Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 01/Jun/2021; 1st round notif.: 10/Aug/2021; New version: 07/Sep/2021; 2nd round notif.: 08/Oct/2021;
Camera ready: 13/Dec/2021; Edition review: 15/Dec/2021; Available online: 23/Dec/2021; Published: 23/Dec/2021;

Estudo Exploratório através de Análises Longitudinais aplicado à Ciência da Computação a partir da Base de Dados do ENADE

Title: Exploratory Study Through Longitudinal Analysis applied to Computer Science from ENADE Database

Eliene Ribeiro Rosa
Universidade Federal de Goiás
ln.ribeiro.rosa@gmail.com

Deller James Ferreira
Universidade Federal de Goiás
ORCID: 0000-0002-4314-494X
deller@ufg.br

Nádia Félix Felipe da Silva
Universidade Federal de Goiás
ORCID: 0000-0002-3875-2211
nadia.felix@ufg.br

Alfredo Assis
Universidade Federal de Goiás
alfredo.mat.ufg@gmail.com

Resumo

Em busca da qualidade na educação, diversas pesquisas no âmbito educacional são realizadas, no entanto, análise de grandes bases de dados em busca de informações úteis é uma tarefa desafiadora. Nesta pesquisa, foram identificados aspectos relacionados ao desempenho acadêmico dos alunos, utilizando como base as provas do ENADE aplicados ao curso de Ciência da Computação. A pesquisa foi regida pela metodologia que tange o Estudo Longitudinal Transversal Repetido, buscando analisar os dados ao longo do tempo, para verificar seu comportamento nos anos estudados. Os anos analisados foram 2011, 2014 e 2017. Foram aplicadas técnicas de mineração de dados nos microdados fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. A contribuição desta pesquisa é mostrar que dentre os algoritmos utilizados nos experimentos, o que melhor classificou os dados em relação ao desempenho dos alunos foi a árvore de decisão, a qual possibilitou identificar que algumas características socioeconômicas como por exemplo renda familiar, escolaridade do pai, situação de trabalho do discente em conjunto com a categoria administrativa e o turno de graduação impactam no desempenho acadêmico.

Palavras-chave: *INEP, ENADE, Estudo Longitudinal, Mineração de Dados Educacionais, Aprendizagem de Máquina, Árvore de Decisão*

Abstract

In search of quality in education, several researches in the educational scope are carried out, however, analyzing large databases in search of useful information is a challenging task. In this research, we seek to identify aspects related to students' academic performance, using the ENADE tests applied to the Computer Science course as a basis. The research was governed by the methodology that concerns the type of Longitudinal Analysis, of the Repeated Transversal type, seeking to analyze the data over time, to verify their behavior in the years studied. The years analyzed were the years 2011, 2014 and 2017, were data mining techniques were applied to microdata provided by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. The contribution of the research was to show that among the algorithms used in the experiments, the algorithm that best classified the data in relation to the students' performance, was the decision tree, which made it possible to identify that some socioeconomic characteristics such as family income, education level father, the student's work situation in conjunction with the administrative category and the undergraduate shift have an impact on academic performance.

Keywords: *INEP, ENADE, Longitudinal Study, Educational Data Mining, Machine Learning, Decision Tree*

Cite as: Rosa, E. R., Ferreira, D. J., Silva, N. F. F., & Assis, A. (2021). Estudo Exploratório através de Análises Longitudinais aplicado à Ciência da Computação a partir da Base de Dados do ENADE. Revista Brasileira de Informática na Educação, 29, 1463-1486. DOI: 10.5753/rbie.2021.2073.

1 Introdução

A busca por qualidade na educação é um dos principais parâmetros almejados pelo sistema educacional. Para averiguar a qualidade no sistema educacional brasileiro, existe no Brasil uma autarquia vinculada ao ministério da educação, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)[1]. O Exame Nacional de Desempenho de Estudantes (ENADE) busca por melhorias no sistema educacional do Ensino Superior, sendo utilizado tanto para o desenvolvimento de políticas públicas da educação, quanto para garantir maior transparência dos dados (Lima, 2018).

O INEP possui uma extensa base de dados proveniente de levantamentos estatísticos e avaliativos em diversos níveis e modalidades de ensino. O INEP tem como missão promover estudos, pesquisas e avaliações sobre o sistema educacional brasileiro. Com o intuito de elaborar e implementar políticas públicas, o INEP produz informações passíveis de serem utilizadas por pesquisadores, educadores e a sociedade em geral (da Fonseca, 2016; Nóbrega, 2016).

Ao realizar avaliações periódicas em todos os níveis educacionais, o INEP detém informações importantes do percurso escolar dos alunos. As avaliações iniciam-se no 3º ano do Ensino Básico, com a prova do Sistema Nacional de Avaliação da Educação Básica (SAEB), que também avalia alunos do 5º e 9º anos. O Exame Nacional do Ensino Médio (ENEM) avalia alunos no fim do Ensino Médio e o ENADE avalia os concluintes dos cursos de graduação.

O conjunto de dados disponibilizado pelo INEP tem potencial para tornar-se uma fonte de extração de diversas informações relevantes (Silva, Morino, & Sato, 2014). A análise dos dados educacionais provenientes desses dados permite examinar o comportamento de estudantes e instituições, possibilitando o desenvolvimento de critérios de qualidade educacional para que, dentro de um contexto, ocorram decisões de melhorias do processo de ensino e aprendizagem, investimentos, planejamentos estratégicos, entre outros.

Com publicações em seu site oficial, o INEP disponibiliza informações relevantes, tais como resumos e dados que podem ser utilizados para pesquisas educacionais. As informações disponibilizadas em relatórios geralmente apresentam estatísticas descritivas, que demonstram resumos das informações coletadas, as quais podem ser utilizadas por gestores, educadores e ou demais interessados no âmbito educacional (da Fonseca, 2016; Lima, 2018). São disponibilizados ainda pelo INEP os microdados das provas aplicadas. Esses microdados são dados abertos ao público em geral no formato csv e podem ser baixados e utilizados para realizar diversas pesquisas para complementar as informações fornecidas.

Fazendo uso dos dados fornecidos, pode-se realizar, por exemplo, pesquisas que possam contribuir para identificação de fatores associados ao desempenho dos alunos, permitindo às instituições, ou mesmo o próprio sistema educacional, trabaralhem de forma preventiva, possibilitando a atuação em grupos onde se identifique algum fator de risco de baixo desempenho, o que pode garantir a busca constante pela qualidade na educação.

Com o intuito de acompanhar o desempenho nas provas aplicadas para o curso de Ciência da Computação ao longo dos anos, um Estudo Longitudinal Transversal Repetido foi realizado. Segundo (Fontelles, Simões, Farias, & Fontelles, 2009), as pesquisas podem ser classificadas de

[1] http://portal.inep.gov.br/microdados./ Acessado 01/12/2020

acordo com o período de tempo em que os dados são analisados, podendo ser classificadas como Estudo Transversal (ET) ou Estudo Longitudinal (EL). Enquanto o Estudo Transversal observa um determinado evento em um único instante de tempo, o Estudo Longitudinal observa os dados em um determinado intervalo de tempo.

Segundo (Hedeker & Gibbons, 2006), o EL possui algumas vantagens se comparado ao ET. Por exemplo, podem fornecer mais informações que o ET, uma vez que as medidas repetidas de um único sujeito são repetidas várias vezes em EL, enquanto que no ET são medidas uma única vez. Outra vantagem apresentada diz respeito a permitir que se separe os efeitos de mudanças ao longo do tempo dentro dos indivíduos analisados.

Segundo (Ruspini, 2003), alguns exemplos de Estudo Longitudinal comumente utilizados são:

• Estudo Longitudinal Transversal Repetido (ELTR): Estudos realizados regularmente, cada vez usando uma amostra diferente ou uma amostra completamente nova;

• Estudo Longitudinal Prospectivo (painel): São realizadas entrevistas repetidamente utilizando os mesmos sujeitos durante um período de tempo;

• Estudo Longitudinais retrospectivos (história de eventos ou dados de duração): São realizadas entrevistas onde os entrevistados são solicitados a lembrar e reconstruir eventos e aspectos de seus próprios cursos de vida.

Assim sendo, o objetivo da pesquisa é analisar os microdados do ENADE compreendendo os anos de 2008, 2011, 2014 e 2017, em particular, os dados do curso de Ciência da Computação, com o propósito de investigar a influência das características sociais dos alunos sob o seu desempenho acadêmico. A questão de pesquisa que norteou este estudo é: "É possível extrair relações entre as características socioeconômicas dos alunos de ciência da computação e o seu desempenho acadêmico, a partir da base do ENADE, nos anos 2008, 2011, 2014 e 2017?". Ao analisar os dados em um determinado período de tempo, acompanhando os perfis dos alunos, pode-se compreender como eles vem se portando ao longo do tempo e medidas preventivas podem ser tomadas para melhorar o ensino de computação. Para tanto, torna-se necessário o uso das técnicas e algoritmos para auxiliar nesse processo.

Este trabalho está estruturado em 6 seções. Na Seção 2, é apresentada a fundamentação teórica. Na Seção 3, são abordados os trabalhos relacionados com a presente pesquisa. Na Seção 4, é descrita a metodologia utilizada para alcançar os resulta-dos Desse trabalho. Na Seção 5, são apresentados os resultados desta pesquisa. Finalmente, na Seção 6, são apresentadas as discussões e conclusões.

2 Fundamentação Teórica

Analisar grandes volumes de dados sem o auxílio de ferramentas apropriadas pode ser inviável. Nesse contexto, utilizar o processo de descoberta de conhecimento em base de dados, do inglês *Knowledge Discovery in Databases* (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) para auxiliar na descoberta de informações úteis é uma alternativa de automatização desse processo. Neste trabalho, foram aplicadas as etapas do KDD em conjunto com o Estudo Longitudinal Trans-

versal Repetido às provas do ENADE. As informações coletadas pelo INEP em todos os anos, viabilizaram a realização do experimento, assim, puderam ser analisados os mesmos atributos em públicos distintos ao longo do tempo.

(Stephens & Sukumar, 2006) afirmam que o KDD é composto por diversas etapas: Entrada, Pré-Processamento, Mineração de Dados e Pós-processamento. Para (Fayyad et al., 1996), KDD é composto pelas etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação ou interpretação dos resultados, conforme é mostrado na Figura 1.

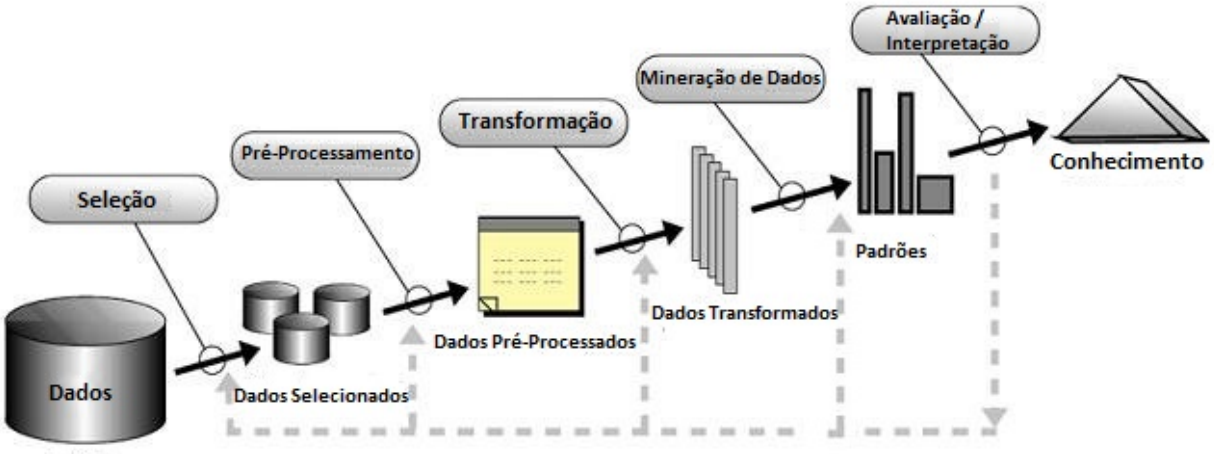

Figura 1: Processo de descoberta do conhecimento em bases de dados. Adaptação de (Fayyad et al., 1996).

Independentemente da nomenclatura ou etapas adotadas, o KDD inicia-se com a entrada dos dados, passa por procedimentos que visam ajustar os dados para que a Mineração de Dados possa ser aplicada e gere a descoberta de informações úteis.

A etapa de entrada consiste no processo de aprender o domínio da aplicação, possibilitando o entendimento a cerca dos dados estudados. Os dados obtidos podem ser oriundos de diversos formatos e fontes de armazenamento tais como planilhas, arquivos de texto, Tabelas relacionais entre outros.

Para (Stephens & Sukumar, 2006), o pré-processamento de dados é uma área ampla e consiste em uma série de estratégias e técnicas diferentes para preparação dos dados a serem minerados abrangendo: Integração de dados de múltiplas fontes, limpeza de dados para remover ruídos, remoção de dados duplicados, seleção de atributos relevantes para a tarefa de mineração desejada, transformação de dados entre outras.

Os dados ruidosos são aqueles que possuem valores discrepantes em relação aos demais valores presentes, ou seja, não apresentam o comportamento geral da maioria das informações presentes para um determinado atributo. (Stephens & Sukumar, 2006) afirma que a remoção de dados ruidosos presentes no conjunto de dados, pode melhorar o resultado da análise gerando resultados mais precisos.

A Mineração de Dados consiste na etapa do KDD em que se combina métodos tradicionais de análise de dados com algoritmos sofisticados, com o intuito de processar grandes volumes de

dados (Stephens & Sukumar, 2006). Enquanto que, a etapa de avaliação/interpretação consiste em interpretar os padrões encontrados, possibilitando ao usuário a tomada de decisão a cerca dos padrões encontrados. (Stephens & Sukumar, 2006) afirmam que pode retornar aos processos anteriores caso necessário.

3 Trabalhos relacionados

Com o intuito de identificar os tipos de análises que vêm sendo abordadas a partir do ponto de vista longitudinal, transversal e/ou por coortes em dados disponibilizados pelo INEP, foi realizada uma revisão da literatura, buscando identificar quais dados estão sendo explorados e quais tipos de estudos estão sendo desenvolvidos. Foram identificadas três categorias onde há maior ênfase na literatura: (i) acesso ao ensino superior; (ii) avaliação do desempenho do aluno; e (iii) correlação entre a formação docente e o desempenho do estudante.

Com respeito ao acesso ao ensino superior, como retratado por (Oliveira & Silva, 2017; Picanço, 2016), os trabalhos buscaram investigar a democratização do ensino superior brasileiro a partir do perfis socioeconômicos dos alunos. A evasão escolar também foi um dos aspectos identificados, que impede o ingresso no ensino superior. Para este caso, é mencionado o estudo sobre evasão realizado por Vieira (2012), onde é avaliada a ocorrência de retenção escolar até os 11 anos de idade.

Foram encontradas na literatura análises que objetivam investigar o desempenho e avaliar os resultados dos estudantes ou das instituições no exame. Dentre essas análises destaca-se Melguizo (2016) que fizeram a comparação entre os ingressantes e concluintes da prova do ENADE, com intuito de averiguar o conhecimento adquirido durante os anos de estudo.

Foram também encontrados trabalhos que visam a formação docente, onde os dados dos exames são analisados a fim de aprimorar a formação do docente e relacioná-la com o desempenho dos estudantes nas provas. Nessa perspectiva, encontra-se o trabalho de Medeiros (2014).

Os trabalhos supracitados utilizaram diferentes bases de dados disponibilizadas pelo INEP, tais como ENEM, ENADE, SAEBE, Prova Brasil, entre outros. É importante salientar que alguns trabalhos, como o de (Banni, Oliveira, & Bernardini, 2021), analisam diversos tipos de dados e em bases diferentes. Contudo, na maioria dos trabalhos, as análises disponibilizadas limitam-se a estatísticas descritivas dos dados, que visam descrever e resumir as informações coletadas. Análises mais sofisticadas, envolvendo estatística inferencial ou mineração de dados, principalmente com relação a dados socioeconômicos são menos comuns.

No que diz respeito ao impacto dos fatores socioeconômicos no desempenho dos estudantes, Banni et al. (2021) apresentaram uma análise experimental baseada em Mine-ração de Dados Educacionais, a partir dos dados do ENEM, para identificar os atributos mais relacionados ao desempenho dos estudantes. Os resultados mostraram que os atributos socioeconômicos, de fato, apresentam uma relação significativa com o resultado dos estudantes no ENEM.

Estudos apontam que alunos de instituições privadas apresentam melhores notas no ENEM, em contrapartida, escolas com maiores taxas de evasão apresentam piores notas no ENEM (Hoed & Saraiva, 2019). Resultados similares foram encontrados do estudo de (Carmo, Heckler, F., &

Carvalho, 2020), mostrando que há melhor desempenho dos candidatos com maior poder aquisitivo, provenientes de escolas privadas, evidenciando os reflexos das diferencas socio-econômicas em seu desempenho.

No trabalho de Silva et al. (2014), foi utilizada uma tarefa da Mineração de Dados conhecida por Associação de Dados para encontrar padroes de regras nos resultados de provas e questionários socioeconômicos do Exame Nacional de Ensino Médio (ENEM) de 2010. A partir dos resultados e do conhecimento extraído, a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas que moram com o estudante são atributos que diminuem o desempenho do aluno.

Considerando mais especificamente cursos de computação, com o vasto volume de dados fornecidos pelo ENADE, a Mineração de Dados permite análises de perfil ou de predição de resultados, que correlacionam dados socioeconômicos dos alunos, notas da prova, além de informações sobre a estrutura das instituições. Contudo, poucos trabalhos utilizando Mineração de Dados em cursos de computação são identificados na literatura atual.

A partir da edição do ENADE 2017, (Capelari & Schwerz, 2021) identificaram os perfis socio-econômicos dos alunos de graduação em computação, considerando diversos cursos de informática da região Sul do Brasil. Os resultados mostraram que o perfil socioeconômico dos estudantes é formado por alunos brancos, do sexo masculino, com idade entre 18 e 26 anos, renda familiar de 3 a 6 salários mínimos, a maioria dos alunos não tem bolsa de estudos, seus pais cursaram até o ensino médio, concluído em escola pública e estudam no turno noturno de instituições privadas.

Um outro estudo forneceu informações que podem ser úteis para diretores e coordenadores que queiram melhorar a qualidade de seus cursos. Por meio da geração de resultados da Universidade Federal do Pará (UFPA), foi detectado que o desempenho do curso de ciência da computação na parte de componente específico melhorou ao longo dos anos, que o desempenho de formação geral diminuiu e que os alunos da UFPA possuem baixo desempenho nos temas de Inteligência Artificial, Teoria da Computação, Compiladores e Banco de Dados (Cunha, Sales, & Santos, 2021).

No estudo de (Lima, Ambrósio, Oliveira, & Carvalho, 2021), foram identificadas, por meio de análise de conteúdo de dados do Enade, as áreas de conhecimento preditoras do sucesso no resultado do exame. Os dados analisados apontam que os temas mais recorrentes são Algoritmos, Estrutura de Dados e Programação, Linguagens Formais e Automatos, Compiladores e Computabilidade, Lógica Matemática, Matemática Discreta, Estatística e Grafos. Esses temas são congruentes com a matriz curricular do curso de Bacharelado em Ciência da Computação.

Além da escassez de estudos abordando a mineração de dados na base do ENADE, foi detectada uma lacuna na literatura no que tange estudos longitudinais envolvendo alunos do curso de Ciência da Computação, a fim de avaliar o impacto de dados socioeconômicos no desempenho acadêmico, ou seja, não foi encontrado estudo longitudinal que avalie relações entre dados socioeconômicos e o desempenho de estudantes de computação, considerando o escopo nacional, nos anos 2011, 2014 e 2017. Desse modo, há um campo de pesquisa em aberto, no que diz respeito a investigações a partir de informações importantes provenientes da base do ENADE, para compreender melhor o perfil sócio-econômico dos alunos de Ciência da Computação em função de seu desempenho, permitindo a definição de ações preventivas que visem mitigar possíveis

problemas que venham ser detectados.

4 Metodologia

Para alcançar o objetivo Desse trabalho, foi utilizada uma metodologia mista abordando métodos qualitativos e quantitativos, obedecendo os passos metodológicos de acordo com as etapas revisão e identificação, pré-processamento, mineração de dados e apresentação dos resultados, descritas na Figura 2.

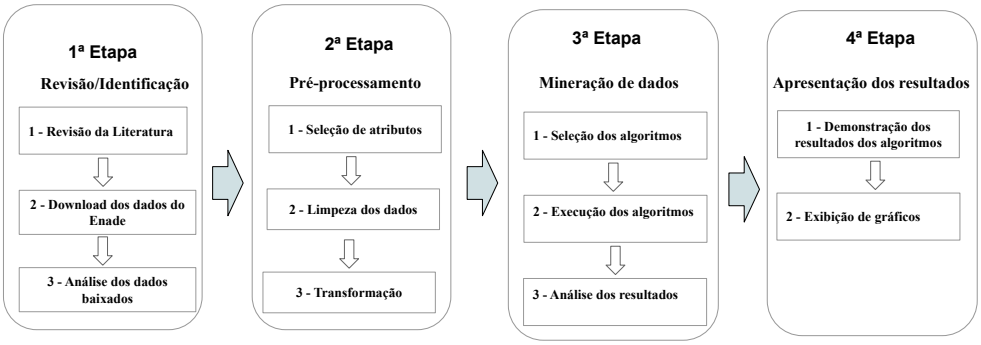

Figura 2: Etapas da metodologia.

4.1 1ª Etapa: Revisão/Identificação

4.1.1 Revisão da Literatura

Essa etapa consistiu em realizar uma busca na literatura para identificar quais trabalhos haviam sido publicados utilizando as bases de dados do INEP e posteriormente, realizar o *download* das provas e, assim, realizar um levantamento dos dados a serem utilizados na pesquisa. Os resultados provenientes da revisão da literatura reforçam a originalidade e importância desse trabalho, uma vez que em sua grande maioria, os trabalhos retornados realizam somente estatística descritiva.

4.1.2 Download dos dados do ENADE

Para realizar a pesquisa, foi feito o *download* dos microdados e dicionários de dados das provas do ENADE fornecidos pelo INEP em seu site oficial[2]. Foram baixadas as informações dos anos 2008, 2011, 2014 e 2017. Os dicionários foram necessários para realizar o estudo da aplicação, tornando possível o entendimento dos atributos existentes. Os microdados foram necessários para realizar as análises das características socioeconômicas dos alunos.

4.1.3 Análise dos dados

Após realizar o download dos dados, uma análise no dicionário de dados disponibilizado juntamente com os arquivos baixados foi realizada. Após fazer um levantamento das informações

[2]Disponível em: http://portal.inep.gov.br/microdados. Acessado em: 01/05/2021.

presentes no dicionário de dados, foi possível compreender os atributos e o domínio de aplicação dos dados. Uma vez analisados os atributos, foi possível mapear as características socioeconômicas dos alunos nos microdados de cada ano do experimento. Esses atributos serviram como base para realizar a pesquisa.

4.2 2ª Etapa: Pré-processamento

Nessa etapa, os dados foram preparados para que a etapa de mineração de dados pudesse ser realizada. Foram adotadas algumas técnicas do pré-processamento de dados tais como: limpeza de dados, seleção de atributos e transformação de dados, conforme (Stephens & Sukumar, 2006).

4.2.1 Seleção de atributos

Para realizar a seleção dos atributos, inicialmente foi feito um levantamento de todos os atributos pertencentes ao questionário dos alunos em todos os anos do experimento. Uma vez realizada uma previa análise nos atributos, utilizou-se uma das abordagens do Estudo Longitudinal Transversal Repetido, que é realizar o estudo de forma a utilizar amostras diferentes em períodos distintos, assim, uma análise em todos os anos do experimento foi realizada e somente os atributos presentes em cada um dos anos abordados no estudo foram selecionados.

Após selecionar os atributos presentes em todos os anos da pesquisa, a ferramenta Weka do inglês *Waikato Environment for Knowledge Analysis* foi utilizada para extrair os atributos relevantes (Hall et al., 2009). Foram utilizados os algoritmos de seleção de atributos presentes na ferramenta Weka:

CfsSubsetEval, GainRatioAttributeEval, ClassifierAttributeEval, OneRAttributeEval, InfoGainAttributeEval, SymmetricalUncertAttributeEval. Para os experimentos foram considerados os valores padrão de parâmetros dos respectivos algoritmos mencionados. Os atributos selecionados foram aqueles retornados por pelo menos 50% dos algoritmos.

Para melhor entendimento, alguns pontos importantes dos algoritmos utilizados são apontados. O *CfsSubsetEval*, seleciona os atributos relevantes, considerando a capacidade preditiva individual de cada recurso. O *InfoGainAttributeEval*, avalia o valor de um atributo medindo a taxa de ganho de informação em relação à classe, enquanto que o *SymmetricalUncertAttributeEval*, avalia o valor de um atributo medindo a incerteza simétrica em relação à classe.

4.2.2 Limpeza dos dados

É possível verificar na Figura 3 que existe grande percentual de dados faltantes no ano de 2008. Por exemplo, Estado Civil tanto no grupo baixo quanto no grupo alto desempenho, apresentam dados faltantes, sendo 30% do grupo baixo e aproximadamente 34% do grupo alto. Outras características também apresentam grande número de dados faltantes, como a renda familiar, escolaridade do pai, escolaridade da mãe entre outras. Devido ao alto índice de dados faltantes no ano de 2008, neste trabalho, as análises realizadas tiveram como fonte de entrada o ano de 2011, desprezando o ano de 2008.

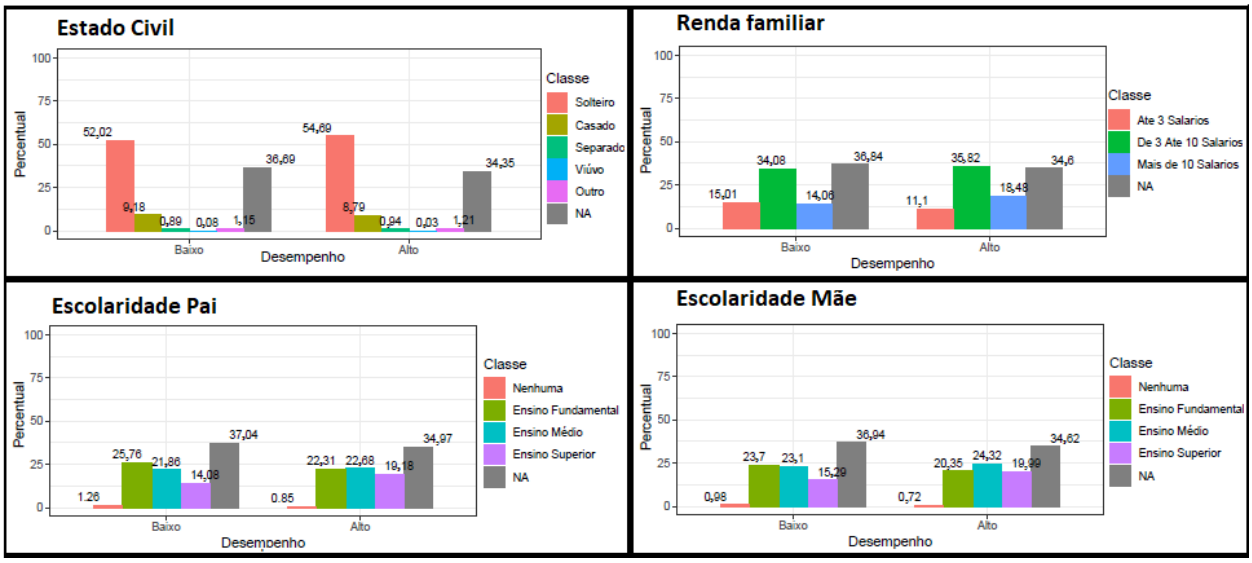

Figura 3: Dados do ano de 2008.

4.2.3 Transformação

Com levantamento realizado observando o dicionário de dados, foi identificado que muitos atributos possuem nomenclatura diferente de um período para o outro. Desse modo, foi realizada uma padronização na nomenclatura dos atributos em todos os períodos, tornando possível realizar o ELTR e aplicar a mineração de dados.

Ainda foi observado que, no dicionário de dados, não existia uma coluna que retratasse a situação do aluno específica de seu desempenho acadêmico. Assim, foi criada coluna para armazenar o desempenho de cada aluno tomando como base a coluna NT_GER [3]. Para a criação dessa coluna, os dados ruidosos presentes na coluna nota bruta foram excluídos e, posteriormente, as notas foram normalizadas.

4.3 3ª Etapa: Mineração de dados

Para realizar a mineração de dados nos dados selecionados para a pesquisa, buscou-se na literatura, um levantamento de possíveis técnicas comumente utilizadas na classificação de dados na literatura.

4.3.1 Seleção dos algoritmos

Como a pesquisa busca identificar se os atributos socioeconômicos impactam no respectivo desempenho acadêmico dos alunos, foram utilizados alguns algoritmos de classificação de dados. Embora fosse possível ter escolhido outros classificadores, os adotados aqui têm sido amplamente utilizados na prática, portanto, são adequados como prova de conceito. Os algoritmos utilizados buscaram classificar os dados em um dos dois possíveis rótulos presentes na coluna desempenho: Baixo/Alto.

[3]Média ponderada da formação geral (25%) e componente específico (75%), Nota de (0 a 100)

4.3.2 Execução dos algoritmos

Foram utilizados alguns dos algoritmos de classificação de dados durante o processo de mineração de dados: Árvore de decisão, *Random Forest* e o SVM (Han, Pei, & Kamber, 2011), (Alpaydin, 2020), (Stephens & Sukumar, 2006).

4.3.3 Análise dos resultados

A acurácia de cada um foi analisada após a realização da execução de cada um dos algoritmos. Essa análise foi necessária para averiguar qual algoritmo melhor classificou os dados para que o resulta-do da análise pudesse ser feita e apresentada na pesquisa.

4.4 4ª Etapa: Apresentação dos resultados

Uma vez executados os algoritmos, os resultados foram apresentados na forma de Tabelas e gráficos.

4.4.1 Demonstração dos resultados dos algoritmos

Foi mostrado, em forma de Tabela, a acurácia de cada um dos algoritmos. Desse modo, foi possível a visualização do algoritmo que foi capaz de melhor classificar os dados.

4.4.2 Exibição de gráficos

Após a obtenção dos resultados, os atributos que melhor representaram a classificação dos dados foram mostrados em forma de gráficos. Todos os experimentos foram realizados utilizando a ferramenta de mineração de dados Weka (Hall et al., 2009) e a linguagem de programação R [4].

5 Resultados

Nesta Seção, são apresentadas os resultados com respeito ao tratamento e análise de dados para efetivação do Estudo Longitudinal Transversal Repetido.

5.1 Entrada de dados

Uma vez obtidos os dados a partir do *download* no site oficial e posteriormente analisar as informações presentes nos dicionários de dados disponibilizados juntamente com os arquivos dos microdados, foi possível se ter entendimento do domínio da aplicação. O passo seguinte consistiu em realizar a importação dos dados na aplicação para realizar as demais etapas do KDD. A relação da quantidade de registros e atributos podem ser vistos na Tabela 1.

[4]Disponível em: https://www.r-project.org/. Acessado em: 01/12/2019.

Tabela 1: Relação de registros e atributos.

Ano	Quantidade de registros	Número de atributos	Questionário do aluno
2011	376.180	115	54
2014	484.720	154	26
2017	537.436	150	26
Total	1.398.336	419	106

5.2 Pré-processamento

Na etapa de pré-processamento, foram aplicadas algumas medidas para tornar possível realizar a etapa de mineração de dados: limpeza dos dados, transformação, Estudo Longitudinal Transversal Repetido e seleção de atributos relevantes.

5.3 Limpeza dos dados

Uma vez selecionados os atributos a serem utilizados na pesquisa, alguns registros foram desconsiderados, permanecendo somente os registros pertencentes ao curso ciência da computação, o aluno esteve presente na realização da prova e a nota bruta contenha informação válida.

Foi aplicada ainda, a remoção dos valores *outliers* presentes na coluna de nota bruta, uma vez que essa serviu de parâmetro para criar uma coluna que passou a conter o desempenho acadêmico dos alunos. Os valores considerados como *outliers* foram aqueles valores que discrepantes em relação as demais informações presentes na coluna analisada.

A Figura 4 apresenta o gráfico dos valores presentes na coluna Nota Bruta. A informação apresentada é um diagrama de caixa ou *box plot*, que é uma ferramenta utilizada para representar a variação de dados observados de uma variável numérica.

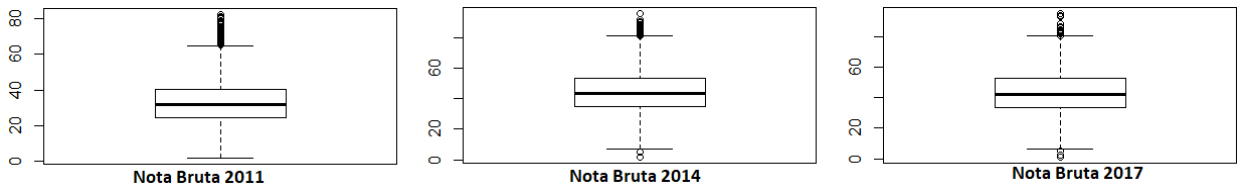

Figura 4: Gráfico Nota Bruta.

A relação dos dados antes e após a aplicação dos filtros pode ser vista na Tabela 2. Antes de realizar a seleção dos registros, tinha-se o total de 1.398.336 registros, após realizar a seleção permaneceram 26.247 registros no experimento.

Tabela 2: Total de Registros.

Ano	Registros antes da filtragem	Registros após a filtragem
2011	376.180	9.580
2014	484.720	8.224
2017	537.436	8.443
Total	1.398.336	26.247

5.4 Análise longitudinal transversal repetido aplicando estudo de coortes

Como pode ser visto na Tabela 1, a quantidade de atributos pertencentes ao questionário do aluno, pode variar de um período para o outro. Para realizar o Estudo Longitudinal Transversal Repetido, uma análise em todos os anos foi realizada e somente os atributos presentes em todos os anos da pesquisa foram selecionados. A padronização da nomenclatura dos atributos está descrita na Tabela 3. Na Tabela 4 é mostrada a legenda de alguns dos novos atributos os quais foram mapeados na Tabela 3. Desse modo, os atributos utilizados como parâmetro na pesquisa, foram os atributos listados na Tabela 5.

5.5 Transformação dos dados

Após analisar os dicionários de dados, alguns ajustes foram necessários para tornar possível realizar a pesquisa.

5.5.1 Padronização de nomenclatura dos Atributos

Alguns atributos continham nomenclaturas e informações distintas de um período para o outro. Desse modo alguns ajustes foram necessários para normalizar as nomenclaturas dos atributos. Por exemplo, a coluna *Escolaridade do Pai* no ano de 2011 era armazenada na coluna *QE_I13*, enquanto que em 2014 e 2017 era armazenada na coluna *QE_I04*. Foi criada uma coluna *Escolaridade_Pai* que passou a conter a informação independente do período. Mais detalhes são apresentados na Tabela 3.

Tabela 3: Padronização da nomenclatura dos atributos.

Atributo	2011	2014	2017
Bolsa	QE_I09	QE_I11	QE_I11
Escolaridade_Mae	QE_I14	QE_I05	QE_I05
Escolaridade_Pai	QE_I13	QE_I04	QE_I04
Leu_Livros	QE_I19	QE_I22	QE_I22
Mora_Sozinho	QE_I04	QE_I07	QE_I07
Renda_Familiar	QE_I05	QE_I08	QE_I08
Situacao_Trabalho	QE_I07	QE_I10	QE_I10
Atributos normalizados pela equipe ENADE			
Atributo	2011	2014	2017
Categoria_Administrativa	CO_CATEGAD	CO_CATEGAD	CO_CATEGAD
Escola_Ensino_Medio	QE_I17	QE_I17	QE_I17
Estado_Civil	QE_I01	QE_I01	QE_I01
Modalidade_Ensino_Medio	QE_I18	QE_I18	QE_I18
Raca	QE_I02	QE_I02	QE_I02
Sexo	TP_SEXO	TP_SEXO	TP_SEXO

Mais informações sobre os campos disponibilizados pelo INEP, podem ser obtidas através do site oficial dos microdados do ENADE.

Tabela 4: Legenda dos atributos.

Atributo	Descrição
Bolsa	Possuí Bolsa de Estudos?
Escolaridade_Mae	Escolaridade da Mãe
Escolaridade_Pai	Escolaridade do Pai
Leu_Livros	Leu livros além da bibliografia?
Mora_Sozinho	Mora sozinho?
Renda_Familiar	Renda Familiar
Situacao_Trabalho	Esta trabalhando?
Categoria_Administrativa	Categoria Administrativa
Escola_Ensino	Tipo de escola Enisno Médio
Estado_Civil	Estado Civil
Modalidade_Ensino_Medio	Modalidade Ensino Médio
Raca	Raça
Sexo	Sexo

5.5.2 Criação de Coluna

Conforme mencionado na Seção 4.2.3, a ausência de um atributo que armazenasse o desempenho acadêmico tornou necessária a criação de uma coluna para comportar o desempenho. A coluna Desempenho foi criada a partir da coluna Nota Bruta removendo os valores *outliers* e realizando a normalização das notas.

A normalização foi feita selecionando a maior nota por ano. Posteriormente, foi calculado 60% dessa nota, para que a mesma se tornasse o limiar de classificação. A exemplo de (Corso & Resende, 2018), que utilizaram a média menor que 60 para classificar o desempenho como ruim, as notas menores ou iguais ao limiar calculado foram consideradas como desempenho "Baixo"e as notas maiores foram consideradas como desempenho "Alto". Por exemplo, no ano de 2017, a maior nota foi 80,5, calculando 60% dessa nota, há o limiar de 48,3. As notas menores ou iguais a 48,3 receberam o rótulo de desempenho "Baixo", enquanto que as notas acima foram rotuladas como desempenho "Alto".

5.6 Seleção dos atributos relevantes

Uma vez realizada a normalização dos atributos, cujo exemplo de padronização pode ser visto na Tabela 3, alguns atributos permaneceram no experimento, enquanto outros não. A metodologia utilizada para realizar a primeira etapa na seleção dos atributos relevantes, pode ser vista nos passos descritos na Seção 4.2.1, posteriormente, somente os atributos selecionados em todos os anos seguiram na pesquisa.

Assim, os atributos *Situação de Trabalho*, *Mora Sozinho*, *Sexo*, *Bolsa* e *Modalidade do Ensino Médio* não permaneceram na pesquisa, pois, embora tenham sido selecionados pelos algoritmos em alguns anos, não foram selecionados em todos os anos do experimento. Por outro lado, os atributos *Escolaridade do Pai*, *Escolaridade da Mãe*, *Categoria Administrativa*, *Turno da Graduação*, *Bolsa*, *Renda Familiar*, *Estado Civil* e *Situação de Trabalho* foram selecionados pelos algoritmos e também em todos os anos do experimento, como pode ser visto na Tabela 5.

Tabela 5: Seleção de atributos.

Ano	Atributos
2011	Estado Civil, Sexo, Escolaridade do Pai Escolaridade da Mãe, Renda Familiar, Situação de Trabalho, Categoria Administrativa, Turno da Graduação e Mora Sozinho
2014	Estado Civil, Sexo, Escolaridade do Pai, Escolaridade da Mãe, Renda Familiar, Situação de Trabalho, Categoria Administrativa, Turno da Graduação e Bolsa.
2017	Estado Civil, Escolaridade do Pai, Escolaridade da Mãe, Renda Familiar, Categoria Administrativa, Turno da Graduação, Bolsa, Modalidade do Ensino Médio, Situação de Trabalho
Atributos selecionados	Escolaridade do Pai, Escolaridade da Mãe Categoria Administrativa, Turno da Graduação, Bolsa, Renda Familiar, Estado Civil, Situação de Trabalho

5.7 Mineração de dados

Na etapa de mineração de dados foram executados alguns algoritmos de classificação nos dados selecionados. A execução foi realizada com o intuito de averiguar qual dos algoritmos era capaz de melhor classificar os dados em um dos dois rótulos de classe gerados no campo desempenho: “Baixo” ou “Alto”. A relação das acurácias obtidas é apresentada na Tabela 6.

Tabela 6: Acurácia por algoritmo.

Ano	J48	Random Forest	SVM
2011	73,26%	70.65%	72,36%
2014	71,13%	70,49%	70,49%
2017	71,14%	69,43%	67,57%
Todos	71,54%	70,91%	70,72%

5.8 Pós-processamento

Após a execução dos algoritmos, os resultados demonstram que o algoritmo que melhor classifica os dados é a árvore de decisão, neste trabalho retratada com o algoritmo J48, pois, em todos os anos, este apresentou a melhor acurácia conforme apresentação da Tabela 6.

Em se tratando da análise dos atributos, Carvalho, Faceli, Lorena, and Gama (2011) afirmam que, na árvore de decisão, quando os atributos não são listados é um indício de que os atributos são irrelevantes. Por outro lado, os atributos que compõe a J48 são selecionados conforme o cálculo de ganho de informação.

O cálculo de ganho de informação é realizado para cada um dos atributos presentes no conjunto de dados, onde o atributo com maior ganho de informação é utilizado para dividir os

dados, sendo o atributo que melhor particiona ou classifica os dados (Carvalho et al., 2011). Este atributo é utilizado como o primeiro atributo da árvore de decisão, tornado-se o nó raiz da árvore.

Desse ponto de vista, o atributo que melhor classifica os dados utilizados na pesquisa é a categoria administrativa, uma vez que este atributo foi a raiz de todas as árvores geradas.

Conforme pode ser visto na distribuição gráfica apresentada na Seção 5.9, a categoria administrativa teve grande impacto no desempenho acadêmico. Os demais atributos também são apresenta-dos nessa seção.

5.9 Visualização gráfica dos resultados por ano

Nesta Seção, são demonstrados os resultados em forma de gráficos, bem como as árvores de decisão resultantes do algoritmo J48 aplicado aos dados, possibilitando melhor entendimento das informações obtidas.

5.9.1 Visualização gráfica dos resultados do ano 2011

Conforme pode ser visualizado na árvore de decisão gerada no ano de 2011 da Figura 5, a categoria administrativa foi o atributo com maior relevância, seguido pelo atributo turno em que o aluno cursou sua graduação. A seguir, na Figura 5 é mostrada uma adaptação de parte da árvore de decisão gerada pela ferramenta Weka.

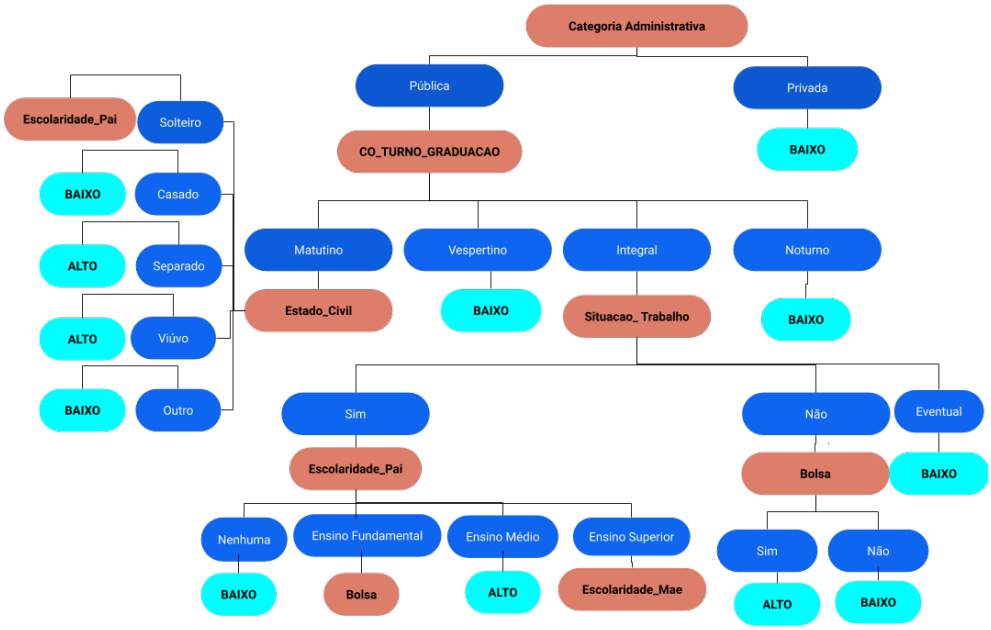

Figura 5: Árvore de decisão Ano 2011.

A categoria administrativa no ano de 2011 teve grande influência no desempenho. Na árvore de decisão na Figura 5, quando a instituição é privada, o desempenho foi classificado como baixo, assim como, pode ser observado na Figura 6, que na distribuição gráfica, aproximadamente 77%

dos desempenhos acadêmicos do grupo de baixo estão nas instituições privadas.

Seguindo a construção da árvore de decisão, o segundo atributo elencado na árvore como mais relevante foi o turno em que o aluno cursou. Nas instituições públicas, quando o turno da graduação é matutino e o estado civil é casado, o desempenho foi classificado como baixo, assim como quando o turno é vespertino, ou noturno, o desempenho também tendeu a ser baixo. Com respeito ao turno integral, quando o aluno declarou não estar trabalhando e afirmou ter bolsa de estudos, o desempenho tendeu a ser alto e, nessas mesmas condições quando o aluno declarou não ter bolsa, teve o desempenho baixo.

Os gráficos de distribuição dos atributos categoria administrativa e do turno de graduação podem ser visualizados na Figura 6 e na Figura 7, respectivamente.

A Figura 6 demonstra que aproximadamente 77% dos alunos de instituições privadas no ano de 2011, tiveram o desempenho baixo, levando em conta a análise da categoria administrativa isoladamente.

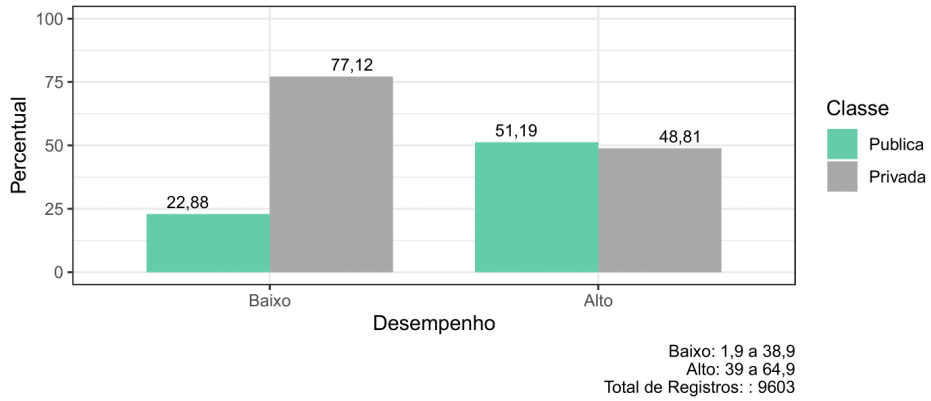

Figura 6: Categoria Administrativa Ano 2011.

A Figura 7 demonstra que aproximadamente 76% dos alunos que cursaram o curso de ciência da computação no turno noturno, no ano de 2011, tiveram o desempenho baixo.

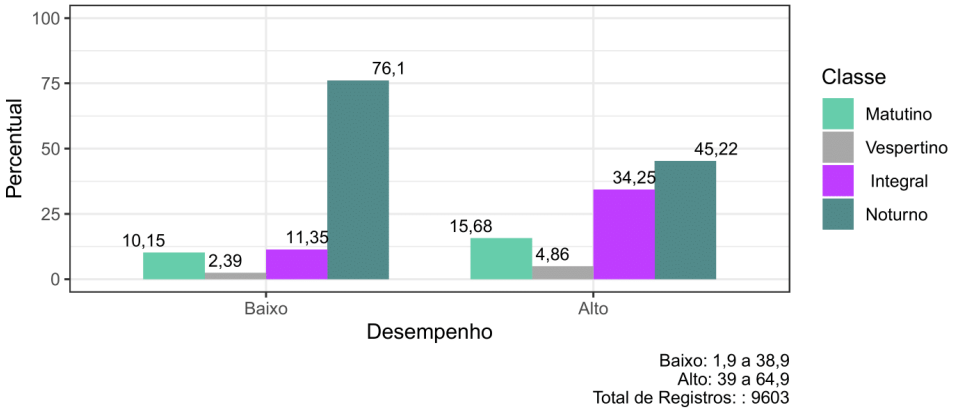

Figura 7: Turno Graduação Ano 2011.

5.9.2 Visualização gráfica dos resultados ano 2014

Conforme pode ser visualizada na árvore de decisão gerada no ano de 2014, assim como no ano de 2011, a categoria administrativa foi o atributo com maior relevância, seguido pelo atributo turno em que o aluno cursou a graduação. A seguir, é mostrada uma adaptação de parte da árvore de decisão gerada pela ferramenta Weka representada na Figura 8.

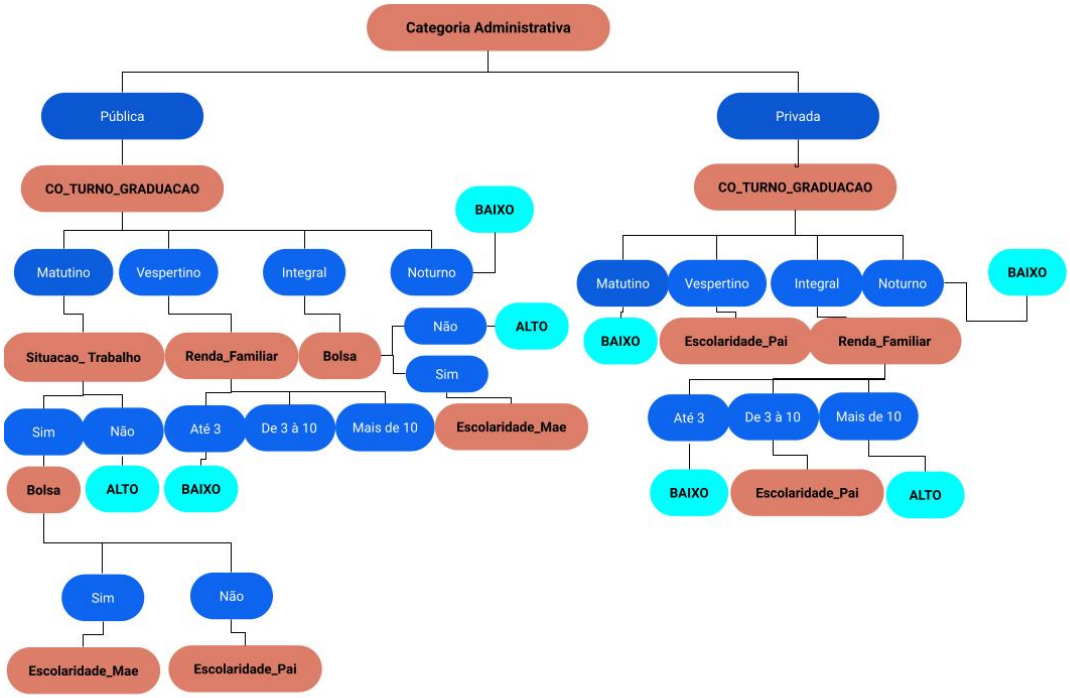

Figura 8: Árvore de decisão Ano 2014.

Assim como no ano de 2011, o ano de 2014 elencou como segundo atributo mais relevante o turno em que o discente cursou a graduação, assim, algumas informações podem ser observadas.

Quando a instituição é pública, o turno é matutino e o aluno declarou não estar trabalhando, o desempenho foi classificado como alto, assim como, para o turno vespertino, quando a renda familiar é de até 3 salários, o desempenho foi classificado como baixo. Em se tratando do turno integral, quando o aluno não teve bolsa, o desempenho foi classificado como alto.

Nas instituições privadas, em se tratando dos turnos matutino ou vespertino, o desempenho foi classificado como baixo. Considerando-se o turno integral, o desempenho foi classificado dependendo da renda familiar, onde até 3 salários foi classificado como baixo e mais de 10 salários como alto.

Os gráficos de distribuição dos atributos categoria administrativa, turno de graduação e renda familiar podem ser visualizados na Figura 9 e na Figura 10, respectivamente.

A Figura 9 demonstra que aproximadamente 75% dos alunos de instituições privadas no ano de 2014, tiveram o desempenho baixo.

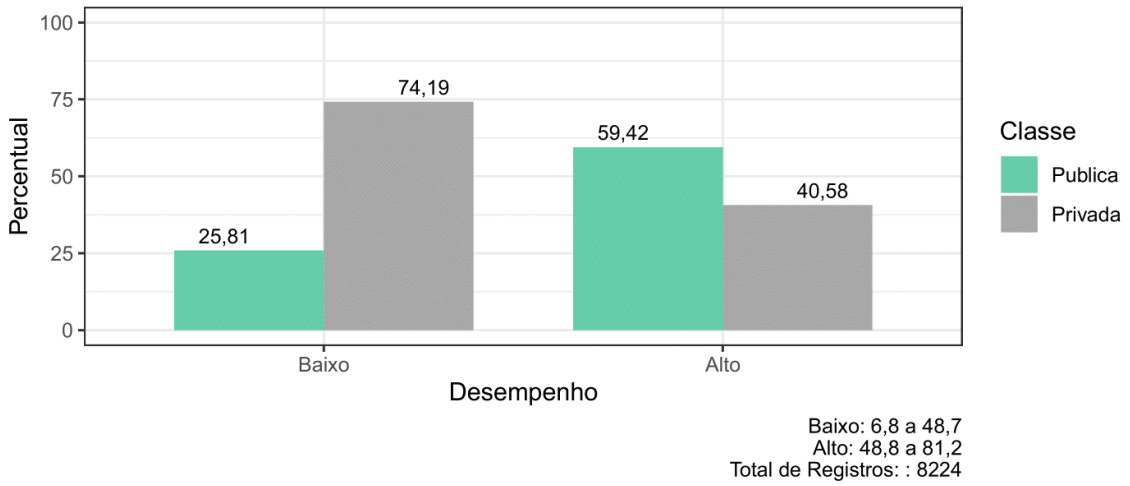

Figura 9: Categoria Administrativa ano de 2014

A Figura 10 apresenta que aproximadamente 62% dos alunos que cursavam o curso de ciência da computação no turno noturno tiveram desempenho baixo.

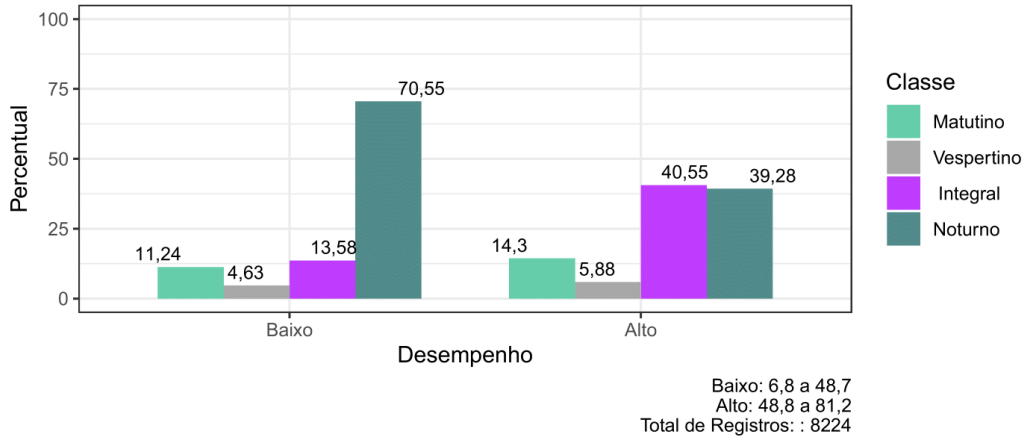

Figura 10: Turno Graduação ano de 2014.

5.9.3 Visualização gráfica dos resultados Ano 2017

Como pode ser visualizado na árvore de decisão gerada no ano de 2017 na Figura 11, a categoria administrativa foi o atributo com maior relevância, seguido pelo atributo renda familiar.

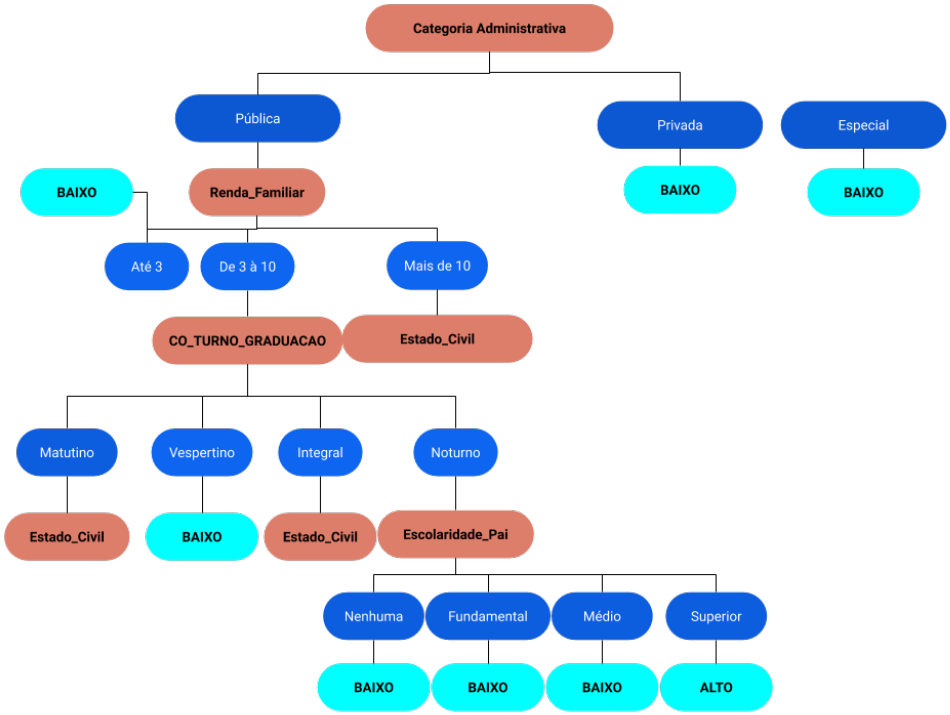

Figura 11: Árvore de decisão Ano 2017.

Com base nos resultados apresentados pela árvore de decisão, quando a categoria administrativa é privada ou especial, o desempenho foi classificado como baixo. Porém, quando a categoria administrativa é pública é observado o atributo renda familiar, onde com a renda familiar em até 3 salários, o desempenho foi classificado como baixo e quando a renda familiar é de 3 a 10 salários, a classificação do desempenho depende do turno em que o aluno cursou a graduação.

Nas Figura 12 e Figura 13, são apresentados os gráficos contemplando a distribuição do desempenho no ano de 2017 por categoria administrativa e turno de graduação respectivamente. Analisando os dados do gráfico na Figura 12, é possível notar, que em se tratando de instituições privadas, aproximadamente 65% dos estudantes tiveram desempenho baixo, enquanto que aproximadamente 68% dos estudantes tiveram desempenho alto em se tratando de instituições públicas.

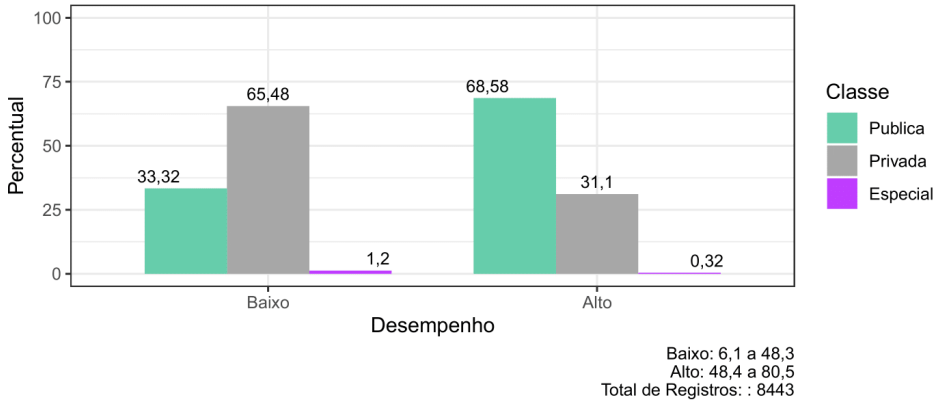

Figura 12: Categoria Administrativa Ano 2017.

A Figura 13, demonstra que aproximadamente 62% dos estudantes que cursaram o turno noturno, tiveram o desempenho baixo.

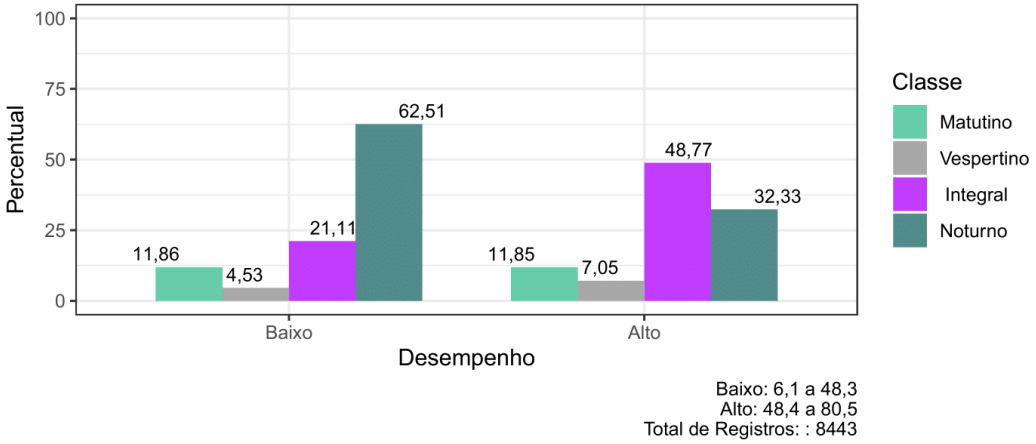

Figura 13: Turno Graduação Ano 2017.

6 Conclusões

Os resultados obtidos nos experimentos mostrados neste trabalho respondem à questão de pesquisa "É possível extrair relações entre as características socioeconômicas dos alunos de ciência da computação e o seu desempenho acadêmico, a partir da base do ENADE, nos anos 2008, 2011, 2014 e 2017?", demonstrando que algumas características socioeconômicas dos alunos são relevantes no diz respeito ao desempenho acadêmico, no entanto, não isoladamente, uma vez que estes dependem da categoria administrativa da instituição em todos os anos do estudo e nos anos 2011 e 2014, dependem ainda do turno em que o discente cursou a universidade.

Como pode ser visto nas árvores de decisão geradas em cada um dos anos, mostrados nas Figuras 5, 8 e 11, as características socioeconômicas impactaram de formas distintas de um

período para o outro. Por exemplo, no ano de 2011 e 2017 quando a instituição é privada, o desempenho foi classificado como baixo sem ser necessário analisar outros atributos, porém, no ano de 2014, em se tratando das instituições privadas, foi necessário analisar o turno de graduação em que o discente cursou seguidos pela situação de trabalho e escolaridade do pai.

Outra situação que comprova a diferença da classificação de um período para o outro, foi para as instituições públicas, que nos anos 2011 e 2014, o desempenho dependeu do turno de graduação em que o discente cursou, enquanto que no ano de 2017, o desempenho dependeu da renda familiar.

O Estudo Longitudinal Transversal Repetido evidenciou que mesmo de formas distintas, de um período para o outro, as informações socioeconômicas em conjunto com a categoria administrativa e o turno em que o discente cursou impactam no desempenho acadêmico. Estudo Longitudinal Transversal Repetido demonstrou ainda que, ao analisar a distribuição de desempenho baixo e alto em relação ao conjunto de dados presentes no estudo, é notável que o desempenho baixo foi maior em todos os períodos. É possível verificar que, entre todos os períodos analisados, o público de alunos do sexo masculino é consideravelmente maior do que o público feminino.

Assim, com base na análise exploratória realizada nas bases do INEP seguindo a metodologia aplicada, pode-se concluir que algumas informações socioeconômicas em conjunto com a categoria administrativa e o turno em que o discente cursa o curso de ciências da computação, impactam sim no desempenho acadêmico.

Os resultados desta pesquisa formam uma base empírica que ressalta desigualdades sociais em cursos de computação no Brasil, revelando diferentes desempenhos de estudantes de acordo com diferentes grupos sociais, a partir, principalmente, de indicadores de escolaridade da mãe e de dependência administrativa da instituição. Desse modo, este estudo corrobora a relação existente entre a desigualdade social e desigualdade educacional no Brasil, reafirmando que as origens sociais e heranças familiares afetam os destinos educacionais dos alunos de computação.

A partir dos resultados apresentados sobre escolaridade das mães, advoga-se a elaboração e aplicação de políticas que promovam redução das desigualdades econômicas, assim como políticas e ações educacionais especificamente voltadas à aquisição de "capital cultural" ao longo da trajetória acadêmica do estudante, não somente para alunos do ensino superior, mas desde a educação básica, visando mitigar a falta oportunidades para o acesso à universidade e de adultos pouco escolarizados. Pois, em ambientes familiares onde os pais tiveram oportunidades escolares, constrói-se um ambiente onde existe o gosto pelos estudos, influenciando os filhos positivamente ao desenvolvimento das potencialidades educacionais.

Com base na dependência administrativa da instituição, os resultados desta pesquisa apontam que o desempenho depende do ambiente em que o aluno realizou sua graduação. Embora o Ensino Superior esteja mais acessível aos estratos sociais mais baixos, isto ocorre pelo aumento da quantidade de instituições de menor prestígio social. Assim, há uma demanda para o desenvolvimento de estratégias, sejam elas provenientes de políticas públicas ou oriundas de instituições de ensino, que consigam melhorar o ambiente universitário para os estudantes, promovendo a formação de profissionais mais qualificados.

6.1 Ameaças à Validade

Nos experimentos realizados neste trabalho foi considerado que a distribuição de dados era uniforme entre os anos estudados. Entretanto, é importante frisar que os anos considerados não obedecem uma distribuição uniforme, ou seja, existe variação no números de instâncias de um período para o outro, o que poderia ocasionar um desbalanceamento entre as classes, por exemplo. Outra decisão que pode ter impacto nos resultados é o fato de termos desconsiderado o ano de 2008 (devido à grande quantidade de dados faltantes).

Não menos importante que as questões levantadas anteriormente é a suposição feita em relação ao limiar de classificação – em que a exemplo de (Corso & Resende, 2018), utilizamos o valor da média menor que 60 para classificar o desempenho como ruim (as notas menores ou iguais ao limiar calculado foram consideradas como desempenho "Baixo" e as notas maiores foram consideradas como desempenho "Alto"). Faz-se importante posicionar esse aspecto do trabalho como uma ameaça à validade, tendo em vista que ocasionalmente outros limiares podem ser adotados.

6.2 Limitações e Trabalhos Futuros

A pesquisa foi realizada levando em conta somente o curso de ciência da computação. Para selecionar as variáveis e executar os algoritmos e também inferir árvores de decisão, foi necessário o intermédio da ferramenta Weka.

Foi identificado que nos anos de 2005 e 2008, existem diversos dados faltantes, e por esse motivo, esses anos não foram considerados na pesquisa. Outro aspecto identificado foi que, existem mais informações do desempenho baixo do que no desempenho alto, ocasionando um desbalanceamento de classes que repercutiu no desempenho dos algoritmos de classificação.

Para tratar essas limitações e dar continuidade a esta pesquisa são sugeridos como trabalhos futuros: (i) Possibilitar a seleção do curso a ser analisado ao longo do tempo dentro da ferramenta; (ii) Criar um processo para automatizar tanto a seleção de variáveis quanto a execução dos algoritmos; (iii) Aplicação de algoritmos para tratamento de dados faltantes; (iv) Para este primeiro estudo, o problema de inferência do desempenho escolar foi tratado como um problema de classificação. Outra vertente Desse mesmo trabalho poderia explorar algoritmos de regressão, uma vertente mais realista e conveniente com o modelo atual de análise de dados educacionais.

References

Alpaydin, E. (2020). *Introduction to machine learning*. Londres, Inglaterra: MIT Press. [GS Search]

Banni, M., Oliveira, M., & Bernardini, F. (2021). Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade* (pp. 57–66). Porto Alegre, RS, Brasil: SBC. doi: 10.5753/wics.2021.15964 [GS Search]

Capelari, L., & Schwerz, A. (2021). O Perfil Socioeconmico dos Concluintes de Computação do Sul do Brasil. In *Anais do computer on the beach* (p. 133-140). Itajai, SC, Brasil. doi:

10.14210/cotb.v12.p133-140 [GS Search]

Carmo, R. V., Heckler, W., F., & Carvalho, J. V. (2020). Uma Análise do Desempenho dos Estudantes do Rio Grande do Sul no ENEM 2019. *Revista Novas Tecnologias na Educação*, *18*(2), 378–387. doi: 10.22456/1679-1916.110257 [GS Search]

Carvalho, A., Faceli, K., Lorena, A., & Gama, J. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Rio de Janeiro, RJ, Brasil: LTC.

Corso, G. T., & Resende, M. D. G. S. (2018). *Mineração de dados aplicada ao enade* [Trabalho de Conclusão de Curso (Licenciatura em Ciência da Computação) - Universidade de Brasília]. Recuperado de: https://bdm.unb.br/handle/10483/25017.

Cunha, R., Sales, C., & Santos, R. (2021). Análise Automática com os Microdados do ENADE para Melhoria do Ensino dos Cursos de Ciência da Computação. In *Anais do computer on the beach* (p. 208-217). doi: 10.5753/wei.2021.15912 [GS Search]

da Fonseca, A. A. N. (2016). Mineração em bases de dados do INEP: Uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, *32*(1). doi: 10.1590/0102-4698140742 [GS Search]

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37. doi: 10.1609/aimag.v17i3.1230 [GS Search]

Fontelles, M. J., Simões, M. G., Farias, S. H., & Fontelles, R. G. S. (2009). Metodologia da pesquisa científica: diretrizes para a elaboração de um protocolo de pesquisa. *Revista Paraense de Medicina*, *23*(3), 1–8. [GS Search]

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18. doi: 10.1145/1656274.1656278 [GS Search]

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. [GS Search]

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley & Sons. [GS Search]

Hoed, R., & Saraiva, P. F. (2019). Desempenho das Instituições de Ensino Brasileiras no ENEM: uma abordagem Usando Mineracao de dados. In J. Sánchez (Ed.), *Nuevas ideas en informática educativa* (Vol. 15, pp. 106–113).

Lima, P. (2018). Análise das Provas do Enade Agrupadas por Tema: Um Estudo de Caso para Estudantes de Ciência da Computação. *Dissertação (Mestrado) - Universidade Federal de Goiás*. [GS Search]

Lima, P., Ambrósio, A. P. L., Oliveira, J., & Carvalho, C. (2021). Análise de conteúdo das provas do Enade para os alunos do curso de Bacharelado em Ciência da Computação. *Revista Brasileira de Informática na Educação*, *29*, 385–413. doi: 10.5753/rbie.2021.29.0.385 [GS Search]

Medeiros, G. (2014). A valorização do professor do ensino médio em Santa Catarina e Minas Gerais: limites e possibilidades. *Dissertação (Mestrado) - Universidade do Sul de Santa Catarina*. [GS Search]

Melguizo, T. (2016). Toward a set of measures of student learning outcomes in higher education: evidence from brazil. *Higher Education (00181560)*, *72*(3), 381–402. doi: 10.1007/s10734-015-9963-x [GS Search]

Nóbrega, J. M. (2016). Educação Superior no Brasil e avaliação da qualidade do ensino-o ENADE e sua contribução: O que os indicadores revelam. *Dissertação (Mestrado) - Universidade*

Lusófona de Humanidades e Tecnologias Faculdade de Ciências Sociais,Educação e Administração Instituto de Educação.

Oliveira, A., & Silva, I. R. (2017). Social Inclusion Policies in Brazilian Higher Education: A study on the socioeconomic profile of students in the years 2010-2012. *Educação em Revista*, *33*. doi: 10.1590/0102-4698153900 [GS Search]

Picanço, F. (2016). Juventude e acesso ao ensino superior no brasil: Onde está o alvo das políticas de ação afirmativa. *Latin American Research Review*, *51*(1), 109–131. doi: 10.1353/lar.2016.0001 [GS Search]

Ruspini, E. (2003). *An introduction to longitudinal research*. Londres, Inglaterra: Routledge. [GS Search]

Silva, L. A., Morino, A. H., & Sato, T. M. C. (2014). Prática de mineraçao de dados no exame nacional do ensino médio. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 3, p. 651). doi: 10.5753/cbie.wcbie.2014.651 [GS Search]

Stephens, C., & Sukumar, R. (2006). Introduction to Data Mining. *Handbook of Marketing Research, Thousand Oaks: Sage*, 455–485. doi: 10.4135/9781412973380.n22 [GS Search]

Vieira, M. F. (2012). Prevalência de retenção escolar e fatores associados em adolescentes da coorte de nascimentos de 1993 em Pelotas, Brasil. *Revista Panamericana de Salud Publica*, *31*(4), 303–310. doi: 10.1590/S1020-49892012000400006 [GS Search]