Mineração de Dados Abertos - ENEM 2018

Alessandro Aparecido Barcellos, Prof. Dr. Seiji Isotani, Carlos Diego Nascimento Damasceno

Abstract

The National High School Exam (ENEM) is a standardized Brazilian exam taken by students to enter public and private universities. In 2018 alone, ENEM had approximately 5.5 million candidates with different socioeconomic profiles. Given this database and the information it contains, the process of analysis and discovery of knowledge becomes quite challenging and its automation essential. The origin base, for this study, is made up of 5.5 million records with an approximate size of 3.4GB. After phases of data selection, pre-processing and transformation, our final file included 950 thousand records and approximately 20MB in size. It is worth mentioning that the majority of students (about 76.19%) have access to the internet. Regarding parents' education, only 13.75% attended higher education. The difference, 86.25% is divided between parents who attended elementary and high school. As a result of one of the scatter plots, students who have access to the internet performed more satisfactorily at ENEM (2018).

Resumo

O Exame Nacional do Ensino Médio (ENEM) é um exame padronizado brasileiro realizado por estudantes para ingressar em universidades públicas e privadas. Somente em 2018, o ENEM teve aproximadamente 5,5 milhões de candidatos com diferentes perfis socioeconômicos. Diante desta base de dados e das informações nela contida, o processo de análise e descoberta de conhecimento torna-se bastante desafiador e sua automação essencial. A base origem, para este estudo, é composta por 5.5 milhões de registros com um tamanho aproximado de 3.4GB. Após fases de seleção dos dados, pré-processamento e transformação, nosso arquivo final contemplou 950 mil registros e aproximadamente 20MB de tamanho. Vale ressaltar que a maioria dos estudantes (cerca de 76,19%) tem acesso à internet. Com relação à escolaridade dos pais, apenas 13,75% cursaram o ensino superior. Sendo que a diferença, 86,25% divide-se entre os pais que cursaram ensino fundamental e médio. Conforme resultado de um dos gráficos de dispersão, os estudantes que possuem acesso à internet, desempenharam resultados mais satisfatórios no ENEM (2018).

Cite as: Barcellos, A. & Isotani, S & Damasceno, C. (2020). Mineração de Dados Abertos - ENEM 2018. Anais dos Trabalhos de Conclusão de Curso. Pós-Graduação em Computação Aplicada à Educação Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo.

1. Introdução

Desde os tempos pré-históricos, há registros de extração de minerais para geração de ligas metálicas, destinadas às mais diversas aplicações: ferramentas, panelas, caldeiras, mesas, portões, etc. No entanto, o termo "mineração" só apareceu no século XVI, significa um processo industrial cuja finalidade é extrair substâncias ou metais valiosos de grandes depósitos ou minerais. Hoje em dia, na era da transformação digital, vemos um conceito semelhante ao que apareceu pela primeira vez: diante de dados cada vez mais importantes em escala global, a mineração de dados tem se tornado personagem principal com informações cruciais para decisões corporativas e governamentais.

Enquanto na mineração de metais preciosos, se peneira a terra até encontrar o metal precioso, a mineração de dados é o termo que se popularizou para denominar o processo de descoberta de conhecimento em base de dados. Trata-se da utilização de ferramentas computacionais a fim de descobrir informações valiosas, potencialmente úteis, descritas na forma de padrões, a partir dos volumes de dados que estão sendo coletados e armazenados pelas organizações atualmente. A obtenção desses conhecimentos implícitos tem sido útil, sobretudo, para as empresas conhecerem melhor seu público-alvo e tomarem decisões mais acertadas ao objetivarem aumentar a competitividade [Fayyad 1996]. Segundo [Carvalho 2002], para ocorrer aprendizado sobre uma base de dados, uma série de informações de diferentes formatos e fontes precisa ser organizada de maneira consistente na grande memória empresarial (*Data Warehouse*). Após isto, métodos de análise estatística, Inteligência Artificial e de Aprendizado de Máquina precisam ser aplicados sobre esses dados e novas e úteis relações à empresa devem ser descobertas, ou seja, os dados devem ser minerados. Sobre o enfoque do autor, a mineração de dados consiste em descobrir relações entre produtos, classificar consumidores, prever vendas, localizar áreas geográficas potencialmente lucrativas, prever ocorrências ou inferir necessidades.

O processo de descoberta de conhecimento em banco de dados também é adotado no ambiente educacional, aqui nos concentramos na tecnologia de mineração de dados educacional (EDM). Tais técnicas ganham destaque no cenário educacional, proporcionando aos gestores educacionais, educadores e inclusive ao governo subsídios para tomada de decisões, além de perspectivas para mitigar oportunidades(desafios) habituais e não habituais como os impactos da pandemia (COVID-19). Neste estudo, foi utilizado o arquivo do ENEM (Exame Nacional do Ensino Médio), disponibilizado pelo INEP, datado de 2018. Usamos a tecnologia de mineração de dados de agrupamento, para descobrir o perfil dos estudantes, que fazem o exame e verificar se podem ou não impactar positivamente nas notas dos mesmos. Essas técnicas amparam na identificação do perfil desses estudantes a partir de dados de outros estudantes que não foram bem-sucedidos em determinados exames [Baker 2011].

Neste trabalho apresentaremos os fundamentos básico para o uso da Mineração de Dados Educacionais, em especial dados abertos. Na seção 2, apresentaremos a base teórica. Na seção 3, trabalhos relacionados. Na seção 4, apresentaremos o método usado. Na seção 5 apresentaremos a avaliação. A seção 6 apresentaremos as discussões do estudo de caso. Na seção 7, apresentaremos a conclusão.

2. Fundamentação Teórica

Processo KDD

KDD (*Knowledge Discovery in Database*) é um importante processo de identificação de padrões em um conjunto de dados, potencialmente úteis e compreensíveis de serem analisados pelo usuário. [Fayyad 1996]. Inclui os seguintes passos: conforme mostrado na Figura 2.1.

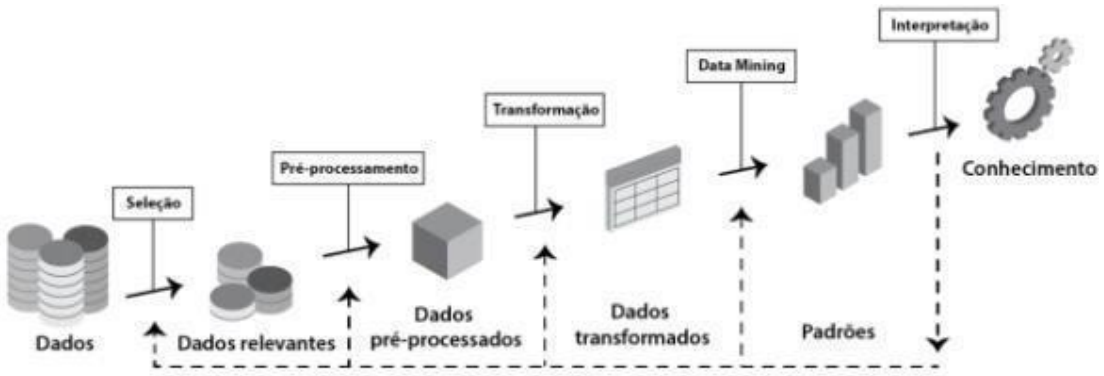

Figura 2.1. Passos do processo KDD – Adaptado de Fayyad.

Seleção de dados: consiste em selecionar um conjunto de dados ou mesmo destacar um subconjunto onde a higienização deve ser realizada;

Pré-processamento: esta fase é responsável pela remoção de ruídos, seleção de atributos relevantes, tratamento de campos ausentes e formatação dos dados;

Transformação: ocorre logo após a fase de pré-processamento, quando os dados são transcritos para um formato adequado;

Mineração de Dados: nesta fase alguns algoritmos são executados a fim de identificar padrões relevantes nos dados;

Interpretação e Avaliação: os padrões extraídos são interpretados e avaliados de acordo com o critério definido pelo usuário, a fim de obter novos padrões de informações encontrada;

Implantação do Conhecimento: após a validação dos padrões extraídos pelas etapas anteriores, o conhecimento pode ser utilizado pelo usuário.

Mineração de Dados

Em mineração de dados, segundo [Fayyad 1996], existem duas formas de descobrir padrões de dados: automático ou semiautomático.

Mencionaremos na Figura 2.2 abaixo, algumas tarefas relacionadas às respectivas metas para o modelo preditivo e descritivo. Na Figuura 2.3 detalharemos alguns exemplos destas tarefas, as quais fazem parte da mineração de dados [Alpaydin 2004].

Figura 2.2. Modelos preditivos e descritivos.

[adaptada/traduzida em Alpaydin 2004]

Tarefas	Descrição	Exemplos	Algoritmos
Classificação	É utilizada para separar os dados em classes definidas de acordo com a necessidade do usuário.	Classificar estudantes, classificar pacientes, classificar pedidos de crédito, classificar operações fraudulentas.	*J48, Id3, C45, ADTree, UserClassifier, PredictionNode, Splitter, ClassifierTree, M5Primer, Prism, Part, OneR.*
Regressão/Previsão	É uma função usada para prever o valor de uma variável desconhecida no seu modelo.	Prever o índice de evasão escolar, prever o valor de vida de um equipamento, prever o desempenho do aluno.	*Cubist, RETIS, M5, CART.*
Associação	Busca encontrar relação dentro de um conjunto de valores onde é possível identificar padrão de comportamento entre seus atributos.	Determinar quais disciplinas o aluno tem desempenho semelhante, determinar quais clientes compram dois produtos distintos.	*Apriori, FPGrowth, PredictiveApriori, Tertius.*
Agrupamento/Clustering	Seu objetivo é dividir um conjunto de dados em subconjuntos que apresentem alguma característica similar.	Identificar grupos de escolas (para investigar as diferenças e similaridades entre escolas), achar grupos de alunos (para investigar as diferenças e similaridades entre alunos).	*K-Means, Cobweb, EM, X-Means.*
Sumarização	Consiste em encontrar uma descrição mais simples para um conjunto de dados menor do que o seu conjunto de dados original.	Ofertar para potenciais clientes um novo modelo de SUV, sumarizando o perfil típico de famílias com 3 ou mais filhos, acima de 40 anos nível superior.	*Algoritmo de Luhn, Algoritmo de GistSumm.*

Figura 2.3. Detalhamento das tarefas.

[Evandro Costa 2012, Daniel Gomes Dosualdo 2003]

Ferramentas de Mineração de Dados

No mercado, atualmente, existem diversas ferramentas desenvolvidas, para apoiar a mineração de dados. Mencionaremos algumas ferramentas com foco nesta tarefa.

Weka

Weka é um pacote desenvolvido pela *Waikato University* em 1993, possui vários algoritmos para as tarefas de mineração de dados [*Holmes 1994*].

O software é distribuído sob a Licença Pública Geral, portanto seu código-fonte pode ser alterado. Weka é escrito em linguagem Java, e tem como foco ser uma ferramenta simples de utilizar, podendo ser usada por não especialista em mineração de dados. Possui uma série de heurísticas para serem usadas com grande volume de dados relacionadas a regras de classificação, regressão, agrupamento, associação e visualização, incluindo: *BayesNet*, *SimpleLinearRegression*, *IBk*, *Stacking*, *MultiScheme*, *ZeroR*, *RandomTree*, *Apriori*, *SimpleKMeans*. Em comparação às outras ferramentas, como as mencionadas abaixo, possui uma baixa curva de aprendizagem, passando a impressão de que a mineração de dados é algo simples. A difícil implementação de novas bibliotecas, é uma de suas principais desvantagens, exigindo do usuário conhecimentos avançados em JAVA [Suporte ao Software Livre 2020].

A Figura 2.4 ilustra a execução de um algoritmo de agrupamento realizada através do WEKA.

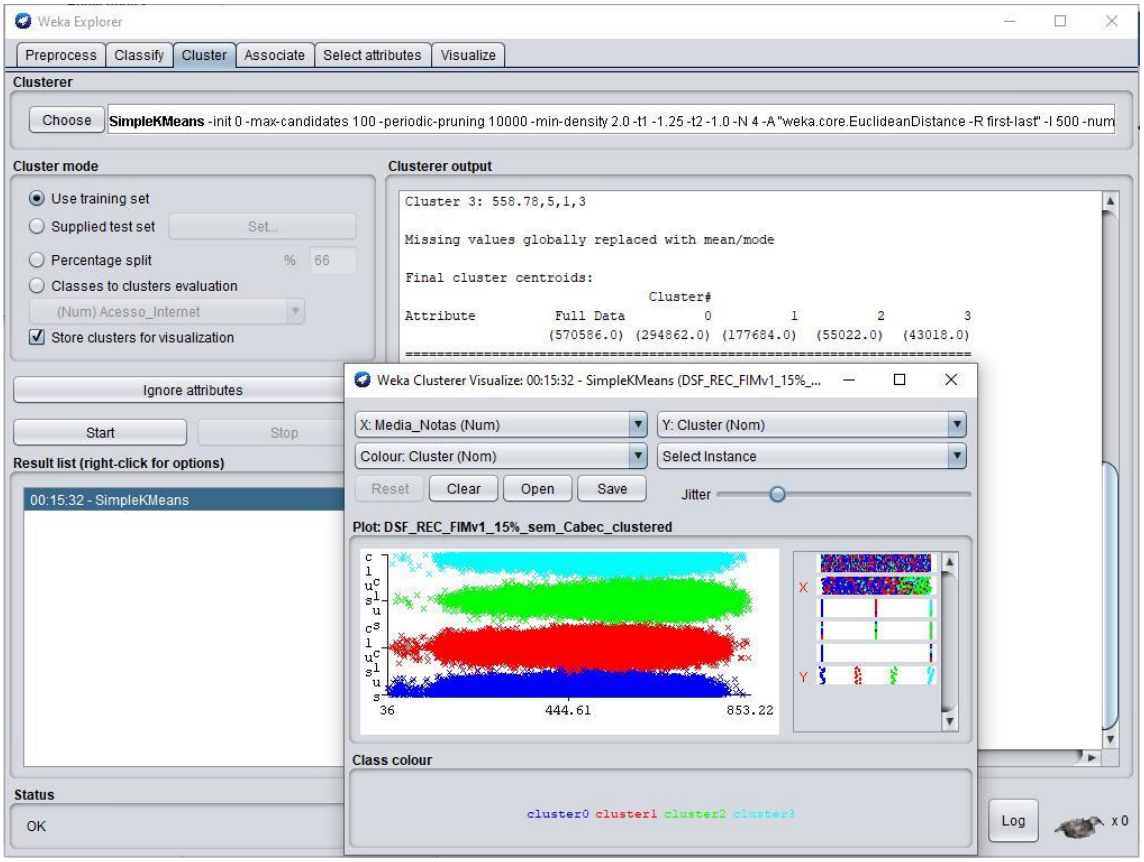

Figura 2.4. Execução do algoritmo de agrupamento e visualização do gráfico de dispersão.

R

Com a vantagem de ser software livre, R é uma linguagem orientada a objetos criada por *Ross Ihaka* e *Robert Gentleman* em 1996. Combinada com um ambiente integrado, pode realizar cálculos, manuseio de dados e construção de gráficos [*The R Project 2020*].

Possui uma grande variedade de recursos estatístico-computacionais tendo como uma vantagem a sua difusão. Uma curva de aprendizagem média, quanto a absorção de conhecimentos, relacionados a sua linguagem de programação.

RStudio é uma IDE (*Integrated Development Environment*) de código aberto implementada em C++/Qt dispondo de uma interface que facilita seu uso [*RStudio 2020*].

R é essencialmente uma linguagem de programação funcional, o que a torna menos amigável que o Weka.

A Figura 2.5 ilustra a interface do *RStudio*.

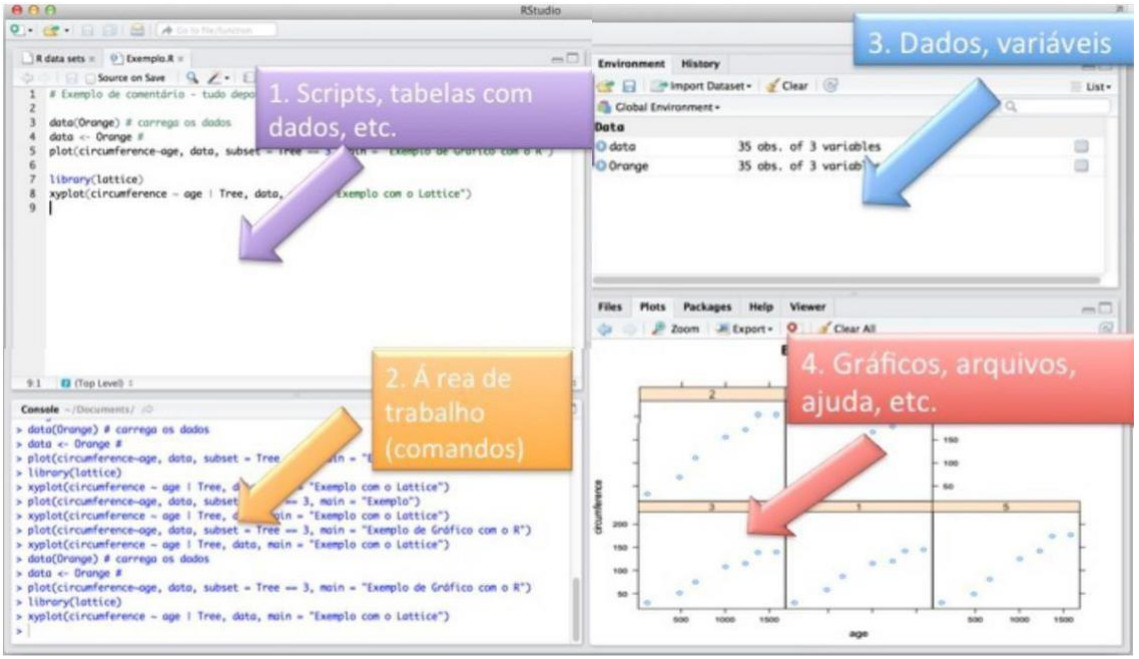

Figura 2.5. Interface *Rstudio* contemplando, gráficos, tabelas e área de Trabalho.

[Ciência Prática 2015]

RapidMiner

RapidMiner é um pacote de mineração de dados, que possui versões gratuitas e pagas, escrito em Java pela empresa homônima. Inclui análise preditiva, aprendizado de máquina e mineração de texto, numa plataforma integrada. Possui uma baixa curva de aprendizagem, além de produzir gráficos de qualidade. Aplicações com diversas finalidades como por exemplo: pesquisa, educação, treinamentos, comerciais, negócios e aprendizado de máquina. Possui como principal desvantagem que algumas funcionalidades estão disponíveis somente nas versões pagas [*Sourceforge 2015*].

A figura 2.6 ilustra a execução de um algoritmo de agrupamento realizada através do *RapidMiner*.

A figura 2.7 ilustra a plotagem do gráfico de dispersão de Média de Notas x Renda Família e seus respectivos agrupamentos.

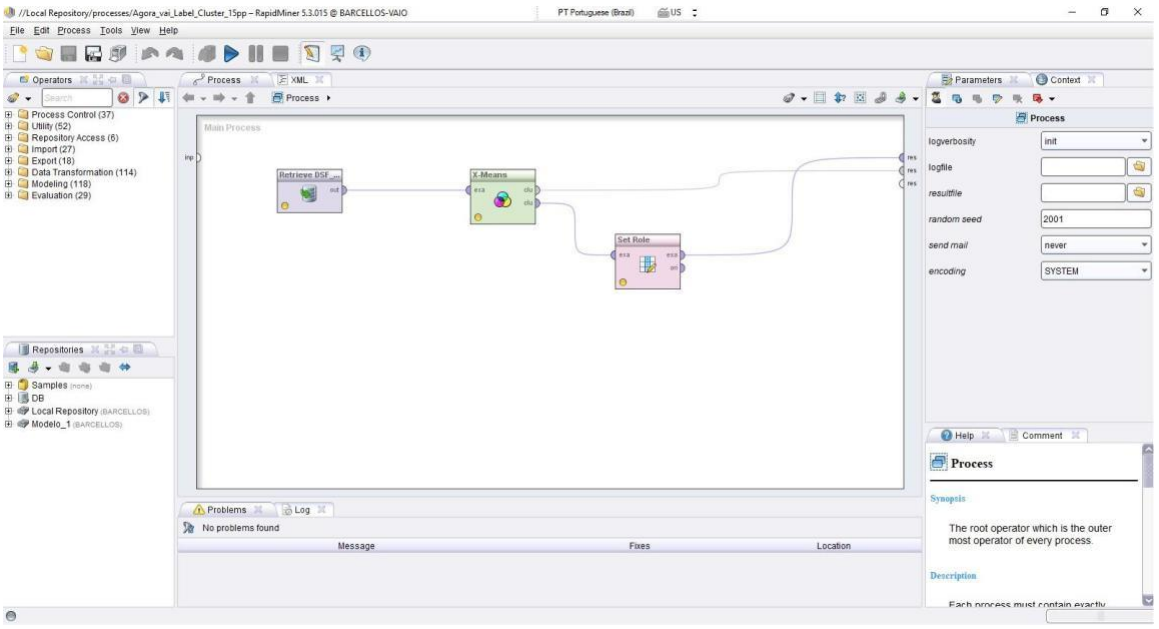

Figura 2.6. Processo contemplando, leitura de arquivo, execução do algoritmo de agrupamento.

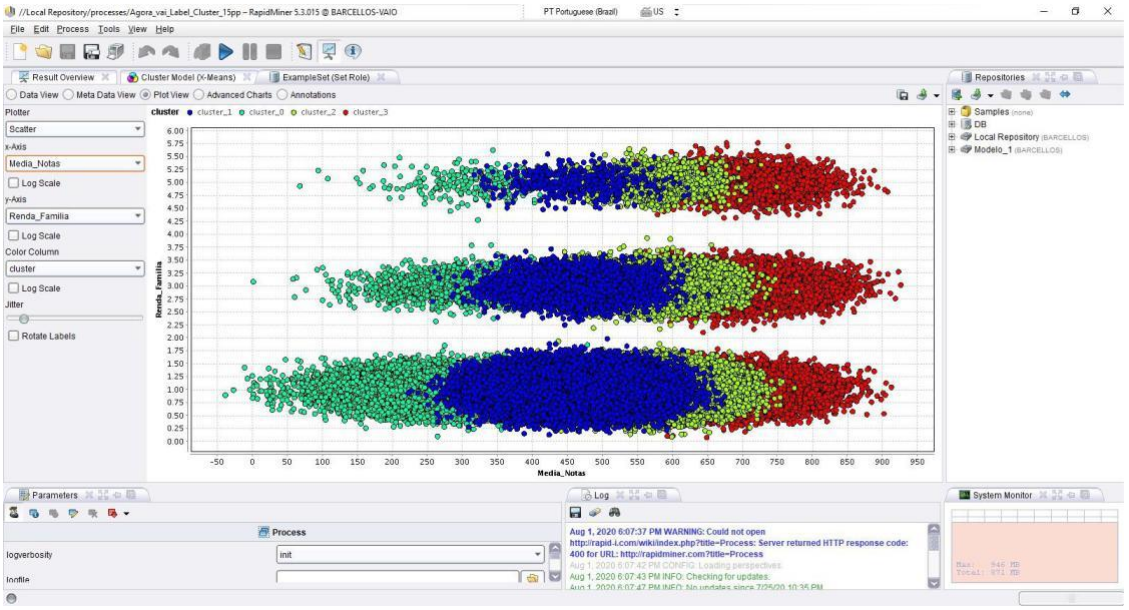

Figura 2.7. Plotagem do gráfico de dispersão.

ORANGE

É uma ferramenta, que trabalha com aprendizado de máquina e visualização de dados, por meio de seu *software* de código aberto, para conduzir o trabalho de mineração de dados de uma forma interessante e proveitosa. Tanto usuários novatos quanto experientes podem usá-lo. Dispõe de uma sequência de passos baseado num fluxo de trabalho interativo e uma ampla gama de ferramentas, incluindo várias técnicas como modelagem de dados, pré-processamento, exploração e visualização. Possui uma área de interação atrativa e intuitiva, pode ser usado como um módulo da linguagem *Python*, por usuários mais avançados. Oferece uma variedade de meios com foco em estatística para o processamento das informações, atingindo objetivos como identificar padrões de similaridade ou de não conformidade nos conjuntos de informações [*LARHUD 2018*].

Data Sampler, tem a função de gerar uma amostra aleatória dos dados, e nos permite definir qual o percentual de dados para geração da nova amostra. É uma saída, quando não temos poder de processamento para base de dados origem.

A figura 2.8 ilustra a execução de alguns *Data Sampler's* gerando uma amostra de 7%, 15%, 25%, 30%, 35% e 40% da Base de Dados original realizada pelo *Orange*.

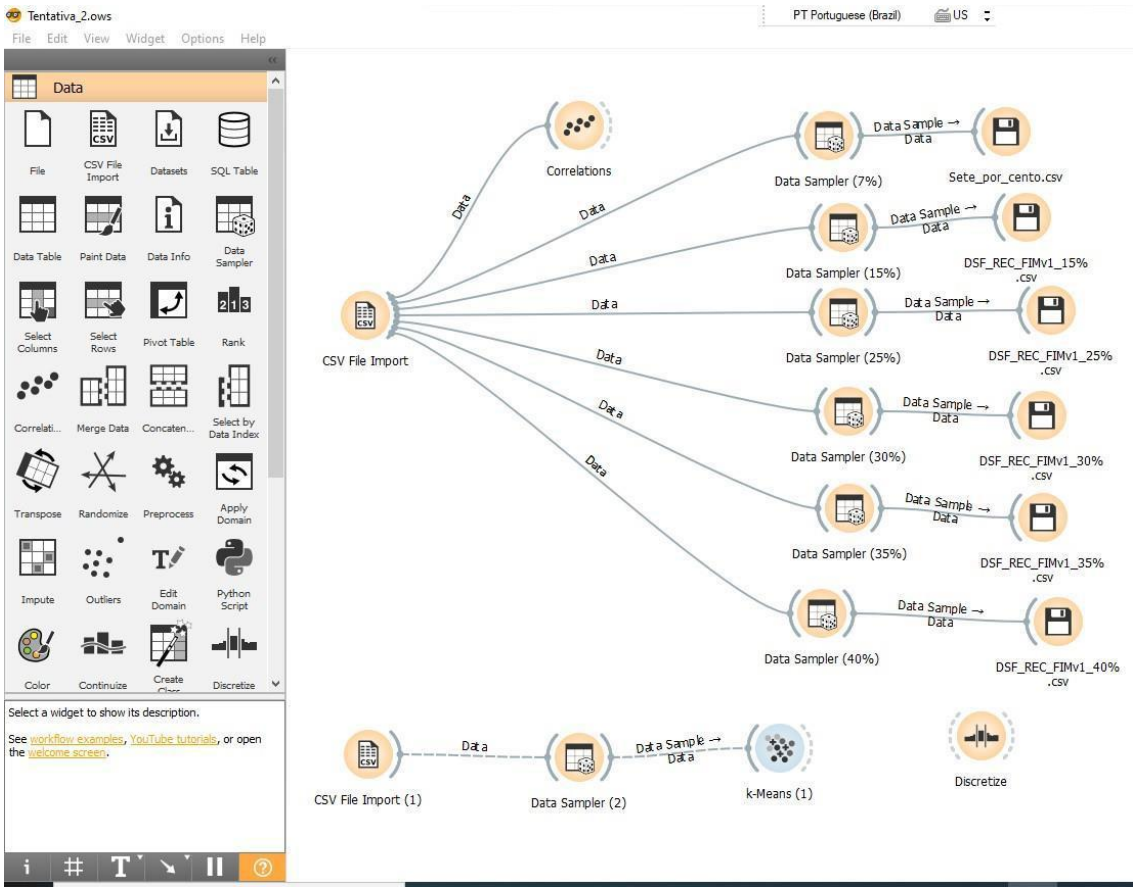

Figura 2.8. Gerando amostras de dados com *Data Sampler*.

3. Trabalhos Relacionados

Apresentaremos um levantamento dos trabalhos correlatos pesquisados e selecionados que fazem o uso da descoberta do conhecimento no contexto educacional, mostrando as técnicas de mineração que são aplicadas e traz algumas abordagens que se relacionam com o objetivo desta proposta de trabalho.

Romero (2008) tem o objetivo de criar um instrumento de Mineração de Dados Educacionais incorporado no ambiente virtual de aprendizagem Moodle. Comparam diferentes técnicas de mineração de dados para classificar os alunos com base em suas informações registradas no Moodle durante sua interação e a nota final conquistada no curso. Foram utilizados dados reais de sete cursos de engenharia, de estudantes da Universidade de Córdoba, usuários do Moodle.

Prass (2004) apresentou um estudo que confronta os principais algoritmos de análise de agrupamento presentes na literatura e também em aplicativos, tendo como objetivo o seu uso no processo de descoberta de conhecimentos em bancos de dados. O autor destacou que os algoritmos são diferenciados de acordo com o seu método de formação (baseado em grade, baseado em densidade, baseado em modelo, partição, hierárquico) e também pela medida de distância que expressa a similaridade ou dissimilaridade entre os objetos. A conclusão que se pode tirar desta pesquisa é que a análise de agrupamento é uma técnica de mineração de dados eficaz, mas seu uso requer conhecimento teórico e domínio da base de dados.

Faria (2014) com base na análise de agrupamento, gráficos de controle e métodos de regressão logística, foi proposto um modelo de análise de dados. A finalidade deste trabalho é obter conhecimento relacionado à previsão do desempenho escolar e ajudar o docente de educação *online* a monitorar de forma eficaz as atividades e atuação dos alunos. No processo de ensino e aprendizagem, a qualquer momento o docente, pode analisar o sucesso do aluno e a perspectiva de fracasso, trazendo-lhes menos erros e reduzindo o número de eventos de alunos afastados da escola.

França e Amaral (2013) focam no desempenho, introduzindo o uso da metodologia de agrupamento para fins de mineração de dados, tendo como objetivo formar uma turma de alunos, que apresentam dificuldades similares de aprendizagem no ensino de programação. Fundamentado nisso, espera-se que seja possível desenvolver estratégias de ensino apropriadas a turma de alunos para melhorar seu desempenho.

4. Metodologia

Esta seção apresenta, uma base de dados do Exame Nacional do Ensino Médio (ENEM) 2018, disponibilizada pelo Instituto Nacional de Pesquisas Educacionais Anísio Teixeira-INEP, utilizada para este estudo de caso.

O propósito deste estudo de caso consiste em aplicar a mineração de dados para averiguar o perfil dos alunos. Para atingir o alvo, são utilizadas técnicas de agrupamento. Um dos objetivos é saber quais fatores impactam nas notas finais dos alunos participantes do ENEM (2018), ou quais resultados podem responder a tais questões:

- Alunos cujos pais possuem ensino superior, têm possibilidades de obterem melhores notas no ENEM?
- Quanto a renda da família, os alunos categorizados como alta, tiram melhores notas?
- Os alunos categorizados como baixa renda e que não possuem acesso à internet não tiram boas notas no ENEM?
- Quais elementos terão efeitos significativos no desempenho do aluno?

A referida base de dados do ENEM abrange informações socioeconômicas, notas obtidas nas avaliações, inclusive redação, cidade de origem, escolas onde estudaram, entre outras. Essas informações foram preenchidas pelos candidatos no momento da inscrição no ENEM 2018.

Esse arquivo consiste em aproximadamente 5.500.000 (cinco milhões e quinhentos mil) registros, disponibilizado em formato de texto (TXT), com cerca de 3,4 GB. No entanto, não tínhamos a capacidade necessária para tal processamento das informações, contudo, eliminamos por volta de 1.300.000 (1,3 milhão) registros, referentes aos estudantes que zeraram na somatória das provas. Excluímos também aproximadamente 350.000 (trezentos e cinquenta mil) registros, referente a resposta "Não sei" para formação escolar do Pai. Ficamos com aproximadamente 3.800.000 (três milhões e oitocentos mil) registros, dos quais, através da função *Data Sampler* do *Orange*, foram selecionados uma amostra de 25% destes registros para a execução do algoritmo de agrupamento.

O arquivo de dados do ENEM (2018) contém 137 colunas, 8 destas, achamos mais importante para esse estudo, cujo 5 representam as notas (Redação, Linguagens e Códigos, Ciências da Natureza, Ciências Humanas e Matemática), que foram utilizadas para gerar a média do estudante. Resultando um total de quatro colunas para este estudo, sendo elas: (acesso à internet, formação escolar do Pai, renda familiar, média das notas). Na Tabela 4.1 apresentamos as colunas utilizadas, assim como a disposição dos resultados apontados pelos estudantes.

		Base Completa 3.803.906 registros		Base Parcial (25%) 950.977 registros	
Questão	Resposta	Qtde.	%	Qtde.	%
Em sua casa, tem acesso a internet?	Não	906,168	23,82%	226,410	23,81%
	Sim	2.897.738	76,18%	724,567	76,19%
Até quando seu Pai estudou?	Ensino Fundamental	1.686.247	44,33%	421,922	44,37%
	Ensino Médio	1.595.752	41,95%	398,263	41,88%
	Ensino Superior	521,907	13,72%	130,792	13,75%
Quanto aproximadamente é a renda mensal da família?	Até 2 salários mínimos - até R$ 1.908,00	2.296.579	60,38%	573,658	60,32%
	Entre 2 e 12 salários mínimos de R$ 1.908,01 até 11.448,00	1.384.373	36,39%	346,352	36,42%
	Acima de 12 salários mínimos - acima de R$ 11.448,01	122,954	3,23%	30,967	3,26%
Nota (Média das Notas Redação, Linguagens e Códigos, Ciências da Natureza, Ciências Humanas e Matemática)	< 500 pontos	1.647.272	43,30%	412,158	43,34%
	>=500 e <700 pontos	2.028.999	53,34%	507,116	53,33%
	>= 700 pontos	127,635	3,36%	31,703	3,33%

Figura 4.1. Perguntas selecionadas, com representatividade das respostas.

Investigando as informações, podemos verificar, que grande parte dos alunos (cerca de 76,19%) tem acesso à internet. No que tange a formação do pai, apenas 13,75% cursaram o ensino superior. O complemento, 86,25% divide-se entre os pais que cursaram ensino fundamental e médio, praticamente o mesmo percentual. No que se refere a renda familiar, um maior número de família de estudantes, por volta de 60,32%, mantém-se com menos de 2 salários mínimos. No tocante a nota, podemos evidenciar que a maior parte dos alunos atingiram médias inferiores a 699 pontos, e que somente 3,33% atingiram médias maiores ou iguais a 700 pontos, no exame. O registro de menor nota(média) nas bases completa e parcial é de 36 pontos. Enquanto que o registro de maior nota(média) na base completa é 858,18 pontos e na base parcial (amostra) é 853,22 pontos.

De acordo com o método mostrado na Figura 2.1, a primeira etapa é usar a linguagem de programação *python*, para escolher entre mais de 5.500.000 (cinco milhões e quinhentos mil) registros, 3.800.000 (três milhões e oitocentos mil) registros. Em virtude de não possuirmos poder de processamento para a base de dados origem, outra etapa deste primeiro passo, foi utilizar o *software Orange* e seu "*widget* (componente)" *Data Sampler* para retirar uma amostra de 25%, resultando em 950.977 (novecentos e cinquenta mil e novecentos e setenta e sete) registros, utilizados para este estudo de caso.

Apresentamos na Figura 4.2, as 3 questões e a média (Nota) dos estudantes, retiramos dos registros selecionados na etapa anterior, dentre os 137, existentes no formulário socioeconômico e dados da prova objetiva, para tal, foi necessário desenvolver um código em *python* para executar essa triagem. A rotina criada selecionava apenas os registros necessários no arquivo original de (3.4 GB) e adicionava tais informações em um *DataFrame* específico (*DataFrame* é uma estrutura de dados bidimensional alinhados em formato de tabela por linhas e colunas, podendo sofrer alterações dinâmicas e potencialmente heterogêneo, semelhante às pastas de trabalho MS-EXCEL), no final do processo tínhamos um novo arquivo de 60 MB. Com a seleção de 25% de uma amostra dos dados no *Orange* finalizamos com um arquivo de 20 MB.

No estágio de pré-processamento, como o algoritmo que utilizamos *K-means*, não é capaz de manipular atributos qualitativos, é necessário converter os atributos qualitativos das variáveis (acesso à internet, formação escolar do Pai, renda familiar) em atributos quantitativos. Na sequência foram discretizados os dados das variáveis (formação escolar do Pai e renda familiar). Desenvolvemos em *python*, processos que transformam atributos qualitativos em quantitativos ao conteúdo das respostas do formulário socioeconômico. Na Figura 4.2 estão representados os resultados deste estágio.

		Após o pré-processamento	
Questão	**Atributos Originais**	**Transformação de atributos**	**Discretização de atributos**
Em sua casa, tem acesso a internet?	(A) Não	A=1	Não se aplica
	(B) Sim	B=3	
Até quando seu Pai estudou?	(A) Nunca estudou.	Não se aplica	(A,B e C) = 1
	(B) Não completou a 4ª série/5° ano do Ensino Fundamental.		
	(C) Completou a 4ª série/5° ano, mas não completou a 8ª série/9° ano do Ensino Fundamental.		
	(D) Completou a 8ª série/9° ano do Ensino Fundamental, mas não completou o Ensino Médio.		(D e E) = 3
	(E) Completou o Ensino Médio, mas não completou a Faculdade.		
	(F) Completou a Faculdade, mas não completou a Pós-graduação.		(F e G) = 5
	(G) Completou a Pós-graduação.		
	(H) Não sei.		(H) Eliminados da Base
Quanto aproximadamente é a renda mensal da família?	(A) Nenhuma renda	Não se aplica	(A,B,C e D) = 1
	(B) Até R$ 954,00.		
	(C) De R$ 954,01 até R$ 1.431,00.		
	(D) De R$ 1.431,01 até R$ 1.908,00.		
	(E) De R$ 1.908,01 até R$ 2.385,00.		(E,F,G,H,I,J,K,L,M e N) = 3
	(F) De R$ 2.385,01 até R$ 2.862,00.		
	(G) De R$ 2.862,01 até R$ 3.816,00.		
	(H) De R$ 3.816,01 até R$ 4.770,00.		
	(I) De R$ 4.770,01 até R$ 5.724,00.		
	(J) De R$ 5.724,01 até R$ 6.678,00.		
	(K) De R$ 6.678,01 até R$ 7.632,00.		
	(L) De R$ 7.632,01 até R$ 8.586,00.		
	(M) De R$ 8.586,01 até R$ 9.540,00.		
	(N) De R$ 9.540,01 até R$ 11.448,00.		
	(O) De R$ 11.448,01 até R$ 14.310,00.		(O,P e Q) = 5
	(P) De R$ 14.310,01 até R$ 19.080,00.		
	(Q) Mais de R$ 19.080,00.		
Nota (Média)	0 - 1000	Não se aplica	Não se aplica

Figura 4.2. Transformação e discretização de atributos.

5. Avaliação

Nesta fase, utilizamos o aplicativo *RapidMiner*, que permite uma fácil interação com o usuário. Ao finalizar o processo de tratamento dos dados no *python*, importamos o arquivo final do tipo csv para o *Orange*, gerando uma amostra com 25% dos dados, o qual foi importado para o *RapidMiner* e executado o algoritmo *K-means*.

Arquivo tipo CSV, das iniciais em inglês de Comma-Separated Values, traduzindo "valores separados por vírgulas". Trata-se de uma forma compacta de representar dados em tabelas. Tem-se um arquivo texto com vírgulas separando os valores das colunas.

K-means, é um algoritmo de agrupamento de dados não-hierárquico, que avalia e agrupa os dados de acordo com suas características. Este algoritmo, busca minimizar a distância dos elementos, em conjunto de dados com k centros de forma iterativa.

6. Discussão

Ao executarmos o algoritmo de agrupamento, resultou na alocação dos dados em diferentes grupos, exatamente quatro (*cluster* 0, *cluster* 1, *cluster* 2 e *cluster* 3), conforme as características presentes nas informações fornecidas pelos estudantes, tais como: escolaridade do pai, rendimento familiar e se possui internet em casa. Os grupos identificados são exibidos na Figura 6.1.

	Formação Escola Pai	Renda Familiar	Acesso a Internet	Média	Qtde.	%
cluster 0	Ensino Fundamental	Classe Baixa	Não possui Internet	491	176.996	18,61%
cluster 1	Ensino Fundamental	Classe Baixa	Possui Internet	530	280.313	29,48%
cluster 2	Ensino Superior	Classe Média	Possui Internet	475	146.953	15,45%
cluster 3	Ensino Médio	Classe Baixa	Possui Internet	533	346.715	36,46%
					950977	

Figura 6.1. *Clusters* identificados.

Ao gerarmos os gráficos, os dados representados no eixo horizontal secundário, foram rotulados conforme figura 6.2.

Cores			
Eixo horizontal secundário	**1**	**3**	**5**
Renda Familiar	Classe Baixa	Classe Média	Classe Alta
Formação Escola Pai	Ensino Fundamental	Ensino Médio	Ensino Superior

Cores			
Eixo horizontal secundário	**1**	**Não se aplica**	**3**
Acesso a Internet	Não possui Internet	Não se aplica	Possui Internet

Figura 6.2. Rótulo dos dados – eixo horizontal secundário.

A Figura 6.3 ilustra, a distribuição das médias representadas no eixo x, os respectivos grupos gerados com base no algoritmo de agrupamento no eixo y e levando em conta o rendimento familiar dos estudantes representados no eixo horizontal secundário pelas cores (azul, verde e vermelho) distribuídos conforme figura 6.2.

A quantidade de estudantes, designados nos aglomerados (grupos 3, 1 e 2) de classe alta e média (vermelho=classe alta e verde=classe média) que conquistaram médias mais elevadas é superior, comparado com o aglomerado (grupo 0) de classe baixa (azul=classe baixa). Outra evidência é a escassez de estudantes de classe alta (vermelho) que conquistaram médias baixas no exame, contudo a aglomeração elevada de estudantes da classe média (verde) e baixa (azul) que obtiveram médias muito baixa é evidente, conforme aglomerado (grupo 2).

Considerando as médias em relação à renda familiar do estudante, percebemos que embora a renda seja um fator importante, não é decisiva no desempenho do estudante. Comprovando a semelhança de médias entre os estudantes das classes alta, média e baixa. No entanto, como expusemos anteriormente, a dimensão de estudantes aglomerados em classe alta e média que conquistaram boas médias é maior, comparados à classe baixa. Salientamos também que poucos estudantes da classe alta conquistaram médias baixas no exame, contudo a aglomeração de estudantes da classe baixa que tiveram médias muito baixas no ENEM (2018) é representativa.

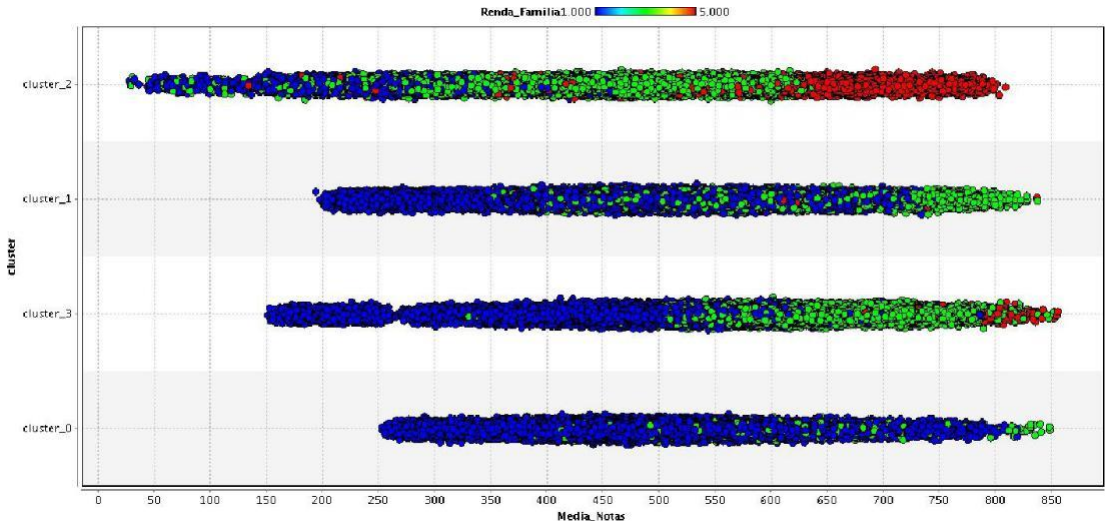

Figura 6.3. Gráfico de dispersão (Média Notas x *cluster* x Renda Familiar).

A Figura 6.4 ilustra, a distribuição das médias representadas no eixo x, os respectivos grupos gerados com base no algoritmo de agrupamento no eixo y e levando em conta a escolaridade do pai dos estudantes representados no eixo horizontal secundário pelas cores (azul, verde e vermelho) distribuídos conforme figura 6.2.

A quantidade de estudantes, designados nos aglomerados (grupos 0, 3, 1 e 2) de pai com ensino médio e superior (vermelho=ensino superior e verde=ensino médio) que conquistaram médias acima de 700 pontos é superior, comparado com os estudantes de pai com ensino fundamental (azul=ensino fundamental). Contudo ao analisarmos as médias inferiores a 50 pontos, temos estudantes cujo formação do pai encontra-se nas 3 faixas (vermelho=ensino superior, verde=ensino médio e azul=ensino fundamental).

Relacionando as médias com base na escolaridade do pai, trata-se de outro ponto importante analisado neste trabalho. Conforme representado na Figura 6.4, percebemos que um bom alicerce cultural familiar, consegue interferir positivamente, no resultado do estudante no ENEM (2018). Essa situação é esperada, já que possivelmente o pai com nível escolar superior, importa-se mais com a educação do filho. Revelando indícios de que a formação escolar do pai, pode render aos estudantes, o alcance de resultados positivos, ou seja, notas mais elevadas.

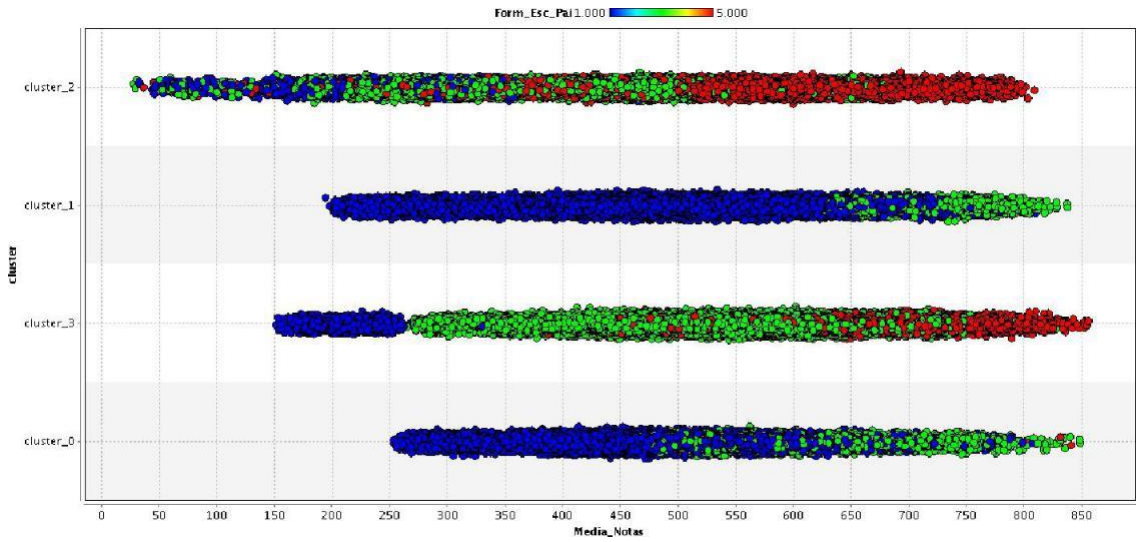

Figura 6.4. Gráfico de dispersão (Média Notas x *Cluster* x Escolaridade do pai).

Atualmente para qualquer usuário, o acesso à internet, tornou-se um recurso essencial, independentemente da finalidade, quer seja usado para promover a expansão entre indivíduos geograficamente distantes, assuntos de cunho cultural, escolar ou ciência, simples consulta ou solucionar questões particulares.

Finalmente, em nosso estudo, revelamos a ligação entre possuir acesso à internet e a média dos estudantes. A Figura 6.5 ilustra, a distribuição das médias representadas no eixo x, os respectivos grupos gerados com base no algoritmo de agrupamento no eixo y e levando em conta o acesso ou não dos estudantes à internet representados no eixo horizontal secundário pelas cores (azul e vermelho) distribuídos conforme figura 6.2.

A quantidade de estudantes, designados nos aglomerados (grupos 0, 3, 1 e 2) com acesso à internet (vermelho=Possui internet) que conquistaram médias acima de 500 pontos é superior, comparado com os estudantes sem acesso à internet (azul=Não possui internet), revelando indícios de que tal instrumento pode render aos estudantes, frutos mais eficientes no resultado do ENEM (2018).

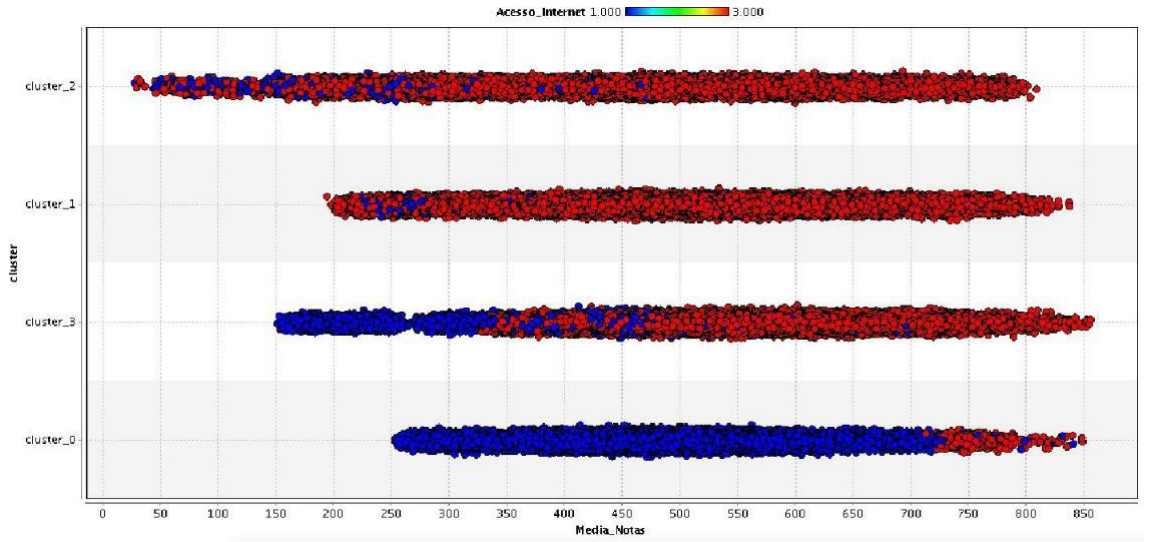

Figura 6.5. Gráfico de dispersão (Média Notas x *cluster* x Acesso à Internet).

7. Conclusão

Este estudo final de curso, apontou fundamentos da conceituação de mineração de dados, suas principais técnicas, a relevância da atividade de extração de conhecimento em dados abertos e as ferramentas que possibilitam realizar mineração de dados, bem como suas características principais. Por fim, foi apresentado um trabalho que revelou um modelo de mineração de dados que perpetramos. Nele detalhamos etapas como, manipulação de bases, o emprego do algoritmo de mineração (*K-Means*), finalizando com a verificação dos resultados obtidos. No trabalho, buscamos ponderar a presença de peculiaridades dos estudantes que interferiram em sua média final, e descobrimos que apesar de não ser condição essencial, aspectos como escolaridade do pai, se possui acesso à internet e renda da família, tem indícios de afetarem a média do estudante.

Como pesquisas consecutivas, alunos que tiverem interesse, poderão aplicar algum outro algoritmo de mineração de dados para comparação dos resultados obtidos e também os representantes do governo, poderão utilizar os dados sobre o baixo desempenho de estudantes sem acesso à internet, para auxiliar na tomada de decisão, desenvolvimento de programas de auxílio e de investimentos, disponibilizando tal recurso para os mesmos, a fim de que, todos se beneficiem, atingindo melhores notas.

Referências

[1] Fayyad, U.; Piatesky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press, (1996).

[2] Romero, C., Ventura, S., Espejo, P.G., and Hervás, C. (2008). Data mining algorithms to classify students. In EDM.

[3] Prass, F. S. (2004). Estudo comparativo entre algoritmos de análise de agrupamentos em datamining. 2004. 71 f. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, SC.

[4] Faria, S. M. S. M. L. et al. (2014). Educational data mining e learning analytics na melhoria do ensino online.

[5] França, R. S. d. and Amaral, H. J. C. d. (2013). Mineração de dados na identificação de grupos de estudantes com dificuldades de aprendizagem no ensino de programação. RENOTE, 11(1).

[6] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 281-297. (1967).

[7] Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. Brazilian Journal of Computers in Education, 19(02):(2011).

[8] Alpaydin, E. (2004). Introduction to Machine Learning.

[9] Evandro Costa, Ryan S.J.d. Baker, Lucas Amorim, Jonathas Magalhães, Tarsis Marinho. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Anais da Jornada de Atualização em Informática na Educação (JAIE 2012). <https://www.br-ie.org/pub/index.php/pie/article/view/2341>.

[10] Daniel Gomes Dosualdo, Solange Oliveira Rezende. Análise da Precisão de Métodos de Regressão. (2003) <https://web.icmc.usp.br/SCATUSU/RT/BIBLIOTECA_113_RT_197.pdf>.

[11] Holmes, F., et al, (1994). "Weka: A machine learning workbench", pg 1: "The Weka is a cheeky, inquisitive native New Zealand bird about the size of a chicken.

[12] Suporte ao Software Livre (2020). <https://www.gnu.org/licenses/licenses.pt-br.html>.

[13] The R Project for Statistical Computing (2020). <https://www.r-project.org/>.

[14] RStudio, PBC. (2020). <https://rstudio.com/>.

[15] Ciência Prática (2015). R – Um ambiente de trabalho gratuito para análise e visualização de dados. <https://cienciapratica.wordpress.com/2014/12/02/r-uma-linguagem-gratuita-para-analise-e-visualizacao-de-dados/>.

[16] Sourceforge (2015). RapidMiner. <https://sourceforge.net/projects/rapidminer/>.

[17] LARHUD (2018). Orange. <http://www.larhud.ibict.br/index.php?title=Orange>.

[18] Carvalho, Luis Alfredo V. A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. São Paulo: Érica, 2002. 242p.

[19] Pandas 2020. <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>.