Universidade Federal de Pernambuco
Centro de Ciências Sociais Aplicadas
Departamento de Ciências Administrativas
Mestrado Profissional em Administração

Alexandre Magno Gurgel Fialho

Mineração de dados educacionais dos alunos de graduação da UFPE: um estudo de caso da mobilidade e internacionalização

RECIFE

2020

UNIVERSIDADE FEDERAL DE PERNAMBUCO
MESTRADO PROFISSIONAL EM ADMINISTRAÇÃO

CLASSIFICAÇÃO DE ACESSO A TESES E DISSERTAÇÕES

Considerando a natureza das informações e compromissos assumidos com suas fontes, o acesso a monografias do Programa é definido em três graus:

- "Grau 1": livre (sem prejuízo das referências ordinárias em citações diretas e indiretas);
- "Grau 2": com vedação a cópias, no todo ou em parte, sendo, em consequência, restrita à consulta em ambientes de biblioteca com saída controlada;
- "Grau 3": apenas com autorização expressa do autor, por escrito, devendo, por isso, o texto, se confiado a bibliotecas que assegurem a restrição, ser mantido em local sob chave ou custódia.

A classificação desta dissertação se encontra, abaixo, definida por seu autor.

Solicita-se aos depositários e usuários sua fiel observância, a fim de que se preservem as condições éticas e operacionais da pesquisa científica na área de administração.

Título da dissertação: Mineração de dados educacionais dos alunos de graduação da UFPE: um estudo de caso da mobilidade e internacionalização.

Nome do autor: Alexandre Magno Gurgel Fialho

Data da aprovação: 29/10/2020

Classificação, conforme especificação acima:

Grau 1 [X]

Grau 2 [ ]

Grau 3 [ ]

Recife, 26 de novembro de 2020:

Assinatura do autor

Alexandre Magno Gurgel Fialho

Mineração de dados educacionais dos alunos de graduação da UFPE: um estudo de caso da mobilidade e internacionalização

Dissertação de Mestrado apresentada como requisito complementar para obtenção do grau de Mestre em Administração, do Programa de Mestrado Profissional em Administração da Universidade Federal de Pernambuco.

Orientadora: Dra. Taciana Barros Jerônimo

Recife
2020

Catalogação na Fonte
Bibliotecária Ângela de Fátima Correia Simões, CRB4-773

F438m Fialho, Alexandre Magno Gurgel
Mineração de dados educacionais dos alunos de graduação da UFPE: um estudo de caso da mobilidade e internacionalização / Alexandre Magno Gurgel Fialho. - 2020.
106 folhas: il. 30 cm.

Orientadora: Prof.ª Dra. Taciana Barros Jerônimo.
Dissertação (Mestrado em Administração) – Universidade Federal de Pernambuco, CCSA, 2020.
Inclui referências e apêndices.

1. Mineração de dados. 2. Tomada de decisão. 3. Mobilidade internacional. I. Jerônimo, Taciana Barros (Orientadora). II. Título.

658 CDD (22. ed.) UFPE (CSA 2020 – 112)

Universidade Federal de Pernambuco
Centro de Ciências Sociais Aplicadas
Departamento de Ciências Administrativas
Mestrado Profissional em Administração

Mineração de dados educacionais dos alunos de graduação da UFPE: um estudo de caso da Mobilidade e internacionalização

Alexandre Magno Gurgel Fialho

Dissertação submetida ao corpo docente do Curso de Mestrado Profissional em Administração da Universidade Federal de Pernambuco e aprovada em 29 de outubro de 2020.

Banca Examinadora:

Profa. Dra. Taciana Barros Jerônimo, UFPE (Orientadora)

Prof. Dr. Fernando Gomes de Paiva Junior, UFPE (Examinadora interna)

Prof. Dr. Jadielson Alves de Moura, UFPE (Examinador Externo)

“Otimismo é esperar pelo melhor. Confiança é saber lidar com o pior.”.

(Roberto SIMOSEN, 1938)

Agradecimentos

A minha esposa, Roberta Macedo, pelo companheirismo, paciência e apoio diário.

A filha, Liz Macedo Gouveia Gurgel, por me dar alegrias diária de sua existência e se tornar mais um motivo de melhorar meu desenvolvimento profissional.

A minha mãe, Miracema Gurgel de Almeida, por me inspirar em ser uma pessoa boa e de caráter.

A professora Taciana Barros Jerônimo, pelas orientações, paciência, dedicação a este trabalho e me motivar nas horas certas.

Aos professores examinadores, Fernando Gomes de Paiva Junior e Jadielson Alves de Moura, por suas contribuições e incentivo em melhorar e concluir esta pesquisa.

Aos professores do Mestrado Profissional em Administração, pelos ensinamentos passados em suas disciplinas e suas experiências profissionais compartilhadas.

Aos amigos de trabalho do setor de Relações Internacionais e todos colegas de trabalhos de toda reitoria UFPE, pela colaboração, apoio e exemplos de vidas.

Todos vocês contribuíram com a execução desta dissertação e sem ela não conseguiria concluir. Minha humilde gratidão a todos.

A Deus, pelos momentos de iluminação para continuar em momentos difíceis.

Resumo

Esta dissertação de mestrado propôs um modelo de tomada de decisão utilizando uma adaptação do *framework CRISP-DM* e as fases do processo *KDD* da mineração de dados para tratar dados de alunos de graduação que participaram de programas de mobilidade internacional na UFPE. Bem como estudar sobre a internacionalização da instituição e contribuir com novas estratégias para futuros Planos de desenvolvimentos Institucionais (PDI). Baseou-se, principalmente, em referências sobre gestão da informação, processo decisório, mineração de dados e suas aplicações na administração pública. A pesquisa é um estudo de caso com natureza descritiva, o método de abordagem predominante quantitativa por meio de pesquisa bibliográfica, coleta de dados e realizado no setor de Diretoria de Relações Internacionais da UFPE. O trabalho é norteado em atender os objetivos específicos: (a) avaliar procedimento, técnicas e algoritmos para mineração de dados; (b) estruturar dados para mineração; (c) encontrar padrões e descobrir conhecimentos nos dados de mobilidade internacional; (d) realizar levantamento sobre ações de internacionalização; (e) realizar diagnóstico de gestão. Os dados coletados dos alunos que participaram de mobilidade foram registrados de maneira descentralizada, manual e consequentemente ocorrendo perdas de dados. Este estudo tratou tais dados usando ferramentas de mineração, fez junção destes dados com dados da instituição, demonstrou na prática o uso da adaptação do *framework CRISP-DM* e possibilita que outros setores ou instituições possam usar tal método. Esta pesquisa contribui de maneira inovadora, extraindo conhecimento de dados para as novas tomadas de decisões de instituições públicas e privadas, propondo aplicar metodologia em lugares que podem estar com dados descentralizados, incorrendo em “esconder” conhecimentos valiosos.

Palavras-chave: Mineração de dados. Tomada de decisão. Mobilidade internacional.

KDD. CRISP-DM.

Abstract

This master's thesis proposed a decision-making model using an adaptation of the CRISP-DM framework and the phases of the KDD process of data mining to treat data from undergraduate students who participated in international mobility programs at UFPE. As well as studying the institution's internationalization and contributing with new strategies for future Institutional Development Plans (PDI). It was mainly based on references on information management, decision making, data mining and its applications in public administration. The research is a case study with a descriptive nature, the method of predominant quantitative approach through bibliographic research, data collection and carried out in the sector of International Relations Directorate at UFPE. The work is guided in meeting the specific objectives: (a) to evaluate procedures, techniques and algorithms for data mining; (b) structuring data for mining; (c) find patterns and discover knowledge in international mobility data; (d) conduct a survey on internationalization actions; (e) perform management diagnosis. The data collected from students who participated in mobility were recorded in a decentralized, manual manner and, consequently, data loss occurred. This study treated such data using mining tools, merged these data with data from the institution, demonstrated in practice the use of the adaptation of the CRISP-DM framework and allows other sectors or institutions to use this method. This research contributes in an innovative way, extracting knowledge from data for new decision-making by public and private institutions, proposing to apply methodology in places that may have decentralized data, incurring in “hiding” valuable knowledge.

Keywords: Data mining. Decision making. International mobility. KDD. CRISP-DM.

Lista de Figuras

Lista de Quadros

Lista de Abreviaturas e Siglas

AAI - Assessorias para assuntos internacionais
AM - Aprendizado de máquina
AUF - Agence Universitaire de la Francophonie
AULP - Associação das Universidades de Língua Portuguesa
BI - Business Inteligence
CA - Centro do agreste
CAC - Centro de artes e comunicação
CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CCB - Centro de ciências biológicas
CCEN - Centro de ciências exatas e da natureza
CCJ - Centro de ciências jurídicas
CCS - Centro de ciências da saúde
CFCH - Centro de filosofia e ciências humanas
CIN - Centro de informática
CTG - Centro de tecnologias e geociências
CGU - Controladoria-Geral da União
CRISP-DM - *Cross Industry Standard Process for Data Mining*
CSF - Ciências sem fronteiras
DPU - Defensoria Pública da União
DRI - Diretoria de relações internacionais
ENADE - Exame Nacional de Desempenho dos Estudantes
ENAP - Escola Nacional da Administração Pública
FAUBAI - Associação Brasileira de Educação Internacional
GCUB - Grupo Coimbra de Universidades Brasileiras
IES - Instituições de ensino superior
IFES - Instituições Federais de Ensino Superior
INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IO - Inteligência organizacional
KDD - *Knowledge Discovery in Databases*
MEC - Ministério da Educação e Cultura
MDE - Mineração de dados educacionais
MPDG - Ministério do Planejamento, Desenvolvimento e Gestão
NTI - Núcleo de tecnologia de informação (Atual STI - Superintendência de tecnologia da informação)
NUCLI - Núcleo de Línguas – Idiomas sem fronteiras UFPE
OECD - Organização para a Cooperação e Desenvolvimento Econômico
OLAP - *Online Analytical Processing*
OUI - Organização Universitária Interamericana
PARU - Programa de avaliação de reforma universitária
PDI - Plano de desenvolvimento Institucional
PEI - Plano estratégico institucional UFPE - Universidade Federal de Pernambuco
PG&C - Perspectivas em Gestão & Conhecimento
PGFN - Procuradoria-Geral da Fazenda Nacional
RCE - Associação de Causa e Efeito
RMR - Região Metropolitana do Recife
SAMU - serviço de atendimento móvel de urgência
SERPRO - Serviço federal de processamento de dados

SPSS - *Statistical Package for the Social Sciences*
TCE - Tribunal de contas do estado
TCU - Tribunal de contas da união
TIC - Tenconologias de informação e comunicação
UDUAL - Unión de Universidades de América Latina y el Caribe
UFPE - Universidade Federal de Pernambuco

Sumário

1 Introdução

As Diretorias de Relações Internacionais (DRI) ou Assessorias para Assuntos Internacionais (AAI) fazem parte das instituições públicas de ensino superior no Brasil, sendo responsáveis por oferecer e gerenciar programas de intercâmbio acadêmicos. Hoje a internacionalização das universidades é uma realidade e a tendência de crescimento de estudantes interessados em realizar intercâmbio acadêmico, devido à oferta de bolsas em instituições estrangeiras por meio de programas, como por exemplo, durante o programa Ciência Sem Fronteiras (CAPES, 2011) que ofertou mais de 100 mil bolsas para estudantes de ensino superior entre os anos de 2012 à 2016 e programas semestrais do Santander. Apesar da reformulação do programa Ciência sem fronteiras em 2017 para focar em alunos de pós-graduação, segundo a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a internacionalização necessita de ajustes para torná-lo mais eficiente (CAPES, 2017). O projeto de internacionalização da Universidade federal de Pernambuco (UFPE) foi aprovado em 2018 e junto com outras 24 Institutos de Ensino Superior (IES) que foram selecionadas para programa que vai ampliar a internacionalização destas instituições (UFPE, 2018). Contudo, as DRIs ou AAIs de algumas Instituições de Ensino Superior (IES), como é o caso da DRI/UFPE, ainda podem estar registrando suas informações sobre mobilidade internacional estudantil de forma descentralizada, pouco informatizada, com pouca padronização e fora de sistemas informatizados da instituição, dificultando a recuperação automática e eficiente da informação.

Diante dessas informações iniciais, o projeto propôs a utilização do processo *Knowledge Discovery in Databases* (KDD) – extração de conhecimento em bases de dados – para o desenvolvimento de um ambiente computacional analítico de informações sobre intercâmbio acadêmico. Sendo o intercâmbio acadêmico um período que o aluno passa fora do seu país de origem ou residência, conhecendo outro lugar com objetivos acadêmicos. O processo KDD é importante na busca para encontrar e interpretar padrões/regras mediante obtenção e integração de diversas fontes de dados. O intuito é extrair de bases de dados, com ou sem formulação prévia de hipóteses, informações desconhecidas a priori, factíveis, válidas e acionáveis, que poderão ser úteis para a tomada de decisão (FRAWLEY et al., 1992; FAYYAD et al., 1996). O processo KDD determina

as etapas que produzem conhecimentos a partir dos dados e, principalmente, define a etapa de Mineração de Dados (*Data Mining*), fase esta que transforma dados em conhecimentos (WITTEN et al., 2016). Outra forma de extrair conhecimento com mineração de dados é o uso do *framework* CRISP-DM desenvolvidos por Wirth e Hipp. Tal *framework* foi desenvolvido para mineração de dados entre setores de indústria e sua abordagem cíclica remete a teorias e conceitos mais familiarizados com a administração. A partir dos conceitos do processo KDD e o *framework* CRISP-DM é proposto um método adaptado para minerar os dados encontrados na pesquisa e extrair conhecimento utilizando a mineração de dados. Após o conhecimento adquirido, o especialista do domínio do conhecimento deve aplicá-lo no órgão responsável, neste caso as DRIs ou AAIs, proporcionando a gestão eficiente dos métodos e processos. Cada fase da execução do processo KDD e do framework CRSIP-DM possuem interseção com as demais, deste modo, os resultados produzidos em uma fase podem ser utilizados para melhorar os resultados das próximas fases. Este cenário revela um processo interativo, que busca aprimorar os resultados a cada interação.

O sistema computacional de análise e diagnóstico do cenário atual de mobilidade internacional estudantil foi utilizado visando à integração com planejamento estratégico institucional, as ações e metas de internacionalização e utilizaram os dados da Diretoria de relações internacionais e dados solicitados ao NTI (Núcleo de Tecnologia de Informação) da UFPE. Neste sentido, este trabalho elaborou framework de mineração de dados locais da DRI e da UFPE, e como consequência, a descoberta de padrões que geram novos conhecimentos para que medidas preventivas e corretivas possam ser empregadas, a fim de mitigar, entre outros problemas, o cumprimento das ações e metas voltadas a programas de mobilidade e internacionalização da UFPE. Ao constatar deficiências e pontos fracos nos programas de intercâmbio, as ações poderão ser tomadas por parte da instituição e, mediante políticas públicas, visa solucionar tais deficiências e dar apoio a tomada decisão.

A mineração de dados foi realizada pelos algoritmos do Aprendizado de Máquina (AM), e para alcançar os propósitos da pesquisa foram utilizados métodos do Aprendizado de Máquina Supervisionado e Não Supervisionado (HAN e KAMBER, 2011). Também aplicados os fundamentos de Administração Pública, Gestão Estratégica, Banco de Dados, *Data Science* e *Business Intelligence* (MOTTA, 2006; SILBERSCHATZ et al., 2012; BAHGA e MADISETTI, 2016).

Diante do exposto, organizar e tratar os dados educacionais relacionados aos programas de mobilidade acadêmica e internacionalização da UFPE, a fim de apresentar um diagnóstico de gestão do setor de relações internacionais da UFPE e uma nova visão dos dados registrados, meios de atender melhor os alunos, servidores e professores no setor, gerando uma base de dados sólida para apoio a decisão e aperfeiçoamento da gestão.

Visando investigar o que foi exposto na introdução, no que concerne ao processamento de dados educacionais, gestão do conhecimento e apoio a tomada de decisão da DRI da Universidade Federal de Pernambuco, questionamos: Como tratar dados educacionais da UFPE, com uso de mineração de dados para ajudar na tomada de decisão relacionada aos programas de mobilidade acadêmica e internacionalização?

1.1 Objetivos

Para responder à pergunta de pesquisa apresentada anteriormente foram estabelecidos os seguintes objetivos de pesquisa geral e específicos:

1.1.1 Objetivo Geral

O objetivo geral da pesquisa consiste em propor modelo de tomada de decisão usando mineração de dados para tratar dados relacionados aos alunos de graduação dos programas de mobilidade acadêmica e internacionalização da UFPE.

1.1.2 Objetivos Específicos

Para atingir o objetivo proposto, têm-se os seguintes objetivos específicos:

- Estruturar conjunto de dados para mineração de dados;
- Avaliar procedimentos, técnicas e algoritmos (classificação, associação e agrupamento) por meio de ferramentas de Mineração de dados, métodos de aprendizado de máquina e estatística descritiva para analisar os dados estudados;
- Encontrar padrões e descobrir novos conhecimentos nos dados de mobilidade internacional estudantil;

- Realizar levantamento sobre as ações de internacionalização realizadas pelo setor para atender as metas propostas no PDI 2014-2018 da UFPE;
- Realizar diagnóstico de gestão no setor de relações internacionais da UFPE.

1.2 Justificativa e Contribuições do Estudo

Existe uma tendência de internacionalização das universidades públicas, sendo um tema discutido na UFPE desde 2014, incluído no planejamento estratégico da UFPE e suas pró-reitoras. Um dos setores que lidam diariamente com questões relacionadas à internacionalização é a Diretoria de Relações Internacionais, sendo registrados e acompanhados os alunos estrangeiros que chegam à Instituições de ensino superior (IES), bem como os alunos que vão para o exterior estudar em instituições de países diversos. A resolução 04/2014 – UFPE regulamenta sobre as mobilidades nacionais e internacionais, sendo esta última atribuição da Diretoria de Relações Internacionais. A DRI registra desde 1999 dados relacionados aos intercâmbios, que totalizam, atualmente, aproximadamente 5.600 alunos, locais e estrangeiros, que participaram de intercâmbio. Sendo relevante esta pesquisa por aspectos práticos e teóricos.

Este estudo teve como desafio, identificar padrões de dados educacionais de intercâmbio internacional estudantil da UFPE por meio do uso de mineração de dados, que é uma das fases do *KDD (knowledge-discovery in databases*) – processo de extração de informações de base de dados – visto que os dados se encontram em planilhas eletrônicas descentralizadas e sem padronização, dificultando a recuperação automática e eficiente da informação para apoio a decisão. Esta pesquisa permitiu observar tendências; detectadas por meio de dados mais frequente e revelaram deficiências ou beneficies nos programas de mobilidade acadêmica; classificou padrões nos dados de mobilidade acadêmica e entre outros diagnósticos inerentes ao escopo do projeto. Tais benefícios proporcionarão melhorias na gestão do setor, melhor acompanhamento dos alunos nos programas de mobilidades e ajudar os professores e envolvidos a focarem em diminuir ou eliminar as deficiências.

A consequência prática da proposta estudada consiste em analisar esses dados e encontrar novas perspectivas por meio de técnicas computacionais de mineração de dados, analisar o planejamento institucional e setorial, visando aprimorar a gestão de informação e tomada de decisão para o setor e contribuir com aprimoramento do processo

de intercâmbio da DRI da UFPE. Como o PDI 2014-2018 da UFPE continha 23 metas relacionadas com internacionalização, a interação de análise de dados com o processo de tomada de decisão pode proporcionar uma nova visão dos resultados para a instituição e contribuir para o cumprimento dos objetivos do novo plano de internacionalização 2017-2027 lançado no início de 2018. As atividades demonstradas neste estudo também poderão ser replicadas em outros setores ou em outras instituições e utilizada como base para novas ações futuras. Melhorando a gestão, irá consequentemente, gerar benefícios aos principais usuários do setor: os alunos, os pesquisadores e a comunidade acadêmica.

Um estudo publicado em 2015 por Cobbe et al na revista Perspectivas em Gestão & Conhecimento (PG&C), disponibilizado pelo portal de periódicos científicos eletrônicos da UFPB, analisou a inteligência organizacional aplicada às instituições de ensino superior, mostrando sua importância como área do conhecimento em pleno desenvolvimento, também usada para melhorias dos processos e a capacidade possuída por organizações para coletar informações, analisá-las, criar conhecimentos, inovar e agir com base nos conhecimentos adquiridos (MORESI, 2001). Ao melhorar a gestão teremos boas tomadas de decisões, otimização da utilização dos recursos e tendo mais informações poderão diminuir os custos não previstos.

A contribuição teórica para a academia é propor uma nova perspectiva na utilização de mineração de dados educacionais, ajudando atores da administração pública na tomada de decisão. Ao propor modelo de tomada de decisão para tratar dados educacionais relacionados aos programas de mobilidade acadêmica e internacionalização da UFPE, o pesquisador contribui também para que o *Business Intelligence* (BI - inteligência empresarial) seja mais explorado nas instituições educacionais, visando maior eficiência na aplicação dos recursos utilizados e entendendo mais detalhes de como os programas e os estudantes se comportam ao ingresso nas mobilidades internacionais.

Resumindo, as contribuições centrais estão baseadas em três pontos importantes: a criatividade em fazer a depuração dos constructos (variáveis e atributos), pois organiza e se beneficia dos dados; a mineração por ter seu esforço em encontrar as ferramentas corretas para ser aplicada; e a analise e interpretação dos padrões descobertos na mineração.

Para concretizar o seu propósito, este trabalho de pesquisa está organizado em 5 capítulos, sendo o primeiro a introdução, incluído a justificativa e o contexto da pesquisa, o segundo o embasamento teórico do estudo, o terceiro sua metodologia, em seguida resultados e discussões e o último, as conclusões incluindo as recomendações gerenciais.

1.3 Contexto da Pesquisa

Esta seção apresenta o contexto do estudo, apresentando o cenário atual da problematização do estudo para dar introdução ao referencial teórico.

1.3.1 A Internacionalização na UFPE

Segundo De Wit (2013), internacionalização e educação internacional são conceitos distintos. Educação internacional são atividades internacionais isoladas e internacionalização da educação é algo mais amplo que além das instituições, países, pesquisa e extensão. Englobam relações culturais, relacionamento e aprimoram o processo de formação do estudante. A UFPE está trabalhando para romper a barreira da educação internacional e evoluir para internacionalização como estratégia institucionalizada.

O planejamento da Internacionalização nas universidades públicas está sendo mais relevante nos últimos anos devido ao crescente movimento de globalização e ampliação de ações que visam expansões das IES brasileiras. Segundo a CAPES (2017, p.6) internacionalização é um processo amplo e dinâmico envolvendo ensino, pesquisa e prestação de serviços para a sociedade, além de construir um recurso para tornar a educação superior responsiva aos requisitos e desafios de uma sociedade globalizada. A CAPES afirma que as IES brasileiras precisam se tornar mais proativas para melhorar o impacto de suas ações no assunto.

A UFPE elaborou seu Plano de desenvolvimento Institucional (PDI) para o período de 2014 a 2018 (PDI – UFPE, 2014-2018), que na fase inicial propôs como objetivo estratégico implantar uma política de internacionalização, nas fases seguintes firmou a internacionalização com um dos nove eixos temáticos do plano e dando mais importância formal ao assunto institucionalmente. Neste plano, para a internacionalização foram desenvolvidas 23 metas, e dentre elas, 10 metas são diretamente de responsabilidade do setor de Diretoria de relações internacionais da UFPE e sua verificação de conclusão contribui para atender objetivos deste estudo, seguem elencadas abaixo:

- Aprimorar a estrutura normativa para possibilitar a equivalência dos créditos resultantes de mobilidade nacional, internacional e interna (inter campi);

- Estruturar e consolidar o Núcleo de Formação em Línguas para todos os cursos de graduação. Incrementar em 50% o ensino de língua estrangeira para a comunidade acadêmica;
- Adotar as medidas necessárias para estimular a admissão, na pós-graduação, de alunos provenientes de outros países (alunos estrangeiros);
- Divulgar os programas de pós-graduação junto às embaixadas sediadas no Brasil;
- Estruturar e realizar missões internacionais (pelo menos uma missão por ano);
- Incentivar a capacitação de professores no exterior (pos doc);
- Ampliar a oferta de material de divulgação da UFPE em português, inglês, francês e espanhol;
- Ofertar sistematicamente disciplinas em inglês na pós-graduação começando com pelo menos uma disciplina por ano, por programa;
- Investir em marketing institucional interno e externo;
- Ampliar em 30% o auxílio a idiomas para os estudantes de graduação fazerem cursos de inglês.

1.3.2 Cenário Atual da Internacionalização na UFPE

Em 2018 a UFPE aplicou e foi aprovado, junto a CAPES, seu projeto de internacionalização. A CAPES está investindo R$ 300 milhões anuais a partir de 2019 e contemplando cinco grandes temas que abrangem diversos campos do que são importantes para a UFPE, são: Biodiversidade e Conservação de Recursos Naturais, Inovação nas Ciências Básicas, Estado e Sociedade na Contemporaneidade Global, Dinâmicas de Desigualdade e Desenvolvimento, Inovação em Saúde e Modelagem de Sistemas (UFPE, 2018). Em abril de 2018, a UFPE divulgou seu Plano de Internacionalização 2017-2027 no portal da internet da instituição e tem como propósito integrar o Plano Estratégico Institucional (PEI) 2013/2027 e o Plano de Desenvolvimento Institucional 2014/2018 e contemplando objetivos, estratégias e a ações para implementar e expandir a internacionalização da UFPE. Sendo a internacionalização presento em dois objetivos estratégicos do PEI: tornar a UFPE uma das 100 melhores universidades do mundo e implantar uma política de internacionalização.

Segundo a Organização para a Cooperação e Desenvolvimento Econômico (OECD, 2017) estudantes matriculados em instituições universitárias fora de seus países

cresceram de 800 mil no final da década de 1970 para 4,6 milhões em 2015. De acordo com o Plano de internacionalização da UFPE 2017-2027 (2018), as tendências do ensino superior no contexto da internacionalização observam os principais aspectos: expansão e diversificação das instituições de ensino superior; Valorização do ensino superior como instrumento de mobilidade social; Valorização social e científica da interdisciplinaridade; Formação com base em competências, conteúdos transversais e uso das Tecnologias de informação e comunicação (TIC) no processo de ensino e aprendizagem e Profissionalização da gestão universitária.

Os principais desafios identificados inicialmente no setor são: a necessidade de maior interação com as coordenações de cursos, alinhamento de ações com as pró-reitorias e acompanhamento dos alunos em mobilidade quanto a parte de aproveitamento de disciplinas cursadas em outras instituições e dificuldades de adaptação nos países estrangeiros.

Para que todos os desafios sejam superados e a internacionalização da instituição seja mais eficiente, o plano descreve cinco objetivos, cinco eixos estratégicos e três eixos transversais visualizados, conforme pode se observar na Figura 1.

Figura 1 - Interseção entre eixos estratégicos e eixos transversais

Fonte: Plano de internacionalização UFPE 2017-2027 (2018).

A seguir os objetivos deste Plano de Internacionalização:

- Incorporar dimensões internacionais e interculturais no ambiente universitário;

- Ampliar a capacidade de comunicação internacional da comunidade universitária;
- Dar visibilidade nacional e internacional às atividades de ensino, pesquisa, extensão e inovação;
- Fortalecer e adensar a produção do conhecimento e da pesquisa realizada na UFPE;
- Promover um ambiente intercultural e internacional de ensino-aprendizagem e trabalho, que traga benefícios para o processo democrático de formação de qualidade.

Eixos estratégicos e ações para atingir os objetivos deste plano:

- Mobilidade universitária – refere-se à mobilidade in e out de estudantes, professores, pesquisadores e técnicos administrativos;
- Internacionalização da graduação, da pós-graduação, da pesquisa, da extensão e da inovação – refere-se ao conjunto de medidas que levem à ampliação da capacidade de diálogo da comunidade acadêmica com o mundo nos seus processos de ensino e aprendizagem e de trabalho;
- Internacionalizar em casa – refere-se a ações que ampliem e fortaleçam as possibilidades de trocas de conhecimentos interculturais e acadêmicos, prioritariamente ocorridas no ambiente da UFPE, a exemplo do ambiente virtual de aprendizagem, salas interativas e da promoção de eventos de caráter internacional na UFPE;
- Missões institucionais e participação em redes – refere-se à presença institucional da UFPE em associações e redes internacionais de ensino superior;
- Desenvolvimento de capacidades - de estudantes, professores e técnicos administrativos que estarão envolvidos com a internacionalização universitária.

E os três eixos transversais que apoiam e dão sustentação para os eixos estratégicos:

- Habilidades em línguas estrangeiras - oferta de cursos de línguas para a comunidade acadêmica;
- Tecnologia da informação e marketing institucional – divulgação por meio de mídias digitais e meios tecnológicos, o acesso multilíngue às informações sobre a instituição e as atividades universitárias;
- Regulamentação - definir as diretrizes legais para ações de internacionalização institucional.

2 Referencial Teórico

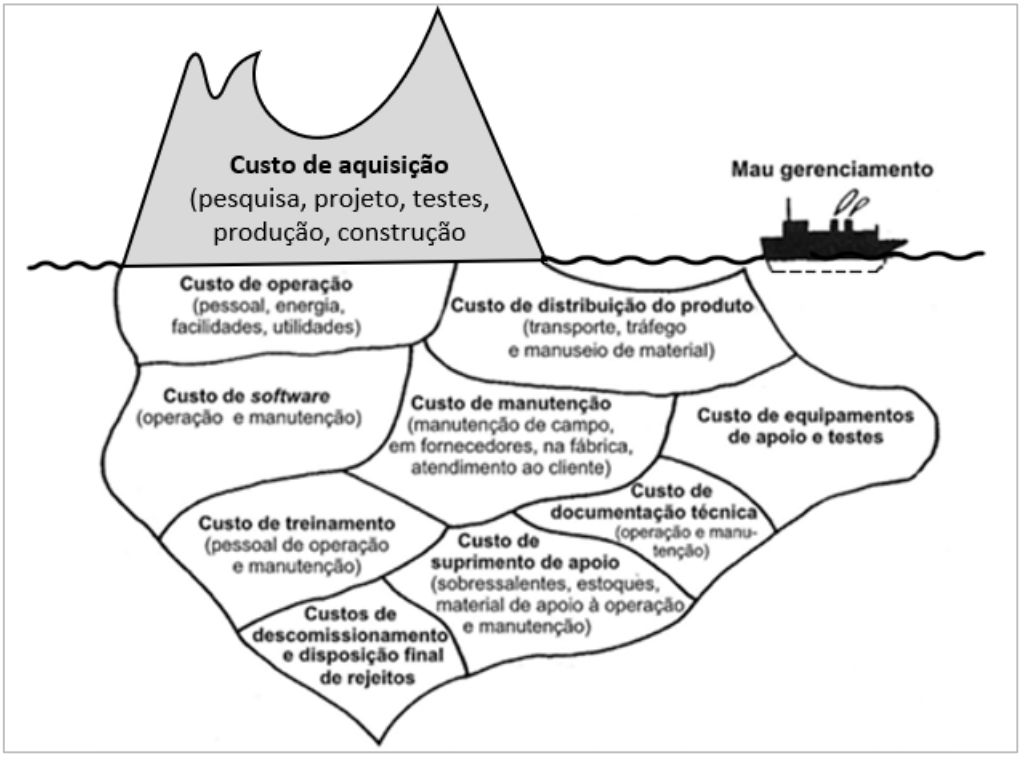
Este capítulo versa sobre os pontos relevantes a respeito da gestão estratégica da informação, *Business Intelligence*, tomada de decisão e mineração de dados, sendo apresentado os principais conceitos acerca do processo decisório: informação, atores, estilos decisórios e processo decisório na Administração Pública.

2.1 Gestão Estratégica da Informação

Sobre gestão estratégica na administração pública podemos destacar a criação do Programa Nacional de Gestão Pública e Desburocratização (GESPÚBLICA), criado em 23 de fevereiro de 2005 pelo decreto 5.378 e descontinuado em 18 de julho de 2017. Tal programa fez levantamentos sobre indicadores, processos e no ano de 2014 criou um modelo de excelência em gestão pública. Portanto, mesmo sendo revogado, o programa gerou diversos registros que podem ser explorados e usados como ponto de partida no estudo sobre gestão estratégica na administração pública.

O Planejamento Estratégico, segundo Mintzberg (2010) é distribuído em três grandes grupos: prescritivas, descritivas e integrativas. A abordagem prescritiva é a mais influente da estratégia organizacional, pois a formulação da estratégia se dá por um processo de concepção, focado em objetivos definidos. A tecnologia está em processo dinâmico e veloz de mudanças, sua variedade torna a escolha dependente de que os gerentes sejam mais bem informados, para tomar decisões inteligentes sobre novos produtos e tecnologias de processos (MATOS e GUIMARÃES, 2005). Uma tecnologia apropriada é aquela que impulsiona a estratégia e dá à empresa uma vantagem sustentável. Numa realidade de crise, a busca por redução de custos é procurada constantemente e tal redução é justificativa para vários projetos. A fase de planejamento na concepção deve ser bem estudada e verificar todos os possíveis custos para evitar “surpresas” ocultas na fase de implementação e execução dos possíveis projetos. A Figura 2 a seguir demonstra os possíveis custos mencionados por (MATOS e GUIMARÃES, 2005). Um mau gerenciamento não considera os custos ocultos e que podem prejudicar o andamento e conclusão de um projeto. Portanto em um planejamento é de suma importância para prevê e fazer levantamento do maior número de informações.

Figura 2 - Custos ocultos da tecnologia



Fonte: Adaptado de Blanchard (1988).

O maior desafio é verificar como uma nova tecnologia irá ajudar a empresa alcançar vantagens competitivas, gerir de forma eficiente para evitar custos “extras” em seus projetos e quais as prioridades são mais vantajosas para atender seus objetivos. Segundo Laudon e Laudon (2010, p. 05), “novos negócios e setores aparecem enquanto os antigos desaparecem, e empresas bem-sucedidas são aquelas que aprendem como usar as novas tecnologias”.

Existem três mudanças inter-relacionadas na área de tecnologia: (1) a plataforma digital móvel que está surgindo; (2) o crescimento do *software on-line* como serviço; (3) o crescimento computacional em nível de conhecimento, no qual um número crescente de *software* empresarial é executado na internet (LAUDON e LAUDON, 2010, p. 06).

As novas tecnologias utilizam sistemas de informação, que segundo Laudon e Laudon (2010), é um conjunto de componentes inter-relacionados que coletam, processam, armazenam e distribuem informações, destinadas a apoiar a tomada de decisões. Portanto existe uma entrada de dados ao sistema de informações, esses dados são processados e tratados conforme estratégias definidas pela organização e geram novos dados que servirão como conhecimento em sua distribuição.

2.1.1 Inteligência Empresarial (Business Intelligence)

O *Business Intelligence* (BI), procura organizar e analisar grandes porções de informações de forma eficiente para apoiar a gestão por meio de uma abordagem no uso das tecnologias da informação e utilizado por organizações para melhorar o acesso e a compreensão de suas das informações (MOSS e ATRE, 2003).

No início dos anos 70, segundo Watson e Wixom (2007), os sistemas de apoio a decisão foram os primeiros aplicativos projetados para tal suporte, como por exemplo: processamento de transações operacionais, entrada de pedidos, controle de estoque e folha de pagamento. No decorrer dos anos surgiram novos aplicativos de suporte à decisão para informações de executivos, análise preditiva e processamento analítico online, do inglês *Online Analytical Processing* (OLAP).

O pioneiro a usar o termo *business inteligence* foi Howard Dressner, analista do Gartne Group nos anos 90. Hoje a maior utilidade da BI é ser um instrumento para impulsionar a eficácia e a inovação das empresas. O BI é um processo que inclui duas atividades principais: entrada e saídas de dados. Na entrada consistem em plataformas heterogêneas de dados e por meio de uma *Data Warehouse* (armazém de dados) extrai dados dos sistemas de origem e transforma em informações significativas para suporte à decisão (WATSON e WIXOM, 2007). Por exemplo, registros de várias disciplinas em uma instituição de ensino podem ser combinados e consolidados com base em número de identificação de alunos para agrupar de acordo com critérios de desempenho.

2.1.2 Uso da Informação no Processo Decisório

O uso de informações no processo decisório é necessário porque tomar uma decisão quando se está mediante a um problema que apresenta mais de uma alternativa plausível para solucioná-lo, havendo a influência de ao menos dois fatores conflitantes (GOMES; GOMES, 2012). É importante destacar que as consequências da decisão podem ocorrer de imediato, no curto ou longo prazo ou ainda pode ocorrer um impacto multidimensional.

Igualmente, a tomada de decisão não é um evento simples e único, no entanto é um produto de um processo social complexo que se estende sobre um período (SIMON, 1965). À vista disso, decidir envolve um processo de coleta de informações que vão subsidiar alternativas possíveis de solução e posteriormente a eleição da escolha mais

adequada (GOMES; GOMES, 2012). De acordo com Andrade e Amboni (2010, p. 200), “a tomada de decisão é um processo contínuo que permeia toda a atividade organizacional”.

Para Igor Ansoff - considerado pai da gestão estratégica por meio da obra clássica *Estratégia Corporativa* publicada em 1965 e influenciado pelas ideias de Peter Drucker e Alfred D. Chandler - a tomada de decisão pode ocorrer em três níveis: Estratégico (decisões voltadas ao alcance dos objetivos da instituição e de longo prazo), táticas (Voltadas ao atingimento de metas estratégicas e de médio prazo) e operacionais (voltadas às atividades do dia-a-dia e são de curto prazo). Ansoff diz que as análises são importantes, mas elas não podem produzir uma paralisia na empresa, pois as atividades não podem travar (MAXIMIANO, 2012).

Este processo decisório envolve os procedimentos de definição de problemas, avaliação das possíveis alternativas e a escolha das soluções plausíveis (BRAGA, 1987). Assim, considera o aspecto da racionalidade do decisor e depende de suas crenças, ou seja, as opiniões que são influenciadas pelas informações disponíveis no processo decisório. Neste sentido podem se tornar informações errôneas, surgidas pelas crenças do decisor, mas também podem gerar comportamento racional a partir do momento em que as ações sejam amparadas por estas informações (MELO e FUCIDJI, 2016).

Neste sentido, o comportamento racional depende de um conhecimento amplo, mas o ser humano possui um conhecimento incompleto, baseado na sua percepção da realidade e dos fenômenos que cercam sua ação (PEREIRA, LOBLER e SOMONETTO, 2010). Esta percepção da realidade advém do conhecimento presente na memória de longo prazo do decisor, conhecida como intuição, que é um processo de pensar relevante tanto no julgamento como na tomada de decisão (BETSCH, 2008).

Dessa forma, não apenas o comportamento racional do decisor pode ser suficiente para a tomada de decisão, na prática, a tomada de decisão racional sofre entraves, por meio do “choque de interesses entre sócios da empresa, pelas barganhas e negociações entre grupos e indivíduos, pelas limitações e idiossincrasias que envolvem as decisões, pela falta de informações e assim por diante” (CHOO, 2003, p. 29).

No entanto, os atores desse processo apresentam atribuições relevantes para que a tomada de decisão ocorra de forma satisfatória. Considerando, nesse âmbito, processo decisório como “procedimentos de definição de problemas, avaliação de alternativas e escolha de uma diretriz de ações e/ou soluções” (BRAGA, 1987, p.47). O Quadro 1 lista os principais atores que agem nesse processo.

Quadro 1 - Atores do processo decisório

Tipos de Atores	Atribuições
Decisor	Pessoa ou grupo a quem o processo decisório é voltado, apresentam a incumbência de validar uma decisão e arcar com as consequências
Facilitador	Um ou mais líderes que apresentam um papel de coordenar os decisores motivando-os e mantendo postura neutra durante o processo decisório
Agidos	Pessoas afetadas ou a quem o programa é imposto. Elas apenas participam e sofrem as consequências da decisão
Intervenientes	Possuem ação direta, tomam as decisões diretamente
Analista	Analisa, assiste os decisores na visualização do problema

Fonte: Adaptado de Gomes e Gomes (2012).

A partir desse quadro, pode-se identificar que vários atores fazem parte do processo de decisão, visto que por diversas vezes o decisor necessita de apoio de uma equipe que o assista, como o analista, bem como de um facilitador que coordene as atividades e faça a mediação durante o processo. Outros atores presentes neste processo são os intervenientes, que tomam as decisões diretamente, no entanto, o decisor que validará a decisão, e, por último, os agidos que são afetados pelas decisões que são impostas. Para nosso estudo é possível identificar que os agidos são: estudantes, pesquisadores, professores e técnicos da instituição estudada. Ou seja, os atores afetados pelo processo decisório sobre mobilidade estudantil e internacionalização da instituição estudada.

No assunto em questão, o decisor de estratégias institucionais lida com situações não estruturadas e incorpora elementos estruturáveis, sendo as decisões programadas as que são rotineiras e as não programadas, conhecidas como não estruturadas, são as que apresentam problemas novos e de consequências importantes (MINTZBERG, 1976; SIMON, 1963).

As decisões programadas estão mais voltadas ao nível operacional, enquanto que as decisões não programadas estão voltadas ao nível tático e estratégico (gerência e alta chefia). James D. Thompson (1967) apud Christensen (2007) afirma que, na administração pública, o nível superior - considerado um nível complexo que é responsável pelas tomadas de decisão - é composto pelos cargos mais altos, estando neles: cargos públicos, políticos e administrativos; enquanto que a liderança a nível tático é considerada administrativa, dado que os decisores auxiliam na implementação das decisões estabelecidas pelo nível estratégico no nível operacional, sendo este nível também responsável pela mediação, adaptação de normas e valores. O último nível é o operacional – os atores apresentam pouca liberdade e tem suas tarefas programadas.

Dessa forma, pode-se inferir que no processo decisório, a ação de cada ator é fundamental para a tomada de decisão, sendo as informações provenientes da racionalidade fundamentais para a realização desse processo.

2.1.3 Processo Decisório em Programas de Mobilidade e Internacionalização de IES

O uso de informações no processo decisório não é um processo simples, como já mencionado por Simon (1963) anteriormente. A base para uma boa tomada de decisão é como os dados são aproveitados para gerar informação importante para a tomada de decisão (GOMES, GOMES, 2012).

Desde a década de 80 o MEC tem aplicados instrumentos de avaliação nas instituições de ensino superior, visando garantir padrões mínimos de qualidade e incentivar as instituições com instrumentos de autogestão. Inicialmente com o Programa de avaliação da reforma universitária (PARU) em 1983, o Ministério da Educação (BRASIL/MEC) desenvolveu um conjunto de ferramentas de avaliação, resultando no atual Sistema Nacional de Avaliação da Educação Superior (Sinaes). O Sinaes utiliza como fontes de dados o Censo da Educação Superior (INEP, 2019), os resultados do Exame Nacional de Desempenho dos Estudantes (ENADE) e os relatórios das comissões de avaliação das IES brasileiras in loco para pontuar a qualidade do ensino das instituições brasileiras.

Este processo desde a década de 80 vem proporcionado um aprendizado organizacional e um aprimoramento coletivo das instituições. Este aprimoramento é considerado inteligência organizacional (IO), que segundo Choo (2002), constroem meios de aproveitamento das inteligências individuais e coletivas da organização e se torna "inteligente" na medida em que ela identifica, captura, disponibiliza e usa de forma extensiva a informação e o conhecimento. O lado prático da IO é realizado com o apoio da *Business Inteligence*, onde busca organizar e analisar maiores quantidades de informações de forma eficiente em apoio a gestão por meio do uso intensivo das tecnologias da informação e é amplamente utilizado por organizações para simplificar o acesso e a compreensão de suas das informações (MOSS, ATRE, 2003).

A internacionalização possibilita desenvolver uma cooperação entre instituições de ensino superior, bem como colaboração científica, tecnológica ou cultural, formando

equipes de pesquisa, formação de dupla titulação, acolhimento de alunos de graduação e pós-graduação, e mobilidade técnica de servidores e docentes. A internacionalização de uma IES envolve não apenas um conjunto de políticas, mas também estratégias, ações e atores (OLIVEIRA, FREITAS, 2016). O estudo realizado por Oliveira e Freitas (2016), verificou a motivação de alunos e professores universitários para a realização de mobilidade acadêmica internacional. A pesquisa empírica foi feita com 30 estudantes (brasileiros e estrangeiros) e professores (brasileiros) que escolheram a mobilidade acadêmica internacional como parte da sua formação.

De forma ampla, os resultados mostraram motivações pessoais, acadêmicas e profissionais em todos os entrevistados. Nos alunos brasileiros os fatores pessoais foram predominantes, também vinculados com a idade e ao estágio de formação em que estavam vivenciando; já entre os alunos estrangeiros e professores, o que prevaleceu foram os fatores acadêmicos e profissionais. Já no estudo de Freitas (2015), teve como resultados é necessário desenvolver vários aspectos para alcançar a internacionalização como: conhecer os grandes eixos da internacionalização; ter visão internacional; promover estratégia para a internacionalização; saber das características de um centro institucional de internacionalização; e conhecer as vantagens institucionais dessa ação. E o autor também afirma que a internacionalização é imprescindível para a pós-graduação brasileira, pois contribui com o aumento da vitalidade e capacidade de inovação e na atualidade, é impossível imaginar ciência sem internacionalização.

2.2 Mineração de Dados

Como mencionado na introdução, o processo *Knowledge Discovery in Databases* (KDD) determina as etapas que produzem conhecimentos a partir dos dados e, principalmente, define a etapa de Mineração de Dados (*Data Mining*), fase esta que transforma dados em conhecimentos (WITTEN et al., 2016). “É o processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis, de uma fonte de dados.” (FAYYAD, 1996).

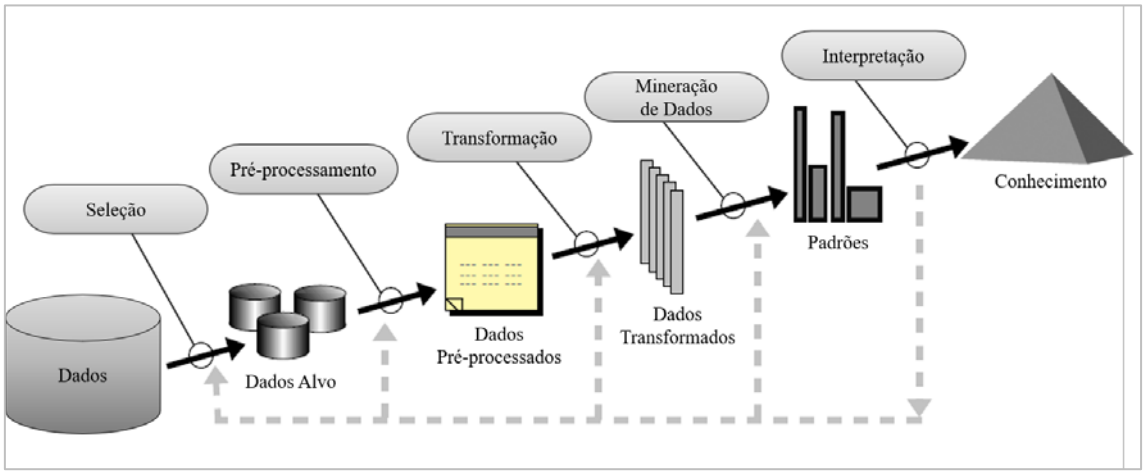
> [...] a mineração de dados são processos para explorar e analisar grandes volumes de dados em busca de padrões e normalmente está associada ao aprendizado de máquina: uma área da inteligência artificial que desenvolve algoritmos capazes de fazer com que o computador aprenda a partir de dados do passado (AMARAL,2016, P.2).

Na mineração existem vários métodos de Aprendizado de Máquina, dentre eles: classificação, predição, regressão, associação e agrupamento (MÜLLER e GUIDO, 2016). Neste projeto os métodos de classificação, associação e agrupamento foram pesquisados, a fim de selecionar os algoritmos mais propícios a apresentar as respostas/análises desejadas para que seja encontrado o modo mais adequado para os dados estudados e obter soluções mais viáveis. Também serão aplicadas técnicas de estatística com a finalidade de encontrar correlações e padrões nos dados. Já as ferramentas de *Business Intelligence* auxiliaram nas decisões estratégicas, visto que permitem a criação de consultas não triviais, facilitando a análise e visualização dos dados (SHARDA, 2017).

O uso de tecnologias *Online Analytical Processing* (OLAP) - processamento analítico *on-line* - possibilita agregações e sumarizações nos dados, gerando informações úteis ao processo decisório e oferecendo uma análise mais detalhada do cenário da mobilidade acadêmica. As ferramentas de visualização de dados proporcionam maneiras diferentes e intuitivas de apresentar os dados, de modo que sejam fáceis de entender. Esses recursos visuais serão fundamentais, tanto nas fases iniciais do projeto – permitindo entender os dados explorados (análise descritiva) e observar discrepâncias e *outliers* (fora do escopo principal) com maior facilidade, quanto nas fases finais com a obtenção dos *insights* (compreensões intuitivas).

A mineração gera regras de associação, descrevendo assim, padrões de relacionamento entre itens da base de dados. Uma base de dados pode conter informações importantes que só na mineração pode-se obter algo a mais que o ser humano é capaz de interpretar (CARVALHO, 1999). Nesta base de dados pode-se aplicar algumas etapas do processo de descoberta de conhecimento e identificar informações que a capacidade humana não “encontraria”. Portanto, KDD (*Knowledge Discovery in Databases*) ou descoberta de conhecimento em base de dados é definida como: processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, inseridos em grande volume de dados (FAYYAD, 1996). A Figura 3 demonstra as etapas do processo KDD, iniciando com a coleta e tratamento de dados, transformação, mineração e interpretação, até gerar conhecimento.

Figura 3 - Processo KDD



Fonte: Adaptado de Fayyad et al (1996).

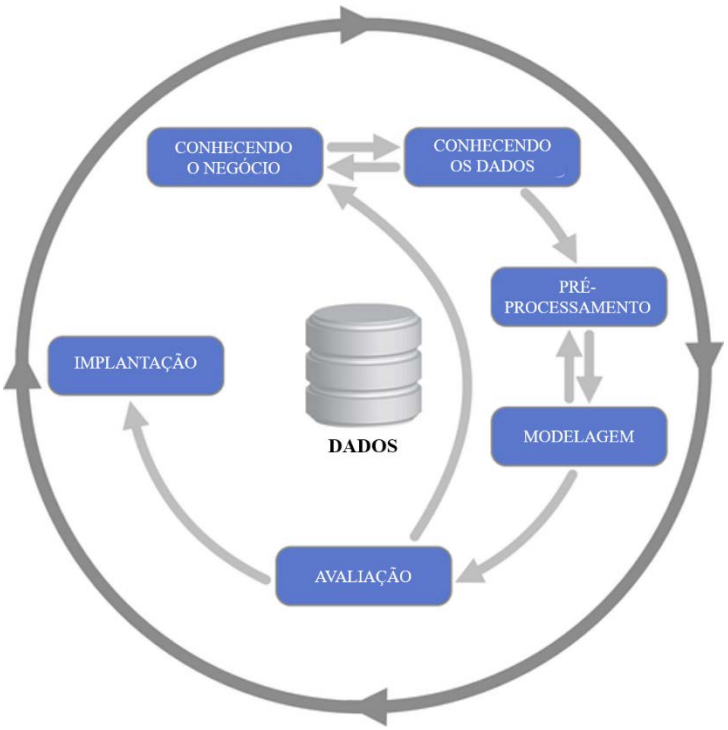
Uma das etapas da mineração de dados utilizadas nesta pesquisa é a seleção automática de características (atributos), afinal, sabe-se que explorar grandes quantidades de atributos, muitas vezes impossibilita a criação de modelos eficientes pelos algoritmos do aprendizado de máquina, fazendo-se necessário escolher quais atributos são realmente relevantes (BRAMER, 2013). Entende-se como atributo o mesmo que variável ou cabeçalho de uma coluna, campo da primeira linha de uma coluna, em uma planilha de eletrônica de dados ou base de dados. Dessa forma, faz-se necessário a redução da dimensionalidade, que consiste na redução dos atributos, porém, sem perder a representatividade dos dados originais, permitindo que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado. Para auxiliar nessa tarefa, serão analisados e aplicados algoritmos capazes de determinar o menor subconjunto de atributos com a melhor taxa de acurácia, em conformidade com custos computacionais razoáveis.

2.2.1 CRISP-DM

Outro processo de mineração e também muito usado é o *framework CRISP-DM* de Wirth e Hipp (2000). *CRISP-DM* é a abreviação de *Cross Industry Standard Process for Data Mining* que, trazendo para o português, pode ser entendida como processo padrão entre setores da indústria para mineração de dados. Essa é uma metodologia capaz de transformar os dados da empresa em conhecimento e informações de gerenciamento. Desenvolvido para projetos de mineração de dados e é dividido em seis fases: (a)

Conhecendo o negócio, (b) Conhecendo os dados, (c) Pré-processamento dos dados, (d) Modelagem da mineração, (e) Avaliação dos resultados e (f) Implantação, de acordo com a Figura 4. O formato circular representa a natureza cíclica do *framework*, pois em cada fase o aprendizado é constante e geram novas consequências nas demais fases.

Figura 4 - Metodologia CRISP-DM



Fonte: Adaptado de Wirth e Hipp (2000).

Cada fase tem seu principal propósito e “conversam” com as demais fases e são definidas a seguir:

a) Conhecendo o negócio: buscar informações sobre o ambiente estudado, entender suas necessidades, definir os objetivos e os critérios de sucesso;
b) Conhecendo os dados: organizar e estruturar os dados disponíveis;
c) Pré-processamento dos dados: preparação das bases de dados e escolha de técnicas de mineração;
d) Modelagem da mineração: aplicar a mineração com base nos objetivos definidos;
e) Avaliação dos resultados: interpretar os resultados obtidos na mineração e preparar processo de tomada de decisão;
f) Implantação: implementar processos definidos com base nas informações colhidas nas fases anteriores.

Portanto seu formato circular assemelha-se a processos usados na administração, como o método PDCA de Walter Andrew Shewart da década de 20 e os círculos de controle de qualidade de Ishikawa no final da década de 40.

2.2.2 Mineração de Dados Educacionais

Nesta seção foram abordados trabalhos que estudaram sobre mineração de dados educacionais e os tópicos importantes sobre avaliação de desempenho de alunos para efetuar comparativo com dados abertos governamentais disponíveis no Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e dados disponíveis na instituição estudada.

Os trabalhos acadêmicos na área mostra uma maior utilização da Mineração de Dados Educacionais (MDE) voltados para estudo na modalidade de ensino a distância como no estudo de Maschio et. al (2018), que buscou resultados de quais técnicas mais utilizadas; o que é mais investigado; e os dados mais relevantes para a pesquisa. As técnicas mais utilizadas foram de aprendizagem de máquina e agrupamento, a investigação mais realizada foi a de desempenho dos estudantes e os dados mais relevantes foram os números de interações entre alunos e professores que se dão nas avaliações, fóruns, *chats*, e outras formas de avaliação quantitativa. Este estudo colabora com a introdução sobre mineração de dados na seção anterior, pois reforça a utilização de algoritmos de aprendizagem de máquina e a hipótese de investigar o desempenho dos alunos como uma das aplicações da mineração de dados.

A MDE é considerada uma área recente, trata em desenvolver métodos para explorar conjuntos de dados coletados em ambientes educacionais e atualmente forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino (BAKER, et al, 2011). Os métodos utilizados na MDE são originalmente da área de mineração de dados, porém, Baker afirma que muitas vezes eles precisam ser modificados, porque é preciso considerar a hierarquia dos níveis da informação. Isso é necessário devido à falta de independência estatística nos tipos de dados encontrados em ambientes educacionais.

Neste estudo os principais métodos e modelos de mineração de dados mais aplicados a este estudo de acordo com o que é proposto nos objetivos. Segundo Baker, citando categorias dos métodos de mineração de dados estudados por Moore (2005), os

métodos: Predição, Agrupamento e Mineração de relações são de interesse tanto da área de MDE quanto da mineração de dados.

O método de predição, utilizando classificação e regressão são algoritmos mais populares na MDE e geram informações sobre o construto examinado, como em curvas de aprendizagem, e podem predizer os benefícios educacionais para um conjunto de estudantes. Caso o número de dados seja muito extenso, o modelo pode ser construído por uma parte dos dados, auxiliar no desenvolvimento e estimar os benefícios educacionais previamente sendo aplicadas com os alunos.

No agrupamento, o princípio é encontrar dados que se agrupem naturalmente, desconhecidos inicialmente e classificando por grupos e/ou categorias diferentes. No caso de dados educacionais, é possível agrupar alunos por diferenças e similaridades de desempenho ou comportamentos.

E em mineração, o objetivo é encontrar relações possíveis entre variáveis dos dados. Pode-se aprender quais variáveis se associam fortemente com uma determinada variável isolada. Estas identificações de relações podem ser de quatro tipos: regras de associação, correlações, sequências ou causas.

2.2.3 Mineração de Dados na Administração Pública

A mineração de dados na administração pública no Brasil está em processo de evolução de disseminação de experiências em diversos órgãos públicos e em diferentes áreas do conhecimento. Nesta seção foram citados trabalhos utilizando mineração de dados realizados em órgãos públicos.

Em pesquisas nos portais de trabalhos científicos, inicialmente a mineração de dados foi utilizada para pesquisas sobre os alunos da universidade aberta do Brasil com o ensino a distância instituída pelo Decreto 5.800, de 8 de junho de 2006 do governo federal e gerenciada pela CAPES. Como por exemplo, um trabalho sobre motivação de alunos na educação a distância usando combinação de técnicas de mineração de dados e obtiveram como resultados bons níveis de motivação, o que é bastante intrigante, pois na literatura evidencia dificuldades na adaptação à modalidade, altos índices de evasão e problemas decorrentes de pouca habilidade com os recursos computacionais (CAVALCANTI et al, 2014).

Tal trabalho foi elaborado usando técnicas de mineração de dados educacionais e o modelo de visualização de informações baseado em nuvem de *tags*. Nuvem de tags ou nuvem de etiquetas é uma lista de palavras organizadas hierarquicamente para melhor visualizar as palavras que mais ocorrem e que menos ocorrem em determinados textos estudados. Devido ao estudo levantar dados derivados em textos, a nuvem de *tags* ajuda visualizar de maneira hierárquica as palavras de maior ocorrência e analisando de maneira prática e rápida os níveis de motivação ou frustação de alunos recém-ingressantes na modalidade de ensino a distância.

Os dados foram coletados por questionários *online*, um no início e outro no final do curso com questões de respostas em faixas de motivação e texto livre e gerou duas planilhas eletrônicas. Para mineração foi usado o *software RapidMiner* 5.3[1], ferramenta de mineração de dados paga com modelos automatizados e que se ajusta aos dados que estão sendo analisados. E devido aos dados estudados foram utilizados para mineração regras de associação, algoritmos de frequência de subconjuntos, alguns agrupamentos e arvores de dados e sendo melhor visualizados em nuvem de *tags* com o usao da ferramenta gratuita *online* Wordle[2]. Este trabalho gerou contribuição no processo decisório por gerar dados de alunos frustrados ou indiferentes, mesmo que em menor número (16,5% dos alunos), mas a organização pode utilizar os dados para planejar ações de combate aos motivos deste fenômeno e monitorar se estas ações estão surtindo efeito.

Por sua vez, trabalhos de outras áreas do conhecimento já foram feitos, como por exemplo um trabalho sobre mineração de dados no serviço de atendimento móvel de urgência (SAMU) em Curitiba que analisou um conjunto de 5.839 eventos, onde foram atendidas 4.946 pessoas, sendo 12 pessoas demandaram seis ou mais vezes que os outros e que esta descompensação era devido a doenças crônicas que necessitavam um tratamento com maior acompanhamento (GOMES et al., 2014). Neste caso foi usado técnicas de mineração de dados através do processo KDD, passando pelas seguintes etapas: pré-processamento, processamento de estatísticas simples, extração de conhecimento da base de dados por meio da mineração de dados e análise dos padrões descobertos. No estudo menciona os seguintes algoritmos de mineração usados: Apriori[3] para extrair regras de associação; CHRONOASSOC de Gomes e Carvalho (2011) para identificar as sequências entre os eventos associados; e o algoritmo que descobre Regras

1 http://rapidminer.com
2 http://www.wordle.net
3 http://www.borgelt.net/apriori.html

de Associação de Causa e Efeito (RCE) de Gomes e el al. (2010) para extrair regras de associação considerando a janela de tempo entre os eventos associados. RCE é uma técnica bastante utilizada na área da saúde, pois nas pesquisas de doenças mede-se indicadores de causas para inúmeras patologias. E para a estatística simples usada o software SPSS versão 15. A base de dados utilizados foi disponibilizada pela Divisão de Ensino e Pesquisa da Secretaria Municipal de Curitiba e sendo utilizada a descoberta de regras de associação devido a natureza da aplicação para identificar os eventos que estão fortemente relacionados. O estudo estabeleceu critérios para preparar subconjuntos de dados para a etapa de mineração, a saber: o número de chamados pelo mesmo indivíduo e a escolhe de uma patologia capaz de exemplificar múltiplos chamados. Onde resultou em três subconjuntos: CONJ1 mais de um chamado no ano; CONJ2 complicações do câncer; e CONJ3 pessoas com mais de seis chamados no ano. Sabendo-se, segundo Amaral (2016) que para regras de associação são utilizados algoritmos de frequência de subconjuntos e/ou árvore de dados. Este trabalho gerou contribuição no processo decisório na gestão dos SAMUs em criar ações para melhor padronizar o preenchimento dos chamados, formas de identificação do paciente para facilitar o rastreamento de sequências de atendimentos para um mesmo indivíduo e melhorar o monitoramento de pacientes atendidos.

Outro estudo relevante encontrado foi sobre detecção de casos suspeitos de fraudes em licitações em municípios do estado da Paraíba no período de 2005 a 2016 e que os resultados gerados fornecem norteamentos importantes para otimização de processos e auditorias em órgão de controle (FRAGA, 2017). As técnicas de mineração de dados utilizadas foram para descoberta de regras de associação considerando tendências com padrões de estratégias cooperativas em jogos repetidos para mensuração da probabilidade de vitória de grupos de empresas com atuações cíclicas e da média de concorrentes (FRAGA, 2017). Para a descoberta de regras de associação o autor utilizou o algoritmo Apriori com limiares mínimos de suporte de 0,02% e confiança para identificação de regras de associação de 80% para produzir o maior número possível de indícios de irregularidades (suporte e confiança foram abordados na seção 2.3). Os dados foram coletados junto ao Tribunal de Contas do Estado da Paraíba para os períodos de 2005 a 2016 e totalizando 140.303 licitações ocorridas nos 223 municípios no estado. Devido a ter encontrado vários padrões de associação entre licitantes os resultados deste estudo contribuíram para a tomada de decisão fornecendo norteamentos importantes para

otimização de processos de fiscalização e auditorias em órgãos de controle aconselhados pelo autor.

Na administração pública federal, o Serviço federal de processamento de dados (SERPRO), maior empresa pública de prestação de serviços em tecnologia da informação do Brasil, segundo a revista Exame em 2012, faz uso da mineração de dados na administração pública para melhorar a transparência dos dados em prol dos serviços para a sociedade. Em setembro de 2019 aconteceu a quinta edição do seminário sobre análise de dados na administração pública organizado pelo Tribunal de Contas da União (TCU), Controladoria Geral da União (CGU) e pela Escola Nacional da Administração Pública (ENAP) com participação de 500 pessoas, entre técnicos e gestores públicos que participaram de oficinas e palestras.

O SERPRO fez presente e compartilhou boas práticas de mineração de dados na administração pública apresentando trabalhos sobre otimização na análise e classificação dos requerimentos da Procuradoria-Geral da Fazenda Nacional (PGFN), apoiando procuradores na classificação e análise de um grande volume de requerimentos submetidos, auxiliando na tomada de decisão. Este trabalho utilizou ferramentas de mineração de classificação para os requerimentos. Também foram usados algoritmos de aprendizagem de máquina. Neste caso a contribuição da tomada de decisão foi negativa, pois os resultados obtidos não foram suficientes para a tomada de decisão. Porém o trabalho sobre uma solução que realiza o processamento de amostras de petições iniciais da Defensoria Pública da União (DPU) que auxilia o defensor nas necessidades de assistência jurídica dos cidadãos do referido órgão (SERPRO, 2020). Sendo este em fase de elaboração do projeto e propõe utilizar ferramentas de mineração para classificação e agrupamento, gerando categorias de seções e subseções, como também, utilizando algoritmos para técnicas de aprendizado supervisionado, chamado pelo autor de “Rede Semântica” para classificação de textos. A contribuição prevista é de melhorar o processo de peticionamentos da DPU, aumentando a capacidade na composição das petições e melhorar o tempo para outras atividades dos defensores.

Por fim, é importante mencionar o INEP, já abordado em seções anteriores, onde são encontrados dados sobre os dados educacionais do ensino superior do Brasil (INEP, 2019). E o portal da transparência do governo federal, pois nele são apresentados dados de diversos sistemas utilizados pelo governo para auxiliar a gestão financeira e administrativa e prover transparência da gestão pública e proporcionar a sociedade realizar o controle social (Portal da Transparência, 2020).

A fim de apresentar uma visão geral dos casos exitosos de mineração de dados abordados nesta seção, segue o Quadro 2 com as informações dos tipos de mineração e suas contribuições no processo decisório.

Quadro 2 - Mineração de dados e processo decisório

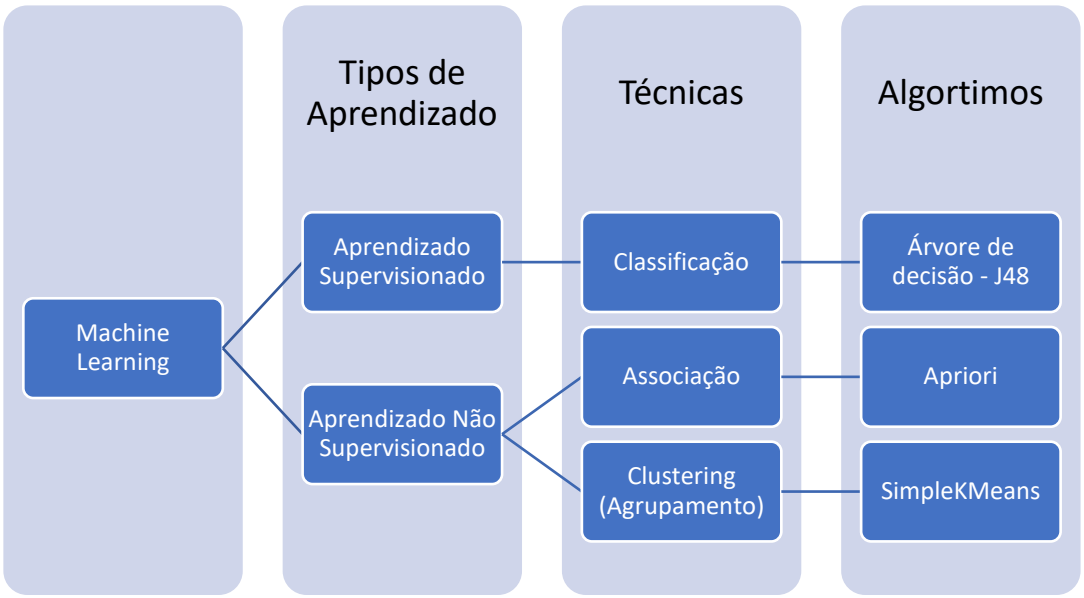
Casos estudados	**Técnicas de mineração**	**Contribuição no Processo decisório**
CAPES	Associação de texto	Melhoria de motivação dos alunos de ensino a distância.
SAMU (Curitiba)	Associação	Melhorar padronização do preenchimento dos chamados e facilitar o rastreamento de sequências de atendimentos para um mesmo indivíduo.
TCE - PB	Associação	Otimização de processos de fiscalização e auditorias de licitações.
SERPRO (PGFN)	Classificação	Sem contribuição.
SERPRO (DPU)	Classificação e Agrupamento	Melhorar processo de peticionamento da DPU e melhorar o tempo para outras atividades dos defensores.

Fonte: Elaborado pelo autor (2020).

2.3 Algoritmos de Mineração de Dados

Segundo Amaral (2016), a mineração de dados está associada ao aprendizado de máquina (*Machine Learning*) e que é uma área da inteligência artificial que desenvolve algoritmos capazes de aprender com os dados do passado. No aprendizado de máquina os dados estão dispostos como em uma tabela, que é composta por linhas e colunas. As colunas representam os atributos, semelhantes as variáveis, e as linhas são as instâncias, ou seja, onde os dados ocorridos são registrados. Para trabalhar com os dados precisamos entender como eles estão organizados e escolher quais tipos ou técnicas utilizar para minerar. O aprendizado de máquina é subdividido em supervisionado, informado que existe um atributo principal ou classe, e o não supervisionado que não levam em consideração um atributo principal. Após o tipo de aprendizado vem as técnicas de mineração, que neste trabalho foram utilizadas devido as características dos dados estudados: árvore de decisão, associação e agrupamento. E enfim os algoritmos das técnicas: J48, *Apriori* e *SimpleKmeans*. Segue a Figura 5 abaixo ilustrando as subdivisões do aprendizado de máquina.

Figura 5 - Visão geral de Aprendizado de Máquina



Fonte: Elaborado pelo autor (2020).

A técnica de árvore de decisão utilizada por meio do algoritmo de classificação J48, faz parte do aprendizado supervisionado. Ela cria uma estrutura no formato de árvore onde cada ligação entre os nós é uma instância e percorre até chegar no topo que é a classe (AMARAL, 2016). Segundo Harrison (2020), a arvore de decisão serve para fazer uma séria de perguntas com a finalidade de determinar a causa. Esta pesquisa optou em usar o software por ser um popular *open source* (código aberto), ou seja, fornecido gratuitamente desenvolvido em Java e mantido pela Universidade de Waikato na Nova Zelândia (AMARAL, 2016). Ao utilizar o *software WEKA* para aplicar algoritmo de classificação podemos analisar as reações do grau de precisão das classes do framework proposto neste trabalho. Geralmente quando se faz um framework não se tem análise interclasse para organizar os valores das classes e por isso a necessidade de eles serem independentes. Dois eventos independentes significam que a ocorrência de um não depende e não é influenciado pela ocorrência do outro, ou seja, são mutuamente exclusivos e ocorrem de forma separada. Portanto nesta pesquisa ocorre a independência.

A seguir um exemplo de uma árvore de decisão, que visa classificar se haverá jogo de tênis, considerando as condições climáticas do dia. Conforme ilustra a Figura 6, essa base de dados possui 16 registros e 4 atributos: Tempo {ensolarado, nublado, chuvoso}; Temperatura {quente, amena, fria}; Umidade {alta, normal}; Ventando {verdadeiro, falso}; e Jogar {sim, não}.

Figura 6 - Exemplo sobre condições climáticas para jogo de tênis

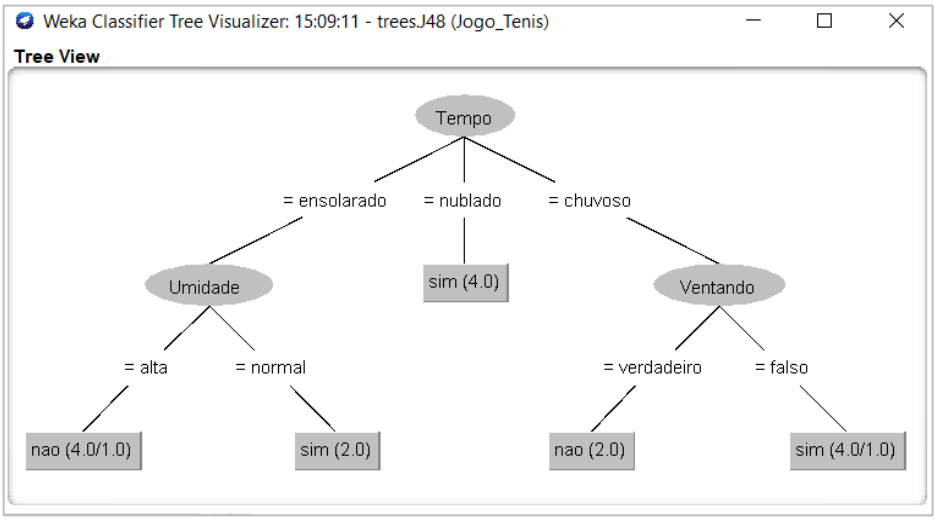
Relation: Jogo_Tenis

No.	1: Tempo Nominal	2: Temperatura Nominal	3: Umidade Nominal	4: Ventando Nominal	5: **Jogar** Nominal
1	ensolarado	quente	alta	falso	nao
2	ensolarado	quente	alta	verdadeiro	nao
3	nublado	quente	alta	falso	sim
4	chuvoso	amena	alta	falso	sim
5	chuvoso	fria	normal	falso	sim
6	chuvoso	fria	normal	verdadeiro	nao
7	nublado	fria	normal	verdadeiro	sim
8	ensolarado	amena	alta	falso	nao
9	ensolarado	fria	normal	falso	sim
10	chuvoso	amena	normal	falso	sim
11	ensolarado	amena	normal	verdadeiro	sim
12	nublado	amena	alta	verdadeiro	sim
13	nublado	quente	normal	falso	sim
14	chuvoso	amena	alta	verdadeiro	nao
15	ensolarado	quente	alta	verdadeiro	sim
16	chuvoso	fria	normal	falso	nao

Fonte: Elaborado pelo autor (2020).

Após execução do algoritmo de árvore de decisão, que neste exemplo foi o J48, obtém-se a árvore ilustrada na Figura 2. A visualização da árvore é invertida, isto é, a raiz (atributo Tempo) encontra-se no topo da árvore, e as folhas (atributo Jogar – sim ou não) encontra-se na parte inferior. No topo o algoritmo encontrou o atributo com maior ganho de informação, e as folhas representam a classe.

Figura 7 - Exemplo de Árvore de Decisão gerada pelo *software WEKA*



Fonte: Elaborado pelo autor (2020).

Na árvore de decisão cada nó representam os atributos, as ligações entre os nós representam os valores dos atributos e as folhas representam as classes. Cada caminho do nó raiz até as folhas representa uma regra, ou seja, um padrão detectado pelo algoritmo de mineração de dados. As árvores de decisão representam os padrões, que conduzem a

uma classe, de forma bastante intuitiva e autoexplicativa, por meio de um conjunto de regras (galhos), onde cada nó não terminal representa um teste ou decisão.

No exemplo da Figura 7, tem-se 5 regras do tipo: SE (valores dos atributos) → ENTÃO (classe). Segue exemplo de uma regra indicando que poderá ter jogo de tênis considerando as seguintes condições climáticas: SE Tempo = “ensolarado”, umidade = “normal” → ENTÃO “sim” (haverá jogo de tênis com essas condições climáticas). Em contrapartida, tem-se a regra que indica que não poderá ter jogo de tênis: SE Tempo = “chuvoso”, ventando = “verdadeiro” → ENTÃO “não”.

Portanto, uma das principais vantagens da aplicação de árvores de decisão é o fato de o modelo gerado apresentar regras explícitas, facilitando o entendimento do analista e especialista do domínio.

Para os algoritmos de classificação tem-se a avaliação de seu desempenho por meio de quatro opções de testes: *Use training test set, Supplied test set, Cross-validation e Percentage split*. No *Use training test set* todos os registros são utilizados para a construção do modelo e os mesmo registros usados para avaliar sua performance. No *Supplied test set* o pesquisador escolhe um arquivo externo para testar o modelo, ou seja, será treinado com os próprios dados e avaliado com os dados do arquivo externo. A *Cross-validation* os dados serão divididos em n partições definidas pelo pesquisador e serão feitos n testes sobre o modelo. Por fim o *Percentage split* testa apenas um percentual restante do que for definido. Nesta opção de teste o padrão do *WEKA* é definido em 66% e quer dizer que 34% dos dados que serão usados para teste de avaliação de desempenho.

A técnica de associação utilizada por meio do algoritmo *Apriori* faz parte do aprendizado não supervisionado. É mais utilizado para encontrar um conjunto de itens frequentes que consequentemente um subconjunto destes itens também será frequente e este princípio inverso também é válido (AMARAL, 2016).

Para criação de regras de associação existem dois importantes parâmetros: suporte (*support*) e confiança (*confidence*). Suporte calcula quantas transações contém todos os itens da transação e confiança aponta a porcentagem de vezes que uma transação contém um elemento e também contém outro. O algoritmo *Apriori* gera regras de acordo com a frequência com maior confiança na seguinte estrutura: antecedente A ==> consequente B <conf:(X.XX)>. A interpretação da regra é o antecedente causa (==>) o consequente com X.XX de confiança. O número apresentado para confiança é em formato decimal, exemplo 0.98, que significa 98%. Antes do símbolo ==> antecedente A tem um número que indica a cobertura absoluta da regra e o número após o consequente B, antes do

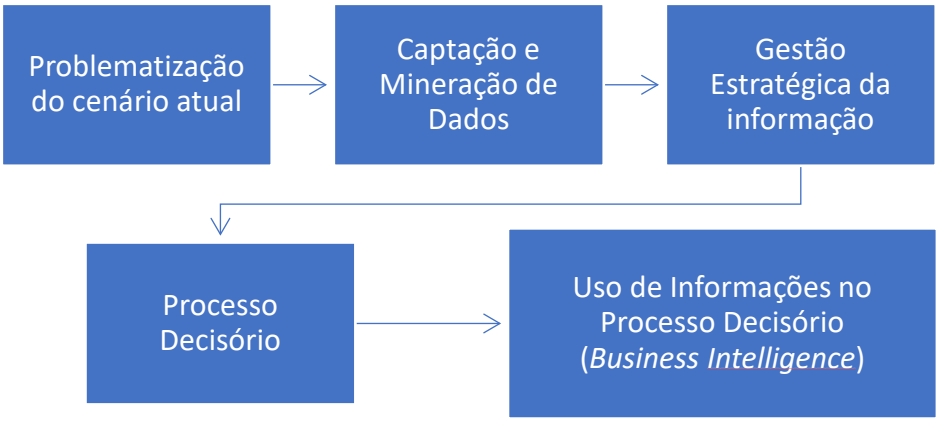
<conf:(X.XX)>, indica o número de instâncias que apresentam tal regra (AMARAL, 2016). Um exemplo na prática é a seguinte regra: “APROVACOES = '57_a_63` CENTRO = CTG **373** ==> POS-GRADUAÇÃO **366** conf:(0.98)”. Cuja interpretação é: Se alunos com número de aprovações entre 57 a 63, centro igual a CTG consequentemente fez Pós-graduação. Esta regra possui uma confiança (conf) de 98%, isto é, 373/366 = 0.98. Os parâmetros são configurados pelo pesquisador dependendo dos tipos de dados e o tipo de pesquisa que está sendo realizada. O padrão do software é de suporte entre 0,1 e 1 a confiança mínima de 0,9.

Por fim, a técnica de agrupamento (*clustering*) que também é um aprendizado não supervisionado, utilizada por meio do algoritmo *SimpleKMeans* tem a finalidade de dividir um conjunto de dados em um determinado número de grupos. Ele inspeciona os atributos e determina quais são semelhantes para pertencer a um grupo (*cluster)*. O pesquisador escolhe, baseado em alguma hipótese, o número de cluster a ser dividido e o algoritmo escolhe aleatoriamente um conjunto de dados com métricas de distanciamento mais semelhante (HARRISON, 2020).

3 Procedimentos Metodológicos

Esta seção apresenta as etapas necessárias para a consecução deste estudo. Os dados levantados pela mobilidade internacional na instituição geram informações que serão usadas de forma estratégica para poder obter conhecimento e inteligência empresarial. A estratégia utilizada neste estudo é captar os dados, aplicar mineração de dados dos estudantes de mobilidade, utilizar gestão estratégica da informação, entender o processo decisório da instituição e utilizar as informações obtidas para dar mais eficiência ao processo decisório. Tal sequência lógica está ilustrada na Figura 8.

Figura 8 – Etapas da metodologia do trabalho



Fonte: Elaborado pelo autor (2020).

A Figura 8 ilustra a sequência lógica de realização desta pesquisa. A problematização do cenário atual é a verificação das necessidades que o problema estudado apresenta aos pesquisadores. Após a problematização o foco é a busca dos dados que irão ajudar na utilização e apoio a gestão estratégica do negócio. Os dados trabalhados mostraram informações para a gestão o que será necessário para futuros processos decisórios. O uso de ferramentas de *Business Intelligence* gera informações que se tornaram em conhecimento e base para aprimorar as ações futuras do negócio estudado.

Esta pesquisa propõe organizar e tratar os dados educacionais relacionados aos programas de mobilidade acadêmica e internacionalização da IES, a fim de apesentar uma nova visão dos dados registrados, gerando uma base de dados sólida para apoio a decisão e aperfeiçoamento da gestão. Sem apresentar o intuito de generalizar os resultados para outras populações de outras universidades, sendo o paradigma que rege o trabalho em questão é o pós-positivista, também conhecido por alegação de conhecimento pós-

positivista, é uma filosofia determinista. Pois, estudam-se as causas que determinam os resultados da pesquisa, pautando-se pela observação, mensuração e verificação da teoria que está exposta no referencial teórico (CRESWELL, 2010). Diante do exposto, o sujeito epistemológico identificado é o estudante que participa da mobilidade acadêmica na instituição.

A seção a seguir apresenta a natureza da pesquisa e as diferentes abordagens metodológicas presentes na literatura e tipifica a abordagem metodológica adotada na pesquisa.

3.1 Natureza e Método da Pesquisa

A metodologia da pesquisa, segundo Santos (2002), construiu o conhecimento para o entendimento da realidade, por meio de pesquisas, análises e apresentando os fatos de forma correta e verídica.

O método científico é um conjunto de procedimentos técnicos, sistemáticos e racionais adotados na investigação de um fenômeno (CERVO, BERVIAN, SILVA, 2007), e a escolha do método deve estar correlacionado ao objetivo da pesquisa, neste trabalho.

Quanto à metodologia utilizada neste trabalho, caracteriza-se, quanto aos fins, como descritiva, pois, tem o intuito de descrever características do fenômeno (GIL, 2010). Corroborando com esta definição, os autores Cervo, Bervian e Silva (2007, p.61) afirmam que “a pesquisa descritiva observa, registra, analisa e correlaciona fatos ou fenômenos (variáveis) sem manipulá-los”. E por ser um estudo de caso com finalidades práticas, sua natureza é de pesquisa aplicada (SAMPIERI, 2013).

Ao organizar o método da pesquisa, o presente estudo é caracterizado como uma pesquisa de (i) abordagem quantitativa, (ii) natureza aplicada e (iii) caráter descritivo a ser realizado por meio de (iv) pesquisa bibliográfica e (v) estudo de caso (GIL, 2008; CERVO; BERVIAN; SILVA, 2007; SILVEIRA; CÓRDOVA, 2009; ASSIS, 2008; FONSECA, 2002; YIN, 2016).

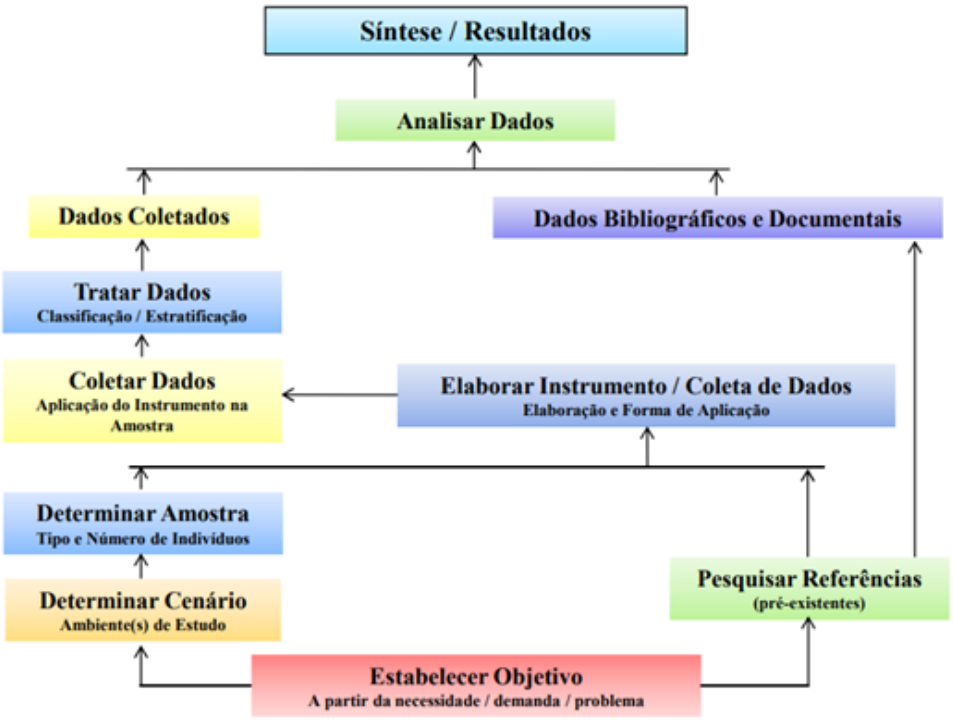
A abordagem quantitativa utiliza a coleta de dados para testar hipóteses, baseando-se na medição numérica e na análise estatística para comprovar teorias (SAMPIERI, 2013). O caráter descritivo concerne às pesquisas cujo objetivo seja descrever fenômenos em um contexto (CERVO; BERVIAN; SILVA, 2007; GIL, 2008; 2017). Os estudos

descritivos buscam especificar características de grupos de pessoas, comunidades, processos ou qualquer outro fenômeno que se submeta a uma análise (SAMPIERI, 2013).

Neste sentido, a pesquisa opta pela amostra não aleatória e intencional, em que os atores da pesquisa foram escolhidos intencionalmente pelo pesquisador. Por se tratar de uma pesquisa descritiva, além de relacionar variáveis de um pressuposto, se fez relevante obter a evidência da aplicabilidade de uma técnica para diferentes fins e em uma realidade específica (GIL, 2008; 2017; COOPER; SCHINDLER, 2016).

O estudo seguiu uma sequência lógica do método científico proposto por Rozenfield et al (2006) ilustrado na Figura 9, iniciado no problema da pesquisa estudada e estabelecendo seus objetivos. Determinar amostra neste estudo foi equivalente a verificação inicial dos dados e em seguida elaborar instrumento do início da preparação dos dados para posteriormente trata-los e em seguida aplicar a mineração para analisa-los. Praticamente em todo o processo foi necessário realizar pesquisas para as referências, bem como os dados bibliográficos e documentais para que o estudo tivesse andamento e fosse evoluindo.

Figura 9 - Desenho do estudo do método científico



Fonte: Rozenfeld et al (2006).

Para atingir os objetivos específicos da pesquisa seguiu-se a sequência lógica de Rozenfield elencando-os juntamente com a sequência do framework proposto na próxima seção.

Portanto a síntese dos procedimentos metodológicos da pesquisa é demonstrada no Quadro 3. Método descritivo, aplicado, determinista, quantitativo, pesquisa bibliográfica incluindo levantamento de dados e estudo de caso.

Quadro 3 - Procedimentos metodológicos da pesquisa

Natureza da Pesquisa	Aplicada
Filosofia	Determinista
Abordagem	Quantitativa
Caráter e Meio	Descritivo, pesquisa bibliográfica, levantamento de dados e estudo de caso

Fonte: Elaborado pelo autor (2020).

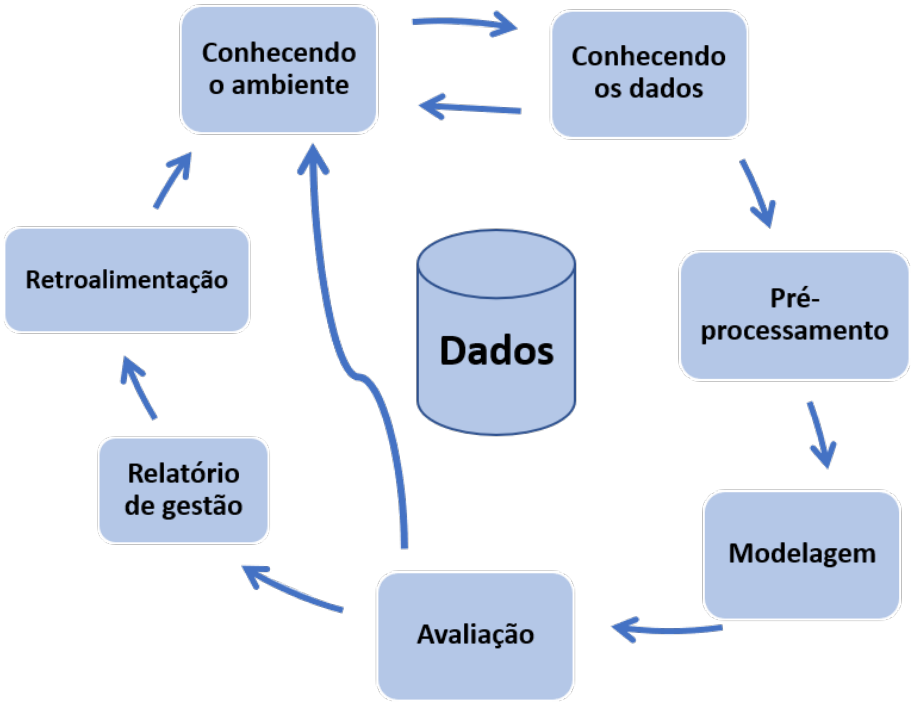
3.2 Framework proposto: adaptação do CRISP-DM

Para a análise dos dados está sendo proposto a utilização de uma adaptação do *framework* CRISP-DM de Wirth e Hipp, 2000 e integrando a sequência de atividades do KDD. Ele foi desenvolvido para projetos de mineração de dados e é dividido em seis fases: a) Conhecendo o negócio, b) Conhecendo os dados, c) Pré-processamento dos dados, d) Modelagem da mineração, e) Avaliação dos resultados e f) Implantação. O formato circular representa a natureza cíclica do *framework*, pois em cada fase o aprendizado é constante e geram novas consequências nas demais fases.

Porém a fase (a) conhecendo o negócio, renomeamos para “Conhecendo o ambiente” por motivo de poder aplicar em ambientes públicos ou privados. A fase (e) de implantação, é mais utilizada por profissionais de computação, e este trabalho propõe adaptá-la em duas novas fases para se fazer mais útil sua aplicação em administração, planejamento e tomada de decisões. Estas novas fases nomeadas como fase de “Relatório de gestão”, como a nova fase (f) voltado para planejamento e “Retroalimentação” acrescentando mais uma fase (g) voltada para execução de ações e seus feedbacks.

Ficando na seguinte ordem: (a) Conhecendo o negócio; (b) conhecendo os dados; (c) pré-processamento de dados; (d) modelagem da mineração; (e) avaliação dos resultados; (f) relatório de gestão; (g) retroalimentação. Segue abaixo a Figura 10 com as adaptações propostas nas fases do framework CRISP-DM.

Figura 10 - Fases do framework CRISP-DM adaptados para administração



Fonte: Adaptação proposta de Wirth e Hipp (2000).

A seguir tem-se a descrição detalhada de cada fase do modelo proposto, tomando como base teórica o framework CRISP-DM de Wirth e Hipp, 2000:

a) Conhecendo o ambiente;

Fase inicial onde é identificado os objetivos e metas para a mineração de dados. Neste caso o negócio é o setor de Relações internacionais da UFPE e seu foco de trabalho os estudantes que participam da mobilidade acadêmica para cursar disciplinas em instituições de ensino no exterior. Os objetivos e metas são as etapas realizadas nesta pesquisa para atender os objetivos gerais e específicos. Também são definidos os critérios de sucesso do projeto que neste caso foram definidos em: acurácia acima de 85% e confiança próximo ou maior que 90%.

b) Conhecendo os dados;

Esta fase é responsável pela coleta inicial de dados e familiarização como os tipos de dados específicos do projeto. Os dados utilizados para a mineração foram coletados no setor de relações internacionais por meio de planilhas onde são registrados os dados dos alunos que participaram de mobilidade internacional estudantil e dados solicitados ao NTI-UFPE de discentes matriculados na instituição constando dados de alunos com desempenho acadêmico.

Posteriormente, na fase de analises descritivas, foi feito uma comparação de desempenho entre os alunos que participaram de mobilidade acadêmica *versus* alunos que não participaram de mobilidade na base dados da instituição.

c) Pré-processamento;

Nesta fase é feita a construção do conjunto de dados a ser utilizado na modelagem (fase seguinte), organizando as linhas e colunas, segundo Amaral (2016, p. 05), cada coluna é um atributo e cada linha é uma instância. O atributo é análogo a uma característica ou variável e a instância pode ser comparada a um fato. Ou seja, os dados propriamente ditos pois cada linha é composta por um aluno e seu desempenho. Os dados atribuídos serão separados, categorizados e servirão como base ao andamento do projeto de mineração, com instâncias consideradas de boa qualidade e sem valores nulos. Em seguida é verificado se existem campos nulos ou com dados incoerentes a seus atributos e por fim ajustar a dimensionalidade dos atributos para evitar super ajustes de modelo ou atributos com campos de alta cardinalidade (quantidade muito alta de valores diferentes).

d) Modelagem;

Esta fase é onde define o modelo que será usado na mineração de dados, envolvendo as partes práticas das atividades específicas da mineração. Aprendizado de máquina, classificação, associação, agrupamento, supervisionados ou não e quais os tipos de algoritmos serão utilizados para gerar informações mais precisas. As atividades utilizaram a ferramenta *WEKA* (*Waikato Envoriment for Knowledge Analysis*) pacote de software para análise computacional e estatística dos dados fornecidos recorrendo a técnicas de mineração de dados (HALL et al. 2009). Sendo realizados ensaios preliminares ou experimentos para verificação de qual técnica e algoritmos mostram resultados mais relevantes. Após escolher os aprendizados, técnicas e algoritmos aplica-os gerando os experimentos.

e) Avaliação dos resultados;

Nesta fase é avaliada os resultados da modelagem e se os critérios de sucesso definidos na primeira fase foram atingidos. Caso algo deu errado é necessário determinar um novo escopo e tentar novamente. Ou caso a modelagem apresentar bons resultados segue para a próxima fase.

f) Relatório de gestão;

Esta fase propõe organizar a avaliação dos resultados para serem apresentados como relatório de gestão e serve como base para um planejamento de ações para a última fase. O relatório pode ser feito de forma convencional respondendo as questões a seguir:

Para quem o relatório será feito? Qual é a finalidade do relatório? Quais informações serão coletadas? Como os dados podem ajudar na tomada de decisões?

Também podem ser utilizadas ferramentas mais sofisticadas e automatizadas que criam relatórios gerenciais. Como por exemplo: Google Data Studio, Primavera e ferramentas de *Business Intelligence* que fazem a coleta, organização, classificação e análise dos dados.

g) Retroalimentação.

Nesta fase é proposto que seja feito um plano de ação para resolver eventuais fragilidades detectadas nos dados sobre o ambiente estudado. Sendo feito o acompanhamento do que foi sugerido no relatório de gestão e servirá como retroalimentação (*feedbacks*) para aprimoramentos futuros e/ou iniciando um novo projeto de mineração de dados com novos problemas identificados.

A mineração de dados em si, faz parte de um processo maior denominado de extração de conhecimento em bases de dados (WITTEN et al., 2016). Como mencionado na introdução e no referencial teórico, as etapas do processo KDD, iniciando com a coleta e tratamento de dados, transformação, mineração e interpretação, até gerar conhecimento. E para ficar mais familiar aos procedimentos utilizados na administração propomos na metodologia um framework adaptado baseado no CRISP-DM de Wirth e Hipp, 2000. O Quadro 4 a seguir demonstra comparativo entre as duas teorias e suas semelhanças entre as fases.

Quadro 4 - Comparativo *KDD x CRISP-DM* Adaptado

	Fase preparatória	Mineração	Descoberta de conhecimento
Processo KDD	Selecionando os dados e pré-processamento	Transformação e mineração de dados	Interpretação
Framework CRISP-DM Adaptado	Conhecendo o negócio, os dados e pré-processamento	Modelagem e avaliação	Relatório de gestão e *feedbacks*

Fonte: Elaborado pelo autor (2020).

Para melhor entendimento entre as duas teorias, neste trabalho a análise dos dados foi dividida em três fases: preparação, mineração e descoberta de conhecimento. Sendo a fase de preparação dos dados apresentada na próxima seção (3.3), e as fases de mineração de dados e descoberta de conhecimento constam no capítulo de resultados e discussões.

3.3 Seleção e Pré-processamento (Preparação dos Dados)

Os dados utilizados na mineração foram coletados no setor de relações internacionais, DRI/UFPE, por meio de planilhas onde são registradas informações dos alunos que participaram de mobilidade internacional. Também foram obtidos dados dos discentes oriundos do sistema acadêmico da instituição, disponibilizados pelo Núcleo de Tecnologia de Informações da UFPE (NTI).

Na etapa de pré-processamento e limpeza de dados são apresentados como os dados foram tratados, organizados e limpos para que não ocorra vieses, inconsistências ou campos com dados vazios ou atributos com grandes quantidades de valores diferentes.

O estudo de caso é válido por apresentar caráter de profundidade e do detalhamento na análise quantitativa da pesquisa. Segundo Yin (2015), o estudo de caso, é uma investigação empírica que conta com diversas fontes de evidência, utilizada para auxiliar na construção do conhecimento. Quanto aos indivíduos: servidores que atuam na área e alunos que participaram ou podem participar de programas de mobilidade na diretoria de relações internacionais. Para isso, será realizada o cruzamento dos dados institucionais dos alunos matriculados (dados NTI) e dados presentes em planilhas eletrônicas existentes na DRI/UFPE.

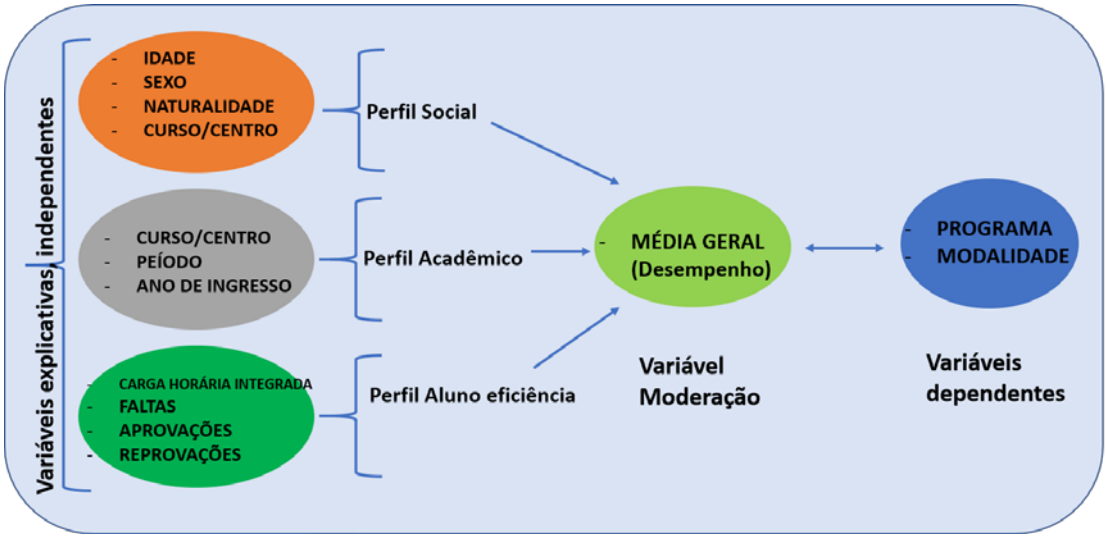
As atividades da fase de preparação de dados consistem em preparar as bases de dados, organizando, limpando e formatando para poder ser inserido num software de banco de dados e no software de mineração de dados. Equivale a cerca de 70% do tempo de todo processo passa-se nesta fase.

Inicialmente é feita uma avaliação nos tipos de dados que estão disponíveis, que neste caso eram planilhas eletrônicas com dados registrados no setor de relações internacionais no período de 1999 a 2019 e dados de alunos ativos solicitados ao NTI no lapso temporal de até final de 2019. Nesta avaliação foi identificado que continham dados de 4234 alunos que participaram de programas de mobilidade internacional estudantil em instituições estrangeiras e que a instituição tinha em 2019 mais de 42 mil alunos ativos segundo dados do NTI.

A princípio foi feito uma prévia de como ficariam distribuídas as variáveis, que futuramente são chamadas de atributos, separadas em perfis dos alunos estudados e suas

relações de causas como demonstrados na Figura 11 a seguir. Sendo esta prévia usada como base para a organização dos atributos nesta fase, moldando futuros perfis e que neste momento foram separados em três perfis: Social, acadêmico-social e acadêmico-eficiência.

Figura 11 - Perfis previstos na preparação dos dados



Fonte: Elaborado pelo autor (2020).

O passo seguinte foi verificar se existiam campos em branco, duplicados, ou com caracteres errados nas planilhas registradas no setor. Devido ter campos em branco e inconsistências de dados, na preparação esta lista foi reduzida de 4234 para 2432 alunos.

Em seguida foi feito uma avaliação dos tipos de variáveis estavam nas planilhas com o objetivo de eliminar as variáveis que não iriam contribuir para o estudo e para a mineração. Por exemplo, os campos de identificação dos alunos não devem e não precisam constar na base de dados para a mineração, porque não são considerados variáveis, e em outros casos, continham campos distintos com quantidade muito alta (Ex.: acima de 15). Neste trabalho, a partir daqui o termo atributo é utilizado no lugar de variável, já que o termo é mais utilizado na área de mineração de dados.

Dentre os atributos retirados tem-se: nome, e-mail principal, e-mail secundário, passaporte, tipo de ingresso, coordenação do programa e curso. Quanto ao CPF, este atributo foi retirado após a execução da junção entre as duas planilhas estudadas, em momento anterior à mineração. E quanto ao atributo curso, ele foi retirado porque tinham inconsistências quanto a quantidade de cursos contidos na planilha DRI e que não coincidiam com a lista de cursos da planilha NTI. Desta forma optou-se em usar o centro para identificar a área de conhecimento e evitar erros na mineração.

Avançando mais no estudo dos atributos a serem minerados, viu-se a necessidade de adaptar alguns deles para que não criem vieses ou dados inúteis e foi preciso diminuir a cardinalidade (quantidade de campos distintos) de alguns atributos para que se tornem dados úteis. Optando em categorizar os campos com muitos valores distintos. Devido a necessidade destas adaptações, segue abaixo quais atributos foram adaptadas e sua justificativa:

- Atributo DATA_NASCIMENTO foi adaptado para IDADE_MOBILIDADE porque a data de nascimento em si não iria contribuir e foi preciso modificar para saber qual foi a idade que o aluno fez a mobilidade. Para saber a idade mobilidade foi inserido uma função utilizando os atributos DATA_NASCIMENTO e ANO_MOBILIDADE para calcular a IDADE_MOBILIDADE.
- Atributo NATURALIDADE foi adaptado para que a quantidade de campos distintos fosse reduzida e não ocorresse o efeito viés. Este atributo continha cidades de aproximadamente todo estado de Pernambuco e inúmeras cidades de todo o país e foi categorizado para RMR (Região Metropolitana do Recife), INTERIOIR_PE e OUTROS_ESTADOS.
- Atributo PERCENTAGEM_INTEGRALIZADO, que informa o percentual de integralização do curso, onde 100% seria o caso de o aluno ter concluído todos os créditos. Vimos que seria uma informação isolada e talvez poderia ser melhor aproveitada. Por isso utilizamos o atributo ANO_MOBILIDADE junto ao PERCENTAGEM_INTEGRALIZADO e usamos uma função para saber por quanto tempo o aluno que ainda não tinha integralizado 100% passou após ter realizado a mobilidade e retornou resultados em um novo atributo chamado INTEGRALIZACAO_X_MOBILIDADE para saber se o aluno já integralizou 100% ou por quantos anos ainda não integralizou com casos de 1 a 7 anos.
- Atributo MEDIA_GERAL tinha campos em branco e foi preciso fazer uma investigação minuciosa "*in-loco*" (na base de dados) para verificar o motivo de tal fenômeno estar ocorrendo. Foi constatado que 100% dos casos eram devido a alunos que haviam sido incluídos no sistema recentemente, pois os dados dos atributos: FALTAS, APROVACOES E REPROVACOES, estavam "zerados", em branco, e para não "enviesar" os resultados dos dados, foi calculada a mediana (excluídos os zeros) e preenchidos com o valor da mediana nesses casos. Tal técnica é chamada de imputação de dados, faz parte de tratamento de valores nulos

do processo de limpeza e aconselhado por Harrison (2020, p. 54) o uso média ou mediana para dados numéricos.

- Atributo MEDIA_GERAL foi categorizado e criado um novo atributo, DESEMPENHO_ACADEMICO, sendo classificado como insuficiente médias entre 0 a 4,9, regular médias entre 5 a 6,9, bom com médias entre 7 a 8,5 e ótimo para alunos com médias entre 8,6 a 10. Esta categorização foi feita através de funções matemáticas na própria planilha eletrônica.
- Atributo CENTRO foi otimizado para que visualmente os centros fiquem mais fáceis de identificar e que ocupem menos espaço, portanto foi trocado o nome completo do centro por suas siglas. Isto é útil devido ao minerar e aparecer os centros nos resultados e regras as siglas ocupam menos espaços e visualmente mais fácil de enxergar por completo.
- Atributo PROGRAMA foi criado duas categorias programas nacionais CAPES e programas estrangeiros para diminuir a cardinalidade do atributo. Sendo Programas_nacionais_CAPES: Capes, dupla titulação, FACEPE, PLI, Marca e Brafitech. E Programas_estrangeiros: FINBRATECH, Bracol, ELAP, Erasmus, Mitacs.

Após preparar as planilhas para que as variáreis/atributos sejam mais bem aproveitadas pelas ferramentas de mineração de dados, o próximo passo foi inserir as planilhas eletrônicas limpas em um sistema de banco de dados, para que seja possível efetuar organizações e junções dos dados antes da mineração. As ferramentas escolhidas para esta etapa de estudo foram o *MySQL Community Server* (servidor de banco de dados) e o *MySQL WorkBench*[4] (interface gráfica) versões 8.0.20, devido ter uma interface gráfica para bancos de dados e que facilita sua utilização por profissionais que não sejam especialistas em computação. Nele é possível executar consultas SQL (Linguagem de Consulta Estruturada, do inglês *Structured Query Language*), administrar, modelar, manter e criar bases de dados em um ambiente integrado e de fácil compreensão do usuário. O *MySQL* foi usado por ser um sistema de gerenciamento de banco de dados que funciona em mais de 20 sistemas operacionais, incluindo Linux, Windows e Apple MacOs, é uma ferramenta gratuita e muito utilizado na comunidade. Estas ferramentas *MySQL* foram utilizadas para inserir planilhas eletrônicas em um banco de dados e

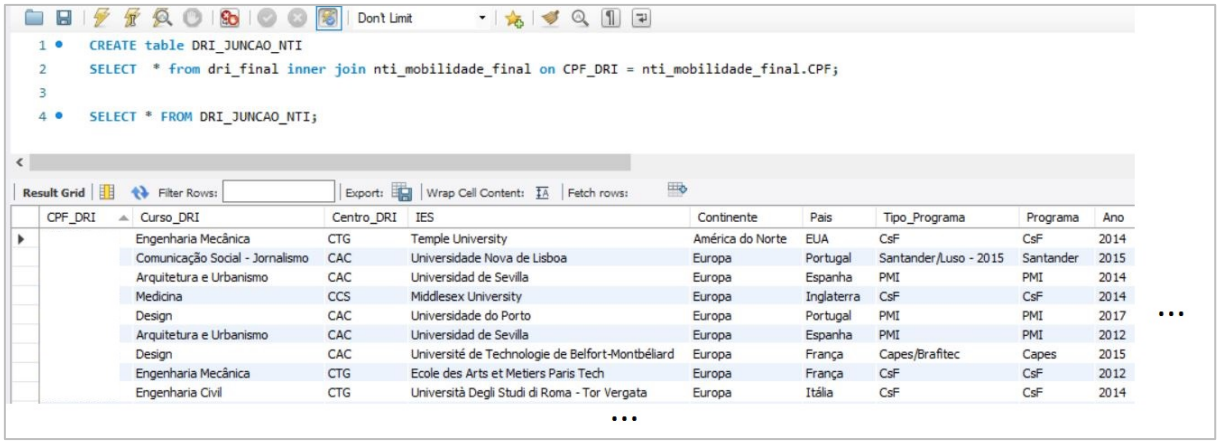
[4] https://dev.mysql.com/downloads/mysql/

realizar a junção dos dados em uma única base de dados, ou seja, transformar várias planilhas e uma só e possibilitar inseri-la em ferramentas de mineração de dados.

Para importar planilhas eletrônicas é necessário que o arquivo seja salvo com extensão .CSV separados por vírgula e para não ocorrer erros os nomes dos atributos, que são as primeiras linhas das colunas nas planilhas, não podem conter espaços, cedilhas e acentos. Por isso que estamos demonstrando os nomes de atributos utilizando o caractere _ (sublinha) nos nomes dos atributos, como por exemplo DESEMPENHO_ACADEMICO. Portanto as planilhas salvas com extensão .CSV são inseridas no *MySQL WorkBench*, selecionadas e com o uso do comando INNER JOIN, escolhendo um atributo que faz parte das duas ou mais planilhas que serão unidas. No nosso caso usamos o atributo CPF, pois é o atributo que estava presente nas planilhas estudadas. Obs.: Após este processo de junção, não foi mais necessário utilizar o atributo CPF.

Segue abaixo Figura 12 da tela do *MySQL Workbench* executando Script, ou seja, conjunto de instruções para que uma função seja executada em determinado aplicativo, com a utilização do comando para a criação de um banco de dados unindo duas planilhas.

Figura 12 - Tela do MySQL WorkBench executando Script



Fonte: Elaborado pelo autor (2020).

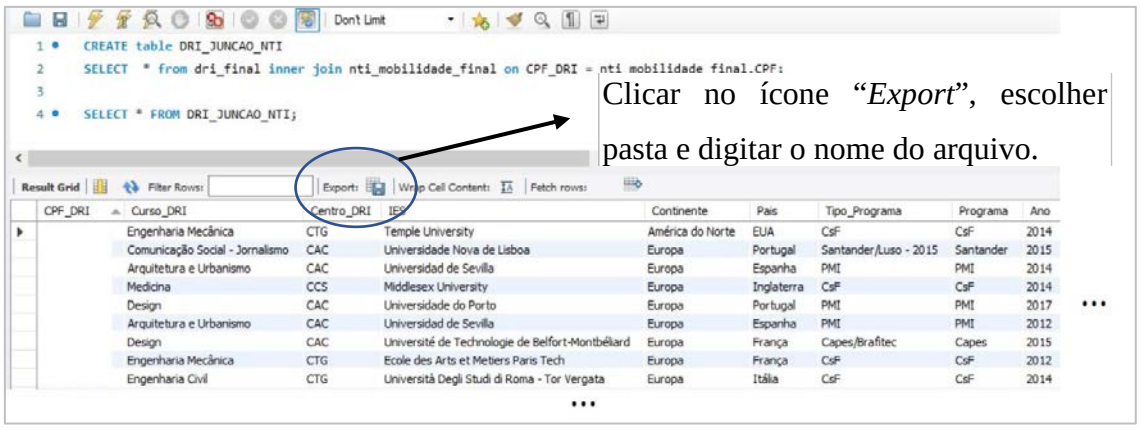
Detalhes das linhas de comandos do script:

```
CREATE TABLE dri_juncao_nti
SELECT * FROM dri_final INNER JOIN nti_mobilidade_final
ON cpf_dri = nti_mobilidade_final.cpf
```

O detalhamento deste script utilizado serviu para a criação de uma nova planilha de dados (*table*) com o nome DRI_JUNCAO_NTI, selecionando (*select*) duas planilhas dri_final e nti_mobilidade_final, onde em ambas planilhas contém o atributo CPF.

Em seguida foi exportado o resultado desta junção e salva em arquivo. Sendo também um arquivo com extensão .CSV, como demostrado na Figura 13 a seguir.

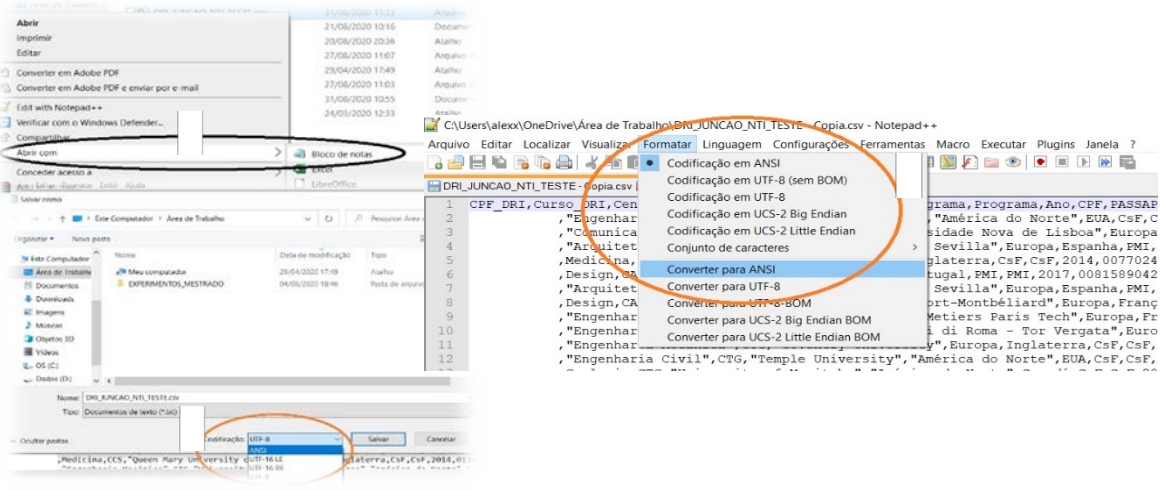
Figura 13 - Exportando resultado de junção no MySQL WorkBench



Fonte: Elaborado pelo autor (2020).

O arquivo gerado pelo *MySQL WorkBench* vem por padrão com codificação binaria dos caracteres em *UTF-8* e para que o *Weka* reconheça tal arquivo, sem que ocorra erros, precisamos modificar tal codificação para *ANSI*. Para isto, basta abrir (1) o arquivo a ser modificado usando o "Bloco de Notas", escolher o comando "Salvar como" e modificar a codificação de *UTF-8* para *ANSI* (2). Ou abrir o arquivo usando o software livre Notepad++[5] versão 7.8.6, clicando na aba formatar e escolhendo a opção "Converter para *ANSI*" (3) como demonstrado na Figura 14 a seguir.

Figura 14 - Modificando codificação de *UTF-8* para *ANSI*



Fonte: Elaborada pelo autor, 2020.

[5] http://notepad-plus-plus.rog

As Figuras 13 e 14 foram incluídas para que não ocorra dúvidas nos procedimentos realizados e por serem atividades atípicas.

Caso seja necessário voltar para corrigir algo no arquivo, incluir uma função ou criar um novo atributo, como aconteceu algumas vezes neste estudo, a melhor solução foi utilizar o editor de planilha eletrônica Calc do pacote de software livre LibreOffice[6]. Para abrir o arquivo .CSV no Calc, deve-se utilizar o conjunto de caracteres Europa ocidental (ISO-8859-1), idioma Padrão – Português (Brasil), opções de separadores marcado em separados por vírgula, delimitador de texto por " (aspas) e deixando marcado formatar campos entre aspas como texto. Estas opções são acessadas quando escolhe salvar como e marca a opção editar as configurações do filtro que está abaixo do nome e tipo de arquivo (grifo nosso). O arquivo para ser trabalhado no

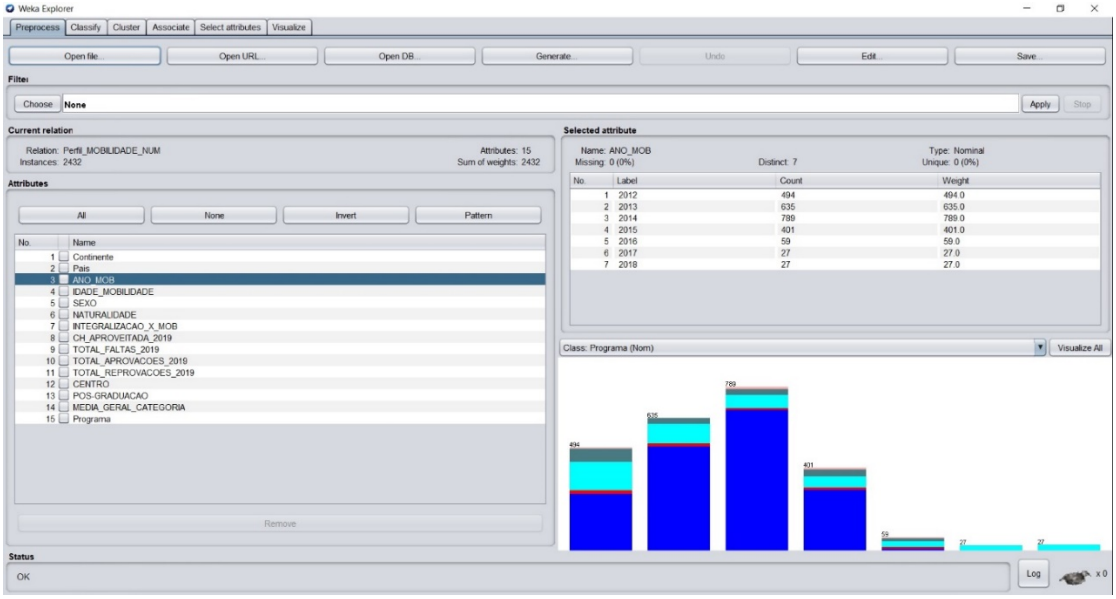
Para finalizar esta fase vamos descrever como inserir o arquivo que foi gerado no *MySQL WorkBench* na ferramenta de mineração *Weka*[7] (*Waikato Enviroment for Knowledge Analysis*) versão 3.8.4. O *Weka* é uma ferramenta desenvolvida na Universidade de Waikato na Nova Zelandia em 1993 usando Java[8]. Segundo o site oficial do *Weka*, ele é um software de aprendizagem de máquina, usado para mineração de dados e de código aberto. Um software de código aberto tem licenciamento livre, acessível a toda comunidade e sem a necessidade de pagar alguma taxa. O *Weka* tem inúmeras vantagens em relação a outras ferramentas de mineração de dados, a primeira é ser grátis, basta acessar seu site, baixar o arquivo instalador e começar a usar. Outra vantagem é que seu ambiente de relação com o usuário (interface) é gráfico com menus de opções de fácil acesso, sua aplicação não necessita o conhecimento de programação e em seu próprio site é possível ter acesso livre a vídeos, cursos, livros e documentos sobre a utilização da ferramenta.

Após todo o processo de preparação e a criação do arquivo de extensão .CSV, é possível acessar a ferramenta de mineração de dados *Weka*, abrir o arquivo para a análise dos dados, realização dos experimentos de mineração e obter a descoberta de conhecimentos. A Figura 15 apresenta a tela inicial do *Weka* com um arquivo .CSV aberto.

---

[6] Disponível em: <https://pt-br.libreoffice.org/>. Acesso em: 26 ago. 2020.
[7] Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 26 ago. 2020.
[8] Disponível em: <https://www.java.com/pt_BR/>. Acesso em: 26 ago. 2020.

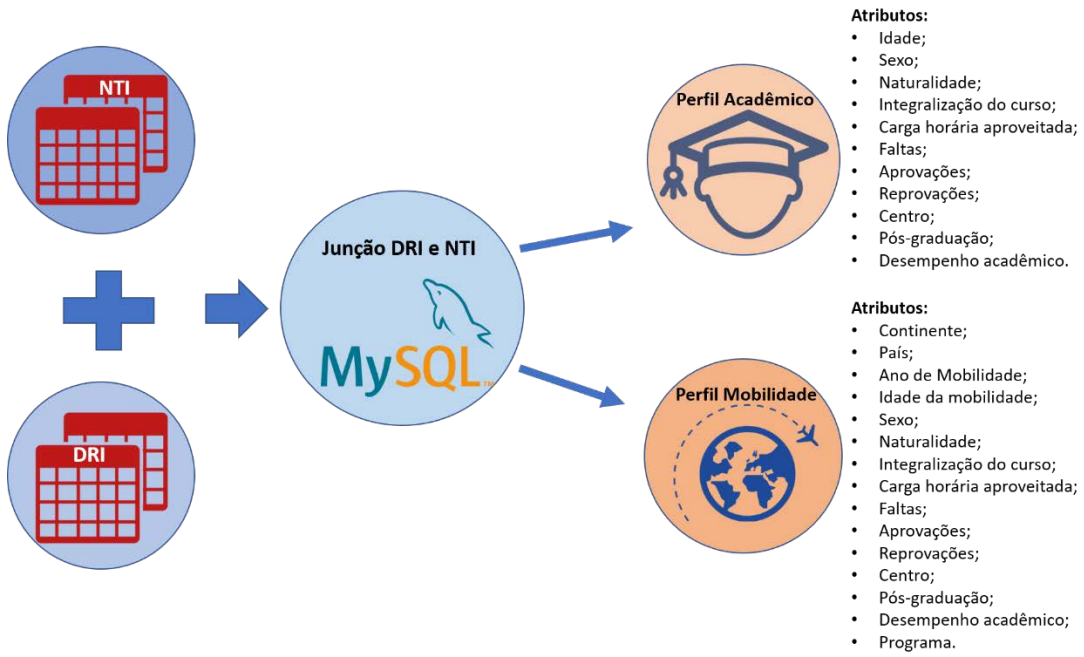
Figura 15 - Tela inicial do *Weka*



Fonte: Elaborado pelo autor (2020).

O resultado da fase de preparação consiste em finalizar a construção do conjunto de dados que será utilizado na próxima fase, a fase de mineração dos dados. Sendo um dos passos mais importante ter organizado os atributos, feito tratamentos para evitar vieses ou campos nulos, criado a base de dados para mineração e separados em perfis para análises futuras. A seguir segue Figura 16 que ilustra didaticamente a fase de preparação.

Figura 16 - Junção das planilhas, organização de atributos e criação de perfis



Fonte: Elaborado pelo autor (2020).

A criação dos perfis foi feita associando dois conjuntos de atributos de acordo com cada perfil criado. Os perfis escolhidos são pertinentes ao estudo deste trabalho sobre alunos e mobilidade internacional estudantil, ou seja, perfil acadêmico e perfil mobilidade. A planilha da DRI continha nove atributos e 2.432 alunos, a planilha do NTI continha 29 atributos e 42.280 alunos e que geraram uma base de dados unidade de 38 atributos e 2.432 alunos. Ao organizar os atributos, eliminar os que estavam duplicados e os que não eram foco do estudo, ao final desta etapa ficaram 15 atributos que são analisados nas próximas fases.

3.4 Experimentos de Mineração de Dados

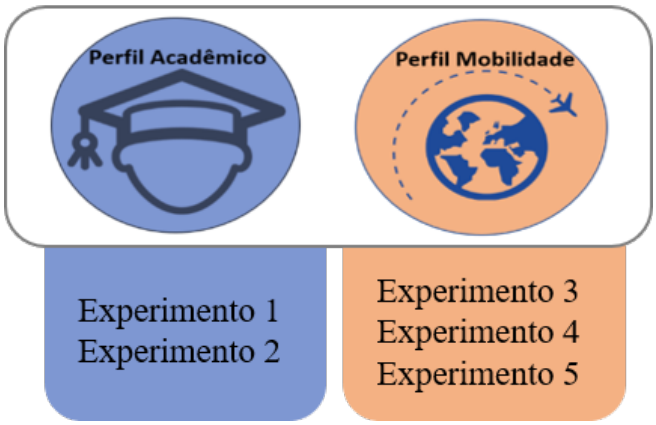
A fim de encontrar conhecimentos sobre os perfis dos estudantes da UFPE que fizeram mobilidade (intercâmbio) foram aplicados dois tipos de aprendizagem de máquina (*Machine Learnning*): (I) Aprendizado supervisionado e (II) Aprendizado não supervisionado. Para o aprendizado supervisionado foi aplicado a técnica de classificação chamada arvore de decisão por meio do algoritmo J48. Para o aprendizado não supervisionado foram aplicadas as técnicas de associação e agrupamento (*clustering*) por meio dos algoritmos *Apriori* e *SimpleKMeans* respectivamente.

O resultado da fase de preparação nos retornou duas bases de dados para mineração, Perfil Acadêmico e Perfil Mobilidade com os atributos demonstrados na seção anterior.

Foram realizados cinco experimentos, sendo dois para o perfil acadêmico e três para o perfil mobilidade. O perfil mobilidade contém os atributos do perfil acadêmicos, somados aos atributos que envolvem informações sobre a mobilidade e por isso que foi feito um experimento a mais. No perfil acadêmico o experimento 1 foi aplicado o tipo de aprendizado de máquina supervisionado, utilizando a técnica de classificação árvore de decisão e o algoritmo J48. E no experimento 2 foi aplicado o tipo de aprendizado de máquina não supervisionado, utilizando a técnica de associação e o algoritmo *Apriori*. Para o perfil mobilidade os experimentos 3 e 4 foram aplicados o mesmo que nos experimentos 1 e 2, aprendizado de máquina supervisionado, técnica de classificação árvore de decisão e o algoritmo J48 e aprendizado de máquina não supervisionado, técnica de associação e o algoritmo *Apriori*. E o experimento 5 foi aplicado o tipo de aprendizado de máquina não supervisionado, utilizando a técnica de agrupamento e o

algoritmo *SimpleKMeans*. Este último se fez necessário apenas no perfil mobilidade pois busca dividir os dados em grupos e a partir desta divisão que se faz uma análise a fim de encontrar características distintas sobre os grupos encontrados pelo algoritmo. O perfil mobilidade foi escolhido para ser aplicado a técnica de agrupamento, já que este perfil tem mais atributos e consequentemente mais dados, sendo repetitivo aplicar esta técnica em ambos perfis.

Figura 17 - Perfis e experimentos



Fonte: Elaborado pelo autor (2020).

Os experimentos feitos foram em quantidades maiores que os descritos acima, cerca de 20 experimentos. Porém foram aproveitados àqueles que geraram resultados dentro de parâmetros de confiança e/ou acurácia esperados, com relevância e que contribuíram para o estudo. Para este estudo foram usados acurácia acima de 85%, e confiança próximas de 90% como parâmetros válidos. Acurácia de 85% e confiança de 90% representam critérios rígidos para a mineração e por isso que muitos experimentos não foram aproveitados. Um exemplo prático obtido foi usando o algoritmo de arvore de decisão que teve acurácia próxima de 60%, ou seja, neste modelo erra 40% das vezes. Em outro exemplo prático usando o algoritmo *Apriori* que não mostrava nenhuma regra com confiança em 90%. Para cada um dos perfis existe duas bases de dados, um com dados nominais e outra com dados nominais e numéricos. Devido a este detalhe, foram feitos mais experimentos, dos que estão indicados na Figura 17. Isso se deu pelo fato de serem analisados diversos parâmetros dos algoritmos e também a variedade de opções de testes de avaliação de desempenho apresentados na seção 2.3: *Use training test set, Supplied test set, Cross-validation e Percentage split*.

Na fase de mineração de dados são feitos inúmeros experimentos usando vários algoritmos a fim de verificar seus resultados, avaliar quais algoritmos retornam resultados

mais relevantes, podendo configurar seus parâmetros e permitir a comparação de diferentes estratégias de aprendizagem.

Os experimentos foram realizados de forma que a ferramenta retorne seus resultados e apresentados juntos com suas métricas. Na seção 4.3 sobre descoberta de conhecimento que está informado a interpretação dos resultados dos experimentos. Pois lá está uma compilação dos dados encontrados que geraram novos conhecimentos e mais contribuíram para este estudo.

3.4.1 Qualidade e confiabilidade dos dados

Segundo Sampieri (2013, p. 218), a coleta de dados implica preparar um plano com procedimentos que levam a reunir dados com um propósito específico. Neste plano deve-se atender quatro critérios: qual fonte; localização; método de coleta e forma de preparo para análise. A reunião dos dados deve reunir três requisitos essenciais: confiabilidade, validade e objetividade. Confiabilidade é o grau em que os dados produzem resultados consistentes e coerentes. Validade é o nível em que os dados realmente mensuram seu ponto alvo. E a objetividade refere-se ao grau de permeabilidade dos dados à influência dos vieses e/ou tendências do pesquisador. Ao aplicar uma padronização na avaliação dos resultados procura-se reduzir ao mínimo possível influências características de cada pesquisador. Tais critérios foram aplicados na prática para garantir imparcialidade da pesquisa e garantir qualidade e confiabilidade dos dados.

Ao trabalhar com mineração de dados a organização e limpeza dos dados requer seu devido cuidado e segundo Harrison (2020, p. 49), não há resposta única sobre como tratar dados ausentes. Esta ausência pode implicar em vieses, respostas aleatórias ou falsas correlações.

Para Harrison existem diversas maneiras de lidar com os dados ausentes:

- Remover qualquer linha com dados ausentes;
- Remover qualquer coluna com dados ausentes;
- Imputar dados (média, mediana, máximo ou mínimo) aos valores ausentes, quando possível;
- Criar uma coluna para informar que os dados estavam ausentes.

Como exemplo mais marcante neste estudo, ocorreu um episódio que vale ressaltar. Ele foi mencionado na seção 3.3, logo após a Figura 11. Onde houve uma redução de

4234 para 2432 na lista de alunos estudados. Um dos principais motivos foi que na lista inicial (4234) continha a coluna CPF, que neste caso é o identificador principal para buscar dados em outra base de dados e efetuar a junção de mais variáveis antes não conhecidas. Tal redução foi necessária, pois existiam dados ausentes (campos um branco) na coluna CPF, impossibilitando aproveitar o restante da linha e forçando a exclusão destes alunos. Apesar da redução o estudo seguiu seu propósito.

3.4.2 Medidas de Desempenho

Em se tratando de métricas, que são as medidas de desempenho dos algoritmos de classificação do aprendizado de máquina supervisionado, tem-se a matriz de confusão.

No aprendizado de máquina supervisionado existem diversas medidas de desempenho que podem ser utilizadas para avaliar os modelos gerados. Dentre as métricas utilizadas nesse trabalho, tem-se: Taxa Acurácia, Taxa de Erro, Precisão (*Precision*), *Recall* (Cobertura, Lembrança ou Sensibilidade), *F-Measure*, entre outras. Tais métricas são obtidas por meio da Matriz de Confusão, apresentada abaixo, no Quadro 5 - Matriz de Confusão.

Quadro 5 - Matriz de Confusão

	Valor previsto classe positiva↓	Valor previsto classe negativa↓
Valor real classe positiva→	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Valor real classe negativa→	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaborado pelo autor (2020).

A Taxa de Acerto (Acurácia) avalia o modelo treinado de uma maneira geral, informando o percentual de instâncias classificadas corretamente sobre o total de instâncias. Já a Taxa de Erro informa o percentual de instâncias classificadas incorretamente.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \qquad Taxa\ de\ Erro = \frac{FP + \mathrm{FN}}{VP + VN + FP + FN}$$

A Precisão (*precision*) refere-se às amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas. A Precisão mede o quão eficiente é o modelo na previsão de valores positivos.

$$Precision\ (Precisão) = \frac{VP}{VP + FP}$$

Recall, também conhecida como Cobertura, Sensibilidade, Lembrança ou taxa de verdadeiro positivo (*TP Rate*), refere-se às amostras positivas classificadas corretamente sobre o somatório do valor real da classe positivas. A Cobertura descreve a taxa de acerto real para a previsão de verdadeiros positivos.

$$Recall\ (\text{Cobertura, Lembrança ou Sensibilidade}) = \frac{VP}{VP + FN}$$

F-Measure, também conhecida como *F1-Score*, utiliza uma combinação de *Precision* e *Recall*, calculando uma média harmônica ponderada. Portanto, *F-Measure* leva em consideração tanto os FP quanto os FN, sendo indicada na tentativa de não enviesar a métrica para as classes mais populosas, ou seja, é indicada quando há distribuição de classe desigual (desbalanceamento da classe).

$$F\_Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

A Área ROC, do inglês *Area Under the Curve - Receiver Operating Characteristic*, relaciona a Sensibilidade (ou taxa de verdadeiro positivo) e a Especificidade (ou taxa de verdadeiro negativo) = VN / (VN + FP). Na prática, a Sensibilidade e Especificidade variam em direções opostas, ou seja, quando o modelo é muito sensível, tende a gerar muitos Falso Positivo (FP), e quando o modelo é muito específico, tende a gerar muitos Falso Negativo (FN). O valor do AUC varia de 0,0 até 1,0, desta forma um modelo com previsões 100% erradas tem uma AUC de 0, enquanto um modelo com previsões 100% corretas tem uma AUC de 1. Portanto quanto mais perto de 1 o valor da Área ROC, melhor é o modelo.

Visando exemplificar tais métricas, tem-se a matriz de confusão abaixo (ver Tabela 1) referente ao modelo gerado para a base de dados sobre condições climáticas para jogo de tênis, apresentado na Figura 7 (Seção 2.3).

Tabela 1 - Matriz de Confusão do exemplo sobre condições climáticas para jogo de tênis.

<table>
<tr><td>

```
=== Confusion Matrix ===

 a b   <-- classified as
 9 1 | a = sim
 1 5 | b = nao
```

</td><td>

	Valor previsto (SIM) ↓	Valor previsto (NÃO) ↓
Valor real (SIM) →	VP (9)	FN (1)
Valor real (NÃO) →	FP (1)	VN (5)

</td></tr>
<tr><td colspan="2">

```
=== Summary ===

Correctly Classified Instances          14               87.5      %
Incorrectly Classified Instances         2               12.5      %
```

</td></tr>
</table>

Fonte: Elaborado pelo autor (2020).

A acurácia do referido modelo foi de 87,5%, e a taxa de erro 12,5%. Os cálculos estão apresentados abaixo:

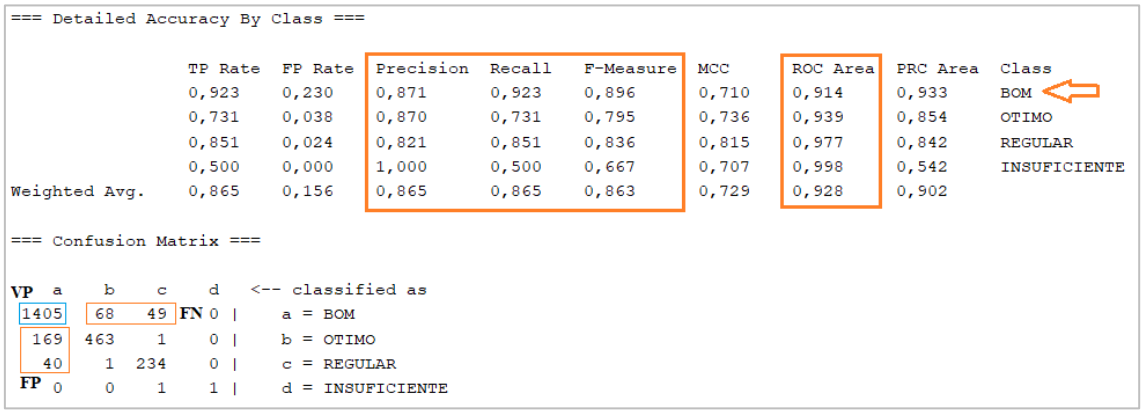
$$Acurácia = \frac{9+5}{9+5+1+1} = \frac{14}{16} = 0{,}875 = 87{,}5\%;$$

$$Taxa\ de\ Erro = \frac{1+1}{9+5+1+1} = \frac{2}{16} = 0{,}125 = 12{,}5\%.$$

Interpretação: *Correctly Classified Instances* (Acurácia) significa instâncias classificadas corretamente e *Incorrectly Classified Instances* (Taxa de Erro) são as instâncias classificadas incorretamente. As métricas calculadas de cada são de acordo a soma dos números apresentados, 14 acertos, 2 erros e sua soma é 14+2=16. Consequentemente seus percentuais são determinados pela fração de seus números divididos pela soma, 14/16=0,875 (87,5%) e 2/16=0,125 (12,5%).

Para exemplificar as demais métricas, segue na Figura 18 o resultado de um modelo gerado neste estudo.

Figura 18 - Métricas *Precision, Recall, F-measure e ROC Area* do exemplo



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0,923     0,230     0,871       0,923   0,896       0,710    0,914     0,933     BOM
                 0,731     0,038     0,870       0,731   0,795       0,736    0,939     0,854     OTIMO
                 0,851     0,024     0,821       0,851   0,836       0,815    0,977     0,842     REGULAR
                 0,500     0,000     1,000       0,500   0,667       0,707    0,998     0,542     INSUFICIENTE
Weighted Avg.    0,865     0,156     0,865       0,865   0,863       0,729    0,928     0,902

=== Confusion Matrix ===

VP  a    b    c    d   <-- classified as
 1405   68   49 FN 0 |   a = BOM
  169  463    1    0 |   b = OTIMO
   40    1  234    0 |   c = REGULAR
FP  0    0    1    1 |   d = INSUFICIENTE
```

Fonte: Elaborado pelo autor (2020).

Os cálculos correspondentes às métricas utilizadas no exemplo da figura anterior são:

$$Precision\ (\frac{VP}{VP+FP}) = \frac{1405}{1405+169+40} = 0{,}871\ (classe\ BOM);$$

$$Recall\ \left(\frac{VP}{VP+FN}\right) = \frac{1405}{1405+68+49} = 0{,}923\ (classe\ BOM);$$

$$F\_Measure(2 * \frac{Precision*Recall}{Precision+Recall}) = 2 * (\frac{0{,}871*0{,}923}{0{,}871+0{,}923}) = 0{,}896\ (classe\ BOM).$$

4 Resultados e Discussão

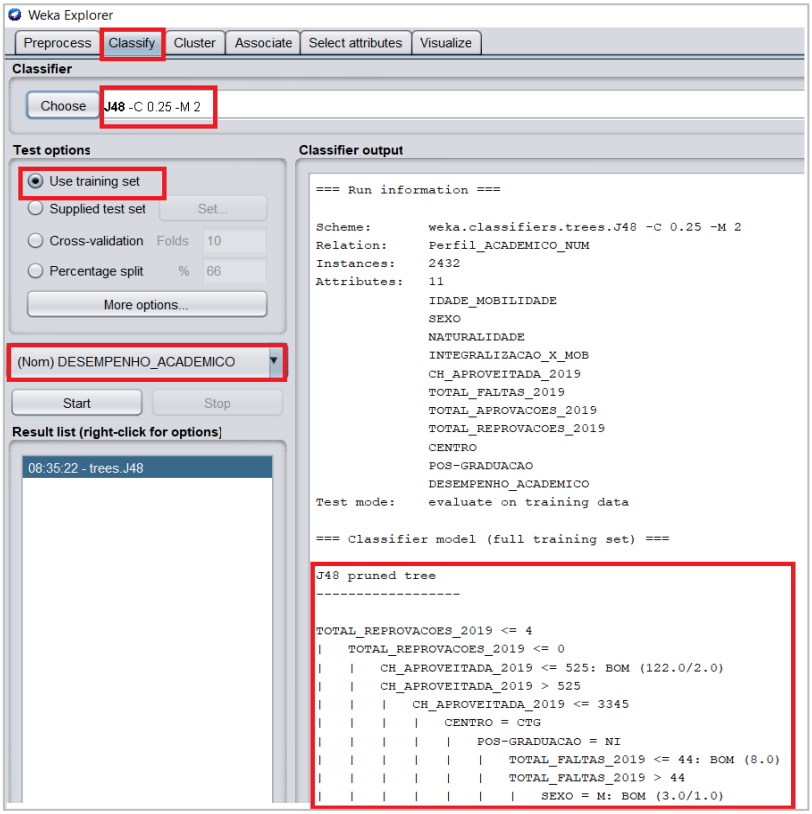
4.1 Mineração de Dados do Perfil Acadêmico

Esta seção apresenta os resultados dos dados analisados nos experimentos aplicados no perfil acadêmico.

4.1.1 Experimento 1 – Classificação (Árvore de Decisão J48)

Os experimentos buscam encontrar o melhor percentual de instâncias classificadas corretamente no perfil acadêmico. Usando o aprendizado supervisionado, a técnica de arvore de decisão e o algoritmo J48 com a opção de teste configurado em *Training set* e que obteve o melhor resultado. O perfil em questão contem 11 atributos, sendo o atributo “desempenho acadêmico” escolhido como classe. A Figura 19 mostra a tela da ferramenta com os resultados, lista de atributos e os pontos mencionados em destaque.

Figura 19 - Tela com resultado da execução do algoritmo



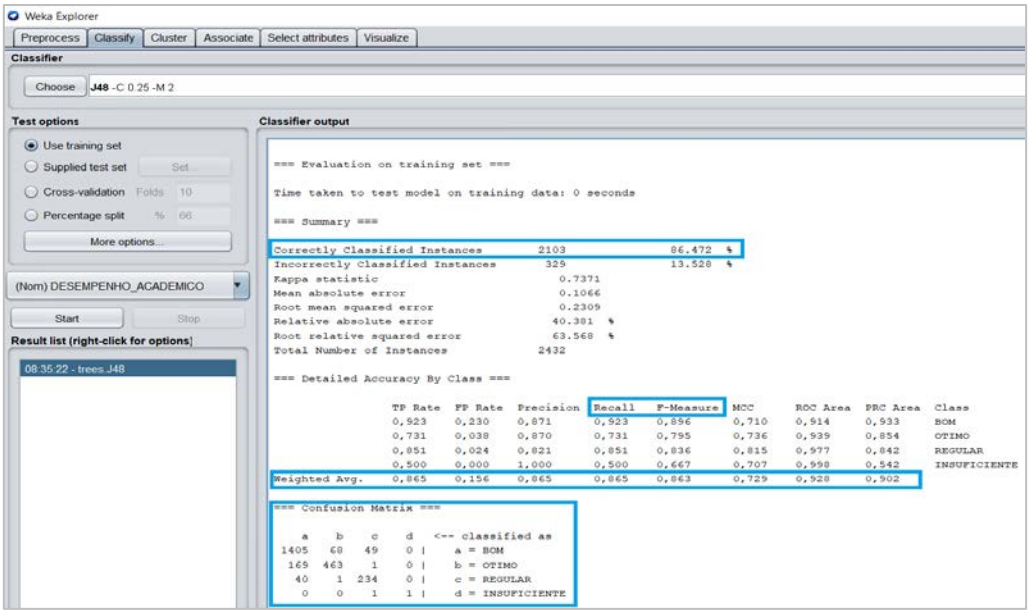
Fonte: Elaborado pelo autor (2020).

- Obs.1: Foram utilizados três tipos de testes disponíveis no *Weka* para o algoritmo J48: *Training Test, Cross-validation e Percentage split*. Nas bases de dados para o perfil acadêmico com atributos numéricos e com atributos nominais.

O algoritmo gera arvore textual, como demarcado na parte inferior da Figura 19, onde inicia com um atributo e uma regra como o nó principal da arvore de decisões, que neste caso foi “total de reprovações menor igual a 4 (<=4)”. Este nó principal significa a raiz de uma arvore e que são divididas em vários outros nós e até chegando a outra extremidade que representam as folhas. A ferramenta também pode gerar a representação gráfica da arvore de decisão e nela é possível visualizar a raiz (nó principal) no topo, os outros nós que dividem em ramos ou galhos e as extremidades como se fossem as folhas, como demonstrado no Apêndice A. ambas apresentam valores próximos de 1, o que indica um modelo com bom desempenho.

Figura 20 aponta os destaques das métricas do experimento 01, onde demonstra um percentual de instâncias classificadas corretamente (acurácia) de 86,472%. Este percentual é calculado por meio da matriz confusão, que encontramos na parte inferior da imagem, ou seja, quantidade de acertos por classe (bom, ótimo, regular e insuficiente) dividido pela somatória de todos acertos e erros (1405+463+234+1)/2432=86,472%. As outras métricas importantes de mencionar são as médias (*Weighted average*) do *F-measure*=0,863 (média harmônica entre precisão e *Recall)* e *ROC área*=0,928 (representa sensibilidade e especificidade), ambas apresentam valores próximos de 1, o que indica um modelo com bom desempenho.

Figura 20 - Métricas do experimento 1



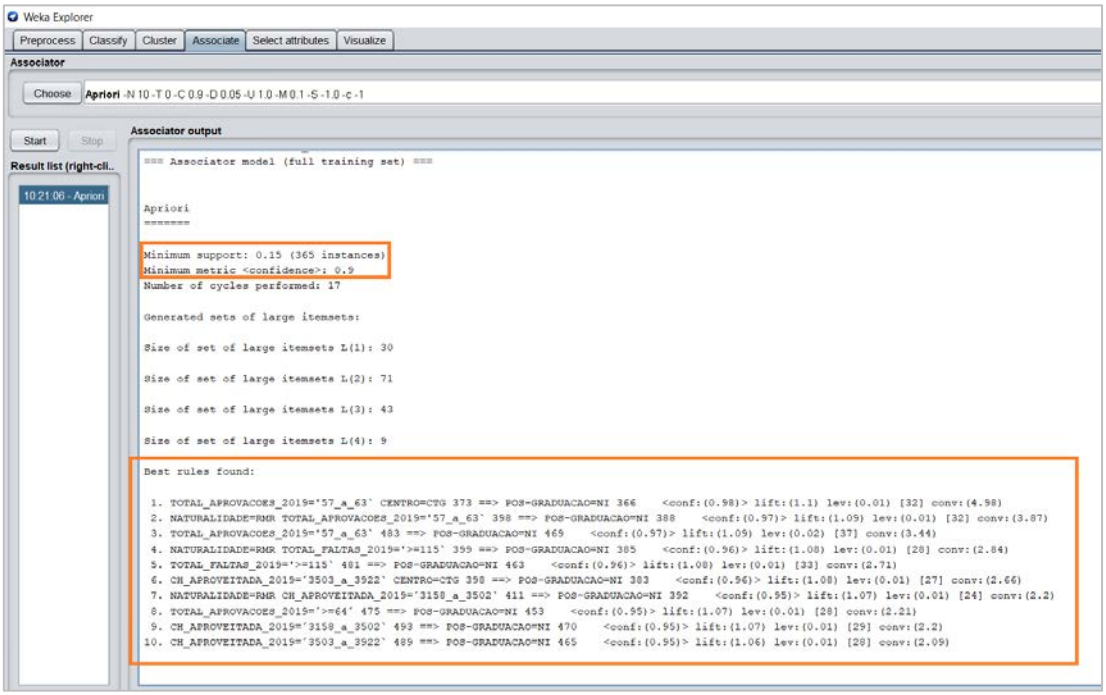
Fonte: Elaborado pelo autor (2020).

Na seção 4.2 são apresentados e discutidos os padrões/regras e a descoberta de conhecimento obtidos pelo algoritmo J48.

4.1.2 Experimento 2 – Associação (Apriori)

Ao rodar o algoritmo *Apriori*, o primeiro resultado retornou regras com viés para o atributo POS-GRADUACAO e com o valor “NI”, que significa “não informado”, ou seja, na fase de pré-processamento dos dados foram identificados alunos que não havia a informação se fez ou não pós-graduação e por ser uma quantidade bastante significativa, 2172 “NI” de 2432 que significa 89,3% dos alunos da relação de alunos estudada. O viés é confirmado pois todas as regras encontradas pelo algoritmo ficaram com o atributo em questão inseridos como demonstrado na Figura 21 abaixo. O algoritmo *Apriori* retorna itens mais frequentes e o valor “NI” ter frequência muito alta, este atributo enviesou as regras.

Figura 21 – Primeiro resultado experimento 2

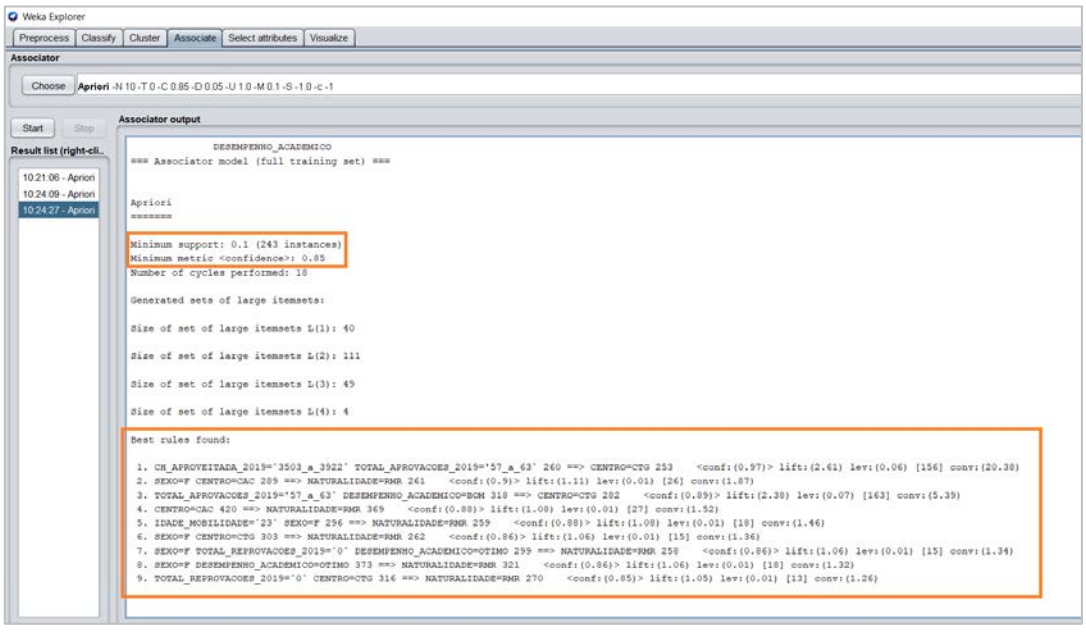


Fonte: Elaborado pelo autor (2020).

As métricas apresentadas do primeiro resultado do experimento 2 são: *support* (suporte) de 0,15 em 365 instâncias, *confidence* (confiança) de 0,90 (90%) e a primeira regra apresentou confiança de 98% (conf:(0.98)).

A técnica de associação do algoritmo *Apriori* é de aprendizado de máquina não supervisionado, porém não é escolhido um atributo como classe e se um atributo está presente em todas regras encontradas pelo algoritmo isto demonstra um viés. Para evitar o ocorrido no primeiro resultado do experimento, e verificar se o algoritmo é capaz de encontrar regras importantes com os outros atributos, foi retirado o atributo POS-GRADUACAO, configurada a confiança para 0,85, porque no padrão de 0,9 só retornou duas regras, o que é muito abaixo das expectativas e com confiança em 0,85 foram retornadas 9 regras que se tornaram relevantes para o estudo. Incluindo as duas regras com confiança acima de 0,9 e mais sete regras com confiança entre 0,89 e 0,85. Porém as regras encontradas com confiança mais próximas de 0,9 foram consideradas na interpretação das regras na seção de descoberta de conhecimento. Esta redução foi feita apenas em caráter experimental para verificar a existência de alguma regra muito relevante, o que não ocorreu neste caso. Para esta nova configuração o segundo resultado encontrado é demonstrado na Figura 22 abaixo.

Figura 22 - Segundo resultado experimento 2



Fonte: Elaborado pelo autor (2020).

As métricas apresentadas do segundo resultado do experimento 2 são: *support* (suporte) de 0,1 em 243 instâncias, *confidence* (confiança) de 0,85 (90%) e a primeira regra apresentou confiança de 97% (conf:(0.97)).

4.2 Descoberta de Conhecimento dos Experimentos 1 e 2

Nesta seção foi analisado os resultados dos experimentos 1 e 2 que foram extraídos dos algoritmos J48 e *Apriori* e compreender o perfil acadêmico (11 atributos) a fim de descobrir conhecimento. Esses experimentos apresentaram dezenas (30) de regras, dentre elas foram selecionadas as que tinham maior índice de assertividade e dando preferência as que ficaram mais próximas de 100% de acurácia ou confiança. Seguem abaixo as explicações das principais delas:

- **Regra 1**: Se aluno não teve reprovações, carga horária do curso menor ou igual a 3345 (provável conclusão do curso), pertencente ao CTG, com desempenho acadêmico ótimo ingressaram no mestrado e/ou doutorado.
- **Regra 2**: Se aluno não teve reprovações, carga horária do curso menor ou igual a 3345 (provável conclusão do curso), pertencente ao CTG e total de faltas menor ou igual a 44 então seu desempenho acadêmico foi bom.
- **Regra 3**: Se aluno não teve reprovações, carga horária do curso entre 525 e 3345 horas, pertencente ao CAC e teve aprovações menor ou igual a 45 então seu desempenho acadêmico foi ótimo.
- **Regra 4**: Se aluno não teve reprovações, carga horária do curso entre 525 e 3345 horas, pertencente ao CCB e teve aprovações menor ou igual a 39 então seu desempenho acadêmico foi bom.
- **Regra 5**: Se aluno não teve reprovações, carga horária do curso entre 525 e 3345 horas, pertencente ao CCJ e teve faltas menor ou igual a 26 então seu desempenho acadêmico foi ótimo.
- **Regra 6:** Se aluno não teve reprovações, carga horária do curso entre 2040 e 3345 horas, pertencente ao CCJ, teve mais de 26 faltas, não integralizou o curso entre 1 a 4 anos após a mobilidade e é do sexo feminino então seu desempenho acadêmico foi ótimo. (Regra muito específica e boa para investigar in loco).
- **Regra 7:** Se aluno teve entre 5 e 7 reprovações, aprovações menor ou igual a 81, pertencente ao CAC e não integralizou o curso 7 anos após a mobilidade então seu desempenho acadêmico foi bom. (Regra muito específica e boa para investigar in loco).

- **Regra 8:** Se aluno teve entre 5 e 7 reprovações, aprovações menor ou igual a 81, pertencente ao CA então seu desempenho acadêmico foi bom. (Regra muito específica e boa para investigar in loco).
- **Regra 9:** Se aluno teve mais de 12 reprovações, aprovações menor ou igual a 81 e pertencente ao CTG então seu desempenho acadêmico foi regular. (Regra com 100% de acerto e boa para investigar in loco).
- **Regra 10:** Se aluno teve mais de 12 reprovações, aprovações menor ou igual a 81 e pertencente ao CA então seu desempenho acadêmico foi regular. (Regra com 100% de acerto e boa para investigar in loco).
- **Regra 11:** Se aluno teve carga horária do curso entre 3503 e 3922, aprovações entre 57 e 63 então pertencente ao CTG. (253 de ocorrência com 97% de confiança).
- **Regra 12:** Se a aluna é do sexo feminino, pertencente ao CAC então é natural da região metropolitana de Recife. (261 de ocorrência com 90% de confiança).
- **Regra 13:** Se aluno teve entre 57 a 63 aprovações e desempenho acadêmico bom então são do CTG. (282 de ocorrência com 89% de confiança).
- **Regra 14:** Se aluno é do CAC então é natural da região metropolitana de Recife (369 de ocorrência com 88% de confiança).
- **Regra 15:** Se aluna do sexo feminino tinha idade de 23 anos quando fez mobilidade então é natural da região metropolitana de Recife. (259 de ocorrência com 88% de confiança).

A descoberta do conhecimento para o perfil acadêmico, como visto nas regras, teve como classe (atributo principal) o desempenho acadêmico e evidenciou sua importância em relacionar com outros atributos: aprovações, reprovações, carga horária do curso e faltas. O algoritmo ao encontrar as regras de 1 a 7 priorizou o número de reprovações no início da regra, sendo que a maioria incluiu carga horária, aprovações ou faltas, dando destaque aos alunos do centro de tecnologia e geociências (CTG), com desempenho acadêmico ótimo e que fizeram mestrado ou doutorado, comprovando que os melhores alunos ingressaram na pós-graduação. Porém os alunos com mais de 12 reprovações, do CTG e do centro do agreste (CA), mencionados nas regras 8 e 9, tiveram desempenho regular, ou seja, alunos com mais dificuldades que precisam de maior acompanhamento. E por fim ao analisar o atributo naturalidade, encontra uma disparidade entre alunos natural de Pernambuco e alunos de outros estados. Apenas 7,5% dos alunos que

participaram de mobilidade são naturais de outros estados e 92,5% do estado de Pernambuco, sendo 81,2% da região metropolitana de Recife e 11,3% do interior do estado.

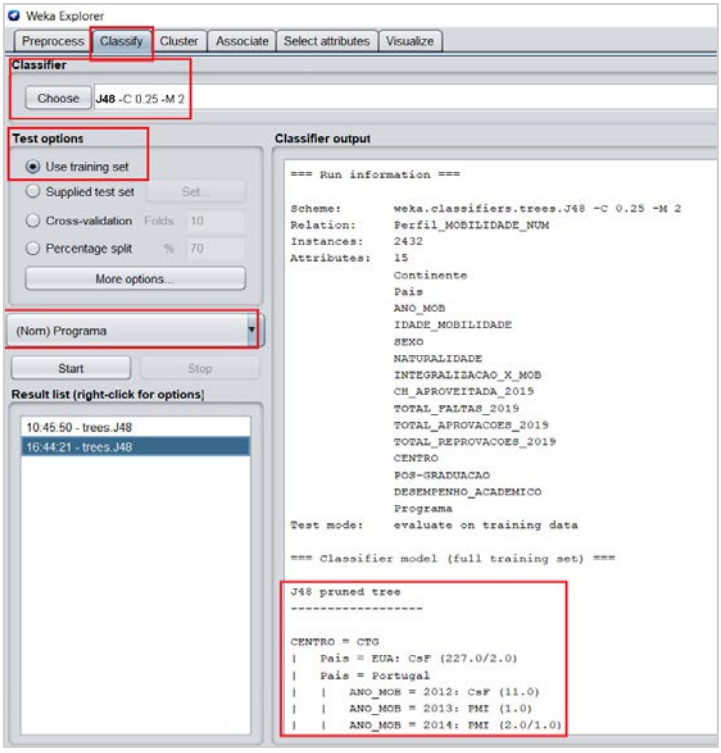
4.3 Mineração de Dados do Perfil Mobilidade

Esta seção apresenta os resultados dos dados analisados nos experimentos aplicados no perfil mobilidade.

4.3.1 Experimento 3 – Classificação (Árvore de Decisão J48)

Como no experimente 01, buscamos encontrar o melhor percentual de instâncias classificadas corretamente no perfil mobilidade. Usando o aprendizado supervisionado, a técnica de arvore de decisão e o algoritmo J48 com a opção de teste configurado em *Training set* e que obteve o melhor resultado. Sendo o atributo “programa” escolhido como classe. A Figura 23 a tela da ferramenta com os resultados, lista de atributos e os pontos mencionados em destaque.

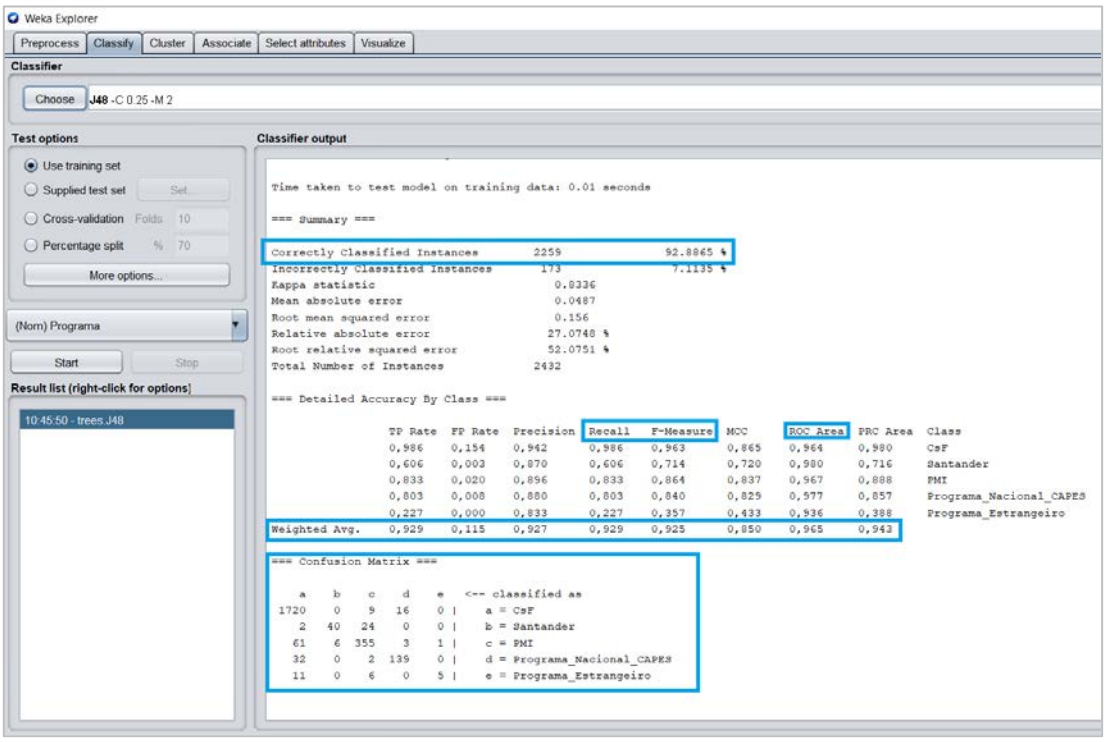
Figura 23 - Tela com resultado da execução do algoritmo



Fonte: Elaborado pelo autor (2020).

A arvore gerada pelo algoritmo, como demonstrado no Apêndice B, onde inicia com o nó principal da arvore de decisões, que neste caso foi “centro”. O atributo “centro” foi sinalizado pelo algoritmo como a raiz da arvore e as folhas, extremidade oposta a raiz, o atributo “programa” e que também é a classe do conjunto de atributos do perfil mobilidade. Figura 24, a seguir, aponta os destaques das métricas do experimento 04, onde demonstra um percentual de instâncias classificadas corretamente (acurácia) de 92,8865%. Este percentual é calculado por meio da matriz confusão (ver seção 4.3.1), que encontramos na parte inferior da figura em questão, ou seja, quantidade de acertos por classe (bom, ótimo, regular e insuficiente) dividido pela somatória de todos acertos e erros (1720+40+355+139+5)/2432=92,8865%. As outras métricas importantes de mencionar são as médias (*Weighted average*) do *F-measure*=0,925 (média harmônica entre precisão e *Recall)* e *ROC área*=0,965 (que representa sensibilidade e especificidade), ambas apresentam valores próximos de 1, o que indica um modelo com ótimo desempenho.

Figura 24 - Métricas do experimento 4



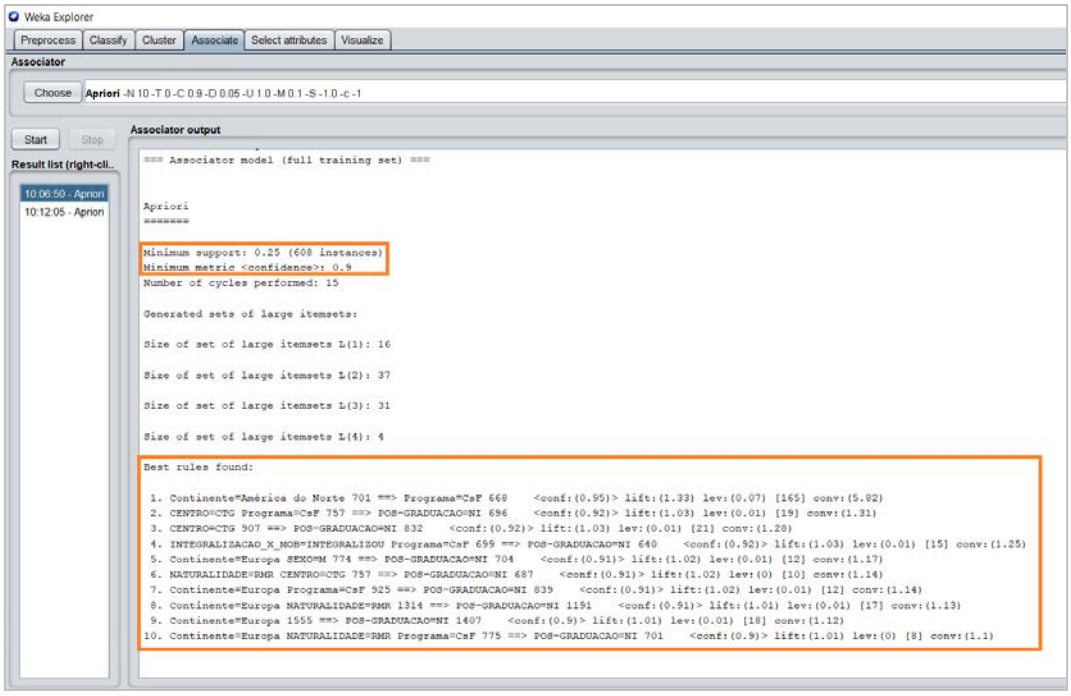
Fonte: Elaborado pelo autor (2020).

Na seção 4.4 são apresentados e discutidos os padrões/regras e a descoberta de conhecimento obtidos pelo algoritmo J48.

4.3.2 Experimento 4 – Associação (*Apriori*)

O primeiro resultado do algoritmo *Apriori* retornou regras com viés do atributo POS-GRADUACAO, presente em nove das dez regras encontradas, e com o valor “NI”, que significa “não informado”, ou seja, na fase de pré-processamento dos dados foram identificados alunos que não havia a informação se fez ou não pós-graduação e por ser uma quantidade bastante significativa, 2172 “NI” de 2432 que significa 89,3% dos alunos da relação de alunos estudada. Segue abaixo a Figura 25 com as regras encontradas.

Figura 25 - Primeiro resultado experimento 5

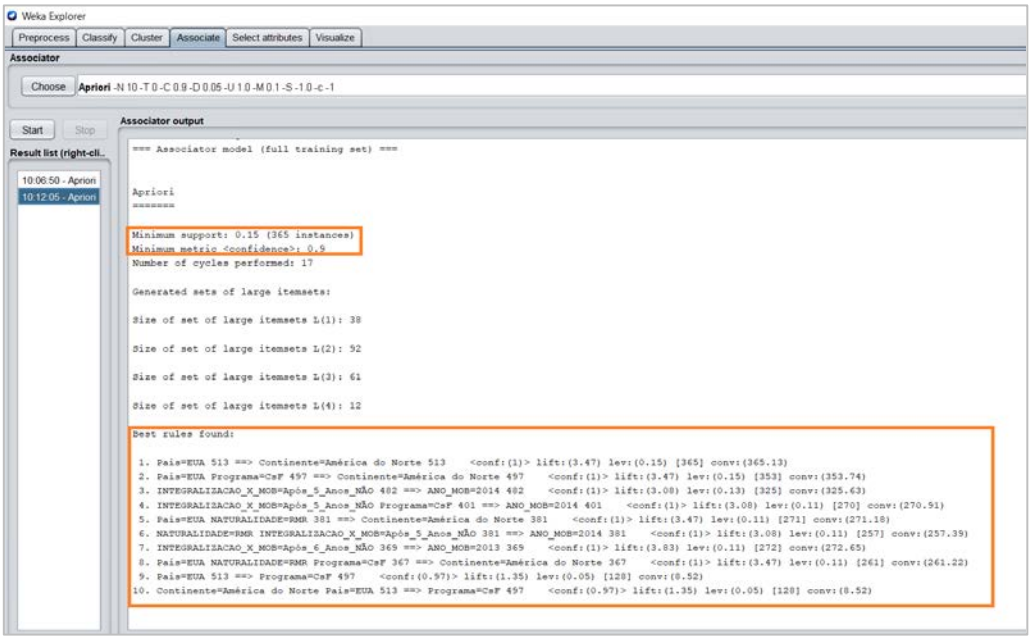


Fonte: Elaborado pelo autor (2020).

As métricas apresentadas do primeiro resultado do experimento 2 são: *support* (suporte) de 0,25 em 608 instâncias, *confidence* (confiança) de 0,90 (90%) e a primeira regra apresentou confiança de 95% (conf:(0.95)).

No segundo resultado, feito para evitar o viés ocorrido no primeiro resultado do experimento, foi retirado o atributo POS-GRADUACAO, configurada a confiança no padrão de 0,9 e o algoritmo retornou 10 melhores regras como visto da Figura 26 abaixo.

Figura 26 - Segundo resultado experimento 5



Fonte: Elaborado pelo autor (2020).

Sendo as métricas apresentadas do primeiro resultado do experimento 2 são: *support* (suporte) de 0,15 em 365 instâncias, *confidence* (confiança) de 0,90 (90%) e a primeira regra apresentou confiança de 100% (conf:(1)).

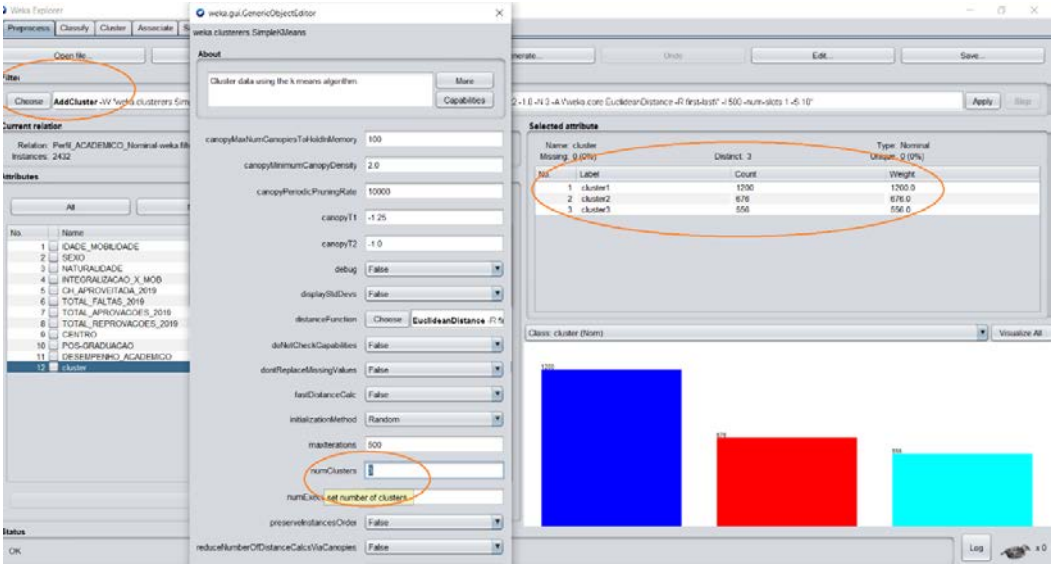
4.3.3 Experimento 5 – Clustering (Agrupamento)

O algoritmo mais utilizado é o *SimpleKmeans* e nele é possível escolher a quantidade de grupos que serão gerados. Devido ao estudo ter previsto na fase de seleção e pré-processamento três perfis para os alunos estudados, o experimento foi feito com a criação de três clusters, ou seja, a base de dados foi separada.

O uso do algoritmo de agrupamento foi feito usando o perfil mobilidade, pois nele contém todos os atributos e conjunto de dados e o agrupamento é mais recomendado usando um conjunto maior possível de atributos.

Foi aplicado tipo de aprendizado não supervisionado, técnicas de agrupamento (*clustering*) e o algoritmo *SimpleKmeans*. O algoritmo tem um filtro (*addCluster*) que adiciona um novo atributo chamado de "*cluster*" e nele podemos configurar a quantidade grupos que queremos criar demonstrados na Figura 27 a seguir.

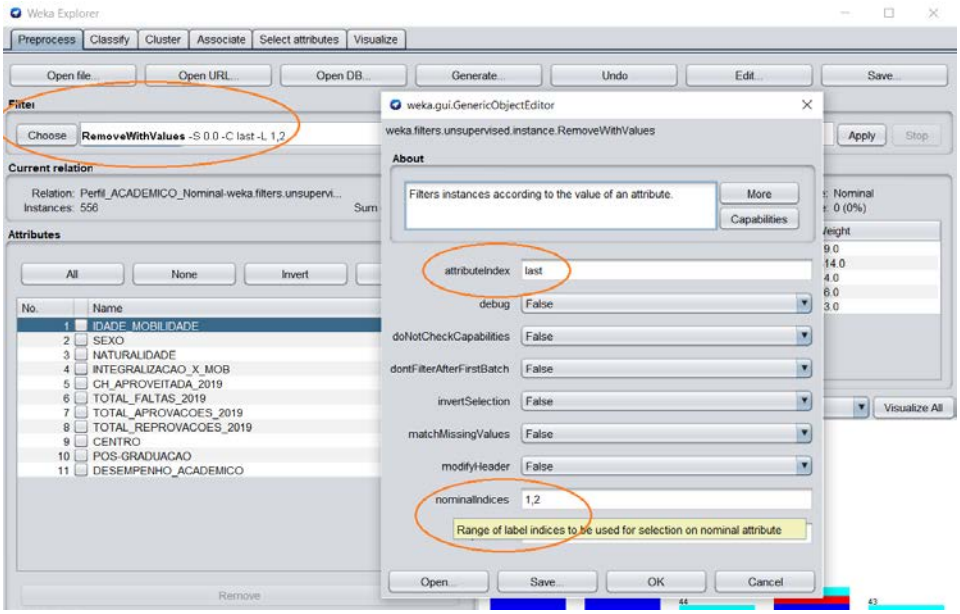
Figura 27 - Algoritmo *SimpleKmeans*



Fonte: Elaborado pelo autor (2020).

Para separar os dados por grupo (*cluster*) foi usado outro filtro chamado *RemoveWithValues*, que remove valores de acordo com o atributo escolhido e quais campos quer eliminar. E no exemplo usado da Figura 27, foi adicionado um atributo *cluster* separados em três grupos, que para desmembrar em três arquivos com dados separados, usamos o filtro *RemoveWithValues*, escolhendo o último atributo como índice e removendo dois campos para que reste um e salve-o como arquivo. Este processo é visto na Figura 28 abaixo e foi repetido três vezes para restar apenas os dados de cada grupo de cluster.

Figura 28 - Separando os *clusters*



Fonte: Elaborado pelo autor (2020).

Na seção 4.5.1 consta a análise descritiva dos arquivos gerados neste experimento, ou seja, análises dos grupos separados a fim de identificar suas características.

4.4 Descoberta de Conhecimento dos Experimentos 3 e 4

Nesta seção foi analisado os resultados dos experimentos 3 e 4 que foram extraídos dos algoritmos J48 e *Apriori* e compreender o perfil acadêmico a fim de descobrir conhecimento. Esses experimentos apresentaram dezenas (45) de regras, dentre elas foram selecionadas as que tinham maior índice de assertividade e dando preferência as que ficaram mais próximas de 100% de acurácia (no caso do J48) ou 100% de confiança (no caso do *Apriori*). Quanto mais atributos (15), mais regras podem ser encontradas e mais complexo sua análise. Seguem abaixo as explicações das principais delas:

- **Regra 1**: Se aluno pertencente ao CTG e fez mobilidade nos Estados Unidos então participou do programa ciência sem fronteiras. (227 de ocorrências com 99,13% de acurácia).
- **Regra 2**: Se aluno pertencente ao CTG e fez mobilidade na Inglaterra então participou do programa ciência sem fronteiras. (109 de ocorrências com 99,1% de acurácia).
- **Regra 3**: Nenhum aluno pertencente ao CTG que fez mobilidade pelo CSF na Itália e com carga horária aproveitada maior que 3810 teve desempenho acadêmico insuficiente.
- **Regra 4**: Todos alunos pertencentes ao CTG que fizeram mobilidade na Austrália ou Irlanda participaram do programa ciência sem fronteiras. (74 de ocorrências com 100% de acurácia).
- **Regra 5**: Se aluno pertencente ao CAC, fez mobilidade em Portugal, teve mais de 32 aprovações e desempenho acadêmico bom então participou do programa de mobilidade internacional da UFPE. (28 de ocorrências com 96,6% de acurácia).
- **Regra 6**: Nenhum aluno pertencente ao CAC que fez mobilidade pelo programa ciência sem fronteiras na Espanha teve desempenho acadêmico regular ou insuficiente.
- **Regra 7:** O aluno pertencente ao CAC, que fez mobilidade nos países: Inglaterra, Itália, Austrália ou Irlanda, participou do programa CSF. (122 de ocorrências com acurácia acima de 94%).

- **Regra 8:** Se aluno pertencente ao CCS então participou do programa ciência sem fronteiras (233 de ocorrências com 94,33% de acurácia). Outro ponto de vista, 5,66% dos alunos do CCS (14 alunos) fizeram mobilidade por outro programa.
- **Regra 9:** Nenhum aluno pertencente ao CFCH que fez mobilidade pelo programa de mobilidade internacional da UFPE na Espanha teve desempenho acadêmico regular ou insuficiente.
- **Regra 10:** Nenhum aluno pertencente ao CFCH fez mobilidade pelo programa de mobilidade internacional da UFPE nos países: Inglaterra, Austrália e Irlanda (Bom para investigar carência de mobilidade).
- **Regra 11:** Todos os alunos do CA que participaram do programa ciência sem fronteiras foram do sexo masculino (48 de ocorrências com 100% de acurácia).
- **Regra 12:** 93,3% dos alunos (251) pertencentes ao CIN que fizeram mobilidade internacional foram pelo programa ciência sem fronteiras. Portanto, 6,7% dos alunos (18) pertencentes ao CIN fizeram mobilidade por outro programa.
- **Regra 13:** 95,5% dos alunos do CCB que fizeram mobilidade internacional foram pelo programa CSF. Portanto, 4,5% dos alunos do CCB fizeram mobilidade por outros programas.
- **Regra 14:** 91,7% dos alunos (22) pertencentes ao CCEN que fizeram mobilidade internacional foram pelo programa ciência sem fronteiras. Portanto, 8,3% dos alunos (dois) pertencentes ao CCEN fizeram mobilidade por outro programa.
- **Regra 15:** Nenhum aluno pertencente ao CCSA fez mobilidade pelo programa de mobilidade internacional da UFPE nos continentes: Oceania, Ásia e América Central (Bom para investigar carência de mobilidade).
- **Regra 16:** Se o aluno fez intercâmbio para América do Norte, foi por meio do programa CSF (668 de ocorrência com 95% de confiança).
- **Regra 17:** Todos os alunos que participaram do programa ciência sem fronteiras em países do continente europeu são naturais da região metropolitana de Recife.
- **Regra 18:** 482 alunos que participaram de programas de mobilidades internacionais no ano de 2014 e após cinco anos não integralizaram o curso.
- **Regra 19:** 401 alunos que participaram do programa ciência sem fronteiras no ano de 2014 e após cinco anos não integralizaram o curso.

- **Regra 20:** 381 alunos naturais da região metropolitana de Recife participaram de programas de mobilidades internacionais no ano de 2014 e após cinco anos não integralizaram o curso.
- **Regra 21:** 369 alunos que participaram do programa ciência sem fronteiras no ano de 2013 e após seis anos não integralizaram o curso.

A descoberta do conhecimento para o perfil mobilidade, como visto nas regras encontradas, teve como classe (atributo principal) os programas de mobilidade e evidenciou sua importância em relacionar com outros atributos: centro, país, continente, naturalidade, carga horária aproveitada, desempenho acadêmico e integralização após a mobilidade. O algoritmo ao encontrar as regras de 1 a 14 priorizou os alunos pelo centro que pertencem e quais programas participaram, sendo uma clara polarização entre os dois que tiveram mais alunos: CSF e PMI. Fica evidente o efeito da instituição em usar como critério de participar de programas de mobilidade a média geral igual ou acima de 6 (regular), pois as regras 3, 5 e 8 apontam alunos dos centros CTG, CAC e CFCH que participaram de mobilidade e não tiveram desempenho acadêmico insuficiente, ou seja, abaixo de média 5. Outro ponto importante é que as regras encontraram dados relevantes sobre o impacto do programa ciência sem fronteiras teve na instituição, pois mostram países, continentes e centros que tiveram ou não alunos em mobilidade. Dando destaque ao grande número de alunos dos centros CTG, CCS, CIN, CCB e CCEN que fizeram mobilidade pelo CSF (para alunos de graduação) e a pequena participação dos alunos em outros programas de mobilidade. Como também a falta de alunos do CCSA em mobilidade aos continentes: Oceania, Ásia e América Central.

As regras 15 a 20 foram encontradas pelo algoritmo de associação *Apriori* e nelas também estão presentes dados do CSF e PMI. Apontando grande número de alunos que participaram do CSF em países do continente América do Norte e que Todos os alunos que participaram do programa ciência sem fronteiras em países do continente europeu são naturais da região metropolitana de Recife. As regras 17 a 20 mostram informações relevantes, pois aponta uma quantidade significativa de alunos (1186) que participaram de programas de mobilidades e após cinco, seis ou até sete anos ainda não haviam integralizados seus cursos, ou seja, provavelmente não concluíram. Os 1186 alunos representam 28.01% do universo de 4234 estudados. Pois sabe-se que a instituição usa como padrão, para que os alunos participarem de programas de mobilidades internacionais, o critério de ter cursado um ano acadêmico (dois semestres), a duração da

mobilidade de até dois semestres e que grande parte dos cursos tem duração de cinco anos.

É importante ressaltar que as regras encontradas pelos algoritmos são encontradas com o uso de aprendizado de máquina de forma automatizada de acordo com o tipo de aprendizado e suas utilidades. Algumas vezes as regras podem parecer confusas ou óbvias demais, mas elas servem para que sejam interpretadas por um ator decisório inserido ao trabalho junto aos dados, e assim as regras podem ser melhor aproveitadas e gerarem a descoberta de conhecimento.

4.5 Análises Descritivas

Esta seção mostra uma forma complementar de análise dos dados a fim de contribuir usando estatística descritiva adicionada a mineração e dar apoio a tomada de decisão.

4.5.1 Análises Descritivas do Experimento 5

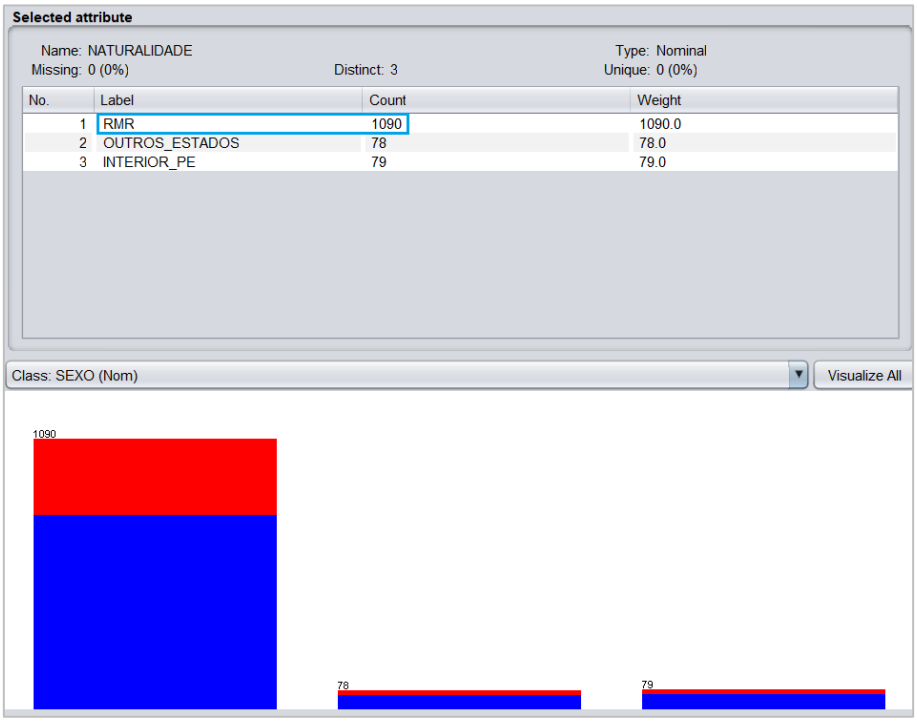
Na seção 4.3.3 foi feita o agrupamento por meio do algoritmo *SimpleKmeans* divididos em três grupos e que cada grupo resultou em um arquivo para análise descritiva. O algoritmo determina automaticamente as características para cada grupo e com esta análise descobrimos quais são. E para esta análise foi feita uma verificação dos dados com ajuda das ferramentas gráficas e de estatísticas do *Weka* com a finalidade de identificar características diferentes entres os grupos.

01) Análise descritiva do grupo 01.

Verificando a fundo os dados do grupo 01 foi possível separar suas principais características nos atributos: naturalidade, integralização após a mobilidade e programas.

- A naturalidade é de 87,41% da região metropolitana de Recife, ou seja, 1090 alunos do grupo, como mostrado na Figura 29 a seguir.

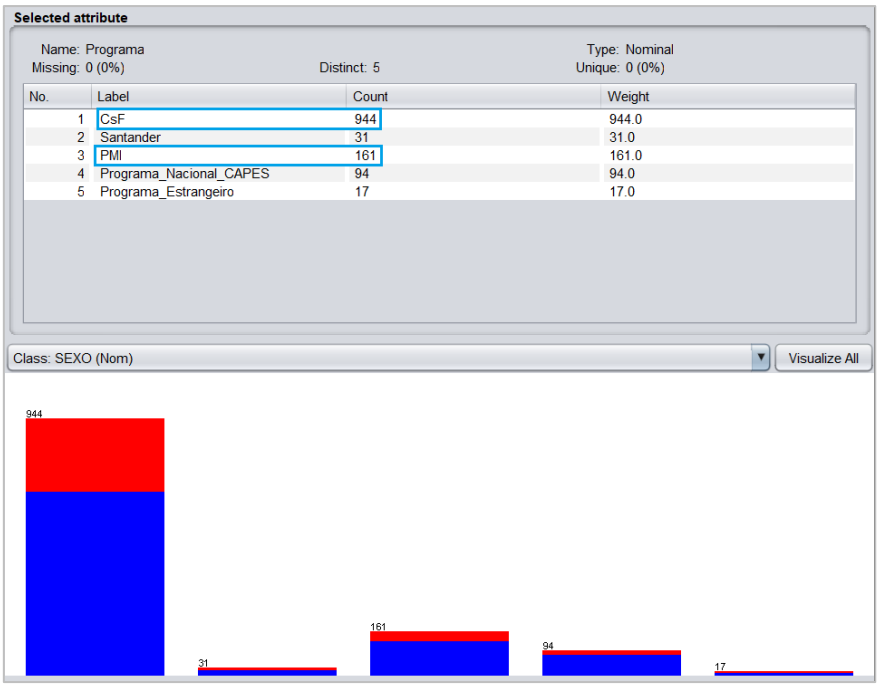
Figura 29 - Naturalidades grupo 01



Fonte: Elaborado pelo autor (2020).

- 51,16% dos alunos do grupo integralizaram o curso, enquanto os grupos dois e três tiveram um percentual bem menor, 21,72% e 21% respectivamente.
- O programa CSF representou 75,7% dos alunos deste grupo e o programa PMI 12,91% vistos na Figura 30 abaixo.

Figura 30 - Programas grupo 01



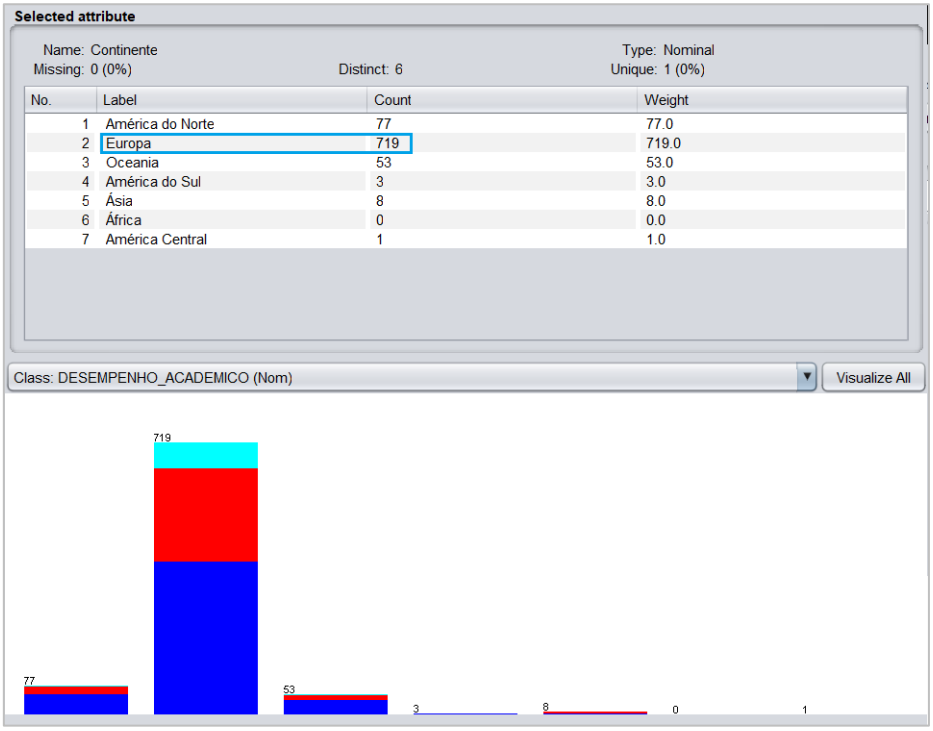
Fonte: Elaborado pelo autor (2020).

02) Análise descritiva do grupo 02.

Analisando melhor os dados do grupo 02 foi possível separar suas principais características nos atributos: continente, naturalidade, sexo, total de reprovações, programa e desempenho.

- O continente europeu representa 83,51% dos alunos neste grupo (Figura 31), enquanto os grupos um e três tiveram resultados 64,88% e 8,33% respectivamente.

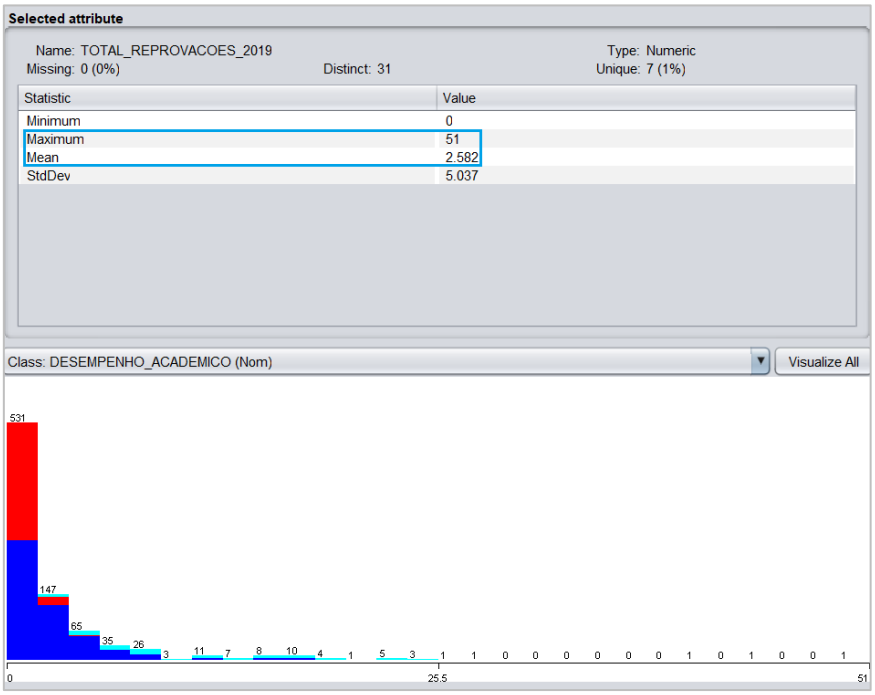
Figura 31 - Continentes grupo 02



Fonte: Elaborado pelo autor (2020).

- A naturalidade dos alunos foi 88,27% para região metropolitana de Recife.
- Predominou o sexo feminino com 83,62% dos alunos neste grupo.
- Apesar de apresentar o valor máximo do total de reprovações maior que os outros grupos, 51 reprovações, obteve a menor média de reprovações em 2,582 reprovações demonstrados na Figura 32 a seguir.

Figura 32 - Total de reprovações grupo 02



Fonte: Elaborado pelo autor (2020).

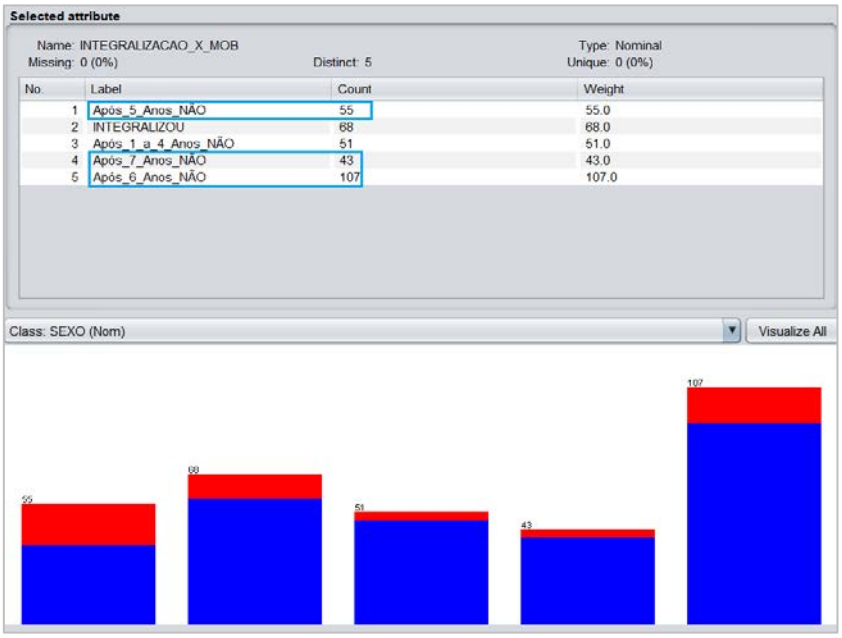
- Neste grupo foi o maior percentual de alunos no programa PMI com 29,62%, bem maior que os grupos 01 e 03, 12,91% e 3,09% respectivamente.
- O desempenho dos alunos foi de 33% neste grupo, sendo o maior percentual para desempenho ótimo. Grupo 01 foi 24,3% e o grupo 03 14,2%.

03) Análise descritiva do grupo 03.

Mergulhando nos dados do grupo 03 foi possível separar suas principais características nos atributos: continente, sexo, naturalidade, integralização após mobilidade, total de faltas, total de reprovações, programas.

- O continente América do Norte representou 79,63% dos alunos deste grupo, sendo bem mais alto que os grupos um e dois que apresentaram 29,35% e 8,94% respectivamente.
- O grupo ficou representado por 83,33% de homens.
- Grupo com maior percentual de alunos, 48,46%, com a naturalidade interior de Pernambuco. Grupo 01 obteve 6,33% e grupo 02 4,53%.
- 63,27% dos alunos deste grupo não integralizaram o curso 5 anos ou mais após a mobilidade internacional. Sendo 52,2% destes não integralizaram após 6 anos. Destacadas na Figura 33 a seguir.

Figura 33 - Integralização grupo 03



Fonte: Elaborado pelo autor (2020).

- Este grupo obteve menor valor máximo, 676, para total de faltas, 853 para grupo 01 e 864 o grupo 02. A maior média de faltas 86,728, sendo 74,804 do grupo 01 e 72,798 do grupo 02. Detalhes demonstrados na Figura 34.

Figura 34 - Análises de faltas dos grupos

Statistic	Grupo 01	Value
Minimum		0
Maximum		676
Mean		86.728
StdDev		77.585

Statistic	Grupo 02	Value
Minimum		0
Maximum		864
Mean		72.798
StdDev		67.037

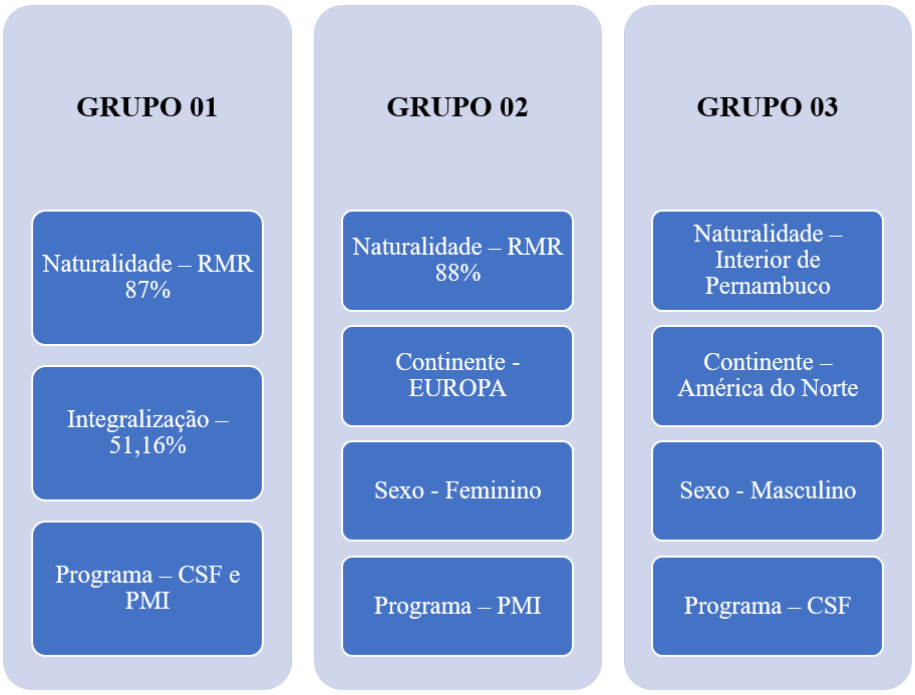
Statistic	Grupo 03	Value
Minimum		1
Maximum		853
Mean		74.804
StdDev		72.806

Fonte: Elaborado pelo autor (2020).

- O grupo obteve 37 de valor máximo do total de reprovações, sendo o menor que os grupos 01 e 02, 48 e 51 respectivamente. Mas apresentou a maior média de faltas com 4,114, contra 3,172 do grupo 01 e 2,582 do grupo 02.
- 94,13% do grupo foi de alunos que participaram do programa CSF.

Segue a Figura 35 com as principais características de cada grupo para melhor visualizar as informações encontradas nos grupos após a análise descritiva e evidenciar o agrupamento feito pelo algoritmo na fase de mineração.

Figura 35 - Principais características dos grupos



Fonte: Elaborado pelo autor (2020).

Apesar de inicialmente ter previsto três grupos, após a análise dos dados do grupo, evidenciou uma intercessão entre os grupos 01 e 02 com os atributos “Naturalidade” e “Programa”. E tais conhecimentos serão úteis para o especialista do domínio do conhecimento e os decisores utilizá-las em futuras tomadas de decisão.

4.5.2 Análise Descritiva

Nesta seção foi feita a análise descritiva (análise exploratória) da base completa de alunos UFPE a fim de comparar com a base dos alunos que fizeram mobilidade. Portanto, ter uma visão geral dos dados e um comparativo entre os dados da base geral com 42.280 alunos e a base de 2432 alunos que fizeram mobilidade internacional.

Para este comparativo foi utilizado a seguinte lista de atributos de desempenho acadêmico: média geral, percentual de integralização do curso, carga horária aproveitada, total de faltas, total de aprovações e total de reprovações. Segue a Tabela 2 contendo os números analisados.

Tabela 2 - Atributos de desempenho acadêmico

	TIPO_INGRES	MEDIA_GERAL	PERCENT_INTEGR	CH_APROVEITADA	TOTAL_FALTAS	TOTAL_APROVACOES	TOTAL_REPROVACOES
Todos Alunos	Mediana	7.68	0.4804	1080	51	17	0
	Media	7.42	0.60	1403.30	84.82	22.90	3.91
	Maximo	10	3180	8505	1449	123	151
	Mínimo	0.04	0.0054	15	0	1	0
	Desvio padrão M	1.046854418	0.375221502	996.3150455	69.86511841	16.26074195	4.883243966
	Desvio padrão	1.349613624	17.56490022	1183.376087	105.2513397	19.74723638	7.512125672
Alunos Mobilidade	Mediana	8	100	3300	55	52	1
	Media	7,99	94,62	3280,64	75,68	50,44	3,09
	Maximo	9	100	8505	864	123	51
	Mínimo	6	0	0	0	0	0
	Desvio padrão	0,3003	18,7197	1312,6479	71,5965	21,2947	5,4973
	Desvio padrão Méd	0.0888	9.3130	806.5485	48.9555	15.6882	3.5697

Fonte: Elaborado pelo autor (2020).

As informações mais importantes colhidas na análise destes dados são:

- A média geral é maior (7,99>7,42), 7,7% maior, para os alunos que fizeram mobilidade e tem valor mínimo de seis;
- A média de faltas dos alunos (75,68<84,82) que fizeram mobilidade é 12,08% menor que dos demais alunos;
- A média do total de aprovações dos alunos que fazem mobilidade é 120,27% maior que de todos os alunos e a mediana do percentual de integralização é de 100% e 48,04% dos demais alunos, ou seja, maior que o dobro;
- A média do total de reprovações dos alunos de mobilidade é de 3,09 e dos demais alunos de 3,91, são próximo, mas o valor máximo de reprovações é de 51 e dos demais alunos de 151, ou seja, praticamente o triplo.

Estas informações finalizam a análise de uma forma geral para uma futura interpretação na fase final de conclusão do estudo.

4.6 Levantamento das ações para internacionalização

Nesta seção são apresentadas as ações realizadas, através de dados coletados no setor DRI, que buscaram atender as metas do Plano de desenvolvimento Institucional (PDI) para o período de 2014 a 2018. Segue abaixo, ponto a ponto, as 23 metas identificadas na subseção “1.3.1 A Internacionalização na UFPE” e ações realizadas no setor. Este levantamento foi realizado pelo pesquisador no papel de analista do processo decisório. Analisando as metas elaboradas pelos facilitadores e decisores, as ações realizadas pelos intervenientes com as atividades propriamente ditas para atingir os agidos, pois são as pessoas que participam diretamente com as consequências das decisões tomadas.

- **Meta:** Aprimorar a estrutura normativa para possibilitar a equivalência dos créditos resultantes de mobilidade nacional, internacional e interna (Inter campi);
  Ação: Formulação da RESOLUÇÃO Nº 09/2019 – CEPE sobre disciplinas internacionalizadas. Parceria entre a DRI e PROACAD para ofertar disciplinas eletivas com código próprio que envolvam a participação de uma IES parceira em cooperação com a UFPE em cursos de graduação.
- **Meta:** Estruturar e consolidar o Núcleo de Formação em Línguas para todos os cursos de graduação. Incrementar em 50% o ensino de língua estrangeira para a comunidade acadêmica;
  Ações: Ampliação de oferta de cursos de língua estrangeira: Turmas de inglês ofertadas inicialmente no Centro Acadêmico do Agreste (UFPE em Caruaru) em 2017, espanhol a partir de 2018 e italiano e francês em 2019.
- **Meta:** Adotar as medidas necessárias para estimular a admissão, na pós-graduação, de alunos provenientes de outros países (alunos estrangeiros);
  Ações: Realizado divulgação da qualidade do ensino, pesquisa, extensão e inovação da instituição pelas redes sociais e portal institucional na internet. Participou das redes e encontros nacionais e internacionais (Tordesilhas, GCUB, AUF, AULP, Nafsa, EIAE e outros) de internacionalização da educação superior participando destes eventos divulgando in loco a instituição.
- **Meta:** Divulgar os programas de pós-graduação junto às embaixadas sediadas no Brasil;
  Ação: Agenda consolidada de reuniões com embaixadas de outros países sediadas no Brasil. Exemplo: Embaixada Italiana em Recife.
- **Meta:** Estruturar e realizar missões internacionais (pelo menos uma missão por ano);
  Ação: Realizadas missões ao exterior: Japão, china, França, Nova Zelândia, Reino Unido, Rússia e grandes eventos internacionais (NAFSA, EAIE, AIEA) de ensino superior no período de 2014 a 2018.
- **Meta:** Incentivar a capacitação de professores no exterior (Pós-doc);
  Ação: Os convênios firmados com instituições estrangeiras atendem docentes e pesquisadores devido a um acordo entre a DRI, PROACAD, PROPESQ e a Procuradoria desde 2016 com a finalidade de incentivar a participação de professores em capacitações no exterior.

- **Meta:** Ampliar a oferta de material de divulgação da UFPE em português, inglês, francês e espanhol;

  Ação: Criação de folders em inglês para distribuir em eventos internacionais. Meta atendida parcialmente, porém o setor está em processo de admitir pessoal de apoio que tenha condições para traduzir material de divulgação para francês e espanhol.
- **Meta:** Ofertar sistematicamente disciplinas em inglês na pós-graduação começando com pelo menos uma disciplina por ano, por programa;

  Ação: Comunicação constante, com verificação todo semestre, entre DRI e coordenações dos cursos (graduação e pós) para colher e divulgar disciplinas ofertadas em inglês desde 2016. A divulgação é feita junto as instituições que irão enviar alunos estrangeiros para cursar mobilidade na UFPE.
- **Meta:** Investir em marketing institucional interno e externo;

  Ação: Criação de material de divulgação feita no setor desde 2014 e o setor está presente nas redes sociais Facebook e Instagram como meio divulgação e interação com a comunidade acadêmica.
- **Meta:** Ampliar em 30% o auxílio a idiomas para os estudantes de graduação fazerem cursos de inglês;

  Ação: Criação de sala com computadores em 2018 para apoio a curso de idiomas e provas de exames de proficiência em inglês TOEFL ITP realizados pelo NUCLI (Núcleo de Línguas – Idiomas sem fronteiras UFPE) no Centro de Artes e Comunicação no campus Recife da UFPE.

Portanto estas informações corroboram em atender o objetivo específico “Realizar levantamento sobre as ações de internacionalização realizadas pelo setor para atender as metas propostas no PDI 2014-2018 da UFPE”, e vai ajudar com a elaboração das recomendações gerenciais e o relatório executivo.

5 Conclusão

A presente pesquisa se motivou em propor um modelo de tomada de decisão utilizando técnicas de mineração de dados para tratar dados relacionados aos alunos de graduação que participam dos programas de mobilidade internacional estudantil e internacionalização da Universidade Federal de Pernambuco. É um estudo de caso do setor diretoria de relações internacionais da instituição e tem abordagem quantitativa e descritiva.

Dessa forma, segundo pesquisas, este estudo aborda uma maneira diferente de contribuir com a academia, as teorias sobre tomada de decisão gerencial e utilizando ferramentas de tecnologia de informação disponíveis no mundo digital.

Por consequência da pesquisa bibliográfica realizada neste estudo, foi encontrada duas teorias utilizadas principais para mineração de dados, o processo *KDD* de Frawley et al (1992) e Fayyad et al (1996) e o modelo *CRISP-DM de* Wirth e Hipp (2000). Propôs uma adaptação utilizando estas teorias unidas a teorias da área de administração e processo decisório. Esta pesquisa resultou no framework proposto adaptando as fases do *CRISP-DM* com uma finalização visando aplicação na administração com elaboração de relatórios de gestão e feedback.

A fim de contemplar o objetivo da pesquisa foram realizadas atividades referentes a atender os objetivos específicos na ordem que eles estão dispostos em sua subseção. Primeiramente foram coletados os dados e estruturados em banco de dados na fase de pré-processamento da mineração. Este objetivo requereu boa parte de todo estudo, pois inclui coleta, conhecer os dados, organizar, limpar e estruturar em atributos e instâncias em um banco de dados sólido e que possibilita a mineração. Em seguida foi feito uma avaliação dos procedimentos, técnicas e algoritmos utilizados para mineração de dados. Esta etapa inclui o estudo teórico, os trabalhos relacionados sobre mineração de dados.

A etapa de encontrar padrões e descobrir novos conhecimentos nos dados de mobilidade internacional estudantil foi contemplada no capítulo de análise da dados. Para encontrar padrões foi aplicada as ferramentas de mineração, seus algoritmos nos dados estudados e chamados de experimentos. Ao analisar os resultados dos experimentos fez-se uma compilação dos padrões e seus resultados obtidos as descobertas de conhecimentos, ou seja, as regras encontradas. As regras interpretadas serviram para descobrir conhecimento que antes não eram visíveis e contribuem para o processo de

tomada de decisão para aplicar futuras ações e atividades estratégicas. Sendo este o ponto alto do estudo, pois na descoberta do conhecimento foram encontrados diversos elementos reveladores e que servirão para beneficiar os estudantes.

Por fim foi realizado levantamento sobre ações feitas que correspondem à metas para o setor estudado e propostas no PDI 2014-2018 da instituição. Este levantamento foi feito com informações colhidas no setor sobre os atores envolvidos diretamente nas atividades, demonstraram o profundo empenho dos envolvidos em atender todas as metas presentes no PDI, ajudando o setor a se desenvolver e preparar a instituição para enfrentar novos desafios.

Esta pesquisa teve a intenção de colaborar para o desenvolvimento do campo do conhecimento voltado ao processo de tomadas de decisão utilizando ferramentas de tecnologia de informação de mineração de dados e contribuir com a academia incluindo novos possibilidades de análise de dados na área de administração. Também contribuir com outras instituições, que registram grandes volumes de dados, em aprimorar suas atividades setoriais de acompanhamento das metas elaboradas em seus planos e melhorar a qualidade dos serviços prestados a toda comunidade.

Portanto o *framework* mostrou-se possível sua aplicação, tanto para análise de dados descentralizados, como também, sua contribuição com o processo de tomada de decisão. Em diversos casos pode-se estar registrando dados de maneira descentralizada, sem um sistema de gerenciamento de banco de dados, por consequente, tais dados não sendo analisados adequadamente e perdendo a oportunidade de descobrir conhecimentos pertinentes para a melhoria setorial e institucional. Apesar de ser um estudo de caso, esta pesquisa contribui também para que profissionais de administração, sistema de informação ou computação tomem como ideia para aplicar em benefício para a sociedade, pois pode-se investigar dados e buscar melhor atender usuários de diversos segmentos onde não ainda não foi aplicado a mineração de dados.

Para possíveis trabalhos futuros seria interessante explorar dados que necessitem outros algoritmos não usados neste trabalho, ou testar o *framework* para segmentos não educacionais e verificar sua validade. Por fim, verificar a aplicação em conjuntos de dados maiores que necessitam de computadores mais potentes e explorar outras ferramentas de mineração de dados.

5.1 Recomendações gerenciais (Diagnóstico de gestão)

Nesta seção são apresentadas posicionamento da pesquisa sobre o estudo de caso do setor estudado, sugestões, separados por focos alvos, considerando a descoberta de conhecimento encontrada após a mineração dos dados e fatores importantes relacionados à pesquisa realizada para execução deste trabalho.

Devido a dificuldades encontradas durante a pesquisa, segue lista de recomendações gerais para o setor:

- Incluir no sistema de gestão acadêmica da instituição um módulo informatizado que possibilite o registro centralizado de informações dos estudantes que participam de mobilidade internacional, a fim de evitar que os dados sejam registrados manualmente ou perdidos;
- Por meio da recomendação acima é possível garantir que os todos os estudantes estejam regularmente cadastrados e evite-se casos de mobilidades que não sejam de conhecimento da instituição;
- Ampliar divulgação para os centros e coordenações de cursos de que o estudante necessita estar regularmente registrado na instituição para participar de mobilidades internacionais, com registro no sistema de gestão acadêmica, sendo estes pré-requisitos para que as disciplinas cursadas em outra instituição possam ser aproveitadas pela instituição de origem e também contribuir para evitar que o estudante faça mobilidade sem devido registro;
- Melhorar a participação dos colaboradores, incluindo-os no processo de elaboração de metas e ações para o setor;
- Elaborar um cronograma específico de acompanhamento de atividades estratégicas e divulga-lo a todos envolvidos;
- Ampliar a rotina de reuniões específicas para metas e ações estratégicas (pelo menos uma por mês).

Um setor bem comandado terá menos problemas para executar ações e atividades estratégicas.

Considerando a descoberta de conhecimento segue abaixo recomendações:

- Informar centros para tomar ciência sobre os pontos positivos e negativos encontrados nas regras dos experimentos sobre questões acadêmicas apresentados no apêndice C;

- Informar aos decisores de futuros PDIs sobre a grande diferença entre alunos do estudo natural do estado de Pernambuco em comparação de alunos natural de outros estados somam apenas 7,5% dos alunos que fizeram mobilidade internacional.
- Informar centros sobre a carência de certos centros na participação de alunos na mobilidade internacional apresentados no apêndice C;
- Informar as Pró-reitorias responsáveis pelos alunos de graduação e pós-graduação e aos seus centros a existência de alunos de uma quantidade significativa de alunos que fizeram mobilidade e após cinco, seis e até sete anos ainda não integralizaram seus cursos;
- Criar programa de bolsas de estudos no exterior com recursos próprios ou com parceiros.

Verificando o levantamento das ações de internacionalização segue as recomendações:

- Ampliar agenda de relacionamento com as embaixadas estrangeiras sediadas em Recife e buscas embaixadas que não estão na cidade para realizar visitas visando divulgação de programas de mobilidade;
- Diversificar missões ao exterior para participar de novos eventos internacionais;
- Ampliar o relacionamento com instituições do continente América do Norte, pois após o fim do CSF para alunos de graduação a mobilidade internacional para tal continente é praticamente zero;
- Ampliar incentivo a capacitação no exterior de professores e incluir os técnicos administrativos;
- Criar programa de incentivo para publicações no exterior;
- Ampliar parcerias com instituições estrangeiras para realizar mais trabalhos colaborativos;
- Criar evento na instituição sobre mobilidade internacional, visando integrar alunos estrangeiros que participam de mobilidade na instituição com os alunos regulares que pretendem realizar mobilidade em outros países;
- Criar ações com finalidade de melhorar a mobilidade para mais grupos de alunos com base nos conhecimento encontrados nos dados do experimento 5 demonstrados na seção 4.5.1 e na Figura 35: alunos com naturalidade de outros

estados, ampliar PMI para o continente américa do norte e incentivar os alunos integralizarem seus cursos;

- Criar programa de preparação para os alunos selecionados para mobilidade de acordo com os países de destino. Exemplos: Curso rápido sobre cultura, curso rápido de defesa pessoal para grupo feminino;
- Ofertar, ou melhor divulgar, serviço de tradução para pesquisadores que pretendem submeter trabalhos em eventos internacionais.

E por fim, recomendações para utilização do framework por outros setores ou instituições:

- Criar um documento para registrar as atividades realizadas em cada fase do estudo, separando em: conhecendo o negócio, conhecendo os dados, pré-processamento, modelagem, avaliação, relatório de gestão e *feedback*;
- Definir os objetivos, verificar se existem problemas e os atores (decisor e analista);
- Coletar os dados e iniciar sua estruturação para a mineração;
- Escolher quais algoritmos serão utilizados e qual ferramenta de mineração;
- Iniciar experimentos aplicando mineração;
- Selecionar experimentos válidos e dentro dos critérios;
- Avaliar e interpretar informações encontradas;
- Gerar relatório com as descobertas de conhecimento;
- Sugerir ações ao decisor;
- Verificar feedback das atividades.

Quanto aos eixos estratégicos e eixos transversais, apresentado na Figura 1, da seção “1.3.2 Cenário Atual da Internacionalização na UFPE”. Este trabalho contemplou três dos cincos eixos estratégicos do Plano de internacionalização da UFPE 2017-2027: Mobilidade universitária, internacionalização da graduação e desenvolvimento de capacidades. Pois estudou diretamente sobre a mobilidade internacional, a internacionalização da instituição e é uma forma de melhorar o desenvolvimento de capacidades dos estudantes, professores e técnicos administrativos que estão envolvidos com a internacionalização universitária. Também comtemplou dois dos três eixos transversais: Tecnologia de informação e marketing institucional e Regulamentação. Pois o estudo fez uso de várias ferramentas da tecnologia da informação para encontrar novos

conhecimentos sobre o assunto e ajudar nas futuras regulamentações sobre o tema na instituição melhorando a qualidade dos serviços.

Portanto, é importante lembrar que o *framework* proposto da adaptação do *CRISP-DM* é composto por fases interligas e que realizam um ciclo, complementando com o relatório de gestão, o *feedback* do que foi proposto, comparando com a realidade e gerando novos objetivos, metas e ações futuras. Após um determinado tempo, que pode coincidir com o planejamento estratégico, pode-se fazer uma nova bateria de experimentos com os dados do período depois do que já foi minerado e verificar os efeitos das tomadas de decisões.

Referência

AMARAL, Fernando. **Aprenda mineração de dados**: teoria e prática. 1ª ed. Rio de Janeiro: Atlas Books, 2016.

ANDRADE, AMBONI, N. **Estratégias de gestão** – processos e funções do administrador. Rio de Janeiro: Elsevier/Campus, 2010.

BAHGA, A.; MADISETTI, V. **Big Data Science & Analytics**: A Hands-On Approach. VPT, 2016.

BAKER, S. J. D.; CARVALHO, M. J. B. D.; ISOTANI,. **Mineração de Dados Educacionais**: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 2011.

BLANCHARD, Benjamin S.; BLYLER, John E. **Systems engineering management**. New York: John Wiley & Sons, 1988.

BRAMER, Max. **Principles of data mining**. 2nd ed. London: Springer, 2013.

CAPES, **A internacionalização na Universidade Brasileira**: resultados do questionário aplicado pela Capes, 2017. Disponível em: <https://www.capes.gov.br/images/stories/download/diversos/A-internacionalizacao-nas-IES-brasileiras.pdf>. Acesso em: 22 de jul. de 2019.

UFPE, **Plano de Internacionalização UFPE 2017-2027**, 2018. Disponível em: < encurtador.com.br/pJKU5>. Acesso em: 20 de jul. de 2020.

CHOO, C. W. **Information Management for the Intelligent Organization**. The Art of Scanning the Environment. 3. ed. Medford: Information Today, 2002.

CAVALCANTI, A. G. G. et al. **Mineração e Visualização de Dados Educacionais**: Identificação de Fatores que Afetam a Motivação de Alunos na Educação a Distância, Workshop de Educação e Informática Bahia, Alagoas e Sergipe e XIV Escola Regional de Computação Bahia, Alagoas e Sergipe, 2014.

COBBE, P. R. C. O. et al. **A inteligência organizacional como instrumento de autoavaliação em instituições de ensino superior, Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 5, n. 2, p. 111-126, 2015.

DE WIT, H. **Repensando o conceito de internacionalização**, In: Revista Ensino Superior Unicamp, 2013, p. 69-71. Disponível em: www.revistaensinosuperior.gr.unicamp.br

ELMASRI, R. E.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 6a Ed. São Paulo: Pearson, 2011.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH. Padhraic. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37, 1996.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. **Knowledge discovery in databases**: An overview. AI magazine, v. 13, n. 3, p. 57, 1992.

FREITAS, D. de. **Strategies in Search for InternationalPartnerships**. Revista do Colégio Brasileiro de Cirurgiões, Rio de Janeiro, v. 42, supl. 1, p. 81-82, 2015.

GOMES, D. C. et al. **Mineração de Dados no Serviço de Atendimento de Urgências**. Revista Journal of Health Informatics, Outubro-Dezembro 2014.

GOMES, L. F. A. M.; GOMES, C. F. S. **Tomada de Decisão Gerencial**: enfoque multicritério. 4 ed. São Paulo: Atlas, 2012.

GOMES H. M.; Carvalho D. R. ***A hybrid Data Mining Method****: Exploring Sequential Indicators Over Association Rules.* Iberoamerican Journal of Applied Computing 2011; 1(1):40-60.

GOMES H. M.; CARVALHO D. R. **A hybrid Data Mining Method**: Exploring Sequential Indicators Over Association Rules. Iberoamerican Journal of Applied Computing 2011.

GOMES H. M.; HAUGT L. G.; CARVALHO D. R. **Mineração de Dados Temporal**: Descobertas de Regras De Causa e Efeito. In: Anais do V Congresso Sul Brasileiro de Computação; 2010.

HALL, M. et al. **The weka data mining software**: an update. ACM SIGKDD explorations newsletter 11 (1): 10–18, 2009.

HAN, Jiawei; PEI, Jian; KAMBER, M. **Data mining**: concepts and techniques. 3 Ed. Elsevier, 2011.

HARRISON, Guy. **Next Generation Databases**: NoSQL, NewSQL and Big Data. Apress, 2015.

INEP. **Censo de Educação Superior**, 2019. Disponível em: <http://inep.gov.br/censo-da-educacao-superior>. Acesso em: 10 de out. de 2019.

INMON, W. H. **Building the Data Warehouse**: Gettiing Started. 4a Ed. Editora: Wiley Publishing, 2005.

JARZABKOWSKI, P.; KAPLAN, S. **Strategy tools-in-use**: a framework for understanding "technologies of rationality" in practice. Strategic Management Journal, v. 36, p. 537-558, 2015.

KAMPFF, Adriana J. Cerveira. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente**. Tese de Doutorado, Programa de Pós-Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul – UFRS, Porto Alegre, 2009.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit**: the definitive guide to dimensional modeling. 3rd ed. John Wiley & Sons, 2013.

LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 5ª ed. São Paulo: Atlas, 2003.

LANTZ, Brett. **Machine Learning with R**: Expert techniques for predictive modeling to solve all your data analysis problems. Packt Publishing, 2015.

LAUDON, K.; LAUDON, J. **Sistemas de Informação Gerenciais**. São Paulo: Pearson Prentice Hall, 2010.

LINOFF, Gordon S.; BERRY, Michael J. A. **Data Mining Techniques**: For Marketing, Sales, and Customer Relationship Management. 3a Edition. Wiley, 2011.

MASCHIO, Pedro et al. **Um Panorama acerca da Mineração de Dados Educacionais no Brasil**. SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO – SBIE/ Congresso Brasileiro de Informática na Educação – CBIE, 2018.

HARRISON, Matt. **Machine Learning**: Guia de referência rápida. 1ª ed. São Paulo: Novatec, 2020.

MAXIMIANO, Antonio Cesar Amaru. **Teoria geral da administração** – 2ª ed. São Paulo: Atlas, 2012.

MINTZENBERG, H.; AHLSTRAND, B.; & LAMPEL, J. **Safári de Estratégia**: um roteiro pela selva do planejamento estratégico (2nd ed., p. 392). Porto Alegre: Bookman, 2010.

MOORE, A. Statistical Data Mining Tutorials. Disponível em: <https://www.cs.cmu.edu/~./awm/tutorials/list.html />, 2005. Acesso em: 15 de out. de 2019.

MORESI, E. A. D. **Inteligência Organizacional**: um referencial integrado. Ciência da Informação, Brasília, v. 30, n. 2, p. 35-46, maio/ago. 2001. Disponível em: www.scielo.br/pdf/ci/v30n2/6210.pdf. Acesso em: 04 jun. 2010.

MOSS, L. T.; ATRE, S. **Business Intelligence Roadmap**: The Complete Project Lifecycle for Decision-Support Applications. Boston: Addison Wesley, 2003.

MOTTA, F. C. P.; Vasconcelos, I. F. G. **Teoria Geral da Administração**. Thomson Pioneira Ed. Cengage. 3ª Edição, 2006.

MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to machine learning with Python**: a guide for data scientists. O'Reilly Media, 2016.

O'BRIEN, James A. **Sistemas de informação e as decisões gerenciais na era da internet**. 2º. ed. São Paulo: Saraiva, 2005. p. 143.

OLIVEIRA, A. L. de; FREITAS, M. E. de. **Motivações para Mobilidade Acadêmica Internacional**: A Visão de Alunos e Professores Universitários. Educação em revista, Belo Horizonte, v. 32, n. 3, p. 217-246, 2016.

PATI, Camila. **36 universidades do Brasil entraram no maior ranking educacional do mundo**. Exame, Você S/A, São Paulo, 26 de set. de 2018. Disponível em: <https://exame.abril.com.br/carreira/36-universidades-do-brasil-entraram-no-maior-ranking-educacional-do-mundo>. Acesso em: 20 de ago. De 2019.

ROZENFELD, H. et al. **Gestão de desenvolvimento de produtos**: uma referência para a melhoria do processo. São Paulo: Saraiva, 2006.

SAMPIERI, Roberto H.; CALLADO, Carlos F.; LUCIO, Maria del Pilar B. **Metodologia de pesquisa**. 5º. Ed. Porto Alegre: Penso, 2013.

SEEL, Norbert M. (Ed.). **Encyclopedia of the Sciences of Learning**. Springer, 2012.

SERPRO. **Serpro compartilha boas práticas de mineração de dados na administração pública**, 2019. Disponível em: < https://www.serpro.gov.br/menu/noticias/noticias-2019/serpro-boas-praticas-mineracao-dados-administracao-publica>. Acesso em: 05 de mar. De 2020.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business intelligence, analytics, and data science**: a managerial perspective. 4rd ed. Pearson, 2017.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN S. **Sistema de Banco de Dados**. 6a Edição. Rio de Janeiro: Editora Campus, 2012.

TAN, Pang-Ning et al. **Introduction to Data Mining**. 2nd ed. Pearson, 2018.

UFPE. **Projeto de internacionalização da UFPE é aprovado na Capes**, 2018. Disponível em: < https://www.ufpe.br/agencia/noticias/-/asset_publisher/VQX2pzmP0mP4/content/projeto-de-internacionalizacao-da-ufpe-e-aprovado-na-capes/40615>. Acesso em: 22 de jul. de 2019.

UFPE, **Plano de Internacionalização 2017-2027**, 2018. Disponível em: <https://www.ufpe.br/documents/40788/506683/PLI+UFPE+versão+port+Final+0405.pdf/e4fe9157-930b-4098-89c0-78cf23a48546>. Acesso em: 19 de ago. de 2019.

WANG, J. **Encyclopedia of Data Warehousing and Data Mining**. Montclair State University, USA: Idea Group Inc, 2006.

WATSON, Hugh J.; WIXOM, Barbara H. **The Current State of Business Intelligence**. IEEE Computer Society Online, pag. 96-99, setembro de 2007.

WHITTINGTON, R. **Strategy as practice. Long Range Planning**, v. 29, n. 5, 731-735, 1996.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for Data Science**: import, tidy, transform, visualize, and model data. O'Reilly Media, 2017.

WIRTH, R.; HIPP; J. **CRISP-DM**: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Citeseer, pp. 29–39, 2000.

WITTEN, Ian H; FRANK, Eibe; HALL, Mark A. **Data mining**: practical machine learning tools and techniques. 4rd ed. Morgan Kaufmann - Elsevier, 2016.

WREMBEL, Robert; KONCILIA, Christian. **Data Warehouses and OLAP**: Concepts, Architectures and Solutions. Edição: IRM Press. Hershey PA: Idea Group Inc., 2007.

YIN, Robert K. **Estudo de caso: planejamento e métodos**. Tradução: Daniel Grassi. Porto Alegre: Bookman, 2004.

APÊNDICE

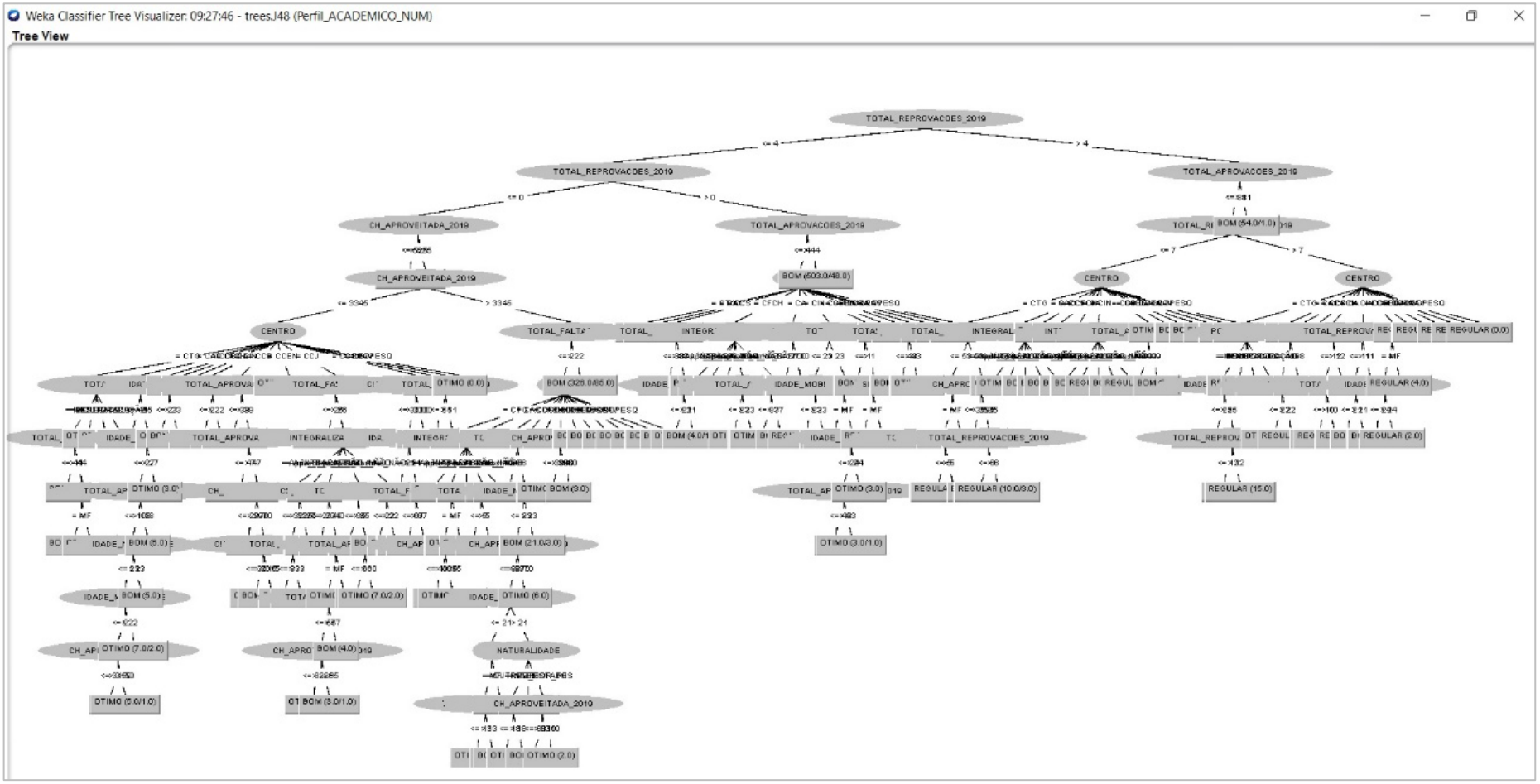
A - Árvore de decisão completa do experimento 1

B - Figura da árvore de decisão completa do experimento 4

C - Pontos positivos e negativos das regras encontrados na descoberta de conhecimento separadas por centro

APÊNDICE A - Árvore de decisão completa do experimento 1

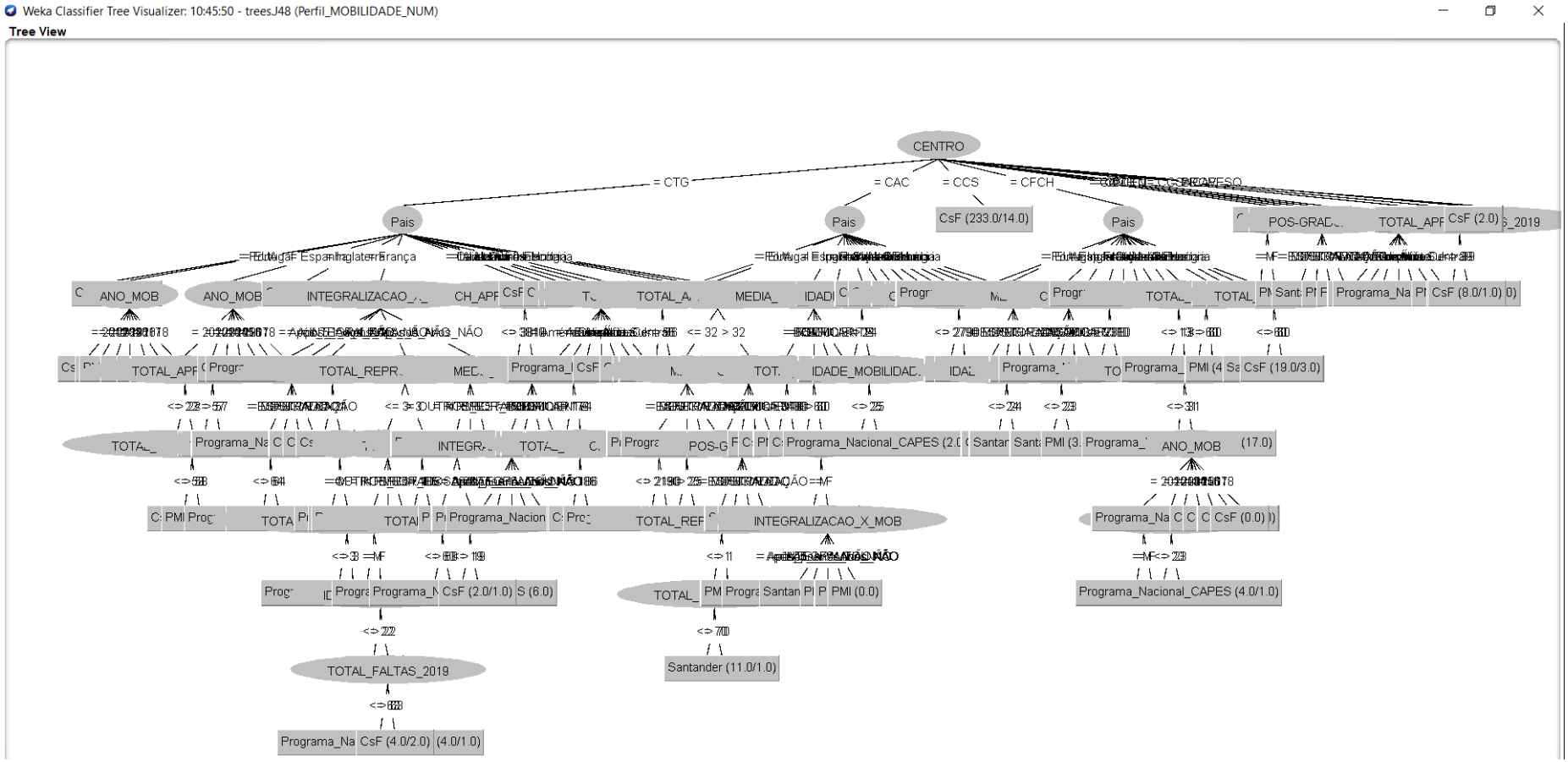
Figura - Árvore de decisão completa do experimento 1



Fonte: Elaborado pelo autor (2020).

APÊNDICE B - Árvore de decisão completa do experimento 4

Figura - Árvore de decisão completa do experimento 4



Fonte: Elaborado pelo autor (2020).

APÊNDICE C - Pontos positivos e negativos das regras encontrados na descoberta de conhecimento separadas por centro

Centros	Experimentos	Regras	Pontos mais positivos	Pontos menos negativos
CTG	1 e2 3 e 4	1, 2, 9, 11 e 13 1, 2, 3 e 4	Aluno com caga horária <=3345 e com desempenho acadêmico ótimo ingressaram no mestrado e/ou doutorado.	Alunos que teve mais de 12 reprovações tiveram desempenho acadêmico regular.
CAC	1 e2 3 e 4	3, 12 e 14 5, 6 e 7	90% são femininas e natural de RMR.	Alunos com desempenho acadêmico bom, com reprovações entre 5 e 7 mas ainda não integralizaram o curso 7 após a mobilidade.
CCB	1 e2 3 e 4	4 13	Alunos que não tem reprovações tem desempenho acadêmico bom.	95,5 % dos alunos que fizeram mobilidade neste centro foram pelo CSF. Apenas 4,5% fizeram mobilidade por outros programas.
CCJ	1 e 2	5 e 6	Alunos que não tiveram reprovações tiveram desempenho acadêmico ótimo.	

Centros	Experimentos	Regras	Pontos mais positivos	Pontos menos negativos
CA	1 e 2 3 e 4	8 e 10 11	Alunos que tiveram entre 5 e 7 reprovações obtiveram desempenho acadêmico bom.	Alunos que tiveram mais de 12 reprovações obtiveram desempenho acadêmico regular.
CCS	3 e 4	8	94,33% dos alunos que fizeram mobilidade foram pelo programa CSF.	Apenas 5,66% dos alunos fizeram mobilidade por outros programas de mobilidade.
CFCH	3 e 4	9 e 10	Nenhum aluno que fez mobilidade na Espanha teve desempenho acadêmico regular ou insuficiente.	Nenhum aluno deste centro fez mobilidade nos países: Inglaterra, Austrália e Irlanda.
CIN	3 e 4	12,	93,3% dos alunos que fizeram mobilidade foram pelo programa CSF.	Apenas 6,7% dos alunos fizeram mobilidade por outros programas de mobilidade.
CCEN	3 e 4	14	91,7% dos alunos que fizeram mobilidade foram pelo programa CSF.	Apenas 8,3% dos alunos fizeram mobilidade por outros programas de mobilidade.

Fonte: Elaborado pelo autor (2020).