UNIVERSIDADE DO VALE DO TAQUARI
CURSO DE ENGENHARIA DA COMPUTAÇÃO

Análise e predição de evasão na Educação a Distância de uma Universidade Comunitária utilizando técnicas de mineração de dados

Vinícius Rockenbach

Lajeado, 29 de Junho de 2020

Vinícius Rockenbach

Análise e predição de evasão na Educação a Distância de uma Universidade Comunitária utilizando técnicas de mineração de dados

Monografia apresentada na disciplina de Trabalho de Conclusão de Curso, do curso de Engenharia da Computação, da Universidade do Vale do Taquari - Univates, como parte da exigência para a conclusão do título de Bacharel em Engenharia da Computação.

Orientador: Edson Moacir Ahlert.

Lajeado, 29 de Junho de 2020

AGRADECIMENTOS

Primeiramente agradeço aos meus pais, Valdir e Gemanir, por estarem ao meu lado durante este período, tanto nos momentos de alegria quanto nos momentos de dificuldade e por sua dedicação ao longo de todos estes anos, auxiliando no meu crescimento pessoal e profissional dia após dia.

Faço um agradecimento muito especial a minha namorada, Ana Luiza, pela compreensão e companheirismo durante esse período, sempre me motivando e me dando forças.

Por fim, quero agradecer a todos professores e colegas que auxiliaram de uma forma ou outra na construção deste trabalho, em especial aos colegas Matheus e Artur e ao Sr. Edson Moacir Ahlert, que me deu a honra de sua orientação, se mostrando sempre prestativo.

RESUMO

A Mineração de dados vem crescendo nos últimos anos juntamente com o interesse das mais diversas áreas de atuação na descoberta de dados que auxilie a alcançar uma vantagem no mercado e, dentro da área da educação, o comportamento dos estudantes muitas vezes indica seus interesses e motivações. Esta pesquisa tem o objetivo de relacionar estes dois extremos aplicando técnicas de mineração de dados e aprendizado de máquina em dados provenientes do banco de dados de uma universidade comunitária do interior do Rio Grande do Sul explorando os conceitos da mineração de dados educacional para verificar a tendência dos alunos a evadirem dos cursos da modalidade EaD, dentro de um período de dois anos. Os resultados obtidos através dos experimentos utilizando diferentes técnicas de mineração de dados são comparados para se encontrar a forma mais eficiente de prever a evasão dos alunos. Após o treinamento dos modelos preditivos e a aplicação deles sobre o conjunto de teste, ficou constatado que os algoritmos Random Tree e Decision Tree obtiveram os melhores resultados, atingindo percentuais superiores a 98% nos melhores resultados. Os resultados também são comparados com outros resultados observados em estudos dentro do mesmo campo de pesquisa ou que também se utilizam das técnicas de mineração de dados educacionais, verificando-se resultados semelhantes em alguns deles.

Palavras-chave: Educação a distância, Ambiente Virtual de Aprendizagem, Mineração de dados.

ABSTRACT

Data mining has been growing in recent years along with the great interest of the most diverse areas of activity in the discovery of data that helps on achieving some advantage in the market and, within the area of education, the behavior of students often indicates their interests and motivations. This research aims to relate these two extremes by applying data mining and machine learning techniques to data from the database of a community university in the interior of Rio Grande do Sul state exploring the concepts of educational data mining to verify the trend of students to evade on distance education courses, within a period of two years. The results obtained through the experiments using different data mining techniques are compared to find the most efficient way to predict students' dropout. After training the predictive models and applying them to the test set, it was found that the Random Tree and Decision Tree algorithms obtained the best results, reaching percentages greater than 98% in the best results. The results are also compared with other results observed in studies within the same research field or studies that also use educational data mining techniques, with similar results being seen in some of them.

Key words: Distance Education, Virtual Learning Environment, Data Mining.

LISTA DE ILUSTRAÇÕES

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

AD	Árvore de Decisão
AVA	Ambiente Virtual de Aprendizagem
DB	*Database*
DM	*Data Mining*
EaD	Educação a Distância
EDM	*Educational Data Mining*
GIGO	Garbage In, Garbage Out
GPL	*General Public License*
GPU	*Graphical Processor Unit*
GUI	*Graphical User Interface*
IA	Inteligência Artificial
IES	Instituição de Ensino Superior
KDD	*Knowledge Discovery in Databases*
K-NN	*K-Nearest Neighbors*
ML	*Machine Learning*
MLP	*Multilayer perceptron*
NB	Naive Bayes
RNA	Redes Neurais Artificiais
SGBD	Sistema de Gestão de Banco de Dados
SMO	*Sequential Minimal Optimization*
SVM	*Support Vector Machines*
TCC	Trabalho de Conclusão de Curso
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

Com a tecnologia da informação já enraizada na sociedade moderna e a emergente ascensão da computação em nuvem e das mídias digitais, instituições de ensino, tanto no Brasil como mundo afora, estão buscando adaptar-se a essas mudanças nos métodos de acesso à informação e ao material de ensino. Seguindo essa tendência, educadores estão mudando de forma significativa suas maneiras de ensinar, assim como os estudantes o seu jeito de aprender. A utilização da internet facilitou de forma exponencial o acesso à informação e a expansão da educação por meio de ferramentas e softwares hospedados na web e novas e diferentes metodologias educacionais, que vêm a compor a Educação a Distância (EaD).

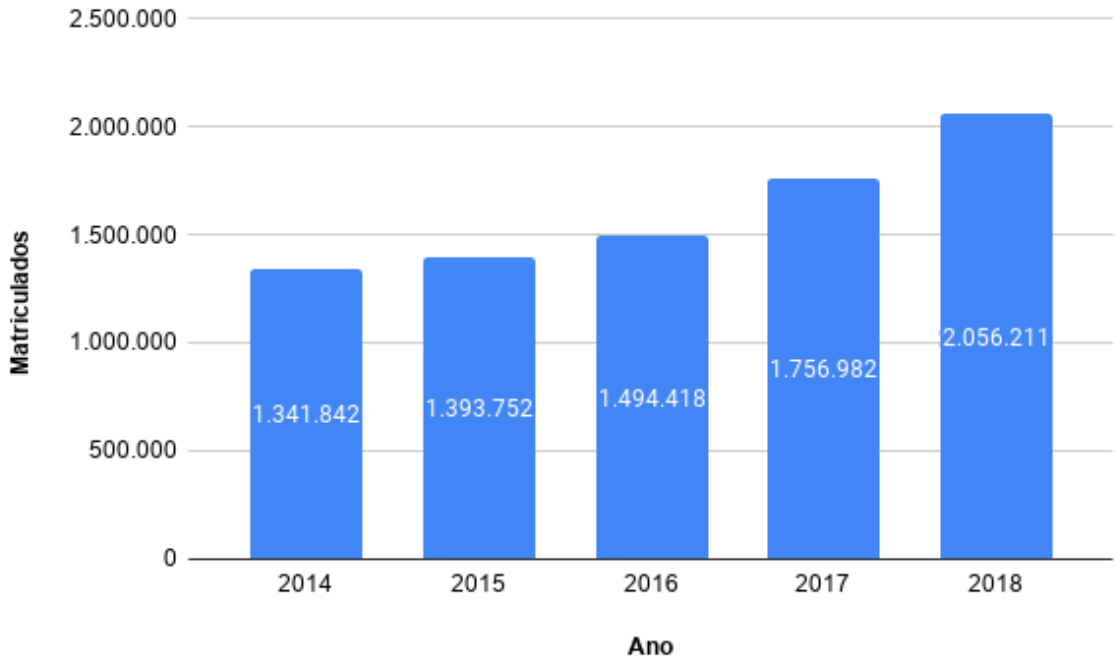
Diversas pesquisas já realizadas por institutos especializados, como o INEP, apontam para uma direção onde a adesão a modalidade a distância só tende a crescer, levando em conta tanto a facilidade que as ferramentas de tecnologia da informação proporcionam no acesso à informação, quanto uma possível questão financeira ou logística para a pessoa.

Atualmente, um em cada três estudantes matriculados no ensino superior estão realizando um curso da modalidade a distância, enquanto o ensino presencial apresentou quedas no número de estudantes, as matrículas no ensino a distância apresentaram o maior salto desde 2008, chegando a um total de 17,6% de aumento do ano de 2016 para 2017 e de 14,55% de 2017 para 2018,

totalizando um número de estudantes na casa de 2 milhões no ano de 2018, o que equivale a 24,33% do total de matriculados em todo o cenário do ensino superior.

Outro dado interessante de observar é que, no intervalo de 10 anos que compreende o período entre 2007 e 2017, o ensino a distância obteve um aumento de 375,2% nas matrículas, enquanto a modalidade presencial cresceu apenas 33,8% (INEP, 2017). Esse crescimento no número de matriculados é perceptível de um ano para o outro, como podemos observar em âmbito nacional no Gráfico 1.

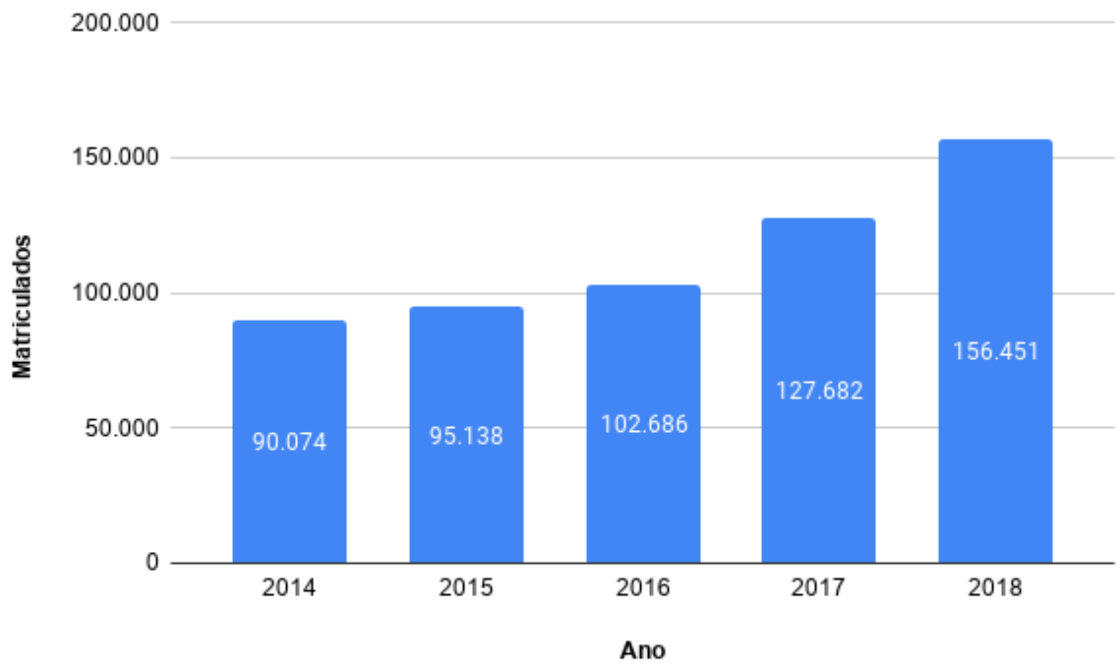
Gráfico 1 - Total de matriculados na modalidade EaD no Brasil.



Fonte: Adaptado de INEP (2018).

Considerando somente no Rio Grande do Sul, o crescimento na adesão à modalidade a distância é semelhante nos anos anteriores a 2016, porém, o aumento da adesão de 2016 para 2017 e 2018 é ainda maior que a porcentagem do país, como se pode observar no Gráfico 2.

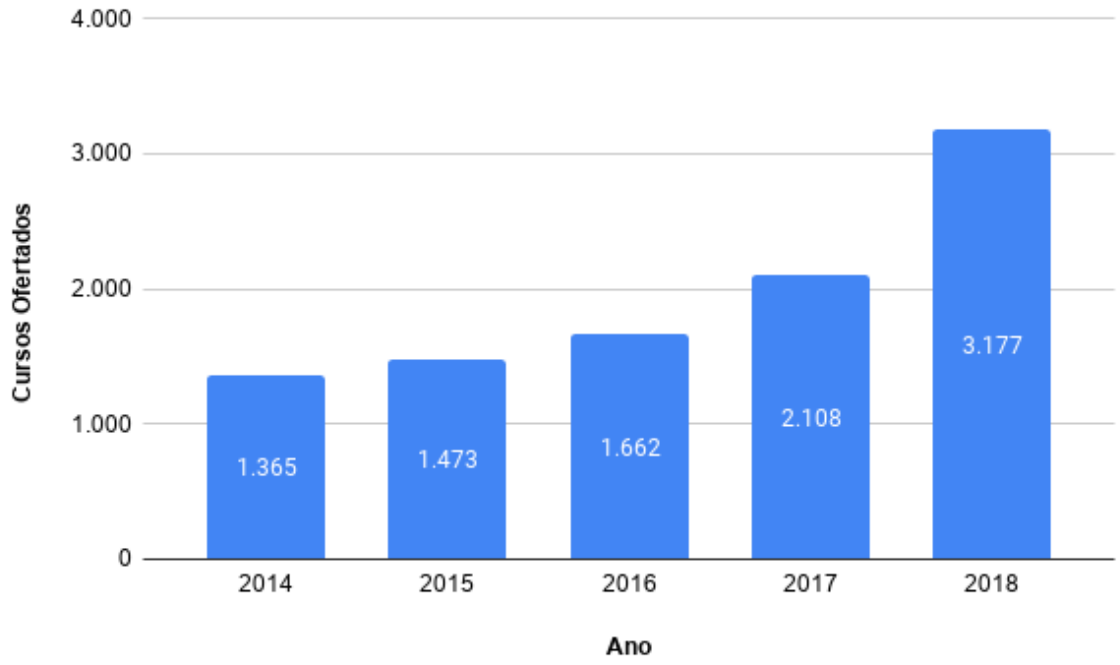
Gráfico 2 - Total de matriculados na modalidade EaD no Rio Grande do Sul.



Fonte: Adaptado de INEP (2018).

Como as grandes instituições sempre estão atentas nos censos do INEP, baseando-se nessa crescente da adesão a EaD e também, levando em consideração a queda nas matrículas dos cursos presenciais, muitas IES estão apostando no ensino a distância e concentrando suas forças no aperfeiçoamento e melhoria dos cursos já ofertados, assim como na busca da oferta de novos cursos dentro da modalidade. Sendo assim, dados oficiais tornam possível observar como é o interesse das Instituições de Ensino (IES) na modalidade a distância, já que o número de cursos ofertados no país na modalidade passou de 1.662 para 2.108 cursos do ano de 2016 para 2017 e de 2.108 para 3.177 em 2018, apresentando um aumento de 33,64%, batendo o crescimento de 26,8% de 2017 que era o maior crescimento registrado desde 2009 (INEP, 2018). Podemos observar no Gráfico 3 esse crescimento dentro do território nacional.

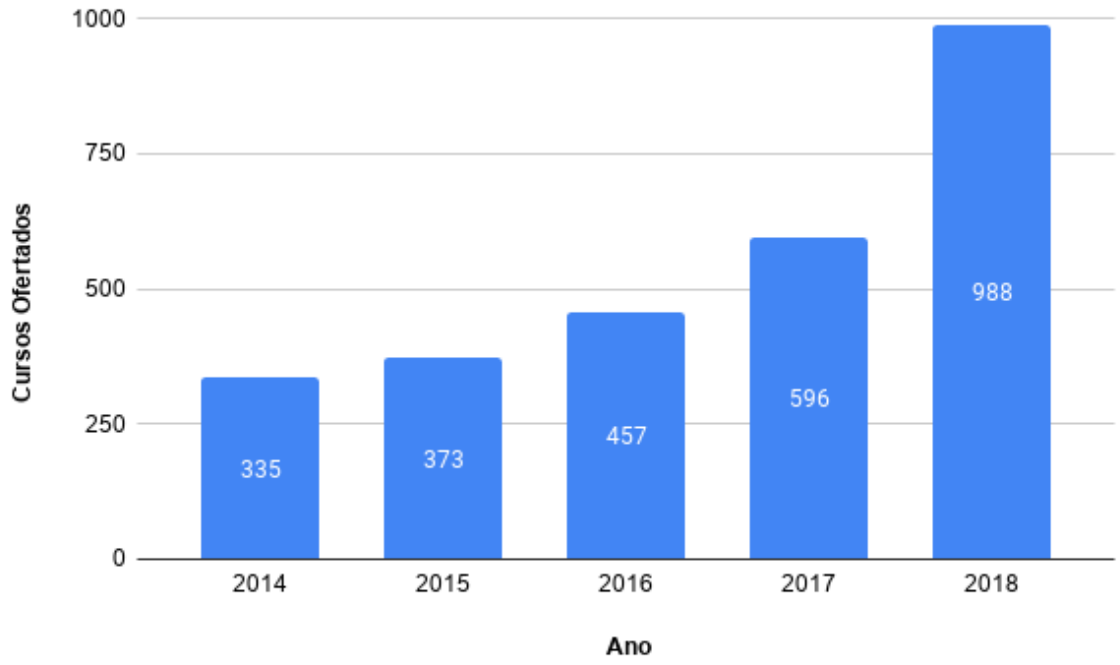
Gráfico 3 - Total de cursos ofertados na modalidade EaD no Brasil.



Fonte: Adaptado de INEP (2018).

A região sul do país fica em segundo lugar com mais ofertas de cursos da modalidade EaD, ficando atrás apenas da região sudeste que conta com grandes metrópoles como São Paulo, Rio de Janeiro e Belo Horizonte. Nessa região também foi observado um aumento substancial, se comparado aos anos anteriores, conforme mostra o Gráfico 4.

Gráfico 4 - Total de cursos ofertados na modalidade EaD na região Sul.



Fonte: Adaptado de INEP, 2014, 2015, 2016 e 2017.

Com os dados concretos registrados pelo INEP (2018), podemos trazer essas observações para um cenário mais próximo e observar que a Universidade do Vale do Taquari - Univates, a qual havia registrado somente um único curso da modalidade EaD no Censo de 2017, apareceu com um total de 14 cursos ofertados no Censo de 2018.

1.2 Justificativa

Altos índices de evasão na educação da distância já se tornaram tema frequente de estudos dentro da área de pesquisa educacional e, resultados diferentes são apresentados em pesquisas semelhantes a esta nos últimos anos.

Conforme Iaralham (2009), a EaD é um processo educativo, organizado e sistemático, exigindo dos alunos uma comunicação que leva a utilização de diferentes meios tecnológicos de informação e comunicação para alcançar a aprendizagem de forma efetiva e por isso, a EaD é considerada a maior referência para a mudança significativa que vem ocorrendo no ensino superior. Sendo assim, as IES vêm buscando uma forma de garantir um *feedback* para saber o que é necessário para se aproximar e atender da forma mais eficiente possível as

necessidades de seus discentes e as exigências da sociedade e do mercado de trabalho e é isso que a utilização de um Ambiente Virtual de Aprendizagem (AVA) garante.

Diversas pesquisas vêm sendo desenvolvidas nos últimos anos buscando compreender o efeito, avaliando as diferentes variáveis que causam este fenômeno, buscando atenuar o problema. Uma destas linhas de pesquisa é a utilização de Técnicas de Extração do Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD), com a finalidade de extrair conhecimento das bases de dados disponíveis e analisar o comportamento do estudante dentro do AVA através de modelos preditivos que são treinados para aprender e entender esse estudante.

Com os dados em mãos se busca, através da aplicação das técnicas de mineração e do desenvolvimento de modelos preditivos eficazes, identificar os alunos que têm propensão a evadir dos cursos da modalidade a distância e oferecer subsídios para que os responsáveis pelo ensino possam criar medidas preventivas afim de mitigar esse problema.

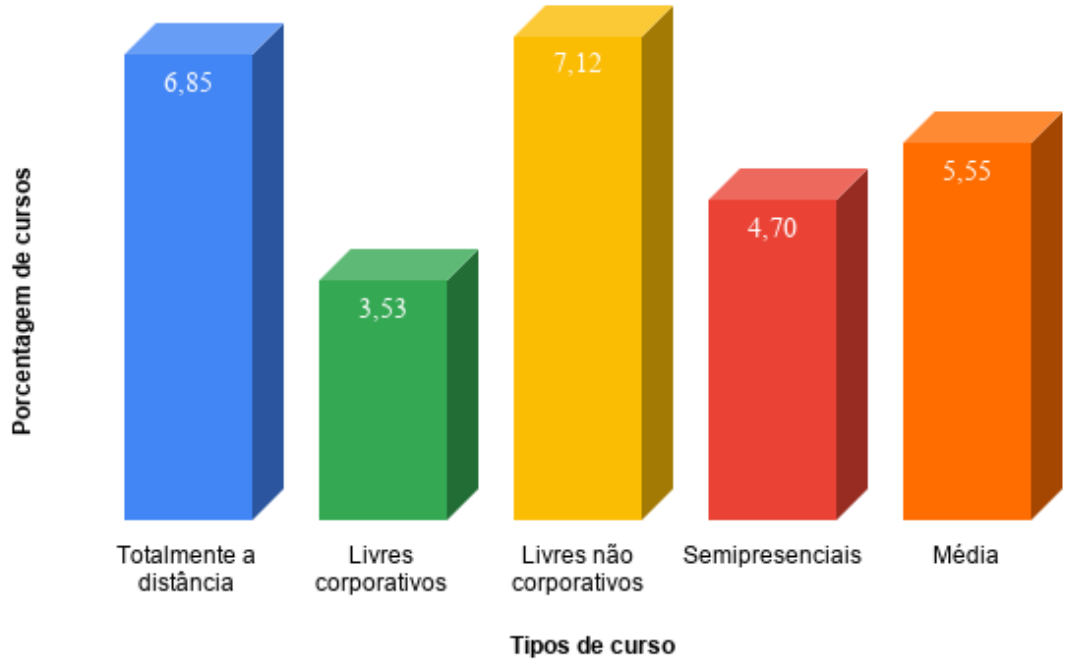
1.3 Problema

Apesar do crescimento no número de ingressantes nos cursos na modalidade a distância, dados extraídos no CensoEaD.br, disponibilizados por ABED (2018), demonstram como a evasão continua sendo a pedra no sapato das instituições participantes, sendo apontada como o maior problema enfrentado dentro da gestão da EaD. Estes índices de evasão tem uma grande variação, levando em conta os tipos de EaD praticados, sendo estes semipresenciais, totalmente a distância, cursos livres ou cursos livres corporativos.

Considerando os tipos de cursos EaD e os dados do CensoEaD.br ABED (2018), verifica-se que os cursos livres apresentam a maior porcentagem de cursos com taxas de evasão entre 26 e 50% com um número superior a 7%. Os cursos totalmente a distância vem logo atrás com porcentagens perto dos 7%. Já os cursos corporativos são os que apresentam o menor número. No total, é

possível verificar que uma média de 5,55 dos cursos apresentam taxas de evasão superiores a 26%, conforme é possível observar no Gráfico 5.

Gráfico 5 - Taxas de cursos à distância com porcentagens de evasão entre 26% e 50%.



Fonte: Adaptado de CensoEaD.br (ABED, 2018).

É possível que estes índices possuam esta variância por tratar-se de diferentes públicos. No tipo de curso totalmente a distância, como já sugere o nome, as aulas acontecem totalmente a distância e assim, os estudantes são expostos a diversos outros estímulos concorrentes em suas residências, no seu trabalho ou qual seja o outro ambiente escolhido para estudar. No segundo tipo, os semipresenciais, geralmente são cursos presenciais que contam com disciplinas ministradas *on-line* e fazem parte da grade curricular dos alunos que frequentam um curso presencial que, por apresentar um contato do estudante com a instituição, apresentam um número menor de evasão. O terceiro tipo, que são os cursos livres e não corporativos, também apresentam um valor expressivo na taxa de evasão e isso ocorre por estes serem cursos sem nenhuma interação que são em grande parte gratuitos e o estudante não possui um compromisso financeiro ou institucional de concluí-lo. Já o quarto e último tipo, são os cursos livres e corporativos, os quais apresentam a menor taxa de evasão, por se tratarem de

cursos empresariais onde os estudantes recebem o curso e o fazem para capacitação profissional, treinamento, aperfeiçoamento e/ou desenvolvimento de novas técnicas de gestão e/ou trabalho, as quais são totalmente relevantes para a parte administrativa ou operacional da organização.

Assim como no restante do país, a evasão também é um dos maiores problemas enfrentados na EaD da Univates, a qual, pelo pouco tempo de implantação, ainda carece de volume de dados e ferramentas que possam fornecer uma inteligência através da análise destes.

Dentro deste cenário, chegamos ao problema de pesquisa questionando se é possível, através das técnicas de mineração de dados, apresentar um modelo de predição eficiente na previsão dos alunos com tendência a evadir dos cursos da modalidade a distância também realizando uma exploração dos dados adquiridos até o momento oferecendo assim, subsídios para que os responsáveis pelo ensino tomem decisões e pensem em ações para mitigar este problema.

1.4 Objetivo Geral

O objetivo geral do estudo é extrair dados provenientes do AVA disponibilizado pelo Setor de Educação da Distância da Instituição e também do sistema de gestão da IES, aplicar diferentes técnicas de mineração de dados e modelos preditivos e identificar possíveis tendências existentes entre os acadêmicos dos Cursos de modalidade a distância da Universidade do Vale do Taquari – Univates que os levem a evadir do curso para então, fornecer subsídios que irão auxiliar na tomada de decisão e na formulação de medidas preventivas a fim de diminuir ao máximo a evasão dos estudantes.

1.4.1 Objetivos específicos

- Obter as informações oriundas da base de dados do AVA da Univates, as quais permitam uma correlação com a tendência dos estudantes dos cursos na modalidade EaD e informações referentes às motivações que levam a evasão dentro da modalidade;

- Analisar e gerar modelos preditivos capazes de identificar a tendência de diferentes estudantes e detectar possíveis comportamentos e motivos que indiquem a evasão na EaD da Univates;
- Avaliar o impacto de diferentes técnicas de mineração de dados na elaboração da classificação dos estudantes da EaD da Univates;

1.5 Estrutura do trabalho

O presente estudo foi dividido em 5 capítulos macros. No primeiro capítulo é inserido o contexto do tema Educação a Distância, a proposta de se analisar e fazer a avaliação dos dados de um AVA e do sistema de gestão da IES através das técnicas de KDD, os objetivos geral e específicos e, por fim, possíveis benefícios que o estudo deverá trazer para a instituição.

No segundo capítulo são apresentadas as referências teóricas. Este capítulo irá contextualizar a Educação a distância, apresentar conceitos dos Ambientes Virtuais de Aprendizagem, conhecidos como AVAs. Também serão introduzidos conceitos sobre a Mineração de Dados Educacionais, sobre a Predição do Desempenho de Estudantes e uma revisão sobre Sistemas de Apoio a Decisão, tópicos essenciais para uma plena compreensão deste trabalho.

O terceiro capítulo descreve trabalhos que, assim como este, utilizam de métodos de mineração de dados educacionais, seja para predizer desempenho dos estudantes, seja para apresentar motivos para as taxas de evasão ou seja para a avaliar um determinado AVA utilizado no ensino a distância.

O quarto capítulo tem a responsabilidade de esclarecer as metodologias utilizadas para a realização do trabalho, apresentando as técnicas aplicadas em cada etapa do processo de KDD e apresentar as ferramentas utilizadas no decorrer do desenvolvimento do trabalho para se alcançar o objetivo final.

O capítulo cinco descreve os experimentos realizados, avaliando os métodos propostos, a acurácia dos modelos preditivos obtidos e a sua capacidade na identificação e predição da evasão dos alunos e as motivações que os levam a ter esse comportamento, apresentando uma análise dos resultados obtidos.

E por final, são apresentadas as considerações finais do estudo e são discutidos desafios para o futuro.

2 REFERENCIAL TEÓRICO

Com uma pesquisa aprofundada nos motores de busca disponíveis pode se constatar a existência de diversos artigos, teses, dissertações e entre outros trabalhos, os quais possuem relação com esse trabalho, seja de forma direta ou indireta. Neste estudo, buscou-se argumentar e realizar questionamentos quanto a avanços teóricos e práticos citados em outras publicações de uma forma que pudesse ser proveitoso para alcançar os motivos corretos que viriam a sedimentar a importância de se abordar diferentes formas e buscar soluções para a identificação e o tratamento efetivo do problema da evasão na modalidade EaD.

2.1 A Educação a Distância

No Brasil, definido em 2005, a EaD foi descrita no Decreto de número 5.622 como:

> [...] a modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorre com a utilização de meios e tecnologias de informação e comunicação, com estudantes e professores desenvolvendo atividades educativas em lugares ou tempos diversos (BRASIL, 2005, p.1).

Dentre os diversos autores pesquisados e que estudam a EaD, notou-se uma opinião unânime quanto à evolução histórica e a classificação das diferentes gerações da EaD. Muitos deles já citam e referem a EaD como algo que já havia surgido em civilizações antigas, considerando diferentes manuscritos utilizados para a difusão do cristianismo como pioneiros na educação da distância.

No que é considerado a EaD moderna, o êxito veio com os cursos implantados pela Open University da Inglaterra, a qual surgiu no final dos anos 1960, tendo o início efetivo dos cursos em 1970, tornando-se referência na visão de vários autores como sendo o marco da EaD contemporâneo (NUNES, 2009).

2.1.1 A EaD na Universidade do Vale do Taquari – Univates

Dentro da Univates os cursos na modalidade a distância possuem um funcionamento semelhante ao de outras universidades, onde os cursos são organizados por profissionais e disponibilizados via ambiente virtual Moodle. O estudante tem a possibilidade de acessar o material quando e onde quiser. Para estudar, o aluno lê o material sob orientação de tutores e realiza as atividades sempre a distância, contando com videoconferências semanais com professores que buscam uma maior interação com os mesmos. No final do módulo, de 200 horas, o estudante comparece para a avaliação presencial na data e hora marcados, junto ao polo o qual o mesmo possui vínculo de matrícula, já que, semelhante a Universidade Aberta do Brasil (UAB), a Univates conta com um sistema de polos, os quais estão distribuídos pelo estado do Rio Grande do Sul.

2.2 Ambientes Virtuais de Aprendizagem

Docentes e discentes atualmente fazem a utilização dos Ambientes Virtuais de Aprendizagem (AVAs) para o ensino e aprendizagem a distância os quais, como definido por Dias (2008), é um sistema responsável por fornecer suporte a atividade de qualquer natureza realizada pelo estudante, podendo ser considerado então, um conjunto de ferramentas de comunicação e interação utilizadas em diferentes partes do processo de aquisição do conhecimento, encontrando-se, atualmente, em uma plataforma totalmente *on-line*, podendo ser acessadas em qualquer computador ou dispositivo móvel pessoal que o estudante possua.

Segundo Gonzalez (2005), os AVAs contemplam um grande número de ferramentas as quais dão suporte a diferentes tipos de abordagem. Ele as classifica em 4 grandes categorias, sendo elas: ferramentas de comunicação;

ferramentas de coordenação; ferramentas de produção dos estudantes ou de cooperação e ferramentas de administração, conforme Quadro 1.

Quadro 1 - Grupos de ferramentas dos AVAs.

Grupo de ferramentas	**Descrição**
Ferramentas de Coordenação	Possuem a finalidade de organizar o curso, sendo utilizadas pelos professores para disponibilizar diferentes informações para os estudantes, tanto a respeito da metodologia do curso (avaliações, objetivos, duração) e da estrutura do ambiente de aprendizagem (recursos, dinâmica, agenda, etc), quanto a respeito das informações da parte pedagógica, como, material de apoio, material de leitura, entre outros recursos de perguntas frequentes).
Ferramentas de Comunicação	Estas ferramentas concentram diversos recursos como, fóruns para discussão, bate-papo, correio eletrônico e conferência entre os participantes. Possui o objetivo de facilitar a interação e o processo de ensino-aprendizagem entre os participantes, estimulando a colaboração e o aprendizado contínuo.
Ferramentas de Produção dos Estudantes ou de Cooperação	Nesta parte os AVAs oferecem um espaço para os estudantes organizarem seus trabalhos e/ou grupos através de portfólio, diário, mural ou perfil dos estudantes, podendo realizar publicações.
Ferramentas de Administração	Nas ferramentas de administração podemos encontrar recursos para o gerenciamento, do curso (cronograma, ferramentas que serão disponibilizadas, inscrições, etc), de estudantes (relatórios de acesso, frequência no ambiente, utilização das ferramentas, etc.) e de apoio a aprendizagem (inserir material didático, atualizar agenda, habilitar diferentes ferramentas no ambiente, etc.). Com isso é possível oferecer ao professor informações sobre a participação e o progresso dos alunos no decorrer do curso, podendo então visualizar quando é necessário apoiá-los e motivá-los durante o processo de aquisição do conhecimento.

Fonte: Gonzalez (2005).

Já Sabbatini (2007) faz essa classificação em apenas 3 conjuntos, são eles: ferramentas de comunicação; ferramentas de interação e ferramentas de avaliação. O conceito básico entre as duas diferentes classificações é o mesmo, sendo somente agrupadas e classificadas de forma diferente pelos autores, como mostra o Quadro 2.

Quadro 2 - Recursos de TI implementados nos AVAs.

Grupo de ferramenta	**Descrição**
Ferramentas de Comunicação	- Páginas simples de texto; - Páginas em HTML; - Acesso para arquivo de qualquer formato (PDF, DOC, PPT, áudio, vídeo, etc.) ou a qualquer link externo; - Acesso a diretórios (pastas ou arquivos no servidor); - Rótulos; - Lições interativas; - Livros eletrônicos (e-books); - Glossários (estático); - Perguntas frequentes.
Ferramentas de Interação	- Bate-papo (chat); - Fórum de discussão; - Diários; - Wikis (conteúdo colaborativo); - Glossários (colaborativo).
Ferramentas de Avaliação	- Avaliação do Curso; - Questionários de avaliação; - Ensaios corrigidos; - Tarefas e exercícios; - Enquetes.

Fonte: Sabbatini (2007).

Segundo Rodrigues (2016), o formato que será adotado para a disponibilização destas ferramentas pelo professor e, sua utilização, possibilita que dentro de um mesmo AVA seja possível se abordar o processo de ensino e aprendizagem em diferentes enfoques. Isso abre a possibilidade de se trabalhar dentro de um modelo de ensino linear, com uma sequência preestabelecida, com muito pouca ou até nenhuma possibilidade de o estudante ser o protagonista de seu ensino, até outros modelos que tragam uma proposta metodológica que possibilita a construção cooperativa do conhecimento entre os diferentes atuantes, primando pelo protagonismo do estudante.

Rodrigues (2016), ainda complementa que o professor que atua na modalidade EaD, que faça uso do AVA com a determinação de promover a educação, poderá escolher dentre as diversas teorias de aprendizagem, e aplicar o conteúdo juntamente do conjunto de ferramentas de uma forma que ela seja evidenciada.

Dentro da educação a distância, encontram-se diversos AVAs disponíveis e, entre os de software livre[1] os mais difundidos são o Amadeus[2], Moodle[3], e-ProInfo[4] e Eureka[5]. O Moodle é o AVA abordado nesta pesquisa por se tratar da ferramenta utilizada na instituição.

2.2.1 Moodle

Conforme Ribeiro *et al.* (2007), o Moodle é uma plataforma de Software Livre, que está sob licença GPL[6], podendo ser instalado, utilizado, modificado e até mesmo distribuído gratuitamente. Possui o objetivo de gerenciar o aprendizado, oferecendo trabalho colaborativo no ambiente virtual. Também permite a criação e a administração de cursos *on-line* com grupos de trabalho e comunidades de aprendizagem.

A forma que o Moodle utilizada para armazenar as interações realizadas pelos usuários se dá em forma de *logs*. Nestes podemos encontrar informações úteis que auxiliam na tomada de decisão quando utilizados com a técnica de mineração mais adequada.

A ferramenta Moodle conta, atualmente, com 105660 ambientes registrados em 229 países, sendo o Brasil o quarto país em que ela é mais utilizada. Mais de 17 milhões de cursos foram ofertados na plataforma, que conta com mais de 156 milhões de usuário espalhados no mundo todo (MOODLE, 2019).

2.2.2 Interações no AVA

Segundo Rodrigues (2016), mesmo com a farta oferta de ferramentas disponibilizadas pelos AVAs, detectar as deficiências na aprendizagem não é algo simples, mostrando que não basta simplesmente abrir uma sala com conteúdo e

[1] Softwares que estão sob as licenças reconhecidas pela Free Software Foundation.
[2] Projeto Amadeus. http://amadeus.cin.ufpe.br/.
[3] Moodle.org. https://moodle.org/.
[4] e-Proinfo. http://e-proinfo.mec.gov.br/.
[5] Eureka - PUCPR. https://eureka.pucpr.br/.
[6] GPL - Licença Pública Geral (do Inglês *General Public License*), criada por Richard Stallman, fundador da Free Software Foundation. Entre os Softwares Livres, a GPL atualmente é a licença de Software Livre mais difundida. (LICENSE, 2019).

ferramentas síncronas e assíncronas, mas sim, preciso realizar um acompanhamento periódico do estudante até o fim, para que ele se mantenha interessado e participativo nas diversas atividades propostas.

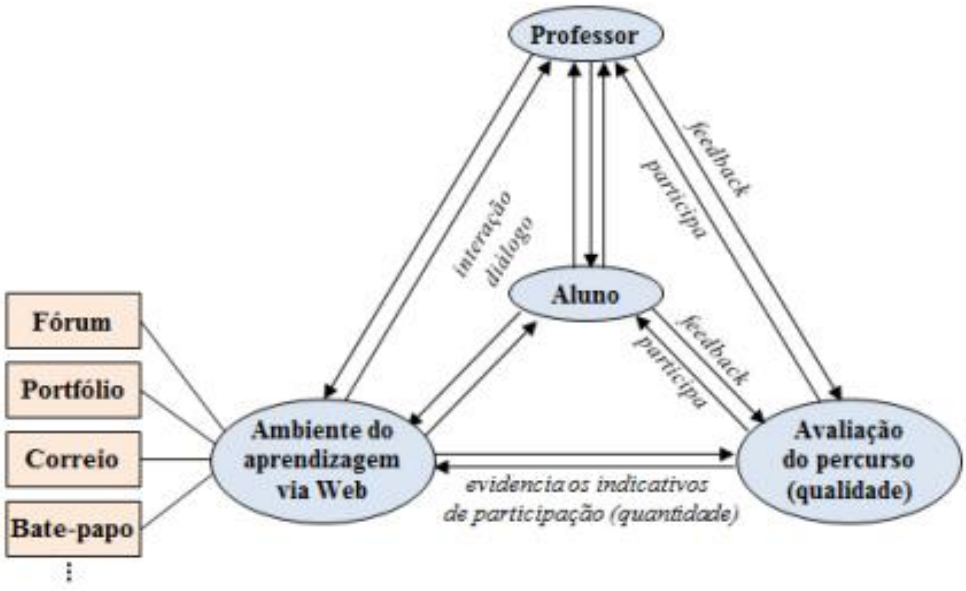
Complementando, Moran (2003) ainda afirma que, se a manutenção dessa motivação nos ambientes presenciais já merece uma grande atenção, nos ambientes virtuais a tarefa é ainda mais crítica, pois, é necessário envolver os estudantes em processos participativos, que os inclua, inspirando confiança. Em cursos que se limitam à transmissão da informação e do conteúdo, mesmo que de forma brilhante, há um grande risco de os alunos se desmotivarem no longo prazo, tornando-se necessário buscar o equilíbrio entre o teórico e o prático nas estratégias adotadas. Na sala de aula, professores obtêm mais facilmente um *feedback* da ocorrência de problemas no aprendizado e em cima disso avaliar e aplicar diferentes estratégias pedagógicas. Já no cenário virtual, isso é mais difícil de observar, pois normalmente o estudante só está acessível por e-mail, mensageiro instantâneo ou ferramentas disponíveis no AVA.

Diversas tecnologias da informação, como por exemplo, agentes de softwares[7] e KDD, estão sendo difundidas com a finalidade de auxiliar os professores da modalidade a distância no acompanhamento dos alunos, não somente em aspectos objetivos do seu desempenho, como avaliações e exercícios, mas também em aspectos subjetivos, como a motivação (LACHI *et al.*, 2002). Com isso, é buscado explorar melhor os registros das interações as quais os estudantes realizam nos ambientes virtuais e prover o suporte aos professores na coleta, identificação, seleção e análise de diferentes informações relevantes à avaliação formativa do aluno.

Na Figura 1 é possível se observar as diversas formas de interação entre estudantes, professores e os diversos recursos envolvidos em uma atividade de ensino e aprendizagem via Internet, todas elas alimentando as bases de dados, matéria prima para a aplicação do KDD.

[7] Agente de software é um programa que executa em segundo plano e tem como principais requisitos: um ciclo de vida contínuo no tempo, um ambiente de atuação, sensores para recolher informações do ambiente, atuadores que alteram o ambiente e têm autonomia, ou seja, funcionamento independente da interferência do usuário.

Figura 1 - Mapa das interações dentro do AVA.



Fonte: Souza (2007).

Conforme Rodrigues (2016), as possibilidades de interação nos ambientes virtuais de aprendizagem são muitas: a interação entre o estudante e o professor, entre o estudante e a turma e entre o estudante e os diferentes recursos disponíveis no AVA.

2.3 Evasão na EaD

Assim como no ensino superior em geral, na EaD não é diferente, pois os índices de evasão têm se mostrado um grande desafio para os gestores das IES e também para pesquisadores que trabalham para buscar, identificar causas e sugerir soluções para a diminuição deste fenômeno, pois, a evasão, é o problema de maior preocupação e o que mais atinge a modalidade a distância (DAUDT; BEHAR, 2013).

Outro fator muito apontado por diversos autores é o próprio desconhecimento dos estudantes a respeito da utilização das ferramentas e o envolvimento necessário para o estudo a distância. Muitos que procuram a

modalidade, acreditam estar adentrando em um curso mais fácil do que o presencial, pensando pelo lado da comodidade ou pelo fácil acesso, já que o material do curso é facilmente acessado de qualquer lugar em que se possua uma conexão de internet, sem a necessidade de se sair do espaço de trabalho ou familiar. Apesar dessa comodidade, a interação entre os participantes, seja estudante-estudante quanto estudante-educador, empobrece, aumentando as tendências de desmotivação e consequentemente de evasão (MORAN, 2003).

Oliveira (2014), reforça a importância de monitorar, de forma mais próxima, os alunos nas disciplinas iniciais do curso e, para um monitoramento eficaz, é necessária a aquisição de uma consistente rede de indicadores que disponibilizem os dados necessários sobre o desempenho e a atuação dos alunos no AVA, de forma frequente e rotineira. Estes dados necessitam chegar aos professores e gestores, para que possam ser tomadas medidas preventivas no combate à evasão (FERNANDES, 2014).

2.3.1 Aspectos que levam à evasão na EaD

Ao que concerne às principais causas da evasão, o Censo EaD.BR (2015) mostra em seus resultados que os maiores motivos que levam a evasão na modalidade a distância são, a falta de tempo, seguida por questões financeiras e a falta de adaptação à modalidade. Se observou também o fator de escolha equivocada do curso, porém em uma escala menor.

Na visão de Lacerda e Espíndola (2013), a modalidade a distância apresenta diversos aspectos que propiciam flexibilidade aos estudantes, mas, em contrapartida, apresentam desafios a serem superados, como, por exemplo, uma dificuldade no acompanhamento do cronograma dos estudos, serem obrigados a ter autonomia discente para gerir seu aprendizado e as dificuldades com o manuseio da tecnologia para um melhor aproveitamento. Estes aspectos têm forte impacto nas decisões dos alunos, levando-os muitas vezes a desistir do curso sem concluí-lo.

Já no levantamento realizado pelas pesquisadoras Baltar e Silva (2017, p. 67), os principais motivos que levam os alunos a evadirem do ensino a distância foram definidos em cinco categorias:

- Fator situacional: Aqui se encaixam a falta de apoio no trabalho ou na família, problemas de saúde ou algum outro problema familiar que dificulta o foco nos estudos do aluno;
- Falta de apoio acadêmico: Se incluem a falta de interação entre os pares (aluno e professor) e a falta de *feedback* ou de apoio do tutor.
- Dificuldades com a tecnologia: Nesta categoria se encaixam a falta de conhecimento técnico no manuseio das ferramentas tecnológicas, o envio de tarefas via fax ou correio ou até a falta de computador e/ou acesso à internet.
- Falta de apoio administrativo: Quando há uma má logística de distribuição de materiais, se têm prazos curtos para a entrega de trabalhos e tarefas, entre outros problemas no recebimento dos módulos.
- Sobrecarga de trabalho: Quando o aluno está com falta de tempo para dedicar ao estudo e curso, não possui organização e autonomia para o estudo ou tem dificuldades em conciliar trabalho, família e estudo.

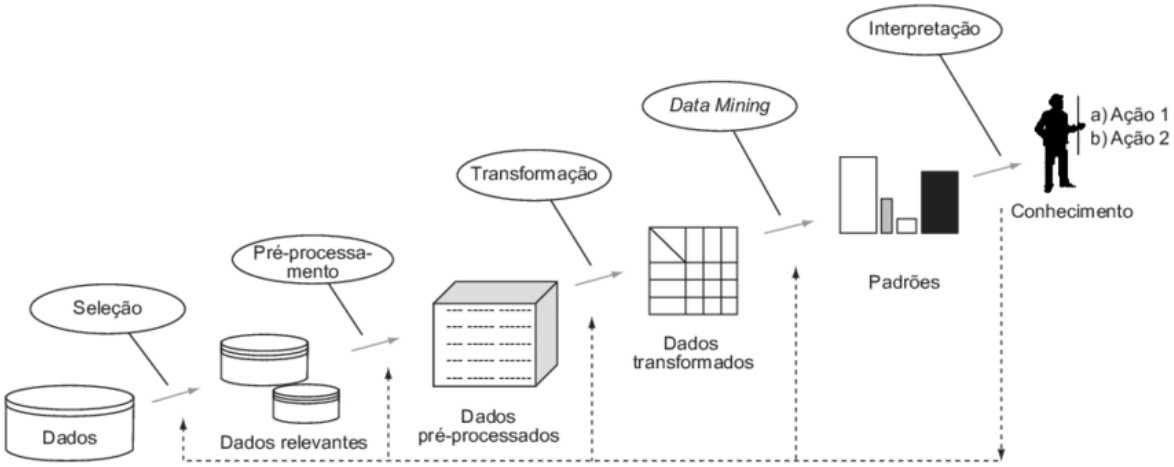
2.4 Mineração de Dados

Desde o surgimento dos sistemas operacionais de forma geral, um dos principais desafios e objetivos das grandes organizações tem sido o de armazenar os dados, conforme sugerem Camilo e Silva (2009). Essa disposição ficou ainda mais em evidência quando observarmos as últimas décadas, pois, com a queda dos preços dos hardwares, ficou possível armazenar uma quantidade cada vez maior de dados e, como observado por Barbosa (2014), em escala mundial, esse volume já é gigantesco e continua crescendo de forma exponencial.

Também segundo Camilo e Silva (2009), com a finalidade de oferecer maiores opções de utilização e com o intuito de tornar a extração destes dados em algo útil, no final da década de 80 foi proposta a Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD).

Já Fayyad *et al.* (1996, p. 40 e 41), formularam uma definição para o KDD como um “processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”. Seguem ainda afirmando que, para que o conhecimento seja obtido através de um processo de KDD, é necessário se passar por 5 etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação, conforme podemos observar na Figura 2.

Figura 2 - Etapas do KDD.



Fonte: Fayyad *et al.* (1996) e Julio Nievola (2006).

Todo o processo se dá de forma iterativa e interativa, pois, todas as etapas descritas se conectam e contêm decisões e tarefas a serem realizadas pelos usuários. Caso alguma delas não obter resultado satisfatório, pode se regredir para a etapa de preparação dos dados.

Segundo Goldschmidt e Passos (2005), tem se mostrado bastante útil o conhecimento que se tem adquirido através das aplicações do KDD, e nas mais diversas áreas, esse conhecimento vem auxiliando no aperfeiçoamento dos processos, como na educação, na parte da saúde, na área de finanças, em telecomunicações, meteorologia, agropecuária, bioinformáticas e muitas outras.

O processo de KDD possui 5 etapas, sendo uma delas a Mineração de Dados (*Data Mining* - DM), porém, diversos autores se referem ao KDD como DM, mas conforme descreve Silva (2004), KDD é todo o processo de descoberta de conhecimento em base de dados e DM é uma das cinco etapas do KDD. Fayyad

et al. (1996) descreve os procedimentos que envolvem o KDD de forma mais detalhada como:

- **Entendimento do domínio**: Conhecimento prévio dos dados, do domínio da aplicação e dos objetivos do KDD, são importantes para obtenção de bons resultados.
- **Seleção dos dados:** Consiste na seleção de um subconjunto de dados que represente todos os dados disponíveis. Esta etapa é importante pois é computacionalmente inviável trabalhar com todos os dados, visto que estes possuem terabytes de informações.
- **Pré-processamento dos dados:** Nesta etapa é realizado uma limpeza nos dados por ser necessário a remoção de informações inconsistentes ou duplicadas. Também são verificados os casos onde há falta de informação.
- **Modelagem dos dados:** Com o objetivo de reduzir a dimensionalidade dos dados são aplicadas técnicas de seleção de atributos mais relevantes, assim os dados terão a maior representatividade com a menor quantidade de atributos possíveis.
- **Mineração de dados:** Considerada uma das mais complexas etapas do KDD, a mineração de dados também é um processo que contém subetapas. Nesta etapa é escolhido o algoritmo que será utilizado, de acordo com os dados e os objetivos do KDD
- **Interpretação dos dados:** A análise dos dados obtidos pode ser feita através de métodos de visualização de informações. Caso os resultados não atendam às expectativas do que se espera após o processo de KDD, devem-se realizar algumas etapas anteriores novamente.
- **Validação dos resultados:** Os resultados obtidos na etapa de Mineração dos Dados devem ser validados com dados não utilizados pelos algoritmos. Os padrões descobertos serão válidos se atenderem um grau de certeza definido pelo responsável do processo de KDD.

Uma das principais tarefas da mineração de dados é a classificação. As técnicas de classificação buscam analisar um conjunto de dados as quais possuem características e classes conhecidas para então criar modelos que serão capazes de classificar novas instâncias a partir de suas características. Essa

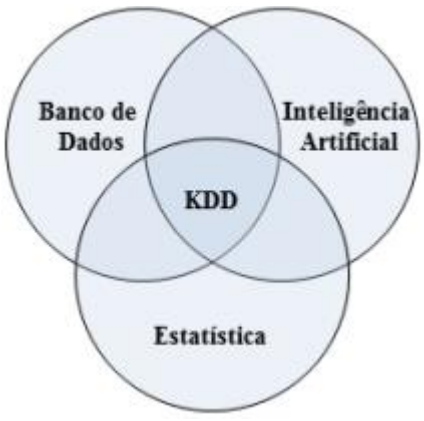
tarefa é definida como preditiva, visto que são realizadas inferências sobre os dados para se predizer a classe de uma nova instância (MERSCHMANN, 2007). Esta é a abordagem principal deste trabalho.

Conforme afirmado por Cardoso e Machado (2008), e, corroborado por Costa *et al.* (2012) e Rodrigues (2016), uma instituição que emprega o processo de KDD para adquirir conhecimento através de seus dados é capaz de:

1. Desenvolver e aplicar parâmetros que irão entender o comportamento dos dados, podendo ser referentes a pessoas envolvidas nos processos da organização;
2. Identificar semelhanças entre os dados, que podem ser, por exemplo, entre pessoas e produtos ou serviços;
3. Realizar a predição de hábitos e comportamentos das pessoas, identificando também, hábitos fora do padrão.

Portanto, aplicar técnicas de KDD com a finalidade de investigar os dados pode revelar informações úteis para toda e qualquer organização, trazendo mais inteligência e conhecimento para auxiliar na tomada de decisão e no alcance de diferenciais competitivos em relação a outras organizações que não façam uso da técnica.

Figura 3 - Áreas do KDD.



Fonte: Lin e Cercone (1997).

Sendo assim, o KDD pode ser definido como uma forma de explorar e analisar bancos de dados, com o objetivo de encontrar padrões, regras e desvios. Ele tem sido objeto de estudo em múltiplas disciplinas, principalmente nas áreas

de Estatística, Inteligência Artificial (IA) e Banco de Dados (DB) conforme observado na Figura 3.

Tan, Steinbach e Kumar (2009) também afirmam que as tarefas da mineração de dados em geral são separadas em duas categorias:

• Tarefas de Previsão: tem como objetivo prever o valor de um atributo baseado nos demais atributos, o atributo a ser previsto é conhecido como o atributo alvo, já os demais atributos são conhecidos como as variáveis explicativas;

• Tarefas Descritivas: têm como objetivo prover padrões de correlações, agrupamentos e tendências, as tarefas descritivas são frequentemente utilizadas de forma exploratória, necessitando técnicas de pós processamento para a validação dos dados.

2.4.1 Mineração de dados educacionais (EDM)

Conforme descrito por Pinheiro *et al.* (2009), diferentes técnicas de KDD vem sendo utilizadas com a finalidade de investigar questões científicas dentro da área da educação, buscando alcançar respostas para questões como, fatores que afetam a aprendizagem, como desenvolver um sistema educacional mais eficaz, e também buscar uma relação entre a abordagem pedagógica e a motivação do estudante em continuar buscando conhecimento. Com este contexto, a definição Mineração de Dados Educacionais (*Educational Data Mining* - EDM) surgiu como praticamente um sinônimo do KDD, mas possuindo um enfoque total na extração de conhecimento em base de dados educacionais.

Como afirmado por Baker (2009), com a expansão e disseminação dos cursos a distância e também de outros cursos que necessitam suporte computacional, diversos pesquisadores da área de Informática na Educação, especialmente, a Inteligência Artificial aplicada à Educação, vem estudando diferentes formas de aplicar e utilizar a mineração de dados com esta finalidade.

Baker (2009) e Gottardo *et al.* (2012) também complementam o conceito do EDM afirmando que, o grande foco do EDM é o desenvolvimento de métodos para

realizar a descoberta de conhecimento em bases de dados educacionais. Baker também apresenta outra definição:

> A mineração de dados educacionais (EDM)[...] tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino (BAKER, 2011, p.1).

Já na definição alcançada por Pinhero (2014), o grande foco do EDM é o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais e assim, compreender de forma correta os estudantes em como eles aprendem o papel do contexto no qual a aprendizagem acontece e entre outros fatores que irão influenciar a aprendizagem.

Com a aplicação do EDM, por exemplo, é possível observar a relação entre uma certa abordagem pedagógica e o aprendizado alcançado pelo estudante. A partir dessa informação, o educador poderá verificar se a metodologia de ensino aplicada está realmente ajudando o aluno, e, no caso de necessidade, considerar métodos alternativos para tornar o ensino mais eficaz (BAKER, 2009).

Conforme Romero e Ventura (2010), o EDM possui diversos papéis de interesse, os principais estão mapeados no Quadro 3.

Quadro 3 - Os papéis mais importantes do EDM.

Atores	Objetivos dentro da mineração
Estudantes	- Personalização do *e-learning*; - Sugerir atividades, recursos e tarefas alternativas que auxiliarão na melhora da aprendizagem; - Facilitar o acesso a material de apoio, como livros, artigos, etc.
Professores	- Obter feedback seja da instituição, seja dos alunos; - Analisar o desempenho de estudantes, buscando encontrar padrões, sejam comuns ou irregulares; - Encontrar estratégias mais eficazes para transmitir o conhecimento.
Demais agentes (Gestores, Desenvolvedores de materiais, Instituições)	- Analisar toda a estrutura do curso, avaliando o conteúdo programático do mesmo e a eficácia das estratégias pedagógicas; - Definir parâmetros de melhora, seja nas

	ferramentas, seja nos tutores; - Aperfeiçoar os processos de tomadas de decisão.

Fonte: Adaptado de Romero e Ventura (2010) e Rodrigues (2016).

Embora o conceito de EDM seja algo relativamente novo, ele já possui vários segmentos de atuação. Baker (2011) apresentou a taxonomia das principais subáreas da EDM e Rodrigues (2016) adaptou em categorias conforme apresentado na Quadro 4.

Quadro 4 - Principais categorias do EDM.

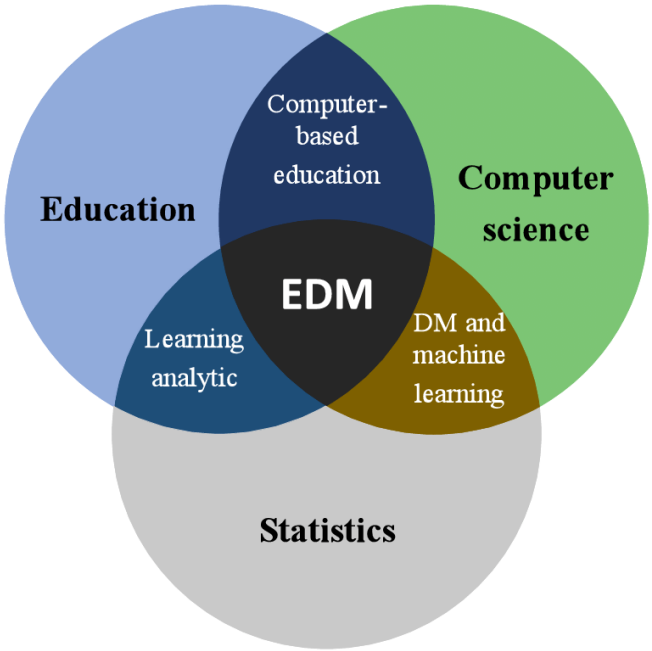
Atividades/Métodos	**Objetivos do Método**	**Aplicação Principal**
Predição - **Classificação** - **Regressão** - **Estimação**	O principal objetivo deste método é desenvolver modelos que percebam características específicas dos dados, conhecidas como variáveis preditivas (*predicted variables)*, realizando a análise e a junção das características encontradas nos dados, chamadas de variáveis preditoras (*predictor variables).*	Encontrar comportamentos dos estudantes, por exemplo, comportamentos fora do padrão em possíveis resultados educacionais. Também encontrar possíveis estudantes propensos à evasão.
Agrupamento	Achar dados que irão se agrupar naturalmente, dividindo o total dos dados nestes conjuntos de categorias.	Foco principal na descoberta de novos padrões e na investigação de semelhanças e diferenças entre os grupos.
Associação **- Mineração de Regras de associação** **- Mineração de Correlações;** **- Mineração de Padrões Sequenciais;** **- Mineração de Causas.**	Encontrar relações entre variáveis.	Achar associações curriculares na sequência do curso e analisar quais as estratégias pedagógicas que aumentem a eficácia e facilitem a aprendizagem.
Definição de Modelos	É um modelo de um fenômeno que é encontrado com a predição, aglomeração, ou conhecimento de engenharia; complementa a predição.	Revelação de relações entre comportamento e/ou características dos estudantes ou variáveis contextuais.
Clarificação de dados para facilitar as decisões humanas	Os dados são refinados para permitir o ser humano identificar e classificar rapidamente diferentes características destes dados.	Identificação humana de padrões na aprendizagem dos alunos, no comportamento ou nos padrões de colaboração; são dados de rotulagem desenvolvidos posteriormente ao modelo de previsão.

Fonte: Adaptado de Baker (2011) e Rodrigues (2016).

Apesar de contar com diversas subáreas, o objetivo final do EDM é basicamente converter os dados brutos oriundos das bases de dados dos sistemas educacionais em informações úteis, que devem ter grande impacto na prática educacional (ROMERO; VENTURA, 2010).

Ainda segundo Romero e Ventura (2013), a EDM combina as 3 principais áreas do conhecimento: Computação, Educação e Estatística. O cruzamento entre estas três áreas ainda gera mais três subáreas: *E-learning*, *Data Mining* e *Machine Learning* e a *Learning Analytics* que também se relacionam diretamente com a EDM, conforme observado na Figura 4.

Figura 4 - Áreas relacionadas ao EDM.



Fonte: Romero e Ventura (2013).

No contexto de EDM o processo não difere muito da aplicação do Aprendizado em Base de Dados em outras áreas de negócio ou na área de genética e medicina, pois, ele segue as mesmas etapas: pré-processamento, mineração de dados e pós-processamento, como é sugerido por Romero *et al.* (2004). Apesar disso, é importante reforçar que o termo KDD possui uma definição

muito mais abrangente que o EDM, o qual se limita a aplicação de técnicas típicas do KDD dentro da área educacional.

2.5 Pré-processamento dos dados

Conforme concluído por Castro e Ferrari (2016), a etapa de pré-processamento consiste em identificar problemas e prepará-los para que os resultados da etapa da mineração dos dados não sejam comprometidos devido à problema nos dados, levando em consideração a filosofia de GIGO. Sendo assim, o objetivo desta etapa é realmente preparar os dados com a finalidade de identificar os tipos de atributos presentes na base, a existência de dados ausentes, ruidosos ou sem consistência e identificar se há atributos irrelevantes.

Segundo Han, Kamber e Pei (2012) a etapa de pré-processamento dos dados é crucial para a obtenção dos melhores resultados e dos melhores tempos de processamento na mineração dos dados.

2.5.1 Limpeza dos dados

O processo de limpeza consiste em uma das técnicas dentro da mineração de dados responsável por solucionar possíveis problemas que podem ser observados nos dados. Han, Kamber e Pei (2012) definem a etapa de limpeza de duas maneiras:

- Valores ausentes: no caso de valores ausentes podem ser aplicadas algumas técnicas para o preenchimento do valor ausente, podendo ser feita a inclusão de um valor constante para todos os atributos faltantes, técnicas estatísticas também podem ser utilizadas como média e moda de acordo com tipo de valor, outra opção é a dedução do valor através da aplicação de técnicas de aprendizado de máquina como árvores de decisão ou modelos de regressão;

- Dados ruidosos: um problema difícil de identificar pois pode representar um erro aleatório ou um erro de variação em uma variável de medição, para solucioná-los podem ser aplicadas técnicas como a de aproximação, técnica essa em que funções de aproximação vão substituir os valores reais ou aplicação de média ou moda como valor para substituição. Algoritmos de agrupamento e regressão também são alternativas para tratar problemas de ruído nos dados.

2.5.2 Integração dos dados

Segundo afirmação de Castro e Ferrari (2016), o processo de integração de dados consiste no cruzamento de dados de fontes diferentes em uma única base de dados. Três problemas que podem ocorrer nesta etapa podem ser evidenciados:

- Redundância: na mineração de dados a redundância significa que um atributo ou objeto é capaz de ser alcançado de um ou mais atributos ou objetos, como exemplo podemos citar um atributo de idade e outro de data de nascimento, nesse caso a base de dados não irá necessitar dos dois atributos. Para a descoberta destas situações é necessário realizar uma análise de correlação;

- Duplicidade: situação nas quais atributos ou objetos estão repetidos na base de dados, o que pode ocasionar anomalias e distorções nos dados. A prevenção pode ser realizada através da normalização da base de dados;

- Conflitos: cenário onde a mesma entidade está diferente nas diferentes fontes de dados, o problema é comum em atributos que possuem unidade de medida, sendo assim as diferentes bases apresentam diferentes unidades de medida e por consequência valores.

2.5.3 Redução dos dados

Conforme Han, Kamber e Pei (2012) o processo de redução tem como objetivo reduzir o tamanho do volume na base de dados mantendo a integridade original deles. Com isso o processo de mineração torna-se mais eficiente e vai produzir um resultado igual ou muito semelhante ao da base original quando aplicadas as técnicas de mineração.

Outra forma de reduzir a dimensionalidade dos dados, conforme Tan, Steinbach e Kumar (2009) é a seleção de um subconjunto de característica, já que descartar atributos irrelevantes ou redundantes muitas vezes necessita de um tratamento sistemático. Isso pode ser alcançado através de abordagens internas onde o algoritmo de mineração de dados vai naturalmente escolher quais atributos utilizar e ignorar. Outro método que pode ser aplicado é o de pesagem de características nas quais as mais importantes recebem um peso maior, isso pode ser feito com base no conhecimento de domínio relativo das características ou de forma automática, através da aplicação de modelos de classificação como máquinas de vetor de suporte.

2.5.4 Transformação dos dados

A etapa de transformação tem o objetivo de alterar e ajustar os dados de forma que o mesmo seja interpretado de forma adequada no processo de mineração, esse processo abrange a padronização dos dados nos quais são resolvidos problemas de formatos, conversão de unidades, remoção de caracteres especiais e capitalização (CASTRO; FERRARI, 2016).

Conforme Han, Kamber e Pei (2012) a normalização dos dados tem como objetivo principal a disposição dos dados dentro de intervalos menores ou comuns como, por exemplo, escolher unidades de medidas nas quais os intervalos entre o maior e menor valor seja a menor possível, assim atribuindo um peso semelhante a todos os atributos, sendo essa técnica especialmente útil para algoritmos de classificação e agrupamento que utilizam redes neurais artificiais. Algumas técnicas de normalização de destaque são Max-min, Z-score e escalonamento decimal.

2.5.5 Discretização dos dados

Segundo Castro e Ferrari (2016), alguns algoritmos não possuem a capacidade de trabalhar com atributos numéricos, sendo necessário nessa situação o uso da discretização. Ela pode ser realizada através de diversos métodos como o encaixotamento, a análise de histograma ou então distribuindo os valores em intervalos que serão representados pela mediana ou média dos valores. Já para Han, Kamber e Pei (2012), um dos métodos de discretização mais eficiente é o de agrupamento, resultando em uma hierarquia de conceito em formas de nós.

2.6 Aprendizado de máquina

Conforme afirmado por Mitchell (1997), a área de pesquisa do aprendizado de máquina é focada no desenvolvimento de programas de computadores que são capazes de aperfeiçoarem-se em determinadas atividades através da experiência.

Já Artero (2008) complementa que, contando com a sua capacidade de extrair informações e conhecimento, as técnicas de aprendizado de máquina estão sendo utilizadas de forma recorrente no processo de mineração de dados procurando extrair essas informações da forma mais automatizada possível.

2.6.1 Classificando o Aprendizado de Máquina

Conforme Artero (2008), é possível se classificar o aprendizado de máquina de diversas maneiras, entretanto as definições mais comuns utilizadas são o agrupamento nas categorias supervisionado e não supervisionado.

2.6.1.1 Aprendizado supervisionado

Segundo Coppin (2013), os algoritmos de aprendizado supervisionado aprendem ao serem executados em dados pré-classificados, possuindo a

capacidade de trabalharem com diferentes pesos aos atributos de acordo com os dados de entrada e saída.

Aprendizado supervisionado é sinônimo de classificação onde a supervisão do aprendizado tem como origem os exemplos presentes nos dados explorados, sendo eles os responsáveis pelo aprendizado (HAN; KAMBER; PEI, 2012).

2.6.1.2 Aprendizado não supervisionado

Os métodos de aprendizado não supervisionado possuem a capacidade de aprender sem qualquer auxílio humano. São de grande utilidade para as técnicas de agrupamento e classificação de dados que não são conhecidos previamente (COPPIN, 2013).

Já Han, Kamber e Pei (2012) afirmam que o aprendizado não supervisionado pode ser considerado uma técnica de agrupamento. Os dados apresentados não estão classificados e fica a cargo da máquina identificar diferentes padrões e tendências presentes nos dados. Porém, como os dados não são classificados o modelo de aprendizado gerado não possui a capacidade de indicar o significado destes resultados.

2.7 Métodos de Classificação

A classificação pode ser descrita como uma técnica de predição baseada em registros que possuem valores de saída. Com uma análise histórica desse conjunto de dados é possível se criar modelos capazes de prever o valor de saída de um determinado atributo (CASTRO; FERRARI, 2016).

Neste cenário, Han, Kamber e Pei (2012) afirmam que diferentes métodos de classificação são propostos pelos pesquisadores dentro da área de aprendizados de máquina, estatística e reconhecimento de padrões. Também descrevem que atualmente o desenvolvimento na área de mineração de dados faz com que seja possível se criar modelos de classificação sobre grandes conjuntos

de dados, onde o processo de classificação é dividido em duas etapas, sendo em um primeiro momento aplicada a etapa de aprendizado ou treinamento que consiste na construção do modelo e, em um segundo momento é aplicada a classificação ou teste, que consiste na avaliação do modelo desenvolvido.

2.7.1 Árvores de Decisão

A definição de Árvores de Decisão consiste em se utilizar de modelos estatísticos os quais são submetidos a um treinamento supervisionado com a finalidade de fazer a classificação e a previsão dos dados, ou seja, sua construção se dá utilizando um conjunto de treinamento que é formato por entradas e saídas, sendo estas últimas as classes.

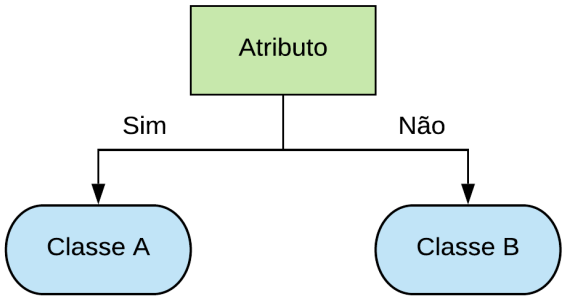
A estratégia adotada por este modelo é a do dividir para conquistar, onde uma questão complexa é dividida em subproblemas mais simples e então, de forma recursiva, a técnica é aplicada a cada um desses subproblemas (GAMA, 2004).

Elas estão entre os algoritmos de inferência mais populares e a técnica vem sendo útil em diversas áreas, além do EDM, como para diagnósticos médicos e análises de risco de crédito, sendo possível se extrair delas regras do tipo “se-então” de fácil compreensão (MITCHELL, 1997).

Para medir a capacidade de discriminação de uma árvore, é realizada a divisão do espaço definido por seus atributos em subespaços, e então, para cada subespaço, uma classe é associada.

Já segundo Castro e Ferrari (2016), o processo de classificação em uma árvore de decisão acontece de maneira recursiva, conforme Figura 5, de modo que o nó inicial representa um conjunto de dados e em seguida deve ser feita uma avaliação para conferir se os objetos são da mesma classe. Caso sim, o nó é considerado um nó da folha, caso não, outro atributo é necessário para fazer a divisão dos dados. Este processo é executado recursivamente podendo ser descontinuado caso faltarem atributos para realizar testes de divisão ou caso todos os registros forem da mesma classe.

Figura 5 - Representação de uma árvore de decisão.



Fonte: Castro e Ferrari (2016, p. 166).

O critério que é utilizado para definir as partições é o da utilidade daquele atributo para a classificação. Dentro deste critério, são determinados os ganhos de informação a cada atributo e, o atributo que é escolhido como o atributo teste para o corrente nó, deverá ser o que possui o maior ganho de informação. Após esta aplicação, irá se iniciar um novo processo de partição.

Quando utilizada para classificação, os critérios para a partição da árvore mais difundidos são os baseados na entropia e no índice de Gini, onde, no primeiro, o cálculo do ganho de informação é baseado em uma medida utilizada na teoria da informação e, o segundo, foi desenvolvido por Conrado Gini em 1912 com a finalidade de medir a heterogeneidade dos dados, ou seja, pode ser utilizado para medir a impureza de um nó (ONODA, 2001).

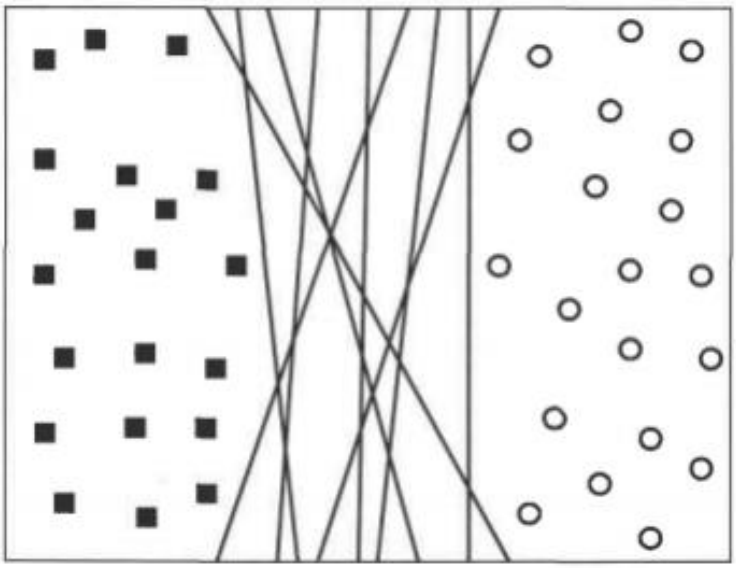
2.7.2 *Random Forests*

As *Random Forests* ou em português, florestas aleatórias são um conjunto de árvores de decisão as quais juntas formam uma floresta. Estas árvores são geradas baseadas em um atributo aleatório o qual é responsável pela divisão da árvore em nós. A precisão dessa floresta é definida de acordo com a força de cada classificador da árvore assim como o nível de dependência entre eles e sendo assim o melhor jeito de alcançar essa precisão é mantendo a força desses classificadores e não aumentar a correlação entre eles (HAN; KAMBER; PIN, 2012).

2.7.3 *Support Vector Machines*

As *Support Vector Machines* (SVM) ou Máquinas de vetores de suporte, possuem como fundamento principal o aprendizado a partir da estatística, já que esse tipo de algoritmo apresenta ótimo desempenho sendo aplicado em dados de alta dimensionalidade. O funcionamento do mesmo se dá sobre um hiperplano, onde neste é definido um limite linear para realizar a classificação, conforme ilustrado na Figura 6. A função do algoritmo é detectar o hiperplano de margem máxima, sendo aquele que possui a maior margem de separação entre as classes, com a finalidade de apresentar o menor número de erros de generalização em relação a margens menores (TAN; STEINBACH; KUMAR, 2009).

Figura 6 - Classes separadas de forma linear em uma SVM.



Fonte: Tan, Steinbach e Kumar (2009, p. 257).

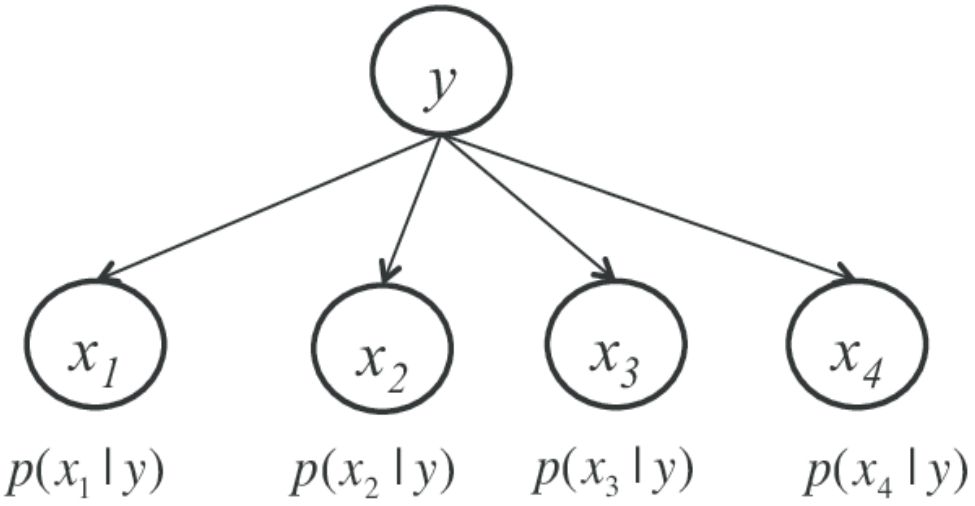
Han, Kamber e Pei (2012) também afirmam que o algoritmo SVM pode ser utilizado para a previsão de dados numéricos e classificação e apresentar um alto índice de acurácia, porém, a etapa de treinamento é considerada lenta. O SVM é aplicado em diversas áreas, com destaque para o reconhecimento de voz e de objetos.

2.7.4 Classificadores Bayesianos

Como definido por Tan. Steinbach e Kumar (2009), os classificadores bayesianos têm a função de classificar se um determinado registro faz parte de uma determinada classe. Essa tarefa é feita se aplicando o teorema de Bayes, um princípio estatístico que faz uso de conhecimentos previamente conhecidos das classes combinando com conjuntos de novos dados.

Classificadores possuem um desempenho e uma acurácia muito alta quando aplicados em grandes bases de dados. O Naive Bayes é exemplo de um algoritmo bayesiano já que o mesmo assume que o valor de um atributo de determinada classe tem efeito independente em relação aos valores dos demais atributos, esse antecedente é conhecido como independência condicional da classe que tem o propósito de simplificar cálculos (CASTRO; FERRARI, 2016).

Figura 7 - Exemplo de modelo Naive Bayes para a integração de bases de dados.



Fonte: Thomas (2015).

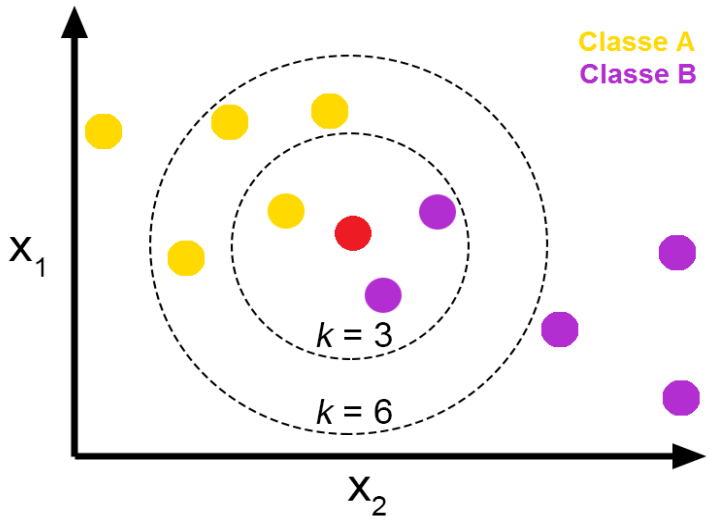
Conforme Russell e Norvig (1995), apesar de se encontrarem entre os modelos mais simples das redes Bayesianas, estes classificadores são altamente escaláveis, exigindo um número de parâmetros linear no número de variáveis (características/preditores) em um trabalho de aprendizado. Um treino de probabilidade máxima pode ser feito utilizando um cálculo de forma-fechada, que leva um tempo linear, ao invés de se utilizar um método iterativo que muitas vezes

possui um alto custo de tempo, como utilizado em diversos outros tipos de classificadores.

2.7.5 Classificadores KNN

A definição KNN vem do inglês *K-nearest neighbors*, que significa K vizinhos mais próximos e conforme definido por José (2018), é um dos muitos algoritmos disponíveis para aprendizagem supervisionada dentro do campo de mineração de dados e aprendizado de máquina, como mostra a Figura 8.

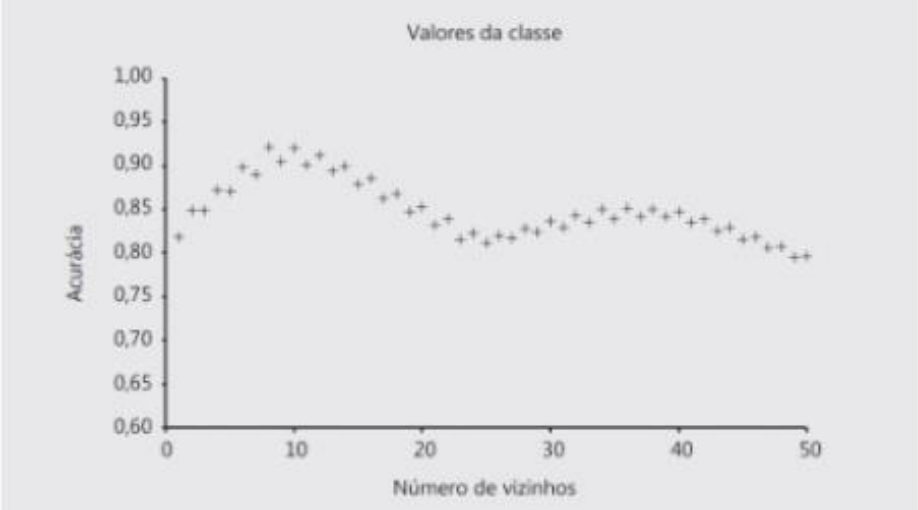
Figura 8 - Exemplo de estrutura do KNN.



Fonte: José (2018).

A classificação k vizinhos mais próximos tem como modo de aprendizagem a analogia, realizando comparações entre um objeto de teste e objetos semelhantes. Estes objetos estão distribuídos em um espaço de N dimensões no qual todos os objetos de treinamento são inseridos. A partir do momento que um objeto desconhecido é inserido no espaço, o classificador procuro dentro deste espaço de padrões outros objetos de treinamento que possuam a maior semelhança com o objeto desconhecido de acordo com sua proximidade (HAN; KAMBER; PEI, 2012).

Figura 9 - Acurácia do classificador KNN.



Fonte: Castro e Ferrari (2016, p. 169).

Castro e Ferrari (2016), afirmam que o método de classificação KNN deve ser considerado um método baseado em instâncias, ou seja, ele vai determinar a classe de um objeto desconhecido através da classe de outras instâncias. Conforme a Figura 9, um número maior de vizinhos reduz os ruídos na classificação, porém tornam as fronteiras entre as classes maiores.

2.7.6 Avaliação do desempenho dos classificadores

Conforme Han, Kamber e Pei (2012), um jeito de avaliar o desempenho de um algoritmo de classificação é avaliando sua capacidade preditiva. Isso precisa ser realizado expondo o modelo de classificação a dados não visualizados durante seu treinamento, se não, ele não é capaz de identificar se há ruídos ou irá encontrar dificuldades para generalizar os dados.

Para classificar problemas binários, que é a predicação entre duas classes, é utilizada uma de matriz de confusão, indicando os dados com colunas representando as classes de previsão e linhas as classes atuais dos dados, conforme ilustrado na Figura 10 (TAN; STEINBACH; KUMAR, 2009).

Figura 10 - Ilustrando Matriz de confusão.

	Classe Prevista

		0	1
Classe Atual	**0**	**Verdadeiro Negativo**	**Falso Positivo**
	1	**Falso Negativo**	**Verdadeiro Positivo**

Fonte: Adaptado de Tan, Steinbach e Kumar (2009, p.351).

Conforme os autores os termos utilizados na composição de uma matriz de confusão são subsequentes:

- Verdadeiro Positivo (TP): número de exemplos positivos classificados corretamente;
- Falso Negativo (FN): número de exemplos negativos classificados incorretamente;
- Falso Positivo (FP): número de exemplos positivos classificados incorretamente;
- Verdadeiro Negativo (TN): número de exemplos negativos classificados corretamente.

Segundo Han, Kamber e Pei (2012) há outras métricas que podem ser utilizadas para a avaliação do desempenho de classificadores conforme listado a seguir:

- **Acuracidade:** taxa dos objetos de testes classificados corretamente pelo classificador;
- **Taxa de erro:** percentual de objetos classificados incorretamente pelo classificador;
- **Revocação:** taxa de objetos positivos que realmente são verdadeiros e foram classificados corretamente;
- **Especifidade:** taxa de objetos falsos verdadeiros classificados de forma correta;
- **Precisão:** taxa de objetos positivos classificados corretamente, podendo ser um falso positivo ou um verdadeiro positivo;

- **Medida F:** média harmônica entre as medidas de precisão e revocação.

Na seção seguinte são apresentadas técnicas e algoritmos de estimação, assim como métodos de avaliação de desempenho dos algoritmos.

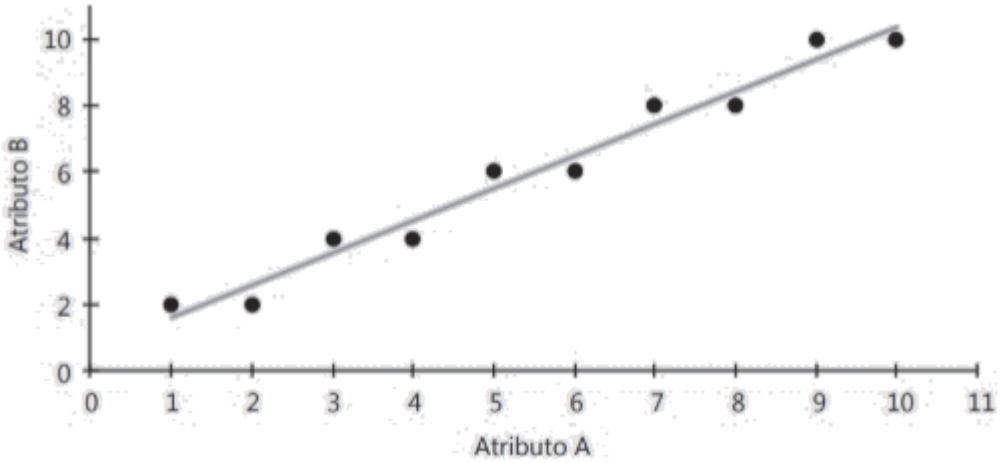
2.8 Técnicas de estimação

Técnicas de estimação possuem a finalidade de prever um valor contínuo de uma variável alvo. Há muito em comum entre técnicas de estimação e as técnicas de classificação listadas na Seção 2.5.1, de modo que praticamente todos os algoritmos de estimação podem ser utilizados como classificadores, porém não o contrário. Ainda assim se utilizando de algumas adaptações, técnicas como árvores de decisão e classificadores bayesianos podem atender a necessidade de estimação. A principal diferença entre as técnicas de classificação e estimação encontra-se no método de avaliação dos algoritmos (CASTRO; FERRARI, 2016).

2.8.1 Regressão Linear

Conforme Castro e Ferrari (2016), a regressão linear tem o objetivo de modelar a associação entre uma ou mais variáveis de saída e entrada. Este processo pode ser definido em duas categorias, sendo as paramétricas, no qual o relacionamento entre as variáveis é conhecido, e as não paramétricas, que é quando não se possui um conhecimento prévio entre as variáveis. As técnicas de regressão linear buscam relacionar duas variáveis através de uma equação em uma linha reta. A relação entre as variáveis pode ser observada na Figura 11.

Figura 11 - Regressão linear em um conjunto de dados bidimensional.



Fonte: Castro e Ferrari (2016, p.207).

Os autores ainda seguem afirmando que, a regressão polinomial também pode ser considerada uma técnica de regressão linear e, a grande diferença está na relação entre a variável dependente e as variáveis independentes, de modo que sua relação acaba sendo não linear e sim um polinômio de grau N.

2.8.2 Regressão Logística

Segundo Witten e Frank (2005), regressão logística é uma técnica utilizada para se estimar uma variável de natureza binária, estimando o valor em 0 ou 1, onde as variáveis independentes podem ser de natureza categórica ou não. Semelhante a regressão linear, na regressão logística é necessário aplicar pesos aos dados de treinamento do algoritmo, porém no final ao invés de buscar a melhor reta, a regressão logística busca a melhor curva. A regressão logística realiza o cálculo da razão de probabilidade da variável alvo, a qual após isso é convertida em uma variável de base logarítmica, permitindo assim a classificação com base na aproximação de um dos valores.

2.8.3 Redes Neurais Artificiais

Redes neurais artificiais (RNA) consistem em modelos computacionais que foram projetados embasados na estrutura do cérebro humano. A rede neural é composta de unidades simples de processamento, assim, tornando-a em um

processador distribuído e paralelizado. Cada uma dessas unidades de processamento é denominada neurônio, o qual possui uma tendência natural de armazenar qualquer conhecimento experimental e então disponibilizar os mesmos para uso (HAYKIN, 1999).

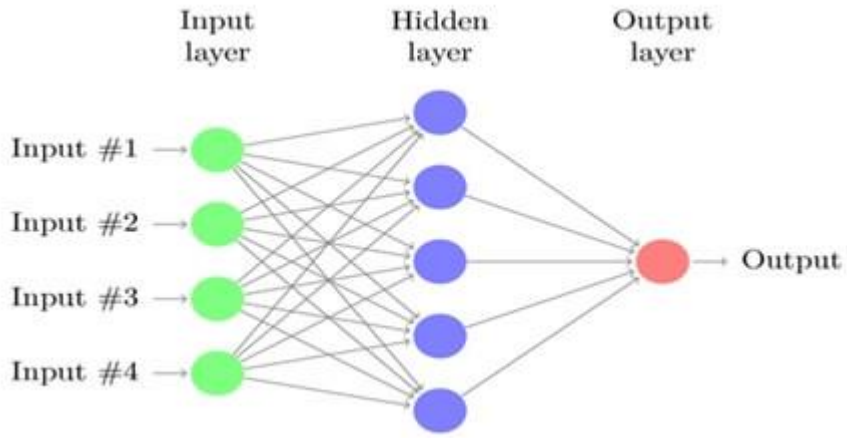
As RNAs contam com uma capacidade de aprendizado através de um grupo reduzido de dados e então generalizar essa informação aprendida e até identificar, automaticamente, padrões variados em um conjunto de dados complexo (BRAGA *et al.*, 2000).

Elas se assemelham ao cérebro humano em dois aspectos principais:

- Todo o conhecimento da rede é construído através de um processo de aprendizagem;
- Para armazenar o conhecimento gerado na rede, são definidos pesos nas conexões que ocorrem entre os neurônios, conhecidos como pesos sinápticos.

A estrutura de uma rede neural pode ser observada na Figura 12.

Figura 12 - Diagrama da rede neural artificial (entrada, camada escondida e saída).



Fonte: Lattaro (2017).

Parecido com o cérebro humano, a rede inicialmente passa por um processo de aprendizagem, onde uma amostra do conjunto é apresentada para a rede e, a partir deles, ela extrai as características necessárias para após representar a informação fornecida. Através desse aprendizado, a rede será capaz

de reconhecer diversos padrões nos objetos que inicialmente não estavam na amostra.

Também é importante reforçar que as Redes Neurais na sua maioria possuem o propósito da estimação, porém, dentro da classificação elas também se aplicam inclusive apresentando uma eficácia muito grande.

2.8.3.1 Redes Neurais Artificiais do tipo Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP) é uma rede neural artificial (RNA) de múltiplas camadas que tem como principal objetivo solucionar problemas linearmente separáveis. Possui como características pelo menos uma camada intermediária e alto grau de conectividade. O treinamento de uma rede MLP geralmente é realizado através de um algoritmo de retro propagação. O funcionamento do algoritmo se dá em duas etapas, no primeiro passo é realizada a propagação do sinal funcional com os pesos fixados e no segundo passo é executada a retro propagação do erro onde os pesos serão ajustados de acordo com o erro (CASTRO; FERRARI, 2016).

2.8.4 Avaliação das técnicas de estimação

Castro e Ferrari (2016) definem o resultado de um estimador como um valor numérico que deve ser aproximado do valor alvo desejado, ou seja, a diferença entre o valor alvo e o estimado proporciona uma medida de erro de estimação. Diversas medidas permitem estimar o tamanho deste erro, com destaque para os seguintes métodos: Soma dos erros quadráticos; Erro quadrático médio; Erro absoluto médio; Erro quadrático relativo; Erro absoluto relativo e Coeficiente de correlação. Já o método correto a se utilizar na avaliação do modelo é definido de acordo com os dados aplicados (CASTRO; FERRARI, 2016).

3 TRABALHOS RELACIONADOS

Em uma pesquisa rápida é possível encontrar diversos estudos dentro da literatura que buscam prever o desempenho acadêmico dos estudantes e até a sua propensão à evasão, utilizando técnicas de mineração de dados. Sendo o problema da evasão no ensino, principalmente no superior, uma preocupação frequente tanto para os cursos presenciais quanto os cursos EaD, na última década, diversos trabalhos de pesquisa em todo o planeta vem abordando de forma mais profunda esta problemática e buscando, com o auxílio da mineração de dados e do aprendizado de máquina, apresentar cenários, motivações e comportamentos dos estudantes que possuem um mal desempenho acadêmico e que estejam propensos à evasão, oferecendo para os gestores educacionais informações úteis que auxiliem na tomada de decisão, na construção de estratégias e no aperfeiçoamento da abordagem pedagógica, visando uma melhora no desempenho acadêmico e no aumento do interesse dos participantes.

Uma pesquisa desenvolvida por Yükselürk *et al.* (2014), mapeou e investigou a participação de 189 estudantes em um programa de certificação em TI (Tecnologia da Informação) na Turquia entre 2007 e 2009. Os autores analisaram quatro técnicas de mineração de dados/aprendizado de máquina, sendo eles: k-nearest neighbors (k-NN), Árvore de Decisão (AD), Naive Bayes (NB) e Redes Neurais Artificiais (RNa). Para alimentar esses algoritmos classificadores, ele separou nove atributos relacionados aos estudantes: Gênero, Idade, Nível de educação, Experiências *on-line* anteriores, Profissão/Ocupação,

Auto-eficácia, Disponibilidade, Conhecimento Prévio e Lócus de controle. Por final foi apresentado que o algoritmo k-NN alcançou uma taxa de acerto significativamente maior do que as taxas mostradas pelos outros métodos: k-NN (87%), AD (79,7%), RN (76,8%) e NB (73,9%).

No estudo realizado por Márquez-Vera et al (2013), foram analisados um total de 670 estudantes oriundos da Universidade Autónoma de Zacatecas, no México, entre 2009 e 2010. Os autores realizaram experimentos utilizando 13 algoritmos, sendo eles: Jrip, NNge (Redes Neurais), OneR, Prism, Ridor, ADTree, Árvore de Decisão (J48), RandomTree, REPTree, SimpleCart, ICRM v1, ICRM v2 e ICRM v3). Seus resultados indicaram que o algoritmo ICRM (do Inglês *Classification Rule Interpretable Mining*) v3 superou em desempenho os algoritmos restantes. O ICRM alcançou uma taxa de acerto de 98,7%.

Em dissertação de Pós-graduação realizada por Weslley Rodrigues (2016), da Universidade Católica de Brasília, que buscou fazer a predição da evasão dos estudantes nos cursos EaD, foram mapeados um total de 150 estudantes matriculados/ativos, sendo 117 deles estudantes regulares do curso EaD e 33 sendo estudantes do presencial, mas que optaram por alguma disciplina a distância. No estudo foram experimentados 5 algoritmos, sendo eles: Rede Bayesiana, SMO, Naive Bayes, Árvore de Decisão (J48) e Multilayer Perceptron. Em 6 experimentos realizados, o algoritmo Naive Bayes obteve a maior acurácia em 4 deles.

Em trabalho desenvolvido por Humberto Rabelo, Aquiles Medeiros Filgueira Burlamaqui, Ricardo Alexsandro de Medeiros Valentim, Danieli Silva de Souza Rabelo e Soraya Roberta dos Santos Medeiros na UFRN e apresentado no CBIE 2017, foi feita a predição de desempenho de alunos da EaD dentro dos AVAs utilizando de técnicas de mineração de dados. Na pesquisa, foram captados da base de dados do Moodle da instituição um número de 514 usuários com perfil de aluno em 13 Turmas de Cursos de Graduação da UFRN. O experimento contou com dois algoritmos de classificação baseados em árvore de decisão, sendo o ID3

e o J48. Os resultados obtidos na taxa de acerto dos dois algoritmos foram de 93,97% e 96,50%, respectivamente.

No estudo realizado por Tayná Costa Gonçalvez, Josenildo Costa da Silva e Omar Andres Carmona Cortes, em 2018, no Instituto Federal do Maranhão, foi realizado um estudo de caso da evasão no ensino superior dentro do IFMA. Para isso, foram utilizados tantos métodos manuais, quanto métodos automatizados no pré-processamento dos dados, os métodos utilizados foram o *Information Gain* (InfoGain) e o *Correlation Based Feature Selection* (CSF), dando origem a 3 datasets. Após a análise, foram aplicados sobre os dados 3 algoritmos de classificação, o Naive Bayes, J48 e o SVM (Support Vector Machine). As 3 tabelas básicas utilizadas foram: Pautas, Histórico e Matrículas. Como resultado final, a acurácia do Naive Bayes foi de 94% na seleção manual e no método CSF e 93% no InfoGain, a do algoritmo SVM foi de 96% na seleção manual e 97% em ambos métodos automatizados e, por último, o J48 apresentou uma acurácia de 97% na seleção manual e no InfoGain e um total de 98% de acurácia no método CSF.

4 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos consistem na análise das ferramentas as quais são necessárias para a construção e produção da pesquisa e dos experimentos, sendo assim, eles são cruciais para o pleno desenvolvimento deste estudo. Neste capítulo serão apresentadas as técnicas e ferramentas que servirão de base para a realização das cinco etapas do KDD que esta pesquisa se submeteu.

4.1 Tipo de pesquisa

A presente pesquisa possui o objetivo de analisar dados oriundos da base de dados de um AVA Moodle e do Sistema de Gerência Interno da IES estudada, minerar estes dados utilizando algoritmos de aprendizado de máquina e apresentar cenários e comportamentos que demonstrem fatores que levam a evasão nos cursos da modalidade a distância na instituição. Para o pleno cumprimento desse objetivo, foi feita uma pesquisa de caráter exploratório dos temas relacionados.

Segundo Selltiz *et al.* (1965), se encaixam dentro do panorama dos estudos exploratórios todos os estudos que buscam desvendar ideias e intuições, visando sempre uma maior familiaridade com o caso da pesquisa. A formulação de hipóteses nestes estudos não é sempre necessária. Já conforme Marconi e Lakatos (2001), a pesquisa de caráter exploratório se discorre através de três etapas: o desenvolvimento de uma hipótese, a construção de um referencial

teórico de consistência para aumentar o conhecimento sobre o assunto a ser estudado e, por último, a identificação do fato ou fenômeno.

Com o intuito de agregar ao estudo, foi realizado uma síntese de informações relevantes na literatura acerca deste tema, o que levou a necessidade de empregar o método de pesquisa bibliográfica. Conforme Gil (2007), os principais exemplos que caracterizam o tipo de pesquisa bibliográfico são as investigações sobre ideologias ou as pesquisas que têm a finalidade de analisar de diferentes pontos de vista um determinado problema. O autor ainda complementa afirmando que, este método de pesquisa faz o uso de referências em bibliografias a respeito do tema da pesquisa, para uma melhor fundamentação do objeto científico (GIL, 2007).

Também foi necessária discorrer de uma pesquisa documental a respeito das ferramentas e da realização do estudo de caso, que é o cenário da IES estudada. Segundo May (2004), a pesquisa documental não se encaixa em uma categoria distinta e bem reconhecida como o levantamento (*survey*) e a observação. Raramente pode se dizer que constitui um método, pois ao dizer que serão utilizados documentos, não especifica como eles serão aplicados no trabalho. Para Tsikriktsis, Voss e Frohlich (2002), estudo de caso é uma história de um fenômeno passado ou atual, que é desenvolvida a partir de múltiplas fontes de provas, que pode incluir dados da observação direta e entrevistas sistemáticas, bem como pesquisas em arquivos públicos e privados.

Já na abordagem da pesquisa foi utilizado o modo quantitativo. Conforme Mattar (2001), a pesquisa quantitativa visa validar hipóteses através da utilização de dados estruturados, estatísticos, com análise de um grande número de casos representativos, os quais irão recomendar um curso final da ação. Ela quantifica os dados com o intuito de generalizar os resultados da amostra para os interessados.

4.2 Ferramentas Utilizadas

Nesta seção será feita uma apresentação detalhada das principais ferramentas e tecnologias as quais foram utilizadas no discorrer deste estudo.

4.2.1 PostgreSQL

O PostgreSQL[8] é um sistema de gerência de banco de dados que trabalho com o modelo relacional, de código aberto e que amplia a linguagem SQL, associando a mesma com diversos outros recursos para facilitar o armazenamento e o dimensionamento de dados. É dito que ele trabalha com o modelo relacional pois os dados são organizados em forma de tabelas, ou seja, linhas e colunas relacionadas através de chaves estrangeiras (POSTGRESQL, 2019).

4.2.2 LibreOffice

O Libreoffice[9] é uma suíte de aplicativos livre para utilização em escritório a qual é disponibilizada para os sistemas operacionais Windows, Unix, Solaris, Linux e MacOS. Ela utiliza o formato OpenDocument, sendo também compatível com os formatos do Microsoft Office, além de diversos outros formatos legados (LIBREOFFICE, 2020).

4.2.3 Microsoft Excel

O Microsoft Excel[10] é um software produzido e escrito pela Microsoft e é baseado no manuseio de planilhas eletrônicas. O sistema é utilizado geralmente para a realização de cálculos, gráficos, relatórios, construção de formulários e entre outras atividades que são rotinas nas empresas, seja na área econômica, administrativa ou até na rotina doméstica.

4.2.4 WEKA

[8] PostgreSQL. https://www.postgresql.org/. Acessado em 2 nov. 2019.
[9] Libreoffice https://pt-br.libreoffice.org/sobre-nos/historia-do-libreoffice/. Acessado em 27 jun. 2020
[10] Excel – Microsoft https://products.office.com/pt-br/excel. Acessado em 2 nov. 2019.

Desenvolvido pela Universidade de Waikato na Nova Zelândia, o Weka[11] é um software gratuito e de código aberto, com licença no GNU (GNU/GPL)[12]. Ela implementa a linguagem Java e contém uma GUI (Interface Gráfica de Usuário, do Inglês *Graphical User Interface*) a qual permite interagir com diferentes arquivos de dados e produzir os resultados visualmente (WEKA, 2015). Esta ferramenta conta com diversos métodos de classificação, associação e clusterização, além de ser customizável e expansível, permitindo que novos métodos sejam inseridos ou removidos de forma simples. Suporte somente manipulação de arquivos do tipo ARFF.

4.2.5 Google Colaboratory

O Google Colaboratory ou simplesmente Google Colab é um produto desenvolvido e mantido pelo Google Research. Ele possibilita e facilita que qualquer um execute qualquer código arbitrário em linguagem Python através do navegador e é especialmente recomendo para aprendizado de máquina, análise de dados e educação em geral. Em termos mais técnicos, o Colab é basicamente um serviço Jupyter Notebook que não necessita de nenhuma configuração para usar, enquanto fornece acesso livre a diversos recursos do computador, inclusive a GPU[13].

4.2.6 Python

Python[14] é uma linguagem de programação de alto nível, orientada a objetos com semânticas dinâmicas. Possui uma sintaxe de fácil aprendizagem, com ênfase na legibilidade e na redução do custo de manutenção dos programas. A linguagem suporta diversos módulos, pacotes e bibliotecas que estendem os seus recursos e encorajam a modularidade e ao reuso de trechos de código. O interpretador do Python e toda sua biblioteca estão todos disponíveis de forma

[11] Weka https://www.cs.waikato.ac.nz/ml/weka/. Acessado em 2 nov. 2019.
[12] GNU/GPL - https://www.gnu.org/licenses/gpl-3.0.pt-br.html. Acessado em 27 jun. 2020.
[13] GPU - https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/. Acessado em 27 jun. 2020.
[14] Python.org. https://www.python.org/. Acessado em 2 nov. 2019.

gratuita, tanto o seu código fonte quanto na forma binária, para todas as grandes plataformas (PYTHON, 2019).

Recentemente o Python vem se tornando uma das referências em linguagens de programação quando se trata da área de *machine learning*, sendo tanto em ambientes acadêmicos quanto na indústria. Um dos principais motivos desse crescimento é a atenção e as melhorias que bibliotecas como o Scikit-learn e Pandas vem recebendo, tornando sua utilização muito popular para diversas atividades ligadas a análise e ciência dos dados (MCKINNEY, 2012).

4.2.7 SCIKIT-Learn

O Scikit-learn[15] é uma biblioteca em Python utilizado para aprendizado de máquina e mineração de dados de código aberto. A biblioteca inclui diversos algoritmos de classificação, regressão e agrupamento incluindo máquinas de vetores de suporte, florestas aleatórias, gradient boosting, k-means e DBSCAN. Foi projetada para interagir com as bibliotecas Python numéricas e científicas NumPy e SciPy (COURNAPEAU, 2011).

4.2.8 Pandas

Segundo Mckinney (2012), a biblioteca Pandas é utilizada para facilitar o trabalho e a análise de estruturas de dados, fornecendo para isso técnicas para agrupamento de dados, transformação e limpeza. A biblioteca também oferece possibilidades para a manipulação de dados dentro de planilhas e banco de dados relacionais.

Estruturas de dados dentro do Pandas também podem ser separadas em duas categorias, *series* e *dataframes*, onde as *series* são semelhantes a vetores, contando com apenas uma dimensão e os *dataframes* possuem uma estrutura semelhante a uma planilha eletrônica, com um conjunto de colunas possuindo índice e podendo receber diversos tipos de dados. Sendo assim, há a

[15] Scikit-learn. https://scikit-learn.org/. Acessado em 2 nov. 2019.

possibilidade de leitura de arquivos como CSV e XLSX e manter suas estruturas replicadas de forma idêntica para o *dataframe* (MCKINNEY, 2012).

4.2.9 NumPy

Conforme McKinney (2012), a biblioteca NumPy é uma base para computação cientifica no Python. Oferecendo funções para realizar operações matemáticas entre diferentes vetores, números aleatórios, álgebra linear e também oferecendo a possibilidade de integração nativa com diversas outras linguagens de programação, como o C e o C++.

4.2.10 Matplotlib

A biblioteca Matplotlib é uma biblioteca muito utilizada para a visualização dos dados em gráficos. Ela se destaca por ser facilmente integrada com o restante do ambiente Python, onde outras bibliotecas famosas a utilizam para visualização dos dados, como por exemplo a Pandas (MCKINNEY, 2012).

4.3 Coleta e análise dos dados

O conjunto de dados extraído para ser utilizado nesta monografia são dados e informações registrados durante os anos de 2018 e 2019, sendo coletados dados que correspondem ao período de duração dos trimestres listados na Tabela 1.

Tabela 1 - Períodos analisados.

Trimestre	Período
2018A-EAD1	Início: 05-03-2018 Encerramento: 12-05-2018
2018A-EAD2	Início: 14-05-2018 Encerramento: 21-07-2018
2018B-EAD1	Início: 30-07-2018 Encerramento: 06-10-2018

2018B-EAD2	Início: 08-10-2018 Encerramento: 15-12-2018
2019A-EAD1	Início: 18-02-2019 Encerramento: 05-05-2019
2019A-EAD2	Início: 06-05-2019 Encerramento: 13-07-2019
2019B-EAD1	Início: 22-07-2019 Encerramento: 06-10-2019
2019B-EAD2	Início: 07-10-2019 Encerramento: 14-12-2019

Fonte: Do Autor (2020).

Na análise dos dados, foram analisados um total de 2487 alunos diferentes dentro do ambiente virtual de aprendizagem, dados esses que foram adquiridos de duas fontes distintas, do sistema de gerenciamento interno da instituição e da plataforma Moodle, a qual as tabelas utilizadas no estudo se encontram listadas na Tabela 2.

Tabela 2 - Descrição de tabelas do Moodle.

Tabela	Descrição
mdl_logstore_standard_log	Tabela responsável pelo registro de todos os eventos que acontecem dentro do AVA
mdl_user	Tabela responsável pelo registro de todos usuários cadastrados dentro do AVA
mdl_course	Tabela responsável pelo registro de todos os cursos cadastrados dentro do AVA

Fonte: Do Autor (2020).

Posteriormente foram aplicadas as técnicas de pré-processamento de dados para realizar o cruzamento entre eles. Um conjunto de dados organizado torna a recuperação e o pré-processamento dos atributos mais eficiente para

então ser iniciado o processo de extração do conhecimento (CASTRO; FERRARI, 2016).

4.4 Pré-processamento

Segundo Castro e Ferrari (2016), pré-processamento é a etapa em que aplicadas técnicas com a finalidade de preparar os dados para o processo de aprendizagem, técnicas essa que incluem a limpeza, integração, transformação e a seleção dos atributos mais relevantes para o estudo.

Sendo assim, a primeira técnica a ser aplicada é a de integrar e cruzar os dois conjuntos de dados. Primeiramente foi desenvolvida uma função na base de dados do sistema de gerenciamento interno para a aquisição dos dados necessários para o estudo. Após, os dados da plataforma Moodle foram inseridos em uma tabela de um banco de dados PostgreSQL e então foram extraídos somente os dados necessários conforme os dados presentes no CSV anteriormente extraído do sistema de gerenciamento interno da instituição através de uma consulta no banco de dados. Com isso, os dados foram inseridos em outro documento CSV que contém o cruzamento das duas bases de dados.

Neste processo foi também necessário a transformação de diversos atributos presentes na base de dados para alcançar um total de interações e de tempo de acesso dentro da plataforma durante o período analisado, já que o objetivo é classificar a tendência a evasão do curso levando em consideração o comportamento do mesmo dentro do AVA e os registros da base de dados são por eventos realizados dentro da plataforma. Para isso foram desenvolvidas funções na base de dados adquirida do AVA para se chegar em um total de interações e tempo de acesso na plataforma Moodle.

Para então se aplicar os algoritmos de classificação foi necessária a conversão de atributos categóricos para numéricos, como a categoria do curso e o trimestre cursando e para isso foi utilizado o pacote *preprocessing* da biblioteca Scikit-learn.

4.5 Aprendizado

Nesta etapa é realizada a exploração dos dados pré-processados, com o intuito de extrair tendências e influências relacionadas as estatísticas e informações presentes no conjunto de dados. Para a seleção dos atributos foi realizado uma análise estatística dos atributos de maior impacto citados em trabalhos relacionados a essa área de estudo e através de uma análise realizada pelos responsáveis pelo ensino a distância da instituição. Durante o desenvolvimento também foi verificado o peso dos atributos com algoritmos específicos para isso.

Para se alcançar um resultado mais preciso nas análises, o conjunto total de dados foi separado em três conjuntos, onde o primeiro deles contempla a totalidade dos dados, o segundo conta com os dados de alunos possivelmente evadidos somente no ano de 2019 com mais de 500 interações no AVA e o terceiro contempla somente os atributos referentes ao AVA do segundo conjunto, ignorando dados originários do sistema de gerenciamento da instituição que podem ter uma baixa confiabilidade. Posteriormente, os atributos selecionados foram aplicados nos algoritmos citados nas seções 2.7 e 2.8.

Na aplicação dos algoritmos é feito o uso da biblioteca Scikit-learn em um ambiente de desenvolvimento do Google Colaboratory com a linguagem Python, levando em conta que ela possui todos os algoritmos necessários de classificação para o estudo. Neste mesmo ambiente é utilizada também a biblioteca Pandas para a manipulação dos dados. A divisão dos dados foi feita em conjuntos de treinamento e teste, onde o conjunto de treinamento representa 80% dos dados.

4.6 Avaliação

O processo de avaliação dos algoritmos tem a finalidade de, após o processo de treinamento utilizando-se dos atributos de entrada originários das interações no AVA, tempo de acesso no AVA e outras características do aluno nos

trimestres anteriores no começo do trimestre, comparar qual dos algoritmos apresenta o maior índice de acuracidade na classificação da tendência dele evadir ou não do curso. A avaliação dos algoritmos também é realizada em métricas como precisão, revocação e Medida F.

A partir destas análises é possível concluir qual dos algoritmos possui o melhor desempenho na classificação da evasão para então utilizá-lo como modelo de previsão nos próximos trimestres dos cursos EaD e possivelmente em outras frentes do processo de ensino-aprendizagem da instituição.

No capítulo a seguir desta monografia será apresentado o desenvolvimento do trabalho, apresentando detalhadamente as tarefas realizadas durante o mesmo e os resultados obtidos.

5 DESENVOLVIMENTO

Neste capítulo são descritos os processos realizados durante o desenvolvimento da monografia buscando alcançar os objetivos propostos. Nas seções a seguir, são mostrados os dados de entrada, detalhando a origem destes dados, os atributos e todos os processos de pré-processamento em que estes dados foram submetidos. No capítulo também são detalhados os resultados obtidos no processo de mineração de dados, comparando o desempenho entre os diferentes algoritmos aplicados.

5.1 Dados de Entrada

A primeira ação realizada foi um estudo estatístico dos dados que seriam relevantes para se alcançar o objetivo do trabalho de predizer através de algoritmos de aprendizado de máquina se um estudante tem a tendência de evadir ou não do curso EaD e baseando-se nos processos adotados pelos autores estudados nos trabalhos relacionados e em variáveis de interesse alinhadas em reunião com o setor responsável pelo processo de ensino-aprendizagem em EaD da instituição, os atributos apresentados na Tabela 3 foram os utilizados no desenvolvimento do trabalho e foram os que apresentar os melhores resultados após a aplicação dos algoritmos.

Tabela 3 - Atributos para previsão da evasão no EaD.

submissions	Total de submissões realizadas no Moodle
course_views	Total de visualizações de disciplinas relativas ao EaD na plataforma Moodle
video_int	Total de interações e participações nas aulas ao vivo ou gravadas
forum_int	Total de interações realizadas nos fóruns das disciplinas
external_int	Total de interações e submissões de arquivos fora da plataforma Moodle
time_med	Tempo de acesso médio na plataforma Moodle em minutos
trimester	Último trimestre que o aluno frequentou
course	Categoria do curso
age	Idade
reprate	Número de reprovações no histórico do aluno
polo	Polo EaD frequentado
incentive	Tem incentivo ou não
evasion	Evadido ou não

Fonte: Do Autor (2020).

Os dados contidos nessa base de dados tiveram origem no cruzamento entre duas bases de dados maiores, originarias do Sistema de gerenciamento interno da instituição e da plataforma Moodle da instituição. A exportação destes dados foi feita nos formatos CSV e SQL respectivamente e eles abrangem o período dos anos 2018 e 2019 e seus respectivos trimestres como citado na seção 4.3.

Os dados referentes aos atributos *trimester, course, age, reprate, polo e incentive* e *evasion* foram extraídos do sistema ERP da instituição através de um relatório montado no sistema Adianti Reports[16] através de funções de banco de dados desenvolvidas em conjunto com o DBA responsável. A função desenvolvida para a extração dos dados buscou todos os alunos matriculados entre o período de 2018 e 2019 e os dados referentes aos trancamentos de matrícula referentes aos currículos EaD, junto a todos os campos de informações agregadas nos registros. Após a geração do relatório, ele foi descarregado em formato CSV para

[16] Adianti Reports - https://www.adianti.com.br/reports - Acessado em 28 Jun. 2020.

os dados serem manipulados e tratados através da ferramenta de planilhas Libre Office.

Os dados referentes aos atributos *submissions, course_views, vídeo_int, fórum_int, external_int* e *time_med* foram retirados da plataforma Moodle aplicada na instituição e são originários de três tabelas distintas da base de dados geral, conforme citado na seção 4.3. Todos estes atributos contemplam dados totais de cada aluno, por exemplo, total de submissões durante o período, total de visualizações dos cursos, total de interações em vídeo, fórum e externas e também o tempo médio que cada aluno está dentro do ambiente virtual. Esse conjunto de tabelas foi fornecida pelo DBA responsável através de arquivos no formato SQL e elas foram importadas para o PostgreSQL e manipuladas através da ferramenta PgAdmin4[17]. Nenhum dos atributos analisados foi extraído em formato original da base pois todos concentram totais de valores alcançados na etapa de pré-processamento, que será detalhada posteriormente.

Após a extração inicial, o conjunto de dados concentrava 2635 resultados de alunos que se encaixavam dentro dos parâmetros buscados e após o pré-processamento, o número baixou para 2486 resultados totais, com 13 atributos cada um.

Para se alcançar um resultado mais preciso nas análises, o conjunto de dados foi separado em três conjuntos específicos e que, nos dois últimos, diminuiu o número total de atributos, sendo eles:

- **Experimento 1 – Totalidade dos dados:** O primeiro conjunto contempla a totalidade dos dados durante os anos de 2018 e 2019, incluindo dados do ano de 2018, que foi o primeiro ano do EaD na instituição, ano em que os registros possuíam muitos ruídos e baixa confiabilidade, já que diversas informações ainda não estavam incluídas no sistema de protocolos da instituição;

[17] PgAdmin4 - https://www.pgadmin.org/docs/pgadmin4/development/index.html - Acessado em 27 Jun. 2020.

- **Experimento 2 – Dados de 2019 com mais de 500 interações:** o segundo contempla os dados de alunos possivelmente evadidos somente no ano de 2019 com mais de 500 interações no AVA, excluindo os dados ruidosos de 2018 e também excluindo alguns registros que possuíam uma baixa contagem de interações no AVA, podendo influenciar de forma errônea na precisão dos resultados;
- **Experimento 3 – Dados de 2019 com mais de 500 interações e somente atributos relacionados ao AVA:** o terceiro conjunto de dados contempla somente os atributos referentes ao AVA do segundo conjunto, ignorando dados originários do sistema de gerenciamento da instituição que possuem certo ruído e podem ter uma influência negativo as análises dos algoritmos.

A etapa de captura de dados não se consolidou totalmente antes de iniciar a etapa do pré-processamento, pois como descrito por Goldschmidt e Passos (2005), no processo de descoberta do conhecimento é possível que ao se iniciar uma etapa subsequente do processo, perceba-se algum problema ou ruído gerado em uma etapa anterior assim sendo necessário revisar e retrabalhar algum atributo ou variável. Durante a etapa de pré-processamento foi percebido a falta de alguns atributos essenciais para o trabalho e eles foram buscados e retrabalhados diversas vezes até se alcançar o conjunto de dados ideal.

Na Figura 13 é possível analisar os dados antes da etapa de pré-processamento.

Figura 13 - Dados da análise antes do pré-processamento.

evasion	submissions	course_views	bigbluebutton	fórum	external	access_median	course_type	work	age	dpp_rate	polo	incentivo	trimestre
1	85	1061	51	63	12	01:15:20	gestao	NAO	53	0	Lajeado	0,10	1
0	89	967	143	55	19	01:19:29	gestao	SIM	53	8	Encantado	0,10	4
1	27	58	10	11	2	00:41:34	gestao	NAO	49	1	Lajeado		2
0	150	2352	69	71	44	00:33:59	gestao	NAO	46	0	Encantado		4
0	112	828	42	108	51	01:21:16	gestao	NAO	45	0	Lajeado	0,10	4
0	216	3196	131	514	141	01:33:46	humanas	NAO	45	0	Lajeado		4
0	61	421	4	11	4	00:42:37	gestao	NAO	45	3	Lajeado		4
0	3	37	9	3	1	00:10:21	ti	NAO	54	0	Teutônia	0,10	4
0	10	240	2	35	3	00:36:22	gestao	NAO	57	24	Lajeado	0,20	4

Fonte: Do Autor (2020).

Na Figura 13 também é possível verificar alguns atributos que no final não foram incluídos na pesquisa por não serem dados precisos e apresentarem ruídos

em demasia. Eles foram removidos pois acreditou-se que os mesmos afetariam negativamente os resultados.

Na seção seguinte é abordada a etapa de pré-processamento dos dados, onde é demonstrado como os dados foram preparados para posteriormente aplicar as técnicas de mineração de dados.

5.2 Pré-processamento dos dados

Nesta etapa do trabalho, foi buscada a integração entre os dois conjuntos de dados e para isso foram extraídos os dados dos alunos aptos a estarem incluídos na análise e então, a partir dos códigos de aluno e códigos dos cursos EaD destes foram construídas diversas funções de banco de dados para extrair os dados dos respectivos alunos e cursos dos registros da plataforma Moodle os quais constituíram a base de dados analisada.

A etapa de pré-processamento contou com diversas fases onde foram necessárias aplicar técnicas de limpeza dos dados pois eles possuíam muitos ruídos, valores errados e valores ausentes, ruídos esses que foram verificados tanto na base do sistema de gerenciamento interno da instituição quanto na base da plataforma Moodle.

Neste processo, no conjunto de dados extraídos do sistema de gerenciamento interno, também foi necessária a transformação e discretização de diversos atributos categóricos, enquanto outros ou tornaram-se redundantes por estarem presentes nas duas bases ou não se mostraram relevantes para o estudo. Os atributos *course* e *incentive* por exemplo, possuíam o nome exato do curso e o total do incentivo em porcentagem de desconto nos valores das disciplinas e foram transformados e discretizados em classes de áreas do curso e em se o aluno contava com incentivo ou não, respectivamente, buscar uma maior performance na análises posteriores.

Já no pré-processamento dos dados referentes a plataforma Moodle, o processo de extração foi feito de forma mais automatizada, sendo criadas funções de banco de dados com tabelas temporárias que buscaram concentrar os dados necessários para análise.

O mesmo ocorreu em duas etapas, onde a primeira fez a extração, limpeza e transformação dos valores referentes a interação dentro da plataforma levando em conta as ações tomadas dentro da plataforma como submissão de trabalhos, visualizações de disciplinas do curso, interações com vídeos de aulas, interações no fórum e *upload* e *download* de arquivos da plataforma, seguindo a referência da Figura 6. Para a concentração destes dados foi desenvolvida uma rotina *PrepareStatement* em SQL que recebia o código do aluno e a último data em que ele teve registros antes da evasão do curso, sendo usado o último dia de 2019 de referência para os não evadidos. Essa rotina inseriu os dados transformados em uma tabela auxiliar temporária para visualização e exportação destes dados, parte da função pode ser vista na Figura 14.

Figura 14 - Parte da rotina PrepareStatement para extração dos dados do Moodle.

```
#PrepareStatement to capture everything
PREPARE usrmdldate (bigint, date) AS
SELECT usuario, vw.VIEWED, it.INTERACTS, bb.BIGBLUE, ff.FORUM
FROM
    (
    SELECT mus.username as usuario, mus.id,
    count(mlog.action) FILTER (WHERE mlog.action IN ('viewed')) as VIEWED
    FROM mdl_logstore_standard_log mlog, mdl_user mus, mdl_course mcu
    WHERE mlog.userid IN (
    SELECT id
    FROM mdl_user
    WHERE username IN ($1) )
    AND courseid IN (
    SELECT id
    FROM mdl_course
    WHERE category IN (127, 137, 205, 167, 188, 200) )
    AND mus.id = mlog.userid
    AND mcu.id = mlog.courseid
    AND mlog.target = 'course'
    AND to_timestamp(mlog.timecreated) > '2017-12-31'
    AND to_timestamp(mlog.timecreated) < $2
    group by 1, 2
```

Fonte: Do Autor (2020).

A segunda etapa consistiu em verificar e extrair o tempo médio de acesso a plataforma de cada aluno, etapa essa que foi automatizada por uma função de banco de dados que analisou o horário da primeira interação após a ação de *login* na plataforma até o horário da última interação antes da ação de *logout* somando o intervalo de tempo entre cada clique e dividindo pelo total de interações que o aluno teve no ambiente, ignorando a ação *failed*, que registra as tentativas falhas de *login*. A função desenvolvida é apresentada na Figura 15.

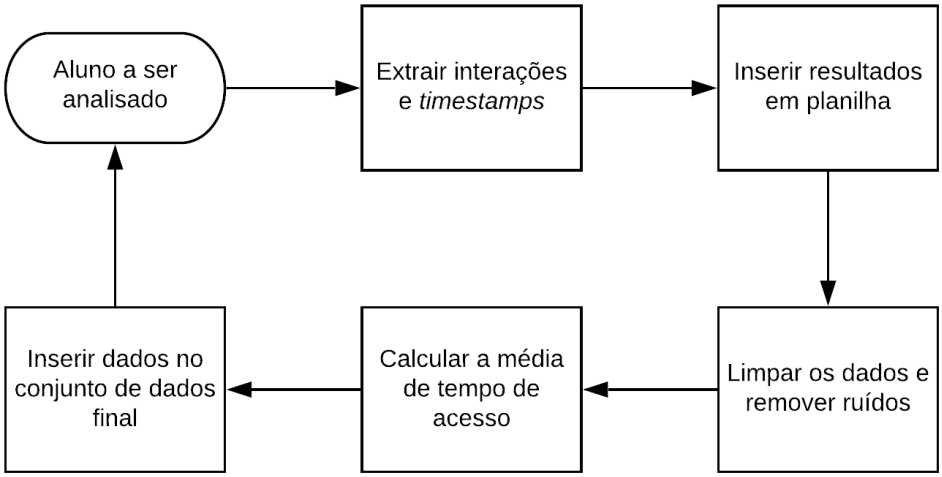
Figura 15 - Função para extração do tempo médio de acesso a plataforma.

```
SELECT accesstime.username, AVG(accesstime.total)
FROM ( SELECT
  username,
  userid,
  action,
  to_timestamp(timecreated)::TIMESTAMP,
  to_timestamp(timecreated)::TIMESTAMP - lag(to_timestamp(timecreated)::TIMESTAMP, 1)
      OVER (ORDER BY timecreated) delta,
  to_timestamp(timecreated)::TIMESTAMP - first_value(to_timestamp(timecreated)::TIMESTAMP)
      OVER (PARTITION BY c ORDER BY timecreated) total
FROM (
  SELECT
    COUNT(*) FILTER (WHERE ACTION = 'loggedin')
            OVER (ORDER BY timecreated) c,
    timecreated,
  action,
  userid,
  mus.username
  FROM mdl_logstore_standard_log, mdl_user mus
  WHERE userid IN (
  SELECT id
  FROM mdl_user
  WHERE username IN ('usercode') )
AND mus.id = userid
AND to_timestamp(timecreated) > '2017-12-31'
AND to_timestamp(timecreated) < '2019-12-31'
) mdl_logstore_standard_log
ORDER BY username, timecreated ) accesstime
group by 1
order by 1;
```

Fonte: Do Autor (2020).

Por questões de ruídos e algumas falhas nos registros de *logout*, alguns alunos ficaram com os tempos de acesso a plataforma muito discrepantes e em função disso foi realizado um trabalho manual de tratamento desses registros para cada aluno, inserindo as interações em uma planilha e removendo manualmente discrepâncias de tempo.

Figura 16 - Fluxograma do processo de tratamento das interações.



Fonte: Do Autor (2020).

Após o tratamento destes dados, eles foram discretizados de horas para minutos e foram inseridos no arquivo CSV. Finalizando o pré-processamento dos dados foi feita a discretização de alguns atributos em classes para exploração dos mesmos, atributos como tempo médio de acesso e taxa de reprovação foram discretizados utilizando o pacote *preprocessing* do Scikit-learn em classes que vão de muito baixo a muito alto. A exploração realizada nos dados é detalhada na seção 5.3.

Depois dos procedimentos realizados nesta seção, o conjunto de dados está pronto para ser aplicado nos algoritmos de aprendizado de máquina. Os resultados obtidos são detalhados na seção 6.

6 RESULTADOS

Nesta seção é feita uma exploração das informações e estatísticas presentes no conjunto de dados após o processo de integração, onde se buscou identificar tendências e quais os atributos que possam ser mais influentes na hora de identificar a evasão.

6.1 Exploração dos dados

Em uma observação inicial dos dados já foi possível se verificar que a taxa de evasão na realidade da instituição é praticamente o dobro da observada na pesquisa realizada pela ABED em 2018, como se pode observar na Tabela 4.

Tabela 4 - Taxa de evasão no EaD entre 2018 e 2019.

Total Analisados	Total Evadidos	%
2487	982	39,49%

Fonte: Do Autor (2020).

Aprofundando a análise, foi feita uma exploração de outros atributos que deveriam impactar no comportamento dos alunos dentro do curso EaD e que pudessem influenciar eles a desistir do curso ou alguma disciplina, culminando na

evasão. Os atributos foram escolhidos tomando por base as referências estudadas e algumas observações iniciais realizadas pelos responsáveis pelo processo de ensino-aprendizagem da instituição.

O primeiro fator observado foi a faixa etária em que se encontravam os alunos analisados com a finalidade de se alcançar um público que possuísse uma maior tendência a desistir do curso, onde variáveis como problemas familiares, dificuldades no manuseio de tecnologias ou falta de tempo teriam um impacto maior em faixas etárias superiores. Porém, como se pode observar na Tabela 5, todas faixas etárias possuem resultados semelhantes nas taxas de evasão onde a maior faixa etária observada possui um resultado levemente superior as demais.

Tabela 5 - Taxas de evasão por faixa etária.

Faixa etária	Analisados	Evadidos	%
<19	192	74	38,54%
20-24	826	331	40,07%
25-29	530	217	40,94%
30-34	436	166	38,07%
35-39	283	106	37,45%
40-45	138	55	39,85%
45>	81	34	41,97%

Fonte: Do Autor (2020).

Após isso foi feita uma observação nas áreas distintas de atuação dos cursos ofertados e já pode se verificar que os cursos de gestão e humanas possuem a superioridade de alunos matriculados concentrando um montante de 88,18% dos alunos analisados no estudo. Finalizando os cálculos das taxas de evasão foi possível observar que os cursos de gestão possuem as menores taxas, com 37,85% de alunos evadidos no total e os cursos de tecnologia ficaram com os maiores valores, com mais de 10% superior aos cursos de gestão totalizando em

um valor de 48,34% de desistência observada. A totalidade dos dados pode ser observada na Tabela 6.

Tabela 6 - Taxas de evasão por área de atuação do curso.

Curso	Analisados	Evadidos	%
Ciências	82	34	41.46%
Gestão	1527	578	37,85%
Humanas	666	269	40,39%
TI	211	102	48,34%

Fonte: Do Autor (2020).

Algo que é muito discutido e afirmado em diversos analises como um dos fatores mais influentes na predição da tendência que um aluno possui para evadir ou não do curso é o tempo em que o aluno frequenta o ambiente virtual de aprendizado e nas análises deste estudo foi muito semelhante. As classes foram definidas através de uma análise estatística dos dados e fazendo o uso de medianas e por final ficaram definidas de forma que a classe Muito Baixo eram alunos que possuíam menos de 20 minutos de tempo médio de acesso, Baixo uma média entre 20 e 29 minutos, Médio entre 30 e 45 minutos, Alto entre 46 e 69 minutos e acima de 70 minutos se considerou a classe Muito Alto.

A tendência de o aluno evadir do curso diminui gradativamente conforme o tempo médio de acesso do aluno na plataforma aumenta, conforme pode ser observado na Tabela 7.

Tabela 7 - Taxas de evasão por tempo de acesso médio ao AVA.

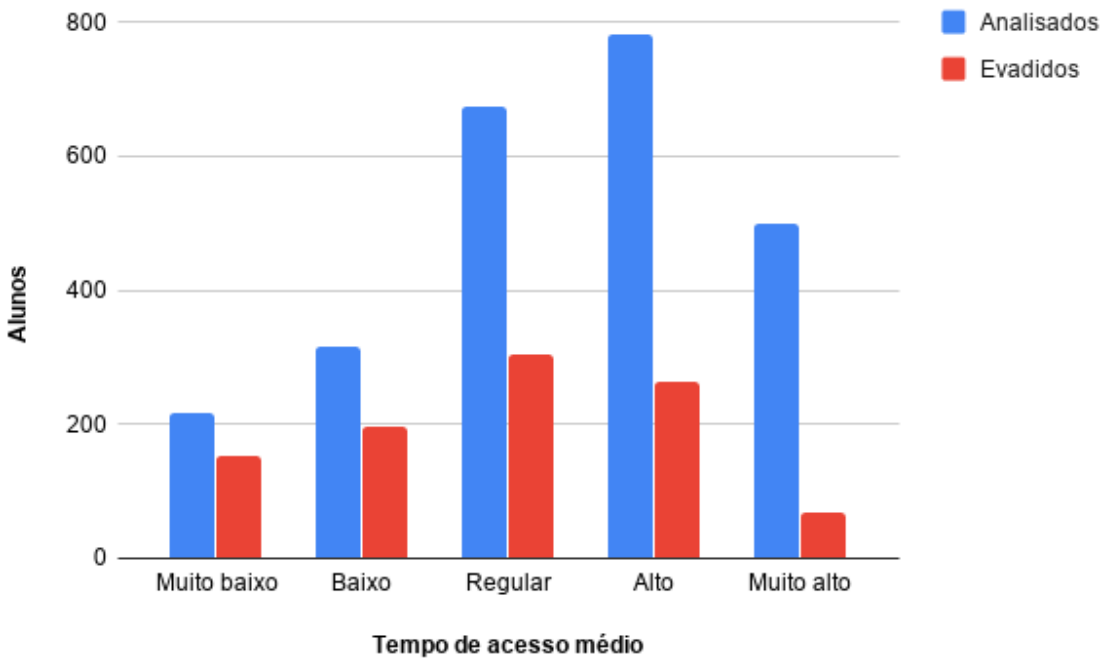
Tempo de acesso médio	Analisados	Evadidos	%
Muito baixo	217	152	70,05%
Baixo	316	195	61,70%
Regular	674	305	45,25%

Alto	781	263	33,67%
Muito alto	498	67	13,45%

Fonte: Do Autor (2020).

Alunos com médias muito baixas e baixas de tempo de acesso ficam com valores muito superiores a 50% de desistência e como contraponto alunos com médias regulares e altas ficam em menos de 50% e 40% respectivamente e, alunos que frequentam assiduamente o ambiente e possuem médias muito altas apresentam taxas de menos de 15% de tendência a evasão. No Gráfico 6 é possível observar estes dados.

Gráfico 6 - Relação de evasão por tempo de acesso médio.



Fonte: Do Autor (2020).

Dentro da exploração dos dados foram buscados também verificar se havia uma época em que os estudantes mostrassem uma tendência à evasão, e para isso, foram analisadas taxas de evasão por período trimestral, levando em conta que os módulos EaD são divididos em quatro trimestres, conforme citado na seção 4.3.

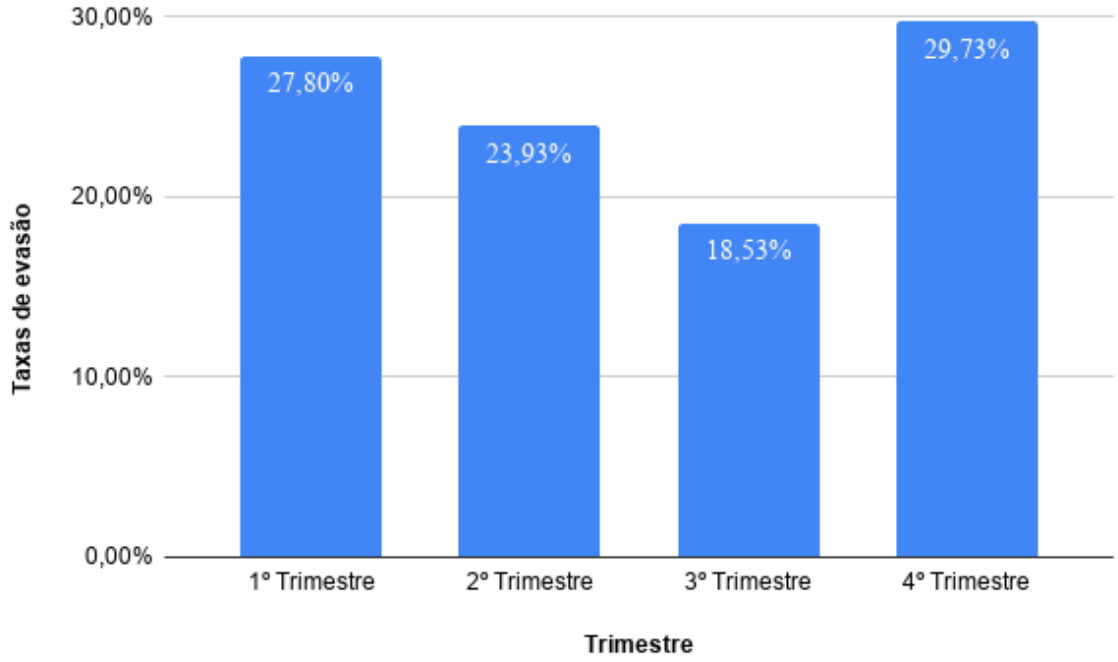
Tabela 8 - Taxas de evasão por período trimestral.

Trimestre	Número de evadidos	%
1º Trimestre	273	27,80%
2º Trimestre	235	23,93%
3º Trimestre	182	18,53%
4º Trimestre	292	29,73%

Fonte: Do Autor (2020).

Observando os dados da Tabela 8 é possível se verificar que as taxas são observadas no primeiro e no último trimestre, muito provavelmente pelos períodos analisados incluírem o pré início e o pós fim das aulas respectivamente. O segundo trimestre vem logo após e leva ao encontro de uma tendência já verificada anteriormente onde caso os alunos concluíssem o primeiro semestre com êxito, os mesmos geralmente seguiam até o final do módulo e caso não, já sairiam do curso. No Gráfico 7 é apresentado de forma mais clara as taxas trimestrais.

Gráfico 7 - Relação das taxas de evasão por trimestre.



Fonte: Do Autor (2020).

Outro fator que deve se mostrar importante para estudos futuros e que foi observado neste estudo, são as taxas de evasão por polo ativo da instituição onde os valores variam bastante, sendo que alguns polos apresentam valores baixos e outros altos. Um comportamento observado neste fator é o de polos localizados em cidades mais distantes possuírem as taxas mais elevadas e os polos localizados em cidades vizinhas da matriz possuírem os alguns dos menores valores, conforme a Tabela 9 exibe.

Tabela 9 - Taxas de evasão por polo EaD.

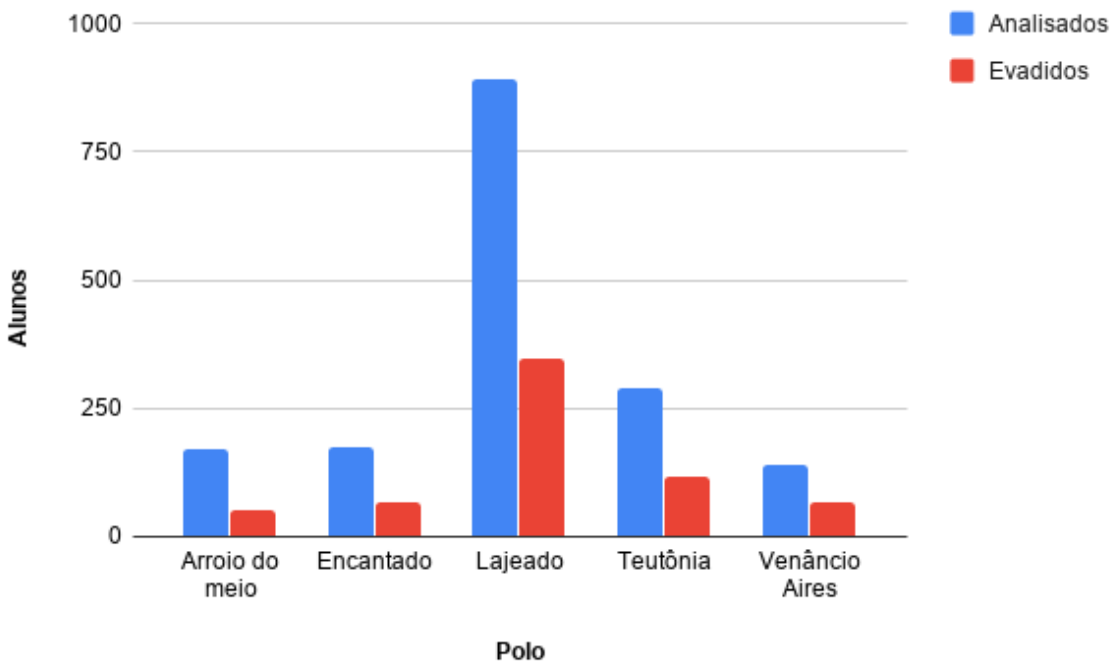
Polo	Estudantes	Evadidos	%
Arroio do meio	170	53	31,17%
Arvorezinha	73	24	32,87%
Bom Retiro do Sul	87	28	32,18%
Carlos Barbosa	136	61	44,85%
Cruz Alta	33	17	51,51%
Encantado	176	66	37,50%
Estrela	102	34	33,33%
Garibaldi	54	19	35,18%
Guaporé	75	36	48,00%
Ibirubá	2	0	0,00%

Lajeado	890	345	38,76%
Montenegro	3	2	66,67%
Porto Alegre	5	0	0,00%
Serafina Corrêa	51	22	43,13%
Soledade	67	31	46,26%
Taquari	100	43	43,00%
Teutônia	289	116	40,13%
Triunfo	11	11	100,00%
Venâncio Aires	141	65	46,09%
Veranópolis	21	9	42,85%

Fonte: Do Autor (2020).

Também pode se afirmar que o maior volume de público da modalidade EaD ainda se encontra no polo matriz na cidade de Lajeado e por isso ela conta com o maior número de alunos evadidos, conforme constatado pelo Gráfico 8. Muitos dos polos descritos na análise estavam em implantação até pouco antes ou durante o período em que os dados foram extraídos, por este motivo em estudos futuros estes valores devem sofrer alterações.

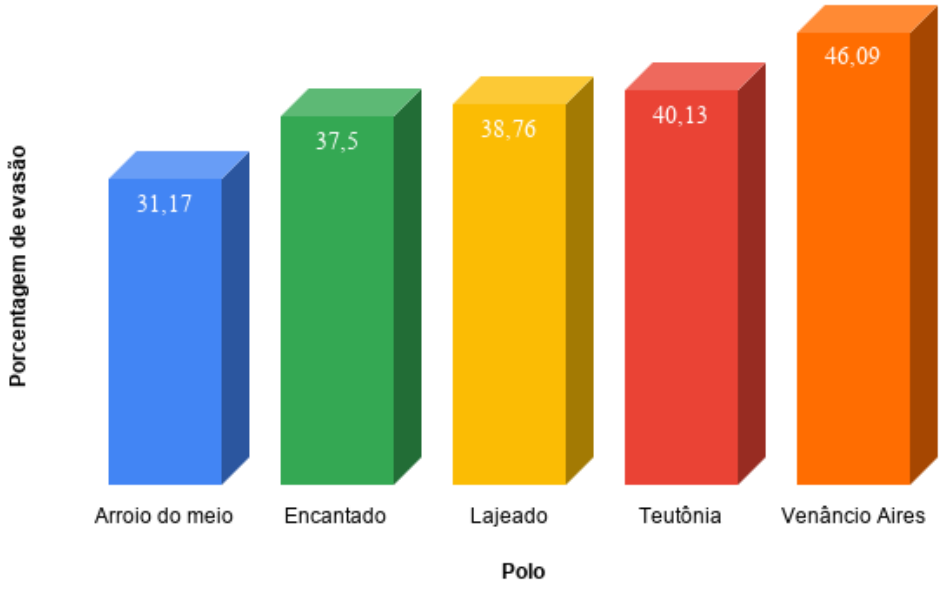
Gráfico 8 - Relação dos polos que possuem mais alunos matriculados.



Fonte: Do Autor (2020).

Observando os mesmos dados no ponto de vista da porcentagem, percebesse que apesar do polo de Lajeado possuir o maior número de evasão de alunos, a taxa se mantém semelhante se comparado aos outros, podendo se perceber inclusive que o polo de Venâncio Aires possui uma taxa de evasão mais elevado que a do polo de Lajeado, conforme é possível observar no Gráfico 9.

Gráfico 9 - Taxas de evasão por polo.



Fonte: Do Autor (2020).

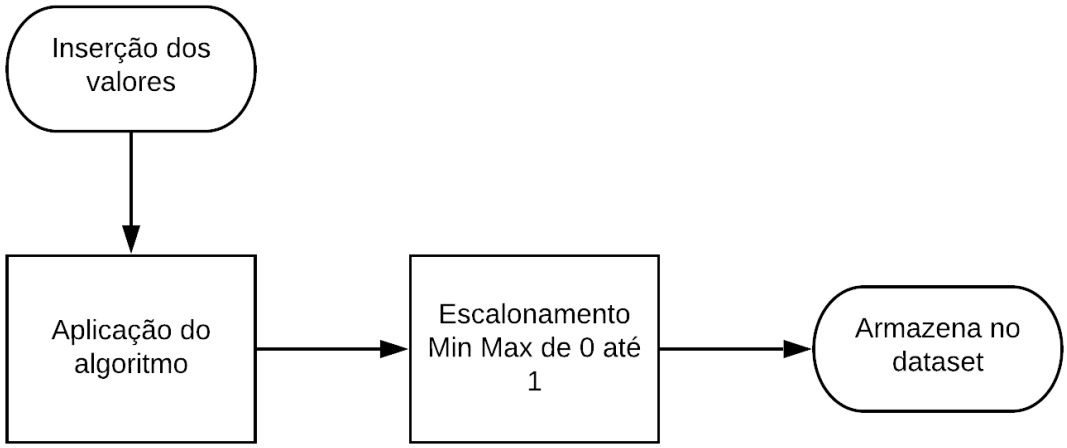
6.2 Resultados das análises com as técnicas de mineração de dados

Os resultados apresentados a seguir foram obtidos através da aplicação dos algoritmos pertencentes a biblioteca Scikit-Learn. Foi possível se realizar uma análise exploratória aprofundada utilizando-se de diversos testes com a ferramenta para validação dos dados. De forma exploratória, os atributos apresentados na seção 5.1 deste estudo foram aplicados nas ferramentas da biblioteca em três experimentos.

Antes de fazer o armazenamento dos valores no *dataset* criado pela biblioteca Pandas para se aplicar no algoritmo, foi necessária a aplicação do algoritmo MinMaxScaler com a finalidade de escalonar os números dos resultados

e transformar os atributos categóricos em classes numéricas visando uma melhora de desempenho e para evitar a limitação de alguns algoritmos que só trabalham com valores numéricos. A Figura 17 apresenta o fluxograma desta etapa.

Figura 17 - Fluxograma para armazenamento dos atributos.



Fonte: Do Autor (2020).

Os dados antes de serem tratados pelo MinMaxScaler se encontravam em intervalos esparsos, os quais variavam muito entre um aluno e outro, conforme a Figura 18.

Figura 18 - Dados esparsos antes do tratamento.

	submissions	course_views	video_int	forum_int	external_int	time_med
2185	100	1623	115	127	54	67
1431	336	591	50	31	18	96
1472	6	557	0	23	9	19
2306	54	951	120	58	18	30
1187	91	967	127	19	27	46

Fonte: Do Autor (2020).

Após a aplicação do algoritmo, os dados foram escalonados em intervalos de 0 a 1 nos valores esparsos, facilitando o reconhecimento dos algoritmos e melhorando o desempenho, a estrutura é exemplificada na Figura 19.

Figura 19 - Dados escalonados após o tratamento.

	submissions	course_views	video_int	forum_int	external_int	time_med
1980	0.098868	0.122447	0.049407	0.075956	0.176692	0.251082
1101	0.041509	0.045871	0.009881	0.006810	0.048872	0.186147
1560	0.187925	0.088288	0.037549	0.009429	0.048872	0.134199
1578	0.003774	0.002327	0.000000	0.001048	0.000000	0.181818
660	0.415849	0.062913	0.084980	0.052907	0.244361	0.134199

Fonte: Do Autor (2020).

O conjunto de dados de entrada foi dividido em dois para ser aplicado nos algoritmos, um conjunto de treinamento e um conjunto de testes utilizando-se a função *train_test_split* da biblioteca Scikit-Learn. A divisão dos dados foi de 80% para treinamento e 20% para teste no experimento 1 e de 70% para treinamento e 30% para teste nos experimentos seguintes.

Após separar os conjuntos de treino e teste, foi aplicada a função *GridSearchCV*, que é uma técnica de validação cruzada que realiza uma busca exaustiva através dos valores dos parâmetros para encontrar a melhor combinação de parâmetros, a qual irá gerar o melhor modelo preditivo possível. Na sequência, os conjuntos de testes foram submetidos ao modelo preditivo gerado.

A aplicação dos mesmos se deu utilizando-se da abordagem de aprendizado supervisionado, onde os classificadores tiveram de indicar se conforme o comportamento dentro do AVA o estudante evadiu ou não do curso e se os outros parâmetros analisados possuem ou não influência nesse comportamento. Para fins de comparação, todos os algoritmos aplicados utilizaram seus parâmetros padrões conforme a biblioteca Scikit-Learn possui documentado.

Assim se considera o problema de classificação como binário, classificando como 0 o aluno que se manteve na instituição até o final de 2019 e como 1 caso o aluno evadiu do curso ou disciplina durante o período analisado. Para se avaliar o

desempenho dos algoritmos, métricas de acuracidade, precisão, revocação e Medida F foram utilizadas, realizando a comparação entre os modelos dentro destas métricas para se concluir qual o modelo apresentou o melhor desempenho dentro da proposta do trabalho. Na Tabela 10 é possível a visualização da acuracidade dos modelos para os 3 experimentos.

Tabela 10 - Acuracidade dos modelos.

Algoritmo	SVM	Random Forest	NaiveBayes	Decision Tree	KNN
Experimento 1	93,57%	94,37%	90,56%	96,38%	92,97%
Experimento 2	97,60%	98,26%	97,60%	98,47%	95,65%
Experimento 3	93,04%	97,39%	85%	96,73%	95,65%

Fonte: Do Autor (2020).

É possível afirmar que todos os algoritmos apresentaram resultados excelentes no primeiro experimento, com destaque para o Decision Tree que apresentou uma acuracidade superior a 95% nos três experimentos. Os outros algoritmos apresentaram um desempenho levemente inferior, com destaque para o NaiveBayes que apresentou uma queda de 7% no primeiro e um valor muito inferior, de 85% no terceiro experimento, o que demonstra que ele possui dificuldades em trabalhar com os valores esparsos do conjunto.

Para uma apresentação mais detalhada dos algoritmos que apresentaram as melhores acuracidades, foi feito o uso de matrizes de confusão, conforme as Tabelas 11, 12, 13, 14, e 15, as quais apresentam o total de instâncias classificadas de forma correta por classe nos diferentes experimentos.

Tabela 11 - Matriz de confusão do algoritmo Random Forest para o experimento 2.

	Classe predita	
Classe atual	**0**	**1**
0	340	1

1	7	112

Fonte: Do Autor (2020).

Tabela 12 - Matriz de confusão do algoritmo NaiveBayes para o experimento 3.

	Classe predita	
Classe atual	**0**	**1**
0	285	56
1	13	106

Fonte: Do Autor (2020).

Tabela 13 - Matriz de confusão do algoritmo SVM para o experimento 3.

	Classe predita	
Classe atual	**0**	**1**
0	341	0
1	32	87

Fonte: Do Autor (2020).

Tabela 14 - Matriz de confusão do algoritmo Decision Tree para o experimento 1.

	Classe predita	
Classe atual	**0**	**1**
0	306	5
1	13	174

Fonte: Do Autor (2020).

Tabela 15 - Matriz de confusão do algoritmo KNN para o experimento 1.

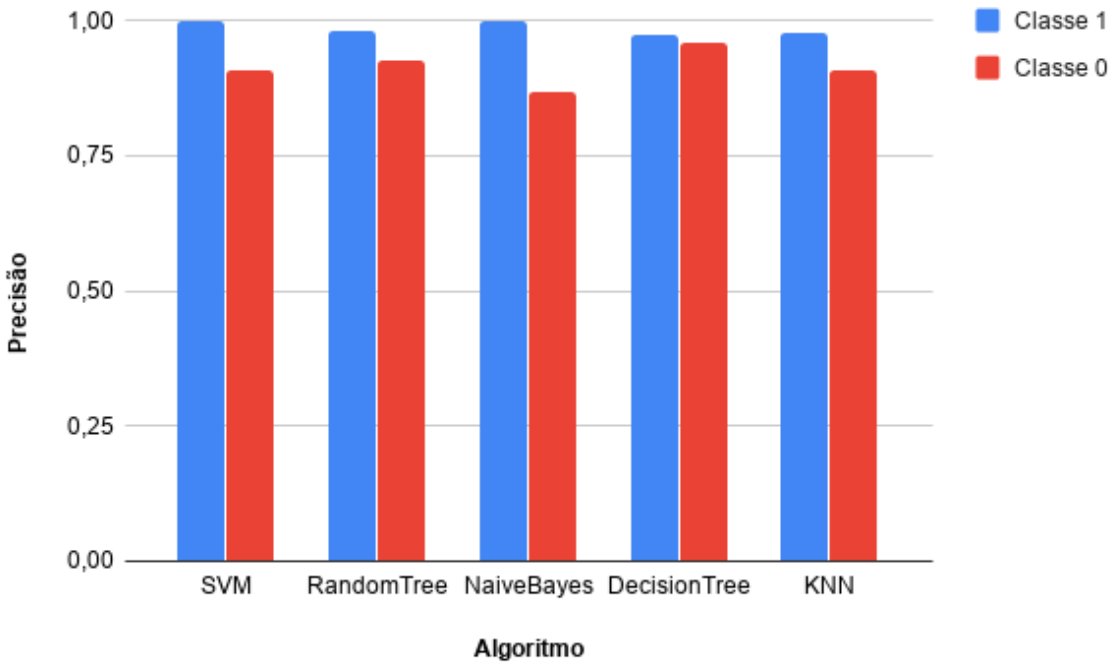
	Classe predita

Classe atual	0	1
0	307	4
1	31	156

Fonte: Do Autor (2020).

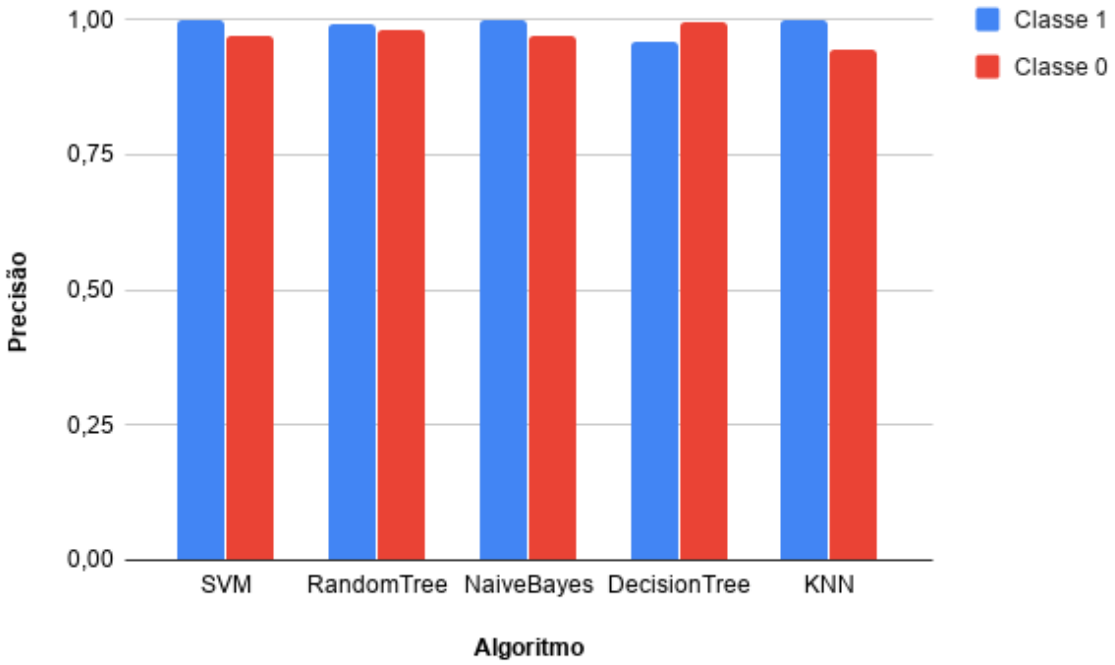
A partir dos Gráficos 10, 11 e 12 é possível verificar que a maioria dos modelos possui precisão superior na hora de predizer a classe 1, pois é verificado que somente o algoritmo Decision Tree possui um desempenho superior para a classe 0, que define que o aluno não possui tendência a evadir do curso, nos dois últimos experimentos além do NaiveBayes, que apresentou uma diferença grande de desempenho no experimento 3, sendo muito superior na classe 0 onde novamente podemos reforçar a dificuldade do algoritmo com os dados que possuem valores esparsos. Lembrando que no gráfico a medida está escalonada entre 0 e 1 representando de 0 a 100 na precisão.

Gráfico 10 - Precisão dos modelos por classe no experimento 1.



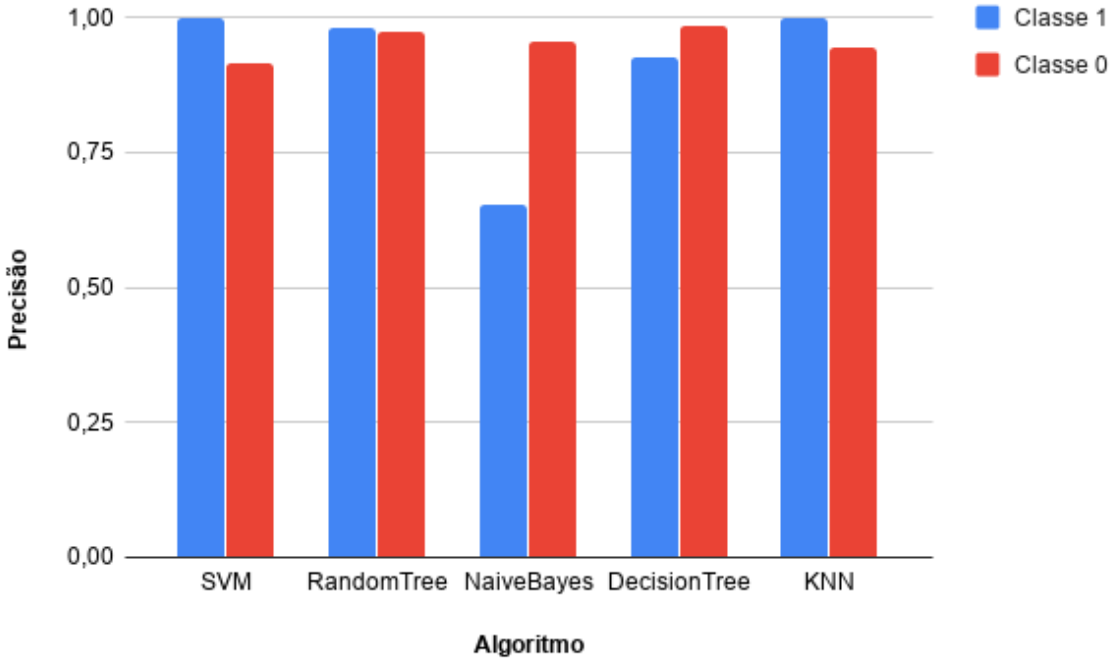
Fonte: Do Autor (2020).

Gráfico 11 - Precisão dos modelos por classe no experimento 2.



Fonte: Do Autor (2020).

Gráfico 12 - Precisão dos modelos por classe no experimento 3.



Fonte: Do Autor (2020).

É possível se verificar que os algoritmos SVM foi o único que obteve uma precisão de 100% nos três experimentos, que quer dizer que os algoritmos não

classificaram nenhum aluno como possível evasão se o mesmo realmente não possuía grandes tendências a evadir do curso, conforme a Tabela 16, sendo possível entender o porquê das máquinas de vetores de suporte serem conhecidas por terem um alto desempenho ao trabalharem com grandes volumes de dados.

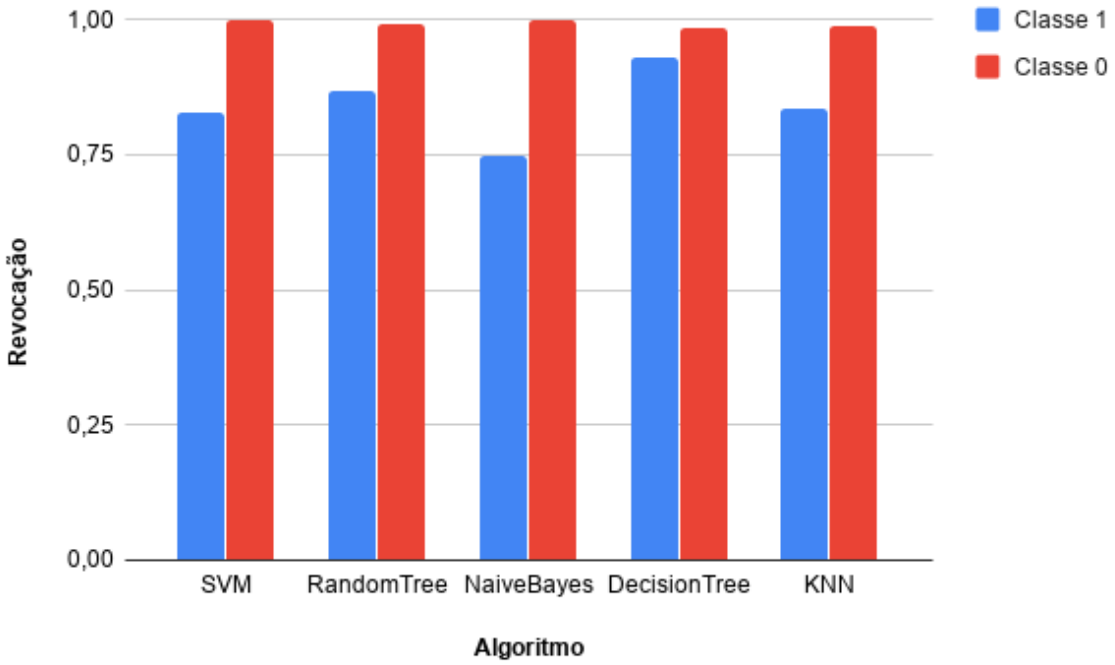
Tabela 16 - Precisão dos modelos através dos 3 experimentos.

Algoritmo	SVM	Random Forest	NaiveBayes	Decision Tree	KNN
Experimento 1	100,00%	98,18%	100,00%	97,21%	97,50%
Experimento 2	100,00%	99,12%	100,00%	95,90%	100,00%
Experimento 3	100,00%	98,20%	65,43%	92,62%	100,00%

Fonte: Do Autor (2020).

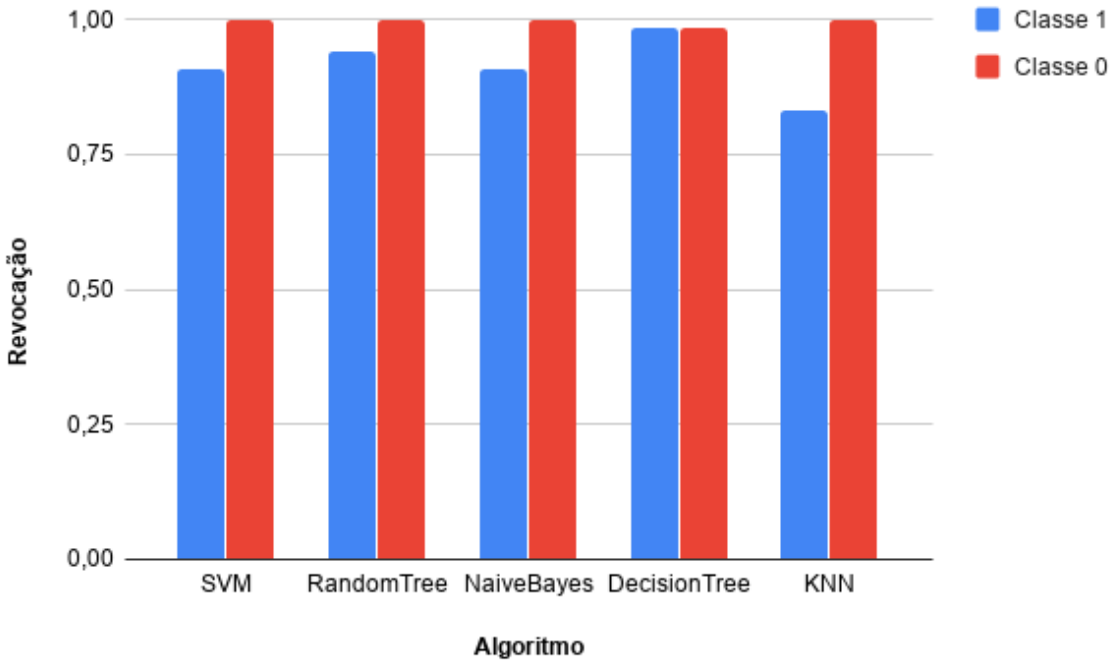
Nas análises de revocação dos modelos a tendência se inverte, conforme os Gráficos 13, 14 e 15, tendo a classe 0 com resultados superiores em relação a classe 1. Outro detalhe que é possível observar é a maior discrepância de valores entre as duas classes, onde somente o algoritmo Decision Tree manteve um equilíbrio entre as duas classes. O KNN apresentou o menor desempenho na classificação da classe 1 e o algoritmo SVM apresentou grande variância através dos três experimentos na classificação da classe 1. Quanto a classe 0, é possível verificar que todos os algoritmos obtiveram valores altos de revocação.

Gráfico 13 - Revocação dos modelos por classe no experimento 1.



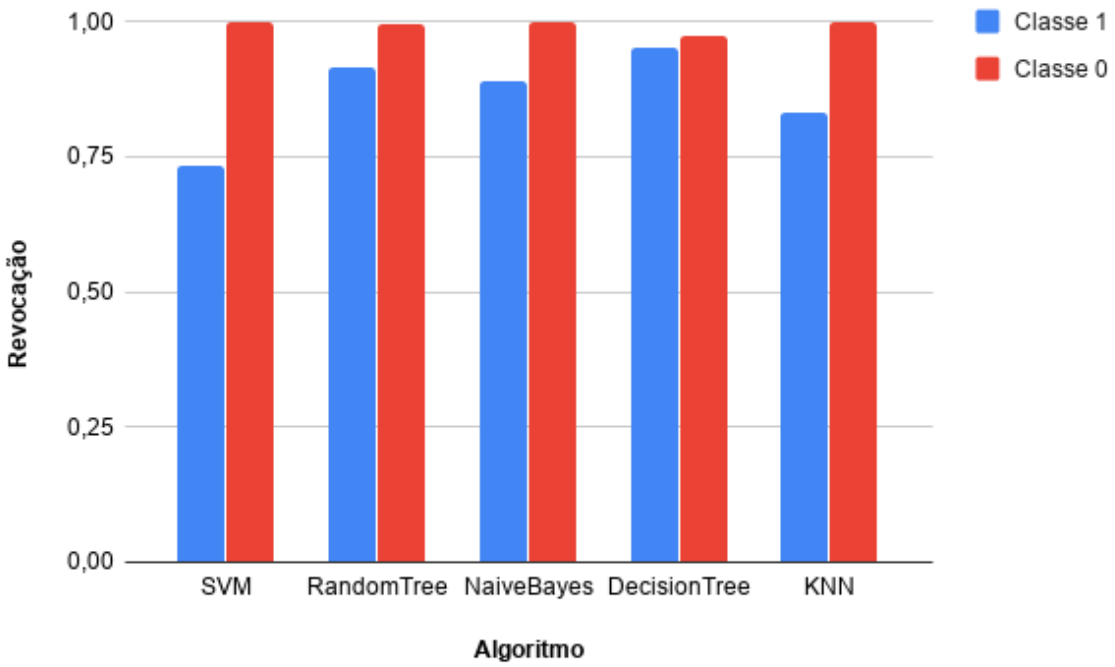
Fonte: Do Autor (2020).

Gráfico 14 - Revocação dos modelos por classe no experimento 2.



Fonte: Do Autor (2020).

Gráfico 15 - Revocação dos modelos por classe no experimento 3.



Fonte: Do Autor (2020).

No índice geral de revocação o algoritmo Decision Tree se destacou por apresentar valores acima dos 90% em todas as análises, demonstrando que o algoritmo foi o que apresentou a melhor segmentação dos dados na hora da escolha dos pontos corretos dentro da nuvem dispersa dos dados e assim, possuindo o melhor desempenho na classificação dos alunos que realmente evadiram e dos que realmente não possuem qualquer tendência de evasão, conforme apresenta a Tabela 17.

Tabela 17 - Revocação dos modelos através dos 3 experimentos.

Algoritmo	SVM	Random Forest	NaiveBayes	Decision Tree	KNN
Experimento 1	82,89%	86,63%	74,87%	93,05%	83,42%
Experimento 2	90,76%	94,12%	90,76%	98,32%	83,19%
Experimento 3	73,11%	91,60%	89,08%	94,96%	83,19%

Fonte: Do Autor (2020).

Na etapa de avaliação da Medida F, os algoritmos Random Forest, Decision Tree se sobressaíram sobre o restante dos algoritmos alcançando valores superiores a 0,9 na Medida F nos três experimentos, o que é um excelente desempenho, levando em consideração a máxima 1,0. Os outros algoritmos também mostraram um desempenho bom, destacando o algoritmo SVM que mostrou o seu melhor desempenho no experimento 2 que conta com um grande volume de dados com valores espalhados. Visualizando os dados é possível afirmar que a harmonia entre a precisão e a revocação foi excelente, ou seja, a probabilidade de haver distorções nos valores da matriz de confusão é mínima. Tudo isso pode ser observado melhor na Tabela 18.

Tabela 18 - Medida F dos modelos através dos 3 experimentos.

Algoritmo	SVM	Random Forest	NaiveBayes	Decision Tree	KNN
Experimento 1	0,9064	0,9205	0,8563	0,9508	0,8991
Experimento 2	0,9515	0,9655	0,9515	0,9710	0,9083
Experimento 3	0,8447	0,9478	0,7544	0,9378	0,9083

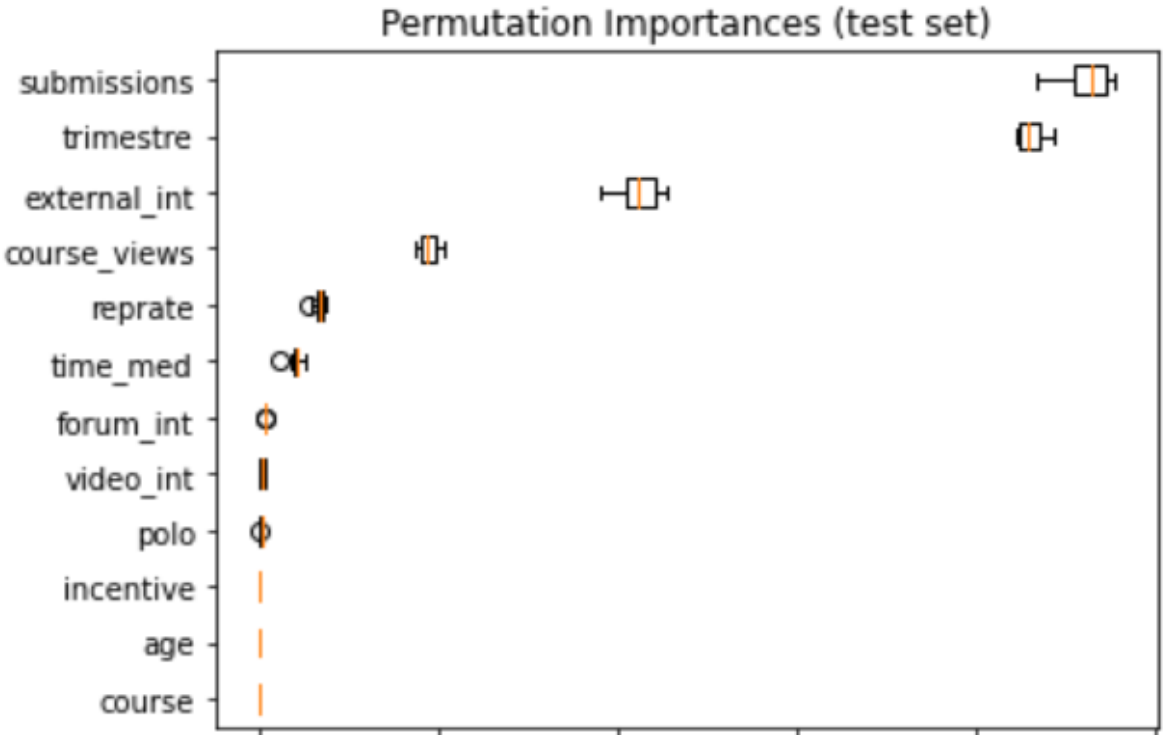
Fonte: Do Autor (2020).

Também foram feitas análises utilizando os recursos *feature_importances* localizado na biblioteca do Random Forest e o *permutation_importance* da biblioteca de inspeção do Scikit-Learn, sendo ambas funções para verificar a importância dos diferentes atributos para a predição dos algoritmos. Foram feitas as análises nos algoritmos Decision Tree no experimento 1 e SVM no experimento 2 com o permutation_importance e com o feature_importances no Random Forest no experimento 3.

Verificando os Gráficos 16, 17 e 18, que foram gerados pela biblioteca matplotlib no Scikit-Learn, é visto que para os algoritmos o atributo de maior impacto foram as submissões, seguido do atributo do trimestre em que ocorreu a evasão do aluno nos dois primeiros experimentos. É verificado também que após estes dois, os atributos de visualizações ao curso no AVA, interações externas,

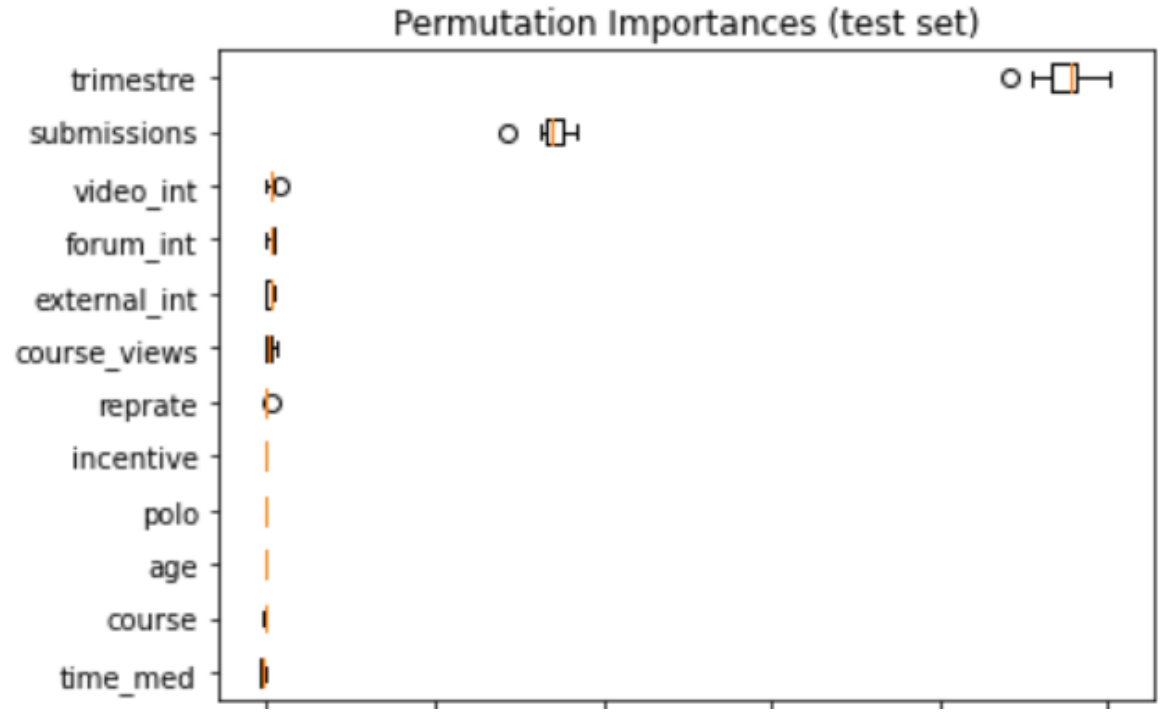
que abrangem *upload* e *download* de arquivos, taxa de reprovação e o tempo de acesso médio a plataforma são impactantes para as análises.

Gráfico 16 - Importância das permutações no Decision Tree no Experimento 1.



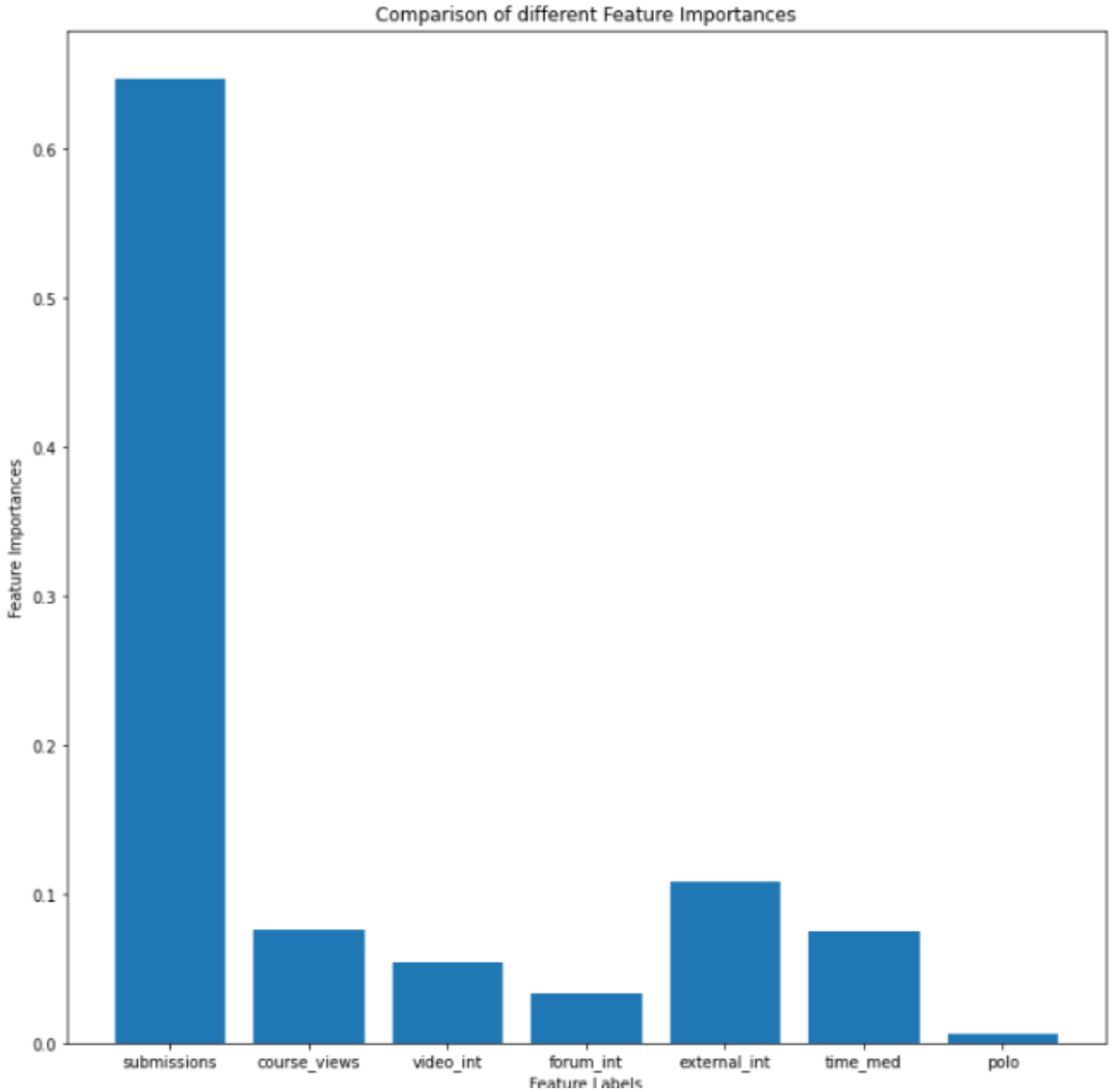
Fonte: Do autor (2020).

Gráfico 17 - Importância das permutações no NaiveBayes no Experimento 2.



Fonte: Do autor (2020).

Gráfico 18 - Importância dos atributos no Random Forest no Experimento 3.



Fonte: Do Autor (2020).

Em comparação com os trabalhos relacionados, os resultados obtidos nesse estudo se mostraram superiores em diversas das análises, onde somente o estudo realizado por Gonçalvez (2018) alcançou resultados próximos utilizando técnicas de InfoGain e CSF para o pré-processamento dos dados. Outro fator que pode ter influenciado nos resultados é o conjunto de dados, em que todos os estudos é totalmente distinto um do outro, além das bibliotecas utilizadas para a aplicação da mineração e a utilização da técnica de validação cruzada *GridSearch* que também contribuiu para o excelente desempenho dos modelos de predição

apresentados neste estudo e atualmente é exclusiva da biblioteca Scikit-Learn. Na Tabela 19 é possível ver uma comparação mais detalhada.

Tabela 19 - Comparação da acuracidade dos modelos preditivos.

Algoritmo	Gonçalvez, Silva, Cortês (2018)	Yükselürk (2014)	Rodriguez (2016)	Este trabalho
SVM	97%	Não aplicado	79,78%	97,60%
RandomForest	Não aplicado	Não aplicado	Não aplicado	98,26%
NaiveBayes	94%	73,90%	58,51%	97,60%
RN/MLP	Não aplicado	76,80%	83,63%	Não aplicado
DecisionTree	98%	79,70%	80%	98,47%
KNN	Não aplicado	87%	Não aplicado	95,65%

Fonte: Do Autor (2020).

No próximo capítulo são apresentados alguns modelos de predição sugeridos os quais podem auxiliar na tomada de decisão a partir da aplicação dos modelos preditivos nos dados da modalidade EAD.

6.3 Modelo de predição de evasão baseado em dados do AVA (1º modelo)

O modelo detalhado no Quadro 5 foi construído a partir das análises do Experimento 3 e baseado no estudo de Rodriguez (2016), fazendo uso das variáveis exclusivas dos cursos EaD da instituição e que foram extraídas do AVA e foram identificadas como as mais relevantes após o estudo estatístico. No Experimento 3, os classificadores de árvore de decisão que possuem uma estrutura de funcionamento mais simples conseguiram classificar de forma correta, a partir da amostra de teste, 109 estudantes propensos a evadir de um total de 111 classificados no algoritmo Random Tree e 113 estudantes de 122 classificados no algoritmo Decision Tree, alcançando precisões de 98,19% e 92,62%, respectivamente. Com a aplicação destes modelos, os resultados apresentados são passados para que os responsáveis pelas tomadas de decisão na parte de EaD possam iniciar o Modelo de Tomada de Decisão e Ações de Combate à Evasão.

Quadro 5 - Modelo de predição inicial de evasão em EaD.

Passos	Ação	Detalhamento	Objetivo
1	Extração dos dados para mineração de dados.	**Atributos:** - Quantidade de submissões realizadas no curso; - Quantidade de visualizações realizadas nas disciplinas do curso; - Quantidade de interações e participações nos vídeos das aulas; - Quantidade de postagens e interações nos fóruns das disciplinas; - Quantidade de interações com arquivos externos a plataforma; - Tempo de acesso médio na plataforma.	Extrair do banco de dados do AVA os dados necessários para o processo de mineração de dados.
2	Transformação e formatação dos dados	- Extrair os dados do AVA concentrando a totalidade dos valores; - Formatar os dados em formato CSV; - Adequar e transformar os dados conforme requisitos dos algoritmos utilizando as técnicas de MinMaxScaler e GridSearch.	Deixar os dados no formato adequado conforme os requisitos dos algoritmos para a mineração de dados.
3	Aplicação da Mineração de Dados.	- Classificadores RandomTree ou DecisionTree.	Obter informações para combate da evasão.

Fonte: Do Autor (2020).

6.4 Modelo para tomada de decisão (2º modelo)

Com a aplicação do modelo apresentado no Quadro 5 que foi baseado no estudo de Rodriguez (2016), são geradas informações importantes para que os responsáveis pelas tomadas de decisão no EaD, providos destas informações, possam tomar decisões e dar início em ações para mitigar a evasão. Para dar um acompanhamento a estas ações durante os semestres subsequentes, sugere-se o modelo do Quadro 6. As diferentes ações que compõem este conjunto podem conter elementos como: (*I*) se colocar à disposição dos estudantes que apresentem alguma dificuldade através das mais diversas plataformas, (*II*)

contatar os estudantes e buscar um maior empenho dos mesmos caso não acessem com frequência o AVA, (*III*) oferecer lembretes aos alunos quanto aos prazos das atividades e encontros presenciais avaliativos, (*IV*) prover ações que estimulem a participação dos estudantes nas atividades acadêmicas, e/ou (*V*) enviar informações que incentivem os estudantes de forma motivacional e que reforcem o vínculo deles com a instituição, (*VI*) provocar os professores para que estimulem os estudantes e os mantenham motivados através de conteúdos que envolvam inovações e assuntos atuais da área de atuação do curso, (*VII*) buscar integração e parcerias da Universidade com diferentes plataformas já difundidas e outras que estão crescendo no mercado com o intuito de aplicá-las dentro dos padrões do cursos da instituição, (*VIII*) premiar professores e cursos que apresentarem os menores índices de evasão.

Quadro 6 - Modelo para tomada de decisão e ações para mitigar a evasão.

Passos	Ação	Objetivo
1	Acompanhamento das matrículas nos cursos.	Registrar a quantidade de estudantes matriculados no semestre.
2	Recebimento das informações a serem geradas no 1º Modelo	Obter informações para facilitar o processo decisório e dar início as ações.
3	Análise das informações.	Avaliação das informações que forem geradas no 1º Modelo.
4	Ação 1 - Gerar convite através das diversas plataformas sociais da instituição para retorno às atividades acadêmicas.	Obter retorno dos estudantes às atividades acadêmicas.
5	Ação 2 - Participação dos professores no convite e no estímulo dos estudantes para retornarem às atividades e acessarem os sistemas.	Obter maior engajamento dos professores para apresentar inovações a fim de combater a evasão.
6	Após 7 dias, solicitar nova execução do 1º Modelo.	Obter informações atualizadas dos estudantes propensos a evadir.
7	Análise das informações.	Avaliação das informações que forem geradas no 1º Modelo.
8	Cruzar tais informações a fim de visualizar casos persistentes, novos casos e taxas totais de retorno.	Obter dados refinados.
9	Ação 3 - Contato direcionado aos estudantes, envolvendo professores das disciplinas ou coordenador do curso.	Obter retorno dos estudantes quanto às atividades acadêmicas, dificuldades e motivação.
10	Após 7 dias, solicitar nova execução do 1º Modelo.	Obter informações atualizadas dos estudantes propensos a evadir.
11	Análise das informações.	Avaliação das informações que forem geradas no 1º Modelo.

12	Cruzar tais informações a fim de visualizar casos persistentes, novos casos e taxas totais de retorno.	Obter dados refinados.
13	Ação 4 - Contato direcionado aos estudantes diretamente pelo coordenador do curso, a fim de verificar os motivos da eminente evasão e tentar a reversão.	Obter retorno dos estudantes quanto às atividades acadêmicas, dificuldades e os motivos específicos que levaram a evasão, oferecer alternativas que a instituição possui, analisar e solicitar correções no curso caso necessário.
14	Durante o desenrolar dos trimestres manter contato com os professores e melhorar os pontos detectados no que couber.	Diminuir focos que levem a evasão.

Fonte: Do Autor (2020).

Baseado na experiência deste autor, bem como considerando a interação e informações junto ao Setor de Educação da Distância da Instituição é possível lançar algumas ideias, que podem ser utilizadas como possibilidades de estratégias junto aos gestores, coordenadores, professores e tutores dos cursos EaD da Univates, com o objetivo de diminuir a evasão nos cursos:

- Ensinar habilidades de disciplina e motivação aos estudantes EaD, logo no início do curso, incentivando a gestão do tempo, uso das ferramentas disponíveis, preparação do ambiente de estudo entre outros, com isso sanando também dificuldade com a plataforma em si, como de ferramentas e recursos que venham a ser explorados durante o curso;
- Incentivar junto ao grupo de professores o uso de diferentes estratégias didáticas, visando uma participação mais ativa dos estudantes nas atividades, bem como considerar experiências de inovação nos cursos, testando práticas como ensino híbrido, gamificação, realidade aumentada, uso maior de podcast, vídeos curtos, videoconferências com maior foco na discussão e menos expositivas utilizando-se das diversas plataformas disponíveis com estes recursos no mercado e que são frequentadas em massa pelos estudantes (Spotify, Youtube, Soundcloud, Twitch, etc.) fornecendo assim o conteúdo em diferentes formatos, garantindo que todos os perfis de alunos sejam atendidos;
- Reavaliar constantemente a metodologia e a proposta pedagógica e a eficácia do método de ensino empregado, investigando o que os alunos têm a dizer sobre o currículo, se as aulas e atividades agregam conhecimento,

despertam seu interesse e sua curiosidade e ajudam no desenvolver das suas habilidades técnicas;

- Para estudantes com dificuldades acadêmicas em conteúdos básicos (Ex.: Matemática, Computação), além do envolvimento do NAP - Núcleo de Apoio Pedagógico da Instituição, investir estratégias de materiais extras, considerando a curadoria de recursos disponíveis, como a Khan Academy[18], Coursera[19] ou Udemy[20] e outras maneiras de incentivar o aprendizado desses alunos, evitando que eles fiquem desmotivados com o curso.
- Testar e avaliar outras ferramentas e recursos que possibilitem maior interação nas discussões entre professor x tutor x estudante, além dos obtidos por meio dos fóruns de discussão, disponíveis no Univates Virtual, tais como o Google Chat, Google Apps ou o Google Colaboratory como utilizado neste trabalho, que permitem uma colaboração de forma eficiente com mensagens diretas e chats em grupo protegidos.
- Maior participação e interação por parte do polo, que quando acionado, pudesse contatar o estudante, questionando sobre dificuldades e auxílio que este pudesse necessitar, também organizando encontros periódicos das turmas nos polos de apoio, estimulando a socialização, entre outros.
- Realizar projetos interdisciplinares, como forma de estimular os estudantes a ter uma compreensão do curso como um todo, obtendo também uma visão da aplicação do que se aprende dentro do contexto no qual ele vive.
- Por meio de sistemas associados à inteligência artificial mensurar indicadores e saber antecipadamente quais alunos têm maior propensão a deixar o curso e por quê. Também, por meio da IA, permitir feedbacks de atividades e exercícios, permitindo que maior tempo do tutor seja dedicado a atenção e empatia com o estudante.
- Além do AVA, outros canais podem ser utilizados. Mensagens por telefone, como o WhatsApp, podem ser usadas como forma de verificar o que atrapalha o estudante, e, dessa maneira, ajudá-lo a voltar aos estudos. A criação de grupos de Telegram para disponibilização de conteúdo para os

[18] Khan Academy - https://pt.khanacademy.org/ - Acessado em 29 Jun. 2020
[19] Coursera - https://pt.coursera.org/ - Acessado em 29 Jun. 2020
[20] Udemy - https://www.udemy.com/ - Acessado em 29 Jun. 2020

estudantes também é uma sugestão. Redes sociais também podem funcionar para promover a sensação de pertencimento ao curso. Ao mesmo tempo, manter uma presença marcante no Facebook, Instagram ou outra rede utilizada entre os estudantes, com publicações que promovam a interação e o compartilhamento de conteúdo, podem contribuir na manutenção dos estudantes e inclusive na ingressão de novos alunos.

No próximo capítulo são apresentadas as considerações finais após o desenvolvimento do trabalho, oferecendo sugestões para trabalhos futuros que visem aperfeiçoar a predição da evasão utilizando a mineração de dados.

7 CONCLUSÕES

Através da pesquisa realizada e dos resultados apresentados, fica comprovado que a aplicação de algoritmos de aprendizado de máquina para a predição de evasão de alunos na educação a distância é muito recomendada, pois os algoritmos conseguem de forma automatizada rapidamente entender o cenário e retornar resultados muito satisfatórios que são praticamente impossíveis de visualizar a olho nu. As técnicas de classificação podem oferecer valiosos conhecimentos aos gestores e responsáveis pelo processo de ensino-aprendizagem e ajudar na tomada de decisão e a compreender melhor as tendências e o comportamento dos diferentes perfis de alunos que acabam por evadir de algum curso ou disciplina, seja qual for o motivo.

No estudo foram aplicados algoritmos de classificação distintos em três conjuntos de dados semelhantes, mas que tinham seu volume de dados diminuído respectivamente entre os três experimentos para melhorar o desempenho. Apesar disso, desde o primeiro experimento os algoritmos foram capazes de alcançar resultados ótimos, mostrando que é possível se aplicar algoritmos de mineração de dados tanto nos dados do Sistema de Gerenciamento Interno da instituição quanto nos dados extraídos do AVA e produzir grandes conhecimentos. Os algoritmos se mostraram levemente mais eficientes na classificação dos alunos que não evadiram, muito provavelmente pelo motivo do volume destes ser maior. Apesar disso, a classificação dos alunos evadidos também mostrou um resultado excelente.

Os atributos mais influentes para a classificação do estudante que tende ou não a evadir foram, a quantidade de submissões que o mesmo fez durante o período analisado, o trimestre de trancamento do mesmo, mostrando um certo padrão de época em que os estudantes tendem a evadir, assim como outras interações dentro da plataforma e o tempo de acesso médio em que o estudante frequenta a plataforma.

Ao se referir aos trabalhos relacionados que foram tomados como base para o desenvolvimento desta pesquisa, ficou constatado que dentro da área de mineração de dados educacional, sejam quais forem os conjuntos de dados analisados, os algoritmos de árvore de decisão possuem um desempenho excelente, assim como as máquinas de vetores de suporte, que performam ainda melhor quando há um volume de dados maior.

Já em relação as ferramentas utilizadas para a aplicação das técnicas de classificação e mineração de dados, a biblioteca do Scikit-Learn, que trabalha com a linguagem Python, se mostrou apropriada e eficiente por contar com diversos recursos para as diferentes etapas do processo da mineração de dados, partindo do pré-processamento até a aplicação dos algoritmos de classificação e avaliação dos resultados gerados. Essa biblioteca possui uma fácil integração com diversas outras bibliotecas para manipulação dos dados, como a Pandas e a NumPy, além de contar com integração com a matplotlib para exibição de gráficos e visualização dos dados. Juntando as bibliotecas, podemos executar as mesmas na plataforma de desenvolvimento compartilhado Google Colab, que conta com diversos recursos para melhor o desempenho e facilitar o desenvolvimento, como por exemplo execução diretamente na GPU e busca dos erros direto na página StackOverflow[21].

Dentro dos algoritmos utilizados para a análise, foi concluído que os algoritmos Random Forest e Decision Tree que são dois classificadores que usam o conceito de árvore de decisão apresentaram os melhores resultados com porcentagens de 98,26% e 98,47% de acuracidade nos melhores resultados e

[21] StackOverflow - https://pt.stackoverflow.com/ - Acessado em 27 Jun. 2020.

valores de 97,39% e 96,73% de acurácia no experimento com dados somente do ambiente virtual. Apesar da alta acuracidade, os valores de revocação, que representam a eficácia na real predição dos alunos evadidos, ficaram levemente menores, em 91,60% e 94,96% respectivamente no experimento 3, porém ainda assim excelentes resultados.

A partir da predição a evasão e das informações adquiridas através dos modelos preditivos, os gestores de curso, sendo apoiados ou não por outros setores da instituição, podem ser articuladas medidas envolvendo os estudantes que se enquadram no grupo de risco de evasão, intervindo, com o objetivo de diminuir seus índices.

Considerando os resultados obtidos, acredita-se que o processo de aplicação das técnicas de mineração de dados possa ser focalizado em diferentes turmas com diferentes atributos e também, possam ser estendidas aos demais cursos da instituição ou até mesmo de outras instituições onde esse problema de evasão é observado, sendo flexível ajustá-lo às especificidades de diferentes frentes.

Esta pesquisa que buscou detectar de forma preditiva os estudantes com tendências a evasão, através da aplicação de técnicas de mineração de dados, mostrou-se eficaz dentro do conjunto de dados analisado, obtendo boa precisão, revocação e acuracidade conforme os resultados apresentados anteriormente. Apesar disso, este pode ser tomado somente como um ponto de partida para futuros estudos e aplicações que possam ser desenvolvidas tais como se sugere na seção a seguir.

7.1 Trabalhos futuros

Para aperfeiçoar ainda mais o desempenho dos modelos preditivos, seja em futuros estudos ou aplicações na prática, são propostas algumas sugestões:

- Realizar uma minuciosa análise nas bases de dados dos diferentes sistemas em funcionamento dentro da instituição, a fim de aperfeiçoar o registro dos dados, removendo ruídos e dados inconsistentes;

- Treinar os modelos de predição com um maior número de resultados de diferentes períodos em que o ensino EaD será aplicado;

- Complementar o conjunto de dados com dados de notas de alunos e situação financeira;

- Implementar os modelos de predição em outros cursos e verificar os resultados alcançados;

- Realizar experimentos com dados gerados em um intervalo menor de tempo, por trimestre ou mensalmente, de forma a verificar o mínimo de dias necessários para se detectar um estudante propenso a evadir;

- Experimentar algoritmos de redes neurais e verificar como eles se comportam nas análises dos conjuntos de dados;

- Com auxílio de especialistas das diferentes áreas de aprendizado dentro dos cursos EaD, buscar variáveis distintas nas diferentes ementas de disciplinas e cursos que venham a influenciar na forma como se comportam os alunos;

- Analisar e desenvolver uma interface amigável para a aplicação do modelo de predição, oferecendo aos gestores do ensino uma forma de acompanhar de forma sistemática os estudantes com tendência a evadir e, assim, gerenciar as medidas tomadas para a prevenção da evasão.

REFERÊNCIAS

ALVES, J. R. M. Educação à Distância e as Novas Tecnologias de Informação e Aprendizagem. **Novas Tecnologias na Educação**. 2015. Disponível em: <http://www.clam.org.br/bibliotecadigital/uploads/publicacoes/186_1700_alvesjoaoroberto.pdf>. Acesso em: 11 set. 2019.

ARTERO, Almir Olivette. **Inteligência Artificial Teórica e Prática**. 1º Edição. São Paulo. Editoria Livraria da Física, 2008.

BALTAR, P. C.; SILVA, S. S. **Um olhar acerca da evasão na educação a distância**. Revista Uniabeu, 10, 4, 61-73, 2017.

BAKER, R. S. J. D.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **JEDM-Journal of Educational Data Mining**, v. 1, n. 1, p. 3-17, 2009.

BAKER, R. S. J. D. et al. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, vol. 19, no. 2, p. 2-13, 2011.

BARBOSA, W.; MÁXIMO, D.; JATOBÁ, A.; LEITE, A.; SOARES, E. Uma Proposta para Identificação de Causas da Evasão na Educação a Distância através de Mineração de Dados. **Anais da ERBASE-Escola de Computação Bahia-Alagoas-Sergipe**, 2014. Disponível em: <http://erbase2014.uefs.br/artigos/125801.pdf>. Acesso em: 10 out. 2019.

BRAGA, A. d. P; CARVALHO, A; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. Livros Técnicos e Científicos, 2000.

BRASIL. Decreto n° 5.622, de 19 de dezembro de 2005. 2005. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5622.htm.>. Acesso em: 15 set. 2019.

CAMILO, C. O; SILVA, J. C. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de administração pública**, v. 42, n. 3, p. 495-528, 2008.

CASTRO de, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. 1a Edição. São Paulo. Saraiva, 2016.

Censo EAD.BR. **Relatório analítico da aprendizagem a distância no Brasil.** São Paulo: ABDR Education do Brasil, 2013, 2014, 2015, 2016, 2017. Disponível em: <http://www.abed.org.br/site/pt/midiateca/censo_ead/1554/2018/10/censoeadbr_-_2017/2018>. Acesso em: 24 ago. 2019.

COELHO, Maria L. **A Evasão nos Cursos de Formação Continuada de Professores Universitários na Modalidade de Educação a Distância Via Internet**. Em: CONGRESSO INTERNACIONAL DE EDUCAÇÃO A DISTÂNCIA, e. 8, 2001, Brasília. Disponível em: <http://www.abed.org.br/site/pt/midiateca/textos_ead/626/2004/12/a_evasao_nos_cursos_de_formacao_continuada_de_professores_universitarios_na_modalidade_de_educacao_a_distancia_via_internet_> Acesso em: 05 out. 2019.

COPPIN, Ben. **Inteligência Artificial**. 1º Edição. Rio de Janeiro. Livros Técnicos e Científicos Editora Ltda, 2013.

COSTA, S. S. da; CAZELLA, S; RIGO, S. J. Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. **RENOTE**, v. 12, n. 2, 2012.

COURNAPEAU, D. **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 2011. Disponível em: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf> Acesso em: 15 out. 2019.

DAUDT, S. I. D.; BEHAR, P. A. **A gestão de cursos de graduação a distância e o fenômeno da evasão**. Educação, v. 36, n. 3, p. 412-421, 2013.

DIAS, A. A. S. E-Learning para E-formadores. **Guimarães: TecMinho / Gabinete de Formação Contínua.** Universidade do Minho, 2008. Disponível em: <https://repositorium.sdum.uminho.pt/bitstream/1822/8723/3/dos%20lms%20aos%20objectos.pdf> Acesso em: 10 out. 2019.

DORE, R.; LÜSCHER, A. Z. **Permanência e Evasão na Educação Técnica de Nível Médio em Minas Gerais**. Cadernos de Pesquisa**,** v. 41, n. 144, p. 770-789, 2011.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. John Wiley & Sons, 2012. Disponível em: <https://www.researchgate.net/publication/228058014_Pattern_Classification> Acesso em: 25 set. 2019.

ESPÍNDOLA, R.M.; LACERDA, F.K.D. **Evasão na Educação a Distância: um estudo de caso**. Revista EAD em foco. Fundação CECIERJ, v. 3, n. 1, dezembro de 2013, p. 96-108, 2013.

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

FERNANDES, J. et al. **Identificação de Fatores que Influenciam na Evasão em um Curso Superior de Ensino a Distância**. Perspectivas OnLine 2007-2010, v. 4, n. 16, 2014.

GAMA, J.; MEDAS, P. **Learning Decision Trees from Dynamic Data Streams**. Journal of Universal Computer Science, vol. 11, no. 8, 2005.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data preprocessing in data mining**. Springer, 2015.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: Um Guia Prático-Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. Rio de Janeiro: Campus, v. 1, 2005.

GONÇALVES, T.; SILVA, J; CORTES, O. **Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão**. Revista Brasileira de Computação Aplicada**,** vol. 10, n. 3, p. 11-20, 2018.

GONZALEZ, M. **Fundamentos da tutoria em educação a distância.** Avercamp, 2005.

GOTTARDO, E. et al. **Previsão de Desempenho de Estudantes em Cursos EAD utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais**. Em: Anais do Simpósio Brasileiro de Informática na Educação, 2012. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/view/1758>. Acesso em: 27 out. 2019.

HAN, J.; PEI J.; KAMBER, M. **Data Mining Concepts and Techniques**. Morgan Kaufmann Publishers. USA, 2001.

HAYKIN, S. **Neural networks a comprehensive introduction**. Prentice Hall, New Jersey, 1999.

IARALHAM, L. C. **Contribuição da Tecnologia da Informação na Educação a Distância no Instituto Universal Brasileiro: um Estudo de Caso**. Revista Científica da Faculdade das Américas**,** ano 3, p 1-9, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Sinopse Estatística da Educação Superior.** Brasília, DF: Ministério da Educação 2014, 2015, 2016, 2017, 2018. Disponível em: <http://inep.gov.br/sinopses-estatisticas-da-educacao-superior>. Acesso em: 24 ago. 2019.

JOSÉ, I. **KNN (K-Nearest Neighbors).** 2018. Disponível em: <https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e>. Acesso em: 15 out. 2019.

LACHI, R. L. et al. Uso de agentes de interface para auxiliar a avaliação formativa no ambiente TelEduc. **XIII Simpósio Brasileiro de Informática na Educação.** São Leopoldo-RS, novembro, p. 2-9, 2002.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos metodologia científica**. 4.ed. São Paulo: Atlas, 2001.

LATTARO, A. **Redes neurais artificiais: o que são? Onde vivem? Do que se alimentam?**. 2017. Disponível em: <https://imasters.com.br/devsecops/redes-neurais-artificiais-o-que-sao-onde-vivem-do-que-se-alimentam>. Acesso em: 20 out. 2019.

MAIA, M. de C et al. Análise dos índices de evasão nos cursos superiores a distância do Brasil. **Anais do XI Congresso Internacional de Educação à Distância.** Salvador, Bahia. 2004.

MÁRQUEZ-VERA, C.; CANO, A.; ROMERO, C.; VENTURA, S. **Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data**. Appl. Intell. v. 38, n. 3, p. 315-330, 2012.

MARTINS, C. B. N. **Evasão de alunos nos cursos de graduação em uma instituição de ensino superior.** Montes Claros, 2007. Disponível em: <https://fpl.edu.br/2018/media/pdfs/mestrado/dissertacoes_2007/dissertacao_cleidis_beatriz_nogueira_martins_2007.pdf>. Acesso em: 11 set. 2019.

MATTAR, F. N. **Pesquisa de marketing**. 3.ed. São Paulo: Atlas, 2001.

MAY, Tim. **Pesquisa social: questões, métodos e processos**. Porto Alegre: Artmed, 2004.

MCKINNEY, Wes. **Python for Data Analysis**. 1ª edição. Sebastopol: O'Reilly Media, 2012.

MERSCHMANN, L. de C. **Classificação probabilística baseada em análise de padrões.** Tese (Doutorado), UFF - Universidade Federal Fluminense, Brasil, 2007. Disponível em: <http://www.ic.uff.br/PosGraduacao/frontend-tesesdissertacoes/download.php?id=357.pdf&tipo=trabalho>. Acesso em: 11 set. 2019.

MITCHELL, T. M. **Machine Learning.** McGraw-Hill, Inc., New York, NY, 1997.

MORAN, J. M. **Contribuições para uma pedagogia da educação on-line. Educação online: teorias, práticas, legislação, formação corporativa**. São Paulo: Loyola, 2003. p. 39-50.

NUNES, I. B. **A história da EaD no mundo**. Educação a Distância. O estado da arte. São Paulo: Pearson Education do Brasil, 2009, p. 2-8.

OLIVEIRA, A. E. F; FRANÇA R. M. **Avaliação da Produtividade no Processamento dos Dados em um Curso de Pós-graduação Lato Sensu da UNASUS/UFMA na Modalidade a Distância utilizando o Sistema de Monitoramento (SIM).** 2014. Disponível em: <http://www.telessaude.uerj.br/resource/goldbook/pdf/27.pdf>. Acesso em 15 set. 2019.

ONODA, M.; EBECKEN, N. **Implementação em Java de um Algoritmo de Árvore de Decisão Acoplado a um SGBD Relacional**. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2001. Disponível em: <https://www.researchgate.net/publication/221535942_Implementacao_em_Java_de_um_Algoritmo_de_Arvore_de_Decisao_Acoplado_a_um_SGBD_Relacional>. Acesso em: 17 set. 2019.

PINHEIRO, M. F. et al. **Identificação de Grupos de Estudantes em Ambiente Virtual de Aprendizagem: Uma Estratégia de Análise de Log Baseada em Clusterização.** Em: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2014. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/3282>. Acesso em 27 out. 2019.

RABELO, H.; BURLAMAQUI, A. M. F.; VALENTIM, R. A. M.; DANIELI S. S. R. **Uso de Técnicas de Mineração de Dados Educacionais para Prever o Desempenho do Professor de EAD em Ambientes Virtuais de Aprendizagem**. VI Congresso Brasileiro de Informática na Educação**,** 2017. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/view/7684>. Acesso em: 25 out. 2019.

RIBEIRO, E. N. et al. **A importância dos ambientes virtuais de aprendizagem na busca de novos domínios da EAD**. Em: Congresso da Associação Brasileira da Educação a Distância, Goiás, 2007. Disponível em: <http://www.abed.org.br/congresso2007/tc/4162007104526am.pdf>. Acesso em: 25 out. 2019.

RIBEIRO, M. A. **O projeto profissional familiar como determinante da evasão universitária: um estudo preliminar**. Revista Brasileira de Orientação Profissional, v. 6, n. 2, p. 55-70, 2005. Disponível em: <http://pepsic.bvsalud.org/pdf/rbop/v6n2/v6n2a06.pdf>. Acesso em: 26 out. 2019.

ROBERTS, L. E. **Not now, maybe later, and often not at all: situational, institutional, dispositional, epistemological, and technological barriers to business-based online training.** 2004. Graduate Faculty of North Carolina State University, Raleigh, 2004. Disponível em: <https://www.semanticscholar.org/paper/Not-now%2C-maybe-later%2C-and-often-not-at-all%3A-and-to-Roberts/8c216c0a0ca269df5774f09a402c6d2c6e3c365d>. Acesso em: 24 out. 2019.

RODRIGUES W. S. **Predição de Evasão na Educação a Distância como Subsídio à Tomada de Decisão.** Universidade Católica de Brasília, 2016. Disponível em: <https://bdtd.ucb.br:8443/jspui/handle/tede/2318>. Acesso em: 10 set. 2019.

ROMERO, C. et al. **Knowledge discovery with genetic programming for providing feedback to courseware authors**. User Modeling and User-Adapted Interaction**,** v. 14, n. 5, p. 425-464, 2004.

ROMERO, C; VENTURA, S. **Data Mining in Education**. WIREs Data Mining and Knowledge Discovery, v. 3, p. 12-27, 2013.

RUSSELL, S. J; NORVIG, P. **Artificial Intelligence: A modern approach.** Prentice-Hall, Inc., Upper Saddle River, NJ, 1995.

SABBATINI, R. M. E. **Ambiente de Ensino e Aprendizagem via internet: a plataforma moodle**. Campinas: Instituto Edumed, 2007. Disponível em: <https://www.researchgate.net/publication/260385940_Ambiente_de_Ensino_e_Aprendizagem_via_Internet_A_Plataforma_Moodle>. Acesso em: 15 set. 2019.

SANTOS, A. P. **A predição da evasão de estudantes de graduação como recursos de apoio fornecido por um assistente inteligente.** 2014. 64f. Dissertação (Mestrado em Gestão do Conhecimento e Tecnologia da Informação) - Universidade Católica de Brasília, Brasília, 2014. Disponível em: <https://www.aedb.br/seget/arquivos/artigos13/52618669.pdf>. Acesso em: 20 set. 2019.

SELLTIZ, C.; WRIGHTSMAN, L. S.; COOK, S. W. **Métodos de pesquisa das relações sociais**. São Paulo: Herder, 1965.

SILVA FILHO, R. L. L. et al. **A evasão no ensino superior brasileiro**. Cadernos de Pesquisa**,** v. 37, n. 132, p. 641-659, 2007.

STEINER, Maria Teresinha Arns; SOMA, Nei Yoshihiro; SHIMIZU, Tamio; NIEVOLA, Julio Cesar; STEINER NETO, Pedro José. **The Influence of**

Multivariate Data Analysis on the KDD Process: An Application to Medical Diagnosis. International Journal of Operations and Quantitative Management, v. 12, p. 73-83, 2006.

TAN, Pang – Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Datamining: mineração de dados**. 1 a Edição. Rio de Janeiro: Editora Ciência Moderna, 2009.

TAYLOR, J. **Fifth generation distance education**. Em: 20th ICDE WORLD CONFERENCE ON OPEN LEARNING AND DISTANCE EDUCATION, 2001. Disponível em: <https://www.researchgate.net/publication/246182977_5th_Generation_Distance_Education>. Acesso em: 20 set. 2019.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition, academic press**. New York, 1999.

THEODORIDIS, S.; KOUTROUMBAS, K. **Clustering: basic concepts**. Pattern Recognition, p. 483–516, 2006.

THOMAS, Jaya. **Overview of integrative analysis methods for heterogeneous data.** 2015. Disponível em: <https://www.researchgate.net/publication/282282568_Overview_of_integrative_analysis_methods_for_heterogeneous_data>. Acesso em: 27 out. 2019

WITTEN, Ian H.; FRANK, Eibe. **Data Mining Pratical Machine Learning Tools and Techniques**. 2ª Edition. Elsevier 2005.

YÜKSELTÜRK, E.; ÖZEKEŞ, S.; TÜREL, Y. K. **Predicting Dropout Student: An Application of DataMining Methods in an Online Education Program**. European Journal of Open, Distance and E-Learning – EURODL, 17(1), 118-133, 2014. Disponível em: <https://content.sciendo.com/view/journals/eurodl/17/1/article-p118.xml>. Acesso em: 20 out. 2019.

VOSS, C.; TSIKRIKTSIS, N.; FROHLICH, M. **Case research in operations management**. International Journal of Operations & Production Management, v.22, n.2, p.195-219, 2002.