UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Data Mining e *Data Analytics* para Apoio à Gestão Estratégica e Mitigação da Evasão Escolar

Dissertação de Mestrado

Nathanael Oliveira Vasconcelos

São Cristóvão – Sergipe

2019

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Nathanael Oliveira Vasconcelos

Data Mining e *Data Analytics* para Apoio à Gestão Estratégica e Mitigação da Evasão Escolar

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Dr. Methanias Colaço Rodrigues Júnior

São Cristóvão – Sergipe

2019

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: Nathanael Oliveira Vasconcelos

Em 22 dias do mês de agosto do ano de dois mil e dezenove, com início às 14h00min, realizou-se na Sala de Seminário do DCOMP da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Nathanael Oliveira Vasconcelos**, que desenvolveu o trabalho intitulado: ***" Data Mining e Data Analytics para Apoio à Gestão Estratégica e Mitigação da Evasão Escolar"***, sob a orientação do Prof. Dr. **METHANIAS COLACO RODRIGUES JUNIOR**. A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **André Britto de Carvalho** (PROCC/UFS) e, em seguida, a Profª. **Sônia Virginia Alves França** (UFRPE). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) APROVADO *"(aprovado/reprovado)"*. Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 25/2014/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária *"Prof. José Aloísio de Campos"*, 22 de agosto de 2019.

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Prof. Dr. André Britto de carvalho
(PROCC/UFS)
Examinador Interno

Profª. Drª. Sônia Virginia Alves França
(instituição)
Examinador Externo

Nathanael Oliveira Vasconcelos
Candidato

Dedico esta dissertação a toda a minha família, amigos e professores que me deram o apoio necessário para chegar até aqui.

Agradecimentos

Enfim, mais uma meta concluída. Apesar dos desafios enfrentados (conciliar trabalho, estudar, dar atenção a família e amigos), o mestrado foi bastante prazeroso.

Este trabalho não teria sido possível sem a orientação, o apoio e o encorajamento de várias pessoas que me acompanham nesta jornada.

Primeiramente, gostaria de agradecer ao meu orientador, Prof. Dr. Methanias Colaço Rodrigues Júnior, pela confiança, paciência, por todo o aprendizado durante a condução da orientação deste mestrado, e sobretudo, pela amizade e parceria, as quais espero que perdurem por vários anos.

Aos meus pais, que sempre estiveram do meu lado, dando-me apoio, conselhos, amor e por serem meu maior exemplo de vida.

Aos meus irmãos e cunhados pelo apoio, conselhos e companheirismo de todas as horas.

A todos os professores que tive durante toda a jornada da vida, que apesar das dificuldades, sempre incentivaram e deram o melhor para transmitir os conhecimentos.

Aos amigos que fiz na Superintendência de Tecnologia de Informação da UFS e na Stefanini.

A todos amigos que estiveram presente durante essa jornada, pela força, carinho, amizade e compreensão nos momentos de "ausência".

E, sobretudo, agradeço a Deus, que sempre iluminou meu caminho, me fez persistir e lutar pelos meus sonhos.

Resumo

Contexto: A evasão é, certamente, um dos grandes problemas que afligem as instituições de ensino em geral, uma vez que as perdas ocasionadas pelo abandono do aluno são desperdícios sociais, acadêmicos e econômicos. A busca de suas causas tem sido objeto de muitos trabalhos e pesquisas educacionais em todo mundo. No campo prático, diversas organizações de ensino norteiam as suas decisões estratégicas para o controle da taxa de evasão, entretanto, no Brasil, ainda são poucos os trabalhos publicados nesta área de pesquisa. Como consequência, fica evidente a necessidade de aumentar a compreensão do problema e de suas causas, com a adoção de medidas mais eficazes para identificar e entender os principais fatores que podem contribuir com o insucesso dos estudantes. **Objetivo:** Este trabalho teve por propósito fazer duas análises experimentais dos algoritmos de mineração de dados mais utilizados na área de educação, avaliando o que melhor se adequa ao contexto de abandono do ensino em duas instituições federais, bem como implementar um método de uso do melhor modelo, o qual auxiliará o processo de apoio à decisão e à mitigação da evasão escolar. **Método:** Foram planejados e executados dois experimentos controlados "in vivo", para comparar a eficácia dos classificadores selecionados. Em seguida, foi realizado um estudo de caso com interface criada para aplicar o algoritmo que obteve a melhor eficácia. **Resultados:** Os resultados evidenciaram que existem diferenças significativas entre os algoritmos utilizados, e que, apesar do SVM possuir a maior média das métricas de eficácia, estatisticamente, após a metanálise dos experimentos, os algoritmos MLP e *Random Forest*, respectivamente, obtiveram resultados semelhantes de acurácia (85,38%, 84,40% e 84,13%). Para medida-F, a significância estatística foi igual apenas para o MLP (84,42%, 83,44%). **Conclusões:** Esta dissertação expôs a necessidade de aumentar a adoção de medidas para identificar e entender os principais fatores que podem contribuir com o insucesso dos estudantes. Após duas análises experimentais, foi evidenciado que existem diferenças significativas entre os três primeiros colocados e os demais algoritmos avaliados, sendo o SVM, pelo seu pequeno destaque, selecionado para ser aplicado em um estudo de caso de atendimento a um dos itens do planejamento estratégico da Universidade Federal de Sergipe – UFS.

Palavras-chave: Evasão; Mineração de Dados Educacionais, Algoritmos de Classificação; Experimentação.

Abstract

Context: Dropout is certainly one of the major problems that plague educational institutions in general, since, as a result of student dropout, they are social, academic and economic assets. A search for their actions has been the subject of much educational work and research around the world. In the practical field, the various American high school associations had their guidelines focused on the dropout rate control, however, in Brazil, there is still little work done in this area of research. As a result, there is the ability to increase understanding of the problem and its causes by adopting more effective measures to identify and understand the key factors that may contribute to student failure. **Objective:** This work had the purpose of conducting experiments of more advanced data mining algorithms in the area of education, aiming to improve the educational context of the high school dropout of two federal institutions, as well as to implement a method of using the best model, which supports the decision support process and the school dropout mitigation. **Method:** Two in vivo controlled experiments were developed and performed to compare the selected classifiers. Then, a case study with interface created to apply the algorithm that obtained the best result was performed. **Results:** The results showed the differences between the algorithms used and, despite the SVM, had a higher average of the accumulation metrics, statistically, after a meta-analysis of the experiments, the algorithms MLP and Random forest, respectively, obtained similar accuracy results (85.38 %, 84.40 % and 84.13 %). For F-measure, a statistical significance was equal only for the MLP (84.42 %, 83.44 %). **Conclusions:** This dissertation exposes the need to increase the admission of measures to identify and understand the main factors that may contribute to students' failure. The two experimental tests were examined and introduced by the success criteria, being the SVM, the selected differential, selected to be applied in a case study of one of the planning items. Federal University of Sergipe - UFS.

Keywords: School Dropout; Educational Data Mining, Classification Algorithms; Experimentation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	Application Programming Interface
BI	Business Intelligence
DM	Data Mart
DW	Data Warehouse
DM	Data Mining
DS	Data Sources
EDM	Educational Data Mining
ETL	Extract, Trasnform and Load
IFS	Instituto Federal de Sergipe
MLP	Multi Layer Perceptron
OLAP	Online Analytical Processing
PEN	Plano Estratégico de Negócio
PETI	Plano Estratégico de Tecnologia de Informação
REUNI	Reestruturação e Expansão das Universidades Federais
RF	Random Forest
SVM	Support Vector Machines
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SPSS	Statistical Package for Social Science
TI	Tecnologia da Informação
TAM	Termo de Acordo de Metas e Compromissos
UFS	Universidade Federal de Sergipe

Sumário

1 Introdução

De acordo com o Censo da Educação Superior 2015, dos alunos que entraram na graduação em 2010, 11% desistiram já no primeiro ano. Até 2014, quase metade (49%) dos estudantes saíram dos cursos que haviam optado em 2010 (TEIXEIRA, 2015). Sendo assim, a evasão é, certamente, um dos grandes problemas que afligem as instituições de ensino em geral.

Segundo Filho et al. (2007), diversas situações podem ser caracterizadas como evasão, tais como: o trancamento de um curso por um estudante, a desistência por falta de interesse, a falta de recursos financeiros do aluno, motivo de doença, gravidez precoce, ou até mesmo a desistência devido à incompatibilidade de horários das aulas com o mercado de trabalho ou ainda quando os estudantes dão início a carreira profissional.

Diversas organizações governamentais como o Programa REUNI - Reestruturação e Expansão das Universidades Federais - e o TAM - Termo de Acordo de Metas e Compromissos - buscam por decisões estratégicas quem visam o controle da taxa de evasão (FILHO et al., 2007). As perdas ocasionadas pelo abandono do aluno são desperdícios sociais, acadêmicos e econômicos. No setor público, recursos são investidos sem o devido retorno, em paralelo, no setor privado, as taxas de evasão representam uma importante perda de receita. Como consequência, fica evidenciada a necessidade de aumentar a compreensão do problema e de suas causas, com a adoção de medidas mais eficazes para identificar e entender os principais fatores que podem contribuir com o insucesso dos estudantes.

Nessa direção, o Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica sugere que cada instituição da Rede Federal elabore e desenvolva um Plano Estratégico de Intervenção e Monitoramento para Superação da Evasão e Retenção (MEC, 2014).

Diante do exposto, para suportar as decisões dos gestores, as organizações buscam mecanismos que viabilizem informações sumarizadas e interligadas. Estas informações podem ser oriundas dos diversos sistemas transacionais existentes, bem como podem ser extraídas de

um único sistema transacional, limitadas ao escopo do mesmo (JUNIOR, 2004). Independente da origem, torna-se essencial ter informações gerenciais relevantes que retratem as necessidades reais das organizações, para o atingimento de suas metas e objetivos.

Com o crescimento de ambientes virtuais de aprendizagem e tecnologias educacionais de apoio ao ensino, um volume expressivo de dados são gerados das interações entre alunos e professores, bem como do comportamento dos estudantes no processo de ensino e aprendizagem.

Como proposta para tratar estes volumes de dados, uma ferramenta que vem sendo empregada nas instituições é o *Business Intelligence* (BI), que segundo Khan e Quadri (2012), pode ser apresentado como uma arquitetura, uma ferramenta, uma tecnologia ou um sistema que coleta e armazena dados, analisa-os utilizando ferramentas analíticas, propicia a criação de relatórios e consultas, e entrega informação ou o conhecimento com a finalidade de melhorar a tomada de decisão das organizações. Para Hans e Mnkandla (2013), a utilização de BI pode viabilizar uma melhoria para os processos de tomada de decisão.

Todavia, a implantação de um sistema de BI nas organizações, muitas vezes, não consegue influenciar a tomada de decisão. Segundo Olszak e Ziemba (2012), muitos projetos de BI falham ou não são entregues em sua totalidade. Eliminando as questões políticas e micropolíticas de interesses, muito comum em instituições públicas brasileiras, a principal razão para o fracasso é o conhecimento limitado das organizações sobre as oportunidades geradas e os seus benefícios.

As principais ferramentas de BI incluem técnicas de mineração de dados (*Data Mining* – DM), as quais, tem como objetivo descobrir informações ocultas, por meio da análise de grandes quantidades de dados (WITTEN et al., 2016). O termo "ocultas" refere-se ao processo de identificar relações entre dados que podem produzir novos conhecimentos e gerar novas descobertas científicas (BAKER; ISOTANI; CARVALHO, 2011). Para Turban e Volonimo (2013), busca extrair informação e conhecimento, por meio de relações entre os dados, que permitam inferências sobre o que pode ocorrer (análise preditiva) ou correlações entre o que já ocorreu.

No âmbito da educação, a mineração de dados é conhecida como *Educational Data Mining* (EDM) (BAKER et al., 2010). EDM é definida pelo site da comunidade [1], como uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar os tipos únicos de dados que vêm do ambiente educacional e usar esses métodos para entender melhor os alunos e as características de seu aprendizado.

Neste contexto, para Costa et al. (2012), a mineração de dados educacionais surge como uma ferramenta para auxiliar as instituições de ensino a entenderem o perfil de seus estudantes possibilitando a personalização das interações pedagógicas visando um processo de aprendizagem mais adequado a cada estudante e como consequência a diminuição dos índices de evasão dos discentes, tanto na educação a distância como na presencial.

[1] www.educationaldatamining.org

Sendo assim, é possível minerar dados de alunos com o objetivo de identificar relações entre os diversos fatores que os levam a abandonar os cursos e identificar inferências sobre o que pode ocorrer. Neste trabalho, é proposta uma análise comparativa de algoritmos de mineração de dados, avaliando o que melhor se adequa ao contexto de abandono do ensino em duas instituições federais - Universidade Federal de Sergipe (UFS) e Instituto Federal de Sergipe (IFS), bem como a implementação de um método de uso do melhor modelo, visando auxiliar o processo de apoio à decisão estratégica.

1.1 Problemática e Hipótese

Presente no sistema educacional brasileiro nos seus diversos níveis e modalidades, a evasão escolar tem atingido desde a educação básica até a superior, gerando prejuízos sociais, econômicos, políticos, acadêmicos e financeiros a todos os envolvidos no processo educacional, desde o estudante até os órgãos governamentais. Diagnosticar suas causas, compreender como esse processo ocorre nas instituições de ensino e conhecer a visão dos gestores educacionais sobre esta problemática pode auxiliar na compreensão deste fenômeno social e nas ações preventivas para sua redução (SANTOS, 2017).

Segundo Lobo (2012), diminuir a evasão escolar custa seis vezes menos do que trazer um novo estudante até a instituição de ensino. Combater a evasão é o modo mais eficiente de aumentar o número de matrículas e, ainda, mostrar a efetividade do processo, tendo em vista que fazer o aluno concluir seu curso, com qualidade, significa que a instituição atingiu seu objetivo.

Por outro lado, a falta de informação, informações imprecisas ou a incapacidade de reconhecer o valor da informação que lhe foi entregue ou está disponível é a principal causa de decisões equivocadas. Frota (2009), afirma que existem várias explicações para essa falta de informação, destacando-se a ausência de integração entre dados de diferentes sistemas, além da inexistência de um ambiente propício e ferramentas adequadas para o tratamento dos dados.

Para Silva, Silva e Gome (2016), a dificuldade de gerir dados e informações que sejam analisadas e interpretadas de forma adequada e eficaz, objetivando auxiliar um processo de tomada de decisão, é comum nos ambientes organizacionais. Muitas empresas obtem diversos dados sobre o seu negócio e mercado que está inserido, no entanto, não consegue transformar em informações relevantes e estratégicas para melhores decisões.

O *survey* realizado por Lima, Júnior e Nascimento (2017), com empresas brasileiras, evidenciou que 72% das empresas entrevistadas não utilizavam um método específico para o desenvolvimento de aplicações alinhado ao planejamento estratégico da organização. A ausência de uma metodologia alinhada à estratégia da empresa sugere que os gestores podem estar tomando decisões com base em informações não relevantes à instituição ou desalinhadas às estratégias de negócio.

Os sistemas são utilizados pelas empresas para agilizar o processo de tomada de decisão, disponibilizando informações oportunas e em tempo hábil aos tomadores de decisão. Portanto, torna-se fundamental a utilização de sistemas aliados aos objetivos estratégicos da empresa para auxiliar a tomada de decisão gerencial. Dessa forma, precisa-se reconhecer que os sistemas de informações são necessários para todas as organizações e que as informações geradas suportam o ambiente para a tomada de decisão dos gestores (FERNANDES et al., 2012).

Essas evidências endossam que adotar uma abordagem que seja orientada aos objetivos organizacionais, visando auxiliar o processo de apoio à decisão estratégica para a mitigação da evasão escolar, pode ser promissor. Recentemente, um número razoável de pesquisas foram conduzidas para aplicar técnicas de mineração de dados na área de educação, ordenadas para classificar e prever o desempenho dos alunos em vários institutos de educação (MáRQUEZ-VERA et al., 2016) (GOPALAKRISHNAN et al., 2017) (MANHÃES et al., 2012) (KANTORSKI et al., 2016) (PASCOAL et al., 2016).

Estes trabalhos serão explanados na Seção 2.5, vale a pena mencionar que não foram encontrados trabalhos que realizaram uma análise comparativa de algoritmos aplicados ao contexto da evasão escolar, considerando uma abordagem experimental, com a validação estatística da significância dos dados, como é proposto neste trabalho.

Sendo assim, este trabalho vislumbra a condução de um processo experimental, seguindo as diretrizes de Wohlin (2012), para também efetuar um comparativo entre os resultados encontrados e as evidências disponíveis na literatura.

Para guiar o estudo, foi elaborada a seguinte questão de pesquisa, cuja resposta visa cumprir um dos objetivos deste trabalho. No contexto da Evasão Escolar de duas instituições de ensino superior, entre os algoritmos selecionados na área de EDM, qual o melhor em termos de Eficácia?

Para avaliar esta questão, foram utilizadas quatro métricas: Acurácia (2.1), *Recall* (2.2), Precisão (2.3) e Medida-F (2.4). Sendo assim, com os objetivos e métricas definidas, será considerada a hipótese a seguir (para cada métrica):

- H_0: Os algoritmos$_{(1,2..n)}$ possuem médias iguais para a métrica.
  μ1(metrica) = μ2(metrica) ... = μn(metrica);

  Hipótese Alternativa:

- H_1: Os algoritmos$_{(1,2..n)}$ possuem médias diferentes para a métrica.
  μ1(metrica) $\neq$ μ2(metrica) ... $\neq$ μn(metrica);

1.2 Justificativa

Uma pesquisa realizada em 2016 pela Deloitte, com 1.200 executivos de TI (CIO), afirma que 78% das empresas entrevistadas veem o alinhamento estratégico de atividades de TI com a estratégia de negócios como fundamental para o sucesso de uma organização (DELOITTE, 2017). A transformação digital da maioria dos setores de uma organização requer a integração eficaz e eficiente dos recursos de software, em todos os tipos de produtos e serviços, como um pré-requisito para a construção de modelos de negócios bem-sucedidos. Como consequência, os gestores das organizações precisam entender como usar o software e como medir (mensurar) as atividades deste software, aliadas aos objetivos de alto nível da organização, tais como metas de negócios (MÜNCH, 2013).

Neste sentido, segundo o Plano de Desenvolvimento Institucional da Universidade Federal de Sergipe (UFS) 2016-2020 - documento que expressa um diagnóstico da Universidade, visão, políticas acadêmicas e administrativas, fundamentadas em sua realidade institucional, estabelecendo objetivos e metas estratégicas para o período de 2016 a 2020 - a taxa de evasão nos cursos de graduação em 2014 foi de 13,4%, no mesmo documento, é explicitado que a meta da instituição é reduzir a taxa para 3% até 2020.

Assim, reforça-se a necessidade de implementação de meios que ajudem na superação desse fenômeno, de modo a possibilitar a realização de diagnósticos apurados em relação às causas da evasão, e que ajudem na definição de políticas institucionais e a adoção de ações administrativas e pedagógicas que contribuam para a mitigação da evasão escolar.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo deste trabalho é fazer duas análises experimentais de algoritmos de mineração de dados, avaliando o que melhor se adequa ao contexto de abandono do ensino em duas instituições federais, bem como implementar um método de uso do melhor modelo, o qual auxiliará o processo de apoio à decisão e à mitigação da evasão escolar.

1.3.2 Objetivos Específicos

Para possibilitar a realização do objetivo geral, podemos enumerar os seguintes objetivos específicos:

- Realizar um mapeamento sistemático da literatura, com a finalidade de identificar os principais algoritmos de classificação encontrados nos trabalhos de EDM;

- Análise experimental de algoritmos de mineração de dados no Instituto Federal de Sergipe - IFS.
- Análise experimental de algoritmos de mineração de dados na Universidade Federal de Sergipe - UFS.
- Realizar um Estudo de Caso na UFS para atender ao planejamento estratégico, no que diz respeito à evasão escolar, com o propósito de aplicar o modelo selecionado.

1.4 Metodologia

O método utilizado para o trabalho em questão consiste, em termos de classificação, de uma pesquisa exploratória (SEVERINO, 2017), pois foi realizado um mapeamento sistemático, no qual foi definido os atributos utilizados para a geração dos arquivos utilizados pela ferramenta de mineração e bibliotecas de *Machine Learning*. A seleção de atributos foi realizada após análise dos trabalhos relacionados, como, por exemplo, o trabalho (MARQUEZ-VERA; MORALES; SOTO, 2013) (JÚNIOR et al., 2015), bem como considerando os atributos disponíveis na base de dados.

Além disso, foi utilizado algoritmos de seleção de atributos (*CfsSubsetEval* e *Random Forest*), os quais avaliam o valor preditivo de cada atributo individualmente, gerando um ranking em que aqueles atributos que possuírem maior relação com a classe e menor correlação com os demais atributos recebem maior pontuação (SILVA; BATISTA, 2009).

Após a definição dos atributos, foram gerados modelos de mineração, com o objetivo de realizar uma comparação entre os algoritmos utilizados. Foram escolhidos os principais classificadores encontrados nos trabalhos de EDM (JÚNIOR et al., 2015)(MARQUEZ-VERA; MORALES; SOTO, 2013): árvore de decisão (BREIMAN et al., 1984), Naive Bayes (JOHN; LANGLEY, 1995), vizinho mais próximo (CAMPOS et al., 2016), redes neurais (MLP) (NüRNBERGER; PEDRYCZ; KRUSE, 2002), máquina de vetores de suporte (SVM) (CORTES; VAPNIK, 1995) e *ensemble methods (Random Forest)*(BREIMAN, 2001a).

Para realização do objetivo principal desta pesquisa e consequente coleta de dados, serão executados dois experimentos controlados, *in vivo*, envolvendo as bases de dados de duas instituições de nível superior. De acordo com Wohlin et al. (2012), uma experimentação não é uma tarefa simples, pois envolve preparar, conduzir e analisar dados corretamente. Os autores destacam como uma das principais vantagens da experimentação o controle dos sujeitos, objetos e instrumentação, o que torna possível extrair conclusões mais gerais sobre o assunto investigado. Outras vantagens incluem a habilidade de realizar análises estatísticas, utilizando métodos de teste de hipóteses e oportunidades para replicação. Juristo e Moreno (2001) também afirmam que a pesquisa científica não pode ser baseada em opiniões ou interesses comerciais. Investigações científicas são representadas por estudos baseados em observação e/ou experimentação acerca

do mundo real e seus comportamentos mensuráveis, como nesta pesquisa. Estes aspectos devem ser levados em consideração também na construção e avaliação de algoritmos e softwares.

Na execução dos experimento, com a definição dos algoritmos e atributos, as bases de treinamentos foram submetida à ferramenta livre e open-source de mineração de dados *Weka* e utilizando Python e suas bibliotecas, com isto, foram gerados modelos de conhecimento com o objetivo de realizar testes dos algoritmos e comparar a eficácia. Neste contexto, este trabalho também foi classificado como de laboratório e experimental, devido ao planejamento e à execução de um experimento controlado.

Para auxiliar nos cálculos e verificar se existiam diferenças significativas na eficácia dos algoritmos, foi utilizada a ferramenta para análise de dados *Statistical Package for Social Science* - SPSS (IBM, 2017). O SPSS é um software estatístico internacionalmente utilizado há muitas décadas, desde suas versões para computadores de grande porte (MUNDSTOCK, 2006).

Sumarizando, os experimentos podem ser divididos em três etapas principais: (1) planejamento e execução dos experimentos no Instituto Federal de Sergipe (IFS) e na Universidade Federal de Sergipe (UFS); (2) análise quantitativa do comportamento dos algoritmos selecionados; e (3) implementar um método de uso do melhor modelo, o qual auxiliará o processo de apoio à decisão e à mitigação da evasão escolar.

1.5 Organização da Dissertação

Este documento está organizado da seguinte maneira. Este capítulo discutiu a motivação principal para o trabalho proposto, além de descrever de forma geral como ele foi desenvolvido. O Capítulo 2 apresenta os conceitos e termos necessários para o entendimento do trabalho. Os Capítulos 3 e 4 apresentam as avaliações experimentais dos principais algoritmos utilizados em trabalhos de EDM. O Capítulo 5 apresenta um estudo de caso com interface criada para aplicar o algoritmo que obteve a melhor eficácia. Por fim, no Capítulo 6, discutem-se as conclusões, contribuições e dificuldades encontradas, bem como são apresentados os possíveis trabalhos futuros relacionados com esta pesquisa.

2

Fundamentação Teórica

2.1 Alinhamento Estratégico

Na literatura, existem muitos conceitos que detalham o significado de alinhamento estratégico, no entanto, serão apresentados aqueles que possuem maior vínculo com o assunto desta proposta. (1) O alinhamento entre o plano estratégico de negócio (PEN) e o plano estratégico de tecnologia de informação (PETI) é alcançado quando o conjunto de estratégias de sistemas (objetivos, obrigações e estratégias) é derivado do conjunto estratégico organizacional (missão, objetivos e estratégias) (KING, 1988); (2) O elo entre PEN-PETI corresponde ao grau no qual a missão, os objetivos e os planos de TI refletem, suportam e são suportados pela missão, pelos objetivos e pelos planos de negócio (REICH B. H. E BENBASAT, 1996); (3) É a forma como os negócios e a TI trabalham em conjunto para alcançar o objetivo comum (CAMPBELL, 2005) e (4) O alinhamento entre PEN-PETI é a adequação da orientação estratégica do negócio com a de TI (CHAN, 1997).

Para Lima (2017), o problema das organizações, atualmente, é que nem sempre as empresas conseguem declarar os objetivos estratégicos de forma explícita ou suficientemente clara, para que se possa verificar se tais objetivos têm realmente alcançado as metas e estão alinhados à TI. Este desafio não é ter a área de TI como um suporte, mas sim como parte de uma plataforma de negócio, servindo como elemento essencial à estratégia de negócio. Essa nova visão deve vincular o alinhamento estratégico da TI ao negócio da organização. Estes dois elementos precisam relacionar-se entre si, em busca da melhoria contínua e do sucesso da organização.

2.2 Alinhamento estratégico e sua relação com Tecnologia da Informação - TI

Como introduzido na Seção 2.1, alinhamento estratégico, refere-se à harmonia entre a estratégia de uma empresa e o ambiente externo (CHI-AN; CHANDRA, 2012). O alinhamento estratégico implica construção dos sistemas de informação da organização, que devem estar alinhados ao desenvolvimento das estratégias de negócios e das metas organizacionais. Para Velcu (2010), o alinhamento estratégico é a diferenciação entre os conceitos de estratégia planejada e realizada, enfatizando a importância de analisar não só o que planejam as empresas, mas também o que se tem de resultado.

Fica claro o papel fundamental da tecnologia da informação nos processos estratégicos, com avanços que têm redefinido como as empresas os realizarão, permitindo-lhes reduzir significativamente seus custos e melhorar desempenho por meio da automação (MARCH; HEVNER, 2007). As aplicações de *Business Intelligence* são uns desses avanços, sendo classificada como um dos principais temas no meio executivo ao longo dos últimos anos (ISIK et al., 2013).

O uso de BI pelas organizações é visto como uma previsão de acontecimentos, por meio das tendências apontadas pelas informações existentes. Algumas organizações fracassam na manutenção da qualidade dos dados advindos dos sistemas transacionais, que alimentam o sistema de BI, levando-as a uma situação em que não é possível perceber que suas decisões estão sendo prejudicadas por dados errados (ISIK et al., 2013).

Decidir o futuro de uma organização com base em informações equivocadas poderá levar à extinção ou ao fracasso da empresa. Definir quais caminhos deve ser seguido exige conhecimento e experiência das partes envolvidas, a fim de identificar quais os principais elementos são valiosos para o sucesso da empresa. Isto exige um planejamento de ações para garantir que quaisquer mudanças nos objetivos estratégicos possam refletir nas aplicações de tomada de decisão.

Essa vinculação é importante, de modo que todo esforço adotado no desenvolvimento da aplicação possa contribuir com os objetivos estratégicos (LIMA, 2017).

2.3 *Data Mining* e *Data Analytics*

A mineração de dados ou *data mining* (DM) é o processo de análise de dados para extrair informações e conhecimentos que não são visualizados claramente em sistemas comuns. Busca extrair informação e conhecimento, por meio de relações entre os dados, que permitam inferências sobre o que pode ocorrer (análise preditiva) ou correlações entre o que já ocorreu (BARBIERI, 2011).

Para Sharda, Delen e Turban (2014), *data mining* é um termo utilizado para descrever

a descoberta (ou mineração) de conhecimento a partir de grandes quantidades de dados. Tecnicamente falando, consiste em um processo que utiliza técnicas de estatística, matemática e inteligência artificial para extrair e identificar informações úteis e consequentemente, conhecimento (ou padrões) a partir de grandes volumes de dados, sendo que estes padrões podem ser apresentados como tendências, regras de negócio, correlações ou modelos preditivos. Esses modelos e padrões podem ser utilizados para guiar o processo decisório, bem como prever o efeito dessas escolhas (LAUDON; LAUDON, 2014).

A quantidade de dados disponíveis cresce a cada dia e desafia a capacidade de armazenamento, seleção e uso. Essas tecnologias permitem a "mineração" desses dados, a fim de gerar um real valor do dado, transformando-o em informação e na maioria das vezes, em conhecimento (REZENDE; ABREU, 2003). Neste viés, os mesmos autores afirmam que o *Data Mining* (DM) é uma tecnologia que tem a capacidade de fazer a seleção de dados relevantes, no intuito de gerar informações e conhecimentos para as organizações. Este tem a capacidade "de aprender com base nos dados, extrair deduções, gerar informações com hipóteses, correlacionar coisas aparentemente desvinculadas, fazer previsões, revelar os atributos importantes, gerar cenários, relatar e descobrir conhecimentos" importantes para a gestão empresarial.

Data analytics consiste basicamente na aplicação de tratamento estatístico em dados coletados, com o intuito de gerar previsões e insights dando sentido a esses dados, transformando-os em informação que ajudam na tomada de decisões e planejamento estratégico das empresas (CHEN; CHIANG; STOREY, 2012). Davenport e Patil (2012) definem *Analytics* como o uso extensivo de dados, modelos estatísticos e análises quantitativas, modelos de análise preditiva e exploratória, além de decisões e ações gerenciais baseadas em fatos.

Para Harriott (2013), *Analytics* se relaciona com o "como", ou seja, como desenvolver e prover a informação de forma a realmente resolver os desafios da organização. Para isso, o autor alerta que o mais importante é sempre focar nos objetivos finais dos usuários da informação e prosseguir com a análise com o intuito de atingi-los.

Em mineração de dados temos as tarefas e as técnicas que norteiam o que queremos e como obtermos tais informações. As tarefas apontam quais os objetivos, ou seja, quais as informações pretende-se obter. As técnicas de mineração apontam para os métodos empregados para se obter padrões desejados (CAMILO; SILVA, 2009). Duan e Xu (2012) classificam as técnicas de mineração em dois abrangentes grupos: aprendizado supervisionado e aprendizado não supervisionado. Os métodos de aprendizado supervisionado constroem modelos para predizer um atributo não conhecido de acordo com atributos observados, enquanto os métodos de aprendizado não supervisionados extraem padrões, como agrupamentos, gráficos de processo e correlações entre os dados.

Os métodos supervisionados necessitam de um conjunto de dados que possuem uma variável alvo principal pré-definida e os registros são categorizados em relação a ela, ou seja aplica-se a regra programada com precisão. Nos métodos não-supervisionados não há a necessi-

dade de uma pré-categorização para os registros, isso significa que não é preciso um atributo como alvo principal (FILHO et al., 2007).

As técnicas mais comuns de aprendizado supervisionado são a classificação (que também pode ser não-supervisionado) e a regressão (MCCUE, 2007). Porém, em mineração de dados permite-se que várias técnicas sejam aplicadas ou até mesmo que haja uma combinação entre elas visando a busca de melhor resultado esperado (COSTA et al., 2012). Neste trabalho, foram escolhidos os principais classificadores encontrados nos trabalhos de EDM: árvore de decisão, Naive Bayes, vizinho mais próximo, redes neurais, máquina de vetores de suporte (SVM) e *ensemble methods (Random Forest)*.

2.3.1 Árvores de Decisão (*Decision Trees*)

Consiste em um conjunto de condições, organizados em uma estrutura hierárquica. É um modelo preditivo em que um exemplo é classificado, seguindo o caminho de condições satisfeitas, a partir da raiz da árvore até atingir uma folha, que vai corresponder a um rótulo de classe. As árvores de decisão são consideradas modelos de fácil compreensão, porque um processo de raciocínio pode ser dado para cada conclusão, exceto se a árvore obtida é muito grande (uma série de nós e folhas) (ROMERO et al., 2008).

2.3.2 Redes Neurais (*Neural Networks*)

É uma técnica que tem origem na psicologia e na neurobiologia. Consiste basicamente em simular o comportamento dos neurônios. De maneira geral, uma rede neural pode ser vista como um conjunto de unidades de entrada e saída conectados por camadas intermediárias e cada ligação possui um peso associado. Durante o processo de aprendizado, a rede ajusta estes pesos para conseguir classificar corretamente um objeto. Essa técnica que necessita de um longo período de treinamento, ajustes dos parâmetros e é de difícil interpretação, não sendo possível identificar de forma clara a relação entre a entrada e a saída. Em contrapartida, as redes neurais conseguem trabalhar de forma que não sofram com valores errados e também podem identificar padrões para os quais nunca foram treinados (CAMILO; SILVA, 2009).

2.3.3 *Random Forest*

O algoritmo de *Random Forest* (RF) é um termo geral para métodos de *ensemble*, os quais caracterizam-se pela geração de muitos classificadores e combinação de seus resultados utilizando classificadores do tipo árvore (DIETTERICH, 2000). O RF foi introduzido por Breiman (2001b), consistindo em construir uma grande quantidade de árvores de decisão para fora do sub-conjunto de dados a partir de um treinamento único definido. Tal treinamento é realizado usando *bagging* (um meta-algoritmo para melhorar a classificação e a regressão de modelos de acordo com a estabilidade e a precisão da classificação).

Este procedimento extrai casos aleatoriamente a partir de conjuntos de dados de treinamento originais e os conjuntos são usados para construir cada uma das árvores de decisão que compõe a RF. Cada árvore classificadora é apontada como um componente preditor. A RF constrói sua decisão por meio da contagem dos votos dos componentes preditores em cada classe e, em seguida, seleciona a classe vencedora em termos de número de votos acumulados (CAMILO; SILVA, 2009).

2.3.4 Máquina de Vetores de Suporte (*Support Vector Machines* - SVM)

É um método definido inicialmente para dados com separação linear. O objetivo é encontrar o hiperplano de maior margem que separa as classes. As SVMs têm a restrição de os dados devem serem numéricos contínuos (ou quantificados). O modelo não é facilmente interpretável, e a seleção dos parâmetros adequados pode ser difícil (HAMALAINEN W., 2011).

2.3.5 Naive Bayes

Naive Bayes é um algoritmo de classificação ingênuo probabilístico, baseado na aplicação do teorema de Bayes para determinar a classe de maior probabilidade para cada nova instância a ser classificada (JOHN; LANGLEY, 1995).

Para classificar uma nova instância, o algoritmo determina a classe mais provável, dados os atributos (a_1,a_2,...,a_M) que descrevem a instância (MITCHELL, 1997). Esse método é denominado ingênuo porque ele assume que os atributos contribuem de forma independente para formar probabilidades estimativas da ocorrência de uma determinada classe do problema durante o processo de classificação (JOHN; LANGLEY, 1995). Contudo, o algoritmo tem obtido bons resultados em situações complexas do mundo real (FRANK; BOUCKAERT, 2006).

2.3.6 Vizinho Mais Próximo

É uma técnica baseada em aprendizagem por analogia, ou seja, comparando uma determinada tupla teste com tuplas de treinamento que são semelhantes. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional. Desta forma, todas as tuplas de formação são armazenadas num espaço padrão de n dimensões. Quando uma dada tupla é desconhecida, um classificador k-vizinho mais próximo procura o espaço padrão para as tuplas de treinamento k que estão mais próximas da tupla desconhecida. Estas tuplas de treinamento k são os k "vizinhos mais próximos" da tupla desconhecida (HAN; KAMBER, 2012).

2.4 Matriz de Confusão

Dentre as diversas formas de avaliar a capacidade de predição de um classificador para determinar a classe de vários registros, a matriz de confusão é a mais simples dessas formas (HAN; PEI; KAMBER, 2011).

Para n classes, a matriz de confusão é uma tabela de dimensão $n \times n$. Para cada classificação possível, existe uma linha e coluna correspondente, ou seja, os valores das classificações serão distribuídos na matriz de acordo com os resultados, assim gerando a matriz de confusão para as classificações realizadas (BASILI; WEISS, 1983). As linhas correspondem às classificações corretas e as colunas representam as classificações realizadas pelo classificador (HAND; MANNILA; SMYTH, 2001).

Quando existem apenas duas classes, uma é considerada como *positivo* (no contexto desse trabalho, "Evadido") e a outra como *negativo* ("Não Evadido") (HAND; MANNILA; SMYTH, 2001). Assim, podemos ter quatro resultados possíveis:

- *Verdadeiro Positivo* (TP): uma instância de classe positivo é classificada corretamente como positivo;
- *Falso Negative* (FN): uma instância de classe positivo é classificada incorretamente como negativo;
- *Verdadeiro Negativo* (TN): uma instância de classe negativo pode ser classificada corretamente como negativo;
- *Falso Positivo* (FP): uma instância de classe negativo é classificada incorretamente como positivo.

As predições corretas e errôneas para as duas classes podem ser dispostas em uma única matriz, conforme é apresentada no Quadro 1 (HAND; MANNILA; SMYTH, 2001).

Quadro 1 – Matriz de confusão.

Classe Verdadeira	**Classe Prevista**	
	Positivo	**Negativo**
Positivo	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Negativo	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: Adaptada Hand, Mannila e Smyth (2001).

2.4.1 Métricas de Qualidade

Neste trabalho, foram utilizadas as métricas acurácia, *recall*, precisão e medida-F.

2.4.1.1 Acurácia

É o percentual de instâncias classificadas corretamente.

$$acuracia = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

2.4.1.2 *Recall*

É o percentual de instâncias que foram classificadas corretamente como positivo.

$$recall = \frac{TP}{TP+FN} \tag{2.2}$$

2.4.1.3 Precisão

É o percentual de instâncias classificadas como positivo que são realmente positivo.

$$precisao = \frac{TP}{TP+FP} \tag{2.3}$$

2.4.1.4 Medida-F

Também conhecida como média harmônica, pois combina a precisão e *recall*, ponderando uniformemente.

$$medida-F = \frac{2.precisao.recall}{precisao+recall} \tag{2.4}$$

2.5 Trabalhos Relacionados

Recentemente, um número razoável de pesquisas foram conduzidas para aplicar técnicas de mineração de dados na área de educação, ordenadas para classificar e prever o desempenho dos alunos em vários institutos de educação.

Gopalakrishnan et al. (2017) propõem a aplicação de técnicas de mineração de dados para prever insucesso escolar e abandono, em um estudo de caso com dados de 1552 estudantes do ensino superior da Universidade de São Francisco, Califórnia. As acurácias obtidas nos experimentos variaram de 58% a 87% nos cinco classificadores utilizados. Os autores concluem que algoritmos de classificação podem ser utilizados com sucesso, a fim de prever um desempenho escolar do aluno.

Márquez-Vera et al. (2016) utiliza em seu trabalho dados coletados de 419 estudantes do ensino médio no México, vários experimentos foram realizados para prever o abandono em diferentes etapas do curso, com o objetivo de selecionar os melhores indicadores de abandono e para comparar um algoritmo proposto com alguns algoritmos de classificação clássicos. Os resultados mostram que o algoritmo foi capaz de prever a evasão de alunos nas primeiras quatro e seis semanas do curso e ser confiável o suficiente para ser usado em um sistema de alerta. As acurácias obtidas variaram de 74% a 97%, nos seis classificadores utilizados. Vale a pena salientar que apesar de possuir uma acurácia alta em alguns ensaios, o contexto aplicado é diferente ao proposto neste trabalho.

No âmbito do Brasil, Manhães et al. (2012) compararam 6 algoritmos classificadores e verificaram problemas com alunos que não conseguem completar os seus cursos de graduação. A amostra dos dados é composta por 7304 alunos do ensino superior da UFRJ. Os dados são classificados em três classes: alunos que concluíram o curso obtendo o diploma, alunos que não conseguiram concluir o curso, e alunos que possuíam matrícula ativa depois do prazo médio de conclusão do curso de graduação na IFES. O estudo obteve uma precisão de acerto em torno de 80%.

Kantorski et al. (2016) abordam a evasão existente na graduação de uma instituição pública de ensino superior, considerando os cursos de Administração e Zootecnia. A pesquisa desenvolvida utiliza vários métodos de aprendizagem de máquina para a previsão e apresenta como resultados experimentos que alcançaram uma acurácia em torno de 70%, na previsão de alunos que abandonaram o curso.

Machado et al. (2015) apresentaram no artigo "Estudo bibliométrico em mineração de dados e evasão escolar", uma pesquisa bibliométrica dos artigos científicos publicados, entre 2005 e março de 2015, os quais abordam os temas mineração de dados e evasão escolar, nas bases de dados *Scopus*, *Web Of Science*, *ScienceDirect* e *Scielo*. Na busca realizada nas bases de dados, retornaram 16 artigos na *Scopus*, 6 artigos na *Web Of Science*, 3 artigos na *ScienceDirect* e, na *Scielo*, não retornaram artigos. Assim, fizeram parte do escopo deste estudo a análise de 24 artigos científicos.

Machado et al. (2015) também listaram nove métodos de mineração de dados encontrados em sua pesquisa, entre eles, *Decision Trees* (Árvores de Decisão), *Neural Network* (Redes Neurais), Naive Bayes e *Support Vector Machine* (Máquina de vetores de suporte).

Foi feito um mapeamento sistemático com o objetivo de analisar trabalhos publicados entre os anos de 2014 a 2018 na área de EDM aplicados à evasão escolar, nas bases de dados *ACM Digital Library*, *IEEEXplore*, *SpringerLink* e *WileyInterScience*.

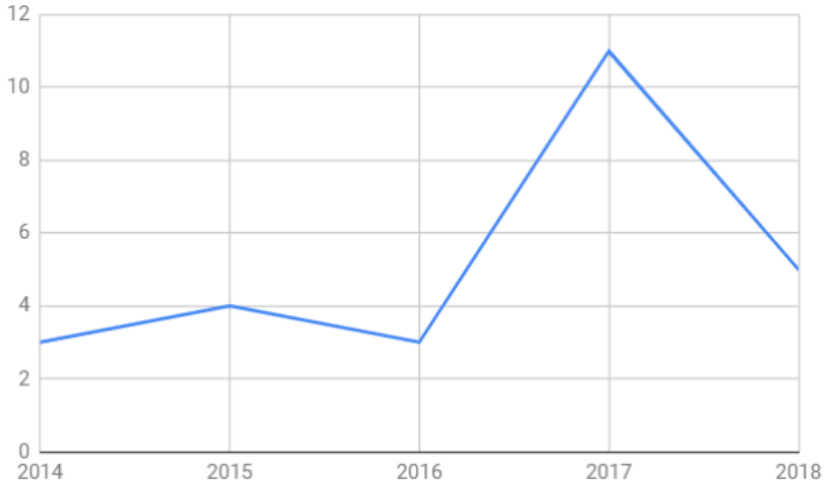
Visando encontrar artigos inseridos na área de EDM com foco no problema da evasão escolar, foram utilizadas as palavras-chave "educational data mining + student" e algumas palavras associadas a evasão escolar. Para este significado, foi utilizado o termo "drop out" e

as diferentes grafias (“dropout” e “drop-out”) observadas em trabalhos científicos sobre o tema. Também foi utilizada a palavra “evasion”, visto que esta também designa o ato de evadir, embora menos comum em trabalhos científicos no contexto da evasão escolar.

Após a definição da estratégia, foi realizada a busca primária, que retornou 156 artigos. Em seguida foi realizado a leitura dos títulos e resumos onde o número inicial foi reduzido para 38 estudos selecionados. Em última revisão e aplicação os critérios de inclusão e exclusão estabelecidos neste trabalho, foram definidos os 26 artigos escolhidos para compor este mapeamento sistemático.

Como é possível observar na Figura 1, durante o período analisado (2014 a 2018), houve um acréscimo nas publicações referente a área de mineração de dados aplicadas à evasão escolar. O ano de 2017 destaca-se com a maior quantidade de contribuições ao tema. Os Estados Unidos se destaca em publicações de artigos na área, com 8 contribuições, o Brasil aparece com 2 contribuições no período mencionado.

Figura 1 – Publicações por ano, no período entre 2014 e 2018.



Fonte: Própria (2019).

No Quadro 2, é possível observar uma lista dos trabalhos citados anteriormente, seus classificadores e as acurácias obtidas. Importante notar que os trabalhos utilizam bases de dados diferentes para os seus estudos, impossibilitando uma comparação direta dos resultados. No entanto, alguns destes trabalhos também utilizaram dados de uma instituição pública de ensino superior, podendo ajudar, em conjunto com este trabalho, em uma análise geral secundária.

Quadro 2 – Trabalhos relacionados, classificadores utilizados e acurácias obtidas.

Autores	Classificadores	Acurácia
(MáRQUEZ-VERA et al., 2016)	J48, Naive Bayes, SMO, JRip, J48, IBK, ICRM	Entre 74% e 97%

(GOPALAKRISHNAN et al., 2017)	Naive Bayes, SVM, AdaBoost, K Neighbours, Extra trees	Entre 58% e 87%
(MANHÃES et al., 2012)	J48, SVM, AdaBoost, Naïve Bayes, SimpleCart, MLP	Média = 80%
(KANTORSKI et al., 2016)	J48, IBK, CART, Naive Bayes, MLP	Média = 74%

Fonte: Própria (2019).

Vale a pena mencionar que não foram encontrados trabalhos que realizaram uma análise comparativa de algoritmos aplicados ao contexto da evasão escolar, considerando uma abordagem experimental, com a validação estatística da significância dos dados, como é proposto neste trabalho. Uma base de conhecimento robusta só poderá ser gerada com as replicações de verdadeiros experimentos controlados que validem estatisticamente seus trabalhos, as quais poderão servir de insumo para verdadeiras meta-análises dos dados.

3

Primeira Avaliação Experimental

Neste e no próximo capítulo, este trabalho é apresentado como um processo experimental. O mesmo segue as diretrizes de (WOHLIN et al., 2012). A Figura 2 ilustra as etapas do trabalho, as quais serão mais detalhadas ao decorrer do capítulo.

Foram considerados os dados de todos alunos de graduação de duas instituições de nível superior, IFS e UFS, no período entre 2003 a 2017. Neste primeiro experimento, para o processo de mineração de dados, foi utilizada a ferramenta *Weka* (HALL EIBE FRANK; WITTEN, 2009), com os dados do IFS, no segundo experimento foi utilizada a linguagem Python e suas bibliotecas na base do IFS e da UFS.

Figura 2 – Etapas do trabalho.

Fonte: Própria (2019).

3.1 Definição do Objetivo

Esta e a próxima seção irá focar na etapa 1, ou seja, definição do objetivo e no planejamento dos experimentos. O objetivo deste estudo é avaliar os principais classificadores encontrados nos trabalhos de EDM, identificando o melhor algoritmo em termos eficácia, com foco na evasão escolar, no âmbito de duas instituições de ensino superior. Os principais modelos serão usados para formar um metamodelo de predição que ajudará à gestão da evasão.

O experimento terá como alvo os dados da graduação do Instituto Federal de Sergipe (IFS). O objetivo foi formalizado utilizando o modelo GQM proposto por (BASILI; WEISS, 1983): **Analisar**, por meio de experimento controlado, os principais algoritmos de mineração de dados aplicados ao contexto da educação, com **foco** na evasão escolar, **com a finalidade de** confirmar e/ou identificar o melhor algoritmo em termos de eficácia, **com respeito à** acurácia, *recall*, precisão e medida-F, **do ponto de vista de** pesquisadores e profissionais de *Data Analytics*, **no contexto** dos dados sobre evasão de duas instituições públicas de ensino superior.

3.2 Planejamento

3.2.1 Seleção de Contexto

O experimento foi "in vivo"(análises foram feitas em ambiente real, no qual foram gerados e serão usados os modelos de conhecimento para apoio ao planejamento da redução da evasão) e considerou os dados de alunos de todos os cursos de Graduação, de uma instituição pública de nível superior.

Foram considerados os alunos ingressantes entre 2003, ano no qual se iniciaram os primeiros cursos de Graduação, e 2017. O ano da instituição no período da elaboração deste experimento era 2018, o qual, por este motivo, não fez parte deste estudo. A seleção de dados levou em consideração atributos pessoais, acadêmicos e sociais.

3.2.2 Formulação de Hipóteses

Para guiar o estudo, foi elaborada a seguinte questão de pesquisa, cuja resposta visa cumprir o objetivo do trabalho. No contexto da Evasão Escolar de uma instituição de ensino superior, entre os algoritmos selecionados na área de EDM, qual o melhor em termos de Eficácia?

Para avaliar esta questão, foram utilizadas quatro métricas: Acurácia (2.1), *Recall* (2.2), Precisão (2.3) e Medida-F (2.4). Sendo assim, com os objetivos e métricas definidas, será considerada a hipótese a seguir (para cada métrica):

- H_0: Os algoritmos$_{(1,2..n)}$ possuem médias iguais para a métrica.
  μ1(metrica) = μ2(metrica) ... = μn(metrica);
- H_1: Os algoritmos$_{(1,2..n)}$ possuem médias diferentes para a métrica.
  μ1(metrica) $\neq$ μ2(metrica) ... $\neq$ μn(metrica);

3.2.3 Seleção de Participantes

A base de dados analisada foi a do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), o qual armazena toda a vida acadêmica dos estudantes da instituição. As

instituições cederam a base de dados para o experimento em questão.

Após o pré-processamento, o qual consistiu na remoção de registros com dados nulos, foram selecionadas 6672 instâncias, que representam todos os alunos de graduação da instituição, dentro do período mencionado anteriormente. Dos dados selecionados, 3212 (48,1%) representam alunos evadidos e 3460 (51,9%) representam alunos não evadidos.

3.2.4 Variáveis Independentes

Para a tarefa de classificação, foram considerados 17 atributos, da base descrita na subseção 3.2.3, os quais são apresentados no Quadro 3.

Quadro 3 – Atributos considerados para a análise.

Atributo	**Descrição**
sexo	Gênero do discente
idade	Idade do discente no início do curso
inst_seg_grau	Tipo de instituição de conclusão do segundo grau
raca	Etnia do discente
est_civil	Estado civil
qtd_tranc	Quantidade de trancamentos no curso
reab_matricula	Indica se o aluno realizou reabertura de matrícula no curso
qtd_ap_med_p	Quantidade média de disciplinas aprovadas por período
qtd_ap_1p	Quantidade de disciplinas aprovadas no primeiro período
qtd_rep_med_p	Quantidade média de disciplinas reprovadas por período
qtd_rep_1p	Quantidade de disciplinas reprovadas no primeiro período
qtd_per_cur	Quantidade de períodos cursados pelo discente
cra	Coeficiente de rendimento acadêmico
perc_aprov	Percentual de disciplinas aprovadas pelo discente
media_geral	Média geral do discente no curso
media_Faltas	Média de faltas do discente no curso
cotista	Indica se o aluno ingressou no curso por sistema de cotas

Fonte: Própria (2019).

Os algoritmos utilizados: árvore de decisão (J48), Naive Bayes, vizinho mais próximo (IBK), redes neurais (MLP), máquina de vetores de suporte (SMO) e *ensemble methods (RandomForest)*, com os parâmetros apresentados na Tabela 1.

3.2.5 Variáveis Dependentes

Acurácia (2.1), *Recall* (2.2), Precisão (2.3) e Medida-F (2.4).

Tabela 1 – Parâmetros utilizados por algoritmo (*Weka*).

Algoritmo	Naive Bayes	J48	IBk	MLP	Random	SMO
bacthSize	100	100	100	100	100	100
doNotCheckCap	false	-	-	-	-	-
debug	false	-	-	-	-	-
folds	- 3	3	-	-	-	-1
numDecimalPlaces	2	-	2	2	2	2
supervisedDisc	false	-	-	-	-	-
seed	-	1	-	0	1	1
bacthSize	100 -	100	-	-	-	-
binarySplits	-	false	-	-	-	-
confidenceFactor	-	0.25	-	-	-	-
minNumObj	-	2	-	-	-	-
SubtreeRaising	-	true	-	-	-	-
useMDLcorrection	-	true	-	-	-	-
KNN	-	-	3	-	-	-
meanSquared	-	-	false	-	-	-
autoBuild	-	-	-	true	-	-
GUI	-	-	-	false	-	-
learningRate	-	-	-	0.3	-	-
momentum	-	-	-	0.2	-	-
trainingTime	-	-	-	500	-	-
validationThreshold	-	-	-	20	-	-
c	-	-	-	-	-	1.0
epsilon	-	-	-	-	-	1.0E-12
filterType	-	-	-	-	-	Normalize training data
kernel	-	-	-	-	-	PolyKernel -E1.0-C250007
toleranceParameter	-	-	-	-	-	0
calibrator	-	-	-	-	-	logistic -R1.0E-8-M-1
bagSizePercent	-	-	-	-	100	-
maxDepth	-	-	-	-	0	-
numExecutionSlots	-	-	-	-	1	-
numFeatures	-	-	-	-	0	-
numIterations	-	-	-	-	100	-

Fonte: Própria (2019).

3.2.6 Projeto do Experimento

Uma das métricas utilizadas neste trabalho foi a acurácia, a qual exige o balanceamento dos dados das classes. Como nossa base já é balanceada (MACHADO; MARCELO, 2007), não foi necessário planejar a adoção de um método de balanceamento.

Foi utilizada a abordagem 10-*Fold Cross-validation* (HASTIE; TIBSHIRANI; FRIEDMAN, 2001), em que os dados são divididos em 10 partes, mantendo suas proporções. Assim, são realizados 10 testes, nos quais uma parte dos dados é separada para ser testada posteriormente

e as demais são usadas para serem treinadas.

3.2.7 Instrumentação

Para o processo de mineração de dados, foi utilizada a ferramenta *Weka* (HALL EIBE FRANK; WITTEN, 2009), a qual possui diversos algoritmos de aprendizado de máquina que podem ser utilizados para extrair informações relevantes de uma base de dados. A ferramenta foi adotada devido ao fato de ser *open-source* e livre de custos, além da facilidade de uso, para um experimento piloto.

Os dados utilizados para a análise provêm do SIGAA, o qual possui como SGBD (Sistema de Gerenciamento de Banco de Dados) o *PostgreSQL*. Foi criando um ETL (*Extract, Transform and Load*), para extrair, limpar e carregar os dados em um *Data Warehouse* específico, o qual é a base para geração dos modelos de conhecimento, levando em consideração as variáveis detalhadas na subseção 3.2.4.

3.3 Operação do Experimento

3.3.1 Preparação

Consistiu na execução do ETL implementado para carga do *Data Warehouse*, representado nas etapas 2 e 3 da Figura 2.

3.3.2 Execução

Consistiu na realização do processo classificatório nos dados dos discentes, planejado na Seção 3.2.6, para cada algoritmo de mineração selecionado, utilizando o dicionário discutido na subseção 3.2.4, correspondendo assim, à etapa 4 da Figura 2.

3.3.3 Validação dos Dados

Como auxílio para análise, interpretação e validação - etapa 5 da Figura 2, foram utilizados quatro tipos de testes estatísticos, Teste Shapiro-Wilk, Teste Levene, Teste Anova e Teste Tukey.

O Teste Anova foi utilizado por ser necessário comparar mais de dois grupos de valores. Como este teste possui os pressupostos de que a distribuição deve ser Normal e de que haja homocedasticidade entre os tratamentos (variâncias homogêneas) (FIELD, 2009), foi utilizado o Teste Shapiro-Wilk (SHAPIRO; WILK, 1965) para o teste de Normalidade e o Teste de Levene (LEVENE, 1960) para o teste de homocedasticidade.

O Teste Anova evidencia que ao menos um algoritmo se diferencia dos demais, mas não é possível afirmar qual é mais discrepante. Para isso, foi utilizado o Teste Tukey, que

segundo Anjos (ANJOS, 2009), permite testar qualquer contraste, sempre, entre duas médias de tratamentos, sendo possível verificar quais são estatisticamente iguais ou diferentes.

Todos os testes estatísticos foram feitos utilizando a Ferramenta SPSS – IBM (IBM, 2017).

3.4 Resultados

3.4.1 Análise e Interpretação de Dados

Após a execução dos algoritmos, utilizando a abordagem 10-*Cross-validation*, foram obtidos os resultados das classificações, os quais serão apresentados abaixo.

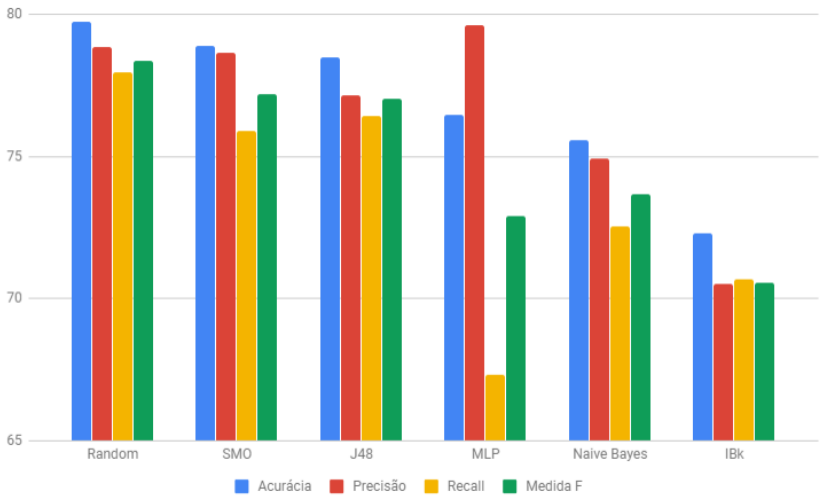
Na Tabela 2 e na Figura 3, são apresentadas as médias das métricas obtidas por cada algoritmo.

Tabela 2 – Comparativo das métricas dos algoritmos - Experimento 1.

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
Radom Forest	79,75%	78,86%	77,95%	78,37%
SMO	78,89%	78,65%	75,91%	77,18%
J48	78,50%	77,15%	76,43%	77,01%
MLP	76,45%	79,63%	67,31%	72,92%
Naive Bayes	75,58%	74,93%	72,54%	73,67%
IBk	72,28%	70,53%	70,69%	70,57%

Fonte: Própria (2019).

Figura 3 – Gráfico comparativo das métricas dos algoritmos - Experimento 1.



Fonte: Própria (2019).

Estes resultados foram utilizados para responder a questão de pesquisa. Como é perceptível, os algoritmos obtiveram médias de acurácias distintas e o algoritmo *RandomForest* obteve a maior média, seguido pelo SMO e J48, que conseguiram médias bastante semelhantes e próximas ao *RandomForest*. Todavia, não é possível fazer essas afirmações sem evidências estatísticas suficientemente conclusivas.

Como já mencionado anteriormente, o Teste Anova foi aplicado para validar a hipótese, e, por possuir os pressupostos da normalidade e da homocedasticidade, primeiramente foi feito o Teste Shapiro-Wilk e em seguida o Teste de Levene.

Foi definido um nível de significância de 0,05 em todo o experimento. Ao aplicar o Teste de Shapiro-Wilk, para análise da normalidade da distribuição dos dados, foram obtidos os *p-values* apresentados na Tabela 3, na qual, observa-se valores acima do nível de significância adotado, concluindo-se que as distribuições são normais.

Tabela 3 – Resultado do Teste de Shapiro-Wilk, para análise da normalidade dos dados - Experimento 1.

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
IBK	0,208	0,531	0,573	0,998
MLP	0,295	0,833	0,447	0,232
SMO	0,145	0,651	0,254	0,717
J48	0,353	0,676	0,48	0,52
Naive Bayes	0,433	0,148	0,182	0,973
Random Forest	0,374	0,951	0,936	0,975

Fonte: Própria (2019).

Em seguida, foi feito o Teste de Levene, obtendo-se os resultados apresentados na Tabela 4, como pode ser observado, os *p-values* obtidos são maiores que o nível de significância adotado, validando o pressuposto da homogeneidade de variâncias.

Uma vez que os pressupostos foram atendidos, foi possível aplicar o Teste Anova, em que, verificou-se *p-values* fortemente menores que o nível de significância adotado, como pode ser observado na tabela 4. Desta forma, confirmou-se a evidência de diferença entre as médias, ou seja, a hipótese (H_0), de que os algoritmos possuem a mesma acurácia foi rejeitada, dentro do contexto do experimento realizado.

Sendo assim, com o Teste Anova, foi evidenciado que ao menos um algoritmo se diferencia dos demais, mas não é possível afirmar qual é mais discrepante. Para isso, foi utilizado o Teste Tukey, a Tabela 5 a seguir, apresenta as médias das acurácias dos algoritmos agrupados, formando quatro grupos homogêneos. É possível observar que a maior média foi a do algoritmo *RandomForest*, com de 79,75%. Porém, considerando o nível de significância de 5%, os algoritmos *RandomForest*, SMO e J48 apresentaram médias similares de acurácia. O IBK apresentou a média mais baixa, com 72,28%. Estes resultados confirmam evidências anteriores encontradas

Tabela 4 – *P-values* dos Testes de Levene e Anova - Experimento 1.

Métricas	Levene	Anova
Acurácia	0,744	$4,65^{-13}$
Precisão	0,163	$3,61^{-11}$
Recall	0,306	$3,50^{-13}$
Medida-F	0,438	$7,68^{-16}$

Fonte: Própria (2019).

nos trabalhos descritos em (MARQUEZ-VERA; MORALES; SOTO, 2013) e (DEKKER; J., 2009).

Tabela 5 – Valores obtidos pelo Teste de Tukey para Acurácia - Experimento 1.

Algoritmo	Subconjunto para alfa = 0.05			
	1	2	3	4
IBk	72,28			
Naive Bayes		75,58		
MLP		76,45	76,45	
J48			78,5	78,5
SMO				78,89
Random Forest				79,75
Sig.	1	0,865	0,101	0,597

Fonte: Própria (2019).

Algo semelhante aconteceu com as demais métricas (precisão, *recall* e medida-F), a Tabela 6 apresenta o resultado do Teste de Tukey da medida-F. Para evitar repetição, o Teste de Tukey da precisão e *recall* serão omitidos neste trabalho, já que a medida-F faz a combinação destas métricas. Os resultados evidenciam que os algoritmos *RandomForest*, SMO e J48 apresentaram médias similares para a medida-F, sendo 78,37%, 77,18% e 77,01%, respectivamente. O IBK apresentou a média mais baixa, com 70,57%.

Utilizando o algoritmo *CfsSubsetEval*, para seleção de atributos, percebe-se que alguns possuem maior relevância e outros poderiam ser eliminados sem influenciar nos resultados. Os 3 atributos de maior relevância foram, em ordem, partindo do mais relevante: "perc_aprov", "raca"e "cotista". Os atributos que apresentaram menor relevância foram "sexo"e "qtd_ap_med_p", os quais foram retirados do modelo e não influenciaram nas médias das métricas avaliadas.

Isto evidencia, por exemplo, que o problema não é privilégio de um sexo específico e traz um alerta para análises mais aprofundadas das áreas sociais da instituição, considerando as etnias e tipo de ensino na educação básica que estão sendo mais afetados.

Tabela 6 – Valores obtidos pelo Teste de Tukey para a Medida-F - Experimento 1.

Algoritmo	Subconjunto para alfa = 0.05		
	1	2	3
IBk	70,57		
MLP		72,92	
Naive Bayes		73,67	
J48			77,01
SMO			77,18
Random Forest			78,37
Sig.	1	0,902	0,422

Fonte: Própria (2019).

3.5 Ameaças à Validade

Embora os resultados do experimento tenham se mostrado satisfatórios, o mesmo apresenta ameaças à sua validade que devem ser comentadas.

Ameaças à validade interna: O sistema acadêmico atual está presente na Instituição desde 2017, o qual herdou a base do sistema acadêmico legado, com várias informações inconsistentes, principalmente até meados de 2007. Esta ameaça foi mitigada com a realização do processo de limpeza dos dados, diminuindo a probabilidade do uso de informações mais antigas incorretas.

Ameaças à validade de construção: No experimento desta dissertação, a instituição não possuía algumas informações bastante relevantes em relação à evasão, como, por exemplo, dados socioeconômicos dos alunos e seus familiares. A inclusão dessas informações pode influenciar nos resultados do desempenho dos algoritmos, aumentando as eficácias destes. Para mitigar esta ameaça e incrementar o Sistema de Apoio à Decisão a ser desenvolvido, estas informações serão sugeridas para captação por parte da instituição e levadas em consideração em trabalhos futuros.

4

Segunda Avaliação Experimental

Neste experimento, para o processo de mineração de dados, foi utilizada a linguagem Python e suas bibliotecas, com os dados do IFS e da UFS.

4.1 Definição do Objetivo

De forma análoga ao descrito na subseção 3.1, o objetivo deste experimento é avaliar os principais classificadores encontrados nos trabalhos de EDM, identificando o melhor algoritmo em termos eficácia, com foco na evasão escolar, no âmbito de duas instituições de ensino superior. Os principais modelos serão usados para formar um metamodelo de predição que ajudará à gestão da evasão.

O experimento terá como alvo os dados da graduação do IFS e da UFS. O objetivo foi formalizado utilizando o modelo GQM proposto por (BASILI; WEISS, 1983): **Analisar**, por meio de experimento controlado, os principais algoritmos de mineração de dados aplicados ao contexto da educação, com **foco** na evasão escolar, **com a finalidade de** confirmar e/ou identificar o melhor algoritmo em termos de eficácia, **com respeito à** acurácia, *recall*, precisão e medida-F, **do ponto de vista de** pesquisadores e profissionais de *Data Analytics*, **no contexto** dos dados sobre evasão de uma instituição pública de ensino superior.

4.2 Planejamento

4.2.1 Seleção de Contexto

O experimento foi "in vivo"(análises foram feitas em ambiente real, no qual foram gerados e serão usados os modelos de conhecimento para apoio ao planejamento da redução da evasão) e considerou os dados de alunos de todos os cursos de Graduação, de duas instituições públicas de nível superior.

Com dados do IFS, foram considerados os alunos ingressantes entre 2003, ano no qual se iniciaram os primeiros cursos de Graduação, e 2017. Para UFS foi considerado o mesmo período, o ano da elaboração deste experimento era 2018, o qual, por este motivo, não fez parte deste estudo. A seleção de dados levou em consideração atributos pessoais, acadêmicos e sociais.

4.2.2 Formulação de Hipóteses

Para guiar o estudo, foi elaborada a seguinte questão de pesquisa, cuja resposta visa cumprir o objetivo do trabalho. No contexto da Evasão Escolar de uma instituição de ensino superior, entre os algoritmos selecionados na área de EDM, qual o melhor em termos de Eficácia?

Para avaliar esta questão, foram utilizadas quatro métricas: Acurácia (2.1), *Recall* (2.2), Precisão (2.3) e Medida-F (2.4). Sendo assim, com os objetivos e métricas definidas, será considerada a hipótese a seguir (para cada métrica):

- H_0: Os algoritmos$_{(1,2..n)}$ possuem médias iguais para a métrica.
  μ1(metrica) = μ2(metrica) ... = μn(metrica);
- H_1: Os algoritmos$_{(1,2..n)}$ possuem médias diferentes para a métrica.
  μ1(metrica) $\neq$ μ2(metrica) ... $\neq$ μn(metrica);

4.2.3 Seleção de Participantes

A base de dados analisada foi a do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), o qual armazena toda a vida acadêmica dos estudantes das instituições. As instituições cederam a base de dados para o experimento em questão.

4.2.3.1 Base do IFS

Após o pré-processamento, o qual consistiu na remoção de registros com dados nulos, foram selecionadas 6672 instâncias, que representam os alunos de graduação da instituição, dentro do período mencionado anteriormente. Dos dados selecionados, 40.571 (48,1%) representam alunos evadidos e 44.048 (51,9%) representam alunos não evadidos.

4.2.3.2 Base da UFS

Na base da UFS foram selecionadas 84.619 instâncias, que representam os alunos de graduação da instituição, dentro do período mencionado anteriormente. Dos dados selecionados, 40.571 (47,95%) representam alunos evadidos e 44.048 (52,05%) representam alunos não evadidos.

4.2.4 Variáveis Independentes

Para a tarefa de classificação, na base do IFS, foram considerados 17 atributos, apresentados na subseção 3.2.4. Na base da UFS, além destes atributos, foi considerado o atributo "auxilio", para indicar se o discente possui algum auxílio. Este atributo não foi considerado na base do IFS por não possuir esta informação na respectiva base.

Os algoritmos utilizados foram os correspondentes aos utilizados na ferramenta *Weka*: árvore de decisão, Naive Bayes, k-Nearest Neighbor (KNN), Neural Networks (MLP), máquina de vetores de suporte (SVM) e *ensemble methods (Random Forest)*, com os parâmetros apresentados na Tabela 7.

4.2.5 Variáveis Dependentes

Acurácia (2.1), *Recall* (2.2), Precisão (2.3) e Medida-F (2.4).

4.2.6 Projeto do Experimento

Uma das métricas utilizadas neste trabalho foi a acurácia, a qual exige o balanceamento dos dados das classes. Como as nossas bases já são balanceadas (MACHADO; MARCELO, 2007), não foi necessário planejar a adoção de um método de balanceamento.

Foi utilizada a abordagem 10-*Fold Cross-validation* (HASTIE; TIBSHIRANI; FRIEDMAN, 2001), em que os dados são divididos em 10 partes, mantendo suas proporções. Assim, são realizados 10 testes, nos quais uma parte dos dados é separada para ser testada posteriormente e as demais são usadas para serem treinadas.

Os algoritmos podem receber vários parâmetros de ajuste, não sendo possível identificar previamente quais valores para o parâmetros irão produzir os melhores modelos a partir dos dados de treinamento, sendo necessário o teste de uma gama de valores para esses parâmetros. Tal atividade é onerosa e maçante, dificultando a criação de modelos de alta performance. Para reduzir esse problema, foi utilizada uma técnica denominada *Grid Search*, onde se pode construir vários modelos a partir da combinação iterativa de diferentes valores para cada parâmetro (COOK, 2017). Após esse processo, os parâmetros que apresentaram os melhores resultados são apresentados na Tabela 7.

4.2.7 Instrumentação

Para o processo de mineração de dados, foi utilizada a linguagem Python e suas bibliotecas, as quais possuem diversos algoritmos de aprendizado de máquina que podem ser utilizados para extrair informações relevantes de uma base de dados. Segundo o 16º anuário do uso de software de análise e mineração de dados (POLL, 2015), Python foi considerada a linguagem de programação mais usada pela comunidade de profissionais de Mineração de dados e Big data.

Python tem inúmeras razões de ter atraído interesse como linguagem para análise de dados, é open-source e livre de custos, possui um variado conjunto de bibliotecas para trabalhar com diversas áreas, permitindo a comparação de desempenho dos algoritmos e apresentando diversos recursos para análise dos dados. Além de oferecer uma sintaxe simples e objetiva, permitindo o programador focar no problema a ser resolvido, sem se preocupar tanto com detalhes de implementações.

Os dados utilizados para a análise provêm do SIGAA, o qual possui como SGBD o *PostgreSQL*. Foi criado um ETL (*Extract, Transform and Load*), para extrair, limpar e carregar os dados em um *Data Warehouse* específico, o qual é a base para geração dos modelos de conhecimento, levando em consideração as variáveis detalhadas na subseção 3.2.4.

4.3 Operação do Experimento

4.3.1 Preparação

Consistiu na execução do ETL implementado para carga do *Data Warehouse*, representado nas etapas 2 e 3 da Figura 2.

Nesta etapa, foi feita a transformação de alguns atributos, em que foi aplicado a abordagem *"One Hot"*, consistindo em representar uma variável categórica de forma binária.

Após esse pré-processamento, na base do IFS, os atributos passaram de 17 para 65, e na base da UFS de 18 para 86, segundo (HAND H. M., 2000), nem todos os atributos da base de dados são necessários para discriminar a classe de maneira precisa, e incluí-los no modelo de classificação pode até mesmo gerar resultados inferiores do que seriam obtidos se eles fossem removidos da base. Sendo assim, em seguida foi feita a seleção de atributos.

Para isso, existem diversas técnicas, neste trabalho foi utilizada a de Ranqueamento de Atributos, em que, consiste em utilizar uma métrica para avaliar a qualidade de cada um dos atributos individualmente. Dessa forma, é possível ordenar o conjunto de atributos em um ranking, que pode então ser utilizado para selecionar todos os atributos cujo valor da métrica esteja acima de um determinado limiar (PEREIRA, 2009). Neste trabalho, o classificador utilizado foi o *Random Forest*, em que através de testes, na base da IFS, foram selecionados os 41 primeiros atributos, e na da UFS, os 60 primeiros, mantendo a acurácia e em alguns casos apresentando melhores resultados.

4.3.2 Execução

Consistiu na realização do processo classificatório nos dados dos discentes, planejado na Seção 3.2.6, para cada algoritmo de mineração selecionado, utilizando o dicionário discutido na subseção 3.2.4, com o acréscimo do atributo auxilio, o qual indica se o aluno possui algum tipo

de auxílio, e utilizando os parâmetros apresentados na 7, correspondendo assim, à etapa 4 da Figura 2.

4.3.3 Validação dos Dados

Como auxílio para análise, interpretação e validação - etapa 5 da Figura 2, foram utilizados quatro tipos de testes estatísticos, Teste Shapiro-Wilk, Teste Levene, Teste Anova e Teste Tukey.

O Teste Anova foi utilizado por ser necessário comparar mais de dois grupos de valores. Como este teste possui os pressupostos de que a distribuição deve ser Normal e de que haja homocedasticidade entre os tratamentos (variâncias homogêneas) (FIELD, 2009), foi utilizado o Teste Shapiro-Wilk (SHAPIRO; WILK, 1965) para o teste de Normalidade e o Teste de Levene (LEVENE, 1960) para o teste de homocedasticidade.

O Teste Anova evidencia que ao menos um algoritmo se diferencia dos demais, mas não é possível afirmar qual é mais discrepante. Para isso, foi utilizado o Teste Tukey, que segundo Anjos (ANJOS, 2009), permite testar qualquer contraste, sempre, entre duas médias de tratamentos, sendo possível verificar quais são estatisticamente iguais ou diferentes.

Todos os testes estatísticos foram feitos utilizando a Ferramenta SPSS – IBM (IBM, 2017).

4.4 Resultados

4.4.1 Análise e Interpretação de Dados

Após a execução dos algoritmos, utilizando a abordagem 10-*Cross-validation*, foram obtidos os resultados das classificações, os quais serão apresentados a seguir.

4.4.1.1 Base do IFS

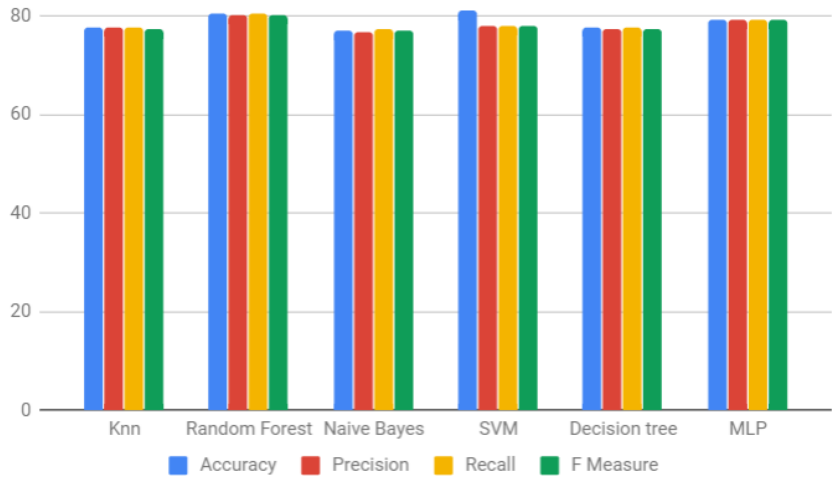
Na Tabela 8 e na Figura 4, são apresentadas as médias das métricas obtidas por cada algoritmo com todos atributos.

Tabela 8 – Comparativo das métricas dos algoritmos com todos atributos - Experimento 2 (IFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
SVM	81,01%	77,88%	78,02%	77,91%
Random Forest	80,63%	80,28%	80,33%	80,3%
MLP	79,15%	79,3%	79,25%	79,25%
Decision tree	77,62%	77,44%	77,7%	77,49%
Knn	77,52%	77,56	77,51%	77,46%
Naive Bayes	77,13%	76,84%	77,45%	76,9%

Fonte: Própria (2019).

Figura 4 – Gráfico comparativo das métricas dos algoritmos com todos atributos - Experimento 2 (IFS).



Fonte: Própria (2019).

Utilizando o algoritmo *Random Forest* para seleção de atributos, foi possível notar que alguns tem mais relevância e outros podem ser eliminados sem ter forte influencia nos resultados. Os quatro atributos de maior relevância partindo dos mais relevantes foram: "idade", "sexo", "qtd_ap_1p"(quantidade de disciplinas aprovadas no primeiro período) e "media_geral". Os atributos que apresentaram menor relevância foram "qtd_rep_med_p_9:10"(número médio de reprovações por período), "qtd_per_cur_15:19"(quantidade de períodos cursados pelo discente) e "qtd_ap_1p_11:13"(quantidade de disciplinas aprovadas no primeiro período).

Isso mostra, por exemplo, que a idade e o gênero dos alunos influenciam no problema e podem estar relacionados com as responsabilidades eles tem além dos estudos. Também traz um alerta para posterior análise das áreas sociais da instituição, considerando o gênero e faixa etária na educação básica mais afetada. Além disso, é evidenciado que o desempenho do aluno no primeiro período é bastante relevante.

Em seguida foram removidos os atributos menos relevantes, e os algoritmos foram

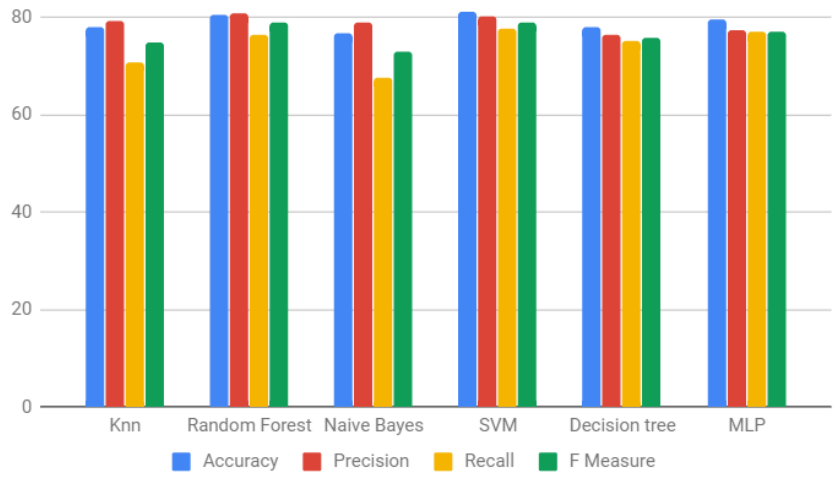
executados novamente. As médias de cada algoritmo são apresentadas na Tabela 9 e na Figura 5 abaixo.

Tabela 9 – Comparativo das métricas dos algoritmos com a seleção de atributos - Experimento 2 (IFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
SVM	81,18	80,25	77,51	78,81
Random Forest	80,36	80,79	76,50	78,86
MLP	79,43	77,34	76,92	77,06
Decision tree	78,06	76,51	75,15	75,79
Knn	78,01	79,09	70,83	74,69
Naive Bayes	76,8	79,00	67,59	72,81

Fonte: Própria (2019).

Figura 5 – Gráfico comparativo das métricas dos algoritmos com a seleção de atributos - Experimento 2 (Python).



Fonte: Própria (2019).

Esses resultados foram utilizados para responder a questão de pesquisa. Os algoritmos obtiveram médias distintas e o algoritmo SVM obteve os maiores, seguido do *Random Forest*, que alcançaou médias muito semelhantes e próximas ao SVM. No entanto, conforme discutido anteriormente na subseção 3.4.1, não é possível fazer tais suposições sem evidências estatísticas suficientemente conclusivas.

Ao aplicar o Teste de *Shapiro-Wilk*, para análise da normalidade da distribuição dos dados, foram obtidos os *p-values* apresentados na Tabela 10, na qual, observa-se valores acima do nível de significância adotado, concluindo-se que as distribuições são normais.

Tabela 10 – Resultado do Teste de *Shapiro-Wilk*, para análise da normalidade dos dados. - Experimento 2 (IFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
Knn	0,167	0,736	0,600	0,338
Random Forest	0,706	0,633	0,018	0,837
Naive Bayes	0,089	0,807	0,566	0,051
SVM	0,637	0,560	0,306	0,579
Decision tree	0,735	0,980	0,482	0,698
MLP	0,592	0,870	0,453	0,289

Fonte: Própria (2019).

Em seguida, foi feito o Teste de Levene, obtendo-se os resultados apresentados na Tabela 11, como pode ser observado, os *p-values* obtidos são maiores que o nível de significância adotado, validando o pressuposto da homogeneidade de variâncias.

Uma vez que os pressupostos foram atendidos, foi possível aplicar o Teste Anova, em que, verificou-se *p-values* fortemente menores que o nível de significância adotado, como pode ser observado na tabela 11. Desta forma, confirmou-se a evidência de diferença entre as médias, ou seja, a hipótese (H_0), de que os algoritmos possuem a mesma acurácia foi rejeitada, dentro do contexto do experimento realizado.

Tabela 11 – P-*values* dos Testes de Levene e Anova - Experimento 2 (IFS).

Métricas	Levene	Anova
Acurácia	0,807	$4,10^{-6}$
Precisão	0,795	$3,02^{-6}$
Recall	0,978	$4,29^{-13}$
Medida-F	0,469	$4,01^{-9}$

Fonte: Própria (2019).

Os resultados do Teste Tukey são apresentados na Tabela 12, a seguir, em que é possível observar quatro grupos homogêneos. Da análise dessa tabela, percebe-se que a maior média foi a do algoritmo SVM, com média de 79,75%. Porém, considerando o nível de significância de 5%, pode-se observar que os algoritmos SVM e *RandomForest*, apresentaram uma mesma média de acurácia. O *Naive Bayes* apresentou a média mais baixa, com 76,80%.

Tabela 12 – Valores obtidos pelo Teste de Tukey para Acurácia - Experimento 2 (IFS).

Algoritmo	Subconjunto para alfa = 0.05			
	1	2	3	4
Naive Bayes	76,80			
KNN		78,01		
Árvore de decisão		78,06	78,06	
MLP			79,43	
Random Forest				80,36
SVM				81,18
Sig.	1,000	,865	,101	,597

Fonte: Própria (2019).

De forma análoga, a Tabela 13 apresenta o resultado do teste de Tukey da Medida-F. Para evitar a repetição, o teste Tukey da precisão e *recall* será omitido, uma vez que a Medida-F faz a combinação dessas métricas. Os resultados mostram que os algoritmos *Random Forest*, SVM e MLP apresentaram médias semelhantes para a Medida-F, sendo 78,86%, 78,81% e 75,99%, respectivamente. *Naive Bayes* teve a menor média com 72,81%.

Tabela 13 – Valores obtidos pelo Teste de Tukey para a Medida-F - Experimento 2 (IFS).

Algoritmo	Subconjunto para alfa = 0.05			
	1	2	3	4
Naive Bayes	72,81			
KNN	74,69	74,69		
Decision Tree	75,79	75,79	75,79	
MLP		75,99	75,99	75,99
SVM			78,81	78,81
Random Forest				78,86
Sig.	0,056	0,802	0,051	0,073

Fonte: Própria (2019).

4.4.1.2 Base da UFS

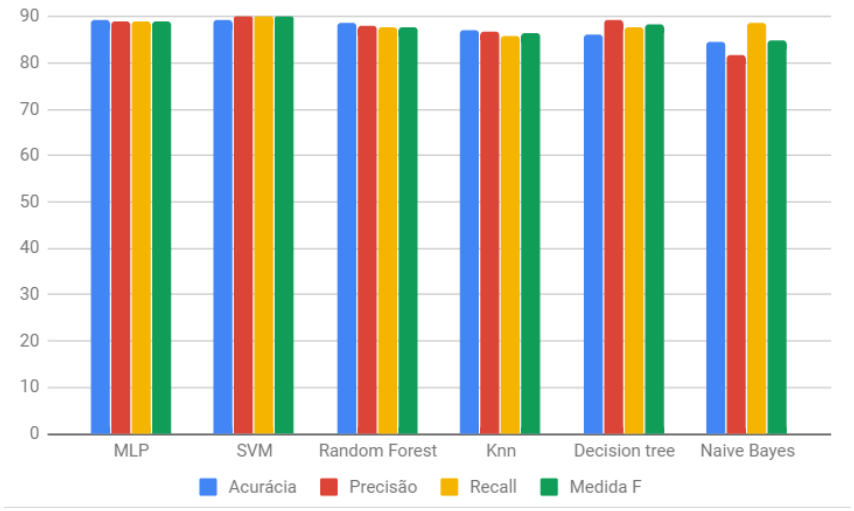
Na Tabela 14 e na Figura 6, são apresentadas as médias das métricas obtidas por cada algoritmo com todos atributos.

Tabela 14 – Comparativo das métricas dos algoritmos com todos atributos - Experimento 2 (UFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
MLP	89,30%	88,91%	89,04%	88,96%
SVM	89,26%	90,33%	90,07%	90,19%
Random Forest	88,62%	87,94%	87,70%	87,81%
Knn	86,97%	86,71%	85,82%	86,25%
Decision tree	86,17%	89,13%	87,64%	88,37%
Naive Bayes	84,57%	81,70%	88,75%	84,80%

Fonte: Própria (2019).

Figura 6 – Gráfico comparativo das métricas dos algoritmos com todos atributos - Experimento 2 (UFS).



Fonte: Própria (2019).

Utilizando o algoritmo *Random Forest* para seleção de atributos, foi possível notar que alguns tem mais relevância e outros podem ser eliminados sem ter forte influencia nos resultados, e em alguns casos, influenciando positivamente nos resultados. Os quatro atributos de maior relevância partindo dos mais relevantes foram: "qtd_rep_med_p_0"(quantidade média de disciplinas reprovadas por período), "media_geral_0:5"(média geral do discente no curso), "est_civil"(estado civil) e "raca"(etnia do discente). Os atributos que apresentaram menor relevância foram "qtd_rep_med_p_8:10"(número médio de reprovações por período), "qtd_ap_med_p_11:12"(quantidade média de disciplinas aprovadas por período) e "qtd_rep_1p_9:10"(quantidade de disciplinas reprovadas no primeiro período).

Isso mostra, por exemplo, que a quantidade de reprovações e a média entre 0 e 5 influenciam no problema e podem estar relacionados com as responsabilidades eles tem além dos estudos. Também traz um alerta para posterior análise das áreas sociais da instituição, considerando o estado civil e a etnia do discente.

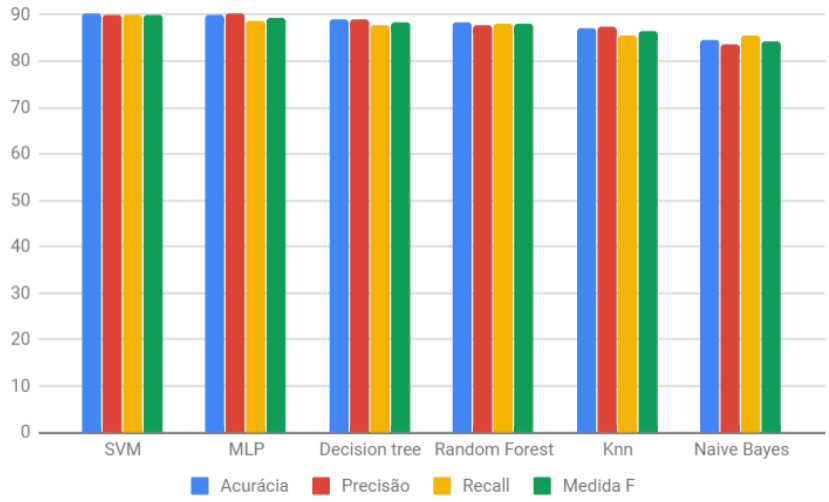
Em seguida foram removidos os atributos menos relevantes, e os algoritmos foram executados novamente. As médias de cada algoritmo são apresentadas na Tabela 15 e na Figura 7 abaixo.

Tabela 15 – Comparativo das métricas dos algoritmos com a seleção de atributos - Experimento 2 (UFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida F
SVM	90,43	90,01	90,06	90,03
MLP	89,90	90,20	88,62	89,39
Decision tree	88,92	88,96	87,82	88,38
Random Forest	88,48	87,88	88,18	88,02
Knn	87,09	87,38	85,47	86,41
Naive Bayes	84,69	83,50	85,67	84,36

Fonte: Própria (2019).

Figura 7 – Gráfico comparativo das métricas dos algoritmos com a seleção de atributos - Experimento 2 (UFS).



Fonte: Própria (2019).

Estes resultados foram utilizados para responder a questão de pesquisa. Como é perceptível, os algoritmos obtiveram médias de acurácias distintas e o algoritmo SVM obteve a maior média, seguido pela MLP, que conseguiu média bastante semelhante e próxima ao SVM. Todavia, não é possível fazer essas afirmações sem evidências estatísticas suficientemente conclusivas.

Como já mencionado anteriormente, o Teste Anova foi aplicado para validar a hipótese, e, por possuir os pressupostos da normalidade e da homocedasticidade, primeiramente foi feito o Teste Shapiro-Wilk e em seguida o Teste de Levene.

Foi definido um nível de significância de 0,05 em todo o experimento. Ao aplicar o Teste de Shapiro-Wilk, para análise da normalidade da distribuição dos dados, foram obtidos os

p-values apresentados na Tabela 16, na qual, observa-se valores acima do nível de significância adotado, concluindo-se que as distribuições são normais.

Tabela 16 – Resultado do Teste de Shapiro-Wilk, para análise da normalidade dos dados - Experimento 2 (UFS).

Algoritmo	Acurácia	Precisão	*Recall*	Medida-F
Knn	0,004	0,006	0,209	0,008
Random Forest	0,11	0,007	0,944	0,016
Naive Bayes	0,105	0,003	0,090	0,492
SVM	0,015	0,007	0,814	0,023
Decision tree	0,005	0,009	0,068	0,005
MLP	0,009	0,013	0,576	0,012

Fonte: Própria (2019).

Em seguida, foi feito o Teste de Levene, obtendo-se os resultados apresentados na Tabela 4, como pode ser observado, os *p-values* obtidos são maiores que o nível de significância adotado, validando o pressuposto da homogeneidade de variâncias.

Uma vez que os pressupostos foram atendidos, foi possível aplicar o Teste Anova, em que, verificou-se *p-values* fortemente menores que o nível de significância adotado, como pode ser observado na tabela 17. Desta forma, confirmou-se a evidência de diferença entre as médias, ou seja, a hipótese (H_0), de que os algoritmos possuem a mesma acurácia foi rejeitada, dentro do contexto do experimento realizado.

Tabela 17 – *P-values* dos Testes de Levene e Anova - Experimento 2 (UFS).

Métricas	Levene	Anova
Acurácia	0,032	$9,10^{-10}$
Precisão	0,005	$6,19^{-7}$
Recall	0,001	$6,72^{-11}$
Medida-F	0,116	$1,87^{-12}$

Fonte: Própria (2019).

Os resultados do Teste Tukey são apresentados na Tabela 18, a seguir, em que é possível observar três grupos homogêneos. Da análise dessa tabela, percebe-se que a maior média foi a do algoritmo SVM, com média de 90,43%. Porém, considerando o nível de significância de 5%, pode-se observar que os algoritmos SVM, MLP e *Decision Tree*, apresentaram uma mesma média de acurácia. O *Naive Bayes* apresentou a média mais baixa, com 84,69%.

Tabela 18 – Valores obtidos pelo Teste de Tukey para Acurácia - Experimento 2 (UFS).

Algoritmo	Subconjunto para alfa = 0.05		
	1	2	3
Naive Bayes	84,69		
KNN	87,09	87,09	
Random Forest		88,48	
Decision Tree		88,92	88,92
MLP			89,90
SVM			90,43
Sig.	0,074	0,289	0,228

Fonte: Própria (2019).

De forma análoga, a Tabela 19 apresenta o resultado do teste de Tukey da Medida-F. Para evitar a repetição, o teste Tukey da precisão e *recall* será omitido, uma vez que a Medida-F faz a combinação dessas métricas. Os resultados mostram que os algoritmos SVM, MLP e *Decision Tree* apresentaram médias semelhantes para a Medida-F, sendo 90,03%, 89,39% e 88,38%, respectivamente. *Naive Bayes* teve a menor média com 84,36%.

Tabela 19 – Valores obtidos pelo Teste de Tukey para a Medida-F - Experimento 2 (UFS).

Algoritmo	Subconjunto para alfa = 0.05		
	1	2	3
Naive Bayes	84,36		
KNN	86,41	86,41	
Random Forest		88,02	
Decision Tree		88,38	88,38
MLP			89,39
SVM			90,03
Sig.	0,133	0,158	0,148

Fonte: Própria (2019).

4.4.2 Meta-análise

Como o experimento apresentou resultados diferentes entre as bases, foi feita uma meta-análise para agregar os resultados estatísticos, ou seja, o resultado esperado é um valor que represente a força da associação entre a variável estudada, gerando uma resposta padrão generalizável (BREI; VIEIRA; MATOS, 2014).

Desta forma, o resultado dos algoritmos que se destacaram nas duas bases foram agregados e o teste de TuKey executado novamente. O resultado para acurácia é apresentado na Tabela 20, em que é possível perceber que o SVM continuou com a maior média, porém, estatistica-

mente semelhante com o MLP e *Random Forest*, apresentando as seguintes médias: 85,38%, 84,40% e 84,13%, respectivamente.

Tabela 20 – Valores obtidos pelo Teste de Tukey para Acurácia após a meta-análise.

Algoritmo	Subconjunto para alfa = 0.05	
	1	2
Decision Tree	83,15	
Random Forest		84,13
MLP		84,40
SVM		85,38
Sig.	0,173	0,599

Fonte: Própria (2019).

A Tabela 21 apresenta o resultado do teste de Tukey da Medida-F após a meta-análise. Os resultados mostram que os algoritmos SVM e MLP destacaram-se, apresentando médias semelhantes, sendo 84,42% e 83,44%, respectivamente.

Tabela 21 – Valores obtidos pelo Teste de Tukey para Medida-F após a meta-análise.

Algoritmo	Subconjunto para alfa = 0.05	
	1	2
Decision Tree	82,09	
Random Forest	82,69	
MLP		83,44
SVM		84,42
Sig.	0,254	0,648

Fonte: Própria (2019).

4.5 Ameaças à Validade

Embora os resultados do experimento tenham se mostrado satisfatórios, o mesmo apresenta ameaças à sua validade que devem ser comentadas.

Ameaças à validade interna: O sistema acadêmico atual do IFS está presente na Instituição desde 2017, o qual herdou a base do sistema acadêmico legado, com várias informações inconsistentes, principalmente até meados de 2007. Esta ameaça foi mitigada com a realização do processo de limpeza dos dados, diminuindo a probabilidade do uso de informações mais antigas incorretas.

Ameaças à validade de construção: No experimento desta dissertação, as instituições não possuíam algumas informações bastante relevantes em relação à evasão, como, por exemplo, dados socioeconômicos dos alunos e seus familiares. A inclusão dessas informações pode influenciar nos resultados do desempenho dos algoritmos, aumentando as eficácias destes. Para mitigar esta ameaça e incrementar o Sistema de Apoio à Decisão a ser desenvolvido, estas informações serão sugeridas para captação por parte da instituição e levadas em consideração em trabalhos futuros.

Tabela 7 – Parâmetros utilizados por algoritmo (Python).

Algoritmo	KNN	Random	Naive Bayes	SVM	Decision Tree	MLP
n_neighbors	25	-	-	-	-	-
random_state	-	0	-	-	0	-
criterion	-	entropy -	-	-	gini	-
n_estimators	-	75 -	-	-	-	-
max_depth	-	10 -	-	-	None	-
n_jobs	-	-1 -	-	-	-	-
max_features	-	0.3 -	-	-	None	-
bootstrap	-	-	-	-	-	-
C	-	-	-	0.001	-	-
cache_size	-	-	-	200	-	-
class_weight	-	-	-	None	None	-
max_iter	-	-	-	-1	-	200
probability	-	-	-	False	-	-
random_state	-	-	-	None	-	1
shrinking	-	-	-	True	-	-
tol	-	-	-	0.001	-	0.0001
verbose	-	-	-	False	-	False
coef0	-	-	-	0.0	-	-
decision_function_shape	-	-	-	ovr	-	-
degree	-	-	-	3	-	-
gamma	-	-	-	1	-	-
kernel	-	-	-	poly	-	-
max_leaf_nodes	-	-	-	-	None	-
min_impurity_decrease	-	-	-	-	0.0	-
min_impurity_split	-	-	-	-	None	-
min_samples_leaf	-	-	-	-	1	-
min_samples_split	-	-	-	-	39	-
min_weight_fraction_leaf	-	-	-	-	0.0	-
presort	-	-	-	-	False	-
splitter	-	-	-	-	best	-
activation	-	-	-	-	-	logistic
alpha	-	-	1	-	-	1e-05
batch_size	-	-	auto	-	-	auto
beta_1	-	-	-	-	-	0.9
beta_2	-	-	-	-	-	0.999
early_stopping	-	-	-	-	-	False
epsilon	-	-	-	-	-	1e-08
hidden_layer_sizes	-	-	-	-	-	(3, 2)
learning_rate	-	-	-	-	-	constant
learning_rate_init	-	-	-	-	-	0.002
momentum	-	-	-	-	-	0.9
shuffle	-	-	-	-	-	True
solver	-	-	-	-	-	adam

Fonte: Própria (2019).

5
Estudo de Caso

Este capítulo apresenta o estudo de caso realizado na Universidade Federal de Sergipe - UFS. A escolha se deu por conveniência, pelo acesso do pesquisador à instituição e consequente acesso aos dados necessários para o estudo de caso. Conforme explicitado na Seção 1.2, na UFS, a taxa de evasão nos cursos de graduação em 2014 foi de 13,4%. Neste contexto, a instituição possui como objetivo e meta estratégica para o período de 2016 a 2020 reduzir a taxa para 3% até 2020.

5.1 Etapas e diretrizes para o Estudo de Caso

As principais etapas para a realização do estudo de caso foram as seguintes:

- Definição do objetivo – Definir os objetivos do estudo de caso;
- Planejamento – Realizar planos do estudo de caso;
- Operação do Estudo de caso – Definir a preparação e execução do estudo de caso;
- Consolidação e resultados.

O detalhamento de cada uma dessas etapas será realizado nas próximas seções.

5.1.1 Definição do objetivo

O objetivo deste estudo é aplicar o algoritmo que, nos experimentos sobre predição de evasão, obteve a melhor eficácia, fornecendo uma interface para apoiar a definição de políticas institucionais e a adoção de ações administrativas e pedagógicas que contribuam para o atendimento de um dos itens do planejamento estratégico da UFS: a redução da taxa de evasão.

5.1.2 Planejamento

O planejamento do estudo de caso iniciou em fevereiro de 2018 e finalizou em julho de 2019, tendo a participação dos Gestores da Superintendência de Tecnologia da Informação (STI), Pró-Reitor de Graduação e representantes dos Departamento de Psicologia, Estatística e Computação.

Inicialmente, foram feitas reuniões semanais. Em seguida, por motivos de mudança de posto de trabalho do autor desta dissertação, a comunicação foi mantida prioritariamente por meio de E-mail.

5.1.2.1 Seleção de participantes

Devido ao conhecimento mais aprofundado dos pesquisadores deste estudo, sobre os alunos do Curso de Sistemas de Informação - Campus de Itabaiana, o qual que permite uma avaliação qualitativa e etnográfica dos resultados, foi solicitado ao Chefe do Departamento, que o estudo de caso fosse feito com os dados dos alunos do referido curso. Sendo assim, foram selecionados 373 alunos ativos, com ingresso até o ano/período 2018/2.

5.1.2.2 Instrumentação

Para o processo de classificação, foi utilizada a linguagem Python e suas bibliotecas. Os critérios de decisão foram mencionados na subseção 4.2.7.

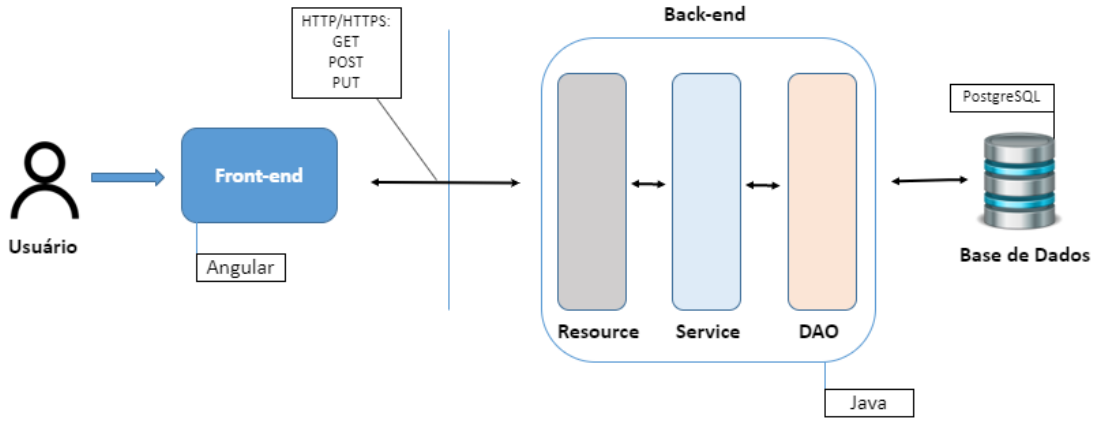
Os dados utilizados para a análise provêm do SIGAA, o qual possui como SGBD o *PostgreSQL*. Foi criado um ETL (*Extract, Transform and Load*), para extrair, transformar e limpar os dados, sendo estes utilizados na aplicação do modelo selecionado, definido após o processo experimental apresentado nos capítulos 3 e 4.

Foi criada uma interface para validar o modelo selecionado. Para o *front-end*, foi utilizada a plataforma e framework Angular 6, adotados e mantidos pelo Google. O Angular 6 é um *framework open-source* que possibilita gerar produtos interativos para multiplataformas (GOOGLE, 2019). Para o *back-end*, foi utilizada a linguagem java, apoiada em sua API, acrônimo para *Application Programming Interface* (Interface de Programação de Aplicações), a qual proporciona a integração entre sistemas que possuem linguagem totalmente distintas. Esta linguagem também foi selecionada por conveniência, pela familiaridade do pesquisador, além de ser uma tecnologia com comunidade global e atuante. A representação da arquitetura é apresentado na Figura 8.

5.2 Operação do Estudo de Caso

Conforme é apresentado na Figura 9, para o gestor ter acesso às informações dos discentes com risco de evasão, deverá selecionar um arquivo do tipo .csv, na tela principal da interface

Figura 8 – Representação arquitetural da aplicação.



Fonte: Própria (2019).

criada. Este arquivo será gerado através de uma consulta personalizada, disponível no SIGAA, em uma área restrita aos gestores, contendo os dados dos discentes, discutidos na subseção 4.3.2.

Figura 9 – Interface criada para aplicar o modelo proposto.

Fonte: Própria (2019).

Em seguida, a aplicação executa um pré-processamento, para transformar os dados de acordo com modelo proposto, conforme foi apresentado no capítulo 4. Alguns atributos passam a adotar o padrão *One Hot*, bem como é feita a remoção dos atributos com baixa relevância. Após este processo, a aplicação efetua a classificação dos registros, utilizando a base descrita na Seção 4.2.3.2 e o algoritmo SVM, selecionado após o processo experimental apresentado nos capítulos anteriores.

Durante toda a descrição dos resultados deste estudo, algumas informações estarão ocultas, para preservar a identidade dos envolvidos. Estes resultados serão apresentados na próxima Seção.

5.3 Resultados

Conforme descrito na Seção 5.1.2.1, foram selecionados todos discentes ativos, com ingresso até o período informado, do DSI - Campus Itabaiana, totalizando 373 alunos. O resultado após o processo classificatório é apresentado na Figura 10.

Figura 10 – Risco de evasão dos discentes do curso de Sistemas de Informação - Campus Itabaiana.

Nome	Curso	Ano/período	Risco de Evasão
[illegible]	SISTEMAS DE INFORMAÇÃO	2018/2	97.34
[illegible]	SISTEMAS DE INFORMAÇÃO	2018/2	97.3
[illegible]	SISTEMAS DE INFORMAÇÃO	2014/1	97.01
[illegible]	SISTEMAS DE INFORMAÇÃO	2016/2	96.89
[illegible]	SISTEMAS DE INFORMAÇÃO	2018/2	96.03
[illegible]	SISTEMAS DE INFORMAÇÃO	2018/2	95.49
[illegible]	SISTEMAS DE INFORMAÇÃO	2017/1	95.28
[illegible]	SISTEMAS DE INFORMAÇÃO	2016/1	94.36
[illegible]	SISTEMAS DE INFORMAÇÃO	2018/1	93.45
[illegible]	SISTEMAS DE INFORMAÇÃO	2016/1	93.43

Fonte: Própria (2019).

Conforme é mencionado na Seção 2.3, os métodos supervisionados necessitam de um conjunto de dados que possuem uma variável alvo principal pré-definida e os registros são categorizados em relação a esta, sendo assim, no caso deste trabalho, os registros seriam classificados em "Risco de evadir"e "Sem risco de evadir".

Porém, como é possível observar na Figura 10, os discentes foram separados em 3 grupos: Risco Baixo, Risco Médio e Risco Alto. Ou seja, foram destacados os discentes que estão mais propícios a evadir. Para esta classificação, foram utilizados os seguintes limiares (importante destacar que os limiares são parametrizados, podendo ser ajustados na interface):

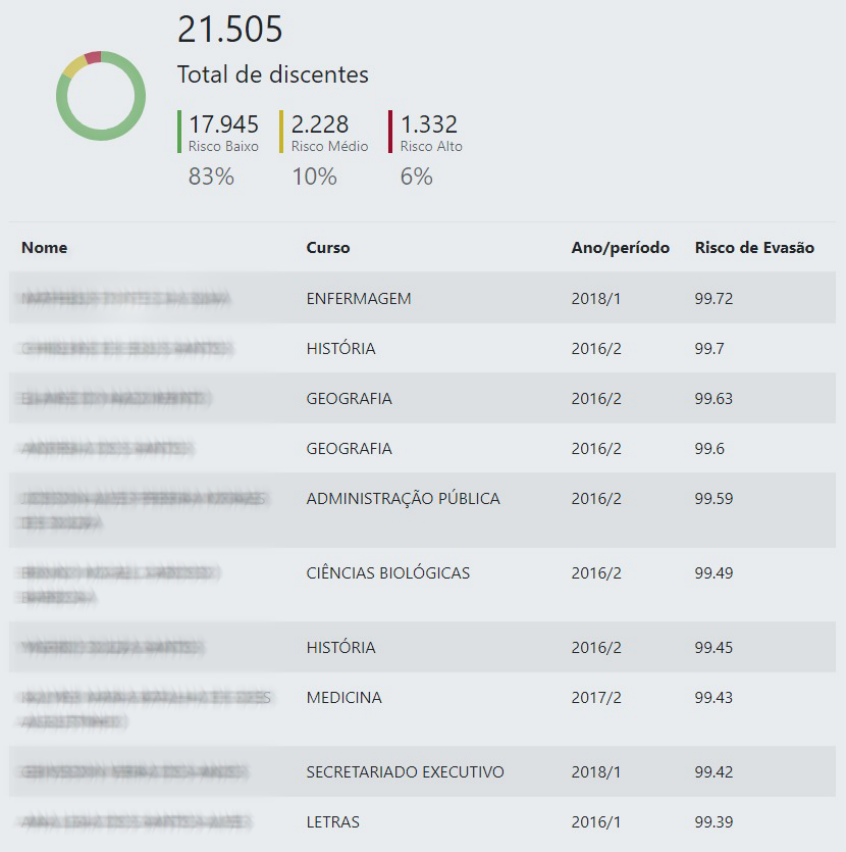
- Risco Baixo: Probabilidade abaixo de 50%;
- Risco Médio: Probabilidade entre de 50% e 85%;

- Risco Alto: Probabilidade acima 85%;

Com isto, dos 373 discentes do curso em questão, 34 (9%) destacaram-se com alto risco de evadir, e 49 (13%) apresentaram risco médio. Importante salientar que os discentes que estão acima do nível baixo (50%), estão propensos a evadir, totalizando 83 (22%) discentes. Sendo esta divisão em grupos para destacar os que estão com risco mais alto, exigindo assim, mais atenção.

A Figura 11 a seguir, apresenta o resultado com todos os discentes de graduação ativos da UFS. É possível observar que 1.332 (6%) se destacaram com alto risco de evadir, e 2.228 (10%) apresentaram risco médio.

Figura 11 – Risco de evasão dos discentes da UFS.



Nome	Curso	Ano/período	Risco de Evasão
[redacted]	ENFERMAGEM	2018/1	99.72
[redacted]	HISTÓRIA	2016/2	99.7
[redacted]	GEOGRAFIA	2016/2	99.63
[redacted]	GEOGRAFIA	2016/2	99.6
[redacted]	ADMINISTRAÇÃO PÚBLICA	2016/2	99.59
[redacted]	CIÊNCIAS BIOLÓGICAS	2016/2	99.49
[redacted]	HISTÓRIA	2016/2	99.45
[redacted]	MEDICINA	2017/2	99.43
[redacted]	SECRETARIADO EXECUTIVO	2018/1	99.42
[redacted]	LETRAS	2016/1	99.39

Fonte: Própria (2019).

O relatório permite o gerenciamento por exceções e a identificação de possíveis alunos com risco de evasão, permitindo ações administrativas e pedagógicas mais acertivas que contribuam para a mitigação deste problema.

5.3.1 Ameaças à validade

Embora os resultados do estudo de caso tenham se mostrado satisfatórios, o mesmo apresenta ameaças à sua validade que devem ser comentadas.

Ameaça à validade de conclusão e externa: O estudo de caso foi feito com apenas um curso, limitando as opiniões sobre a interface, como também impedindo a avaliação de significância para a UFS e outras universidades.

6
Conclusão

Esta dissertação expôs a necessidade de aumentar a adoção de medidas para identificar e entender os principais fatores que podem contribuir com o insucesso dos estudantes.

O trabalho foi consolidado com a realização de duas análises experimentais dos algoritmos de mineração de dados mais utilizados na área de educação, avaliando o que melhor se adequa ao contexto de abandono do ensino em duas instituições federais. Com o propósito de aplicar o modelo selecionado, foi realizado um Estudo de Caso na UFS, para atender a um dos itens de seu planejamento estratégico, o qual gerou uma interface que auxiliará o processo de apoio à decisão e à mitigação da evasão escolar.

6.1 Principais Conclusões

Dentre as principais contribuições deste trabalho, destacam-se:

- A avaliação experimental de seis algoritmos mais utilizados, no contexto da EDM, em termos de acurácia, precisão, *recall* e medida-F, identificando fatores influenciadores da evasão escolar. O primeiro experimento foi submetido ao SBBD (Simpósio Brasileiro de Banco de Dados) 2019, qualis B2, sob o título "Análise Comparativa de Algoritmos de Mineração de Dados Aplicados ao Contexto da Evasão Escolar".
- Os resultados do segundo experimento, com a base do IFS, foram aprovados na conferência internacional FedCSIS (Federated Conference on Computer Science and Information Systems) 2019, qualis B1, sob o título *"Comparative Analysis of Data Mining Algorithms Applied to the Context of School Dropout"*.
- Estudo de Caso na UFS, para atender a uma parte do planejamento estratégico, o qual gerou uma interface que auxiliará o processo de apoio à decisão.

Como consequência destas contribuições, foi possível responder à questão principal de pesquisa, elaborada no início do trabalho:

No contexto da Evasão Escolar de uma instituição de ensino superior, entre os algoritmos selecionados na área de EDM, qual o melhor em termos de Eficácia?

O SVM destacou-se, obtendo a maior média da métricas de eficácia, porém, estatisticamente, os algoritmos MLP e Random Forest obtiveram resultados semelhantes.

No primeiro experimento, utilizando a ferramenta *Weka* com a base do IFS, os resultados mostraram que existem diferenças significativas entre os algoritmos utilizados, e que, apesar do *Random Forest* possuir a maior média de eficácia em termos de acurácia, verificou-se, estatisticamente, que possui similaridade com os algoritmos SMO e J48. Portanto, foi evidenciado que apresentam médias similares de acurácia (79,75%, 78,89% e 78,50%), precisão (78,86%, 78,65% e 77,15%), *recall* (77,95%, 75,91% e 76,43%) e medida-F (78,37%, 77,18% e 77,01%), na predição de alunos com insucesso acadêmico.

No segundo experimento, utilizando a linguagem Python e suas bibliotecas, nas bases do IFS e da UFS, os resultados evidenciaram que o SVM possui a maior média de eficácia em algumas métricas, em ambas as bases, mas que, estatisticamente, possui similaridade com outros algoritmos. Na base do IFS, o SVM apresentou similaridade com o *Random Forest*, apresentando médias similares de acurácia (81,18% e 80,36%), precisão (80,25%, 80,79), *recall* (77,51%, 76,50) e medida-F (78,51%, 78,86%), na predição de alunos com insucesso acadêmico.

Na base da UFS, o SVM apresentou a maior média de eficácia em quase todas métricas, porém, estatisticamente, os algoritmos MLP e *Decision Tree* obtiveram um resultado semelhante: acurácia (90,43%, 89,90% e 88,42%), precisão (90,01%, 90,20% e 88,96%), *recall* (90,06%, 88,62% e 87,82%) e medida-F (90,03%, 89,39% e 88,38%), respectivamente.

Após a metanálise dos experimentos, apesar do SVM possuir a maior média das métricas de eficácia, estatisticamente os algoritmos MLP e *Random Forest*, respectivamente, obtiveram resultados semelhantes de acurácia (85,38%, 84,40% e 84,13%). Para medida-F, a significância estatística foi igual apenas para o MLP (84,42%, 83,44%).

Após as duas análises experimentais, o SVM, pelo seu pequeno destaque, selecionado para ser aplicado em um estudo de caso de atendimento a um dos itens do planejamento estratégico da UFS: a redução da taxa de evasão. Após o processo classificatório na interface criada, dos 373 discentes do curso de Sistemas de Informação - Campus Itabaiana, 34 (9%) destacaram-se com alto risco de evadir, e 49 (13%) apresentaram risco médio.

Por conseguinte, com estes resultados é possível identificar possíveis alunos com risco de evasão, permitindo ações administrativas e pedagógicas mais acertivas que contribuam para a mitigação deste problema.

Os trabalhos publicados sobre o tema possuem algumas lacunas científicas, se conside-

rarmos que não houve validação rigorosa dos resultados, o que permitiria uma combinação mais assertiva de evidências experimentais. Neste contexto, este trabalho valida seus resultados e confirma algumas evidências anteriores encontradas nos trabalhos descritos em (MáRQUEZ-VERA et al., 2016) e (GOPALAKRISHNAN et al., 2017).

6.2 Limitações e Perspectivas

Pretende-se realizar a análise dos algoritmos em outros níveis de ensino, bem como adicionar outros tipos de variáveis para a análise, como, por exemplo, informações socioeconômicas. A inclusão destas informações pode influenciar nos resultados do desempenho dos algoritmos, aumentando as eficácias destes.

Essas informações serão sugeridas para captação por parte da instituição e levadas em consideração em trabalhos futuros, como, por exemplo, no aperfeiçoamento do sistema preditivo para a gestão de ensino, melhorando o auxílio nas tomadas de decisões de combate à evasão e à retenção escolar.

O estudo de caso foi feito com apenas um curso, limitando as opiniões sobre a interface, como também impedindo a avaliação de significância para a UFS e outras universidades. Sendo assim, pretende-se fazer entrevistas com Chefes de Departamento de outras áreas.

Destaque-se que os resultados integrados desta dissertação serão submetidos ao periódico *Informatics in Education*, ISSN 2335-8971, qualis A2, em razão de seu escopo estar diretamente relacionado ao presente trabalho.

Referências

ANJOS, A. Análise de variância. *Universidade Federal do Paraná, Departamento de Estatística - UFPR, Curitiba*, http://www.est.ufpr.br/ce003/material/apostilace003.pdf, p. Capítulo 7, 2009. Citado 2 vezes nas páginas 35 e 43.

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 14.

BAKER, R. et al. Data mining for education. *International encyclopedia of education*, Oxford, UK: Elsevier, v. 7, n. 3, p. 112–118, 2010. Citado na página 14.

BARBIERI, C. *Business Intelligence: modelagem e qualidade*. [S.l.]: Information and management [0378-7206], 2011. 392 p. Citado na página 21.

BASILI, V. R.; WEISS, D. M. *A methodology for collecting valid software engineering data.* [S.l.], 1983. Citado 3 vezes nas páginas 25, 31 e 39.

BREI, V. A.; VIEIRA, V. A.; MATOS, C. A. Meta-análise em marketing. *Revista Brasileira de Marketing e-ISSN: 2177-5184*, 2014. Citado na página 51.

BREIMAN, L. Machine learning. *Kluwer Academic Publishers*, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 18.

BREIMAN, L. Random forests. machine learning. p. 5–32, 2001. Citado na página 23.

BREIMAN, L. et al. *Classication and regression trees*. [S.l.]: Monterey, CA: Wadsworth and Brooks, 1984. Citado na página 18.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Instituto de Informática Universidade Federal de Goiás - Relatório Técnico, 2009. Citado 3 vezes nas páginas 22, 23 e 24.

CAMPBELL, B. Alignment: Resolving ambiguity with bounded choice. PACIS 2005 Proceedings, 2005. Citado na página 20.

CAMPOS, G. O. et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, v. 30, n. 4, p. 891–927, Jul 2016. Citado na página 18.

CHAN, Y. E. e. a. Business strategic orientation, information system strategic orientation, and strategic alignment. Information Systems Research, v. 8, p. 125–150, 1997. Citado na página 20.

CHEN, H.; CHIANG, R. H. L.; STOREY, V. C. Business intelligence and analyics: From big data o big impact. MIS Quarterly, 2012. Citado na página 22.

CHI-AN, C.; CHANDRA, A. Impact of owner's knowledge of information technology (it) on strategic alignment and it adoption in us small firms. *Journal of Small Business and Enterprise Development*, v. 19, p. 114–131, 2012. Citado na página 21.

COOK, D. Practical machine learning with h2o - powerful, scalable techniques for ai and deep learning. 1st ed. Sebastopol, CA: O'Reilly Media, 2017. Citado na página 41.

CORTES, C.; VAPNIK, V. . Support-vector networks. machine learning, 20. p. 273–297, 1995. Citado na página 18.

COSTA, E. et al. Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicaçõe. ornada de Atualização em Informática na Educação, p. 1–29, 2012. Citado 2 vezes nas páginas 14 e 23.

DAVENPORT, T. H.; PATIL, D. J. Data scientist.harvard business review. p. 70–76, 2012. Citado na página 22.

DEKKER, P. M.; J., V. Predicting students drop out: A case study. *Proceedings of the International Conference on Educational Data Mining*, p. 41–50, 2009. Citado na página 37.

DELOITTE. Survey 2016-2017 trajetória entre perfis: No rumo da geração de valor ao negócio. 2017. Disponível em: <https://www2.deloitte.com/br/pt/pages/technology/articles/cio-survey.html>. Citado na página 17.

DIETTERICH, T. G. Ensemble methods in machine learning. Springer Berlin Heidelberg, p. 1–15, 2000. Citado na página 23.

DUAN, L.; XU, L. D. Business intelligence for enterprise systems: a survey. *Advances in knowledge discovery and data mining*, IEEE Transactions on Industrial Informatics, v. 8, p. 679–687, 2012. Citado na página 22.

FERNANDES, E. R. et al. O uso do sistema de informação contábil como ferramenta para a tomada de decisão nas empresas da região de contagem. SEGET, 2012. Citado na página 16.

FIELD, A. *Descobrindo a estatística usando o SPSS*. [S.l.]: 2.ed. Porto Alegre: Artned, 2009. Citado 2 vezes nas páginas 34 e 43.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007. Citado 2 vezes nas páginas 13 e 23.

FRANK, E.; BOUCKAERT, R. R. Naive bayes for text classification with unbalanced classes. knowledge discovery in databases: Pkdd 2006. Lecture Notes in Computer Science, 4213, 503-510, 2006. Citado na página 24.

FROTA, L. C. M. Inteligência nas organizações públicas de saúde: soluções e informações estratégicas para gestão. Dissertação (Mestrado em Saúde) – Escola Nacional de Saúde Pública Sergio Arouca, 2009. Citado na página 15.

GOOGLE. Angular. Disponível em: <https://angular.io/>, 2019. Citado na página 56.

GOPALAKRISHNAN et al. A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. p. 372–381, July 2017. Citado 4 vezes nas páginas 16, 26, 29 e 63.

HALL EIBE FRANK, G. H. B. P. P. R. M.; WITTEN, I. H. The weka data mining software: An update. *SIGKDD Explorations*, v. 10, n. 1, 2009. Citado 2 vezes nas páginas 30 e 34.

HAMALAINEN W., V. M. Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman and Hall/CRC Data Mining and Knowledge*, Discovery Series, p. 57–71, 2011. Citado na página 24.

HAN, J.; KAMBER, M. Data mining concepts and techniques. Morgan Kaufmann Publishers, v. 3, 2012. Citado na página 24.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 25.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. [S.l.]: MIT press, 2001. Citado na página 25.

HAND H. M., S. P. D. Principles of data mining. *Bradford Book*, Cambridge, 2000. Citado na página 42.

HANS, R. T.; MNKANDLA. Modeling software engineering projects as a business: A business intelligence perspective. AFRICON, p. 5, 2013. Citado na página 14.

HARRIOTT, J. 7 pillars for successful analytics implementation. Marketing Insights, spring, 2013. Citado na página 22.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: Data mining, inference, and prediction. *Proceedings of the International Conference on Educational Data Mining*, Springer, 2001. Citado 2 vezes nas páginas 33 e 41.

IBM. Spss. *IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp*, 2017. Citado 3 vezes nas páginas 19, 35 e 43.

ISIK et al. Business intelligence success: The roles of bi capabilities and decision environments. Information and management [0378-7206], v. 50, p. 13–23, 2013. Citado na página 21.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, pp. 338-345, 1995. Citado 2 vezes nas páginas 18 e 24.

JÚNIOR, J. G. d. O. et al. *Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2015. Citado na página 18.

JUNIOR, M. C. Projetando sistema de apoio À decisão baseados em data warehouse. Axcel Books Do Brasil, 2004. Citado na página 14.

JURISTO, N.; MORENO, A. Software engineering experimentation. Kluwer Academic Press, 1ª edição, 2001. Citado na página 18.

KANTORSKI, G. et al. Predição da evasão em cursos de graduação em instituições públicas. In: *Brazilian Symposium on Computers in Education*. [S.l.: s.n.], 2016. v. 27, n. 1, p. 906. Citado 3 vezes nas páginas 16, 27 e 29.

KHAN, R.; QUADRI, S. M. K. Business intelligence: an integrated approach. business. Intelligence Journal, v. 5, p. 64–70, 2012. Citado na página 14.

KING, W. R. How effective is your is planning? long range planning. v. 21, p. 103–122, 1988. Citado na página 20.

LAUDON, K.; LAUDON, J. P. Sistemas de informação gerenciais. São Paulo: Pearson Education do Brasil, v. 11, 2014. Citado na página 22.

LEVENE, H. Robust tests for equality of variances. *International Journal of Machine Learning and Cybernetics*, Contributions to Probability and Statistics: essays in honor of harold hotelling. Stanford University, p. 278–292, 1960. Citado 2 vezes nas páginas 34 e 43.

LIMA, A.; JúNIOR, M.; NASCIMENTO, A. Um survey com empresas brasileiras acerca da utilização de business intelligence (bi) e um diagnóstico sobre a infraestrutura e metodologias associadas. 2017. Citado na página 15.

LIMA, A. S. d. Proposta e avaliação da combinação de uma metodologia Ágil e gqm+strategies para o desenvolvimento de aplicações de business intelligence dirigido à estratégia. Dissertação - Universidade Federal de Sergipe, 2017. Citado 2 vezes nas páginas 20 e 21.

LOBO, M. B. D. C. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. ABMES Cadernos, v. 1, p. 09–58, 2012. Citado na página 15.

MACHADO, E.; MARCELO, L. Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes. *XXVII Congresso da Sociedade Brasileira de Computação*, 2007. Disponível em: <https://www.cos.ufrj.br/~ines/enia07_html/pdf/28076.pdf>. Citado 2 vezes nas páginas 33 e 41.

MACHADO, R. D. et al. Estudo bibliométrico em mineração de dados e evasão escolar. *XI Congresso Nacional de Excelência em Gestão*, 2015. Citado na página 27.

MANHÃES, L. M. B. et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*, 2012. Citado 3 vezes nas páginas 16, 27 e 29.

MARCH, S. T.; HEVNER, A. R. Integrated decision support systems: A data warehousing perspective. Decision support systems [0167-9236], v. 43, p. 1031–1043, 2007. Citado na página 21.

MARQUEZ-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 18 e 37.

MCCUE, C. Data mining and predictive analysis - intelligence gathering and crime analysis. Elsevier, 2007. Citado na página 23.

MEC, B. Documento orientador para a superação da evasão e retenção na rede federal de educação profissional, científica e tecnológica. 2014. Citado na página 13.

MITCHELL, T. Machine learning. McGraw Hill, 1997. Citado na página 24.

MUNDSTOCK, E. Introdução à análise estatística utilizando o spss 13.0. cadernos de matemática e estatística série b. Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2006. Citado na página 19.

MÜNCH, J. e. a. The effects of gqm+strategies on organizational alignment. Proceedings of the DASMA Software Metric Congress, 2013. Citado na página 17.

MáRQUEZ-VERA, C. et al. Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, v. 33, n. 1, p. 107–124, 2016. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12135>. Citado 4 vezes nas páginas 16, 27, 28 e 63.

NüRNBERGER, A.; PEDRYCZ, W.; KRUSE, R. . *Handbook of data mining and knowledge discovery. Chapter data mining tasks and Methods: Classi- cation: Neural network approaches*. [S.l.]: New York, NY, USA: Oxford University Press, 2002. 304-317 p. Citado na página 18.

OLSZAK, C. M.; ZIEMBA, E. Critical success factors for implementing business intelligence systems in small and medium enterprises on the example of upper silesia, poland. Interdisciplinary Journal of Information, Knowledge, and Management, v. 7, 2012. Citado na página 14.

PASCOAL, T. A. B. et al. Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socieconômicos. *V Congresso Brasileiro de Informática na Educação (CBIE 2016)*, Anais do XXVII Simpósio Brasileiro de Informática na Educação, p. 926–935, 2016. Citado na página 16.

PEREIRA, R. B. SeleÇÃo lazy de atributos para a tarefa de classificaÇÃo. *Bradford Book*, UNIVERSIDADE FEDERAL FLUMINENSE - Niterói, 2009. Citado na página 42.

POLL, K. S. Analytics, data mining software used. Disponível em: <https://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html>, 2015. Citado na página 41.

REICH B. H. E BENBASAT, I. Measuring the linkage between business and information technology objectives. MIS Quarterly, v. 20, p. 55–81, 1996. Citado na página 20.

REZENDE, D.; ABREU, A. Tecnologia da informação: aplicada a sistemas de informação empresaria. São Paulo Atlas, v. 3, 2003. Citado na página 22.

ROMERO, C. et al. Data mining algorithms to classify students. EDM, p. 8–17, 2008. Citado na página 23.

SANTOS, J. S. d. Business intelligence : uma proposta metodológica para análise da evasão escolar em instituições federais de ensino. Dissertação - Universidade Federal do Paraná, 2017. Citado na página 15.

SEVERINO, A. J. *Metodologia do trabalho científico*. [S.l.]: Cortez editora, 2017. Citado na página 18.

SHAPIRO, S.; WILK, M. An analysis of variance test for normality (complete samples). *International Journal of Machine Learning and Cybernetics*, Biometrika, v. 52, p. 591–611, 1965. Citado 2 vezes nas páginas 34 e 43.

SHARDA, R.; DELEN, D.; TURBAN, E. Business intelligence and analytics. *Systems for decision support*, New Jersey: Pearson Education Inc., v. 10, p. 1031–1043, 2014. Citado na página 21.

SILVA, D. F.; BATISTA, G. E. de A. P. A. Uma comparação experimental de métodos de imputação de valores desconhecidos. *ICMC - Instituto de Ciências Matemáticas e de Computação, São Paulo*, 2009. Citado na página 18.

SILVA, R. A. d.; SILVA, F. C. A.; GOME, C. F. S. O uso do business intelligence (bi) em sistema de apoio à tomada de decisão estratégica. GEINTEC, 2016. Citado na página 15.

TEIXEIRA, I. Instituto Nacional de Estudos e P. E. A. Censo da educação superior 2015. 2015. Citado na página 13.

TURBAN, E.; VOLONIMO, L. *Business Intelligence e Suporte à Decisão. In: Tecnologia da Informação para Gestão: em busca do melhor desempenho estratégico e operacional*. [S.l.: s.n.], 2013. v. 8. 468 p. Citado na página 14.

VELCU, O. Strategic alignment of erp implementation stages: An empirical investigation. Information and Management, v. 47, p. 158–166, 2010. Citado na página 21.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 14.

WOHLIN, C. Experimentation in software engineering. Springer Science and Business Media, 2012. Citado na página 16.

WOHLIN, C. et al. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012. Citado 2 vezes nas páginas 18 e 30.