MINERAÇÃO DE DADOS NOS MICRODADOS ENADE COMPUTAÇÃO

Wander Luis Francelino[1]; Lucas de Souza Machado[2]

[1]Universidade do Sul de Santa Catarina, Tubarão, Brasil

[2]Universidade do Sul de Santa Catarina, Tubarão, Brasil

wandluis@gmail.com, just.lucasmachado@gmail.com

Abstract. *The large amount of data found in the ENADE database is the collection of information from all students who took the exam and its results. Discovering new knowledge from this mass of data can guarantee a better understanding of the students, courses and universities they attend. The focus was on the characteristics of students and educational institutions that impact the final exam performance. For that, KDD data mining techniques and the k-means clustering algorithms were applied.*

Resumo*. A grande quantidade de dados, encontradas na base do ENADE, é a coleta das informações de todos os alunos que fizeram o exame e seus resultados. Descobrir um conhecimento novo a partir dessa massa de dados pode garantir um melhor entendimento sobre os alunos, os cursos e universidades que frequentam. O foco foi nas características dos alunos e instituições de ensino que tem impacto no desempenho final do exame. Para isso, foram aplicadas a técnicas de mineração de dado KDD e o algoritmos de clusterização k-means.*

1. Introdução

A Mineração de Dados é o centro deste trabalho, que resumidamente, é um conjunto de técnicas aplicadas em grandes massas de dados que visa obter novas informações úteis, como também dá a capacidade de prever resultados de uma observação futura, ou seja, traz para a superfície informações não contidas explicitamente na base original de dados. Quanto mais informações, ou seja, quanto maior a base de dados de uma empresa ou de uma pesquisa estatística, por exemplo, maior o leque de possibilidades de descobrir novas informações ou de prever com mais exatidão um evento (TAN; STEINBACH; KUMAR, 2009).

O ENADE (Exame Nacional de Desempenho de Estudantes) é um dos métodos de avaliação do Sinaes (Sistema Nacional de Avaliação da Educação Superior) que é realizado pelo Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), o qual é uma autarquia federal vinculada ao MEC (Ministério da Educação), com base nas diretrizes definidas pelo CONAES (Comissão Nacional de Avaliação da Educação Superior), órgão colegiado de coordenação e supervisão do Sinaes.

De acordo com o resultado da instituição o Ministério da Educação, quem aplica o exame, pode tomar medidas administrativas e até mesmo suspender o curso que não teve um rendimento aceitável. Os cursos recebem o Conceito Enade, um indicador de qualidade que avalia os cursos.

Participam da prova do ENADE, todo ano, em torno de 500 mil estudantes do ensino superior espalhadas em várias universidades de financiamento público e privado, que somadas, chegam a quase 10 mil instituições. A cada ano, cursos diferentes são avaliados por meio dos seus alunos, para descobrir a capacidade de formação acadêmica das instituições de ensino superior do Brasil (BRASIL, 2005).

A aplicação de técnicas de mineração de dados pode descobrir associações ou possibilitar previsões relevantes na base de dados do ENADE. Todavia, essa é uma das características da mineração de dados, expor uma característica, um padrão não perceptível ao analisar da forma tradicional os dados (AMO, 2014).

Tem-se a disposição os resultados dos alunos de Ciência da Computação de todo o Brasil. Conseguir descobrir novas informações sobre os alunos inscritos pode garantir um melhor entendimento sobre os alunos e os cursos e universidades que frequentam. Garante um novo ponto de vista sobre a situação.

A vasta quantidade de dados, somada a dificuldade de extração de novas informações a partir dela justificam a escolha por técnicas de mineração de dados. Algoritmos sofisticados processam os dados visando o descobrimento de padrões, características importantes para um melhor entendimento do que antes era simplesmente uma grande massa de dados.

Diante desse cenário, define-se a seguinte questão de pesquisa: Pode-se utilizar o método de descoberta de conhecimento, aplicado à base de dados do ENADE, para encontrar indicadores ou relações entre os dados, relevantes, não triviais e determinantes para melhorar o entendimento do resultado da avaliação?

O artigo segue uma estrutura de tópicos e subtópicos da seguinte forma: a segunda seção apresenta os trabalhos correlatos a este. A terceira seção refere-se à contextualização teórica da base do ENADE, de KDD, mineração de dados e do algoritmo k-means. A quarta seção descreve os materiais e métodos utilizados na descoberta de informações e descreve a metodologia de como chegou-se aos resultados interpretados. A quinta seção apresenta os resultados obtidos, bem como as discussões e pontos de vistas relacionadas à estes resultados. Por fim a sexta seção apresenta as conclusões deste artigo.

2. Trabalhos correlatos

Na contribuição de Silva *et al.* (2015), havia disponível uma massa de dados do ambiente virtual da Universidade Federal do Vale do São Francisco, contendo informações sobre acessos e interações dos alunos com as diversas disciplinas cursadas e também o seu desempenho, e a partir disso obter conhecimento relevante. Eles seguiram as principais etapas da mineração: Preparação dos dados, extração dos dados e o pós-processamento. Utilizando o software R para implementar um método de clusterização, chamado Método de Ward (WITTEN *et al*., 2017), eles apresentaram os

dados finais e suas interpretações e relacionando a quantidade de interações realizadas no ambiente virtual com o desempenho dos alunos, por exemplo.

No artigo de Fonseca e Namen (2016) é apresentado um estudo sobre a base de dados, retirada do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), com informações estatísticas e avaliações da Educação Básica no Brasil. Os autores utilizaram o método KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em Banco de Dados) para identificar os fatores que contribuem com a proficiência dos alunos, relacionado ao perfil dos professores de Matemática. Todo o processo do KDD é explicado e documentado, focando na etapa de mineração dos dados e no algoritmo utilizado para fazer a classificação. Tudo isso dentro da ferramenta WEKA que já disponibiliza versões de grande parte dos algoritmos utilizados.

No artigo de Silva e Nunes (2015), é minerada uma base de dados de uma escola particular dos alunos do Ensino Médio contendo informações dos alunos entre 2011 e 2014. Todas as informações são tratadas e importadas para o WEKA. O algoritmo escolhido foi o J48, que gera árvores de decisão e por essa razão são utilizadas como classificação dos dados de treinamento submetidos. O resultado do trabalho tentou elucidar as características formadoras do resultado final de turma da escola: aprovação ou reprovação.

3. Contextualização

3.1. Enade

O Exame Nacional de Desempenho dos Estudantes (Enade) é um dos instrumentos de avaliação do Sistema Nacional de Avaliação da Educação Superior (Sinaes). A prova é componente curricular obrigatório nos cursos de graduação, conforme determina a Lei nº 10.861/2004 (BRASIL, 2004). É aplicado periodicamente aos estudantes de todos os cursos de graduação, durante o primeiro (ingressantes) e último (concluintes) ano do curso.

A avaliação tem o objetivo de acompanhar a aprendizagem e o desempenho acadêmico dos estudantes, em relação aos conteúdos programáticos previstos nas diretrizes curriculares de cada curso. Os resultados podem produzir dados por instituição de educação superior, categoria administrativa, organização acadêmica, município, estado, região geográfica entre outros. Estes resultados são disponibilizados por meio do portal do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira).

3.2. KDD

Conforme Castanheira (2008), o KDD (Knowledge Discovery in Databases) é um processo de descoberta de conhecimento em bases de dados que tem como objetivo principal extrair conhecimento a partir de grandes bases de dados. Isto engloba diversas áreas de conhecimento, tais como: estatística, matemática, bancos de dados, inteligência artificial, visualização de dados e reconhecimento de padrões. São utilizadas técnicas, em seus diversos algoritmos, oriundos dessas áreas.

Corroborando com este pensamento, Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 40) afirmam que "KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". Para Goldschimidt e Passos (2005) o termo "iterativo" sugere a possibilidade de repetições integrais ou parciais do processo de KDD e a expressão não trivial alerta para a complexidade normalmente presente na execução de processos de KDD.

Composto de 5 (cinco) etapas, a primeira fase do KDD é o agrupamento de dados, ou seleção. Segundo Castanheira (2008), a etapa da limpeza dos dados vem a seguir, por meio de um pré-processamento dos dados, visando adequá-los aos algoritmos que serão utilizados a partir deste momento que se chega à fase de mineração de dados começa com a escolha das ferramentas (algoritmos) a serem utilizadas. Essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, agrupamento, regras associativas ou desvio. De acordo com o algoritmo utilizado será gerado um arquivo de descobertas [...] este arquivo deve ser interpretado, gerando as conclusões que fornecem o conhecimento da base de dados estudada.

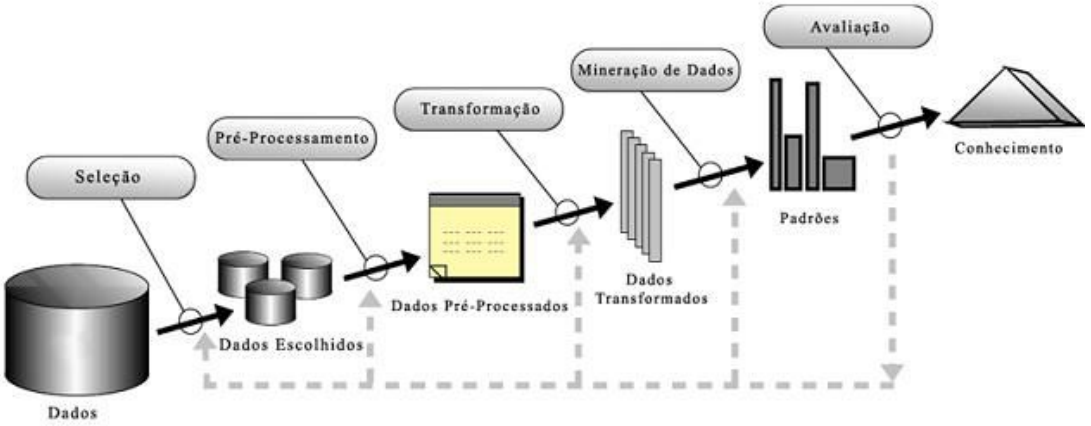

Figura 1 - Processo de KDD (Fonte: FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

3.3. Mineração de dados

A mineração de dados envolve encontrar padrões em dados. A maioria dos métodos de mineração é baseada em técnicas testadas de aprendizado de máquina e reconhecimento de padrões e estatística, como classificação, associação, regressão, clustering, sumarização, modelagem de dependência, entre outros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Por esta perspectiva percebe-se que a mineração de dados é o ponto central da descoberta de conhecimento, o KDD, como enfatiza Han, Kamber e Pei (2012, p. 8, tradução nossa) "mineração de dados é uma etapa no processo de descoberta de conhecimento, embora seja a etapa essencial por revelar padrões ocultos para avaliação".

Na fase de mineração dentro do processo de KDD, é escolhida a tarefa e definido o algoritmo a ser utilizado, podendo ser executado mais de uma vez já que esta etapa é um processo iterativo, para que haja a extração de padrões (GARCIA, 2012).

De acordo com Witten *et al.* (2017), aprendizado de máquina fornece bases técnicas à mineração de dados com algoritmos especializados em encontrar padrões em grandes massas de dados.

3.4. K-means

Segundo Tan, Steinbach e Kumar (2009) o k-means é uma técnica particional de agrupamento baseada em protótipos, que iterando sobre os objetos, tenta encontrar um número específico de grupos (k) definido pelo usuário, representados pelos seus centróides.

O k-means é um dos algoritmos mais antigos e amplamente usados para fazer análise de grupos devido a sua simplicidade e eficácia. A análise de grupos é muitas vezes chamada de classificação não supervisionada, pois encontra o rótulo para cada classe usando apenas os próprios dados, sem nenhum tipo de referência de fora (TAN; STEINBACH; KUMAR, 2009).

Em agrupamento, o objetivo é dividir o conjunto de dados em grupos, de forma que os objetos contidos na base de dados fiquem agrupados naturalmente de acordo com a semelhança entre eles. Os algoritmos de agrupamento são técnicas de aprendizado não-supervisionado, logo os grupos ou categorias, e até mesmo suas quantidades, não são conhecidos inicialmente (COSTA, 2012).

O algoritmo calcula centróides imaginários para cada cluster, que são médias da posição dos objetos que ele está clusterizando. A cada iteração os centróides movem-se dentro do plano dos dados sendo influenciados pela posição dos objetos. Uma resposta satisfatória é obtida quando a movimentação do centróide deixa de ser brusca ou cessa.

Na Figura 2 está o pseudocódigo do algoritmo k-means.

Entrada:

Obrigatórios: k (número de clusters), D (conjunto de dados contendo p objetos)

Opcional: n (número de iterações)

Método:

-Aleatoriamente escolha k objetos de D para que suas posições sejam definidas como a posição do cluster.

-Repita:

-Para cada objeto em D: calcular a distância em relação a cada cluster e classifique o objeto como pertencente ao cluster mais próximo.

-Para cada cluster k: recalcule a sua posição obtendo a média entre os objetos pertencentes a ele.

-Acrescenta contador i

Até o cluster não mudar ou i=n

Fim.

Figura 2 - Pseudo-código do K-means (Fonte: Autores).

O cálculo de distância utilizado no algoritmo é o Euclidiano, definido pela fórmula na figura 3. O cálculo de distância é utilizado para identificar qual cluster está mais próximo de cada objeto pertencente a base. Outro cálculo importante do algoritmo é feito para determinar a nova posição dos clusters, que são uma média dos atributos de cada objeto pertencente a ele, aqueles objetos que estiveram mais próximos dele (MACQUEEN, 1967). A fórmula de definição dessa média está na figura 4.

$$DistEuc(O_i, C_j) = \sqrt{\sum_{r=1}^{m} (O_{i,r} - C_{j,r})^2}$$

Onde:

m é a quantidade de atributos dos objetos e centróides

O_i é o vetor de atributos do objeto, assim $O = (O_{i,1}, O_{i,2}, \ldots, O_{i,m})$

C_j é o vetor de atributos do centroide, assim $O = (C_{i,1}, C_{i,2}, \ldots, C_{i,m})$

Figura 3 - Fórmula da distância Euclidiana aplicada ao k-means (Fonte: Autores).

$$C_{i,k} = \frac{\sum_{n \in S_k} x_{i,n}}{Q(C_k)}$$

Onde:

k é o número de clusters
n é o objeto pertencente ao cluster k
$Q(C_k)$ é o número de objetos pertencentes ao cluster C_k
$C_{i,k}$ é a dimensão i do cluster k, para i = {1, …, n} e j = {1, …, k}
S_k é o conjunto de objetos do cluster k
$x_{i,n}$ é a dimensão i do objeto n

Figura 4 - Fórmula média dos atributos (Fonte: Autores).

4. Materiais e métodos

4.1. Base do enade

A base do enade é uma estrutura tabular onde cada coluna é um atributo de cada estudante inscrito no exame e, obviamente, cada linha representa cada um dos estudantes. Os atributos são: instituição do aluno (1), sigla da instituição (2), categoria administrativa da IES (3), organização acadêmica da IES (4), município da IES (5), UF da IES (6), região da IES (7), área do curso (8), município do curso (9), UF do curso (10), total de inscritos (11), total de participantes (12), nota bruta da formação geral (13), nota padronizada da formação geral (14), nota bruta do componente específico (15), nota padronizada do componente específico (16), nota bruta geral (17), nota padronizada geral (18), conceito ENADE do curso (19), faixa ENADE do curso (20), idade do aluno (21), sexo do aluno (22), ano de conclusão do 2º grau (23), ano de ingresso na graduação (24), semestre de ingresso (25), turno matutino (26), turno vespertino (27), turno noturno (28), indicativo de concluinte ou ingressante (29), presença na prova (30), gabarito da formação geral (31), gabarito da formação específica (32), acertos da prova objetiva da formação geral (33), acertos da prova objetiva da formação específica (34), nota da prova objetiva da formação geral (35), nota da prova discursiva da formação geral (36), nota da formação geral (37), nota da prova objetiva da formação específica (38), nota da prova discursiva da formação específica (39), nota da formação específica (40), nota da prova (41), estado civil (42), escolaridade do pai (43), escolaridade da mãe (44), renda familiar (45), trabalho (46), escola ensino médio (47), modalidade ensino médio (48), escolha curso (49), escolha IES (50).

São 50 atributos que caracterizam cada um dos 9822 estudantes de Ciência da Computação que por todo país foram submetidos ao mesmo exame com o intuito de serem avaliados e serem colocados dentro de uma métrica comum. Os dados estão em formato CSV (Comma-Separeted Values), uma forma simples porém eficaz de apresentar dados tabulares, podendo ser editados por diversas ferramentas ou importados para um banco de dados.

4.2. Ferramentas

4.2.1. Weka

O *Waikato Environment for Knowledge Analysis*, ou WEKA, é uma das principais ferramentas de mineração de dados.

> [...] Muito popular no meio acadêmico, o WEKA (Waikato Environment for Knowledge Analysis) é um software livre escrito em Java e desenvolvido pela Universidade de Waikato, Nova Zelândia, formado por uma coleção de algoritmos de diversas técnicas de mineração de dados e ferramentas de pré-processamento (WEKA, [2019], p. 1).

Utilizou-se o WEKA como principal ferramenta na mineração dos dados.

4.2.2. Python

Escolheu-se como a ferramenta de programação a linguagem Python pois possui uma estrutura de código facilitada e de rápido aprendizado com muitas bibliotecas úteis dos principais algoritmos de mineração, inclusive o escolhido para este trabalho. E também bibliotecas capazes de gerar a visualização dos dados gerados pela mineração.

4.3. Modelo computacional

A partir do modelo de KDD, uniu-se algumas etapas em um único estágio, para adequar o método aos dados aqui analisados. Composto de quatro (4) etapas nossa metodologia apresenta Levantamento Inicial (1), Pré-Processamento dos Dados (2) que abrange as sub etapas de Seleção, Limpeza e Formatação dos Dados, Mineração dos Dados (3) e Resultados (4), conforme fluxograma da figura 5.

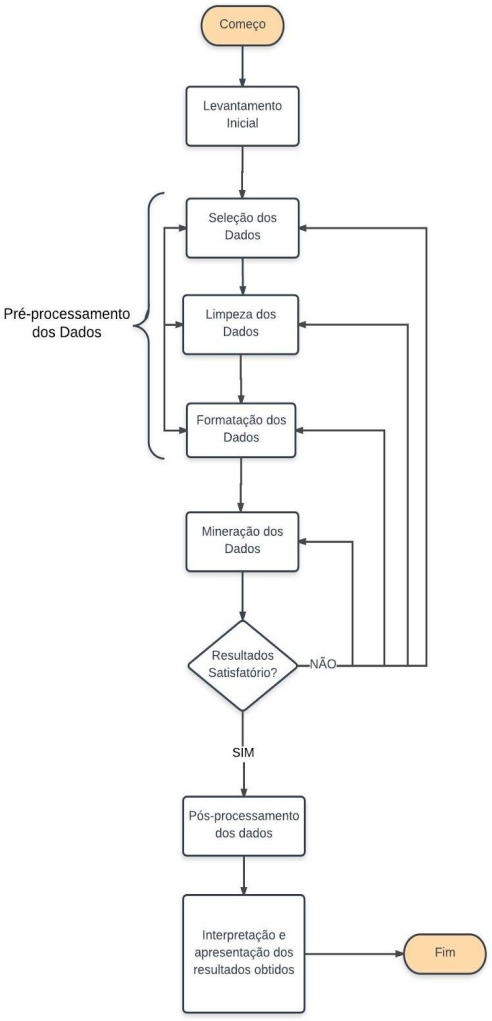

Figura 5 - Metodologia da Descoberta de Conhecimento em BDs (Fonte: Autores).

4.3.1. Levantamento inicial

O início do trabalho partiu da coleta do conjunto de dados da base de dados do ENADE aplicado em 2013, no portal no INEP[1], na seção de Microdados. O arquivo do tipo CSV contém todas as informações relevantes para o ENADE, desde informações fornecidas pelo aluno no momento da inscrição, relativos a dados pessoais e socioeconômicos, como também os relativos à realização da prova e seu resultado: notas e se o aluno compareceu ou não a aplicação do exame.

4.3.2. Pré-processamento

---

1 Microdados - INEP. Disponível em: http://portal.inep.gov.br/microdados. Acesso em: 07 nov. 2018.

Dividiu-se o pré-processamento dos dados em outras etapas para melhor sistematização e entendimento dos estágios. Os algoritmos e técnicas de mineração já estão prontas e implementadas para a utilização na ferramenta escolhida, o WEKA, porém a manipulação de dados varia de acordo com os dados.

As etapas dessa fase são a Seleção, Limpeza e Formatação dos Dados. Nestas etapas desconsiderou-se todos os alunos que não participaram no ENADE, facilmente reconhecidos por meio da coluna "presença", pois o foco desta pesquisa seria sempre com relação ao resultado. Identificou-se também alunos com dados incompletos e colunas sem valores, considerados ruídos e também foram descartados para a execução do algoritmo. Demais alunos foram deixados na base e o que antes abrangia 9.822 alunos, após as primeiras limpezas, restaram 8.264 alunos.

Continuando o processo, iniciou-se a identificação de quais dados de cada aluno seriam interessantes como predicado de cada um dos alunos que será agrupado com o algoritmo de clusterização. Das 50 colunas disponíveis, descartou-se 38 delas, ficando somente 12 disponíveis para as próximas fases.

Os 38 campos descartados são: nome da instituição de ensino (1), sigla da instituição (2), categoria administrativa (3), organização acadêmica (4), município da instituição (5), estado da instituição (6), área do curso (7), munícipio do curso (8), estado do curso (9), total de inscritos no ENADE da instituição por curso (10), total de participantes do ENADE da instituição por curso (11), nota bruta da formação geral por curso (12), nota padronizada da formação geral por curso (13), nota bruta por componente específico por curso (14), nota padronizada do componente específico por curso (15), nota geral do componente específico por curso (16), nota bruta geral por instituição por curso (17), nota padronizada geral por instituição por curso (18), nota objetiva da formação geral (19), nota discursiva formação geral (20), nota discursiva formação específica (21), nota formação específica (22), faixa ENADE do curso (23), conceito ENADE do curso (24), se o curso tem turno matutino na instituição (25), se o curso tem turno vespertino na instituição (26), se o curso tem turno noturno na instituição (27), semestre do aluno (28), ano de conclusão do ensino médio (29), ano de ingresso na graduação (30), presença do aluno (31), gabarito geral (32), gabarito específico (33), acertos da objetiva geral (34), acertos da objetiva específica (35), se aluno é concluinte ou ingressante (36), escola do ensino médio (37) e escolha do ensino médio (38).

Escolheu-se 12 colunas para a continuação do pré-processamento dos dados: região da instituição (1), idade (2), sexo do aluno (3), nota final do aluno (4), estado civil do aluno (5), escolaridade do pai (6), escolaridade da mãe (7), renda familiar do aluno (8), quantidades de hora do contrato de trabalho do aluno (9), modalidade do ensino médio do aluno (10), escolha do curso (11) e a escolha da instituição de ensino (12).

Nossas escolhas foram baseadas no fato de alguns itens já classificarem os alunos como sendo de uma escola ou de uma universidade e terem valores repetidos para vários alunos. O interesse nesta pesquisa é que outros dados que escolheu-se manter sejam mais decisivos para a clusterização. Outros dados foram retirados por serem iguais para todos os alunos, como o campo que definia a presença, o campo que

definia se ele era concluinte ou ingressante, pois todos eram concluintes e os dois campos que informava o gabarito da prova. Todos são irrelevantes para o algoritmo.

Na etapa de formatação de dados, alterou-se a representação de alguns dos atributos dos alunos com o intuito de categorizar faixas dados que podem ser considerados de certa forma como iguais. Agrupou-se a idade em cinco conjuntos de A até E, onde o grupo A abrange a faixa de alunos de 18 até 24 anos, B contém alunos dos 25 aos 30 anos, C os alunos dos 31 aos 40 anos, D os alunos com 41 aos 50 anos e por fim o grupo E com alunos com mais de 51 anos de idade.

A partir da coleta e o pré-processamento dos dados, iniciou-se a análise exploratória dos elementos coletados, para assim ter um completo entendimento do contexto das informações e suas dimensões. Novamente usou-se as ferramentas Excel e Weka com intuito de realizar análises básicas e ter uma visão mais aprofundada agora que passou-se pelas fases iniciais processamento de dados e com mais clareza identificar informações que poderiam ser mais relevante e com potencial para obter resultados relevantes. Principalmente com o auxílio de gráficos:

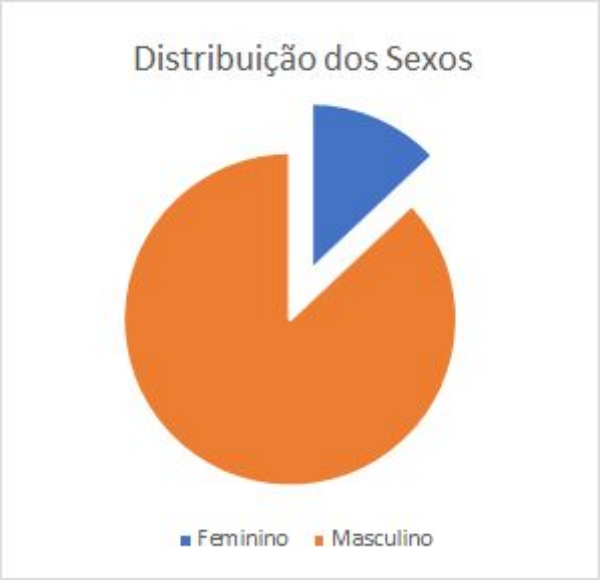

Gráfico 1 - Distribuição do Sexo dos Alunos no Enade (Fonte: Autores).

No gráfico 1 nota-se que a maioria dos alunos de Ciência da Computação, 87%, são homens, enquanto as mulheres somam 13%.

No próximo gráfico, observa-se a distribuição dos alunos pelas regiões e suas notas médias, onde é possível notar uma diferença de até 6 pontos entre a região de menor média, a região Nordeste, comparada a região de maior média, região Sul.

O número de alunos na região Sudeste é notoriamente maior que as demais regiões.

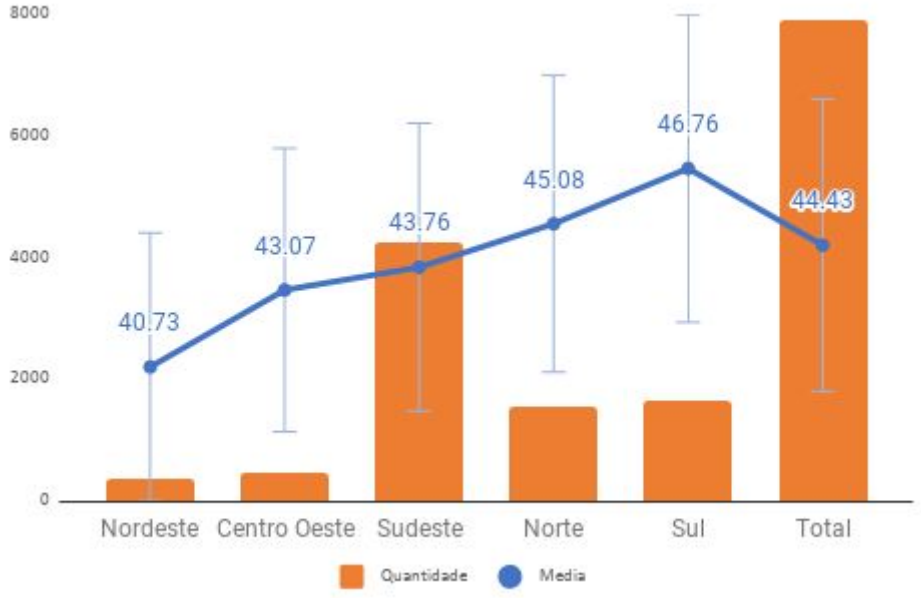

Gráfico 2 - Média das Notas por Região Geográfica do Brasil (Fonte: Autores).

Apresenta-se no gráfico 3 o demonstrativo da relação dos alunos e seus respectivos regimes de trabalho. Esta análise é interessante para demonstrar que menos da metade dos alunos tem a oportunidade de não trabalhar ou trabalhar meio período (20 horas) e ter mais tempo para se dedicar aos estudos, enquanto grande parcela, a maioria, precisa se dedicar à faculdade conciliando o trabalho integral.

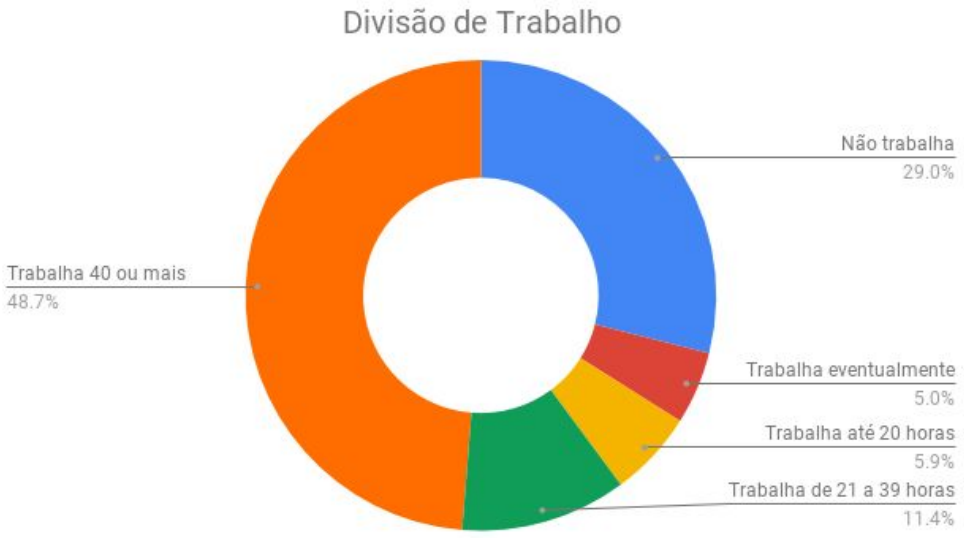

Gráfico 3 - Divisão de Trabalho (Fonte: Autores).

A média da nota e de cinco atributos escolhidos, serão apresentadas no quadro a seguir.

Nota (de 0 até 100)	44,43
Estado Civil	Solteiro
Escolaridade do Pai	Ensino médio
Escolaridade da Mãe	Ensino médio
Renda Familiar	De 3 até 4,5 salários mínimos
Horas de Trabalho (Semanal)	40 horas ou mais

Quadro 1 - Média dos Atributos para Todos os Alunos (Fonte: Autores).

4.3.3. Mineração de dados

No Weka foram selecionados os atributos de interesse, como nota da prova e escolaridade dos pais, por exemplo. Após a importação, na opção *cluster*, o algoritmo utilizado para gerar os grupos é o *Simple K-means*. Entre os parâmetros de inicialização do algoritmo é possível configurar a função de distância como Euclidiana, geometria do táxi, *Manhattan* entre outras. O método de inicialização pode ser randômico, mais distante, entre outros. O número máximo de iteração e o número de *clusters* são de livre escolha e afetam diretamente o resultado da análise.

Dentro da ferramenta Weka ao final do algoritmo, há uma avaliação da qualidade da distribuição dos dados nos clusters, indicando o quanto podemos confiar no clusterização e o quão sólido é cada cluster. Esse foi o principal método de validação dos dados gerados após a aplicação do algoritmo de mineração k-means dentro da ferramenta.

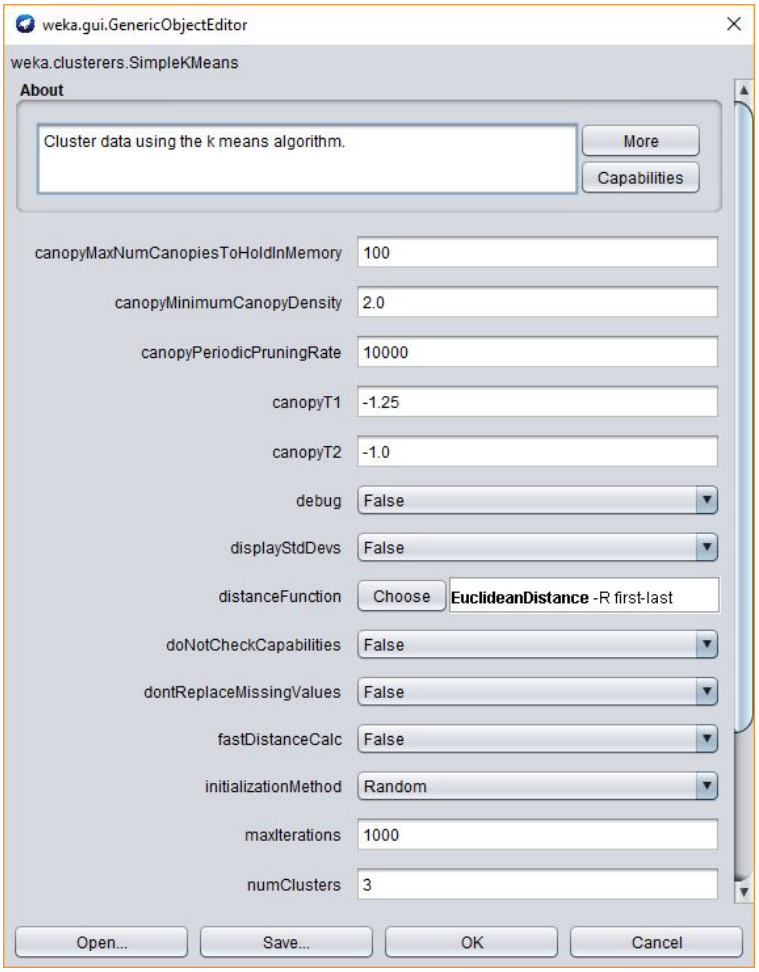

Figura 6 - Janela de parametrização do algoritmo no WEKA (Fonte: Autores).

Chama-se A1 e A2 cada clusterização gerada, onde os atributos são: nota prova, estado civil, escolaridade pai, escolaridade mãe, renda familiar e trabalho, configurou-se os parâmetros do algoritmo da seguinte forma:

Parâmetros	A1	A2
Número de Iterações	500	500
Distância do Centróide	Euclidiana	Euclidiana
Número de Instâncias	8264	8264
Número de Clusters	4	3

Quadro 2 – Parâmetros do algoritmo (Fonte: Autores).

Escolheu-se o número de iterações igual a 500 por não haver alteração significativa nos resultados extraídos utilizando-se um número maior. Enquanto a

escolha pela distância euclidiana é devida aos dados não serem padronizados, tornando o resultado final insensível a *outliers.*

5. Resultados e discussões

Após o desenvolvimento do experimento cabe expor os resultados obtidos bem como discorrer sobre os mesmos. Sempre visando relacionar os microdados do ENADE com as notas obtidas pelos alunos. Depois todos os passos da mineração obteve-se os resultados que serão discutidos nos parágrafos a seguir, com o auxílio de tabelas e gráficos manufaturados que melhoram a apresentação e por consequência o entendimento dos mesmos.

Na primeira mineração de dados que chama-se A1 nota-se que 31% do total, ou 2.562 alunos, no grupo (cluster #3) de maior nota, sendo esta de 52.13 pontos. O grupo de menor nota (cluster #2) soma 562 alunos ou apenas 7% do total analisado, sendo sua nota de 35.60 pontos. O maior grupo (cluster #1) com 3.361 alunos, ou 41% do total, alcança uma nota de 43.18, é também o grupo de maior renda familiar, 4,5 até 6 salários. O último grupo (cluster #0) desta análise, alcança uma nota de 38.48.

Atributo	**Cluster #0**	**Cluster #1**	**Cluster #2**	**Cluster #3**
Nota (de 0 até 100)	38,4752	43,1849	35,598	52,1286
Estado Civil	Solteiro	Solteiro	Casado	Solteiro
Escolaridade do Pai	Fundamental	Ensino médio	Fundamental	Ensino médio
Escolaridade da Mãe	Fundamental	Ensino médio	Fundamental	Graduação
Renda Familiar	De 1,5 até 3 salários	De 4,5 até 6 salários	Até 1,5 salários	De 3 até 4,5 salários
Horas de Trabalho (Semanal)	40 horas ou mais	40 horas ou mais	Não trabalha	Não trabalha

Quadro 3 – Resultados de A1 (Fonte: Autores).

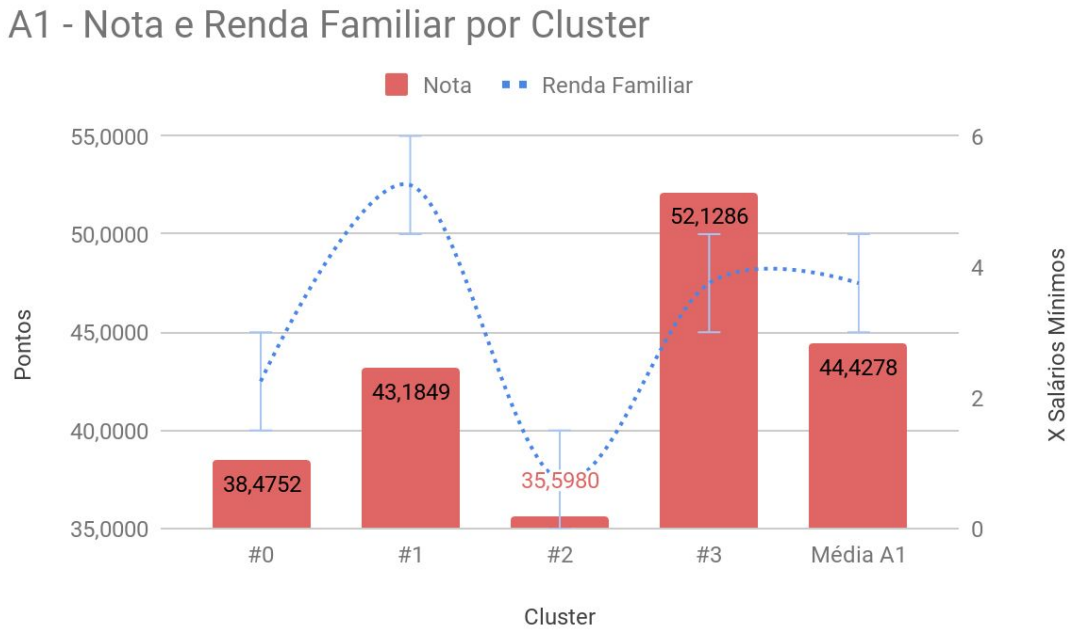

Figura 7 – Nota e renda familiar por Cluster (Fonte: Autores).

No segundo experimento que chama-se A2, os resultados foram diferentes. O maior grupo, cluster (#1), compreende 51% do total de alunos analisados. Na análise A1, o maior grupo, também o cluster (#1), representou 41% do total. Entretanto a nota em A2 é ligeiramente maior, 47.12, quando comparado com A1, 43.18. Aquela também é a maior nota entre os três *clusters* gerados na análise A2. Curiosamente, as demais características socioeconômicas foram idênticas nestes dois clusters (#1), tanto em A1 como em A2. A diferença entre as notas dos clusters (#0) e (#2) foi menor que 1 ponto, 42.41 e 41.06 respectivamente. Na análise A1, houve uma melhor distribuição das notas entre os grupos.

Atributo	**Cluster #0**	**Cluster #1**	**Cluster #2**
Nota (de 0 até 100)	42,4094	47,1177	41,0565
Estado Civil	Solteiro	Solteiro	Solteiro
Escolaridade do Pai	Fundamental	Ensino médio	Fundamental
Escolaridade da Mãe	Graduação	Ensino médio	Fundamental
Renda Familiar	De 1,5 até 3 salários	De 4,5 até 6 salários	De 3 até 4,5 salários

Horas de Trabalho (Semanal)	40 horas ou mais	40 horas ou mais	Não trabalha

Quadro 4 – Resultados de A2 (Fonte: Autores).

Na primeira clusterização, A1, a escolaridade do pai foi o ensino fundamental para os dois grupos que apresentaram as menores notas, clusters (#0) e (#2). Nos clusters (#1) e (#3), de maior nota, a escolaridade do pai também foi melhor, sendo esta o ensino médio. Em A2 o resultado foi semelhante. Para as duas menores notas, clusters (#0) e (#2) a escolaridade do pai também foi o ensino fundamental. E para o único grupo de maior nota, cluster (#1), a escolaridade também o ensino médio. Portanto não houve alteração na escolaridade do pai nas análises A1 e A2.

Para a escolaridade da mãe os resultados foram distintos. Em A1 os dois clusters de menor nota, também resultam em uma menor escolaridade da mãe, sendo esta o ensino fundamental. Semelhante a escolaridade do pai, em A1, os dois grupos de maior nota também resultam e uma melhor escolaridade da mãe. Para o cluster (#1) tem-se o ensino médio, e para o cluster (#2), o de maior nota entre ambos, a escolaridade é a graduação.

Entretanto em A2 a maior escolaridade da mãe, a pós-graduação, não é encontrada no cluster de maior nota, (#1), e sim no grupo com a segunda menor nota, cluster (#0). Apesar de a escolaridade da mãe ser o ensino fundamental no grupo de menor nota (#2), a diferença de nota entre o cluster (#0), onde a escolaridade é a pós-graduação, e o cluster (#2), foi menor que um ponto. Ou seja, a maior escolaridade da mãe, não resulta em uma nota melhor para o aluno em A2. A renda familiar em A2 mostrou-se proporcional a nota, ou seja, quanto maior a renda, melhor a nota do aluno na prova. A mesma relação não é observada em A1.

A última mineração dos dados realizada, relaciona o resultado obtido com as cidades brasileiras. O objetivo deste experimento por meio da localização geográfica de cada instituição é verificar se há uma divisão perceptível entre as regiões do Brasil. Utilizou-se todos os campos de nota da base: nota bruta da formação geral por curso (1), nota padronizada da formação geral por curso (2), nota bruta por componente específico por curso (3), nota padronizada do componente específico por curso (4), nota geral do componente específico por curso (5), nota bruta geral por instituição por curso (6), nota padronizada geral por instituição por curso (7), nota objetiva da formação geral (8), nota discursiva formação geral (9), nota discursiva formação específica (10) e nota formação específica (11). Fez-se uma média simples para cada cidade distinta, mesmo que seja a mesma Universidade a ofertar o curso.

Criou-se os códigos em Python com a biblioteca de clusterização da Scikit-Learn que tem implementado o algoritmo k-means e também a biblioteca Folium para interpretação dos resultados gerados em mapa.

Realizou-se tentativas de clusterização entre 2 e 5 clusters e com diferentes, porém não obteve-se resultados satisfatórios que pudessem determinar uma relação entre as notas dos alunos sumarizadas por instituição de ensino e sua localização no

território nacional. A distribuição dos clusters em mapa do território nacional é exibida na figura 8.

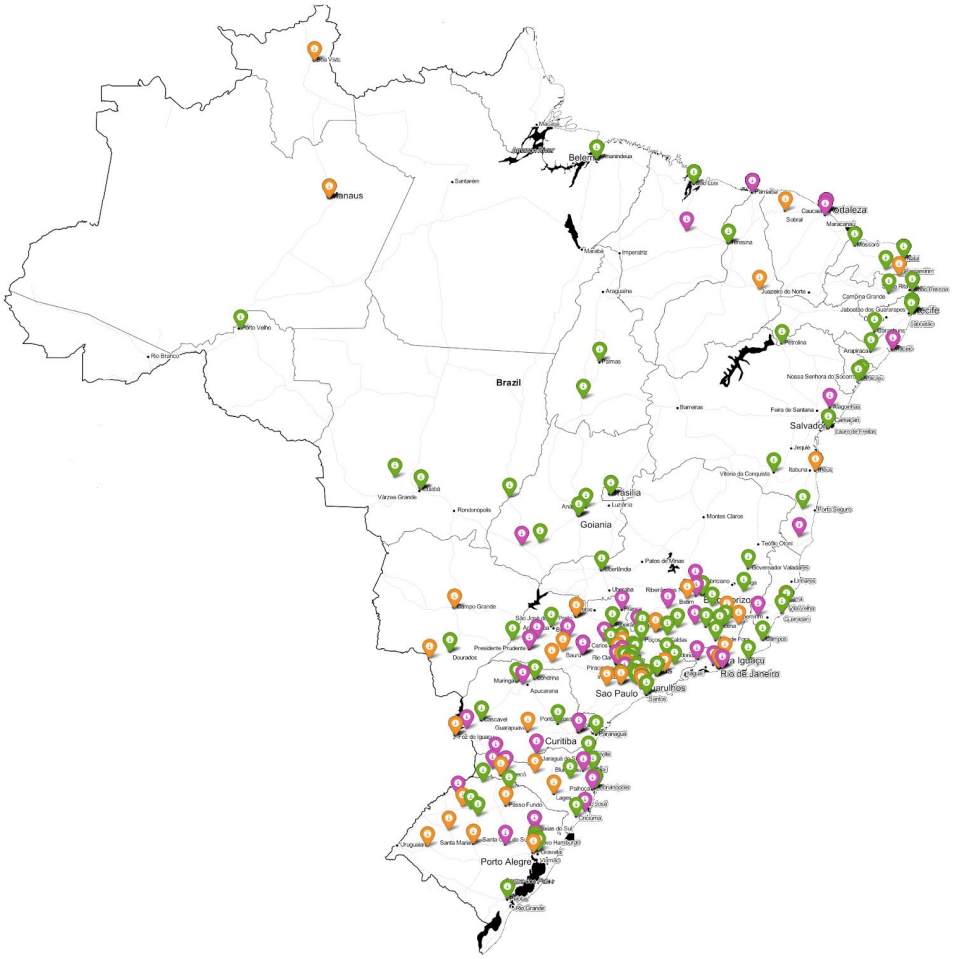

Figura 8 - Distribuição dos clusters (Fonte: Autores).

A figura 8 mostra o resultado da clusterização com 3 clusters, que foi validado como a melhor modelagem dos dados utilizando algoritmo k-means. A validação foi realizada de maneira manual nos dados e constado como a melhor apresentação de agrupamento das instituições. Os tokens laranja são as instituições com os melhores resultados e o verde são as aquelas com os piores resultados de seus alunos.

6. Conclusão

O principal objetivo deste trabalho foi relacionar os atributos dos alunos com a nota final obtida no ENADE se valendo da técnica do KDD de descoberta de conhecimento e utilizando o K-means como algoritmo de mineração de dados. Em um primeiro momento, o foco foi o questionário sócio-econômico de cada aluno e tentar clusterizar os grupos, estabelecendo a nota média de cada grupo. Iniciou-se a execução do trabalho, ao mostrar as diferenças de duas

clusterizações com os mesmos dados, porém configurado para gerar menos clusters em seus resultados. Vê-se uma diferença bastante grande de resultado que pode levar a interpretações enganosas.

No primeiro experimento (A1), percebeu-se indícios que não há uma relação entre renda da família do estudante e seu desempenho no exame, porém no segundo experimento (A2) foi possível ter essa conclusão. Fazendo validações simples dos dados gerados, viu-se que em A1 o agrupamento foi mais assertivo e tem-se uma melhor distribuição dos alunos, ajudando na conclusão de que 4 *clusters* é a quantidade ideal para o agrupamento para os dados de entrada selecionados.

No segundo experimento, utilizou-se os resultados de cada aluno em cada segmento da prova, inclusive a nota final para fazer os agrupamentos. Utilizou-se cada nota de cada instituição para plotar no mapa o agrupamento realizado e ver isso distribuído no território nacional. Novamente, a mineração realizada nos dados direciona à uma conclusão diferente do senso comum sobre o rendimento dos alunos de acordo com a região que ele estuda. Se isso fosse verdade, seria possível observar uma divisão mais clara dos clusters, mas as cores ficaram bastante homogêneas na imagem para obter essa conclusão.

Portanto, sobre a ótica do KDD há a indicação da existência de um descolamento do que possa acreditar ser uma verdade sem analisar com minúcia os dados. Este projeto então apresenta uma possível desconstrução de falácias relacionadas ao desempenho e desenvolvimento de alunos de computação nas instituições de ensino superior do país.

7. Referências

AMO, S. Técnicas de mineração de dados. 2014. Disponível em: http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf. Acesso em: 08 mar. 2020.

BRASIL. Lei nº 10.861, de 14 de abril de 2004. Institui o Sistema Nacional de Avaliação da Educação Superior – SINAES e dá outras providências. Brasília, DF: Presidência da República. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm. Acesso em: 10 mar. 2020.

BRASIL. Instituto Nacional de Estudos e Pesquisa Educacionais Anísio Teixeira. Microdados. 2015. Disponível em: http://portal.inep.gov.br/microdados. Acesso em: 07 nov. 2018.

CASTANHEIRA, L. G. Aplicações de técnicas de mineração de dados em problemas de classificação de padrões. 2008. 95 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica, UFMG, Belo Horizonte, 2008. Disponível em: https://www.ppgee.ufmg.br/defesas/349M.PDF. Acesso em: 10 nov. 2018.

COSTA, E. *et al*. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *In*: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO, 1., 2012, Rio de Janeiro. **Anais eletrônicos...** Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2012. p. 1-29. Disponível em: https://www.br-ie.org/pub/index.php/pie/article/view/2341/2096. Acesso em: 15 jan. 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. FROM DATA MINING TO KNOWLEDGE DISCOVERY IN DATABASES. AI Magazine, v. 17, n. 3, p. 37-54, 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131. Acesso em: 08 mar. 2020.

FONSECA, S. O.; NAMEN, A. A. Mineração em bases de dados do Inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educ. rev., Belo Horizonte, v. 32, n. 1, p. 133-157, jan./mar. 2016. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982016000100133&lng=pt&nrm=iso. Acesso em: 20 nov. 2019.

GARCIA, A. C. MINERAÇÃO DE DADOS APLICADA A SISTEMAS DE RECOMENDAÇÃO. 2012. TCC (GRADUAÇÃO) - Universidade de Santa Cruz do Sul, Santa Cruz do Sul, 2012.

GOLDSCHMIDT, R.; PASSOS, E. Data mining: um guia prático. Rio de Janeiro: Campus, 2005.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3. ed. Waltham: Elsevier, 2012. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=data+mining+articles&ots=tzHr1TmE-3&sig=XUOSPAgnNh4uTLm0J1346cUHZVo#v=onepage&q=processo&f=false. Acesso em: 08 mar. 2020.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: PROCEEDINGS OF THE FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Berkeley. Califórnia. **Anais eletrônicos...** Berkeley: University of California, 1967. p. 281-297. Disponível em: https://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512992. Acesso em: 16 out. 2019.

SILVA, J. L. D.; NUNES, I. D. Mineração de dados educacionais como apoio a classificação de alunos do ensino médio. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 26., 2015, Maceió. Anais eletrônicos... Disponível em: http://www.br-ie.org/pub/index.php/sbie/article/view/5430/3789. Acesso em: 20 nov. 2019.

SILVA, R. E. D. *et al.* Mineração de dados educacionais na análise das interações dos alunos em um ambiente virtual de aprendizagem. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 26., 2015, Maceió. **Anais eletrônicos...** Disponível em: http://www.br-ie.org/pub/index.php/sbie/article/view/5448/3807. Acesso em: 20 maio 2020.

TAN, P. N.; STEINBACH, M.; KUMAR, V. Introdução ao data mining: mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS. The workbench for machine learning. [2019]. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 15 nov. 2019.

WITTEN, I. H. *et al.* Data mining: practical machine learning tools and techniques. 4. ed. Cambridge: Elsevier, 2017. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=data+mining+articles&ots=8IDJwdlBxc&sig=yCa6mkhOQyBF-n-xkOufXzE-bFU#v=onepage&q=data%20mining%20articles&f=false. Acesso em: 12 mar. 2020.