- Plataforma Microsoft Windows 10;

- Banco de dados PostgreSQL (9.4.24) http://postgresql.org, gratuito.
- DBeaver, Version 21.1.0.202106070736, com licença educacional gratuita e renovada anualmente;
- Rapid Miner Studio, Version 9.10.001 (ver: 9d03d3, platform: WIN64), com licença educacional gratuita e renovada anualmente;
- Pacote Office, para uso de planilhas eletrônicas nos formatos xlxs e csv, e criação de gráficos.

4.2.1 Ferramenta Rapid Miner para Aprendizado de Máquina

O Rapid Miner é um ambiente integrado de procedimentos de MD e AM, incluindo: carregamento e transformação de dados, pré-processamento e visualização de dados, análise preditiva e modelagem estatística, avaliação, validação, otimização e implantação. A ferramenta é usada para aplicações comerciais e industriais, bem como para a pesquisa, a educação, a formação, a prototipagem rápida e desenvolvimento de aplicativos (JALLOULE *et al.*, 2014).

O Rapid Miner contém operadores "*drag and drop*" para todas as tarefas de análise de dados, desde o particionamento de dados até a análise baseada no mercado. Dentro da ferramenta é configurada a conexão com o banco de dados Postgres do Moodle, e após a conexão bem-sucedida, pode-se realizar *queries*[25] (APÊNDICE H). Os resultados podem ser compostos em um *dataset* (conjunto de dados) na forma de uma tabela atributo-valor, que pode ser reutilizada em outros modelos ou experimentos.

Além disso, foi utilizada a extensão Auto Model do Rapid Miner. O Auto Model é uma funcionalidade simples de Auto ML (HE *et al.*, 2021) que testa diferentes combinações de modelos e hiperparâmetros, utilizando dados de validação para estimar o desempenho desses modelos. O Auto Model seleciona aqueles que obtiveram o menor erro de validação para inferir nos dados de teste.

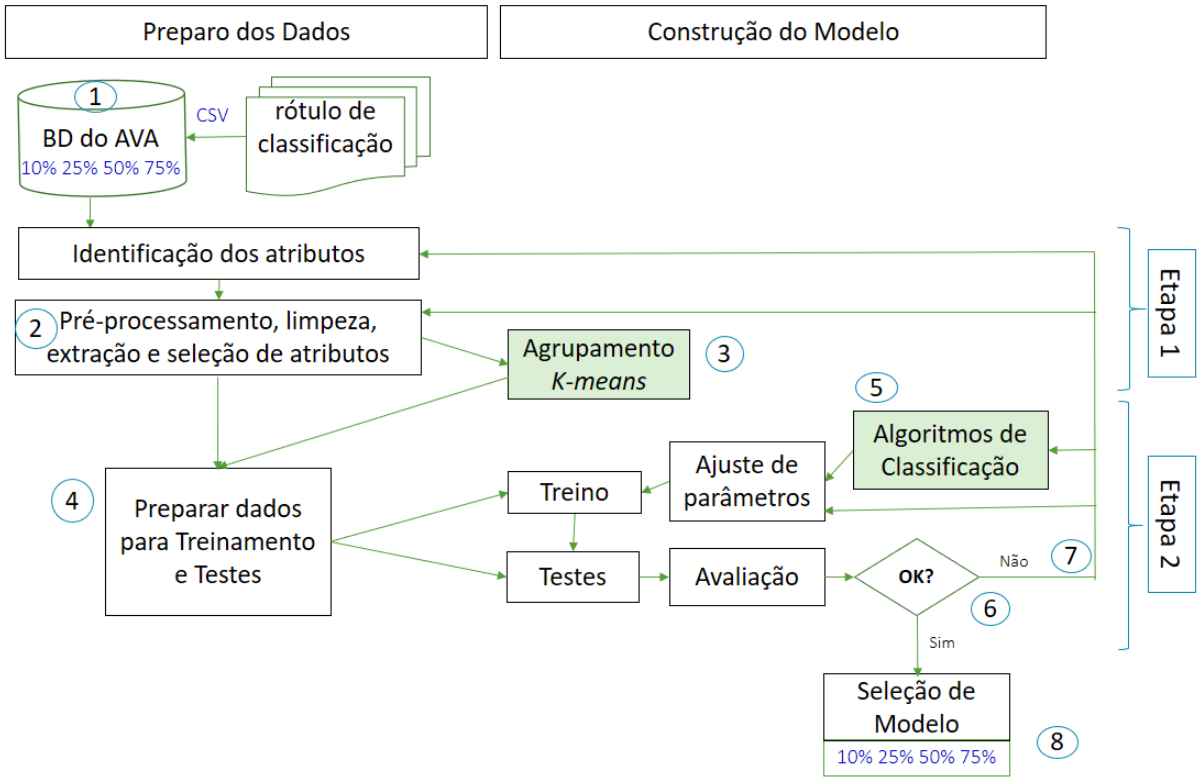
4.3 Modelo de Predição

Cada acesso e todas as ações dos usuários no Moodle são monitorados e armazenados em uma tabela de log no banco de dados, gerando um grande volume de dados sobre navegação no sistema, visualização, *upload/download*, participação em chat, fóruns, questionários e

[25] linguagem de consulta e manipulação no banco de dados.

outros. Esses dados são agrupados por alunos, processados e transformados em atributos. A esses dados, são acrescentados atributos sobre dados pessoais (idade, gênero, estado civil etc), dados socioeconômicos (cidade, renda familiar, membros na família etc) e o rótulo de classificação, com importação para a base de teste do Moodle de arquivo em formato .CSV, via comandos SQL (APÊNDICE E).

Figura 18 - Preparação, treino e construção do modelo de predição.



A Figura 18 apresenta a metodologia adotada nesta pesquisa, que foi elaborada e conduzida em duas etapas com os seguintes passos:

Etapa 1: 1) integrar os dados rotulados no banco de dados do Moodle, analisar os registros de logs e organizar a coleta de dados em diferentes momentos (10%, 25%, 50%, 75% da conclusão do curso), a fim de validar a proposta em diferentes cenários de tempo já concluído do curso, conforme explicação na seção de amostragem; 2) preparação dos dados, limpeza e seleção dos atributos; 3) construção de modelos de aprendizado de máquina com técnicas de agrupamento K-*means*.;

Etapa 2: 4) utilização dos grupos resultantes como atributo agregado aos dados de entrada no modelo de classificação; 5) seleção dos algoritmos de classificação para treino e teste; 6) após os testes analisar o desempenho das métricas do modelo, analisar os resultados no contexto e verificar possíveis ajustes voltando às etapas iniciais ou intermediárias; 7) caso

haja ajustes, pode ser em qualquer etapa anterior desta metodologia, desde uma nova seleção de atributos, voltando à etapa 1, até outras técnicas nos algoritmos de classificação; 8) caso não haja ajustes, o modelo para predizer o risco de abandono nos 4 ciclos (10%, 25%, 50%, 75%) é finalizado e encerra-se a Etapa2.

Todo o processo é realizado em quatro cenários diferentes. Cada cenário é construído coletando-se os dados do log do Moodle de um intervalo distinto, que sempre coincide com o início do curso, mas termina em momentos diferentes. O primeiro corresponde às primeiras 3 ou 4 disciplinas do curso, que na matriz curricular corresponde às Etapas 1 e 2 do Módulo 1. Isso representa, dependendo do curso, de 8 a 12% da duração do curso (considerou-se um valor médio e aproximado de 10% para facilitar a sua referência). Esse cenário é identificado como S* neste trabalho, e. Os demais períodos analisados correspondem a 25%, 50% e 75% da duração do curso, que ao ser um curso de 2 anos de duração, equivalem ao primeiro, segundo e terceiro semestres, respectivamente. Portanto, definiu-se uma nomenclatura para cada ciclo ou período para facilitar a sua referência durante o trabalho, chamando-as de S*, S1, S2 e S3.

4.4 Coleta de Dados

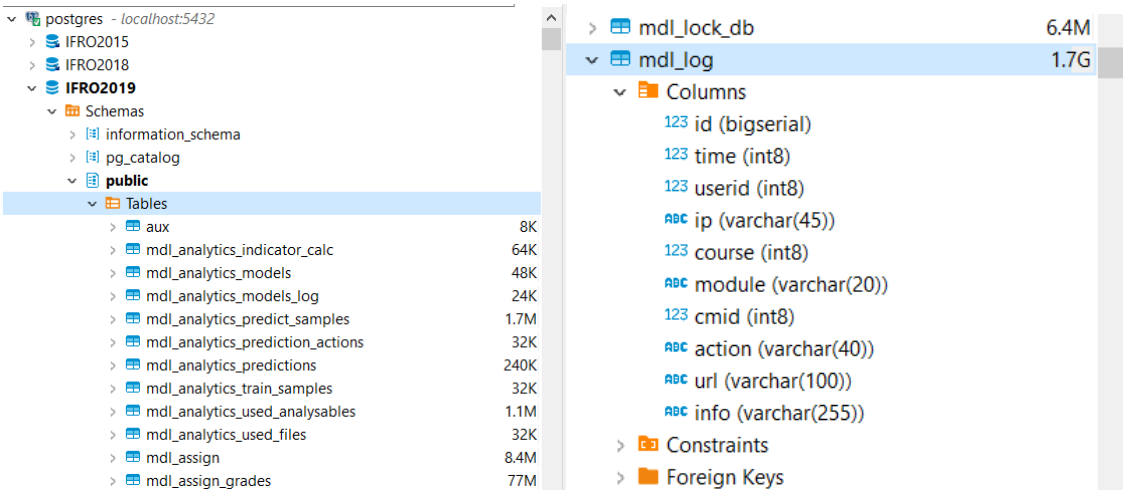
A figura 18 mostra que as fontes dos dados secundários deste trabalho são o Moodle e uma base integrada via arquivo em formato CSV. O Moodle armazena todos os materiais disponibilizados pelo docente, videoaulas, atividades, avaliações, notas, trocas de mensagens em chat e fóruns, log de navegação, encontros remotos etc. Por outro lado, as informações sobre a conclusão ou não do curso (rótulo de classificação no método supervisionado) estavam, no momento da pesquisa, na base de dados da CRA (Coordenação de Registro Acadêmico).

4.4.1 Fonte de Dados

O banco de dados utilizado pelo Moodle é o Postgres. Neste trabalho foi restaurado um *backup* de julho de 2019, quando era utilizado o Moodle versão 2.6.11. Um problema encontrado foi a falta de documentação técnica de apoio sobre as descrições da estrutura e dicionário de dados do banco de dados naquela versão.

O prefixo padrão para nomenclatura das tabelas é mdl_ (Figura 19a) e a tabela ‘mdl_log’ (Figura 19b) é a tabela para pesquisa dos dados históricos. O Quadro 7 mostra o dicionário de dados dessa tabela, e os exemplos de registros estão na Figura 20.

Figura 19 - Captura de tela da base de dados: (a) tabelas; (b) colunas da mdl_log.



(a) (b)

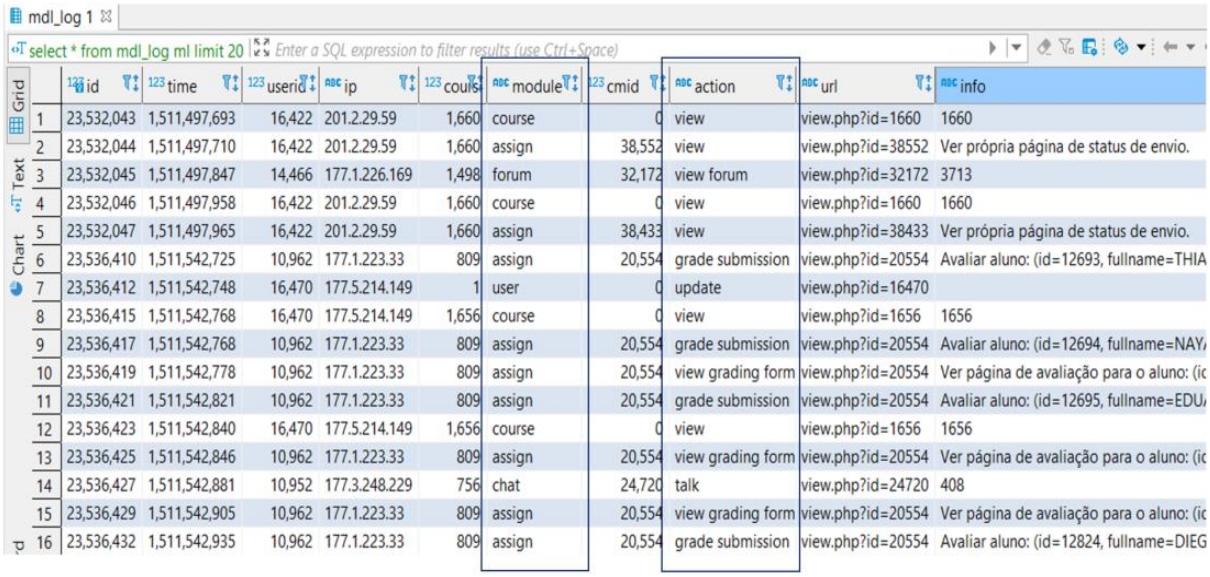
Quadro 7 - Dicionário de Dados da tabela mdl_log.

Atributo	Tipo	Descrição	Referência	Exemplo
id	bigint (10)	Identificador do registro		23532043
time	bigint (10)	*Timestamp* da inclusão do registro.		1511497693
userid	bigint (10)	id do usuário que gerou o log	mdl_user	16422
IP	varchar(45) utf8_bin	Endereço IP do equipamento		201.2.29.59
course	bigint (10)	Código da disciplina é um número único que identifica curso, turma e disciplina	mdl_course	1660
module	varchar(20) utf8_bin	O AVA é separado por módulos como fórum, questionário, tarefas, materiais e outros	Junto com o atributo info determina o acesso tabela de fórum (mdl_forum), questionário (mdl_quiz), e outros	*referência no Anexo D com os 33 tipos de módulos. Exemplo: forum, assign, course, quiz etc
cmid	bigint (10)			38552
action	varchar(40)	Ações relacionadas ao atributo module, como visualização, acesso, envio		*referência no Anexo E com os 131 tipos de ações que são registradas

url	varchar(100)	url que gerou o log		view.php?id=38552
info	varchar(255)	Informação adicional		

A Figura 20 é um exemplo com valores nas colunas "*module*" e "*action*" da tabela "mdl_log". A coluna "*action*" pode registrar 135 tipos de ações (ANEXO E), que dependem dos valores atribuídos à coluna "module" (ANEXO D).

Figura 20 - Captura de tela com destaque nos 2 atributos principais da tabela de log.



	id	time	userid	ip	coul	module	cmid	action	url	info
1	23,532,043	1,511,497,693	16,422	201.2.29.59	1,660	course	0	view	view.php?id=1660	1660
2	23,532,044	1,511,497,710	16,422	201.2.29.59	1,660	assign	38,552	view	view.php?id=38552	Ver própria página de status de envio.
3	23,532,045	1,511,497,847	14,466	177.1.226.169	1,498	forum	32,172	view forum	view.php?id=32172	3713
4	23,532,046	1,511,497,958	16,422	201.2.29.59	1,660	course	0	view	view.php?id=1660	1660
5	23,532,047	1,511,497,965	16,422	201.2.29.59	1,660	assign	38,433	view	view.php?id=38433	Ver própria página de status de envio.
6	23,536,410	1,511,542,725	10,962	177.1.223.33	809	assign	20,554	grade submission	view.php?id=20554	Avaliar aluno: (id=12693, fullname=THIA
7	23,536,412	1,511,542,748	16,470	177.5.214.149	1	user	0	update	view.php?id=16470	
8	23,536,415	1,511,542,768	16,470	177.5.214.149	1,656	course	0	view	view.php?id=1656	1656
9	23,536,417	1,511,542,768	10,962	177.1.223.33	809	assign	20,554	grade submission	view.php?id=20554	Avaliar aluno: (id=12694, fullname=NAY
10	23,536,419	1,511,542,778	10,962	177.1.223.33	809	assign	20,554	view grading form	view.php?id=20554	Ver página de avaliação para o aluno: (id
11	23,536,421	1,511,542,821	10,962	177.1.223.33	809	assign	20,554	grade submission	view.php?id=20554	Avaliar aluno: (id=12695, fullname=EDU
12	23,536,423	1,511,542,840	16,470	177.5.214.149	1,656	course	0	view	view.php?id=1656	1656
13	23,536,425	1,511,542,846	10,962	177.1.223.33	809	assign	20,554	view grading form	view.php?id=20554	Ver página de avaliação para o aluno: (id
14	23,536,427	1,511,542,881	10,952	177.3.248.229	756	chat	24,720	talk	view.php?id=24720	408
15	23,536,429	1,511,542,905	10,962	177.1.223.33	809	assign	20,554	view grading form	view.php?id=20554	Ver página de avaliação para o aluno: (id
16	23,536,432	1,511,542,935	10,962	177.1.223.33	809	assign	20,554	grade submission	view.php?id=20554	Avaliar aluno: (id=12824, fullname=DIEG

4.4.2 Dados Integrados

As atividades avaliativas têm as suas notas disponibilizadas na base de dados do Moodle. Entretanto, o histórico escolar e o status final, que informa conclusão ou não do curso, estão na base de dados da CRA (Coordenação de Registros Acadêmicos). Essas informações foram disponibilizadas em arquivo .CSV. Nesse arquivo (Figura 21) há os dados do cadastro do estudante (nome completo, CPF, RG, e-mail etc.) e dados sobre o status do aluno. Os dados do arquivo e os dados do Moodle foram integrados utilizando como chave única o CPF de cada discente.

Figura 21 - Planilha eletrônica (parcial) com status final.

B	C	D	E	F	G
STATUS	CERT	ESTÁGIO	SEX	DATA DE NASCIMEN	CPF
Evadido	-	-	o	22/10/2001	048. .332-08
Aprovado	Não	Não	a	19/01/2002	817. .572-91
Aprovado	Não	Não	o	15/08/2002	043. .992-08
Reprovado	-	-	o	18/03/2003	039. .912-00
Concluído	Sim	Sim	a	21/04/2001	032.. .382-44
Desistente	-	-	a	04/02/2002	052. .962-96
Evadido	-	-	o	03/05/2000	066. .072-97
Reprovado	-	-	o	30/08/2002	043. .042-10

4.5 Pré-processamento de dados

As técnicas de pré-processamento de dados são utilizadas para melhorar a qualidade dos dados por meio da eliminação ou minimização de problemas. Os valores dos atributos de um conjunto dados podem ser numéricos ou simbólicos e certos modelos podem demandar que os valores sejam convertidos em um ou outro tipo. Os dados podem conter ruídos e imperfeições, valores incorretos, inconsistentes, duplicados ou ausentes, precisando ser substituídos ou removidos. E os conjuntos de dados podem apresentar poucos ou muitas instâncias, que por sua vez podem ter poucos ou muitos atributos, casos nos quais pode ser necessário coletar ou gerar artificialmente novos exemplos e eliminar ou elaborar novos atributos (FACELI *et al.*, 2017).

As técnicas de pré-processamento são frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso de algoritmos de AM. Essas técnicas podem ser agrupadas nos seguintes grupos de tarefas: eliminação manual de atributos, integração de dados, amostragem de dados, balanceamento de dados, limpeza de dados, redução de dimensionalidade e transformação de dados (FACELI *et al.*, 2017). As técnicas utilizadas neste trabalho são discutidas nas próximas subseções

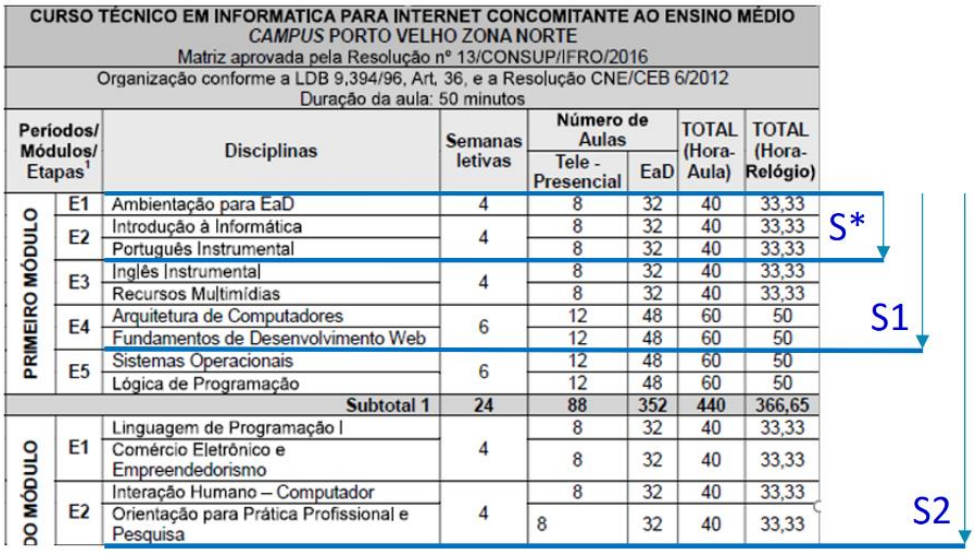
4.5.1 Amostragem de Dados

Analisaram-se três cursos técnicos concomitantes: Informática para Internet, Administração e Finanças. Esses cursos estão divididos em 3 módulos (Primeiro, Segundo e Terceiro Módulo), a serem concluídos de forma regular em 2 anos. Cada módulo é composto por 3 a 5 etapas, denominadas E1, E2, E3, E4 e E5 (Figura 22).

Cada etapa tem duas disciplinas, cursadas concomitantemente. Excepcionalmente, a disciplina “Ambientação para EaD”, uma disciplina introdutória e obrigatória no início de todos os cursos EaD, pode acontecer de forma avulsa na E1 do Primeiro Módulo (Figura 22). Nos

cursos de Informática para Internet e Finanças, que têm carga horária de 1.000 horas, a E1 tem essa disciplina avulsa. Já no curso de Administração, cuja carga horária total é de 1.100 horas (Figura 23), a etapa E1 inicia com duas disciplinas concomitantes.

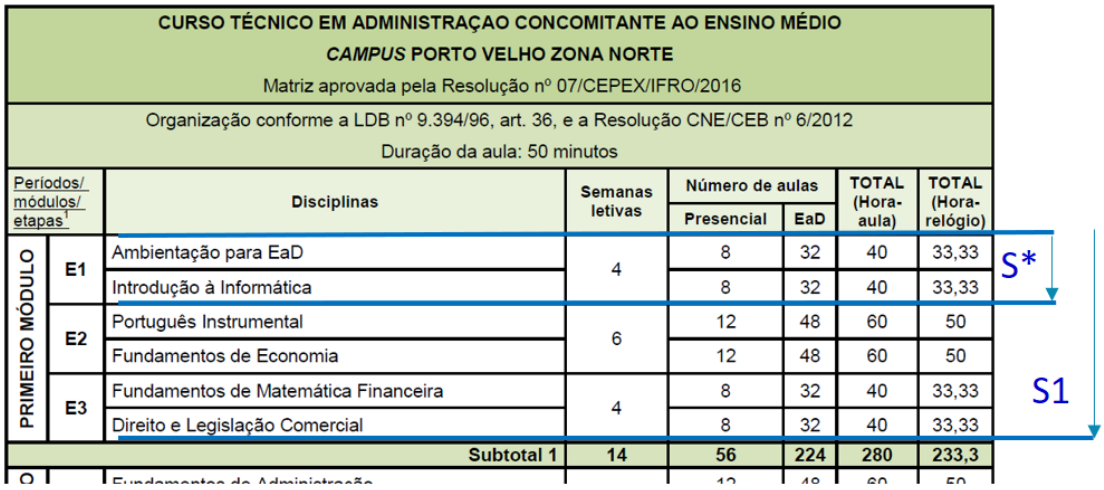
Figura 22 - Primeiro módulo do Curso Técnico em Informática para Internet Concomitante ao Ensino Médio.



CURSO TÉCNICO EM INFORMATICA PARA INTERNET CONCOMITANTE AO ENSINO MÉDIO
***CAMPUS* PORTO VELHO ZONA NORTE**
Matriz aprovada pela Resolução nº 13/CONSUP/IFRO/2016
Organização conforme a LDB 9,394/96, Art, 36, e a Resolução CNE/CEB 6/2012
Duração da aula: 50 minutos

Períodos/ Módulos/ Etapas[1]		Disciplinas	Semanas letivas	Número de Aulas		TOTAL (Hora-Aula)	TOTAL (Hora-Relógio)
				Tele - Presencial	EaD		
PRIMEIRO MÓDULO	E1	Ambientação para EaD	4	8	32	40	33,33
	E2	Introdução à Informática	4	8	32	40	33,33
		Português Instrumental		8	32	40	33,33
	E3	Inglês Instrumental	4	8	32	40	33,33
		Recursos Multimídias		8	32	40	33,33
	E4	Arquitetura de Computadores	6	12	48	60	50
		Fundamentos de Desenvolvimento Web		12	48	60	50
	E5	Sistemas Operacionais	6	12	48	60	50
		Lógica de Programação		12	48	60	50
		Subtotal 1	**24**	**88**	**352**	**440**	**366,65**
DO MÓDULO	E1	Linguagem de Programação I	4	8	32	40	33,33
		Comércio Eletrônico e Empreendedorismo		8	32	40	33,33
	E2	Interação Humano – Computador	4	8	32	40	33,33
		Orientação para Prática Profissional e Pesquisa		8	32	40	33,33

Figura 23 - Primeiro módulo do Curso Técnico em Administração Concomitante ao Ensino Médio.



CURSO TÉCNICO EM ADMINISTRAÇAO CONCOMITANTE AO ENSINO MÉDIO
***CAMPUS* PORTO VELHO ZONA NORTE**
Matriz aprovada pela Resolução nº 07/CEPEX/IFRO/2016
Organização conforme a LDB nº 9.394/96, art. 36, e a Resolução CNE/CEB nº 6/2012
Duração da aula: 50 minutos

Períodos/ módulos/ etapas[1]		Disciplinas	Semanas letivas	Número de aulas		TOTAL (Hora-aula)	TOTAL (Hora-relógio)
				Presencial	EaD		
PRIMEIRO MÓDULO	E1	Ambientação para EaD	4	8	32	40	33,33
		Introdução à Informática		8	32	40	33,33
	E2	Português Instrumental	6	12	48	60	50
		Fundamentos de Economia		12	48	60	50
	E3	Fundamentos de Matemática Financeira	4	8	32	40	33,33
		Direito e Legislação Comercial		8	32	40	33,33
		Subtotal 1	**14**	**56**	**224**	**280**	**233,3**

Neste trabalho, analisaram-se estudantes de nove turmas iniciadas entre 2016 e 2018, sendo três turmas do curso de Informática para Internet, quatro turmas de Administração e duas turmas de Finanças. É importante ressaltar que a conclusão de todos os cursos ocorreu antes de dezembro de 2019, portanto, os dados não foram afetados pela pandemia de Covid-19. Cada turma tinha, inicialmente, aproximadamente 40 estudantes matriculados. Até o enceramento do período trabalhado alguns estudantes não tinham completado o período de integralização do

curso, mas neste trabalho analisou-se até os 75% do andamento regular do curso e o status final de cada estudante, rótulo de classificação, foi obtido com informação atualizada da situação final (Figura 21).

O *status* do estudante, que informa a conclusão ou não do curso assume um dos seguintes valores:

- "Aprovado": o estudante foi aprovado em todas as disciplinas.
- "Apto a se formar": o estudante foi aprovado em todas as disciplinas, completou o estágio obrigatório e está com a documentação necessária para dar entrada no processo da obtenção de diploma.
- "Concluído": o estudante se formou e recebeu um diploma.
- "Evasão": o estudante abandonou o curso (não compareceu e não confirmou desistência).
- "Reprovado": o estudante ficou reprovado em uma ou mais disciplinas, seja por nota ou frequência, até esgotar o prazo de integralização do curso.
- "Desistente": o estudante confirmou a desistência do curso.

O estudante que obtém aprovação em todas as disciplinas também precisa ter o estágio concluído para estar apto à solicitação do diploma. Se no sistema da CRA (Figura 21) o estudante tem número do processo de expedição do diploma significa que concluiu o curso, caso contrário existe uma dificuldade em definir se um estudante é evadido ou não, pois tem um prazo que transcorre até o tempo de integralização curricular do curso, na qual os estudantes contam com diversas oportunidades extras até encerrar o prazo para jubilar e ser desligado.

O problema de detecção de evasão é tratado neste trabalho como um problema de classificação binária. Como é usual em tarefas de classificação com duas classes, o caso de maior interesse para este trabalho é chamado de classe positiva (P), enquanto o outro é a classe negativa (N). Assim, a **classe positiva (P) refere-se aos estudantes que não se formaram após três anos**, prazo normal de integralização do curso. A base de dados é composta por informações de 381 estudantes, com 127 estudantes da classe N e 254 estudantes da classe P (Tabela 1).

Um resumo do sucesso/fracasso para as nove turmas é apresentado na Tabela 1. As linhas estão relacionadas aos três cursos e as colunas referem-se ao momento da coleta em 10%, 25%, 50% e 75% de conclusão do curso, ou seja, todo o procedimento de construção do modelo

(Figura 18) é repetido quatro vezes, uma para cada ciclo avaliado durante a duração do curso. Efetivamente, desenvolveram-se quatro versões dos dados de treinamento e testes, de acordo com S*, S1, S2 e S3.

Os 3 cursos concomitantes selecionados têm um total de 25 disciplinas e cada disciplina tem entre 40 e 60 horas-aula e, por isso, a quantidade de disciplinas por semestre pode variar por curso e semestre. Para ter uma estimativa, considerou-se que o momento S1 ocorreu após o estudante ter concluído entre 5 e 8 disciplinas, S2 após 12 disciplinas e S3 após, aproximadamente, 18 disciplinas.

Tabela 1- Distribuição das aulas para cada curso e momentos.

	S* (10%) S1 (25%)		**S2 (50%)**		**S3 (75%)**	
	N	P	N	P	N	P
IPI (Informática)	34	95	34	95	34	63
ADM (Administração)	65	105	65	88	65	65
FIN (Finanças)	28	54	28	54	28	47
Total	127	254	127	237	127	175
%	33,33	66,67	34,89	65,11	45.05	57.95
Total de inscritos	381		369		304	

A Tabela 1 mostra que a informação sobre a turma inteira está no S*, pois nela contém todos os alunos que se matricularam no início do curso, e cada coluna à direita mostra a diminuição de alunos à medida que avança o curso.

4.5.2 Limpeza de Dados

Os dados coletados do AVA raramente ficariam incompletos, pois o log é gerado automaticamente, porém os dados das fichas cadastrais dos estudantes podem conter informações incompletas ou inconsistentes devido a erros de digitação ou valores ausentes.

Um atributo com erro de digitação encontra-se na informação sobre a renda familiar, que por ser preenchido de forma livre contém valores de naturezas distintas, como ‘1,5’, (referente a salários mínimos), ‘1015’ (referente a um valor bruto em reais), por extenso etc. Esses ruídos e informações incompletas podem levar a uma predição imprecisa no

processamento posterior. Nesses casos exemplificados não foi feita a conversão para salários mínimos da época e, simplesmente, foram deixados nulos.

O efeito adverso de valores ausentes pode ser reduzido, por exemplo, preenchendo com o valor mais frequente (mediana/moda) ou valor médio para o recurso ausente. Neste trabalho, no caso do atributo que armazena a ‘idade’ foi adotado o valor da mediana quando encontrados valores nulos ou ausentes. Como cada instância representa informações de um estudante, essa substituição seria um dado falso, mas priorizou-se a análise e resultado do conjunto para não impactar a distância na técnica de agrupamentos.

4.5.3 Eliminação Manual de Atributos

Quando um atributo claramente não contribui para a estimativa do valor do atributo alvo, ele é considerado irrelevante e pode ser descartado. Dessa forma, decidiu-se que atributos como Nome e CPF do estudante são importantes para identificação e descrição da instância, porém desconsiderados para dados de entrada.

Outra situação em que o atributo resulta irrelevante é quando ele não contém informação que ajude a distinguir as instâncias. Isso acontece com o atributo ‘sit_familia’ (Tabela 1), um atributo categórico que identifica a participação do estudante na renda familiar com valores 1-provedor de renda; 2- dependente; 3-compõe a renda. Nas turmas analisadas, todos os alunos dessa amostra têm o mesmo valor, portanto o atributo foi removido, mas é importante mencionar que um atributo não precisa ter exatamente o mesmo valor para todas as instâncias para ser considerado irrelevante.

4.5.4 Transformação de Atributos

Foi feita a normalização dos atributos, necessária na entrada de dados para métodos de agrupamento. A seguir é detalhado o tratamento e seleção dos atributos como passo final antes de executar os algoritmos de AM.

Foram selecionados 14 atributos para o modelo de AM, dos quais 7 são dados institucionais e 7 são dados de rastreamento.

Tabela 2 - Atributos de Dados Institucionais.

Dados institucionais			
código	**nome**	**tipo**	**Descrição**
idade	continuo	atributo calculado	14 a 19 anos
gênero	categórico		
renda_familiar	continuo	renda familiar em salários mínimos	
qtd_membros_familia	continuo	número de pessoas na família	
sit_trabalho	categórico	situação de trabalho familiar	1-Desempregado, 2-Profissional liberal, 3-Empregado, 4-Aposentado, 5-Empresário, 6-Autônomo, 7-Cooperado, 8-Estudante
sit_familia	categórico	participação na renda familiar	1 - Provedor de Renda; 2 - Dependente; 3 - Compõe a renda
status	categórico	estado final	"aprovado", "concluido", "apto a graduar", "evadido", "reprovado", "desistente"

Os dados institucionais idade, gênero, renda_familiar, qtd_membros_familia, sit_trabalho e sit_familia são coletados dos estudantes na admissão e o status é um dado na sua saída do curso (Tabela 2). Esses primeiros 6 atributos são dados socioeconômicos extraídos da ficha do aluno armazenados no Moodle.

Tabela 3 – Atributos de Dados de Rastreamento.

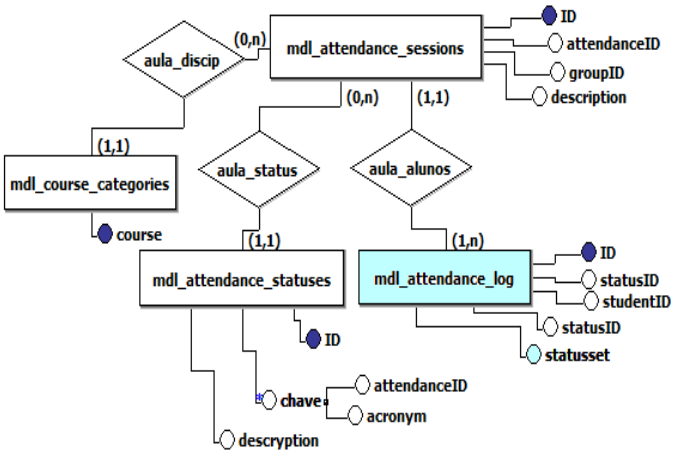
Dados de rastreamento			
nome	tipo	Descrição	
FREQ	continuo	Frequencia do aluno	0-100
tGPA	continuo	média das notas finais das disciplinas pelo tempo de previsão definido na duração do curso.	0-100
pGPA	continuo	média das notas das atividades submetidas. São necessárias pelo menos 2 atividades de cursos por disciplina	0-100
VIS	continuo	porcentagem de visualização de módulos, tarefas, fóruns, quizzes, etc.	0-100
QST	continuo	porcentagem de visualização, acesso, tentativas de resposta e resposta a questionários	0-100
UAA	contínuo	porcentagem de acesso a materiais externos e tarefas enviadas	0-100
CIR	contínuo	porcentagem de acesso ao curso e recursos/materiais disponibilizados pelo docente	0-100

Os dados de rastreamento (Tabela 3) são extraídos e transformados a partir do log do Moodle. FREQ, tGPA, pGPAsão atributos calculados com base em colunas de diversas tabelas, e VIS, QST, UAA e CIR são atributos agregados ou transformados, que juntam atributos para formar um novo. Segue a descrição sobre a obtenção de cada atributo:

a) FREQ é a frequência total por estudante e disciplina. Essa informação deve ser extraída e calculada com base em 4 tabelas do banco de dados (mdl_course_categories, mdl_attendance_sessions, mdl-attendance_statuses, mdl_attendance_log) e a mdl_attendance_log.statusset armazena os logs das frequências. A Figura 24 mostra

um modelo conceitual com as tabelas utilizadas para compor as anotações de frequência para este trabalho.

Figura 24 - Modelo conceitual das tabelas relacionadas a frequência.



Para calcular a taxa de frequência, adotou-se como referência o valor máximo de frequências encontrado para a semana, por turma e disciplina, pois dependendo da semana pode haver quantidade de aulas diferentes, seja por feriados, recessos ou reposição. Uma falta justificada não é contabilizada como falta no histórico escolar do estudante, porém neste trabalho foi deixado como ausência.

Foi calculada a taxa de participação em referência ao máximo da turma e disciplina. Esse atributo foi normalizado considerando a normalização dos dados min-max, usando uma escala que vai de 0,0 para o menor valor e 1,0 para o maior valor e, neste trabalho, foi considerada uma escala de 0 a 100.

$$normalized = 100 \times \frac{x - min(x)}{max(x) - min(x)}$$

b) Os atributos tGPA e pGPA são calculados a partir das notas do estudante. O atributo tGPA é a média acumulada das notas finais das disciplinas e pGPA é a nota média acumulada das atividades submetidas. Para tGPA a nota máxima é 100 e o estudante é aprovado com 60 pontos e frequência mínima de 75%. No caso da pGPA cada atividade tem uma nota máxima que o docente determina e registra no Moodle. Portanto, foi necessário calcular a taxa da nota obtida sobre o máximo possível da atividade.

c) Os atributos VIS, QST, UAA e CIR são relacionados às atividades registradas no log (ANEXO D) e as ações (ANEXO E). Dentre as mais de 100 ações previstas para

registro consideraram-se 9 delas, por possuírem os maiores números de registros no log, ou seja, pela contagem da quantidade de registros por módulo e ação, que são: visualização de URL, visualização de atribuição de tarefa, visualização de curso, visualização de recursos, visualização de questionário, tentativa de resposta de questionário, envio das respostas do questionário, visualização de fórum e submissão ou envio de tarefa.

Essas 9 ações foram contabilizadas (1 por registro encontrado no log), totalizadas por ações, estudante e disciplina em cada período analisado, e agrupadas em quatro novos atributos: VIS abrange as visualizações de todos os módulos (*URLview, assignview, courseview, resourceview, quizview, forumview*); QST está relacionado apenas às atividades com questionários, que considera as ações visualizar (*quizview)*, tentativas de resposta (*quizattempt*) e envio para avaliação (*quizclose*); o atributo UAA representa o acesso ao material externo, como vídeos (*URLview)*, mais as tarefas, com visualização (*assignview)* e tarefas concluídas para avaliação (*assignsubmit*); o atributo CIR é o acesso ao curso (*courseview*) e aos materiais fornecidos pelo docente, como fazer download de apresentações (*resourceview*).

Para cada atributo, calculou-se o percentual de participação de cada estudante em cada ação sobre o máximo registrado por cada turma e disciplina nos momentos S*, S1, S2 e S3, pois os valores absolutos podem variar dependendo da disciplina, turma e docente que propõe as atividades. Ao calcular as taxas, os valores ficam normalizados entre 0 e 100.

Embora o Moodle tenha mais de dez tipos de recursos (fórum, *quizzes*, tarefa, wiki, URL, aula, livro, lição e outros) para diversificar as atividades, os professores preferem utilizar o questionário como atividade avaliativa, o que indica a importância deste atributo. Em relação ao número de registros, os questionários respondem por mais da metade das atividades propostas no AVA.

4.5.5 Redução e Seleção de Dados

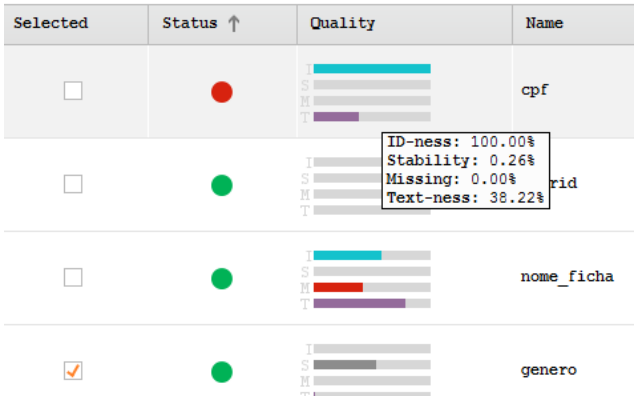
Em MD, de acordo com Bellman (1961), o número de classificadores que devem ser considerados aumenta exponencialmente com o número de atributos do conjunto de dados, ficando mais difícil para o algoritmo de AM encontrar um modelo adequado durante a fase de treinamento. Uma das formas de evitar esse problema é efetuando a redução do número de atributos por meio de técnicas de seleção de atributos ou transformação de atributos.

Neste trabalho, foi realizada seleção de atributos a fim de remover aqueles que não agregam informação útil ao modelo. Analisou-se a qualidade da informação baseada em correlação de um atributo com outros, *ID_ness*, estabilidade, falta de dados e *text-ness*. O significado desses critérios é o seguinte.

- Correlação: Fator de correlação linear entre o atributo e a classe;
- ID-ness: Propriedade na qual um atributo tende a ter valores únicos para todos os exemplos (ou poucas ocorrências de um mesmo valor), comportando-se como um ID;
- Estabilidade: Propriedade na qual um atributo tende a ter o mesmo valor para todos os exemplos (ou muitas ocorrências de um mesmo valor);
- Falta de dados: Colunas com excesso de valores ausentes;
- *Text-ness:* Propriedade de que um atributo é composto por "texto livre" (e.g., nome) em vez de ser um atributo categórico, numérico etc.

Para o cálculo desses critérios, foi usada a ferramenta Rapid Miner, conforme ilustrado na Figura 25. Nesse exemplo, mostra-se que o atributo CPF possui 100% de *ID_ness* e um valor bastante elevado de *Text-ness*. Isso é compatível com um atributo proveniente de uma chave única que não é um valor numérico. Por esses motivos, o Rapid Miner sugere que esse atributo é pobre em informação marcando-o em vermelho na coluna "Status" (Figura 25).

Figura 25 - Tela de seleção de atributos no Rapid Miner.



Em geral, em AM, os atributos devem ter valores baixos para dados ausentes, estabilidade e ID-ness. Atributos com altas correlações geralmente são preferidos, mas não se a alta correlação for devido a uma relação direta de causa e efeito com o atributo preditivo (classe). Com auxílio da ferramenta, os atributos foram analisados caso a caso.

O CPF foi mantido na base para permitir estudar a evolução dos grupos gerados em S*, S1, S2 e S3. Entretanto, ele não foi utilizado no treinamento dos modelos, pois é um identificador e não tem qualquer relação com o desempenho de um estudante.

Os atributos renda_familiar, sit_familia, qtd_membros_familia e sit_trabalho apresentaram 42,26% de ausência de dados (Figura 26). O atributo renda_familiar também apresentou os erros de digitação abordados na Seção 4.5.2. No caso do atributo sit_familia foi encontrado só o valor 2 (Dependente), quando não nulo, e no sit_trabalho foi encontrado, na sua maioria, o valor 8 (Estudante), quando não nulo e, por isso, ambos apresentam estabilidade próxima a 90%. O atributo idade possui alta estabilidade (são alunos entre 15 e 19 anos) e baixa correlação com a classe. Com base nessa análise, os atributos mencionados foram descartados. Os demais atributos, considerados os conjuntos de valores para gerar os modelos em AM (Figura 27), foram mantidos. Estes são os 7 atributos FREQ, tGPA, pGPA, VIS, QST, UAA e CIR.

Figura 26 - Qualidade dos dados para seleção de atributos.

Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness
	genero	0.86%	0.52%	53.28%	0.00%	0.62%
	idade	3.48%	1.57%	43.80%	0.52%	0.00%
	renda	1.46%	?	55.00%	42.26%	0.00%
	sit_familia	0.02%	0.79%	89.55%	42.26%	0.00%
	qtd_familia	0.39%	2.10%	29.09%	42.26%	0.00%
	sit_trabalho	0.01%	0.79%	88.64%	42.26%	0.00%

Figura 27- Qualidade dos dados dos atributos selecionados.

Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness
	freq	14.23%	?	22.83%	0.00%	0.00%
	tgpa	23.81%	?	2.37%	0.26%	0.00%
	pgpa	28.60%	?	2.37%	0.26%	0.00%
	uaa	7.93%	?	1.62%	2.62%	0.00%
	qst	13.31%	?	2.62%	0.00%	0.00%
	vis	13.02%	?	1.08%	2.36%	0.00%
	cir	11.82%	?	3.15%	0.00%	0.00%

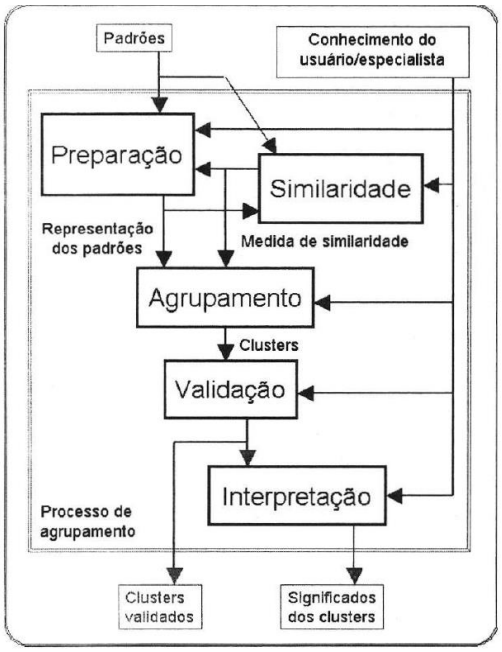
4.6 Primeira Etapa: Agrupamento

Clustering ou agrupamento é uma tarefa de AM não supervisionado. O objetivo do agrupamento é justamente encontrar estruturas de grupos nos exemplos. Uma boa estrutura de grupos é aquela na qual todos os exemplos de um grupo são bastante próximos aos demais exemplos do seu próprio grupo e distantes dos exemplos dos outros grupos.

Normalmente, essa proximidade entre os exemplos é calculada por meio de uma função de distância ou uma função de similaridade, tais como a distância euclidiana e a similaridade cosseno. O agrupamento pode ser muito útil para estudar os dados e obter *insights* sobre o domínio. Apesar de ser uma técnica não supervisionada, o conhecimento de um especialista no domínio dos dados pode ser necessário nas várias etapas do processo de agrupamento, utilizando a capacidade interpretativa de dados para diagnosticar acerca das observações que estejam presentes nas amostras e interpretar os grupos gerados.

A sequência de passos típica do processo de agrupamento é mostrada na Figura 28. Resumidamente, a figura mostra que existem padrões “escondidos” nos dados e que o objetivo do agrupamento é tornar esses padrões explícitos. Assim como no AM supervisionado, é necessário realizar preparação dos dados, eliminando atributos com pouca informação, baixa qualidade etc. A validação pode ser subjetiva ou com métricas, como o coeficiente de silhueta (FACELI *et al*., 2017).

Figura 28 - Etapas do processo de agrupamento.



Fonte: Faceli *et al*. (2017, p.197)

Neste trabalho, o pré-processamento dos dados, para a tarefa de agrupamento, foi o mesmo empregado na tarefa de classificação, como apresentado na Seção 4.5. A proximidade entre os exemplos foi calculada através da **distância Euclidiana**. É importante normalizar os dados ou escolher uma proximidade que não dependa da magnitude dos atributos, para um atributo não ser dominante sobre o outro (FACELI *et al*. 2017) e todos contribuam de maneira aproximadamente igual ao calcular a medida de proximidade. Dentre as diferentes técnicas de normalização, como Min-Max, Z-Score, Tanh e Soma, adotou-se a primeira. O algoritmo de agrupamento escolhido para este trabalho foi o **K-means**. O K-means é um algoritmo de agrupamento baseado em centroides. Primeiro, são escolhidos *k* centroides. Em seguida, cada exemplo é atribuído a exatamente um grupo, de acordo com o centroide mais próximo. Em seguida, a posição do centroide é atualizada de acordo com os exemplos do grupo. Esses passos são repetidos até convergir.

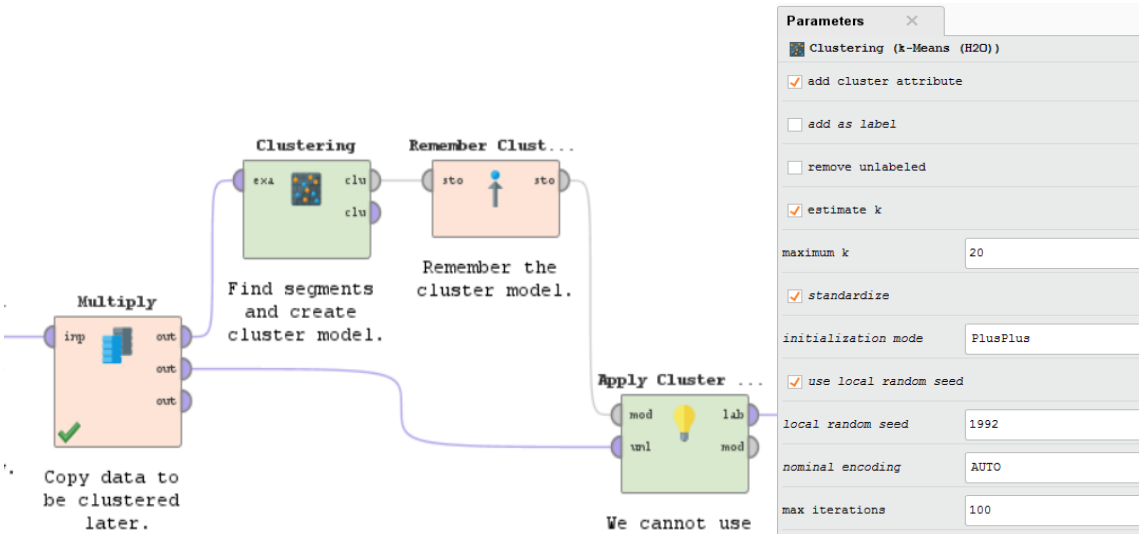
Neste trabalho, o número de grupos *k* não é fixo, pois entende-se que em contexto dinâmico de curso em andamento, podem ser encontrados diferentes grupos para acompanhar o dinamismo do contexto educacional. Em vez disso, utilizou-se a ferramenta X-means do Rapid Miner para encontrar um valor apropriado de *k* utilizando um conjunto de validação.

A classe (situação final do aluno) é desconsiderada para este experimento. No entanto, ela foi usada posteriormente para verificar a qualidade dos grupos encontrados.

4.6.1 Operador X-means

O Operador X-means é a uma ferramenta de Auto ML para agrupamento. Ela procura o melhor número de grupos *k* em um intervalo pré-definido usando um conjunto de validação. Neste trabalho, o X-means procurou o melhor valor de *k* entre 2 e 20 (Figura 29). É relevante dizer que diferentes modelos de agrupamento foram obtidos nos momentos S*, S1, S2 e S3. Além disso, o número ideal de grupos não foi o mesmo para todos eles.

Figura 29 - Operador Clustering no X-means e seus parâmetros.



Este operador realiza o agrupamento usando o algoritmo K-means H2O[26]. As posições iniciais dos centroides são determinadas por alguma heurística:

- Aleatório: escolhe k exemplos aleatórios.
- PlusPlus: Escolhe um centro inicial aleatoriamente e pondera a seleção aleatória de centros subsequentes para que os pontos mais distantes do primeiro centro tenham maior probabilidade de serem escolhidos.
- Mais distante: Escolhe aleatoriamente um centro inicial e, em seguida, escolhe o próximo centro para ser o ponto mais distante em termos de distância euclidiana.

---

[26] H2O usa redução proporcional no erro (PRE) para determinar quando parar de dividir. O PRE é calculado com base na soma dos quadrados dentro de (SSW).

$$PRE = (SSW[\text{before split}] - SSW[\text{after split}]) SSW[\text{before split}]$$

H2O para de se dividir quando PRE cai abaixo de um *threshold*, que é uma função do número de variáveis e do número de casos, e assume omenor desses dois valores descrito abaixo: 0,8 ou

$$0.02 + 10 number\_of\_training\_rows + 2.5 number\_of\_model\_features$$

Fonte: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/k-means.html

Neste trabalho o ponto inicial foi determinado pelo PlusPlus (Figura 29).

O resultado do método de agrupamento é um conjunto de grupos que pode ser usado para estudar os dados. Neste trabalho, o número do grupo ao qual cada exemplo pertence foi utilizado na etapa seguinte, de classificação, como um atributo transformado e descritivo.

4.7 Segunda Etapa: Classificação

Como explicado na Seção 4.5, o problema de detecção de evasão é tratado neste trabalho como um problema de classificação binária. Os atributos extraídos a partir dos logs do Moodle e de dados socioeconômicos, junto com a informação de grupoobtida na primeira etapa, são usados para treinar um modelo que prediz se o estudante irá abandonar o curso ou não.

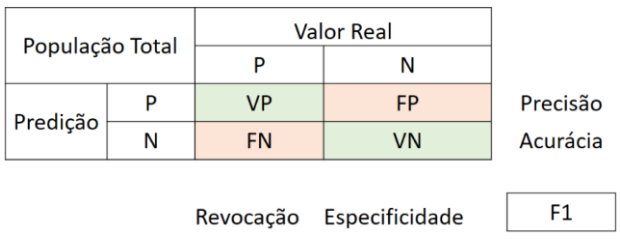
Os algoritmos de classificação com parâmetros padrão executados são: floresta aleatória (RF: random forest), árvore de decisão (DT: decisão tree), árvores com gradiente (GBT: gradiente boost tree), regressão logística (LR: logistic regression), Naive Bayes (NB) e máquina de vetores de suporte (SVM: support vector machine).

Para avaliar os modelos, foi considerada a matriz de confusão para classificação binária. Dada uma instância de teste, cuja classe verdadeira pode ser positiva ou negativa, a saída do classificador pode ser categorizada como:

- Verdadeiro positivo (VP): uma instância positiva é classificada corretamente;
- Falso negativo (FN): uma instância positiva é classificada incorretamente;
- Verdadeiro negativo (VN): uma instância negativa é classificada corretamente;
- Falso positivo (FP): uma instância negativa é classificada incorretamente.

Dado um conjunto de teste, o número de ocorrências de VP, FN, VN e F normalmente é representado em uma matriz de Confusão 2 x 2. Essa matriz (Figura 30) forma a base para muitas métricas comuns de avaliação para os classificadores binários.

Figura 30 - Matriz de Confusão para Classificadores.



Neste trabalho, dentre as várias medidas para avaliar os testes, utilizou-se Precisão, Revocação e F1-score.

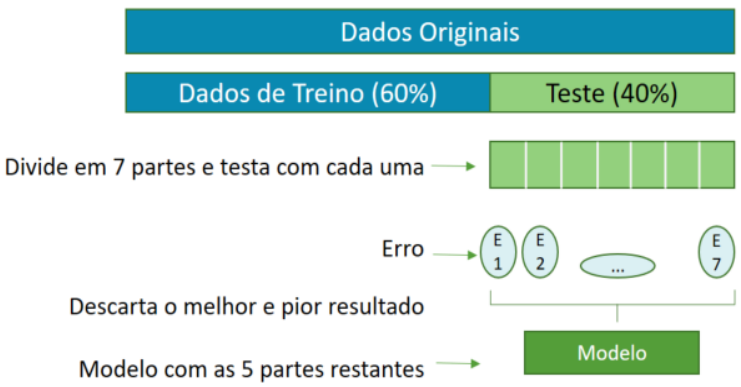
- Acurácia (Accuracy): Representa a taxa de acerto geral do classificador e é obtida pela expressão: Acurácia = (VP + VN) / (VP + VN + FP + FN).
- Precisão (Precisão): Representa a preditividade positiva, que é o percentual de acertos de verdadeiros positivos dentre todos os exemplos classificados como positivos. Sua expressão é: Precisão = VP / (VP + FP). Quanto maior a precisão, menor o erro de falsos positivos cometidos pelo classificador.
- Revocação ou Sensibilidade (Recall): Indica a taxa de verdadeiros positivos, ou seja, o percentual de VP previstos corretamente. É obtida por Revocação = VP / (VP + FN). Uma Revocação alta indica que o classificador produziu poucos exemplos positivos classificados como falso negativos.
- F1 ou F-measure ou F-score: a média harmônica ponderada de Precisão e Revocação.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Para testar os modelos foi utilizado validação *multi hold-out*. Esse procedimento é semelhante ao *hold-out*, no qual a amostra é dividida em um conjunto de treino e um conjunto de teste. Entretanto, no *multi-hold-out* no Auto ML, o conjunto de teste é dividido em 7 sub-conjuntos disjuntos. O modelo é testado nos 7 subconjuntos, mas o melhor e o pior resultado são descartados.

O *multi-hold-out* foi executado com proporção 60-40 (isto é, 60% dos exemplos usados para treinamento e o restante para teste). O *multi-hold-out* (Figura 31) é diferente da validação cruzada porque apenas um modelo é treinado. Também é diferente do *hold-out* porque tenta desconsiderar "coincidências estatísticas" quando descarta o melhor e o pior resultados.

Figura 31 - Validação *multi-hould-out.*



Neste trabalho foi utilizada a métrica F para comparar os desempenhos dos modelos, e também se deu atenção aos falsos negativos, que correspondem aos estudantes que foram erroneamente classificados como concluídos com sucesso quando não eram e, de certa forma, não foi detectado como classe positiva para poder receber algum tipo de intervenção ou acompanhamento.

4.8 Considerações Finais sobre o Capítulo

Neste capítulo a Figura 18 apresenta os passos do método para predição de estudantes em risco de não concluir o curso, que consiste na coleta de dados, pré-processamento dos dados, transformação e seleção de atributos para dar entrada na construção do modelo usando técnicas de agrupamento, e ao obter os grupos dos estudantes pelas interações no AVA, esse resultado é utilizado como entrada na construção do modelo usando as técnicas de classificação de AM para predição de evasão.

Foram detalhados os critérios para a limpeza de dados, eliminação manual de atributos, transformação e seleção de 14 atributos e a normalização dos dados. Utilizou-se a ferramenta RapidMiner para auxiliar na seleção de atributos e executar os processos de AM para analisar os modelos em 10%, 25%, 50% e 75% de conclusão do curso, ou seja, todo o procedimento de construção do modelo é repetido quatro vezes, gerando quatro versões de treinamentos e testes, chamados de S*, S1, S2 e S3, para predição das classes P (positiva) e N (negativa). A classe P refere-se ao caso de maior interesse, ou seja, aos estudantes que não conseguiram concluir o curso dentro do período da sua integralização.

A maior parte do tempo despendido nos experimentos foi no pré-processamento dos dados. Estima-se que ele consumiu em torno da metade do tempo e esforço total da análise de dados, e envolveram conhecimentos de linguagem de manipulação em banco de dados.

Capítulo 5
Resultados e Discussão

Este capítulo apresenta os resultados da metodologia proposta, com a descrição dos passos efetuados para agrupamento e para predição, os resultados parciais, as comparações entre os resultados de cada modelo aplicado com algoritmos diferentes, as interpretações, os achados e discussões sobre os resultados finais.

5.1 Primeira Etapa: Agrupamento

5.1.1 Análise dos Dados e Resultados

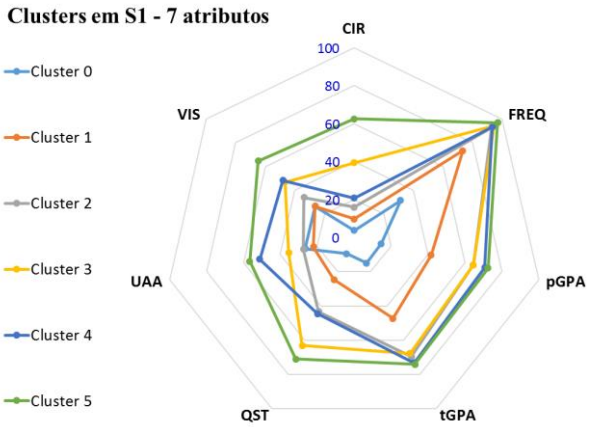
As técnicas de agrupamento são, normalmente, empregadas em tarefas não supervisionadas, nas quais o objetivo é descobrir alguma estrutura nova nos dados. Entretanto, neste trabalho, o objetivo é encontrar padrões que estejam relacionados ao abandono do curso. É importante lembrar que cada grupo pode conter estudantes das classes positiva (evasão) e negativa (conclusão). Portanto, o rótulo de classificação não está necessariamente relacionado aos grupos. Todavia, analisar as características dos estudantes em cada grupo e entender as correlações desses grupos com a evasão é uma parte importante deste trabalho.

O agrupamento foi realizado como explicado no Capítulo 4. Ressalta-se aqui que os atributos CPF e o status de conclusão do curso não foram usados para calcular a proximidade entre os exemplos da base de dados. Após o pré-processamento, o primeiro resultado de agrupamento é mostrado na Figura 32 como um gráfico de radar representando as coordenadas de cada grupo. Nesse gráfico, os eixos representam os atributos e cada polígono representa os centroides de um grupo. Esse gráfico dá uma noção de quais centroides estão mais próximos uns dos outros, e pode ser usado para comparar os grupos encontrados em diferentes momentos.

A Figura 32 mostra seis grupos numerados arbitrariamente de 0 a 5. Observa-se que existem alguns atributos em que os centroides diferem mais entre os grupos. Por outro lado, todos os grupos possuem valores muito semelhantes para os atributos tGPA (nota final da disciplina) e pGPA (média acumulada das atividades). Como as notas médias acumuladas das atividades podem conter mais informações que a nota final da disciplina, e também o vértice do atributo pPGA tem melhor correlação com as classes (Figura 27), optou-se por descartar o

atributo tPGA. Os experimentos seguintes com agrupamento foram feitos com seis atributos: FREQ , pGPA , VIS , QST , UAA e CIR.

Figura 32 - Clusters em S1.



Foram gerados agrupamentos em S*, S1, S2 e S3. Os valores de todos os grupos encontrados são apresentados na Tabela 4. Gráficos de radar para os grupos encontrados nesses momentos são apresentados nas figuras 33 e 34.

Tabela 4 –X-means – Tabela de centroides para S*, S1, S2 e S3.

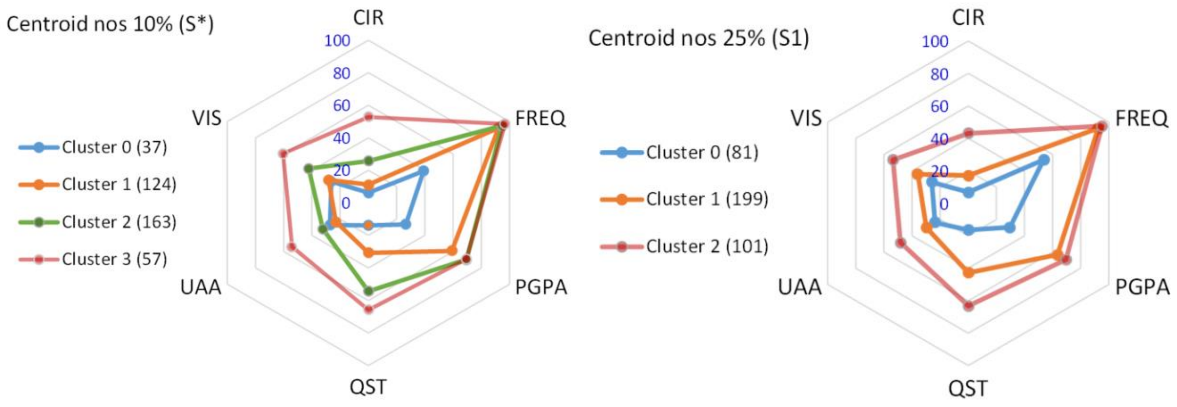
Qtd	S*	CIR	FREQ	PGPA	QST	UAA	VIS
37	Cluster 0	6.19	39.17	26.50	13.79	27.32	26.30
124	Cluster 1	11.15	92.18	59.16	30.72	23.16	28.31
163	Cluster 2	25.76	94.85	69.59	54.47	32.21	42.30
57	Cluster 3	52.82	96.90	69.30	65.86	54.13	60.52
381							

Qtd	S1	CIR	FREQ	PGPA	QST	UAA	VIS
81	Cluster 0	6.83	53.59	29.60	16.53	23.72	26.03
199	Cluster 1	17.12	92.44	63.39	42.82	29.82	35.86
101	Cluster 2	43.04	95.18	69.48	63.28	48.26	53.61
381							

Qtd	S2	CIR	FREQ	PGPA	QST	UAA	VIS
77	Cluster 0	5.22	44.36	31.20	12.23	26.20	29.31
111	Cluster 1	13.71	86.04	52.86	33.83	23.08	29.54
121	Cluster 2	25.11	90.74	65.69	46.11	36.25	41.65
55	Cluster 3	49.27	95.45	70.67	60.88	53.51	57.77
364							

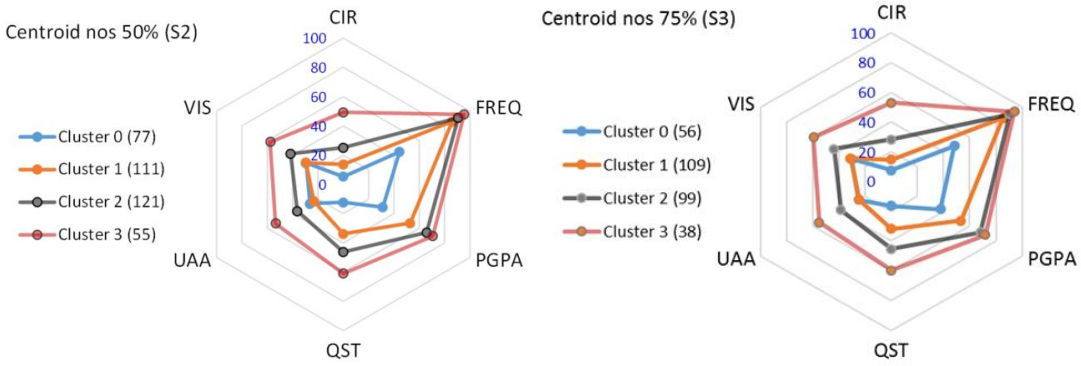
Qtd	S3	CIR	FREQ	PGPA	QST	UAA	VIS
56	Cluster 0	7.60	48.07	37.52	16.32	24.92	30.06
109	Cluster 1	14.82	85.32	53.17	31.93	24.53	31.21
99	Cluster 2	28.14	90.34	68.28	45.50	38.62	43.82
38	Cluster 3	53.11	93.99	71.70	59.76	55.09	59.57
302							

Figura 33 - Agrupamento em S* e S1.



(a) **(b)**

Figura 34 - Agrupamento em S2 e S3.



(a) **(b)**

Embora a numeração dos grupos seja arbitrária e os grupos encontrados em um momento não estejam necessariamente relacionados aos grupos encontrados em outros momentos, é possível observar algumas similaridades. Na Figura 34 observa-se que os centroides dos grupos encontrados em S2 são bastante semelhantes aos grupos encontrados em S3. Mais precisamente, os cluster_0 e cluster_3 são semelhantes em S2 e S3. Isso sugere que eles representam padrões de comportamento comuns entre os estudantes.

Nas quatro figuras fica nítido que existe um grupo com altas taxas (anel mais externo) para todos os atributos. Em cada caso, esse é o único grupo no qual o número de alunos que não concluíram o curso é menor do que o número de alunos concluintes. Então os cluster_3 em S*, S2 e S3, além do cluster_2 em S1, representam os estudantes que possuem bom desempenho. Por sua vez, o cluter_0 é o que possui o centroide com valores mais próximos de zero. Nos quatro casos, ele foi o único grupo no qual todos os alunos evadiram ou reprovaram.

O atributo FREQ é referente ao encontro presencial semanal e o estudante precisa de um mínimo de 75% de presença para aprovação. Nos quatro momentos, o cluster_0 é o único para o qual o centroide ficou abaixo de 75. Mais do que isso, analisando-se os valores de cada estudante nesse grupo, é possível observar que a maioria dos estudantes tem frequência abaixo de 75. Esse grupo identifica, principalmente, o estudante que seria reprovado por falta. Nos demais grupos FREQ está próximo do valor máximo. O atributo referente a frequência do aluno nas aulas, normalmente, é analisado em qualquer curso de qualquer modalidade e nível, e nos resultados obtidos confirma sua relevância, porém limita-se aos casos de reprovação por falta e para outras causas de evasão deve ser analisada em conjunto com outros atributos.

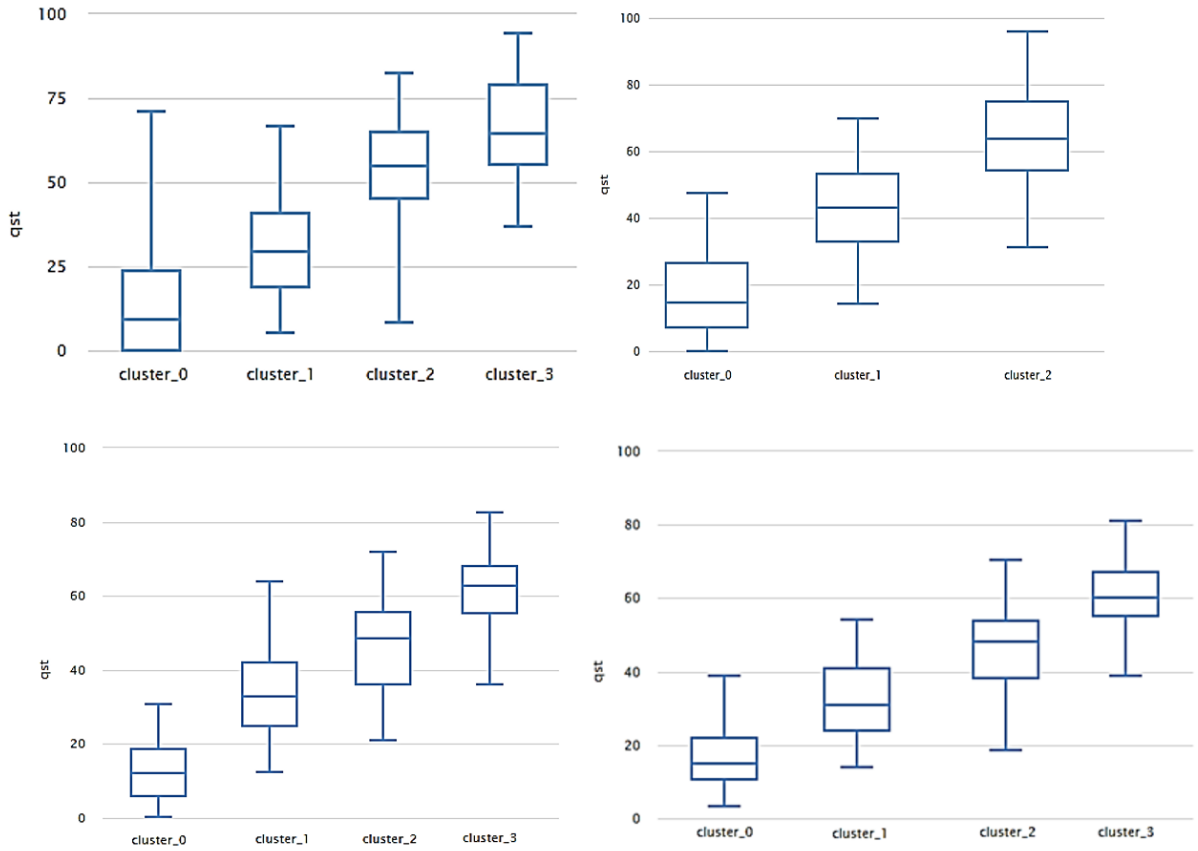
Para continuar a análise de cada atributo por grupo, extraiu-se sua mediana, mínimo e máximo em S*, S1, S2 e S3. Os resultados estão na Tabela 5. Nesta é possível ver numericamente as informações sobre a dispersão e a tendência central dos conjuntos de dados e percebe-se que os atributos QST e CIR têm valores inter-clusters mais distantes.

Tabela 5 – Mediana, mínimo e máximo de cada atributo por cluster em cada período.

		pGPA			FREQ			UAA			QST			VIS			CIR		
		mediana	min.	máx.	mediana	min.	máx.	mediana	min.	máx.	mediana	min.	máx.	mediana	min.	máx.	mediana	min.	máx.
S*	cluster_0	26.25	0.00	61.97	40.50	0.00	82.50	30.50	6.00	50.00	9.00	0.00	70.75	23.25	9.25	38.92	3.25	0.00	29.25
	cluster_1	60.03	29.57	85.83	95.00	60.00	100.00	22.42	1.00	64.50	29.29	5.00	66.50	28.33	6.67	44.67	9.33	0.00	38.25
	cluster_2	71.32	28.00	97.33	100.00	55.25	100.00	31.00	6.50	65.00	54.75	8.00	82.33	42.00	28.00	66.33	25.50	1.00	55.50
	cluster_3	68.42	35.50	90.63	100.00	66.67	100.00	51.50	25.00	100.00	64.67	36.67	94.25	59.67	45.33	84.50	51.50	15.33	96.00
S1	cluster_0	32.78	0.00	70.89	56.14	0.00	100.00	24.33	1.00	50.80	14.67	0.00	47.17	25.00	6.67	52.40	4.29	0.00	29.00
	cluster_1	64.53	33.53	89.57	94.29	62.29	100.00	29.60	3.00	54.20	43.00	14.14	69.50	35.86	18.43	58.29	16.14	1.14	45.14
	cluster_2	70.03	30.25	93.86	97.14	58.57	100.00	45.75	23.00	100.00	63.57	31.00	95.67	50.83	38.43	84.00	42.00	5.29	92.17
S2	cluster_0	32.45	0.00	58.11	48.67	0.00	77.31	27.00	3.00	54.20	12.00	0.00	30.50	27.60	10.56	57.40	4.08	0.00	19.92
	cluster_1	53.40	25.21	79.74	86.92	50.17	100.00	22.38	9.44	41.56	32.69	12.08	63.75	29.55	16.36	43.55	13.46	1.62	35.92
	cluster_2	65.42	26.17	91.08	94.17	50.75	100.00	36.08	17.71	56.54	48.50	20.85	71.77	40.92	29.92	58.15	25.46	1.85	47.92
	cluster_3	71.79	48.19	88.43	97.31	74.67	100.00	51.77	33.10	81.00	62.62	35.85	82.50	55.92	42.75	75.75	48.23	20.62	80.62
S3	cluster_0	37.90	14.05	58.11	48.88	10.00	78.00	23.89	5.83	46.86	14.91	3.06	38.53	29.85	12.13	57.40	6.27	0.53	33.40
	cluster_2	53.59	31.11	79.94	86.24	61.76	99.47	24.65	11.50	47.00	30.74	13.74	54.05	31.11	16.78	45.18	14.00	1.89	32.47
	cluster_1	70.01	41.77	90.29	93.05	52.95	100.00	37.77	19.25	64.09	48.05	18.24	70.25	43.26	29.74	64.00	28.42	4.00	50.95
	cluster_3	72.33	55.74	89.88	94.34	83.53	100.00	54.39	39.76	79.00	60.19	38.74	81.00	58.45	48.84	75.21	51.98	25.63	77.42

Para uma melhor visualização da dispersão desses dois casos apresenta-se em *boxplot* ou diagrama de caixa. Os *boxplots* dos outros atributos estão nos apêndices D, E, F e G.

Figura 35 - *Boxplot* do atributo QST por grupo nos momentos S* (acima e à esquerda), S1, S2 e S3 (abaixo e à direita). Os grupos estão ordenados pelo valor da mediana.



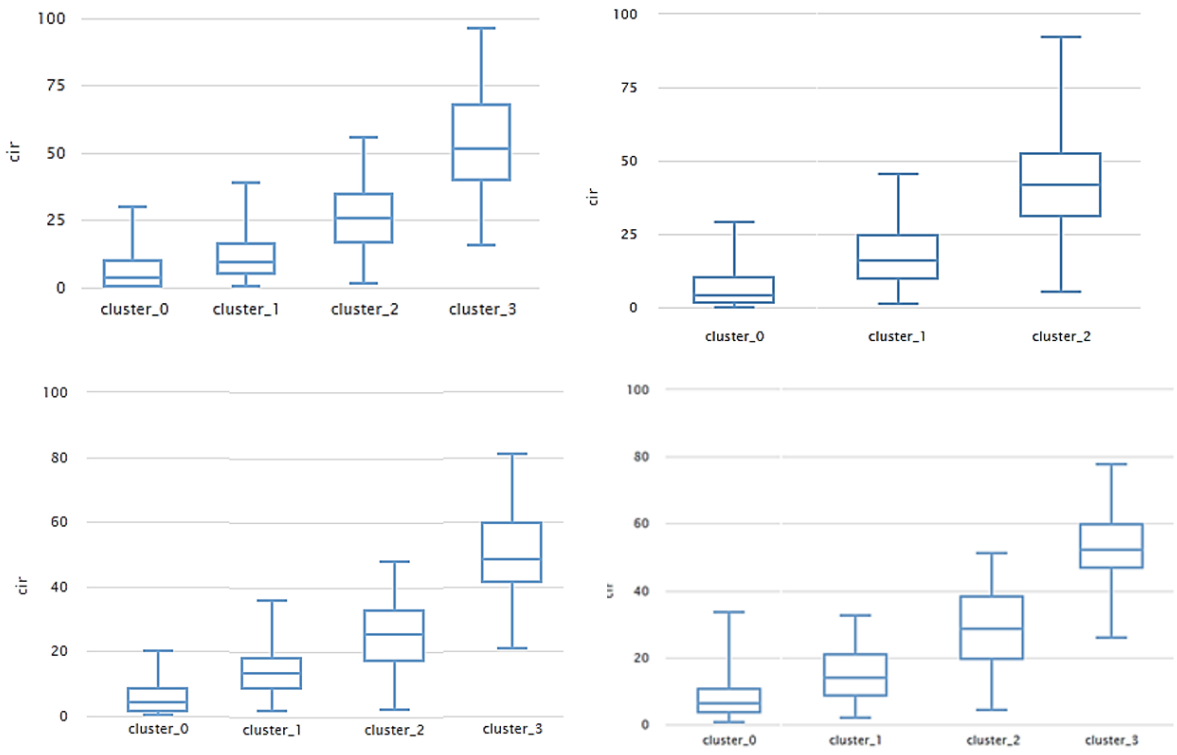
A Figura 35 mostra a distribuição do atributo QST em S*, S1, S2 e S3, com a mediana e os quartis[27]. O atributo QST corresponde às ações relacionadas aos questionários (visualização, tentativa de resposta e envio para correção). Apesar de vários instrumentos de avaliação no AVA os docentes preferem utilizar o questionário, e chega a ser metade das atividades avaliativas por permitir incluir vários tipos de elaboração de questões numa única atividade, como exercícios de associação, de verdadeiro/falso, arrastar e soltar na imagem ou no texto, questões dissertativas, selecionar palavras que faltam, além da mais utilizada múltipla escolha. Em todos os momentos, o cluster_0 contém estudantes com baixa interação com os

[27] Um *boxplot* tem 3 quartis que dividem um determinado conjunto de dados numéricos reais em 4 grupos, e cada grupo inclui aproximadamente 25% (ou um quarto) de todos os valores incluídos no conjunto de dados. Esse método proporciona a localização visual da posição, dispersão, simetria, caudas e os valores extremos (*outliers*) dos dados.

questionários. Uma vez que esse grupo contém apenas estudantes reprovados e evadidos, fica claro que a baixa interação com questionários é um sinal precoce de risco de evasão.

O seguinte atributo relevante é o CIR (Figura 36), um percentual calculado em referência ao máximo da turma, que informa a taxa de visualizações do curso e dos materiais disponibilizados pelo docente.

Figura 36 - *Boxplot* do atributo CIR por grupo em S* (acima e à esquerda), S1, S2 e S3 (abaixo e à direita). Os grupos estão ordenados pelo valor da mediana.



As taxas baixas no CIR indicam pouco acesso ao conteúdo do curso e baixar os materiais (*resources*) enviados pelo professor. De maneira geral, todos os atributos indicam comportamento inadequado dos estudantes quando seus valores são baixos. A relevância do algoritmo de agrupamento na análise desses atributos é justamente encontrar limiares desses atributos para categorizar os estudantes que possuem comportamento semelhante. Por exemplo, como será argumentado no Capítulo 6, o cluster_0, em todos os cenários, concentra estudantes com interação muito baixa com o AVA (todos reprovados ou evadidos), enquanto o grupo que possui mediana mais alta para todos os atributos concentra os estudantes com melhor desempenho.

O Quadro 8 exibe uma contagem dos grupos em classe Positiva, com os status E (evadidos e desistentes) e R (Reprovado), e classe Negativa, que inclui os aprovados e aptos a colar grau. É importante dizer que essa informação não foi utilizada para gerar os grupos.

Quadro 8 - Grupos e instâncias por status e classe nos 4 períodos (S*, S1, S2 e S3). Os grupos estão ordenados como aparecem nas figuras anteriores.

S*				S1				S2				S3			
cluster	classe	status	qtd	cluster	classe	status	qtd	cluster	classe	status	qtd	cluster	classe	status	qtd
cluster_0	P	E	33	cluster_0	P	**E**	**71**	cluster_0	P	**E**	**73**	cluster_0	P	**E**	**43**
cluster_0	P	R	4	cluster_0	P	R	9	cluster_0	P	R	4	cluster_0	P	R	13
cluster_1	N		25	cluster_1	N		66	cluster_1	N		18	cluster_1	N		29
cluster_1	P	**E**	**51**	cluster_1	P	**E**	**56**	cluster_1	P	E	30	cluster_1	P	E	14
cluster_1	P	**R**	**48**	cluster_1	P	**R**	**78**	cluster_1	P	**R**	**63**	cluster_1	P	**R**	**66**
cluster_2	N		69	cluster_2	N		61	cluster_2	N		65	cluster_2	N		66
cluster_2	P	**E**	**48**	cluster_2	P	E	15	cluster_2	P	E	19	cluster_2	P	E	9
cluster_2	P	**R**	**46**	cluster_2	P	R	25	cluster_2	P	R	37	cluster_2	P	R	24
cluster_3	N		33					cluster_3	N		44	cluster_3	N		32
cluster_3	P	E	10					cluster_3	P	E	3				
cluster_3	P	R	14					cluster_3	P	R	8	cluster_3	P	R	6
estudantes			381				381				364				302

Esse quadro resume as informações obtidas nos experimentos. É possível confirmar que o cluster_0 só contém instâncias da classe P. Esse é o grupo que aparece mais interno nos gráficos de radar. O grupo que aparece mais externo nos gráficos (cluster_2 em S1 e cluster_3 nos outros momentos) tem maioria de classe N Em S* observa-se que nos clusters intermediários 1 e 2 existem instâncias tanto da classe P como N. Estes representam os casos de estudantes com comportamentos iniciais semelhantes, porém parecem evoluir diferentes até finalizar o curso, e reúnem os casos que precisam ser acompanhados com mais atenção.

Em S2 encontram-se 4 grupos. Nesse momento, distinguem-se melhor os casos dos evadidos (cluster_0), o de maioria reprovada (cluster_1) e o de maioria aprovada (cluster_3). O total de estudantes difere porque há uma transição do 1º ao 2º semestre entre S1 e S2 e alguns estudantes abandonam o curso, sem renovar a matrícula. Essa situação se acentua no começo do S3, mostrando que ao finalizar o primeiro ano do curso, na metade da duração regular, 21% dos estudantes não renovaram a matrícula.

5.1.2 Descobertas com técnicas de Agrupamento

A Tabela 6 é um resumo com os quantitativos das classes Negativas e Positivas.

Tabela 6 - Estudantes nas classes N e P.

classe	S*	S1	S2	S3
NEGATIVO	127	127	127	127
POSITIVO	254	254	237	175
	381	381	364	302

Além das informações apresentadas, é relevante lembrar que cada instância possui uma chave identificadora, o CPF. Essa chave foi ignorada na construção dos grupos e no treinamento do classificador, mas pode ser usada para analisar a quais grupos cada estudante pertence em S*, S1, S2 e S3.

Usando o CPF, foi feita uma análise de transição dos estudantes entre cada período, de (S* para S1), de (S1 a S2) e de (S2 a S3). Essas ocorrências podem ser observadas no Quadro 9 para os estudantes da classe P (reprovado ou evadido) e no Quadro 10 para os estudantes da classe N. Com essa importante informação, foi possível detectar mudanças comportamentais entre os estudantes. Por exemplo, todos os estudantes que foram agrupados no cluster_0 em S* também foram agrupados no cluster_0 em S1. Isso não significa que os dois grupos são iguais em S* e em S1, mas que esses estudantes possuem um padrão de comportamento recorrente. Além da contagem de casos de transição, os quadros trazem algumas considerações sobre o comportamento observado nos estudantes desses grupos.

Quadro 9 –Transição dos estudantes entre grupos da classe P.

Classe P (Possitiva)								
de S* para S1			de S1 para S2			de S2 para S3		
cluster_0cluster_0	37	nem acessa AVA	cluster_0cluster_0	57	nem acessa AVA	cluster_0cluster_0	33	evadido
			cluster_0cluster_1	7	alguma melhora	cluster_0	44	não renovou matrícula
			cluster_0cluster_2	1				
cluster_1cluster_0	33	risco alto	cluster_0	15	não renovou matri	cluster_1cluster_0	20	evadido
cluster_1cluster_1	65	risco médio				cluster_1cluster_1	59	risco muito alto
cluster_1cluster_2	1		cluster_1cluster_0	19		cluster_1cluster_2	1	
			cluster_1cluster_1	83	risco alto	cluster_1	13	não renovou matrícula
			cluster_1cluster_2	30	risco médio			
cluster_2cluster_0	10	risco alto	cluster_1	2	não renovou matri	cluster_2cluster_0	3	risco alto
cluster_2cluster_1	64	risco médio				cluster_2cluster_1	21	risco alto
cluster_2cluster_2	20		cluster_2cluster_0	1	risco alto	cluster_2cluster_2	27	estável
			cluster_2cluster_1	3	risco médio	cluster_2	5	não renovou matrícula
			cluster_2cluster_2	25				
cluster_3cluster_1	5	queda abrupta	cluster_2cluster_3	11		cluster_3cluster_2	5	queda
cluster_3cluster_2	19					cluster_3cluster_3	6	
	254			254			237	

Quadro 10 - Transição dos estudantes entre grupos da classe N.

ClasseN (Negativa)							
de S* para S1			de S1 para S2			de S2 para S3		
cluster_1cluster_1	25	baixo	cluster_1cluster_1	17	baixo	cluster_1cluster_1	16	baixo
			cluster_2cluster_1	1		cluster_1cluster_2	2	melhorou
cluster_2cluster_1	41	baixou	cluster_1cluster_2	47	melhorou	cluster_2cluster_1	13	baixou
cluster_2cluster_2	28	permanece alto	cluster_2cluster_2	18	estável	cluster_2cluster_2	51	bom
						cluster_2cluster_3	1	alto
cluster_3cluster_2	33	permanece alto	cluster_1cluster_3	2	melhorou			
			cluster_2cluster_3	42	permanece alto	cluster_3cluster_2	13	permanece alto
						cluster_3cluster_3	31	permanece alto
	127			127			127	

Quadro 11: Detalhes na transição dos estudantes entre os grupos.

				FREQ				pGPA
Classe	Status	Estudantes	Transição	S0	S1	S2	S3	S0
P	D	3	00	41.83	40.11			31.35
P	E	12	00	40.85	39.18			16.04
P	D	5	000	41.75	28.71	22.59		26.97
P	E	12	000	37.62	30.96	25.66		32.95
P	R	2	000	10.00	10.00	27.44		30.99
P	E	1	0000	40.00	24.29	13.08	10.00	49.17
P	R	1	0000	73.33	56.14	61.31	41.95	18.00
P	R	1	0010	40.00	33.33	58.92	63.26	34.74
P	D	2	100	100.00	51.43	44.33		62.36
P	E	1	3222	100.00	95.71	74.62	57.65	71.99
P	R	4	3222	98.33	94.94	94.69	90.16	65.52
N	A	21	2122	96.79	95.84	94.42	93.40	74.69
N	C	9	2122	95.46	93.84	93.33	92.26	77.17

No Quadro 11 são apresentados alguns casos de estudantes na transição entre os grupos. Como exemplo, a transição 00 mostra os 15 estudantes que desistiram no primeiro semestre do curso (círculo laranja); a transição 3222 mostra os estudantes que ficaram em grupos com maior desempenho durante o curso, porém não conseguiram concluir com sucesso (círculo celeste) e por outro lado os que conseguiram sucesso (círculo verde). Essa informação permite analisar melhor as características usando os valores dos atributos durante o andamento do curso. Quadros mais detalhados mostrando todas as possíveis combinações de transições entre os grupos são apresentados no Apêndice I (classe P) e Apêndice J (classe N).

Com base nisso, as seguintes descobertas são apresentadas.

a) Embora o status final não tenha sido usado como parte do agrupamento, a maioria dos grupos mostra uma distribuição de estudantes que reflete sua condição futura (aprovado, desistente ou evadido);

b) O cluster_0 é aquele cujos valores dos atributos são menores. Ele contém os estudantes com frequência baixa, histórico de visualização das atividades ruim e notas, em geral,

baixas. Estudantes que se agrupam nesse cluster logo no início do curso dificilmente se recuperam e tendem a permanecer com esse comportamento até desistirem ou serem reprovados;

c) Todos os 37 casos estudantes no grupo de baixo desempenho em S* (cluster_0) não concluíram o curso, sendo que 15 desses estudantes abandonaram ainda no primeiro semestre, 19 no segundo e 3 no terceiro. Esses casos parecem ser os mais evidentes na sua detecção e ao mesmo tempo resulta mais desafiador conseguir reverter esse comportamento. Seria necessária uma intervenção imediata quando registradas as primeiras faltas contínuas no primeiro mês, pois a baixa frequência, assim como outras características com valores baixos, são os sinais que apresentam quem permanece nos grupos de baixo desempenho;

d) Estudantes que começam a demonstrar comportamento semelhante ao dos estudantes no cluster_0 apresentam risco extremamente elevado de não completarem o curso, mesmo que esses sinais não sejam imediatos. Dos 80 estudantes agrupados no cluster_0 em S1, apenas 8 conseguiram recuperar e não foram novamente agrupados com estudantes de baixo desempenho em S3;

e) Estudantes que iniciam nos grupos de maior desempenho (cluster_3 em S* e cluster 2 em S1) não necessariamente mantêm o desempenho durante todo o curso, e os estudantes que perdem o rendimento têm dificuldade em recuperar. Dos 33 estudantes aprovados que iniciaram no cluster_3 em S*, apenas 3 não foram agrupados com seus colegas no cluster_3 em S2. Além disso, a taxa de reprovação entre os estudantes que estavam no cluster_3 em S* e foram atribuídos a outro cluster em S2 foi de 84,21%.

f) As instâncias pertencentes aos grupos “intermediários” (cluster_1 e cluster_2 em S* e cluster_1 em S1) concentram a maior parte dos estudantes. Há uma quantidade semelhante de instâncias das classes N e P, mas a partir de S2 parece haver características que sugerem um estudante com maior risco de reprovação ou de evasão. O cluster_2 em S2 compartilha muitas semelhanças com o cluster_2em S3 e ambos possuem maior quantidade de reprovações do que evasões dentre os casos da classe positiva. Estudantes nesses grupos possuem FREQ (frequência) relativamente elevada (mediana acima de 90%). Os atributos de visualização de questionários (QST) e atividades (CIR), por outro lado, são baixos. Mas, observando-se casos específicos desses dois grupos, existe uma diferença significativa com respeito à nota. Os estudantes desses grupos que terminaram sendo reprovados ou que evadiram tiveram

pPGA médio 60,10, já os estudantes aprovados tiveram pPGA médio 72,98. Estudantes com alta frequência e baixa atividade no AVA são casos suspeitos, mas aqueles que vem mantendo notas razoáveis estão em menor risco de abandonarem o curso subitamente;

g) O melhor desempenho é obtido pelos estudantes que apresentam resultados elevados na avaliação, frequência e acesso aos recursos disponibilizados pelo docente (CIR), simultaneamente, desde o início do curso. Apenas parte deles não seria suficiente;

h) O cluster_1 em S1 pode ser considerado o dos evadidos com baixa interação. Dos 134 casos, 19 deles (14%) pararam de ter interações com o AVA, 83 estudantes (33%) frequentaram as aulas, mas em outros atributos têm alcances abaixo da média;

i) O cluster_2 em S1 chama a atenção que ao finalizar o primeiro ano, este continha 10 estudantes em grupo com suposto bom desempenho. Entretanto, ao acompanhar a evolução individual, viu-se que o estudante desistiu de forma voluntária ou teve queda abrupta no desempenho e não concluiu o curso.

As conclusões que não podem ser tiradas do sumário apresentado nos quadros 8 e 9 foram obtidas analisando-se a evolução dos alunos nos quatro momentos, não apenas na transição de um momento para o seguinte.

5.2 Segunda Etapa: Classificação

A informação obtida sobre os grupos formados em cada momento S*, S1, S2 e S3 foi utilizada como atributo qualitativo agregado no conjunto de atributos de entrada para classificação na predição de estudante com risco de evasão.

Os algoritmos de classificação são: Random Forest (RF), Decision Tree (DT), Gradient Boost Trees (GBT), Logistic Regression (LR), Naïve Bayes (NB) e Support Vector Machine (SVM).

A qualidade dos modelos foi avaliada de acordo com Precisão, Revocação e F1-score. A Precisão está relacionada à capacidade do classificador em identificar corretamente amostras da classe Positiva. No contexto deste trabalho, alta Precisão significa que, quando o classificador identifica que um estudante está em risco de evasão, há uma alta probabilidade de que esse estudante específico esteja realmente em risco. Mas a Precisão por si só não é uma métrica definitiva, e a Revocação, relacionada à capacidade do modelo de detectar todas as amostras da classe Positiva, também foi considerada. Neste contexto, alta Revocação significa

que, se um estudante estiver em risco de evasão, há uma alta probabilidade de que o classificador o detecte.

Em geral, há um equilíbrio entre Precisão e Revocação. Maximizar a Precisão tende a tornar o modelo mais focado nos casos positivos “mais fáceis”, diminuindo os falsos positivos em detrimento dos verdadeiros positivos. Por outro lado, maximizar o Revocação tende a criar modelos que “correm riscos maiores”, aumentando os verdadeiros positivos, mas também os falsos positivos. Um equilíbrio entre essas duas medidas pode ser obtido com métricas como o F1-Score.

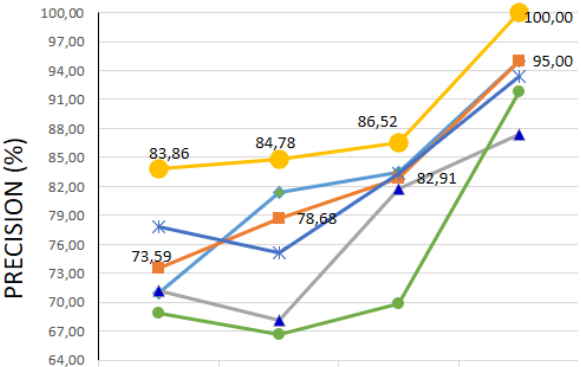
Existem duas questões importantes a serem analisadas. Primeiro, se é verdade que um método pode ser criado para predizer o desempenho futuro de um estudante a partir dos dados no Moodle. Em segundo lugar, é relevante verificar com quanta antecedência esse método pode ser empregado para produzir respostas adequadas, e direcionar os esforços para os estudantes que mais precisam de assistência.

5.2.1 Análise dos Dados e Resultados

Considerou-se o desempenho de diferentes modelos de classificação treinados para maximizar Precisão, Revocação e F1-Score (individualmente). Todo o processo é repetido quatro vezes, uma para cada momento, S*, S1, S2 e S3. Os resultados dos seis algoritmos para classificação nos quatro momentos são mostrados na Figura 37 e na Figura 38.

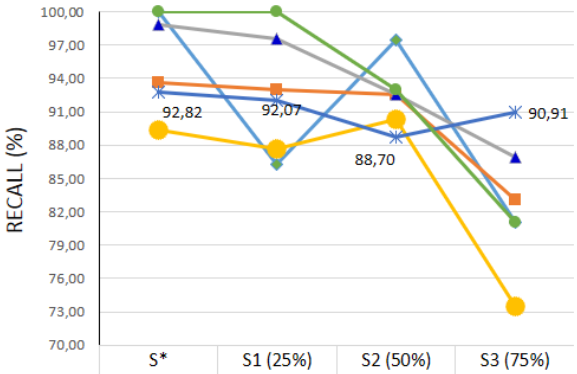
Nas Figuras 37a e 37b, o NB e RF são as mais equilibradas nas mudanças de Precisão e Revocação, o que se traduz em uma excelente métrica F1 (Figura 38) durante o curso. Observa-se também que o NB, algoritmo baseado em cálculos probabilísticos, destaca-se por apresentar F1 com melhoria constante e significativa à medida que o curso avança.

Figura 37 - Métricas dos algoritmos de classificação:
(**a**) Métrica de Precisão; (**b**) Métrica de Revocação.



	S*	S1 (25%)	S2 (50%)	S3 (75%)
GBT	71,00	81,40	83,47	95,00
RF	73,59	78,68	82,91	95,00
DT	71,24	68,17	81,70	87,36
LR	83,86	84,78	86,52	100,00
NB	77,84	75,12	83,37	93,33
SVM	68,87	66,71	69,85	91,74

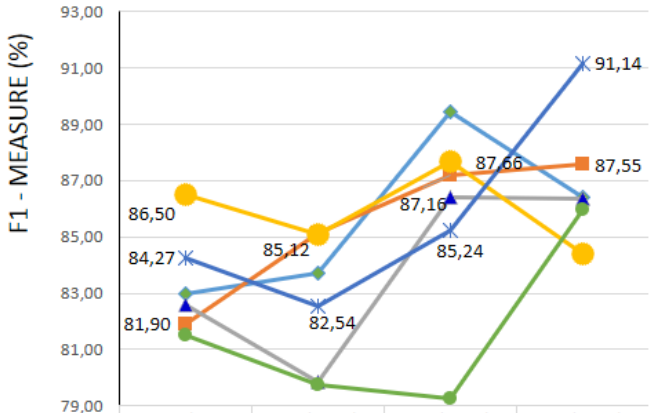
(a)



	S*	S1 (25%)	S2 (50%)	S3 (75%)
GBT	100,00	86,28	97,42	81,09
RF	93,65	92,96	92,52	83,09
DT	98,82	97,57	92,52	86,91
LR	89,39	87,65	90,30	73,45
NB	92,82	92,07	88,70	90,91
SVM	100,00	100,00	93,02	81,09

(b)

Figura 38 - Desempenho da métrica F1 com 6 atributos + cluster,
em S*, S1, S2, S3.



	S*	S1 (25%)	S2 (50%)	S3 (75%)
GBT	82,99	83,71	89,44	86,38
RF	81,90	85,12	87,16	87,55
DT	82,58	79,87	86,38	86,35
LR	86,50	85,09	87,66	84,39
NB	84,27	82,54	85,24	91,14
SVM	81,50	79,74	79,28	85,94

Quadro 12 - Métrica F1 no momento S* com
(a) 11 atributos; (b) 6 atributos.

F1	S*	S1 (25%)	S2 (50%)	S3 (75%)
GBT	78.83	81.10	85.27	82.70
RF	78.52	82.72	87.71	87.65
DT	79.88	80.48	85.35	82.70
LR	77.95	79.78	85.22	84.78
NB	78.01	81.54	84.53	87.45
SVM	79.01	80.54	80.55	83.83

F1	S*	S1 (25%)	S2 (50%)	S3 (75%)
GBT	76.74	80.47	89.47	85.33
RF	78.93	80.01	88.72	84.29
DT	79.40	82.31	84.20	83.84
LR	66.89	79.72	86.52	85.11
NB	77.82	76.35	84.97	86.61
SVM	77.79	76.26	83.12	82.52

Comparou-se o desempenho da métrica F1 desses seis algoritmos (Figura 38) com outros dois casos:

a) Modelos treinados e testados com os 11 atributos selecionados inicialmente antes da Primeira Etapa (Quadro 12a). São os 13 atributos iniciais menos os atributos sit_familia (maioria com valor 8-Estudante) e sit_trabalho (maioria com valor 2-Dependente), pois estes 2 tinham 90% dos dados com valor único;

b) Modelos treinados e testados com os 6 atributos selecionados na Primeira Etapa, porém sem incluir a informação de grupo.

É possível observar que o F1 obtido com os 7 atributos (Figura 38), quando inclui o número do grupo como atributo, em qualquer situação obtém uma taxa F1 acima de 80%, enquanto o S* com 11 atributos ou com 6 atributos ficam abaixo desse valor. Isso revela um aspecto muito importante do método, que evidenciou uma boa métrica de predição nas semanais iniciais dos cursos em S*. A classificação em duas etapas permite que os modelos tenham, logo nas primeiras semanas, desempenho equivalente ao obtido no final do curso sem a etapa de agrupamento.

Também é possível observar que em qualquer outro período a melhor métrica F1 do método proposto com duas etapas na Figura 38 supera os resultados do melhor encontrado com os métodos do Quadro 12. Em S* a melhor métrica F1 é Logistic Regression (LR) com 86.50%, em S1 os melhores são Logistic Regression 85.09% e Random Forest (RF) com 85.12%, em S2 é o Gradient Boost Tree (GBT) com 89.44% e en S3 o Naive Bayes (NB) com 91.14%.

Outra informação comparada é a matriz de confusão para cada algoritmo que obteve o melhor desempenho na métrica F1 em S*, considerando os dois casos de seleção de atributos (11 e 6 atributos), com e sem o agrupamento. Esse resultado é mostrado no Quadro 13. Observe que o melhor algoritmo não foi o mesmo em todos os casos (o melhor algoritmo está indicado

em cada subquadro). Os valores de precisão e revocação mostrados nas tabelas são para cada classe.

É importante que a predição de risco de evasão seja informada o mais cedo possível para ter uma intervenção mais efetiva. Com esse objetivo, comparou-se o desempenho dos indicadores no S*, quando finalizam as primeiras 3-4 disciplinas do curso, correspondente a aproximadamente 10% do curso.

Quadro 13 - matriz de confusão para o algoritmo com melhor F1 para o momento S* com: (a) 11 atributos; (b) 6 atributos; (c) 11 atributos + cluster; (d) 6 atributos + cluster.

S* - 11atrib. **DT** (F1 - 79.88%)	Real N	Real P	Precisão
pred. N	18	10	64.29%
pred. P	19	62	76.54%
Revocação	48.65%	86.11%	

(a)

S* - 6 atrib **DT** (F1 - 79,40%)	Real N	Real P	Precisão
pred. N	5	4	55.56%
pred. P	31	69	69.00%
Revocação	13.89%	94.52%	

(b)

S* - 12 atrib. **NB** (F1- 80.8%)	Real N	Real P	Precisão
pred. N	0	0	0.00%
pred. P	35	74	67.89%
Revocação	0.00%	100.00%	

(c)

S* - 7 atrib. **LR** (F1 - 86.5%)	Real N	Real P	Precisão
pred. N	21	8	72.41%
pred. P	13	68	83.95%
Revocação	61.76%	89.47%	

(d)

Ao analisar a comparação é importante mencionar que a soma das classes Positivas e Negativas não ficaram iguais em todas as matrizes. Isso acontece porque o *multi-hold-out* empregado não garante que os cinco subconjuntos da partição de teste considerados serão os mesmos (o melhor e o pior resultados, de cada modelo, são descartados, como abordado na Seção 4.7).

Ao analisar a Revocação, no caso dos 11 atributos (Quadro 13a) a arvore de decisão teve um resultado de 48.65%. Com 6 atributos (Quadro 13b) sinaliza incorretamente 31 estudantes com risco em S* (estudantes com predição errada de risco de evasão), que corresponde a 13.89% de Revocação, e no caso dos 12 atributos (Quadro 13c) é um classificador majoritário, ou seja, seleciona a classe majoritária no conjunto de dados e classifica todas as instâncias como sendo da classe P, o que leva a 0% de Revocação. No caso de 7 atributos (6 atributos + cluster) (Quadro 13d), o regressor logístico obteve um melhor equilíbrio de Precisão e de Revocação em S*, com 89.47% e 61.76%, respectivamente.

5.2.2 Descobertas com a Classificação

Na figura 38 observa-se que a o algoritmo regressor logístico tem o melhor F1 com 86,5% em S* e junto com a matriz de confusão do Quadro 13d mostram ser a melhor técnica para os 10% do curso. Em S1, que corresponde ao encerramento do 1º. semestre, o regressor logístico continua com um desempenho alto (85,09%), com valor próximo da floresta aleatória (85,12%), que ficou como melhor F1.

Alguns classificadores dão melhores resultados no S3 quando comparados ao S2, enquanto outros caem no desempenho. No entanto, quase todos melhoram de S1 para S2. Nos casos em que Precisão ou Revocação foi menor em S2 do que em S1, isso foi compensado por um aumento na outra métrica.

Isso é provavelmente explicado pela diferença nos perfis dos estudantes à medida que o curso avança. A transição do S1 para o S2 acontece no final do primeiro semestre e início do segundo semestre, e a maioria dos estudantes que, eventualmente, irão desistir não abandona o curso nesse momento. Consequentemente, os estudantes de S1 e S2 são em sua maioria os mesmos (com exceção do curso de Administração), mas em S2 tem mais informações sobre eles e os modelos tendem a ter um desempenho melhor neste ponto do que antes. Por outro lado, entre S2 e S3 os cursos perderam vários estudantes, e há uma diferença maior entre os dados coletados em S2 e em S3, o que pode causar a necessidade de um ajuste maior do modelo com os casos de estudantes que continuam matriculados e explique a queda na Precisão e Revocação na análise dessa nova massa de dados.

É particularmente interessante como essa mudança no comportamento do estudante afeta o NB e RF, os dois casos que melhoram de desempenho à medida que avança o tempo. Para esses 2 algoritmos verificaram-se as matrizes de confusão em S1, S2 e S3.

Figura 39 - Matrizes de confusão para o classificador NB.

S1-NB (F1-82.5%)	true N	true P	S2-NB (F1-85.2%)	true N	true P	S3-NB (F1-91.1%)	true N	true P
pred. N	1	0	pred. N	30	8	pred. N	34	5
pred. P	36	72	pred. P	11	55	pred. P	3	45

Figura 40 - Matrizes de confusão para o classificador RF.

S1 RF (F1-85.1%)	true N	true P	S2-RF (F1-87.16%)	true N	true P	S3-RF (F1 87.55%)	true N	true P
pred. N	19	5	pred. N	27	5	pred. N	36	9
pred. P	18	67	pred. P	12	60	pred. P	2	39

Para NB (Figura 39) sinaliza incorretamente estudantes como em risco e gera um alto número de falsos positivos. Esse erro afeta a precisão do modelo e, à medida que o curso avança

e a proporção de instâncias da classe negativa para positiva aumenta, e causa um impacto maior no F1-score do modelo. A mesma situação acontece com o RF (Figura 40).

O algoritmo NB é um classificador probabilístico, que por ser muito simples e rápido, possui um desempenho relativamente maior do que outros classificadores. Por ser simples, ele só precisa de um pequeno número de dados de treinamento, diferentemente de modelos mais complexos como as redes neurais profundas, que requerem grandes volumes de dados de treinamento para ajustar a enorme quantidade de pesos que esses modelos possuem.

Por fim, analisou-se a correlação de cada atributo com a turma para determinar quais atributos contribuem mais para a avaliação do desempenho de um estudante. O peso de cada atributo é o coeficiente de correlação de Pearson com o alvo. Na Tabela 7 observa-se que os atributos variam de acordo com o momento em que os dados foram analisados, mas os primeiros permanecem na mesma ordem.

Tabela 7 - Pesos dos atributos segundo correlação de Pearson:
(a) em S* (10%); S1 (25%); (b) em S2 (50%); (c) em S3 (75%).

S* (8% -10%)		S1 (25%)		S2 (50%)		S3 (75%)	
Atributo	Peso	Atributo	Peso	Atributo	Peso	Atributo	Peso
pGPA	0.439	pGPA	0.534	pGPA	0.625	pGPA	0.705
cluster_no	0.352	**cluster_no**	0.431	**cluster_no**	0.585	**cluster_no**	0.577
FREQ	0.294	FREQ	0.377	FREQ	0.483	FREQ	0.535
CIR	0.285	QST	0.365	CIR	0.466	QST	0.475
VIS	0.278	VIS	0.359	QST	0.451	VIS	0.465
QST	0.256	CIR	0.344	VIS	0.443	CIR	0.460
UAA	0.211	UAA	0.280	UAA	0.418	UAA	0.447

O pGPA é o atributo com maior correlação com as classes P e N, lembrando que é a nota média acumulada das atividades no AVA e não a nota final na disciplina. Essa nota foi o atributo com maior peso por correlação ao longo do curso e, de certa forma, confirma o conhecimento empírico que as notas da avaliação se correlacionam diretamente com o sucesso ou não no estudo.

O seguinte atributo é o cluster_no, atributo qualitativo resultante da saída do método de agrupamento, seguido do atributo FREQ, que é a frequência. As diferenças começam com o 4º. atributo mais útil em cada período analisado, na qual estão o QST, que está relacionado à taxa com que os estudantes se envolvem com os questionários, e o CIR, que é a taxa de acesso aos materiais disponibilizados no AVA, confirmando o que foi detectado nos grupos encontrados na Etapa 1.

O atributo pGPA torna-se mais crítico à medida que avança o período de predição; um baixo número de acessos aos materiais do professor (CIR) identifica estudantes em risco de evasão e reprovados; e o acesso a materiais externos e envio de tarefas (UAA) tem baixa correlação no desempenho dos estudantes.

5.3 Discussão dos Resultados

Este trabalho adotou a abordagem de duas etapas em cascata, com técnicas não supervisionada e supervisionada e, sem dúvida, demonstra a importância do agrupamento de estudantes em problemas relacionados à predição. A esse respeito, este trabalho demonstrou que o peso da descrição com o grupo formado para treinar os modelos melhora o desempenho.

O método seguiu os passos da Figura 18, em duas etapas: na primeira, com técnicas para agrupar os estudantes pelas características de comportamento no AVA e, na segunda etapa, utilizou esse resultado como um dos dados de entrada na base rotulada para treinar e testar os algoritmos de classificação.

Os dados de treinamento e teste para algoritmos de AM foram montados com recursos a partir de dados de log do Moodle, bem como dados socioeconômicos fornecidos pelos estudantes no ingresso no curso. Foram analisados os dados em 10% da conclusão do curso (S*), 25% (S1), 50% (S2) e 75% (S3).

Na Etapa 1, de agrupamentos, foram encontrados 4 grupos em S*, 3 em S1 e 4 em S2 e S3. As extrações de novas informações sobre os comportamentos dos grupos precisam de uma boa análise e interpretações baseados nos diversos dados que essa técnica permite explorar, gerando os *insights* na seção 5.2.2.

Na Etapa 2, de classificação, foram utilizados 6 algoritmos: DT, RF, GTB, NB, LR e SVM em 4 situações diferentes: a) 11 atributos (dados de rastreamento e socioeconomicos); b) 6 atributos (dados de rastreamento); c) idem a + informação de cluster; d) idem b+ informação de cluster. Na comparação dos resultados, o caso d) apresentou a taxa F1 acima de 80% em S* para todos os classificadores, enquanto nos outros 3 contextos a F1 ficou por debaixo desse valor em todos os algoritmos, constatando a importante melhoria na predição ao trabalhar com o atributo obtido na primeira etapa do método. Nos momentos S1, S2 e S3 o caso d) também obteve os melhores resultados.

Os resultados sugerem que o regressor logístico obteve um melhor equilíbrio de Precisão e de Revocação em S* e oferece o melhor modelo para predizer estudantes em risco

de evasão no momento inicial, com 10% do curso (S*), e a partir do S1, ao completar o 1º. semestre, o NB, que é baseado em cálculos estatísticos, mostrou-se eficiente até o final do curso com acentuada melhoria nos resultados do S1 ao S2, e deste ao S3.

A frequência (FREQ) não é determinante, porém ela junto com a nota média das atividades (pGPA) e a taxa de acesso aos recursos disponibilizados pelo docente (CIR) determinam com maior claridade a predição das classes P e N. Isso é mostrado na alta correlação com as classes (Tabela 7) e pelos agrupamentos encontrados (Tabela 5).

Capítulo 6

Considerações Finais da Tese

Este capítulo apresenta os resultados obtidos no decorrer desta tese para construção de um método de predição de risco de evasão em cursos EaD. São mostradas as dificuldades na etapa de pré-processamento, as contribuições e limitações da pesquisa, os trabalhos futuros e as publicações alcançadas.

Esta Tese de Doutorado mostrou um estudo sobre evasão dos estudantes baseado nas interações no AVA de cursos em andamento ofertados na EaD, contribuindo com a Informática na Educação, sob o tema da evasão usando técnicas de AM.

O ano de 2017 foi marcado pela flexibilização das regras para a criação de instituições e polos de EAD no Brasil, incluindo o nível médio. A pesar que os maiores investimentos e foco continuam sendo no ensino superior, segundo a ABED (2021) destaca um aumento na oferta de conteúdo para a educação básica, tendência verificada pela mudança nas novas Diretrizes Curriculares Nacionais, que, em suma, passa a permitir oferta de um percentual de EAD para o ensino médio.

Vemos que nos últimos 10 anos os investimentos financeiros do Governo Federal no ensino técnico e profissional são altos, iniciado com o Pronatec, porém o retorno esperado ainda está aquém do esperado para chegar em qualidade e quantidade nos níveis de outros países membros da OCDE. No Brasil, o problema da evasão no EPT é um dos desafios destacados no relatório da OCDE de 2021.

A RSL mostrou que a maioria das pesquisas usam classificadores (técnicas supervisionadas), poucos usam técnicas não supervisionadas e são raras as abordagens com ambas, e quando abordada são usadas com objetivos diferentes. Neste trabalho analisaram-se 3 cursos concomitantes ao ensino médio ofertados no IFRO na modalidade EaD, que conta com um encontro presencial semanal. Para uma melhor predição apresentou-se uma técnica em duas etapas. Inicialmente, com abordagem não supervisionada para agrupar estudantes sem fixar o valor de k, e usar essa informação do grupo formado como um atributo na entrada de dados para a técnica supervisionada, analisando as métricas de desempenho com seis classificadores

dos estudantes em risco de evasão. Dessa forma, apresentou-se um método que une duas abordagens de AM, aplicado em cascata ou consecutivas.

Este estudo revelou que técnicas de agrupamento (técnica não supervisionada) para encontrar os grupos pelas características dos estudantes melhoram significativamente o desempenho dos modelos preditivos (técnica supervisionada). Durante a linha do tempo do módulo do curso, o desempenho dos estudantes foi previsto logo no início, com 10% do andamento do curso (S*). Posteriormente, foi previsto o desempenho dos estudantes em 25% (S1), 50% (S2) e 75% (S3) da duração do curso de 2 anos, ou seja, nestes casos coincide com o fechamento de cada semestre. Essa análise em momentos diferentes da duração do curso aborda o dinamismo de um ambiente educacional, que neste estudo 21% dos estudantes abandonaram o curso no primeiro ano ou metade do período normal de duração.

Com aproximadamente 10% da duração do curso, o modelo preditivo de LR produziu resultados promissores com 86.5% de métrica F1, obtido com os 7 atributos (Figura 38), quando inclui o número do cluster como atributo agregado. O uso desse atributo nos seis algoritmos (DT, RF, GBT, LR, NB e SVM) resultou em uma taxa F1 acima de 80%, e quando comparado aos resultados sem o atributo agregado os algoritmos retornaram F1 abaixo dos 80%. Isso revela um aspecto muito importante deste método, que aponta uma boa métrica de predição nas semanas iniciais dos cursos. Também foi analisada a Revocação e os falsos negativos, pois não só importam os acertos, senão também os que não foram detectados com risco de evasão, os que ficariam erroneamente desconsiderados para uma eventual necessidade de intervenção.

Atualmente, os docentes dos cursos analisados acessam só as notas e frequência das suas próprias disciplinas, sem acesso ao importante histórico do estudante, este só disponível aos gestores e também espalhado em mais de um sistema. O modelo proposto integra diversas características do estudante nos 7 atributos selecionados que, além dos requisitos conhecidos como nota média das atividades e frequência, trabalha com 4 atributos resultantes da transformação de 9 variáveis originárias do log, o grupo formado e o rótulo de classificação. Acredita-se que os resultados obtidos a partir da aplicação do modelo apresentado auxiliem os docentes e gestores das instituições que oferecem cursos em EaD.

Dessa forma, prova-se parcialmente a hipótese inicial que a metodologia apresentada permite predizer em 86.5% o sucesso ou não do estudante em EaD na conclusão do curso em até 10% do início do curso, baseado nas interações com a plataforma do AVA.

O cluster_0 em S1, S2 e S3 representa 30 a 36% do total de classe P, ou seja, um terço dos que não concluem o curso tem taxas muito baixas de acesso ao AVA já desde o início do curso e se mantém com esse comportamento nos semestres seguintes, renovando matrícula só para manter o vínculo com a instituição ou abandona o curso. Ante essa situação seria necessário apontar melhorias no processo seletivo ou em alguma ação inicial na entrada dos estudantes, pois eles ocuparam vagas que outros interessados poderiam ocupar.

Também se observou que os atributos dos dados de rastreamento (Tabela 3), relacionados aos registros de log, têm uma relevância notavelmente maior do que os dados institucionais (Tabela 2) para predizer estudantes em risco. Atributos como idade, renda e outros dados socioeconômicos resultaram ser pouco correlacionados com a classe para a amostra analisada. Isso pode ser parcialmente explicado pela ausência de dados por ser uma informação de preenchimento opcional, ou informação com muitas ocorrências de um mesmo valor.

Os *insights* de um trabalho com agrupamento são mais amplos que na classificação, pois podemos analisar características de cada grupo que identificam o comportamento dos estudantes e sua evolução durante a linha do tempo. Além disso, permite analisar quais instâncias/estudantes formam parte em cada grupo, sendo possível uma análise individual de rastreamento de comportamento. Para isso, é necessário a interpretação dos dados de um especialista, diferente de uma classificação que objetiva minimizar a taxa de erros.

Especificamente, a abordagem em duas etapas adotada nesta tese pode, sem dúvida, atentar para a importância do agrupamento de estudantes ao tratar problemas relacionados à predição. Todavia, o peso dos atributos e recursos usados para treinar os modelos pode variar consideravelmente de um grupo de estudantes para outro.

É oportuno lembrar que os estudantes das turmas analisadas frequentam o ensino médio regular e, simultaneamente, fazem um curso para obter o título técnico. O ensino médio é uma etapa de difícil preparação para o vestibular e para os que precisam iniciar no mercado de trabalho, contribuindo para a evasão do curso técnico. Por outro lado, eles são motivados a ter essa formação e o diploma, que os prepara melhor para conseguir um emprego, preparar-se para os estudos superiores e aumentar a autoestima.

Por fim, um detalhe apreendido neste trabalho, além do estado da arte, das técnicas para manipular banco de dados e das técnicas e ferramentas em AM, foi compreender a importância

da escolha de tipos de gráficos e dados exibidos, seguindo Knaflic (2019), com técnicas que representem, da melhor forma possível, a história a ser contada e transmitida através dos dados.

6.1 Limitações do Trabalho

Este estudo deve ser interpretado dentro das seguintes limitações:

(a) os resultados podem estar sujeitos aos mecanismos automatizados da ferramenta Rapid Miner. O *Auto Model* do Rapid Miner cria um modelo de forma automatizada, mas permite abrir o modelo (opção '*open process'*) e acessar ou alterar os operadores e hiperparâmetros;

(b) pode haver algum viés envolvido nos recursos inicialmente selecionados para construir o conjunto de dados a partir dos registros e decisões para lidar com os dados ausentes. Ao extrair os dados direto dos diversos sistemas de gerenciamento de dados, estes podem apresentar problemas de desbalanceamento, ou seja, uma desproporção entre as distribuições das classes P e N que influencie. O conhecimento é extraído estritamente a partir desses dados e depende da qualidade deles;

(c) utilizou-se a versão 2.5 do Moodle. Atualmente, está disponível a versão 3.9, que conta com documentações de apoio mais completas e a tabela que armazena o log mudou.

Neste trabalho foram apresentados os processos de pré-processamento dos dados, uma das etapas já determinadas no KDD, e a granularidade e a normalização de dados são importantes. São passos que requerem execuções cuidadosas em grande volume de dados e esta metodologia proposta não poderia ser utilizada em sistemas em tempo real, e recomenda-se executar, por exemplo, em períodos semanais ou quinzenais para criar um modelo atualizado com os grupos encontrados.

6.2 Dificuldades durante a execução

A primeira grande dificuldade, no início da pesquisa, foi encontrar documentações e um dicionário de dados sobre a base de dados Moodle no IFRO, que na época estava na versão 2.5 do Moodle, com 489 tabelas considerando as tabelas originais, as adicionadas para os *plugins* de terceiros e as utilizadas em customizações locais. O campus não tinha documentação formal referente ao *schema* do banco de dados e das tabelas utilizadas para customizar o AVA. Atualmente, a partir da versão 3.7 do Moodle, está disponível o dicionário de dados[28]. As

[28] https://www.examulator.com/er/

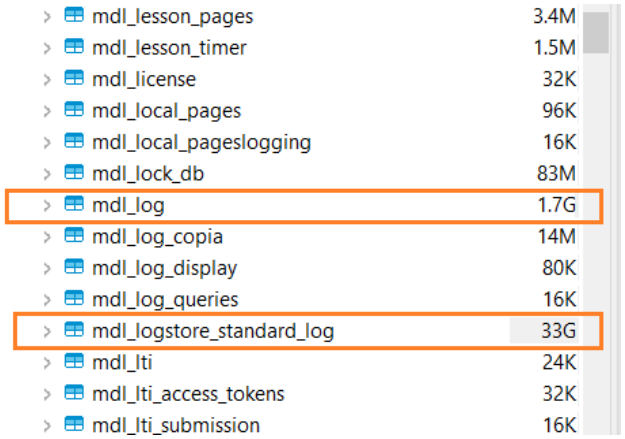
versões anteriores só contam com um diagrama resultante de uma engenharia reversa (ANEXO F).

Na base de dados do Moodle as chaves existentes são do tipo *sequence* para informar o PK (*primary key*) e não existem chaves estrangeiras, o que dificulta entender os relacionamentos entre tabelas sem ter a devida documentação.

Inicialmente, seriam analisadas mais turmas concomitantes anteriores a 2016 e cursos técnicos subsequentes EaD desde 2013, mas os *backups* restaurados tiveram um *gap* de 3 meses em 2015 e, por isso, ficou limitado aos 9 cursos estudados, que contam com o log completo.

A partir da versão 2.7 do Moodle teve atualizações na base de dados com tabelas novas e também colunas novas em tabelas existentes. Dentre essas mudanças a tabela de log mudou de mdl_log para mdl_logstore_standard_log (Figura 41).

Figura 41 – tabelas de log antes e depois da versão 2.7 do Moodle.



Nessa atualização, os logs antigos não foram migrados, pois são dados históricos que não impactam no funcionamento do sistema e, ao mesmo tempo, trata-se da maior tabela com mais registros na base, pois armazena aproximadamente 4 milhões de registros mensais para registrar as ações de todos os usuários, que geram vários GB de dados, enquanto outras tabelas ocupam alguns KB ou MB. Não seria eficiente executar *join* ou *union*, para consulta simultânea dessas tabelas em turmas que ficaram afetadas com essa migração. Portanto, a partir da versão 2.7 as queries criadas neste trabalho precisam de algumas alterações simples com consulta da nova tabela, pois as informações sobre módulos e ações continuam semelhantes e com mais informações armazenadas.

Os cursos analisados neste trabalho finalizaram antes dessa migração e considera a tabela mdl_log. Não foi possível analisar os logs dos cursos que concluíram a partir de 2020 por 2 motivos importantes: 1- a mudança na tabela de log no segundo semestre de 2019 e; 2- o

início da pandemia de Covid-19 em março de 2020, quando a equipe docente e técnica passou a se adaptar aos impactos e ter que trabalhar por videoconferências, em isolamento social completo, gerando, obviamente, impacto no replanejamento das atividades no AVA e nas interações dos estudantes, o que impactaria na análise das características de comportamento.

6.3 Contribuições da Tese

Ao concluir este trabalho, a sua principal contribuição foi o desenvolvimento e a especificação de uma metodologia que acompanha o comportamento do estudante para detecção de risco de evasão. Por causa da característica dinâmica de interações no AVA aplicaram-se técnicas não supervisionadas para identificar o comportamento e usar essa informação na entrada de dados em técnicas supervisionadas para predição de risco de evasão em cursos EaD, de forma que possa acompanhar a evolução de cada turma e curso, e não quando finalizado.

Outras contribuições que podem ser citadas são que:

- Disponibiliza-se em URL todos os títulos que foram pesquisados na revisão sistemática da literatura sobre soluções para predição de evasão em EaD usando técnicas de MDE e AM. Estão disponíveis em site[29] com um menu sobre os critérios de inclusão e exclusão mostrando os títulos selecionados em cada etapa.
- Apresenta-se uma base de conhecimento e referências no tema, utilizados nos capítulos 2 e 3.
- Oferece-se uma metodologia para gestão com mecanismos que automatizam ou auxiliam a predição de grupos de estudantes com risco de evasão. As técnicas sugeridas podem ser adaptadas em outros tipos de cursos, que sejam na modalidade EaD e usem a plataforma Moodle.
- Publicação de artigos em conferências e periódicos.

Dessa forma, acredita-se que este trabalho possa contribuir na tomada de decisões para uma intervenção precoce em estudantes com situação de risco de evasão, utilizando um metodologia que possa ser aplicado em cursos em andamento para realizar os agrupamentos e o modelo de predição seja ajustado a cada turma encerrada. Espera-se que seja uma ferramenta validada na prática trazendo resultados que permitam analisar ajustes e validação pela gestão.

[29] https://goo.gl/1U3AKS

6.4 Trabalhos Futuros

Com perspectivas de trabalhos futuros, pretende-se dar maior relevância aos estudos que pesquisem o ensino na modalidade EaD e disseminem métodos do uso conjunto de técnicas supervisionadas e não supervisionadas de AM.

Vale ressaltar que este estudo foi realizado com dados de turmas concluídas até o final de 2019, assim sendo, antes do início da pandemia de Covid-19. No entanto, novos estudos podem ser realizados para comparar e considerar o impacto do isolamento social ao longo do curso e como a adoção de novas tecnologias em salas de aula remotas afetou o comportamento dos estudantes. No caso dos cursos técnicos concomitantes estudados neste trabalho tinham um encontro presencial semanal e durante a pandemia passaram a ser um ou dois encontros síncronos remotos semanais, com adaptações tecnológicas e que devem ter gerado maior volume de registros no log do AVA.

Pretende-se criar mecanismos automatizados e contínuos de consulta dos resultados com o método proposto para dar suporte na tomada de decisões aos gestores acadêmicos.

Novos estudos podem ser realizados considerando o k=4 para formar grupos por similaridade de comportamento nos intervalos de tempo definidos. A escolha do valor 4 se deve a que neste trabalho descobriram-se 4 grupos em S*, S0, S2 e S3 e só em S1 ficou com 3 grupos. Assim sendo, a mesma quantidade de formação de grupos em cada período facilitaria uma análise comparativa.

Podem ser realizadas novas abordagens nos algoritmos de classificação com ajustes na matriz de custo para colocar peso diferenciado em predições erradas de classes rotuladas. Neste trabalho foram utilizados os valores padrões (Figura 42) e seria interessante trabalhar com mais classes. Por exemplo, separar os reprovados e evadidos dentro da classe P, assim como colocar pesos (custos) diferenciados para predições erradas, como no exemplo da Figura 43. Dessa forma, fazer novas comparações e análises de agrupamentos e classificações.

Figura 42 – Valor padrão para acertos e erros nas predições.

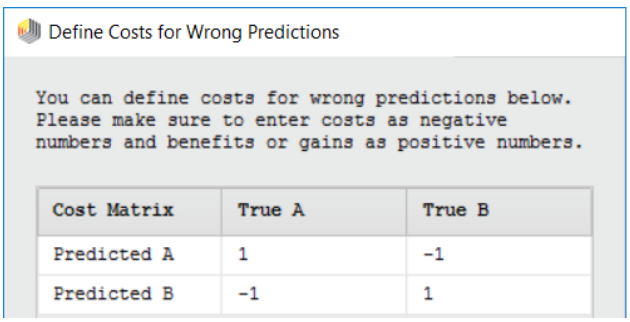
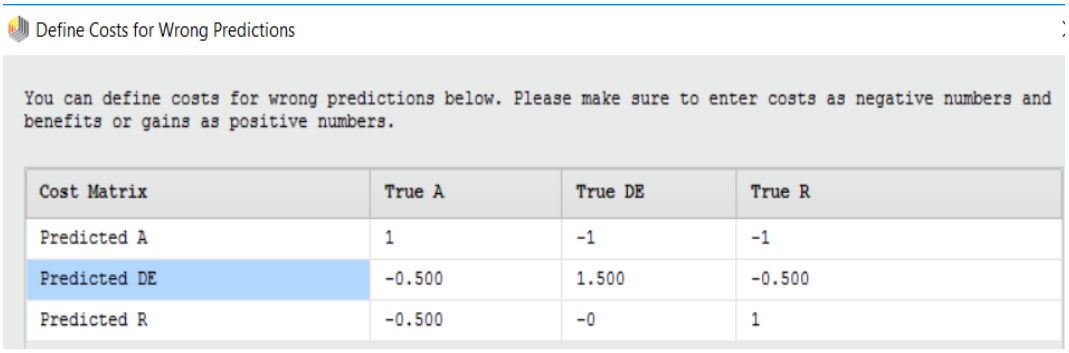


Figura 43 - Definir pesos diferentes nos acertos e erros nas predições.



Existem outras informações na base de dados que podem ser exploradas. Por exemplo, no caso dos dados socioeconômicos do estudante, atualmente, existe um novo sistema que valida os dados e se tornaram de preenchimento obrigatório no ato da matrícula. Esses tipos de atributos hão sido desconsiderados neste trabalho pela falta de dados ou por erro no preenchimento nas fichas que eram utilizadas na época da pesquisa, mas é importante que sejam avaliadas porque levam em consideração a grande diversidade da população brasileira para correlação dos atributos e, também, pode contribuir com mais pesquisas na Região Norte, pois segundo Colpo *et al*. (2020), a maioria das pesquisas desse tipo se concentra nas regiões Sul, Sudeste e Nordeste.

6.5 Lista de Publicações

Esta seção apresenta a lista de publicações obtidas no decorrer desta pesquisa, relacionadas ao tema deste trabalho.

As publicações seguem um delineamento da evolução desta tese, começando com uma revisão sistemática da literatura (RSL) sobre predição e técnicas de AM para redução de evasão na modalidade EaD. Depois foi publicado um trabalho sobre análise de registros de log do AVA para análise de desempenho de estudantes por meio de técnicas supervisionadas e, mais recente, a publicação em periódico em edição especial sobre ‘Aprendizagem de Máquina em Mineração de Dados Educacionais’, cujos editores, Kostopoulos e Kotsiantis, são referências importantes. Esta última publicação detalha a pesquisa sobre os estudantes do ensino técnico usando as técnicas não supervisionadas e supervisionadas, de forma separada. Nesta tese os trabalhos avançaram para experimentar essas duas técnicas em duas etapas consecutivas.

- TAMADA, Mariela Mizota; NETTO, José Francisco de M.; DE LIMA, Dhanielly P. R. **Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review** In: 2019 IEEE Frontiers in Education Conference (FIE), 2019, Cincinnati, USA. (APÊNDICE B)
- TAMADA, Mariela Mizota; GIUSTI, Rafael; NETTO, José Francisco de M. **Predicting Student Performance Based on Logs in Moodle LMS** In: 2021 IEEE Frontiers in Education Conference (FIE), 2021, Lincoln, USA. (APÊNDICE C)
- TAMADA, Mariela Mizota; GIUSTI, Rafael; NETTO, José Francisco de M. **Predicting Students at Risk of Dropout in Technical Course Using LMS Logs**. MDPI ELECTRONICS, Special Issue "Machine Learning in Educational Data Mining", v.11, p.468 - 490, 2022.(APÊNDICE D)

Referências Bibliográficas

ABED. **Censo EAD.BR. 2017-2018**. **Relatório Analítico de aprendizagem a distância no Brasil** Disponível em: < http://www.abed.org.br/site/pt/midiateca/censo_ead/>. Acessado em 30 de nov. de 2018.

_____. (2021) **Censo EaD.BR 2019-2020. Relatório Analítico de aprendizagem a distância no Brasil**. Disponível em <http://abed.org.br/arquivos/CENSO_EAD_2019_PORTUGUES.pdf>

Albreiki B, Zaki N, Alashwal H. (2021). **A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques**. *Education Sciences*. 11(9):552. https://doi.org/10.3390/educsci11090552

Alturki, U.; Aldraiweesh, A. (2021). **Application of Learning Management System (LMS) during the COVID-19 Pandemic: A Sustainable Acceptance Model of the Expansion Technology Approach**. Sustainability 2021, 13, 10991. https://doi.org/10.3390/ su131910991.

Alves, L; Barros, D; Okada, A. (Org.) (2009). **Moodle: estratégias pedagógicas e estudo de caso**. Salvador: EDUNEB, 2009. ISBN: 978.85.7887-001-0.

Amaral, F. (2016) **Aprenda Mineração de Dados: Teoria e prática**. Rio de Janeiro: Alta Books Editora.

Amershi, S., Conati, C. (2009). **Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments**. Journal of Educational Data Mining, 1(1):18-71.

Baas, A. (1991). **Promissing Strategies for at-risk youth**". ERIC Digest, no. 59.

Bacich, L., Neto, A. T., & de Mello Trevisani, F. (2015). **Ensino híbrido: personalização e tecnologia na educação**. Penso Editora.

Baker, R.; Isotani, S.; Carvalho, A. (2011). **Mineração de dados educacionais: Oportunidades para o Brasil**. Brazilian Journal of Computers in Education, 19(02):11, 3-13. http://dx.doi.org/10.5753/RBIE.2011.19.02.03

Barbosa, F. R. (2020**). Análise de dados para identificação do perfil de alunos evadidos do curso Técnico em Informática do IFNMG - Campus Januária**. Dissertação. Universidade Federal dos Vales do Jequitinhonha e Mucuri. http://acervo.ufvjm.edu.br/jspui/handle/1/2526

Bardagi, M. P.; Hutz, C. S. (2009). **"Não havia outra saída": percepções de alunos evadidos sobre o abandono do curso superior**. Psico-USF, v. 14, n. 1, p. 95-105, jan./abr. 2009 p.95. https://doi.org/10.1590/S1413-82712009000100010.

Barnes, T.; Desmarais, M.; Romero, C.; Ventura, S. (2009). **Educational Data Mining** 2009: 2nd International Conference on Educational Data Mining, Proceedings. Cordoba, Spain.

Bellman, R. E. (1961**). Adaptative control processes**. In Adaptive Control Processes. Princeton university press. https://doi.org/10.1515/9781400874668

Bienkowski, M.; Feng, M.; Means, B. (2012). **Enhancing teaching and learning through educational data mining and learning analytics: an issue brief**. Washington, D.C.: U.S.

Department of Education, 2012. Relatório. Disponível em: <https://tech.ed.gov/wp-content/uploads/2014/03/edm-la-brief.pdf>. Acesso em: 13 nov. 2021.

Bittencourt, I. M.; Mercado, L. P. L.(2014). **Evasão nos cursos na modalidade de educação a distância: estudo de caso do Curso Piloto de Administração da UFAL/UAB**. https://doi.org/10.1590/S0104-40362014000200009

BRASIL. MEC. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. Relatório da comissão especial de estudos sobre evasão nas universidades públicas brasileiras. Brasília, 1997. http://www.dominiopublico.gov.br/download/texto/me002240.pdf

BRASIL. **Lei 11.892 de 29 de Dezembro de 2008**. Institui a Rede Federal de Educação Científica e Tecnológica, cria os Institutos Federais de Educação Ciência e Tecnologia, e dá outras providências.

_____. **Decreto no.9.057, de 25 de maio de 2017**, regulamenta o art. 80 da Lei nº 9.394, de 20 de dezembro de 1996 , que estabelece as diretrizes e bases da educação nacional.

_____. **Decreto n° 9.570, de 22 de novembro de 2018**. Lei do Aprendiz.

_____. Ministério da Educação. Conselho Nacional de Educação. **Resolução Nº 6, de 20 de setembro de 2012**. Diretrizes Curriculares Nacionais para a Educação Profissional Técnica de Nível Médio.

_____. Ministério da Educação. Conselho Nacional de Educação. **Resolução Nº 3, de 21 de novembro de 2018**. Atualiza as Diretrizes Curriculares Nacionais para o Ensino Médio.

_____. Ministério da Educação. Conselho Nacional de Educação. **Resolução CNE/CP Nº 1, de 5 de janeiro de 2021**. Diretrizes Curriculares Nacionais Gerais para a Educação Profissional e Tecnológica.

Burgos, C.; Campanario, M.L.D.; Peña, D.L.; Lara, J.A.; Lizcano, D.; Martínez, M.A. (2018). **Data mining for modeling students performance: A tutoring action plan to prevent academic dropout**. Comput. Electr. Eng. 2018, 66, 542–556. https://doi.org/10.1016/j.compeleceng.2017.03.005.

Cambruzzi, W.; Cazella, S.C.; Rigo, S.J.; Barbosa, J.L.V.; (2014). **Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios**. Revista Brasileira de Informática na Educação, Volume 22, Número 1, 2014. http://dx.doi.org/10.5753/rbie.2014.22.01.132

Cerezo, R., Sanchez-Santillan, M., Paule-Ruiz, M.P.; Núñez, J.C. (2016). **Students' LMS interaction patterns and their relationship with achievement: A case study in higher education**. Comput. Educ. 2016, 96, 42–54. doi.org/10.1016/j.compedu.2016.02.006.

Cervo, A. L.; Bervian, P. A.; Da Silva, R. (2007). **Metodologia científica.** 6 ed. São Paulo: Pearson Prentice Hall, 2007.

Chen, Y.; Chen, Q.; Zhao, M; Boyer, S.; Veeramachaneni, K.; Qu, H. (2016). **DropoutSeer: Visualizing learning patterns in Massive Open Online Courses for dropout reasoning and prediction**. 2016 IEEE Conference on Visual Analytics Science and Technology (VAST).

Colpo, M.P., Primo, T.T., Pernas, A.M., Cechinel, C. (2020). **Mineração de Dados Educacionais na Predição de Evasão: uma RSL sob a Perspectiva do Congresso Brasileiro de Informática na Educação**. IX Congresso Brasileiro de Informática na Educação. 24-11-2020. https://doi.org/10.5753/cbie.sbie.2020.1102

Corcovia, L. O.; Alves, R. S. (2019). **Aprendizagem de Máquina e Mineração de Dados: avaliação de métodos de aprendizagem.** Revista Interface Tecnológica, [S. l.], v. 16, n. 1, p. 90-101, 2019. Disponível em: https://revista.fatectq.edu.br/index.php/interfacetecnologica/article/view/562. Acesso em: 17 jul. 2021.

Cunha, J. A.; Moura, E.; Analide, C. (2016).**Data mining in academic databases to detect behaviors of students related to school dropout and disapproval**. Advances in Intelligent Systems and Computing, v. 445, p. 189–198.

Cutler, A. (2010). **Random forest for regression and classification**. Retrieved from internal-pdf://semisupervised-3254828305/semisupervised.ppt

Da Silva, G. J.; Dos Santos, S.C.; Battestin, V.; Zamberlan M.F. (2020). **Diretrizes para Educação a Distância da Rede Federal de Educação Profissional e Tecnológica**: módulo de legislação, CONIF, Vitória, ES: Edifes.

Da Silva, L. M.; Dias, L. P. S.; Rigo, S.; Barbosa, J. L. V.; Leithardt, D. R.; Leithardt, V. R. Q. (2021). **A literature review on intelligent services applied to distance learning**. Education Sciences, 11(11), 666.

EDM. Acessado em http://www.educationaldatamining.org/

Estivil-Castro, V. (2002). **Why so many cluster algorithms** - a position paper. SIGKDD Explorations 4(1): 65-75.

Faceli, K.; Lorena, A. C.; Gama, J.; Almeida, T.A.; de Carvalho André C. P. L. F. (2017). **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. Rio de Janeiro. LTC.

Favero, R.V.M. (2006). **Dialogar ou evadir: Eis a questão! Um estudo sobre a permanência e a evasão na Educação a Distância**. 2006. 167 f. Dissertação (Mestrado em Educação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

Fayyad, U. M. Shapiro, G. P. Smyth, P. (1996). **From Data Mining to Knowledge Discovery in Databases.** American Association for Artificial Intelligence, 17, 37-54.

Fei, M.; Yeung, D.-Y. (2015). **Temporal Models for Predicting Student Dropout in Massive Open Online Courses**. 2015 IEEE 15th International Conference on Data Mining Workshops.

Fonseca, J. J. S. (2002). **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002. Apostila.

Francis, B.; Babu, S.S. (2019). **Predicting Academic Performance of Students Using a Hybrid Data Mining Approach**. Journal of Medical Systems. 43 (6). 162. 10.1007/s10916-019-1295-4.

Friedman, J. Hastie, T., Tibshirani, R. (2001). **The elements of statistical learning 1**, Springer Series in Statistics, New York, NY, USA.

Gerhardt, T.E. e Silveira, D.T. (2009). **Métodos de pesquisa** / [organizado por] coordenado pela Universidade Aberta do Brasil – UAB/UFRGS e pelo Curso de Graduação Tecnológica. Porto Alegre: Editora da UFRGS, 2009.

Gil, A. C. (1999). **Métodos e técnicas de pesquisa social**. 5. ed. São Paulo: Atlas, 1999.

_____. (2007). **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

Goldschmidt, R.R., Passos, E.; Bezerra E. (2015) **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. 2a. ed. Elsevier, Rio de Janeiro.

Han, J.; Pei, J.; Kamber, M. (2011). **Data Mining: Concepts and Techniques**; Elsevier: Amsterdam, The Netherlands, 2011; p. 744.

Handl, J.; Knowles, J. (2007). **An evolutionary approach to multiobjective clustering**. In: IEEE Congress on Evolutionary Computation, 11(1):56-76.

Harrel, F.E. (2011). **Regression modeling strategies.** Rev Esp Cardiol 64(6):501–507. doi:10.1016/j.recesp.2011.01.019.

He, X.; Zhao, K.; Chu, X. (2021). **AutoML: A Survey of the State-of-the-art**. Knowledge-Based Systems 212 (2021), 106622. https://doi.org/10.48550/arXiv.1908.00709

Heykin, S. (2009). **Neural Networks And Learning Machine**. 3 Ed. Prentice Hall.

Hong, B., Wei, Z. and Yang, Y. (2017). **Discovering learning behavior patterns to predict dropout in MOOC**. The 12th International Conference on Computer Science & Education (ICCSE 2017).

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). **Developing early warning systems to predict students' online learning performance**. Computers in Human Behavior, 36, 469–478.

Hung, J.-L.; Zhang, K. (2008). **Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching**. MERLOT J. Online Learn. Teach. 2008, 4, 426–437.

Iatrellis, O.; Savvas, I.K.; Fitsilis, P.; Gerogiannis, V.C. (2021). **A two-phase machine learning approach for predicting student outcomes**. Educ Inf Technol 26, 69–88. https://doi.org/10.1007/s10639-020-10260-x

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira -INEP (2020). **Avaliação da Educação profissional e tecnológica: um campo em construção**. Org. Moraes, G.H. *et al*. Brasília

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (2021). **Censo Escolar**. **Educação profissional.** Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar/resultados. Acessado em: 02/fev./2022.

Isidro, C.; Carro, R.M.; Ortigosa, A. (2018). **Dropout detection in MOOCs: An exploratory analysis**. SIIE 2018 - 2018 International Symposium on Computers in Education.

Jain, A. K. (2010). **Data clustering: 50 years beyond K-means. Pattern Recognition Letters**. 31(8):651-666 < https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323>. Acessado em 20 nov.2021. ISSN: 0167-8655. https://doi.org/10.1016/j.patrec.2009.09.011.

Jalloule, J.; Sallé, J.; Bittencourt, R. (2014) **RapidMiner Aprenda a Usar**. Disponível em: https://prezi.com/-yo8qjamdbbq/rapidminer-aprenda-a-usar/

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013). **An introduction to statistical learning**, Springer.

Jiménez-Gómez, M.; Luna, J. M.; Romero, C.; Ventura, S. (2015). **Discovering Clues to Avoid Middle School Failure at Early Stages**, in Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (New York, NY, USA: Association for Computing Machinery), 300–304. doi:10.1145/2723576.2723597

Kampff, A.J.C. (2009). Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente. Tese (doutorado). UFRGS, Porto Alegre-RS, 189p.

Kemczinski, A. (2005). **Método de Avaliação para Ambientes E-Learning**. Tese Doutorado em Engenharia da Produção da UFSC – Universidade Federal de Santa Catarina, Florianópolis-SC, 205 p.

Kitchenham, B. E.; Charters, S. (2007). **Guidelines for performing Systematic Literature Reviews in Software Engineering**. Version 2.3. Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele, Reino Unido.

Kleinberg, J. (2002). **Learning drifting concepts: Example selection vs. example weighting**. Intelligent Data Analytics, 8(3): 281-300.

Knaflic, C.N. (2019). **Storytelling com dados: um guia sobre visualização**. ISBN: 978-85-508-0468-2. Alta Books, Rio de Janeiro (RJ), 242p.

Kotsiantis, S.B. (2011). **Decision trees: a recent overview**. Artif Intell Rev (2013) 39:261–283. DOI 10.1007/s10462-011-9272-4

Kotsiantis, S.B. (2016). **Supervised Machine Learning: A Review of Classification Techniques**. IJCSIT, Vol. 7 (3), 2016, pp.1174-1179.

Laveti, R.N., Kuppili, S., Pal, J.Ch.S.N., Babu, N.S.C. (2017). **Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing**, 978-1-5386-1922-3/17, 2017.

Liñán, L. C. Perez, A. J. (2015). **Educational Data Mining and Learning Analytics: differences, similarities, and time evolution**. RUSC.Universities and Knowledge Society Journal, 12(3). pp. 98-112. doi: http://dx.doi.org/10.7238/rusc.v12i3.2515.

Linke, E. C.; Nogueira, B. C. (2017). **A evasão escolar no ensino técnico profissionalizantes**. In: SEMINÁRIO INTERINSTITUCIONAL DE ENSINO, PESQUISA E EXTENSÃO, 22., 2017. Cruz Alta. Anais... Cruz Alta: Unicruz, 2017. p. 1-14.

López-Zambrano, J., Lara, J. A., & Romero, C. (2020). **Towards portability of models for predicting students' final performance in university courses starting from Moodle logs**. Applied Sciences, 10(1), 354. https://doi.org/10.3390/app10010354.

Louppe, G. (2014) **Understanding random forests: from theory to practice**. Dissertation, University of Lie`ge. doi:10.13140/2.1.1570.5928

Lu, X., Wang, S., Huang, J., Chen, W., Yan, Z. (2017). **What decides the dropout in MOOCS?**. DASFAA 2017 Workshops, LNCS 10179, pp. 316–327

Lüscher, A. Z.; Dore, R**.** (2011). **Política educacional no Brasil: educação técnica e abandono escolar**. Revista Brasileira de Pós Graduação, Brasília, DF, v. 8, n. 1, p. 147-176. doi:10.21713/2358-2332.2011.v8.244.

Macfadyen, L. P., & Dawson, S. (2010). **Mining AVA data to develop an "early warning system" for educators: A proof of concept**. Computers & Education, 54(2), 588–599.

Mandala, S., Abdullah, A. H. and Ismail, A. S. (2013). **A Survey of E-Learning Security**. In: Proceedings of International Conference on ICT for Smart Society (ICISS), pp. 1-6, Jacarta.

Manhães, L.M.B. Cruz, S.M.S. Zimbrão, G. (2014). **WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM**.UFRJ-RJ.

Marek, M.W.; Chew, C.S. (2021). **Teacher Experiences in Converting Classes to Distance Learning in the COVID-19 Pandemic**. Int. J. Distance Educ. Technol. 2021, 19, 40–60. https://doi.org/10.4018/IJDET.20210101.oa3.

Márquez-Vera, C. A., Cano, A. Romero, C., Noaman, A.Y.M, Fardoun, H. M., Ventura, S. (2016). **Early dropout prediction using data mining: a case study with high school students**. Expert Systems, February 2016, Vol. 33, No. 1. Wiley Publishing Ltd..

Medeiros, J.C. (2019). **Tensões e Desafios no Processo de Institucionalização dos Cursos Técnicos a Distância no Instituto Federal de Brasília**. VI Seminário Internacional sobre Profissionalização Docente – SIPD- Cátedra UNESCO – CIERS-ed/FCC

Mehta, A. A.; Buch, N. J. (2016). **Depth and breadth of educational data mining: Researchers' point of view.** 2016 10th International Conference on Intelligent Systems and Control (ISCO).

Meira, C. A. (2015). **A evasão escolar no ensino técnico profissionalizante**: um estudo de caso no campus Cariacica do Instituto Federal do Espírito Santo. Dissertação (Mestrado em Gestão Pública),Vitória (ES).

Mitchell, T. M. (1997) **Does machine learning really work?.** AI magazine, 18(3), 11-11. https://doi.org/10.1609/aimag.v18i3.1303

MOODLE: Open-source Learning Platform. Disponível em <https://moodle.org/>. Acessado em: novembro de 2018.

Moran, J. (2002). O que é educação a distância. Rio de Janeiro. Disponível em: < http://www2.eca.usp.br/moran/wp-content/uploads/2013/12/dist.pdf>. Acessado em: 20 de nov. de 2021.

Moran, J. (2009). **Modelos e Avaliação do Ensino Superior a distância no Brasil**. ETD – Educação Temática Digital, Campinas, v.10, n.2, p.54-70, jun.

Neville P.G (1999). **Decision trees for predictive modeling**. SAS Institute Inc., pp 1–24

OCDE (2014). Education at a Glance 2015. Disponível em https://www.oecd.org/brazil/Education-at-a-glance-2015-Brazil-in-Portuguese.pdf

OCDE (2017). Education at a Glance 2017. https://www.oecd-ilibrary.org/education/education-at-a-glance-2017_eag-2017-en

OCDE (2021). **Education Policy Outlook: Brazil** - com foco em políticas internacionais. Disponível em: https://www.oecd.org/education/policy-outlook/country-profile-Brazil-2021-INT-PT.pdf. Acessado em 20 de fev. 2022.

Oeda, S; Hashimoto, G. (2017). **Log-Data Clustering Analysis for Dropout Prediction in Beginner Programming Classes**. Procedia Computer Science, Vol. 112, pag. 614-621.

Papert, S. (1997). **A família em rede**: ultrapassando a barreira digital entre gerações. Título original: The Connected Family: bridging the digital generation gap. Lisboa: Relógio D'Água Editores, 1997.

Park, H. A. (2013). **An introduction to logistic regression**: from basic concepts to interpretation with particular attention to nursing domain. J Korean Acad Nurse 43(2):154–164. Retrieved from doi:10.4040/jkan.2013.43.2.154

Piaget, J. (1976). **A equilibração das estruturas cognitivas**: problema geral do desenvolvimento. Rio de Janeiro: Zahar, 1976.

Queiroga, E.M.; Cechinel, C.; Araújo, R. (2017). **Predição de estudantes com risco de evasão em cursos técnicos a distância**. SBIE 2017. http://www.br-ie.org/pub/index.php/sbie/article/view/7686

Qiu, L.; Liu, Yanshen.; Liu, Yi. (2018). **An Integrated Framework with Feature Selection for Dropout Prediction in Massive Open Online Courses.** doi:10.1109/ACCESS.2018.2881275, IEEE Access

RapidMiner Studio. (2014) **RapidMiner Studio Manual**. Boston. Disponível em: <https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf>. Acessado em: 05 de mar. 2021

Rego, F.A.; Rosas, I.R de C.; Prados, R.M.N. (2021). **Educação Profissional e Tecnológica como alternativa de acesso ao mercado de trabalho**. Brazilian Journal of Development, Curitiba, v.7, n.2, p. 14585-14596 feb. 2021. DOI:10.34117/bjdv7n2-198

Riestra-González, M.; Paule-Ruíz, M. del P.; Ortin, F. (2020). **Massive LMS log data analysis for the early prediction of course-agnostic student performance**. Computers & Education 163 (2021) 104108. Elsevier, 2020. https://doi.org/10.1016/j.compedu.2020.104108

Romero, C.; Ventura, S. (2013). **Data mining in education**. Wiley Interdisciplinary. Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Romero, C.; López, M.-I.; Luna, J.-M.; Ventura, S. (2013). **Predicting students' final performance from participation in on-line discussion forums**. Comput. Educ. 2013, 68, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009.

Santos, E. M. dos et al. (2008). **Evasão na Educação a Distância: identificando causas e propondo estratégias de prevenção.** Revista Brasileira de Aprendizagem Aberta e a Distância, São Paulo, p. 1-10, maio 2008.

Sha, L.; Looi, C.-K.; Chen, W. e Zhang, BH (2012). **"Understanding mobile learning from the perspective of self-regulated learning**", J. Comput. Assist. Aprender.vol. 28, 4, pp. 366-378, agosto de 2012. https://doi.org/10.1111/j.1365-2729.2011.00461.x

Sheather, S.J. (2009**). A Modern approach to regression with R**. Bimometrics 67(2):675–677 doi:10.1111/j.1541-0420.2011.01614.x

Seidman, A. (1996). **Retention revisited: R=E, id+E & In**, iv.College and University, 71(4):18–20.

Siemens, G.; Long, P. (2011). **Penetrating the Fog: Analytics in Learning and Education**. EDUCAUSE Review, v46 n5 p30-32, 34, 36, 38, 40 Sep-Oct 2011.

Siemens, G.; Baker, R. SJD. (2012). **Learning analytics and educational data mining: Towards communication and collaboration.** In Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge, LAK '12 ,pp. 252–254.

Silveira, P. D. N.; Cury, D.; Menezes, C.; Santos, O. L. dos (2019). **Analysis of classifiers in a predictive model of academic success or failure for institutional and trace data**. IEEE Frontiers in Education Conference (FIE), 2019, pp. 1-8, doi: 10.1109/FIE43999.2019.9028618.

Talavera, L.; Gaudioso, E. (2004). **Mining student data to characterize similar behavior groups in unstructured collaboration spaces**. In Proceedings of the Workshop on Artificial Intelligence in CSCL, 16th European Conference on Artificial intelligence, Valencia, Spain, 22–27, August 2004; pp. 17–23.

Tamada, M.M.; Netto, J.F.M.; Lima, D.P.R. (2019). **Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review**. In Proceedings of the 2019 IEEE Frontiers in Education Conference (FIE), Covington, KY, USA, 16–19 October 2019. https://doi.org/10.1109/FIE43999.2019.9028545.

Tang, J. K. T.; Xie, H.; Wong, T. (2017). **A Big Data Framework for Early Identification of Dropout Students in MOOC.** Springer-Verlag Berlin Heidelberg. J. Lam *et al*. (Eds.): ICTE 2015, CCIS 559, pp. 127–132, 2015.

Teruel, M; Alonso Alemany, L. (2018). **Co-embeddings for Student Modeling in Virtual Learning Environments**. 26th Conference on User Modeling, Adaptation and Personalization.

Triviños, A. N. S. (1987). **Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação**. São Paulo: Atlas, 1987.

Uddin, M.F.; Lee, J. (2017). **Proposing stochastic probability-based math model and algorithms utilizing social networking and academic data for good fit students prediction**. Journal Social Network Analysis and Mining (2017) 7:29. Springer-Verlag GmbH Austria, 2017.

Vellido, A., Castro, F., Etchells, T.A., Nebot, a., Mugica, F. (2007). **Data mining of virtual campus data. In Evolution of Teaching and Learning Paradigms in Intelligent Environment**. Studies in Computational Intelligence (SCI) 62, series Advanced Information and Knowledge Processing. Springer. 223-254.

Wolpert, D. H. (1996). **The Lack of A Priori Distinctions Between Learning Algorithms**. Neural Computation (1996) 8 (7): 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

Wazlawick, R.S.(2014). **Metodologia de pesquisa para ciência da computação**. 2a. edição. GEN LTC, 2014.168p.ISBN-13 978-8535277821

Witten, I. H., Frank, E., Hall, M. A.; Pal, C. J. (2016). **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann.

Xavier, V.L. (2012). **Resolução do Problema de Agrupamento segundo o critério de Minimização da Soma de Distâncias**. Dissertação (Mestrado) -COPPE, UFRJ, Rio de Janeiro, 2012.

Xing, W.; Du, D. (2018). **Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention.** Journal of Educational Computing Research. Journal of Educational Computing Research. . I-24, 2018, doi: 10.1177/0735633118757015

Yadav, S.; Shukla, S. (2016). **Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification**. In: Proceedings of the 6th international advanced computing conference, IACC 2016, (Cv), pp 78–83. doi:10.1109/IACC.2016.25

Zhou, ZH., Li, M. (2010). **Semi-supervised learning by disagreement**. Knowl Inf Syst 24, 415–439 https://doi-org.ez134.periodicos.capes.gov.br/10.1007/s10115-009-0209-z

APÊNDICE A – FIE 2019

Publicação em Conferência Internacional

Evento: Frontiers in Education (FIE2019). **Local**: Cincinnati, USA

Data: Outubro/2019.

Referência:

Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review

Mariela Mizota Tamada
Federal University of Amazonas (UFAM)
Institute of Computing
Manaus, Brazil
Institute Federal of Rondonia (IFRO)
Porto Velho, Brazil
mariela.tamada@ifro.edu.br

José Francisco de Magalhães Netto,
Dhanielly Paulina R. de Lima
Federal University of Amazonas (UFAM)
Institute of Computing
Manaus, Brazil
{jnetto, dhanielly}@icomp.edu.edu.br

***Abstract*— Context: This Research to Practice Full Paper presents a systematic review of methodologies that propose ways of reducing dropout rate in Virtual Learning Environments (VLE). This generates large amounts of data about courses and students, whose analysis requires the use of computational analytical tools. Most educational institutions claim that the greatest issue in virtual learning courses is high student dropout rates. Goal: Our study aims to identify solutions that use Machine Learning (ML) techniques to reduce these high dropout rates. Method: We conducted a systematic review to identify, filter and classify primary studies. Results: The initial search of academic databases resulted in 199 papers, of which 13 papers were included in the final analysis. The review reports the historical evolution of the publications, the Machine Learning techniques used, the characteristics of data used, as well as identifies solutions proposed to reduce dropout in distance learning. Conclusion: Our study provides an overview of the state of the art of solutions proposed to reduce dropout rates using ML techniques and may guide future studies and tool development.**

Keywords- dropout prediction; machine learning; distance education; online learning; systematic review.

I. INTRODUCTION

Distance education has surged with the gradual progress in information technology over the past decade and the Scholar Dropout phenomenon has been increasing, having repercussions in social, economic, and academic aspects, among others. These changes have created new challenges for different stakeholders in managing the learning process through a virtual platform.

A Virtual Learning Environment (VLE) is a virtual classroom that allows teachers and students to communicate with each other online. Class information, learning materials, and assignments are provided via the Web. The amount of data collected through educational database technologies is increasing rapidly in volume and complexity, which allows for statistical analysis, data mining and predictive actions.

In the last 10 years, Education Data Mining (EDM) has emerged as a new area concerned with the application of Data Mining (DM), Machine Learning (ML), and statistics of information generated from an educational setting [1], and the knowledge discovered may help improve teaching/learning processes. This study analyzed ML techniques in VLE.

There are several studies that compare the performance of algorithms used in student performance prediction or dropout prediction systems, but this study looks for concrete solutions with a method that is designed or implemented using ML techniques. Thus, this paper aims to conduct a systematic review of the information obtained in the literature about solutions that reduce the high dropout rates, leading to the need for early identification of students who are likely to dropout and the possibility of providing teachers, tutors, and managers with strategic information to identify possible dropouts, which helps in decision making about adequate pedagogical intervention.

This paper is organized as follows: Section II presents the background, Section III presents related works, Section IV details the research method, Section V presents data analysis, discusses the results, and presents an overview. Finally, section VI concludes the paper and presents further research and challenges of dropout prediction models using ML techniques.

II. BACKGROUND

This section presents two basic terms that are used in this paper: the learning platforms used in distance learning and the ML concepts, the main prediction techniques that, along with data mining, develop methods to discover patterns that lead to knowledge.

A. Learning Platforms

Among the platforms or VLEs are Moodle[1], which is most commonly used to adapt traditional face-to-face courses to blended learning or entirely online courses, and MOOC[2], which can simultaneously reach thousands of students in several

[1] Modular Object-Oriented Dynamic Learning Environment, a free software created in 2001. moodle.org

[2] Massive Open Online Course, created in 2012 for free and open online courses. mooc.org

978-1-7281-1746-1/19/$31.00 ©2019 IEEE

APÊNDICE B – FIE 2021

Publicação em Conferência Internacional

Evento: Frontiers in Education (FIE2021). **Local**: Lincoln, Nebraska, USA

Data: Outubro/2021.

Referência:

PREDICTING STUDENT PERFORMANCE BASED ON LOGS IN MOODLE LMS

Mariela Mizota Tamada
Institute of Computing
Federal University of Amazonas (UFAM),
Manaus, Brazil
Institute Federal of Rondonia, (IFRO),
Porto Velho, Brazil
mariela.tamada@ifro.edu.br

Rafael Giusti
José Francisco de Magalhães Netto
Institute of Computing
Federal University of Amazonas (UFAM),
Manaus, Brazil
{ rgiusti, jmetto }@icomp.ufam.edu.br

***Abstract*— Context: This innovative practice full paper presents a methodology to predict at-risk students in the context of a course assisted by an LMS (Learning Management System). LMSs generate large amounts of data about courses and students, which allows schools to make useful insights with the help of computational analytical tools. Most educational institutions claim that the most significant issue in virtual learning is high student dropout rates, and school performance is one of its main factors. Objective: Our study aims to use Machine Learning techniques based on logs from the Modular Object-Oriented Dynamic Learning Environment (Moodle). Those data are used to analyze student behavior and create a model that helps detect students at risk. Method: This paper used institutional data and trace data generated by LMS of a Computing education technical courses, blended and distance learning, at high school. We compared 7 algorithms with models trained at 6%, 20%, 40%, and 60% of the course duration, with the intent of exploring the compromise between early and late detection of at-risk students. Our model has 69% positive classe (failed) and 31% negative class (passed), and the false positives cost is important. Results: The results show 7 created models of predicting. The findings for Random Forest performed the best when predicting a student's performance. Conclusion: Our study provides a student at-risk prediction model using ML techniques on logs in Moodle LMS and may guide future studies and tool development to reduce these high dropout rates.**

Keywords- performance prediction; Machine Learning; Learning Management Systems; logs in Moodle.

I. INTRODUCTION

Distance education has improved with the gradual advancement of information technology in the past decade, and to avoid school dropout, there is research on models for predicting academic performance. If the institution detects at-risk students in the early stages of the course, there will be more time to make an intervention that may improve their performance and help them complete the course successfully.

This work presents Machine Learning (ML) techniques for analysis of registration data and logs files of a Learning Management System (LMS), which is a virtual classroom environment. LMSs enables communication between teacher and students, and provides a space for students to access learning materials and course activities.

All interactions in the LMS generate a log, which stores information in a database. With this, the amount of data collected is rapidly increasing in volume and complexity, allowing statistical analysis, data mining, and building predictive models of school performance. Some surveys create predictive models with the log files generated in the course. Still, these models fail to generalize to an early prediction because the actual log information is different from that used to train the models. In this work, we use Machine Learning to predict academic performance in the LMS, analyzing the socio-economic profile and the LMS log files generated up to the forecast time. In other words, a model that fits the characteristics and the progress of the course in the passage of the first subjects.

Concurrent enrollment, more commonly known as dual enrollment, refers to programs where students are enrolled in two schools simultaneously in a high school environment, one in regular high school and another one in technical courses, offered in blended learning (b-learning) or distance learning (e-learning), which duration is 2 years. The educational institution analyzed has more than 600 campuses in Brazil, at different levels of education, ranging from integrated high school to graduate school, and courses take place in the traditional classroom, blended and distance learning formats. Technical education concomitant with high school is present in several of them.

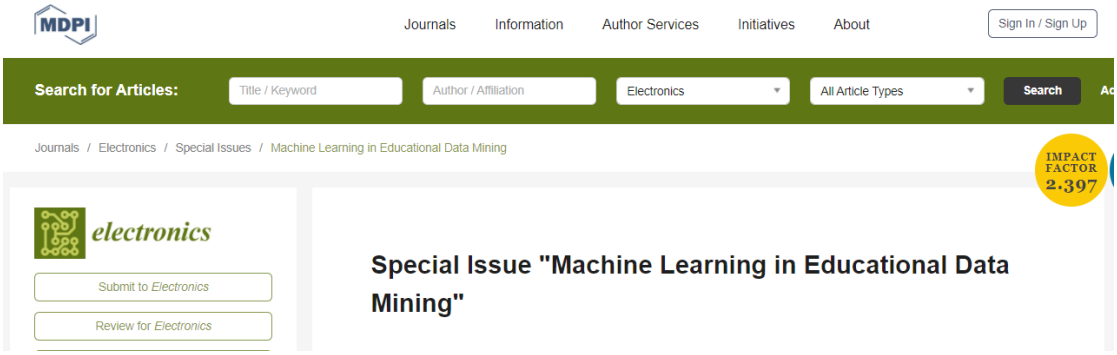
In this work, the evaluated courses had, on average, 69% of failure or dropout, a high rate. Students who are unsuccessful in their studies lose time and effort in their failed searches, and they and their families can suffer financially and emotionally. Institutions also lose the scarce resources they have invested in.

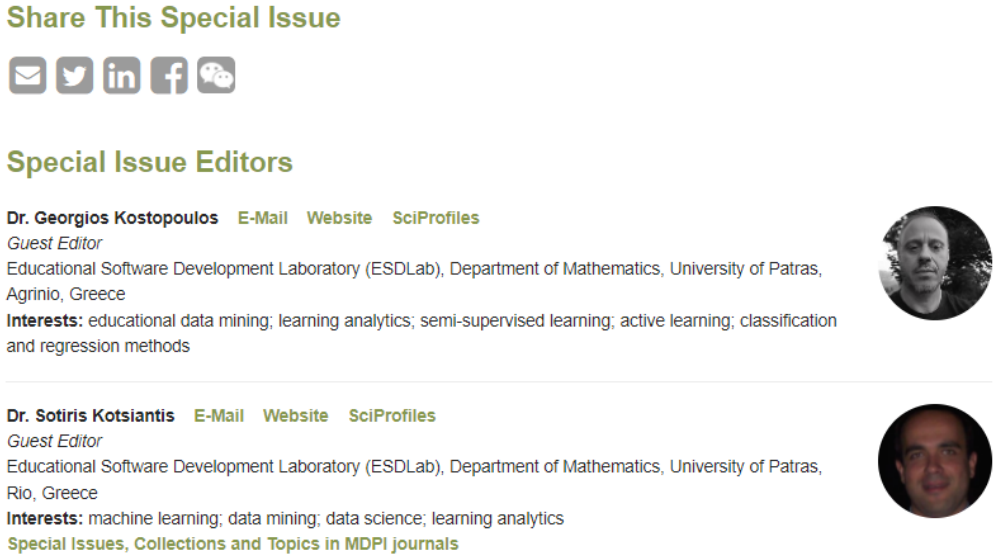
The application of ML and statistics of information generated from an educational setting [1] and the knowledge discovered may help improve teaching/learning processes. This study analyzed ML techniques in LMS with the objective of early identification of students with high probability of evasion. Early identification of at-risk students may provide teachers, tutors, and managers with strategic information that helps in making decisions for appropriate pedagogical interventions.

APÊNDICE C –Journal MDPI, edição especial “MachineLearning in Educational Data Mining”.

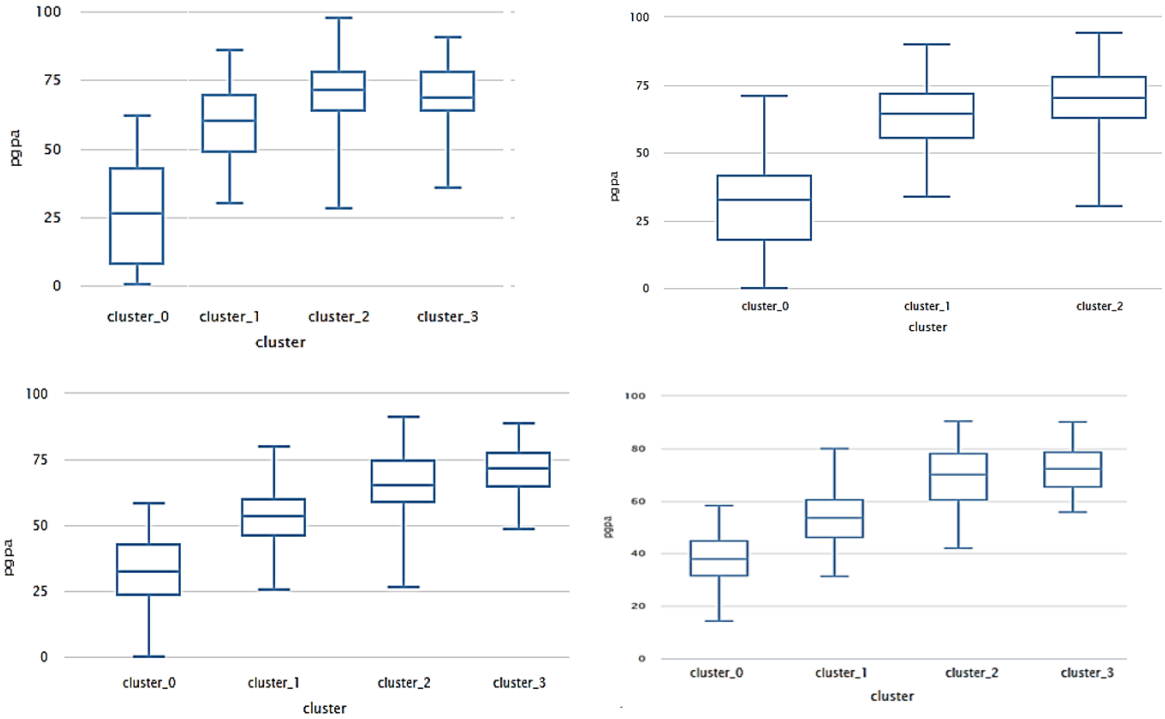
Os editores Kostopoulos e Kotsiantis são referência com dezenas de publicações relevantes no tema.

Disponível em : https://www.mdpi.com/journal/electronics/special_issues/EDM_electronics

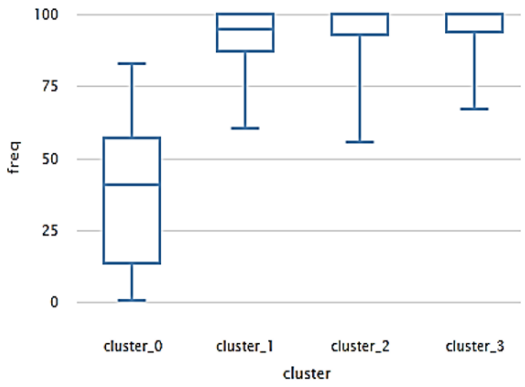


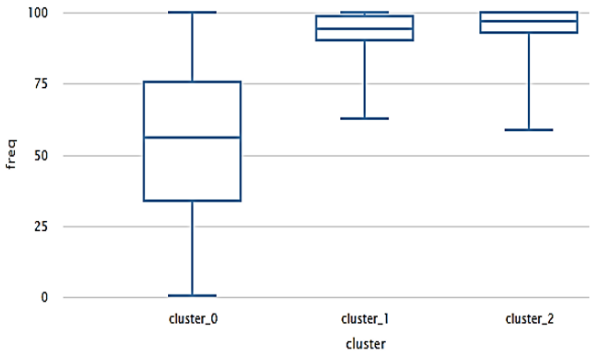


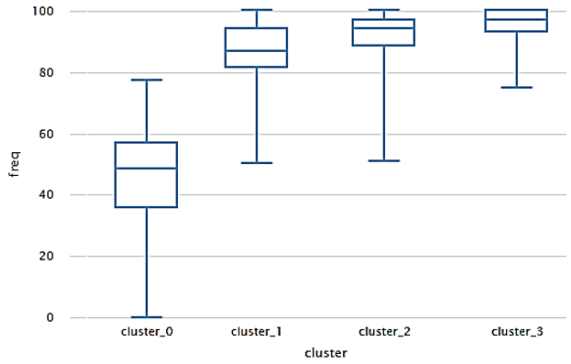
APÊNDICE D – Boxplot de pGPA em S*, S1, S2 e S3.

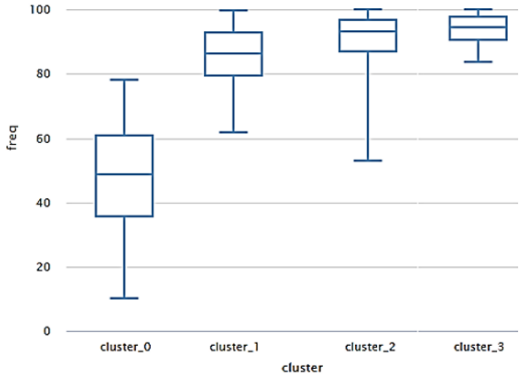


APÊNDICE E – Boxplot de FREQ em S*, S1, S2 e S3.

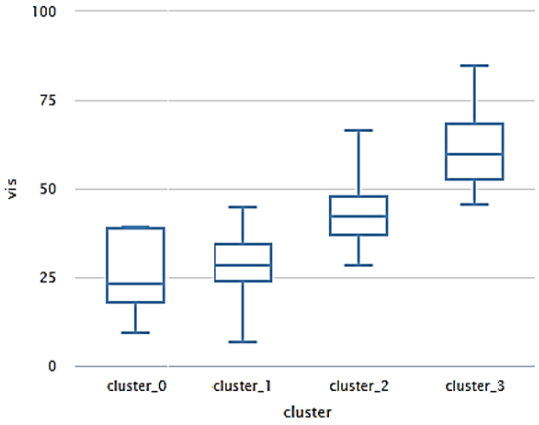


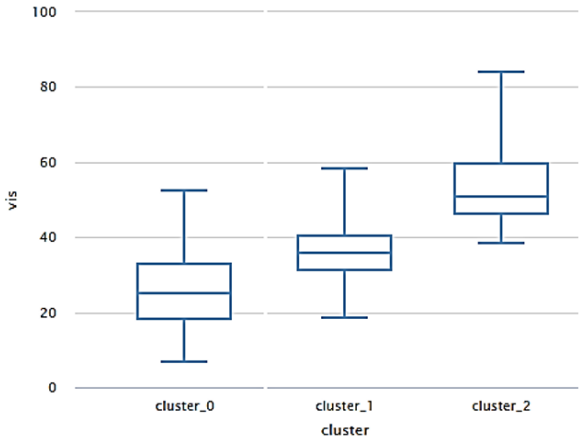


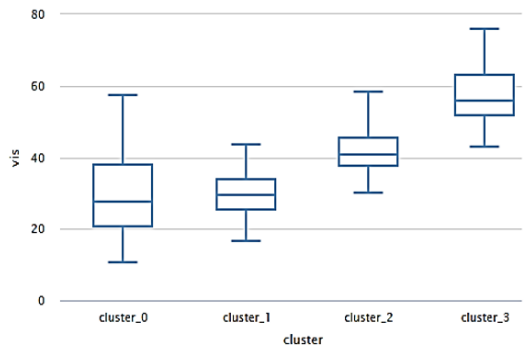


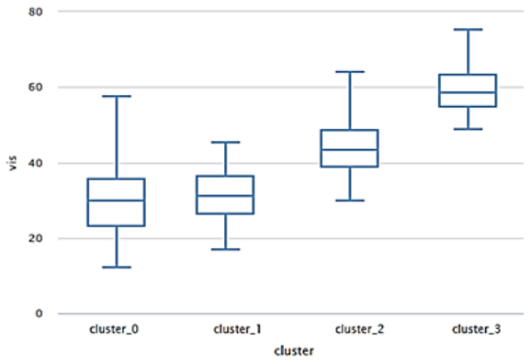


APÊNDICE F – Boxplot de VIS em S*, S1, S2 e S3

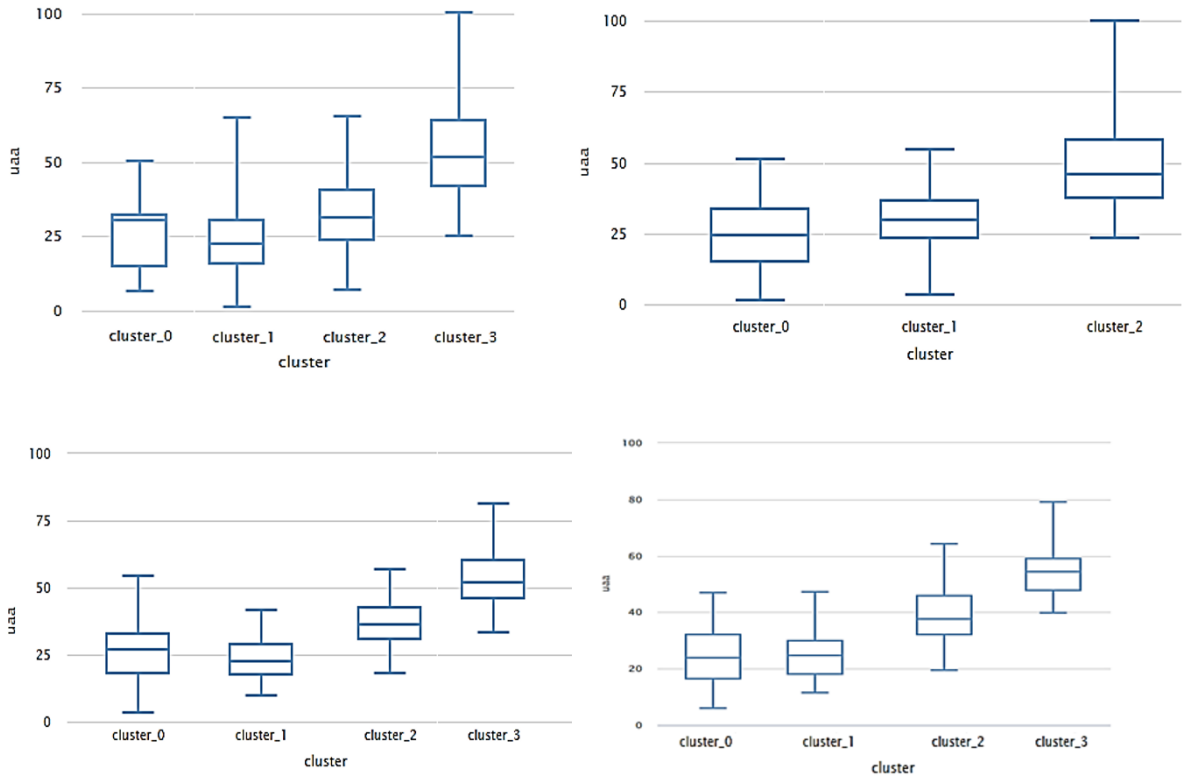








APÊNDICE G –Boxplot de UAA em S*, S1, S2 e S3.



APÊNDICE H – Principais queries no AVA

Base de Dados Postgres.

Queries criadas

1) Tipos de cursos

```
select  distinct parent  from mdl_course_categories

Parent=
1;"Presencial"
2;"Subsequente"
10;"FIC"
15;"Apoio"
19;"Demonstração"
132;"MedioTec"
133;"Concomitante"
134;"Pós Graduação
```

2) Listar os cursos 133= técnicos concomitantes

```
select * from mdl_course_categories where parent =133

Resultado

id |name
---|--------------------------------------------------
 71|Técnico em Informática para Internet [Concomitante]
 97|Técnico em Recursos Humanos [Concomitante]
 78|Técnico em Finanças [Concomitante]
 82|Técnico em Cooperativismo
174|Turma especial Cursos Concomitantes
 94|Técnico em Administração [Concomitante]
100|Técnico em Computação Gráfica [Concomitante]
```

3) Listar as turmas/cursos dos cursos técnicos concomitantes

```
select id, name, parent, to_timestamp(timemodified)
from mdl_course_categories mcc
where parent in (select id from mdl_course_categories mcc2 where parent = 133)
order by id
```

4) Disciplinas do curso

```
DROP TABLE IF EXISTS  t_disciplinas CASCADE;

select  id as courseid, category, sortorder, startdate, startdate AS
startdateCourse, startdate AS startdateSubject, fullname
into    t_disciplinas
from    mdl_course mc
where   enablecompletion=1
--0 significa o curso e 1 a disciplina     --somente dos concomitantes
and     category in (select id from t_courses) --considerar vários cursos
```

```
order by category, id
```

5) Quantidade de alunos por disciplina

```
SELECT
   course.category
 , course.id as course
 , to_timestamp(course.startdate) as startdate
 , course.fullname
 , context.id as context
 , COUNT(course.id) AS Students

FROM mdl_role_assignments AS asg
JOIN mdl_context          AS context  ON asg.contextid = context.id AND
context.contextlevel = 50
JOIN mdl_user             AS u      ON u.id = asg.userid
JOIN mdl_course           AS course ON context.instanceid = course.id

WHERE asg.roleid = 5
and   course.enablecompletion=1   ---0 significa o curso e 1 a disciplina
and   course.category IN (SELECT distinct category FROM t_disciplinas) --_1sem )
GROUP BY course.id, context.id
ORDER BY fullname
```

6) Obter e armazenar a os dados das disciplinas por aluno

```
drop table if exists t_aluno_disciplina;

SELECT

  u.id as userid
, u.username as CPF
, u.firstname  AS Nome
, u.lastname AS Sobrenome
, g.courseid --> se separar por disciplina não daria certo por trabalhar com
módulos e o startdate é por semestre.
, c.fullname
, cat.id as cat
, u.department  as polo
, c.etapa, c.modulo, c.semestre, c.horas_aula
-- 3 atributos criados em t_disciplinas

into t_aluno_disciplina

FROM mdl_user AS u   --usuários
JOIN mdl_role_assignments  AS rs ON u.id = rs.userid
JOIN mdl_context  AS e ON rs.contextid = e.id
--JOIN mdl_course            AS c ON c.id = e.instanceid
JOIN t_disciplinas AS c ON c.courseid = e.instanceid    -- código da disciplina
JOIN mdl_course_categories AS cat ON c.category = cat.id
JOIN mdl_role as r    ON r.id=rs.roleid
JOIN mdl_groups_members as m ON u.id = m.userid
JOIN mdl_groups as g  ON g.id=m.groupid

WHERE rs.roleid =5
AND   e.contextlevel= 50
```

```
and   g.courseid   = c.courseid --importante
AND   cat.id       = c.category
AND   r.id = 5
ORDER BY u.id, g.courseid
```

6) Listar as disciplinas do 1o. semestre de um determinado curso

```
 select  id, category, sortorder, startdate, fullname
 from    mdl_course mc
 where   enablecompletion=1   ---0 significa o curso e 1 a disciplina
      --somente dos concomitantes
 and     category = 129 -- se descomentar acima, informar direto o código do
curso
 and     startdate  =
      (select min(startdate) from mdl_course mc3
           where mc3.category= mc.category and enablecompletion=
mc.enablecompletion)
 order by category, id
```

Resultado

```
d  |category|sortorder|startdate |fullname                                              |
---|--------|---------|----------|------------------------------------------------------|
381|     129|   770010|1488340800|17-1 [IPI-CV-EaD] Lógica de Programação               |
382|     129|   770002|1488340800|17-1 [IPI-CV-EaD] Ambientação em EaD                  |
383|     129|   770003|1488340800|17-1 [IPI-CV-EaD] Introdução à Informática            |
384|     129|   770004|1488340800|17-1 [IPI-CV-EaD] Português Instrumental              |
385|     129|   770005|1488340800|17-1 [IPI-CV-EaD] Inglês Instrumental                 |
386|     129|   770006|1488340800|17-1 [IPI-CV-EaD] Recursos Multimidias                |
387|     129|   770007|1488340800|17-1 [IPI-CV-EaD] Arquitetura de Computadores        |
388|     129|   770008|1488340800|17-1 [IPI-CV-EaD] Fundamentos de Desenvolvimento WEB|
389|     129|   770009|1488340800|17-1 [IPI-CV-EaD] Sistemas Operacionais              |
```

8) Quantidade de interações no LOG
Quantidade de acessos semanais do aluno por disciplina

```
SELECT
  u.id as userid
, u.firstname  AS Nome
, g.courseid
, c.fullname, cat.id as cat
, COUNT(l.id) AS Edits
/* caso análise semanal */
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))<0,1,0)) --antes de começar
+ SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=0,1,0)) AS Week1
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=1,1,0)) AS Week2
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=2,1,0)) AS Week3
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=3,1,0)) AS Week4
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=4,1,0)) AS Week5
, SUM(f_1 ( FLOOR((l.time - c.startdate)/(60*60*24*7))=5,1,0)) AS Week6

FROM mdl_user AS u
JOIN mdl_role_assignments  AS rs ON u.id = rs.userid
JOIN mdl_context           AS e ON rs.contextid = e.id
--JOIN mdl_course            AS c ON c.id = e.instanceid
JOIN t_disciplinas   AS c ON c.courseid = e.instanceid    -- código da disciplina
```

```
JOIN mdl_course_categories AS cat ON c.category = cat.id
JOIN mdl_role r           ON r.id=rs.roleid
JOIN mdl_groups_members m ON u.id = m.userid
JOIN mdl_groups g         ON g.id=m.groupid

LEFT JOIN mdl_log  l ON l.userid = u.id AND l.course = c.courseid
     AND l.action NOT LIKE 'view%' and l.action != 'login' and l.action  !=
'logout' and l.action != 'mail error'

WHERE rs.roleid =5
AND   e.contextlevel= 50
and   g.courseid = c.courseid --importante
AND   cat.id = c.category
--and   l.userid =  --só para filtras aluno específico
AND   r.id = 5
GROUP BY  u.id, u.firstname, g.courseid, c.fullname, cat.id
ORDER BY u.id, g.courseid
```

9) Cálculo da idade do estudante, baseado na data do início do curso e data de nascimento.

```
update t_aluno_ficha as taf
  set   AGE =   (date_part('year', to_timestamp( td.startdate))  -
date_part('year',taf.dt_nasc::date) )
from t_aluno_disciplina tad , t_disciplinas td
   where tad.userid = taf.userid
   and   td.courseid = tad.courseid
```

10) Integrar os dados em .CSV

```
copy t_formados  ( category, status, cert, estagio, nome, sexo, dt_nasc, cpf,
renda,abs)
from 'D:\Formados\IPI-CONCOMITANTE.completo.CSV' delimiter ';' csv header
encoding 'windows-1251';
```

11) Função para média de 2 valores, considerando que pode estar nulo e considerar 0 no lugar de deixar nulo a conta.

```
CREATE OR REPLACE FUNCTION AVERAGE2 (
V1 NUMERIC,
V2 NUMERIC )
RETURNS NUMERIC
AS $FUNCTION$
DECLARE
    COUNT NUMERIC;
    TOTAL NUMERIC;
BEGIN
    COUNT=0;
    TOTAL=0;
    IF V1 IS NOT NULL THEN COUNT=COUNT+1; TOTAL=TOTAL+V1; END IF;
    IF V2 IS NOT NULL THEN COUNT=COUNT+1; TOTAL=TOTAL+V2; END IF;
    RETURN cast(TOTAL/COUNT as integer);
    EXCEPTION WHEN DIVISION_BY_ZERO THEN RETURN NULL;
END
```

```
$FUNCTION$ LANGUAGE PLPGSQL;
```

12) Função para substituir um valor nulo por valor informado

```
CREATE OR REPLACE FUNCTION f_1 (_expr boolean
                              , _true anyelement
                              , _else anyelement
                              , OUT result anyelement)
  RETURNS anyelement LANGUAGE plpgsql AS
$func$
begin
   if   _expr   is null then
     result := _else;
   else
     EXECUTE
     'SELECT CASE WHEN (' || _expr || ') THEN $1 ELSE $2 END'
     USING _true, _else
     INTO result;
   end if;

END
$func$;
```

13) Frequência semanal considerando total por disciplina/etapa

```
drop table if exists t_frequencia cascade;

SELECT
  log.studentid
  , c.fullname
  , c.courseid
  , count(se.sessdate ) as freq

into t_frequencia
from mdl_attendance_log log
  inner join mdl_attendance_statuses st on log.statusid= st.id
  inner join mdl_attendance_sessions se on log.sessionid = se.id
  inner join mdl_groups mg  on se.groupid  = mg.id   --polo, disciplina
(courseid)
  inner join t_aluno_disciplina c on mg.courseid  = c.courseid
  and log.studentid = c.userid
where  st.acronym in ('P') --, 'J') --P= presente; J=Falta justificada
 -- mesmo justificada é falta e não considera.
group by log.studentid, c.fullname, c.courseid
order by studentid ;
```

14) Os dados de saída no RapidMiner são exportados para CSV e importados na base de dados para executar outras queries para outras análises.

```
drop table if exists t_cluster_S1_X;

create table t_cluster_S1_X
```

```
(
id          integer,
CPF         char(11),
cluster_no char(10),
FREQ   numeric (5,2),
pGPA   numeric (5,2),
UAA         numeric (5,2),
QST         numeric (5,2),
VIS         numeric (5,2),
CIR         numeric (5,2),
cStatus char(1), -- A ou B
sStatus char(1), -- A,C, D, E, R
cat     char(3)
);

copy t_cluster_S0_X (id, CPF, cluster_no, FREQ, pGPA, UAA, QST, VIS, CIR )
from 'D:\S0-CPF.csv'
csv header delimiter ';' ;
```

15) Transições de um cluster ao seguinte

```
select * from t_cluster_CPF_serieA
--Média de cada atributo para cada transição de cluster
--que ficaram nos grupos com baixo desempenho

select  s.classe, s.status, count(*), s.cluster,
       avg(s0.FREQ), avg(s1.FREQ), avg(s2.FREQ), avg(s3.FREQ),
       avg(s0.pGPA), avg(s1.pGPA), avg(s2.pGPA), avg(s3.pGPA),
       avg(s0.UAA), avg(s1.UAA), avg(s2.UAA), avg(s3.UAA), avg(s0.QST), avg(s1.QST), avg(s2.QST), avg(s3.QST),
       avg(s0.VIS), avg(s1.VIS), avg(s2.VIS), avg(s3.VIS), avg(s0.CIR), avg(s1.CIR), avg(s2.CIR), avg(s3.CIR)
from   t_cluster_S1_X s1 left outer join t_cluster_S0_X s0 on s1.CPF = s0.CPF
                         left outer join t_cluster_S2_X s2 on s1.CPF = s2.CPF
                         left outer join t_cluster_S3_X s3 on s1.CPF = s3.CPF

                         join t_cluster_CPF_serieB s on s1.CPF = s.CPF
group by s.cluster, s.classe,s.status
order by s.cluster, s.classe,s.status
```

```
SELECT coalesce(s0.cluster_no, '') || coalesce(s1.cluster_no, '' ) as ev, count(*) as qtd
from   t_cluster_S0_X s0 left outer join t_cluster_S1_X s1 on s0.CPF = s1.CPF
                         join t_cluster_CPF_serieA s on s.CPF = s0.CPF

where  substring(s.cluster,1,1) in ('0') --> Informar o primeiro cluster
group by s0.cluster_no, s1.cluster_no, s0.cluster_no || s1.cluster_no
order by s0.cluster_no, s1.cluster_no
```

APÊNDICE I –Classe P-Transição dos clusters em S*, S1, S2 e S3.

Na coluna Transição informa o número de cluster que transitou em S*, S1, S2 e S3 agrupado por status.

Classe	Status	Estudantes	Transição	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3
P	D	3	**00**	41.83	40.11			31.35	37.40			33.83	21.06			13.67	17.72			15.58	16.44			6.42	5.22		
P	E	12	**00**	40.85	39.18			16.04	17.34			22.63	22.79			18.08	17.33			25.28	24.22			8.71	7.03		
P	D	5	**000**	41.75	28.71	22.59		26.97	14.65	11.39		26.05	25.83	25.92		8.07	5.54	3.24		23.83	22.93	22.38		6.62	4.47	2.62	
P	E	12	**000**	37.62	30.96	25.66		32.95	11.92	10.47		28.71	28.26	27.00		9.97	6.67	4.02		27.97	26.93	26.27		3.33	2.44	1.40	
P	R	2	**000**	10.00	10.00	27.44		30.99	28.91	41.23		32.07	33.41	32.72		0.00	0.00	1.70		38.92	38.48	24.64		0.00	0.00	1.20	
P	E	1	**0000**	40.00	24.29	13.08	10.00	49.17	32.78	24.58	14.05	14.00	14.00	14.00	14.00	17.33	7.43	4.00	3.06	23.00	23.00	23.00	23.00	4.33	1.86	1.00	0.76
P	R	1	**0000**	73.33	56.14	61.31	41.95	18.00	16.91	18.12	16.72	49.33	29.57	21.67	21.67	4.00	14.71	13.15	9.00	29.00	29.14	29.69	29.69	8.00	9.71	16.85	11.53
P	R	1	**0010**	40.00	33.33	58.92	63.26	34.74	27.79	42.09	38.47	35.50	35.50	20.80	15.11	70.75	47.17	39.33	34.21	38.50	31.00	25.70	22.88	19.75	13.33	9.92	7.95
P	D	2	**100**	100.00	51.43	44.33		62.36	31.18	28.79		47.75	42.25	42.25		18.34	7.86	6.12		35.17	28.75	28.75		8.50	5.08	3.94	
P	E	13	**100**	88.51	69.63	53.61		50.34	36.23	32.07		18.01	17.59	17.92		23.07	14.53	11.16		25.16	23.82	24.60		6.93	4.64	3.18	
P	D	2	**1000**	78.63	65.56	45.62	33.62	46.48	41.71	34.34	27.50	14.17	15.57	14.57	14.57	36.25	27.81	15.09	10.33	30.09	30.27	28.95	28.95	18.79	15.16	8.54	5.93
P	E	9	**1000**	82.11	76.09	56.48	44.22	48.04	38.77	31.80	29.55	17.33	17.53	16.93	16.93	24.53	21.22	12.68	9.33	25.31	25.00	23.77	24.63	9.95	7.50	4.73	3.32
P	R	1	**1000**	80.00	76.67	61.08	63.00	56.13	50.58	42.20	44.61	24.00	24.00	17.25	21.00	6.00	4.33	7.75	12.26	9.00	7.17	10.56	15.33	5.25	4.67	5.17	9.16
P	D	1	**101**	86.25	84.17	84.17		29.57	34.06	34.06		15.33	15.33	15.33		31.50	32.67	32.67		16.25	17.80	17.80		8.50	6.17	6.17	
P	E	1	**101**	84.25	86.17	82.42		38.15	28.32	25.21		35.00	30.40	24.71		45.00	33.50	24.92		32.25	27.00	21.18		16.00	14.67	12.00	
P	E	1	**1010**	77.50	75.00	83.75	59.21	58.92	54.11	50.69	37.01	22.25	21.00	18.75	15.90	32.00	21.50	22.00	18.16	25.25	19.50	18.18	19.54	28.00	21.00	15.33	10.53
P	R	1	**1010**	93.33	86.86	81.69	69.95	51.00	44.01	41.84	40.76	13.00	10.86	9.44	9.44	9.67	11.14	12.54	10.47	17.33	17.86	19.77	20.07	3.67	2.86	4.23	3.47
P	R	2	**1011**	79.88	76.58	80.96	80.66	47.00	38.81	46.13	44.42	20.25	24.25	22.75	18.39	29.50	27.42	27.38	29.26	16.63	17.69	19.97	20.47	7.88	8.34	10.00	10.32
P	E	1	**11**	94.25	92.83			47.85	44.01			15.33	15.50			45.25	37.17			20.75	22.20			8.75	8.50		
P	E	1	**110**	90.00	90.00	51.83		36.80	33.53	30.20		16.50	44.33	37.25		30.50	32.50	17.92		15.75	18.83	16.88		4.75	8.00	4.25	
P	D	1	**1100**	96.67	90.00	62.31	47.65	69.66	65.57	48.09	37.79	29.50	30.25	30.25	30.25	46.67	36.00	21.00	16.06	28.00	25.86	25.22	25.22	16.33	11.43	6.31	4.82
P	E	5	**1100**	94.67	88.37	56.25	43.09	55.98	59.39	45.48	44.81	30.60	34.39	31.65	31.65	24.74	27.94	15.65	11.96	32.20	37.71	33.06	33.06	9.40	9.57	5.63	4.31
P	D	1	**111**	90.00	90.00	55.67		83.14	80.87	72.62		26.25	23.20	27.00		30.00	35.83	21.58		23.50	24.17	26.00		16.25	16.67	13.17	
P	E	2	**111**	88.38	88.92	75.21		70.83	70.83	60.74		35.88	33.10	27.50		30.88	35.58	26.83		29.00	28.50	23.55		12.25	13.92	10.83	
P	R	1	**111**	100.00	92.86	88.00		59.36	50.01	46.51		4.50	17.67	17.67		42.33	26.71	20.78		44.67	42.80	42.80		2.33	3.71	2.89	
P	D	1	**1110**	100.00	92.86	87.69	60.00	72.67	62.52	49.08	45.31	39.50	40.50	29.18	29.18	22.67	23.14	18.69	12.79	35.33	39.29	32.23	32.23	8.67	9.71	11.38	7.79
P	E	5	**1110**	93.50	88.81	76.09	60.95	63.33	61.15	51.77	48.05	23.47	24.22	24.99	24.99	28.13	32.74	22.72	17.59	28.33	31.26	28.70	28.70	9.07	8.74	6.34	4.88
P	R	5	**1110**	93.95	92.43	83.25	65.03	60.65	57.61	51.50	42.85	23.77	25.28	22.26	20.63	37.32	39.10	29.82	20.99	30.35	32.70	29.56	27.37	12.58	15.54	12.70	9.15
P	E	3	**1111**	100.00	93.86	91.82	81.40	57.42	61.32	55.01	50.76	21.67	33.00	30.84	30.73	26.33	30.91	27.74	23.97	31.56	37.95	36.23	36.25	5.11	10.05	10.31	10.31
P	R	31	**1111**	93.55	92.12	89.85	87.07	60.00	57.27	53.46	50.38	21.40	23.56	20.57	19.86	36.41	38.65	33.89	29.04	28.44	30.01	28.82	27.44	11.76	12.18	12.70	12.24
P	E	1	**1122**	100.00	94.29	94.62	76.47	55.00	57.55	58.03	49.77	3.50	24.75	30.75	29.30	32.33	57.43	59.85	51.00	33.00	46.14	45.08	44.00	29.00	36.86	32.69	26.53
P	R	4	**1122**	92.71	92.23	93.50	94.37	62.87	64.63	62.84	62.90	39.25	40.61	41.59	40.31	30.63	35.77	37.11	37.04	36.90	38.91	41.53	42.55	10.67	9.86	12.69	14.54
P	E	1	**1121**	93.33	87.14	79.23	61.76	43.33	48.14	48.08	41.67	9.50	33.25	34.50	34.50	60.00	55.00	59.23	45.29	36.67	41.29	44.23	44.23	23.33	33.29	33.85	25.88
P	R	2	**1121**	96.67	97.86	90.77	73.92	58.11	66.99	58.59	53.85	25.17	32.52	32.73	29.00	25.67	37.50	39.35	32.42	37.50	40.86	42.89	39.56	22.84	21.86	25.31	20.37
P	R	1	**1210**	100.00	92.86	80.00	57.37	65.44	69.23	50.13	37.60	34.00	58.20	41.56	41.56	28.67	38.00	33.62	23.00	28.00	38.43	31.00	27.07	20.00	32.00	20.69	14.26
P	D	2	**200**	96.67	65.72	51.11		35.56	29.14	29.14		35.00	35.60	35.60		48.17	25.00	19.45		55.34	46.20	46.20		19.17	8.86	6.89	
P	E	5	**200**	90.67	64.91	51.42		58.58	42.24	37.36		27.17	30.45	28.85		48.33	27.57	21.45		46.00	41.08	40.69		16.07	9.83	7.64	
P	E	2	**2000**	80.00	52.86	28.46	22.06	67.61	61.65	48.42	36.44	25.50	28.50	28.50	28.50	54.17	27.79	14.96	11.44	38.00	26.75	26.75	26.75	32.34	17.93	9.66	7.39
P	R	1	**2022**	55.25	53.50	70.92	73.47	56.10	44.90	42.74	47.10	45.00	37.20	39.09	39.61	55.00	43.33	40.00	42.53	42.75	33.00	41.33	42.53	40.25	29.00	40.42	42.21
P	E	1	**21**	100.00	93.33			59.11	46.41			25.00	25.00			80.00	59.67			52.50	39.00			49.75	33.83		
P	D	1	**210**	80.75	80.50	47.33		75.92	72.22	56.79		29.00	31.75	23.33		60.50	60.83	30.50		34.50	36.00	27.88		19.00	18.00	9.33	
P	D	2	**2100**	93.34	74.15	42.93	33.44	77.75	54.76	42.09	38.06	42.84	35.70	35.70	35.70	31.34	27.22	14.66	11.21	44.34	41.06	41.06	41.06	5.67	7.07	3.81	2.92
P	E	8	**2100**	90.83	80.77	48.56	38.21	68.89	54.74	42.89	38.88	38.37	33.86	32.38	31.84	43.83	34.45	18.75	14.57	47.38	41.07	38.83	38.43	25.71	19.71	10.73	8.62
P	E	7	**211**	94.65	87.68	73.69		68.80	60.77	56.46		29.38	29.67	26.42		63.21	49.68	37.98		40.27	35.51	31.61		28.19	22.24	16.16	
P	E	4	**2110**	98.75	95.48	81.78	59.69	58.46	50.61	38.08	30.58	19.25	18.48	16.96	16.49	59.29	52.05	39.00	27.78	35.69	31.97	27.59	26.77	22.85	18.85	12.87	9.26
P	R	1	**2110**	100.00	100.00	78.46	64.74	60.00	59.38	53.40	45.16	34.00	34.25	34.33	29.50	35.33	43.00	33.69	23.05	35.00	38.86	36.09	30.64	39.33	30.86	20.77	15.11
P	D	1	**2111**	86.67	91.43	91.15	71.76	51.00	51.24	44.96	39.46	40.33	43.14	32.69	29.27	38.67	42.14	34.31	28.00	46.67	47.43	40.31	36.80	13.33	12.43	13.46	10.53
P	E	2	**2111**	76.00	74.79	74.73	75.83	75.90	62.36	55.55	56.28	24.50	27.15	20.29	20.29	61.34	51.68	41.77	31.51	42.34	37.11	33.88	30.54	18.67	14.90	16.55	13.05
P	R	17	**2111**	95.98	93.20	88.33	84.25	65.14	57.32	51.48	49.02	30.26	27.93	22.10	21.19	58.83	50.61	40.12	33.10	39.01	35.21	29.97	28.28	20.33	19.47	16.00	14.43

P	E	1	**212**	96.25	97.50	90.42		68.72	65.93	59.44		45.00	38.80	27.70		44.50	52.67	52.83		41.75	41.17	38.92		35.25	33.17	26.83	
P	R	1	**2120**	95.00	93.33	94.17	60.37	63.62	66.24	65.42	44.54	36.25	37.20	36.60	36.60	35.25	42.33	42.75	27.00	34.00	36.00	37.50	37.50	28.00	29.50	29.92	18.89
P	E	1	**2122**	86.67	94.29	94.62	72.35	77.33	77.62	79.83	79.83	26.00	39.57	47.23	47.23	50.33	45.14	37.38	28.59	49.67	52.43	52.62	52.62	18.67	16.29	18.23	13.94
P	R	7	**2122**	98.10	97.07	95.49	94.15	69.61	68.42	65.46	67.18	27.50	29.96	34.04	34.90	53.99	52.14	50.39	45.54	39.87	40.89	41.70	41.59	23.40	22.79	24.03	22.79
P	D	1	**2121**	100.00	100.00	82.31	63.35	65.28	69.41	62.55	57.74	23.67	33.43	37.45	37.45	57.00	53.71	35.62	27.24	47.00	47.86	45.18	45.18	6.33	8.71	6.46	4.94
P	E	1	**2121**	86.67	83.14	80.31	61.82	87.00	80.29	66.94	62.16	48.67	45.00	47.00	47.00	40.00	44.71	33.85	25.88	42.33	43.14	41.15	41.15	4.33	3.71	5.00	3.82
P	R	8	**2121**	96.67	95.63	89.76	84.74	65.54	62.18	59.90	54.50	31.52	36.40	32.70	31.38	49.53	47.38	42.11	37.49	40.60	40.20	37.84	36.43	19.88	19.08	17.89	17.04
P	E	1	**220**	95.00	93.33	50.17		72.74	74.33	55.75		45.75	54.20	54.20		34.25	42.17	21.08		38.25	40.00	40.00		30.50	28.17	14.08	
P	R	1	**2211**	72.25	81.50	85.75	89.42	49.21	52.42	43.92	39.91	41.75	37.20	25.82	22.13	21.25	32.17	29.58	29.47	46.00	44.50	35.50	31.56	50.25	49.83	35.92	28.58
P	D	1	**222**	100.00	100.00	51.17		77.08	77.72	58.29		43.50	38.00	38.00		54.75	61.67	30.83		42.25	45.33	45.33		31.50	34.17	17.08	
P	E	1	**2220**	96.67	92.86	69.23	52.94	80.55	76.05	55.59	47.65	26.50	37.75	28.17	28.17	65.67	66.86	49.85	38.12	40.33	44.71	34.92	34.92	22.33	31.14	25.54	19.53
P	E	3	**2222**	93.33	92.86	86.15	74.61	78.22	78.36	69.91	64.80	39.50	46.94	45.45	45.78	54.44	59.48	51.80	43.90	44.44	48.81	44.82	44.78	16.11	20.72	20.92	19.20
P	R	5	**2222**	97.33	95.57	91.21	90.05	67.17	62.74	56.75	58.50	29.50	34.65	31.01	30.24	72.53	72.17	60.20	53.73	46.45	49.94	44.77	43.97	39.27	41.44	33.76	35.38
P	E	2	**2221**	90.00	93.57	82.69	65.59	70.02	64.92	57.12	49.59	46.25	40.25	36.79	34.67	61.67	58.57	53.74	42.56	41.17	44.36	39.39	38.42	29.50	38.86	29.08	22.85
P	R	3	**2221**	94.25	96.17	93.47	87.33	59.73	57.87	56.06	46.46	17.72	30.18	29.53	24.50	74.33	70.28	55.36	42.96	44.17	47.50	41.58	33.48	39.75	42.72	35.59	26.52
P	E	1	**2232**	86.67	92.86	96.15	82.94	71.84	70.32	66.81	67.30	22.00	32.75	36.75	31.75	63.67	56.00	60.85	50.12	48.33	48.86	52.46	45.65	51.67	49.29	48.00	40.94
P	R	1	**2232**	85.50	90.33	95.17	82.47	59.21	55.47	58.47	48.46	21.67	28.25	33.10	27.38	72.00	76.17	63.17	52.37	54.50	59.67	53.00	43.11	55.50	60.00	52.92	40.21
P	R	1	**2233**	100.00	97.14	93.08	88.95	70.33	70.26	70.37	69.18	39.00	68.25	67.63	68.46	67.00	66.71	69.62	65.42	48.67	50.71	56.85	58.11	37.67	35.71	44.00	45.11
P	E	1	**3100**	66.67	66.29	35.69	27.29	54.78	45.98	35.76	35.76	56.33	40.00	40.00	40.00	57.00	42.86	23.08	17.65	65.33	51.00	51.00	51.00	27.33	20.00	10.77	8.24
P	E	1	**3111**	93.33	95.71	91.54	89.41	60.58	55.77	47.57	45.79	49.00	42.20	33.50	28.30	64.67	46.71	36.62	28.00	47.67	39.29	31.54	26.94	40.00	29.43	20.23	16.59
P	R	1	**3111**	100.00	100.00	83.08	82.11	52.33	43.71	35.27	33.71	46.50	41.67	20.57	16.64	36.67	44.86	30.69	27.16	45.33	39.71	28.75	25.39	54.67	36.14	22.23	19.47
P	E	1	**312**	93.33	72.57	57.22		73.56	69.72	59.76		42.67	43.40	43.40		48.00	31.43	24.44		61.00	45.71	45.71		33.33	19.14	14.89	
P	E	1	**3121**	100.00	100.00	93.08	75.88	70.11	54.62	53.28	46.17	40.50	27.75	29.75	29.75	71.33	55.57	48.15	36.82	51.00	38.57	38.00	38.00	42.33	28.00	26.54	20.29
P	R	1	**3212**	100.00	96.67	84.33	88.26	46.43	45.90	43.08	51.29	49.00	45.75	29.10	38.88	68.25	60.00	49.25	49.58	52.50	46.33	35.83	43.79	45.75	37.00	25.17	35.16
P	D	1	**322**	92.00	71.00	66.33		35.50	30.25	26.17		100.00	100.00	51.50		77.00	75.50	53.67		55.00	49.00	40.33		26.00	14.50	19.67	
P	E	1	**322**	94.25	86.17	50.75		47.04	45.03	33.81		42.67	44.50	44.50		74.50	74.67	37.33		56.50	56.50	56.50		39.50	41.00	20.50	
P	R	1	**3220**	100.00	58.57	55.38	48.00	51.89	37.73	37.70	33.93	71.50	55.33	53.00	46.14	91.67	52.14	44.31	38.53	84.33	50.29	43.00	39.13	88.67	47.43	36.38	33.40
P	E	1	**3222**	100.00	95.71	74.62	57.65	71.99	71.77	50.50	43.76	53.00	49.25	51.33	51.33	62.67	54.71	52.00	39.76	48.00	42.00	44.73	44.73	40.67	32.71	23.54	18.00
P	R	4	**3222**	98.33	94.94	94.69	90.16	65.52	59.88	54.86	51.54	53.88	44.09	37.16	36.28	61.85	51.33	40.93	35.15	61.94	52.34	44.46	42.20	52.67	43.44	36.71	32.86
P	E	1	**3221**	87.50	91.67	78.67	70.42	77.24	76.33	62.43	43.16	63.00	63.00	43.10	30.13	58.00	57.67	45.08	34.74	58.00	56.83	45.08	37.59	40.25	43.83	30.25	22.42
P	R	1	**3221**	90.00	88.00	87.18	76.61	65.88	62.77	53.98	44.78	53.00	46.00	40.09	31.94	76.00	73.60	50.18	34.50	51.25	48.00	36.18	28.18	35.25	31.40	20.18	15.06
P	D	1	**3232**	79.00	76.00	74.67	52.95	60.40	49.65	55.43	47.91	41.50	38.80	58.82	55.93	58.00	46.00	46.17	37.89	51.75	44.50	53.75	52.50	51.50	45.33	54.25	43.05
P	E	1	**3232**	100.00	91.43	80.00	61.76	65.52	65.35	48.19	41.77	54.00	59.00	49.14	49.14	53.33	60.29	59.15	45.24	61.00	60.57	54.58	54.58	69.33	59.43	41.46	31.71
P	R	1	**3232**	100.00	100.00	91.83	78.84	47.69	51.96	49.98	49.47	56.75	49.80	40.45	36.89	39.75	54.00	49.83	43.11	53.25	52.17	50.67	46.95	60.25	53.33	54.67	50.95
P	R	5	**3233**	98.50	97.67	95.65	91.82	68.34	64.01	63.35	60.40	62.00	61.49	55.64	50.28	76.40	76.75	70.36	62.35	69.38	68.65	62.59	57.01	69.72	65.71	59.96	53.76

APÊNDICE J –Classe N-Transição dos clusters em S*, S1, S2 e S3.

Na coluna Transição informa o número de cluster que transitou em S*, S1, S2 e S3 agrupado por status.

				FREQ				pGPA				UAA				QST				VIS				CIR			
Classe	Status	Qtd	Evolução	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3	S0	S1	S2	S3
N	A	1	1112	100.00	100.00	100.00	96.32	76.49	71.71	75.41	78.95	29.50	29.50	31.89	27.69	36.25	28.00	32.33	41.53	27.00	22.33	26.25	29.74	20.25	16.83	15.17	19.21
N	A	9	1111	92.74	93.64	93.57	91.98	64.20	64.65	60.01	59.21	24.71	27.48	24.98	25.49	31.52	32.29	27.65	24.85	30.86	32.75	31.47	31.52	7.50	8.66	11.18	10.57
N	C	3	1111	92.00	94.67	91.92	91.48	74.64	75.37	73.25	74.39	25.38	24.75	22.56	21.02	35.50	45.75	45.42	49.00	24.63	29.67	28.50	29.77	16.75	18.67	16.71	17.03
N	A	3	1122	93.61	93.25	89.44	91.23	74.55	78.31	77.92	78.73	24.17	24.62	38.24	36.72	42.28	45.55	49.60	49.04	26.75	27.71	35.40	38.03	17.33	18.18	26.62	29.97
N	C	2	1122	100.00	98.57	97.40	94.90	78.49	76.82	72.90	71.14	34.75	35.54	37.46	40.26	27.21	38.27	35.20	38.48	30.17	35.16	37.06	39.61	15.67	17.20	19.23	22.76
N	A	4	1121	96.67	94.64	93.46	92.06	69.55	70.38	69.48	69.65	31.50	30.54	32.04	32.17	24.17	30.32	31.35	25.74	31.34	35.25	38.04	36.96	12.58	11.97	16.38	15.08
N	C	3	1121	100.00	97.62	94.02	92.15	77.94	69.91	67.16	65.44	20.33	27.93	35.37	34.82	29.97	30.54	30.00	29.27	28.47	33.98	37.70	37.24	9.25	12.76	16.19	16.85
N	A	1	2112	96.67	94.29	96.92	95.79	68.59	56.86	57.09	60.33	32.50	25.00	29.71	38.27	58.67	44.57	42.46	39.37	43.33	32.71	33.62	34.89	22.00	17.00	17.85	20.42
N	A	2	2111	93.38	92.25	88.00	84.90	66.61	63.99	55.32	55.01	19.67	21.88	27.50	25.11	69.25	66.33	43.00	33.16	34.00	37.67	30.66	26.64	17.00	20.67	16.00	14.82
N	C	1	2111	98.00	95.33	95.17	88.63	70.69	65.13	59.63	59.10	27.00	26.50	19.44	17.31	53.00	60.17	45.33	49.84	29.00	34.67	28.67	32.79	18.50	24.00	19.67	25.42
N	A	21	2122	96.79	95.84	94.42	93.40	74.69	74.76	74.40	74.29	33.70	34.09	36.99	37.32	47.72	48.20	45.82	43.68	41.53	39.85	42.08	43.22	23.23	21.39	24.78	25.51
N	C	9	2122	95.46	93.84	93.33	92.26	77.17	75.75	73.62	72.89	31.12	29.97	32.00	33.04	48.37	49.70	50.06	48.10	36.57	35.43	36.72	38.23	26.55	23.86	24.78	25.97
N	A	4	2121	98.33	98.21	95.91	93.77	74.72	70.88	66.56	62.99	21.92	24.86	27.04	25.93	55.21	52.99	48.07	38.49	40.69	39.32	37.51	33.97	24.40	23.05	18.77	18.07
N	C	1	2121	100.00	95.71	97.69	96.47	80.64	72.80	70.82	64.08	33.50	41.67	36.08	31.06	38.33	40.86	29.85	25.26	45.33	46.43	40.38	36.26	17.33	15.00	14.85	16.37
N	A	2	2132	100.00	100.00	96.19	93.54	61.37	67.86	72.00	73.07	42.67	42.87	54.79	52.68	48.71	46.47	48.86	43.15	47.75	45.11	55.73	54.48	21.00	17.60	31.37	27.73
N	A	1	2211	84.25	89.50	82.83	82.83	72.83	67.89	54.35	54.35	43.67	46.75	28.11	28.11	64.00	56.33	34.67	34.67	40.25	41.33	29.55	29.55	20.50	28.33	18.33	18.33
N	A	10	2222	96.67	95.77	94.52	92.02	73.45	71.76	68.07	67.60	36.54	45.91	42.59	40.01	54.83	53.65	46.47	42.02	44.36	47.48	46.81	45.44	26.43	27.23	30.38	28.31
N	C	3	2222	98.33	96.82	94.81	93.55	81.40	79.92	78.49	78.37	28.00	36.57	34.33	33.47	57.33	55.94	57.82	55.15	43.19	41.91	43.35	42.24	43.61	38.44	38.64	37.40
N	A	1	2221	96.00	94.00	93.67	91.71	71.10	68.96	58.86	54.43	34.67	34.00	20.88	19.67	76.75	77.67	59.17	53.57	47.50	47.33	36.50	33.43	25.00	24.67	20.92	19.29
N	A	1	2223	90.75	93.83	93.58	93.32	78.95	79.19	80.35	81.80	31.00	34.00	43.50	49.65	73.75	71.33	56.67	56.16	50.00	50.50	46.42	48.84	39.50	41.33	35.58	40.68
N	A	4	2232	100.00	96.79	94.90	94.24	73.23	75.21	76.22	76.52	39.46	47.25	47.80	47.37	57.38	63.18	63.61	59.56	41.48	48.68	48.46	46.50	26.00	34.75	36.78	34.45
N	A	5	2233	97.52	95.80	94.97	91.26	71.30	75.63	72.63	73.83	34.92	46.14	50.72	52.37	63.48	63.76	57.38	54.16	49.87	55.40	56.95	57.29	36.30	43.15	47.88	48.35
N	C	3	2233	100.00	99.05	94.80	94.77	76.08	79.47	75.81	76.39	40.00	41.08	52.85	51.65	56.28	59.37	62.17	59.12	45.75	48.10	55.76	58.72	41.22	39.79	48.09	54.19
N	A	3	3222	100.00	97.90	95.99	89.98	72.56	69.14	65.45	63.79	41.06	42.32	34.26	37.03	62.20	59.78	49.33	47.18	56.86	54.79	48.79	47.98	46.64	41.36	37.62	34.81
N	A	6	3232	98.89	98.10	96.60	94.11	76.34	77.84	74.60	74.75	51.81	53.07	50.42	48.15	63.46	62.64	56.69	48.70	58.64	56.94	54.32	51.36	43.42	39.65	35.19	31.51
N	C	1	3232	100.00	100.00	100.00	97.89	81.73	85.82	83.91	80.78	56.00	56.00	49.56	44.00	60.75	67.17	60.92	57.68	49.75	50.83	48.50	45.79	39.25	38.50	38.33	32.84
N	A	15	3233	97.52	97.60	97.83	95.73	72.71	70.54	69.17	70.15	55.78	56.37	56.06	55.56	68.63	69.71	61.69	58.52	65.30	65.31	62.99	61.95	59.18	57.20	56.24	54.91
N	C	8	3233	98.13	97.74	95.31	94.21	77.13	76.79	76.55	77.65	51.72	53.21	59.13	59.20	66.27	69.92	61.98	63.95	57.67	58.04	58.30	59.98	53.65	53.58	53.52	54.43
		127																									

ANEXO A – Matriz curricular dos cursos técnicos em Informática para Internet, Finanças e Administração.

CURSO TÉCNICO EM INFORMATICA PARA INTERNET CONCOMITANTE AO ENSINO MÉDIO
CAMPUS PORTO VELHO ZONA NORTE
Matriz aprovada pela Resolução nº 13/CONSUP/IFRO/2016
Organização conforme a LDB 9.394/96, Art. 36, e a Resolução CNE/CEB 6/2012
Duração da aula: 50 minutos

Períodos/ Módulos/ Etapas[1]		Disciplinas	Semanas letivas	Número de Aulas		TOTAL (Hora-Aula)	TOTAL (Hora-Relógio)
				Tele - Presencial	EaD		
PRIMEIRO MÓDULO	E1	Ambientação para EaD	4	8	32	40	33,33
	E2	Introdução à Informática	4	8	32	40	33,33
		Português Instrumental		8	32	40	33,33
	E3	Inglês Instrumental	4	8	32	40	33,33
		Recursos Multimídias		8	32	40	33,33
	E4	Arquitetura de Computadores	6	12	48	60	50
		Fundamentos de Desenvolvimento Web		12	48	60	50
	E5	Sistemas Operacionais	6	12	48	60	50
		Lógica de Programação		12	48	60	50
Subtotal 1			**24**	**88**	**352**	**440**	**366,65**
SEGUNDO MÓDULO	E1	Linguagem de Programação I	4	8	32	40	33,33
		Comércio Eletrônico e Empreendedorismo		8	32	40	33,33
	E2	Interação Humano – Computador	4	8	32	40	33,33
		Orientação para Prática Profissional e Pesquisa		8	32	40	33,33
	E3	Análise e Projeto de Sistemas I	6	12	48	60	50
		Programação Orientada a Objetos		12	48	60	50
	E4	Banco de Dados	6	12	48	60	50
		Programação para Web I		12	48	60	50
Subtotal 2			**20**	**80**	**320**	**400**	**333,32**
TERCEIRO MÓDULO	E1	Linguagem de Programação II	4	8	32	40	33,33
		Análise e Projetos de Sistemas II		8	32	40	33,33
	E2	*Design* para Web	4	8	32	40	33,33
		Programação para Dispositivos Móveis		8	32	40	33,33
	E3	Segurança da Informação	4	8	32	40	33,33
		Rede de Computadores		8	32	40	33,33
	E4	Ética Profissional e Cidadania	6	12	48	60	50
		Programação para Web II		12	48	60	50
Subtotal 3			**18**	**72**	**288**	**360**	**299,98**
Núcleo Complementar		**Prática Profissional**				240	200
CARGA HORÁRIA TOTAL DO CURSO			**62**	**240**	**960**	**1.440**	**1.200**

Quadro 01 – Matriz curricular do Curso Técnico em Administração Concomitante ao Ensino

CURSO TÉCNICO EM ADMINISTRAÇAO CONCOMITANTE AO ENSINO MÉDIO *CAMPUS* PORTO VELHO ZONA NORTE Matriz aprovada pela Resolução nº 07/CEPEX/IFRO/2016							
Organização conforme a LDB nº 9.394/96, art. 36, e a Resolução CNE/CEB nº 6/2012 Duração da aula: 50 minutos							
Períodos/ módulos/ etapas[1]		Disciplinas	Semanas letivas	Número de aulas		TOTAL (Hora-aula)	TOTAL (Hora-relógio)
				Presencial	EaD		
PRIMEIRO MÓDULO	E1	Ambientação para EaD	4	8	32	40	33,33
		Introdução à Informática		8	32	40	33,33
	E2	Português Instrumental	6	12	48	60	50
		Fundamentos de Economia		12	48	60	50
	E3	Fundamentos de Matemática Financeira	4	8	32	40	33,33
		Direito e Legislação Comercial		8	32	40	33,33
Subtotal 1			**14**	**56**	**224**	**280**	**233,3**
SEGUNDO MÓDULO	E1	Fundamentos de Administração	6	12	48	60	50
		Contabilidade Geral		12	48	60	50
	E2	Orientação para a Pesquisa e Prática Profissional	4	8	32	40	33,33
		Segurança, Meio Ambiente e Saúde		8	32	40	33,33
	E3	Gestão de Pessoas	6	12	48	60	50
		Organização, Sistemas e Métodos		12	48	60	50
Subtotal 2			**16**	**64**	**256**	**320**	**266,7**
TERCEIRO MÓDULO	E1	Comportamento Organizacional	4	8	32	40	33,33
		Matemática Financeira Aplicada		8	32	40	33,33
	E2	Marketing	6	12	48	60	50
		Contabilidade de Custos		12	48	60	50
	E3	Administração da Produção	4	8	32	40	33,33
		Fundamentos do Direito Tributário		8	32	40	33,33
	E4	Ética Profissional e Cidadania	4	8	32	40	33,33
Subtotal 3			**18**	**64**	**256**	**320**	**266,7**
QUARTO MÓDULO	E1	Planejamento Estratégico	6	12	48	60	50
		Fundamentos de Logística		12	48	60	50
	E2	Estatística Aplicada	4	8	32	40	33,33
		Gestão da Qualidade		8	32	40	33,33
	E3	Empreendedorismo	4	8	32	40	33,33
		Técnicas de Recepção, Atendimento e Cobrança		8	32	40	33,33
Subtotal 4			**14**	**56**	**224**	**280**	**233,3**
			62	**240**	**960**	**1200**	**1000**
Núcleo complementar		**Prática Profissional Supervisionada**				**120**	**100**
CARGA HORÁRIA TOTAL DO CURSO						**1.320**	**1.100**

Fonte: IFRO (2016).

Quadro 1: Matriz Curricular do Curso Técnico em Finanças Concomitante ao Ensino Médio

CURSO TÉCNICO EM FINANÇAS CONCOMITANTE AO ENSINO MÉDIO *CAMPUS* PORTO VELHO ZONA NORTE Matriz aprovada pela Resolução nº 14/CONSUP/IFRO/2016 Organização conforme a LDB 9.394/96, Art. 36, e a Resolução CEB/CNE 6/2012 Duração da aula: 50 minutos							
Períodos/ Módulos/ Etapas[1]		Disciplinas	Semanas letivas	Número de aulas		TOTAL (Hora-Aula)	TOTAL (Hora-Relógio)
				Telepresencial	EaD		
PRIMEIRO MÓDULO	E1	Ambientação para EaD[2]	4	8	32	40	33,33
	E2	Português Instrumental	4	8	32	40	33,33
		Introdução à Informática		8	32	40	33,33
	E3	Fundamentos de Matemática Financeira	4	8	32	40	33,33
		Fundamentos de Economia		8	32	40	33,33
	E4	Direito e Legislação Comercial	4	8	32	40	33,33
		Contabilidade Geral		8	32	40	33,33
	E5	Fundamentos de Administração	4	8	32	40	33,33
		Orientação para a Pesquisa e Prática Profissional		8	32	40	33,33
Subtotal 1			**20**	**72**	**288**	**360**	**300**
SEGUNDO MÓDULO	E1	Matemática Financeira Aplicada	4	8	32	40	33,33
		Contabilidade de Custos		8	32	40	33,33
	E2	Ética Profissional e Cidadania	4	8	32	40	33,33
		Redação Científica e Oficial		8	32	40	33,33
	E3	Planejamento Financeiro	4	8	32	40	33,33
		Fundamentos de direito tributário		8	32	40	33,33
	E4	Técnicas de Recepção, Atendimento e Cobrança	4	8	32	40	33,33
		Tópicos de Economia Monetária		8	32	40	33,33
Subtotal 2			**16**	**64**	**256**	**320**	**267**
TERCEIRO MÓDULO	E1	Estatística Aplicada	4	8	32	40	33,33
		Empreendedorismo		8	32	40	33,33
	E2	Segurança, Meio Ambiente e Saúde	4	8	32	40	33,33
		Gestão Tributária		8	32	40	33,33
	E3	Análise das Demonstrações Financeiras	4	8	32	40	33,33
		Análise de Investimento Financeiro		8	32	40	33,33
	E4	Fundamentos de Legislação Trabalhista	4	8	32	40	33,33
		Projetos Empresariais		8	32	40	33,33
Subtotal 3			**16**	**64**	**256**	**320**	**267**
Núcleo Complementar		**Prática Profissional**				240	200
CARGA HORÁRIA TOTAL DO CURSO			**52**	**200**	**800**	**1.240**	**1.034**[3]

Fonte: IFRO (2015)

ANEXO B – Parecer do Comitê de Ética em Pesquisa (CEP)

UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

Plataforma Brasil

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: PREDIÇÃO E REDUÇÃO DE EVASÃO DE ALUNOS EM EAD COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL E LEARNING ANALYTICS

Pesquisador: MARIELA MIZOTA TAMADA

Área Temática:

Versão: 1

CAAE: 88677018.3.0000.5020

Instituição Proponente: Universidade Federal do Amazonas - UFAM

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 2.757.851

Apresentação do Projeto:

Este projeto proposto para tese de doutorado insere-se na linha de pesquisa de Informática na Educação do Programa de Pós-Graduação emInformática da Universidade Federal do Amazonas - UFAM, tendo como tema "Predição e redução de evasão de alunos EaD com técnicas de Inteligência Artificial e Learning Analytics".

Conforme Moran e Valente (2011) a aprendizagem efetiva ocorre a partir da composição de duas concepções: a informação deve ser acessada e o conhecimento deve ser construído pelo aprendiz. O desafio da Educação, de modo geral, e da Educação a Distância (EaD), em particular, está em criar p condições para que a aprendizagem ocorra baseada nestas duas concepções

UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PPGI- PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA

Termo de Compromisso de Utilização de Dados

Eu, **MARIELA MIZOTA TAMADA**, da UNIVERSIDADE FEDERAL DO AMAZONAS, **do curso de Doutorado do PPGI (Programa de Pós-graduação em Informática)**, no âmbito do projeto de pesquisa intitulado: **"Predição e redução de evasão de alunos EaD com técnicas de Inteligência Artificial e *Learning Analytics*"**, comprometo-me com a utilização dos dados contidos **no banco de dados dos cursos EaD do IFRO (Instituto Federal de Rondônia),** a fim de obtenção dos objetivos previstos, e somente após receber a aprovação do sistema CEP-CONEP.

Comprometo-me a manter a confidencialidade dos dados coletados no **banco de dados,** bem como com a privacidade de seus conteúdos.

Esclareço que os dados a serem coletados se referem aos dados **sobre cursos, disciplinas, material didático, avaliações, estudantes e outros a partir do período de 01/07/2013.**

Declaro entender que é minha a responsabilidade de cuidar da integridade das informações e de garantir a confidencialidade dos dados e a privacidade dos indivíduos que terão suas informações acessadas.

Também é minha a responsabilidade de não repassar os dados coletados ou o banco de dados em sua íntegra, ou parte dele, à pessoas não envolvidas na equipe da pesquisa.

Por fim, comprometo-me com a guarda, cuidado e utilização das informações apenas para cumprimento dos objetivos previstos nesta pesquisa aqui referida. Qualquer outra pesquisa em que eu precise coletar informações serão submetidas a apreciação do CEP/UFAM.

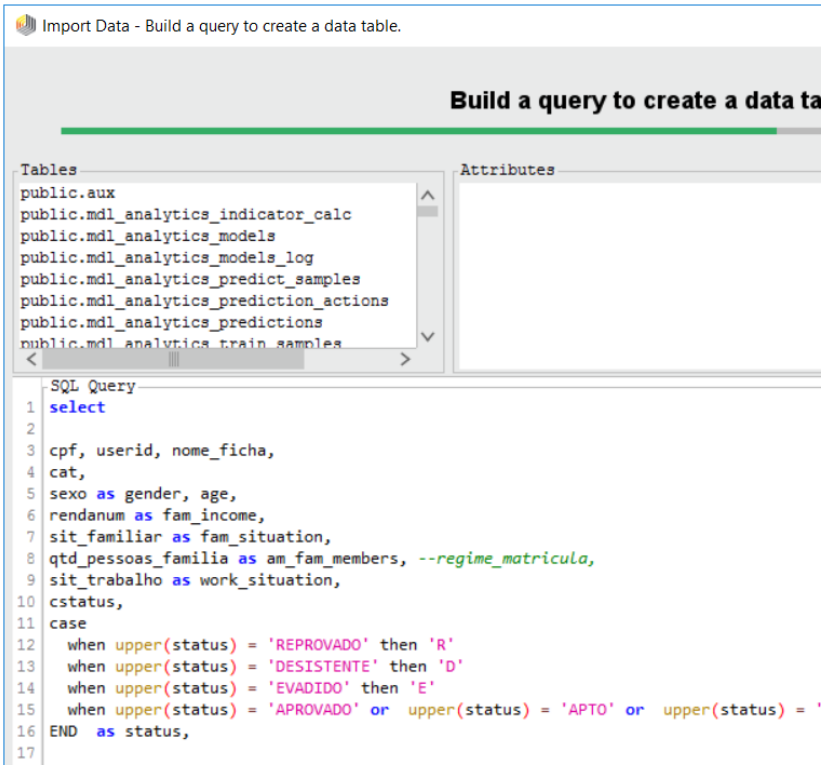
Porto Velho, 02/04/2018.

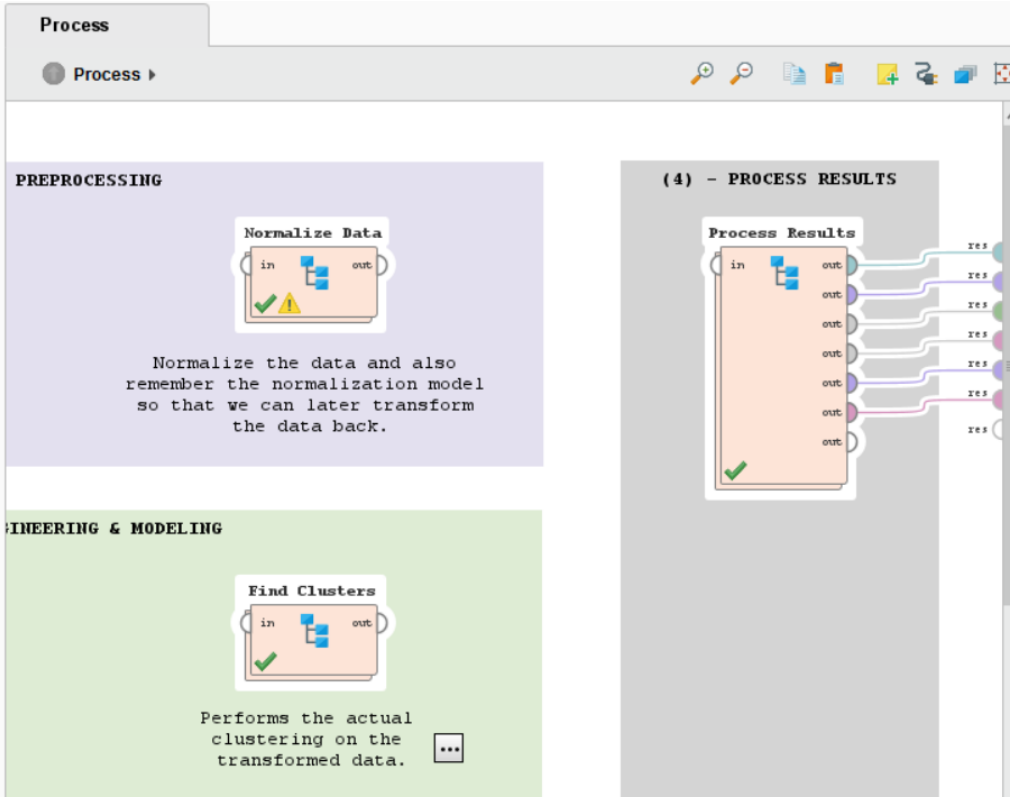
_____

Mariela Mizota Tamada
(pesquisador responsável)

ANEXO C - Rapid Miner

Tela para colocar a query para seleção da entrada de dados.





ANEXO D - Valores para o atributo module da tabela mdl_logtabela

Em negrito foram destacos os módulos utilizados para filtrar no log

admin	lesson
assign	library
attforblock	login
book	lti
calendar	message
category	notes
chat	page
choice	**quiz**
course	**resource**
data	role
dialogue	survey
discussion	tag
folder	**url**
forum	user
glossary	webservice
grade	workshop
label	

ANEXO E – Valores para o atributo action da tabela mdl_log

Destacados em negrito os valores da coluna mdl_log.action utilizados para filtrar no log. Os actions disponíveis dependem da escolha da coluna module.

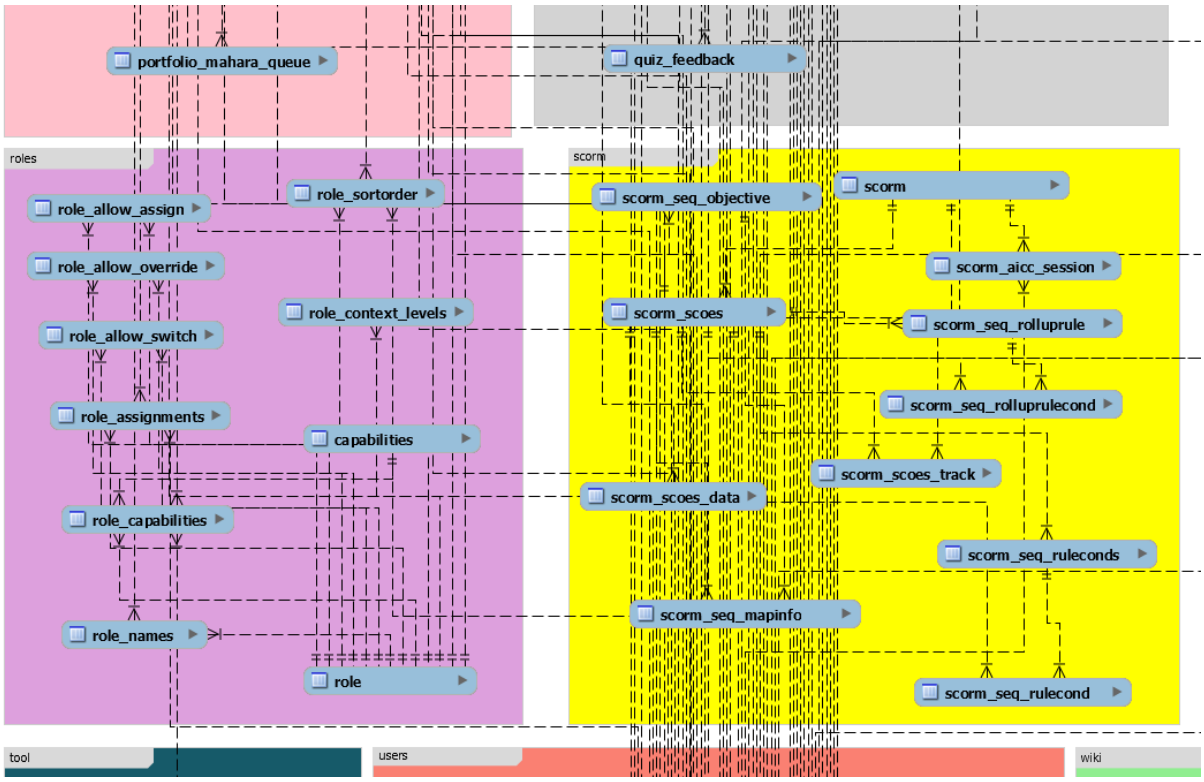
add	lock submission	**submit**
add assessment	login	submit for grading
add chapter	logout	subscribe
add contact	mail digest error	subscribeall
add discussion	mail error	talk
add entry	mailer	templates saved
add mod	manualgrade	templates view
add post	mark read	tool capability
add submission	moodle_enrol_get_enrolled_users	unassign
approve entry	moodle_enrol_get_users_courses	unblock contact
assign	moodle_user_get_users_by_id	unlock submission
attempt	moodle_webservice_get_siteinfo	unsubscribe
attendance taked	move	unsubscribeall
automatically create user token	move discussion	update
block contact	new	update aggregate grades
choose	open conversation	update assessment
choose again	preview	update chapter
close attempt	print	update entry
completion updated	print chapter	update grade
continue attempt	prune post	update mod
core_course_get_contents	remove contact	update post
core_user_add_user_device	reply	update submission
delete	report	update switch phase
delete attempt	report live	user report
delete discussion	report log	**view**
delete entry	report outline	**view all**
delete mod	report participation	**view chapter**
delete post	report questioninstances	**view confirm submit assignment form**
disapprove entry	report stats	**view conversation**
download all submissions	report viewed	**view discussion**
edit	reveal identities	**view feedback**
edit all	revert submission to draft	**view form**
edit override	review	**view forum**
editquestions	search	**view forums**
editsection	sending requested user token	**view grade**
end	session updated	**view grading form**
enrol	sessions added	**view graph**
error	sessions deleted	**view report**
fields add	start	**view section**
fields delete	start tracking	**view submission**
fields update	status added	**view submission grading table**
flag	status updated	**view submit assignment form**
grade submission	stop tracking	**view subscribers**
grant extension	submission statement accepted	**view summary**
launch	submissioncopied	write

ANEXO F – Moodle, *schema* do banco de dados

No site oficial do Moodle tem informação sobre o schema do banco de dados:

https://docs.moodle.org/dev/Database_Schema

Na figura abaixo só tem um recorte das principais tabelas e seu diagrama de relacionamento, mas são tantos relacionamentos que fica difícil de acompanhar e trata-se só de uma figura. Ela não é clickável para obter outras informações. Isso nas versões mais antigas.



Só a partir da versão recente 3.5 que foi disponibilizado uma documentação mais detalhada com descrição das tabelas, significado das colunas e as chaves estrangeiras. E de forma interativa por assunto, ou seja, matrículas, tabelas do fórum, tabelas do questionário etc., porém só tem uma parte das tabelas, consideradas as mais importantes. https://www.examulator.com/er/

https://www.examulator.com/er/output/index.html

https://www.examulator.com/er/output/tables/analytics_predictions.html

ANEXO G – Dicionário de dados - Exemplo

Tabela	Coluna do Tipo de Instância	Coluna da Instância	Descrição
mdl_context	contextlevel Registra tabela de domínio do contexto: 10 – Sistema 30 – Usuário 40 = Cat. de curso 50 – Curso ...	instanceid Registra instância, ou seja, chave estrangeira conforme o contexto definido na coluna contextlevel. Se o contexto for 50, a chave estrangeira será da tabela mdl_course	A tabela mdl_context registra o contexto de curso, usuário, grupo etc para efeito de gerenciamento de permissão.
mdl_course_modules	module Registra chave estrangeira da tabela mdl_modules	instance Registra instância, ou seja, chave estrangeira conforme o valor da coluna module. Se a coluna module tiver id referente ao fórum, a chave estrangeira será da tabela mdl_forum	A tabela mdl_course_modules registra os recursos / atividades cadastrados no curso.
mdl_grade_items	itemmodule Registra o nome do módulo	iteminstance Registra instância, ou seja, chave estrangeira conforme o valor da coluna itemmodule. Se a coluna itemmodule tiver valor forum, a chave estrangeira será da tabela mdl_forum	A tabela mdl_grade_items registra as avaliações geradas por diversos plugins de atividade instanciadas no ambiente do curso.

ANEXO H - Estrutura da tabela de log do Moodle a partir da versão 2.7 (tabela mdl_logstore_standard_log)

Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo
id	bigint(10)			Não
eventname	varchar(255)	utf8_bin		Não
component	varchar(100)	utf8_bin		Não
action	varchar(100)	utf8_bin		Não
target	varchar(100)	utf8_bin		Não
objecttable	varchar(50)	utf8_bin		Sim
objectid	bigint(10)			Sim
crud	varchar(1)	utf8_bin		Não
edulevel	tinyint(1)			Não
contextid	bigint(10)			Não
contextlevel	bigint(10)			Não
contextinstanceid	bigint(10)			Não
userid	bigint(10)			Não
courseid	bigint(10)			Sim
relateduserid	bigint(10)			Sim
anonymous	tinyint(1)			Não
other	longtext	utf8_bin		Sim
timecreated	bigint(10)			Não
origin	varchar(10)	utf8_bin		Sim
ip	varchar(45)	utf8_bin		Sim
realuserid	bigint(10)			Sim

Exemplo de retorno de dados da tabela de log do Moodle a partir da versão 2.7

Resultado da query

```
select * from mdl_logstore_standard_log limit 20
```

mdl_logstore_standard_log 1

```
select * from mdl_logstore_standard_  Enter a SQL expression to filter results (use Ctrl+Space)
```

	id	eventname	component	action	target	objecti	objectid	r	ed	con	co	userid	c	other
1	1	\core\event\user_login_failed	core	failed	user_login	[NULL]	[NULL]	r	0	1	10	0	6,482	0
2	2	\core\event\user_login_failed	core	failed	user_login	[NULL]	[NULL]	r	0	1	10	0	6,482	0
3	3	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
4	4	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
5	5	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
6	6	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
7	7	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
8	8	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
9	9	\core\event\user_login_failed	core	failed	user_login	[NULL]	[NULL]	r	0	1	10	0	0	0
10	10	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
11	11	\core\event\course_viewed	core	viewed	course	[NULL]	[NULL]	r	2	57,078	50	1,268	6,482	1,268
12	12	\tool_usertours\event\tour_started	tool_usertours	started	tour	tool_usert	2	r	2	57,078	50	1,268	6,482	1,268
13	13	\tool_usertours\event\step_shown	tool_usertours	shown	step	tool_usert	7	r	2	57,078	50	1,268	6,482	1,268
14	14	\tool_usertours\event\tour_ended	tool_usertours	ended	tour	tool_usert	2	c	2	57,078	50	1,268	6,482	1,268
15	15	\mod_resource\event\course_modul	mod_resource	viewed	course_mod	resource	13,251	r	2	60,567	70	27,582	6,482	1,268
16	16	\core\event\user_loggedin	core	loggedin	user	user	6,482	r	0	1	10	0	6,482	0
17	17	\core\event\config_log_created	core	created	config_log	config_log	2,340	c	0	1	10	0	6,482	0
18	18	\core\event\config_log_created	core	created	config_log	config_log	2,341	c	0	1	10	0	6,482	0
19	19	\core\event\config_log_created	core	created	config_log	config_log	2,342	c	0	1	10	0	6,482	0

other
a:2:{s:8:"username";s:11:"87112035
a:2:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:1:{s:8:"username";s:11:"87112035
a:2:{s:8:"username";s:7:"2118147";s
a:1:{s:8:"username";s:11:"87112035
N;
a:1:{s:7:"pageurl";s:63:"https://curs
a:3:{s:7:"pageurl";s:63:"https://curs
a:3:{s:7:"pageurl";s:63:"https://curs
N;
a:1:{s:8:"username";s:11:"87112035
a:4:{s:4:"name";s:24:"showdatarete
a:4:{s:4:"name";s:20:"courseendda
a:4:{s:4:"name";s:20:"predictionspr