Mineração de Dados do Sistema Acadêmico do Instituto Federal do Sudeste de Minas Gerais – Campus Juiz de Fora

Aluísio Cardoso Silva, Ricardo Costa Pinto e Santos

[1]Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais - Campus Juiz de Fora

aluisiohg@hotmail.com, ricardo.cpsantos@gmail.com

Abstract. *This work proposes the application of Educational Data Mining in a database from the Integrated Academic Management System of the Juiz de Fora Campus of the Federal Institute of Southeastern Minas Gerais in order to subsidize the internal decision making and the improvement of education by part of the direction. Based on computational intelligence and statistical techniques, we sought to identify patterns among the characteristics of the students (race, type of entrance, municipality of residence and income range) and their performance in the basic disciplines in high school. It was possible to identify punctual influences about some disciplines in relation to certain characteristics, allowing the objective intervention by those in charge.*

Resumo. *Este trabalho propõe a aplicação da Mineração de Dados Educacionais em uma base de dados oriunda do Sistema Integrado de Gestão Acadêmica do Campus Juiz de Fora do Instituto Federal do Sudeste de Minas Gerais a fim de subsidiar a tomada de decisões internas e o aprimoramento da educação por parte da direção. A partir de técnicas da inteligência computacional e da estatística, buscou-se identificar padrões entre as características dos alunos (raça, tipo de ingresso, município de residência e faixa de renda) e seu desempenho nas disciplinas básicas do ensino médio. Foi possível identificar influências pontuais acerca de algumas disciplinas em relação a determinadas características, possibilitando a intervenção objetiva por parte dos responsáveis.*

1. Introdução

A Mineração de Dados tem sido amplamente aplicada nas mais diversas áreas do conhecimento com o intuito de descobrir novas informações e subsidiar as tomadas de decisões. Uma de suas aplicações mais recentes, ainda pouco difundida no Brasil, é definida pelo termo "Mineração de Dados Educacionais". As possibilidades de descobertas de informações, por meio de estudos nesse ramo, são de grande potencial de apoio ao desenvolvimento da educação no país [Baker, Isotani e Carvalho 2011].

A transformação de uma nação depende fundamentalmente de uma cultura que priorize a educação para que maneiras concretas de ascensão profissional-econômica, através do esforço em trabalho e estudo, façam parte da realidade. Portanto, o desenvolvimento de um país está fortemente atrelado à educação de qualidade [Freitas, 2016].

Diante do exposto no parágrafo anterior, o estudo descrito neste trabalho pretende, a partir de ferramentas computacionais, especificamente através da aplicação de técnicas de Mineração de Dados, contribuir para o aprimoramento da educação, dentro do Instituto Federal do Sudeste de Minas Gerais – Campus Juiz de Fora. As informações extraídas pelos métodos propostos permitem que ações práticas sejam planejadas e executadas pontualmente a fim de nivelarem o desempenho escolar entre os grupos de estudantes da instituição.

Este trabalho descreve os processos de Mineração dos Dados utilizados para verificação de influência do perfil dos alunos dos cursos técnicos integrados do IF Sudeste MG – Campus Juiz de Fora em seu desempenho escolar. Os dados trabalhados são oriundos do SIGA (Sistema Integrado de Gestão Acadêmica), utilizado pela instituição. Após criteriosa análise à base de dados e eliminação de possíveis fatores de influência no desempenho acadêmico, considerando critérios de garantia de consistência, quatro fatores foram definidos como objetos de estudo, são eles: raça, faixa de renda familiar, município de residência e tipo de ingresso do aluno.

Atualmente as modalidades de cursos presenciais oferecidos pelo Campus Juiz de Fora estão divididas em Especialização, Graduação, Técnico Integrado, Técnico Modular e Técnico Proeja. Objetivando-se a pluralidade de amostras e a maior amplitude da análise, foram definidas como alvo desse estudo as notas das disciplinas comuns aos alunos dos cursos técnicos integrados, ofertados pela instituição, sendo elas as disciplinas das áreas de: Matemática, Português, História, Geografia, Biologia, Física e Química. Os dados foram coletados de disciplinas ministradas no intervalo de tempo iniciado no primeiro semestre de 2010 até o segundo semestre de 2015.

Os processos descritos englobam desde a etapa de pré-processamento dos dados, incluindo toda a tarefa de seleção e agregação das informações disponíveis em diferentes tabelas, por meio de SQL (Structured Query Language), até a etapa de processamento e análise dos resultados, onde modelos de inferência, oriundos tanto da Inteligência Artificial quanto da Estatística, foram utilizados.

Através da execução de testes estatísticos e da análise de *clusters* (ou análise de agrupamentos), buscou-se extrair informações relevantes e identificar padrões consistentes presentes nos dados explorados. Posteriormente, foram expostos os resultados encontrados pelos métodos propostos, objetivando o fornecimento de informações úteis à tomada de decisões dentro do Campus.

2. Revisão Sistemática e trabalhos relacionados

A revisão sistemática da literatura [Kitchenham 2004] aplicada a este trabalho foi motivada pelo intuito de selecionar pesquisas nacionais, com propósito semelhante àquele fundamentado neste estudo, com o objetivo da realização de possíveis comparações de resultados entre análises semelhantes àquelas propostas neste artigo, pertinentes à realidade da educação brasileira. É importante ressaltar que a metodologia empregada no presente trabalho foi traçada em função dos objetivos em questão e dos desafios particulares relativos à base de dados estudada. Em paralelo observou-se a justificativa para utilização de cada método proposto, encontrada na literatura.

A revisão foi direcionada pela seguinte questão: “Qual o estado da arte para o tema Mineração de Dados Educacionais no Brasil?”.

Buscando a amplitude de resultados, a *string* de busca ideal foi definida como: “Mineração de Dados Educacionais”, que por sua vez foi processada no Portal de Periódicos CAPES, visto que o acervo disponível abrange diversas bases de dados. Obteve-se um total de quatro trabalhos, após a análise dos conteúdos retirou-se um deles pro se tratar de duplicação.

A tese de [Kampff 2009] descreve uma pesquisa que, através da Mineração de Dados gerados em um Ambiente Virtual de Aprendizagem (AVA), busca identificar padrões comportamentais e característicos indicadores de evasão ou reprovação. Considerando que os dados foram rotulados previamente, a extração de tais padrões se deu por meio de regras de classificação (aprendizagem supervisionada). Em seguida foi proposta uma arquitetura para sistemas de alertas em AVA baseados nos padrões inferidos pela mineração dos dados. Através de um experimento preliminar foi comprovado que as intervenções realizadas pelos professores, direcionadas pelos alertas, contribuíram para melhoria dos índices de aprovação e redução dos índices de

evasão. As conclusões foram baseadas em comparações obtidas por análises estatísticas fundamentadas em testes de hipóteses.

[Júnior 2010] apresenta uma abordagem semelhante à do trabalho anteriormente citado. Acerca da dificuldade do educador em compreender suas classes em um ambiente educacional online, a dissertação propõe uma metodologia de trabalho capaz de fornecer subsídio ao docente possibilitando o aprimoramento dos cursos ministrados e a identificação de possíveis problemas. O trabalho define e apresenta um conjunto de etapas a ser seguido para obtenção dessas informações e ao final aplica a metodologia em um conjunto de dados reais obtidos através do sistema utilizado na UNICAMP. Dentre as etapas estão inseridas técnicas de Mineração de Dados Educacionais escolhidas em função das particularidades dos dados apresentados e dos objetivos do trabalho.

O trabalho de [Baker, Isotani e Carvalho 2011] discorre acerca das oportunidades para o Brasil no ramo da Mineração de Dados Educacionais (ou EDM, "*Educational Data Mining*"), destacando que ainda se trata de uma área em expansão no país com grande potencial de apoio na melhoria da educação. Além disso, são apresentados conceitos gerais da EDM, suas possibilidades de aplicação e o levantamento das linhas de pesquisa e técnicas disponíveis.

Dentre os trabalhos selecionados, aqueles desenvolvidos acerca de uma aplicação prática da Mineração de Dados Educacionais foram direcionados predominantemente à educação à distância. A metodologia envolvida considerou as peculiaridades, os desafios e as oportunidades envolvidas em ambientes virtuais. O presente trabalho apresenta, em essência, objetivo semelhante, porém, aplicando-se análises acerca de dados oriundos de cursos presenciais e com o foco do estudo não somente no critério de evasão, e sim no impacto de diversos fatores, externos ao domínio escolar, no desempenho acadêmico do aluno. Dessa forma, fez-se necessária a elaboração de uma metodologia diferente daquelas revisadas, que atendesse as particularidades e os objetivos em questão.

3. Referencial Teórico

Os métodos aplicados neste trabalho estão inseridos no amplo conceito definido por Mineração de Dados, área da ciência que busca a descoberta de novos conhecimentos através da Estatística e da Inteligência Artificial. Muitas vezes, o grande volume de dados trabalhados não permite uma extração trivial de informações relevantes. Nestes casos, as técnicas automáticas, englobadas pela Mineração de Dados ou *Data Mining*, possibilitam a exploração de grandes bases de dados à procura de novos padrões ou relações, que diante da complexidade de processamento, seriam inviáveis através do uso restrito de comandos SQL ou estatística simples [Santos 2006].

A Mineração de Dados Educacionais (ou "*Educational Data Mining*") [Baker, Isotani e Carvalho 2011] é definida pela execução dos processos intrínsecos à Mineração de Dados sobre conjuntos de dados oriundos do ambiente educacional. Os estudos e análises envolvidos nessa área de pesquisa pretendem identificar fatores que influenciam a aprendizagem do aluno, ou ainda, inferir informações úteis para compreensão objetiva do processo de aquisição de conhecimento.

3.1 Análise Estatística

Os testes de hipóteses [Moore e McCabe 2002], bases para a Análise Estatística proposta neste trabalho, compõem os métodos mais comuns de inferência.

Um teste de hipóteses constitui na verificação de evidências estatísticas para se tomar por verdadeira determinada condição. Normalmente, a análise é feita sobre uma amostra de dados e a partir dela busca-se fundamentar tal inferência acerca de toda a população. A realização dos testes tem por objetivo a comprovação ou rejeição da

Hipótese Nula (H0), usualmente estabelecida como “nenhum efeito” ou “nenhuma diferença”. Em caso de rejeição, considera-se que há evidencias estatísticas para que se aceite a Hipótese Alternativa (H1) [Câmara e Silva 2001], [Minitab support].

Anteriormente à execução dos testes [Câmara e Silva 2001], deve-se definir uma probabilidade *alfa*, denominada por “nível de significância”. Através dela, delimita-se a região de rejeição, na qual estão inseridos os valores extremos da amostragem. O valor de *alfa* será o ponto de corte do teste e representará a possibilidade máxima de se rejeitar equivocadamente a Hipótese Nula (H0), sendo comumente definido em 0,05. Dessa forma, o nível de confiança do teste aplicado é de 95%, ou seja, o complemento do valor *alfa* [Sousa, Junior e Ferreira 2012].

Após a execução dos testes, o *p-value* (valor-p), denominado por “probabilidade de significância”, determinará a probabilidade de H0 ser verdadeira. Caso o seu valor seja maior do que o nível de significância estabelecido, então aceita-se H0. Do contrário, se o *p-value* obtido for menor ou igual a *alfa*, então rejeita-se a Hipótese Nula [Câmara e Silva 2001], [Minitab support].

A escolha do teste estatístico adequado deve ser feita com base em critérios objetivos que indiquem primeiramente a opção pelos testes paramétricos ou pelos não paramétricos [Reis e Junior 2007]. Os paramétricos compreendem àqueles que, para sua validação, possuem algumas pressuposições a serem atendidas, já que os respectivos cálculos estatísticos se baseiam em parâmetros pré-verificados. Dentre estes parâmetros, na maioria dos casos, inclui-se a comprovação da normalidade na distribuição das amostras [Siegel e Castellan 2006], [Reis e Junior 2007].

Os testes não paramétricos, por sua vez, apresentam-se como alternativa aos paramétricos. Os cálculos estatísticos, nesses casos, consideram os postos atribuídos aos dados ordenados (*ranks*) e não os seus valores reais. Por esse motivo, garantem maior liberdade em relação às pré-exigências dos testes paramétricos, especialmente em relação à normalidade e homogeneidade de variâncias das amostras. Tais verificações, nesse caso, não são requisitos [Pontes e Corrente 2001], [Reis e Junior 2007].

Ainda que os cálculos em testes paramétricos sejam executados de maneira diferente daqueles executados pelos não paramétricos [Campos 2001], em ambos os casos, em termos práticos diante dos resultados, pode-se argumentar acerca das médias das amostras.

Identificado o grupo de testes adequado aos pressupostos exigidos, deve-se partir efetivamente para a escolha do teste a ser adotado. A Tabela 1 expõe os testes comumente utilizados e, para cada um deles, as características a serem atendidas pelo conjunto de dados trabalhado [Campos 2001].

Tabela 1 - Testes Estatísticos [Campos 2001]

Testes Estatísticos			
Paramétricos		Não-Paramétricos	
Independentes	Vinculados	Independentes	Vinculados
2 amostras	2 amostras	2 amostras	2 amostras
Teste *t* (Student)	Teste *t* (Student)	Mann-Whitney T. da Mediana χ^2 (2 x 2) Proporções Exato (Fisher)	Wilcoxon T. dos sinais Mac Nemar Binomial
Mais de duas	Mais de duas	Mais de duas	Mais de duas
Análise de variância	Análise de variância	Kruskal-Wallis Mediana (m x n) χ^2 (m x n) Nemenyi	Cochran Friedman

3.2 Análise de *Clusters*

O outro método utilizado neste trabalho para inferência de informações foi o de clusterização (*clustering*). Sua utilização é justificada no momento em que se pretende agrupar os dados baseando-se em suas características e não em um rótulo ou classe pré-definidos, por isso a técnica é definida por classificação não supervisionada. A execução do processo de clusterização resulta justamente na rotulação dos dados, no entanto, os rótulos serão definidos pelos padrões identificados pelo próprio algoritmo [Moscato e Von Zuben 2002].

O agrupamento resultante da clusterização é direcionado pelo intuito de se distribuir dados heterogênios em grupos (*clusters*), considerando em cada um deles a inserção dos elementos mais similares possíveis. Paralelamente a isto, esses elementos devem apresentar a maior diferenciação possível daqueles integrantes de outros grupos [Santos 2006].

Anteriormente ao processo efetivo de clusterização [Santos 2006], é necessário validar o número de *clusters* pelo qual o conjunto de dados será dividido. Tal validação ocorre através da comparação da eficiência obtida com a separação em diversas quantidades de grupos. Para tanto, o índice PBM constitui-se em uma das formas mais populares de validação de *clusters*, considerando tanto a separação entre os grupos, quanto à compactação dos dados em cada um deles. Aplicado neste trabalho, o índice PBM é compreendido pela Equação 1:

Equação 1 - Índice PBM [SANTOS,2006]

$$PBM(K) = \left(\frac{1}{k} \times \frac{E_1}{E_K} \times D_K\right)^2$$

Sendo que K representa o número de *clusters*. O fator E_i é a soma das distâncias de todos os elementos ao centro geométrico do conjunto de dados $\mathbf{w}_0$ e é calculado através da Equação 2, onde *t* varia de 1 a *N* (número de elementos):

Equação 2 – Fator E_1

$$E_1 = \sum_{t=1..N} d(\boldsymbol{x}(t), \boldsymbol{w}_0)$$

O fator E_κ corresponde à soma das distâncias dos registros ao centro do respectivo *cluster*, ou a soma das distâncias intra-grupos de *K* agrupamentos, ponderada pelo valor de pertinência do registro ao *cluster*. O cálculo é definido pela Equação 3 em que o ponto geográfico de cada elemento é representado por x_i:

Equação 3 - Fator E_K

$$E_K = \sum_{t=1}^{n} u_n d(\boldsymbol{x}_t, \boldsymbol{w}_i)$$

E D_κ representa a máxima separação entre os grupos, calculado a partir da Equação 4:

Equação 4 - Fator D_K

$$D_K = \max_{i,j=1..K} \left(d(\boldsymbol{w_i}, \boldsymbol{w_j})\right)$$

Cumprida a etapa anterior, utiliza-se o número ótimo de *clusters* encontrado como parâmetro fixo no algoritmo de clusterização propriamente dito. Dentre as diferentes abordagens possíveis para efetivação do processo, existe aquela em que os dados são rotulados com números absolutos, indicando que cada elemento pertence unicamente a determinado grupo, denominada usualmente por "*Crisp*". Entretanto, no presente trabalho, adotou-se a abordagem *fuzzy*, em que os dados podem ser atrelados a vários *clusters*, com diferentes graus de associação em cada um deles. Através desta última, é possível obter uma análise mais rica uma vez que sua flexibilidade permite maior grau de detalhamento nos resultados [Yonamine, Specia, Carvalho e Nicoletti 2002].

Dentre os métodos *fuzzy* existentes, o algoritmo *fuzzy c-means* é o que apresenta maior popularidade em estudos da área, principalmente por sua aplicabilidade e seus resultados satisfatórios [Yonamine, Specia, Carvalho e Nicoletti 2002], [Junior 2006]. Dada sua relevância e ampla utilização optou-se por sua aplicação na proposta de análise deste trabalho.

Proposto por *Bezdek* (1981) o algoritmo *fuzzy c-means* representa os *clusters* através de pontos estabelecidos como centros ou centroides no espaço, com posições definidas normalmente de forma aleatória. A pertinência dos elementos do universo amostral em relação a cada *cluster* é definida por sua distância euclidiana em relação aos respectivos centroides. Cada iteração executada pelo algoritmo procura reajustar os centros para que a distância entre os elementos do grupo e o seu centro seja cada vez menor. Esta etapa ocorre através da minimização da função objetiva dada pela Equação 5. Ao fim do processo é gerada uma matriz indicando a pertinência de cada ponto em relação aos centros das classes [Santos 2006].

Equação 5 - Determinação das pertinênicias W_{ij} do FCM

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_j\|}{\|\mathbf{x}_i - \mathbf{c}_k\|} \right)^{\frac{2}{m-1}}}$$

Onde, "$\mathbf{W}_{ij}$" é o grau de pertinência do elemento "x_i" ao grupo "j"; "**m**" (m>1) é o parâmetro de fuzzificação onde quanto mais próximo de 1, mais parecido com

conjuntos Crisp o FCM ficará; "**c**" reúne as coordenadas dos centros dos clusters, calculado a partir das médias das coordenadas dos elementos que compõem os clusters

4. Metodologia

Para que o trabalho pudesse ser desenvolvido, foi requerido oficialmente à Diretoria de Ensino o acesso ao Banco de Dados do SIGA (Sistema Integrado de Gestão Acadêmica) do Instituto Federal do Sudeste de Minas Gerais – Campus Juiz de Fora. A disponibilização do acesso ocorreu mediante a assinatura do Termo de Confidencialidade e Sigilo sobre as informações obtidas. Como consequência desse acordo, a Coordenação de Tecnologia da Informação da instituição providenciou o acesso a uma cópia temporária do Banco de Dados do sistema, alocada no servidor da instituição.

Os dados do sistema estavam armazenados em um SGBD (Sistema de Gerenciamento de Banco de Dados) PostgreSQL versão 8.3.17. Foram fornecidas as permissões de criação, edição, deleção e exportação de tabelas dentro da base de testes criada. Permissões essas, necessárias à execução da mineração dos dados educacionais da instituição. A manipulação e o acesso aos dados se deram através da ferramenta web phpPgAdmin versão 4.2.2, utilizada para administrar bancos de dados baseados em PostgreSQL por meio de uma interface gráfica.

4.1 Estudo e Seleção dos Dados

A primeira etapa deste trabalho constituiu-se no levantamento de questões relevantes a serem respondidas através da mineração dos dados educacionais dos alunos do IF Sudeste MG – Campus Juiz de Fora. Para tanto, fez-se necessária a observação dos dados disponíveis no banco de dados do sistema e a identificação daqueles potencialmente úteis para a posterior extração de informações relevantes, bem como na procura de padrões consistentes.

Considerando que o SIGA é constituído por diversas tabelas e que existe a necessidade de integração e consolidação dos dados para que uma análise possa ser efetuada, o software Navicat for PostgreSQL foi utilizado para gerar o Modelo Entidade-Relacionamento, com o objetivo de proporcionar uma maior capacidade de análise e entendimento da base a ser manipulada.

Definido o objeto de estudo, seguiu-se com o levantamento dos possíveis fatores influentes no desempenho dos alunos. A investigação foi direcionada por questionamentos como: "qual a influência da raça do aluno no desempenho escolar?", ou "qual a influência da renda familiar no desempenho do aluno?". Os demais questionamentos foram elaborados de acordo com os dados disponíveis e passíveis de análise detalhados ao longo do trabalho.

A partir da análise do modelo de dados foram escolhidas as tabelas e campos em função de sua relevância para o problema a ser estudado, tratando-se da identificação de fatores potencialmente relevantes na determinação do rendimento escolar dos alunos em função dos questionamentos levantados, e da qualidade e quantidade de dados inseridos em cada um dos campos em questão.

Feito o levantamento dos fatores de interesse, iniciou-se o processo de agregação dos dados em uma única tabela utilizando-se queries SQL. Inicialmente obteve-se uma tabela para cada área: Matemática, Português, História, Geografia, Biologia, Física e Química. Cada linha de registro contendo as colunas nota final, disciplina, matrícula e os possíveis fatores de influência selecionados: raça, tipo de ingresso, faixa de renda familiar, município de residência, área de procedência, procedência escolar e quantidade de pessoas em casa.

Diante da primeira seleção de dados, julgou-se necessária a checagem da consistência dos registros em questão. Através da observação e execução de queries SQL, foram encontradas algumas anormalidades na amostragem que poderiam interferir nas inferências. A primeira relacionada aos dados ausentes ou discrepantes. Alguns alunos apresentavam na coluna da nota os valores "*NULL*", "0" (zero) ou ainda valores extremamente baixos, em se considerando o valor "100" como nota máxima. O estudo desses casos, dada a observação e repetição da análise estrutural do banco de dados, permitiu que se atentasse para o campo "situação" pertencente à tabela "matrícula". Tabela esta, que contém além de outros dados, a nota, a disciplina e a matrícula do aluno.

Com base na execução de *Joins* SQL, identificou-se que os discrepantes ou nulos apresentam os valores "Cancelado", "Transferido" ou "Trancado" no campo "situação" da respectiva disciplina. Naturalmente, o cancelamento, a transferência ou o trancamento podem ocorrer ao longo do ano letivo, o que implica em registros fora do padrão desejado para este trabalho, uma vez que a nota informada não corresponde à nota final do aluno naquele ano. Buscando-se garantir a integridade dos dados trabalhados, foi atribuído o valor "*NULL*" para notas das disciplinas do ano em questão.

Outra anormalidade encontrada diz respeito à duplicação de registros. Por meio da execução de novos comandos SQL constatou-se que algumas matrículas apresentaram mais de um lançamento em uma mesma disciplina. O cruzamento de tabelas por meio de *Joins* SQL permitiu confirmar que a maioria dos casos tratava-se de reprovações. De acordo com o regulamento acadêmico dos cursos técnicos integrados da instituição, ao ser reprovado em uma disciplina o aluno deve repetir todo o ano letivo, justificando a existência de registros duplicados. Considerando que a matrícula do aluno representa a chave primária da tabela final criada neste trabalho, apenas um dos registros poderia ser mantido. Logo, optou-se pelos registros do ano de aprovação do aluno. Vale destacar que, os lançamentos duplicados representam aproximadamente 5% da amostragem total, ou seja, são pouco impactantes sob o ponto de vista global da análise. Garantida a consistência dos dados, retomou-se o processo de agregação dos mesmos.

Neste ponto, a chave primária das tabelas obtidas era composta pela matrícula do aluno e a disciplina, ou seja, para a área de Matemática, por exemplo, o aluno poderia ter até três linhas de registro, considerando as disciplinas de Matemática que compõem os três anos do ensino médio (Matemática I, II e III). No processo seguinte as notas de cada uma das disciplinas foram transformadas em colunas de um mesmo registro, identificadas pelo nome da respectiva disciplina. Dessa forma, a chave primária da tabela de cada uma das áreas passou a ser somente a matrícula do aluno.

O procedimento anterior permitiu que se criasse posteriormente uma única tabela unindo as notas de todas das disciplinas cursadas pelo aluno em uma só linha de registro. A considerar que, para as análises aplicadas neste trabalho, o desempenho do aluno deva ser expresso para cada uma das sete áreas, fez-se necessário o cálculo da média aritmética sobre notas obtidas nas disciplinas associadas a cada uma delas. Para tanto, foram indispensáveis alguns ajustes prévios acerca da estrutura do banco de dados.

A falta de restrições em relação à entrada de dados do sistema resultou na existência de notas em que a separação das casas decimais estavam sendo representados tanto por vírgulas quanto por pontos, tornando-se necessária a execução do comando *replace* para cada um dos 21 campos, trocando as vírgulas por pontos.

Foi preciso ainda, tratar o fato dos campos terem sido registrados como tipo "*character(4)*" o que impediu a realização de operações matemáticas. Neste último caso, após repetidas tentativas, chegou-se ao comando "*TYPE real USING CAST('campo' AS real)*", sintaxe para *casting* aceita pela versão 8.3 do PostgreSQL.

Alterado o tipo dos campos "nota" para "*real*", efetuou-se uma pesquisa por um comando capaz de realizar a média aritmética entre as colunas de uma tabela, posto que, a função "avg('expressão')" é comumente utilizada para o cálculo sobre as linhas de uma mesma coluna. Diante do exposto, chegou-se ao comando exemplificado pela Querie 1:

```
CREATE TABLE 'tabela_final' AS SELECT
        (SELECT AVG(c)
         FROM   (SELECT 'nota1'
                 UNION ALL
                 SELECT 'nota2'
                 UNION ALL
                 SELECT 'nota3'
                 ) T (c)) AS 'media_aritmetica'
```

Querie 1 – Média aritmética entre colunas

Finalmente, executou-se o comando para criação da última tabela, mantendo-se a mesma seleção de colunas, porém, substituindo-se as notas das disciplinas pela média aritmética associada a cada uma das áreas envolvidas.

Os possíveis fatores de influência no desempenho do aluno, contidos na seleção final, foram averiguados um a um para que se confirmasse a viabilidade de sua utilização nas etapas seguintes. A decisão de se manter ou não cada coluna selecionada foi baseada na quantidade de valores nulos dentro da amostragem e na verificação da confiabilidade dos valores preenchidos. Tal verificação se deu mediante questionamento à Secretária Acadêmica, responsável pela inserção dos dados dos alunos no sistema, e à Coordenação de Tecnologia da Informação, responsável pela manutenção do SIGA.

Diante da averiguação, os parâmetros passíveis de análise foram definidos da seguinte forma: Município de Residência, Tipo de Ingresso, Raça e Faixa de Renda Familiar. Apesar da integridade dos dados, alguns ajustes ainda foram necessários, a serem descritos a seguir:

- Definiu-se que o estudo acerca do parâmetro Município de Residência seria feito confrontando-se o desempenho dos alunos moradores de Juiz de Fora com o dos alunos residentes em outras cidades.
- Os representantes da raça 'Indígena' e da raça 'Amarela' somavam apenas quatro do total de 967 registros. Diante da mínima representatividade a análise do parâmetro raça ficou restrita aos valores "Branca", "Parda" e "Negra".
- Tratando-se da coluna Tipo de Ingresso, somente aqueles definidos como "Exame de Seleção - Escola Pública" ou "Exame Seleção - Ampla Concorrência" foram considerados, por serem mais representativos e se tratarem de inserções mais confiáveis.

O parâmetro Faixa de Renda não apresentou irregularidades e nem registros nulos. Todos os alunos estão atrelados a uma das seis faixas de renda definidas em quantidade de salários mínimos.

4.2 Análise Estatística

Após o processo de sumarização e agregação dos dados seguiu-se com a análise estatística. A tabela resultante do processo de seleção foi exportada com a extensão .CSV, por se tratar de um dos formatos compatíveis aos softwares utilizados nos processos seguintes. Os testes estatísticos foram aplicados através do software Minitab

versão 17.1.0, que conta com ferramentas de fácil utilização para análise de dados, além de guias para utilização dos testes estatísticos adequados às características das amostras.

Os testes estatísticos foram realizados abordando-se cada disciplina separadamente. O primeiro passo se deu pela verificação da normalidade das amostras através do Teste de Kolmogorov-Smirnov, adotando-se as hipóteses abaixo:

H0 (hipótese nula): as amostras possuem distribuição normal.
H1 (hipótese alternativa): as amostras não possuem distribuição normal.

A grande maioria das amostras atestou a não normalidade de sua distribuição, já que os testes apontaram um *p-value* < 0,05 (nível de significância definido). Considerando-se que ao constatar a não normalidade para uma das amostras de um fator deva-se optar por um método estatístico não paramétrico, os casos de normalidade não impactaram nesta decisão, pois, para todos os quatro fatores em estudo, pelo menos um dos tratamentos apresentou distribuição não normal.

Nos demais testes estatísticos os dados foram analisados de acordo com o seguinte teste de hipóteses, considerando-se o nível de significância igual a 5%:

H0 (hipótese nula): não há diferenças entre as medianas.
H1 (hipótese alternativa): existem diferenças entre as medianas.

Foi aplicado o teste de Kruskal-Wallis para os fatores Raça e Faixa de Renda, por serem compostos por mais de duas amostras. Havendo rejeição da hipótese nula, como no caso da Figura 1, aplicou-se em seguida o teste de Kruskal-Wallis Multiple Comparisons para indicar quais tratamentos apresentavam diferenças entre si – resultados exemplificados na Figura 2.

Kruskal-Wallis Test: bio versus raca

578 cases were used
389 cases contained missing values

Kruskal-Wallis Test on bio

raca	N	Median	Ave Rank	Z
1	422	71,98	295,1	1,33
2	31	69,10	216,5	-2,50
3	125	71,47	288,7	-0,06
Overall	578		289,5	

H = 6,40 DF = 2 P = 0,041
H = 6,40 DF = 2 P = 0,041 (adjusted for ties)

Figura 1 - Resultado do Teste de Kruskal-Wallis entre notas de Biologia e o parâmetro Raça.

Kruskal-Wallis Multiple Comparisons: Conclusions

The following groups showed significant differences (adjusted for ties):

Groups	Z vs. Critical value	P-value
1 vs. 2	2,58601 >= 1,834	0,0097
2 vs. 3	2,18162 >= 1,834	0,0291

Figura 2 - Resultado do Teste de Kruskal-Wallis Multiple Comparisons entre notas de Biologia e o parâmetro Raça.

No caso dos parâmetros Município de Residência e Tipo de Ingresso, foi aplicado o teste de Mann-Whitney, com as respectivas saídas exemplificadas pelas Figura 3 e Figura 4, já que possuem apenas duas amostras para cada disciplina, bastando a informação de que as medianas são semelhantes ou não.

Mann-Whitney Test and CI: bio_1111; bio_4733

```
           N  Median
bio_1111  167  73,000
bio_4733  800  70,915

Point estimate for η1 - η2 is 1,970
95,0 Percent CI for η1 - η2 is (0,580;3,400)
W = 89814,0
Test of η1 = η2 vs η1 ≠ η2 is significant at 0,0062
The test is significant at 0,0062 (adjusted for ties)
```

Figura 3 - Resultado do Teste de Mann-Whitney entre notas de Biologia e o parâmetro Município de Residência.

Mann-Whitney Test and CI: bio_54; bio_55

```
          N  Median
bio_54  502  71,035
bio_55  274  71,575

Point estimate for η1 - η2 is -0,350
95,0 Percent CI for η1 - η2 is (-1,630;0,900)
W = 193357,5
Test of η1 = η2 vs η1 ≠ η2 is significant at 0,5760
The test is significant at 0,5760 (adjusted for ties)
```

Figura 4 - Resultado do Teste de Mann-Whitney entre notas de Biologia e o parâmetro Tipo de Ingresso.

4.3 Análise de *Clusters*

Acerca do método de clusterização, por se tratar de um processo de análise multivariado, ou seja, que analisa, nesse caso, todas as notas conjuntamente, somente os alunos que possuíam consistência nas notas em todas as áreas permaneceram para o processamento. Diferentemente da análise estatística, as nulidades provocariam erros nos processos envolvidos neste método, sendo necessário mais um esforço no pré-processamento.

Diante do exposto, foi gerada uma nova tabela no banco de dados totalizando 762 registros passíveis de análise. Os dados foram exportados com a extensão .CSV e posteriormente importados em uma planilha eletrônica. As colunas referentes as notas foram separadas dos demais parâmetros e importadas para o software Matlab.

Em seguida, executou-se o algoritmo de Validação de *Clusters* a fim de identificar a quantidade ideal de *clusters* a ser trabalhada pelo atual processo. O resultado apontou que a separação dos registros em dois grupos se trata da melhor opção, como mostrado na Figura 5, onde foi testada a eficiência na divisão de 2 até 10 grupos. Para a leitura do índice PBM, utilizado para validação, entende-se que a menor nota obtida pelo processo corresponde a melhor distribuição de grupos.

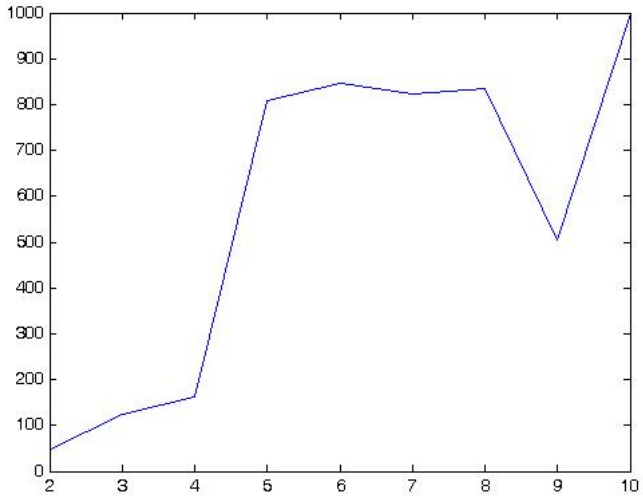

Figura 5 - Resultado da validação de *clusters* através do índice PBM

Finalmente, executou-se o algoritmo Fuzzy C-Means, fixando-se o número de *clusters* em 2 e efetivando-se o processo de clusterização.

Após a execução do algoritmo e a separação dos registros em dois *clusters*, foi feito o estudo estatístico das características de cada um deles, baseando-se no percentual de recorrência para cada tratamento. Tal levantamento ocorreu através da planilha eletrônica.

Confrontando-se as características dos dois grupos, foi considerado como diferença significativa aquelas que ultrapassaram 5%.

5. Resultados

5.1 Resultados Análise Estatística

A análise estatística feita acerca das notas de cada área apresentou como resultado geral poucas diferenças de desempenho escolar entre os grupos, especificamente no que tange aos parâmetros utilizados neste trabalho. Entretanto, destacam-se os pontos em que foi possível constatar a diferenciação das médias em relação aos fatores selecionados, baseando-se nos testes estatísticos. O detalhamento dos resultados e a tabela-resumo para cada um deles são expressos a seguir. As tabelas contém a informação da respectiva amostra em relação às demais testadas para a mesma área. Os casos de médias estatisticamente semelhantes foram representados pela sigla "E.S.". As medidas que apresentaram diferenças estão devidamente representadas pelos sinais matemáticos "<" e ">":

Município de Residência. Os testes de Mann-Whitney para a maioria dos tratamentos confirmou a hipótese alternativa, quando existe diferença significativa entre as amostras. Nesses casos, do ponto de vista estatístico, existem evidências de que as médias dos alunos residentes em Juiz de Fora são menores em relação aos alunos moradores de outras cidades. Tal superioridade só não foi verificada pelos testes nas disciplinas de Matemática e História, quando o *p-value* foi maior que 0,05, aceitando-se a hipótese nula, como mostra a Tabela 2.

Tabela 2 - Resultados dos testes estatísticos para o fator Município de Residência x Área

Município x Área	Biologia	Física	Química	Português	Matemática	História	Geografia
Juiz de Fora	<	<	<	<	E.S.	E.S.	<
Outras cidades	>	>	>	>	E.S.	E.S.	>

Tipo de Ingresso. Novamente aplicando-se os testes de Mann-Whitney, foi possível afirmar que as médias das duas categorias de ingresso nas disciplinas de Física e Geografia são estatisticamente diferentes, como expresso na Tabela 3, quando se obteve um *p-value* < 0,05. Os alunos que ingressaram no curso pelo processo seletivo na categoria "Ampla Concorrência", do ponto de vista estatístico, apresentaram médias superiores em relação aos alunos ingressantes pela categoria "Escola Pública". Nas demais disciplinas os testes não verificaram diferenças significativas.

Tabela 3 - Resultados dos testes estatísticos para o fator Tipo de Ingresso x Área

Tipo de Ingresso x Área	Biologia	Física	Química	Português	Matemática	História	Geografia
Ampla Concorrência	E.S.	>	E.S.	E.S.	E.S.	E.S.	>
Escola Pública	E.S.	<	E.S.	E.S.	E.S.	E.S.	<

Raça. Em relação ao parâmetro Raça, o teste de Kruskal-Wallis indicou rejeição da hipótese nula (não existe diferença significativa entre as amostras) somente para as médias em Biologia. O teste de Kruskal-Wallis Mutiple Comparisons, executado em seguida, indicou que estatisticamente os alunos das raças branca e parda possuem médias semelhantes nessa área, já os alunos da raça negra apresentam médias inferiores em relação aos grupos anteriores, como pode ser conferido através da Tabela 4.

Tabela 4 - Resultados dos testes estatísticos para o fator Raça x Área

Raça x Área	Biologia	Física	Química	Português	Matemática	História	Geografia
Branca	> Negra	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Negra	< Branca < Parda	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Parda	> Negra	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.

Faixa de Renda Familiar. Exposto através da Tabela 5, o cruzamento entre as faixas de renda familiar e as notas de cada uma das sete áreas, através do teste de Kruskal-Wallis, mostrou que as médias não apresentam diferenças significativas do ponto de vista estatístico.

Tabela 5 - Resultados dos testes estatísticos para o fator Faixa de Renda x Área

Faixa de Renda x Área	Biologia	Física	Química	Português	Matemática	História	Geografia
Faixa 1 (Menor que 0,5 salário mínimo)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Faixa 2 (Maior que 0,5 e menor que 1 salário mínimo)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Faixa 3 (Maior que 1 e menor que 1,5 salário mínimo)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Faixa 4 (Maior que 1,5 e menor que 2,5 salários mínimos)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Faixa 5 (Maior que 2,5 e menor que 3 salários mínimos)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.
Faixa 6 (Maior que 3 salários mínimos)	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.	E.S.

5.2 Resultados Análise de *Clusters*

Após a execução do algoritmo de clusterização, os registros foram divididos em dois *clusters* de acordo com as características, nesse caso, compreendidas pelas notas nas sete áreas do conhecimento. Um dos grupos composto pelos registros com os valores das maiores notas, alunos com desempenho superior, e o outro contendo os alunos com desempenho inferior. Apesar da maioria dos tratamentos não apresentarem um percentual de recorrência significativamente diferente entre os dois *clusters*, deve-se ressaltar aqueles em que foi possível constatar uma diferença acima dos 5%.

Os alunos que estão enquadrados na faixa de renda familiar mais alta (maior que três salários mínimos) representam 29,32% do grupo dos portadores de melhor desempenho escolar, sendo a faixa mais recorrente no grupo. Em contrapartida, dentre os alunos inseridos no grupo com desempenho inferior, apenas 15,38% deles situam-se dentro dessa faixa. Trata-se de uma diferença considerável que não foi verificada pelos testes estatísticos, nos quais o fator em questão foi o único a não apresentar diferenças relevantes entre as médias das amostras. Já a faixa de renda número 3 (maior que um e menor que um e meio salário mínimo) representa a maior das seis amostras do grupo com desempenho escolar inferior, 33,33%. Por outro lado, no grupo dos alunos com desempenho superior esse número é 8,44% menor.

Além do exposto anteriormente, verificou-se diferença significativa em relação ao parâmetro raça. A recorrência da raça Parda foi 9,82% maior no *cluster* composto pelas maiores notas, já a raça Branca foi 7,14% menos recorrente nesse grupo. Para os demais parâmetros, os dois *clusters* não apresentaram diferenças expressivas.

A análise completa do processo de clusterização pode ser conferida através das tabelas: Tabela 6, Tabela 7, Tabela 8 e Tabela 9.

Tabela 6 - Resultados da Análise de Clusters para o fator Faixa de Renda

Faixa de Renda por *cluster*	Grupo 1 – Desempenho inferior	Grupo 2 – Desempenho superior	Diferença
Faixa de Renda 1	7,69%	5,12%	-2,57%
Faixa de Renda 2	10,26%	9,54%	-0,71%
Faixa de Renda 3	33,33%	24,90%	-8,44%
Faixa de Renda 4	17,95%	18,12%	0,17%
Faixa de Renda 5	15,38%	13%	-2,38%
Faixa de Renda 6	15,38%	29,32%	13,94%

Tabela 7 - Resultados da Análise de Clusters para o fator Raça

Raça por *cluster*	Grupo 1 – Desempenho inferior	Grupo 2 – Desempenho superior	Diferença
Branca	80,77%	73,63%	-7,14%
Negra	7,69%	5%	-2,69%
Parda	11,54%	21,36%	9,82%

Tabela 8 - Resultados da Análise de Clusters para o fator Tipo de Ingresso

Tipo de Ingresso por *cluster*	Grupo 1 – Desempenho inferior	Grupo 2 – Desempenho superior	Diferença
Ampla Concorrência	69,23%	67,73%	-1,50%
Escola Pública	30,77%	32,27%	1,50%

Tabela 9 - Resultados da Análise de Clusters para o fator Município de Residência

Município de Residência por *cluster*	Grupo 1 – Desempenho inferior	Grupo 2 – Desempenho superior	Diferença
Juiz de Fora	79,49%	84,09%	4,61%
Outras Cidades	20,51%	15,91%	-4,61%

6. Discussão dos Resultados

Acerca de uma análise pontual dos resultados obtidos, no tocante à análise estatística, pode-se dizer que alguns se contrapuseram ao senso comum. A superioridade das

médias, na maioria das áreas, verificada para alunos moradores de outras cidades, trata-se de um dado estatístico curioso. Considerando que esses alunos se deslocam diariamente até Juiz de Fora para estudarem na instituição, poderiam eventualmente ter seu desempenho prejudicado, seja pelo desgaste físico extra no deslocamento ou mesmo pelo tempo hábil reduzido em relação àqueles residentes na própria cidade do Campus. Por outro lado, a necessidade de um esforço maior para comparecer às aulas pode servir de estímulo para os alunos nessa condição, ocasionando, possivelmente, uma maior dedicação por parte deles.

A semelhança estatística verificada entre as médias dos alunos sob a ótica do parâmetro Faixa de Renda, também merece destaque. Supostamente, os alunos que dispõem de mais recursos financeiros, teriam o acesso facilitado à *internet* e outras ferramentas úteis ao aprendizado como livros e computador. No entanto, deve-se considerar que tais ferramentas são disponibilizadas pelo instituto, o que pode contribuir para o nivelamento desse aspecto entre os alunos.

Analogamente aos resultados obtidos para o parâmetro Faixa de Renda, constatou-se para os fatores Raça e Tipo de Ingresso a predominância de resultados positivos ao se testar a semelhança entre as médias escolares das amostras. Baseando-se na política de cotas raciais, cotas para estudantes de escola pública e as discussões que se tem sobre o assunto no país, os resultados foram surpreendentes, uma vez que não foi detectada diferença estatística para maioria dos testes.

Neste ponto, poderia se teorizar, por exemplo, que a instituição está cumprindo com eficácia seu papel de nivelar os alunos, que chegam na instituição com níveis variados de aprendizado. Ou ainda, pode-se levantar a suposição de que o ensino público fundamental não estaria defasado em relação ao ensino particular. Outra hipótese que não se pode descartar é a de que os alunos estejam se dedicando a ponto de compensarem possíveis discrepâncias na comparação de desempenho escolar.

Os resultados encontrados pela análise de *clusters* confirmaram em grande parte as inferências evidenciadas pela análise estatística. Deve-se destacar o fator Faixa de Renda, que nesse caso, correspondendo às expectativas iniciais, apresentou-se como característica de influência no desempenho escolar. É importante frisar que o contraste de resultados é plausível em se tratando de métodos diferentes. Ou seja, abordando-se as áreas separadamente (análise estatística), não houve diferença significativa entre as médias. Entretanto, sob a ótica global de desempenho escolar dos alunos (análise de *clusters*), verificou-se influência da Faixa de Renda na caracterização dos *clusters*.

6.1 Desafios encontrados

Em primeiro lugar, a ausência de restrições por parte do SIGA no momento da entrada de dados resultou na falta de padronização dos registros no banco de dados. Para o desenvolvimento deste trabalho, tal questão teve de ser contornada algumas vezes. A título de exemplificação, nas primeiras etapas, ao executar o comando para seleção dos registros das disciplinas de Matemática, por exemplo, deparou-se com as variações de nomenclaturas, como por exemplo: “Matemática ”, “Matemática 1” e “Matemática I”. Os valores possivelmente referiam-se à mesma disciplina, no entanto, trabalhou-se apenas com aqueles contendo numerais ou algarismos romanos. A necessidade de se optar pela desconsideração de registros em casos como este pode ter contribuído para a discrepância entre o total de registros de cada disciplina.

Além disso, ficou evidente na execução deste trabalho que o desenvolvimento do sistema e a própria modelagem do banco de dados não consideraram a manutenibilidade dos mesmos. Por esse motivo, as modificações, naturalmente necessárias ao longo do tempo, foram claramente improvisadas resultando em desnecessárias inconsistências nos dados.

Diante do exposto nesta seção, o sistema pode apresentar ainda perdas de desempenho e falhas de segurança, além de complexidade e redundâncias dispensáveis na base de dados. Ademais, análises como as operadas neste estudo ficam altamente prejudicadas, tornando-se excessivamente custosas. No entanto, acredita-se que as possíveis inconsistências encontradas, passíveis de serem gerenciadas, foram devidamente contornadas a fim de maximizar a confiabilidade dos resultados diante das falhas do sistema.

Deve-se destacar ainda que os resultados obtidos pelas análises dependem fundamentalmente da veracidade das informações inseridas no sistema, da qual foi buscada a comprovação por meio do questionamento aos responsáveis pelo preenchimento dos dados, o que ainda sim, não exclui a possibilidade de inconsistência gerada por inserções não confiáveis.

7. Conclusão

De uma forma geral, a identificação das causas para os resultados encontrados não se trata de uma tarefa simples e não coube a este trabalho precisá-las. No entanto, estudos como este podem embasar e direcionar futuras análises nesse sentido, principalmente quando atreladas aos anseios da diretoria de ensino. Ademais, o presente trabalho, por meio da extração de informações não triviais contidas nos dados educacionais, pode fornecer ainda subsídio à tomada decisões e ao planejamento estratégico pedagógico do Campus Juiz de Fora – IF Sudeste MG.

As ocorrências pontuais de discrepância estatística entre as médias dos alunos podem servir de base para intervenções pedagógicas, não só da coordenação, mas também por parte dos professores. Se atendo aos resultados da análise estatística, é possível identificar em qual disciplina obteve-se o cenário de desigualdade de desempenho e qual o fator influente neste caso. Dessa forma, a condução de ações de nivelamento, por exemplo, pode se apoiar precisamente ao que foi exposto neste trabalho.

Quanto a futuras análises, sugere-se a aplicação da Mineração de Dados Educacionais focando-se especificamente em reprovações. A partir da base de dados trabalhada, identificar padrões inerentes aos alunos reprovados a fim de subsidiar possíveis intervenções de maneira objetiva na prevenção desses casos. Outra possibilidade de estudo, com intuito semelhante, pode ser feito em relação às evasões, buscando colaborar na tarefa de prevenção de tal fenômeno dentro do Campus Juiz de Fora – IF Sudeste MG.

8. Referências Bibliográficas

Baker, R. S. J. D, Isotani, S e Carvalho, A. M. J. B. (2011). Mineração de dados educacionais: oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19 (2).

Câmara, F. G., Silva, O. (2001) Estatística não paramétrica: testes de hipóteses e medidas de associação. Departamento de Matemática-Universidade dos Açores, Ponta Delgada. 121p.

Campos, G. M. (2001) Estatística prática para docentes e pós-graduandos. Disponível em: <http://143.107.206.201/restauradora/gmc/gmc_livro/gmc_livro_cap14.html>. Acesso em: 22 dez. 2016.

Freitas, Eduardo de. "Educação, base do desenvolvimento"; *Brasil Escola*. Disponível em <http://brasilescola.uol.com.br/geografia/educacao-base-desenvolvimento.htm>. Acesso em 03 de janeiro de 2017.

Júnior, G. R. F. (2010). Metodologia de mineração de dados para ambientes educacionais online. (Dissertação de Mestrado – UNICAMP)

Júnior, N. L. C. (2006) Clusterização baseada em algoritmos *fuzzy*. Recife: Centro de Informática, Universidade Federal de Pernambuco.

Kampff, A. J. C. (2009). Mineração de dados educacionais para geração de alertas emambientes virtuais de aprendizagem como apoio à prática docente. (Tese de Doutorado -UFRGS)

Kitchenham, B. (2004). *Procedures for performing systematic reviews*. Keele, UK, Keele University, 33(2004), p. 1-26.

Minitab, support "O que é um teste de hipótese?", Disponível em: <http://support.minitab.com/pt-br/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/what-is-a-hypothesis-test> Acesso em: 22 dez. 2016

Moore, D. S. e McCabe, G. P. (2002) Introdução à prática da estatística. 3ª ed. Rio de Janeiro: Editora LTC.

Moscato, P. e Von Zuben, F. J. (2002) Uma visão geral de clusterização de dados. Unicamp,Campinas, SP. Disponível em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia368_02/topico5_02.pdf> . Acesso em: 23 de dez. 2016.

Pontes, A. C. F., Corrente, J. E. (2001). Comparações múltiplas não-paramétricas para delineamento com um fator de classificação simples. Revista de Matemática e Estatística, 19, 179-197.

Reis, G. M., Junior, J. I. R. (2007) Comparação de testes paramétricos e não paramétricos aplicados em delineamentos experimentais. III SAEPRO, UFV.

Santos, R. C. P. e. (2006) "Avaliação de Métodos Baseados em Sistemas Fuzzy para Mineração de Dados Georeferenciados", Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, RJ.

Siegel, S., Castellan JR., N. J. (2006) Estatística não-paramétrica para ciências do comportamento. Tradução de Sara Ianda Correa Carmona. 2ª ed. Porto Alegre: Editora Artmed.

Sousa, C. A., Junior, M. A. L., Ferreira, R. L. C. (2012) Avaliação de testes estatísticos de comparações múltiplas de médias. Revista Ceres, Vol. 59 n. 3, Viçosa.

Yonamine, F. S., Specia, L., Carvalho, V. O., Nicoletti, M. C. (2002) Aprendizado não supervisionado em domínios fuzzy Algoritmo fuzzy c-means. São Carlos: UFSCAR, 18p.