Uma Abordagem Baseada em Ranking para a Predição de Evasão Escolar: Um Estudo de Caso na Universidade Federal de Mato Grosso do Sul

Gregório Takashi Higashikawa

Gregório Takashi Higashikawa

Orientador: *Profº Dr. Rafael Geraldeli Rossi*
Coorientador: *Profº Dr. Luciano Gonda*

UFMS - Campo Grande
agosto/2022

Resumo

A evasão acadêmica é um problema discutido em várias instituições de ensino, pois acarreta um dispêndio para a instituição e um grande prejuízo educacional para o aluno. Além disso, causas definidas por meio de percepções humanas podem estar incorretas e consequentemente levar à ações preventivas incorretas. Por outro lado, observa-se o emprego com sucesso de técnicas de mineração de dados em diversas instituições de ensino no Brasil e no mundo para a extração automática de hipóteses, e a geração de modelos para classificar se um aluno tem propensão para evadir ou não. O uso da classificação neste cenário pode gerar resultados não satisfatórios, principalmente quando a confiança de classificação entre as classes evasão e não evasão são próximas. Outro ponto a considerar é quando existe um desbalanceamento entre as classes, conhecido por gerar modelos de classificação inacurado na classe minoritária. Ademais, as causas de evasão podem variar de acordo com a localidade e curso. Como alternativa, classificar instâncias de testes baseada em sua propensão à pertencer à uma determinada classe significativa e ranqueando-as, direcionaria esforço de forma mais precisa aos alunos com maior chance de evasão. Dado isso, este trabalho apresenta um método para detecção de evasão baseado em *ranking* e tem como conjunto de treinamento os dados dos históricos escolares dos alunos. É então avaliado se a evasão ocorre nos *top* alunos ranqueados com potencial de evasão. Ademais, o método foi empregado em diferentes cursos e diferentes áreas do conhecimento da Universidade Federal de Mato Grosso do Sul (UFMS) para avaliar o seu comportamento em diferentes cenários. Os resultados mostram uma precisão de $0,97$, $0,99$ e $0,98$ para os 50 primeiros alunos ranqueados das áreas de ciências Exatas, Biológicas e Humanas. E para os 10 primeiros a precisão foi de $0,99$ à $1,0$. Por outro lado, com o uso de modelos de classificação, os resultados obtidos foram de $0,72$, $0,79$, $0,78$ na acurácia e $0,68$, $0,65$, $0,64$ para a precisão. Vale ressaltar que apesar da avaliação do *ranking* englobar apenas os *top-k* alunos da classe de evasão, a assertividade na classificação faz com

que os esforços sejam mais bem direcionados e esperançosamente mais efetivos. Com isso, dada a previsão assertiva do método proposto para os alunos com maior potencial de evasão, os interessados poderão aplicar estratégias mais eficientes no contensão à evasão.

Palavras-Chave: algoritmos de classificação, evasão escolar, mineração de dados, predição de evasão, *ranking*.

Abstract

Academic dropout is a problem discussed in several educational institutions, as it entails an expense for the institution and a significant educational loss for the student. In addition, causes defined by human perceptions can be incorrect and consequently lead to wrong preventive actions. On the other hand, the successful use of data mining techniques is observed in several educational institutions in Brazil and in the world for the automatic extraction of hypotheses and the generation of models to classify whether a student has a propensity to drop out or not. The use of classification in this scenario can generate unsatisfactory results, especially when the classification confidence between the evasion and non-evasion classes are close. Another point to consider is when there is an imbalance between classes, known to generate inaccurate classification models in the minority class. In addition, the causes of dropout may vary by location and course. Alternatively, ranking test instances based on their propensity to belong to a certain meaningful class and ranking them would more precisely direct effort to students most likely to drop out. Given that, this work presents a method for detecting dropout based on ranking and having as a training set the data from students' school records. It is then evaluated whether dropout occurs in the top ranked students with dropout potential. Furthermore, the method was used in different courses and different areas of knowledge at the Federal University of Mato Grosso do Sul (UFMS) to evaluate its behavior in different scenarios. The results show an accuracy of 0.94, 0.98 and 0.96 for the first 50 ranked students in the exact, biological and humanities areas. And for the first 10 the precision was from 0.99 to 1.0. On the other hand, with the use of classification models, the results obtained were 0.72, 0.79, 0.78 for accuracy and 0.68, 0.65, 0.64 for precision. It is worth mentioning that despite the ranking evaluation only encompassing the top-k students in the dropout class, the assertiveness in the classification makes the efforts better targeted and hopefully more effective. Therefore, given the assertive prediction of the proposed method for students with greater dro-

pout potential, those interested will be able to apply more efficient strategies to combat dropout.

Key-Words: classification algorithms, data mining, evasion prediction, ranking, school dropout.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AA Análise de Aprendizado

Agetic Agência de Tecnologia da Informação e Comunicação

AD Árvore de Decisão

AM Aprendizado de Máquina

ANN *Artificial Neural Network*

AUC-ROC *Area Under the Curve - Receiver Operating Characteristic*

AUV *Area Under the Curve*

AVA Ambiente Virtual de Aprendizagem

CAAE Certificado de Apresentação de Apreciação Ética

CEP Comitê de Ética em Pesquisa com Seres Humanos

CONEP Comissão Nacional de Ética em Pesquisa

CPTL Campus de Três Lagoas

CPCG Cidade Universitária – Campo Grande

CRA Coeficiente de Rendimento Acadêmico

CRISP-DM *Cross-Industry Standard Process for Data Mining*

DT Decision Tree

EAD Educação a Distância

EDM *Educational Data Mining*

EDRM Modelo de Referência de Dados Educacionais

FN Falso Negativo

Forplad Fórum Nacional de Pró-Reitores de Planejamento e Administração

FP Falso Positivo

HTTP Hypertext Transfer Protocol

IFES Instituto Federal do Espírito Santo

JSON JavaScript Object Notation

KDD *Knowledge Discovery in Databases*

LA *Learning Analytics*

LGPD Lei Geral de Proteção de Dados

LR Logistic Regression

MDE Mineração de Dados Educacionais

MGA Média Geral Acadêmica

MLP *Multilayer Perceptron*

NCR *National Cash Register Corporation*

NB *Naive Bayes*

PASSE Programa de Avaliação Seriada Seletiva da UFMS

PESFARM *Probabilistic Ensemble Simplified Fuzzy Adaptive Resonance Theory*

Proaes Pró-Reitoria de Assuntos Estudantis

Prograd Pró-Reitoria de Graduação

Proplan Pró-reitoria de Planejamento

RGA Registro Geral do Acadêmico

Reuni Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

SISU Seleção Unificada

Siscad Sistema Acadêmico

SPSS *Statistical Package for the Social Scien- ces*

SVM *Support Vector Machines*

TSG Taxa de Sucesso na Graduação

UFMS Universidade Federal de Mato Grosso do Sul

UFPB Universidade Federal da Paraíba

VN Verdadeiro Negativo

VP Verdadeiro Positivo

WEKA *Waikato Environment for Knowledge Analysis*

CAPÍTULO

1

Introdução

Nos últimos anos, o governo brasileiro proporcionou um aumento significativo no número de vagas nas universidades. Através da Seleção Unificada (SISU), as universidades se beneficiaram de um aumento expressivo de ingressos e mobilidade acadêmica, por outro lado houve uma redução na Taxa de Sucesso na Graduação (TSG) após a adesão de universidades ao Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (Reuni) (BRASIL, 2014). De acordo com a Secretaria de Educação Superior, o número de alunos em tempo integral, no período de 2009 a 2018, teve um aumento de $49,6\%$, e a Taxa de Sucesso na Graduação (TSG) em 2009 foi de $57,05\%$ para $43,71\%$ em 2018 (BRASIL, 2019). Pode-se observar que com o aumento da oferta também houve aumento no número de evasões e alterações nos perfis desses alunos (BRASIL, 2019, 2014; Gilioli, 2016).

No momento em que o aluno decide evadir-se do ambiente educacional, todo o investimento financeiro aplicado pela instituição, pelos pais ou qualquer outra fonte se torna falho, uma vez que o objetivo dos investimentos educacionais é concretizar o estudo do aluno, sendo uma realidade presente não só no Brasil como no mundo Lamb (2011). Durante o Fórum Nacional de Pró-Reitores de Planejamento e Administração (Forplad), foi criada uma comissão de planejamento a fim de apurar o custo das universidades federais no Brasil BRASIL (2018b). De acordo com essa comissão, em 2009, o governo teve um gasto de R$ $24.572,50$ (vinte e quatro mil e quinhentos e setenta e dois reais e cinquenta centavos) por aluno, e em 2016 o valor gasto subiu para R$ $37.551,20$ (trinta e sete mil quinhentos e cinquenta e um reais e vinte centavos). A cada ano, tem-se um aumento na média de dois mil reais no custo para se

manter um aluno[1].

A evasão no ensino superior pode ocorrer por diferentes motivos. A seguir, serão apresentados alguns trabalhos encontrados na literatura que buscaram inferir estas diferentes motivações que levam o aluno à evasão, após a realização de estudos sobre o tema em diferentes instituições. Em Cunha et al. (2001) são apresentados tipos de causas deduzidas em um estudo realizado no curso de Química da Universidade de Brasília:

- Deficiências acumuladas do ensino básico, público ou privado.
- O desamparo sentido na chegada ao curso.
- Frustração das expectativas sobre o curso.
- Despreparo para lidar com as diferenças entre o segundo grau e o sistema universitário.

Já em Silva e Franco (2014) são apresentados os motivos citados pelos alunos do curso de Física da Universidade Estadual de Maringá, que desmotivam a continuidade do curso:

- A desvalorização do profissional.
- Base matemática frágil.

Por fim, em Hoed (2016), os autores atribuem como características dos alunos evadidos da área da computação:

- Baixa qualidade das aulas.
- Critérios de avaliações rígidos.
- Dificuldades em disciplinas com base matemática.
- Dificuldades em disciplinas relacionadas a algoritmos e programação.

Como pode-se observar, há vários motivos que podem causar a evasão de modo diferente entre as áreas (Exatas, Humanas, Biológicas, Saúde), cursos, modalidades (presencial, semipresencial, Educação a Distância - EaD). Características individuais como, idade, renda familiar, escolaridade dos pais, entre várias outras também são atributos que podem torná-lo mais suscetível à evasão (Marques, 2020). Logo, a tarefa de caracterização dos alunos evadidos torna-se mais trabalhosa e passível de erro por meio da formulação de hipóteses subjetivas ou ainda gerais, ou seja, a mesma causa para qualquer tipo de curso.

[1] Gastos do governo na educação podem ser consultados pelo portal da transparência (BRASIL, 2018a).

Porém, indicar as causas da evasão de maneira subjetiva, sem auxílio de ferramentas ou métodos específicos, se torna passível de erro e pode levar a ações errôneas, infrutíferas e dispendiosas. Estes equívocos podem causar prejuízos financeiros à instituição e a ineficácia na resolução do problema. Por este motivo, a necessidade do suporte aos dados é fundamental, não só para validar hipóteses elencadas, bem como para automaticamente inferir as hipóteses (Baker e Yacef, 2009; Rodrigues et al., 2013; Nascimento et al., 2018).

Para a identificação da evasão escolar, vários artigos relatam o uso de técnicas de mineração de dados (Lykourentzou et al., 2009; Márquez-Vera et al., 2013; Berens et al., 2019). A mineração de dados é o processo de obter informações e padrões automaticamente de um grande conjunto de dados, sendo necessário um pré-processamento de forma a transformar os dados brutos em um formato apropriado (Han et al., 2011; Aggarwal, 2015; Tan et al., 2019). A mineração de dados tem sido aplicada em várias áreas de conhecimento, principalmente na área relacionadas ao marketing e a sistemas de suporte à tomada de decisão (Tan et al., 2019). Já na área educacional houve um elevado interesse na utilização de técnicas de mineração de dados nos últimos anos (Lemay et al., 2021). Assim, surgiu uma nova área de pesquisa chamada de "Mineração de Dados Educacionais", do inglês, *"Educational Data Mining"* (EDM). Como referência pode-se citar a Sociedade Científica da EDM (*International Educational Data Mining Society*).

As principais aplicações da EDM são (Costa et al., 2013; Baker e Yacef, 2009):

- Modelagem do estudante: representa informações elencadas das características do aluno;
- Modelagem do domínio: modelos da estrutura de conhecimento de um domínio;
- Suporte pedagógico;
- Descoberta científica;
- Modelagem do processo de compreensão de aprendizagem;

Em Costa et al. (2013), os autores discorrem sobre essa área e demonstram várias técnicas para a EDM, entre eles a Árvore de Decisão (*Decision Tree*), Máquina de Vetores de Suporte (*Support Vector Machine*), Regressão Linear (*Linear Regression*), Algoritmo k-*Means*, Algoritmo Genético e Regras de Associação.

Os algoritmos utilizados nas técnicas de EDM necessitam dos atributos dos alunos, podendo ser atributos variáveis com o tempo, como disciplinas cursadas, notas e frequência da disciplina. Também pode-se utilizar os atributos

invariantes com o tempo, como gênero, estado de origem, e fonte de ingresso na universidade. Em Alvarez et al. (2020); Brito et al. (2014); Antonio e Luiz (2020); Souza (2020); Díaz et al. (2021), são apresentados trabalhos relacionados à identificação da evasão escolar utilizando técnicas de Mineração de Dados, aplicados nos dados socioeconômicos e acadêmicos. Apesar dos trabalhos possuírem o mesmo objetivo, utilizam diferentes técnicas e atributos para alcançá-lo.

O Brasil tem se destacado com inúmeras publicações em artigos relacionados à problemas de evasão escolar utilizando técnicas em mineração de dados educacionais, evidenciados nos trabalhos de Manhães et al. (2012b); Gottardo et al. (2012); Paz e Cazella (2017); Vasconcelos et al. (2018); Júnior et al. (2019); Jesus et al. (2021); Lenon et al. (2021). Porém, vale ressaltar que nesses trabalhos o problema da evasão é tratado como um problema de classificação, ou seja, um problema de classificação binário ("evade" ou "não evade"). Em muitos casos, os resultados das métricas trazem valores baixos devido à uma geração de erros ao classificar exemplos de baixa confiança ou devido ao problema do desbalanceamento de classes, ou seja, quado há muita informação da classe mais incidente e baixa informação da classe minoritária.

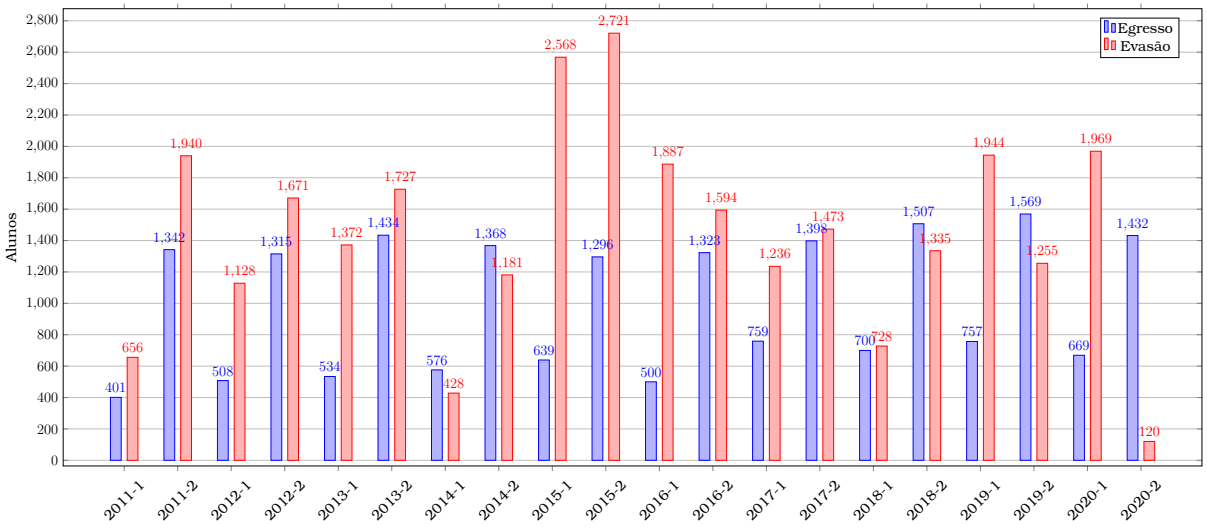
Dado isso, esse trabalho propõe a predição de evasão como uma tarefa de *ranking*, isto é, ao invés de apenas rotular todos os alunos a serem classificados como potencial de evasão ou não, os alunos são ranqueados de acordo com a confiança para a classe "evade" e são analisadas as assertividades dos modelos nos *top-*k alunos do *ranking*. O uso do *ranking* torna menos crítico o impacto do desbalanceamento de classes, além de evitar o equívoco na classificação ao considerar alunos cujas confianças de classificação são similares.

Em síntese, a motivação para o desenvolvimento deste trabalho consiste em produzir uma ferramenta que possa ser utilizada como subsídio para as políticas institucionais da UFMS, que tenham como objetivo minimizar a evasão dos estudantes. Tal ferramenta será capaz de ranquear os alunos mais propensos à evasão, utilizando apenas as notas finais das disciplinas cursadas pelos alunos ao final de cada semestre, pois visa avaliar o impacto no desempenho de cada discente nas disciplinas para a análise da evasão.

A Figura 1.1 ilustra o quantitativo de estudantes egressos e evadidos no período entre 2011 e 2020, de acordo com os dados a média da evasão foi de quase dois mil e seiscentos alunos evadidos por ano para o período. Outro fator motivador consiste na contribuição para retomada no índice da TSG, indicador de desempenho adotado pelas instituições de ensino superior que em 2019 alcançou a marca de $47,72\%$ UFMS (2020a) muito longe da meta global de 90% instituída no decreto que constitui o programa REUNI BRASIL (2007). Por fim, o fato de não haver estudos utilizando conceitos de mineração de

dados para a predição de evasão no contexto dessa universidade foi também um incentivo para o desenvolvimento deste trabalho.

Figura 1.1: Dados de egresso e evasão da UFMS.



Fonte: UFMS (2020b)

O modelo de classificação proposto utilizou dados de 8.195 acadêmicos do campus de Três Lagoas (CPTL) de um total de 15 cursos. As análises demonstraram ótimos resultados ao ranquear os alunos, apresentando uma média das precisões para os 50 primeiros alunos de $0,97\%$ para os cursos da área de Exatas, $0,99\%$ para a área de Biológicas e $0,98\%$ para a área de Humanas, alcançando até 100% de precisão em alguns casos. Tão logo, ao utilizar a classificação de uma determinada instância baseada na propensão à qual pertence e ranqueando-as, torna mais assertivo o esforço para a redução da evasão de modo a se obter um resultado muito mais preciso.

O restante deste texto está dividido da seguinte forma: no Capítulo 2 será apresentado o embasamento teórico das técnicas utilizadas neste projeto, a revisão bibliográfica, as principais características dos artigos mais relacionados ao trabalho bem como a metodologia utilizada para a construção da revisão bibliográfica. No Capítulo 3 será apresentado o método de pesquisa utilizado no trabalho proposto e o detalhamento específico de cada etapa. No Capítulo 4 serão apresentados os resultados obtidos neste trabalho utilizando dados dos alunos do CPTL. E por fim, a discussão acerca da contribuição deste trabalho para a UFMS, ideias para trabalhos futuros juntamente com o setor de tecnologia da informação e aqueles com relação direta com os estudantes serão apresentados no Capítulo 5, na etapa de conclusões deste trabalho.

CAPÍTULO

2

Embasamento Teórico e Revisão Bibliográfica

Nessa seção será apresentado um embasamento teórico sobre o processo de mineração de dados, e as técnicas aplicadas neste projeto bem como os parâmetros utilizados para a obtenção de trabalhos relacionados. Este por sua vez, serão apresentados de forma sumarizada e com uma breve descrição.

2.1 Mineração de Dados

De acordo com Amo (2004), a Mineração de Dados possui o objetivo de extrair e explorar conhecimento de grandes volumes de dados em busca de padrões até então desconhecidos, além de criar modelos para automatizar tarefas, com auxílio de técnicas que envolvem métodos matemáticos, algoritmos e heurísticas, resultando em descoberta de conhecimentos. A aplicação de técnicas de mineração de dados tem o propósito de formular e validar peguntas de pesquisa bem como descobrir novos padrões de forma autônoma.

Existem dois tipos de tarefas na mineração de dados: tarefas descritivas e preditivas (Tan et al., 2019). As tarefas descritivas visam, como o próprio nome diz, apresentar as descrições dos padrões existentes nos dados, como por exemplo a descrição de grupos ou associações entre atributos. Estas tarefas são responsáveis por identificar e agrupar registros com similaridades entre si, e posteriormente identificar a associação entre valores de atributos em uma base de dados. As tarefas preditivas por sua vez, permitem utilizar os padrões para fazer predições, sejam estes valores discretos (classificação) ou contínuos (regressão). Deste modo, podem identificar a qual classe per-

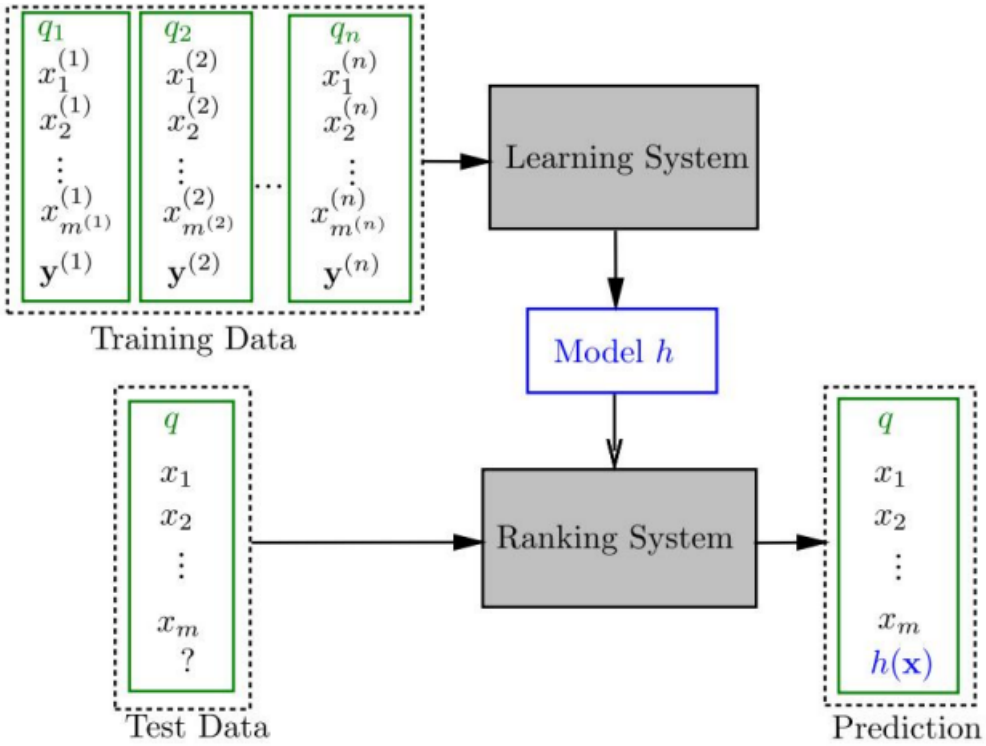
tence um determinado registro (ex: um *e-mail spam* ou não *spam*), ou prever um valor contínuo (ex: um preço de um imóvel) (Camilo e Silva, 2009). Os diferentes autores subdividem o processo de mineração de dados em etapas distintas de acordo com seus respectivos trabalhos e as definem. Embora a subdivisão de etapas bem como suas definições não seja homogênea entre os diversos autores, alguma semelhança entre elas pode ser identificada (Braga, 2005; Calixto et al., 2017; Nascimento et al., 2018).

2.1.1 Learning-to-Ranking

Learning-to-Ranking é uma técnica que utiliza o aprendizado de máquina, com a finalidade de construir modelos de classificação para sistemas de recuperação de informação usando dados de treinamento, de modo que o modelo possa classificar novos objetos de acordo com seus graus de relevância, preferência ou importância, análoga à classificação dos dados do treinamento (Liu, 2011).

As características do fluxo do *Learning-to-Ranking* é demonstrado na Figura 2.1, que exemplifica a classificação de uma determinada consulta, onde o conjunto de treinamento é representado por n *queries* de treinamentos $q_i (i = 1, ..., n)$, com seus documentos associados constituídos por vetores $X^{(i)} = \{x_j^{(i)}\}_{j=1}^{m^{(i)}}$ ($m^{(i)}$ corresponde ao número de documentos associados à *query* q_i) e os julgamentos de relevância correspondentes.

Figura 2.1: *Framework Learning-to-Ranking.*



Fonte: Liu (2011).

Em seguida, para que o modelo de classificação possa prever um novo rótulo, um algoritmo de aprendizado específico é empregado para aprender e combinar os recursos disponíveis da melhor forma cabível, em termos de uma função de perda. Já na fase de teste, com chegada de uma nova *query*, o modelo classifica os documentos com base na sua relevância para a consulta, retornando uma lista classificada com a resposta à sua consulta.

Entre as classificação dos algoritmos de *Learning-to-Ranking*, pode-se destacar duas abordagens:

- ***Pointwise approach***: são otimizados para prever uma métrica chave. Por exemplo, classificar recomendações de produtos em que o usuário tem a maior probabilidade de clicar em um item (modelos de classificação). Também é possível avaliar o modelo através das métricas comuns como, *accuracy*, *precision*, *recall*, F_1 para uma determinada posição k. Por exemplo, quantos alunos estão com probabilidade de evasão acima dos 80% para os top 50 alunos da turma de Ciências da Computação do segundo ano do campus CPTL?

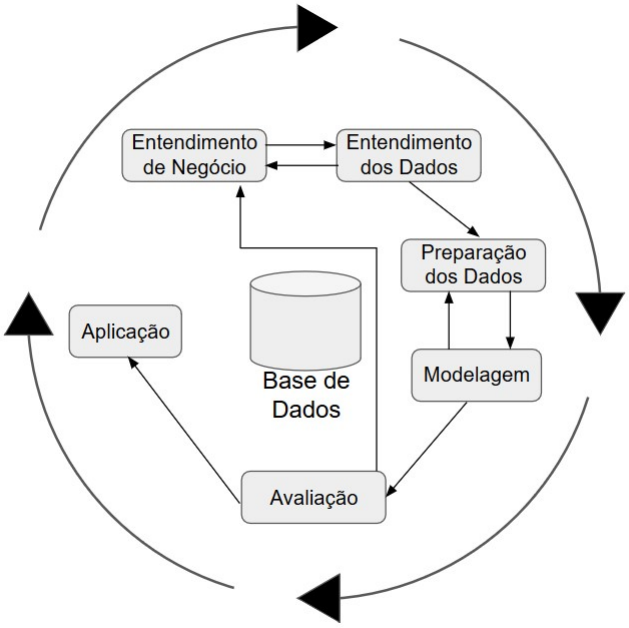
- ***Pairwise approach***: objetiva-se ordenar de forma mais otimizado a ordem de dois item por vez. O melhor cenário para esta abordagem é ter o número máximo de pares na ordem correta, pois se o item mais relevante estiver no topo, não será necessário adicionar perda, caso contrário, aumentaria a função de perda. A avaliação do modelo baseado nesta abordagem é feita calculando as métricas por pares, por exemplo, divisão do número de pares em ordem correta pelo total de pares. Como exemplo de modelo, pode-se citar o RankSVM, na qual é fundamentado no *Support Vector Machine*, em que se calcula a perda minimizando os pares destoantes do modelo com a ordem ideal, punindo-as (Lee e Lin, 2014).

Diante do exposto, este projeto utiliza a abordagem baseado em *Pointwise*, pois possui o objetivo de calcular, de forma individual, a probabilidade de evasão do aluno e ao final gerar uma lista ordenada, na qual o topo representa aquele com maior probabilidade de evasão.

2.2 Metodologia Crisp-DM

Para este trabalho, optou-se por adotar a metodologia Crisp-DM (*Cross-Industry Standard Process for Data Mining*) (Provost e Fawcett, 2018). Segundo Marban et al. (2009), o modelo se revela o mais utilizado para a mineração de dados e descoberta de conhecimento. Além disso, também foi observado o uso desse modelo em vários trabalhos durante a revisão bibliográfica.

Figura 2.2: Fases do Crisp-DM.



Fonte: Figura baseada em Shafique e Qaiser (2014).

A metodologia Crisp-DM foi desenvolvida inicialmente por meio de um consórcio composto por Daimler Chryrsler, *Statistical Package for the Social Sciences* (SPSS) e *National Cash Register Corporation* (NCR). A metodologia possui quatro níveis: **Fases, Tarefas Genéricas, Tarefas Especializadas e Instâncias de Processo**. A metodologia parte de uma visão mais abstrata e afunila para processos mais específicos. As seis fases do Crisp-DM encontram-se representadas na Figura 2.2, pode-se observar a ciclicidade existente entre as fases deste modelo, sendo possível percorrer determinada fase mais de uma vez até obter o resultado desejado. Nas subseções seguintes serão apresentados os detalhes de cada etapa da metodologia Crisp-DM.

2.2.1 Entendimento do Negócio (Business Understanding)

Através de uma perspectiva de negócio, a fase inicial é focada em compreender quais são os objetivos e os requisitos do projeto. Em seguida, todo o conhecimento abstraído é convertido para um problema de mineração de dados e por fim realiza-se o planejamento estratégico preliminares para alcançar tal objetivo (Wirth e Hipp, 2000).

Segundo López (2021), para o entendimento do negócio, podemos ter as seguintes ações:

- **Determinar os objetivos de negócio:** através de uma perspectiva de negócio, compreender de maneira completa os anseios do cliente e estabelecer os critérios de sucesso do negócio.

- **Avaliar a situação:** investigar a disponibilização de recursos necessários, avaliar os riscos e ações de contingências e por fim analisar o custo-benefício do projeto.

- **Determinar o objetivo da mineração de dados:** transcrever para termos técnicos os objetivos na terminologia de negócios e descrever as saídas pretendidas de maneira subjetivo ao projeto.

- **Produzir um plano:** especificar as etapas do plano intencional para alcançar o objetivo de negócio, especificando as ferramentas e técnicas iniciais.

2.2.2 Entendimento dos Dados (Data Understanding)

Após entender o problema e definir os objetivos, esta etapa se inicia pela busca dos dados relacionados ao problema. As informações necessárias podem estar em bancos de dados, documentos físicos ou até mesmo armazenadas na memória de uma pessoa. Os formatos de dados encontrados podem ser estruturados, semiestruturados e até mesmo não estruturados.

Além do esforço em se obter os dados, nesta etapa, estes são explorados quanto aos tipos, qualidade, dados ausentes, forma de distribuição, volumetria e intervalos. Após a familiarização com as informações com base no problema a ser tratado, concebendo os primeiros *insights* sobre os dados coletados e então os subconjuntos são detectados para a formação de perguntas de pesquisa para as informações ocultas.

Para Shearer (2000), a segunda etapa consiste das seguinte tarefas:

- **Coleta de dados iniciais:** obtenção dos dados necessários, seja por carregamento dos dados ou por integração. Deve-se relatar os problemas encontrados e as soluções aplicadas para auxiliar na futura replicação do projeto.

- **Descrição dos dados:** examina-se as propriedades “grossa” ou "superficial"dos dados, como o formato, a quantidade, número de registros e qualquer outra informação relacionado aos dados obtidos.

- **Exploração dos dados:** Consultas, visualizações e relatórios, são tarefas que abordam questões de mineração de dados nesta etapa.

- **Verificação da qualidade dos dados:** deve-se verificar a qualidade abordando questões como: tipos de dados, ausência dos dados, atributos ausentes, plausibilidade dos valores, entre outros.

2.2.3 Preparação dos Dados (Data Preparation)

Esta etapa consiste na estruturação dos dados por meio da seleção e análise das instâncias e atributos para a próxima etapa, eliminação de dados incorretos, avaliação dos dados ausentes, e também são aplicadas técnicas como normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Essa etapa é necessária a fim de garantir a qualidade dos dados a serem utilizados posteriormente na próxima etapa, em geral é a etapa que mais morosa para López (2021) que pode consumir até 90% do projeto.

Shearer (2000) considera cinco etapas na preparação de dados:

- **Selecionar dados:** a seleção dos dados é baseada em vários critérios, como a relevância dos atributos para os objetivos da mineração de dados, restrições de qualidade, limites de volume ou tipo de dados, entre outros.

- **Limpar os dados:** etapa de fundamental importância para os resultados do projeto. Deve-se selecionar subconjuntos limpos de dados ou estimar dados ausentes, através de técnicas como análises de modelagem.

- **Construir os dados:** finalizada a limpeza de dados, é necessário a preparação dos dados, seja construindo novos registros ou através atributos derivados.

- **Integrar os dados:** combinação de diferentes tabelas para criar novos registros ou valores para um mesmo objeto. Também faz parte desta tarefa à agregação, operação responsável pela geração de novos valores originários de registros e/ou tabelas.

- **Formatação de dados:** a formatação de dados parte da necessidade de tonar os dados adequados para o uso em ferramentas de modelagem específica.

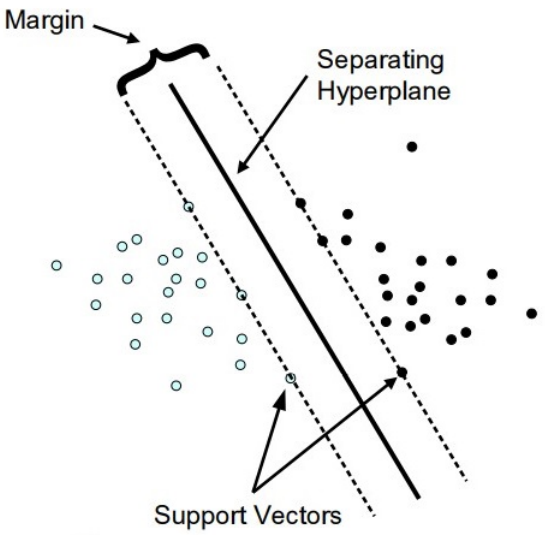
2.2.4 Modelagem (Modeling)

Nesta fase é onde se iniciam as construções dos modelos. São definidos quais serão os tipos de tarefas e técnicas de mineração de dados a ser utilizado, como os modelos de regressão, agrupamento, classificação, e associação. Os resultados obtidos dependem fortemente da etapa anterior, sendo necessário em muitos casos, retroceder para a etapa anterior e definir novas escolhas para pré-processar os dados. Neste projeto, esta etapa é focada no uso de modelos de classificação, tanto para a extração de conhecimento quanto para a previsão de alunos com potencial de evasão.

Neste trabalho pretende-se inicialmente utilizar os modelos de classificação recorrentemente utilizados nos trabalhos relacionais. A seguir são apresentadas as descrições de tais modelos:

- ***Support Vector Machine (SVM)*:** conhecido como Máquina de Vetores de Suporte, foi desenvolvido por *Vapnik* Vapnik (1999). É baseado na Teoria da Aprendizagem Estatística para resolver problemas relacionados à classificação de padrões, utilizando o conceito de hiperplano de separação de margem máxima entre as classes (Viana et al., 2007). Sendo o hiperplano gerado pela SVM, definido através de um subconjunto dos pontos das duas classes, na qual é chamado de vetor de suporte (Gevert et al., 2010). Na Figura 2.3 é possível visualizar o hiperplano separando os conjuntos de dados.

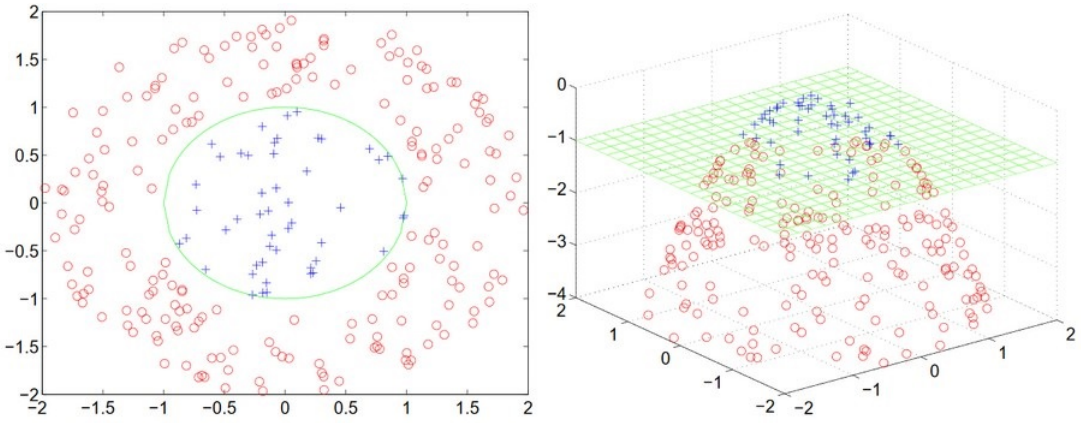
Figura 2.3: Exemplo de um caso linear separável.



Fonte: Adaptado de Meyer e Wien (2001)

Quando os dados não são linearmente separáveis, isto é, não podem ser separados por um hiperplano, utiliza-se o *Kernel Trick*, também conhecido como *Kernel Methods* (Smola et al., 1999). *Kernel Trick* são técnicas que utilizam uma função *kernel*, calculando o produto interno entre os pares dos dados de treinamento no espaço de recurso, fazendo com o que os dados sejam mapeados para um novo espaço onde sejam linearmente separáveis (Hofmann, 2006). Na Figura 2.4 (a), uma ilustração de dados dispostos em duas dimensões e não linearmente separáveis pode ser observada, enquanto na Figura 2.4 (b), os dados após a aplicação da função *kernel*, agora mapeados em um espaço de três dimensões onde é possível separar as duas classes de dados por meio de um hiperplano.

Figura 2.4: Exemplo de Aplicação de *Kernel Trick*.



(a) Espaço com dados iniciais (b) Espaço projetado

Fonte: Adaptado de Wu et al. (2005)

Desde 2007 tem sido observado uma crescente adesão do SVM em aplicações de aprendizagem de máquina, com resultados equiparados e até mesmo superior à outros algoritmos como *Artificial Neural Networks* (Lorena e de Carvalho, 2007). Dentre os trabalhos relatados na literatura que visaram comparar diferentes algoritmos para encontrar àqueles com a melhor performance, o SVM destacou-se pela frequência com que ocorre entre os melhores algoritmos de classificação nos diversos trabalhos publicados[1].

Como exemplos de trabalhos utilizando SVM, pode-se apontar os trabalhos de Díaz et al. (2021) e Manhães e da Cruz (2020), nos quais apresentam propostas de utilização de SVM com variação no *kernel (Poly*, *Radial Basis Function* (RBF)).

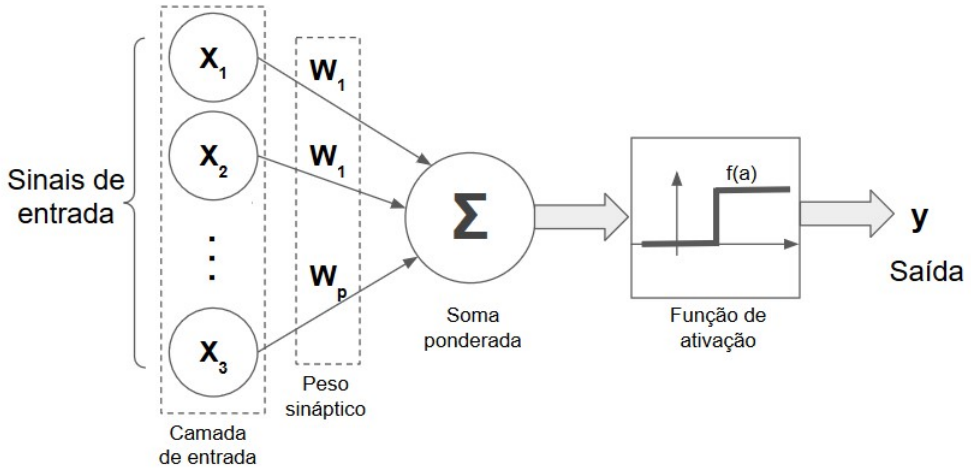
- ***Artificial Neural Networks* (ANN)**: baseada no funcionamento de um cérebro, adquire conhecimento em processo de aprendizagem (algoritmo de aprendizagem) e utiliza conexões sinápticas para armazenar o conhecimento obtido. Possui muitas unidades de processamentos simples (neurônios) em paralelo para aprender superfícies de separação entre as classes (Haykin, 2011).

  Em geral, as redes neurais artificiais são mecanismos simples com entradas e saídas organizados em camadas (Noriega, 2005). Na Figura 2.5 é ilustrado uma unidade de processamento *Perceptron*, onde X representa os sinais de entrada, W representa os pesos sinápticos ou a influência na saída da unidade, e a função de ativação corresponde uma função que

[1]Lista dos trabalho que apresentaram o SVM como o melhor algoritmo de classificação: (Delen, 2010), (Manhães et al., 2012a), (Tekin, 2014), (Santana et al., 2015), (Costa et al., 2017), (Marques, 2020), (Díaz et al., 2021), (Filho et al., 2020), (Del Bonifro et al., 2020), (Gonçalves e Beltrame, 2020)

definirá o valor que será propagado. A rede aprende ajustando os pesos w nas conexões de forma a minimizar o erro dos exemplos de treinamento, isto é, minimizar o valor predito pela rede e o valor desejado.

Figura 2.5: Unidade de processamento de uma *Artificial Neural Networks*.



Fonte: Adaptado de Haykin (2011)

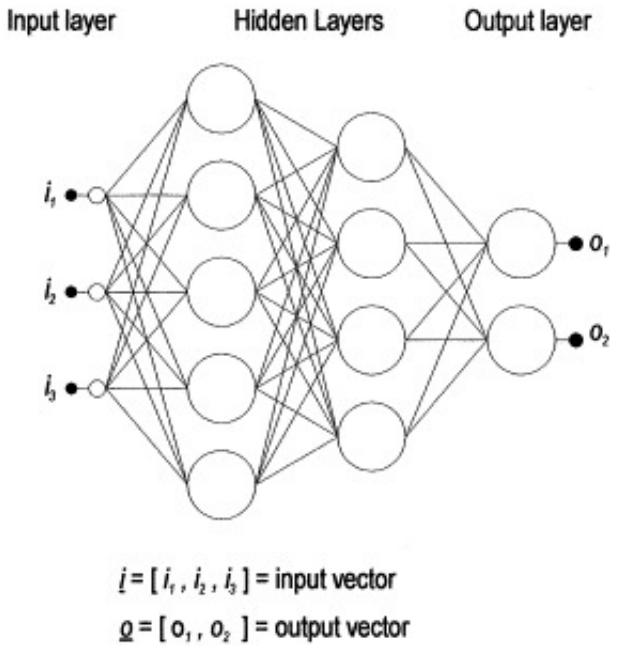
Entretanto o *Perceptron* é limitado à resolver funções linearmente separáveis, ou seja, não consegue gerar um hiperplano separando os dados em duas classes. Diante desse problema surge o *Multilayer Perceptron* (MLP), Figura 2.6, com vários nós interconectados representando um mapeamento não linear entre o vetor de entrada e o vetor de saída, com nós intermediários conectados por peso alimentando uma função de transferência ou ativação não linear (Gardner e Dorling, 1998).

Vários trabalhos utilizam o *Multilayer Perceptron* para a predição de evasão, com ótimos resultados, como em RAMESH et al. (2013); Antonio e Luiz (2020); Rigo et al. (2014); Delen (2010)

Além das unidades de processamento e funções de ativação apresentadas, existem outras redes formadas por outros tipos de neurônios e funções de ativação, destaca-se as redes com camadas recorrentes e camadas de convolução, e funções de ativação como a tangente hiperbólica e a *Rectified Linear Unit* (ReLU) (Aggarwal et al., 2018).

- **Regressão Logística, do inglês *Logistic Regression* (LR):** A regressão logística amplamente empregado na área de aprendizado de máquina, baseia-se na técnica estatística de análise de regressão, através de um modelo adequadamente ajustado, objetiva-se a mensurar dois tipos de relação, ou entre uma variável resposta e um conjunto variável binária alvo ou entre variáveis e um conjunto de características explicativas. Pode-se exemplificar em termos de classificação binária (0 ou 1) para a

Figura 2.6: Exemplo de uma rede neural *Multilayer Perceptron*.

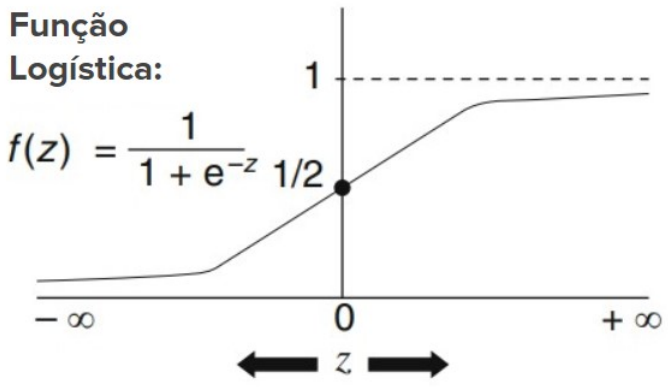


Fonte: (Gardner e Dorling, 1998)

doença cardíaca coronária (DCC), com a classificação dos pacientes em 0 (sem DCC) ou 1 (com DCC), baseadas em variáveis de controle como, idade, corrida, sexo. Sendo assim, a regressão logística é uma modelagem matemática que pode descrever a relação de vários atributos com uma variável dicotômica (Kleinbaum et al., 2002).

Na Figura 2.7 é apresentado a fórmula matemática do modelo logístico, em que quando $z = -\infty$, a função logística é 0 e para $z = +\infty$ a função é 1, logo o intervalo de possibilidade de $f(z)$ está sempre entre 0 e 1.

Figura 2.7: Função logística do modelo logístico.



Fonte: Adaptado de Kleinbaum et al. (2002)

Assim, para obter o modelo logístico, da qual a regressão logística se baseia, é necessário uma função que consiga separar de maneira correta as classes, por exemplo com DCC e sem DCC. Considera-se $z = \alpha + \beta_1 x_1 +$

$\beta_2 x_2 + ... + \beta_k x_k$, onde x são as variáveis de interesses independentes e α e β_i são as constantes que representa os parâmetros desconhecidos. De maneira geral, o modelo logístico pode ser escrito como:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{2.1}$$

Trabalhos como Rovira et al. (2017); Martins et al. (2017); Burgos et al. (2018), obtêm bons resultados utilizando a regressão logística para a predição de evasão dos alunos.

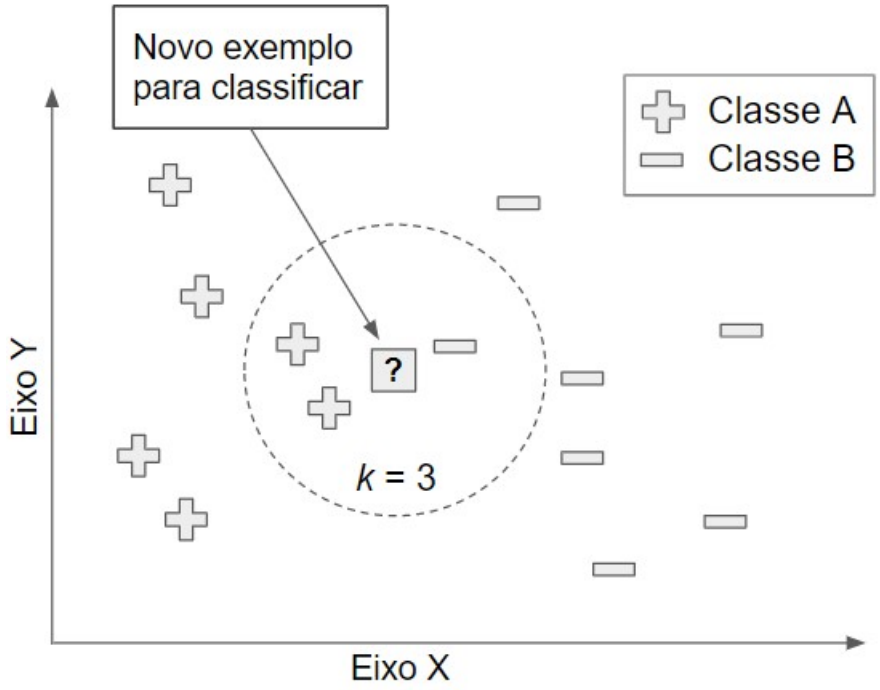
- **K-vizinhos mais próximos (do inglês k-*Nearest Neighbors* – k-NN):** É um modelo de classificação que utiliza a votação majoritária para a classificação de um dado t, baseada na recuperação dos k vizinhos mais próximo, entretanto é de fundamental importância a escolha do conjunto dos k-vizinhos para uma boa classificação (Guo et al., 2003). A Figura 2.8 demonstra a classificação de um novo exemplo à partir do $k = 3$ vizinhos mais próximo, na qual o exemplo irá pertencer à classe A. Há uma preocupação na escolha do valor de k, se k for insuficiente, então o classificador pode estar sujeito a *overfitting*, relacionado ao ruído do conjunto de treinamento. Também o cenário oposto pode ocorrer, *underfitting*, se k for grande demais, o classificador poderá rotular erroneamente a instância de teste, pois o maior conjunto de vizinhos mais próximos estão longe da sua vizinhança (Tan et al., 2019).

  Em Filho et al. (2020) com valor de $k = 5$, destaca o resultado na especificidade em comparação à outros classificadores. No trabalho de Antonio e Luiz (2020), o k-NN obtém ótimos resultados quando ao utilizar o atributo frequência escolar.

- **Árvores de Decisão, do inglês *Decision Tree* (DT):** são formas simples de representação do conhecimento e um modo de construir classificadores que predizem classes baseados nos testes condicionais nos valores dos atributos de uma instância. É baseada na representatividade de uma árvore real com raiz, nós, ramos e folhas. São considerados nós internos de decisão, a raiz (primeiro nó) da árvore e todas as bifurcações, onde cada nó realiza um teste sobre algum atributo e o resultado origina uma aresta para uma subárvore. Nas extremidades das árvores, estão localizados os nós folhas que são os valores de predição e representam as classes de um conjunto de dados (Meira et al., 2008).

  A partir de uma árvore de decisão, é possível derivar regras, que são escritas considerando o trajeto do nó raiz até uma folha da árvore (Shiba et al., 2013). Na Figura 2.9 é apresentado um exemplo de uma árvore

Figura 2.8: Exemplo de uma classificação de vizinho mais próximo com valor *k* grande.



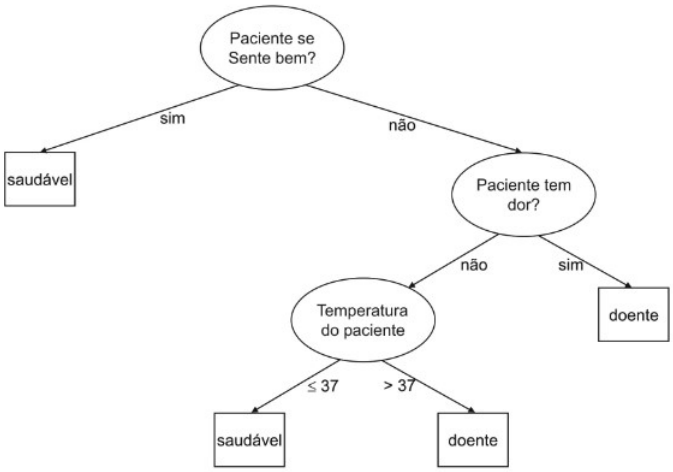
Fonte: Adaptado de Tan et al. (2019).

de decisão para um simples diagnóstico de um paciente, onde cada bifurcação apresenta uma pergunta (um valor de um atributo) e as folhas representam um diagnóstico (classe).

Normalmente adota-se um processo guloso para viabilizar a construção de uma árvore de decisão, isto é, como raiz, escolhe-se o atributo cujos valores melhor separam os dados. A partir daí, cada nó de cada ramo segue a mesma estratégia, até que os conjuntos dos testes de um caminho da árvore recupere apenas nós puros, ou até que um critério de parada seja atingido. Para definir quais atributos melhor separam os dados, normalmente são utilizadas medidas de teoria da informação, como Entropia, Taxa de Ganho e *Gini Index* (Tan et al., 2019).

Esta técnica de classificação é o algoritmo mais utilizado entre todos os trabalhos selecionados durante a revisão bibliográfica. Bunkar et al. (2012) apresenta no trabalho de previsão de melhoria de desempenho, apenas algoritmos baseados em árvores de decisão, $ID3, C4.5$ e $CART$. Já o autor Soares et al. (2020) utilizou apenas o algoritmo floresta aleatória, resultando em ótimos resultados nas métricas de aprendizado de máquina. Para uma definição de floresta aleatório, Cutler et al. (2012) descreve que uma floresta aleatória é uma coleção de classificadores estruturados em árvore $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$, na qual, $\{\Theta_k\}$ são vetores aleatórios distribuídos identicamente independentes e cada árvore escolhe através de voto unitário para a classe mais popular na

Figura 2.9: Exemplo de uma Árvore de Decisão.



Fonte: Adaptado de Monard e Baranauskas (2003).

entrada **x**.

2.2.5 Avaliação (Evaluation)

Nesta fase existem duas avaliações possíveis: uma chamada avaliação subjetiva e outra chamada avaliação objetiva. A avaliação subjetiva consiste na apresentação dos padrões extraídos ou de resultados à especialistas de domínio, os quais atribuirão uma nota ou um grau de qualidade. Já na avaliação objetiva, os padrões extraídos são comparados com um conjunto verdade, ou são utilizadas medidas para analisar a distribuição ou correlações entre os padrões. Durante a fase de avaliação, pode-se refazer todas as fases anteriores ou alguma outra intermediária de forma a obter uma qualidade desejável.

No caso do presente projeto, que utilizará modelos de classificação, têm-se esquemas e métricas para avaliar a qualidade dos modelos de classificação. Entre os esquemas mais utilizados, têm-se:

- ***Holdout***: foi o esquema de avaliação mais utilizado nos trabalhos relacionados. Os dados são divididos em duas partes não sobrepostas e, sendo uma parte utilizada para treinamento do modelo e outra para teste do modelo. A parte que é utilizada para o teste é a resistência ou *holdout*, pois é oferecida para testar o modelo e o restante dos dados é utilizado para aprender (Yadav e Shukla, 2016).

- k***-Fold Cross-Validation***: em segundo lugar está a técnica k-*Fold Cross-Validation*, em que o conjunto de dados é dividido randomicamente em k conjuntos disjuntos (pastas) e são realizadas k iterações, sendo que em cada iteração uma pasta é utilizada para teste a as demais $k - 1$ pastas

são utilizadas para treino (Kohavi, 1995).

Uma vez definido o esquema de avaliação, os rótulos preditos dos exemplos de teste pelo modelo de classificação são comparados com os rótulos reais, formando assim a matriz de confusão, ilustrada na Tabela 2.1. A partir dessa matriz, pode-se derivar métricas para estimar a performance de classificação. A utilização de métricas como a acurácia, precisão, revocação e F_1, são amplamente utilizadas para a avaliação dos resultados.

Tabela 2.1: Matriz de confusão.

Matriz de Confusão		**Classe Real**	
		Positivo	**Negativo**
Classe Predita	**Positivo**	VP	FP
	Negativo	FN	VN

Fonte: Adaptado de Rossi (2011).

Dentre as métricas de avaliação, a acurácia foi a mais utilizada durante a pesquisa dos trabalhos relacionados. A acurácia, segundo Gottardo et al. (2012), é dada pela Equação 2.2:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.2}$$

na qual:

- **VP** é o número de Verdadeiros Positivos, isto é, quando o modelo previu o exemplo como sendo da classe positiva (por exemplo, `evasão`) e acertou a previsão.
- **VN** é o número de Verdadeiros Negativos, isto é, quando o modelo previu o exemplo como sendo da classe negativa (por exemplo, `não evadiu`), porém, o exemplo era da classe positiva.
- **FP** é o número de Falsos Positivos, isto é, quando o modelo previu como sendo da classe positiva, porém, o exemplo pertencia à classe negativa.
- **FN** é o número de Falsos Negativo, isto é, quando o prevê o exemplo como sendo da classe negativa, porém, o exemplo pertence à classe positiva.

A segunda métrica de avaliação mais utilizada nos trabalhos foi a precisão, o qual o trabalho Mueen et al. (2016) considera a precisão conforme descrito na Equação 2.3. Também será considerada a medida de revocação, do inglês *Recall* (conhecida como cobertura ou sensibilidade), conforme apresentado na Equação 2.4. Já para as Equações 2.5, 2.6 e 2.7, são baseadas em calcular cada rótulo individualmente e depois calcula-se a média entre os rótulos,

sendo os rótulo minoritário por influenciar mais nas métricas macro média. E a média harmônica entre a precisão e revocação, denominada F_1 ou F_1-*Score*, é apresentado na Equação 2.8.

$$\textit{Precisão} = \frac{VP}{VP \ + \ FP} \tag{2.3}$$

$$\textit{Revocação} = \frac{VP}{VP \ + \ FN} \tag{2.4}$$

$$\textit{Macro Revocação} = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i+FN_i}}{n} \tag{2.5}$$

$$\textit{Macro } F_1 \frac{\sum_{i=1}^{n} F_{1i}}{n} \tag{2.6}$$

$$\textit{Macro Precisão} = \frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i+FP_i}}{n} \tag{2.7}$$

$$F_1 \textit{ da Classe de Interesse:} = 2 \times \frac{\textit{Precisão} \times \textit{Revocação}}{\textit{Precisão} + \textit{Revocação}} \tag{2.8}$$

Para as métricas do *ranking* são consideradas a precisão@k, revocação@k e F_1@k, as fórmulas são baseadas nas equações 2.3, 2.4 e 2.8, variando no quantitativo de amostra k, ou seja $k = n$, onde k representa o k primeiros alunos ranqueados.

A precisão avalia apenas os acertos dos positivos em relação ao total de casos que o modelo previu como sendo positivo. Já a métrica de revocação, avalia o quanto o modelo acertou nos casos positivos em relação ao total de casos que são realmente positivos. E a medida F_1 calcula a média harmônica entre a precisão e a revocação.

Ainda nesta fase de avaliação, é possível o retorno para a fase inicial do Crisp-DM, caso os resultados não estejam satisfatórios. Sendo assim, será necessário rever mudanças no projeto, como utilização de novos classificadores ou novos atributos.

2.2.6 Aplicação (Deployment)

Após alcançar performances satisfatória com o modelo de classificação, o trabalho tem como o próximo objetivo criar uma aplicação do modelo, última etapa do Crisp-DM. Serão duas partes referentes a aplicação do modelo. A primeira consiste em desenvolver um serviço que interaja com o Siscad-Admin

(Sistema Acadêmico da UFMS, plataforma específica para docente lançar as notas e frequências), tanto para a coleta de dados, quanto para a geração de relatórios oriundo das predições.

A segunda etapa da aplicação consiste na extração de conhecimento via obtenção de regras ou padrões descritivos para o entendimento do que leva o aluno a evadir. Esse conhecimento extraído será apresentado e compartilhado aos interessados, como coordenadores de cursos, diretores de faculdades e institutos, pró-reitores e reitor. Vale ressaltar que o modelo pode ser constantemente atualizado conforme obtenção de novos dados ou outros entendimento do negócio sejam definidos.

2.3 Revisão Bibliográfica

A etapa de revisão bibliográfica realizada neste projeto seguiu quatro passos: 1) definição dos parâmetros de busca; 2) sumarização dos resultados, 3) definição e separação dos principais trabalhos relacionados; e 4) obtenção dos principais algoritmos de classificação e métricas de avaliação. A seguir, são apresentados os detalhes de cada passo dessa etapa.

2.3.1 Parâmetros de Buscas

Para a realização da revisão bibliográfica, foi definido a utilização da plataforma de pesquisa acadêmica Google chamada *Google Scholar*. Esta é uma ferramenta de pesquisa virtual, gratuita, e com grande capacidade de executar buscas por meio de textos ou metadados em diversos formatos de publicações, em múltiplas línguas, disponibilizadas em repositório na *web* ou sites acadêmicos, na qual a plataforma indexa os mais utilizados e conceituados repositórios científicos do mundo. Também é possível ordenar as buscas pelos trabalhos mais citados e também pelos mais recentes, sendo esses dois filtros utilizados para fazer a pesquisa dos trabalhos.

Foram definidas as *queries* de pesquisa com as seguintes combinações:

- *"School dropout"* + *"machine learning"*
- *"School dropout"* + *"data mining"*
- *"School dropout"* + *"forecast"*
- *"University dropout"* + *"data mining"*
- *"Evasão"* + *"mineração de dados"*
- *"Evasão"* + *"aprendizado de máquina"*

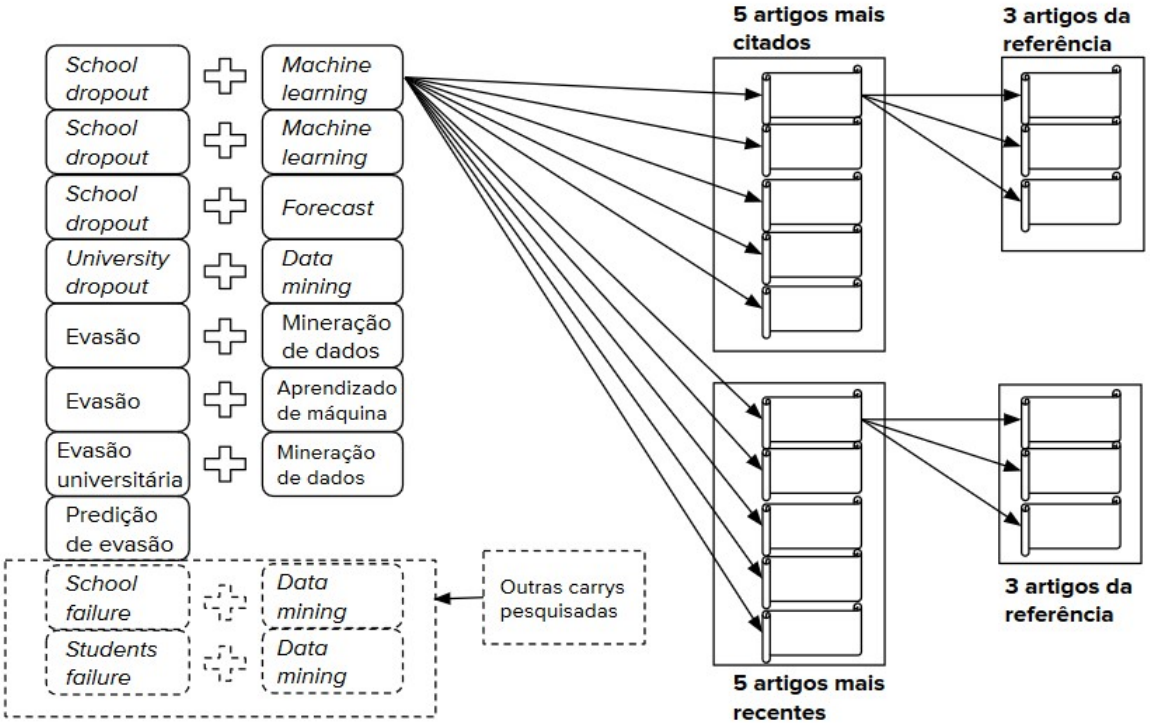
- *"Evasão universitária"* + *"mineração de dados"*
- *"Predição de evasão"*

Também houve, durante a busca, a inclusão de mais duas *queries* conforme a leitura dos trabalhos obtidos na etapa anterior:

- *"School failure"* + *"data mining"*
- *"Students failure"* + *"data mining"*

Foram selecionados os 5 trabalhos mais citados e os 5 trabalhos mais recentes para cada *query* de busca. Além disso, para cada trabalho selecionado na etapa anterior, foram também selecionados os 3 trabalhos citados dentro do trabalho selecionado e que foram considerados os mais relevantes. Na Figura 2.10 é apresentada a ilustração do procedimento para a obtenção dos trabalhos relacionados. No total foram consultados aproximadamente 300 trabalhos durante esta etapa do projeto.

Figura 2.10: *Queries* de pesquisas dos trabalhos.



Fonte: Elaborado pelo autor.

2.3.2 Sumarização dos Resultados da Busca

Para facilitar a divisão dos trabalhos mais relacionados e para dar uma ampla visão das técnicas de mineração de dados e esquemas de validação utilizados na literatura, os trabalhos relacionados foram agrupados considerando: (i)

o país onde o estudo foi realizado, (ii) a modalidade de ensino considerada, (iii) modalidade de ensino considerando apenas Brasil, e (iv) as métricas utilizadas na avaliação. A revisão bibliográfica obteve um elevado número de trabalhos duplicados em diferentes *queries*. Além disso, com as *queries* de busca foram obtidos resultados nos quais pode-se perceber que vários países ao redor do mundo têm o interesse na mineração de dados educacionais.

Na Tabela 2.2 é apresentada a distribuição geográfica dos trabalhos onde o estudo foi realizado, a qual apresenta 31 trabalhos com dados de estudantes brasileiros de um total de 55 trabalhos pesquisados.

Tabela 2.2: Distribuição de trabalhos por país, onde o estudo foi aplicado

País	**Trabalhos**
Brasil	31
Espanha	6
EUA	3
Grécia	2
Índia	2
Alemanha	1
Arábia Saudita	1
China	1
Cuba	1
Holanda	1
Itália	1
Malásia	1
Marrocos	1
Peru	1
Sérvia	1
Turquia	1
Total	55

Fonte: Elaborado pelo autor.

Já na Tabela 2.3, os trabalhos apresentam dados apenas de estudos com alunos do Brasil, categorizado de acordo com a modalidade de ensino, sendo a maioria dos estudos focados em alunos da graduação presencial. Os trabalhos pesquisados apresentaram uma imensa variedade de atributos, em geral, o número de cai ao longo dos estudos de cada trabalho, pois alguns atributos observa-se que não há muita relevância para a predição.

É apresentado na Tabela 2.4 o resumo dos dados coletados dos trabalhos pesquisados apenas com alunos da Graduação Presencial. A fim de trazer maior confiabilidade nos resultados apresentados nos trabalhos, a ordenação na tabela foi baseada no volume de dados que o trabalho utilizou. Vale ressaltar que apenas três trabalhos utilizaram quantitativo acima de 10.000 alunos.

Tabela 2.3: Distribuição de trabalhos por modalidade de ensino aplicado nos alunos do Brasil.

Modalidade	**Trabalhos**	**Quantidade Média de Atributos**
Graduação Presencial	19	22,7
Graduação EAD	5	20,33
Curso EAD	1	21,0
Curso Presencial	1	17,0
Curso Técnico EAD	1	Não informado
Curso Técnico Presencial	1	41,0
Ensino Fundamental Presencial	1	Não informado
Técnico e Graduação Presencial	1	10,0
Técnico Presencial	1	8,0

Fonte: Elaborado pelo autor.

Tabela 2.4: Estatísticas dos trabalhos na modalidade graduação presencial com as acurácias, ordenado pelo maior volume de dados.

Modalidade	***Melhores Resultados***		**Métrica de Avaliação**	**Volume Base de Dados**
Graduação Presencial	*Support Vector Machine*	87,00%	Acurácia	15.000
Graduação Presencial	*Support Vector Machine*	86,35%	Acurácia	14.024
Graduação Presencial	*Naive Bayes*	87,16%	Acurácia	14.000
Graduação Presencial	*Artificial Neural Network*	92,00%	Acurácia	5.854
Graduação Presencial	*Artificial Neural Network*	93,00%	Acurácia	4.560
Graduação Presencial	*Naive Bayes*	90,14%	Acurácia	1.359
Graduação Presencial	*Naive Bayes*	75,00%	Acurácia	300
Graduação Presencial	*Support Vector Machine*	92,30%	Acurácia	162
Graduação Presencial	*Support Vector Machine*	98,00%	Acurácia	100
Graduação Presencial	*Support Vector Machine*	91,6%	Acurácia	95

Fonte: Elaborado pelo autor.

A maioria dos trabalhos, selecionados durante a revisão bibliográfica, possui como característica da métrica de avaliação mais utilizada a acurácia, sendo que esta métrica pode apresentar resultados distorcidos frente à classes desbalanceadas, que é o caso da predição de evasão.

2.3.3 Principais Trabalhos Relacionados

A seguir será apresentado um resumo dos trabalhos relacionados encontrados durante a pesquisa bibliográfica. Para definir os trabalhos mais relacionados a serem descritos aqui, foram consideradas as seguintes características: atributos de pré-registro, ou seja, características da vida escolar do aluno antes de ingressar na instituição, assim como as características dos dados socioeconômicos e da vida acadêmica, como notas e frequência das disciplinas.

Em Alvarez et al. (2020), os autores apresentam uma abordagem para previsão de evasão utilizando atributos como:

- **Gênero**;

- **Cidade de origem**;
- **Fonte de ingresso na universidade**: conhecido como Pré-Universitário e são ministradas por Institutos Pré-Universitários Urbanos - IPU na Cuba;
- **Índice acadêmico anterior à admissão**, ou seja, a média geral das disciplinas cursadas no pré-universitário;
- **Notas dos exames de matemática**: são pré-requisitos para ingressar nas universidades ter concluído o pré-universitário e obter 60% de aproveitamento das provas de Matemática, Espanhol e História;
- **Classificação de opção de graduação**: consistem na lista das 10 carreiras que o aluno escolhe ordenados por preferência.

A partir da compreensão desses atributos de pré-registro, foi possível adaptar no modelo a ser trabalhado. Os autores utilizaram o *software Waikato Environment for Knowledge Analysis* (WEKA) (Hall et al., 2009), por ser uma das ferramentas mais difundidas para o propósito. Para os algoritmos de classificação, foram escolhidos a árvore de decisão J48 e o MLP. Durante o experimento foram utilizados apenas os atributos de pré registro, o trabalho apresenta uma taxa de $68,86\%$ de acurácia sobre os alunos recém ingressantes.

No trabalho de Brito et al. (2014), o autor utiliza três atributos de pré-registro, Média Geral, Média de Matemática e a Média de Física, dos alunos do curso de Ciência da Computação da Universidade Federal da Paraíba (UFPB), e a nota do vestibular. Também foram utilizadas quatro notas relacionadas às disciplinas do primeiro ano do curso: Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica. O trabalho divide a base de dados dos alunos em duas classes:

- Classe A: alunos aprovados em todas as disciplinas;
- Classe B: alunos reprovados em pelo menos uma disciplina;

Em seguida, utilizou-se a ferramenta WEKA para a realização dos testes com cinco classificadores (Naive Bayes, IBk, SMO, *Random Forest* e *Multilayer Perceptron*). Ao final o trabalho mostra o algoritmo *Naive Bayes* com a melhor acurácia entre todos, 75%. Ainda este mesmo algoritmo resulta para a classe A, a porcentagem de verdadeiros positivos de $73,2\%$ e $76,5\%$ para a classe B. Ao final, os autores concluem que o desempenho no vestibular não é o único fator influenciador para o aluno ir bem no primeiro período do curso, entretanto, é uma característica de extrema importância para o seu desempenho acadêmico.

Já no trabalho de Antonio e Luiz (2020), o autor realiza um experimento com os dados socioeconômicos dos alunos do Instituto Federal do Espírito Santo (IFES) - Campus Guarapari. São utilizados 11 atributos do qual oito são relacionados à fatores socioeconômicos (Cidade, Instituição de ensino anterior, Cor/Raça, Renda familiar, Idade, Faixa etária, Tipo de instituição de ensino anterior, Gênero). O trabalho comprova a importância do uso dos atributos relacionados à frequência do aluno, onde o classificador SVM obteve $91,6\%$ contra $73,7\%$ de acurácia sem o uso de dois atributos: Faixa (Frequência), Frequência (Comparecimento). Também é importante destacar a relevância do atributo cidade, o motivo, segundo o autor, é relacionado a dificuldade de transporte para alunos que residem longe do local de estudo. Os autores ressaltam que o trabalho apresenta uma alta porcentagem de falso positivo entre 40 à 50%, à ser melhorado em trabalhos futuros.

Em Souza (2020), o autor do trabalho inicia com 43 atributos, o qual ao longo dos testes da pesquisa reduziu para 32 atributos, sendo 18 atributos relacionados à vida acadêmica e 14 atributos relacionados à questões socioeconômicas. O trabalho evidencia os motivos pelos quais cada atributo foi descartado para obter o melhor resultado possível durante os testes, como atributos com valores iguais para quase todos os alunos, por exemplo, 85% dos alunos não trabalhavam, identificação do estado origem com 95% dos alunos do mesmo estado, atributos redundantes em diferentes tabelas, entre outros. Para o trabalho foram utilizados seis algoritmos de classificação: *Support Vector Machine*, *Logistic Regression*, *Locally-deep VSM*, *Decision Jungle*, *Neural network* e *Boosted Decision Tree*, sendo este último classificador o que apresentou o melhor desempenho para as métrica acurácia, revocação, precisão, medida-F_1 e AUC-ROC (*Area Under the Curve - Receiver Operating Characteristic*), através dos valores 0.964, 0.951, 0.943, 0.947 e 0.994 respectivamente. Na Tabela 2.5, são apresentados os valores obtidos para o preenchimento da matriz confusão.

Tabela 2.5: Matriz de confusão obtido do algoritmo *Two-Class Boosted Decision Tree*.

Matriz de Confusão		**Classe Real**	
		Positivo	**Negativo**
Classe Predita	**Positivo**	8.433	505
	Negativo	435	17.069

Fonte: Adaptado de Souza (2020).

O último trabalho relacionado Díaz et al. (2021), o autor faz a utilização de diferentes atributos para diversas áreas para obter o melhor resultado. A utilização de atributos iguais para alunos de diferentes cursos não resultam em acurácias parecidas, uma vez que cada área possui particularidades nas

características dos alunos. Por isso, o autor necessitou utilizar diferentes atributos para cada área. Alguns atributos são iguais para as 5 áreas, que são os créditos (aprovados, frequentados e matriculados) em 2011, também são comum os atributos créditos (aprovados, frequentados e matriculados) em 2010. O atributo frequência das aulas e crédito frequentado em 2010 está presente em todas as áreas, exceto para a área de Ciências da Saúde. O restante dos atributos está presente na Tabela 2.6. Neste trabalho, o autor utilizou seis algoritmos de classificação: C4.5, CART, Naive Bayes, *Multilayer Perceptron*, k-NN e SVM-*Poly*, sendo este último classificador o qual apresentou o melhor desempenho para a Medida-F_1 em todas as áreas: Artes e Humanas (0.92), Engenharia (0.94), Ciências da Saúde (0.99), Ciências (0.97), Ciências Sociais e Direito (0.91).

Tabela 2.6: Atributos específicos e em comum de cada área.

Área do Curso	**Atributos Específicos**	**Atributos em Comum**
Artes e Humanas	-Relação com o professor; -Escolaridade da mãe; -Relação nota/esforço;	-Créditos aprovados em 2011; -Créditos frequentados em 2011; -Créditos matriculados em 2011; -Créditos aprovados em 2010; -Créditos frequentados em 2010; -Créditos matriculados em 2010;
Engenharia	-Idade; -Nota; -Gênero; -Modalidade de admissão e nota; -Frequência das aulas;	
Ciências da Saúde	-Nota; -Razão da escolha do curso; -Crédito matriculado em 2010; -Gênero; -Modalidade de admissão e nota; -Relação nota/esforço; -Utilidade em atividade de integração;	
Ciências	-Gênero; -Modalidade de admissão e nota; -Prioridade; -Nota; -Frequência das aulas; -Participação nas atividades dos calouros;	
Ciências Sociais e Direito	-Bolsa em 2011; -Nota; -Gênero; -Modalidade de admissão e nota;	

Fonte: Adaptado de Díaz et al. (2021).

Pode-se destacar o trabalho Lykourentzou et al. (2009), o qual faz a utilização de combinações de técnicas de Aprendizado de Máquina. Neste trabalho a autora utiliza três classificadores que são: *Feed-Forward Neural Networks* (FFNN), *Support Vector Machine* (SVM) e *Probabilistic Ensemble Simplified Fuzzy* ARTMAP (PESFARM), em conjunto com a técnica conhecida como *ensemble*, o qual o trabalho chama de esquema. Esta técnica, é conhecida como *ensemble* de classificadores por voto majoritário Aggarwal (2014). Os atributos estão classificado em duas categorias: (i) invariáveis, que são gênero, residência, experiência de trabalho, nível educacional e (ii) atributos variáveis com o tempo, nota no teste de múltipla escolha, nota do projeto, data de submissão do pro-

jeto, após o prazo de envio e seção de atividade. O estudo é aplicado para dois cursos à distância, sendo que os cursos possuem 7 etapas de avaliações de múltipla escolha, o trabalho utiliza primeiramente apenas os atributos invariantes para a predição do aluno e posteriormente apenas os dados variantes com o tempo. O processo de predição é feito através dos três classificadores inicialmente. O aluno possui chances de evasão se: (i) ao menos um classificador considerar que o aluno irá evadir (esquema 1); (ii) se dois classificadores consideram que o aluno irá evadir (esquema 2); e (iii) se três classificadores considerarem que o aluno irá evadir (esquema 3). O texto apresenta a acurácia do *esquema 1*, durante a seção de atividade avaliativa 1 até a 4, como sendo a melhor dentre todos os classificadores e esquemas. A partir da seção de atividades de avaliação 5, o *esquema 1* e o *esquema 2* obtém a melhor classificação, juntamente com outro classificador. Com isso, a utilização do esquema 1 se destaca desde o início do curso até o final do curso. Vale ressaltar que o desempenho com outros classificadores e esquemas se assemelham. Os resultados do trabalho, é segmentado através do uso de atributos invariantes e variantes com o tempo. Os resultados utilizando atributos invariantes no tempo, foram para precisão geral, ou seja, ambas as classes, sensibilidade e precisão da classe evasão: $41 \sim 50\%$, $60 \sim 63\%$ e 43%, respectivamente. Já para os atributos variantes no tempo os resultados apresentados foram $75 \sim 82\%$, $70 \sim 74\%$ e $64 \sim 88\%$, sendo esses resultados alcançados na primeira etapa de avaliação. Na última avaliação do curso o *esquema 1* atinge 97%, $95 \sim 100\%$ e 100%.

Na Tabela 2.7, são apresentados um resumo das principais características como, atributos, país onde o trabalho foi realizado, volume da base de dados utilizado, algoritmos de classificação, esquema de avaliação, métrica de avaliação e os resultados obtidos dos cinco trabalhos relacionados selecionados durante a revisão bibliográfica .

Apensar da acurácia ser uma métrica muito utilizado, devido à sua simplicidade e de fácil interpretação, a utilização desta métrica em determinados casos não é adequado, ou seja, mesmo com taxas elevadas na acurácia a performance do modelo pode estar inadequado.

Na Tabela 2.8, é um exemplo de uma matriz de confusão para um modelo que classifica exames de câncer em positivo ou negativo. O modelo é ingênuo o suficiente para classifica todos os exames como negativo para câncer, com um conjunto de dados disposto por 1.000 exames, destes apenas 10 são positivos para câncer. Ao calcular a acurácia deste modelo, é obtido uma performance de 99%, que aparenta ser um bom resultado, entretanto o conjunto de dados desbalanceado resulta na falsa sensação de bons resultados.

Tabela 2.8: Exemplo de uma matriz de confusão para pacientes com câncer e sem câncer.

Matriz de Confusão		**Classe Real**	
		Positivo	**Negativo**
Classe Predita	**Positivo**	0	0
	Negativo	10	990

Fonte: Adaptado de Souza (2020).

Um outro comportamento que mereça atenção em relação à acurácia, pode-se destacar a atribuição de pesos iguais para ambos os erros, ou seja, quando classificado como falso positivo e falso negativo. Como exemplo, considerando o caso dos exames de câncer, suponha que o modelo acerte 950 exemplos, os outros 50 exemplos pertence à classe positiva (falso positivo) ou à negativa (falso negativo) resultando na acurácia de 95% para os dois casos. Entretanto um exame classificado como um falso negativo possui uma gravidade maior para este caso, tal característica não atentado à métrica. Como alternativa aos problemas relatados no uso da acurácia para calcular métricas de classificadores, pode-se elencar a análise do problema como *ranking*, desta maneira, é possível identificar o quão próximo de uma classe um determinado exemplo está.

Nos últimos anos surgiram novas áreas, conceitos e tecnologias, voltado para a manutenção do aluno no âmbito escolar, destaca-se o trabalho Romero e Ventura (2020), na qual descreve as novas áreas de pesquisa que surgiram baseadas nos dados educacionais. Na Figura 2.11 são apresentadas as diversas áreas correlatas envolvendo os dados educacionais, sendo a Mineração de Dados Educacionais (MDE)/Análise de Aprendizado (AA), do inglês *Education Data Mining* (EDM)/*Learning Analytics* (LA) uma área interdisciplinar que faz o uso combinado de outras três áreas, Ciências da Computação, Educação e Estatística, que também geram outras três subáreas intimamente tocante à EDM/LA. Responsável pela transformação de todo o dado primitivo em informações que possam ser utilizado para uma tomada de decisão em um processo educacional (Borges, 2017).

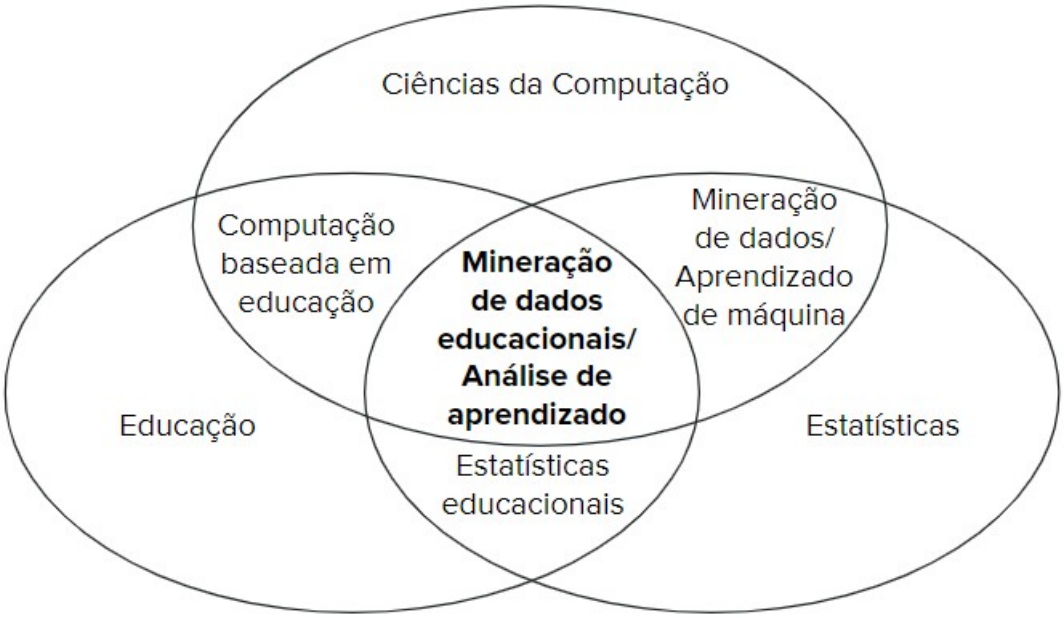
Na revisão sistemática na literatura sobre MDE, o trabalho de Bin Roslan e Chen (2022), apresenta uma busca entre o período de 2015 à 2021, 58 trabalhos publicados, entretanto nenhum de autoria brasileira. Contudo 49% dos trabalhos selecionados focam na identificação de atributos que influenciam o desempenho dos alunos, 42% no desempenho dos algoritmos de mineração de dados e apenas 9% relacionado à mineração de dados no sistema do tipo *e-Learning*, ou seja, no aprendizado não presencial. O autor identificou também a superioridade dos trabalhos imerso no ensino superior, constatou que a técnica de árvore de decisão foi a mais utilizada na mineração de dados e

identificou os atributos utilizados em cada trabalho.

Os pesquisadores Santos et al. (2021) trazem uma abordagem de análise dos trabalhos nacionais e internacionais na área de MDE, juntamente com suas ferramentas/bibliotecas e seus melhores algoritmos que apresentaram os melhores resultados. Consequentemente foram selecionados 22 trabalhos nacional e 28 internacional, destacando o uso da ferramenta *Weka* e a biblioteca *Scikit-learn*[2], também é observado que os trabalhos com os melhores resultados utilizaram pelo menos dois algoritmos de aprendizado de máquina, na qual os algoritmos *Decision Treee* e *Support Vector Machine* apresentam melhores desempenhos.

Já em Maschio et al. (2018) um mapeamento sistemático da literatura é apresentado sobre o cenário brasileiro na MDE desde 2001. Ao aplicar os critérios de seleção, resultaram em 49 trabalhos, dos quais a maioria dos trabalhos avaliam dados de alunos do ensino superior. É evidenciado, que o aprendizado de máquina, é a técnica mais utilizada nos trabalhos avaliados. Entretanto como a maior parte dos trabalhos são oriundos de ambientes virtuais, o tipo de dado mais relevante para as pesquisas são os números de interações na plataforma virtual.

Figura 2.11: Principais áreas da mineração de dados educacionais.



Fonte: Adaptado de Romero e Ventura (2020).

apresentar uma análise crítica discutindo o (mal) uso da acurácia e a vantagem de analisar o problema como ranking.

[2]Biblioteca específica para aprendizado de máquina, possui código aberto especificamente para a linguagem Python. https://scikit-learn.org/

Tabela 2.7: Trabalhos relacionados ao projeto e suas características.

Trabalho	***Características***	***Atributos***	***País***	***Volume Base de Dados***	***Algoritmo***	***Esquema de Avaliação***	***Métrica***	***Resultados***
Alvarez et al. (2020)	Atributos pré-registro	Gênero, Província, Fonte de ingresso, Índices acadêmicos anteriores, Nota em matemática no exame de admissão, Classificação de opções do curso;	Cuba	456 estudantes;	J48	Cross-Validation: 5-Fold	Acurácia	**Pré-registro:** 68,86%, **Primeiro Semestre:** 93.85%, **Primeiro ano:** 96.71%
Brito et al. (2014)	Nota do vestibular, Nota de disciplinas do primeiro período	**Pré-Registro:** Média Geral, Média de Matemática, Média de Física; **Primeiro período:** Cálculo Diferencial e Integral I, Física Aplicada a Computação I, Cálculo Vetorial, Geometria Analítica;	Brasil	300 estudantes;	*Naive Bayes*	*Cross-Validation*: 10-*fold*	Acurácia	75%
Antonio e Luiz (2020)	Dados socioeconômico	Faixa de frequência($0 \sim 25\%$, $25 \sim 50\%$, $50 \sim 75\%$, $75 \sim 100\%$), Frequência (comparecimento), Cidade, Instituição de ensino anterior, Cor/Raça, Renda familiar, Idade, Faixa etária, Tipo de instituição de ensino anterior, Gênero, Tipo de inscrição	Brasil	95 estudantes;	*Support Vector Machine*	*Cross-Validation*: 10-*fold*	Acurácia	$91,6\%$
Souza (2020)	Dados socioeconômico	Necessidade Especial, Sexo, Idade, Grupo Idade, Origem Escola Ensino Médio, Cidade Aluno, Estado Civil, Titulação, Cor Raça, Ocupação, Empregado, Faixa Renda Mensal, Origem Escola Ensino, Fundamental, Situação Aluno (Evadido ou não);	Brasil	26.442 estudantes	*Two-Class Boosted Decision Tree*	*Holdout*	Acurácia	$96,4\%$
Díaz et al. (2021)	Granularidade por área: Artes e Humanas; Engenharia e Arquitetura; Ciências da Saúde; Ciências; Direito e Ciências Sociais;	Os 10 melhores atributos para cada área são diferentes;	Espanha	1055 estudantes;	*Support Vector Machine-Poly*	*Cross-Validation*: 10-*fold*	Acurácia	**Artes e Humanas:**92%, **Engenharia e Arquitetura:** 94%, **Ciências da Saúde:** 99%, **Ciências:** 97%, **Direito e Ciências Sociais:** 91%

Fonte: Elaborado pelo autor.

CAPÍTULO

3

Método de Pesquisa

O método de pesquisa utilizado neste trabalho é o Crisp-DM, apresentado no Capítulo 2. Com o objetivo de seguir as seis fases do modelo, serão apresentadas a seguir as instanciações para cada etapa do processo.

3.1 Entendimento de Negócio

A primeira etapa do método de pesquisa, consistiu no entendimento do negócio. Nesta etapa do projeto, a proposta consiste na utilização de técnicas de mineração de dados com o propósito de reduzir o número de alunos que abandonam a universidade. Para isso, utilizou-se as informações educacionais dos estudantes como fonte de dados. Definiu-se como objetivo a criação de *ranking* de alunos com maiores chances de evasão, visando tomar medidas preventivas para minimizar a taxa de alunos evadidos ao longo dos anos.

Para atingir tal objetivo, iniciou-se a construção de um modelo de classificação, utilizando dados do histórico escolar os quais foram separados por área, campus e curso. Nesta etapa, considerou-se a utilização de diferentes tipos de algoritmos para a construção dos modelos de classificação, tanto visando uma melhor performance de classificação, para que possam ser incorporados em sistemas para a emissão de alertas e geração de relatórios, quanto os que visam a extração de conhecimento, como os algoritmos baseados em regras, para a apresentação dos padrões de evasão aos órgãos e pessoas relacionadas ao tema.

3.2 Entendimento dos Dados

Para o entendimento dos dados, segunda etapa do Crisp-DM, foi coletado os dados acadêmicos no Siscad. Esperava-se utilizar os dados do Ambiente Virtual de Aprendizagem (AVA) da UFMS, entretanto menos de 10% das disciplinas de graduação utilizavam o AVA até 2019. Vale ressaltar que para a utilização dos dados dos alunos da UFMS, foi necessário a aprovação no Comitê de Ética em Pesquisa com Seres Humanos (CEP). Para tal fim, foi necessário cadastrar o projeto de pesquisa na Plataforma Brasil[1]. Para esta pesquisa foi gerado o Certificado de Apresentação de Apreciação Ética (CAAE), identificação do projeto de pesquisa que entra para apreciação ética, CAAE 47952321.8.0000.0021. Após aprovação do CEP, foi gerado o número do parecer 4.955.047 com o *status* informando que o CEP da UFMS aprovou a pesquisa. Assim, posterior à autorização do CEP para uso dos dados dos acadêmicos da graduação da UFMS, foi solicitado os dados do Siscad para à Agência de Tecnologia da Informação e Comunicação (Agetic) e os dados do questionário socioeconômico à Pró-Reitoria de Assuntos Estudantis (Proaes). Com sucesso na obtenção dos dados de apenas uma das base de dados, o Siscad, sendo estes dados anonimizado conforme regulamentado pela Lei Geral de Proteção de Dados (LGPD), inciso IV do art. 7º da referida lei (Brasil, 2018).

Após a obtenção dos dados, ainda nesta etapa, foi necessário a organização e a documentação dos dados. Na Tabela 3.1 são apresentadas informações dos dados fornecidos pela Agetic. É possível observar o quantitativo de alunos, classificados por tipo de ingresso e suas respectivas porcentagens. Entretanto as modalidades **Vestibular, Programa de Avaliação Seriada Seletiva da UFMS (Passe)** e **SiSU**, possuem sub-categorias, essas são decorrentes das chamadas recorrentes das vagas não preenchidas. Vale ressaltar que a modalidade **Quero Ser UFMS**[2] não foi inserido na tabela, devido à forma de ingresso ter iniciado no primeiro semestre de 2021, sendo esta nova forma de ingresso amparada na utilização da nota do Vestibular, Exame Nacional de Ensino Médio (Enem) ou Passe, feito nos anos anteriores.

Por meio dos bancos de dados foram extraídos de cada aluno os seguintes atributos:

- Registro Geral do Acadêmico (RGA);
- Ano de ingresso;
- Pontuação de ingresso;
- Ano de saída;

[1] https://plataformabrasil.saude.gov.br/
[2] https://ingresso.ufms.br/formas-de-ingresso/quero-ser-ufms/

Tabela 3.1: Tipos de ingressos e respectivos quantitativo dos alunos na UFMS (2002/1-2020/2).

Tipo de Ingresso	**Total**	
Aluno especial	259	0,0449%
Portador de curso superior	3.995	6,929%
Revalidação de diploma	18	0,031%
Convênio cultural	54	0,094%
Mobilidade	107	0,186%
Movimentação interna compulsória	40	0,069%
Permuta	271	0,47%
Processo seletivo de reingresso	53	0,092%
Transferência	2.699	4,681%
Vestibular	9.845	17,075%
Refugiados	1	0,002%
Via judicial	290	0,503%
Via Passe	374	0,649%
Via SiSU	38.713	67,144%
Movimentação interna	938	1,627%
Total	**57.657**	**100,000%**

Fonte: Elaborado pelo autor.

- Tipo de saída;
- Tipo de entrada;
- Coeficiente com reprovação;
- Coeficiente rendimento;
- Percentual cursado;

Também foram utilizados as informações do curso e das disciplinas matriculadas pelos alunos, coletando as seguintes informações:

- Curso;
- Semestre;
- Ano;
- Disciplina;
- Notas das avaliações;
- Nota final da disciplina;
- Frequência;
- Situação;

3.3 Preparação dos Dados

De posse dos dados, foi feita a preparação da terceira etapa do Crisp-DM, a qual consiste em verificar as possíveis inconsistências nos dados do Siscad como notas ausentes, selecionando apenas os históricos completos e exclusões das colunas como o RGA, Curso, Semestre e Ano. Optou-se pela utilização da nota final da disciplina ao invés de todas as avaliações da disciplina, pois o método de avaliação empregado em cada disciplina não é padronizado, cada docente responsável pela disciplina escolhe a melhor formas de avaliação disponível (provas, seminários, trabalhos) para atribuir notas ao aluno.

Inicialmente utilizou-se o atributo falta, após análises, foi desconsiderado o uso pela não confiabilidade dos dados cadastrados, uma vez que muitos docentes cadastram as faltas no término da disciplina. Observou-se que há correlação entre frequência e nota, sendo redundantes em muitos cenários. Além disso, através das notas é possível identificar disciplinas "problemáticas" de um curso e que estão associadas com à evasão.

Outro processo para a transformação dos dados utilizados foi a técnica chamada *One Hot Encoding* para as disciplinas, responsável por binarizar os dados de entrada e alimentar os algoritmos de classificação em forma de vetor do espaço vetorial (Yu et al., 2020). Na Figura 3.1, são demonstrados exemplos de disciplinas após a aplicação da técnica.

Figura 3.1: Exemplo de Transformação de uma Tabela de Disciplinas Utilizando *One Hot Encoding*.

Disciplinas	Nota Final	Situação Final
Algoritmos e Programação I	4.5	Não Evasão
Algoritmos e Programação II	6.8	Não Evasão
Análise de Algoritmos	3.8	Evasão
Arquitetura de Computadores I	0.0	Evasão
Banco de Dados I	6.0	Não Evasão
Banco de Dados II	8.2	Não Evasão
Algoritmos e Programação I	4.2	Não Evasão
Banco de Dados II	7.6	Não Evasão

Algoritmos e Programação I	Algoritmos e Programação II	Análise de Algoritmos	Arquitetura de Computadores I	Banco de Dados I	Banco de Dados II	Nota Final	Situação Final
1	0	0	0	0	0	4.5	**Não Evasão**
1	0	0	0	0	0	6.8	**Não Evasão**
0	1	0	0	0	0	3.8	**Evasão**
0	0	1	0	0	0	0.0	**Evasão**
0	0	0	1	0	0	6.0	**Não Evasão**
0	0	0	0	1	0	8.2	**Não Evasão**
0	0	0	0	0	1	4.2	**Não Evasão**
0	0	0	0	0	1	7.6	**Não Evasão**

Fonte: Elaborado pelo autor.

O término da preparação dos dados é o mapeamento da situação final do acadêmico, onde de acordo com a definição da evasão, na primeira etapa do processo Crisp-DM, cada aluno é classificado como **"Evasão"** ou **"Não Evasão"** e consequentemente as disciplinas cursada por ele, conforme demos-

trado na Tabela 3.2. Como resultado, foram gerados 15 arquivos com os respectivos dados educacionais dos alunos de cada curso do CPTL.

Tabela 3.2: Tabela com a Classificação da Situação Final do Aluno.

Evasão	**Não Evasão**
Exclusão solicitada pelo aluno	Regularmente matriculado no período
Exclusão por desistência	Matrícula automática
Exclusão por reprovação	Exclusão por diplomação
Exclusão por jubilação	Matrícula em período especial
Exclusão por transferência para outra IES	Afastamento por mobilidade acadêmica no exterior
Transferência interna	Exclusão por conclusão de disciplinas
Exclusão por permuta de turno	Afastamento por mobilidade acadêmica interna
Exclusão por permuta de curso	Trancamento para regularização do ENADE
Exclusão por permuta institucional	Falecimento
	Exclusão via judicial
	Exclusão por ingresso irregular na instituição
	Matrícula para CCND
	Afastamento por trancamento de matrícula

Fonte: Elaborado pelo autor.

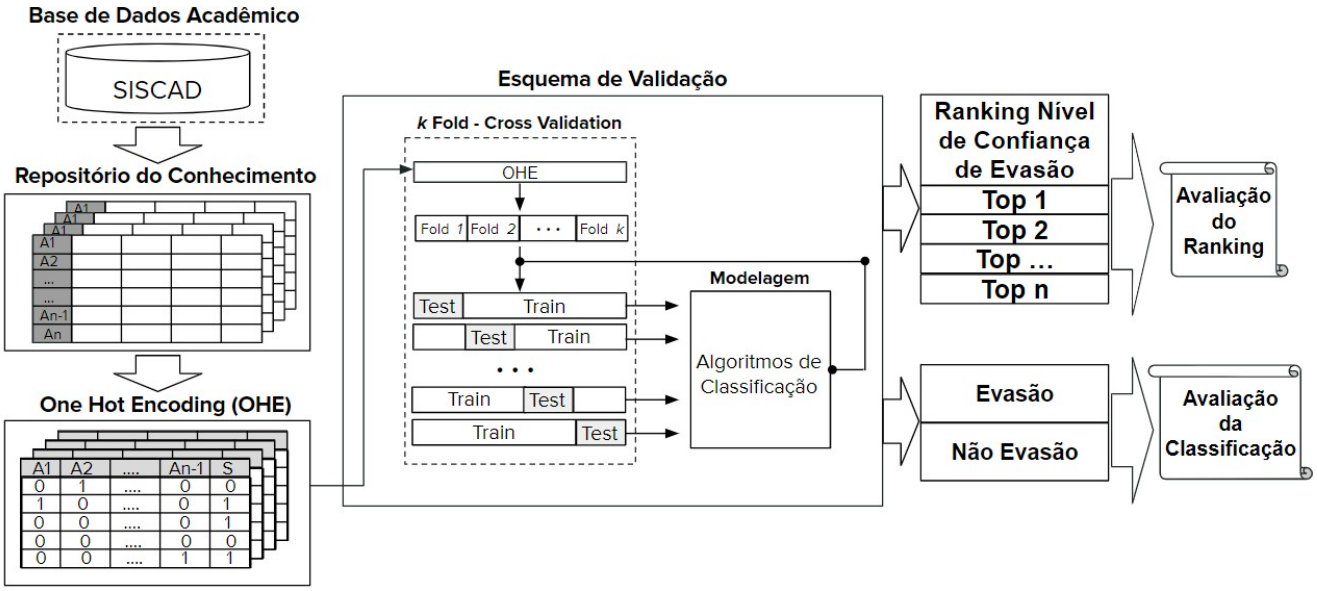
3.4 Modelagem

Após as devidas preparações dos dados, na fase de modelagem foi iniciado a construção do modelo proposto. Na Figura 3.2 é apresentado o desenho do projeto, na parte superior esquerdo da imagem, está localizado o banco de dados utilizado pelo sistema acadêmico, este banco de dados possui a maior parte da informação da vida universitária dos alunos da UFMS, na qual os dados solicitados serão utilizados para criar um repositório do conhecimento. Ainda na fase de preparação dos dados, ao repositório de conhecimento será aplicado a técnica de *One Hot Encoding*, esta técnica poderá ser executada anterior à validação cruzada, pois não será acrescentado nenhuma outra disciplina referente aos cursos. Vale ressaltar que a coleta dos dados acadêmicos é efetuado sempre após a finalização de um semestre, para que nenhuma disciplina excedente seja inserido no histórico do acadêmico e que possa alterar o resultado. Findo a preparação dos dados acadêmicos, foi gerado um arquivo no formato *Comma-separated values* (CSV). As informações consolidadas são utilizados pelos seguintes algoritmos de aprendizado de máquina e seus respectivos parâmetros, sendo estes elencados durante a revisão bibliográfica:

- SVM (Linear, *Poly* e RBF): $C = (10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2})$;
- *Multilayer Perceptron* (MLP): 32 e 64, 32;
- Regressão Logística (LR): $C = (10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2})$;
- $K-$vizinhos mais próximos (k-NN): $N = (7, 9, 11, 13)$;

- Árvores de Decisão (DT): $max\_depth = (2, 5, None)$.

Figura 3.2: Esboço proposto do projeto.

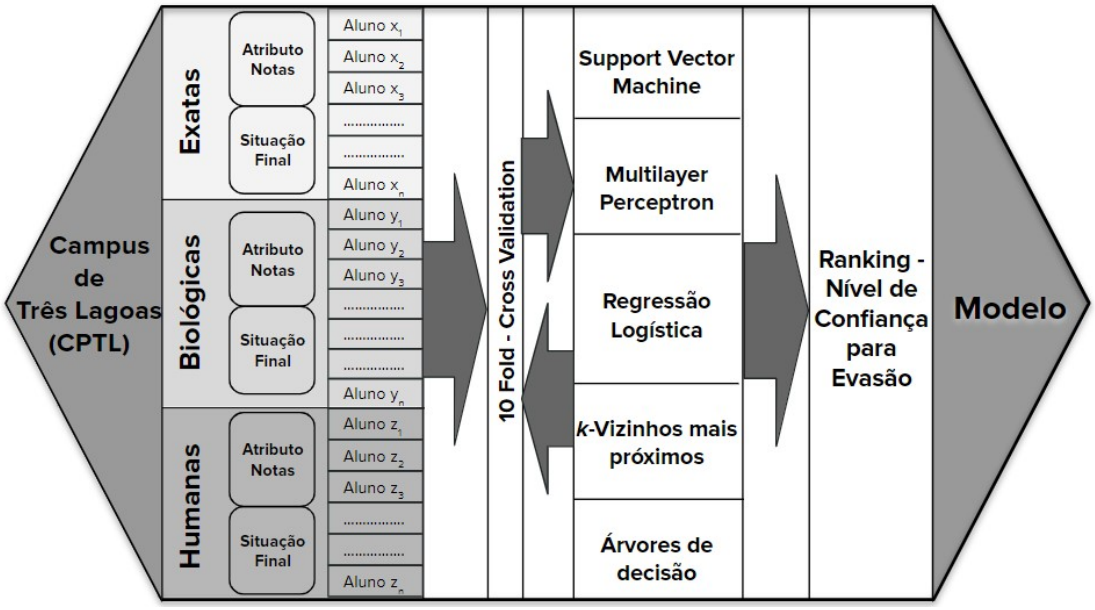


Fonte: Elaborado pelo autor.

Na Figura 3.3 é demonstrada através uma visão macro, como será o modelo para o ranqueamento. Para cada curso das grandes áreas, serão utilizados as notas finais de cada disciplina cursada pelo ano e a situação final do aluno perante o curso, aplicado para todos os campus da instituição. O trabalho utiliza como esquema de validação, 10 *Fold - Cross Validation* juntamente com os 5 algoritmos apresentados na subseção 2.2.4, para finalmente gerar o ranqueamento dos alunos. O código da implementação do modelo está disponível na plataforma de hospedagem de código *GitHub*[3]. A codificação foi desenvolvida na linguagem *Python*, por ser uma linguagem fácil de aprender e amplamente utilizado em ciências de dados (Raschka et al., 2020).

[3]https://github.com/gregoriohigashikawa/Ranking.git

Figura 3.3: Objetivo geral do projeto.



Fonte: Elaborado pelo autor.

3.5 Avaliação

Os classificadores utilizaram como esquema de avaliação o k-*Fold Cross-Validation*, para este trabalho foi adotado o valor para $k = 10$.

Para a avaliação foram utilizadas 16 métricas, dentre elas a acurácia, precisão, revocação, *macro* revocação, *macro* precisão, *macro* F_1 e F_1. Optou-se por não utilizar *micro* F_1 por apresentar valores iguais à acurácia (Takahashi et al., 2021). Já entre as métricas de avaliação do *ranking*, as mais importantes para o projeto são: precisão@k, revocação@k, F_1@k, onde $k = 5, 10, 50$, correspondendo ao *ranking* dos *top* k alunos com maior probabilidade de evasão.

3.6 Aplicação

Após alcançar performances satisfatória com modelo de classificação, o trabalho tem como o próximo objetivo desenvolver uma aplicação do modelo, última etapa do Crisp-DM. Serão duas partes referentes a aplicação do modelo. A primeira consiste em desenvolver um serviço que interaja com o Siscad-Admin, tanto para a coleta de dados, quando para a geração de relatórios oriundo das classificação baseada no *ranking*. Será discutido com a equipe de desenvolvimento da Agetic a melhor forma de implementação deste módulo. A segunda etapa da aplicação consiste na extração de conhecimento, na tentativa de obter regras ou padrões descritivos para o entendimento do que leva o aluno a evadir. Esse conhecimento extraído será apresentado e comparti-

lhado aos interessados, como professores, coordenadores de cursos, diretores, pró-reitores e reitor.

Devido à alta demanda do setor de Desenvolvimento de *Software* da Agetic, foi desenvolvido uma aplicação *web* simples do modelo de classificação, baseada no *ranking* dos alunos de acordo com o nível de confiança da classe evasão. Essa aplicação utilizou o *Flask*[4], um *microframework* desenvolvido em Python para aplicações *web*.

Com uma interface simples, o usuário irá apenas anexar uma lista no formato CSV contendo os históricos dos alunos na qual deseja verificar o ranqueamento. O modelo de classificação utilizado na aplicação, identifica o curso na qual os alunos estão relacionados para então selecionar o modelo de aprendizado de máquina específico do curso. Em seguida é feito a transformação do arquivo de entrada em binarizado, utilizando a técnica de *One Hot Encodming*. Assim, a aplicação está pronto para calcular a probabilidade de evasão de cada aluno baseado no histórico acadêmico. Ao final da execução, a aplicação gera uma listagem contendo o *ranking* de evasão de cada aluno e disponibilizado para *download*. Atualmente a aplicação utiliza os arquivos no formato CSV para a entrada e saída dos dados, posteriormente será adaptada para interagir com o usuário através do formato de dado *JavaScript Object Notation* (JSON) [5], um dos formatos de dados mais popular utilizado na troca de dados estruturado entre servidor e aplicação *web* (Pezoa et al., 2016).

[4]https://flask.palletsprojects.com/en/2.2.x/
[5]https://www.json.org/json-en.html

CAPÍTULO

4

Avaliação Experimental

Neste capítulo são descritos os experimentos e apresentados os resultados obtidos durante a execução do modelo proposto neste trabalho. Ao finalizar a preparação dos dados, escolher os melhores algoritmos de classificação e seus atributos, o modelo está pronto para ser executado. Assim, o modelo proposto apresentou bons resultados na avaliação do *ranking* por utilizar poucos atributos do aluno.

4.1 Configuração Experimental

Para o experimento foi selecionado o campus de Três Lagoas, sendo a base de dados fracionada por cursos, conforme Tabela 4.1, totalizando 8.915 alunos, destaca-se a área de Humanas, que concentra quase a metade dos alunos do campus com 8 cursos. Também é possível observar o quantitativo de alunos que evadiram e concluíram os cursos. Atualmente o CPTL possui 14 cursos de graduação[1] e foram encontrados 15 cursos cadastrados no banco de dados do Siscad. Vale ressaltar que o elevado número de disciplinas nos curso é ocasionado devido ao aluno se matricular em disciplinas de diferentes áreas para completar as disciplinas optativas.

Como forma de registro histórico da evasão dos últimos anos (2011 2020), a Figura 4.1, contém o registro do quantitativo de alunos evadidos a cada semestre. Para tal período, o número médio de evasão por ano corresponde à $400,9$ alunos. Espera-se após a implantação da aplicação deste projeto e com devidas aplicações de políticas para combater à evasão, os indicadores de

[1] https://cptl.ufms.br/

Tabela 4.1: Quantitativo de alunos e disciplinas por curso do CPTL.

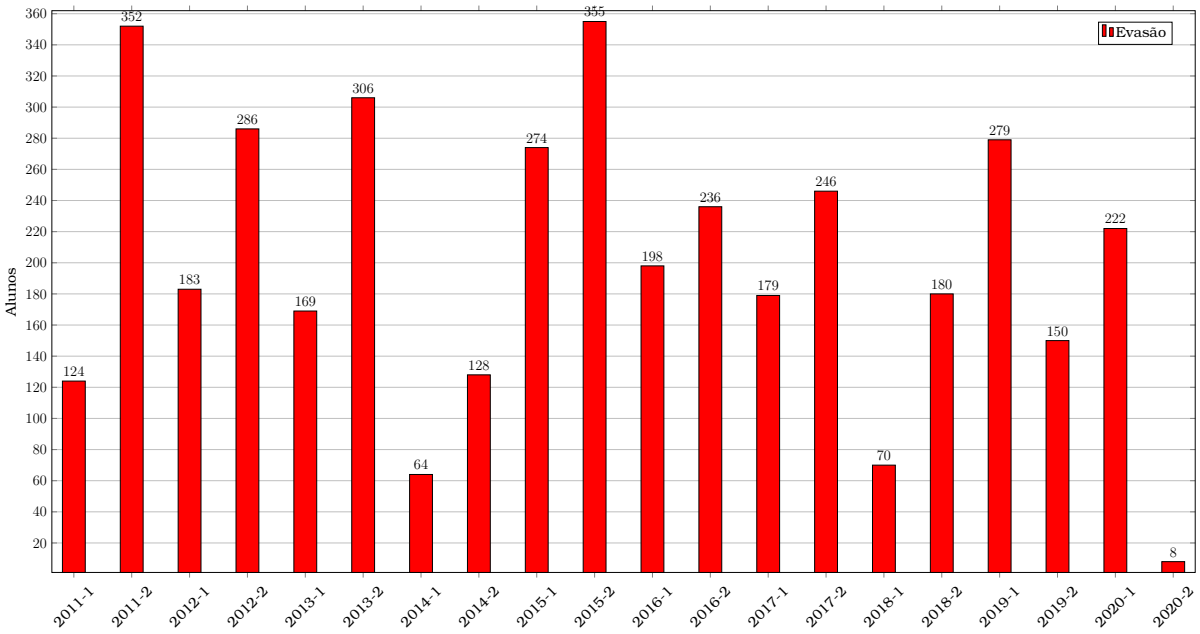
	Curso	Disciplinas	Alunos	Evadiu	Não Evadiu
Exatas	Sistemas de Informação	88	622	345 (55,47%)	277 (44,53%)
	Matemática	88	626	369 (58,95%)	257 (41,05%)
	Engenharia de Produção	145	680	280 (41,18%)	400 (58,82%)
Biológicas	Enfermagem	142	572	161 (28,15%)	411 (71,85%)
	Ciências Biológicas	150	522	207 (39,66%)	315 (60,34%)
	Medicina	101	477	48 (10,06%)	429 (89,94%)
Humanas	Ciências Contábeis	145	731	334 (45,69%)	397 (54,31%)
	Direito	170	1.807	380 (21,03%)	1.427 (78,97%)
	Geografia	123	625	302 (48,32%)	323 (51,68%)
	Letras	102	284	183 (64,44%)	101 (35,56%)
	Letras - Português/Espanhol	91	210	73 (34,76%)	137 (65,24%)
	Letras - Português/Literatura	108	364	156 (42,86%)	208 (57,14%)
	Letras - Português/Inglês	88	271	85 (31,37%)	186 (68,63%)
	História	132	561	195 (34,76%)	366 (65,24%)
	Pedagogia	113	563	262 (46,54%)	301 (53,46%)
Total de alunos			8.915	3.380 (37,91%)	5.535 (62,09%)

Fonte: Elaborado pelo autor.

evasão do Cptl possam reduzir consideravelmente com o decorrer do tempo.

Os algoritmos utilizados para o experimentos estão descritos na Seção 3.4, assim como os respectivos parâmetros de cada classificador e o esquema de validação utilizado para o modelo de classificação. Já nas métricas de avaliação, na Seção 3.5 é demonstrado quais são as métricas utilizadas para a avaliação de classificação e para a avaliação do *ranking*.

Figura 4.1: Dados de evasão do Cptl.



Fonte: UFMS (2020b)

4.2 Resultados e Discussões

A seguir serão apresentados os resultados do experimento para os cursos da áreas de Exatas, Biológicas e Humanas. Os valores apresentados são resultados dos melhores valores alcançados durante a execução dos algoritmos de classificação exposto na Seção 3.4.

Cursos de Exatas

A área de Exatas possui apenas três cursos e, em geral, com os resultados ligeiramente inferior dentre as grades áreas. Já entre os cursos de Exatas, Sistemas de Informação apresenta as menores métricas. Por outro lado, os valores apresentados na avaliação do *ranking* estão todos acima de $0,94$. Já na métrica de avaliação da classificação a acurácia apresenta valores entre $0,72$ à $0,83$ entre os cursos da área.

Observa-se uma particularidade no curso de Sistemas de Informação, na qual possui uma disciplina chamada Algoritmos e Programação I, com o índice de matrícula por aluno de $1,76$, considerado muito elevado em relação à outras disciplinas de outros cursos de áreas diferentes. Os índice de matrícula é calculado pela equação: número de matrículas da disciplina divido pelo número total de aluno, conforme apresentado nas Tabelas dos Apêndices A.1, B.1 e C.1 para todos os cursos. Assim, compreende-se que no curso de Sistemas de Informação, a disciplina de Algoritmos e Programação I, foi cursado por cada aluno pelo menos $1,76$ vezes.

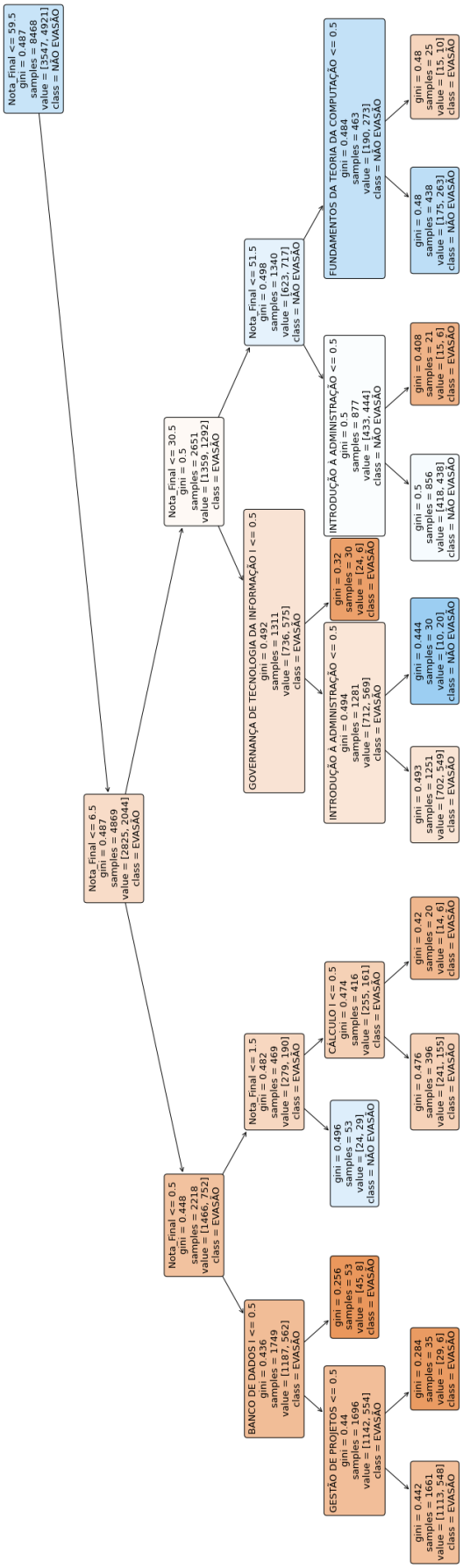
Tabela 4.2: Métricas de avaliação dos cursos da área de Exatas.

		Exatas		
		Sistemas de Informação	**Matemática**	**Engenharia de Produção**
Avaliação da Classificação	**Acurácia**	0,723475	0,774378	0,834059
	Precisão	0,680363	0,758478	0,680138
	Revocação	0,794681	0,845093	0,511132
	Macro Revocação	0,713976	0,773040	0,704467
	Macro Precisão	0,717778	0,772637	0,769709
	Macro $\mathbf{F}_1$	0,714486	0,772552	0,718408
	$\mathbf{F}_1$	0,683237	0,753151	0,543229
Avaliação do *Ranking*	**Precisão** $@k = 5$	1,000000	1,000000	0,980000
	Precisão $@k = 10$	0,990000	1,000000	0,980000
	Precisão $@k = 50$	0,948000	0,996000	0,984000
	Revocação $@k = 5$	1,000000	1,000000	1,000000
	Revocação $@k = 10$	1,000000	1,000000	1,000000
	Revocação $@k = 50$	1,000000	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 5$	1,000000	1,000000	0,988889
	$\mathbf{F}_1$ $@k = 10$	0,994737	1,000000	0,989474
	$\mathbf{F}_1$ $@k = 50$	0,972772	0,997980	0,991899

Vale ressaltar que a disciplina mais cursada no curso de Sistemas de Informação não é representado como um nó na árvore de decisão, ou seja, obter nota baixa ou reprovar nesta disciplina, não é um fator decisivo para a evasão

do aluno. Na Figura 4.2 é apresentado apenas o lado esquerdo da representação da árvore de decisão para o curso de Sistemas de Informação, em que as folhas com a coloração laranja mais escuro tende a ser mais puro para a classe evasão. Ao analisar os nós, conclui-se que os alunos que reprovam na disciplina de Banco de Dados I, à maioria evadem o curso, por outro lado, os alunos que são aprovado em Banco de Dados I, mas reprova em Gestão de Projetos, tendem a evadir o curso. Assim como aqueles alunos que possuem nota final inferior à $30,5$ e reprovam na disciplina de Governança de Tecnologia da Informação I, tendem a evadir. Para a geração da árvore de decisão para todos os cursos do CPTL, foi utilizado como atributos *criterion* = *gini*, ou seja, o critério utilizado para gerar a árvore de decisão é concentrar num ramo a classe mais frequente e *min_samples_leaf* = 20, significa que a folha deve conter no mínimo 20 amostrar para se tornar uma folha.

Figura 4.2: Representação Parcial da Árvore de Decisão do Curso de Sistemas de Informação.



Fonte: Elaborado pelo autor.

Já na avaliação do *ranking*, o menor valor foi de $0,94$, sendo este valor apresentado no curso de Sistemas de Informação para a métrica precisão para $k = 50$.

Cursos de Biológicas

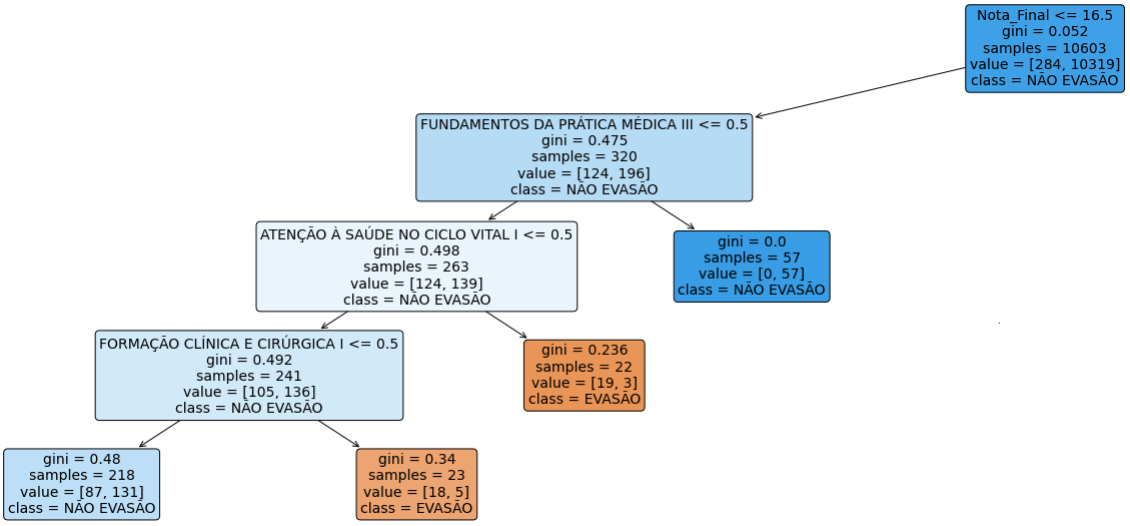
Na área de Biológicas, os três cursos em geral apresentam superioridade nas métricas de avaliação da classificação em relação aos cursos de Exatas, mesmo com quantitativo de alunos inferior. As exceções foram as métricas revocação e F_1, que apresentaram valores a baixo dos $0,5$ para todos os cursos da área de Biológicas. Para as métricas da avaliação do *ranking*, os valores apresentados ficaram acima dos $0,98$ para todas as métricas. Destaca-se o curso de Medicina que além de ter o menor percentual de reprovação, apresentou 100% para todas as métricas na avaliação do *ranking* e também o melhor resultado na acurácia, $0,97$, dentre todos os cursos do CPTL.

Tabela 4.3: Métricas de avaliação dos cursos da área de Biológicas.

		Biológicas		
		Enfermagem	**Ciências Biológicas**	**Medicina**
Avaliação da Classificação	**Acurácia**	0,839015	0,792916	0,978402
	Precisão	0,675047	0,650701	0,792222
	Revocação	0,323675	0,404172	0,344828
	Macro Revocação	0,641235	0,661015	0,669943
	Macro Precisão	0,762999	0,734389	0,885962
	Macro F_1	0,668648	0,679065	0,719504
	F_1	0,431153	0,490494	0,450417
Avaliação do *Ranking*	**Precisão** $@k = 5$	1,000000	1,000000	1,000000
	Precisão $@k = 10$	0,990000	1,000000	1,000000
	Precisão $@k = 50$	0,984000	0,986000	1,000000
	Revocação $@k = 5$	1,000000	1,000000	1,000000
	Revocação $@k = 10$	1,000000	1,000000	1,000000
	Revocação $@k = 50$	1,000000	1,000000	1,000000
	F_1 $@k = 5$	1,000000	1,000000	1,000000
	F_1 $@k = 10$	0,994737	1,000000	1,000000
	F_1 $@k = 50$	0,991815	0,992888	1,000000

A árvore de decisão gerado para o curso de Medicina, apresentado na Figura 4.3, demonstra uma grande diferença em comparação ao curso de Sistemas de Informação, por concentrar apenas em uma ramificação as folhas das evasões. A subárvore com o nó Atenção à Saúde no Ciclo Vital I, possui o nó folha mais pura da árvore de decisão, e a medida que se percorre esta subárvore os nós folhas vão se tornando menos puro para a evasão. A medida que era gerado mais níveis na árvore de decisão para o curso de medicina, à partir de 6 níveis, a árvore resultante limitava-se a apenas 4 nós para a classe evasão.

Figura 4.3: Representação Parcial da Árvore de Decisão do Curso de Medicina.



Fonte: Elaborado pelo autor.

Cursos de Humanas

Nos cursos de Humanas a avaliação da classificação, a métrica acurácia manteve o valor mínimo de $0,78$, apresentado no curso de Geografia, e valor máximo de $0,91$ identificado no curso de Direito. Vale destacar que curso de Direito obtém o maior número de alunos do CPTL, 1.087, e o maior número de disciplinas dentre todas as áreas. Já o curso de Letras - Português/Inglês, apresentou o menor valor para a precisão, $0,64$, sendo o curso com o segundo menor quantitativo de alunos. Já na avaliação do *ranking*, o curso de Pedagogia alcançou 100% para todas as métricas igualando o feito do curso de Medicina. Dentre todos os cursos de Humanas, o menor valor alcançado na avaliação do *ranking* foi de $0,96$, oriundo do curso de Letras.

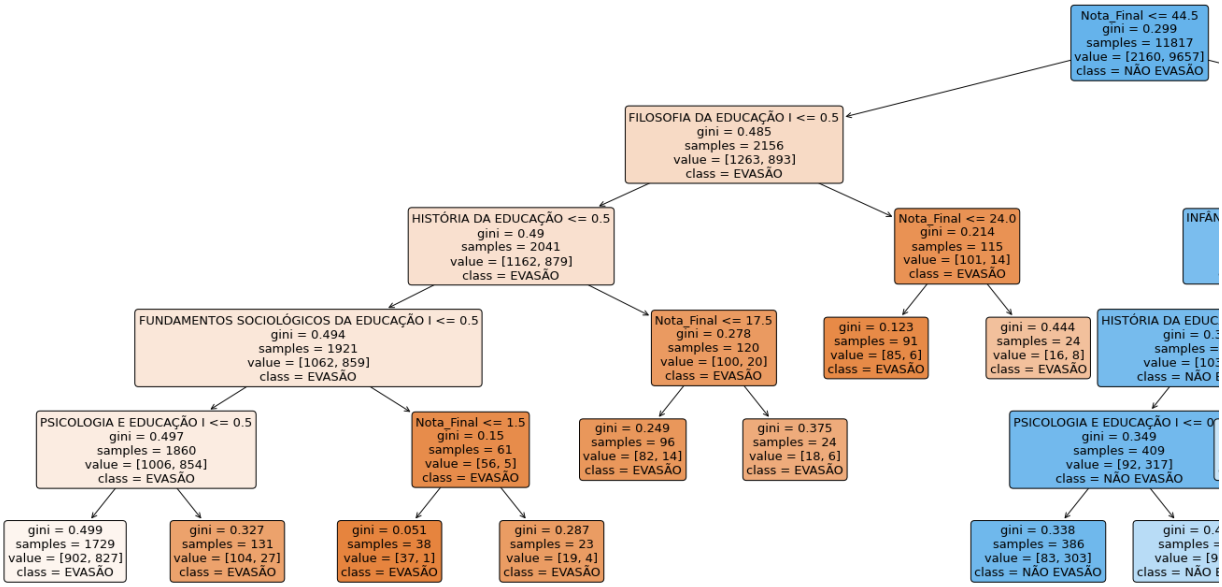
Tabela 4.4: Métricas de avaliação dos cursos da área de Humanas.

		Humanas								
		Ciências Contábeis	**Direito**	**Geografia**	**Letras**	**Letras - Português/ Espanhol**	**Letras - Português/ Literatura**	**Letras - Português/ Inglês**	**História**	**Pedagogia**
Avaliação da Classificação	**Acurácia**	0,791409	0,915387	0,785507	0,800278	0,838213	0,824802	0,870492	0,795681	0,871457
	Precisão	0,714709	0,783755	0,789582	0,818407	0,650905	0,727070	0,643965	0,710566	0,745383
	Revocação	0,777047	0,349360	0,628821	0,717854	0,418861	0,625806	0,480912	0,712266	0,533333
	Macro Revocação	0,777047	0,656117	0,728677	0,774545	0,676626	0,728013	0,709085	0,757802	0,729631
	Macro Precisão	0,777047	0,847346	0,787044	0,802234	0,753125	0,785916	0,770146	0,768726	0,816730
	Macro $\mathbf{F}_1$	0,716869	0,670925	0,741633	0,779781	0,697470	0,736768	0,724955	0,751996	0,756695
	$\mathbf{F}_1$	0,575031	0,394680	0,641430	0,720246	0,493732	0,589282	0,529647	0,666271	0,589793
Avaliação do *Ranking*	**Precisão** $@k = 5$	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
	Precisão $@k = 10$	1,000000	1,000000	0,980000	0,980000	0,990000	0,980000	1,000000	1,000000	1,000000
	Precisão $@k = 50$	1,000000	0,998000	0,964000	0,960000	0,962000	0,968000	0,998000	0,988000	1,000000
	Revocação $@k = 5$	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
	Revocação $@k = 10$	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 5$	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 10$	1,000000	1,000000	0,989474	0,989474	0,994737	0,989474	1,000000	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 50$	0,991840	0,998990	0,981568	0,979355	0,980516	0,983673	0,998990	0,993919	1,000000

A árvore de decisão do curso de Pedagogia, Figura 4.4, é apresentado uma subárvore com a disciplina chamada Filosofia da Educação I, como a principal disciplina para a identificação de outras folhas categorizados como evasão.

Na Tabela 4.5 são apresentados as médias das métricas de avaliação dos

Figura 4.4: Representação Parcial da Árvore de Decisão do Curso de Pedagogia.



Fonte: Elaborado pelo autor.

cursos de cada área correspondente, os valores médios apresentados são obtidos dos melhores resultados dentre os algoritmos e seus parâmetros correspondentes de cada curso, em seguida calcula-se a média entre os cursos de cada área. É possível observar que a precisão da classe de interesse, ficou acima de 70% para todas as áreas, e parcialmente a revocação da classe de interesse obteve valores acima de 70%. E na medida F_1 da classe de interesse, não foram alcançando valores satisfatórios, pois se mantiveram a baixo dos $0,66$. Em se tratando de métricas *micro-averaging*, as classes agregaram valores superiores à 77%, cujos resultados são decorrentes dos valores obtidos na classe majoritária (não evasão). Já quando dá-se peso iguais para ambas as classes (*macro-averaging*), os resultados são em geral superiores à 70%. Portanto ambos os resultados denotaram um maior acerto em influência da classe majoritária. Já quando se combinam a Precisão e Revocação, os resultados são em geral superior à 70% (considerando pesos iguais para ambas as classes - *macro-averaging*), ou superiores à 77% com pesos proporcionais ao número VPs para cada classe (*micro-averaging*).

Em se tratando do problema com o ranqueamento, a precisão, revocação e F_1 apresentam valores sempre superiores à 97% e 100% para $k = 5$ nas áreas de Biológicas e Humanas. Com isso, percebe-se que o tratamento da abordagem de *ranking* ao invés de classificação para o problema de evasão escolar é mais adequado, uma vez que as precisões, revocações, ou combinação de ambas obtiveram resultados próximos a 100%. Posto isso, tem-se mais acertos nas previsões de alunos que irão evadir e um melhor redirecionamento de esforços.

Tabela 4.5: Média dos Resultados por Área.

	Métrica de Avaliação	Exatas	Biológicas	Humanas
Avaliação da Classificação	**Acurácia**	0,777304	0,870111	0,832581
	Precisão da Classe de Interesse	0,706326	0,705990	0,731594
	Revocação da Classe de Interesse	0,716969	0,357558	0,582696
	Macro Revocação	0,730494	0,657398	0,726394
	Macro Precisão	0,753375	0,794450	0,789813
	Macro $\mathbf{F}_1$	0,735149	0,689072	0,730788
	$\mathbf{F}_1$ da Classe de Interesse	0,659872	0,457355	0,577790
Avaliação do *Ranking*	**Precisão** $@k = 5$	0,993333	1,000000	1,000000
	Precisão $@k = 10$	0,990000	0,996667	0,992222
	Precisão $@k = 50$	0,976000	0,990000	0,982000
	Revocação $@k = 5$	1,000000	1,000000	1,000000
	Revocação $@k = 10$	1,000000	1,000000	1,000000
	Revocação $@k = 50$	1,000000	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 5$	0,996296	1,000000	1,000000
	$\mathbf{F}_1$ $@k = 10$	0,994737	0,998246	0,995907
	$\mathbf{F}_1$ $@k = 50$	0,987550	0,994901	0,989872

Fonte: Elaborado pelo autor.

Assim, a manutenção do aluno pode se tornar específica de acordo com a posição do *ranking*, ou seja, para um aluno com chances eminente de evasão, deve-se ter muito mais atenção para mitigar a chance de evasão, e para um aluno com baixa porcentagem de chance de evasão, utilizaria-se métodos mais simples para a manutenção.

Na Tabela 4.6, são apresentados os melhores classificadores de cada curso e seus respectivos parâmetros para a medida *precision* para $k = 50$. Observa-se que o classificador SVC e LR destacam-se em quase todos os cursos, variando apenas os parâmetros.

4.3 Ferramenta Computacional para o Ranqueamento de Alunos com Potencial de Evasão.

Na Figura 4.5, é apresentado a tela da aplicação desenvolvida para calcular as chances de evasão dos alunos de uma determinada turma. É possível observar as opções para o envio e submissão do arquivo para que a aplicação possa processar os dados. Após clicar no botão *submit*, a aplicação automaticamente retorna logo à baixo duas planilhas para *download*, uma sendo o arquivo originário e a outra uma planilha resultante do processamento com o *ranking* de evasão, conforme demonstrado na Figura 4.6. Assim, a aplicação torna possível a consulta, através de uma planilha, para todos os interessados em visualizar o nível de confiança de evasão de uma determinada turma de

Tabela 4.6: Os Melhores Classificadores e Parâmetros por Curso Segundo Métrica *Precision* $@k = 50$.

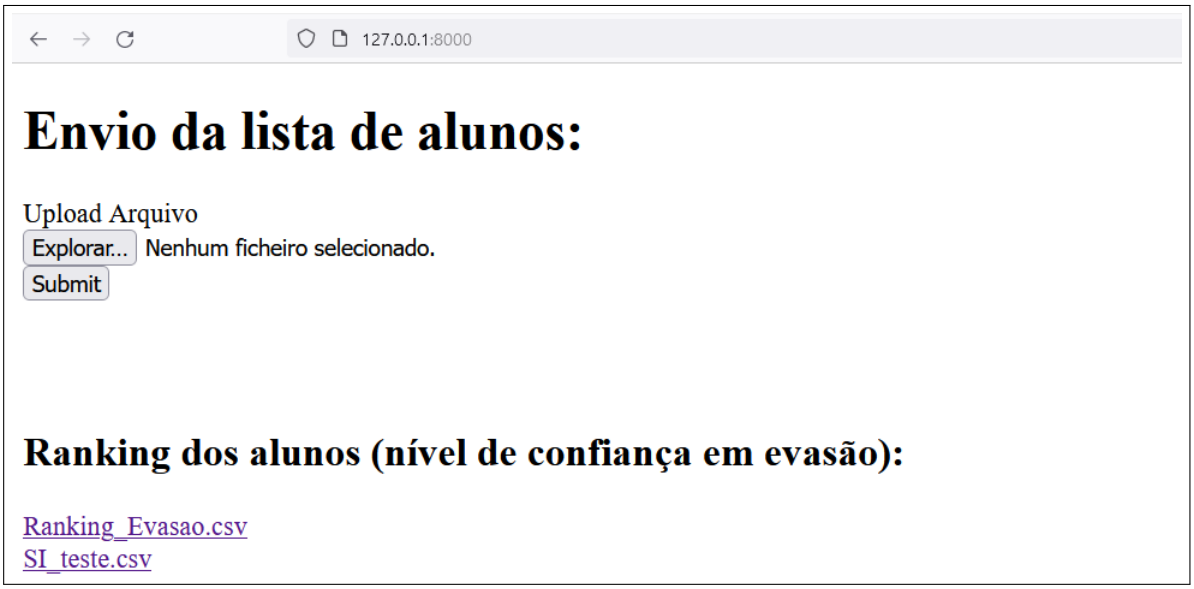
		Curso	**Melhor Classificador**
Avaliação do *Ranking*	**Exatas**	Sistemas de Informação	SVC-*Poly*, $C = 10^2$
		Matemática	MLP $(64, 32)$
		Engenharia de Produção	SVC-*Poly*, $C = 10^0$
	Biológicas	Enfermagem	SVC-*Poly*, $C = 10^{-1}$
		Ciências Biológicas	SVC-*Poly*, $C = 10^0, 10^2$
		Medicina	MLP (32), $(32, 64)$ LR, $C = 10^-2$, $C = 10^-1$, $C = 10^0$, $C = 10^1$, $C = 10^2$
	Humanas	Ciências Contábeis	SVC-RBF, $C = 10^1$
		Direito	MLP (32), $(64, 32)$
		Geografia	MLP (32) LR, $C = 10^1$, $C = 10^2$ SVC-*Linear*, $C = 10^1$, $C = 10^2$
		Letras	LR, $C = 10^{-1}$
		Letras - Português/Espanhol	LR, $C = 10^{-1}$
		Letras - Português/Literatura	LR, $C = 10^1$
		Letras - Português/Inglês	SVC-*Linear*, $C = 10^0$, $C = 10^2$
		História	LR $C = 10^1$
		Pedagogia	MLP (32), $(64, 32)$ LR $C = 10^{-2}$, $C = 10^{-1}$, $C = 10^0$, $C = 10^1$, $C = 10^2$ k-NN-*Euclidean* $= 7$, 9, 11, 13 SVC-*Poly*, $C = 10^{-2}$, $C = 10^{-1}$, $C = 10^0$, $C = 10^1$

Fonte: Elaborado pelo autor.

alunos. A aplicação utiliza modelos de específicos de cada curso, na qual os algoritmos de classificação e seus respectivos parâmetros são fundamentadas na Tabela 4.6.

Vale ressaltar que a aplicação considera um serviço desenvolvido em *Flask* e que carrega os modelos selecionados bem como técnicas de pré-processamento utilizadas na avaliação experimental. Porém, a aplicação pode ser facilmente adaptado para atuar como uma API REST (*Application Programming Interface Representational State Transfer*) para receber requisições HTTP (*Hypertext Transfer Protocol*) com conteúdo em formato JSON (*JavaScript Object Notation*) e também responder no mesmo formato, facilitando assim a integração com outros sistemas (Relan, 2019).

Figura 4.5: Imagem da Aplicação do Modelo.



Fonte: Elaborado pelo autor.

Figura 4.6: Imagem da Tabela com o *Ranking* de Evasão dos Alunos.

	A	B	C
1	Name	Evasão(%)	
2	Auno 13	0.9556144105206453	
3	Auno 12	0.9164714439499088	
4	Auno 11	0.9089834314526649	
5	Auno 6	0.8999591773990903	
6	Auno 1	0.8890774136063935	
7	Auno 3	0.8578485308326563	
8	Auno 10	0.7911485540937807	
9	Auno 4	0.7794789159524206	
10	Auno 9	0.7523306426924701	
11	Auno 2	0.7523035145537399	
12	Auno 14	0.7523026215312995	
13	Auno 7	0.7522954791113063	
14	Auno 8	0.7297926601303142	
15	Auno 5	0.6702371182687606	

Fonte: Elaborado pelo autor.

CAPÍTULO

5

Conclusões

A evasão escolar é um problema conhecido mundialmente desde ensinos fundamentais até cursos superiores, sendo inúmeros os motivos que levam os alunos a evadirem, como cursos, características individuais, e métodos de ensino. A evasão causa prejuízos de tempo e financeiros, não só o aluno, mas também a instituição de ensino.

Elencar os motivos de forma geral ou intuitiva podem ser passíveis de erro e gerando, portando, soluções ineficientes. No encargo de extração de conhecimento de base de dados, as instituições vêm adotando técnicas de mineração de dados, sendo essa estratégia ainda não utilizada pela UFMS e sendo o objetivo deste trabalho. Assim, este trabalho torna possível identificar alunos com maiores chances de evasão, utilizando apenas os atributos de nota e a situação final, também apresenta variações dos melhores algoritmos de classificação para cada curso do CPTL. Este trabalho propõe uma nova forma de identificar os alunos com mais chances de evasão através da tarefa de ranquear os alunos de acordo com a confiança para a classe "evade", analisando as assertivas dos modelos nos *top-k* alunos do *ranking*. O experimento nos dados educacionais dos alunos do CPTL obteve bons resultados, com valores médios entre os cursos superiores à $0,97$ para a área de Exatas, $0,99$ para os cursos de Biológicas e $0,98$ para a área de Humanas.

Após o término deste projeto, espera-se que a performance de classificação alcançados pelo modelo, possa ser integrado ao Siscad-Admin para a geração de relatórios e conhecimento, os quais poderão ser acessados por professores, coordenadores e diretores. Espera-se também que com base nos resultados apresentados, medidas preventivas posam ser adotadas na tentativa da diminuição da evasão. Por fim, caso este projeto tenha efeitos positivos, poderá

ser um modelo para outras universidades para combater a evasão acadêmica, consequentemente, ajudando mais pessoas à alcançarem o objetivo de possuir uma formação no curso superior.

5.1 Contribuições

Foi apresentado durante uma reunião os resultados obtidos deste trabalho. Estavam presentes representantes da Pró-Reitoria de Planejamento (Proplan) e da Pró-Reitoria de Graduação (Prograd). Foi despertado o interesse na aplicabilidade das técnicas utilizadas deste trabalho junto ao Siscad, com cobertura em todos os cursos e geração do relatório do *ranking* dos alunos com mais chances de evasão por curso. Houve uma extensa discussão das possibilidades de intervenção da universidade baseada no ranqueamento provido pelos modelos obtidos nesse trabalho. Vale ressaltar que devido à extensa área de cobertura, com diversas frentes necessárias para a mitigação da evasão, o trabalho tornaria pioneiro com abrangência para todos os cursos da graduação na modalidade presencial. Foi compreendido pelos interessados que o trabalho é apenas uma ferramenta para o auxílio na tomada de decisão, que as intervenções são baseadas nas ocorrências do aluno, não sendo utilizado nenhum instrumento para auxiliar na tomada de decisão antecipada. Tão logo, após a adoção deste projeto, a UFMS estaria evoluindo no tangente a manutenção do aluno no ambiente escolar de forma antecipada. Assim, abre-se um leque com várias possibilidades de intervenções que setores responsáveis podem agir de forma mais assertiva.

5.2 Limitações

Para a execução do trabalho, seriam utilizadas três bases de dados, AVA, Siscad e Questionário Socioeconômico. Entretanto, antes de solicitar os dados, era necessário a autorização do CEP, processo responsável por consumir uma boa parte do tempo. Após a devida autorização, as interações com a Agetic para obtenção dos dados do Siscad percorreu longos períodos até a consolidação da base de dados. Já nos dados do sistema Questionário Socioeconômico, o setor responsável não enviou de forma satisfatória os dados solicitados. Já no AVA, os cursos não utilizavam em sua totalidade o ambiente virtual, embora a situação tenha mudado a partir do ano de 2020 devido à pandemia, logo, optou por não utilizar a base de dados do AVA. Assim, o projeto beneficiou-se apenas da base de dados do Siscad dos alunos do campus de Três Lagoas, correspondendo à menos de 15% do total dos alunos da UFMS. A execução para todos os campus tornaria o dinamismo do trabalho inviável

devido ao alto volume de dados e escassez de tempo, mesmo o campus de Três Lagoas sendo o segundo maior campus da UFMS, com $14,85\%$ dos alunos, ainda há uma grande diferença da população estudantil com o campus de Campo Grande que corresponde à $54,14\%$ do total de alunos da UFMS.

5.3 Trabalhos Futuros

Com o objetivo de ampliar a utilização de técnicas de mineração de dados no âmbito institucional, pode-se sugerir trabalhos futuros como:

- Categorizar o *ranking* de evasão em grupos de riscos como por exemplo alto, médio e baixo;
- Identificação das disciplinas que possuem um maior impacto para à evasão do aluno. Diante dessas informações, os projetos políticos-pedagógicos dos cursos poderão sofrer alterações fundamentadas nas estratégias das disciplinas "evasores de alunos".
- Como alternativa de algoritmos de classificação para este projeto, destaca-se a árvore de decisão baseado em técnicas adaptativas Pistori e Neto (2002), com o objetivo das características do aluno evasor se adaptando à medida que a aplicação é alimentado com novos dados. Também para respeitar a linha temporal dos dados, pode-se aplicar a técnica de *Walk-Forward Validation*, ou seja, utilizar os dados das disciplinas do primeiro ano apenas para turmas do primeiro ano, dados do segundo ano para turmas do segundo ano e assim por diante, assim tornando mais realista a predição de evasão.
- Identificação do perfil do aluno que evade: através dos dados do questionário socioeconômico, traçar quais características dos alunos que abandonam o curso mais se destacam.
- Predição de evasão interna, ou seja, o aluno não deixa a instituição, entretanto migra para um outro curso, seja na mesma área ou não. Identificar quais são as características que motivam o aluno a mudar de curso.
- Aplicar técnicas baseadas em aprendizado simbólico para extrair regras interpretáveis.
- Criar uma fonte única de dados educacionais e utilizar técnicas de mineração de dados para rastrear as tendências dos alunos com chances de evasão, auxiliando tomada de decisão e manutenção do acadêmico, utilizando como referência um trabalho publicado anteriormente que desen-

volveu um Modelo de Referência de Dados Educacionais (EDRM) Borges (2017).

- Desenvolver, junto à Agetic, aplicativo para visualização do perfil do aluno egresso, perfil do aluno em situação de risco, *ranking* dos alunos com chances de evasão e acompanhamento do comportamento da vida acadêmica. Iniciativa semelhante ao da Universidade Federal de Santa Maria, com o projeto Integra UFSM[1].

- Mapear cursos oferecidos pelas outras instituições concorrentes à UFMS e analisar os impactos relacionados à baixa procura pelos cursos e até mesmo à desistência do curso.

Estes são algumas das possibilidades de trabalho voltado para o âmbito educacional com o objetivo de formar o aluno, identificando antecipadamente dificuldades que tornariam o sonho de terminar um curso superior impossível.

[1] https://www.ufsm.br/orgaos-suplementares/cpd/ufsm-integra

APÊNDICE

A

Disciplinas com Mais Matrículas: Cursos de Exatas

Na Tabela A.1 são apresentados o número de matrículas por disciplinas de cada curso da área de Exatas, na última coluna nomeada como Matrícula / Nº Total de Alunos, representa o quantitativo de vezes que um aluno do curso se matriculou na respectiva disciplina.

Tabela A.1: Disciplinas mais matriculadas dos cursos de Exatas.

	Curso	Matrículas	Matrícula / Nº Total de Alunos
Exatas	**Sistemas de Informação**		
	ALGORITMOS E PROGRAMAÇÃO I	1.093	1,76
	INTRODUÇÃO A SISTEMAS DIGITAIS	734	1,18
	FUNDAMENTOS DA TEORIA DA COMPUTAÇÃO	626	1,01
	ALGORITMOS E PROGRAMAÇÃO II	421	0,68
	FUNDAMENTOS DE TECNOLOGIA DA INFORMAÇÃO	416	0,67
	Matemática		
	INTRODUÇÃO AO CÁLCULO I	658	1,05
	HISTÓRIA E FILOSOFIA DA MATEMÁTICA	604	0,97
	ELEMENTOS DE GEOMETRIA	595	0,95
	EDUCAÇÃO ESPECIAL	584	0,93
	PRÁTICA DE ENSINO DE MATEMÁTICA I	561	0,90
	Engenharia de Produção		
	CÁLCULO I	906	1,33
	FÍSICA I	758	1,11
	QUÍMICA GERAL	732	1,08
	GEOMETRIA ANALÍTICA	701	1,03
	INTRODUÇÃO À CIÊNCIA DA COMPUTAÇÃO	699	1,03

Fonte: Elaborado pelo autor.

APÊNDICE

B

Disciplinas com Mais Matrículas: Cursos de Biológicas

Na Tabela B.1 são apresentados o número de matrículas por disciplinas de cada curso da área de Biológicas, na última coluna nomeada como Matrícula / Nº Total de Alunos, representa o quantitativo de vezes que um aluno do curso se matriculou na respectiva disciplina.

Tabela B.1: Disciplinas mais matriculadas dos cursos de Biológicas.

	Curso	Matrículas	Matrícula / nº Total de Alunos
Biológicas	**Enfermagem**		
	BIOQUÍMICA	472	0,83
	ANATOMIA HUMANA I	453	0,79
	SAÚDE E SOCIEDADE	445	0,78
	FISIOLOGIA I	422	0,74
	ANATOMIA HUMANA II	414	0,72
	Ciências Biológicas		
	BIOLOGIA CELULAR I	687	1,32
	GENÉTICA BÁSICA	590	1,13
	MATEMÁTICA	580	1,11
	GEOLOGIA	543	1,04
	GENÉTICA APLICADA	476	0,91
	Medicina		
	BASES BIOLÓGICAS DA PRÁTICA MÉDICA I	470	0,99
	BASES PSICOSSOCIAIS DA PRÁTICA MÉDICA I	414	0,87
	BASES PSICOSSOCIAIS DA PRÁTICA MÉDICA II	371	0,78
	BASES BIOLÓGICAS DA PRÁTICA MÉDICA III	364	0,76
	BASES BIOLÓGICAS DA PRÁTICA MÉDICA II	359	0,75

Fonte: Elaborado pelo autor.

APÊNDICE

C

Disciplinas com Mais Matrículas: Cursos de Humanas

Na Tabela C.1 são apresentados o número de matrículas por disciplinas de cada curso da área de Humanas, na última coluna nomeada como Matrícula / Nº Total de Alunos, representa o quantitativo de vezes que um aluno do curso se matriculou na respectiva disciplina.

Tabela C.1: Disciplinas mais matriculadas dos cursos de Humanas.

	Curso	Matrículas	Matrícula / nº Total de Alunos
Humanas	**Ciências Contábeis**		
	CONTABILIDADE INTRODUTÓRIA I	735	1,01
	MATEMÁTICA FINANCEIRA	590	0,81
	CONTABILIDADE INTRODUTÓRIA II	590	0,81
	MATEMÁTICA	552	0,76
	METODOLOGIA DE PESQUISA	495	0,68
	Direito		
	DIREITO PENAL I	1530	1,41
	HISTÓRIA DO DIREITO	1379	1,27
	DIREITO CIVIL I	1308	1,20
	INTRODUÇAO AO ESTUDO DO DIREITO	1292	1,19
	FILOSOFIA	1188	1,09
	Geografia		
	INTRODUÇÃO À CIÊNCIA GEOGRÁFICA	460	0,74
	GEOLOGIA GERAL	386	0,62
	FUNDAMENTOS DE CLIMATOLOGIA	364	0,58
	FUNDAMENTOS DE DIDÁTICA	356	0,57
	GEOGRAFIA URBANA	340	0,54
	Letras		
	PRÁTICA DE PRODUÇÃO DE TEXTOS EM LÍNGUA PORTUGUESA	259	0,91
	INTRODUÇAO À LINGUÍSTICA	225	0,79
	INTRODUÇÃO À TEORIA LITERÁRIA	214	0,75
	LITERATURA E SOCIEDADE	214	0,75
	INTRODUÇÃO À PESQUISA	211	0,74
	Letras - Português/Espanhol		
	INTRODUÇÃO À TEORIA LITERÁRIA	214	1,02
	GRAMÁTICA NORMATIVA DA LÍNGUA PORTUGUESA	213	1,01
	LÍNGUA ESPANHOLA I	207	0,99
	INTRODUÇAO A PESQUISA	204	0,97
	LITERATURA E SOCIEDADE	202	0,96
	Letras - Português/Literatura		
	INTRODUÇÃO À TEORIA LITERÁRIA	309	0,85
	PRÁTICA DE PRODUÇAO DE TEXTOS EM LÍNGUA PORTUGUESA	304	0,83
	LITERATURA E SOCIEDADE	299	0,82
	INTRODUÇAO A PESQUISA	297	0,81
	INTRODUÇAO À LINGUÍSTICA	288	0,79
	Letras - Português/Inglês		
	LÍNGUA INGLESA I	266	0,98
	GRAMÁTICA NORMATIVA DA LÍNGUA PORTUGUESA	264	0,97
	INTRODUÇAO À TEORIA LITERÁRIA	263	0,97
	LITERATURA E SOCIEDADE	258	0,95
	INTRODUÇÃO À PESQUISA	255	0,94
	História		
	HISTÓRIA DA AMÉRICA PORTUGUESA I	455	0,81
	HISTÓRIA DA AMÉRICA COLONIAL	442	0,79
	ANTIGUIDADE ORIENTAL	437	0,78
	PRÁTICA DE ENSINO E PESQUISA EM HISTÓRIA MULTICULTURALISMO, POVOS INDÍGENAS E DIVERSIDADE	401	0,71
	HISTÓRIA DA AMÉRICA PORTUGUESA II	357	0,63
	Pedagogia		
	PSICOLOGIA E EDUCAÇÃO I	478	0,85
	HISTÓRIA DA EDUCAÇÃO	445	0,79
	FILOSOFIA DA EDUCAÇÃO I	410	0,73
	INFÂNCIA E SOCIEDADE	405	0,72
	TRABALHO ACADÊMICO	395	0,70

Fonte: Elaborado pelo autor.

Referências Bibliográficas

Aggarwal, C. (2014). *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press. Citado na página 28.

Aggarwal, C. (2015). *Data Mining: The Textbook*. Springer International Publishing. Citado na página 3.

Aggarwal, C. C. et al. (2018). Neural networks and deep learning. *Springer*, 10:978–3. Citado na página 15.

Alvarez, N. L., Callejas, Z., e Griol, D. (2020). Predicting computer engineering students' dropout in cuban higher education with pre-enrollment and early performance data. *Journal of Technology and Science Education*, 10(2):241–258. Citado nas páginas 4, 25, e 32.

Amo, S. D. (2004). Técnicas de mineração de dados. *Jornada de Atualização em Informática na Educação*, páginas 1–43. Citado na página 7.

Antonio, R. B. W. e Luiz, G. O. (2020). Socioeconomic data mining and student dropout. *International Journal for Innovation Education and Research*, páginas 1–15. Citado nas páginas 4, 15, 17, 27, e 32.

Baker, R. S. J. D. e Yacef, K. (2009). The state of educational data mining in 2009 : A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17. Citado na página 3.

Berens, J., Schneider, K., Görtz, S., Oster, S., e Burghoff, J. (2019). Early detection of students at risk-predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Data Mining*, 11(3):1–41. Citado na página 3.

Bin Roslan, M. H. e Chen, C. J. (2022). Educational data mining for student performance prediction: A systematic literature review (2015-2021).

International Journal of Emerging Technologies in Learning (iJET), 17(05):pp. 147–179. Citado na página 30.

Borges, A. (2017). *Definição de um modelo de referência de dados educacionais para a descoberta de conhecimento*. PhD thesis, Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-09112018-103702/publico/VanessaAraujoBorges.pdf. Acesso em mai. de 2022. Citado nas páginas 30 e 56.

Braga, L. (2005). *Introdução à Mineração de Dados - 2a edição: Edição ampliada e revisada*. e-papers. Citado na página 8.

BRASIL (2007). Decreto no 6.096. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6096.htm>. Acesso em: 12 de abr. de 2021. Citado na página 4.

Brasil (2018). Lei 13.709 de 14 de agosto de 2018. *Diário Oficial da República Federativa do Brasil, 15 ago. 2018*. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709>. Acesso em 29 mai. de 2022. Citado na página 34.

BRASIL, C.-G. d. U. (2018a). Portal da transparência do governo federal. Disponível em: <http://www.portaltransparencia.gov.br/funcoes/12-educacao>. Acesso em: 10 de abr. de 2021. Citado na página 2.

BRASIL, Ministério da educação, S. d. E. S. (2014). Balanço social sesu 2003-2014. Disponível em <http://portal.mec.gov.br/busca-geral/191-secretarias-112877938/sesu-478593899/20954-arquivos-sesu>. Acesso em: 9 de mai. de 2021. Citado na página 1.

BRASIL, Ministério da educação, S. d. E. S. (2019). Análise crítica sobre os indicadores de gestão das instituições federais de ensino superior 2016. Disponível em <http://portal.mec.gov.br/docman/marco-2019-pdf/110171-analise-critica-indicadores-tcu-2018/file. Citado na página 1.

BRASIL, Ministério da educação, S. e. (2018b). Apuração do custo das universidades federais, e sua relação com os respectivos quantitativos de alunos. Disponível em: <http://www.forplad.andifes.org.br/sites/default/files/forplad/comissaoplanejamento/NT_04-2018_e_anexos_-_

apura%C3%A7%C3%A3o_do_custo_das_universidades.pdf>. Acesso em: 9 de abr. de 2021. Citado na página 1.

Brito, D. M. d., Júnior, I. A. d. A., Queiroga, E. V., e Rêgo, T. G. d. (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 25(1):882–890. Citado nas páginas 4, 26, e 32.

Bunkar, K., Singh, U. K., Pandya, B., e Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. In *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, páginas 1–5. Citado na página 18.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., e Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66:541–556. Citado na página 17.

Calixto, K., Segundo, C., e Gusmão, R. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 28(1):1447. Citado na página 8.

Camilo, C. e Silva, J. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, páginas 1–29. Citado na página 8.

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., e Marinho, T. (2013). Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. *Anais da Jornada de Atualização em Informática na Educação*, 1(1):1–29. Citado na página 3.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., e Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73:247–256. Citado na página 14.

Cunha, A. M., Tunes, E., e Silva, R. R. d. (2001). Evasão do curso de Química da universidade de Brasília: a interpretação do aluno evadido. *Química Nova*, 24:262 – 280. Citado na página 2.

Cutler, A., Cutler, D. R., e Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, páginas 157–175. Springer. Citado na página 18.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., e Zingaro, S. P. (2020). Student dropout prediction. In *International Conference on Artificial Intelligence in Education*, páginas 129–140. Springer. Citado na página 14.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506. Citado nas páginas 14 e 15.

Díaz, I., Bernardo, A. B., Esteban, M., e Rodríguez-Muñiz, L. J. (2021). Variables influencing university dropout: A machine learning-based study. In Herrero, Á., Cambra, C., Urda, D., Sedano, J., Quintián, H., e Corchado, E., editors, *The 11th International Conference on European Transnational Educational - 2020)*, páginas 94–103, Cham. Springer International Publishing. Citado nas páginas 4, 14, 27, 28, e 32.

Filho, F. H., Siqueira, D., e Leal, B. (2020). Predição de evasão utilizando técnicas de classificação: Um estudo de caso do instituto federal do Ceará. In *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*, páginas 141–148, Porto Alegre, RS, Brasil. SBC. Citado nas páginas 14 e 17.

Gardner, M. e Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14):2627–2636. Citado nas páginas 15 e 16.

Gevert, V. G., da Silva, A. C. L., Gevert, F., e Ales, V. T. (2010). Modelos de regressão logística, redes neurais e support vector machine (svm's) na análise de crédito a pessoas jurídicas. *RECEN-Revista Ciências Exatas e Naturais*, 12(2):269–293. Citado na página 13.

Gilioli, R. d. S. P. (2016). *Evasão em instituições federais de ensino superior no Brasil: expansão da rede, Sisu e desafios*. Câmara dos Deputados, Consultoria Legislativa. Consultoria Legislativa da Área XV - Educação, Cultura e Desporto. https://bd.camara.leg.br/bd/handle/bdcamara/28239, acesso em 09/04/2021. Citado na página 1.

Gonçalves, O. e Beltrame, W. (2020). Socioeconomic data mining and student dropout: analyzing a higher education course in Brazil. *International Journal for Innovation Education and Research*, 8:505–518. Citado na página 14.

Gottardo, E., Kaestner, C., e Noronha, R. (2012). Previsão de desempenho de estudantes em cursos ead utilizando mineração de dados: uma estratégia baseada em séries temporais. *Simpósio Brasileiro de Informática na Educação*, 23(1):1–10. Citado nas páginas 4 e 20.

Guo, G., Wang, H., Bell, D., Bi, Y., e Greer, K. (2003). Knn model-based approach in classification. In Meersman, R., Tari, Z., e Schmidt, D. C., editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, páginas 986–996, Berlin, Heidelberg. Springer Berlin Heidelberg. Citado na página 17.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., e Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18. Citado na página 26.

Han, J., Pei, J., e Kamber, M. (2011). *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. Citado na página 3.

Haykin, S. (2011). *Neural Networks and Learning Machines*. Pearson Education. Citado nas páginas 14 e 15.

Hoed, R. M. (2016). Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação. Dissertação (Mestrado Profissional em Computação Aplicada - Universidade de Brasília). Citado na página 2.

Hofmann, M. (2006). Support vector machines—kernels and the kernel trick. *Universität Bamberg*. <https://cogsys.uni-bamberg.de/teaching/ss07/hs_rc/slides/SVM_Seminarbericht_Hofmann.pdf>. Acesso em: 12 de abr. de 2021. Citado na página 13.

Jesus, H. d., Rodriguez, L., e Junior, A. C. (2021). Predição de evasão escolar na licenciatura em Computação. *Revista Brasileira de Informática na Educação*, 29(0):255–272. Citado na página 4.

Júnior, O. d. G. F., Rodrigues, W. R. M., Barbirato, J. C., e Costa, E. d. B. (2019). Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de Maceió. *RENOTE*, 17(1):296–305. Citado na página 4.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., e Klein, M. (2002). *Logistic regression*. Springer. Citado na página 16.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference of Artificial Intelligence*, páginas 1–7. Citado na página 20.

Lamb, S. (2011). *Pathways to School Completion: An International Comparison*, páginas 21–73. Springer Netherlands, Dordrecht. Citado na página 1.

Lee, C.-P. e Lin, C.-J. (2014). Large-scale linear ranksvm. *Neural Computation*, 26(4):781–817. Citado na página 9.

Lemay, D. J., Baek, C., e Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, 2:100016. Citado na página 3.

Lenon, G., Silva, B., Carvalho, J. A., Magno, A., e Maciel, A. (2021). Desenvolvimento de um learning analytics dashboard para modelos de mineração de dados educacionais. *Revista de Engenharia e Pesquisa Aplicada*, 6(3):59–69. Citado na página 4.

Liu, T. (2011). *Learning to Rank for Information Retrieval*. Springer Berlin Heidelberg. Citado na página 8.

López, C. (2021). *DATA MINING. The CRISP-DM METHODOLOGY. The CLEM language and IBM SPSS MODELER*. Lulu.com. Citado nas páginas 10 e 12.

Lorena, A. C. e de Carvalho, A. C. P. L. F. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67. Citado na página 14.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., e Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education*, 53(3):950–965. Citado nas páginas 3 e 28.

Manhães, L. M. B. e da Cruz, S. M. S. (2020). Predição do desempenho acadêmico de alunos da graduação utilizando mineração de dados. *pt. In: Simpósio Pesqui. Operacional e Logística da Mar.-Publicação Online. São Paulo: Editora Blucher*, páginas 2050–2064. Citado na página 14.

Manhães, L., da Cruz, S., Costa, R., Zavaleta, J., e Zimbrão, G. (2012a). Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. In *Anais do VIII Simpósio Brasileiro de Sistemas de Informação*, páginas 284–295, Porto Alegre, RS, Brasil. SBC. Citado na página 14.

Manhães, L. M. B., Serra da Cruz, S. M., Macário Costa, R. J., Zavaleta, J., e Zimbrão, G. (2012b). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Anais do XXII SBIE XVII WIE*, volume 1. Citado na página 4.

Marban, O., Mariscal, G., e Segovia, J. (2009). A data mining & knowledge discovery process model. In Ponce, J. e Karahoca, A., editors, *Data Mining*

and Knowledge Discovery in Real Life Applications, chapter 1. IntechOpen, Rijeka. Citado na página 9.

Marques, L. T. (2020). Mateo: Uma abordagem de descoberta de conhecimento para desvendar as causas da evasão escolar. Dissertação (Mestrado em Ciência da Computação) - Uni-versidade Federal Rural do Semi-Árido. Disponível em: <https://repositorio.ufersa.edu.br/bitstream/prefix/5425/1/LeonardoTM_DISSERT.pdf>. Acesso em: 12 de abr. de 2021. Citado nas páginas 2 e 14.

Martins, L. C. B., Carvalho, R. N., Carvalho, R. S., Victorino, M. C., e Holanda, M. (2017). Early prediction of college attrition using data mining. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, páginas 1075–1078. Citado na página 17.

Maschio, P., Vieira, M., Costa, N., Melo, S., e Júnior, C. (2018). Um panorama acerca da mineração de dados educacionais no Brasil. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 29(1):1936. Citado na página 31.

Meira, C. A., Rodrigues, L. H., e Moraes, S. A. (2008). Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. *Tropical Plant Pathology*, 33:114 – 124. Citado na página 17.

Meyer, D. e Wien, F. T. (2001). Support vector machines. *R News*, 1(3):23–26. Citado na página 13.

Monard, M. C. e Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In *Sistemas inteligentes: fundamentos e aplicações*, páginas 89–114. Manole Ltda, Barueri-SP, 1 edition. Citado na página 19.

Mueen, A., Zafar, B., e Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11):36–42. Citado na página 20.

Márquez-Vera, C., Romero Morales, C., e Ventura Soto, S. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14. Citado na página 3.

Nascimento, R. L. S. d., Junior, G. G. d. C., e Fagundes, R. A. d. A. F. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do INEP. *RENOTE*, 16(1). Citado nas páginas 3 e 8.

Noriega, L. (2005). Multilayer perceptron tutorial. *School of Computing. Staffordshire University*. Citado na página 14.

Paz, F. e Cazella, S. (2017). Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 6(1):624. Citado na página 4.

Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., e Vrgoč, D. (2016). Foundations of json schema. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pagina 263–273, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. Citado na página 40.

Pistori, H. e Neto, J. J. (2002). Adaptree-proposta de um algoritmo para indução de árvores de decisão baseado em técnicas adaptativas. In *Anais Conferência Latino Americana de Informática-CLEI*. Citado na página 55.

Provost, F. e Fawcett, T. (2018). *Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados*. Alta Books. Citado na página 9.

RAMESH, V., PARKAVI, P., e Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. *International Journal of Computer Applications*, 63:975–8887. Citado na página 15.

Raschka, S., Patterson, J., e Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 11(4):193. Citado na página 38.

Relan, K. (2019). *Building REST APIs with Flask: Create Python Web Services with MySQL*. Apress. Citado na página 50.

Rigo, S., Cambruzzi, W., Barbosa, J., e Cazella, S. (2014). Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, 22(01):132. Citado na página 15.

Rodrigues, R. L., De Medeiros, F. P. A., e Gomes, A. S. (2013). Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 24(1):607–616. Citado na página 3.

Romero, C. e Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355. Citado nas páginas 30 e 31.

Rossi, R. (2011). Representação de coleções de documentos textuais por meio de regras de associação. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo. Citado na página 20.

Rovira, S., Puertas, E., e Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLOS ONE*, 12(2):1–21. Citado na página 17.

Santana, M. A., de Barros Costa, E., dos Santos Neto, B. F., Silva, I. C. L., e Rego, J. B. (2015). A predictive model for identifying students with dropout profiles in online courses. In *EDM (Workshops)*. Citado na página 14.

Santos, V., Saraiva, D., e Oliveira, C. (2021). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, páginas 1196–1210, Porto Alegre, RS, Brasil. SBC. Citado na página 31.

Shafique, U. e Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222. Citado na página 10.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22. Citado nas páginas 11 e 12.

Shiba, M. H., Santos, R. L., Quintanilha, J. A., e Kim, H. Y. (2013). Classificação de imagens de sensoriamento remoto pela aprendizagem por árvore de decisão: uma avaliação de desempenho. In *Anais XII Simpósio Brasileiro de Sensoriamento Remoto*, páginas 4319–4326, Goiânia, Goiás, Brasil. Citado na página 17.

Silva, M. B. S. d. e Franco, V. S. (2014). Um estudo sobre a evasão no curso de física da universidade estadual de Maringá: modalidade presencial versus modalidade a distância. *Revista Brasileira de Aprendizagem Aberta e a Distância*, 13:337–360. Citado na página 2.

Smola, B., Soentpiet, R., Schölkopf, B., Burges, C., Mika, S., Smola, A., e Scholkopf, M. (1999). *Advances in Kernel Methods: Support Vector Learning*. MIT Press. Citado na página 13.

Soares, L. C. C. P., Ronzani, R. A., de Carvalho, R. L., e da Silva, A. T. R. (2020). Aplicação de técnicas de aprendizado de máquina em um contexto

acadêmico com foco na identificação dos alunos evadidos e não evadidos. *Humanidades & Inovação*, 7(8):223–235. Citado na página 18.

Souza, A. M. d. (2020). Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão. Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento) - Fundação Mineira de Educação e Cultura. Disponível em: <https://repositorio.fumec.br/xmlui/bitstream/handle/123456789/420/alex_souza_mes_sigc_2020.pdf?sequence=1>. Acesso em: 12 de abr. de 2021. Citado nas páginas 4, 27, 30, e 32.

Takahashi, K., Yamamoto, K., Kuchiba, A., e Koyama, T. (2021). Confidence interval for micro-averaged f1 and macro-averaged f1 scores. *Applied Intelligence*. Citado na página 39.

Tan, P., Steinbach, M., Karpatne, A., e Kumar, V. (2019). *Introduction to Data Mining*. What's New in Computer Science Series. Pearson. Citado nas páginas 3, 7, 17, e 18.

Tekin, A. (2014). Early prediction of students' grade point averages at graduation: A data mining approach. *Eurasian Journal of Educational Research*, 54:207–226. Citado na página 14.

UFMS (2020a). Indicadores TCU - UFMS. Disponível em: <https://www.ufms.br/indicadores-tcu-ufms/>. Acesso em: 10/04/2021. Citado na página 4.

UFMS (2020b). UFMS em números. Disponível em: <https://numeros.ufms.br/>. Acesso em: 11 de abr. de 2021. Citado nas páginas 5 e 42.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999. Citado na página 13.

Vasconcelos, R. C., Neto, A. J. M., e Teles, L. (2018). Proposta de um modelo de mineração de dados educacionais para identificar a colaboração entre estudantes da ead. *CIET:EnPED*. Citado na página 4.

Viana, R., Rodrigues, R. B., Alvarez, M. A., e Pistori, H. (2007). Svm with stochastic parameter selection for bovine leather defect classification. In Mery, D. e Rueda, L., editors, *Advances in Image and Video Technology*, pagina 600–612. Springer Berlin Heidelberg. Citado na página 13.

Wirth, R. e Hipp, J. (2000). Crisp-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, páginas 29–39. Manchester. Citado na página 10.

Wu, G., Chang, E. Y., e Panda, N. (2005). Formulating distance functions via the kernel trick. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pagina 703–709. Association for Computing Machinery. Citado na página 14.

Yadav, S. e Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, páginas 78–83. Citado na página 19.

Yu, L., Zhou, R., Chen, R., e Lai, K. K. (2020). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and Trade*, 0(0):1–11. Citado na página 36.