UNIVERSIDADE FEDERAL DO PAMPA

Karina Casola Fernandes

Estudo da evasão de alunos de graduação utilizando Educational Data Mining

Alegrete
2019

Karina Casola Fernandes

Estudo da evasão de alunos de graduação utilizando Educational Data Mining

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr.Alessandro Bof de Oliveira

Alegrete
2019

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

F363e Fernandes, Karina Casola
Estudo da evasão de alunos de graduação utilizando Educational Data Mining / Karina Casola Fernandes.
89 p.

Trabalho de Conclusão de Curso(Graduação)-- Universidade Federal do Pampa, CIÊNCIA DA COMPUTAÇÃO, 2019.
"Orientação: Alessandro Bof de Oliveira".

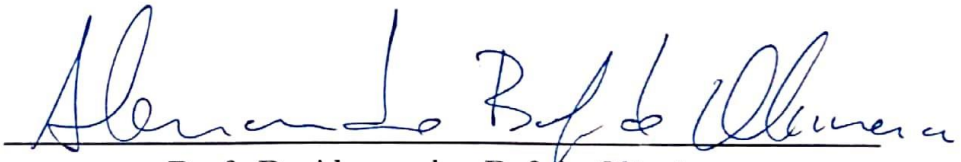
1. Evasão no Ensino Superior. 2. Mineração de Dados Educacionais. 3. Métodos Supervisionados. I. Título.

Karina Casola Fernandes

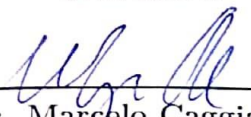
Estudo da evasão de alunos de graduação utilizando Educational Data Mining

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

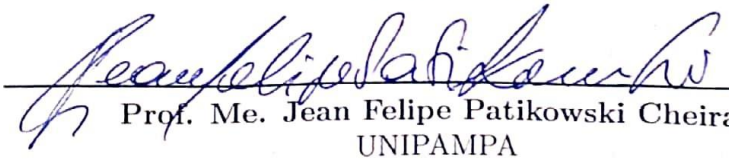
Trabalho de Conclusão de Curso defendido e aprovado em 26 de Junho de 2019
Banca examinadora:



Prof. Dr.Alessandro Bof de Oliveira
Orientador
UNIPAMPA



Prof. Dr. Marcelo Caggiani Luizelli
UNIPAMPA



Prof. Me. Jean Felipe Patikowski Cheiran
UNIPAMPA

AGRADECIMENTOS

Primeiramente, gostaria de agradecer ao Orientador prof. Dr. Alessandro Bof de Oliveira por despender seu tempo para compartilhar seus conhecimentos nessa trajetória de construção de conhecimento durante a execução do Trabalho de Conclusão de Curso. A oportunidade de vivenciar dentro dos grupos PampaGD e GEinfoEdu, experiências dentro de ações de pesquisa e extensão da Unipampa respectivamente, que sem dúvida são essenciais para uma formação mais ampla e engajada com a realidade. A própria existência de Instituições Públicas de Ensino, com sua democratização do acesso ao conhecimento, pesquisa e ensino, pilares fundamentais para uma sociedade igualitária, com a possibilidade de geração de riqueza e crescimento econômico. A família e todas as pessoas que contribuiriam direta ou indiretamente com a formação do meu pensamento crítico, e a tudo que sou hoje. A Jaqueline Moura, minha companheira de trajetória pelo apoio e compreensão.

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform...
But it is likely to exert an indirect and reciprocal influence on science itself."
(Ada Lovelace - 1815-1852).

RESUMO

A problemática da evasão é o objeto de estudo de diversas áreas e uma preocupação recorrente em Instituições Federais de Ensino Superior (IFESs), pois, está associada com a perda social e de recursos de todos os envolvidos no processo de ensino. A análise de maneira ágil e contundente de dados que as Instituições dispõem é de suma importância para se efetivar ações preventivas para o problema. Nesse sentido, esse trabalho objetiva o estudo da evasão utilizando a Educational Data Mining (EDM) do curso de Ciência da Computação da Universidade Federal do Pampa (Unipampa), sob dois aspectos: a análise do perfil socioeconômico dos ingressantes através dos dados Sistema de Seleção Unificada (SiSU)/Exame Nacional do Ensino Médio (ENEM), referente aos anos de 2010 a 2018 e a situação dos alunos matriculados no primeiro ano nos componentes curriculares referentes aos dois primeiros semestres no eixo temporal de 2009 a 2018. As análises do perfil ingressante e discente foram feitas de maneira separada não tendo, portanto, o cruzamento de informações. Na análise socioeconômica foi apurado que mesmo que a evasão de maneira geral tenha um alto índice, existem grupos de maior risco que necessitam de um olhar mais atendo da Instituição. As notas nas competências por área da avaliação ENEM/SiSU, não tem um impacto tão considerável na permanência desse discente.
Na análise do perfil do discente do curso, foi demonstrado que existe um padrão que pode ser mapeado utilizando os modelos preditivos explicitados nesse trabalho, com a utilização de métodos supervisionados obtendo uma excelente acurácia.

Palavras-chave: EDM. Evasão. Métodos Supervisionados.

ABSTRACT

The problem of evasion is the object of study of several areas and a recurring concern in Federal Institutes of Higher Education in Brazil, because it is associated with the social and resources losses of all those involved in the teaching process. The agile and conclusive analysis of data that the Institutions have is extremely important to carry out preventive actions for the problem. In this sense, this work aims at the study of dropout using the EDM of the Computer Science program of the Federal University of Pampa (Unipampa), under two aspects: the analysis of the socioeconomic profile of the entrants through the data SiSU/ENEM, referring to the years 2010 to 2018 and the situation of the students enrolled in the first year in the curricular components referring to the first two semesters in the time axis from 2009 to 2018.The analyses of the entrants and enrolled students profiles were done separately and therefore it was not possible cross the information. In the socioeconomic analysis it was found that even if the dropout in general has a high index, there are groups of higher risk that need a closer look of the Institution. The mark in the competences by area of evaluation ENEM/SiSU, does not have such a considerable impact on the permanence of this student. In the analysis of the profile of the student of the program, it was demonstrated that there is a pattern that can be mapped using the predictive models explained in this work, with the use of supervised methods obtaining an excellent accuracy.

Key-words: EDM. Evasion. Supervised Methods.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ABI Área Básica de Ingresso

API Application Programming Interface

CRISP-DM Cross Industry Standard Process for Data Mining

CSV Comma Separated Values

DM Data Mining

DT Decision Tree

EAP Estrutura Analítica do Projeto

EDM Educational Data Mining

ENEM Exame Nacional do Ensino Médio

EPM Educational Process Mining

FN Falso Negativo

FP Falso Positivo

GURI Gestão Unificada de Recursos Institucionais

IFES Instituto Federal de Ensino Superior

KDD Knowledge Discovery in Databases

kNN k-Nearest Neighbor

LR Logistic Regression

LRM Linear Regression Model

MATLAB MATrix LABoratory

MLP Multi Layer Perceptron

Moodle Modular Object-Oriented Dynamic Learning Environment

NB Naive Bayes

NN Neural Network

Numpy Numerical Python

Pandas Python Data Analysis

PCA Principal Component Analysis

RF Random Forest

RM Rapid Miner

ROC Receiver Operating Characteristic

RS Rio Grande do Sul

SAS Statistical Analysis System

SEMMA Sample Explore Modify Model and Assess

SiSU Sistema de Seleção Unificada

SPARQL SPARQL Protocol and RDF Query Language

SPSS Statistical Package for the Social Sciences

SQL Structured Query Language

SVM Support Vector Machine

TCC Trabalho de Conclusão de Curso

UF Unidade da Federação

VPN Valor Preditivo Negativo

VPP Valor Preditivo Positivo

WEKA Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

O principal objetivo das IFESs é proporcionar uma educação de qualidade a seus alunos e melhorar a qualidade das decisões gerenciais. Conforme aponta Bardagi e Hutz (2005), entre os estudos existentes, a questão da evasão ou permanência no curso universitário desponta como um dos principais interesses de investigação de diversas áreas quando o tema é a Universidade. A Evasão pode ser compreendida, conforme descreve Bonneau (2006), como sendo o abandono por parte do estudante antes do término de um programa de estudos seja em uma escola ou em uma outra instituição de ensino, sem registro de transferência para outra instituição.

A evasão impacta diretamente em uma gestão de qualidade efetiva de quaisquer instituição de ensino, uma vez que o planejamento de vagas a serem ofertadas está atrelado ao índice de vagas preenchidas que muitas vezes contabiliza erroneamente vagas que estão evadidas. Ainda, outro fator importante a ser mencionado é de acordo com Costa et al. (2012b), a baixa taxa de sucesso nos cursos de graduação e do conceito CAPES-MEC[1], dos cursos de pós-graduação, representaram fatores de ineficiência das IFESs. Sumariamente, Lobo (2012) salienta a evasão como sendo a representação da perda social e de recursos de todos os envolvidos no processo de ensino.

Para evidenciar de maneira efetiva os índices reais da evasão em IFES, uma medida adotada pelo Governo Federal brasileiro foi a Lei Nº 12.089 de novembro de 2009 Brasil (2009). Essa lei proíbe a ocupação de duas vagas, simultaneamente, pela mesma pessoa em cursos de graduação de IFESs. Infere-se que a lei visa minimizar os casos nos quais discentes, por desinteresse ou vários outros motivos, abandonem um dos cursos ou demorem mais que o tempo normal para concluírem os estudos.

Diversos trabalhos teóricos estudaram os fatores que levam um discente a abandonar o ensino superior. O primeiro e mais comumente usado modelo na literatura de retenção de estudantes é o modelo de Tinto (1975), Tinto (1987), onde a probabilidade de um estudante se retirar do ensino superior é vista como determinada por atributos individuais, atributos familiares, qualificações prévias, integração social, acadêmica. integração, compromisso individual, compromisso institucional e fatores familiares e sociais externos. Esses trabalhos teóricos baseiam-se essencialmente na concepção de questionários direcionados aos estudantes para se averiguar pontos importantes de análise, para se delimitar um perfil dentro do aspecto que procura ser investigado. Porém diversos trabalhos nesse sentido não obtiveram um rigor e empenho analítico para melhor entendimento dessa temática (BARDAGI; HUTZ, 2005).

Embora tenham sido usados com sucesso, as pesquisas baseada em questionários têm algumas desvantagens, como baixa taxa de participação e o custo associado à organização e administração da pesquisa. Além disso, o estudo baseado somente em pesquisas é demorado. Assim, é fundamental para os profissionais e pesquisadores terem meios sufici-

[1] <http://www.capes.gov.br/>

entes para avaliar as tendências nas circunstâncias de evasão de alunos em sua instituição, a fim de desenvolver ou ajustar os programas de apoio de acordo com essa realidade. Uma ferramenta muito promissora para atingir esse objetivo é o uso de técnicas de *Data Mining (DM)* nos dados comumente disponíveis. Segundo Castro e Ferrari (2016), a mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados KDD. A utilização da mineração de dados na área da educação culminou com o surgimento da *Educational Data mining* EDM (mineração de dados educacionais), uma nova área de estudos que herdou diversas técnicas comumente utilizadas na mineração de dados, para detectar padrões em grandes coleções de dados educacionais - padrões que de outra forma seriam difíceis ou impossíveis de analisar devido ao enorme volume de dados existentes (ROMERO et al., 2010).

A EDM usa muitas técnicas, como Árvores de Decisão - Decision Tree (DT), Redes Neurais NN, Naive Bayes (NB), Regressão, entre outras. As técnicas de mineração de dados são usadas para operar uma grande quantidade de dados para descobrir padrões e relacionamentos ocultos, o que é útil na tomada de decisões. Instituições de ensino superior estão coletando maior volume de dados do que nunca, sobre seus membros, como estudantes e professores, suas instalações e currículos (BICHSEL, 2012; SIEMENS; LONG, 2011).

Conforme Romero et al. (2010) a EDM pode ser visualizada como sendo uma técnica de avaliação formativa. Ou seja, consiste essencialmente em uma ferramenta de avaliação de um programa educacional enquanto ainda está em desenvolvimento, tendo como meta a melhoria continua, possibilitando a concepção de uma base pedagógica para decisões ao projetar ou modificar a pedagogia utilizada pela Instituição de ensino. Nesse sentido, o principal objetivo desse trabalho é analisar os índices de evasão de estudantes de graduação na Universidade Federal do Pampa - Unipampa focado no curso de Ciência da Computação, através de técnicas de EDM, aplicadas para estudar os principais fatores que podem desencadear a evasão e auxiliar no processo de tomada de decisão e assim para mitigar o processo de evasão. A análise de evasão é obtida pelo estudo do comportamento dos estudantes no primeiro semestre utilizando técnicas de classificação de dados.

1.1 Motivação

A problemática da evasão é uma demanda importante da sociedade, envolvendo tanto Universidades públicas quanto privadas, que em seu âmbito institucional, apresentam elevados índices de desistência nos seus mais diversos cursos. A evasão atinge toda a cadeia produtiva do país, uma vez que está estritamente atrelada à questões sociais e financeiras, impactando o resultado econômico das Instituições de ensino, além da perca de papel social do discente que não consegue se colocar economicamente ativo no cenário social-econômico.

Conforme salienta Santos (2002), uma das métricas que podem mensurar a eficácia

de uma Instituição é seu resultado econômico, sendo portanto estendida para a esfera pública. As conjunções, relativas as origens e os motivadores para ações de evasão são altamente diversas e envolvem fatores tanto externos como internos, aspectos pessoais e institucionais. Por se tratar de um problema complexo, Lobo (2012) enfatiza que quanto mais cedo for tratado, maior é a chance de sucesso.

Neste sentido, a utilização da EDM possui como objetivo desenvolver um ambiente que possibilite a descoberta de conhecimento em bases de dados a partir de técnicas de mineração de dados educacionais e utilize esse conhecimento para identificar tendências de enquadramento de perfis, que podem identificar precocemente o risco de evasão, contribuindo para se crie recomendações úteis e construtivas aos planejadores acadêmicos em institutos de ensino superior para aprimorar seu processo de tomada de decisões, minimizando as incidências desse quadro, e de maneira generalista contribuindo para a evolução dessa área de pesquisa contemporânea e oportuna (MACHADO et al., 2015).

1.2 Objetivos

1.2.1 Objetivos Gerais

Analisar a evasão dos discentes de Ciência da Computação, Unipampa *campus* Alegrete, através de práticas da EDM, sob dois ângulos: o perfil socioeconômico do ingressante através de dados SiSU/ENEM, e o perfil discente do curso de Ciência da Computação.

1.2.2 Objetivos Específicos

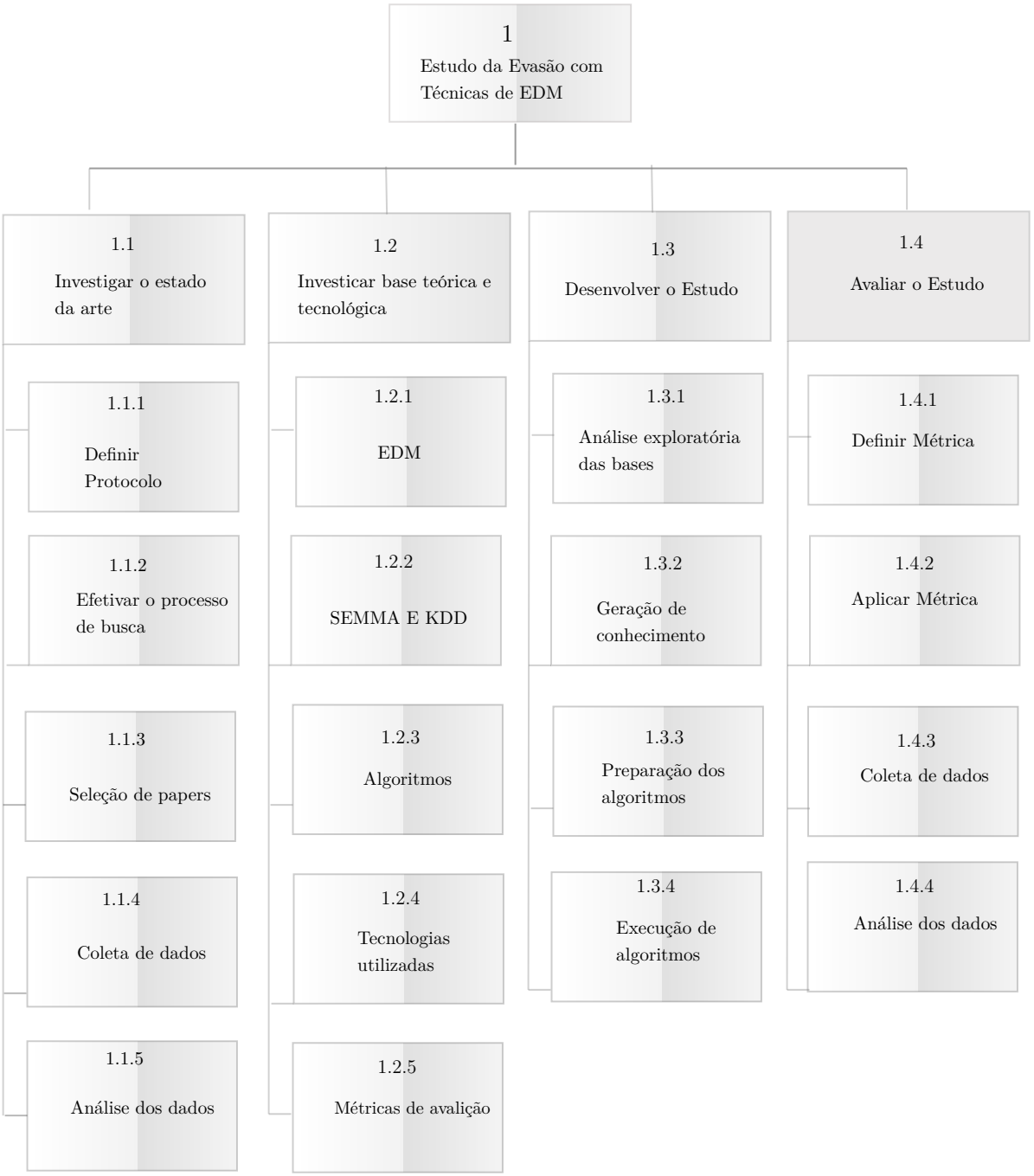
- Propor uma abordagem EDM para a extração de padrões objetivando analisar a evasão de discentes da Universidade Federal do Pampa–Unipampa.
- Criar novos indicadores com o intuito de corroborar o prognóstico da evasão.
- Analisar o desempenho dos indicadores criados.

1.3 Metodologia

Este trabalho se caracteriza como sendo uma pesquisa teórico-prática (WAZLAWICK, 2017). A Figura 1 apresenta a estruturação e o objetivo principal desse estudo que é a análise da evasão do curso Ciência da Computação da Unipampa, *campus* Alegrete-RS utilizando EDM , exemplificando os objetivos específicos de cada tarefa.

A EAP é composta da enumeração de todas atividades de um projeto decompostas hierarquicamente em tarefas curtas e gerenciáveis (TAUSWORTHE, 1979).

Figura 1 – EAP do TCC.



Fonte: Própria Autora.

1.4 Organização do Documento

O Capítulo 1 é apresentado a área focal desse estudo, com uma pequena introdução da EDM e seus preceitos básicos. O Capítulo 2 é retratado o estado da arte dessa área, bem como os trabalhos relacionados selecionados e suas respectivas peculiaridades. O Capítulo 3 é dedicado à fundamentação teórica e tecnológica desse estudo, elencando os conceitos envolvendo as técnicas de EDM que serão empregadas para a realização dessa pesquisa. No Capítulo 4 é apresentado as etapas de desenvolvimento da análise do perfil socioeconômico do ingressante do período de 2010 a 2018 da base SiSU/ENEM . O Capítulo 5 é descrita a análise do perfil discente do período de 2009 a 2018, com a concepção de um modelo preditivo, abordando todas as etapas de desenvolvimento. No Capítulo 6 é discutido os resultados obtidos tanto da análise do perfil socioeconômico do ingressante, quanto da análise do perfil discente.

O Capítulo 7 apresenta as conclusões desse estudo, e no seção 7.1 a projeção de trabalhos futuros para esse estudo.

2 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos relacionados ao projeto de pesquisa aqui proposto, buscando se averiguar o estado da arte desse objeto de estudo e elencar possíveis melhorias que poderiam ser efetivadas nesse campo. A consulta desses trabalhos e triagem se restringiu as bases de dados ACM[1], IEEE[2], Scopus[3], BDTD[4] respectivamente. Para a realização dessa análise preliminar, esse trabalho efetivou uma revisão sistemática que conforme o que salientam Galvão e Pereira (2014), são consideradas estudos secundários, que têm nos estudos primários sua fonte de dados. Entende-se por estudos primários os artigos científicos que relatam os resultados de pesquisa em primeira mão. Neste trabalho foram considerados artigos e teses que tivessem a temática da evasão.

Vale ressaltar que no escopo deste trabalho, optou-se não incluir a predição da retenção, sendo, portanto, uma das possibilidades de trabalhos futuros. Esta pesquisa fornece subsídios para inclusão desta situação, uma vez que o algoritmo identifica rendimento acadêmico abaixo do esperado, é fácil associar este padrão com a ocorrência da retenção. Conforme como salienta Manhães (2015), a retenção costuma ser tratada em trabalhos que utilizam dados do histórico acadêmico dos discentes.

Esta revisão sistemática procurou responder às questões de pesquisa levantadas durante o processo analítico de entendimento, seguindo assim um protocolo, que sumariamente agrega informações das strings de buscas levantadas em cada base pesquisada, e critérios de inclusão e exclusão elencados nesse trabalho. Nas próximas seções será abordado o protocolo utilizado neste trabalho de pesquisa, resultados obtidos, bem como um resumo de cada trabalho selecionado e lições aprendidas nesse processo.

2.1 Protocolo

Nesta seção é apresentado o protocolo que foi definido nessa revisão sistemática da literatura. O protocolo é composto essencialmente pelo objetivo que foi promover uma visão geral do estado corrente da pesquisa relacionada a EDM, aplicado ao domínio das Universidades dentro do contexto evasão. Nesse sentido, foram delimitados questões de pesquisa apontadas em Tabela 1, estratégias utilizadas para busca e seleção de estudos, definindo-se critérios de inclusão e exclusão dos estudos dentro desse segmento.

As perguntas objetivaram identificar o procedimento que era mais utilizado em minerações de dados, seja uso de uma ferramenta de mineração de dados ou implementação de algoritmos, em uma linguagem de programação bem como as respectivas bibliotecas utilizadas.

---

[1] <https://dl.acm.org/>
[2] <https://ieeexplore.ieee.org/Xplore/home.jsp>
[3] <https://www.scopus.com/>
[4] <http://bdtd.ibict.br/>

Tabela 1 – Questões de pesquisa

Q1 -	Quais são os algoritmos classificadores utilizados?
Q2 -	Implementa algoritmos?
Q3 -	Quais bibliotecas e linguagem de programação são utilizadas?
Q4 -	Utiliza ferramentas para a mineração dos dados?

Fonte: Própria Autora.

2.1.1 Estratégias para busca e seleção de estudos

Sumariamente, foram montadas strings de busca que obedecessem o que preconiza cada base de dados para a eleição de palavras-chave que elencassem o processo mineração de dados educacionais no contexto da evasão. Se observou que tanto a SCOPUS quanto à IEEE possuem similaridade nos seus mecanismos de busca, ficando da seguinte forma as strings utilizadas para a triagem inicial dos trabalhos nessas duas bases:

Figura 2 – *String* de busca SCOPUS e IEEE

```
("education" AND "DATA MINING" AND (EVASION OR RETENTION))
```

Fonte: Própria Autora.

Na base ACM a *String* de busca ficou da seguinte maneira:

Figura 3 – *String* de Busca base ACM.

```
((acmdlTitle:("education")) OR
(keywords.author.keyword:("education")) AND (acmdlTitle:("data
mining")) OR (recordAbstract:("data mining")) OR
(keywords.author.keyword:("data mining")) AND (acmdlTitle:("evasion"
"retention")) OR (recordAbstract:("evasion" "retention")) OR
(keywords.author.keyword:("evasion" "retention")) )
```

Fonte: Própria Autora.

Já na base BDTD foram utilizadas palavras-chave que continham os termos evasão, mineração de dados educacionais respectivamente.

2.1.2 Critérios e procedimentos para seleção dos estudos

Nesta subseção serão abordados os critérios elegidos para inclusão e exclusão de trabalhos selecionados desse estudo.

2.1.3 Critérios de inclusão

Os critérios analisados para a inclusão dos trabalhos foram basicamente a utilização de preceitos da EDM dentro do contexto de evasão. Essencialmente se priorizou trabalhos que analisaram o contexto universitário. Quanto ao idioma, foram escolhidos trabalhos preferencialmente em língua inglesa, por ser considerado o idioma mais aceito internacionalmente para artigos científicos na área, mas trabalhos relevantes encontrados na língua portuguesa e espanhola, também foram considerados por se tratar de uma área recente e que portanto, não tem uma produção científica tão extensa levando-se em conta da proximidade cultural desses trabalhos.

2.1.4 Critérios de exclusão

Trabalhos que não utilizaram processo de mineração de dados para coleta e análise de dados, e trabalhos que não tinham o seu objeto de estudo a evasão foram sumariamente excluídos no processo de triagem. Critérios de qualidade também foram analisados, como por exemplo se existiu um processo formal bem fundamentado desde a extração à análise de dados. Estudo de evasão fora do ambiente universitário também foram desconsiderados.

2.1.5 Processo de seleção dos estudos

A primeira etapa de triagem dos estudos se deu através da análise do título. Nessa etapa foram selecionados 5 trabalhos de um conjunto de 44 trabalhos da base IEEE, 20 trabalhos de 98 trabalhos retornados na base SCOPUS, 0 trabalhos selecionados de 13 retornados na base ACM e BDTD 54 trabalhos retornados sendo que 4 trabalhos foram selecionados na primeira filtragem.

A segunda etapa consistiu na análise do *abstract* de cada trabalho selecionado, por ter de maneira mais generalista, a temática e métodos que foram utilizados em cada estudo. Após a seleção dos trabalhos pelo *abstract*, foram lidos em íntegra os trabalhos que serviram como base desse capítulo.

2.2 Resultado

Esta seção apresenta o resultado final da seleção dos trabalhos dessa revisão sistemática, totalizando 6 estudos oriundos das bases IEEE, BDTD e Scopus, respectivamente. Como pode ser observado na Tabela 2, tratam-se essencialmente de trabalhos recentes, utilizando na maioria dos casos ferramentas de mineração de dados gratuitas, que possuem implementações de algoritmos de aprendizado de máquinas e estatísticos utilizados no processo de descoberta de conhecimento. A Tabela 2 relaciona os respectivos autores com as questões de pesquisa respondidas durante a análise dos mesmos.

No trabalho de Pereira e Zambrano (2017) procurou-se evidenciar padrões de abandono universitário analisando dados socioeconômicos dos alunos de graduação da Univer-

sidade de Nariño, da cidade de Pasto (Colômbia), com o apoio da ferramenta Waikato Environment for Knowledge Analysis (WEKA)[5], utilizando o algoritmo de classificação DT. O conhecimento gerado tinha como objetivo de dar apoio à tomada de decisão efetiva do pessoal da Universidade, com foco no desenvolvimento de políticas e estratégias relacionadas aos programas de retenção de alunos dessa instituição. A principal fonte dos dados se deu pelos dados históricos da instituição com complemento de questionários aplicados aos discentes. Com a utilização do algoritmo DT obtiveram uma taxa de acerto(confiança) de 80% na predição do risco de evasão . Os grupos de risco que foram encontrados basicamente estavam relacionados a uma renda *per capita* baixa, associada também a um núcleo familiar composto apenas por um membro (geralmente a mãe). Já o que se refere a notas baixas, estavam diretamente relacionadas ao relacionamento com o corpo docente.

No estudo de Hegde e Prageeth (2018), se procurou responder algumas perguntas de pesquisa sobre evasão, relacionando essas perguntas essencialmente ao fator comportamental que leva o discente a tomar a decisão de evadir e questões relacionadas a saúde que poderiam contribuir para esse cenário. A coleta dos dados foi feita através de questionários e dados institucionais (pré-pesquisa e pós pesquisa) de discentes do primeiro ao 3º semestre. A pesquisa analisou essencialmente o desempenho acadêmico, fatores demográficos, fatores psicológicos, questões de saúde, integração social, mídia social e informações gerais. Através de dados amostrais de uma pequena amostra de 50 indivíduos, com a utilização da linguagem R e WEKA, com os algoritmos NB e DT, obtiveram uma acurácia de 72% do modelo preditivo dos alunos.

Os autores Sarker, Tiropanis e Davis (2014) objetivaram construir um novo modelo preditivo aplicando o método da rede neural MLP das ferramentas SPARQL Protocol and RDF Query Language (SPARQL)[6] e Statistical Package for the Social Sciences (SPSS)[7] em dados comumente disponíveis em fontes abertas de dados internos e externos institucionais, em vez de questionários usados nos modelos preditivos tradicionais de estudantes. Os autores desenvolveram modelos preditivos baseados no modelo de integração estudantil de Tinto, onde um conjunto de questionários chamado Institucional *Integration Scale (IIS)* desenvolvido por Pascarella e Terenzini, foi usado para medir várias dimensões identificadas por Tinto como correspondentes à probabilidade de persistência. Nos modelos utilizando como fonte primária a pesquisa, obtiveram uma acurácia de 69%, Já utilizando os dados institucionais, tiveram uma acurácia de 85,74%.

Nesse estudo o autor Amaral (2016) conduziu experimentos no ambiente da Universidade Federal de Pernambuco (UFPE), utilizando as ferramentas WEKA, Orange, Rapid Miner (RM)[8] respectivamente. Os algoritmos que foram utilizados foram: Support

[5] <https://www.cs.waikato.ac.nz/ml/weka>
[6] <https://www.w3.org/TR/sparql11-query/>
[7] <https://www.ibm.com/products/software>
[8] <https://rapidminer.com/>

Vector Machine (SVM), kNN, Logistic Regression (LR) e Random Forest (RF). Para alguns dos algoritmos classificadores, foi possível obter uma acurácia de classificação de 73,9%, utilizando apenas dados socioeconômicos disponíveis do ingresso do discente na instituição, sem a utilização de nenhum dado dependente do histórico acadêmico. O modelo de processo formal escolhido pelo autor foi o Cross Industry Standard Process for Data Mining (CRISP-DM).

No estudo de Oliveira Júnior (2015) os experimentos foram realizados com dados de alunos da Universidade Tecnológica Federal do Paraná, consolidados em um *Data Warehouse*, que permitiu investigar a evasão entre os anos de 1980 e 2014. Através da ferramenta WEKA, o autor construiu 3 *datasets* aplicando os algoritmos: DT e Linear Regression Model (LRM), tendo obtido uma média de acurácia de 84,29% no primeiro *dataset*, 82,36% no segundo e 84,33% no terceiro.

Os autores Miranda e Guzmán (2017) efetivaram um estudo a partir dos dados fornecidos pelo Carreiras de engenharia na Universidad Católica del Norte em Antofagasta e Coquimbo (Chile) determinar quais as variáveis que melhor explicam a evasão de um aluno tomando por base fatores socioeconômicos e o *Score* de admissão da universidade denominado *PSU*. A qualidade dos classificadores garantem que suas previsões estão corretas com níveis estatísticos da curva Receiver Operating Characteristic (ROC) de 76%, 75% e 83%, com os classificadores: NB, DT e NN, respectivamente.

Tabela 2 – Seleção de *papers*

Autores	**Q1**	**Q2**	**Q3**	**Q4**
Pereira e Zambrano (2017)	DT	Não	N/A	WEKA
Hegde e Prageeth (2018)	NB DT	Sim	R	WEKA
Sarker, Tiropanis e Davis (2014)	NN MLP	Não	N/A	SPARQL e SPSS.
Amaral (2016)	NB SVM kNN LR RF	Não	N/A	WEKA,Orange, RM.
Oliveira Júnior (2015)	DT DT LRM	Não	N/A	WEKA
Miranda e Guzmán (2017)	DT NB NN MLP	Não	N/A	WEKA

Fonte: Própria Autora.

2.3 Lições aprendidas

Através da realização prática da revisão sistemática da literatura se pode evidenciar o reconhecimento e conjunto das evidências disponíveis para tratar certas questões de pesquisa, contribuindo na identificação dessas lacunas dentro dessa área, sendo, portanto, um guia para se posicionar adequadamente diante as novas atividades desse processo. Outro ponto importante que vale ressaltar é que apesar de a base Scopus estar na categoria de indexadores, alguns artigos só foram retornados a partir das bibliotecas digitais das próprias editoras, sendo, portanto, imprescindível a busca direta nas de modalidade editorial.

A possibilidade de fazer uma revisão sistemática em uma determinada área, com a leitura da fundamentação teórica relacionada, auxilia a integralização e idealização do projeto de pesquisa de uma maneira mais ordenada servindo como norte em determinadas linhas que serão seguidas no desenvolvimento, e auxiliando no processo de assimilação dos conhecimentos inerentes.

Os trabalhos relacionados relataram um panorama interessante com relação às pesquisas: a maioria não implementa algoritmos para de aprendizado de máquinas, utilizando-se de ferramentas para esse processo tanto de tratamento dos dados, oriundas de planilhas geradas pelos sistemas educacionais e históricos dos discentes. Isso demonstra que temos uma consolidação, até inclusive de qualidade dessas ferramentas de mineração de dados.

Apesar disso, é um cenário que não se espera encontrar, levando-se em consideração que a linguagem Python já está há muito tempo dentro da pesquisa científica conforme relata Rossum (1995), e mais atualmente R, juntamente com bibliotecas inerentes à análise de dados. Estas bibliotecas são fruto de uma migração do MATrix LABoratory (MATLAB), pela viabilidade maior de criar novas soluções replicáveis uma linguagem *Open Source* contando portanto, com ótima soluções oriundas tanto do desenvolvimento acadêmico/científico quanto da industria.

A linguagem Python e R são multi-paradigma, conforme salienta Sebesta (2011). Na análise de dados elas assumem características de maneira contundente de linguagens orientadas à aplicação, que se traduz em menos linhas de códigos e uma sintaxe mais limpa para a compreensão. Estudos de Pedregosa et al. (2011) sobre a biblioteca Scikit-learn do Python como sendo de alto nível com alguns *trade offs*, expondo uma grande variedade de algoritmos de aprendizado de máquina, supervisionados e não supervisionados, usando uma interface consistente e orientada a tarefas, permitindo uma fácil comparação de métodos para dada aplicação. Como ele depende do ecossistema científico do Python, ele pode ser facilmente integrado com aplicações fora da gama tradicional de análise de dados estatísticos. Com algoritmos implementados em uma linguagem alto nível (PEDREGOSA et al., 2011).

Outra vantagem de se utilizar uma linguagem de programação é a replicação das

tarefas de maneira mais rápida que executar a configuração das ferramentas, e fazer ajustes finos que muitas vezes são inviabilizados pela quantidade de recursos que elas dispõem. Nesse cenário, ferramentas não seriam totalmente indispensáveis, mas sim análises mais rápidas e com uma maior margem de erro tolerável dentro do estudo que se queria iniciar.

3 FUNDAMENTAÇÃO TEÓRICA E TECNOLÓGICA

Nesta seção é abordada a fundamentação teórica deste trabalho, que consiste essencialmente em aprofundar em temáticas e conceitos que formaram a base para a execução desse estudo. A fundamentação teórica é a parte do planejamento do projeto que aponta o desenvolvimento textual sobre um determinado tema, com base nos principais autores consultados, sendo portanto o primeiro levantamento bibliográfico para melhor entendimento da área de pesquisa (SANTOS; DIAS; MOLINA, 2007).

Nas próximas seções será abordado o conceito de EDM 3.1, algoritmos de mineração de dados que serão utilizados nesse trabalho 3.3, bem como as métricas para avaliação dos algoritmos 3.6.

3.1 EDM

Os métodos clássicos de análise de dados sempre se adequaram as regras de negócio existentes dentro de um determinado contexto específico, não seria, portanto, diferente dentro do cenário educacional. A EDM se consolidou como área de pesquisa por diversos *Workshops* organizados no ano de 2005 (BAKER; INVENTADO, 2014), que por conseguinte transformou-se em uma conferência internacional que ocorria anualmente, culminando com o surgimento do *Journal of Educational Data Mining* em 2009[1].

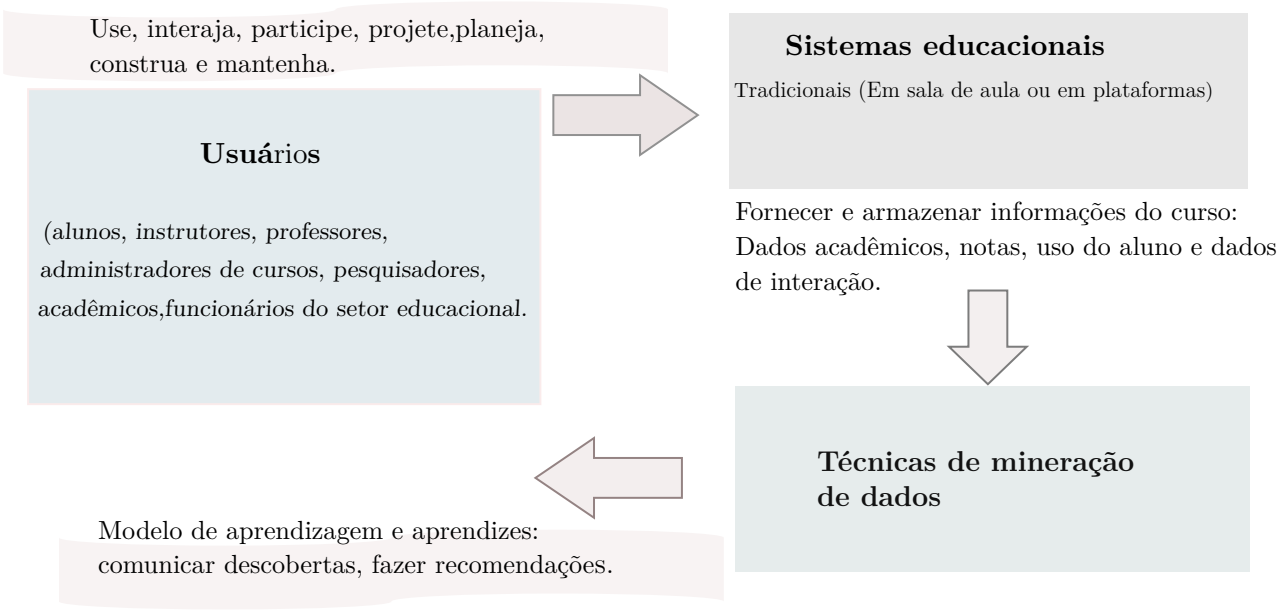
A EDM toma emprestado e amplia campos relacionados, como, por exemplo, *Machine Learning* (o estudo de programas de computador que aprendem e melhoram com dados empíricos), a mineração de texto (abordagens para encontrar padrões em texto em linguagem natural) e estatísticas (BAKER; INVENTADO, 2014).

Outras influências importantes são a psicometria (o estudo de instrumentos psicológicos para medir habilidades e traços humanos) e a análise de registros na *web* (abordagens para identificar perfis de usuários e padrões de navegação de usuários de sites) (SCHEUER; MCLAREN, 2012). A EDM atua diretamente no *design* institucional como pode ser visto na Figura 4, provendo de maneira generalista um ciclo iterativo de formação, teste e refinamento de hipóteses facilitando e melhorando a aprendizagem na totalidade nesse processo. O objetivo não é apenas transformar dados em conhecimento, mas também filtrar a compreensão extraída para tomada de decisões com base nas informações disponíveis sobre cursos, alunos, uso e interação com o conjunto de informações institucionais. As técnicas de mineração de dados podem ser aplicadas para descobrir conhecimentos úteis que ajudem a melhorar os projetos educacionais. O conhecimento descoberto pode ser usado não apenas por projetistas e professores educacionais, mas também por usuários finais — discentes (ROMERO et al., 2010).

A EDM através de sua interdisciplinariedade apresenta uma ligação com diversas áreas de conhecimento, que conforme Romero e Ventura (2013), as principais seriam a

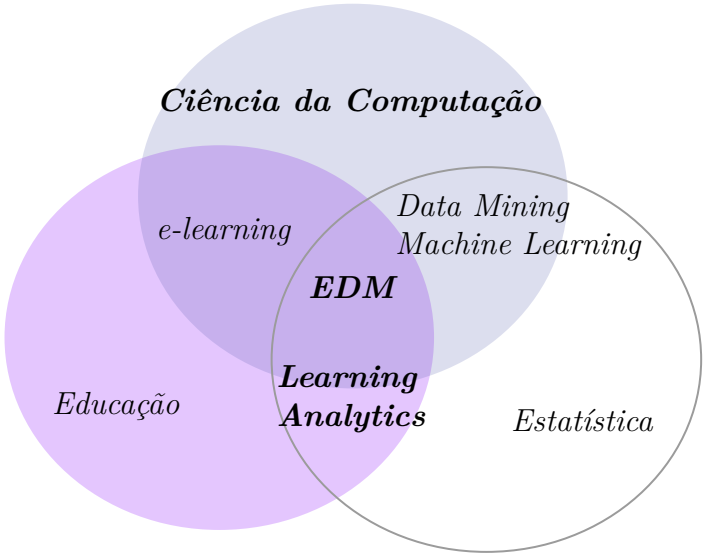
[1] <https://jedm.educationaldatamining.org/index.php/JEDM>

Figura 4 – Aplicando mineração de dados ao *design* de sistemas educacionais



Fonte: Romero et al. (2010).

Figura 5 – Principais áreas relacionadas com a EDM



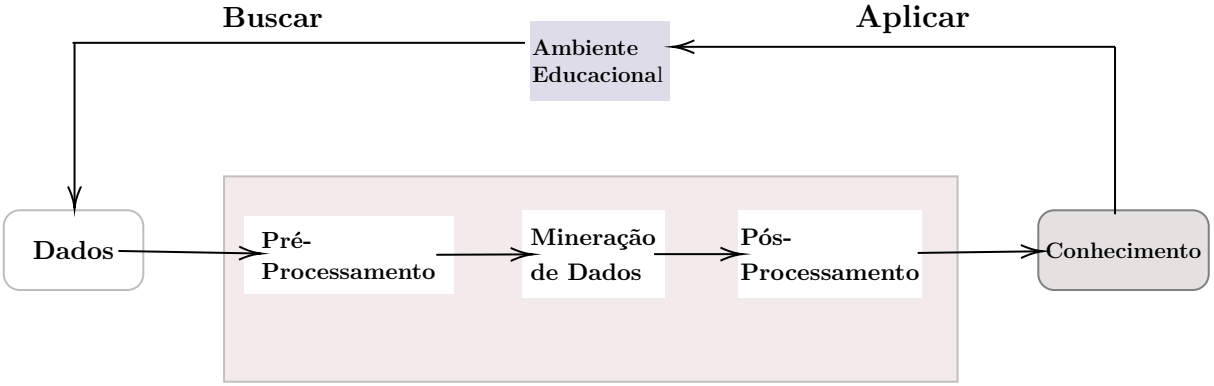
Fonte: Romero e Ventura (2013).

computação, educação e estatística como pode ser visualizado na Figura 5.

A interseção dessas áreas fornece as três subáreas: *E-learning*, *Data Mining* e *Machine Learning* e a *Learning Analytcs* – que estão mais relacionadas com a EDM.

Basicamente, a EDM converte os dados brutos de sistemas educacionais em informação útil que podem ser usados por desenvolvedores de software educacionais, professores, pesquisadores educacionais etc, seguindo o fluxo de processo conforme apresenta Figura 6. Com o objetivo de melhoria dos processos de ensino-aprendizagem e os mecanismos de gestão acadêmica e pedagógica das instituições de ensino (GARCÍA et al., 2011).

Figura 6 – Etapas da EDM



Fonte: García et al. (2011).

Como se pode observar esta é uma área de pesquisa em expansão ao nível mundial, que atua fundamentalmente na predição, agrupamento, mineração de relações, descoberta com modelos e tratamento de dados para apoio à decisão. No Brasil, apesar de ter sido iniciada em meados de 2004 com o evento: *Workshop on Analyzing Student –Tutor Interaction Logs to Improve Educational Outcomes*, que aconteceu em 2004 em Maceió-AL, dentro da programação da *7th International Conference in Intelligent Tutoring Systems* (LESTER; VICARI; PARAGUAÇU, 2004), existem poucas iniciativas nesse segmento, estando o país na retaguarda de pesquisas e aplicações, comparado a outros países. Um marco importante dentro das publicações brasileiras em pesquisas relacionadas a EDM foi o trabalho de Baker, Isotani e Carvalho (2011), onde o autor apontou técnicas e caminhos para a EDM no país, sendo referência a diversos trabalhos com esse enfoque. Dentro do planejamento estratégico das Instituições de ensino, a EDM pode ser agrupada de acordo com seu objetivo final da seguinte maneira:

- **Recomendações aos planejadores educacionais:** Objetiva essencialmente propôr conteúdos de acordo com análise educacional do discente, para por exemplo recupe-

rar o desempenho, ou alguma falha de aprendizagem detectada anteriormente.Dentro dessa temática o uso de associação, sequenciação, 'clusterização'(agrupamento) e classificação, são as técnicas mais usuais para esse diagnóstico (ABEL et al., 2010).

- **Aperfeiçoamento dos cursos:** Tem como foco principal se tornar uma ferramenta estratégica para o aperfeiçoamento dos cursos através de análises de dados Institucionais. Um exemplo nesse sentido é um simples comparativo do desempenho nas disciplinas por professor e notas de vestibular/ENEM dos discentes. Nesse cenário, igualmente ao anterior as técnicas de associação, sequenciação agrupamento e classificação podem ser empregadas (COSTA et al., 2012a).

- **Previsibilidade do desempenho acadêmico:** Como próprio nome diz, procura efetivar uma predição dos resultado de testes e de outras avaliações educacionais, com base na análise das atividades realizadas pelo discente. Novamente técnicas de associação, sequenciação agrupamento e classificação podem ser utilizadas (COSTA et al., 2012a).

- **Avaliação dos processos educacionais:** Tem como foco em uma análise mais holística de todos os processos que contam com a participação do discente, sejam atividades extracurriculares ou participações dentro do curso, a fim de auxiliar os administradores educacionais e professores avaliarem a realização dessas atividades. Mineração de processos educacionais - Educational Process Mining (EPM), geração de relatórios, visualização de dados e a análise estatística de dados, são as técnicas mais utilizadas para esse grupo de aplicações (COSTA et al., 2012a).

- **Mapear perfis de discente:** O ponto focal nesse grupo e traçar perfis através da análise de determinadas características dos discentes. Existe um aumento considerável nesse tipo de segmento de pesquisa em EDM. As técnicas mais utilizadas para esse fim são as análises estatísticas, NB, modelos psicométricos e aprendizado por reforço (ROMERO et al., 2010).

- **Aprendizagem personalizada:** O objetivo, de maneira generalista, é preservar as características fortes e melhorar aquelas abaixo de uma média, após identificá-las. Todas essas premissas se baseiam na análise de dados históricos dos discentes que a Instituição possui.

- **Aprendizagem colaborativa:** Promover o engajamento dos discentes a grupos de estudo, através de identificação de traços de personalidade em comum.

Um ponto bastante importante a ser considerado dentro da EDM, são os valores destoantes, também chamados na literatura de *outliers*, que revelam pontos de dados que não se encaixam no mesmo modelo dos demais, se afastando consideravelmente do

predomínio. Nesse campo de estudo em questão é comum ocorrer *outliers* que caracterizam padrões de discentes que quebram modelos de predição, como por exemplo discentes com sucesso acima da média em determinadas disciplinas ou que falham contra todas expectativas positivas (HÁMÁLÁINEN, 2011).

3.2 SEMMA e KDD

Um dos pontos focais dentro de uma mineração de dados eficiente é determinar um processo a ser seguido que norteará o desenvolvimento desde a concepção a análise de dados. Será utilizado nesse trabalho o processo SEMMA em conjunto com o KDD. O acrônimo SEMMA conforme aponta João (2010) consiste essencialmente em um processo-base da mineração dos dados utilizado pelo Statistical Analysis System (SAS), iniciando com uma representação estatística da amostra dos dados. O SAS é uma organização atuante na tecnologia analítica e pioneira em *Business intelligence*, produzindo soluções para análise avançada, análises multivariadas,gerenciamento de dados e análise preditiva (SAS, 2018).

O SAS Institute considera um ciclo com 5 estágios para o processo SEMMA, como pode ser visto a seguir (AZEVEDO; SANTOS, 2008):

1. Amostra - Esta etapa consiste em coletar amostras dos dados extraindo uma porção de um grande conjunto de dados grande o suficiente para conter as informações significativas, mas pequeno o suficiente para manipular rapidamente. Esta fase é apontada como sendo opcional.

2. Explorar - Esta etapa consiste na exploração dos dados, procurando por tendências e anomalias imprevistas, a fim de obter entendimento e ideias.

3. Modificar - Esta etapa consiste na modificação dos dados, criando, selecionando e transformando as variáveis para focalizar o processo de seleção do modelo.

4. Modelo - Esta etapa consiste em modelar os dados, permitindo que se pesquise automaticamente por uma combinação de dados que prevê com segurança um resultado desejado.

5. Avaliação - Esta etapa consiste em avaliar os dados avaliando a utilidade e a confiabilidade dos resultados do processo de mineração de dados e estimando o desempenho do mesmo.

Embora o processo SEMMA seja independente da ferramenta escolhida DM, ele é vinculado ao software SAS Enterprise Miner e pretende orientar o usuário nas implementações de aplicativos DM (AZEVEDO; SANTOS, 2008).

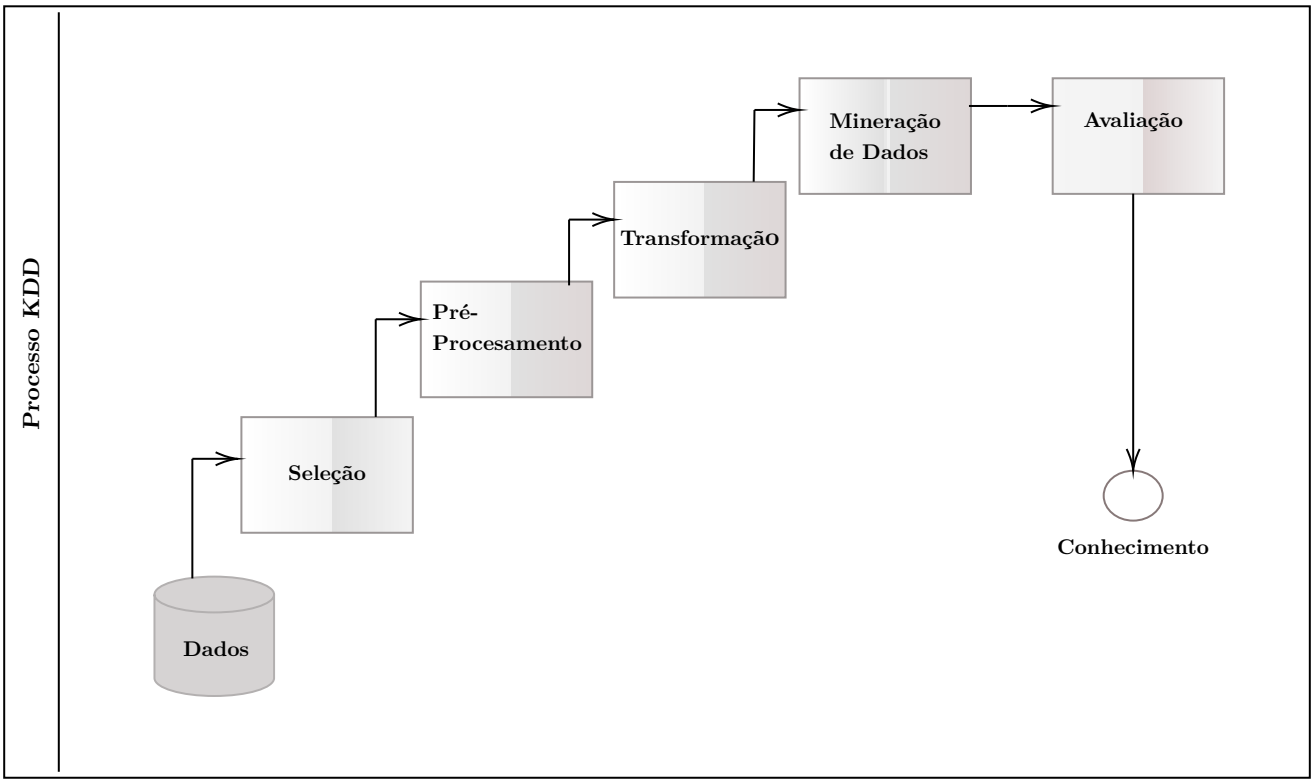
Um fator bastante contundente na escolha desse processo foi a facilidade de compreensão, que permite um desenvolvimento e manutenção organizados e adequados de

projetos de DM, conforme salienta o autor Azevedo e Santos (2008) e portanto confere uma estrutura robusta para a sua concepção, criação e evolução, ajudando a apresentar soluções para problemas de negócios, bem como a encontrar metas de negócio de DM. O KDD está intrínseco em qualquer processo de mineração de dados, e portanto, inexiste qualquer processo novo que em sua essência não carregue os preceitos do KDD. O processo do KDD é interativo e iterativo, envolvendo várias etapas com muitas decisões sendo tomadas pelo usuário, o que viabiliza a adequação de outros processos para auxiliar em uma tomada de decisão eficaz em DM (BRACHMAN; ANAND, 1996).

O KDD compartilha da interseção das áreas que atuam diretamente na mineração de dados. O objetivo principal do KDD é a própria extração do conhecimento de alto nível a partir de dados de baixo nível na conjuntura de grandes agrupamentos de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Conforme Azevedo e Santos (2008), existe uma equivalência entre os estágios do KDD e do SEMMA, Examinando-o com um olhar mais atento, se pode afirmar que os cinco estágios do processo SEMMA podem ser vistos como uma implementação prática dos cinco estágios do processo do KDD. Na Figura 7 é apresentado o processo KDD, com os seus respectivos 5 estágios:

Figura 7 – Etapas do processo de KDD



Fonte: Adaptado: Fayyad, Piatetsky-Shapiro e Smyth (1996).

Os dados utilizados nesse estudo não sofreram modificações conforme aponta as

etapas dos processos SEMMA e KDD respectivamente, devido à própria natureza dos dados já discretizados e estando na forma anônima. O processo exposto será seguido na fase exploratória tentando detectar fatores que contribuem para o surgimento de *outliers*, a concepção de um modelo dentro do processo de mineração de dados, bem como a avaliação desse conhecimento descoberto nessa análise.

3.3 Algoritmos para mineração de dados

3.3.1 kNN

O algoritmo kNN apresentado por Aha, Kibler e Albert (1991) e formalmente estudado por Cover e Hart (1967), é um algoritmo de aprendizado supervisionado do tipo *lazy*, ou seja um algoritmo de classificação mais simples usado para classificar objetos com base em exemplos de treinamento que estão mais próximos no espaço de características. Basicamente esse algoritmo visa encontrar os k exemplos rotulados mais próximos do exemplo não classificado e, com base no rótulo desses exemplos mais próximos, é tomada a decisão relativa à classe do exemplo não rotulado. O kNN relaciona cada um dos exemplos a um ponto em um espaço m-dimensional, sendo m o número de atributos de entrada que descrevem o conjunto de dados (LARRANAGA et al., 2006). Uma das possibilidades de se determinar o vizinho mais próximo de cada ponto no espaço é por meio do cálculo da distância Euclidiana (MICHALSKI; CARBONELL; MITCHELL, 2013), que pode ser visualizado na Equação 3.1 :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{3.1}$$

Onde $d(\overrightarrow{x_i}.\overrightarrow{x_j})$ é a distância entre os pontos $\overrightarrow{x_i}$ e $\overrightarrow{x_j}$. $\alpha(\overrightarrow{x_i})$ é a característica r da instância $\overrightarrow{x_i}$. E η é o número total de coordenadas de cada ponto.

Conforme Michalski, Carbonell e Mitchell (2013) o aprendizado baseado em instâncias apresenta um custo computacional alto para a classificação de novas amostras. Isso se deve ao fato de que todos os cálculos ocorrem no instante da classificação e não quando os exemplos de treinamento são armazenados. Outro ponto que impacta significativamente, é o fato de que todos os exemplos armazenados na memória são utilizados nos cálculos de aproximação para cada nova amostra, quando os mais próximos seriam o suficiente para realização da classificação.

No kNN, a classificação é efetivada conforme a classe mais frequente dentre seus vizinhos mais próximos. Ainda é admissível tornar essa classificação mais eficiente fazendo a atribuição de pesos a cada elemento considerado no processo de classificação, segundo a sua distância em relação a amostra que está sendo classificada, diminuindo a relevância de cada ponto conforme sua distância aumenta (MICHALSKI; CARBONELL; MITCHELL, 2013). Segue abaixo a definição matemática da importância de cada elemento no processo de classificação e a função que define a qual classe uma determinada amostra de entrada

pertence (MICHALSKI; CARBONELL; MITCHELL, 2013) conforme é apresentado na Equação 3.2:

$$f'(x_q) \Leftarrow argmax_{v \epsilon V} \sum_{i=1}^{k} w_i \delta(v, f(x_i)) \tag{3.2}$$

Onde W_i é a importância do exemplo x no processo de definição da classe de amostra x_q. $f(x_q)$ retorna a classe que será atribuída a amostra x_q. V é uma classe no conjunto de V classes possíveis e K é o número de vizinhos mais próximos utilizados para realizar o processo de definição de uma amostra.

3.3.2 NN

Uma NN, é um modelo matemático computacional inspirado no funcionamento de um sistema neural biológico simplificado, tendo como objeto de estudo a topologia das conexões e comportamento conjunto desses elementos de processamento naturais (BRAGA; FERREIRA; LUDERMIR, 2007; SIMON, 2001).

Uma particularidade significativa sobre redes neurais artificiais é que o conhecimento obtido é evidenciado e armazenado através de alterações nas forças das conexões que conectam as unidades de processamento básico (neurônios artificiais). Esta propriedade tem enormes consequências para o processamento e o aprendizado, pois, a representação do conhecimento é modelada para que o conhecimento armazenado através de experiências anteriores influencie no curso de novos processamentos, tornando cada experiência parte do processo desde em que foi adquirida (CASTRO, 2006).

Braga, Ferreira e Ludermir (2007), Simon (2001), Castro (2006) efetivam um comparativo entre as NNs e as redes neurais biológicas:

1. O processamento básico acontece nos neurônios.

2. Estes neurônios podem obter e enviar impulsos a partir de outros neurônios e a partir do ambiente.

3. Neurônios podem ser acoplados uns aos outros, formando uma rede neural.

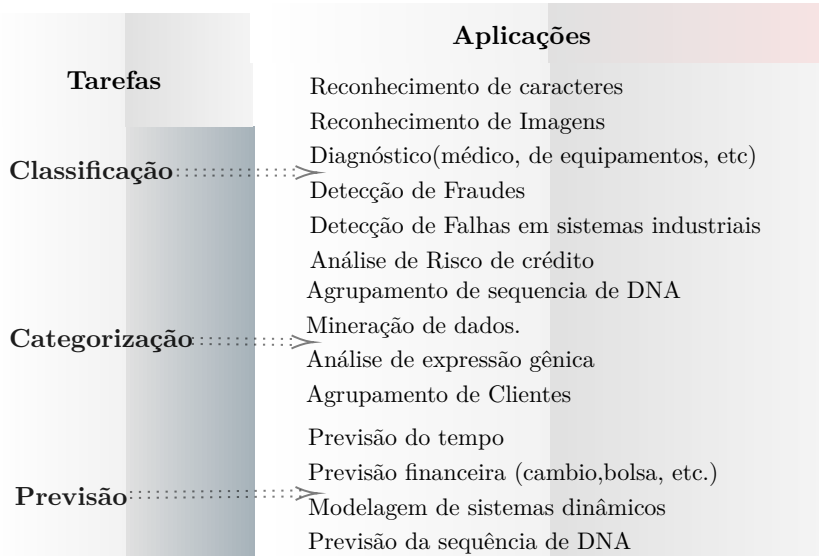
4. O conhecimento é apreendido pela rede a partir do ambiente através de um processo de aprendizado.

5. O processo de aprendizagem ou treinamento é o incumbido por ajustar a força das conexões de acordo com impulsos do ambiente.

6. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são empregadas para armazenar o conhecimento adquirido.

As NNs se aplicam basicamente a problemas em que existem dados, experimentais ou gerados por modelos, por meio dos quais a rede adaptará os seus pesos visando a execução de uma determinada tarefa (BRAGA; FERREIRA; LUDERMIR, 2007). Conforme aponta Braga, Ferreira e Ludermir (2007) os algoritmos dessa classe se aplicam dentro da seguinte categoria: classificação, categorização(agrupamento ou *clustering*), aproximação, previsão e otimização conforme apresenta Figura 8.

Figura 8 – As principais tarefas que as NNs podem executar e alguns exemplos de aplicação



Fonte: Adaptado: Braga, Ferreira e Ludermir (2007).

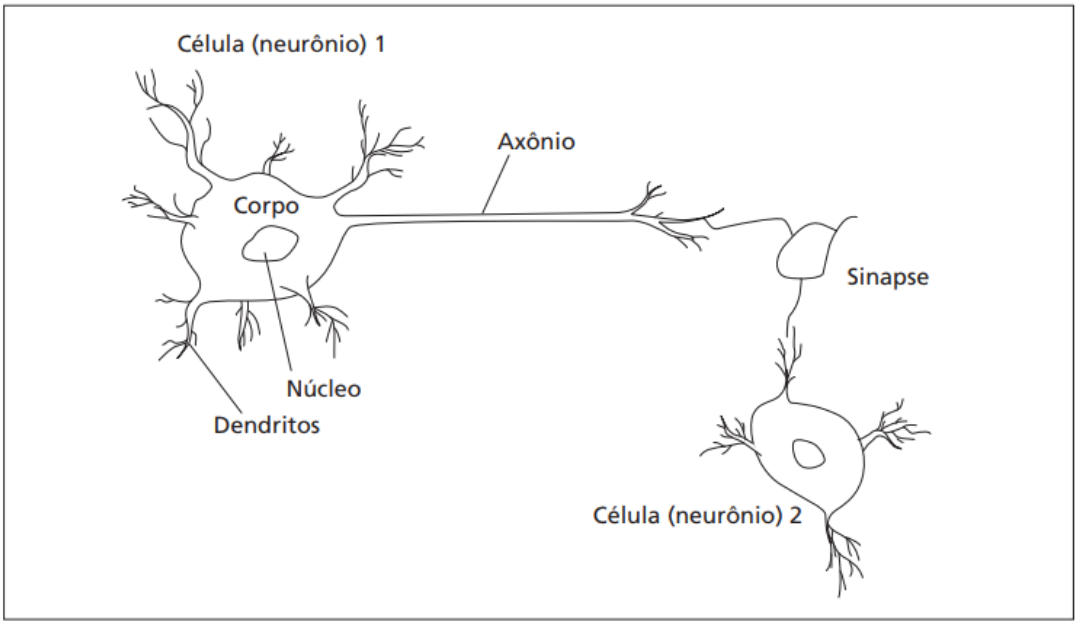
A classificação de uma NN baseia-se em sua disposição, apresentando camada única ou múltiplas camadas, alimentada para frente *feedforward* ou recorrente, sendo total ou parcialmente conectada. Em uma rede *feedforward* o sinal de cada neurônio é ramificado apenas para os neurônios da camada da frente, enquanto que na rede recorrente um neurônio pode propagar seu sinal para um neurônio que não seja o da camada da frente. Em uma rede totalmente conectada cada neurônio fornece sua saída a todas as unidades da camada seguinte, e em uma rede parcialmente conectada estas conexões *forward* não ocorrem inteiramente (HAN; PEI; KAMBER, 2011).

Redes neurais biológicas são compostas de muitos neurônios biológicos primitivos compactamente interconectados. Cada neurônio possui axônios e dendritos, projeções semelhantes a dedos que permitem ao neurônio comunicar-se com seus neurônios vizinhos através da transmissão e do recebimento de sinais químicos e elétricos. Mais ou menos semelhante à estrutura de seus colegas, a NN é composta de elementos de processamento simples e interconectados chamados neurônios artificiais. No processamento da informação, os elementos de processamento em uma NN funcionam de maneira simultânea e coletiva em um modo semelhante aos neurônios biológicos. A NN possui

algumas características desejáveis similares àquelas das redes neurais biológicas, como os recursos de aprendizagem, auto-organização e tolerância ao erro (BRAGA; FERREIRA; LUDERMIR, 2007).

Um modelo de NN emula uma rede neural biológica, fazendo-se uso de uma analogia mais limitada e imprecisa procedente do cérebro humano (BRAGA; FERREIRA; LUDERMIR, 2007).

Figura 9 – Duas células biológicas interconectadas



Fonte: Adaptado de Braga, Ferreira e Ludermir (2007).

Uma parcela de uma rede é mesclada por duas células é mostrada na figura. A célula inclui em si um núcleo (a parte de processamento principal da célula). À esquerda da célula 1, os dendritos geram sinais de entrada para a célula. À direita, o axônio expede sinais de saída para a célula 2 por meio das extremidades do axônio. Esses terminais agrupam-se aos dendritos da célula 2. Os sinais podem ser transmitidos inalterados, ou podem ser alterados pelas sinapses. Uma sinapse é capaz de aumentar ou diminuir a intensidade da ligação entre os neurônios e estimular ou inibir um neurônio subsequente. É dessa forma que a informação é armazenada (BRAGA; FERREIRA; LUDERMIR, 2007).

O limiar b_k na Figura 10 tem o papel de aumentar ou diminuir a influência do valor da entrada líquida para a ativação do neurônio k. Matematicamente, a saída do neurônio k pode ser descrita por (CASTRO; FERRARI, 2016) Equação 3.3, Equação 3.4:

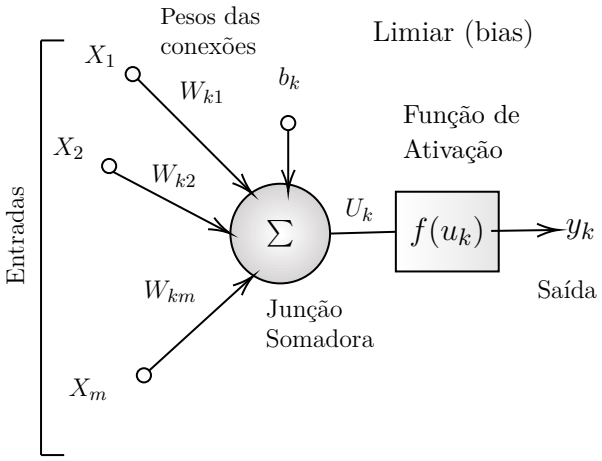
$$y_k = f(u_k) = f(\sum_{j=1}^{m} w_{kj} + b_k) \tag{3.3}$$

ou

$$y_k = f(u_k) = f(\sum_{j=0}^{m} w_{kj} x_j) \tag{3.4}$$

Onde x_0 é um sinal de entrada de valor 1 e peso associado $w_k 0 = b_k$.

Figura 10 – Neurônio artificial genérico



Fonte: Castro e Ferrari (2016).

Existem diversas arquiteturas e implementações de NN porém neste estudo a NN empregada será a rede MLP, uma rede com múltiplas camadas, do tipo *feedforward* totalmente conectada. O algoritmo de treinamento mais usado para redes MLP é o denominado por *backpropagation* (RUMELHART; MCCLELLAND, 1986).

O treinamento de uma MLP é divido essencialmente em dois estágios: o estágio *forward* e o estágio *backward* respectivamente. No estágio *forward* a entrada é expressa à primeira camada da rede, a qual calcula seus sinais de saída e transmite os valores para a camada subsequente, que por sua vez, calcula seus sinais de saída e os transmite para a camada seguinte, e assim sucessivamente, até a camada de saída calcular as saídas da rede que são equiparadas às saídas esperadas (BRAGA; FERREIRA; LUDERMIR, 2007).

Já no estágio *backward* transita de maneira contrária, a partir da camada de saída até a de entrada os pesos dos neurônios vão sendo refinados de forma a diminuir seus erros (os erros dos neurônios das camadas intermediárias são calculados empregando-se o erro dos neurônios da camada subsequente ponderado pelo peso da conexão entre eles). Este procedimento é reforçado até alcançar um determinado critério de parada (BRAGA; FERREIRA; LUDERMIR, 2007).

3.4 Tipos de treinamento

O objetivo do treinamento em uma NN é produzir uma coletânea de saídas esperadas, ou na pior das hipóteses gerar saídas consistentes (GOLDSCHMIDT; PASSOS, 2005).

A realização do treinamento se dá pela execução linear dos vetores de entrada (em algumas situações valendo-se dos de saída também), durante o tempo que os pesos da NN são refinados conforme um mecanismo de treinamento pré-determinado. No decorrer do treinamento, os pesos da NN confluem para determinados valores, fazendo com que os vetores de entrada gerem as saídas fundamentais. As classes de treinamento podem ser subdividas em (FONSECA; NAMEN, 2016; GOLDSCHMIDT; PASSOS, 2005):

- Supervisionado;
- Não Supervisionado.

A diferença primordial entre esses dois tipos de treinamento se dá pela necessidade ou não de um vetor alvo. A utilização de um vetor alvo para efetivar comparativos entre os vetores de entrada e saída é essencialmente a característica marcante do treinamento supervisionado. O processo de treinamento se dá pela aplicação de um vetor de entrada, e então a saída da NN é calculada e comparada com o vetor alvo correspondente. O erro localizado então é realimentado através da rede e os pesos são alterados conforme um determinado algoritmo para efetivar a minimização do erro. Esse procedimento de treinamento é repetido até que o erro para os vetores de treinamento alcance valores pré-estipulados. O treinamento não-supervisionado, por sua vez, modifica os pesos da rede de forma a produzir saídas que sejam consistentes (BRAGA; FERREIRA; LUDERMIR, 2007; GOLDSCHMIDT; PASSOS, 2005). O algoritmo de retropropagação *backpropagation*, será o utilizado nesse estudo para o treinamento das NN.

3.5 Treinamento da MLP com *backpropagation*

Nesta subseção será apresentado a lógica de treinamento da MLP extraída da literatura, e portanto, não se objetiva efetivar um modelo matemático, mas sim apresentar o algoritmo evidenciado no trabalho de Marsland (2011).

1. Comece com os pesos da sinapses w_{ij} atribuindo valores pseudoaleatórios reduzidos positivos e negativos, onde i seja referente ao índice do peso sináptico e j ao índice do neurônio.

2. Enquanto o nível de satisfação não seja atingido faça: pondere a saída para o exemplo de treinamento x utilizando a função de ativação $g(x)$ determinada.

3. Calcule o erro para cada neurônio da camada de saída Equação 3.5:

$$E = (y_k - t_k)y_k(1 - y_k) \tag{3.5}$$

4. Calcule o erro para os neurônios das camadas intermediárias Equação 3.6:

$$E = y_k(1 - y_k)\sum_k wjkEj \tag{3.6}$$

5. Atualize todos os pesos das sinapses da rede Equação 3.7:

$$Wij = Wij + \eta EX_i + \alpha\Delta W_{ij}^{t-1} \tag{3.7}$$

Onde:

η é a taxa de aprendizado, x_i é a entrada de x na posição i. E é o erro na saída do neurônio j. α é o *momentum*. ΔW_{ij}^{t-1} é a variação de Wij nos tempos t e e $t-1$.

Conforme aponta Marsland (2011) o *momentum* é uma técnica que objetiva fazer com que a NN tenha a propensão de superar os mínimos locais através da atribuição de um valor entre zero e um que multiplica a variabilidade dos pesos das conexões sinápticas nos períodos t e $t = 1$, sendo possível efetivar o uso de valores reduzidos para a taxa de treinamento, possibilitando realizar uma regulagem fina na rede.

3.6 Métricas para avaliação algoritmos de mineração de dados

A avaliação de um algoritmo de mineração de dados é associado ao seu desempenho dentro do conceito de modelo de classificação está atrelado a aptidão ou a adequada separação das categorias. Uma estrutura muito usada para essa prática é a matriz de confusão (WITTEN et al., 2016). Em uma matriz de confusão as resoluções das classificações são apresentadas sob uma matriz bidimensional, com uma linha e coluna para cada classe, cada componente da matriz indica o número de instâncias certas ou classificadas erroneamente tomando por base o conjunto de testes usados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Com base em uma matriz de confusão é possível auferir um agrupamento parâmetros para quantificar a performance de um modelo de classificação. Um desses parâmetros é a acurácia, que mensura a taxa de assertividade de maneira íntegra, ou seja, o número de classificações corretas dividido pelo número total de instâncias dos dados a serem classificadas.

A acurácia da classificação em conjunto r é mensurada pela taxa de classificação, que estipula a correlação de registros acertadamente classificados no conjunto r (HÁMÁLÁINEN, 2011).

Além da matriz de confusão, existem outras medidas para se determinar o desempenho de modelos de classificação (WITTEN et al., 2016), como pode ser visualizado na Figura 11, levando-se em consideração o Valor Preditivo Positivo (VPP) e Valor Preditivo Negativo (VPN), ou seja, um cálculo probabilístico para se mensurar um resultado positivo ser de fato positivo, e um resultado negativo ser de fato negativo dentro das análises no momento de predição. É apresentado de maneira sintetizada como **VP** e **VN** respectivamente (MÁRQUEZ-VERA; MORALES; SOTO, 2013; WITTEN et al., 2016).

Figura 11 – Avaliação de algoritmos de mineração de dados

Medida	Fórmula
Acurácia, taxa de reconhecimento	$\frac{VP+VN}{P+N}$
Taxa de erro, taxa de classificação incorreta	$\frac{FP+FN}{P+N}$
Sensibilidade, taxa de verdadeiro positivo, recobrimento	$\frac{VP}{P}$
Especificidade, taxa de verdadeiro negativo	$\frac{VN}{N}$
Precisão	$\frac{VP}{VP+FP}$
F, F_1, *F-score*, média harmônica de precisão, recobrimento	$\frac{2 \times precisao}{precisao+recobrimento}$
F_β, onde β é um número real não negativo	$\frac{(1+\beta^2) \times precisao \times recobrimento}{(\beta^2 \times precisao)+recobrimento}$
Média Geométrica. Indica o equilíbrio entre os desempenhos de classificação nas classes majoritárias e minoritárias	$MG = \sqrt{VP \times VN}$

Fonte: Márquez-Vera, Morales e Soto (2013), Witten et al. (2016).

Na medida **Taxa de Erro** e **Precisão** se utiliza o conceito de Falso Negativo (FN) e Falso Positivo (FP), ou seja a quantidade de exemplos classificados incorretamente como sendo negativos e positivos, respectivamente.

3.6.1 Avaliação de algoritmos com testes estatísticos: Friedman e Nemenyi

O teste não paramétrico de Friedman é utilizado para determinar se existem diferenças significativas entre os modelos (STOJANOVA et al., 2010; FRIEDMAN, 1977). Já o teste *post-hoc* de Nemenyi conforme Calvo e Guzmán (2016), determina se as diferenças existentes são estaticamente significativas.

A hipótese nula do teste de Friedman assume, que as k amostras(colunas) são provenientes da mesma população ou de populações com a mesma mediana (FÁVERO; FÁVERO, 2015).

Para proceder o teste de Friedman (VIEIRA, 2004):

- *Primeiro passo:* A hipótese da nulidade é a de que não há diferença entre grupos. Estabeleça a hipótese alternativa e o nível de significância.
- *Segundo passo:* Atribua um posto a cada dado por bloco.
- *Terceiro passo:* Calcule as somas dos postos de cada grupo, isto é R_1, R_2, ... , R_k. Se as somas dos postos estiverem corretas, então Equação 3.8:

$$\sum R_1 + R_2 + ... + \sum R_i = \frac{1}{2} nk(k+1) \tag{3.8}$$

- *Quarto passo: Calcule.* Equação 3.9:

$$X^2 = \frac{12}{Nk(k+1)} (\sum R_1^2 + \sum R_2^2 + ... \sum R_k^2) - 3N(k+1) \tag{3.9}$$

  Em que: N é igual ao número de unidades e K é o número de grupos.

- *Quinto passo:*

  Sob a hipótese da nulidade, a estatística χ_r^2 tem, aproximadamente, distribuição de χ^2 com $(k-1)$ graus de liberdade. Faça o teste, que consiste em comparar o valor calculado de χ_r^2 com o valor crítico dado na tabela de χ^2, no nível de significância estabelecido e com $(k-1)$ graus de liberdade.

3.6.2 Validação cruzada

O objetivo da validação cruzada é, de forma sistemática, particionar a base de dados em conjuntos de treinamento e teste de modo que os dados de treinamento sejam usados para ajustar os parâmetros livres do modelo e os dados de teste sejam usados para fornecer uma estimativa de como o modelo vai generalizar para dados não usados no treinamento (CASTRO; FERRARI, 2016). A validação cruzada é apropriada quando não se dispõem de uma base de teste separada, servindo de alternativa mais confiável à divisão direta da base, que muitas das vezes podem o problema de que alguns registros que estão na base de teste seriam ótimos previsores, registros que oferecem uma generalização interessante para a geração de um modelo mais efetivo de previsão.

3.7 Tecnologias utilizadas

3.7.1 Python

A linguagem de programação Python (ROSSUM, 1995) está se estabelecendo como uma das linguagens mais populares para a computação científica. Graças à sua natureza interativa de alto nível e seu amadurecido ecossistema de bibliotecas científicas, é uma opção atraente para o desenvolvimento algorítmico e a análise exploratória de dados. Assim como o MATLAB[2], R [3], e similares utilizados na programação científica, a linguagem Python também pode ser utilizada de forma interativa (CODEÇO, 2007).

A sintaxe de Python não é baseada diretamente em nenhuma linguagem comumente usada. Ela é uma linguagem com verificação de tipos, mas tipada dinamicamente. Em vez de vetores, inclui três tipos de estruturas de dados: listas; listas imutáveis, chamadas de tuplas; e dispersões, chamadas de dicionários (SEBESTA, 2011).

[2] <https://www.mathworks.com/products/matlab.html>
[3] <https://www.r-project.org/>

3.7.2 R

O R é um ambiente integrado (linguagem e software) para manipulação de dados, cálculo e exibição gráfica. e esse ambiente inclui (FERREIRA, 2018):

- Instalação eficaz de tratamento e armazenamento de dados;
- Conjunto de operadores para cálculos em numéricos, vetores e matrizes;
- Grande coleção coerente e integrada de ferramentas intermediárias para análise de dados;
- Instalações gráficas para análise de dados e exibição na tela ou em console;
- Linguagem de programação bem desenvolvida.

3.7.3 Scikit-learn

Desde o início do projeto em 2010, o Scikit-Learn [4] se tornou o principal kit de ferramentas de aprendizado de máquina para programadores em Python (MCKINNEY, 2012). Juntamente com o Pandas, Statsmodels [5] e IPython [6]. o Scikit-Learn tem sido fundamental para permitir que o Python seja uma linguagem de programação de uma ciência de dados produtiva (MCKINNEY, 2012).

O Scikit-Learn expõe uma ampla variedade de algoritmos de aprendizado de máquina, supervisionados e não supervisionados, usando uma interface consistente e orientada a tarefas, permitindo uma comparação fácil de métodos para uma determinada aplicação. Como ele depende do ecossistema científico do Python, ele pode ser facilmente integrado a aplicativos fora da faixa tradicional de análise de dados estatísticos. (PEDREGOSA et al., 2011)

A Application Programming Interface (API) do Scikit-Learn foi desenvolvida com os seguintes princípios orientadores, conforme descrito no documento da biblioteca (VANDERPLAS, 2016) :

1. Consistência : Todos os objetos compartilham uma interface comum extraída de um conjunto limitado de métodos, com documentação consistente.
2. Inspeção : Todos os valores de parâmetros especificados são expostos como atributos públicos.
3. Hierarquia de objetos limitados : somente algoritmos são representados por classes Python; conjuntos de dados são representados em formatos padrão (matrizes

[4] <https://scikit-learn.org/>
[5] <https://www.statsmodels.org/>
[6] <https://ipython.org/>

NumPy, Pandas DataFrame, matrizes esparsas SciPy) e nomes de parâmetros usam strings Python padrão.

4. Composição : Muitas tarefas de aprendizado de máquina podem ser expressas como sequências de algoritmos mais fundamentais, e o Scikit-Learn faz uso disso sempre que possível.

5. Padrões sensíveis : quando os modelos exigem parâmetros especificados pelo usuário, a biblioteca define um valor padrão apropriado.

Na prática, esses princípios tornam o Scikit-Learn muito fácil de usar, uma vez que os princípios básicos são compreendidos. Todo algoritmo de aprendizado de máquina no Scikit-Learn é implementado por meio da API *Estimator*, que fornece uma interface consistente para uma ampla variedade de aplicativos de aprendizado de máquina (BUITINCK et al., 2013).

3.7.4 matplotlib

O matplotlib[7] é uma biblioteca para criar gráficos 2D de matrizes em Python. Embora tenha suas origens na emulação do MATLABTM, pode ser usado sem prejuízos no paradigma orientado a objetos. O matplotlib é escrito principalmente em Python puro, faz uso pesado do Numpy e de outros códigos de extensão para fornecer desempenho mesmo para grandes matrizes (HUNTER; DALE, 2007). Um dos pontos particularmente interessantes nessa biblioteca é sua facilidade para a criação de gráficos, através de poucos comandos com uma facilidade considerável para criação de histogramas, o que sem dúvida impacta positivamente na mineração de dados.

A escolha da utilização dessa biblioteca se deve ao fato de que o matplotlib torna a plotagem científica muito direta, combinando facilidade e bom desempenho (DEVERT, 2014).

3.7.5 IPyvolume

O IPyvolume[8] é uma biblioteca Python para visualizar volumes 3D e glifos (por exemplo, gráficos de dispersão 3D) no *Jupyter notebook*, com configuração e esforços mínimos. Atualmente, está na versão pré-1.0. O método *volshow* do IPyvolume é para arrays 3D e o *imshow* do matplotlib é para arrays 2D (BREDDELS, 2016).

[7] <https://matplotlib.org/>
[8] <https://ipyvolume.readthedocs.io/en/latest/>

3.7.6 Pandas

O Pandas [9] é uma biblioteca open source, licenciada pelo BSD [10], que fornece estruturas de dados de alto desempenho e fáceis de usar com ferramentas de análise de dados para a linguagem de programação Python.

Pandas fornece estruturas e funções de dados de alto nível projetadas para tornar a manipulação de dados estruturados ou tabulares rápidos, fáceis e expressivos. Combina as ideias de computação de *array* de alto desempenho do Numpy com o capacidades de manipulação de dados de planilhas e bancos de dados relacionais como Structured Query Language (SQL) (MCKINNEY, 2012).

3.7.7 Numpy

O Numpy[11] é uma biblioteca muito popular no uso em aplicações cientificas, fornecendo muitas estruturas de dados e algoritmos necessários para esse segmento (MCKINNEY, 2012). Numpy contém, entre outras coisas (MCKINNEY, 2012):

- Um *array* de objeto de matriz multidimensional rápido e eficiente.
- Funções para executar cálculos elementares com matrizes ou operações matemáticas de cal entre matrizes.
- Ferramentas para ler e gravar conjuntos de dados baseados em *array* em disco.
- Operações de álgebra linear, transformada de *Fourier* e geração de números aleatórios.
- Uma API C madura para permitir que extensões Python e códigos C ou C ++ nativos acessem Estruturas de dados e instalações computacionais da Numpy.

[9] <https://pandas.pydata.org/>
[10] <https://opensource.org/licenses/BSD-3-Clause>
[11] <https://www.numpy.org/>

4 PERFIL SOCIOECONÔMICO DO INGRESSANTE

Neste capítulo será abordado a análise do perfil do ingressante do curso de Ciência da Computação nos anos de 2010 a 2018 respectivamente. Os dados originaram do processo do SiSU[1], utilizado para o ingresso em qualquer curso da Universidade. Nele constam dados oriundos do ENEM, que basicamente informam a UF e município de origem, bem como a forma de ingresso e a nota por área de competência e final do ENEM.

Os dados possuem o cruzamento de informações da situação atual do discente se, por exemplo, consta como aluno formado, regular, entre outros atributos que podem ser visualizados em Figura 12.

A análise do perfil socioeconômico corrobora um melhor entendimento do domínio que se está inserido um estudo de EDM, captando melhor o panorama dos dados histórico que a instituição dispõe para se gerar o conhecimento. Essa análise irá basicamente seguir uma observação estatística descritiva básica. Estatísticas descritivas frequentemente envolvem descrições de distribuições de dados e relacionamentos (ROMERO; VENTURA, 2013).

Na Figura 13 é possível visualizar cada etapa que foi seguida norteando-se pelo fluxo do processo SEMMA, que serviu como base nos tópicos que serão abordados nas próximas seções. Nessa análise não será efetivada a criação de um modelo, que o processo assume como sendo um algoritmo aprendizado de máquina, mas sim a concepção de algumas inferências que irão contribuir a geração de conhecimento. Apesar de não terem os dados associados a análise do perfil discente, para preservar o anonimato dos discentes.

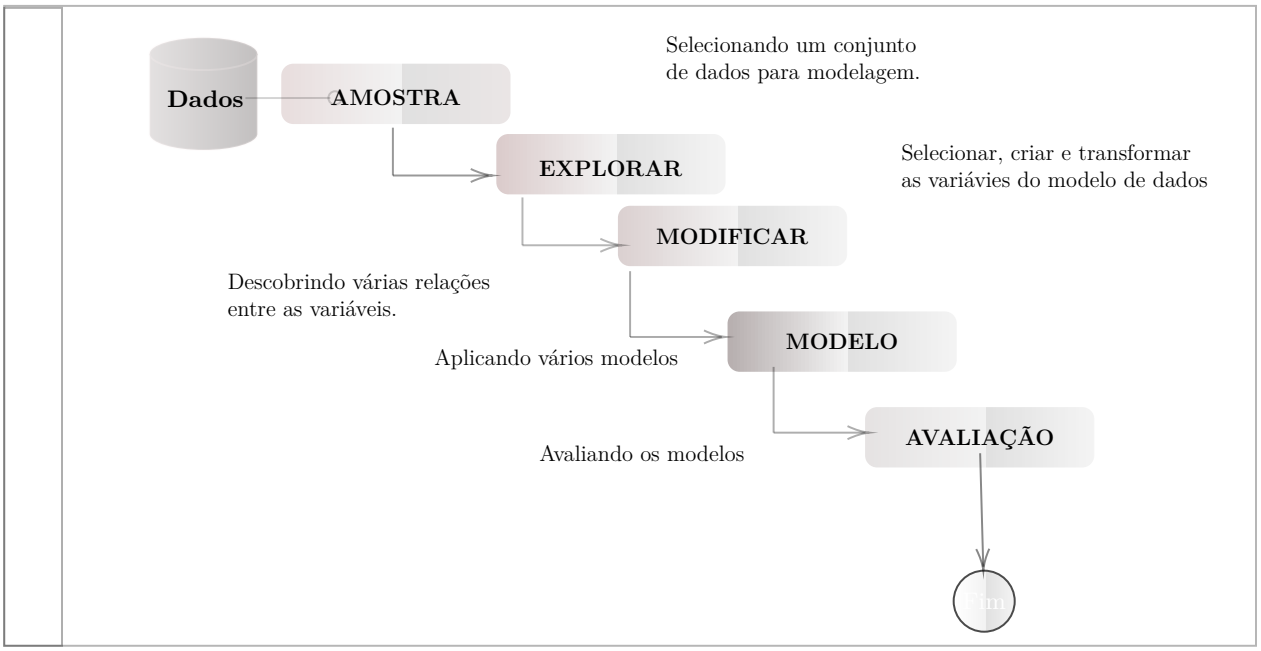
Figura 12 – Situação dos discentes - base: SiSU/ENEM

SITUAÇÃO
Abandono
Aluno Regular
Cancelamento
Cancelamento Sisu
Desligamento
Formado
Reingresso por Novo Vestibular
Transf. Interna Por Reopcao de Curso
Transferido

Fonte: Própria Autora.

[1] <http://www.sisu.mec.gov.br/>

Figura 13 – Processo SEMMA



Fonte: Adaptado: Azevedo e Santos (2008).

4.1 Etapa de amostragem

Os dados para se adequar melhor a biblioteca Pandas e geração de gráficos correspondentes foram pré-processados, ou seja, feito a limpeza e eliminação de quaisquer ruídos que pudessem interferir na análise e geração de gráficos e apontamentos estatísticos, como por exemplo atributos faltantes em quaisquer coluna, atributos com valoração diferente do padrão seguido pela tabela de dados. A estratégia utilizada para a resolução desses problemas pertinentes a dados faltantes foi a amostragem previsto no processo SEMMA, conforme abordado no Capítulo 3, seção 3.2.

Na etapa da amostragem foram eliminados 8 registros que estavam muito inconsistentes pois não haviam quaisquer valoração para as competências específicas abordadas no ENEM, atributos faltantes da nota final, indicação de forma de ingresso, UF e município de origem. Os dados correspondiam um montante de 450 registros antes da eliminação, restando portanto 442 registros.

A falta de padronização dos atributos na rotulagem "Forma de ingresso", denominada como sendo "Ação afirmativa", foi resolvida com a modificação para siglas comumente usadas no processo SiSU, deixando apenas as siglas A1 e V419, adotadas pela Instituição. Eliminando assim a redundância da informação. Na Tabela 3, é explicado as siglas e a ação afirmativa correspondente.

Um outro cuidado com relação aos dados foi a eliminação de quaisquer informações diretas, como nome ou nº do processo ENEM/SiSU que pudessem vincular a um

determinado discente, sendo apresentados portanto de forma anônima.

Tabela 3 – Forma de Ingresso ENEM/SiSU

Nomenclatura	**Ação afirmativa**
AC	Ampla concorrência.
A1	Vagas reservadas a candidatos com necessidades educacionais especiais.
L1	Vagas reservadas a candidatos com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o Ensino Médio em escolas públicas.
L2	Vagas reservadas a candidatos autodeclarados pretos, pardos ou indígenas, com renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o Ensino Médio em escolas públicas.
L3	Vagas reservadas a candidatos que, independentemente da renda, tenham cursado integralmente o Ensino Médio em escolas públicas.
L4	Vagas reservadas a candidatos autodeclarados pretos, pardos ou indígenas que, independentemente da renda, tenham cursado integralmente o Ensino Médio em escolas públicas.
L9	Candidatos com deficiência que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas.
V419	Vagas destinadas a candidatos com deficiência.

Fonte: Adaptado: SISTEMA DE SELEÇÃO UNIFICADA (2019).

4.2 Etapa exploratória

Conforme aponta Azevedo e Santos (2008), a etapa exploratória é uma fase de exploração dos dados para se obter entendimento e *insights* de algumas tendências que podem ser descobertas com essa análise inicial, testando agrupamentos e inferências sobre as informações constantes nesses registros. Nessa etapa foi feito a produção de um algoritmo básico para agrupamento das informações para a seleção de determinados critérios que se queria analisar na base, conforme pode ser visualizado na Figura 14.

Nessa etapa também foi feita a geração de gráficos para melhor entendimento do cenário encontrado, resumindo de maneira rápida e objetiva as principais características

das inferências realizadas sobre o domínio. Na Figura 20 é possível verificar a situação dos discentes em relação ao ano.

Figura 14 – Trecho do algoritmo para exploração dos dados com Pandas

```
# -*- coding: utf-8 -*-
"""
@author: Karina Casola
"""
import pandas as pd
base = pd.read_csv('cc - Plan1.csv')
len(base)

abandono=base['SITUAÇÃO']=='Abandono'

enem=base['NU_NOTA_INSCRITO'] <500
enem3=base['NU_NOTA_INSCRITO'] <700

enem500=base[abandono & enem]
enem700=base[abandono & enem3]
```

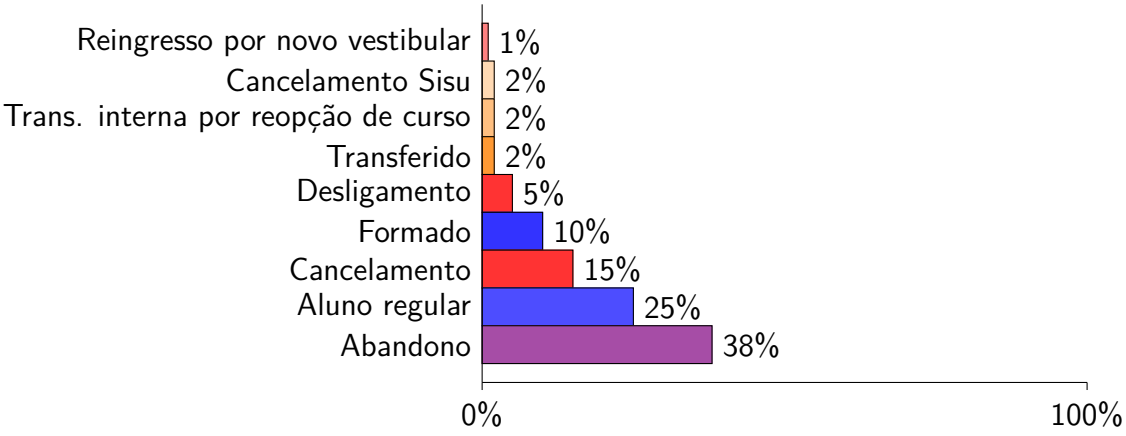
Fonte: Própria Autora.

As Figuras 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27 e 28, representam as ações efetivadas nos dados na busca de relacionamentos causais entre as classes que pudessem incorrer em uma maior incidência da evasão destes discentes. As associações das classes surgiram de inferências comumente apregoadas nessa área de domínio, como por exemplo, a associação do desempenho nas áreas de competência do ENEM com o aumento das chances de evasão, para se comprovar se há de fato alguma relação direta entre elas.

A separação entre as formas de ingresso que estipulam um valor de renda capita também foram analisados para verificar se existia algum impacto econômico inicial incidência de casos de evasão nesse grupo específico. Porém a análise desse fator é sem dúvida um grande desafio: pelo fato de se embasar apenas pela forma de ingresso, que muitas vezes, embora haja o enquadramento, não é a opção do discente devido a burocracia e não necessidade de opção por alguma ação afirmativa que exija maiores trâmites pela pontuação alcançada no ENEM, ou seja, não revela de forma efetiva o quanto esse fator pode impactar na evasão. Entretanto conforme citado em Capítulo 2, e nos estudos de Tinto (1975), esse é um fator agravante na permanência desse discente. Um outro fator é que as universidades não dispõem de dados econômicos dos seus discentes de maneira formal, sem uma pesquisa direta: somente daqueles que possuem bolsas para subsidiar parte das necessidades econômicas que objetivam aumentar a permanência deste discente,

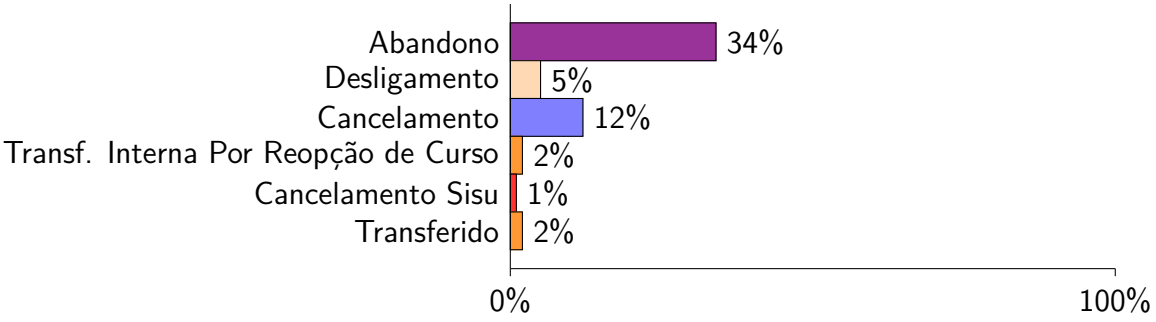
o que inviabiliza observar um panorama geral nesse sentido. No Capítulo 6, seção 6.1 é discutido de maneira detalhada os dados obtidos nessa análise.

Figura 15 – Discentes agrupados pela situação



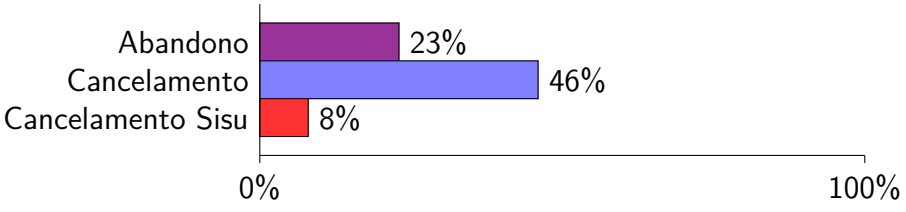
Fonte: Própria Autora.

Figura 16 – Situação da evasão dos discentes UF RS



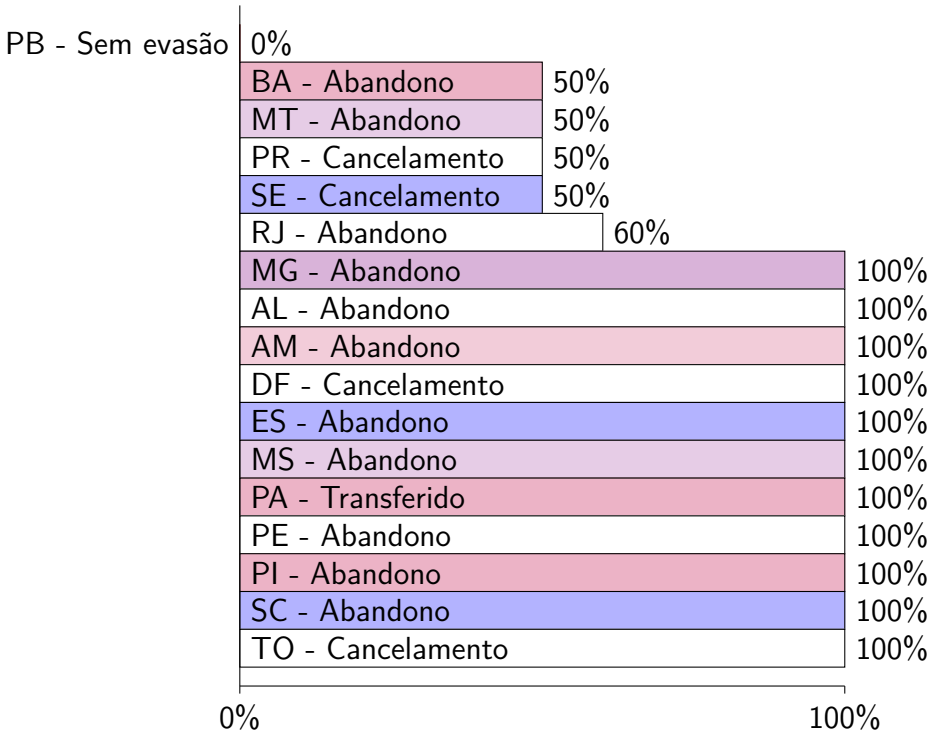
Fonte: Própria Autora.

Figura 17 – Situação da evasão dos discentes UF SP



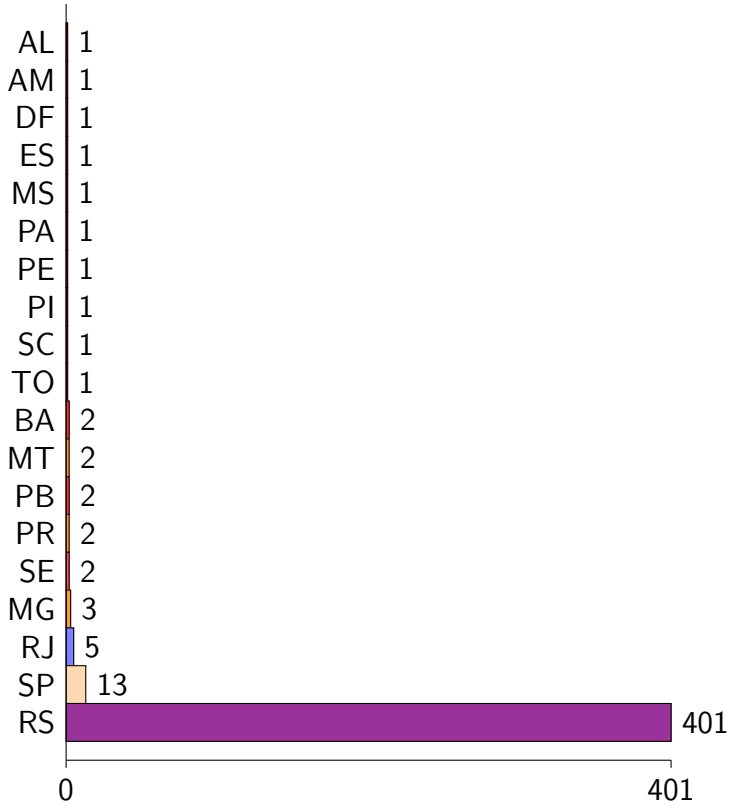
Fonte: Própria Autora.

Figura 18 – Situação da evasão dos discentes demais UFs



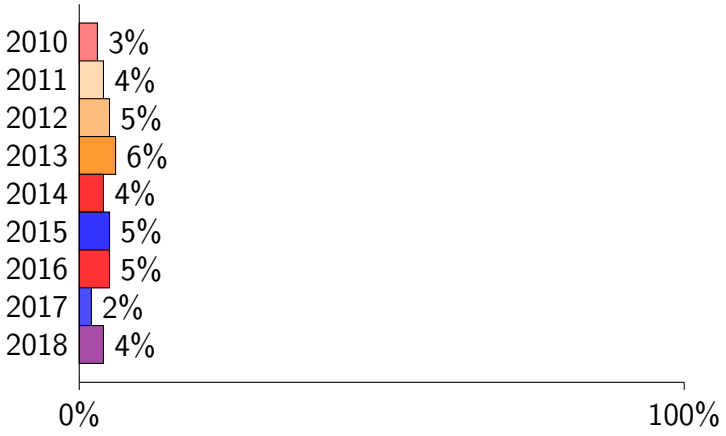
Fonte: Própria Autora.

Figura 19 – Total de discentes por UF



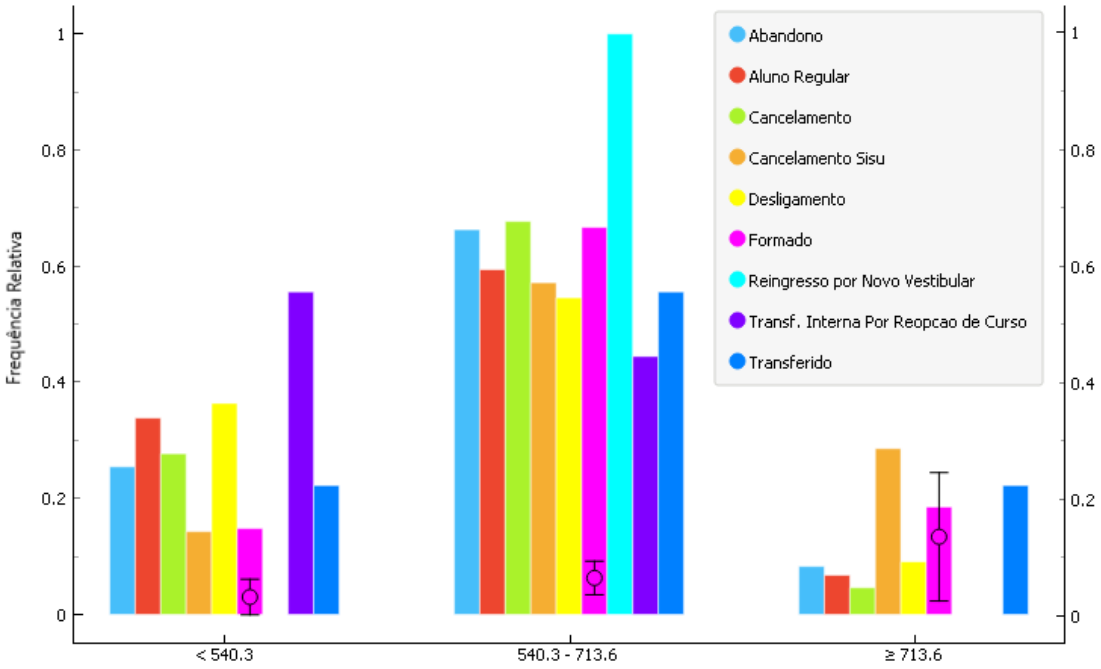
Fonte: Própria Autora.

Figura 20 – Abandono dos discentes por ano - base: SiSU/ENEM



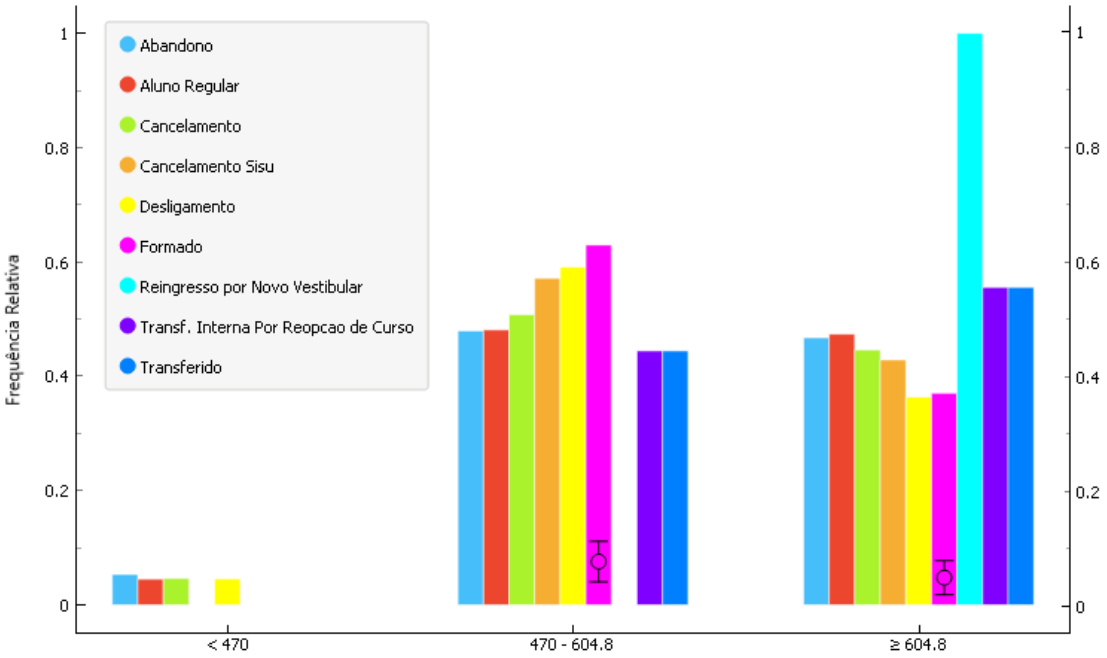
Fonte: Própria Autora.

Figura 21 – Situação dos discentes - por área de competência SiSU/ENEM: Ciências Matemáticas e suas tecnologias



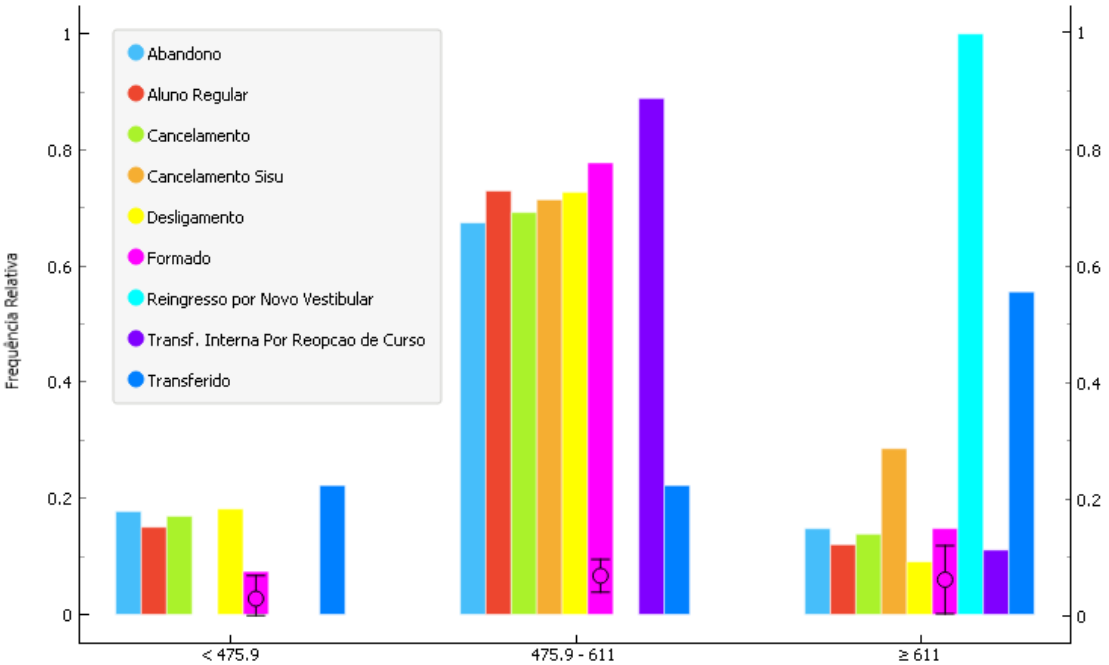
Fonte: Própria Autora.

Figura 22 – Situação dos discentes - por área de competência SiSU/ENEM: Ciências Humanas e suas Tecnologias



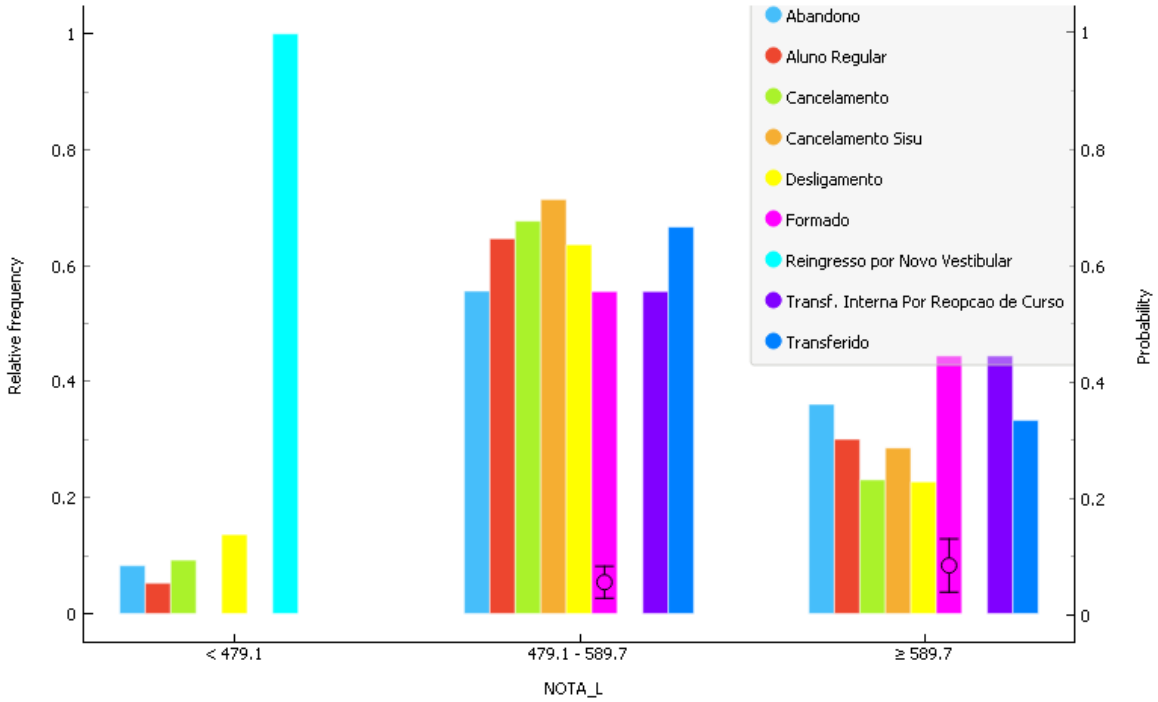
Fonte: Própria Autora.

Figura 23 – Situação dos discentes - por área de competência SiSU/ENEM: Ciências da Natureza e suas Tecnologias



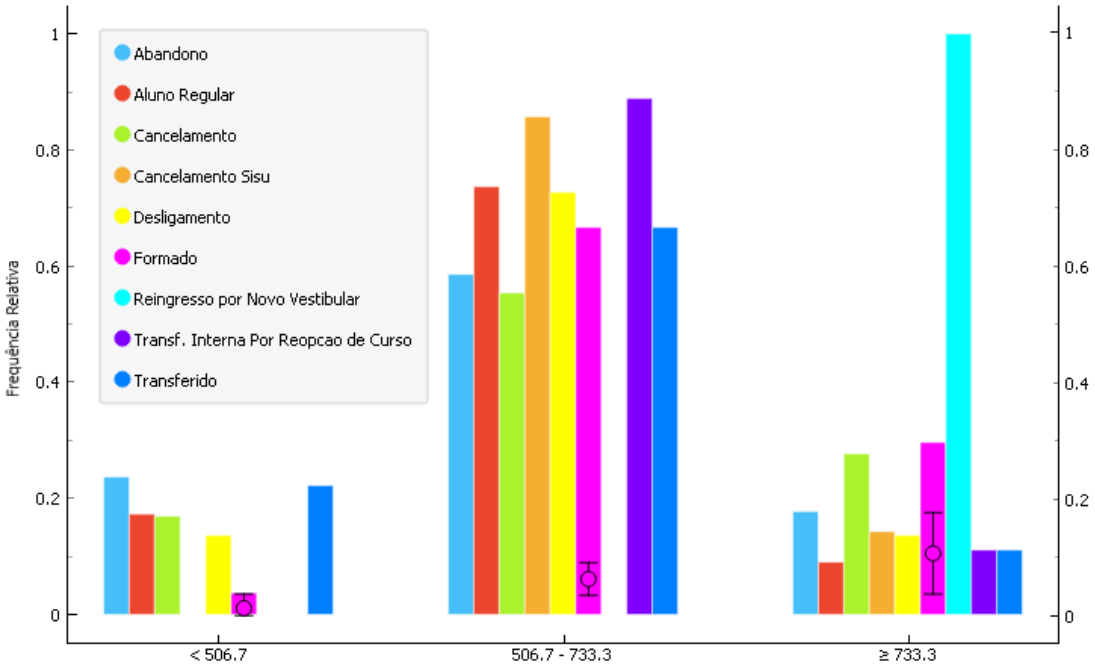
Fonte: Própria Autora.

Figura 24 – Situação dos discentes - por área de competência SiSU/ENEM: Linguagens, Códigos e suas Tecnologias



Fonte: Própria Autora.

Figura 25 – Situação dos discentes - por área de competência SiSU/ENEM: Redação

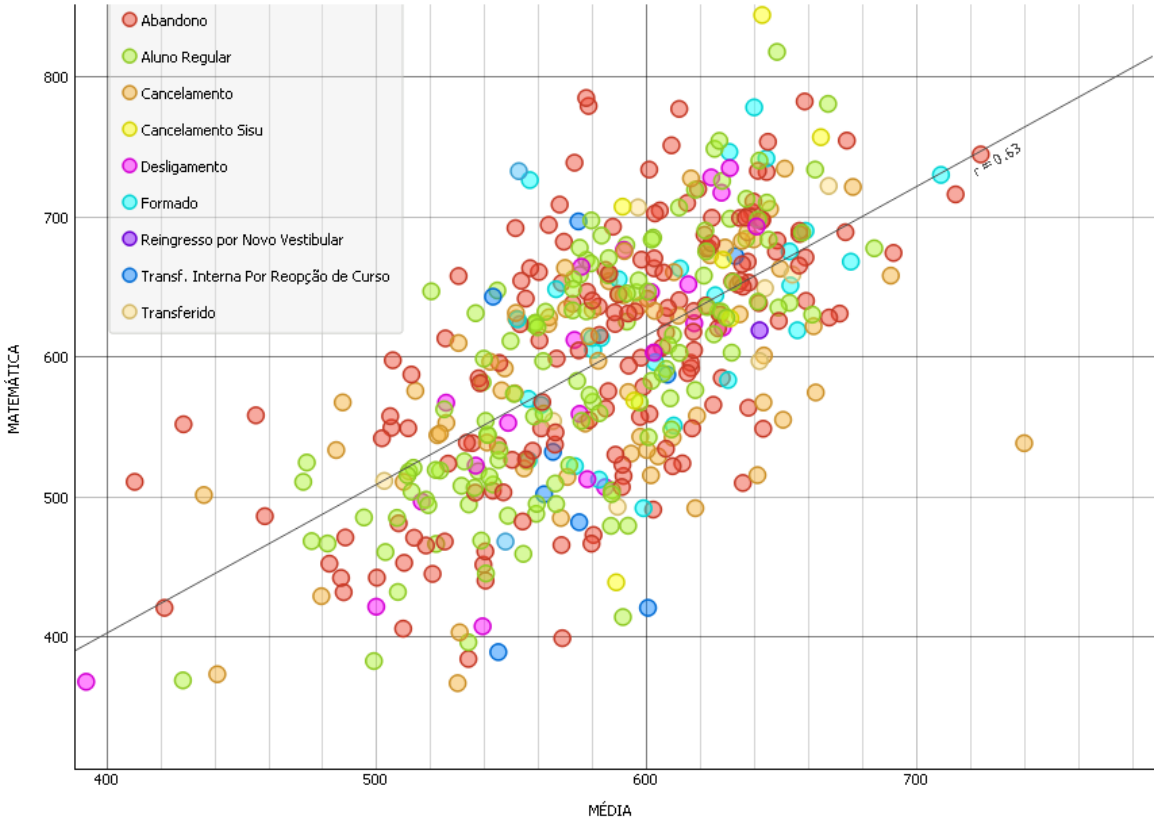


Fonte: Própria Autora.

Figura 26 – Situação dos discentes - Média SiSU/ENEM: Redação

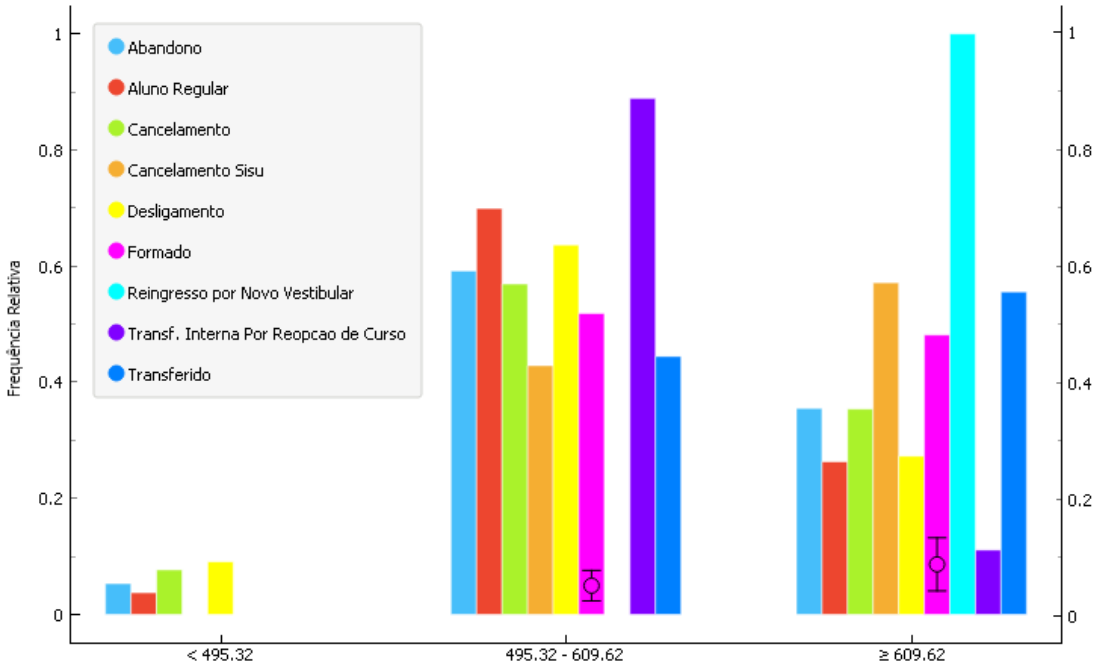
Fonte: Própria Autora.

Figura 27 – Situação dos discentes - Média SiSU/ENEM: Matemática



Fonte: Própria Autora.

Figura 28 – Situação dos discentes - Nota geral SiSU/ENEM



Fonte: Própria Autora.

4.3 Lições do capítulo

Este capítulo procurou demonstrar as etapas de uma análise exploratória seguindo o fluxo do processo SEMMA, com o apoio da biblioteca Pandas, disponível para linguagem de programação Python. A biblioteca Pandas possui uma gama de opções para se efetivar uma análise, com diversos recursos estatísticos que viabilizam juntamente com Python o processo de criação de conhecimento.

A análise exploratória dentro do contexto educacional se torna imprescindível quando não se tem um volume considerável de dados históricos, inclusive quando há a presença de dados faltantes ou se está no meio do período letivo, conforme salienta Romero et al. (2010). A possibilidade de se criar *insights* com essa prática possibilita a transmissão de maneira visual desse conhecimento para toda equipe de planejamento educacional, de acordo com Romero et al. (2010).

No caso deste estudo, o objetivo principal de se efetivar uma análise do perfil socioeconômico, foi basicamente complementar a da base de dados históricos do discente para se tentar criar um panorama sobre outros aspectos não abordados, que extrapolam a conjuntura dos componentes curriculares.

Seguir um fluxo de processo formal, como foi o caso da adoção de maneira adaptada o SEMMA foi fundamental para a melhor organização do trabalho de análise, subdividido-o em etapas-chave a serem cumpridas/observadas.

Os principais desafios encontrados nessa etapa basicamente consiste no tratamento dos dados e atributos com valoração redundante, faltante ou discrepante de uma determinada nomenclatura. O modelo formal estipulado pelo SiSU para designar à forma de ingresso nem sempre é utilizada na sua íntegra pelas IFESs, o que acaba tornando dispendioso a execução da limpeza dos dados nesse segmento.

5 ANÁLISE DO PERFIL DISCENTE

Nesse capítulo são apresentadas as etapas de desenvolvimento dessa análise, seguindo SEMMA um processo lógico e baseado nos preceitos da descoberta de conhecimento em base de dados - KDD (SCHEUER; MCLAREN, 2012; ROMERO et al., 2010), adaptado para o uso do arcabouço tecnológico estipulado e abordado no Capítulo 3. Nas seções seguintes são apresentados características dos dados, bem como cada fase do processo e suas respectivas tarefas.

5.1 Base de dados

A base de dados foi obtida do sistema Gestão Unificada de Recursos Institucionais (GURI), da Universidade Federal do Pampa, Unipampa, *campus* Alegrete-RS. Que é um sistema de Informação capaz de oferecer funcionalidades de controle quanto à dados históricos dos discentes, docentes, cursos e disciplinas (CARVALHO et al., 2012).

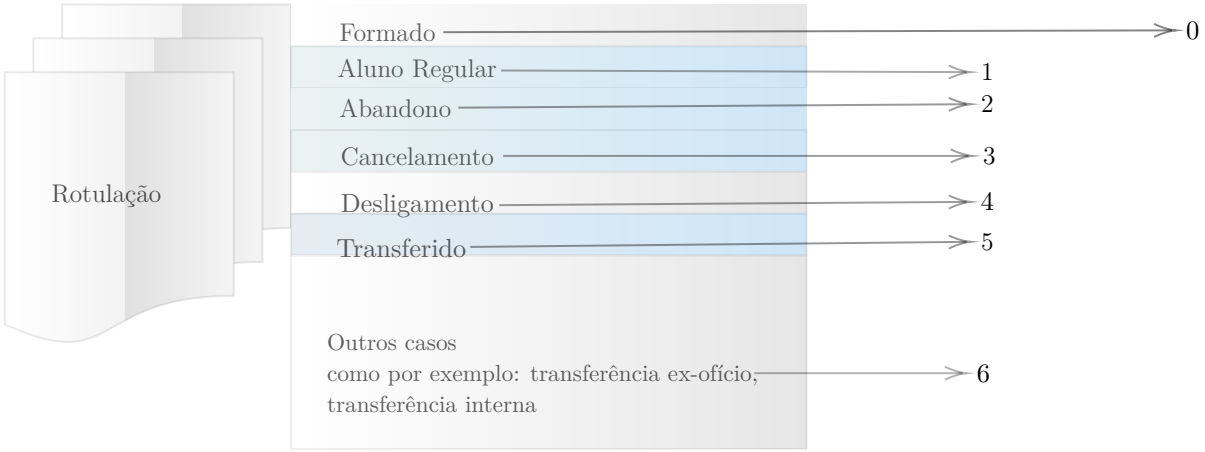
Os dados foram requisitados pelo orientador do TCC, Prof. Dr. Alessandro Bof de Oliveira, que possui acesso ao sistema, assim como qualquer docente da instituição. Os dados foram processados pelo orientador e só posteriormente, após o processo de despersonalização, foram repassados já na forma final para a autora do TCC.

A técnica de extração de características mais comumente utilizada na classificação educacional é discretização. Na discretização, o intervalo de valores numéricos é dividido em intervalos, que será usado como novos valores de atributo (ROMERO et al., 2010).

Os dados originais possuem um eixo temporal de 2009 até 2018 e foram utilizados para gerar uma tabela formada por 522 linhas as quais representam os alunos, e 13 colunas que refletem: 11 colunas representando as disciplinas do primeiro ano do curso de ciência da computação, 1 coluna referente ao trancamento do semestre e a última coluna referente a rotulação da situação dos discentes. A rotulação da situação dos discentes foi definida como mostra a Figura 29, na Figura 30 é possível verificar o procedimento de despersonalização dos dados.

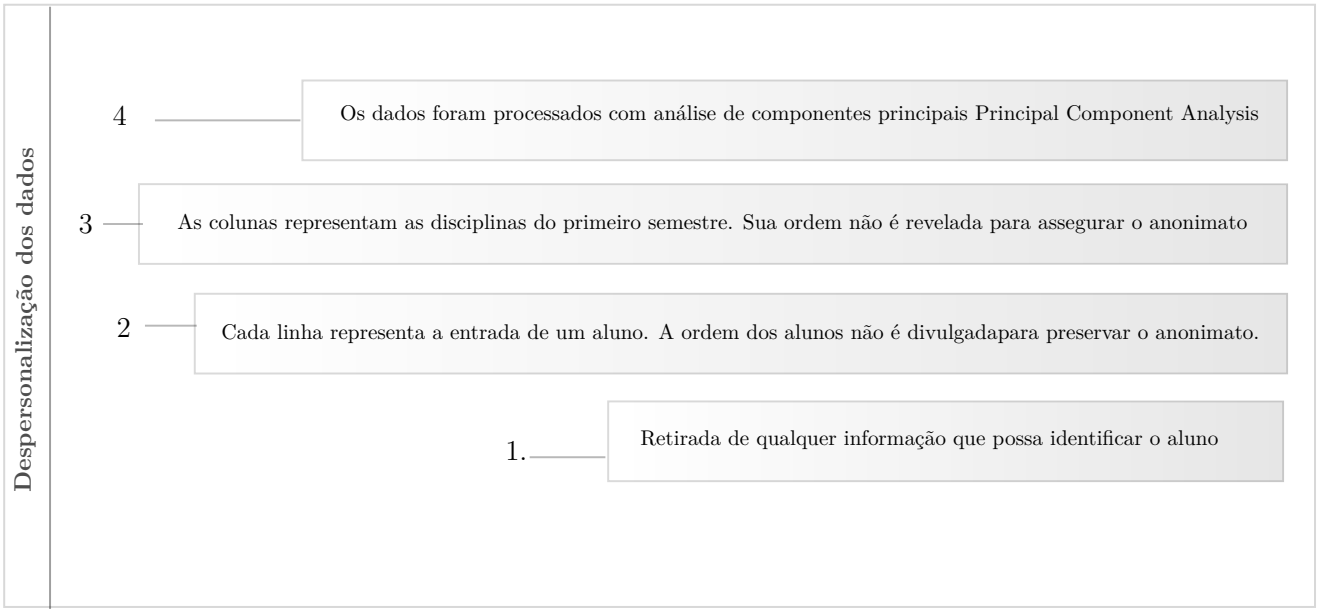
Os dados representam o comportamento dos alunos matriculados no primeiro ano do curso de Ciência da Computação nos componentes curriculares referentes aos dois primeiros semestres. Essa etapa pode ser definida como pré-processamento (CASTRO; FERRARI, 2016; ROMERO et al., 2010) e transformação, a fim de obter conjuntos de dados limpos e prontos para aplicar as técnicas de mineração de dados. A Principal Component Analysis (PCA) efetivada nos dados, objetiva essencialmente reduzir um conjunto original de variáveis, para um conjunto menor de componentes não correlacionados que representam a maioria das informações encontradas nas variáveis originais. A técnica é mais útil quando um grande número de variáveis,e com isso, proíbe a interpretação efetiva das relações entre os objetos (LINTING et al., 2007). Ao reduzir a dimensionalidade, interpretamos alguns componentes em vez de um grande número de variáveis. com o objetivo de transformar os dados para um espaço com maior variância e aumentar a

Figura 29 – Rotulação da situação dos discentes



Fonte: Própria Autora.

Figura 30 – Etapa de despersonalização dos dados



Fonte: Própria Autora.

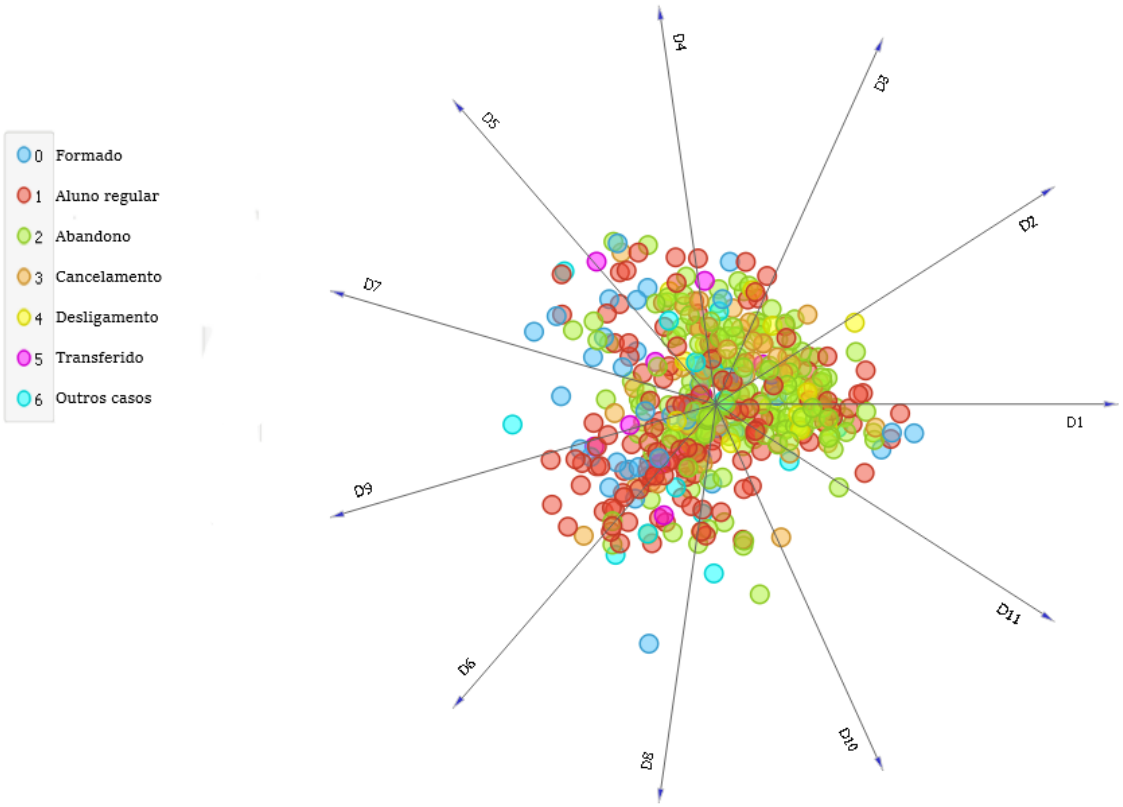
segurança do anonimato.

Além do intuito de manter e aumentar a segurança do anonimato, a discretização suaviza o efeito do ruído e permite modelos mais simples, que são menos propensos à *overfitting* conforme aponta Romero et al. (2010).

5.2 Etapa exploratória

Essa etapa consiste na visualização e entendimento dos dados na busca por padrões antes da criação de um modelo preditivo baseado em aprendizado de máquinas. A base de dados foi analisada utilizando Pandas e Python com a normalização da base no formato: "Comma Separated Values (CSV)", usando a nomenclatura "SIT", para designar a coluna referente à situação dos discentes. Posteriormente a informação foi dispota em um histograma que é uma representação da distribuição dos dados por meio de um gráfico de barras, normalmente de um ou mais atributos da base (CASTRO; FERRARI, 2016).

Figura 31 – Projeção linear das disciplinas com relação à situação



Fonte: Própria Autora.

A escolha da visualização dos dados por meio de gráficos se efetivou, pois conforme Castro e Ferrari (2016), permite uma compreensão mais fácil da distribuição dos valores de

um atributo. A análise feita na base de dados foi com relação à distribuição da rotulagem da situação dos discentes em relação às disciplinas. Os nomes das disciplinas foram definidos como $D1..D11$, pois a base foi discretizada conforme abordado na seção 5.1 para preservar o anonimato dos discentes e qualquer inferência que poderia ser feita nesse sentido. Na Figura 31 é possível verificar essa distribuição.

Na Figura 31 é possível ponderar que a classe **abandono** desponta com maior índice de discentes, seguida da classe **aluno regular**, **cancelamento**,**formado**, **desligamento**, **outros casos** e **transferido**. Com relação à distribuição do **abandono** do curso em relação às disciplinas, vimos que as disciplinas $D1$, $D2$, $D3$ e $D4$, como sendo as mais problemáticas nesse cenário, conforme visto na figura Figura 31.

5.3 Modelo preditivo

Nessa seção será abordado o modelo preditivo adotado para esse estudo, ou seja os algoritmos de aprendizado de máquina utilizados para identificar padrões e prover uma previsão do que de fato pode ocorrer. Os algoritmos selecionados para o estudo foram o kNN supervisionado, e as NN abordados na Capítulo 3. No caso do kNN se objetiva efetivar um comparativo de desempenho de uma das técnicas de aprendizado de máquinas supervisionado mais simples, e as NN sua escolha se deve ao fato de sua relevância no contexto histórico e bom desempenho e grande capacidade de reconhecer padrões.

5.3.1 kNN supervisionado

O algoritmo kNN foi utilizado no modo supervisionado pois apresentou maior desempenho na construção do modelo preditivo da base de dados. Um outro fator bastante importante foi a escolha dos parâmetros para a melhora do desempenho do algoritmo. O valor de k pareceu irrelevante, enquanto o parâmetro *"weights"* ou seja pesos fossem calculados pela distância *"distance"* dos pontos de ponderação pelo inverso da sua distância. Nesse caso, os vizinhos mais próximos de um ponto de consulta terão uma influência maior do que os vizinhos mais distantes. O algoritmo apresentou um pior desempenho com a passagem de parâmetro *"uniform"* que todos os pontos em cada vizinhança são ponderados igualmente. Isso se deve ao fato da própria característica que o tratamento PCA que os dados passaram. Na Figura 33 é possível verificar um pequeno trecho da utilização do algoritmo.

Figura 32 – Trecho do algoritmo kNN supervisionado

```
from sklearn import neighbors
import numpy as np

base=np.load('./daccpca.npy')

previsores = base[:,0:-1]
classe = base[:,-1]

from sklearn.model_selection import train_test_split
prev_tr, prev_teste, classe_tr, classe_teste = train_test_split
(previsores, classe, test_size=0.60, random_state=0)

clf = neighbors.KNeighborsClassifier(9, metric='euclidean',weights='distance')
clf.fit(prev_tr, classe_tr)
previsoes = clf.predict(prev_teste)

clf2 = neighbors.KNeighborsClassifier(5, metric='euclidean',weights='distance')
clf2.fit(previsores, classe)
previsoes2 = clf2.predict(previsores)
```

Fonte: Própria Autora.

5.3.2 Algoritmo NN MLP com *backpropagation*

Na Figura 33 se pode visualizar o algoritmo NN, sob a forma de MLP nas duas variantes: onde há a divisão da base em teste e treinamento, e quando a utiliza na sua totalidade.

A função de ativação geralmente é utilizada com dois propósitos: limitar a saída do neurônio e introduzir não linearidade no modelo (CASTRO; FERRARI, 2016). Para o propósito do estudo, ela tem uma grande importância, pois delimita de maneira mais adequada o espaço de características dos dados, para uma melhor separação entre as classes.

Nesse estudo foram feitos testes com a funções logística e tangente hiperbólica, na tentativa de diagnosticar qual teria maior propensão a aumentar o percentual de precisão da previsão da situação dos discentes. A função logística está vinculada à preocupação em limitar o intervalo de variação da derivada da função pela inclusão de um efeito de saturação, como a função logística apresenta valores de ativação apenas em (0, 1), em muitos casos ela é substituída pela tangente hiperbólica que preserva a forma sigmoidal da função logística, mas assume valores positivos e negativos (CASTRO; FERRARI, 2016). Pelas peculiaridades da disposição dos dados a que se mostrou mais eficiente foi a

função tangente hiperbólica, dividindo a base de maneira mais eficaz.

Na Figura 10 do Capítulo 3 é possível verificar o esquemático básico do funcionamento de um neurônio artificial genérico, com a sua respectiva função de ativação.

Figura 33 – Algoritmo NN MLP com *backpropagation*

```
import numpy as np

#carregamento da base de dados em arquivo npy
base=np.load('./daccpca.npy')
len(base)

#divisão da base em previsores e classe
previsores = base[:,0:-2]
classe = base[:,-1]

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
previsores = scaler.fit_transform(previsores)

from sklearn.model_selection import train_test_split
prev_tr, prev_teste, classe_tr, classe_teste =
train_test_split(previsores, classe, test_size=0.30, random_state=0)

from sklearn.neural_network import MLPClassifier
clf = MLPClassifier(verbose = True,
                                max_iter=1000,
                                tol = 0.0000010,
                                solver = 'adam', #otimizador para base de dados grandes
                                hidden_layer_sizes=(3),#camadas ocultas
                                activation='tanh') #função de ativação
clf.fit(prev_tr, classe_tr)
previsoes = clf.predict(prev_teste)


clf2 = MLPClassifier(verbose = True,
                                max_iter=100000,
                                tol = 0.0000010,
                                solver = 'adam', #otimizador para base de dados grandes
                                hidden_layer_sizes=(50),#camadas ocultas
                                activation='tanh') #função de ativação
clf2.fit(previsores, classe)
previsoes2 = classificador2.predict(previsores)

from sklearn.metrics import confusion_matrix, accuracy_score
precisao = accuracy_score(classe_teste, previsoes)
matriz = confusion_matrix(classe_teste, previsoes)

precisao2 = accuracy_score(classe, previsoes2)
matriz2 = confusion_matrix(classe, previsoes2)
```

Fonte: Própria Autora.

5.4 Matriz de confusão

Após a execução dos algoritmos nas duas formas se obteve os seguintes resultados na matriz de confusão, conforme apontam as Tabelas 4, 5 e as Tabelas 6 e 7, respectivamente. Na diagonal principal da matriz de confusão encontram-se os acertos do algoritmo e os percentuais ao lado correspondem aos enganos/erros de classificação.

Como pode-se verificar, os algoritmos erraram mais quando houve a divisão da base em treinamento e teste, pois, justamente os melhores previsores acabaram não participando do treinamento do algoritmo, e conforme Romero et al. (2010) destaca, se o conjunto de dados já é pequeno, não é aconselhável reduzir o conjunto de treinamento mais justamente por essa problemática.

O kNN obteve uma acurácia de 98,65% e o NN 97,89%. Quanto à precisão, ou seja, o VPP conforme visto na Figura 11 do Capítulo 3, efetivando o cálculo no kNN se obtêm uma precisão de 96,17% em relação à classe 1, ou seja, aluno regular. Já no NN, se obtêm uma precisão 93,63% dessa mesma classe. Como se pode observar, não houve a classificação do VPN, que seria bastante impactante principalmente se fosse com relação à classe 2, da evasão. Na Tabela 8 é apresentado um comparativo entre os dois algoritmos sem a divisão da base.

Tabela 4 – Matriz de confusão do algoritmo kNN sem divisão de base com o percentual de erros e acertos

Rótulo Classe	Predição						
	0	1	2	3	4	5	6
0	100%	0	0	0	0	0	0
1	0	96,17%	3,83%	0	0	0	0
2	0	0	100%	0	0	0	0
3	0	0	2,22%	97,78%	0	0	0
4	0	0	0	0	100%	0	0
5	0	0	0	0	0	100%	0
6	0	0	0	0	0	0	100%

Fonte: Própria Autora.

Já a base sendo usada em íntegra os resultados melhoraram, mostrando que existe um padrão do comportamento da situação dos discentes.

Tabela 8 – Comparativo entre kNN e NN sem divisão da base

Algoritmos		**Qdt. de erros**	**Percentual do erro**	**Zero Rules**
kNN	**Acurácia** 98,65%	7	1,34%	47,87%
NN	**Acurácia** 97,89%	11	2,10%	47,87%

Fonte: Própria Autora.

Tabela 5 – Matriz de confusão do algoritmo NN sem divisão de base com o percentual de erros e acertos

Rótulo Classe	Predição						
	0	1	2	3	4	5	6
0	100%	0	0	0	0	0	0
1	0	93,63%	6,37%	0	0	0	0
2	0	0	100%	0	0	0	0
3	0	0	2,22%	97,78%	0	0	0
4	0	0	0	0	100%	0	0
5	0	0	0	0	0	100%	0
6	0	0	0	0	0	0	100%

Fonte: Própria Autora.

Tabela 6 – Matriz de confusão do algoritmo kNN com divisão de base

Rótulo Classe	Predição						
	0	1	2	3	4	5	6
0	29,54%	25%	9,09%	0	0	0	0
1	5,95%	35,76%	20,38%	0,63%	0	0	0
2	0,43%	10,92%	45,41%	0,43%	0	0	0
3	2,22%	17,78%	42,22%	4,44%	0	0	0
4	4,54%	18,18%	50%	4,54%	0%	0	0
5	20%	30%	10%	0	0	0%	0
6	0	33,33%	33,33%	0	0	0	0%

Fonte: Própria Autora.

Tabela 7 – Matriz de confusão do algoritmo NN com divisão de base

Rótulo Classe	Predição						
	0	1	2	3	4	5	6
0	2,27%	9,09%	0	0	0	0	0
1	0,63%	5,73%	6,36%	0	0	0	0
2	0	3,56%	13,10%	0	0,43%	0	0
3	0	4,45%	8,89%	0%	0	0	0
4	0	4,55%	31,81%	0	0%	0	0
5	0	0	10%	0	0	0%	0
6	0	0	0	0	0	0	0%

Fonte: Própria Autora.

5.5 Validação cruzada

Na Figura 34 pode ser visualizado o o algoritmo desenvolvido para a execução de 30 testes aplicados nos algoritmos kNN e NN, para verificar o quanto os algoritmos conseguem generalizar o aprendizado, e o quanto eles entendem os padrões dos dados. A validação cruzada justamente separada em parte dos dados de maneira aleatória, fechando

os dados em treino e teste.

Figura 34 – Validação cruzada kNN e NN

```
import numpy as np
base =np.load('./daccpca.npy')
previsores = base[:,0:-2]
classe = base[:,-1]

from sklearn.preprocessing import Imputer
imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis = 0)
imputer = imputer.fit(previsores[:,0:-2])
previsores[:,0:-2] = imputer.transform(previsores[:,0:-2])

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
previsores = scaler.fit_transform(previsores)

from sklearn.model_selection import StratifiedShuffleSplit
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPClassifier

result30 = []
for i in range(30):

    kfold = StratifiedShuffleSplit(n_splits=1, test_size = 0.1, random_state = i)

    result1 = []
    for indice_treinamento, indice_teste in
    kfold.split(previsores,
    np.zeros(shape=(classe.shape[0], 1))):

          clf = KNeighborsClassifier(n_neighbors=5, metric='euclidean', p=2)

          clf = MLPClassifier(verbose = True, max_iter = 1000,
                              tol = 0.000010, solver='adam',
                               hidden_layer_sizes=(3), activation = 'tanh',
                              batch_size=200, learning_rate_init=0.001)

clf.fit(previsores[indice_treinamento], classe[indice_treinamento])
        previsoes = clf.predict(previsores[indice_teste])
        precisao = accuracy_score(classe[indice_teste], previsoes)
        resultados1.append(precisao)
    resul1 = np.asarray(resultados1)
    media = resultados1.mean()
    result30.append(media)

result30 = np.asarray(resultados30)
result30.mean()
for i in range(result30.size):
    print(str(result30[i]).replace('.', ','))
```

Fonte: Próprio Autor

Como parâmetro foi utilizado o **StratifiedShuffleSplit** que é uma mesclagem do método **StratifiedKFold** e **ShuffleSplit**, que retornam dobras aleatórias estratificadas. As dobras são feitas preservando a porcentagem de amostras para cada classe (PEDREGOSA et al., 2011). A escolha do **StratifiedShuffleSplit** se deve pelas peculiaridades da base, como, por exemplo, o tamanho ser muito pequeno e o estilo de dimensionamento dos dados. A execução dos testes se deu com a base divida em 70%/30%, e 90%/ 10%, para verificar o comportamento dos algoritmos nesses dois cenários. O valor dos 30 testes foi armazenado em uma planilha, calculando-se a média e o *ranking* dos algoritmos.

Após a execução dos testes, foi feito em R a análise estatística dos algoritmos, ou seja, se existia diferença estatística significante entre eles, aplicando os testes de Friedman e Nemenyi conforme pode ser visto em Figura 35.

Figura 35 – Algoritmo em R Friedman e Nemenyi

```
require(tsutils)

bd <- read.csv("/Users/Karina/Documents/Algoritmos-TCC/dados.csv")
bdl <- as.matrix(bd)
tsutils::nemenyi(bdl,conf.int=0.95, plottype="vline")
```

Fonte: Própria Autora.

6 DISCUSSÃO DOS RESULTADOS

Esse capítulo aborda as discussões dos resultados obtidos nas análises do Perfil socioeconômico no seção 6.1, e a análise do perfil discente no seção 6.2.

6.1 Perfil socioeconômico do ingressante

Para fins de análise da evasão no seu modelo clássico, foi ponderada a situação dos discentes que constavam como "abandono" em um primeiro momento, e posteriormente foi feita a análise da evasão de maneira geral.

Os resultados da análise exploratória do perfil socioeconômico do ingressante do curso de Ciência da computação do período de 2010 a 2018, apontaram que a situação "abandono" obteve um percentual de 38%. Analisando separadamente por competência, se confirma novamente esse alto índice de abandono, tendo pouca variação entre as competências da avaliação do ENEM, apenas a área de Ciências Humanas e suas Tecnologias e Ciências Matemáticas e suas tecnologias a mais em relação as outras, na faixa de 537,4 a 638,5 para Ciências Humanas e 540,3 a 713,6 para Ciências Matemáticas. Esse valor se deve ao alto índice de ingressantes nessa faixa de pontuação, com que se conclui não ser um fator tão significativo na análise da evasão dos discentes.

As competências que apresentaram maior disparidade entre o abandono e o aluno regular foram as de Linguagens, Códigos e suas Tecnologias, com pontuação no exame entre 567 a 634 e Redação na faixa de 620 a 790 pontos.

O índice da evasão por UF, vimos que o Rio Grande do Sul (RS) despontou com a maior taxa, isso se explica pela incidência de discentes do RS ser maior em comparação de outras localidades conforme se pode verificar na Figura 19.

O efeito observado Figura 20 a partir do ano 2017 que se constata a taxa de discentes formados maior que nos anos anteriores, pode ser um indicativo de retenção, sendo recomendado trabalhos futuros nesse sentido.

Na Tabela 9 foi feito a análise da situação abandono pela ABI explicado em Tabela 3, para se tentar diagnosticar se existiam tendências econômicas ou outras impeditivas nesse sentido para a permanência do aluno na Instituição. No âmbito econômico, as ações afirmativas desse grupo correspondem a L1, L2, apresentaram uma alta taxa de evasão/abandono quando se verifica o número de ingressantes nessa modalidade, apesar de corresponder 12% do total geral de alunos com a situação de abandono. A ABI denominada AC foi a segunda que teve maior índice de evasão tanto analisando pela quantidade de ingressantes por essa modalidade, tanto se analisarmos o montante geral nessa modalidade, perdendo apenas para a A1 que teve um percentual de 100% em relação aos ingressantes nessa ABI.

Tabela 9 – Classificação abandono por ação afirmativa SiSU/ ENEM

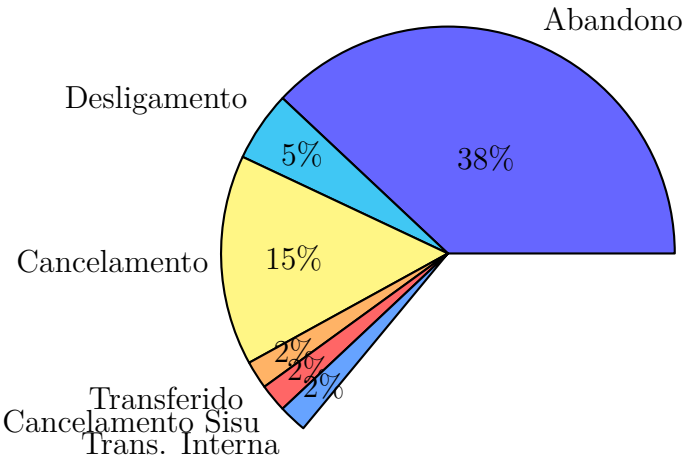
ABI	Abandono por ABI	Total por ABI	% Abandono/ABI	% Abandono/total geral
L1	16	56	29%	9%
L2	5	13	38%	3%
L9	0	1	0%	0%
AC	103	238	43%	61%
L3	32	104	31%	19%
L4	8	19	42%	5%
V419	3	9	33%	2%
A1	2	2	100%	1%

Fonte: Própria Autora.

Para analisar a evasão de um panorama que incluísse o cancelamento, desligamento, cancelamento SiSU e Transferência interna por ABI, foi realizado as estimativas de cada grupo, conforme apontam as Tabelas 10, 11, 12, 13. Na Figura 36 é possível verificar o percentual de cada ABI.

Após a realização da análise de forma individual, foi feita a junção total das estimativas de cada classificação dividida por classe, para se determinar qual das ABI teriam maior índice de evasão englobando todas situações nesse sentido.

Figura 36 – Panorama da evasão



Fonte: Própria Autora.

Tabela 10 – Classificação cancelamento por ação afirmativa SiSU/ ENEM

ABI	Cancelamento por ABI	Total por ABI	% Canc./ABI	% Canc./total
L1	7	56	13%	11%
L2	0	13	0%	0%
L9	0	1	0%	0%
AC	34	238	14%	52%
L3	18	104	17%	28%
L4	4	19	21%	6%
V419	2	9	22%	3%
A1	0	2	0%	0%

Fonte: Própria Autora.

Tabela 11 – Classificação desligamento por ação afirmativa SiSU/ ENEM

ABI	Desligamento por ABI	Total por ABI	% Desl./ABI	% Desl./total
L1	3	56	5%	14%
L2	1	13	8%	5%
L9	0	1	0%	0%
AC	14	238	6%	64%
L3	3	104	3%	14%
L4	0	19	0%	0%
V419	1	9	11%	3%
A1	0	2	0%	0%

Fonte: Própria Autora.

Tabela 12 – Classificação cancelamento SiSU por ação afirmativa SiSU/ ENEM

ABI	Canc.SiSU por ABI	Total por ABI	% Canc.SiSU/ABI	% Canc.SiSU/total
L1	2	56	4%	29%
L2	0	13	0%	0%
L9	1	1	0%	0%
AC	4	238	2%	57%
L3	1	104	1%	14%
L4	0	19	0%	0%
V419	0	9	0%	0%
A1	0	2	0%	0%

Fonte: Própria Autora.

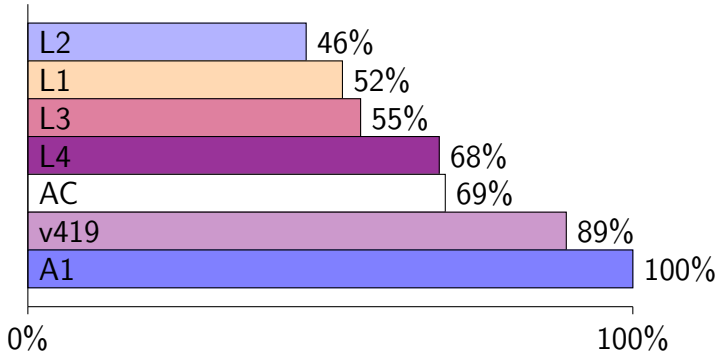
Tabela 13 – Classificação transferência interna por ação afirmativa SiSU/ ENEM

ABI	T.Interna por ABI	Total por ABI	% T.Interna/ABI	% T. Interna/total
L1	1	56	2%	11%
L2	0	13	0%	0%
L9	0	1	0%	0%
AC	5	238	2%	56%
L3	1	104	1%	11%
L4	0	19	0%	0%
V419	2	9	22%	22%
A1	0	2	0%	0%

Fonte: Própria Autora.

O resultado da análise do panorama da evasão pode ser visto na Figura 37. Conforme pode ser observado existe um alto índice de evasão em todos os grupos, levando-se em conta o número de ingressantes. As ABI que tiveram o maior índice de evasão foram a A1 com 100%, e a v419 com 89%. Basicamente podemos verificar que esses grupos necessitam um aporte maior de infraestrutura educacional da Instituição. Além disso, destacam-se AC e L4 com um índice próximo de 70%. A explicação da ação afirmativa correspondente encontra-se na Tabela 3, Capítulo 4.

Figura 37 – Panorama da evasão por ABI



Fonte: Própria Autora.

6.2 Análise do perfil discente

A análise do perfil discente teve basicamente três etapas: a análise exploratória, adoção de modelos preditivos e avaliação desses modelos preditivos. Com a etapa exploratória foi possível verificar a distribuição dos discentes pela sua situação, que posteriormente foi confirmado pelo modelo preditivo. A etapa de avaliação utilizou duas análises: a matriz de confusão e a validação cruzada.

A matriz de confusão permitiu fazer as seguintes conjecturas sobre os dados, tomando por base o algoritmo que teve o melhor desempenho de mapear e prever as classes:

a classe com maior quantidade de registros é a 2, ou seja, cada vez que entrar um novo registro, ele será classificado a *priori* como pertencente à 2. Utilizando a técnica *zero rules* ou zero regras, o acerto mínimo do algoritmo 43,87%, que se obtêm fazendo a seguinte operação Equação 6.1:

$$\frac{ClasseM}{TotalR} X 100. \tag{6.1}$$

Onde a $ClasseM$ representa a classe com quantidade maior de registros e $TotalR$ a quantidade total de registros.

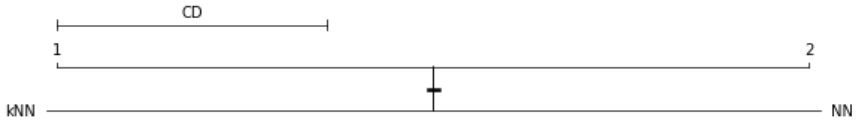
A probabilidade de alguém selecionar a classe correta, levando-se em consideração ao número de classes, ou seja, as "situações" que os discentes podem assumir é de 14,28%. Isso é bem inferior ao acerto mínimo do algoritmo de aprendizado de máquinas, o que se conclui que é interessante sua adoção no planejamento de políticas para inibir a evasão.

Nesse caso podemos deduzir que há um caso de *Overfitting* em relação às NN, ou seja, quando o algoritmo captura o ruído e se ajusta aos dados muito bem (GUPTA, 2015). *Overfitting* acontece quando o modelo é muito complexo em relação ao tamanho dos dados, como é o caso do NN.

Para verificar se esse cenário se repetia, foi feito a validação cruzada dos dois algoritmos executando 30 testes, e posteriormente análise estatística no R de Friedman com Nemenyi. Nas Figuras 38, 39 é possível verificar graficamente esses resultados. Ambas as análises apontaram uma distância crítica de 0,35.

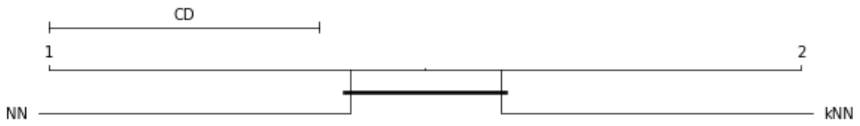
Como se pode observar, não existe uma diferença estatística significativa entre ambos os casos, levando-se em conta a distância crítica calculada.

Figura 38 – Teste 1 diferenças estatísticas entre kNN e NN de 90%/10%



Fonte: Própria Autora.

Figura 39 – Teste 2 diferenças estatísticas entre kNN e NN de 70%/30%



Fonte: Própria Autora.

7 CONSIDERAÇÕES FINAIS

Esse TCC se propôs a efetivar o estudo da evasão utilizando EDM, dos discentes do curso de Ciência da Computação sob dois aspectos: a análise do perfil socioeconômico dos ingressantes através dos dados SiSU/ENEM, referente aos anos de 2010 a 2018 e a situação dos alunos matriculados no primeiro ano nos componentes curriculares referentes aos dois primeiros semestres do curso de 2009 a 2018. Para a realização desse estudo foi utilizado a linguagem de programação Python, a biblioteca Scikit learn, Pandas e demais tecnologias abordadas em Capítulo 3. Observar o aspecto socioeconômico foi importante para testar inferência sobre as informações expostas, servindo também de um olhar mais atento a aspectos que extrapolam a análise puramente dos componentes curriculares do curso. Nessa análise foi constatado que as notas nas competências por área da avaliação ENEM/ SiSU, não tem um impacto tão considerável na permanência desse discente. Mesmo que a evasão de maneira geral tenha um alto índice, existem grupos de maior risco(A1 e v419), que carecem um olhar mais atento da Instituição devido a sua alta vulnerabilidade. Na análise algorítmica da situação dos alunos do curso de Ciência da Computação, através dos algoritmos kNN, NN, o principal enfoque foi a predição: isto é, o quanto o modelo adotado conseguia prever a qual classe estavam os discentes: e consequentemente a possibilidade de evasão. O modelo preditivo obteve uma alta taxa de acurácia e precisão indicando novamente que a classe que representava maior montante era a 2, denominada "abandono" com 43,87%. Isso demonstra que existe um padrão que pode ser mapeado algoritmicamente da evasão na Instituição. Para testar o modelo preditivo, a base foi dividida em de teste e treinamento, e posteriormente foi aplicado a validação cruzada para verificar o poder de generalização do algoritmo.

Conforme é apontado na literatura de acordo com Romero et al. (2010), os primeiros testes não foram eficazes nesse diagnóstico, pois, a base possui um número bastante limitado de registros e muito dos previsores excelentes para ensinar um padrão para o algoritmo, acabavam não sendo selecionados. Na validação cruzada o problema também aconteceu, porém, o algoritmo conseguiu manter um desempenho mais satisfatório em relação ao primeiro teste, mostrando sua capacidade de generalização.

De maneira geral se vê a integração de parte de grandes áreas que estão intrínsecas na composição de uma análise norteada na EDM.

Os principais desafios desse estudo foi dispor de poucos dados históricos que se pudessem fazer inferências do comportamento do discente, como, por exemplo, apontado na literatura (ROMERO et al., 2010), registro *logs* do uso de sistemas educacionais, da plataforma Modular Object-Oriented Dynamic Learning Environment (Moodle). Também foi inviabilizado a extração de traços comportamentais dos discentes em redes sociais por ferir o direito à privacidade e anonimato.

Ainda sim os dados históricos têm menos ruídos que a realização de entrevistas, seja por parte de informações omitidas pelo entrevistado, ou por uma inclinação, ou viés

do entrevistador. Com a mineração dos dados e o modelo preditivo bem sedimentado, é possível atuar em pontos focais evitando casos de evasão, ou seja, em grupos de risco analisados no estudo.

7.1 Trabalhos Futuros

Devido o estudo despontar um grande índice de evasão, é indicado como trabalho futuro a análise nos outros cursos do *Campus*, agrupados por disciplinas dos componentes curriculares em comum a esse curso. Outro ponto que é imprescindível realizar é o teste do modelo preditivo com dados futuros, ou seja, do próximo ano para se mensurar com maior exatidão o poder de generalização desses algoritmos. Como a evasão remete a questões de ordem intimista, é recomendado pesquisas interdisciplinares que explorem e trabalhem na concepção de um modelo preditivo mais amplo.

REFERÊNCIAS

ABEL, F. et al. Recommendations in online discussion forums for e-learning systems. **IEEE transactions on learning technologies**, IEEE, v. 3, n. 2, p. 165–176, 2010. Citado na página 38.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine learning**, Springer, v. 6, n. 1, p. 37–66, 1991. Citado na página 41.

AMARAL, M. G. do. **Mineração de dados aplicada à classificação do risco de evasão de discentes ingressantes em instituições federais de ensino superior**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2016. Citado 2 vezes nas páginas 30 e 31.

AZEVEDO, A. I. R. L.; SANTOS, M. F. KDD, semma and CRISP-DM: A parallel overview. **IADS-DM**, 2008. Citado 4 vezes nas páginas 39, 40, 54 e 55.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 02, p. 03, 2011. Citado na página 37.

BAKER, R. S.; INVENTADO, P. S. Educational data mining and learning analytics. In: **Learning analytics**. Boston: Springer, 2014. p. 61–75. Citado na página 35.

BARDAGI, M.; HUTZ, C. S. Evasão universitária e serviços de apoio ao estudante: uma breve revisão da literatura brasileira. **Psicologia Revista**, v. 14, n. 2, p. 279–301, 2005. Citado na página 21.

BICHSEL, J. **Analytics in higher education: Benefits, barriers, progress, and recommendations**. Louisville: EDUCAUSE Center for Applied Research, 2012. Citado na página 22.

BONNEAU, K. Brief 3: What is a dropout. **North Carolina Education Research Data Center, Center for Child and Family Policy. Retrieved November**, v. 30, p. 2011, 2006. Citado na página 21.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. In: AMERICAN ASSOCIATION FOR ARTIFICIAL INTELLIGENCE. **Advances in knowledge discovery and data mining**. [S.l.], 1996. p. 37–57. Citado na página 40.

BRAGA, A. de P.; FERREIRA, A. C. P. de L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. São Paulo: LTC Editora, 2007. Citado 5 vezes nas páginas 42, 43, 44, 45 e 46.

BRASIL. Lei nº 12.089 de 11 de novembro de 2009. **Diario Oficial da Republica Federativa do Brasil**, nov 2009. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2009/lei/l12089.htm>. Acesso em: 10.08.2018. Citado na página 21.

BREDDELS, M. **IPyvolume**. 2016. Disponível em: <https://ipyvolume.readthedocs.io/en/latest/index.html>. Acesso em: 19.08.2018. Citado na página 51.

BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. **arXiv preprint arXiv:1309.0238**, 2013. Citado na página 51.

CALVO, B.; GUZMÁN, S. scmamp: Statistical comparison of multiple algorithms in multiple problems. **The R Journal, Vol. 8/1, Aug. 2016**, The R Foundation, 2016. Citado na página 48.

CARVALHO, R. S. et al. Integração entre o sistema de gestão acadêmica e o sistema de gestão da aprendizagem: identificando necessidades e prototipando requisitos favoráveis a prática docente. **Revista Brasileira de Computação Aplicada**, v. 4, n. 1, p. 81–91, 2012. Citado na página 65.

CASTRO, L. N. de. **Fundamentals of natural computing: basic concepts, algorithms, and applications**. Boca Ratón: Chapman and Hall/CRC, 2006. Citado na página 42.

CASTRO, L. N. de; FERRARI, D. G. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016. Citado 7 vezes nas páginas 22, 44, 45, 49, 65, 67 e 69.

CODEÇO, F. C. **Computação Científica com Python**. Preópolis: Edição do autor, 2007. Citado na página 49.

COSTA, E. et al. A framework for building web mining applications in the world of blogs: A case study in product sentiment analysis. **Expert Systems with Applications**, Elsevier, v. 39, n. 5, p. 4813–4834, 2012. Citado na página 38.

COSTA, E. M. et al. Eficiência e desempenho no ensino superior: uma análise da fronteira de produção educacional das ifes brasileiras. **Revista de Economia Contemporânea**, scielo, v. 16, n. 3, p. 415–440, 2012. Citado na página 21.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 41.

DEVERT, A. **Matplotlib Plotting Cookbook**. Birmingham: Packt Publishing Ltd, 2014. Citado na página 51.

FÁVERO, L.; FÁVERO, P. **Estatística aplicada: Para cursos de Administração, Contabilidade e Economia com Excel e SPSS**. 1. ed. São Paulo: Elsevier Brasil, 2015. Citado na página 48.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996. Citado 2 vezes nas páginas 40 e 47.

FERREIRA, P. C. **Análise de séries temporais em R**. Rio de Janeiro: Elsevier Brasil, 2018. Citado na página 50.

FONSECA, S. O. da; NAMEN, A. A. Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. **Educação em Revista**, SciELO Brasil, v. 32, n. 1, 2016. Citado na página 46.

FRIEDMAN, J. H. A recursive partitioning decision rule for nonparametric classification. **IEEE Transactions on Computers**, IEEE, n. 4, p. 404–408, 1977. Citado na página 48.

GALVÃO, T. F.; PEREIRA, M. G. Revisões sistemáticas da literatura: passos para sua elaboração. **Epidemiologia e Serviços de Saúde**, SciELO Public Health, v. 23, p. 183–184, 2014. Citado na página 27.

GARCÍA, E. et al. A collaborative educational association rule mining tool. **The Internet and Higher Education**, Elsevier, v. 14, n. 2, p. 77–88, 2011. Citado na página 37.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações**. Rio de Janeiro: Campus, 2005. Citado 2 vezes nas páginas 45 e 46.

GUPTA, A. **Learning Apache Mahout Classification**. Birmingham: Packt Publishing Ltd, 2015. Citado na página 81.

HÁMÁLÁINEN, M. V. W. Classifiers for educational data mining. **Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series**, p. 57–71, 2011. Citado 2 vezes nas páginas 39 e 47.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Waltham: Elsevier, 2011. Citado na página 43.

HEGDE, V.; PRAGEETH, P. Higher education student dropout prediction and analysis through educational data mining. In: IEEE. **2018 2nd International Conference on Inventive Systems and Control (ICISC)**. [S.l.], 2018. p. 694–699. Citado 2 vezes nas páginas 30 e 31.

HUNTER, J.; DALE, D. The matplotlib user's guide. **Matplotlib 0.90. 0 user's guide**, 2007. Citado na página 51.

JOÃO, P. A. de A. **Modelo preditivo da criminalidade–georeferenciação ao concelho de Lisboa**. Tese (Doutorado), 2010. Citado na página 39.

LARRANAGA, P. et al. Machine learning in bioinformatics. **Briefings in bioinformatics**, Oxford University Press, v. 7, n. 1, p. 86–112, 2006. Citado na página 41.

LESTER, J. C.; VICARI, R. M.; PARAGUAÇU, F. (Ed.). **Intelligent Tutoring Systems, 7th International Conference, ITS Proceedings**, v. 3220. Maceió: Springer, 2004. Citado na página 37.

LINTING, M. et al. Nonlinear principal components analysis: introduction and application. **Psychological methods**, American Psychological Association, v. 12, n. 3, p. 336, 2007. Citado na página 65.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos**, n. 25, 2012. Citado 2 vezes nas páginas 21 e 23.

MACHADO, R. D. et al. Estudo bibliométrico em mineração de dados e evasão escolar. In: **XI Congresso Nacional de Excelência em Gestão**. Rio de Janeiro: [s.n.], 2015. Citado na página 23.

MANHÃES, L. M. B. **Predição Do Desempenho Acadêmico De Graduandos Utilizando Mineração De Dados Educacionais**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2015. Citado na página 27.

MÁRQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. **IEEE Revista Iberoamericana de Tecnologias del Aprendizaje**, IEEE, v. 8, n. 1, p. 7–14, 2013. Citado 2 vezes nas páginas 47 e 48.

MARSLAND, S. **Machine learning: an algorithmic perspective**. Boca Ratón: Chapman and Hall/CRC, 2011. Citado 2 vezes nas páginas 46 e 47.

MCKINNEY, W. **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**. California: O'Reilly Media Inc, 2012. Citado 2 vezes nas páginas 50 e 52.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning: An artificial intelligence approach**. New York: Springer Science & Business Media, 2013. Citado 2 vezes nas páginas 41 e 42.

MIRANDA, M. A.; GUZMÁN, J. Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos. **Formación universitaria**, Scielocl, v. 10, n. 3, p. 61–68, 2017. Citado na página 31.

OLIVEIRA JÚNIOR, J. G. d. **Identificação de Padrões para a Análise da Evasão Usando Mineração de Dados Educacionais**. Dissertação (Mestrado), 2015. Citado na página 31.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011. Citado 3 vezes nas páginas 32, 50 e 76.

PEREIRA, R. T.; ZAMBRANO, J. C. Application of decision trees for detection of student dropout profiles. In: **Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on**. [S.l.]: IEEE, 2017. p. 528–531. Citado 2 vezes nas páginas 29 e 31.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013. Citado 3 vezes nas páginas 35, 36 e 53.

ROMERO, C. et al. **Handbook of Educational Data mining**. Boca Ratón: CRC press, 2010. Citado 9 vezes nas páginas 22, 35, 36, 38, 64, 65, 67, 72 e 83.

ROSSUM, G. **Python Reference Manual**. Amsterdam, The Netherlands, The Netherlands, 1995. Citado 2 vezes nas páginas 32 e 49.

RUMELHART, D. E.; MCCLELLAND, J. L. **Parallel distributed processing: explorations in the microstructure of cognition**. Cambridge: MIT Press, 1986. v. 1 foundations. Citado na página 45.

SANTOS, G. D. R. M.; DIAS, V. F.; MOLINA, N. L. **Orientação e dicas práticas para trabalhos acadêmicos**. Curitiba: Ibpex, 2007. Citado na página 35.

SANTOS, L. P. G. dos. Uma contribuição à discussão sobre a avaliação de desempenho das instituições federais de ensino superior: uma abordagem da gestão econômica. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 13, n. 28, p. 86–99, 2002. Citado na página 22.

SARKER, F.; TIROPANIS, T.; DAVIS, H. C. Linked data, data mining and external open data for better prediction of at-risk students. In: IEEE. **2014 International Conference on Control, Decision and Information Technologies (CoDIT)**. [S.l.], 2014. p. 652–657. Citado 2 vezes nas páginas 30 e 31.

SAS. **Company Information**. 2018. Disponível em: <https://www.sas.com/pt_br/company-information/profile.html>. Acesso em: 10.07.2018. Citado na página 39.

SCHEUER, O.; MCLAREN, B. M. Educational data mining. In: **Encyclopedia of the Sciences of Learning**. New York: Springer, 2012. p. 1075–1079. Citado 2 vezes nas páginas 35 e 65.

SEBESTA, R. W. **Conceitos de Linguagens de Programação**. 9. ed. Porto Alegre: Bookman Editora, 2011. Citado 2 vezes nas páginas 32 e 49.

SIEMENS, G.; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE Review**, ERIC, v. 46, n. 5, p. 30, 2011. Citado na página 22.

SIMON, H. **Redes Neurais–Princípios e Prática**. Porto Alegre: Bookman, 2001. Citado na página 42.

SISTEMA DE SELEÇÃO UNIFICADA. **SISU - Tire suas dúvidas**. 2019. Disponível em: <http://www.sisu.mec.gov.br/tire-suas-duvidas>. Acesso em: 10.05.2019. Citado na página 55.

STOJANOVA, D. et al. Estimating vegetation height and canopy cover from remotely sensed data with machine learning. **Ecological Informatics**, Elsevier, v. 5, n. 4, p. 256–266, 2010. Citado na página 48.

TAUSWORTHE, R. C. The work breakdown structure in software project management. **Journal of Systems and Software**, v. 1, p. 181 – 186, 1979. Disponível em: <http://www.sciencedirect.com/science/article/pii/0164121279900189>. Citado na página 23.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. **Review of educational research**, v. 45, n. 1, p. 89–125, 1975. Citado 2 vezes nas páginas 21 e 56.

TINTO, V. **Leaving college: Rethinking the causes and cures of student attrition**. Chicago: ERIC, 1987. Citado na página 21.

VANDERPLAS, J. **Python data science handbook: essential tools for working with data**. Sebastopol: O'Reilly Media, Inc, 2016. Citado na página 50.

VIEIRA, S. **Bioestatística-Tópicos avançados, Testes não-paramétricos, Tabelas de contingência e análise de regressão**. Rio de Janeiro: Campus, 2004. 212 p. Citado na página 48.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. Rio de Janeiro: Elsevier Brasil, 2017. Citado na página 23.

WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. Burlington: Morgan Kaufmann, 2016. Citado 2 vezes nas páginas 47 e 48.

ÍNDICE