ACTA
Biomedica Brasiliensia
ISSN: 2236-0867

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA BASE DE DADOS DO ENADE COM ENFOQUE NOS CURSOS DE MEDICINA

Nícollas Nogueira CRETTON[1]* & Geórgia Regina Rodrigues GOMES[2]

[1]Universidade Candido Mendes. Mestrado em Pesquisa Operacional e Inteligência Computacional, Campos do Goytacazes, Rio de Janeiro, Brasil.
[2]Universidade Federal Fluminense. Departamento de Ciências Exatas, Biológicas e da Terra, Santo Antônio de Pádua, RJ, Brasil.
*Autor para correspondência: nicollas_nc@hotmail.com

http://dx.doi.org/10.18571/acbm.100

RESUMO

O Exame Nacional de Avaliação do Estudante - ENADE tem como função 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. O processo de KDD compreende um conjunto de técnicas capaz de analisar e extrair informações úteis de grandes bases de dados através da criação de padrões. Este artigo tem como objetivo extrair conhecimento da base de dados ENADE do curso de medicina, bem como a resposta sobre o nível de dificuldade do componente específico da prova. Foram utilizadas as etapas do processo de KDD, técnicas de Mineração de Dados e o software WEKA. Após a aplicação das técnicas, foi possível observar a influência da categoria e dos tipos das instituições de ensino superior na criação dos perfis, sendo diretamente ligadas ao nível de desempenho dos estudantes e sua opinião sobre o nível do exame. Foi possível constatar que os estudantes, tanto do Rio de Janeiro, quanto de São Paulo, quando oriundos de universidades sem fins lucrativos, obtiveram em sua maioria, um resultado ruim, com nota menor que sessenta. Ainda em São Paulo, os estudantes de faculdade de instituições municipais, além de receberem um rendimento negativo, também responderam como "fácil" o grau de dificuldade do componente específico do exame. Espera-se que as informações geradas neste trabalho possam ser utilizadas para o aprimoramento dos cursos de medicina, bem como na tomada de decisões referentes aos projetos dos cursos.

Palavras-chave: Mineração de Dados; KDD; WEKA; Medicina; ENADE.

ABSTRACT

Data mining techniques applied in the ENADE database with focus on medicine courses
The National Survey of Student Assessment - ENADE role is to assess the degree of knowledge of students regarding the syllabus provided for in the curriculum guidelines of their respective courses based on their performance on the exam. The KDD process comprises a set of techniques capable of analyzing and extracting useful information from large databases by creating patterns. This article aims to extract knowledge from the medical course of the ENADE database of 2013, and the answer on the level of difficulty of the specific component of the test. The steps were used the KDD process, Data Mining techniques and WEKA software. It was used the KDD process steps, data mining techniques and the WEKA software. After the application of the techniques, it was possible to observe the influence of the category and the types of higher education institutions in the creation of the profiles, being directly linked to the performance level of students and their opinion on the

ISSN: 2236-0867

level of the exam. It was found that students, both from Rio de Janeiro and from São Paulo, when coming from non-profit universities, obtained mainly bad results, with a grade less than sixty. Also in São Paulo, college students of municipal institutions, in addition to receiving a negative result also answered as "easy" the degree of difficulty of the specific component of an examination. It is hoped that the information generated in this work can be used for the improvement of medical courses and in the decisions making regarding the projects of the courses.

Keywords: Data mining; KDD; WEKA; Medicine; ENADE.

1 Introdução

É essencial que existam indicadores para o controle 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 produzirem 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 (PRIMI, 2011).

No Brasil, 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, 2004).

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, baseado nos 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 (MEC, 2010).

Na área da saúde, mais especificamente no curso de medicina, é essencial que este tipo de avaliação seja aplicado, para que possa existir um controle de qualidade das instituições que os possuem, uma vez que este é considerado um curso caro e concorrido.

Neves (2012) diz que, as IES privadas dependem principalmente da cobrança de mensalidades para se sustentar, onde o valor destas mensalidades pode variar drasticamente de acordo com o curso, tipo de instituição (faculdades, universidades e centros universitários) e região. Segundo MEC (2013), o curso de medicina é considerado um dos mais caros do país e também o mais concorrido tanto nas instituições públicas como nas particulares, tendo em média, nas IES públicas duas vagas para cada cem candidatos e, nas IES privadas quatro vagas para cada cem candidatos.

Porém, levando em consideração o nível de concorrência do curso de medicina, os gastos, por parte dos estudantes, começam bem antes da entrada destes no curso. Com o objetivo de se tornarem mais competitivos, os alunos buscam recursos adicionais, como cursos de pré-vestibular, que, da mesma forma que as mensalidades dos cursos superiores, possuem valores bem variados, de acordo com a região, categoria da instituição (Federal, Estadual, Municipal ou privada) e curso pretendido.

ISSN: 2236-0867

Tal gasto acaba sendo muito necessário uma vez que, em cursos mais concorridos, mais de 80% dos estudantes que são aprovados, fizeram cursos de pré-vestibular. Vale ressaltar também que, com o baixo número de vagas, muitos destes alunos acabam não sendo aprovados no primeiro ano, o que normalmente leva estes a fazer o preparatório novamente (BORGES, 2005 e ZAGO, 2006).

A motivação desse trabalho considerou o alto valor de investimento dos estudantes para entrar e se manter no curso de medicina, buscando se tornar profissionais devidamente capacitados e com nível de conhecimento paralelo ao valor aplicado. Para isto, é importante que esta pesquisa aborde não só o desempenho do aluno, mas também seu perfil e opinião sobre o nível de dificuldade do componente específico, uma vez que, quando sua opinião não condiz com seu rendimento, sugere uma possível falta de conhecimento ou confiança por parte do estudante, ambos fatores imprescindíveis para um médico.

Para descobrir o conhecimento necessário para realizar este trabalho, foram utilizadas técnicas de Mineração de Dados, onde esta é uma das etapas mais importantes no processo de busca de conhecimento em bases de dados. Segundo 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.

O objetivo deste trabalho é aplicar técnicas de Mineração de Dados na base de dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), mais especificamente na base do ENADE 2013, utilizando com os dados relativos aos estudantes de medicina, de modo que seja possível ser traçado o perfil destes. Os resultados apresentados neste trabalho, gerados através da mineração, podem ser utilizados na tomada de decisão por parte das instituições, com o objetivo de aprimorar seus cursos e projetos de ensino, além de auxiliar também os futuros estudantes de medicina na hora de escolher sua instituição.

2 Materiais e Métodos

O software utilizado para a realização das tarefas de mineração de dados deste trabalho foi o WEKA 3.7 (Waikato Environment for Knowledge Analysis), que foi desenvolvido na Nova Zelândia, na Universidade de Waikato.

Para Silva (2004), o WEKA é um software intuitivo e com uma interface gráfica amigável que, além de ser gratuito e desenvolvido em Java, o que permite que ele seja utilizado em várias plataformas, também conta com uma grande quantidade de algoritmos, fatores que influenciam no seu alto índice de utilização.

Para que uma base possa ser minerada e ter conhecimentos devidamente extraídos, é necessário que seja feito um tratamento desta base. A base utilizada pode ser encontrada no portal do INEP, onde foi selecionada a base de dados do ENADE 2013, que busca avaliar o desempenho dos estudantes em relação com os conteúdos programáticos previstos nas diretrizes curriculares de seus respectivos cursos de graduação. Nesta base também podem ser encontrados os dados do questionário de percepção da prova e do questionário do estudante.

Com o intuito de extrair conhecimento desta base, foi utilizado o processo de Knowledge Discovery in Databases (KDD), que é constituído por um conjunto de etapas capazes de tratar e gerar informações confiáveis oriundas de uma ou mais bases.

ISSN: 2236-0867

2.1 Etapas do Processo de KDD

Para que a base tivesse seus dados tratados e transformados para serem minerados, foi empregado o processo KDD, que, segundo Fayyad (1996), tem como objetivo encontrar padrões relevantes e desconhecidos a partir de uma base de dados. Este processo possui um conjunto de quatro etapas: seleção, pré-processamento, transformação e mineração de dados, conforme ilustra a Figura 1.

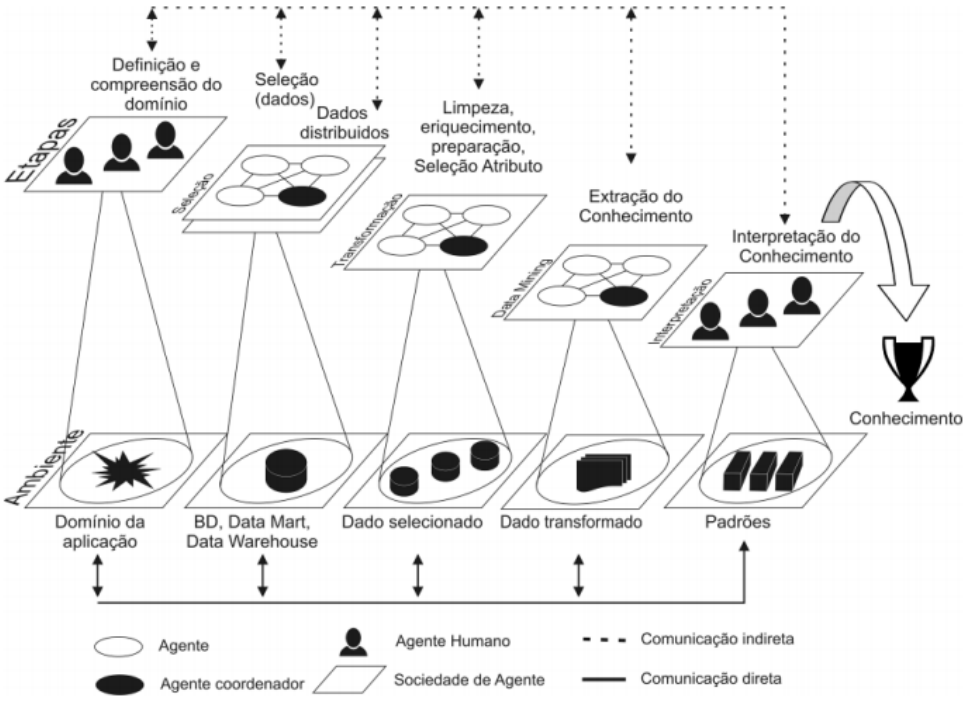

Figura 1: Etapas do KDD (Knowledge Discovery in Databases).
Fonte: (COSTA, 2008 apud FERNANDES, 2010).

Com base na figura apresentada, é possível observar o fluxo das etapas do processo KDD que foram aplicadas no desenvolvimento deste trabalho. Tais etapas são explicadas em detalhe nos itens abaixo.

2.1.1 Seleção

Com a finalidade de traçar um perfil do nível dos estudantes de medicina do país, foi utilizada uma base de dados do INEP, mais especificamente do ENADE 2013, que conta com dados sobre o perfil dos alunos que prestaram o exame e suas respostas, tanto no questionário de percepção da prova, quanto no questionário do estudante. A Figura 2 demonstra estes dados.

ISSN: 2236-0867

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_inscri	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		

Figura 2: Estado inicial da base de dados referente ao ENADE 2013.
Fonte: Portal do INEP.

Nesta base, foi possível obter dados relacionados aos estudantes que prestaram o exame em 2013, ano responsável por avaliar os cursos das áreas de saúde, ciências agrarias e afins, como: medicina, enfermagem, agronomia, farmácia, dentre outros. A base escolhida possui no total de 131 variáveis distintas, tendo dentre elas, idade, sexo, nota no componente específico, categoria da instituição de ensino superior (federal, estadual, municipal, privada sem fins lucrativos e privada com fins lucrativos), etc. A partir destas variáveis, a base contém um total de 196.856 registros. Porém, foram utilizados neste estudo somente 14.142 destes registros, pertencentes ao curso de medicina.

A retirada dos demais cursos foi devido ao objetivo deste trabalho, que é analisar o perfil dos estudantes de medicina que prestaram o exame. Tal separação auxiliou na busca por melhores resultado, além de aprimorar o foco das informações encontradas, uma vez que eliminou dados desnecessários para a pesquisa.

Após seleção da base, os atributos pertencentes a esta precisavam ser filtrados e trabalhados, para que a extração do conhecimento seja feita da melhor forma possível. Tais processos estão descritos no pré-processamento.

2.1.2 Pré-processamento

Segundo Neves (2003), a etapa de pré-processamento é responsável pela análise dos dados, para que estes se tornem consistentes e confiáveis, o que inclui a estrutura das tabelas, valores dos atributos, tipos e formatos dos dados. Outras operações também pertencentes a esta fase são a escolha dos dados pertinentes aos objetivos do usuário, juntamente com o trabalho derivado desta escolha. Além da limpeza e transformação destes dados, para que se torne viável a Mineração de Dados.

Uma vez que os dados presentes na base podem estar em formatos diferentes, a aplicação da segunda etapa do processo de KDD, o pré-processamento é de suma importância, visto que esta é responsável por limpar e formatar os dados da base selecionada, deixando-os de forma padronizada, o que permite que sejam aplicadas, as técnicas de mineração, realizadas por ferramentas especializadas.

ISSN: 2236-0867

Com a base de dados já selecionada, os dados relativos aos estudantes de medicina foram separados dos demais. Tais dados foram removidos a partir do atributo "co_grupo", variável na qual se encontra o código relativo a cada curso, deixando apenas o referente à medicina. Tal trabalho tornou possível a eliminação desta coluna, uma vez que ela só continha um valor.

Após a separação dos demais cursos, ainda era necessária a diminuição dos dados, uma vez que ainda sobravam 130 atributos distintos e, com isso, 2.005.842 milhões de dados, onde, muitos deles não demonstravam relevância para o estudo em questão.

Visando aprimorar o desempenho do algoritmo de análise, outras colunas foram removidas. Estas colunas foram descartadas principalmente por não possuírem importância para o estudo, ou seja, não acrescentavam nenhuma informação que impactasse nos resultados, ou até mesmo, poderiam influenciar negativamente.

Para finalizar esta etapa, foi realizada uma análise dos valores encontrados em cada um dos atributos selecionados, que somavam 16.264 registros até então. Após a realização de tal análise, foram detectados valores vazios e até incorretos, o que tonava inviável a utilização do registro em questão. Com todos os valores inutilizáveis removidos, vazios e incorretos, a base passou a ter um total de 14141 registros.

As alterações realizadas na base de dados primárias tiveram como objetivo, além de tornar a base devidamente estruturada para a aplicação de técnicas de mineração de dados, buscar atributos relevantes para a análise do perfil dos estudantes de medicina que prestaram o exame. A tabela resultante após a etapa de pré-processamento pode ser observada na Figura 3.

cd_catad	cd_orgac	co_uf_curso	nu_idade	tp_sexo	nt_ce	co_rs_i2
1	1	51	25	M	31.1	B
1	1	51	25	F	62.7	C
1	1	51	24	F	54.5	C
1	1	51	25	F	56	C
1	1	51	26	M	19.3	D
1	1	51	25	F	61.3	C
1	1	51	25	M	35.5	C
1	1	51	24	F	54.1	C
1	1	51	24	F	64.8	C
1	1	51	27	M	44.2	C
1	1	51	26	F	46.8	C
1	1	51	23	F	55.9	C
1	1	51	25	M	49.6	C
1	1	51	26	M	33.9	C
1	1	51	23	M	20.4	A
1	1	51	25	M	71.4	C
1	1	51	24	M	56.8	C
1	1	51	24	M	56.6	C
1	1	51	27	F	34.8	D
1	1	51	27	M	37.1	D
1	1	51	24	F	47.3	C
1	1	51	27	F	59.9	C
1	1	51	35	M	70.5	C
1	1	51	24	F	65.3	C

Figura 3: Base de dados pré-processada.

Ao final da etapa de pré-processamento, a base estava completamente reestruturada, contendo somente os atributos considerados relevantes para o estudo. Os atributos resultantes foram: cd_catad, cd_orgac, co_uf_curso, nu_idade, tp_sexo, nt_ce e co_rs_i2. Vale ressaltar que os nomes

ISSN: 2236-0867

dos atributos são os mesmos encontrados na base original do INEP. O Quadro 1 apresenta os atributos selecionados com suas respectivas descrições.

Quadro 1: Relação dos atributos com suas respectivas descrições.

Atributo	Descrição
cd_catad	Código da categoria administrativa da IES
cd_orgac	Código da organização acadêmica da IES
co_uf_curso	Código da UF de funcionamento do curso
nu_idade	Idade do inscrito em 24/11/2013
tp_sexo	Sexo do inscrito
nt_ce	Nota bruta no componente específico
co_rs_i2	Qual o grau de dificuldade desta prova na parte do Componente Específico?

Os atributos descritos no Quadro 1 representam as colunas de maior relevância, encontradas na base de dados do ENADE 2013, para o estudo em questão. Os atributos cd_catad, cd_orgac, co_uf_curso e co_rs_i2 estão com seus valores codificados e a relação entre estes atributos e a descrição de seus valores está representada no Quadro 2.

Quadro 2: Relação dos valores em código dos atributos com suas respectivas descrições.

Atributo	Código	Descrição dos códigos						
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_uf_curso	11	RO	21	MA	28	SE	42	SC
	12	AC	22	PI	29	BA	43	RS
	13	AM	23	CE	31	MG	50	MS
	14	RR	24	RN	32	ES	51	MT
	15	PA	25	PB	33	RJ	52	GO
	16	AP	26	PE	35	SP	53	DF
	17	TO	27	AL	41	PR		
co_rs_i2	A	Muito fácil						
	B	Fácil						
	C	Médio						
	D	Difícil						
	E	Muito difícil						

A descrição tanto dos atributos, quanto de seus valores, pertinentes a base primária, podem ser encontrados junto com a base, dentro do portal do INEP. Este dicionário de variáveis é obtido junto com a base.

2.1.3 Transformação

A etapa de transformação antecede a fase de mineração, nela os dados devem ser devidamente formatados, com a finalidade de melhorar e aprimorar os resultados da mineração.

ISSN: 2236-0867

Como visto na etapa de pré-processamento, muitos atributos possuíam valores codificados. Tal codificação, quando não necessária, pode prejudicar a análise das informações geradas através da mineração. Estas variáveis então passaram por uma transformação, tendo seus valores em código substituídos pelos valores reais, demostrados no Quadro2. Os atributos que passaram por esta transformação são: cd_catad, cd_orgac e co_uf_curso.

Na base, alguns atributos apresentavam uma grande quantidade de valores, o que afeta diretamente, de forma negativa, os resultados das minerações, por isso estes atributos tiveram seus valores transformados. Estes atributos passaram possuir intervalos de valores, no lugar de um valor específico, conforme representado no Quadro 3 e Quadro 4.

Quadro 3: Valores do atributo nu_idade transformados e relacionados com suas respectivas descrições.

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 4: Valores do atributo nt_ce transformados e relacionados com suas respectivas descrições.

Intervalos referentes ao atributo nt_ce	Descrição
< 60	Para todas as notas, do componente específico do exame, menores que sessenta
>= 60 e < 80	Para todas as notas, do componente específico do exame, maiores ou iguais a sessenta e menores que oitenta
>= 80	Para todas as notas, do componente específico do exame, maiores ou iguais a oitenta.

Os intervalos, demonstrados no Quadro 3, foram estruturados de forma que o atributo nu_idade, da qual pertencem, tivessem seus valores distribuídos relativamente iguais. O atributo nt_ce, que representa a nota do coeficiente específico do exame, também teve seus valores transformados, conforme apresentado no Quadro 4. 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.

2.1.4 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 estes serão intensamente analisados e explorados, buscando encontrar padrões e assim extraindo informações úteis.

Segundo Steiner e Bothorel (apud CRETTON, 2015), para a extração de conhecimento em grandes bases de dados, a utilização de técnicas inteligentes, que auxiliem na análise e interpretação, são imprescindíveis, pois, quanto maior o volume de dados, mais difícil e complexa se torna a interpretação humana. A mineração de dados busca justamente isso, extrair conhecimentos e padrões oriundos de grandes bases de dados. Entretanto, não se pode dizer que todo grande volume de dados gera conhecimento, já que, muitas vezes, isto não acontece.

Para Cardoso (2008), a mineração de dados, 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.

ISSN: 2236-0867

Nesta etapa, a extração de conhecimento pode ser realizada de várias maneiras, como: regressão, clusterização, classificação e regras de associação.

2.1.4.1 Tarefa de Classificação

Na etapa de mineração de dados, a base de dados, trabalhada ao longo dos processos anteriores do KDD, foi analisada e trabalhada através de técnica de classificação e utilizou a técnica de árvore de decisões. Para Goldshmidt (2005), a tarefa mais importante e mais utilizada, é a de classificação.

Segundo Tan, Steinbach e Kumar (2009), dentre as técnicas de classificação, técnica de árvore de decisão é a mais intuitiva, uma vez que sua representação do modelo em formato de árvore facilita o entendimento dos padrões encontrados. Esta técnica é muito utilizada para analisar problemas de classificação que envolve um certo grupo, pois podem gerar, os modelos em árvore, onde ambos podem ser utilizados para descoberta de conhecimento úteis derivados de uma base de dados. A utilização desta técnica de classificação apresentou-se mais propícia a obter melhores resultados e para que a meta desta pesquisa fosse alcançada.

A classificação pode ser utilizada com vários objetivos, como, análise de clientes, tendências do mercado financeiro, análise dos produtos mais vendidos, detecção de fraudes, dentre outros (SANTOS, AZEVEDO, 2005).

Neste trabalho, a técnica de classificação utilizada, árvore de decisão, foi aplicada através do algoritmo J48. O algoritmo em questão é apresentado no item a seguir.

2.1.4.2 Algoritmo de Arvore de Decisão J48

Neste algoritmo, a árvore de decisão é modelada baseada no atributo de maior significância, que aparece como a raiz da árvore. A partir desta raiz, são geradas ramificações, que representam a relevância desta ligação. Estas ramificações podem também gerar outras ramificações que funcionariam da mesma forma. Tal estrutura teria então a capacidade de representar, de forma intuitiva, padrões simples e complexos, de onde as informações poderiam ser extraídas.

Goldshmidt (2005), diz que as árvores de decisão também são conhecidas pelos nomes de árvores de regressão ou até árvores de classificação e que elas são representações gráficas de um conjunto de regras, constituídas por raízes, galhos e nós, semelhante a uma árvore, onde a análise destas representações devem ser realizadas do topo para as folhas. Essas árvores de decisão têm como os nós não folha como os valores dos atributos da base e os nós folha como as instâncias destes, ou seja, cada uma das decisões tomadas para a realização desta classificação são pertinentes a um único nó.

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.

3 Resultados e Discussões

As árvores apresentadas são derivadas da base de dados do INEP, puramente 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.

ISSN: 2236-0867

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. Estes resultados tiveram um nível de confiança de 84%, o que demonstra 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, a árvore foi separada em três partes, IES sem fins lucrativos, IES com fins lucrativos e IES municipais. É possível observar que dois outros tipos de IES não foram apresentados, IES estaduais e IES federais. Isto ocorreu, pois ambas não geraram informações interessantes para o estudo em questão.

Em destaque, é possível observar nestas árvores a influência do atributo cd_catad, que representa o tipo das IES, nos resultados gerados, uma vez que este foi utilizado como primeira instância, tornando ele o atributo do qual os ramos seriam formados. Como segunda instância, foi empregado, principalmente, o atributo nt_ce, que possui os intervalos de notas dos estudantes relativas ao componente específico do exame, mostrando a relevância deste na mineração.

A Figura 4 apresenta os principais resultados da árvore relativa às Instituições de Ensino Superior privadas sem fins lucrativos.

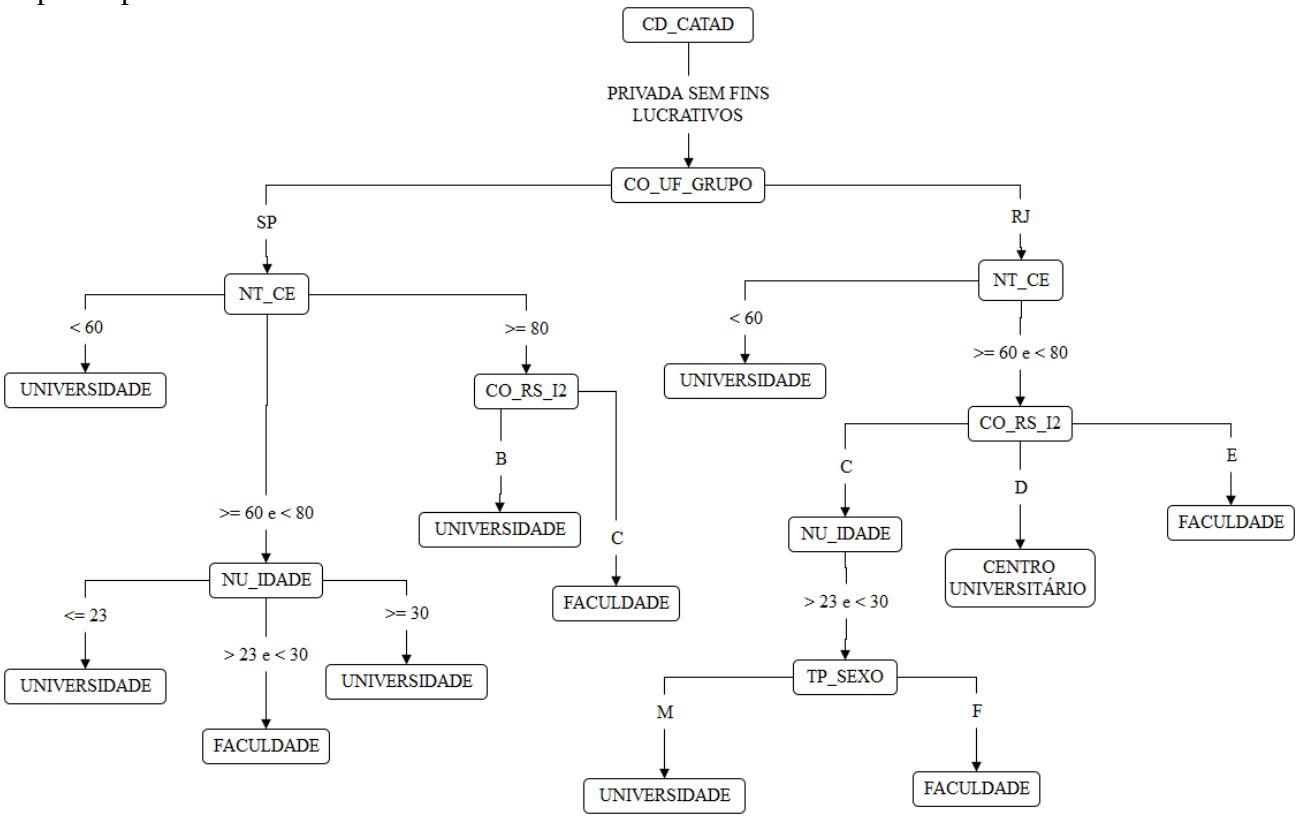

Figura 4: Árvore de decisão referente às instituições privadas sem fins lucrativos.

Observando a Figura 4, é possível analisar os principais padrões referentes às IES privadas sem fins lucrativos, como também, obter conhecimentos importantes relativos a estes padrões.

Nesta categoria, foi descoberto que, no estado de São Paulo, os estudantes pertencentes à universidades tiveram, em sua maioria, nota menor que sessenta no componente específico do exame, independentemente de qualquer outro atributo. Neste mesmo estado, foi observado que estudantes de universidades com idade menor ou igual a vinte e três anos e com idade maior ou

ISSN: 2236-0867

igual a trinta anos, assim como os alunos de faculdade com idade entre vinte e quatro anos e vinte e nove anos, obtiveram nota maior ou igual a sessenta e menor que oitenta.

Os estudantes nota maior ou igual a oitenta, quando de universidades, responderam, no questionário de percepção da prova, como fácil o conteúdo do componente específico, já os alunos de faculdade, como de média dificuldade.

Nas instituições privadas sem fins lucrativos do estado do Rio de Janeiro, assim como em São Paulo, a maior parte dos estudantes oriundos de universidades teve um rendimento menor que sessenta, independente dos outros atributos. Os estudantes com nota, no componente específico da prova, maior ou igual a sessenta e menor que oitenta, quando vindos de uma faculdade, optaram por responder, no questionário de percepção da prova, como muito difícil, o grau de dificuldade deste componente. Os alunos com deste mesmo estado e com a mesma nota, quando oriundos de um centro universitário, marcaram como difícil sobre o grau de dificuldade do componente específico. Foi observado ainda que estudantes com este mesmo rendimento, quando responderam como mediano, sobre o nível de dificuldade da parte específica do exame, e com idade maior que vinte e três anos e menor que trinta anos, foram divididos em dois grupos, sendo os da universidade de sexo masculino e os da faculdade de sexo feminino.

Na Figura 5, são demonstrados os resultados mais significativos pertinentes às IES privadas com fins lucrativos.

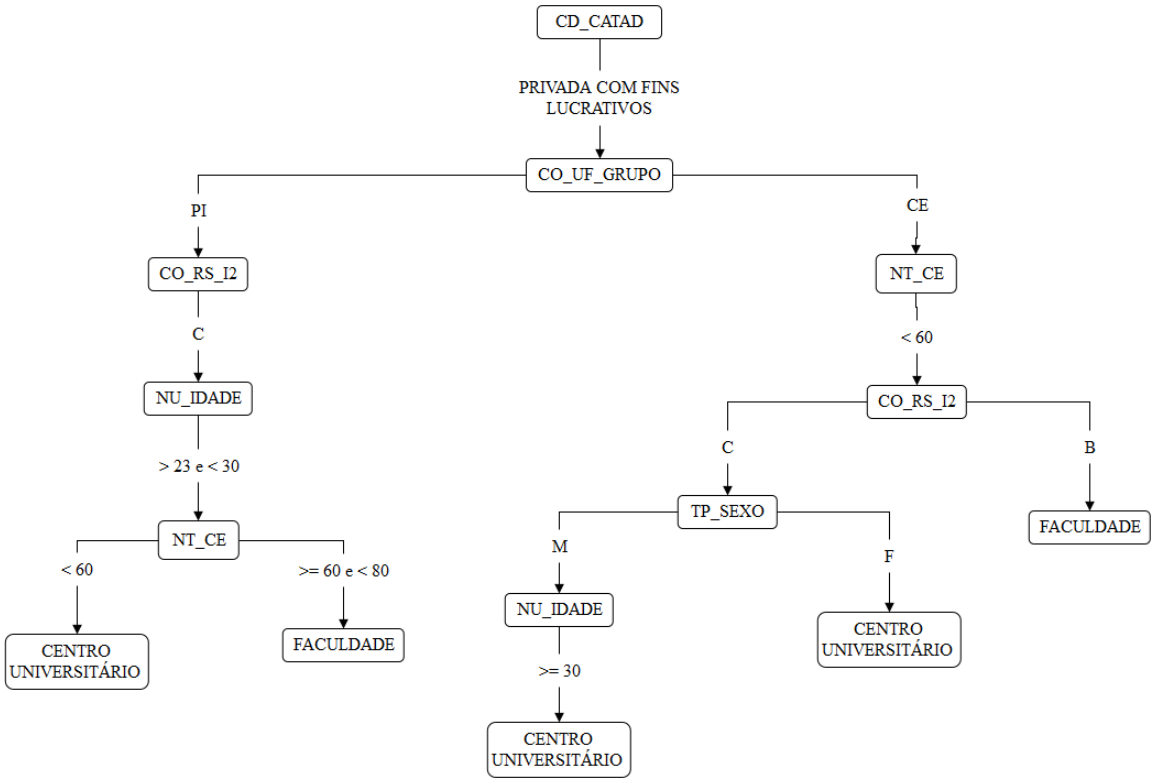

Figura 5: Árvore de decisão referente às instituições privadas com fins lucrativos.

A Figura 5 apresenta a árvore de decisão gerada a partir das principais regras derivadas das IES privadas com fins lucrativos. Tal árvore possibilita uma visualização mais intuitiva dos padrões encontrados, facilitando assim, a obtenção de conhecimento.

ISSN: 2236-0867

Nas IES privadas com fins lucrativos, dois estados se destacaram nos resultados encontrados, o estado do Piauí e o estado do Ceará. No Piauí, os estudantes, quando oriundos de centros universitários e com idade maior que vinte e três anos e menor que trinta anos, obtiveram nota menor do que sessenta no componente específico da prova e marcaram como médio o grau de dificuldade da parte específica. Quando vindos de uma faculdade, os alunos com esta mesma idade tiveram um rendimento maior ou igual a sessenta e menor que oitenta no componente específico do exame e também optaram por marcar como médio o nível de dificuldade deste componente.

No Ceará, os estudantes tiveram, em sua maioria, um rendimento menor que sessenta no componente específico, independente do tipo de instituição. Este resultado, ainda se desmembra de acordo com a resposta dos estudantes no questionário de percepção do exame, onde alunos de faculdade com esta nota, responderam como fácil o grau de dificuldade do componente específico. Enquanto que os estudantes que marcaram como médio o nível de dificuldade desta parte da prova, estes são oriundos de um centro universitário sejam do sexo feminino ou também do masculino, com idade maior ou igual a trinta.

Por fim, a Figura 6 representa as regras de maior relevância, para às IES municipais, por meio de uma árvore de decisão.

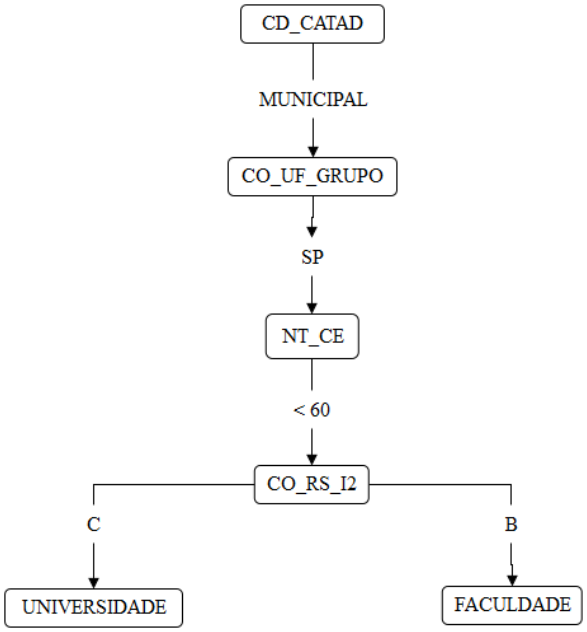

Figura 6: Árvore de decisão referente às instituições municipais.

Na Figura 6, os estudantes das IES municipais pertencentes ao estado de São Paulo, obtiveram notas menores que sessenta no componente específico do exame, sendo os mesmos oriundos de faculdades e responderam como fácil este componente. Os das universidades obtiveram a mesma nota e marcaram como médio o grau de dificuldade da parte específica.

Como resultado final, podem ser observados os perfis dos estudantes que prestaram o exame, com relação a sua nota no componente específico e a sua resposta no questionário de percepção da prova, mais especificamente sobre o grau de dificuldade deste componente. O perfil contou com o tipo de instituição que o estudante frequenta, o estilo desta organização (universidade, faculdade ou centro universitário), o estado desta instituição, o sexo e idade do aluno, sua nota no componente específico e sua resposta sobre o nível de dificuldade desta parte do exame.

ISSN: 2236-0867

Com base nos perfis gerados, é possível observar que a maior parte possui notas inferiores a sessenta e que, quando envolvendo a resposta sobre o grau de dificuldade do componente específico, alguns padrões mostraram que parte destes estudantes não demonstra coerência, o que sugere um domínio ainda menor sobre o conteúdo.

Com a aplicação das metodologias previamente citadas, foram encontrados padrões úteis, que podem ser utilizados pelas instituições nas suas tomadas de decisões. Apesar de todos os estados e tipos de instituição terem sido utilizados neste estudo, não foi possível encontrar padrões relevantes para todos estes elementos.

A mineração de dados não é comumente aplicada nas bases do ENADE, porém, pode-se citar o trabalho 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.

4 Conclusão

Através da utilização dos processos do KDD, juntamente com as técnicas de mineração de dados empregadas, foi possível gerar resultados relevantes que demonstram a importância das análises de bases de dados. Os padrões e conhecimentos derivados desta análise podem auxiliar de forma positiva, tanto para os estudantes de medicina ou candidatos a vestibulares desta área, quanto para as próprias instituições, facilitando nas tomadas de decisões e aprimoramento do curso.

Para este trabalho, foi utilizado o software WEKA para a realização de técnicas de mineração de dados, onde foram feitas classificações através do algoritmo J48, a fim de identificar o perfil dos estudantes de medicina que prestaram o ENADE em 2013. A base apresenta dados sobre todos os estados brasileiros com curso de medicina, porém, São Paulo, Rio de Janeiro, Piauí e Ceará, apresentaram resultados mais relevantes.

Foram levadas em consideração a idade e o sexo dos estudantes, juntamente com suas respectivas notas no componente específico do exame e suas respostas sobre o grau de dificuldade desta parte da prova, além das informações sobre as instituições que estes frequentam, como a categoria da instituição, seu tipo e estado em que se encontra. A partir destes dados e de outras pesquisas sobre mineração de dados, critérios foram estabelecidos para que fosse realizada uma análise mais detalhada do nível dos alunos destas instituições, com o objetivo extrair conhecimento e padrões dos mesmos.

Como resultados relevantes, pode-se ressaltar que os estudantes das instituições privadas sem fins lucrativos, tanto de São Paulo, quanto do Rio de Janeiro, quando pertencentes a organização acadêmica universidade, obtiveram, na sua maioria, nota menor que sessenta, independente dos outros atributos. Já os estudantes das instituições privadas com fins lucrativos do Ceará, quando oriundos de faculdades, além de possuírem um rendimento abaixo da média, de menor que sessenta, também disseram que a prova foi fácil. Por fim, os estudantes das instituições municipais do estado de São Paulo, quando pertencentes a faculdades, tiveram o rendimento menor que sessenta e responderam como “fácil” o grau de dificuldade do componente específico da prova.

Estes resultados demonstram que, além de muitos estudantes não obterem um resultado positivo, algumas vezes, esse grupo ainda aparenta sair do exame com opiniões contrárias às suas notas, afirmando considerar o componente específico fácil.

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 de medicina, objetivando a geração de profissionais devidamente aptos e com um maior nível de conhecimento, tornando-os assim,

ISSN: 2236-0867

melhores médicos. Também é almejado que os futuros estudantes de medicina possam utilizar estas informações para escolher melhor as instituições na qual irão investir.

5 Referências

BORGES, José Leopoldino das Graças; CARNIELLI, Beatrice Laura. Educação e estratificação social no acesso à universidade pública. **Cadernos de Pesquisa,** São Paulo, v. 35, n. 124, p.113-139, abr. 2005. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0100-15742005000100007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-15742005000100007&lng=en&nrm=iso>. Acesso em: 04 set. 2015.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. Gestão do conhecimento usando data mining estudo de caso na Universidade Federal de Lavras. **Revista Brasileira de Administração Pública,** Rio de Janeiro, v. 42, n. 3, p.495-528, jun. 2008. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/184>. Acesso em: 02 set. 2015.

CRETTON, Nícollas Nogueira; FONTANA, Valderedo Sedano; GOMES, Geórgia Regina Rodrigues. 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.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **Ai Magazine,** Palo Alto, v. 17, n. 3, p.37-54, set. 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 15 set. 2015.

FERNANDES, Mauricio Natividade de Oliveira; GOMES, GeÓrgia Regina Rodrigues; SHIMODA, Eduardo. Utilização de Mineração de Dados para Descrição do Perfil de Pacientes Otimistas, Realistas e Pessimistas quanto a Própria Saúde Bucal. In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 17, 2010, Bauru. **Anais... .** Bauru: Simpep, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia Prático:** Conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. 256 p.

Luiza Yoko Taneguti. **PROJETO CNE/UNESCO 914BRZ1136.3 "Desenvolvimento, aprimoramento e consolidação de uma educação nacional de qualidade".** Brasília: Conselho Nacional de Educação, 2013. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=13948-produto-2-oferta-demanda-educ-superior-pdf-pdf&category_slug=setembro-2013-pdf&Itemid=30192>. Acesso em: 06 set. 2015

MARTINS, António Cardoso; MARQUES, João Miguel; COSTA, Paulo Dias. Estudo Comparativo De Três Algoritmos De Machine Learning Na Classificação De Dados Electrocardiográficos. **Trabalho (Mestrado em Informática Médica)** – Universidade do Porto, Porto, mar. 2009. Disponível em:

ISSN: 2236-0867

<http://www.dcc.fc.up.pt/~ines/aulas/0910/MIM/trabs_ano_anterior/noname-1.pdf>. Acesso em: 08 set. 2015.

MINISTÉRIO DA EDUCAÇÃO. **Portaria Normativa MEC nº 40/2007**: Institui o e-MEC, sistema eletrônico de fluxo de trabalho e gerenciamento de informações relativas aos processos de regulação, avaliação e supervisão 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. Brasília: Diário Oficial da União, 2010.

NEVES, Clarissa Eckert Baeta. Ensino Superior no Brasil: expansão, diversificação e inclusão. In: LATIN AMERICAN STUDIES ASSOCIATION, 30., 2012, São Francisco. **Anais... .** São Francisco: Lasa, 2012. Disponível em: <http://flacso.redelivre.org.br/files/2013/03/1114.pdf>. Acesso em: 11 set. 2015.

NEVES, Rita de Cássia David das. **Pré-Processamento no Processo de Descoberta de Conhecimento em Banco de Dados.** 2003. 137 f. Dissertação (Mestrado) - Curso de Programa de Pós-graduação em Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/2701/000375412.pdf?sequence=1>. Acesso em: 12 set. 2015.

NOGUEIRA, Eduardo Dimas Andrino; TSUNODA, Denise Fukumi. Mineração de dados para análise da relação entre as características socioeconômicas de concluintes do ensino superior e o desempenho desses estudantes no enade 2012. **Percurso,** Curitiba, v. 15, n. 1, p.245-268, 2015. Disponível em: <http://revista.unicuritiba.edu.br/index.php/percurso/article/view/1102/761>. Acesso em: 15 set. 2015.

PRESIDÊNCIA DA REPÚBLICA. Congresso. Senado. Lei nº 10.861/2004, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras providências. **Lei no 10.861, de 14 de Abril de 2004.**. Brasília, DF, 15 abr. 2004. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/Lei/L10.861.htm>. Acesso em: 08 set. 2015.

PRIMI, Ricardo; HUTZ, Cláudio S.; SILVA, Marjorie Cristina Rocha da. A prova do ENADE de psicologia 2006: concepção, construção e análise psicométrica da prova. **Aval. Psicol.,** Itatiba, v. 10, n. 3, p.271-294, dez. 2011. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712011000300004&lng=pt&nrm=iso>. Acesso em: 10 set. 2015.

SANTOS, Manuel Filipe; AZEVEDO, Carla Sousa. **Data mining: descoberta de conhecimento em bases de dados.** Lisboa: Fca, 2005. 214 p.

SILVA, Marcelino Pereira dos Santos. **Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka.** 2004. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>. Acesso em: 09 set. 2015.

ISSN: 2236-0867

TAN, Pang-ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Datamining – Mineração de Dados.** Rio de Janeiro: Ciência Moderna, 2009. 928 p.

ZAGO, Nadir. Do acesso à permanência no ensino superior: percursos de estudantes universitários de camadas populares. **Revista Brasileira de Educação,** Rio de Janeiro, v. 11, n. 32, p.226-237, ago. 2006. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s1413-24782006000200003. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-24782006000200003&lng=en&nrm=iso>. Acesso em: 06 set. 2015.