UNIVERSIDADE CANDIDO MENDES – UCAM
PROGRAMA DE PÓS-GRADUAÇÃO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL
CURSO DE MESTRADO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL

Nícollas Nogueira Cretton

MINERAÇÃO DE DADOS APLICADA NA BASE DO ENADE COM ENFOQUE NA CRIAÇÃO DE PERFIS DOS ESTUDANTES QUE PRESTARAM O EXAME UTILIZANDO O ALGORITMO J48

CAMPOS DOS GOYTACAZES,RJ
Dezembro de 2016

UNIVERSIDADE CANDIDO MENDES – UCAM
PROGRAMA DE PÓS-GRADUAÇÃO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL
CURSO DE MESTRADO EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL

Nícollas Nogueira Cretton

MINERAÇÃO DE DADOS APLICADA NA BASE DO ENADE COM ENFOQUE NA CRIAÇÃO DE PERFIS DOS ESTUDANTES QUE PRESTARAM O EXAME UTILIZANDO O ALGORITMO J48

Dissertação apresentada ao Programa de Pós-Graduação em Pesquisa Operacional e Inteligência Computacional, da Universidade Candido Mendes – Campos/RJ, para obtenção do grau de MESTRE EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL.

Orientadora: Prof.ª Geórgia Regina Rodrigues Gomes, DSc.

CAMPOS DOS GOYTACAZES/RJ
Dezembro de 2016

FICHA CATALOGRÁFICA

C924m Cretton, Nícollas Nogueira.

Mineração de dados aplicada na base do enade com enfoque na criação de perfis dos estudantes que prestaram o exame utilizando o algoritmo J48. /. Nícollas Nogueira Cretton – 2017.

127 f. il.

Orientadora: Geórgia Regina Rodrigues Gomes

Dissertação apresentado ao Curso de Mestrado em Pesquisa Operacional e Inteligência Computacional da Universidade Candido Mendes - Campos dos Goytacazes, RJ, 2016.

Bibliografia: f.114-127

1. Educação. 2. Mineração de Dados. 3. Algoritmo J48. 4. KDD (Knowledge Discovery in Databases). 5. Exame Nacional de Avaliação do Estudante (ENADE). I. Universidade Candido Mendes – Campos. II. Título.

CDU – 004.421:37

NÍCOLLAS NOGUEIRA CRETTON

MINERAÇÃO DE DADOS APLICADA NA BASE DO ENADE COM ENFOQUE NA CRIAÇÃO DE PERFIS DOS ESTUDANTES QUE PRESTARAM O EXAME UTILIZANDO O ALGORITMO J48

Dissertação apresentada ao Programa de Pós-Graduação em Pesquisa Operacional e Inteligência Computacional, da Universidade Candido Mendes – Campos/RJ, para obtenção do grau de MESTRE EM PESQUISA OPERACIONAL E INTELIGÊNCIA COMPUTACIONAL.

Aprovada em 16 de dezembro de 2016

BANCA EXAMINADORA

Prof.ª Geórgia Regina Rodrigues Gomes, DSc. - Orientadora
Universidade Candido Mendes

Prof. Helder Gomes Costa, DSc.
Universidade Federal Fluminense

Prof. Ítalo de Oliveira Matias, DSc.
Universidade Candido Mendes

CAMPOS DOS GOYTACAZES, RJ
Dezembro de 2016

Dedico este trabalho à Deus, minha família e amigos.

AGRADECIMENTOS

Agradeço primeiramente à Deus, por ter me dado a força e a capacidade para alcançar meus objetivos e ter sempre iluminado meu caminho.

Agradeço a minha família e amigos por terem me apoiado e me dado a estrutura e confiança para seguir adiante e completar esta fase da minha vida.

Agradeço também aos professores e amigos do Mestrado em Pesquisa Operacional e Inteligência Computacional da Candido Mendes, por terem me provido com o conhecimento necessário para o meu crescimento profissional e educacional.

Em especial ao professor Dr. Ítalo de Oliveira Matias, por sugestões dadas durante as primeiras fases da dissertação e para minha orientadora Dra. Geórgia Regina Rodrigues Gomes, que confiou e acreditou em mim em cada passo do desenvolvimento.

Não há fim para a educação. Não é que você leia um livro, passe um exame e termine com a educação. Toda a vida, desde o momento em que nascemos até o momento em que morremos, é um processo de aprendizagem.

Jiddu Krishnamurti

RESUMO

MINERAÇÃO DE DADOS APLICADA NA BASE DO ENADE COM ENFOQUE NA CRIAÇÃO DE PERFIS DOS ESTUDANTES QUE PRESTARAM O EXAME UTILIZANDO O ALGORITMO J48

A extração do conhecimento, também conhecida como processo KDD (Knowledge Discovery in Databases), engloba um conjunto de técnicas capaz de analisar e extrair padrões e informações potencialmente úteis de gigantescas bases de dados, encontrando ligações entre os dados desta e, gerando assim, regras e padrões. O Exame Nacional de Avaliação do Estudante (ENADE) tem como objetivo avaliar o grau de conhecimento dos estudantes referente aos conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos a partir do desempenho destes no exame. Esta dissertação tem como objetivo extrair conhecimento da base de dados do ENADE do ano de 2013, relacionando o perfil dos estudantes que prestaram a prova com sua nota e respostas no questionário de percepção da prova. A base utilizada foi obtida no portal do INEP, em sua parte de download de micro dados. Para que esta meta fosse alcançada, a base selecionada foi dividida em quatro partes e foram utilizadas as etapas do processo de KDD, técnicas de Mineração de Dados e o software WEKA para a descoberta das informações. A tarefa utilizada para a mineração foi a de Classificação, através da técnica de árvore de decisão utilizando o algoritmo J48. Após a mineração de dados, foi realizada uma análise dos resultados onde foi possível observar que as maiores dificuldades enfrentadas pelos alunos no exame foi, a abordagem do conteúdo, onda na sala de aula, foi cobrada de forma diferente e a falta de motivação para a realização do exame mostraram-se muito relevantes dentre as demais. Além disso, muitos alunos afirmaram não ter tido nenhuma dificuldade ao resolver a prova e ainda sim obtiveram resultados negativos, fator preocupante quando se envolve cursos na área da saúde. Espera-se que as informações geradas a partir deste trabalho possam ser utilizadas para o aprimoramento dos cursos examinados, bem como na tomada de decisões referentes aos projetos dos cursos.

PALAVRAS-CHAVE: Educação. Mineração de Dados. Algoritmo J48. KDD (Knowledge Discovery in Databases). Exame Nacional de Avaliação do Estudante (ENADE).

ABSTRACT

DATA MINING APPLIED IN THE ENADE DATABASE WITH A FOCUS ON THE CREATION OF PROFILES OF STUDENTS THAT TOOK THE EXAM USING J48 ALGORITHM

Knowledge extraction, also known as the KDD (Knowledge Discovery in Databases) process, encompasses a set of techniques capable of analyzing and extracting potentially useful patterns and information from gigantic databases, finding links between the data from this and thus generating rules and standards. The purpose of the National Student Assessment Examination (ENADE) is to evaluate students' knowledge of the contents of the curriculum guidelines of their respective courses, based on their performance in the exam. This dissertation aims to extract knowledge from the database Of ENADE in 2013, relating the profile of the students who took the test with their grade and answers in the questionnaire of perception of the test. The base used was obtained from the INEP portal, in its part of downloading micro data. In order for this goal to be achieved, the selected base was divided into four parts and the KDD process steps, Data Mining techniques and WEKA software were used to discover the information. The task used for the mining was Classification, through the decision tree technique using the algorithm J48. After the data mining, an analysis of the results was performed where it was possible to observe that the greatest difficulties faced by the students in the exam was, the content approach, the wave in the classroom, was charged differently and the lack of motivation for the Examination were very relevant among the others. In addition, many students said they had no difficulty in resolving the test and still obtained negative results, a worrying factor when engaging in courses in health. It is hoped that the information generated from this work can be used to improve the courses examined, as well as to make decisions regarding the course projects.

KEYWORDS: Education. Data mining. J48 algorithm. KDD (Knowledge Discovery in Databases). National Student Assessment Examination (ENADE)

LISTA DE FIGURAS

Figura 1.	Etapas do KDD (Knowledge Discovery in Databases).	49
Figura 2.	Tela inicial do WEKA 3.7.	58
Figura 3.	Tela principal do ambiente gráfico do WEKA.	58
Figura 4.	Ambiente gráfico do WEKA com a 1º base carregada.	72
Figura 5.	Aba de Classificação do WEKA.	73
Figura 6.	Arvore de decisão referente às IES Federais da 1º Base.	78
Figura 7.	Arvore de decisão referente às IES Estaduais da 1º Base.	79
Figura 8.	Arvore de decisão referente às IES Municipais da 1º Base.	80
Figura 9.	Arvore de decisão referente às IES Privadas sem fins lucrativos da 1º Base.	81
Figura 10.	Arvore de decisão referente às IES Privadas com fins lucrativos da 1º Base.	83
Figura 11.	Arvore de decisão referente ao curso de Agronomia nas IES Federais da 2º Base.	84
Figura 12.	Arvore de decisão referente ao curso de Agronomia nas IES Estaduais da 2º Base.	86
Figura 13.	Arvore de decisão referente ao curso de Fisioterapia nas IES Federais da 2º Base.	87
Figura 14.	Arvore de decisão referente ao curso de Fisioterapia nas IES Estaduais da 2º Base.	89
Figura 15.	Arvore de decisão referente ao curso de Fonoaudiologia nas IES Federais da 2º Base.	90
Figura 16.	Arvore de decisão referente ao curso de Fonoaudiologia nas IES Estaduais da 2º Base.	91

Figura 17. Arvore de decisão referente ao curso de Agronomia nas IES Federais da 3º Base. 93

Figura 18. Arvore de decisão referente ao curso de Agronomia nas IES Estaduais da 3º Base. 94

Figura 19. Arvore de decisão referente ao curso de Enfermagem nas IES Federais da 3º Base. 96

Figura 20. Arvore de decisão referente ao curso de Enfermagem nas IES Estaduais da 3º Base. 97

Figura 21 Arvore de decisão referente ao curso de Fisioterapia nas IES Federais da 3º Base. 98

Figura 22. Arvore de decisão referente ao curso de Medicina nas IES Federais da 3º Base. 100

Figura 23. Arvore de decisão referente ao curso de Tecnologia em gestão hospitalar nas IES Federais da 3º Base. 101

Figura 24. Arvore de decisão referente ao curso de Agronomia nas IES Federais da 4º Base. 103

Figura 25. Arvores de decisão referente a vários cursos encontrados nas IES Federais da 4º Base. 104

LISTA DE GRÁFICOS

Gráfico 1. Número de publicações por ano 28

Gráfico 2. Ranking dos artigos mais citados 29

Gráfico 3. Número de artigos por autor. 29

Gráfico 4. Revistas que mais publicam sobre o tema. 30

Gráfico 5. Países que mais publicam sobre o tema. 31

LISTA DE QUADROS E TABELAS

Quadro 1.	Parte do Dicionário de Variáveis do ENADE 2013	62
Quadro 2.	Relação dos atributos pertinente a cada base	63
Quadro 3.	Relação dos atributos com suas respectivas descrições	64
Quadro 4.	Realçao dos Valores em Código dos Atributos que foram Transformados em suas Respectivas Descrições	67
Quadro 5.	Valores do atributo nu_idade transformados e relacionados com suas respectivas descrições.	68
Quadro 6.	Valores dos atributos nt_fg, nt_ce e nt_ger transformados e relacionados com suas respectivas descrições.	68
Quadro 7.	Relação dos valores em código dos atributos não transformados com suas respectivas descrições	69
Quadro 8.	Relação das bases de dados com seus respectivos atributos classificadores.	74
Quadro 9.	Relação das bases de dados com suas respectivas confianças.	77
Tabela 1	Parte da Base de Dados do ENADE 2013.	61
Tabela 2	Parte da Base de Dados com Atributos Selecionados	65
Tabela 3	Parte da Base de Dados Pré-Processada	70

LISTA DE EQUAÇÕES

Equação 1. Fórmula utilizada para calcular o Conceito Preliminar de Curso 44

Equação 2. Fórmula para Utilizar a Propoção fdos Graduados 45

Equação 3. Fórmula utilizada para calcular a proporção dos mestrandos. 45

Equação 4. Fórmula utilizada para calcular o Índice Geral de Cursos. 45

LISTA DE SIGLAS E ABREVIATURAS

CAPES.	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNA	Comissão Nacional de Avaliação
CNRES	Comissão Nacional para a Reformulação do Ensino Superior
ENADE	Exame Nacional de Avaliação do Estudante
ENC	Exame Nacional de Cursos
GERES	Grupo Executivo para a Reformulação do Ensino Superior
IES	Instituição de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	Knowledge Discovery in Databases
MEC	Ministério da Educação
PAIUB	Programa de Avaliação das Universidades Brasileiras
PARU	Programa de Avaliação da Reforma Universitária
SESU	Secretaria de Educação Superior
SINAES	Sistema Nacional de Avaliação do Ensino Superior

SUMÁRIO

1.	**INTRODUÇÃO**	18
1.1.	PROBLEMA	18
1.2.	JUSTIFICATIVA	22
1.3.	OBJETIVOS	23
1.3.1.	**Objetivo Geral**	23
1.3.2.	**Objetivos Específicos**	24
1.4.	ESTRUTURA DO TRABALHO	25
2.	**ESTUDO BIBLIOMÉTRICO**	26
2.1.	DESCRIÇÃO DO ESTUDO BIBLIOMÉTRICO	26
2.2.	O USO DAS PALAVRAS CHAVES E METODOLOGIAS DA BIBLIOMETRIA	27
2.3.	RESULTADOS ENCONTRADOS	28
3.	**HISTÓRIA DA AVALIAÇÃO NO ENSINO SUPERIOR NO BRASIL**	33
3.1.	CONTEXTO HISTÓRICO DA AVALIAÇÃO DA EDUCAÇÃO SUPERIOR	33
3.2.	PROGRAMA DE AVALIAÇÃO DA REFORMA UNIVERSITÁRIA (PARU).	34
3.3.	COMISSÃO NACIONAL PARA A REFORMULAÇÃO DO ENSINO SUPERIOR (CNRES) E GRUPO EXECUTIVO PARA A REFORMULAÇÃO DO ENSINO SUPERIOR (GERES).	35
3.4.	PROGRAMA DE AVALIAÇÃO DAS UNIVERSIDADES BRASILEIRAS (PAIUB).	38
3.5.	EXAME NACIONAL DE CURSOS (ENC).	39
3.6.	SISTEMA NACIONAL DE AVALIAÇÃO DA EDUCAÇÃO SUPERIOR (SINAES).	41

4. KNOWLEDGE DISCOVERY IN DATABASES (KDD). 47
4.1. ETAPA DE DEFINIÇÃO E COMPREENSÃO DO DOMÍNIO 50
4.2. ETAPA DE SELEÇÃO DOS DADOS 50
4.3. ETAPA DE LIMPEZA E TRANSFORMAÇÃO DOS DADOS 51
4.4. ETAPA DE MINERAÇÃO DE DADOS 52
4.4.1. Tarefa de Classificação 53
4.4.2. Tarefa de Regressão 54
4.4.3. Tarefa de Clusterização 55
4.4.4. Tarefa de Regra de Associação 56
4.5. ETAPA DE INTERPRETAÇÃO DO CONHECIMENTO 56
4.6. SOFTWARE WEKA© 57
5. METODOLOGIA 60
5.1. SELEÇÃO DOS DADOS 60
5.2. LIMPEZA E TRANSFORMAÇÃO DOS DADOS 65
5.3. MINERAÇÃO DE DADOS 71
5.3.1. Tarefa de Classificação e Algoritmo J48 72
6. RESULTADOS E DISCUSSÕES 76
7. CONCLUSÃO 110
7.1. TRABALHOS FUTUROS 112
8. REFERÊNCIAS BIBLIOGRÁFICAS 127

1. INTRODUÇÃO

1.1. PROBLEMA

Desde o primórdio da civilização, o conhecimento humano tem sido estudado e auxiliando na evolução desta. Com o passar do tempo, organizações começaram a analisar o seu conhecimento organizacional, baseando-se na sua experiência e na tentativa e erro. As organizações que possuíam uma maior capacidade de geração de conhecimento, de difundi-lo dentro de si mesma e de inserir este nos serviços, produtos e sistemas, se destacavam das demais (NONAKA e TAKEUCHI, 2003).

Portanto, ter acesso a conhecimentos úteis é algo de extrema importância numa organização, uma vez que estes podem auxiliar nas tomadas de decisões e num melhor desempenho de setores desta organização.

Com o avanço tecnológico, tornou-se possível uma coleta dados cada vez mais rápida e um imensurável espaço de armazenamento em meios digitais. Isto proporcionou um grande acúmulo de dados por parte das empresas e instituições, que inviabilizou a visualização de padrões para geração de informações. (DUNHAM, 2002).

Para Camilo e Silva (2009), não é recomendada a utilização das técnicas tradicionais para a extração do conhecimento em grandes bases de dados, uma vez que as chances de insucesso são grandes, devido a grande necessidade de tempo,

recursos e mão-de-obra empenhada para a resolução da análise.

Portanto, somente a coleta e o armazenamento dos dados já não são mais o suficiente. Seria necessária a criação de novos métodos artificiais que fossem capazes de trabalhar e extrair informações úteis destas grandes bases de dados (PANG-NING, STEINBACH e KUMAR, 2009).

Fayyad *et al*. (1996), apresentou um modelo de processos para a extração de conhecimento em bases de dados (*Knowledge Discovery in Databases* ou KDD), que possuí, como um de seus principais processos, a mineração de dados. Esta etapa, por sua vez, é responsável pela extração do conhecimento.

O processo de mineração de dados, por ser capaz de extrair informações de bases de dados previamente processadas, apresentou-se como uma técnica importante para o descobrimento de padrões oriundos destas bases. O Instituto de Pesquisa *GartnerGroup* apontou as ferramentas de mineração de dados como uma das cinco mais importantes tecnologias do século XXI, evidenciando a relevância desta técnica (Oliveira, 2012).

Para Cardoso (2008), a Mineração De Dados ou *Data Mining*, engloba um conjunto de técnicas de bancos de dados, inteligência artificial e estatística utilizada para explorar grandes volumes de dados, com o intuito de descobrir novos padrões que sejam proveitosos para alguém.

Paralelo a este cenário, nas últimas décadas, principalmente após o meado da década de 1990, houve um grande crescimento no número de matrículas e de Instituições de Ensino Superior (IES), criadas para atender a progressiva demanda.

Com esse demasiado crescimento, empresários de diversas áreas viram no ensino superior uma oportunidade de ampliar seus fundos, multiplicando assim seus capitais. A grande maioria destes aventureiros, mesmo não possuindo nenhum conhecimento relevante sobre educação, entrou para o ramo, gerando assim um crescimento significante no número de Instituições, onde, muitas das IES criadas desta maneira não possuíam recursos para atender as exigências do MEC

(Ministério da Educação), o que acarretou sérios problemas no atendimento, com quantidade e qualidade. Desta forma, o crescimento desregulado ocasionou então em uma grande quantidade de Instituições com qualidade duvidosa e de alto custo (Souza, 2009).

Este expansionismo, por outro lado, também auxiliou na democratização do acesso a este nível de educação, que passou a abranger novas camadas sociais e estudantes menos qualificados. Para isto, os empresários do setor privado, criaram uma série de IES privadas capazes de abranger esta nova demanda, onde estas Instituições poderiam variar quanto à duração dos cursos, preços, imagem social, qualidade da educação e na estrutura organizacional e administrativa (DIAS SOBRINHO, 2010).

Diante destas mudanças, surgiu a necessidade da criação de indicadores capazes de auxiliar na avaliação qualitativa destas Instituições e seus cursos, com o objetivo de prestar contas à sociedade.

Segundo Primi (2011), é vital que existam indicadores para o controle da qualidade das instituições de ensino. As avaliações, desenvolvidas por entidades públicas, quando aplicadas em grande escala, podem contribuir na análise de qualidade das instituições. Tais avaliações têm como um de seus objetivos produzir informações sobre a eficiência e qualidade das instituições analisadas. Informações estas que podem ser utilizadas na gestão, a fim de melhorar a qualidade do ensino.

Para Dias Sobrinho (2010), a avaliação pode ser utilizada não somente com o objetivo de esclarecer sobre o nível de capacitação profissional dos cursos e sua qualidade, mas também como ferramenta de auxílio para as instituições educacionais. Quando devidamente aplicadas e com boas informações, estas avaliações são capazes de modificar a metodologia de ensino, a gestão, práticas de formação, dentre outros.

No Brasil, o Exame Nacional de Cursos (ENC), mais amplamente conhecido como Provão, foi o responsável por prestar tal análise. O exame foi aplicado aos estudantes que estavam se formando, entre o período de 1996 e 2003, visando

avaliar a Educação Superior no que se diz respeito aos seus cursos de graduação, com base nos resultados gerados sobre o processo de ensino-aprendizagem.

A partir de 2004, o Sistema Nacional de Avaliação do Ensino Superior (SINAES), instituído pela Lei n° 10.861, é responsável por avaliar as Instituições de Ensino Superior (IES). O sistema tem seus processos avaliativos coordenados e supervisionados pela Comissão Nacional de Avaliação Superior (CONAES) e a operacionalização é de responsabilidade do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (BRASIL, 2003)

O SINAES é constituído por três partes principais: avaliação das instituições, avaliação dos cursos e a avaliação de desempenho dos estudantes, que é feita através do Exame Nacional de Avaliação do Estudante (ENADE).

Este exame busca avaliar o desempenho dos estudantes em relação com os conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos de graduação, bem como suas competências e habilidades oriundas de sua formação.

O ENADE é subdividido em três anos, onde cada ano é composto por um conjunto de áreas de ensino. O ano I abrange as áreas da saúde, ciências agrárias e afins. O ano II é formado pelas áreas de ciências exatas, licenciaturas e afins. O ano III é composto pelas áreas ciências sociais aplicadas, ciências humanas e áreas afins. Após todos os anos serem avaliados, o exame volta novamente a avaliar as áreas relacionadas ao ano I, seguindo posteriormente para os demais anos, formando assim um ciclo, onde cada conjunto de áreas é avaliado em um intervalo de três anos.

Segundo Manfredi (2002), a população, em sua grande maioria, associa o estudo e o grau de escolaridade com melhores empregos e garantia de desejáveis posições em suas respectivas profissões. Porém, segundo o autor, esta analogia não é tão simples como se pensa, uma vez que outras variáveis também devem ser levadas em consideração, o que torna este cenário gradativamente mais complexo do que pressuposto.

Com base no contexto apresentado, torna-se possível observar a importância da utilização de ferramentas e técnicas automáticas ou semiautomáticas capazes de gerir e analisar uma enorme quantidade de dados oriunda de uma ou mais bases de dados. A aplicação de tais técnicas e ferramentas busca, como objetivo principal, encontrar padrões e informações úteis e até então desconhecidas, provenientes das bases de dados analisadas.

Segundo Galvão (2009), grande parte dos procedimentos e funções realizados nas instituições públicas e privadas são gravados computacionalmente, o que gera bancos de dados gigantescos que aumentam exponencialmente. Para Silva e Costa (2015); Steiner (2006), quanto maiores estes repositórios de dados, mais complexa e inviável é a interpretação humana, sendo cada vez mais comum encontrar bases tão grandes que superam a capacidade humana. Desta forma, mostrou-se fundamental que ferramentas computacionais fossem capazes de utilizar técnicas inteligentes para auxiliar neste processo de análise e interpretação dos dados para a geração de informação. O KDD tem como principal objetivo esta procura e extração de conhecimento, onde, na sua etapa de Mineração de Dados, este processo se realiza, buscando obter o maior número de informações possíveis oriundas das bases de dados utilizadas (CARVALHO *et al*, 2012); (BOTHOREL, SERRURIER; HURTER, 2011).

Desta forma, a finalidade deste trabalho é utilizar o processo de KDD e de técnicas de Mineração de Dados para a extração de padrões e informações relevantes sobre a base de dados do ENADE 2013. Acredita-se que com o acesso a um maior número de informações o governo e o público poderão conhecer melhor a estrutura e o nível dos cursos analisados neste ano, auxiliando assim na tomada de decisões, como melhores meditas por parte do governo e na escolha dos cursos que se pretende frequentar, por parte da população.

1.2. JUSTIFICATIVA

Com o crescente número de cursos ofertados pelas instituições, é

fundamental que estudos sejam realizados para acompanhar e avaliar este crescimento. A importância destes estudos vem do âmbito social, uma vez que estes têm o potencial de melhorar a oferta e a qualidade dos cursos brasileiros.

O ENADE fornece uma enorme quantidade de dados variados e concretos sobre os estudantes que participaram do exame, possuindo dados referentes, tanto sobre o perfil dos estudantes que fizeram a prova, quanto seu desempenho, além de suas respostas sobre o questionário de percepção da prova. Esta vasta gama de dados fomenta a possibilidade de serem encontrados padrões e conhecimentos ainda desconhecidos e úteis, que poderiam ser aplicados no aprimoramento da educação superior no país.

Foram encontrados também muito poucos trabalhos relativos ao questionário de percepção da prova, que se encontra presente no exame. Além disso, a utilização de técnicas de Mineração de dados na base de dados do ENADE ainda não é muito comum, o que estimula a realização deste trabalho.

Espera-se que, as informações extraídas desta base possam auxiliar estudantes que pretendem ingressar no ensino superior, na sua tomada de decisão quanto ao curso e região, bem como o governo e as próprias IES, fornecendo padrões e conhecimentos capazes de aprimorar tanto as medidas de qualidade, quanto a estrutura de ensino. Com isso, os futuros estudantes de ensino superior teriam uma maior confiança no investimento feito em sua educação, além da geração de profissionais de maior conhecimento e qualidade.

1.3. OBJETIVOS

1.3.1. Objetivo Geral

O objetivo deste trabalho consiste na aplicação de tarefas e métodos de mineração de dados na base de dados do Instituto Nacional de Estudos e Pesquisas

Educacionais Anísio Teixeira (INEP), especificamente, na base de dados do ENADE 2013, buscando identificar e analisar o perfil dos estudantes que prestaram a prova, bem como suas respostas no questionário de percepção, encontrado no exame.

1.3.2. Objetivos Específicos

Além do objetivo geral, pretende-se obter, como propósitos secundários, os seguintes:

(I). Analisar o perfil dos estudantes, juntamente com sua nota em cada parte do exame, relacionado às respostas sobre a questão do nível de dificuldade de cada uma destas, presente no questionário de percepção do exame;

(II). Verificar qual a maior dificuldade no exame, a partir do questionário de percepção, relacionada ao perfil dos alunos;

(III). Verificar a importância do tempo utilizado para realizar a prova com a nota do estudante no exame e sua opinião sobre a extensão do exame prestado;

(IV). Analisar os conhecimentos obtidos com a aplicação das tarefas de mineração.

1.3. ESTRUTURA DO TRABALHO

Além do presente capítulo, esta dissertação está organizada de acordo com os seguintes capítulos:

(I). Capítulo 2: Apresentação de um estudo bibliométrico realizado sobre o assunto, através da busca de artigos nos repositórios da *SCIELO* e *SCOPUS*;

(II). Capítulo 3: retrata a história da avaliação do ensino superior no Brasil;

(III). Capítulo 4: descreve os conceitos de descoberta de conhecimentos em base de dados, bem como, as suas tarefas para o processo de descoberta de conhecimento, detalhando cada uma de suas etapas. Descreve ainda a ferramenta utilizada neste trabalho

(IV). Capítulo 5: apresenta a metodologia deste estudo de caso. Explica-se o processo KDD para a obtenção do conhecimento por meio da mineração de dados, onde é possível verificar os dados interpretados e a validação dos resultados encontrados. Neste capítulo, a execução de cada etapa do processo de descoberta de conhecimento (*KDD - Knowledge Discovery in Database*s), permitindo que se compreenda como o estudo de caso foi realizado e os resultados obtidos;

(V). Capítulo 6: refere-se enfim, as avaliações dos resultados inferidos pelo algoritmo de mineração de dados J48 a partir da base minerada;

(VI). Capítulo 7: são tiradas as conclusões do trabalho, as contribuições e sugestões de trabalhos futuros;

(VII). Referências Bibliográficas: utilizadas no desenvolvimento da pesquisa.

2. ESTUDO BIBLIOMÉTRICO

2.1. DESCRIÇÃO DO ESTUDO BIBLIOMÉTRICO

Segundo Hood e Wilson, 2001 (apud COSTA, 2010 p.116), a bibliometria busca relacionar a pesquisa bibliográfica com o conhecimento estatístico através do estudo de técnicas, com intuito de elaborar métricas sobre informações, para gerar o conhecimento desejado.

Para o estudo presente neste capítulo da dissertação, foram feitas buscas nas bases de dados *SCIELO* e *SCOPUS.* Entende-se que estas bases são bases muito importantes para o levantamento do conhecimento científico, assim, merecendo atenção no momento da pesquisa dos referenciais, por se tratar de uma das mais volumosas bibliotecas de publicações disponíveis em âmbito acadêmico.

Foram encontradas 30 publicações entre 2004 e 2015 para o assunto discutido neste trabalho. Os resultados encontrados auxiliam amplamente e na pesquisa para o entendimento da mineração de dados aplicada à educação e como referencial teórico para o estudo em questão.

Não foi utilizado filtro de exclusão justamente, para que a ferramenta não reduzisse os resultados encontrados. Neste capítulo pode-se analisar e mapear os

artigos encontrados nas bases de dados, referentes à utilização da mineração de dados na educação, aplicada na criação de perfis. Este estudo apresenta um ponto de vista bibliométrico sobre o assunto, identificando as revistas com maior número de publicações na área, os autores mais relevantes e os artigos mais importantes, baseado no número de citações.

2.2. O USO DAS PALAVRAS CHAVES E METODOLOGIAS DA BIBLIOMETRIA

As bases de artigos utilizadas para o desenvolvimento deste trabalho foram a *SCIELO* e *SCOPUS*. Nestas bases, foram feitas as pesquisas somente nos campos *title* (título), *abstract* (resumo) e *keywords* (palavras-chave), com o intuito de gerar um resultado refinado.

As palavras-chave utilizadas foram: “*data mining*" *AND* ("*education*" *OR* "*e-learning*") *AND* ("*profile*" *OR* "*personalization*”). O intervalo temporal da pesquisa compreendeu o ano de 2000 até o ano de 2015. O estudo bibliométrico, apontou os artigos de maior relevância para o tema proposto neste estudo.

Após tal seleção, foram encontrados trinta e seis artigos de diversas áreas e, para eliminar artigos sem relação, algumas destas áreas foram removidas, deixando somente as áreas de *computer science* (ciência da computação), *social science* (ciência social), *engineering* (engenharia), *mathematics* (matemática), *business*, *management and accounting* (administração e contabilidade), *decision sciences* (ciência de auxílio à decisão), *economics education* (educação) e *materials science* (ciência de materiais), totalizando trinta artigos.

Na base *SCIELO*, com a mesma pesquisa, nenhum resultado foi encontrado. Para garantir a inexistência de artigos relacionados nesta base, ainda foi feita a mesma pesquisa com os termos traduzidos, porém, não houve nenhuma alteração no resultado.

Como resultado final da pesquisa nas bases, foram encontrados 30 artigos

publicados, sendo o primeiro artigo publicado em 2004 e o último em 2015.

2.3. RESULTADOS ENCONTRADOS

A partir dos resultados encontrados, que gerou a coleção de artigos desejados, foi feita uma análise para responder os objetivos do trabalho em questão. O primeiro objetivo é quantificar as publicações do tema com o passar dos anos. O segundo, encontrar os artigos que mais influenciaram nesta área. O terceiro, verificar quais os autores com maior número de artigos. O quarto e último objetivo foi descobrir quais revistas publicam mais artigos da área e quais são os países.

O gráfico 1, apresenta o número de artigos publicados na área pesquisada relacionando-se com o ano, tendo a primeira publicação em 2004 e a última em 2015. O ano de 2013, se destaca por possuir o maior número de publicações.

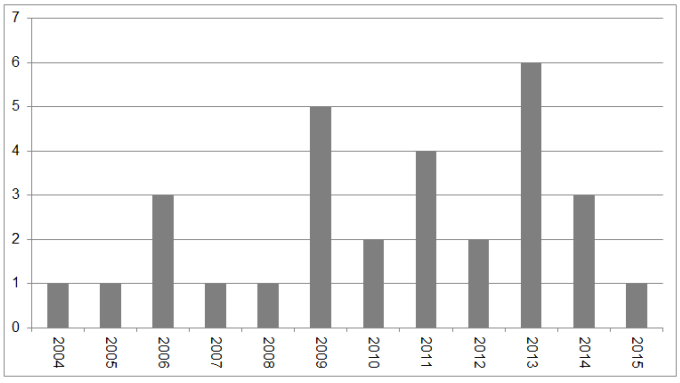

Gráfico 1. Número de publicações por ano.
Fonte: Elaborado pelo Autor (2016).

Nota-se no gráfico 1 que entre os doze anos existentes de publicações, os últimos cinco anos pesquisados (2011 até 2015) são responsáveis por 53% dos artigos publicados.

O Gráfico 2 apresenta os dez artigos mais citados dentre os encontrados na pesquisa, tornando possível observar as publicações que possuem maior influência na área.

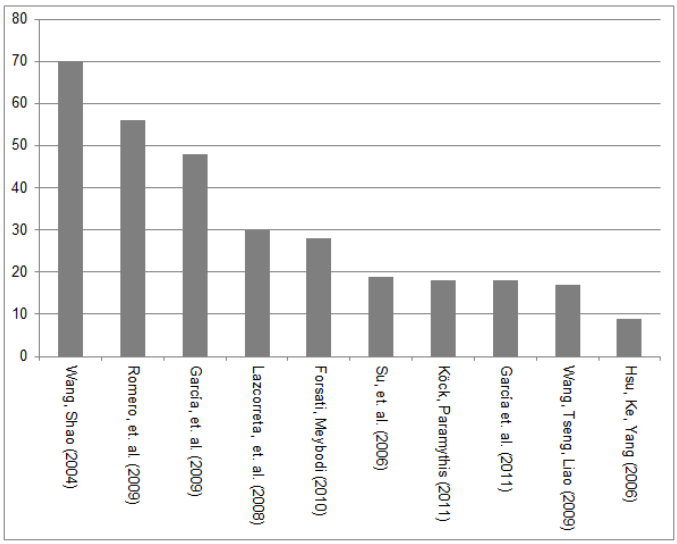

Gráfico 2. Ranking dos artigos mais citados.
Fonte: Elaborado pelo Autor (2016).

Conforme visto no gráfico 2, o artigo mais citado foi o de Wang e Shao (2004), com setenta citações. Em segundo o de Romero, *et. al.* (2009), com cinquenta e seis, e seguido pelo Garcia, *et. al.* (2009) com quarenta e oito citações. Pode-se perceber também que dois dos três artigos mais citados foram publicados em 2009.

Os autores que mais publicaram, dentro do tema pesquisado, estão descritos no gráfico 3.

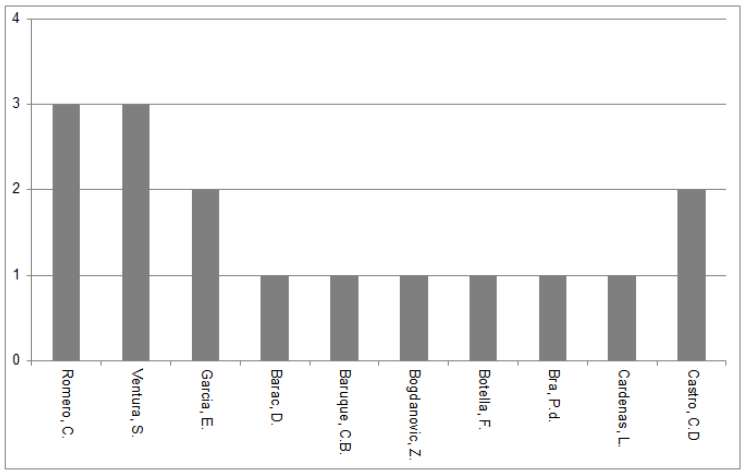

Gráfico 3. Número de artigos por autor.
Fonte: Elaborado pelo Autor (2016).

O gráfico 3 apresentou o gráfico dos autores com maior número de publicações, onde foi possível observar que Romero, C. e Ventura, S., publicaram ambos três trabalhos na pesquisa realizada. Garcia, E. e Castro, C. D., possuem ambos dois trabalhos dentro desta mesma pesquisa.

Neste estudo, foi possível descobrir também quais as revistas científicas que mais aceitaram artigos da área pesquisada, conforme mostra o gráfico presente no Gráfico 4.

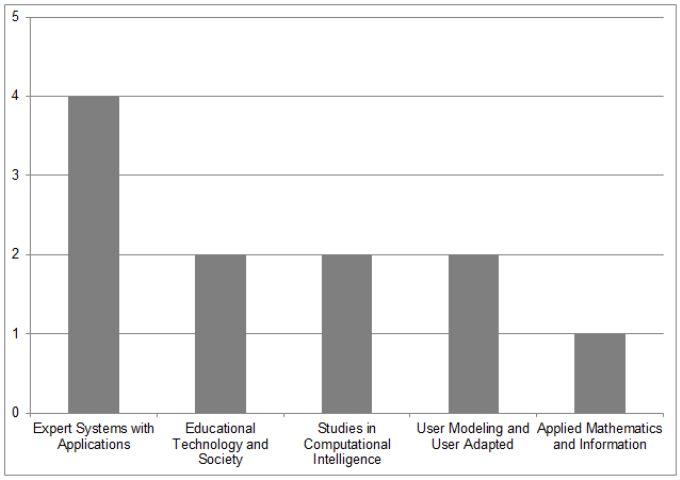

Gráfico 4. Revistas que mais publicam sobre o tema.
Fonte: Elaborado pelo Autor (2016).

Entende-se que a revista *Expert Systems with Applications*, é a revista científica que mais aceita artigos na área, com quatro artigos publicados. A *Educational Technology and Society, Studies in Computacional Inteligence* e a *User Modeling and User Adapted*, possuem cada uma delas dois artigos publicados.

O gráfico 5 apresenta um gráfico com os países que mais publicaram artigos na área pesquisada.

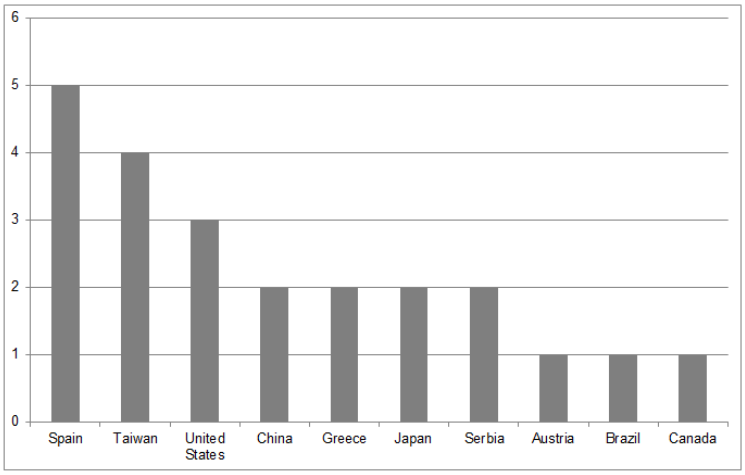

Gráfico 5. Países que mais publicam sobre o tema.
Fonte: Elaborado pelo Autor (2016).

Pode-se observar que a Espanha é o país que mais publica, dentro da pesquisa realizada, com seis artigos publicados. O Taiwan ficou em segundo, com quatro publicações e em terceiro os Estados Unidos, com 3 publicações. O Brasil aparece com apenas uma publicação.

A partir dos principais artigos analisados, é possível encontrar interesses em comum, tendo como objetivo geral personalizar conteúdos e dados para auxiliar tanto professores quanto alunos, na educação.

Nos artigos Garcia *et al.* (2009) e Garcia *et al.* (2011), o enfoque é nos professores, onde estes propõem ferramentas capazes de encontrar, analisar, compartilhar e até sugerir as mudanças mais apropriadas para aprimorar os cursos que ministram. Os dados são adquiridos através dos estudantes e geram um perfil para o professor, que por sua vez pode ser compartilhado e recomendado para professores de perfis semelhantes.

No Su *et al.* (2006), foi apresentado um portfólio de aprendizagem, que utiliza mineração de dados para analisar as características e capacidades dos estudantes, com o objetivo de auxiliar professores no entendimento do porque do mal ou bom rendimento de cada aluno. Para alcançar tal resultado, os dados extraídos dos alunos são divididos para o treinamento e geração de uma árvore de decisão, onde desta forma seriam encontradas as regras personalizadas.

Em Köck e Paramythis (2011) e Wang, Tseng e Liao (2009), ambos propõem aprimoramentos para as sequencias de aprendizagem, buscando encontrar melhores resultados, através da análise dos dados gerada pelos estudantes. Nestes estudos são utilizados respectivamente a criação automatizada de clusters e análise de um algoritmo de árvore de decisão, baseado no perfil dos alunos.

A revisão bibliográfica apresentou quatro artigos que mais se adequaram ao tema, tendo o artigo Garcia *et. al.* (2009) como mais alinhado com o tema.

3. HISTÓRIA DA AVALIAÇÃO NO ENSINO SUPERIOR NO BRASIL

Este capítulo apresenta um estudo sobre a história da avaliação no ensino superior brasileiro, destacando os principais programas e propostas relacionados ao tema.

3.1. CONTEXTO HISTÓRICO DA AVALIAÇÃO DA EDUCAÇÃO SUPERIOR

Historicamente no Brasil, a Campanha Nacional de Aperfeiçoamento de Pessoal de Nível Superior, hoje conhecida como, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), foi fundada em 1951, pelo Decreto nº 29.741. Na década de 70, passa a ser responsável pela elaboração de novas políticas para a pós-graduação e tem sua estrutura alterada pelo Decreto 74.299, se tornando assim, um órgão central superior. Em 1981, através do Decreto nº 86.791, a Capes passa a ser reconhecida como órgão responsável pela elaboração do Plano Nacional de Pós-Graduação Stricto Sensu e, juntamente com isto, passa a ser identificada como Agência Executiva do Ministério da Educação e Cultura, que a torna responsável pela formulação, coordenação, acompanhamento e avaliação das atividades pertinentes ao ensino superior (Capes, 2015).

No que se refere à avaliação dos cursos de graduação, até o começo da década de 80, a Avaliação da Educação Superior era um tema pouco abordado e

com pequeno destaque no âmbito acadêmico. A partir desta década, com um maior número de artigos tratando sobre o tema e demonstrando preocupação com o controle de qualidade das IES, que se deu devido ao repentino aumento no número de matrículas e instituições, surgiram os primeiros ideais sobre o assunto, o Programa de Avaliação da Reforma Universitária, conhecido como PARU (POLIDORI, MARINHO-ARAUJO, BARREYRO, 2006), (MEC, 2003).

3.2. PROGRAMA DE AVALIAÇÃO DA REFORMA UNIVERSITÁRIA (PARU).

O PARU, que surgiu no final do governo militar, a partir da proposta da Associação Nacional de Docentes (ANDES), em 1982, e do MEC, em 1983, teve, também em 1983, sua formalização realizada pelo Conselho Federal de Educação (CFE), que contou com o apoio da Financiadora de Estudos e Projetos (FINEP) e foi coordenado pela Comissão do Aperfeiçoamento de Pessoal de Nível Superior (CAPES). O documento foi produzido e escrito pelo Grupo de Trabalho, que era formado, principalmente, por integrantes da comunidade universitária.

Segundo PARU (1983, apud ALMEIDA JÚNIOR, 2004, p. 84), a educação superior brasileira, na época, apresentava graves problemas, fator que apontava indispensabilidade da geração de novos ideais e estratégias para o seu aprimoramento. Tal necessidade levou então a ideia de que tais metas seriam alcançadas através de uma profunda e metódica avaliação das condições em que se exerce a prática acadêmica.

O programa incorporou estudos específicos e trabalhou com a formulação e implementação de questionários que seriam respondidos pelos dirigentes universitários, docentes e estudantes. Tais ações foram tomadas para que fossem descobertas informações importantes sobre o ensino superior, como: as características do corpo docente e administrativo, a estrutura administrativa, as atividades de ensino, a estrutura e característica das instituições quanto ao aumento exponencial das matrículas, dentre outras (MEC, 2003).

Porém, o programa foi interrompido e desativado em apenas um ano de funcionamento, em razão das discussões e da falta de consenso dentro do Ministério da Educação, sobre quem seria atribuído com a função de realizar e avaliar a Reforma Universitária. Devido a este motivo, muitas instituições sequer passaram das primeiras fases do programa e, os estudos e questionários que possuíam seus dados completos nunca chegaram a ser analisados (CUNHA, 1997).

Em 1985, com a posse do Ministro da Educação Marco Maciel, o PARU, já descartado, foi substituído pela Comissão Nacional para a Reformulação do Ensino Superior (CNRES).

3.3. COMISSÃO NACIONAL PARA A REFORMULAÇÃO DO ENSINO SUPERIOR (CNRES) E GRUPO EXECUTIVO PARA A REFORMULAÇÃO DO ENSINO SUPERIOR (GERES).

A observação de que a avaliação é uma ferramenta vital e necessária para garantir o controle de qualidade da educação superior era cada vez mais evidente, então, com o encerramento do PARU em 1984, foi criada em 1985 a Comissão nacional de reformulação da Educação Superior (CNRES). Esta ficou encarregada de buscar soluções para os problemas urgentes da educação superior, ficando conhecida também como Comissão de Notáveis (ZAINKO, 2008).

A CNRES foi instituída no fim da ditadura militar, pelo Presidente José Sarney, através do Decreto nº 91.177, de 29 de março de 1985. Era constituída de 24 membros, sendo estes professores, sindicalistas, estudantes e industriais (ZANDAVALLI, 2009).

Segundo o MEC (1985), a Comissão, através de contatos formais e informais entre os membros, gerou, no período de seis meses, um relatório final da comissão nacional para reformulação do ensino superior, conhecido como Uma Nova Política Para a Educação Brasileira. Este relatório foi dividido em quatro partes: “1) Princípios e propostas para a nova política e nova universidade; 2) Recomendações feitas pela comissão; 3) Medidas emergenciais e; 4) Declarações de votos de alguns

dos membros da comissão”. (MEC, 1985).

No início do relatório, foram apresentados os principais problemas que afetavam o ensino superior brasileiro:

(I). Professores mal remunerados;

(II). Carência de equipamentos, laboratórios e bibliotecas;

(III). Deficiências na formação profissional dos alunos;

(IV). Descontinuidade das pesquisas;

(V). Discriminação social no acesso às universidades;

(VI). Sistemas antidemocráticos de administração e escolha de quadros dirigentes;

(VII). Crise financeira e pedagógica do ensino privado;

(VIII). Excesso de controles burocráticos nas universidades públicas;

(IX). Pouca clareza na prevalência do sistema de mérito na seleção e promoção de professores (MEC, 1985).

Em seu relatório, o MEC (1985), afirma que a falta de critérios e regras era um dos problemas mais graves do ensino superior brasileiro, pois sem estes dados é impossível responder questões sobre a qualidade do ensino, se este está melhorando ou piorando, quais estados tem as melhores instituições de ensino, se ensinam o necessário para o mercado de trabalho e se o ensino é melhor nas instituições privadas ou públicas. Dessa forma, para o governo isto implica na incapacidade de saber onde melhor direcionar seus recursos. Para os futuros estudantes do ensino superior a impossibilidade de conhecer as instituições com melhor ensino. Para os administradores e professores das instituições o desconhecimento de como aprimorar o ensino.

Contudo, segundo Dotta e Gabardo (2013), igualmente como acontecido com o PARU, à concepção da CNRES sobre a regulação e avaliação do ensino superior não condizia com o ideal esperado pelo Estado, que levou ao MEC a originar o Grupo Executivo para a Reformulação da Educação Superior (GERES), através da Portaria nº 100 de 6 de fevereiro de 1986. O ministro da educação Jorge Bornhausen instituiu o GERES pela Portaria nº 170, de 3 de março de 1986.

O grupo, caracterizado pelo seu teor executivo, contava com apenas cinco membros, todos relacionados com o MEC. O GERES buscava fazer uma reformulação da educação superior do país, através da apresentação de medidas administrativas e legais, além de ficar encarregado de formar propostas. (NOGUEIRA, 2009)

Para alcança seus objetivos, o GERES analisou com profundidade o relatório final da CNRES, além de várias outras propostas disponíveis no ministério. Algumas das medidas abordadas pela CNRES, relacionadas às avaliações, foram vistas como cruciais, pelo grupo, para a educação superior, que apresentou seu suporte para as medidas referentes ao desenvolvimento e implantação de um sistema de avaliação relacionando as instituições e os cursos. (MEC, 1986).

O relatório do GERES foi liberado ainda em 1986 e não contou com o suporte das IES que, em sua maioria, era contra algumas de suas medidas. Seguidamente, o Conselho de Reitores das Universidades Brasileiras (CRUB) e a Associação Nacional dos Docentes do Ensino Superior (ANDES) divulgaram novas propostas de projetos com o intuito de substituir o programa de reformulação do ensino superior, apresentado no relatório do GERES. Tal discordância gerou um vasto número de debates que, como consequência, trouxe o fim ao programa apontado pelo GERES. (ALMEIDA JUNIOR, 2004).

Segundo o MEC (2003), as primeiras experiências relacionadas à avaliação, com um objetivo formativo, foram aplicadas nas instituições públicas entre 1985 e 1986.

3.4. PROGRAMA DE AVALIAÇÃO DAS UNIVERSIDADES BRASILEIRAS (PAIUB).

Os temas de qualidade da educação superior e de avaliação das IES só ressurgiram novamente por parte do governo em 1993, em resposta à insatisfação das IES com relação às propostas feitas até então sobre a avaliação das instituições. (DOTTA E GABARDO, 2013).

A Secretaria de Educação Superior (SESu) fundou então a Comissão Nacional de Avaliação (CNA), que tinha como objetivo elaborar métodos capazes de propiciar um programa de avaliação das instituições de ensino superior do país. Esta comissão sugeriu então a criação do Programa de Avaliação das Universidades Brasileiras (PAIUB). (FRAUCHES, 2014).

Segundo Polidori, Marinho-Araujo, Barreyro (2006), este programa foi caracterizado pelas contribuições feitas pelas IES, pois tinha um teor adoção facultativo e buscava, nas suas avaliações, prezar pela identidade das instituições.

Zandavalli (2009), diz que em 1993, a CNA apresentou o "Documento Básico – Avaliação da Universidade Brasileira: uma proposta nacional", que expõe os ideais desta comissão relativos ao PAIUB. O documento trata de forma clara sobre a fundamentação, os objetivos, os princípios, as formas de elaboração do projeto, as características e os indicadores utilizados na avaliação, explicitando de forma minuciosa os atributos que seriam avaliados.

Com o intuito de melhorar o ensino superior brasileiro, esta proposta de avaliação deverá proporcionar um aprimoramento constante das atividades acadêmicas, gerar uma prestação de contas regrada à sociedade e auxiliar no planejamento de decisões universitárias. Para isto, por esta possuir uma natureza facultativa, as partes envolvidas devem se conscientizar da importância de participar da avaliação, aceitar os critérios e princípios utilizados e auxiliar no desenvolvimento e na execução de medidas capazes de aperfeiçoar o desempenho acadêmico. (MEC, 1994)

Apesar de ter tido um grande apoio das IES brasileiras, o programa sofreu dificuldades em sua implementação. Esta dificuldade surgiu com o fim do apoio por parte do MEC, que significava também um corte no seu financiamento. Tal medida foi tomada em 1996, um ano depois do ingresso do Exame Nacional de Cursos (ENC), que passou a receber o apoio do MEC. Com o fim da parceria, o PAIUB foi aos poucos sendo abandonado, tendo sido capaz apenas de obter as avaliações internas das instituições. (MEC, 2003)

3.5. EXAME NACIONAL DE CURSOS (ENC).

Durante o mandato do presidente Fernando Henrique Cardoso (1994-2001), foi notória a relevância que o governo deu as políticas públicas de avaliação educacional, predominantemente pela criação de um sistema de avaliação voltado para a educação básica. Nesta gestão, ainda foi instituído o “modelo MEC de avaliação”, que era formado pelo Exame Nacional de Cursos (ENC), a Avaliação das Condições de Ensino (ACE) e o Ranking Nacional das IES. (LEITE, 2005)

Os componentes deste modelo foram gradualmente efetivados através da Lei n° 9131/1995, da Lei n° 9394/1996 e do Decreto 2.026, de 10 de outubro de 1996. O resultado destas avaliações foi utilizado para a criação de uma classificação entre as IES que então era fortemente divulgada nas mídias, procurando assim gerar concorrência entre as instituições. Esta abordagem, apresentada pelo MEC, se diferenciou e desconsiderou os programas e processos anteriormente criados. (MEC, 2003); (PEREIRA, 2009)

Segundo Dotta e Gabardo (2013), o ENC tinha como objetivo aplicar provas para estudantes concluintes de cada curso de graduação, a fim de avaliar estes cursos. Após realizados os exames, os cursos recebiam um conceito entre A e E, baseado no desempenho dos alunos, e eram dispostos, em ordem de melhor para o pior, para a criação do ranking nacional de cursos.

O ENC, também conhecido como Provão, era composto então por uma

avaliação dos cursos feita anualmente. O exame era de caráter obrigatório para os estudantes concluintes, uma vez que sem a prestação deste, não era possível a retirada do diploma. Inicialmente, em 1996, o exame foi aplicado somente em três cursos: Direito, Administração e Engenharia Civil. (POLIDORI, MARINHO-ARAUJO, BARREYRO, 2006)

No Decreto nº 3.860, de julho de 2001, a avaliação passa a ser vista, principalmente, como forma de certificação de que uma lista de exigências estava sendo atendida. Estas exigências, por sua vez, são pré-definidas pelo MEC, com o auxílio da comunidade acadêmica. (MEC, 2003)

Dias Sobrinho (2010) diz que, o conjunto de cursos avaliados pelo exame crescia anualmente e chegou a avaliar 26 cursos em 2003, ano em que ocorreu sua última realização. Os resultados gerados pelo Provão eram então unidos aos do ACE, que obtinham seus dados através de visitas feitas *in loco* por especialistas, que observava a infraestrutura física da instituição, a qualificação dos docentes e análise da grade curricular. Os resultados finais gerados eram utilizados para que fosse feito o reconhecimento dos cursos de boa qualidade, além de auxiliar na tomada de decisões referentes aos credenciamentos e recredenciamentos das instituições.

O ENC foi altamente criticado pela sociedade acadêmica e principalmente pelas IES públicas. Este ainda sofreu com protestos e boicotes, que levaram ao MEC a repensar sua proposta inicial sobre a avaliação. (ZANDAVALLI, 2009)

Com a posse de Luiz Inácio Lula da Silva na Presidência da república em 2003, mesmo com os efeitos positivos gerados pelo Provão, como a estruturação formatada do sistema de ensino superior e o sucesso em sua aplicação durante 7 anos, viu-se a necessidade de elaborar um sistema de maior extensão, que fosse capaz de abranger as IES em sua totalidade. Para que tal objetivo fosse alcançado, vários estudos e discussões foram realizados e, ao final, deste processo, foi criado o Sistema Nacional de Avaliação da Educação Superior (SINAES). (ZAINKO, 2008); (DIAS SOBRINHO, 2010); (POLIDORI, MARINHO-ARAUJO, BARREYRO, 2006)

3.6. SISTEMA NACIONAL DE AVALIAÇÃO DA EDUCAÇÃO SUPERIOR (SINAES).

Em 2003, a Comissão Especial de Avaliação (CEA) apresentou uma nova proposta sobre a avaliação do ensino superior brasileiro com o subtítulo de “Bases para uma proposta de avaliação da educação superior”. Este documento era composto de uma proposta base, ou seja, não estava completa, mas serviria de base para o desenvolvimento do sistema de avaliação que existe hoje, o SINAES. (BRASIL, 2003)

O SINAES contou com importantes contribuições, geradas através de grandes debates que contaram com a participação do MEC, dos sindicatos, do parlamento, das sociedades científicas e acadêmicas e da sociedade como um todo. Tais debates acabam causando conflitos entre pensadores de dois paradigmas, o da avaliação baseada nos resultados e no controle externo e o da avaliação formativa e emancipadora. (BRASIL, 2009)

O novo sistema foi estabelecido em 2004, com a Lei 10.861 de 14 de abril de 2004 e regulamentado pela Portaria Nº 2.051 de 9 de julho de 2004. Este é constituído por um conjunto de avaliações, de diferentes estruturas, objetivos e época de utilização e, por um grupo de membros de várias instituições, buscando assim fundamentar por completo o real funcionamento das IES brasileiras. (RIBEIRO, 2015)

Para Oliveira (2013), em sua lei de criação, o SINEAES deixa nítido seu objetivo de garantir que o processo de avaliação das IES, dos cursos de graduação e de desempenho dos estudantes, que sejam cumpridos nacionalmente. Como consequência disto, o sistema tem como propósito a ampliação da oferta e o aperfeiçoamento da qualidade da educação superior, o crescimento contínuo das instituições e das funções sociais e acadêmicas e, principalmente, pela certificação, do respeito à diversidade, reconhecimento de seu dever público, da sua identidade e autonomia e pela realização dos princípios democráticos, por parte das IES.

A CEA, provavelmente por possuir membros ligados ao antigo PAIUB, passou

algumas das características do programa passado para sua proposta base (FRAUCHES, 2014). Segundo Ristoff (2011, apud PINTO, MELLO e MELO, 2016), o novo sistema absorveu alguns dos princípios de seu antepassado, como o comprometimento com a globalidade, a harmonia entre a auto avaliação e a avaliação externa, a estrutura gerada pela avaliação, a continuidade, o reconhecimento da pluralidade do sistema, atuação da comunidade acadêmica e a consideração pela identidade institucional.

Baseando-se nestas características, o SINAES conta com o auxílio de membros da instituição de ensino sendo avaliada e da sociedade que esta faz parte para a realização das avaliações, mantendo assim, a natureza reguladora do Estado (DA SILVA, 2006).

Para que os objetivos avaliativos do sistema fossem alcançados, o processo foi dividido entre duas entidades, a Comissão Nacional de Avaliação da Educação (CONAES), que ficou encarregada da coordenação e organização do processo avaliativo e o Instituto Nacional de Estudos e Pesquisas Educacionais (INEP), que desenvolve e opera as normas avaliativas. (POLIDORI, MARINHO-ARAUJO, BARREYRO, 2006)

Lacerda (2015), afirma que, comparado ao ENC, o SINAES apresenta significativas mudanças na qualidade do método avaliativo das IES. Segundo Dias Sobrinho (2010), isto ocorre, pois, o novo sistema possui um conceito básico, com práticas bem estruturadas, que o possibilitam atingir objetivos lógicos e sólidos. Além disso, outro motivo seria a implantação em escala nacional da avaliação do ensino superior, buscando englobar, por inteiro, as IES, que se baseou em cinco instrumentos de informação e dividiu o sistema em três componentes.

Os instrumentos de informação utilizados como base foram: o sistema de registros CAPES e da Secretaria de Educação Média e Tecnológica (SEMTEC), o censo da educação superior, o projeto político pedagógico específico de cada curso, o cadastro dos cursos e das IES e plano de desenvolvimento institucional. (LEITE, 2005)

Procurando avaliar por completo as IES brasileiras, o SINAES se dividiu em três componentes centrais, a avaliação das IES (AVALIES), a avaliação dos cursos de graduação (ACG) e a avaliação do desempenho dos estudantes, também conhecida como Exame Nacional de Avaliação do Estudante (ENADE). (POLIDORI, MARINHO-ARAUJO, BARREYRO, 2006)

A AVALIES, busca conhecer o perfil das instituições, suas atividades sociais e junto à comunidade e seus programas e projetos. Para isto, conta com a Comissão Própria de Avaliação (CPA), formada com membros da instituição para a realização de uma auto avaliação e uma comissão de especialistas externos, que realizam *in loco*, a avaliação externa; A ACG, através de uma avaliação, que conta com o apoio de uma comissão específica para cada área curso, na sua realização, procura conhecer as condições de ensino disponibilizadas pelas instituições como corpo docente, recursos humanos, estrutura física dos laboratórios e salas de aula e a organização didático pedagógica e; Por fim, o ENADE, que tem como finalidade implantar indicadores de desempenho que levem a uma melhor educação para os estudantes, analisando e moldando as práticas e as atribuições mínimas, além dos conteúdos previstos nas diretrizes curriculares de cada curso, necessários para a criação de um profissional de qualidade. (BRASIL, 2004 L)

Para Ribeiro (2015), as diferentes abordagens do sistema são imprescindíveis, pois a avaliação da educação superior é um tema muito complexo, o que torna o SINAES diferente, por se basear em práticas tradicionais juntamente com outras inovadoras.

O ENADE é segmentado em três anos distintos, onde cada um deles é responsável por um conjunto de áreas de ensino. O ano I abrange as áreas da saúde, ciências agrárias e afins. O ano II é formado pelas áreas de ciências exatas, licenciaturas e afins. O ano III é composto pelas áreas ciências sociais aplicadas, ciências humanas e áreas afins. Ao fim deste processo de três anos, o exame é novamente iniciado, criando um ciclo que avalia as áreas relativas a cada ano periodicamente a cada três anos. (MEC, 2007)

A prova é composta por 10 questões de formação geral, 30 de conhecimento

específico e 9 do questionário de percepção da prova, que não influenciam na nota. O resultado do exame é analisado baseado no Conceito ENADE, que dispõe o resultado numa escala cinco níveis. No ano de 2005, com o intuito de analisar o crescimento do estudante dentro da instituição, foi criado o Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD), que compara os resultados de alunos com até 25% do curso completo, com alunos com no mínimo 75% da grade concluída. Caso o segundo grupo tenha uma nota melhor que o primeiro, o curso receberá uma boa avaliação, porém, caso o primeiro tenha nota próxima ou maior do que o segundo, a avaliação do curso será negativa. (ROTHEN; BARREYRO; 2011).

Segundo Bittencourt *et al* (2009), com o passar dos anos, o MEC passou de um para quatro conceitos utilizados e publicados. No início existia apenas o Conceito ENADE, oriundo da performance dos estudantes ingressantes e concluintes no exame. Em 2005, buscando acabar com as críticas sobre uma possível vantagem por parte das instituições públicas baseada no nível dos estudantes ingressantes, foi criado o IDD para comparar a evolução do aluno durante o curso. No ano de 2008, foram criados o Conceito Preliminar de Curso (CPC) e o Índice Geral de Cursos (IGC). O CPC, que além de contar com o conceito ENADE e IDD, leva em consideração outros elementos em seu cálculo. O IGC, por sua vez busca ampliar os parâmetros analisados e divulga seus resultados sobre as IES anualmente.

O CPC atualmente é calculado através da seguinte fórmula:

$$NCPC_j = 0{,}2\mathrm{NC}_j + 0{,}35\mathrm{NIDD}_j + 0{,}075\mathrm{NM}_j + 0{,}15\mathrm{ND}_j + 0{,}075\mathrm{NR}_j + 0{,}075\mathrm{NO}_j + 0{,}05\mathrm{NF}_j + 0{,}025\mathrm{NA}_j$$

Equação 1. Fórmula utilizada para calcular o Conceito Preliminar de Curso (INEP, 2014)
Fonte: Elaborado pelo Autor (2016).

> Onde NCPC é a Nota contínua do Conceito Preliminar de Curso; NC é a Nota dos Concluintes no ENADE; NIDD é a Nota do Indicador de Diferença entre os Desempenhos Observado e Esperado; NM é a Nota de Proporção de Mestres; ND é a Nota de Proporção de Doutores; NR é a Nota de Regime de Trabalho; NO é a Nota referente à organização didático-pedagógica; e NF é a Nota referente à infraestrutura e instalações físicas (INEP, 2014 a)

O Conceito IGC procura apresentar o nível de qualidade dos cursos de graduação, mestrado e doutorado, em sua totalidade, pertencentes a uma IES, levando em consideração todos os campus e municípios em que esta atua. O indicador utiliza o resultado do CPC e a média dos programas de pós-graduação de cada IES (INEP, 2010). Este índice é medido pela seguinte fórmula:

$$\alpha = \frac{T_G}{T_G + T_{ME} + T_{DE}}$$

Equação 2. Fórmula utilizada para calcular a proporção dos graduandos.
Fonte: INEP (2014).

$$\beta = \frac{T_{ME}}{T_{ME} + T_{DE}}$$

Equação 3. Fórmula utilizada para calcular a proporção dos mestrandos.
Fonte: INEP (2014).

$$IGC_{IES} = \alpha G_{IES} + \frac{(1-\alpha)\beta}{2}(M_{IES} + 5) + \frac{(1-\alpha)(1-\beta)}{3}(D_{IES} + 10)$$

Equação 4. Fórmula utilizada para calcular o Índice Geral de Cursos
Fonte: INEP (2014).

> Onde α é a proporção de graduandos; T_G é o total de matriculados dos cursos de graduação da IES para os quais foi possível calcular o CPC; T_{ME} é o número de mestrandos em termos de graduandos equivalentes da IES; T_{DE} é o número de doutorandos em termos de graduandos equivalentes da IES; β é a proporção de mestrandos; IGC_{IES} é o Índice Geral de Cursos Avaliados da IES; G_{IES} é o conceito médio da graduação da IES; M_{IES} é o conceito médio do mestrado da IES; e D_{IES} é o conceito médio do doutorado da IES. (INEP, 2014 b)

Segundo Bittencourt *et al* (2008), apesar do SINAES possuir um processo avaliativo que engloba as IES por completo, o ENADE é considerado, pela mídia e grande parte das IES, seu elemento mais importante. Ristoff e Giolo (2006) completam dizendo que a maior parte da população e dos estudantes de graduação

supõe que o SINAES e o ENADE são a mesma coisa.

4. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Na década de 80, foi levantada uma preocupação relativa à quantidade de dados armazenados. Segundo Piatetsky-Shapiro (1990), o crescimento da quantidade de banco de dados relativos a um assunto está muito maior e incompatível com a quantidade de conhecimento gerado, surgindo assim à necessidade de uma melhor forma de extração de conhecimento desses bancos de dados.

Segundo Cardoso (2008), para as organizações, o conhecimento é um recurso de muito importante, pois auxilia na tomada de decisões, na criação de planos de negócio e na capacidade de proporcionar produtos e serviços de melhor qualidade. O processo de controle do conhecimento envolve desde a geração e armazenamento dos dados utilizados para o descobrimento, até o ato de identificar e utilizar este conhecimento. Gilbert, Sánchez, Santos (2006), diz que o crescimento da quantidade de dados armazenados aumenta de forma progressiva, uma vez que a tecnologia possibilita a criação de bancos de dados cada vez maiores.

Goes, Steiner (2016), afirma que este crescimento de dados armazenados acontece com organizações das mais variadas áreas e que somente armazenar estes dados não é mais interessante. É necessário então que sejam feitas análises para descobrir se existe algum padrão útil derivado dos dados acumulados.

A prática da procura por padrões interessantes de informação já possuiu

vários nomes. O termo *knowledge discovery in databases* (Extração de Conhecimento em Bancos de dados – KDD) foi aceito como nome para esta atividade em 1989, no primeiro *KDD workshop*, que buscava conscientizar e ressaltar que o objetivo principal desta prática era o conhecimento. (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996)

Durante os anos, surgiram algumas definições e etapas para a realização do KDD. Atualmente, a mais utilizada é a apresentada por Fayyad, Piatetsky-Shapiro e Smyth (1996), que apresenta o KDD como um processo incomum de exploração e descobrimento de diferentes padrões, ainda desconhecidos, que sejam interessante, corretos e de fácil entendimento.

O KDD, para realizar seus processos, utiliza de métodos oriundos de outras áreas de estudo, como da inteligência computacional, estatística, aprendizado de máquina, reconhecimento de padrões, banco de dados, dentre outras. (SASSI, 2012)

O processo de KDD pode possuir um diferente número de etapas, não existindo um único padrão correto. A figura 6 apresenta um destes modelos.

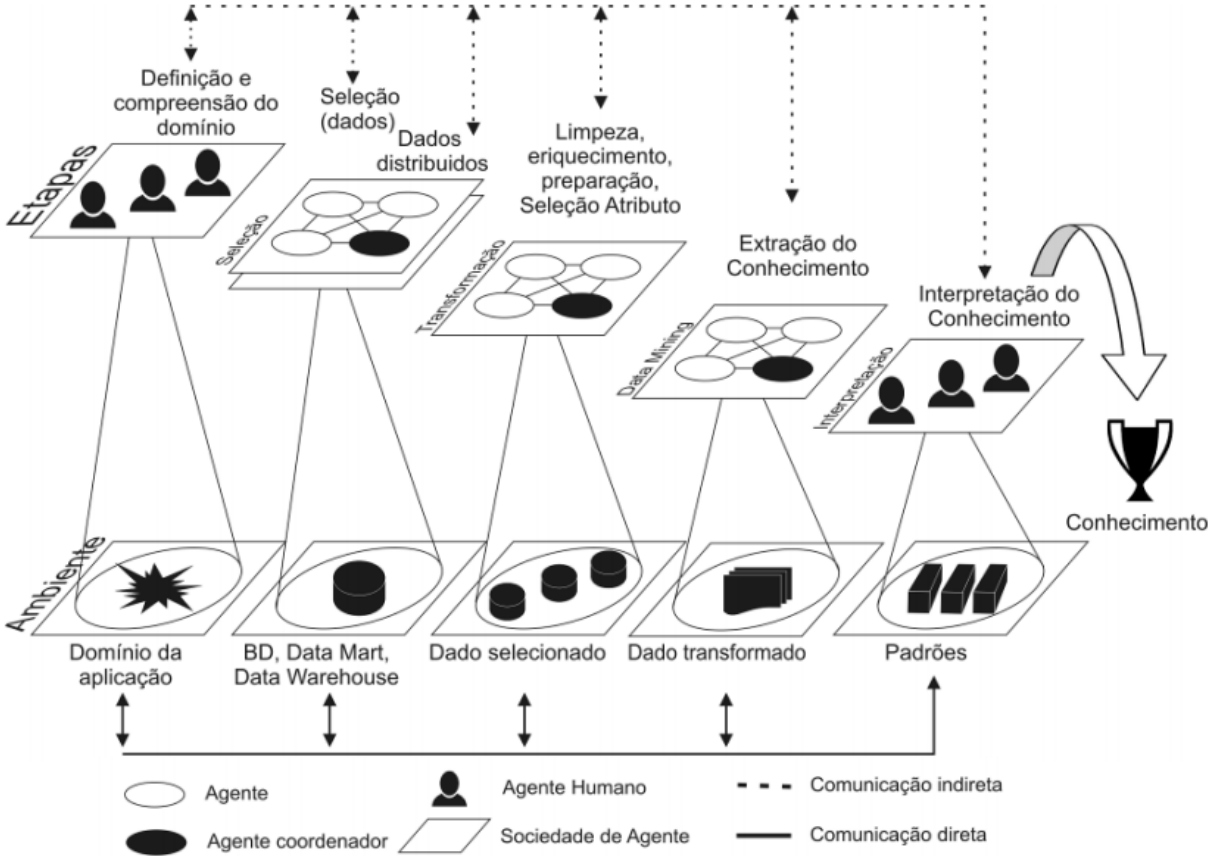

Figura 1. Etapas do KDD (Knowledge Discovery in Databases).
Fonte: Costa ((2008 apud FERNANDES (2010)).

Como pode ser observado na figura 1, as etapas presentes neste modelo são a definição do domínio, a seleção dos dados, a limpeza e transformação dos dados selecionados, a mineração de dados e a interpretação do conhecimento.

Apesar do processo do KDD possuir vários modelos de etapas diferentes, sua ideia e estrutura básica não mudam, uma vez que estes buscam sempre encontrar novos padrões e podem ter sua estrutura reduzida para três etapas, pré-processamento, mineração de dados e pós-processamento.

No modelo adotado, demostrado na Figura 3, as etapas de definição do domínio, seleção dos dados e limpeza dos dados pertencem ao pré-processamento. A mineração de dados é representada pela etapa de extração de conhecimento e o pós-processamento pela etapa final, de interpretação do conhecimento.

Cada uma das etapas citadas será abordada separadamente em futuros

tópicos.

4.1. ETAPA DE DEFINIÇÃO E COMPREENSÃO DO DOMÍNIO

Para que seja possível a seleção, limpeza e mineração dos dados, primeiramente é necessário que seja feita a escolha do assunto a ser estudado, para isto, independente da área de estudo, de vital importância a participação de especialistas para auxiliar neste processo.

Segundo Kasahara, Conceição (2008), para que um domínio seja definido corretamente, é preciso compreender e ter grande conhecimento sobre este. Por esta razão, ter o auxílio de um especialista da área em questão é de suma importância para que um domínio seja estabelecido e, posteriormente, a extração seja feita de forma correta.

Um especialista, que precisa ter um bom conhecimento sobre o domínio, auxiliará na definição do problema, na análise do conteúdo e no desenvolvimento de um escopo geral, além de ajudar na definição de metas, critérios necessários e de passar uma noção básica do que poderá ser descoberto.

Goldschmidt e Passos (2005), diz que os dados podem estar armazenados de duas formas diferente, a natureza real da informação e a representação dos valores relativos a esta informação. No primeiro caso, os dados ainda podem ser interpretados de três maneiras diferentes. Discreto, quando se trata de um atributo sequencial, como um calendário; nominal, utilizado para definir um rótulo ao atributo, como o sexo ou profissão e; contínuo, usualmente utilizado em variáveis numéricas que podem ter um valor finito ou infinito, como o salário ou altura.

4.2. ETAPA DE SELEÇÃO DOS DADOS

Para Dunhan (2002), nesta fase devem ser selecionados os dados que

auxiliarão na descoberta de conhecimento desejado. Azevedo e Santos (2005), completa dizendo que os dados, necessários para a resolução de um problema, podem ser obtidos de diferentes formas, sendo a base de dados e a estatística os mais comuns.

Amaral (2001), afirma que as bases de dados, em geral, possuem uma grande quantidade de dados, relativos a diferentes atributos, se tornando assim indispensável a análise destes para que sejam descobertos quais são os relevantes para a pesquisa em questão. A falha na escolha destes atributos pode afetar indiretamente o resultado da mineração, além de aumentar a complexidade na hora de interpretar o conhecimento gerado.

Os dados ainda podem estar dispersos entre mais de uma base, tabela ou pesquisa estatística, fazendo-se necessária a cópia e junção dos dados e atributos desejados em um único da base ou da tabela. Desta forma, além de evitar problemas com bases de dados em *On-line Transaction Processing* (OLTP), que atualizam o banco constantemente, a análise dos dados selecionados se torna mais intuitiva e simplificada.

Desta forma, a seleção dos dados pode então influenciar na eficiência da mineração, na confiança e qualidade dos resultados gerados e interpretação do conhecimento descoberto.

4.4. ETAPA DE LIMPEZA E TRANSFORMAÇÃO DOS DADOS

Após selecionados, os dados passam pela etapa de limpeza e transformação dos dados. Nela são removidas as redundâncias e as incorreções presentes dos dados. Também é realizada a padronização destes para evitar problemas mineração de dados, fase seguinte do KDD.

Segundo Silva, (2007), esta é uma tarefa trabalhosa e extensa, sendo assim responsável pela maior parte do tempo gasto durante o processo do KDD. Para Dunham (2003), isto acontece devido ao grande número de dados incorretos, ausentes e não padronizados. Estes problemas podem surgir de várias formas,

como: má junção dos dados, diferentes representações métricas, bases mal estruturadas, dentre outros. As incorreções então devem ser trabalhadas e resolvidas, seja por remoção, alteração, codificação ou padronização, para que ao fim esta base esteja estruturada num formato comum entre os dados.

Amaral (2001) ressalta que, nesta etapa deve existir uma preocupação com a transformação dos dados também referente a ferramenta e tarefa utilizada na mineração, uma vez que estas podem variar.

Com isso, a etapa de limpeza e transformação dos dados busca aprimorar a análise que será realizada na fase de mineração de dados, estando assim, diretamente relacionada a qualidade da mineração, do custo e do tempo levado.

4.4. ETAPA DE MINERAÇÃO DE DADOS

Com toda a parte do pré-processamento resolvida, ou seja, com os dados já selecionados, limpos e padronizados, a etapa de mineração de dados pode ser iniciada. Esta etapa é responsável pela extração dos padrões e conhecimentos, independente da quantidade de dados e de sua complexidade.

Para que seja possível realizar a mineração de dados, é necessário que seja selecionado um software, tarefa e algoritmo. Estas escolhas irão definir a forma de realização da mineração e de apresentação dos resultados. (CÔRTES; PORCARO; LIFSCHITZ, 2002).

A etapa de mineração de dados é considerada a mais importante das envolvidas no processo do KDD, sendo até adotada como sinônimo do processo por alguns autores. Dependendo da tarefa e do algoritmo optados, o processo pode ter variações grandes de tempo e até mesmo de desempenho, tornando esta então uma escolha importante para que se tenha um resultado de qualidade. (BOENTE, GOLDSCHMIDT, ESTRELA, 2008)

Segundo Pang-Ning, Steinbach e Kumar (2009), é neste processo que os dados brutos são analisados e possivelmente moldados em informações. Porém, o autor afirma que para se chegar em resultados relevantes, são necessários vários

testes com diferentes tarefas e algoritmos. Porém, Steiner *et al*. (2006) diz que, mesmo com grandes bases de dados e várias análises, a mineração pode não resultar em nenhuma informação nova, ou até mesmo em nenhum conhecimento.

Assim sendo, a mineração de dados pode gerar resultados variados, semelhantes e até vazios, de acordo com a base selecionada e pré-processada, juntamente com a escolha da tarefa e algoritmo. As tarefas mais importantes são: Classificação, Regressão, Clusterização e Associação. Estas tarefas serão descritas nos próximos tópicos, juntamente com o software WEKA©, escolhido para a realização desta etapa.

4.4.1. Tarefa de Classificação

Na tarefa de classificação é necessário que, primeiramente, seja definido um atributo central, escolhido a partir da base selecionada. Os atributos restantes serão analisados com base no atributo central e separados em várias classes de características semelhantes. As técnicas mais utilizadas dentro desta tarefa são: a rede neural, a indução a regras e a árvore de decisão. Dentre todas as tarefas de mineração de dados, a classificação é a mais comumente utilizada. (CHAPMAN et al, 2000).

Carvalho (2005), diz que com o passar dos anos, a tarefa de classificação tem sido cada vez mais estudada e como objetivo descobrir padrões entre a relação dos atributos classe e os atributos previamente definidos, também conhecidos como conjunto de registros.

Segundo Santos e Azevedo (2005), esta tarefa pode ser utilizada para vários propósitos, tendo entre os mais comuns: análise de clientes ou produtos mais vendidos, análise do mercado financeiro, detecção de fraude, dentre outros.

A árvore de decisão, que também pode ser chamada de árvore de classificação, possui características similares a uma árvore comum, como nós,

galhos e raízes. O algoritmo J48 gera modelos de árvores de decisão partindo do topo para base, de forma que, em cada um dos nós, outros atributos sejam avaliados, individualmente, para determinar sua significância na ligação ou até existência nela. (GOLDSHMIDT, PASSOS, 2005); (CRETTON, GOMES, 2016)

Para a realização deste trabalho, foi utilizada a tarefa de classificação, empregando a técnica da árvore de decisão, através do algoritmo J48. Este algoritmo, além de gerar uma árvore de decisão intuitiva, expõe também o conjunto de regras utilizado para o desenvolvimento desta.

Este algoritmo é a transformação do algoritmo C4.5, originalmente escrito na linguagem C, para a linguagem Java, o que tornou possível a utilização deste em ferramentas como o WEKA. Ele aplica o método de divisão e conquista, que divide o problema em vários problemas menores, normalmente com complexidades menores, para assim aumentar a eficiência da árvore de decisão.

O J48 pode ainda trabalhar com atributos contínuos ou discretos, aceita a utilização de valores desconhecidos, através da representação “?” e permite a utilização do método de *prunig*, que reduz significativamente o número de erros encontrados durante a classificação.

4.4.2. Tarefa de Regressão

Esta tarefa se limita na utilização de dados numéricos e é constituída por variáveis de entrada e uma de saída. Nela é feita uma busca por uma função de mapeamento, onde os atributos de entrada são analisados de acordo com o atributo de saída, utilizado como base. A árvore de regressão e a rede neural são as técnicas mais utilizadas para esta tarefa. (HASTIE, TIBSHIRANI, FRIEDMAN, 2001).

Galvão e Marin (2009) diz que esta tarefa tem como estrutura a utilização de técnicas de aprendizado de máquina para analisar novos dados com os aprendidos, desta forma, os novos valores são estimados e preditos com base no aprendizado

anterior. Tem como exemplo de aplicação a probabilidade de recuperação de um paciente, levando em consideração os resultados de uma certa quantidade de exames.

Na tarefa de regressão, quando não se possui dados prévios aos de entrada para a formação da base de análise, parte dos próprios dados de entrada podem ser utilizados para a criação da base. A quantidade de dados utilizados para o treino da base influencia fortemente no nível de predição realizado.

4.4.3. Tarefa de Clusterização

A Clusterização, também chamada de agrupamento, tem como objetivo gerar conjuntos de atributos que se relacionam entre si, possuindo características semelhantes. O número de *clusters* ou agrupamentos pode ser definido e este só possui dados de entrada. Podem ser utilizados para pesquisas de mercado, definição de perfils, etc. O algoritmo mais utilizado desta tarefa é o K-means. (SANTOS, AZEVEDO, 2005).

Segundo Goldschmidt e Passos (2005), esta tarefa busca criar agrupamentos, onde os elementos pertencentes a estes possuam o máximo de propriedades em comum possível, dentro do próprio agrupamento em que fazem parte e um menor número de semelhanças com elementos pertencentes a outros agrupamentos. O Agrupamento, por não possuir dados de saída a serem utilizados como base, pode ser considerado como um processo de aprendizagem não supervisionado.

No algoritmo *K-means*, os dados de entrada são analisados e formam k conjuntos, onde os componentes de cada grupo possuem uma maior afinidade com os outros componentes de seu conjunto. Este algoritmo é muito utilizado quando não se sabe o que esperar como resultado na etapa de mineração de dados.

4.4.4. Tarefa de Regra de Associação

A tarefa de associação é caracterizada pela análise baseada em dependências. Nela os atributos são separados em dependências e consequências. Dessa forma, as dependências se relacionam as consequências como se fossem regras condicionais. Apriori é o algoritmo mais comumente utilizado para a realização desta tarefa. (HAN, HUANG, CERCONE, FU, 1996).

Carvalho (2005) diz que para esta tarefa, não são aceitos dados em formato numérico e que isto deve ser tratado na fase de pré-processamento. Na aplicação do algoritmo apriori, não existe um controle que force um atributo a ser utilizado somente na ação ou reação da regra, o atributo poderá sempre aparecer em ambos os casos, porém não simultaneamente na mesma regra. Pode ser utilizada no auxílio de vendas.

O apriori cria regras baseado no conectivo lógico de condicional, ou seja, as regras são estruturadas afirmando que a primeira parte deve ser cumprida para que se chegue na segunda. Exemplo: Se A e B forem cumpridos, então

$$C.\ A + B \rightarrow C.$$

Equação 5. Conectivo lógico de condicional
Fonte: Elaborado pelo Autor (2016).

4.5. ETAPA DE INTERPRETAÇÃO DO CONHECIMENTO

A etapa de Interpretação do Conhecimento é a última fase do processo do KDD, nele os dados obtidos durante a mineração de dados são analisados, validados e interpretados, com o auxílio do especialista. A confiança dos resultados é de vital importância para que esta análise final seja feita corretamente. (AZEVEDO, 2005).

Segundo Silva (2007), nesta etapa, que também é abordada como pós-

processamento por parte de alguns autores, caso tenha sido gerada uma grande quantidade de resultados, pode ser vantajosa à utilização de análise computacional para facilitar a interpretação dos resultados. O autor ressalta também que ainda nesta fase deve ser feita uma análise do processo de KDD como um todo, começando pelo pré-processamento, até a estrutura escolhida para a mineração dos dados. Caso se veja necessário, o processo deve ser recomeçado, realizando a alterações necessárias, a fim de se obter um melhor resultado.

Pang-Ning, Steinbach e Kumar (2009) afirma que não basta obter resultados com boa confiança e de fácil compreensão, estes ainda precisam úteis e desconhecidos, para só assim ser considerado um resultado vantajoso.

4.6. SOFTWARE WEKA©

O WEKA (*Waikato Enviroment for Knowledge Analysis*) foi desenvolvido na Nova Zelândia, pelo departamento de Ciências da Computação da Universidade de Waikato. É uma ferramenta de mineração de dados composta por uma grande coleção de algoritmos de aprendizagem de máquina. Estes algoritmos são capazes de realizar várias tarefas de mineração de dados, como classificação, regressão, clustering, associação, entre outras.

Maia, Souza (2010) diz que o programa começou a ser desenvolvido em 1993, na linguagem Java. É uma das ferramentas de mineração de dados mais utilizadas e, segundo Vianna, *et al*, (2010) é um software intuitivo e de fácil utilização, que processa os dados de forma rápida e efetiva.

Possui uma interface gráfica amigável e de fácil acesso. A Figura 2 apresenta a tela inicial da ferramenta.

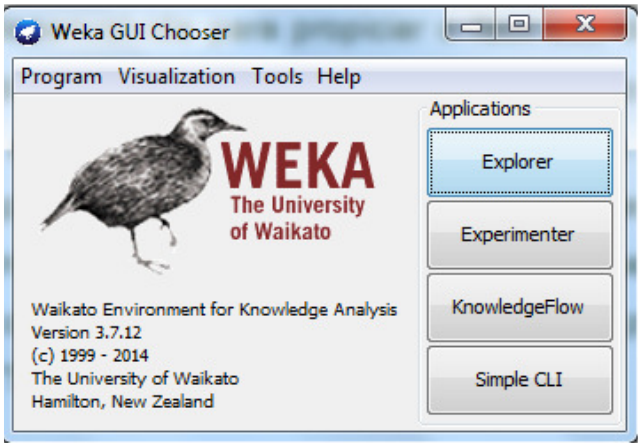

Figura 2. Tela inicial do WEKA 3.7.
Fonte: Elaborado pelo Autor (2016).

A tela inicial do programa, mostrada na Figura 2, permite um fácil acesso ao ambiente gráfico da ferramenta, através da aplicação Explorer. A tela principal pode ser observada na Figura 2.

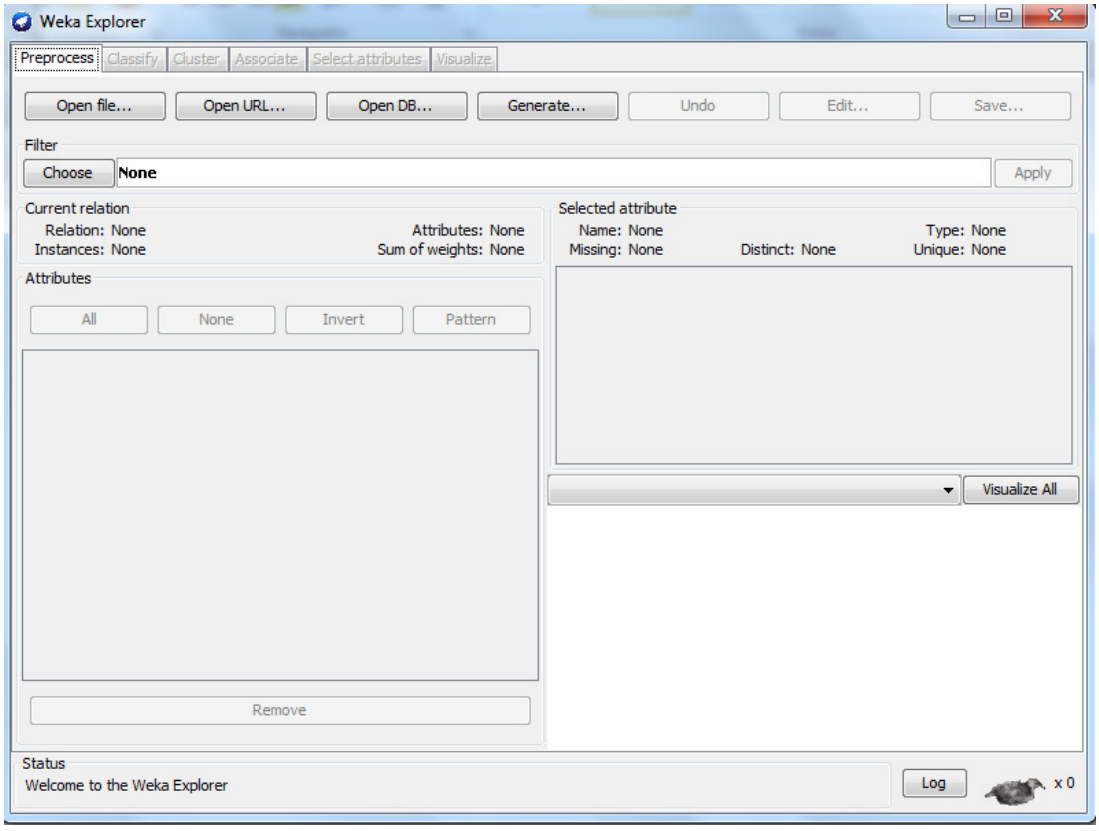

Figura 3. Tela principal do ambiente gráfico do WEKA.
Fonte: Elaborado pelo Autor (2016).

A tela principal do ambiente gráfico da ferramenta, exposta na Figura 3, conterá várias informações pertinentes ao conjunto de dados anexados ao programa. Sendo estas:

Filter: Pode auxiliar na etapa de pré-processo realizando a transformação dos dados.

Current Relation: Provem algumas informações básicas sobre os dados, como o nome da relação, a quantidade de instâncias e atributos e a soma do peso, que caso não alterada, será igual ao número de instâncias.

Attributes: Apresentação de todos os atributos do conjunto de dado anexado. Permitindo que estes possam ser selecionados ou removidos, alterando a estrutura do conjunto.

Selected Attribute: Exibe informações pertinentes aos atributos selecionados, mostrando o nome, o tipo, valores faltantes e únicos em porcentagem, além de gerar gráficos e dados estatísticos sobre estes mesmos dados.

As três abas seguintes a de preprocess, são responsáveis por realizar as tarefas de Classificação, Clusterização e Associação, respectivamente.

5. METODOLOGIA

Segundo FONSECA e NAMEN (2016), existe uma vasta quantidade e dados educacionais e para que estes sejam analisados adequadamente, é necessária a utilização de tecnologias de mineração de dados.

Para a realização deste trabalho, foram utilizados os processos do KDD e o software WEKA 3.7. Esta metodologia foi aplicada para na base do INEP com o objetivo de, encontrar conhecimentos interessantes, através do processo de KDD, realizando os passos necessários para preparar a base, extrair a informação e posteriormente, analisa-la. A ferramenta WEKA foi utilizada para auxiliar na etapa de mineração de dados.

5.1. SELEÇÃO DOS DADOS

A base de dados utilizada para a realização deste estudo foi adquirida no portal do INEP, onde os dados estão disponíveis para o público através de download (INEP, 2016). Os dados escolhidos são oriundos da base de dados do ENADE 2013, que contém dados relativo aos estudantes que prestaram o exame, o questionário de percepção da prova e o questionário do estudante.

Sobre os estudantes é possível encontrar informações como: idade, notas no

exame, tipo de instituição a que pertence, dentre outros. Na parte relativa ao questionário de percepção da prova, são encontradas as respostas relacionadas as questões de caráter perceptivo, respondidas pelos alunos. O questionário do estudante é um questionário socioeconômico realizado previamente a prova e necessário para que seja o local e data desta seja liberada. Contém as respostas sobre questões como a escolaridade dos pais, renda familiar, questões sobre o curso, dentre outras.

Na base escolhida, podem ser encontradas 131 variáveis distintas com dados relativos a quase 200 mil estudantes que realizaram o exame. Relacionando o número de alunos com as variáveis presentes na base, existem mais de 22 milhões da dados dentro da base de dados do ENADE 2013, para serem trabalhados.

Com um número tão grande de dados, os atributos devem ser analisados cuidadosamente, a fim de se descobrir quais possuem real relevância para que os objetivos propostos sejam alcançados.

A base de dados, disposta para *download* no portal do INEP, está em formato de tabela em um arquivo em Microsoft Excel. Uma pequena parte desta base pode ser observada na Tabela 1.

Tabela 1. Parte da base de dados do ENADE 2013

nu_ano	co_grupo	co_ies	cd_catad	cd_orgac	co_munic	co_uf_cur	co_regiao	nu_idade	tp_sexo	ano_fim_	ano_in_gr	tp_semes	in_matut	in_vesper	in_noturn	status	amostra	tp_inscrici	tp_def_fis	tp_d
2013	5	1	1	1	5103403	51	5	22	M	2008	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	21	M	2008	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	28	F	2002	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	25	M	2005	2006	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	26	F	2004	2006	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	22	F	2008	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	M	2007	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	F	2007	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	20	F	2009	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	F	2006	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	21	M	2009	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	22	F	2007	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	24	M	2006	2008	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	F	2007	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	F	2007	2008	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	24	F	2006	2007	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	M	2006	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	22	F	2008	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	22	F	2008	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	34	F	1998	2005	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	30	F	2001	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	26	M	2006	2010	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	23	F	2007	2009	2	1	1	1	1	1	0		
2013	5	1	1	1	5103403	51	5	25	M	2005	2006	2	1	1	1	1	1	0		

Fonte: Elaborado pelo Autor (2016).

Nota-se que, na tabela 1, os dados e atributos estão dispostos de uma forma incompreensível. Tal razão se dá pela codificação das colunas e seus valores, feita por parte do INEP, dificultando assim, o entendimento dos dados.

Para que estes dados se tornassem legíveis, foi necessária a utilização e estudo do Dicionário de Variáveis, arquivo representado por uma tabela, também em formato Excel. O Quadro 1 apresenta o início deste documento.

	DICIONÁRIO DE VARIÁVEIS - ENADE 2013				
	NOME DA VARIÁVEL	TIPO	TAMANHO	DESCRIÇÃO DA VARIÁVEL	DESCRIÇÃO DAS CATEGORIAS
IES DO CURSO / IES	nu_ano	Numérica	8	Ano de realização do exame	-
	co_grupo	Numérica	8	Código da Área de enquadramento do curso no Enade	5 = MEDICINA VETERINÁRIA 6 = ODONTOLOGIA 12 = MEDICINA 17 = AGRONOMIA 19 = FARMÁCIA 23 = ENFERMAGEM 27 = FONOAUDIOLOGIA 28 = NUTRIÇÃO 36 = FISIOTERAPIA 38 = SERVIÇO SOCIAL 51 = ZOOTECNIA 55 = BIOMEDICINA 69 = TECNOLOGIA EM RADIOLOGIA 90 = TECNOLOGIA EM AGRONEGÓCIOS 91 = TECNOLOGIA EM GESTÃO HOSPITALAR 92 = TECNOLOGIA EM GESTÃO AMBIENTAL 3501 = EDUCAÇÃO FÍSICA (BACHARELADO)
	co_ies	Numérica	8	Código da IES (e-Mec)	-
	cd_catad	Numérica	8	Código da categoria administrativa da IES	1 = Pública Federal 2 = Pública Estadual 3 = Pública Municipal 4 = Privada com fins lucrativos 5 = Privada sem fins lucrativos
	cd_orgac	Numérica	8	Código da organização	1 = Universidade 2 = Centro Universitário

Quadro 1. Parte do Dicionário de Variáveis do ENADE 2013.
Fonte: Elaborado pelo Autor (2016).

O Dicionário de Variáveis, descrito na Figura 10, que é obtido juntamente com a base de dados do ENADE 2013, contém as descrições, tanto das colunas, quanto de seus valores, presentes na base de dados do exame. Com isso, os dados da base escolhida puderam começar a ser analisados para que a seleção dos mais relevantes fosse realizada.

Em uma primeira e extensa análise, os atributos que seriam utilizados nas próximas etapas foram reduzidos de 131 para 19, resultando numa diminuição de aproximadamente 85% no número de dados, caindo para 3,4 milhões. Porém, com a realização de vários testes e análises, levando em consideração os resultados, seus respectivos níveis de confiança e os objetivos traçados, os atributos selecionados

foram revisados.

Nesta revisão, surgiu a necessidade de eliminação de 5 destes 19 atributos previamente selecionados e com isso, nestes 14 atributos considerados relevantes eram contidos 2,5 milhões de dados.

Ainda foi realizada nesta fase uma separação de bases. Este procedimento foi feito buscando aprimorar os resultados obtidos e a leitura analítica destes. A base até então construída, possuindo 14 atributos, foi dividida em 4 bases distintas, onde cada uma dessas foi direcionada para o cumprimento de um objetivo específico. Três destas bases passaram a ter 8 atributos e uma, passou a possuir 9. O Quadro 2 mostra como foi festa esta divisão.

Atributo	1º Base	2º Base	3º Base	4º Base
co_grupo				
cd_catad				
cd_orgac				
co_regiao_curso				
nu_idade				
tp_sexo				
nt_fg				
nt_ce				
nt_ger				
co_rs_i1				
co_rs_i2				
co_rs_i3				
co_rs_i7				
co_rs_i9				

Quadro 2. Relação dos atributos pertinente a cada base.
Fonte: Elaborado pelo Autor (2016).

No Quadro 2, é possível observar os atributos que fazem parte de cada uma das quadro bases criadas. Os primeiros 6 atributos podem ser encontrados em todas as quatro bases. São eles o co_grupo, cd_catad, cd_orgac, co_regiao_curso, nu_idade e tp_sexo. Os demais atributos variam, pertencendo somente a algumas destas bases. Na 1º Base, podem ser encontrados também os atributos nt_fg e

co_rs_i1; Na 2º Base, os atributos nt_ce e co_rs_i2; Na 3º Base, os atributos nt_ger e co_rs_i7 e; Na 4º Base, os atributos nt_ger, co_rs_i3 e co_rs_i9.

Estes atributos, ainda com seus nomes originais, codificados pela INEP tem suas respectivas descrições apresentadas no Quadro 3.

Atributo	Descrição
co_grupo	Código da Área de enquadramento do curso no Enade
cd_catad	Código da categoria administrativa da IES
cd_orgac	Código da organização acadêmica da IES
co_regiao_curso	Código da região de funcionamento do curso
nu_idade	Idade do inscrito em 24/11/2013
tp_sexo	Sexo do inscrito
nt_fg	Nota bruta na formação geral
nt_ce	Nota bruta no componente específico
nt_ger	Nota bruta da prova
co_rs_i1	Qual o grau de dificuldade desta prova na parte de Formação Geral?
co_rs_i2	Qual o grau de dificuldade desta prova na parte do Componente Específico?
co_rs_i3	Considerando a extensão da prova, em relação ao tempo total, você considera que a prova foi:
co_rs_i7	Você se deparou com alguma dificuldade ao responder à prova. Qual?
co_rs_i9	Qual foi o tempo gasto por você para concluir a prova?

Quadro 3. Relação dos atributos com suas respectivas descrições.
Fonte: Elaborado pelo Autor (2016).

Os atributos descritos no Quadro 3 representam as colunas de maior relevância, encontradas na base de dados do ENADE 2013, para o estudo em questão.

Com a base de dados e seus atributos devidamente selecionados e descritos, o processo de KDD deste trabalho segue para a próxima etapa, onde será realizada a limpeza e transformação dos dados selecionados, preparando-os para a mineração de dados.

5.2. LIMPEZA E TRANSFORMAÇÃO DOS DADOS

A segunda etapa do processo KDD é responsável pelo tratamento dos dados previamente selecionados, eliminando registros vazios e incorretos, além de formata-los, quando necessário.

A base de dados, já com seus atributos selecionados, tem seus dados parcialmente apresentados na tabela 2. Com o intuito de facilitar a compreensão, a imagem demostra os dados antes de sua divisão em quatro bases.

Tabela 2. Base de dados com atributos já selecionados.

co_grupo	cd_catad	cd_orgac	co_regiao	nu_idade	tp_sexo	nt_fg	nt_ce	nt_ger	co_rs_i1	co_rs_i2	co_rs_i3	co_rs_i7	co_rs_i9
5	1	1	5	22	M	68.6	69.1	69	C	D	C	B	C
5	1	1	5	21	M	63.9	63.7	63.8	C	C	B	B	E
5	1	1	5	28	F	59.7	35.7	41.7	C	C	A	D	D
5	1	1	5	25	M								
5	1	1	5	26	F	75.2	55.9	60.7	B	C	C	B	C
5	1	1	5	22	F	70	67.3	68	C	D	C	B	B
5	1	1	5	23	M	63.3	52	54.8	C	D	B	A	D
5	1	1	5	23	F	47.9	50.9	50.2	C	C	C	B	C
5	1	1	5	20	F	59.8	65.9	64.4	C	C	C	B	.
5	1	1	5	23	F	72.2	43.1	50.4	A	B	C	D	B
5	1	1	5	21	M	0	48	36	C	C	C	C	C
5	1	1	5	22	F	76.2	52.6	58.5	C	C	C	B	D
5	1	1	5	24	M								
5	1	1	5	23	F	45.1	41.9	42.7	C	C	C	D	B
5	1	1	5	23	F	60.7	23.7	33	C	D	B	E	B
5	1	1	5	24	F	46.5	67.8	62.5	C	C	B	B	C
5	1	1	5	23	M	45	43.1	43.6	C	D	B	B	B
5	1	1	5	22	F	53.6	62.7	60.4	C	C	C	E	D
5	1	1	5	22	F	62.7	34.5	41.6	B	C	C	A	B
5	1	1	5	34	F	65.3	44.4	49.6	C	D	A	C	E
5	1	1	5	30	F	54.6	54.5	54.5	C	D	B	A	C
5	1	1	5	26	M	7.2	36.7	29.3	A	A	A	C	A
5	1	1	5	23	F	59.3	44.1	47.9	C	D	B	D	C
5	1	1	5	25	M	62.9	54.3	56.5	C	C	C	B	C

Fonte: Elaborado pelo Autor (201).

Como pode ser facilmente observado na tabela 2, existem vários campos vazios. Uma linha de registro, que contenha qualquer número de campos vazios, deve ser eliminada, já que comprometeria os resultados finais.

Analisando a base e buscando identificar todos os campos vazios, foram encontrados e eliminados quase 28 mil registros com pelo menos um campo em branco.

Após finalizada a eliminação dos campos vazios, foi realizada uma análise dos valores encontrados em cada um dos atributos selecionados, com o intuito de identificar dados incorretos ou desnecessários para a pesquisa. Um dos processos utilizados para encontrar valores equivocados foi analisar os dados os relacionando com as descrições presentes no Dicionário de Variáveis. Caso o valor não existisse no dicionário, este era considerado errado, fazendo com que sua linha de registro fosse descartada.

Alguns dos problemas descobertos foram, por exemplo, valores “N” para o atributo tp_sexo, que deveria conter somente valores “M” e “F” e; valores “.” ou “*” para os atributos co_rs_i1, co_rs_i2, co_rs_i3, co_rs_i7 e co_rs_i9 que, por mais que fossem previstos no Dicionários de Variáveis, estes tem como descrição valor em branco e valores múltiplos, respectivamente, tornando-os desnecessários para o estudo. Nestes atributos foram aceitos somente os valores “A”, “B”, “C”, “D” e “E”.

Com a limpeza dos dados vazios, incorretos e irrelevantes, eliminando completamente os registros a que estes dados faziam parte, a base de dados passou para pouco mais de 2 milhões de dados.

A partir daí a base passou a ter uma credibilidade muito maior, uma vez que os dados inconsistentes foram eliminados. Com isso, a base de dados estava pronta para ter seus valores codificados transformados, a fim de facilitar o entendimento sobre esta.

Por mais que esta seja uma base codificada, nem todos os atributos precisaram ter seus valores transformados. Os atributos que tiveram estão descritos no Quadro 4.

Atributo	Código	Descrição dos códigos
co_grupo	5	Medicina Veterinária
	6	Odontologia
	12	Medicina
	17	Agronomia
	19	Farmácia
	23	Enfermagem
	27	Fonoaudiologia
	28	Nutrição
	36	Fisioterapia
	38	Serviço Social
	51	Zootecnia
	55	Biomedicina
	69	Tecnologia Em Radiologia
	90	Tecnologia Em Agronegócios
	91	Tecnologia Em Gestão Hospitalar
	92	Tecnologia Em Gestão Ambiental
	3501	Educação Física (Bacharelado)
cd_catad	1	Pública Federal
	2	Pública Estadual
	3	Pública Municipal
	4	Privada com fins lucrativos
	5	Privada sem fins lucrativos
cd_orgac	1	Universidade
	2	Centro Universitário
	3	Faculdade
	4	Ifet/Cefet
co_regiao_curso	1	Norte
	2	Nordeste
	3	Sudeste
	4	Sul
	5	Centro-Oeste

Quadro 4. Relação dos valores em código dos atributos que foram transformados em suas respectivas descrições
Fonte. Elaborado pelo autor (2016).

Os atributos presentes no Quadro 4 tiveram seus valores na base alterados pelas suas respectivas descrições. Porém, nem todos os atributos que necessitavam de uma transformação puderam ser alterados desta forma, uma vez que isto afetaria significativamente o desempenho da mineração de dados. Os Quadros 5 e 6

apresentam tais atributos.

Intervalos referentes ao atributo nu_idade	Descrição
<= 23	Para todas as idades menos ou iguais a vinte e três anos
> 23 e < 30	Para todas as idades maiores que vinte e três e menores que trinta anos
>= 30	Para todas as idades maiores ou iguais a trinta anos.

Quadro 5. Valores do atributo nu_idade transformados e relacionados com suas respectivas descrições
Fonte. Elaborado pelo autor (2016).

O atributo referente a idade, não estava com seus valores codificados, então esses não precisaram ser alterados de acordo com suas respectivas descrições, porém, idade é um atributo que possui um grande conjunto de valores e a utilização de cada um destes valores em específico, não seria interessante e prejudicaria uma extração de conhecimento mais aprofundada, durante a fase de mineração de dados. Por estes motivos, A idade, assim como os atributos referentes as notas dos estudantes, foram transformadas em intervalos, que podem ser observados nos Quadros 5 e 6.

Intervalos referentes aos atributos nt_fg, nt_ce e nt_ger	Descrição
< 60	Para todas as notas menores que sessenta
>= 60 e < 80	Para todas as notas maiores ou iguais a sessenta e menores que oitenta
>= 80	Para todas as notas maiores ou iguais a oitenta.

Quadro 6. Valores dos atributos nt_fg, nt_ce e nt_ger transformados e relacionados com suas respectivas descrições.
Fonte. Elaborado pelo autor (2016).

Assim como com a idade, as notas também possuem uma grande quantidade de valores, então para estas notas, foram criados intervalos semelhantes, capazes de agrupar os valores e manter a integridade dos dados. O Quadro 6 representa tal feito. O desenvolvimento dos intervalos desta variável se deu com o intuito de analisar o desempenho do estudante, onde quando, com nota menor que sessenta, este é considerado com um desempenho ruim, com nota maior ou igual a sessenta e

menor que oitenta, é um desempenho regular e com nota maior ou igual a oitenta, o desempenho é bom.

Restaram ainda atributos que não precisaram ser modificados, seja por já apresentarem os valores reais ou por conter uma codificação amigável, que será favorável nas etapas seguintes. O Quadro 7 identifica estes atributos.

Atributo	Código	Descrição dos códigos
tp_sexo	F	Feminino
	M	Masculino
co_rs_i1	A	Muito fácil
	B	Fácil
	C	Médio
	D	Difícil
	E	Muito difícil
co_rs_i2	A	Muito fácil
	B	Fácil
	C	Médio
	D	Difícil
	E	Muito difícil
co_rs_i3	A	Muito longa
	B	Longa
	C	Adequada
	D	Curta
	E	Muito curta
co_rs_i7	A	Desconhecimento do conteúdo
	B	Forma diferente de abordagem do conteúdo
	C	Espaço insuficiente para responder às questões
	D	Falta de motivação para fazer a prova
	E	Não tive qualquer tipo de dificuldade para responder à prova
co_rs_i9	A	Menos de uma hora
	B	Entre uma e duas horas
	C	Entre duas e três horas
	D	Entre três e quatro horas
	E	Quatro horas e não consegui terminar

Quadro 7. Relação dos valores em código dos atributos não transformados com suas respectivas descrições.
Fonte. Elaborado pelo autor (2016).

Os atributos presentes no Quadro 6 apresenta a relação entre os atributos não transformados com suas respectivas descrições dos códigos. Foi optado por manter as opções de A, B, C, D e E para os atributos relativos as questões do questionário de percepção da prova, pois os valores reais são extensos, fator que afetaria a leitura dos resultados e provavelmente a confiança destes.

O resultado final desta etapa pode é apresentado na Tabela 3.

Tabela 3. Base de dados pré-processada

co_grupo	cd_catad	cd_orgac	co_regiao	nu_idade	tp_sexo	nt_fg	nt_ce	nt_ger	co_rs_i1	co_rs_i2	co_rs_i3	co_rs_i7	co_rs_i9
MEDICINA	Federal	Universid	Centro-Oe	<= 23	M	>= 60 e <8	>= 60 e <8	>= 60 e <8	C	D	C	B	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	M	>= 60 e <8	>= 60 e <8	>= 60 e <8	C	C	B	B	E
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	F	< 60	< 60	< 60	C	C	A	D	D
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	F	>= 60 e <8	< 60	>= 60 e <8	B	C	C	B	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	>= 60 e <8	>= 60 e <8	C	D	C	B	B
MEDICINA	Federal	Universid	Centro-Oe	<= 23	M	>= 60 e <8	< 60	< 60	C	D	B	A	D
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	< 60	< 60	< 60	C	C	C	B	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	< 60	< 60	A	B	C	D	B
MEDICINA	Federal	Universid	Centro-Oe	<= 23	M	< 60	< 60	< 60	C	C	C	C	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	< 60	< 60	C	C	C	B	D
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	< 60	< 60	< 60	C	C	C	D	B
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	< 60	< 60	C	D	B	E	B
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	F	< 60	>= 60 e <8	>= 60 e <8	C	C	B	B	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	M	< 60	< 60	< 60	C	D	B	B	B
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	< 60	>= 60 e <8	>= 60 e <8	C	C	C	E	D
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	< 60	< 60	B	C	C	A	B
MEDICINA	Federal	Universid	Centro-Oe	>= 30	F	>= 60 e <8	< 60	< 60	C	D	A	C	E
MEDICINA	Federal	Universid	Centro-Oe	>= 30	F	< 60	< 60	< 60	C	D	B	A	C
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	M	< 60	< 60	< 60	A	A	A	C	A
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	< 60	< 60	< 60	C	D	B	D	C
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	M	>= 60 e <8	< 60	< 60	C	C	C	B	C
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	< 60	< 60	< 60	C	C	D	B	D
MEDICINA	Federal	Universid	Centro-Oe	<= 23	F	>= 60 e <8	>= 60 e <8	>= 60 e <8	C	D	C	E	C
MEDICINA	Federal	Universid	Centro-Oe	> 23 e < 30	M	< 60	< 60	< 60	C	D	B	D	C

Fonte: Elaborado pelo Autor (2016).

Na tabela 3 é possível ser observada a transformação e estruturação dos valores pertencentes aos atributos selecionados e, diferente do que acontece na Figura 8, não são mais vistos valores vazios, uma vez que todos foram eliminados.

Com o fim da etapa de tratamento dos dados, a parte de pré-processamento, do processo de KDD está completa. Tal realização permite que a base avance para a próxima etapa do processo de KDD, a mineração de dados é considerada a principal fase do processo.

5.3. MINERAÇÃO DE DADOS

A etapa de mineração de dados tem como finalidade a aplicação de técnicas e algoritmos de mineração, em grandes bancos de dados, onde esses serão intensamente analisados e explorados, buscando encontrar padrões e assim extraindo informações úteis. (CRETTON, GOMES, 2016)

Com o intuito de aprimorar este processo de mineração e ter um melhor enfoque nos objetivos, a base foi previamente divida em 4 outras bases, cada uma com um foco específico. Os atributos pertencentes a cada uma destas bases pode ser observado no Quadro 2.

Para que a extração do conhecimento fosse realizada, foi necessária a seleção de uma ferramenta capaz de auxiliar neste processo. O WEKA foi esta ferramenta, pois além deste ser *open source*, ter uma interface gráfica amigável e possuir várias tarefas e algoritmos já acoplados, este também é uma das ferramentas mais utilizadas da área.

O WEKA não aceita documentos salvos nos formatos padrões do Excel, como .xls ou .xlsx, então estes precisam ser salvos especificamente no formato CSV (separado por vírgulas), caso isto não seja feito ou feito incorretamente, ocasionará em um erro ou até em uma captura incorreta dos dados.

Depois de feita a mudança do formato, o WEKA deve ser executado e aberto em sua aplicação “Explorer”, que dará acesso ao modo gráfico da ferramenta. Quando este for carregado, clique no botão “Open File...”, procure pela sua base e em “Arquivos do tipo:” escolha “CSV data files”, seleciona seu arquivo e abra-o. A Figura 13 apresenta uma das bases já carregada no programa.

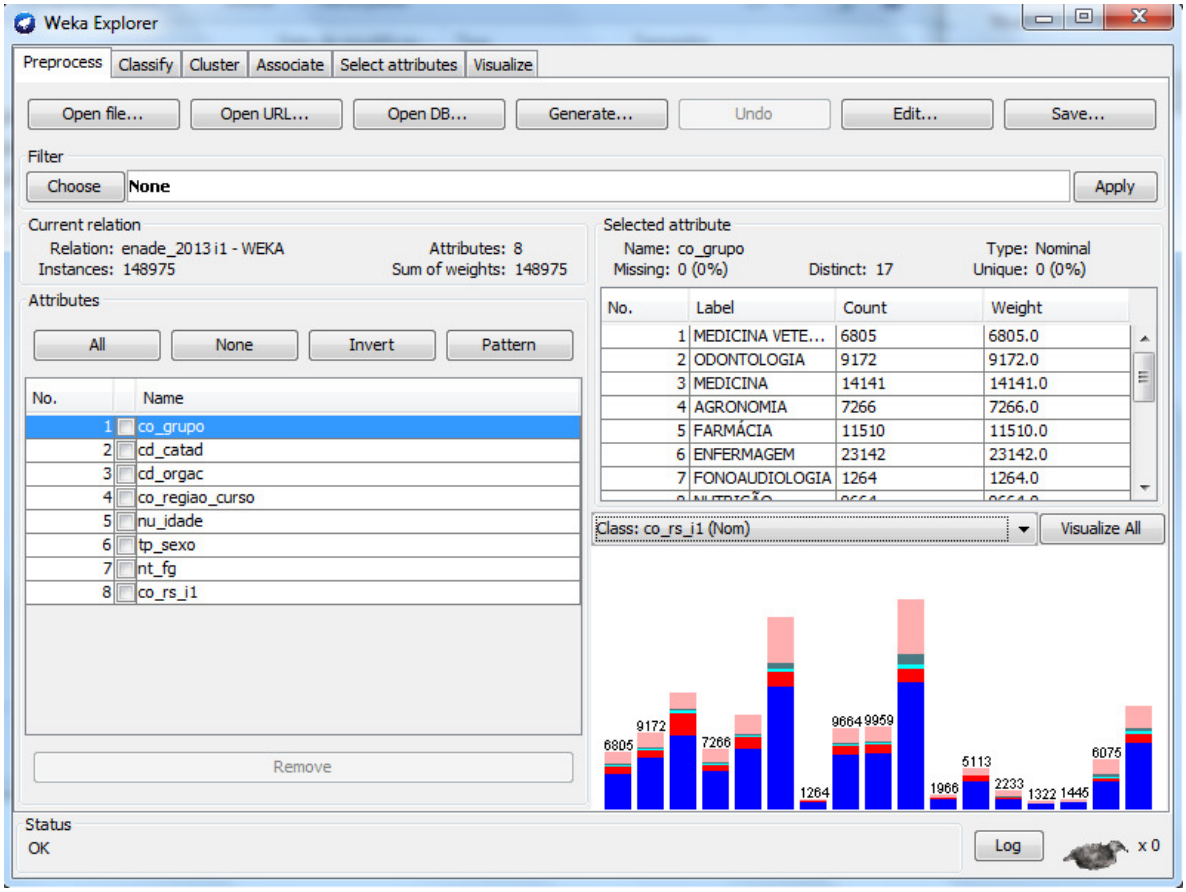

Figura 4. Ambiente gráfico do WEKA com a 1º base carregada.
Fonte: Elaborado pelo Autor (2016).

Na Figura 4 é possível observar a base já carregada, apresentando seus atributos, alguns dados básicos e estatísticos, os valores presentes na no atributo selecionado co_grupo e um gráfico relacionando o atributo co_grupo selecionado com o cs_rs_i1, que pode ser alterado acima deste mesmo gráfico.

A tarefa escolhida para a fase de mineração foi a de Classificação, através do algoritmo de arvore de decisão J48. Estes serão descritos nos tópicos seguintes.

5.3.1. Tarefa de Classificação e Algoritmo J48

Na etapa de mineração de dados, a base de dados, trabalhada ao longo dos processos anteriores do KDD, foi examinada classificada através de técnica de classificação e utilizou a técnica de árvore de decisões.

A técnica de árvore de decisão, presente na tarefa de classificação é uma das mais utilizadas e intuitivas, nela os padrões encontrados são modelados e apresentados em formato de árvore, fator que facilita a observação das ligações de um padrão. (PANG-NING, STEINBACH; KUMAR, 2009).

Por isto, a tarefa de classificação, juntamente com a técnica de árvore de decisões, foram escolhidos, uma vez que esta se apresentou propícia a obtenção de melhores resultados para que os objetivos desta pesquisa fosse cumprido. O algoritmo J48, por sua vez, foi escolhido por ser considerado o melhor entre os de arvore de decisão.

No WEKA, ao selecionar a aba “Classify”, o programa estará em sua parte de classificação, onde pode ser escolhido o algoritmo desejado. A Figura 5 mostra este ambiente inicial.

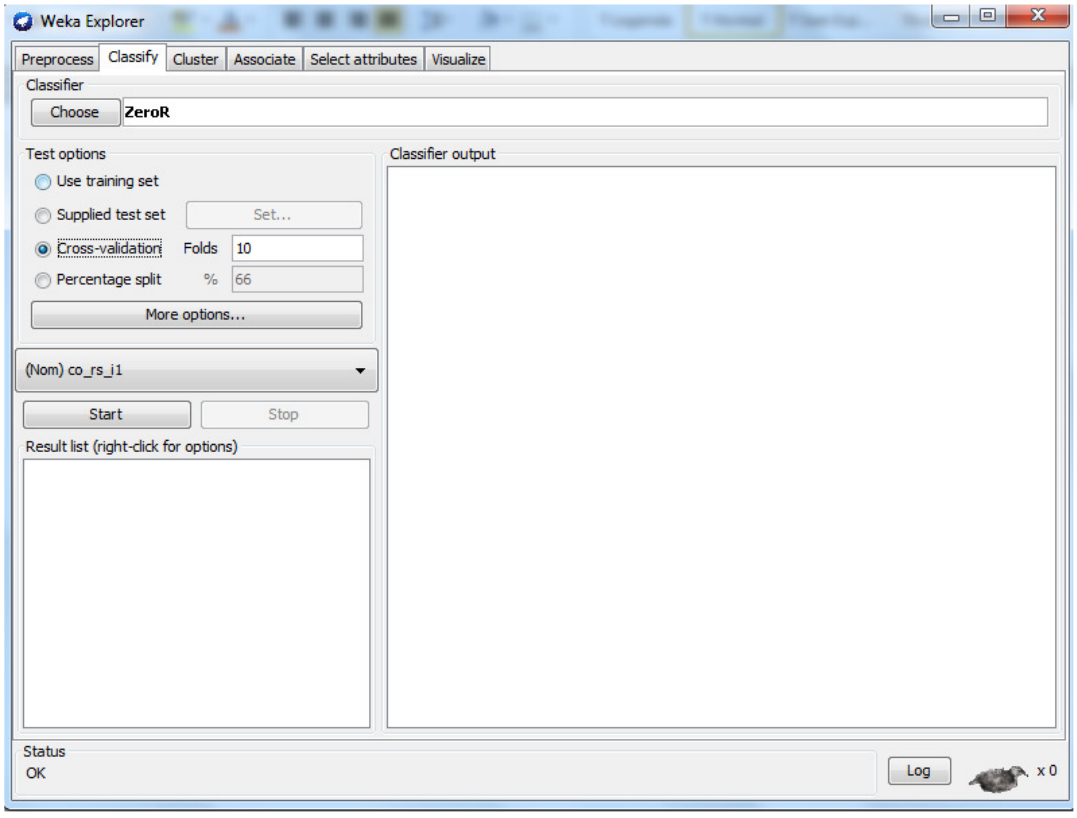

Figura 5. Aba de Classificação do WEKA.
Fonte: Elaborado pelo Autor (2016).

O algoritmo pode ser alterado, selecionando “choose” na parte Classifier, apresentada na Figura 5. Com isso, uma tela de pesquisa irá aparecer, onde será necessário escolher a pasta “trees”, que é onde todos os algoritmos de arvore de decisão se encontram. Nesta parta, opte pelo algoritmo J48.

Depois disso, o processo passa a variar de acordo com a base utilizada. Foi necessário minerar 4 bases de dados, onde cada uma tinha um propósito diferente e, para que este propósito fosse alcançado e como os atributos presentes nestas bases variam entre si, foram escolhidos diferentes atributos para servir como classificador. Tal escolha foi feita buscando atingir o maior nível de confiança possível e melhores resultados a partir de cada base.

Com isso, os atributos selecionados como classificadores foram os tipos de nota presentes em cada uma das 4 bases, uma vez que estas possuíam grande impacto na pesquisa. O Quadro 8 apresenta a relação das bases com seus respectivos atributos classificadores.

Base de Dados	Atributo Classificador
1º Base	nt_fg
2º Base	nt_ce
3º Base	nt_ger
4º Base	nt_ger

Quadro 8. Relação das bases de dados com seus respectivos atributos classificadores..
Fonte: Elaborado pelo Autor (2016).

Como pode ser observado no Quadro 8, para a primeira base foi utilizado o atributo nt_fg, que contém o intervalo das notas sobre a formação geral; a segunda base utilizou o atributo nt_ce, que possui o intervalo das notas sobre o componente específico; a terceira e quarta base adotaram o atributo nt_ger, que compreende o intervalo das notas gerais do exame.

Com a aplicação da tarefa de classificação, pelo algoritmo J48, na etapa de mineração de dados, um modelo classificador foi gerado para da uma das bases, juntamente com suas respectivas arvores de decisão, porém, foram muitos os

resultados e informações extraídas. Estes resultados foram analisados e estruturados, optando pelos padrões de maior referencia com o estudo.

6. RESULTADOS E DISCUSSÕES

As árvores apresentadas são derivadas da base de dados do INEP, do ENADE 2013, que possui uma grande quantidade de dados pertinentes aos estudantes que prestaram a prova. Esta base foi trabalhada através dos processos do KDD, onde os dados foram selecionados, pré-processados, transformados e por fim minerados.

A partir da base tratada, foi feita uma mineração de classificação por meio do algoritmo J48, que apresentou como resultados as regras e a árvore de decisão. Cada uma das quatro bases trabalhadas obteve um nível de confiança distinto, isto acontece, pois, apesar de possuírem atributos iguais, estas ainda diferem em outros, o que as torna únicas, tendo, além de níveis de confiança diferenciados, resultados bem variantes.

O nível de confiança de cada uma das bases pode ser observado no Quadro 8.

Base de dados	Confiança
1º Base	80,86%
2º Base	80,73%
3º Base	83,82%
4º Base	84,05%

Quadro 9. Relação das bases de dados com suas respectivas confianças.
Fonte: Elaborado pelo Autor (2016).

Conforme apresentado no Quadro 9 a 1º base obteve uma confiança de 80,86%, a 2º base de 80,73%, a 3º base de 83,82% e a 4º base de 84,05%, valores que demonstram o potencial dos padrões e informações gerados.

As informações geradas através da técnica de classificação foram ainda analisadas e refinadas, buscando obter resultados diretos e intuitivos. Visando ainda esta meta, as árvores foram separadas de acordo com o curso, categoria da instituição e base de dados a que pertence. Entretanto, será possível observar que nem todas as variáveis presentes nestes atributos serão apresentadas, isto ocorreu, pois estas variáveis não geraram informações interessantes para o estudo em questão.

Em destaque, é possível observar nestas árvores a influência dos atributos cd_catad e co_grupo, que representam a categoria das IES e o curso, respectivamente. Nos resultados gerados, ambos atributos foram utilizados como primeira ou segunda instância, tornando-os em atributos raiz, do qual os ramos seriam formados. Como terceira instância, foi empregado, principalmente, o atributo co_regiao_curso, que possui as divisões regionais do país, isso mostra que existe uma diferença entre estas regiões e expõe a relevância deste na mineração.

Os primeiros resultados apresentados serão pertinentes a 1º Base, que tem como objetivo analisar o perfil dos estudantes, relacionados com sua nota na parte de formação geral do exame e com a resposta da primeira pergunta do questionário de percepção da prova, que busca saber qual o grau de dificuldade desta prova na parte de formação geral.

Na 1º Base, somente um curso foi capaz de gerar padrões detalhados, este foi o curso de medicina. Os demais cursos, presentes no Quadro 3, também geraram informação, porém, o número de instâncias com nota da formação geral menor que sessenta foi tão grande que impossibilitou o descobrimento de maiores detalhes.

A Figura 6 apresenta os principais resultados da 1º base, com uma árvore relativa as IES Federais.

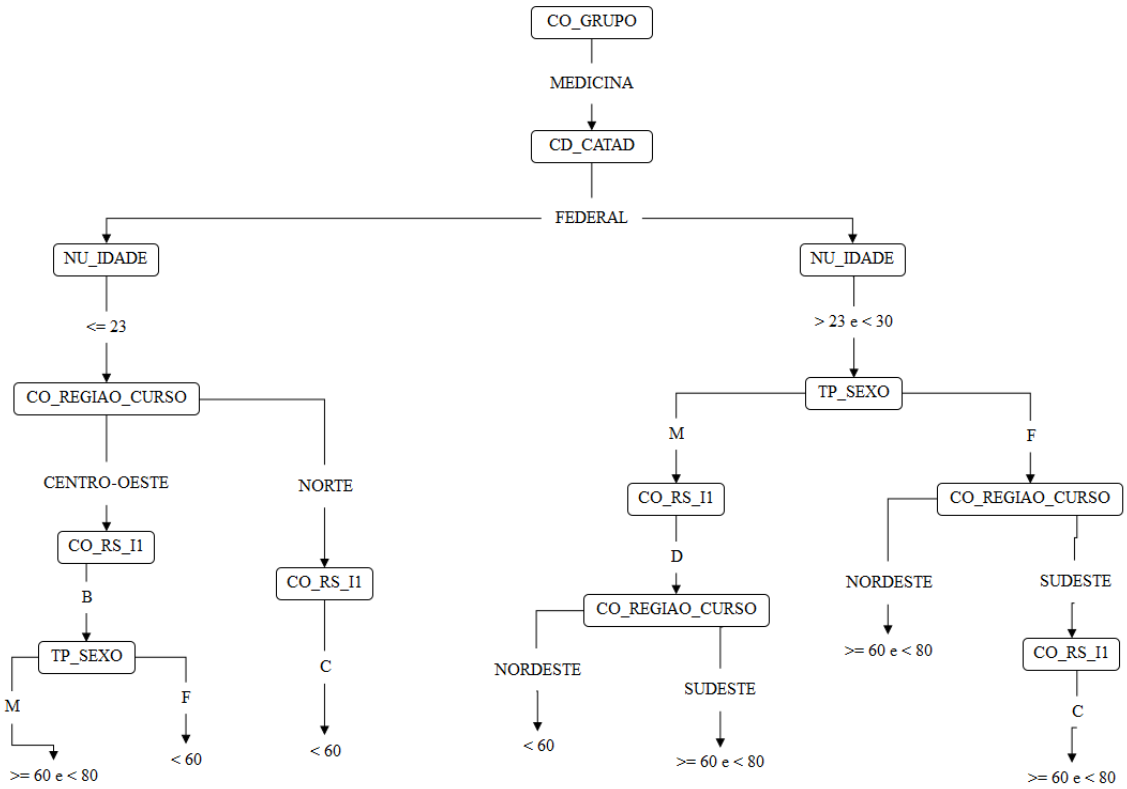

Figura 6. Arvore de decisão referente às IES Federais da 1º Base.
Fonte: Elaborado pelo Autor (2016).

Observando a Figura 6, é possível analisar os principais padrões referentes às IES Federais, como também, obter conhecimentos importantes relativos a estes padrões.

Nas IES Federais, os estudantes do curso de medicina com idade menor ou igual a vinte e três anos, quando da região norte do país, responderam no questionário de percepção da prova, como médio o grau de dificuldade da formação geral do exame e obtiveram uma nota menor que sessenta. Os alunos ainda deste curso, tipo de instituição e idade, quando da região centro-oeste do país, marcam como fácil o grau de dificuldade da formação geral do exame e são do sexo

masculino, tiraram nota maior ou igual a sessenta e menor que oitenta, mas quando estes são do sexo feminino, obtiveram um resultado menor que sessenta na parte de formação geral.

Neste mesmo tipo de instituição e curso, os estudantes com idade maior que vinte e três anos e menor que trinta anos, quando do sexo masculino, respondendo como difícil à parte de formação geral da prova e da região nordeste, receberam nota menor que sessenta, mas quando da região sudeste, tiraram nota maior ou igual a sessenta e menor que oitenta na parte do exame relativa à formação geral. Os alunos ainda com mesma idade, curso e tipo de instituição, quando do sexo feminino e da região nordeste, obteve nota maior ou igual a sessenta e menor que oitenta e quando da região sudeste, respondendo como médio o nível de dificuldade da formação geral da prova, receberam também nota maior ou igual a sessenta e menor que oitenta.

Os resultados mais significativos da 1º base pertinentes às IES Estaduais, em formato de arvore de decisão são mostrados na Figura 7.

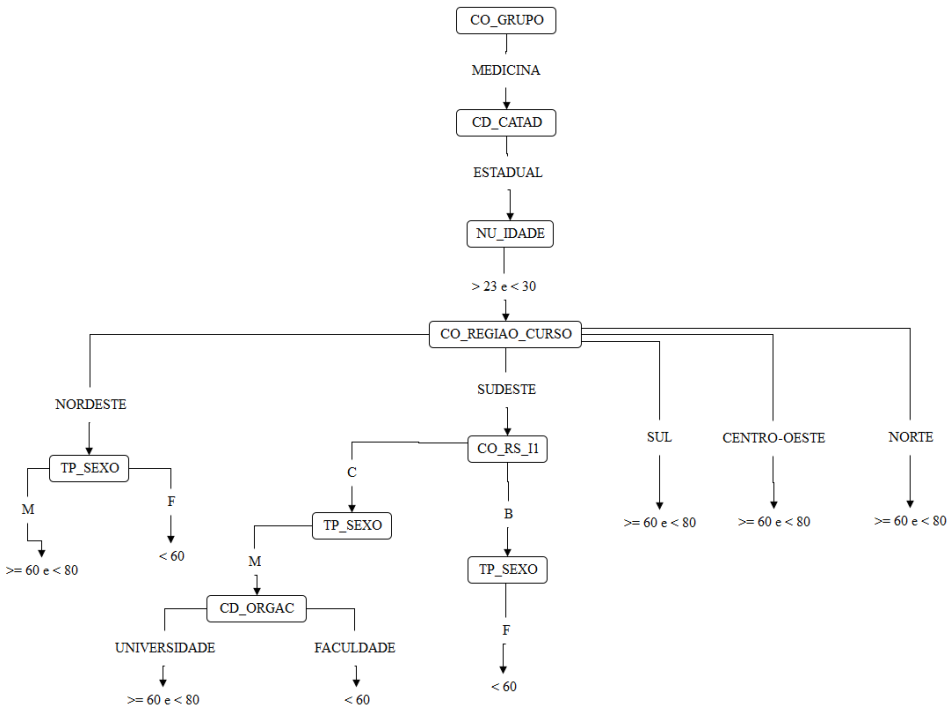

Figura 7. Arvore de decisão referente às IES Estaduais da 1º Base.
Fonte: Elaborado pelo Autor (2016).

Os estudantes do curso de medicina das IES Estaduais com idade maior que vinte e três anos e menor que trinta anos, quando da região nordeste do país e do sexo masculino, receberam nota maior ou igual a sessenta e menor que oitenta, mas quando do sexo feminino, obtiveram nota menor que sessenta. Quando da região sudeste, respondendo como fácil o nível de dificuldade da formação geral do exame e do sexo feminino, receberam nota menor que sessenta, porém ainda nessa região, quando respondem como médio o grau de dificuldade desta parte, são do sexo masculino e assim, vindos de uma universidade, obtiveram uma nota maior ou igual a sessenta e menor que oitenta, mas quando oriundos de uma faculdade, tiraram nota menor que sessenta na parte de formação geral da prova.

Nas regiões norte, sul e centro-oeste, os estudantes do curso de medicina, pertencentes à IES Estaduais e com idade entre vinte e três e trinta anos, independente dos outros atributos, conseguiram nota maior ou igual a sessenta e menor que oitenta.

Na Figura 8, são representados os resultados mais relevantes da 1º base, sendo estes relativos às IES Municipais.

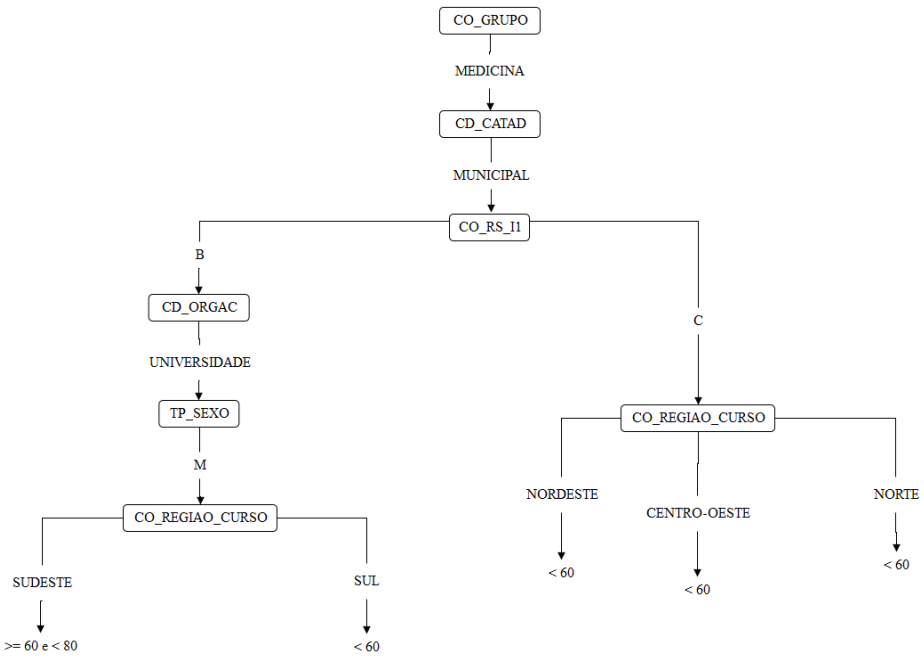

Figura 8. Arvore de decisão referente às IES Municipais da 1º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Municipais, os estudantes do curso de medicina, quando respondem como fácil o nível de dificuldade da parte de formação geral, são oriundos de universidades, do sexo masculino e da região sudeste do país, estes obtiveram resultado maior ou igual a sessenta e menor que oitenta, mas quando da região sul, receberam nota menor que sessenta na parte de formação geral do exame.

Os alunos, ainda deste tipo de instituição e do curso de medicina, quando marcaram como médio o grau de dificuldade e são da região nordeste, centro-oeste e norte, tiraram nota menor que sessenta.

Os principais resultados relacionados às IES Privadas sem fins lucrativos, da 1º base, são mostrados na arvore de decisão da Figura 9.

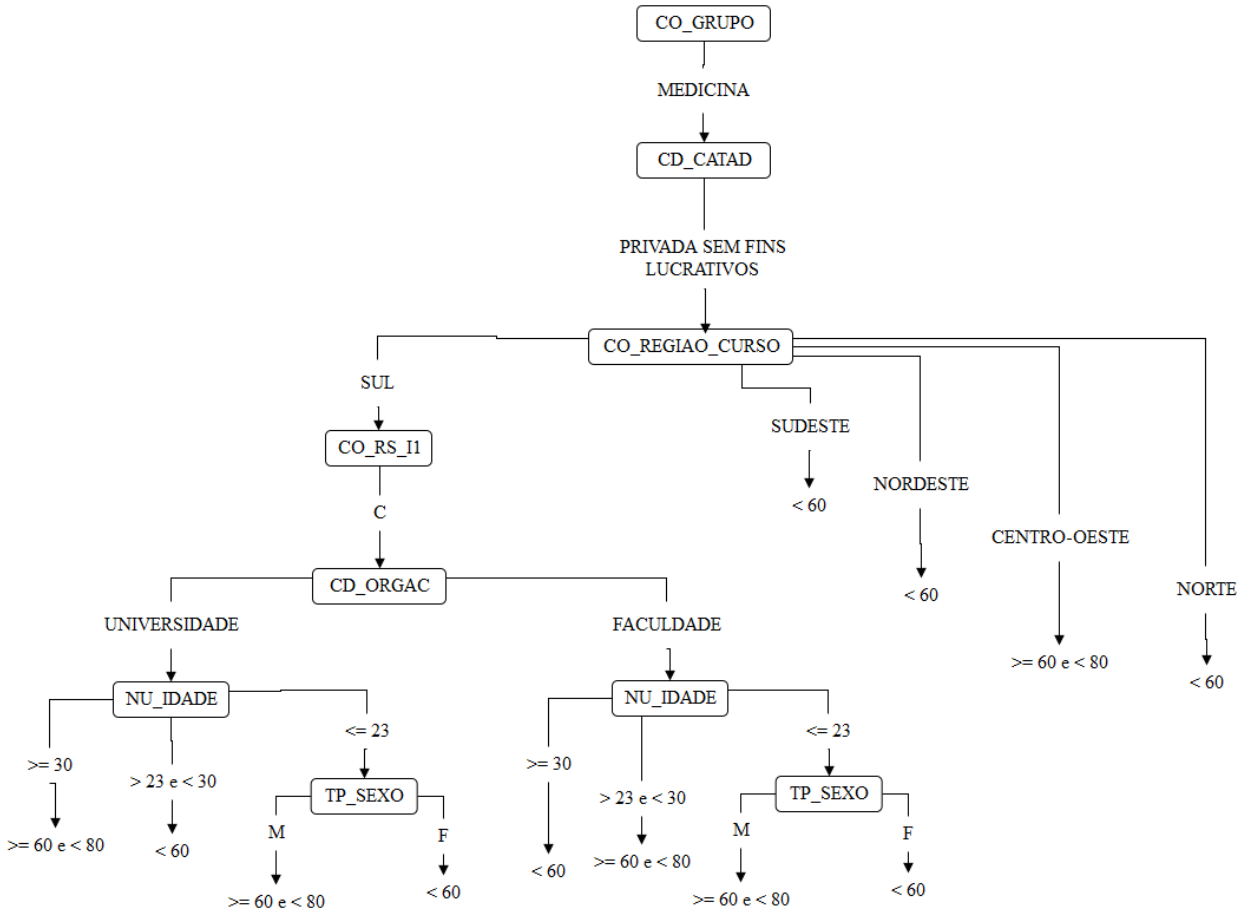

Figura 9. Arvore de decisão referente às IES Privadas sem fins lucrativos da 1º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Privadas sem fins lucrativos, os estudantes do curso de medicina, quando da região sudeste, nordeste e norte do país, obtiveram resultado, na parte de formação geral do exame, menor que sessenta, mas quando do sudeste, estes receberam nota maior ou igual a sessenta e menor que oitenta, independente dos outros atributos.

Neste mesmo tipo de instituição e curso, os alunos do sul, respondendo como médio o grau de dificuldade da parte de formação geral, quando oriundos de universidades, tem seus resultados variando de acordo com a idade, onde com idade maior ou igual a trinta, receberam nota maior ou igual a sessenta e menor que oitenta, com idade entre vinte e três e trinta anos, obtiveram nota menor que sessenta e com idade menor ou igual a vinte e três anos, os estudantes do sexo masculino tiveram um rendimento de maior ou igual a sessenta e menor que oitenta na parte de formação geral da prova e menor que sessenta, quando do sexo feminino.

Os alunos oriundos de faculdade, do mesmo tipo de instituição, curso, região do país e resposta sobre o grau de dificuldade da parte de formação geral da prova, obtiveram resultados diferentes de acordo com a idade, onde com idade maior ou igual a trinta, receberam nota menor que sessenta, com idade maior que vinte e três anos e menor que trinta anos, tiraram nota maior ou igual a sessenta e menor que oitenta e com idade menor ou igual a vinte e três anos, quando do sexo feminino, tiveram resultado menor que sessenta, mas quando do sexo masculino, obtiveram nota maior ou igual a sessenta e menor que oitenta na parte de formação geral do exame.

A arvore de decisão relativa aos resultados mais relevantes das IES Privadas com fins lucrativos, da 1º base é apresentada na Figura 10.

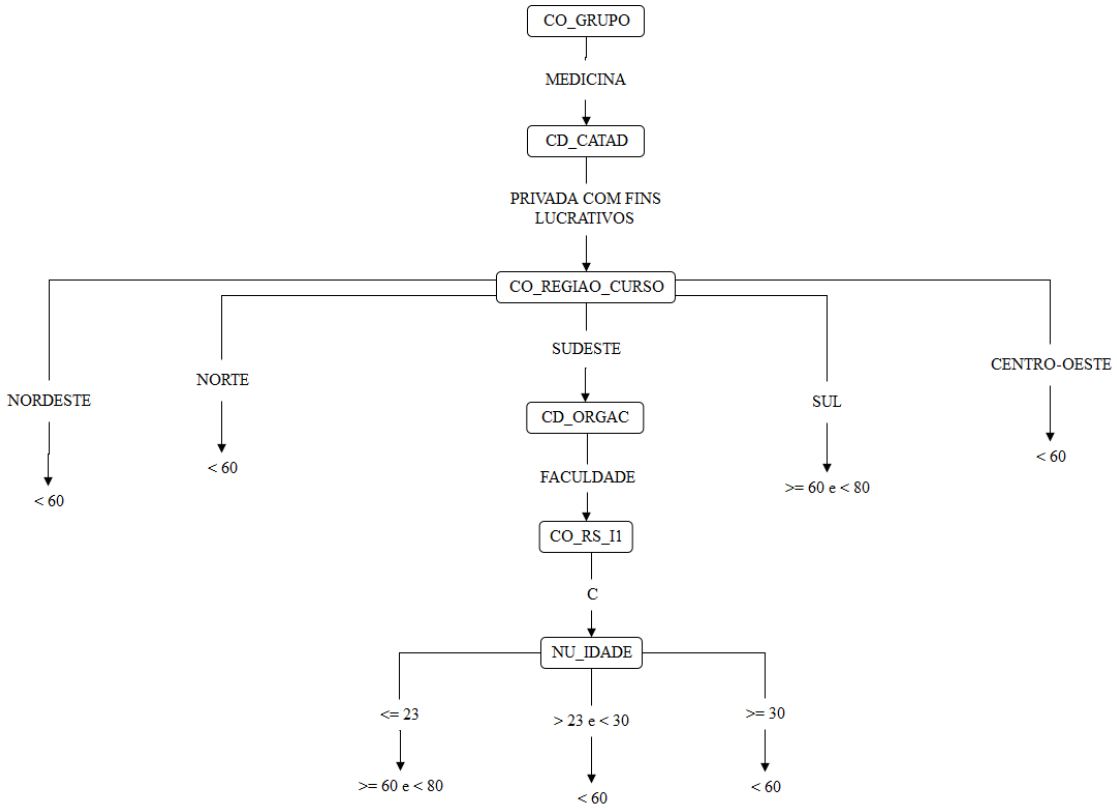

Figura 10. Arvore de decisão referente às IES Privadas com fins lucrativos da 1º Base.
Fonte: Elaborado pelo Autor (2016).

Os estudantes do curso de medicina das IES Privadas com fins lucrativos tem seu desempenho variando principalmente de acordo com as regiões do país, onde quando estes são da região nordeste, norte e centro-oeste, seu rendimento na parte de formação geral da prova são menor que sessenta, mas quando do sul, o resultado nesta parte é maior ou igual a sessenta e menor que oitenta.

Estes estudantes, caso sejam do sudeste e oriundos de faculdades, marcam como médio o nível de dificuldade desta parte da prova e, de acordo com a idade, possuem notas diferentes. Quando, com idade menor ou igual a vinte e três anos, tiraram nota maior ou igual a sessenta e menor que oitenta e, quando com idade entre vinte e três anos e trinta anos ou com idade maior ou igual a trinta anos, obtiveram nota menor que sessenta.

Ao observar os resultados pertinentes à 1º Base minerada, é possível notar a formação do perfil dos estudantes que realizaram o exame, relacionados com suas notas na parte de formação geral e sua resposta sobre o grau de dificuldade deste

elemento.

Os padrões gerados através das arvores de decisão mostram que as notas nesta parte da prova variaram entre menores que sessenta e maiores ou iguais a sessenta e menores que oitenta. Foi revelado ainda que os alunos, em sua maioria, consideraram a parte de formação geral da prova como fácil ou de média dificuldade.

A 2º Base, que tem seus principais resultados retratados nas arvores de decisão posteriores, possui, como meta, relacionar o perfil dos estudantes que prestaram o exame, com sua nota no componente específico da prova e com a resposta da segunda pergunta do questionário de percepção do exame, que procura descobrir qual o nível de dificuldade desta prova, na parte do componente específico.

Nesta base, foram encontrados resultados minuciosos sobre três cursos, o curso de agronomia, de fisioterapia e de fonoaudiologia. Assim como acontecido na 1º Base, os cursos não citados, que podem ser encontrados no Quadro 3, obtiveram um grande número de valores menores que sessenta, como nota do componente específico da prova, o que resultou, no caso destes cursos, em uma mineração pouco detalhada.

A Figura 11 revela os resultados de maior relevância do curso de agronomia, em IES Federais, presentes na 2º Base.

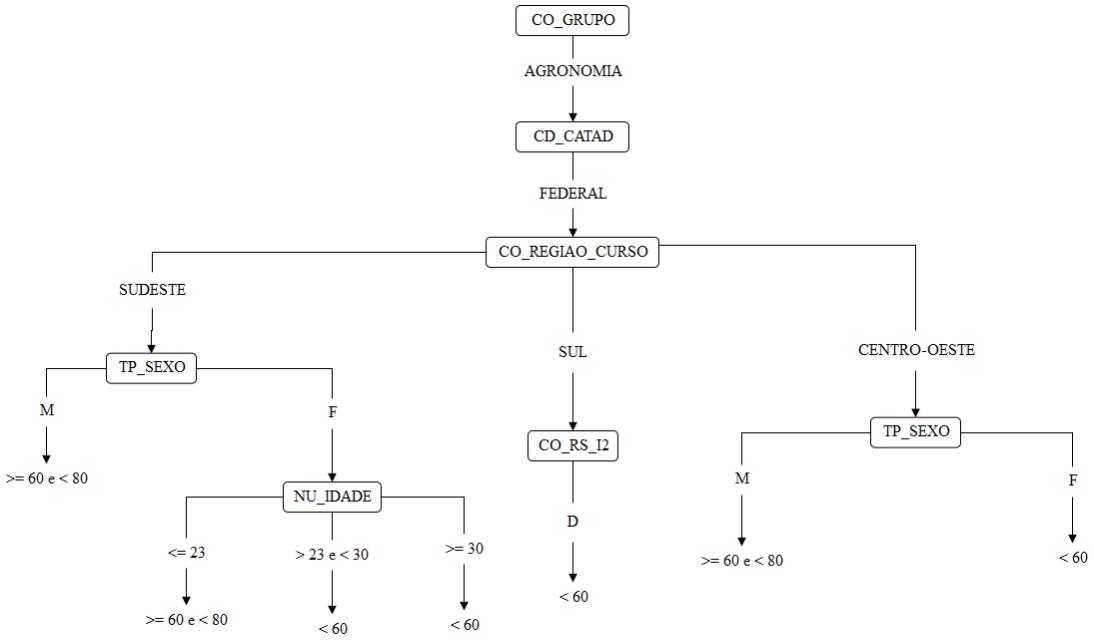

Figura 11. Arvore de decisão referente ao curso de Agronomia nas IES Federais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

Foi apresentada na Figura 11 a árvore de decisão gerada a partir das principais regras derivadas das IES Federais. Tal árvore possibilita uma visualização mais intuitiva dos padrões encontrados, facilitando assim, a obtenção de conhecimento.

Nos cursos de agronomia, das IES Federais, três regiões do país se destacaram nos resultados encontrados, a região sudeste, sul e centro-oeste. No sudeste, os estudantes, quando do sexo masculino, receberam nota maior ou igual a sessenta e menor que oitenta no componente específico da prova, independente dos outros atributos. Quando do sexo feminino, o resultado nesta parte do exame varia com a idade da estudante, sendo quando com idade menor ou igual a vinte e três anos, uma nota maior ou igual a sessenta e menor que oitenta e quando com idade entre vinte e três anos e trinta anos e idade maior ou igual a trinta anos, com resultado menor que sessenta.

No sul, os estudantes responderam como difícil, o grau de dificuldade do componente específico da prova e obtiveram nota menor que sessenta. Já na região centro-oeste, o rendimento dos alunos variou de acordo com o sexo destes. Quando eram do sexo masculino, a nota foi maior ou igual a sessenta e menor que oitenta e,

quando do sexo feminino, o resultado foi menor que sessenta.

Os resultados mais significantes descobertos na 2º Base, sobre o curso de Agronomia das IES Estaduais são apresentados na Figura 12.

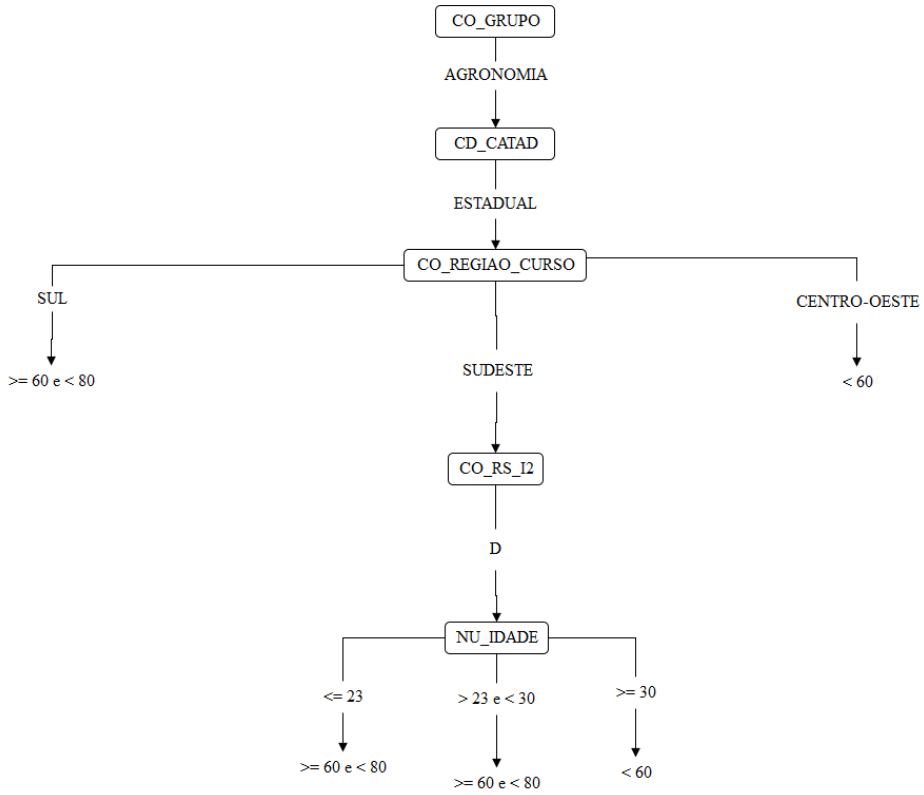

Figura 12. Arvore de decisão referente ao curso de Agronomia nas IES Estaduais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Estaduais, mais especificamente nos seus cursos de agronomia, três regiões se sobressaíram dentre as demais, estas foram às regiões sul, sudeste e centro-oeste. Nas regiões sul e centro-oeste, os estudantes tiveram um rendimento maior ou igual a sessenta e menor que oitenta e menor que sessenta, respectivamente, onde em ambos os resultados, os demais atributos não obtiveram um impacto relevante, capaz de detalhar melhor o perfil destes estudantes.

Na região sudeste, por sua vez, os alunos, em sua maioria, marcou como difícil o nível de dificuldade do componente específico do exame e, dependendo da idade, alcançaram notas diferentes. Quando com idade menor ou igual a vinte e três anos ou com idade entre vinte e três anos e trinta anos, conseguiram um resultado

maior ou igual a sessenta e menor que oitenta, porém, quando com idade maior ou igual a trinta anos, obtiveram nota menor que sessenta no componente específico da prova.

O curso de Fisioterapia, das IES Federais, gerou uma arvore de decisão baseada na 2º Base e mostrada na Figura 13.

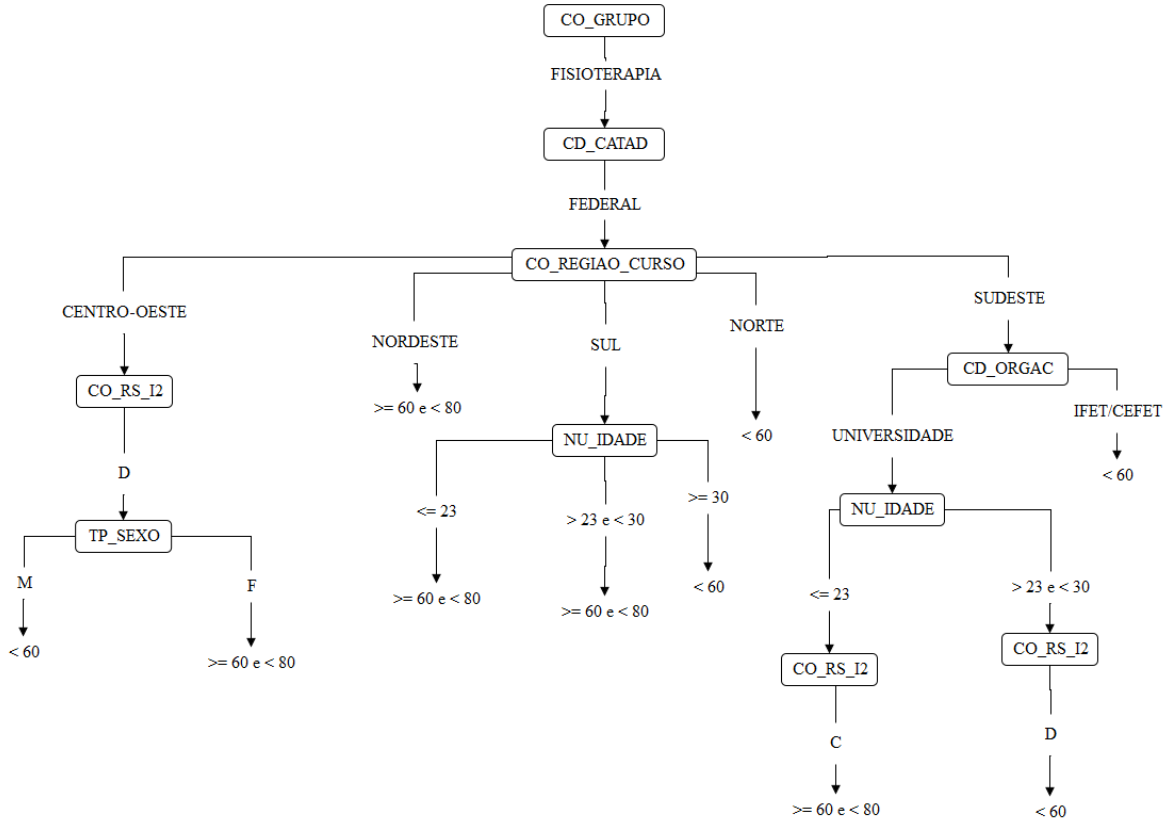

Figura 13. Arvore de decisão referente ao curso de Fisioterapia nas IES Federais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Federais, as notas do componente específico dos estudantes do curso de fisioterapia formaram diferentes padrões de acordo com cada uma das cinco regiões do país. As regiões norte e nordeste obtiveram os resultados mais diretos e com resultados menores que sessenta e entre cinquenta e nove e oitenta, respectivamente. Isto ocorreu, pois, independente dos outros atributos, a nota seria a mesma, uma vez que a maior parte dos estudantes destas regiões obtiveram estes resultados.

Na região centro-oeste, os estudantes responderam como difícil o grau de dificuldade do componente específico do exame e tiraram nota menor que sessenta, quando do sexo masculino e nota maior ou igual a sessenta e menor que oitenta, quando do sexo feminino.

No sul, o rendimento dos estudantes no componente específico da prova variou de acordo com a idade destes, onde, quando com idade menor ou igual a vinte e três anos e com idade entre vinte e três e trinta anos, os alunos receberam nota maior ou igual a sessenta e menor que oitenta, mas quando com idade maior que trinta anos, estes obtiveram um resultado menor que sessenta neste componente.

O sudeste obteve resultados mais detalhados, mostrando que, quando os estudantes são oriundos de universidades, estes possuem notas e opiniões diferentes sobre o componente específico da prova. Estes alunos, quando com idade menor ou igual a vinte e três anos, receberam nota maior ou igual a sessenta e menor que oitenta e responderam como médio o grau de dificuldade do componente específico. Quando com idade entre vinte três e trinta anos, tiveram um resultado menor que sessenta e marcaram como difícil o nível de dificuldade da prova. Ainda foi constatado que os estudantes que vieram do Cefet ou Ifet, possuíram um rendimento menor que sessenta no componente específico da prova, independente da análise com outros atributos.

A arvore de decisão relativa aos principais resultados do curso de Fisioterapia, das IES Estaduais, encontrada na 2º Base, é exibido na Figura 14.

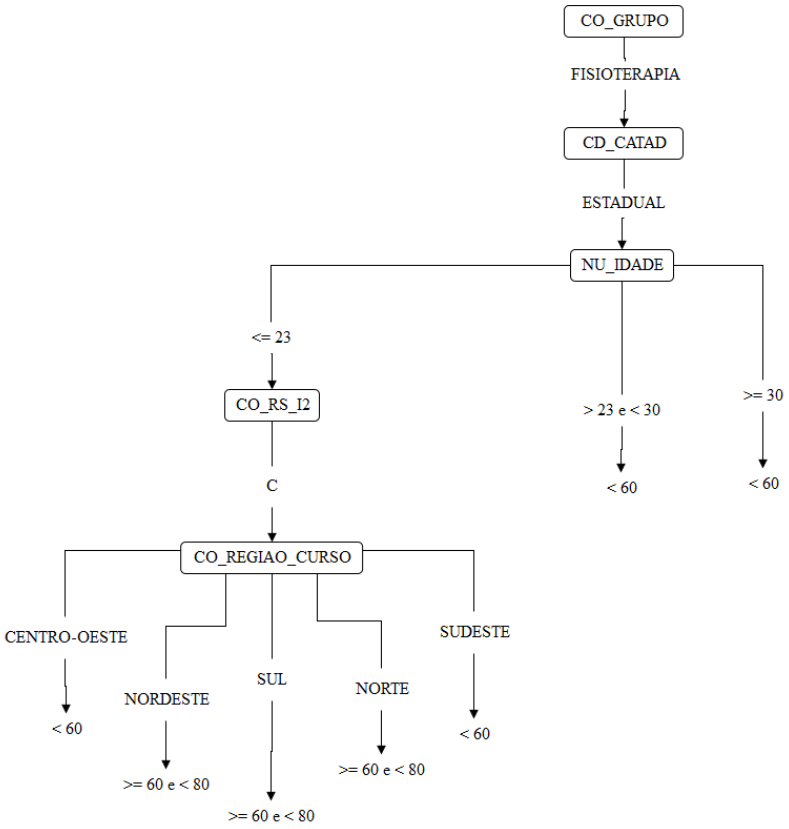

Figura 14. Arvore de decisão referente ao curso de Fisioterapia nas IES Estaduais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

Os estudantes do curso de Fisioterapia, das IES Estaduais, obtiveram resultados no componente específico da prova diferentes de acordo com suas respectivas idades. Estes alunos, quando com idade entre vinte e três e trinta anos e com idade maior ou igual a trinta anos, receberam nota menor ou igual a sessenta, independente da região do país e dos outros atributos analisados.

Quando com idade menor ou igual a vinte e três anos, estes estudantes responderam, em sua maioria, como médio o nível de dificuldade do componente específico do exame e possuíram rendimentos diferentes, neste componente, de acordo com a região. Os oriundos de instituições das regiões norte, sul e nordestes, tiveram um resultado maior ou igual a sessenta e menor que oitenta. Os de instituições do centro-oeste e sudeste receberam nota menor que sessenta.

Na Figura 15, são apresentados os resultados de maior relevância relacionados ao curso de Fonoaudiologia, das IES Federais, presentes na 2º Base.

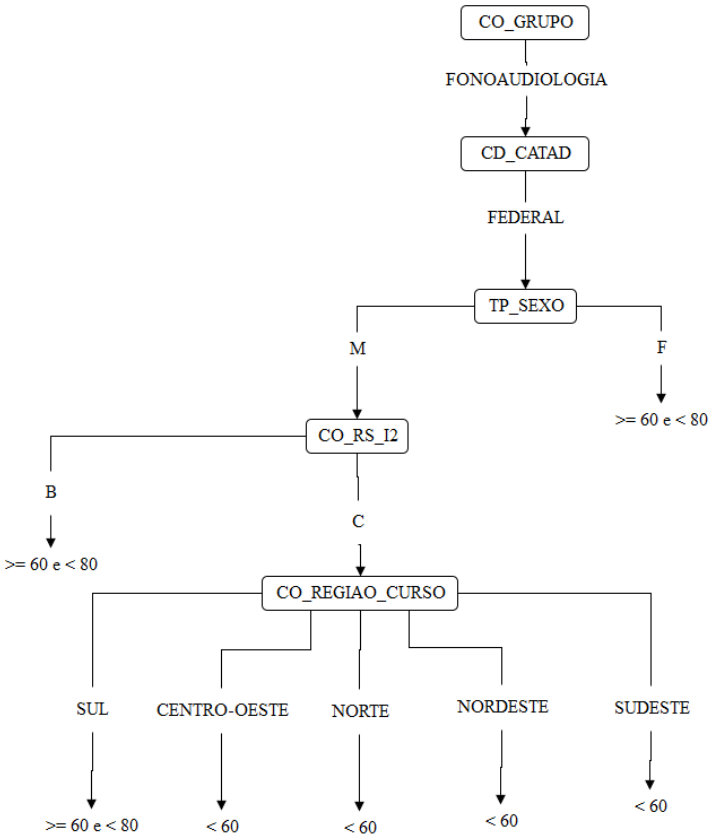

Figura 15. Arvore de decisão referente ao curso de Fonoaudiologia nas IES Federais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

O curso de Fonoaudiologia, das IES Federais, por sua vez, foi primariamente dividido a partir do sexo dos estudantes que prestaram o exame. Sendo que estes, quando do sexo feminino, obtiveram resultado maior ou igual a sessenta e menor que oitenta, independente de qualquer outro atributo analisado em conjunto com estes.

Os estudantes, quando do sexo masculino, responderam, em sua maioria, como fácil e médio, o grau de dificuldade do componente específico da prova, sendo que, os alunos que responderam como fácil, alcançaram uma nota maior ou igual a sessenta e menor que oitenta, porém, os estudantes que marcaram como médio, a dificuldade deste componente, tiveram seus rendimentos neste componente de acordo com a região da instituição que frequentam. Quando são da região sul, o resultado foi maior ou igual a sessenta e menor que oitenta, quando das demais regiões do país, a nota foi menor que sessenta.

Os resultados, pertinentes a 2º Base, mais especificamente do curso de Fonoaudiologia das IES Estaduais, são apresentados na Figura 16.

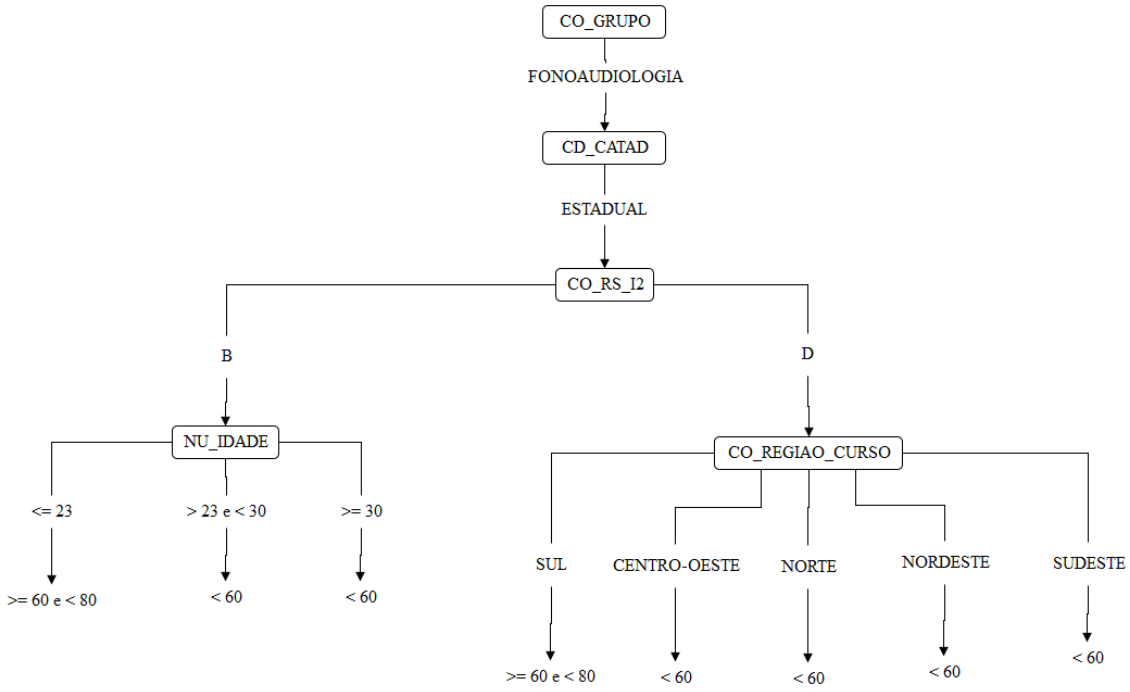

Figura 16. Arvore de decisão referente ao curso de Fonoaudiologia nas IES Estaduais da 2º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Estaduais, os estudantes dos cursos de Fonoaudiologia, tiveram duas respostas, sobre o grau de dificuldade do componente específico do exame, que se destacaram nos resultados encontrados, estas foram: fácil e difícil. Os alunos que marcaram como fácil, nesta parte da prova, tiveram um rendimento diferenciado de acordo com sua idade, onde com idade menor ou igual a vinte e três anos, possuíram, em sua maioria, uma nota maior ou igual a sessenta e menor que oitenta. Quando estes estudantes tem idade entre vinte e três e trinta anos e idade maior que trinta anos, seus resultados foram menor que sessenta no componente específico.

Porém, quando estes alunos responderam como médio o nível de dificuldade deste componente, suas notas variavam, dependendo da região do país em que as instituições, na qual frequentam, se encontram. Os estudantes que frequentam instituições do sul, obtiveram resultado maior ou igual a sessenta e menor que oitenta e, os que estudam nas outras regiões, receberam nota menor que sessenta

no componente específico da prova. Tais resultados foram tão significativos, que mesmo analisados com os outros atributos, não foram afetados.

A visualização dos padrões formador a partir da análise da 2º Base minerada é capaz de demostrar o perfil dos alunos que prestaram o exame e sua relação com nota do componente específico da prova e sua resposta sobre o nível de dificuldade deste mesmo componente.

O resultado analisado e apresentado nas árvores de decisão ainda revela uma certa predominância das notas, relativas ao componente específico da prova, menores que sessenta, porém, nos cursos apontados nestas arvores, também podem ser vistos resultados maiores ou iguais a sessenta e menores que oitenta, dependendo do perfil dos alunos. Podem ser observadas também as respostas sobre grau de dificuldade desta parte do exame, que, na maioria dos casos, variou entre fácil, média e difícil.

A análise e seleção dos resultados concebidos pela mineração de dados da 3º Base foi apresentada em formato de arvores de decisão. Nesta base, o objetivo era identificar o perfil dos estudantes que fizeram o exame, relacionando-os com suas respectivas notas gerais na prova e com a opinião sobre a sétima pergunta do questionário de percepção da prova, que busca saber qual foi a maior dificuldade encontrada, ao realizar a prova.

A partir da análise dos resultados, foram descobertos padrões interessantes sobre cinco cursos, agronomia, enfermagem, fisioterapia, medicina e tecnologia em gestão hospitalar. Os demais cursos foram capazes de gerar a informação de que possuíram um grande número de notas gerais do exame menores que sessenta fator que afetou diretamente seus resultados na mineração e impossibilitou o descobrimento de mais detalhes.

A Figura 17 apresenta os padrões encontrados a partir da 3º Base, mais especificamente no curso de Agronomia das IES Federais.

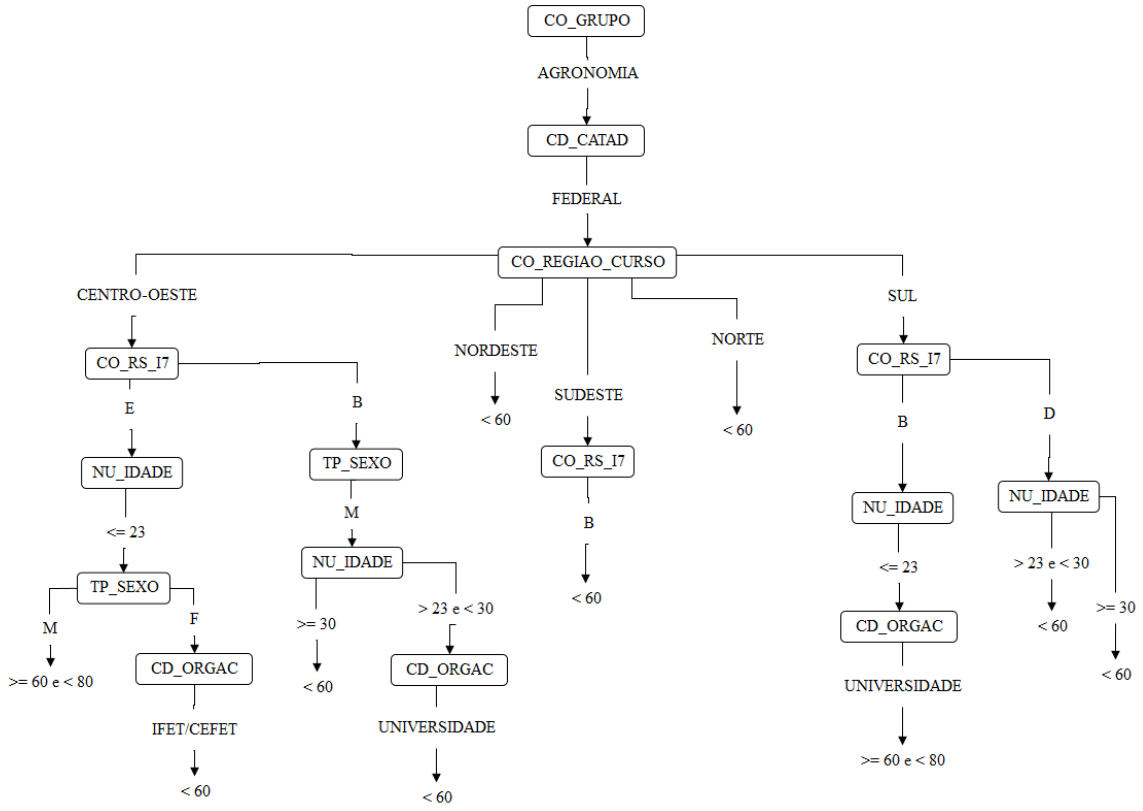

Figura 17. Arvore de decisão referente ao curso de Agronomia nas IES Federais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

Na 3º Base, os estudantes do curso de Agronomia das IES Federais, tiveram seus resultados divididos primeiramente entre todas as regiões do país. Os que frequentam instituições nas regiões norte e nordeste, obtiveram resultados menores que sessenta no exame, independente das relações com outros atributos. Os alunos da região sudeste também receberam um resultado menor que sessenta, mas foi encontrado que a maioria destes respondeu que o conteúdo, presente na prova, foi abordado de forma diferente, na pergunta sobre qual foi a maior dificuldade encontrada ao realizar o exame.

Na região centro-oeste, os alunos foram separados em dois grupos, os que responderam que o conteúdo foi abordado de forma diferente e os que disseram que não tiveram nenhuma dificuldade ao realizar a prova, na questão que abordava qual a maior dificuldade encontrada ao fazer o exame. Os estudantes que afirmaram não ter tido dificuldade tinham, em sua grande maioria, idade menor ou igual a vinte e três anos e, quando do sexo masculino, obtiveram nota maior ou igual a sessenta e menor que oitenta, mas, quando do sexo feminino e frequentando Ifets ou Cefets,

receberam nota menor que sessenta na prova.

Os estudantes da região centro-oeste que responderam que o conteúdo foi abordado de forma diferente eram, principalmente, do sexo masculino e quando com idade maior ou igual a trinta anos ou entre vinte e três e trinta anos, cursando numa universidade, obtiveram resultado menor que sessenta.

Já na região sul do país, os alunos foram divididos de acordo com sua idade e resposta sobre a maior dificuldade que encontraram na prova, onde quando com idade entre vinte e três anos e trinta anos e maior ou igual a trinta anos, estes afirmaram ter tido falta de motivação para fazer a prova e receberam nota menor que sessenta. Os estudantes com idade menor ou igual a vinte e três anos, disseram ter aprendido o conteúdo de forma diferente, eram, em sua maioria, oriundos de universidades e tiraram nota maior ou igual a sessenta e menor que oitenta no exame.

A árvore de decisão contendo os principais resultados sobre os estudantes do curso de Agronomia das IES Estaduais, provenientes da 3º Base, podem ser observados na Figura 18.

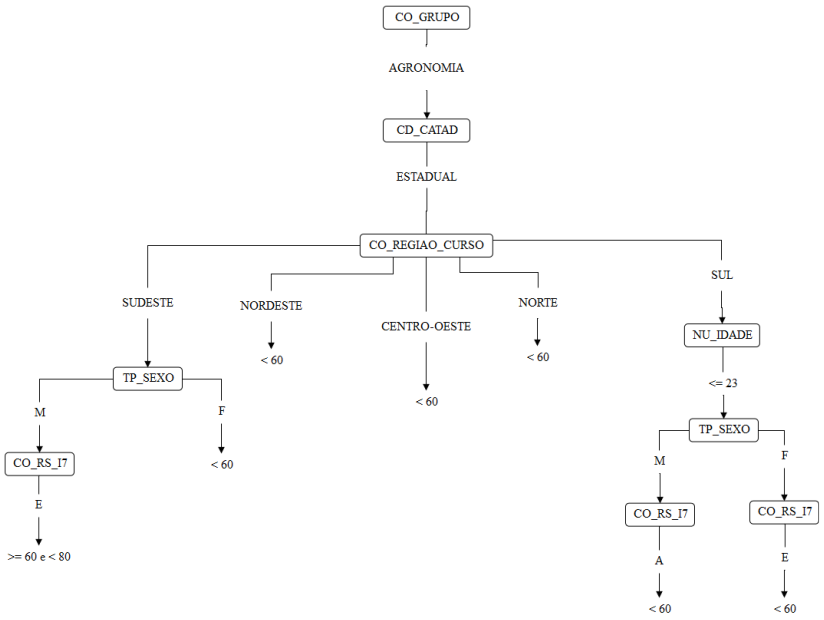

Figura 18. Arvore de decisão referente ao curso de Agronomia nas IES Estaduais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Estaduais, os estudantes do curso de agronomia apresentaram resultados variados de acordo com a região do país em que a instituição que frequentam se encontra. As regiões norte, nordeste e centro-oeste tiveram rendimento na nota geral do exame, sendo ele menor que sessenta. Este resultado foi tão extenso que impediu que maiores detalhes do perfil fossem encontrados, em relação com os outros atributos.

Na região sudeste, os alunos tiveram um rendimento diferente, na nota geral da prova, de acordo com o sexo destes, onde, quando do sexo feminino, estes obtiveram um resultado menor que sessenta, porém, quando do sexo masculino, conquistaram uma nota maior ou igual a sessenta e menor que oitenta e afirmaram não ter tido qualquer dificuldade na realização do exame.

Os estudantes referentes a região sul do país tinham, principalmente, idade menor ou igual a vinte e três anos e, quando do sexo masculino, afirmaram não possuir conhecimento sobre o conteúdo da prova e receberam nota menor que sessenta. As estudantes do sexo feminino, disseram não ter tido nenhuma dificuldade para a realização da prova e também obtiveram um resultado menor que sessenta.

Para o curso de Enfermagem, das IES Federais, presentes na 3º Base, foi gerada uma arvore de decisão baseada nos melhores resultados descobertos. Esta arvore pode ser vista na Figura 19.

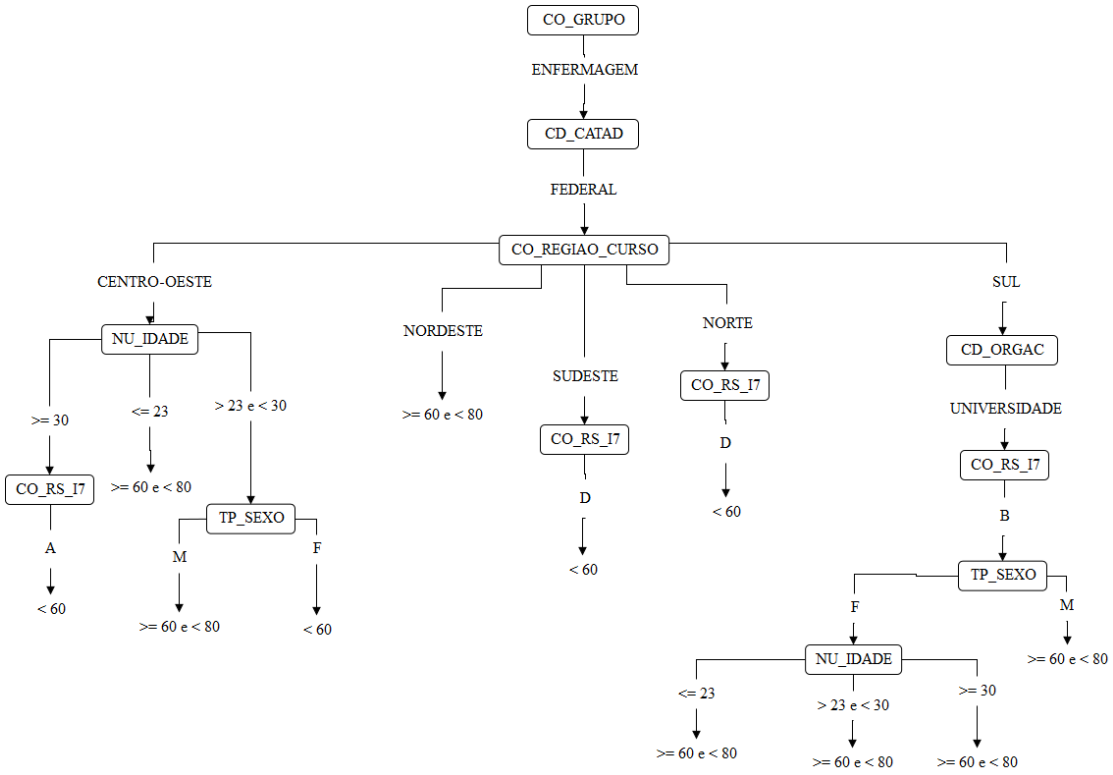

Figura 19. Arvore de decisão referente ao curso de Enfermagem nas IES Federais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

Como pode ser observado na figura 19, no curso de enfermagem, das IES Federais teve o resultado geral dos estudantes separado entre as cinco regiões do país. No nordeste, os alunos receberam, principalmente, uma nota maior ou igual a sessenta e menor que oitenta, sem ter precisado levar em consideração os demais atributos. Nas regiões norte e sudeste, os estudantes obtiveram resultados e opiniões semelhantes, onde, em ambos os casos, os alunos alcançaram um resultado menor que sessenta e responderam ter sofrido com a falta de motivação para a realização do exame, na sétima pergunta do questionário de percepção da prova que busca descobrir qual a maior dificuldade encontrada na realização da prova.

Os estudantes que estudam em instituições pertencentes à região centro-oeste do país tiveram notas diferentes de acordo com sua idade. Quando com idade menor ou igual a vinte e três anos, receberam nota geral no exame maior ou igual a sessenta e menor que oitenta, quando com idade maior ou igual a trinta anos, afirmaram, principalmente, não possuir conhecimento sobre o conteúdo cobrado e obtiveram nota menor que sessenta. Os alunos com idade entre vinte e três e trinta

anos ainda foram separados de acordo com o sexo, onde, quando do sexo masculino, tiraram um resultado maior ou igual a sessenta e menor que oitenta e, quando do sexo feminino, conquistaram um resultado menor que sessenta.

Na região sul, os estudantes frequentavam, em sua maioria, universidades e responderam ter aprendido o conteúdo de forma diferente da cobrada no exame. Além disso, estes alunos, independente do sexo e da idade, alcançaram uma nota geral no exame maior ou igual a sessenta e menor que oitenta.

Ainda sobre os resultados do curso de Enfermagem, quando de IES Estaduais, baseado na 3º Base, foi encontrada a arvore de decisão apresentada na Figura 20.

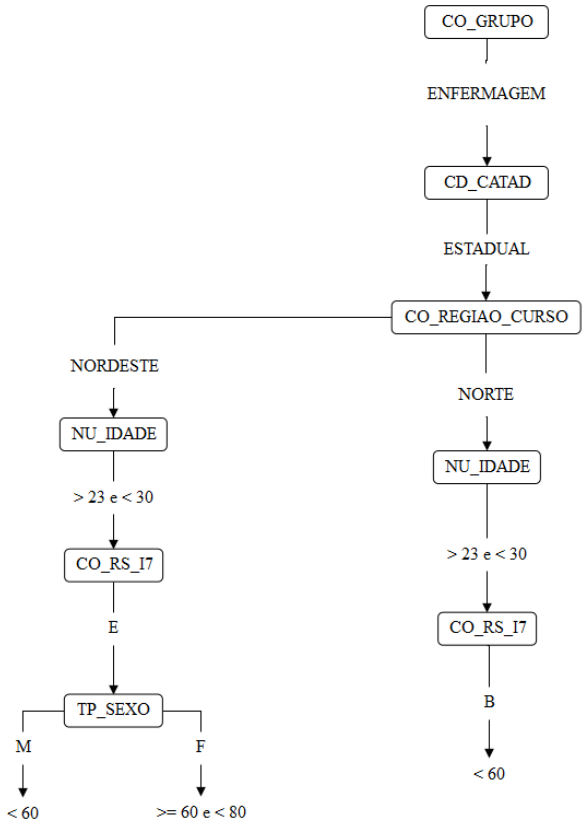

Figura 20. Arvore de decisão referente ao curso de Enfermagem nas IES Estaduais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Estaduais, o resultado referente ao rendimento da nota geral do exame dos estudantes do curso de Enfermagem obteve destaque em duas regiões

do país, norte e nordeste, conforme é mostrado na Figura 29. Quando da região norte, os alunos com idade maior que vinte e três anos e menor que trinta anos, receberam uma nota menor que sessenta e afirmaram que o conhecimento cobrado na prova foi abordado de uma forma diferente da aprendida por estes.

Quando os estudantes frequentavam instituições da região nordeste do país, estes também possuíam, em sua maioria, idade entre vinte e três e trinta anos e responderam não ter tido qualquer tipo de dificuldade ao fazer a prova. Estes alunos ainda obtiveram resultados diferentes relacionados aos seus respectivos sexos. Se eram do sexo feminino, tiveram uma nota maior ou igual a sessenta e menor que oitenta, se eram do sexo masculino, alcançaram uma nota geral no exame menor que sessenta.

Os resultados de maior relevância sobre o curso de Fisioterapia, nas IES Federais, da 3º Base, são mostrados em formato de arvore de decisão na Figura 21.

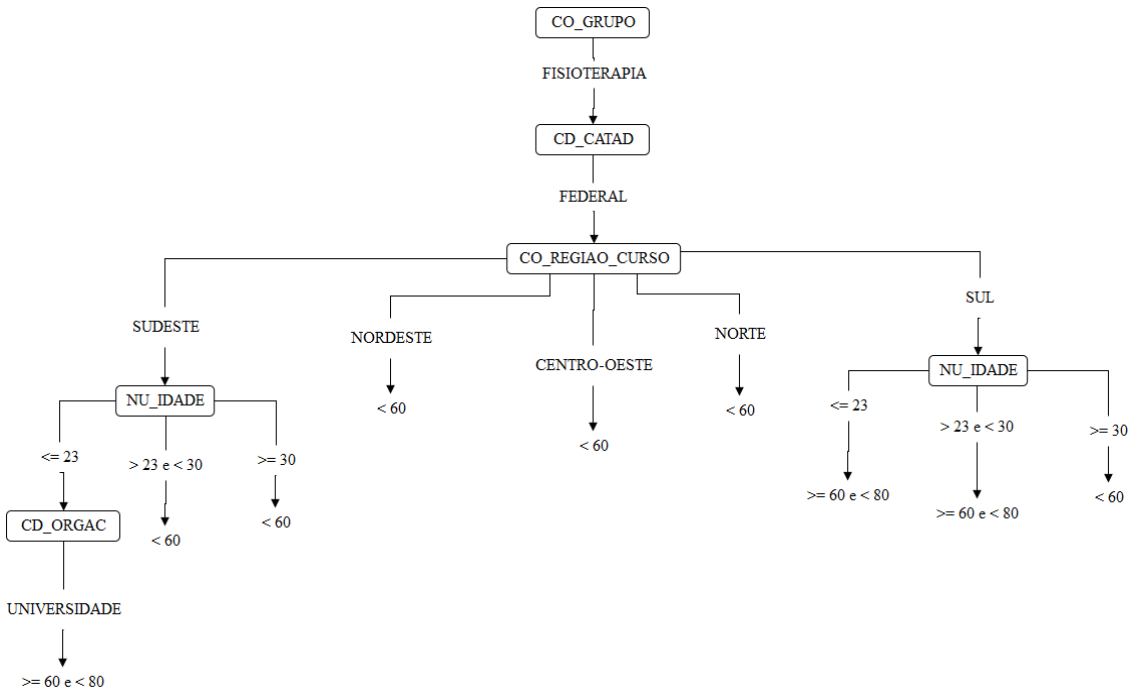

Figura 21. Arvore de decisão referente ao curso de Fisioterapia nas IES Federais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

No curso de Fisioterapia, das IES Federais, o resultado geral dos estudantes desta área foi separado de acordo com a região do país em que se encontram as

instituições que frequentam. Nas regiões norte, nordeste e centro-oeste, os alunos obtiveram, em sua maioria, notas gerais menores que sessenta, independente dos outros atributos envolvidos na mineração.

Na região sudeste, os estudantes ainda foram divididos de acordo com sua idade, onde quando com idade entre vinte e três anos ou maior ou igual a trinta anos, o resultado no exame foi menor que sessenta, porém, quando com idade menor ou igual a vinte e três anos, estes frequentavam, principalmente, universidades e obtiveram uma nota maior ou igual a sessenta e menor que oitenta.

Os estudantes referentes às instituições do sul do país, também receberam notas diferentes de acordo com sua idade, onde com idade menor ou igual a vinte e três anos ou com idade entre vinte e três e trinta anos, a nota foi maior ou igual a sessenta e menor que oitenta e, quando com idade maior ou igual a trinta anos, receberam um resultado menor que sessenta anos.

A 3º Base, também forneceu padrões interessantes referentes ao curso de Medicina das IES Federais, estes padrões se encontram em formato de arvore de decisão e são apresentados na Figura 22.

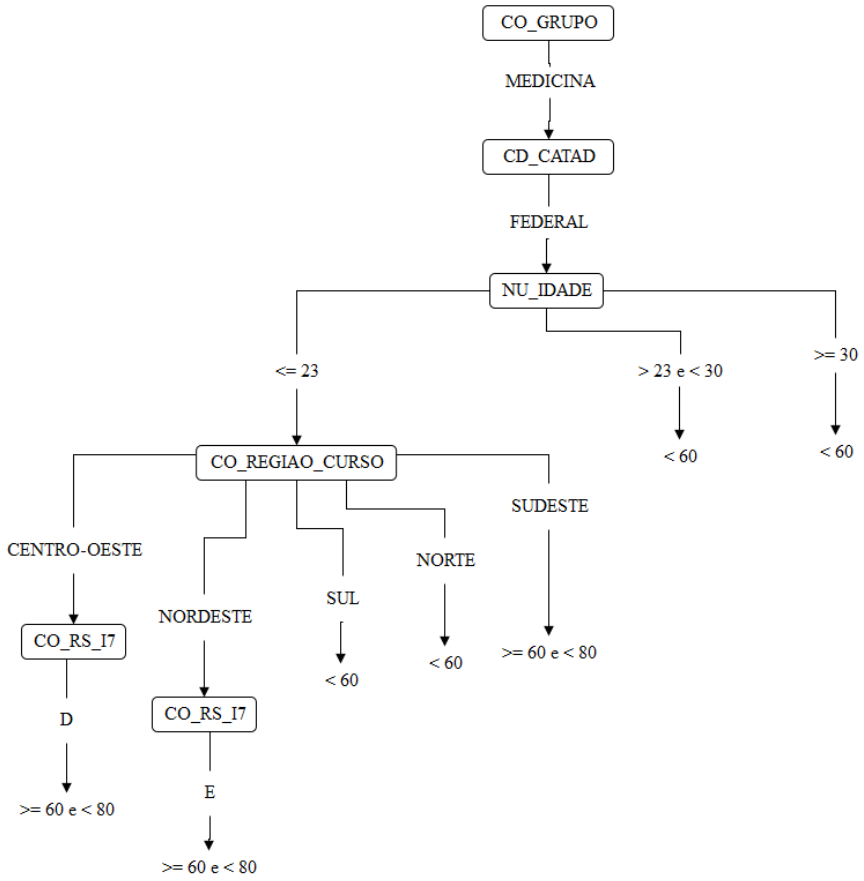

Figura 22. Arvore de decisão referente ao curso de Medicina nas IES Federais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

Como pode ser observado na Figura 22, a arvore de decisão, derivada da 3º base, mostra a influência de certos atributos selecionados na geração do conhecimento.

Os estudantes dos cursos de Medicina, pertencentes à IES Federais, tiveram seus resultados gerais do exame divididos primariamente pela idade dos alunos, fator que se apresentou incomum dentre os resultados descobertos. Estes estudantes, quando com idade entre vinte e três e trinta anos ou com idade maior ou igual a trinta anos, obtiveram nota menor que sessenta, independente dos outros atributos utilizados.

Quando com idade menor ou igual a vinte e três anos, o rendimento destes alunos variou de acordo com a região do país em que se encontra a instituição na qual frequentam. Os referentes as regiões norte e sul, receberam nota menor que sessenta, sendo indiferentes às influencias dos outros atributos. No sudeste, os

estudantes conquistaram, em sua maioria, uma nota maior ou igual a sessenta e menor que oitenta.

Os estudantes desta mesma idade, mas da região centro-oeste, disseram não ter tido motivação para a realização da prova, e receberam uma nota maior ou igual a sessenta e menor que oitenta. Os alunos da região nordeste, por sua vez, afirmaram não ter tido nenhum tipo de dificuldade ao responder a prova e conquistaram uma nota geral no exame maior ou igual a sessenta e menor que oitenta.

Na Figura 23, são mostrados os principais resultados do curso de tecnologia em gestão hospitalar, das IES Federais, encontrados na 3º Base.

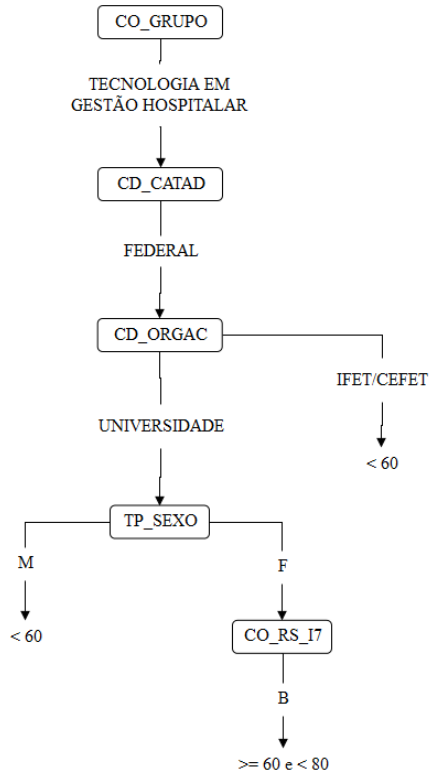

Figura 23. Arvore de decisão referente ao curso de Tecnologia em gestão hospitalar nas IES Federais da 3º Base.
Fonte: Elaborado pelo Autor (2016).

No curso de tecnologia em gestão hospitalar, pertencente às IES Federais, o rendimento dos estudantes, referente à nota geral no exame, dependeu principalmente do estilo da organização institucional, onde dois destes estilos se

destacaram nos resultados descobertos. Quando os alunos eram oriundos de Ifets ou Cefets, seus resultados gerais na prova eram menores que sessenta independente dos demais atributos.

Quando os estudantes frequentaram universidade, seus resultados se dividiram de acordo com seus respectivos sexos. Caso fossem do sexo masculino, receberam nota menor que sessenta, caso do sexo feminino, obtiveram uma nota maior ou igual a sessenta e menor que oitenta e responderam ter aprendido o conteúdo de uma forma diferente da qual foi cobrada na prova.

A 3º Base, após ser minerada e analisada, apresentou o perfil dos estudantes que prestaram o exame, com suas respectivas notas gerais na prova e suas respostas sobre qual foi a maior dificuldade encontrada durante a realização deste exame.

As arvores de decisão, geradas através da mineração de dados e análise dos resultados da 3º Base, demostraram uma grande influência, por parte das regiões do país, nas notas gerais dos estudantes, que obtiveram principalmente notas menores que sessenta e também maiores ou iguais a sessenta e menores que oitenta. Sobre qual foi a maior dificuldade encontrada, durante a realização do exame, todas as respostas foram relevantes, dependendo do perfil dos alunos, porém a resposta referente à forma diferente de abordagem do conteúdo, foi a mais citada.

A 4º Base foi a última base minerada que teve seus dados analisados e extraídos em formato de arvores de decisão. Esta base busca como objetivo encontrar o perfil dos estudantes que realizaram o exame, juntamente com a nota geral da prova, destes estudantes e suas respostas sobre a terceira e nona pergunta do questionário de percepção da prova, que procuram descobrir a opinião do estudante sobre o tempo de prova e quanto tempo este demorou fazendo esta, respectivamente.

Foram encontrados vários padrões úteis, a partir da análise dos resultados desta base. Estes resultados abordaram vários dos cursos examinados pelo ENADE 2013, contendo detalhes relativos aos perfis dos estudantes que prestaram a prova.

Os principais resultados relativos ao curso de Agronomia das IES Federais, presentes na 4º Base, são apresentados em formato de arvore de decisão na Figura 24.

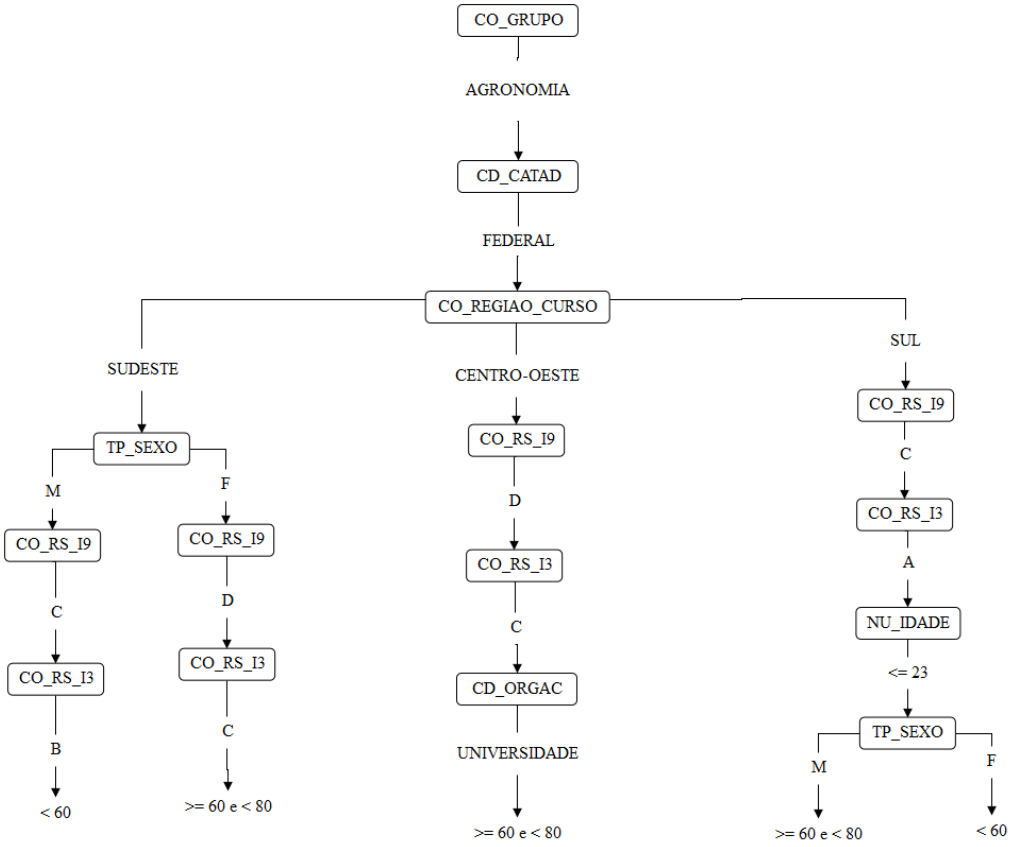

Figura 24. Arvore de decisão referente ao curso de Agronomia nas IES Federais da 4º Base.
Fonte: Elaborado pelo Autor (2016).

Nas IES Federais, os estudantes do curso de Agronomia tiveram seus resultados divididos entre três regiões do país com maior destaque, sudeste, centro-oeste e sul. No sudeste, os alunos do sexo masculino, que obtiveram um resultado geral menor que sessenta, disseram ter demorado entre duas e três horas para realizar a prova e a consideraram longa. As estudantes do sexo feminino tiveram uma nota maior ou igual a sessenta e menor que oitenta, afirmando terem feito a prova em um período entre três e quatro horas e consideraram o tempo de realização desta adequado.

Na região centro-oeste, os estudantes oriundos de universidades alcançaram uma nota geral maior ou igual a sessenta e menor que oitenta, disseram ainda ter

gasto entre três e quatro horas para finalizar a prova e julgaram o intervalo de tempo destinado para a realização do exame adequado.

Os alunos da região sul, por sua vez, tinham, em sua maioria, idade menor ou igual a vinte e três anos. Estes responderam ter terminado a prova entre duas e três horas e ainda declararam achar a prova muito longa, para o período de tempo separado para esta. Com estas opiniões, este padrão de estudantes foi separado ainda em relação ao seu sexo, onde, quando do sexo masculino, receberam uma nota maior ou igual a sessenta e menor que oitenta e, quando do sexo feminino, obtiveram um resultado menor que sessenta.

Na Figura 25, foram apresentados os melhores resultados de vários cursos de IES Federais, da 4º Base.

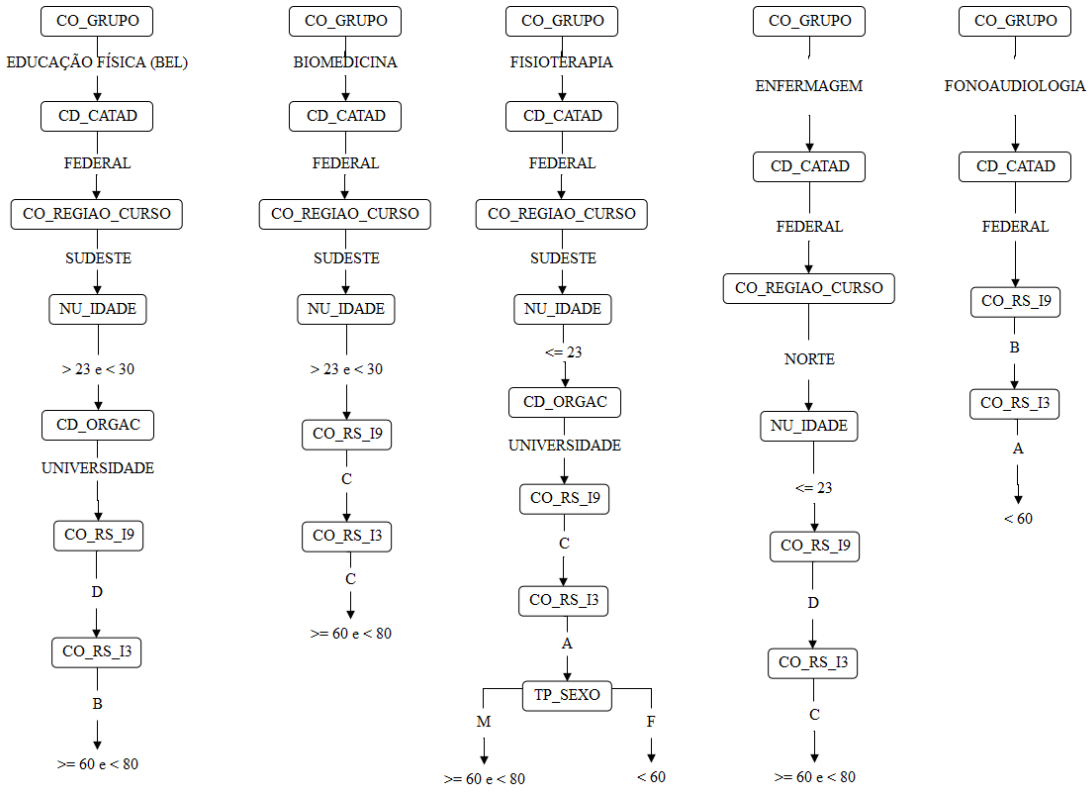

Figura 25. Arvores de decisões referentes a vários cursos encontrados nas IES Federais da 4º Base.
Fonte: Elaborado pelo Autor (2016).

Podem ser observadas na Figura 25, várias arvores de decisão, referentes aos cursos de Bacharelado em Educação Física, Biomedicina, Fisioterapia, Enfermagem e Fonoaudiologia das IES Federais. Cada uma destas arvores forneceu a opinião mais significativa sobre o tempo de realização da prova.

Na arvore relativa ao curso de Bacharelado em educação física, a região sudeste do país se destacou entre as demais, onde os alunos oriundos de universidades, com idade entre vinte e três anos e trinta anos, alcançaram uma nota maior ou igual a sessenta e menor que oitenta, além afirmarem ter feito a prova entre três e quatro horas e de acharem a prova longa, para o período de tempo disposto para a realização desta.

No curso de Biomedicina, também houve um destaque para a região sudeste do país, onde os alunos deste padrão possuíam entre vinte e três e trinta anos e tiveram um resultado maior ou igual a sessenta e menor que oitenta. Estes mesmos alunos ainda afirmaram ter gasto entre duas a três horas para fazer o exame e disseram considerar este com um tempo adequado de realização.

No curso de Fisioterapia, o sudeste do país também obteve resultados mais significativos, onde os estudantes, vindos de universidades, com idade menor ou igual a vinte e três anos disseram ter durado entre duas e três horas para terminar a prova e que esta é muito longa para o tempo determinado de realização. Estes estudantes quando do sexo masculino, receberam uma nota maior ou igual a sessenta e menor que oitenta e, quando do sexo feminino, obtiveram um resultado geral menor que sessenta.

No curso de Enfermagem, os alunos da região norte com idade menor ou igual a vinte e três anos responderam ter feito a prova num período entre duas e três horas e disseram achar o tempo de realização desta prova adequado. Os alunos deste padrão receberam uma nota maior ou igual a sessenta e menor que oitenta.

No curso de Fonoaudiologia, a grande maioria dos alunos, de todas as regiões, idades, sexo e estilos de organização, tiraram uma nota menor que sessenta e declararam ter finalizado a prova entre um período de uma a duas horas

e ainda afirmaram que o exame é muito longo para seu tempo de prova.

Como resultado final, podem ser observados os perfis dos estudantes, formados através de padrões de arvores de decisão, capazes de, em cada uma das quatro bases, analisar uma vertente diferente sobre estes alunos, suas notas e respostas no questionário de percepção da prova.

Na 1º Base, foram encontrados os perfis de estudantes que tiraram notas ruins, menor que sessenta e regulares, maior ou igual a sessenta e menor que oitenta, na parte da prova referente à formação geral, que analisa a capacidade dos alunos de interpretação de texto, análise crítica, ética, dentre outras. Estas informações ainda foram relacionadas com a resposta destes mesmos estudantes sobre o nível de dificuldade deste elemento da prova. Em alguns dos padrões descobertos, foi possível ainda observar contradições por partes dos estudantes, que, apesar de terem obtido um resultado negativo nesta parte do exame, afirmaram considera-la de um nível fácil de dificuldade. Este resultado sugere que estes alunos podem possuir um conhecimento ainda menor sobre o assunto cobrado.

Uma das arvores de decisão que possui tal padrão é a do curso de Medicina das IES Municipais, Figura 8, que demostra que, uma grande quantidade dos estudantes do sexo masculino vindos de universidades da região sul do país, receberam nota inferior a sessenta nesta parte do exame e disseram achar esta fácil.

A 2º Base tratou do componente específico da prova, onde não foi possível encontrar padrões com estudantes que tiveram um bom rendimento neste componente, com nota maior ou igual a oitenta. Ainda que não tenham surgido bons rendimentos, os padrões indicam uma boa quantidade de notas regulares dentre as informações contidas nas arvores de decisão, porém, o número de resultados ruins ainda é ligeiramente maior. Sobre a opinião dos estudantes, a partir deste componente da prova, foram encontradas, principalmente, afirmações em que esta parte é considerada difícil.

Os padrões, desta base, apresentados mostraram que, os estudantes de diferentes partes do país tiveram resultados variados, de acordo com o curso

analisado, o que revela que nenhuma região é ruim em todos os cursos examinados pelo ENADE. Um importante resultado foi sobre o curso de Fisioterapia, onde o desempenho dos estudantes que frequentam Ifets ou Cefets, da região sudeste do país, foram, em sua grande maioria, ruins, independente das outras características analisadas.

A função da 3º Base foi averiguar quais as maiores dificuldades, encontradas pelos alunos, na realização do exame, traçando o perfil destes, com suas notas gerais da prova. Esta extração encontrou padrões interessantes relacionados a quatro das respostas, com exceção da referente ao espaço para desenvolvimento das questões, tendo como maior destaque a resposta referente à forma de abordagem do conteúdo cobrado, que segundo estes, foi diferente de como aprenderam em aula. Este fator é de grande importância para as instituições, que descobrindo este problema, podem analisar a prova e estruturar melhor sua forma de ensino. Outro fator encontrado que pode ser trabalhado pelas instituições é a conscientização da importância do exame, uma vez que uma grande quantidade de alunos afirmou ter falta de motivação para fazer a prova.

Ainda sobre as respostas, foram descobertos perfis de estudantes que afirmam não conhecer sobre o conteúdo cobrado e também outros que disseram não ter tipo nenhuma dificuldade na realização deste exame. Tais resultados são alarmantes, principalmente quando envolvendo os cursos de Medicina e Enfermagem, uma vez que não foram relacionados a uma nota geral de boa, ou seja, maior ou igual a oitenta. Ao declarar que não houve qualquer tipo de dificuldade na realização da prova, estes alunos podem estar equivocados sobre o nível de conhecimento que realmente possuem sobre a área. Estes estudantes, senão aprimorados significativamente, poderão se tornar profissionais de má qualidade e, por serem da área da saúde, um risco para a população.

Um destes importantes resultados é apresentado na Figura 29, onde os estudantes do curso de Enfermagem, em IES Estaduais do nordeste do país, com idade entre vinte e três e trinta anos, tiraram nota menor que sessenta, quando do sexo masculino e maior ou igual a sessenta e menor que oitenta, quando do sexo feminino, afirmando em ambos os casos não ter tido nenhuma dificuldade ao realizar

a prova.

Por fim, na 4º Base foram analisadas as respostas relativas ao tempo gasto para realizar a prova e a opinião dos estudantes sobre a extensão desta com seu tempo designado, que é de quatro horas. Em sua grande maioria, os estudantes disseram ter demorado entre duas e três horas ou entre três e quatro horas, para finalizar a prova. Pode ser observado que, na maioria dos padrões, os estudantes que terminaram a prova com no máximo três horas, afirmaram que a prova era longa ou muito longa para o tempo determinado. Tal conexão torna a opinião destes de pouco valor, uma vez que saíram com no máximo uma hora para o fim do exame. Este fator ainda é comprovado pelos resultados dos estudantes que disseram ter demorado de três até quatro horas, que em sua maioria, acharam o tempo de prova adequado para a extensão desta.

As Figuras 33 e 34 apresentam as arvores que contem os padrões mencionados. Elas ainda mostram que, na maioria dos casos, os estudantes que demandaram mais tempo para a realização da prova obtiveram um melhor resultado geral no exame, com notas maiores ou iguais a sessenta e menores que oitenta.

Com a aplicação das metodologias previamente citadas, foram encontrados padrões úteis, que podem ser utilizados pelas instituições nas suas tomadas de decisões, além de servir como prestação de contas para a sociedade, sobre o nível dos estudantes e seus respectivos cursos que frequentam.

Um ponto negativo envolvendo os resultados foi que, apesar de todos os cursos terem sido utilizados durante a mineração, não foi possível encontrar padrões detalhados e também importantes sobre todos eles. Tal acontecimento porém, não é incomum na extração de conhecimento, uma vez que nem sempre estes padrões existem dentro da base de dados.

A aplicação da mineração de dados e das etapas do KDD não é muito comum nas bases de dado do ENADE, principalmente envolvendo o questionário de percepção da prova.

Alguns trabalhos, porém, podem ser citados, como o de Nogueira e Tsunoda (2015), que analisa a base de dados do ENADE 2012 juntamente com os dados socioeconômicos, buscando descobrir se estes afetam o desempenho dos estudantes. Nele, os autores também utilizaram a tarefa de classificação, através do algoritmo C 4.5, que é o algoritmo em que o J48, do WEKA, se baseia.

Outro trabalho que pode ser citado é o de Cretton, Fontana e Gomes (2015), que, apesar de ser sobre os cursos técnicos do estado do Espírito Santo, também utilizou o KDD e a mineração de dados para descobrir o perfil dos alunos que escolhiam os cursos técnicos. Os resultados foram obtidos através da classificação, feita no WEKA através do algoritmo J48.

7. CONCLUSÃO

A base apresenta dados sobre todos os cursos examinados no Enade de 2013, porém, somente os cursos de medicina, agronomia, fisioterapia, fonoaudiologia, enfermagem, bacharel em educação física, biomedicina e tecnologia em gestão hospitalar apresentaram resultados detalhados e relevantes.

Como resultados de maior importância, pode ser destacada a predominância das notas inferiores a sessenta em todas as partes da prova, seja na formação geral, no componente específico ou na nota total. Este fator mostrou uma grande influência sobre os cursos, impossibilitando inclusive alguns desses de gerar resultados mais detalhados.

Para a 1º Base, pode ser citado o padrão envolvendo os cursos de medicina das IES Estaduais, onde com alunos de idade entre vinte e três e trinta anos, da região sudeste, Quando do sexo feminino, tiraram nota menor que sessenta e disseram considerar a parte de formação geral da prova fácil. Já quando do sexo masculino, os estudantes disseram achar esta parte de um nível médio de dificuldade e, quando vindos de uma universidade, obtiveram nota maior ou igual a sessenta e menor que oitenta, porém, quando oriundos de faculdade, a nota é menor que sessenta.

Na 2º Base, dois padrões se destacaram, o referente ao curso de fisioterapia nas instituições federais da região sudeste que, quando os alunos frequentam um

Ifet ou Cefet, tiraram, principalmente, nota menor que sessenta, mas quando vindos de uma universidade, se separam em dois grupos distintos. O primeiro grupo seria de alunos com idade menor ou igual a vinte e três anos, que receberam nota maior ou igual a sessenta e menor que oitenta, além de dizer que consideram o componente específico da prova de média dificuldade. O segundo grupo com estudantes de idade entre vinte e três e trinta anos, que obtiveram nota menor que sessenta e consideraram esta parte do exame difícil.

O segundo padrão destacado são dos alunos do curso de fonoaudiologia das instituições estaduais, que afirmaram considerar o componente específico da prova como fácil. Estes mesmos estudantes, quando com idade entre vinte e três e trinta anos ou maior ou igual a trinta anos, tiraram uma nota negativa, menor que sessenta, porém, quando a idade destes alunos é menor ou igual a vinte e três anos, estes conseguiram um resultado maior ou igual a sessenta e menor que oitenta. O fato de estudantes considerarem a prova fácil e os mesmos tirarem notas negativas pode ser preocupante, uma vez que sugere um domínio ainda menor sobre o conteúdo cobrado.

Na 3º Base, houve um grande número de padrões em que os estudantes disseram ter aprendido o conteúdo de uma forma diferente e, em outros casos, que não tinham motivação para fazer a prova. Estas informações podem ser utilizadas pelas instituições para que no futuro seus alunos obtenham melhores resultados e se tornem melhores profissionais.

Ainda nesta base, um resultado interessante foi referente ao curso de enfermagem das IES Estaduais do nordeste do país. Nesse curso, os estudantes com idade entre vinte e três anos e trinta anos, responderam não ter tido qualquer problema ao resolver a prova, porém, tal afirmação não condiz com seus resultados, onde, quando estes alunos são do sexo masculino, obtiveram nota menor que sessenta e, quando do sexo feminino, maior ou igual a sessenta e menor que oitenta.

Na 4º Base foram encontrados os últimos resultados da análise, onde foi possível descobrir que, os alunos que demoravam no máximo três horas para

finalizar a prova, tinham como opinião que esta era longa para o tempo atribuído, porém, para os estudantes que gastaram entre três e quatro horas, horário máximo de realização da prova, na maioria dos casos, consideravam o exame com extensão adequada para o tempo determinado. Ainda nesta base foram observados que os padrões com alunos que terminaram a prova com menos de três horas, obtinham, na maioria das vezes, nota menor que sessenta, enquanto os alunos que finalizavam a prova num intervalo de três a quatro horas conquistaram, normalmente, nota maior ou igual a sessenta e menor que oitenta.

Espera-se que, a partir dos padrões e conhecimentos extraídos e apresentados, seja possível auxiliar as instituições nas suas tomadas de decisões, no que se refere as medidas a serem tomadas e melhoria dos projetos de ensino para aprimorar os cursos examinados no ENADE do ano de 2013, objetivando a geração de profissionais devidamente aptos e com um maior nível de conhecimento, tornando-os assim, melhores profissionais. Também é almejado que os futuros estudantes destes cursos possam utilizar estas informações para escolher melhor as instituições na qual irão investir.

Juntamente com o desenvolvimento deste estudo, foi possível publicar um artigo referente ao tema em questão, onde em Cretton e Gomes (2016), foi realizada uma pesquisa na base de dados do ENADE 2013 voltada para o curso de medicina, onde foram descobertos os perfis dos estudantes e relação com uma das perguntas do questionário de percepção da prova. A mineração foi feita através da ferramenta WEKA e aplicando a tarefa de classificação pelo algoritmo J48.

7.1. TRABALHOS FUTUROS

Como trabalhos futuros pretende-se fazer um estudo mais aprofundado, englobando outras bases de dados encontradas no portal do INEP. Estas bases serão referentes aos dados resultantes do ENADE, das áreas já selecionadas, de anos anteriores e posteriores a este trabalho.

As bases então selecionadas serão analisadas e processadas tanto individualmente, quanto em conjunto por cada uma das etapas do processo de KDD. Desta forma objetiva-se averiguar o desenvolvimento dos cursos e seus respectivos alunos. A analise dos dados de diferentes anos permite que seja traçado o nível de desempenho dos estudantes com o passar dos anos, descobrindo assim, a trajetória do curso destes alunos.

Além disso, tal nível de análise gera a possibilidade da realização de uma previsão dos resultados dos alunos dos cursos de diferentes tipos de IES, através da mineração de dados.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA JUNIOR, Vicente de Paula. **O processo de formação das politicas de avaliação da educação superior no Brasil (1983-1996)**. 144. f. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade Estadual de Campinas, 2004. Disponível em: <www.bibliotecadigital.unicamp.br/document/?down=vtls000329214>. Acesso em: 14, Set. 2015.

AMARRAL, Fernanda Cristina Naliato do. **Data Mining**: Técnicas e Aplicações para o Marketing direto. São Paulo: Berkeley, Brasil, 2001.

BARREYRO, Gladys Beatriz; ROTHEN, José Carlos. Para uma história da avaliação da educação superior brasileira: análise dos documentos do PARU, CNRES, GERES e PAIUB. **Avaliação (Campinas)**, Sorocaba , v. 13, n. 1, p. 131-152, Mar. 2008 . Disponível em: < http://www.scielo.br/pdf/aval/v13n1/a08v13n1.pdf >. Acesso em: 01 mar. 2016.

BARUQUE, Cássia Blondet; BARUQUE, Lúcia Blondet; MELO, Rubens Nascimento. Using Data Mining for the Refresh of Learning Objects Digital Libraries. In: INTERNATIONAL CONFERENCE ON COMPUTERS OF THE WORLD SCIENTIFIC AND ENGINEERING ACADEMY AND SOCIETY, 10, 2009, 23-15 mar; Prague, Czech Republic. **Proceedings**… , Wisconsin, USA: WSEAS, 2009. Disponível em: <http://www.icee.usm.edu/ICEE/conferences/ICEE2006/papers/3436.pdf>. Acesso em: 25 Aug. 2015.

BITTENCOURT, Hélio Radke; CASARTELLI, Alam de Oliveira; RODRIGUES, Alziro César de Morais. Sobre o índice geral de cursos (IGC). **Avaliação (Campinas)**, Sorocaba, v. 14, n. 3, p. 667-682, nov; 2009 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772009000300008&lng=en&nrm=iso>. Acesso em: 28 jul. 2016.

http://dx.doi.org/10.1590/S1414-40772009000300008.

_____;_____;_____. Uma análise da relação entre os conceitos ENADE e IDD. **Estudos em Avaliação Educacional**, São Paulo, v. 19, n. 40, p. 247-262, 2008. Disponível em: <http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/2078/2035>. Acesso em: 23 Jul. 2016.

BOENTE, Alfredo Nazareno P.; GOLDSCHMIDT, Ronaldo R.; ESTRELA, Vânia Vieira. **Uma metodologia para apoio à realização do processo de descoberta de conhecimento em bases de dados**. Disponível em: <http://boente.eti.br/publica/seget2008kdd.pdf>. Acesso em: 08 abr. 2016.

BOTHOREL, Gwenael; SERRURIER, Mathieu; HURTER, Christophe. Utilisation d'outils de Visual Data Mining pour l'exploration d'un ensemble de règles d'association. In: **Proceedings of the 23rd Conference on l'Interaction Homme-Machine**. ACM, 2011. p. 12. . Disponível em: <http://dl.acm.org/citation.cfm?id=2044369>. Acesso em: 17 Aug. 2015.

BRASIL. MINISTÉRIO DA EDUCAÇÃO. **Comissão Nacional para Reformulação da Educação Superior:** relatório final: uma nova política para a educação superior brasileira. Disponível em: <http://www.dominiopublico.gov.br/download/texto/me002284.pdf>. Acesso em: 19 maio 2016.

_____._____. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira** . Disponível em: <http://portal.inep.gov.br/basica-levantamentos-acessar> Acesso em: 19 jun. 2016.

_____._____.. **Portaria Normativa n. 4 de 05 de agosto de 2008**. Regulamenta a aplicação do conceito preliminar de cursos superiores, para fins dos processos de renovação de reconhecimento respectivos, no âmbito do ciclo avaliativo do Sistema Nacional de Avaliação da Educação Superior (SINAES). Disponível em: <http://download.inep.gov.br/download/superior/condicoesdeensino/Portaria_N_4_de_5_de_agosto_2008.pdf>. Acesso em: 06 set. 2015.

_____._____. **Portaria normativa n. 40, de 12 de dezembro de 2007**; institui o e-MEC, sistema eletrônico de fluxo de trabalho e gerenciamento de informações relativas aos processos de regulação da educação superior no sistema federal de educação, e o Cadastro e-MEC de Instituições e Cursos Superiores e consolida disposições sobre indicadores de qualidade, banco de avaliadores (Basis) e o Exame Nacional de Desempenho de Estudantes (ENADE) e outras disposições. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=16763-port-norm-040-2007-seres&Itemid=30192>. Acesso em: 06 set. 2015.

_____._____.**Programa de avaliação institucional das universidades brasileiras**.. Disponível em: <http://www.dominiopublico.gov.br/download/texto/me002072.pdf>. Acesso em: 19 Mai. 2016.

_____._____. **Relatório do Grupo Executivo para a Reformulação da Educação Superior (GERES)**. Disponível em: <http://www.schwartzman.org.br/simon/pdf/geres.pdf>. Acesso em: 19 maio 2016.

_____._____. **Sistema Nacional de Avaliação da Educação Superior (SINAES):** bases para uma nova proposta de avaliação da educação superior. Disponível em: <http://portal.mec.gov.br/arquivos/pdf/sinaes.pdf>. Acesso em: 24 fev. 2016.

_____._____. **Sistema Nacional de Avaliação da Educação Superior (SINAES):** da concepção à regulamentação. Disponível em: <http://www.publicacoes.inep.gov.br/portal/download/700>. Acesso em: 24 fev. 2016.

_____._____.INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Cálculo do Índice Geral de Cursos**: nota técnica. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2010/Nota_Tecnica_IGC_2010.pdf>. Acesso em: 06 set. 2015.

_____._____._____. **Nota Técnica n 72:** cálculo do conceito preliminar de curso referente a 2013. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2010/Nota_Tecnica_IGC_2010.pdf>. Acesso em: 06 set. 2015.

_____._____._____. **Nota técnica n. 73:** cálculo do índice geral de cursos avaliados da instituição referente a 2013. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2013/nota_tecnica_n_73_2014_calculo_igc_2013.pdf>. Acesso em: 06 set. 2015.

_____. PRESIDÊNCIA DA REPÚBLICA. **Lei n. 9.131, de 24 de novembro de 1995**: altera dispositivos da Lei n. 4.024, de 20 de dezembro de 1961, e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9131.htm>. Acesso em: 06 set. 2015.

_____._____. **Decreto n. 2.026, de 10 de outubro de 1996**: estabelece procedimentos para o processo e avaliação dos cursos e instituições de ensino superior. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/Antigos/D2026.htm>. Acesso em: 15 set. 2015.

_____._____. **Decreto n. 3.860, de 9 de julho de 2001**: dispõe sobre a organização do ensino superior, a avaliação de cursos e instituições, e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto/2001/D3860.htm>. Acesso em: 21 out. 2015.

_____._____. **Decreto n. 5.773, de 9 de maio de 2006**: dispõe sobre o exercício das funções de regulação, supervisão e avaliação de instituições de educação superior e cursos superiores de graduação e sequenciais no sistema federal de ensino. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/2006/decreto-5773-9-maio-2006-542125-norma-pe.html>. Acesso em: 03 out. 2015.

_____._____. **Decreto n. 91.177, de 29 de março de 1985**: institui comissão nacional visando à reformulação da educação superior e dá outras providências. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/1980-1987/decreto-91177-29-marco-1985-441184-norma-pe.html>. Acesso em: 03 out. 2015.

_____._____. **Decreto n. 29.741, de 11 de Julho de 1951**: institui uma comissão para promover a campanha nacional de aperfeiçoamento de pessoal de nível superior. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/1950-1959/decreto-29741-11-julho-1951-336144-norma-pe.html>. Acesso em: 03 out. 2015.

_____._____. **Decreto n. 74.299, de 18 de Julho de 1974**: Dispõe sobre a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e dá outras providências. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/1970-1979/decreto-74299-18-julho-1974-422808-norma-pe.html>. Acesso em: 03 out. 2015.

_____._____. **Decreto n. 86.791, de 28 de dezembro de 1981**: extingue o Conselho Nacional de Pós-Graduação e dá outras providências. Disponível em: <http://www2.camara.leg.br/legin/fed/decret/1970-1979/decreto-74299-18-julho-1974-422808-norma-pe.html>. Acesso em: 03 out. 2015.

_____._____. **Lei n. 9.394, de 20 de dezembro de 1996:** estabelece as diretrizes e bases da educação. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L9394.htm>. Acesso em: 09 set. 2015.

_____._____**.. Lei nº 10.861, de 14 de abril de 2004:** Institui o Sistema Nacional de Avaliação da Educação Superior (SINAES). Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/Lei/L10.861.htm>. Acesso em: 06 nov. 2015.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: **Conceitos,**

tarefas, métodos e ferramentas. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 13 Mai. 2016.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. **Revista de Administração Pública**, Rio de Janeiro, v. 42, n. 3, p. 495-528, jun; 2008. Disponível em: <http://www.scielo.br/pdf/rap/v42n3/a04v42n3.pdf>. Acesso em: 20 jul. 2016.

CARVALHO, Deborah Ribeiro. **Árvore de decisão/algoritmo genético para tratar o problema de pequenos disjuntos em classificação de dados**. Tese (de Doutorado em Computação de Alto Desempenho / Sistemas Computacionais) - Universidade Federal do Rio de Janeiro, RJ, p. 162, 2005. Disponível em: <http://wwwp.coc.ufrj.br/teses/doutorado/inter/2005/Teses/CARVALHO_DR_05_t_D_int.pdf>. Acesso em: 20 jun. 2016.

CARVALHO, Deborah Ribeiro et al . Mineração de dados aplicada à fisioterapia. **Fisioterapia em Movimento,** Curitiba, v. 25, n. 3, p. 595-605, set. 2012. Disponível em: < http://www.scielo.br/pdf/fm/v25n3/15.pdf >. Acesso em: 07 dez. 2015.

CHAPMAN, Pete; et al. **CRISP-DM 1.0 Step-by-step data mining guide**. Disponível em: <http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf>. Acesso em: 20 maio 2016.

CHRISTUDAS, Beulah Christalin Latha et al. Personalization of e-learning using data mining. **International Journal of Learning**, v. 17, n4, p.585-594, 2010. Disponível em: <http://www.icmlc.org/icmlc2011/009_icmlc2011.pdf>. Acesso em: 20 maio 2016.

COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). **História e missão**. Disponível em: <http://www.capes.gov.br/historia-e-missao>. Acesso em: 24 fev. 2016.

COSTA, H, G. Modelo de webibliominig: proposta e caso de aplicação. **Revista FAE**, Curitiba, v. 13, n.1, p. 115-126, jan-jun. 2010. Disponível em: <https://www.researchgate.net/profile/...Costa/.../57155ee308ae1a840264fa4f?>. Acesso em: 20 maio 2016.

CÔRTES, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de dados-Funcionalidades, técnicas e abordagens**. Disponível em: <ftp://139.82.16.194/pub/docs/techreports/02_10_cortes.pdf>. Acesso em: 16 ago.

2016.

_____; GOMES, Georgia Rodrigues. Aplicação de técnicas de mineração de dados na base de dados do ENADE com enfoque nos cursos de medicina. **Acta Biomedica Brasiliensia**, v. 7, n. 1, p. 72-87, jul, 2016. Disponível em: <http://www.actabiomedica.com.br/index.php/acta/article/view/130/111>. Acesso em: 16 ago. 2016.

_____;_____; FONTANA, Valderedo Sedano. Mineração de dados aplicado à identificação do perfil de alunos inscritos em cursos técnicos oferecidos pela sedu es com relação à predição dos cursos. In: ENCONTRO INTERESTADUAL DE ENGENHARIA DE PRODUÇÃO, 1., 2015, São João da Barra. **Anais**... São João da Barra: EINEPRO, 2015. Disponível em: <http://www.fmepro.org/XP/XP-EasyArtigos/Site/XP-ArtigosSessaoShow.php?idevento=18&id=228&min=0>. Acesso em: 16 ago. 2016.

CUNHA, Luiz Antônio. Nova reforma do ensino superior: a lógica reconstruída. **Cadernos de Pesquisa**, São Paulo, n. 101, p. 20-49, jul. 1997. Disponível em: <http://www.fcc.org.br/pesquisa/publicacoes/cp/arquivos/254.pdf>. Acessado em: 01 mar 2016

DIAS SOBRINHO, José. Avaliação e transformações da educação superior brasileira (1995-2009): do provão ao SINAES. **Avaliação (Campinas)**, Sorocaba, v. 15, n. 1, p. 195-224, 2010. Disponível em: <http://www.scielo.br/pdf/aval/v15n1/v15n1a11.pdf>. Acesso em: 20 jan. 2016.

DOTTA, Alexandre Godoy; GABARDO, Emerson. **A qualidade da educação superior no Brasil:** aspectos históricos e regulatórios da política públicas de avaliação. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/114814/2013183%20-%20A%20qualidade%20da%20educa%C3%A7%C3%A3o%20superior%20no%20Brasil.pdf?sequence=1>. Acesso em: 19 maio 2016.

DOUCEK, P.; MARYŠKA, M.; NOVOTNÝ, O. The analysis of university graduates ICT related study programs. [Analýza souladu obsahu ICT studijních oborů s požadavky praxe v České republice] **E a M: Ekonomie a Management**, v.16, n.3, p. 148-161, 2013. Disponível em: < https://otik.uk.zcu.cz/handle/11025/17518>. Acesso em: 19 maio 2016.

DUNHAM, Margaret. H. **Data mining**: introductory and advanced topics. New Jersey: Pearson Education, 2002.

FAYYAD, Usama. PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **Advances in knowledge discovery and data mining**. California: AAAI Press, 1996.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases**. AI magazine**, v. 17, n. 3, p. 37, 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 21 Feb. 2016. http://dx.doi.org/10.1609/aimag.v17i3.1230

FORSATI, Rana; MEYBODI, Mohammad Reza. Effective page recommendation algorithms based on distributed learning automata and weighted association rules. **Expert Systems with Applications**, v. 37, n. 2, p. 1316-1330, 2010. . Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417409005533>. Acesso em: 19 maio 2016.

FONSECA, Stella Oggioni da; NAMEN, Anderson Amendoeira. Mineração em bases de dados do INEP: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. **Educação em Revista**, Belo Horizonte, v. 32, n. 1, p. 133-157, mar. 2016. Disponível em: < http://www.scielo.br/pdf/edur/v32n1/1982-6621-edur-32-01-00133.pdf>. Acesso em: 11 ago. 2016.

FRAUCHES, Celso da Costa. SINAES: avanços e desafios na avaliação da educação superior. **ABMES Cadernos,** Brasília, n. 29, 2014. Disponível em: <http://www.abmes.org.br/public/arquivos/publicacoes/abmes_cadernos_29.pdf>. Acesso em: 19 maio. 2016.

GALVAO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, São Paulo, v. 22, n. 5, p. 686-690, out, 2009. Disponível em: <http://www.scielo.br/pdf/ape/v22n5/14.pdf >. Acesso em: 19 maio. 2016.

GARCÍA, Enrique; ROMERO, Cristóbal; VENTURA, Sebastián; CASTRO, Carlos de. A collaborative educational association rule mining tool. **The Internet and Higher Education**, v. 14, n. 2, 77-88, 2011. Disponível em: <DOI: 10.1016/j.iheduc.2010.07.006>. Acesso em: 12 maio 2016.

GARCÍA-SAIZ, Diego; PALAZUELOS, Camilo; ZORRILLA, Marta. Data mining and social network analysis in the educational field: An application for non-expert users. In: PEÑA-AYALA, Alejandro (ed.) **Educational data mining**. New York: Springer International Publishing, 2014. p. 411-439. Disponível em: <DOI: 10.1007/978-3-319-02738-8-15>. Acesso em: 12 maio 2016.

GILBERT, Karina; SÁNCHEZ, Roberto Ruiz; SANTOS, José Cristobal Riquelme. Minería de Datos: Conceptos y Tendencias. Inteligencia artificial: **Revista Iberoamericana de Inteligencia Artificial**. v 10, n. 29, p. 11-18, 2006. Disponível em: <http://polar.lsi.uned.es/revista/index.php/ia/article/viewFile/479/463>. Acesso em: 06 set. 2015.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining:** um guia prático - conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Campus, 2005.

HAN, Jiawei; HUANG, Yue; CERCONE, Nick; FU, Yongjian. Intelligent query answering by knowledge discovery techniques. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 3, p. 373-390, 1996. Disponível em: <http://academic.csuohio.edu/fuy/Pub/tkde96.pdf>. Acesso em: 12 ago. 2016. http://dx.doi.org/10.1109/69.506706

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning**: data mining, inference and prediction. New York: Springer-Verlag, 2001. p. 371-406 Disponível em: <https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf>. Acesso em: 14 ago. 2016.

HSU, Tien-Yu; KE, Hao-Ren; YANG, Wei-Pang. Knowledge-based mobile learning framework for museums. **Electronic Library**, v. 24, n. 5, 635-648, 2006. Disponível em: <https://scholar.lib.ntnu.edu.tw/en/publications/knowledge-based-mobile-learning-framework-for-museums >. Acesso em: 14 ago. 2016.

KASAHARA, Cristiane Neves; CONCEIÇÃO, Fernando Wilson Sousa. Análise de ferramentas de mineração de dados. **Universidade Federal do Pará**. Belém–PA, 2008. Disponível em: <http://www.miriam.ufpa.br/arquivos/monog_cristiane_fernando.pdf>. Acesso em: 25 nov. 2015.

KALITA, Oksana et al. Supporting and consulting infrastructure for educators during distance learning process: the case of Russian verbs of motion. In: INTERNATIONAL CONFERENCE ON ENGINEERING APPLICATIONS OF NEURAL NETWORKS, 14, 2013, 1-2set; Greece. **Proceeding…..** New York-USA: Springer, 2014. p. 185-192. Disponível em: <DOI: 10.1007/978-3-642-41016-1_20>. Acesso em: 25 Nov. 2015.

KERDPRASOP, Nittaya et al. Knowledge mining in higher education. **International Journal of Mathematical Models and Methods in Applied Sciences**, v. 6, n. 7, p. 861-872, 2012. Disponível em: <http://www.naun.org/main/NAUN/ijmmas/16-432.pdf>. Acesso em: 25 nov. 2015.

KÖCK, Mirjam; PARAMYTHIS, Alexandros. Activity sequence modelling and dynamic clustering for personalized e-learning. **User Modeling and User-Adapted Interaction**, v.21, n.1-2, p.51-97, 2011. Disponível em: <DOI: 10.1007/s11257-010-9087-z>. Acesso em: 25 nov. 2015.

LACERDA, Leo Lynce Valle de. SINAES: teoria e prática: pressupostos

epistemológicos em oposição. **Avaliação (Campinas)**, Sorocaba, v. 20, n. 1, p. 87-104, Mar. 2015 . Disponível em: < http://www.scielo.br/pdf/aval/v20n1/1414-4077-aval-20-01-00087.pdf >. Acesso em: 02 jan. 2016.

LAZCORRETA, Enrique; BOTELLA, Federico; FERNÁNDEZ-CABALLERO, Antonio. Towards personalized recommendation by two-step modified Apriori data mining algorithm. **Expert Systems with Applications**, v. 35, n. 3, p. 1422-1429, 2008. Disponível em: <DOI: 10.1016/j.eswa.2007.08.048>. Acesso em: 12 maio 2016.

LEITE, Denise. **Reformas Universitárias: Avaliação Institucional Participativa**. Petrópolis: Ed. Vozes, 2005. 141 págs. ISBN 85.326.3120-7 Disponível em: <http://www.ufrgs.br/inov/docs/refrmasuniv_avaliainstpartic>. Acesso em: 19 maio. 2016.

LIU, Xiufeng; WHITFORD, Melinda. Opportunities-to-learn at home: Profiles of students with and without reaching science proficiency. **Journal of Science Education and Technology**, v. 20, n. 4, p. 375-387, 2011. Disponível em: <DOI: 10.1007/s10956-010-9259-y>. Acesso em: 12 maio 2016.

MAIA, Luiz Cláudio; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**., Belo Horizonte , v. 15, n. 1, p. 154-172, abr. 2010. Disponível em: <http://www.scielo.br/pdf/pci/v15n1/09.pdf >. Acesso em: 16 ago. 2016.

MAGDIN, Martin; TURCÁNI, Milan. Personalization of student in course management systems on the basis using method of data mining. Turkish online **Journal of Educational Technology (TOJET)**, v. 14, n. 1, p. 58-67, 2015. Disponível em: <http://www.tojet.net/articles/v14i1/1418.pdf>. Acesso em: 16 ago. 2016.

MANFREDI, Silvia Maria. **Educação profissional no Brasil**. São Paulo: Cortez, 2002.

MARTINOVIC, Dragana; RALEVICH, Victor. Privacy issues in educational systems. **International Journal of Internet Technology and Secured Transactions**, v. 1, n. 1-2, p. 132-150, 2007. Disponível em: <DOI: 10.1504/IJITST.2007.014838>. Acesso em: 16 ago. 2016.

MUNK, Michal; DRLÍK, Martin. Analysis of stakeholders' behaviour depending on time in virtual learning environment. **Applied Mathematics and Information Sciences**, v. 8, n. 2, p.773-785, 2014. Disponível em: <DOI: 10.12785/amis/080238>. Acesso em: 16 ago. 2016.

NEVES, Rita de Cássia David das. **Pré-processamento no processo de descoberta de conhecimento em banco de dados**. 2003. 137 f. Diseertação (Mestrado em Ciência da Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, 2003. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf?sequence=> . Acesso em: 20 jun. 2016.

NOGUEIRA, André Magalhães. **Educação superior na assembleia nacional constituinte: agenda de transição e debates na constituinte**. Rio de Janeiro: Observatório Universitário, 2009. Disponível em: <http://www.observatoriouniversitario.org.br/documentos_de_trabalho/documentos_de_trabalho_85.pdf>. Acesso em: 22 maio 2016.

NONAKA, Ikujirō; TAKEUCHI, Hirotaka. **Criação de conhecimento na empresa**. Rio de Janeiro: Elsevier, 2003. Disponível em: <https://books.google.com.br/books?id=FN_LCwX0s-oC&printsec=frontcover&hl=pt-BR#v=onepage&q&f=false >. Acesso em: 14 nov. 2015.

OLIVEIRA, Ana Paula de Matos et al . Políticas de avaliação e regulação da educação superior brasileira: percepções de coordenadores de licenciaturas no Distrito Federal. **Avaliação (Campinas)**, Sorocaba, v. 18, n. 3, p. 629-655, nov, 2013. Disponível em: < http://www.scielo.br/pdf/aval/v18n3/07.pdf >. Acesso em: 21 ago. 2016.

OLIVEIRA, Déborah. **Data Mining ganha espaço na estratégia empresarial**. Disponível em: <http://computerworld.com.br/tecnologia/2012/03/16/data-mining-ganha-espaco-na-estrategia-empresarial>. Acesso em: 23 nov. 2014.

PANG-NING, Tan; STEINBACH, Michel; KUMAR, Vipin. **Introdução ao Data Mining**. Rio de Janeiro: Ciência Moderna, 2009. 900 p.

PEREIRA, Rodrigo da Silva. Trajetória da avaliação da educação superior de 1980-2008. **Cadernos ANPAE,** Vitória, ES, n. 8. 2009. Disponível em: <http://www.anpae.org.br/congressos_antigos/simposio2009/295.pdf >. Acesso em: 21 maio. 2016.

PEÑA-AYALA, Alejandro; CÁRDENAS, Leonor. How educational data mining empowers state policies to reform education: the Mexican case study. In: _____ (ed). **Educational Data Mining**. New York-USA: Springer International, 2014. p. 65-101. Disponível em: <https://link.springer.com/chapter/10.1007%2F978-3-319-02738-8_3#page-1>. Acesso em: 21 maio. 2016.

PINTO, Rodrigo S; MELLO, Simone P. T. de; MELO, Pedro A. Meta-avaliação: uma década do Processo de Avaliação Institucional do SINAES. **Avaliação (Campinas)**, Sorocaba, v. 21, n. 1, p. 89-108, mar. 2016. Disponível em: <http://www.scielo.br/pdf/aval/v21n1/1414-4077-aval-21-01-00089.pdf>. Acesso em: 13 fev. 2016.

PIATETSKY-SHAPIRO, Gregory. Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. **AI magazine**, v. 11, n. 4, p. 68, 1990. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/view/873/791>. Acesso em: 10 dez. 2015.

POLIDORI, Marlis Morosini; MARINHO-ARAUJO, Claisy M.; BARREYRO, Gladys Beatriz. SINAES: perspectivas e desafios na avaliação da educação superior brasileira. **Ensaio: Avaliação em Políticas Públicas em Edicação,** Rio de Janeiro, v. 14, n. 53, p. 425-436, dez; 2006. Disponível em: <http://www.scielo.br/pdf/ensaio/v14n53/a02v1453.pdf>. Acesso em: 14 fev 2016.

PRIMI, Ricardo; HUTZ, Cláudio S; SILVA, Marjorie Cristina Rocha da. A prova do ENADE de psicologia 2006: concepção, construção e análise psicométrica da prova. **Avaliação Psicológica**, Itatiba, v. 10, n. 3, dez. 2011. Disponível em <http://pepsic.bvsalud.org/pdf/avp/v10n3/v10n3a04.pdf >. Acesso em 07 mar. 2016.

RADENKOVIĆ, Božidar et al. Creating adaptive environment for e-learning courses. **Journal of Information and Organizational Sciences**, v. 33, n. 1, p. 179-189, 2009. Disponível em <https://jios.foi.hr/index.php/jios/article/view/107/72>. Acesso em 07 mar. 2016.

RISTOFF, Dilvo; GIOLO, Jaime. O SINAES como sistema. **Revista Brasileira de Pós-Graduação**, v. 3, n. 6, 2006. Disponível em: <http://ojs.rbpg.capes.gov.br/index.php/rbpg/article/view/106/100>. Acesso em: 11 jun. 2016.

RIGOU, Maria. SIRMAKESSIS, Spiros. Bringing personalization to online learning communities. **WSEAS: Transactions on Information Science and Applications**, v. 12, n. 2, p. 2160-2167, 2005. Disponível em: <http://wseas.org/cms.action?id=7657>. Acesso em: 12 maio 2016.

RIBEIRO, Jorge Luiz Lordêlo de Sales. SINAES: o que aprendemos acerca do modelo adotado para avaliação do ensino superior no Brasil. **Avaliação: Revista da Avaliação da Educação Superior (Campinas),** Sorocaba, v. 20, n. 1, p. 143-161, mar. 2015. Disponível em: < http://www.scielo.br/pdf/aval/v20n1/1414-4077-aval-20-01-00143.pdf >. Acesso em: 08 ago. 2016.

ROTHEN, José Carlos; BARREYRO, Gladys Beatriz. Avaliação da educação superior no segundo governo Lula:" provão II" ou a reedição de velhas práticas?. **Educação e Sociedade**, v. 32, n. 114, p. 21-38, 2011. Disponível em: <http://www.producao.usp.br/bitstream/handle/BDPI/2720/art_BARREYRO_Avaliacao_da_educacao_superior_no_segundo_governo_2011.pdf?sequence=1&isAllowed=y>. Acesso em: 06 mar. 2016.

ROMERO, Cristóbal et al. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. **Computers and Education**, v. 53, n. 3, p. 828-840, 2009. Disponível em: <DOI: 10.1016/j.compedu.2009.05.003>. Acesso em: 06 mar. 2016.

SASSI, Renato José. An hybrid architecture for clusters analysis: rough setstheory and self-organizing map artificial neural network. **Pesquisa Operacional,** Rio de Janeiro, v. 32, n. 1, p. 139-164, abr; 2012. Disponível em: <http://www.scielo.br/pdf/pope/v32n1/aop0512.pdf>. Acesso em: 25 jul. 2016.

SANTOS, Manuel Filipe; AZEVEDO, Carla Sousa. **Data mining:** descoberta de conhecimento em bases de dados. São Paulo: FCA , 2005.

SILVA, Alice Maria Gonçalves. **O desempenho escolar via uma abordagem de descoberta de conhecimento em bases de dados**. 2007. 172 f. Dissertação (Mestrado em Sistemas de Informação) - Universidade do Minho, Braga-Portugal, 2007. Disponível em: <http://repositorium.sdum.uminho.pt/handle/1822/7966>. Acesso em: 12 Abr. 2016.

SILVA, Glauco Barbosa da; COSTA, Helder Gomes. Mapeamento de um núcleo de partida de referências em Data Mining a partir de periódicos publicados no Brasil. **Gestão e Produção**, São Carlos, v. 22, n. 1, p. 107-118, mar. 2015. Disponível em: <http://www.scielo.br/pdf/gp/v22n1/0104-530X-gp-22-01-00107.pdf>. Acesso em: 16 abr 2016.

SILVA, Wender Antônio da. **Contexto do sistema de avaliação da educação superior brasileira**. Disponível em: <http://www.urutagua.uem.br/009/09silva.pdf>. Acesso em: 21 abr. 2016.

SOUSA, José Vieira de. Qualidade na educação superior: lugar e sentido na relação público-privado. **Caderno Cedes**, v. 29, n. 78, p. 242-256, 2009. Disponível em: <http://www.scielo.br/pdf/ccedes/v29n78/v29n78a07.pdf >. Acesso em: 21 abr. 2016.

SPECTOR, J. Michael. Emerging educational technologies and research directions. **Educational Technology and Society**, v. 16, n. 2, p. 21-30, 2013. Disponível em: <www.ifets.info/journals/16_2/3.pdf >. Acesso em: 21 abr. 2016.

STEFANOVIC, Nenad; STEFANOVIC, Dusan; ARSOVIC. Branka. Adaptivity in e-learning LMS platform. **Metalurgia International**, v. 18, n. 3, p. 156-162, 2013. Disponível em: <http://econference.metropolitan.ac.rs/files/pdf/2011/09-branka-arsoovic-adaptivity-in-learning-lms-platform-approaches-and-solutions.pdf>. Acesso em: 12 mar. 2016.

STEINER, Maria Teresinha Arns; et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gestão e Produção,** São Carlos, v. 13, n. 2, p. 325-337, maio 2006. Disponível em: <http://www.scielo.br/pdf/gp/v13n2/31177.pdf >. Acesso em: 09 set. 2015.

SU, Jun-Ming et al. Learning portfolio analysis and mining for SCORM compliant environment. **Educational Technology and Society**, v. 9, n. 1, p. 262-275, 2006. Disponível em: <https://www.jstor.org/stable/jeductechsoci.9.1.262>. Acesso em: 09 set. 2015.

SUMATHI, Sai; SIVANANDAM, S, N. **Introduction to data mining and its applications**. New York-USA: Springer Science and Business Media, 2006. Disponível em: <http://www.csbdu.in/csbdu-old/pdf/Introduction%20to%20Data%20Mining%20and%20its%20Applications.pdf >. Acesso em: 12 abr. 2016.

TETTAMANZI, Andrea; PANNESE, Lucia; SANTALMASI, Mauro. Learner modelling: Optimizing training, assessment and testing. **Journal of E-Learning and Knowledge Society**, v. 5, n. 2, 2009. Disponível em: <http://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/324 >. Acesso em: 12 abr. 2016.

TSURUTA, Setsuo et al. An intelligent system for modeling and supporting academic educational processes. In: PEÑA-AYALA, Alejandro (ed.). **Intelligent and Adaptive Educational-Learning Systems**. Springer, 2013. p. 469-496. Disponível em: <DOI: 10.1007/978-3-642-30171-1_19>. Acesso em: 20 mar. 2016.

VIANNA, Rossana Cristina Xavier Ferreira et al . Mineração de dados e características da mortalidade infantil. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 26, n. 3, p. 535-542, mar. 2010. Disponível em: <http://www.scielosp.org/pdf/csp/v26n3/11.pdf >. Acesso em: 16 ago. 2016.

WEKA. UNIVERSITY OF WAKO. **Data mining software in java**. Disponível em: <http://www.cs.waikato.ac.nz/~ml/weka/>. Acesso em: 20 jun. 2016.

WAN, Xin; JAMALIDING, Qimanguli; OKAMOTO, Toshio. Analyzing learners' relationship to improve the quality of recommender system for group learning support. **Journal of Computers**, v. 6, n. 2, p. 254-262, 2011. Disponível em: <DOI: 10.4304/jcp.6.2.254-262>. Acesso em: 20 jun. 2016.

WANG, Feng-Hsu; SHAO, Hsiu-Mei. Effective personalized recommendation based on time-framed navigation clustering and association mining. **Expert Systems with Applications**, v. 27, n. 3, p. 365-377, 2004. Disponível em: <DOI: 10.1016/j.eswa.2004.05.005/>. Acesso em: 20 jun. 2016.

WANG, Ya-huei; TSENG, Ming-Hseng; LIAO, Hung-Chang. Data mining for adaptive learning sequence in english language instruction. **Expert Systems with Applications**, v. 36, n. 4, p. 7681-7686, 2009. Disponível em: <DOI: 10.1016/j.eswa.2008.09.008>. Acesso em: 20 jun. 2016.

ZAINKO, Maria Amélia Sabbag. Políticas públicas de avaliação da educação superior: conceitos e desafios. **Jornal de políticas educacionais**, v. 2, n. 4, 2008. Disponível em: < http://www.jpe.ufpr.br/n4_2.pdf>. Acesso em: 06 jul. 2016

ZANDAVALLI, Carla Busato. Avaliação da educação superior no Brasil: os antecedentes históricos do SINAES. **Avaliação: Revista da Avaliação da Educação Superior (Campinas),** Sorocaba, v. 14, n. 2, p. 385-438, jul 2009. Disponível em: < http://www.scielo.br/pdf/aval/v14n2/a08v14n2.pdf >. Acesso em: 19 maio 2016.

ZHANG, Ci-zhen; LIU, Hean. Research on data mining technology in education website construction. **International Journal of Digital Content Technology and its Applications**, v. 6, n. 22, p. 222, 2012. Disponível em: <DOI: 10.4156/jdcta.vol6.issue22.24>. Acesso em: 19 maio 2016.

ZHANG, Sonya. An empirical study of the factors affecting weblog success in higher education. **Journal of Information Systems Education**, v. 24, n. 4, p. 267, 2013. Disponível em: <https://eric.ed.gov/?id=EJ1034059>. Acesso em: 19 maio 2016.